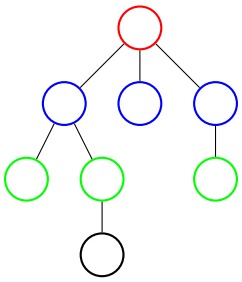
Stochastic Processes



Mark Huber

Fletcher Jones Foundation Associate Professor of Mathematics and Statistics and George R. Roberts Fellow

mhuber@cmc.edu

Contents

Preface

Mathematics is about patterns of relationships between abstract objects. Something that is abstract could be something as simple as the number 2, or as complex as Brownian motion. Roughly speaking, mathematics is about understanding the permissible ways that you can transform problems and ideas to make them easier to work with. One way to think about it is that there are four levels of mathematics.

1: Techniques.

2: Ideas and concepts.

3: Rigorous proof.

4: Automatical proof verification.

The first level is all about using theorems and facts to solve real problems. For instance, in high school you learned an algorithm for how to multiply two numbers written in decimal format. At the second level, we try to understand mathematical facts through a different prism. For instance, why does a times b equal b times a for numbers a and b? Well, a times b is the area of a rectangle that has length a on the horizontal side and b on the vertical side. Whereas b times a is the area of the rectangle with length b on the vertical side and a on the horizontal side. But we got that rectangle just by rotating the first rectangle 90 degrees, which does not change the area.

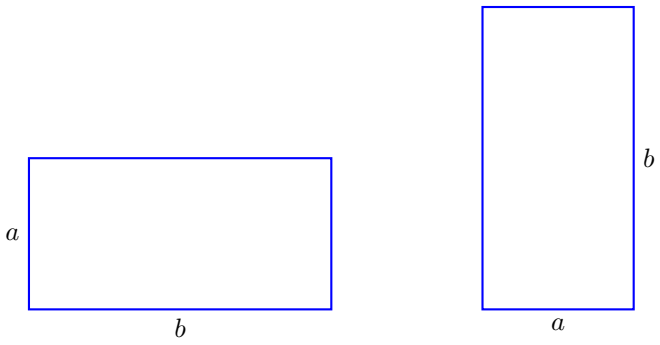


Figure 1: $a \cdot b = b \cdot a$, a geometric perspective.

At the third level, we have the notion of rigorous proof. At this level it is important to define exactly what a particular mathematical object is, so that we can derive true facts that follow logically from the definitions. For instance, for defining $a \cdot b$ where a and b are positive integers, we could build a Turing machine that when fed a tape of a and b in unary, returns $a \cdot b$. This would give a precise definition, and so then it would be possible to logically derive (prove) the result that $a \cdot b = b \cdot a$. This is the level at which most mathematics is done in journals.

At the fourth level, everything has entirely been reduced to symbols, and the permissible steps have been written out precisely enough that it is possible for a computer to check the proof. Very few mathematicians work at this level, but it is the most precise level and the least likely to contain errors.

About this course This text contains a complete semester course in stochastic processes. It starts at the beginning, developing probability from the ground up using set-theory. There are a lot of definitions in the text, covering all the major stochatic processes. These include Markov chains, martingales, Brownian Motion, Harris chains, branching processes, point processes, and drift-diffusions.

There are eight major theorems in this course, most of which are proved in the text. We will begin by laying out the basic definitions and mathematical objects that we will be studying this semester.

1 Introduction to Stochastic Processes

Question of the Day What are stochastic processes?

Random variables A random variable is a variable where you only have partial information about the true value. For instance, the total number of people alive in the continential United States at this moment is a random variable. It is some specific integer value, but you do not have total information about it.

This lack of information can come about because it is simply to difficult to keep track (such as the continental US population example) or because the event has not happened yet! The winner of the next World Series is a random variable, for instance.

It is important to understand that just because we do not exactly know the value, does not mean that we are completely in the dark. For the US population example, I could assign probabilities to the nonnegative integers that represent my lack of knowledge. If I plan to roll two fair six sided dice, then the probability that the sum is 7 is 1/6 whereas the probability the sum will be 2 is only 1/36. Probability is the method we have for codifying partial information.

Stochastic Processes The word stochastic just means random. It comes from the Greek word meaning to aim at a mark and was an archery term. (What could be more random than where an arrow strikes!) For instance, stock prices, location of lightning strikes, your grades at the end of this semester, all of these are random.

Your grade in this class is a single random variable. The grades of you and your classmates is a collection of random variables. Any collection of random variables is a *stochastic process*.

1.1 Measure spaces

In mathematics, a *space* is a set with some additional properties or structure.

- Start with an *outcome space* or *sample space* Ω that represents the possible outcomes of an experiment.
- Now consider a collection of subsets of Ω, call it $\mathcal{F}$. For each $A \in \mathcal{F}$, we can tell if the outcome is in A, or not in A.
- First example: $\Omega = \{1, 2, 3, 4\}$.
  - Then $\mathcal{F}$ could be $\{A_1, A_2, A_3, A_4\} = \{\emptyset, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$.
  - That means for $\omega \in \Omega$, we can answer yes or no to the question: is $\omega \in A_i$ for all i.
- Second example: $\Omega = \mathbb{R}$
  - For all $a \in \mathbb{R}$, want set $(-\infty, a]$ to be in $\mathcal{F}$.
  - That way, can always answer question: is $\omega \leq a$ (equivalently, is $\omega \in (-\infty, a]$) for all real #'s a.
  - The smallest $\mathcal{F}$ that contains all intervals of the form $(-\infty, a]$ is called the **Borel sets**.
- Note: if we can answer "is $\omega \in A$" yes or no, then we can certainly answer "is $\omega \in A^C$" yes or no. So
$$A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}.$$
- Trickier, if we can answer "is $\omega \in A_i$" for all $i = 1, 2, 3, \ldots$, then also assume we can answer "is $\omega \in \cup A_i$".

Definition 1
A nonempty collection of subsets of Ω is a **σ-algebra** or **σ-field** when

1: $(\forall A \in \mathcal{F})(A^{\mathcal{C}} \in \mathcal{F})$.

2: $(\forall A_1, A_2, \ldots \in \mathcal{F})(\cup_{i=1}^{\infty} A_i \in \mathcal{F})$.

Call the pair $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is a σ-algebra for Ω a **measurable space**.

1.2 Probability spaces

- Start with a measurable space $(\Omega, \mathcal{F})$.
- Then $\mathbb{P}$ is a function that takes a measurable set $A \in \mathcal{F}$, and assigns a probability that the outcome falls in A.

Definition 2
A map $\mathbb{P} : \mathcal{F} \to [0,1]$ where $\mathcal{F}$ is a σ-algebra is a **probability measure** or **distribution** if

1: $\mathbb{P}(\omega \in \emptyset) = 0$ and $\mathbb{P}(\omega \in \Omega) = 1$.

2: For $A_1, A_2, A_3, \ldots$ disjoint events (so $(\forall i \neq j)(A_i A_j = \emptyset)$)

$$\mathbb{P}\left(\omega \in \cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(\omega \in A_i).$$

- Call $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space* or *probability triple.*
- Call elements of $\mathcal{F}$ (subsets of Ω that we can assign probabilities to) *measurable events.*
- Notation: $\mathbb{P}(A) = \mathbb{P}(\omega \in A)$.

Example

- Suppose $\Omega = \{1,2,3,4\}$ and $\mathcal{F} = \{\emptyset, \Omega, \{1,2,3\}, \{4\}\}$.
- Then
$$\mathbb{P}_1(\{1,2,3\}) = 1/2, \ \mathbb{P}_1(\{4\}) = 1/2,$$
is one probability measure.
- $$\mathbb{P}_2(\{1,2,3\}) = 3/4, \ \mathbb{P}_1(\{4\}) = 1/4,$$
is another.
- For both $\mathbb{P}_1$ and $\mathbb{P}_2$, must have
$$\mathbb{P}_i(\emptyset) = 0, \quad \mathbb{P}_i(\Omega) = 1.$$

Borel sets

- When $\Omega = \mathbb{R}$, the most common σ-algebra is called the Borel sets. The Borel sets contains all closed, open, half-open intervals, and countable unions and complements of these intervals, and many other sets besides so that they form a σ-algebra.

1.3 Random variables

Definition 3
A **random variable** is a function from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\Omega', \mathcal{F}')$ so that for all $A' \in \mathcal{F}'$, and $A = \{\omega : X(\omega) \in A'\}$, A is in $\mathcal{F}$. This creates a probability distribution on Ω' where for $A' \in F'$,

$$\mathbb{P}_X(A') = \mathbb{P}(X(\omega) \in A').$$

Call $\mathbb{P}_X$ the **distribution** of X.

Example (continued)

- Let $\Omega' = \{0, 1\}$, $\mathcal{F}' = \{\emptyset, \Omega, \{0\}, \{1\}\}$.
- Let $X(\omega) = \mathbb{1}(\omega = 4)$.
- Recall the indicator function of a true or false expression is

$$\mathbb{1}(\text{expression}) = \begin{cases} 1 & \text{if the expression is true} \\ 0 & \text{if the expression is false} \end{cases}.$$

- So $X(1) = X(2) = X(3) = 0$, $X(4) = 1$, and

$$\mathbb{P}_X(1) = \mathbb{P}_1(X = 1) = \mathbb{P}_1(\omega = 4) = 1/2.$$

- Note: $X = \mathbb{1}(\omega = 2)$ is not a random variable! Because to determine if $X = 1$, need to know if $\omega = 2$, but only know if $\omega \in \{1, 2, 3\}$ or in $\{4\}$!

Discrete and continuous A measure is like a probability measure but does not require that the measure be between 0 and 1.

Definition 4
A **measure** μ for a σ-algebra $\mathcal{F}$ satisfies

1: $(\forall A \in \mathcal{F})(\mu(A) \geq 0$.

2: $\mu(\emptyset) = 0$.

3: For $A_1, A_2, A_3, \ldots$ disjoint events (so $(\forall i \neq j)(A_i A_j = \emptyset)$)

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

If the measure satisfies

4: $\mu(\Omega) = 1$

then the measure is a **probability measure** or **distribution**.

The two most important measures are counting measure and Lebesgue mesaure.

- Counting measure just counts the number of elements of a set:

$$\text{count}(\{1, 2, 4\}) = 3, \ \text{count}(\emptyset) = 0, \quad \text{count}(\{1, 2, 3, \ldots\}) = \infty.$$

- Lebesgue measure is 1) length in one dimension, 2) area in two dimensions, 3) volume in three dimensions, 4) hypervolume in higher dimensions (we will not give a formal definition here, since it is a fairly complex idea.)

$$\text{Lebesgue}([3, 8]) = 8 - 3 = 5, \ \text{Lebesgue(unit circle)} = \pi.$$

- For integration, let $\mu_1 =$ counting measure, $\mu_2 =$ Lebesgue measure,

$$\int_{x \in \{1,2,3\}} f(x) \, d\mu_1 = f(1) + f(2) + f(3)$$

$$\int_{x \in A} f(x) \, d\mu_1 = \sum_{a \in A} f(a)$$

$$\int_{x \in [a,b]} f(x) \, d\mu_2 = \text{same as Riemann integral}$$

Definition 5
Say that X has density $f(x)$ with respect to μ if for all $A \in \mathcal{F}$

$$\mathbb{P}(X \in A) = \int_{x \in A} f(x)\, d\mu.$$

Definition 6
X is a **discrete** real valued random variable it has a density with respect to counting measure. Call the density in this case the **probability mass function**.

Definition 7
X is a **continuous** real valued random variable if it has a density with respect to Lebesgue measure.

Any time you have more than one random variable is a stochastic process. For instance, you have a herd of sheep, a box of crayons, a murder of crows, and a stochastic process of random variables.

Definition 8
Any collection of random variables $\{X_t\}$ is called a **stochastic process**.

There are many variations on what the collection could be like. For example:

- t could be discrete, for instance $t \in \{0, 1, 2, \ldots\}$ and S_t is the stock price at the end of the day.
- t could be continuous, for instance $t \in [0, \infty)$ and S_t is the stock price at any time after the present.
- Spatial: The set of points in a city where a disease outbreak has occurred is a stochastic process.

2 Expectation

Question of the Day Suppose $X \sim \mathsf{Exp}(2)$. What is $\mathbb{E}[X]$?

2.1 Informal definition of expectation

The expected value of a random variable is also know as the average, the mean, or the expectation.

Fact 1
For a random variable X with density f_X with respect to μ:

$$\mathbb{E}[X] = \int_{x\in\Omega} s f_X(s)\, d\mu.$$

For a (measurable) function g,

$$\mathbb{E}[g(X)] = \int_{x\in\Omega} g(s) f_X(s)\, d\mu.$$

An important rule to remember is: always apply g to the dummy variable inside the integral rather than to the density.

- Discrete random variables: when $\mathbb{P}(X \in \{x_0, x_1, x_2, \ldots\}) = 1$:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} x_i \mathbb{P}(X = x_i) = \sum_{x:\mathbb{P}(X=x)>0} x\mathbb{P}(X = x).$$

- For example: $\mathbb{P}(X = 1) = 0.2,\ \mathbb{P}(X = 2) = 0.4,\ \mathbb{P}(X = 3) = 0.4$, then

$$\mathbb{E}[X] = (1)(0.2) + (2)(0.4) + (3)(0.4) = 2.200.$$

Definition 9
The **indicator function** $\mathbb{1}(\cdot)$ equals 1 when the argument is true and 0 otherwise.

- Example: for $X \sim \exp(2)$, $f_X(s) = 2\exp(-2s)$ when $s > 0$, and $f_X(s) = 0$ when $s > 0$. Write in one line as:

$$f_X(s) = 2\exp(-2s)\mathbb{1}(s \geq 0).$$

- Qotd: what is $\mathbb{E}[X]$?

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} s f_X(s)\, ds.$$

- Qotd: $f_X(s) = \exp(-2s) \cdot \mathbb{1}(s \geq 0)$ ($X \sim \mathsf{Exp}(2)$).

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} s\exp(-2s) \cdot \mathbb{1}(s \geq 0)\, ds \\ &= \int_{0}^{\infty} s\exp(-s)\, ds \text{ (use indicator to set limits)} \\ &= 1/2 = 0.5000 \text{ (use integration by parts).}\end{aligned}$$

Expectations of functions of a variable

- Example, discrete random variables:

$$\mathbb{E}[X^2] = \sum_{i=0}^{\infty} x_i^2 \mathbb{P}(X = x_i).$$

- Example, continuous random variables:

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} s^2 \underbrace{\mathbb{P}(X \in ds)}_{f_X(s)\ ds}.$$

- Another example: discrete

$$\mathbb{E}[\cos(X)\mathbb{1}(-3 < X < 3)] = \sum_{i=0}^{\infty} \cos(x_i)\mathbb{1}(-3 < x_i < 3)\mathbb{P}(X = x_i).$$

- Another example: continuous

$$\mathbb{E}[\cos(X)\mathbb{1}(-3 < X < 3)] = \int_{-\infty}^{\infty} \cos(s)\mathbb{1}(-3 < s < 3) f_X(s)\ ds.$$

2.2 Formal definition of expectation

The strategy

- Define mean for simple random variables
- Use that defn to define mean for nonnegative r.v.
- Use that defn to define mean for general r.v.

Definition 10
A random variable is **simple** if it takes on a finite number of values in $\mathbb{R}$.

Definition 11
For a simple r.v. X that lies in $\{x_1, \ldots, x_n\}$ with probability 1, the **mean** of X is

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \mathbb{P}(X = x_i).$$

Note everything is nice and finite, so don't have to worry about limits here. Makes proofs easy!

Fact 2
For a real number a and simple r.v. X, $\mathbb{E}[aX] = a\mathbb{E}[X]$.

Proof. Since $X \in \{x_1, \ldots, x_n\}$ is simple, $aX \in \{ax_1, ax_2, \ldots, ax_n\}$ is simple as well.
Case I: $a = 0$. Then $\mathbb{E}[aX] = 0(1) = 0$, and $0\mathbb{E}(X) = 0$, so $\mathbb{E}[0 \cdot X] = 0 \cdot \mathbb{E}(X)$.
Case II: $a \neq 0$:

$$\mathbb{E}[aX] = \sum_{i=1}^{n} ax_i \mathbb{P}(aX = ax_i) = a \sum_{i=1}^{n} \mathbb{P}(X = x_i) = a\mathbb{E}[X]. \quad \square$$

$\square$

Fact 3
For X and Y simple r.v., $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Pf: Suppose $X \in \{x_1, \ldots, x_n\}$ and $Y \in \{y_1, \ldots, y_m\}$.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{z: x_i + y_j = z} z\mathbb{P}(X+Y=z) \\
&= \sum_{z: x_i+y_j=z} \sum_{i \in \{1,\ldots,n\}} \sum_{j \in \{1,\ldots,m\}} z\mathbb{P}(X=x_i, Y=y_i)\mathbb{1}(x_i+y_j=z) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i+y_j)\mathbb{P}(X=x_i, Y=y_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i\mathbb{P}(X=x_i, Y=y_j) + \sum_{j=1}^{m}\sum_{i=1}^{n} y_j\mathbb{P}(X=x_i, Y=y_j) \\
&= \sum_{i=1}^{n} x_i \sum_{j=1}^{m}\mathbb{P}(X=x_i, Y=y_j) + \sum_{j=1}^{m} y_j \sum_{i=1}^{n}\mathbb{P}(X=x_i, Y=y_j) \\
&= \sum_{i=1}^{n} x_i\mathbb{P}(X=x_i) + \sum_{j=1}^{m} y_j\mathbb{P}(Y=y_j) \\
&= \mathbb{E}[X] + \mathbb{E}[Y]. \quad \square
\end{aligned}$$

Definition 12
The **extended real numbers** $\mathbb{R} \cup \{\infty\} \cup \{-\infty\}$ are the real numbers together with a symbol for infinity (∞) and negative infinity ($-\infty$).

Definition 13
The **supremum** of a set of numbers A is the least upper bound on the numbers of A. There are three cases.

1: $A = \emptyset$. Then $\sup(A) = -\infty$.

2: $(\exists b \in \mathbb{R})(A \subseteq (-\infty, b])$. Then

$$\sup(A) = \min\{b : A \subseteq (-\infty, b]\}.$$

3: $(\forall b \in \mathbb{R})(\exists a \in A)(a > b)$ Then $\sup(A) = \infty$.

Mnemonic for supremum

- Superman flies above the buildings, but as low as possible so he can respond to crime quickly

Now define expected value for nonnegative random variables.

Definition 14
Suppose $X \geq 0$. Let $\mathcal{X}$ be the set of random variables Y that are both simple and $Y \leq X$. Then the **mean** of X is

$$\mathbb{E}[X] = \sup_{\{Y \in \mathcal{X}\}} \mathbb{E}[Y].$$

Example

- Suppose $X \sim \mathsf{Unif}([0, 10])$, $Y = \lfloor X \rfloor$.
- (Recall $\lfloor x \rfloor$ means greatest integer at most x.)
- (Example: $\lfloor 3.5 \rfloor = 3$, $\lfloor 3 \rfloor = 3$, $\lfloor -2.8 \rfloor = -3$.)

- Y is simple and $\mathbb{E}[Y] = 4.5$. Hence $\mathbb{E}[X] \geq 4.5$.
- On the other hand, $X \leq 10$ with probability 1, so any simple $Y \leq X$ is also at most 10 with probability 1. Hence

$$4.5 \leq \mathbb{E}[X] \leq 10.$$

Remarks

- This definition works if X is discrete or continuous or neither!
- Every nonnegative r.v. has a mean: but it might be ∞.
- Much more general than Riemann integral! This is called the **Lebesgue integral**
- When both Riemann and Lebesgue exists, they are equal.
- Do calculation with infinite sum/infinite Riemann integral.
- These definitions solely for proving theorems.
- What if X is sometimes negative?

Definition 15
For a random variable X, define $X^+ = X \cdot \mathbb{1}(X \geq 0)$ and $X^- = -X \cdot \mathbb{1}(X < 0)$. (Call X^+ the positive part of X, and X^- the negative part of X.) Note $X^+ - X^- = X\mathbb{1}(X \geq 0) + X\mathbb{1}(X < 0) = X$. When both $\mathbb{E}[X^+] < \infty$ and $\mathbb{E}[X^-] < \infty$, the **mean** of X is

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

Example:

- Suppose $X \sim \mathsf{Unif}([-2, 1])$.
- Then $X^+ = X\mathbb{1}(X \geq 0)$. So

$$\begin{aligned}\mathbb{E}[X^+] &= \int_{-\infty}^{\infty} s\mathbb{1}(s \geq 0)(1/3)\mathbb{1}(s \in [-2, 1])\ ds \\ &= \int_0^1 s/3\ ds \\ &= 1/6.\end{aligned}$$

- So

$$\begin{aligned}\mathbb{E}[X^-] &= \int_{-\infty}^{\infty} s\mathbb{1}(s < 0)(1/3)\mathbb{1}(s \in [-2, 1])\ ds \\ &= \int_{-2}^0 s/3\ ds \\ &= 4/6.\end{aligned}$$

- Hence $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] = 1/6 - 4/6 = -1/2$.

3 Swapping limits and expectation

Question of the Day If $X_t \to X$ with probability 1 as $t \to \infty$, does

$$\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}\left[\lim_{t\to\infty} X_t\right] = \mathbb{E}[X]?$$

Limits of sequences

- Limits easier in *extended reals*, $\mathbb{R} \cup \{-\infty, \infty\}$.
- For example

$$\lim_{t\to\infty} 1/t = 0,\ \lim_{t\to\infty} t = \infty,\ \lim_{t\to\infty} -1/t = 0\ \lim_{t\to\infty} -t = -\infty.$$

Sequences

- Many stochastic processes are sequences of real valued random variables

$$X_0, X_1, X_2, \ldots.$$

- Identically distributed sequence:

$$(\forall i, j)(X_i \sim X_j)$$

- Independent sequence:

$$(\forall n)(\forall A_1, \ldots, A_n)(\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n)).$$

- Call an independent, identically distributed sequence *iid*.
- For most iid sequences, $\lim_{t\to\infty} X_t$ does not exist.
- Example: $X_i \overset{\text{iid}}{\sim} \mathsf{Bern}(1/2)$. Infinitely many 0's and 1's in sequence, so no limit!
- Now suppose

$$U \sim \mathsf{Unif}([0,1]), \quad X_t = U/t.$$

- Then

$$\lim_{t\to\infty} X_t = \lim_{t\to\infty} U/t = 0.$$

Expected value

- Expected value = mean = average = expectation
- For $X \in \{x_0, x_1, x_2, \ldots\}$ with probability 1,

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} x_i \mathbb{P}(X = x_i) = \sum_{x:\mathbb{P}(X=x)>0} x\mathbb{P}(X = x).$$

- Mean of X is sum of outcomes time probabilities of outcomes.
- With probability 1 = wp 1 = almost surely = almost everywhere.
- For $X = U/t$, $U \sim \mathsf{Unif}([0,1])$, $\mathbb{E}[X_t] = \mathbb{E}[U]/t = 1/(2t)$, so

$$\lim_{t\to\infty} \mathbb{E}[X_t] = 0 = \mathbb{E}\left[\lim_{t\to\infty} X_t\right].$$

- So certainly sometimes you can bring limits inside expectation!

Example where we can't bring limits inside expectation

- Suppose
$$U \sim \mathsf{Unif}([0,1]), \quad X_t = t\mathbb{1}(U < 1/t).$$
- Example: $U = 0.3$, $X_0 = 0, X_1 = 1, X_2 = 2, X_3 = 3, X_4 = 0, X_5 = 0, \ldots$
- X_t equals t for a while, then goes to 0 for $t > 1/U$.
- As t goes to infinity:
$$\lim_{t\to\infty} X_t = \lim_{t\to\infty} t \cdot \mathbb{1}(U < 1/t) = \begin{cases} 0 & U > 0 \\ \infty & U = 0 \end{cases}$$
- So
$$\lim_{t\to\infty} X_t = 0 \text{ with probability 1}$$
- So
$$\mathbb{E}\left[\lim_{t\to\infty} X_t\right] = \mathbb{E}[0] = 0.$$
- But $\mathbb{E}[X_t] = t(1/t) + 0(1 - 1/t) = 1$ for all t.
- So $\lim_{t\to\infty} X_t = 1$.

3.1 The Lebesgue Dominated Convergence Theorem

What went wrong in the example: the X_t are unbounded!

Theorem 1 (Lebesgue dominated convergence theorem (DCT))
Suppose $\lim_{t\to\infty} X_t = X$ wp 1 and $|X_t| \le Y$ for all t. Then if $\mathbb{E}[|Y|] < \infty$, then
$$\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}\left[\lim_{t\to\infty} X_t\right].$$

3.2 The Monotone Convergence Theorem

What went wrong: the X_t went from large to small.

Theorem 2 (Monotone convergence theorem (MCT))
Suppose $0 \le X_0 \le X_1 \le X_2 \le \cdots$ wp 1. Then $\lim_{t\to\infty} X_t$ is either a finite real number or ∞ with probability 1, and
$$\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}\left[\lim_{t\to\infty} X_t\right].$$

From real analysis we know that any increasing sequence always has a limit.

Fact 4
Suppose $a_0 \le a_1 \le a_2 \le \cdots$. Then $\lim_{i\to\infty} a_i$ is either a finite real number or ∞.

Two ways to bring limits inside mean

1: If sequence is *dominated* by an integrable random variable.

2: If sequence is nonnegative and increasing.

Example of DCT

- $U \sim \mathsf{Unif}([0,1])$, $X_t = U/t$, $|X_t| \leq 1$. ($Y = 1$ is dominating random variable.)
- So
$$\lim_{t\to\infty} \mathbb{E}[U/t] = \mathbb{E}\left[\lim_{t\to\infty} U/t\right] = \mathbb{E}[0] = 0.$$
- $U \sim \mathsf{Unif}([0,1])$, $W_t = 1 - 1/t$.
- Then $W_t \geq 0$ and $W_t \geq W_{t-1}$, so
$$\lim_{t\to\infty} \mathbb{E}[1 - U/t] = \mathbb{E}\left[\lim_{t\to\infty} 1 - U/t\right] = \mathbb{E}[1] = 1.$$

Example of MCT

- Say $\{B_i\}$ is a sequence of independent r.v.'s with $B_i \sim \mathsf{Bern}((1/2)^i)$. Then $\mathbb{E}[B_i] = (1/2)^i$.
- Let $S_n = B_1 + \cdots + B_n$. Then
$$0 \leq S_1 \leq S_2 \leq S_3 \leq S_4 \leq \cdots$$
wp 1.
- So can apply the MCT,
$$\lim_{n\to\infty} \mathbb{E}[S_n] = \mathbb{E}\left[\lim_{n\to\infty} S_n\right].$$
Now $\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[B_i]$. So this means:
$$\mathbb{E}\left[\sum_{i=1}^{\infty} B_i\right] = \sum_{i=1}^{\infty} \mathbb{E}[B_i] = \sum_{i=1}^{\infty} (1/2)^i = 1$$

4 Properties of Expectation

Question of the Day For integrable X, is always true that $\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$?

Definition 16
A random variable X with $\mathbb{E}[|X|] < \infty$ is called **integrable**.

It's often surprising to people that there can be random variables that are finite with probability 1, yet still do not have finite expectation (are not integrable)! Two canonical examples of this are the Cauchy distribution, and the Zipf (power) law distribution with $\alpha \leq 2$.

4.1 Vector spaces and linear operators

Random variables are an example of a *vector space*. They can be added together: if X and Y are random variables, then $X + Y$ is a random variable. They can be scaled: if X is a random variable, and $c \in \mathbb{R}$, then cX is a random variable. (See the appendix for the formal definition of a vector space.)

Definition 17
For a vector space V with scalars S, we say that $\mathcal{L}$ is a **linear operator** if

$$(\forall x, y \in V)(\forall a, b \in S)(\mathcal{L}[ax + by] = a\mathcal{L}[x] + b\mathcal{L}[y]).$$

Example

- Matrix multiplication is a linear operator over vectors in $\mathbb{R}^n$.

$$A(av + bw) = aAv + bAw$$

- Integration is a linear operator over functions with finite integral.

$$\int_A af(x) + bg(x)\ dx = a\int_A f(x)\ dx + b\int_A g(x)\ dx$$

- Differentiation over functions in C^1

$$[af + bg]' = af' + bg'.$$

Fact 5 (Expectation is a linear operator)
For any integrable random variables X and Y, and real numbers a and b:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

- Important: this holds even when X and Y are not independent!
- Example: If $\mathbb{E}[X] = 4$ and $\mathbb{E}[Y] = 10$, what is $\mathbb{E}[2X - Y]$? Answer: $\mathbb{E}[2X - Y] = 2\mathbb{E}[X] - \mathbb{E}[Y] = 2(4) - 10 = -2$.

4.2 Domination and convexity

An important fact is that bigger random variables have bigger means.

Fact 6 (Domination)
If X and Y are integrable r.v. where $X \leq Y$ with probability 1, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Proof. First consider the case where $X \geq 0$.

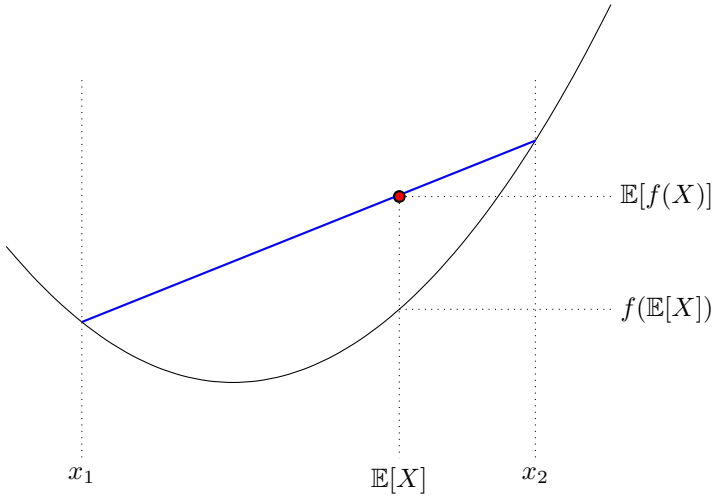
Let $\mathcal{S}_X$ be the set of simple functions dominated by X, and $\mathcal{S}_Y$ be the set of simple functions dominated by Y. Then $\mathcal{S}_X \subseteq \mathcal{S}_Y$, so

$$\sup_{S \in \mathcal{S}_X} \mathbb{E}[S] \leq \sup_{S \in \mathcal{S}_Y} \mathbb{E}[S].$$

The left hand side is just $\mathbb{E}[X]$, and the right hand side is just $\mathbb{E}[Y]$, so we are done with this case.

Now suppose $X \not\geq 0$. Then since $X \leq Y$, $Y - X \geq 0$, so the first cast applies and $\mathbb{E}[Y - X] \geq \mathbb{E}[0] = 0$. By linearity $\mathbb{E}[Y - X] = \mathbb{E}[Y] - \mathbb{E}[X] \geq 0$, so we are done. □

Definition 18
A function is **convex** if the secant line connecting any two points on the graph lies on or above the graph. For example x^2, e^x and $|x|$ are all convex.



Fact 7
For $f \in C^2[a, b]$, if $f''(x) \geq 0$ for all $x \in [a, b]$, then f is convex over $[a, b]$.

Fact 8 (Jensen's inequality)
If $f(x)$ is a convex function then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. In particular:

$$\mathbb{E}[X^2] \geq \mathbb{E}[X]^2, \quad \mathbb{E}[|X|] \geq |\mathbb{E}(X)|.$$

Example:

- In the picture of the convex function above, let $\mathbb{P}(X = x_1) = 30\%$, and $\mathbb{P}(X = x_2) = 70\%$.
- The x-coordinates of the two circles are both $\mathbb{E}[X]$.
- The y-coordinate of the circle on the line is $\mathbb{E}[f(X)]$.
- The y-coordinate of the circle on the curve is $f(\mathbb{E}[X])$.

4.3 Types of convergence for random variables

Convergence with probability 1.

- Start with $\{B_i\}$ iid $\mathsf{Bern}(1/2)$.

- Use the first n values of $\{B_i\}$ and binary digits for X_n.

$$X_n = \sum_{i=1}^{n} \frac{B_i}{2^i}.$$

- Example: $\{B_i\} = 0, 1, 1, 0 \ldots$, $X_1 = 0/2$, $X_4 = 0/2 + 1/4 + 1/8 + 0/16$.
- Note that

$$X_n \to \sum_{i=1}^{\infty} B_i/2^i = U \sim \mathsf{Unif}([0,1])$$

with probability 1.

Recall the limit of a sequence:

Definition 19
Convergence of a sequence. Say $\lim_{t\to\infty} a_t = L$ if

$$(\forall \epsilon > 0)(\exists T)(\forall t \geq T)(|a_t - L| \leq \epsilon).$$

- In words, ϵ represents the difference between the limit and the values of the sequence.
- No matter how small you make ϵ, there is an N such that every term in the sequence past a_N is within distance ϵ of the limit L.

Definition 20
Convergence with probability 1. Say $X_t \to X$ with probability 1 if

$$\mathbb{P}\left(\lim_{t\to\infty} X_t \to X\right) = 1.$$

- What does this mean?
- Remember that random variables are really functions of the outcome. So $X_t = X_t(\omega)$, $X = X(\omega)$.
- For some outcomes ω, it holds that $\lim_{t\to\infty} X_t(\omega) = X(\omega)$. Call this set of ω the event A. So if $\omega \in A$, then the limit statement is true, if $\omega \in A^C$ then it is false.
- Convergence with probability 1 occurs if $\mathbb{P}(\omega \in A) = 1$.

Convergence in probability Suppose that $B_0, B_1, B_2, \ldots$ are independent but not indentically distributed random variables. To be precise, they are independent, but for each i, $B_i \sim \mathsf{Bern}(1/i)$.

Consider the question, what is $\mathbb{P}((\forall i \geq 5)(B_i = 0))$? It is

$$\mathbb{P}(B_5 = 0)\mathbb{P}(B_6 = 0)\mathbb{P}(B_7 = 0)\cdots = \frac{4}{5}\cdot\frac{5}{6}\cdot\frac{6}{7}\cdots = 0$$

The complement rule then gives $\mathbb{P}((\exists i \geq 5)(B_i = 1)) = 1$

There was really nothing special about 5 in the above argument, so in general

$$(\forall n)(\mathbb{P}((\exists i \geq n)(B_i = 1)) = 1)$$

So with probability 1, always see a 1 farther out in the sequence

$$0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, \ldots$$

It is also true that with probability 1, we see 0's infinitely often in the sequence. Taken together, this means that $B_1, B_2, \ldots$ does not converge to either 0 or 1. In other words,

$$\mathbb{P}(\lim_{i\to\infty} B_i \text{ exists}) = 0.$$

But the B_i are closer and closer to 0 as i increases. This is a different type of covergence. It is not convergence with probability 1, we call this type *convergence in probability*.

Now for the formal definition.

Definition 21
Convergence in probability. Say $X_t \to X$ in probability if

$$(\forall \epsilon > 0) \left(\lim_{t \to \infty} \mathbb{P}\left(|X_t - X| > \epsilon \right) = 0 \right).$$

In our example: $B_i \to 0$ in probability

- For $\epsilon > 1$, $|B_i - 0| < 1$ always, so $\mathbb{P}(|B_i - 0| > \epsilon) = 0$
- For $0 < \epsilon < 1$, $\mathbb{P}(|B_i - 0| > \epsilon) = 1/i$, and $\lim_{i \to \infty} 1/i = 0$

Fact 9
If $X_t \to X$ wp 1, then $X_t \to X$ in probability.

- Strong laws: convergence with probability 1.
- Weak laws: convergence in probability.
- Example: Strong Law of Large Numbers: for iid integrable r.v.'s sample average converges to the expected value wp1.
- Our fact means that strong convergence implies weak convergence.
- Turns out we can characterize exactly which weakly convergent sequences we can swap expected value and limits, later in course.

5 Proving limit theorems for expectation

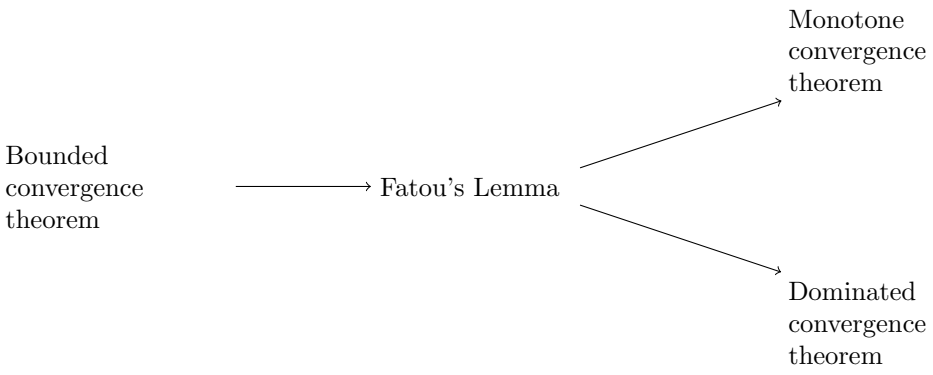
Question of the Day Why are the MCT and DCT true?

When can you swap limits and mean?

1: For $X_t \geq 0$ and $X_0 \leq X_1 \leq X_2 \leq \cdots$ wp 1 (MCT).

2: For $\mathbb{P}(\lim_{t\to\infty} X_t = X) = 1$ and $|X_t| \leq Y$ with $\mathbb{E}[|Y|] < \infty$ (DCT).

Our plan of attack:



Notation: $\min\{a, b\} = a \wedge b$.

5.1 Proving BCT and Fatou's lemma

Fact 10 (Bounded convergence theorem)
For $|X_t| \leq m$ for all t and if $X_t \to X$ in probability then

$$\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}\left[\lim_{t\to\infty} X_t\right] = \mathbb{E}[X].$$

[Recall $X_t \to X$ in prob. if $(\forall \epsilon > 0)(\lim_{t\to\infty} \mathbb{P}(|X_t - X| > \epsilon) = 0$.]

Proof. Let $\epsilon > 0$. Set $G_t = \{\omega : |X_t(\omega) - X(\omega)| \leq \epsilon/2\}$. Note $\mathbb{1}(\omega \in G_t) + \mathbb{1}(\omega \notin G_t) = 1$. So for any r.v.

$$Y = Y\mathbb{1}(\omega \in G_t) + Y\mathbb{1}(\omega \notin G_t).$$

To show that $\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}[X]$, I want to find T large enough that for all $t \geq T$: $|\mathbb{E}[X_t] - \mathbb{E}[X]| \leq \epsilon$.

$$\begin{aligned}
|\mathbb{E}[X_t] - \mathbb{E}[X]| &= |\mathbb{E}[X_t - X]| \leq \mathbb{E}[|X_t - X|] \\
&= \mathbb{E}[|X_t - X|\mathbb{1}(\omega \in G_t) + |X_t - X|\mathbb{1}(\omega \notin G_t)] \\
&= \mathbb{E}[|X_t - X|\mathbb{1}(\omega \in G_t)] + \mathbb{E}[|X_t - X|\mathbb{1}(\omega \notin G_t)]
\end{aligned}$$

At this point note that if $\omega \in G_t$, then $|X_t - X| \leq \epsilon/2$. And since $X_t \to X$ in probability and $|X_t| \leq m$, then $\mathbb{P}(|X| > m) = 0$.

$$\begin{aligned}
|\mathbb{E}[X_t] - \mathbb{E}[X]| &\leq \mathbb{E}[(\epsilon/2)\mathbb{1}(\omega \in G_t) + \mathbb{E}[2m\mathbb{1}(\omega \notin G_t)] \\
&= (\epsilon/2)\mathbb{P}(|X_t - X| \leq \epsilon/2) + 2m\mathbb{P}(|X_t - X| > \epsilon/2) \\
&\leq \epsilon/2 + (2m)\mathbb{P}(|X_t - X| > \epsilon/2).
\end{aligned}$$

Since $\lim_{t\to\infty} \mathbb{P}(|X_t - X| > \epsilon/2) = 0$

$$(\exists T)(\forall t > T)(\mathbb{P}(|X_t - X| > \epsilon/2) < \epsilon/(4m))$$

Hence $|\mathbb{E}[X_t] - \mathbb{E}[X]| \leq \epsilon/2 + 2m\epsilon/(4m) = \epsilon$. □

The next step requires the infimum and infimum limits.

Definition 22
The **infimum** of a set of numbers A is the greatest lower bound on the numbers of A. There are three cases.

1: $A = \emptyset$. Then $\inf(A) = \infty$.

2: $(\exists b \in \mathbb{R})(A \subseteq [b, \infty))$. Then $\inf(A) = \max\{b : A \subseteq [b, \infty)\}$.

3: $(\forall b \in \mathbb{R})(\exists a \in A)(a < b)$ Then $\inf(A) = -\infty$.

Definition 23
The **infimum limit** of a set of numbers $a_0, a_1, a_2, \ldots$ is

$$\liminf(a_0, a_1, \ldots) = \lim_{n\to\infty} \inf\{a_n, a_{n+1}, a_{n+2}, \ldots\}.$$

Example

- $\liminf 2, 3, 2, 3, 2, 3, 2, 3, \ldots = 2$.
- $\liminf 1, 1/2, 1/3, 1/4, \ldots = 0$.
- Note all increasing sequences have limits.
- $\inf\{a_n, a_{n+1}, \ldots\}$ is an increasing sequence.
- So $\liminf a_n$ always exists! (Might be ∞.)
- When $\lim a_n$ exists, $\lim a_n = \liminf a_n$.

Before proving Fatou, we need the following:

Fact 11
Suppose $X \geq 0$ has finite expectation. Then

$$\lim_{m\to\infty} \mathbb{E}[X \wedge m] = \mathbb{E}[X].$$

Pf: Let S_X be all simple functions dominated by X. Then $(\forall Y \in S)(\mathbb{E}[Y] \leq \mathbb{E}[X])$.
Fix m. Let Y be simple with $Y \leq X \wedge m \leq X$. Then $\mathbb{E}[Y] \leq \mathbb{E}[X]$, so $\mathbb{E}[X \wedge m] \leq \mathbb{E}[X]$.
Also $X \wedge m \leq X \wedge (m+1) \Rightarrow \mathbb{E}[X \wedge m] \leq \mathbb{E}[X \wedge (m+1)]$.
That makes $\mathbb{E}[X \wedge 1], \mathbb{E}[X \wedge 2], \ldots$ an increasing sequence bounded above by $\mathbb{E}[X]$.
Hence $\lim_{m\to\infty} \mathbb{E}[X \wedge m]$ exists and is at most $\mathbb{E}[X]$.
Let W be any simple random variable less than X.
Then $W \leq M$ for some M.
Hence for $m > M$, $\mathbb{E}[W] \leq \mathbb{E}[X \wedge m]$
Therefore, for all simple $W \leq X$, $\mathbb{E}[W] \leq \lim_{m\to\infty} \mathbb{E}[X \wedge m]$.
So $\lim_{m\to\infty} \mathbb{E}[X \wedge m]$ is an upper bound on $S = \{\mathbb{E}[W] : W \leq X \text{ and simple}\}$.
Hence $\lim_{m\to\infty} \mathbb{E}[X \wedge m]$ is greater than $\mathbb{E}[X]$, which is the least upper bound on S. □

Fact 12 (Fatou's Lemma)
If $X_t \geq 0$ then

$$\liminf \mathbb{E}[X_t] \geq \mathbb{E}\left[\liminf X_t\right].$$

[Can always bring lim inf inside expectation at the cost of a greater than or equal to sign.]

Pf: Let $Y_t = \inf_{r \geq t} X_t$. [This makes $\liminf X_t = \lim Y_t$.]
The Y_t are increasing and converge to $Y = \liminf X_t$.
Since infima are lower bounds, $Y_t \leq X_t$.
So $\mathbb{E}[Y_t] \leq \mathbb{E}[X_t]$ for all t.
New goal: show $\liminf \mathbb{E}[X_t] \geq \lim \mathbb{E}[Y_t] \geq \mathbb{E}[Y]$.
Fix $m > 0$, so $Y_t \wedge m \leq m$. Now have bounded r.v.!
BCT: $\lim \mathbb{E}[Y_t \wedge m] = \mathbb{E}[\lim Y_t \wedge m] = \mathbb{E}[Y \wedge m]$.
So $\lim \mathbb{E}[Y_t] \geq \lim [Y_t \wedge m] = \mathbb{E}[Y \wedge m]$.
Use first fact from today's lecture: let m go to infinity.
Gives $\lim \mathbb{E}[Y_t] \geq \mathbb{E}[Y]$, and we're done! $\square$

5.2 Proving MCT and DCT

Theorem 3 (Monotone convergence theorem (MCT))
Suppose $0 \leq X_0 \leq X_1 \leq X_2 \leq \cdots$ wp 1. Then $\lim_{t\to\infty} X_t$ exists (and is maybe ∞) with probability 1, and

$$\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}\left[\lim_{t\to\infty} X_t\right].$$

Proof. $\square$

Pf: Since $\liminf X_t = \lim X_t$ for increasing X_t, Fatou gives
$\lim \mathbb{E}[X_t] = \liminf \mathbb{E}[X_t] \geq \mathbb{E}[\liminf X_t] = \mathbb{E}[\lim X_t]$.
Since the X_t are increasing, $X_t \leq \lim X_t$ and $\mathbb{E}[X_t] \leq \mathbb{E}[\lim X_t]$.
Hence $\lim \mathbb{E}[X_t] \leq \mathbb{E}[\lim X_t]$.
That's both directions of inequality, so we have equality! $\square$.

To finish the DCT, recall that the supremum of a set of numbers A is the least upper bound on A.

$$\sup A = S \Leftrightarrow (\forall a \in A)(S \geq a) \wedge (\forall \epsilon > 0)(\exists a \in A)(a > S - \epsilon).$$

Convention: $\sup \emptyset = -\infty$. Just like the infimum, we have a supremum limit, the lim sup.

Definition 24
The **supremum limit** of a set of numbers $a_0, a_1, a_2, \ldots$ is

$$\limsup(a_0, a_1, \ldots) = \lim_{n\to\infty} \sup\{a_n, a_{n+1}, a_{n+2}, \ldots\}.$$

Examples

- $\limsup 2, 3, 2, 3, 2, 3, 2, 3, \ldots = 3$.
- $\limsup 1, 1/2, 1/3, 1/4, \ldots = 0$.
- Note all decreasing sequences have limits (might be $-\infty$).
- $\sup\{a_n, a_{n+1}, \ldots\}$ is an increasing sequence in n.
- So $\limsup a_n$ always exists! (Might be $-\infty$.)
- When $\lim a_n$ exists, $\lim a_n = \limsup a_n$.
- $\liminf -a_n = \limsup a_n$.

Fact 13
If $\liminf a_n = \limsup a_n$, then both are equal to $\lim a_n$.

Theorem 4 (Lebesgue dominated convergence theorem (DCT))
Suppose $\lim_{t\to\infty} X_t = X$ wp 1 and $|X_t| \leq Y$ for all t. Then if $\mathbb{E}[|Y|] < \infty$, then

$$\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}\left[\lim_{t\to\infty} X_t\right].$$

Pf: Since $|X_t| \leq Y$, $X_t + Y \geq 0$. Apply Fatou to get
$\liminf \mathbb{E}[X_t + Y] \geq \mathbb{E}[\lim X_t + Y] = \mathbb{E}[X + Y]$.
By linearity $\liminf(\mathbb{E}[X_t] + \mathbb{E}[Y]) \geq \mathbb{E}[X] + \mathbb{E}[Y]$.
Put the constant out of the lim inf to get $\liminf \mathbb{E}[X_t] \geq \mathbb{E}[X]$.
Use same argument with $-X_t + Y$ to get $\liminf \mathbb{E}[-X_t] \geq -\mathbb{E}[X]$.
So $\limsup \mathbb{E}[X_t] \leq \mathbb{E}[X]$.
Together $\limsup \mathbb{E}[X_t] \leq \mathbb{E}[X] \leq \liminf \mathbb{E}[X_t]$.
Always have $\liminf \mathbb{E}[X_t] \leq \limsup \mathbb{E}[X_t]$.
So they all must be the same value! $\square$

6 Martingales

Question of the Day Suppose I play a fair game where I either win or lose a dollar every play each with probability 1/2. If I start with 15 dollars, what is the expected amount of money I have after 12 plays?

6.1 Intuitive notion of martingale

Fair games

- If I win or lose a dollar with probability 1/2, I am playing a fair game.
- If I win with probability 30% (or 60%) that is an unfair game.
- Martingales are the amount of money you have in fair games.

For question of the day

- $\mathbb{P}(M_1 = M_0 + 1 | M_0) = 1/2,\ \mathbb{P}(M_1 = M_0 - 1 | M_0) = 1/2.$
- What is $\mathbb{E}[M_1 | M_0]$?
- Just apply regular rules for expectation, but treat M_0 as a constant rather than a random variable

$$\begin{aligned}\mathbb{E}[M_1|M_0] &= (1/2)(M_0+1) + (1/2)(M_0-1)\\ &= (1/2)M_0 + 1/2 + (1/2)M_0 - 1/2\\ &= M_0\end{aligned}$$

- Now consider

$$\mathbb{E}[M_2|M_0, M_1] = (1/2)(M_1+1) + (1/2)(M_1-1) = M_1.$$

Intuition: A *martingale* $\{M_i\}$ consists of integrable random variables that form a fair game.

1: $\mathbb{E}[|M_n|] < \infty$

2: for all $n > 0$, $\mathbb{E}[M_{n+1}|M_0, \ldots, M_n] = M_n$.

Qotd

- Is the qotd game a martingale?
- Check condition one. First let

$$D_0, D_1, D_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Unif}(\{-1, 1\}), \text{ and } M_n = 3 + \sum_{i=1}^{n} D_i.$$

- Recall $|a + b| \leq |a| + |b|$ (triangle inequality). So

$$\begin{aligned}\mathbb{E}[|M_n|] &= \mathbb{E}\left[\left|3 + \sum_{i=1}^{n} D_i\right|\right]\\ &\leq \mathbb{E}\left[|3| + \sum_{i=1}^{n} |D_i|\right]\\ &= \mathbb{E}\left[|3|\right] + \sum_{i=1}^{n} \mathbb{E}\left[|D_i|\right]\\ &= 3 + n < \infty.\end{aligned}$$

- Note: okay for $\mathbb{E}[|M_n|]$ to increase in n, the condition is just that $\mathbb{E}[|M_n|]$ be finite, not that the limit as $n \to \infty$ be finite
- Now check condition 2:

$$\mathbb{E}[M_n|M_0, \ldots, M_{n-1}] = (1/2)(M_{n-1} + 1) + (1/2)(M_{n-1} - 1) = M_{n-1}.$$

6.2 Properties of conditional expectation

- Going to be using statements like

$$\mathbb{E}[X|Y]$$

so we need some properties and a formal definition

Fact 14
Properties of conditional expectation:

1: Conditional expectations are linear (like regular expectation)

$$\mathbb{E}[aA + bB|C] = a\mathbb{E}[A|C] + b\mathbb{E}[B|C]$$

2: If X and Y are independent, then

$$\mathbb{E}[X|Y] = \mathbb{E}[X].$$

3: If X is a function of Y, so $X = f(Y)$, then

$$\mathbb{E}[X|Y] = f(Y).$$

4: $\mathbb{E}[X|Y]$ will always be some function of Y.

5: (Fundamental Theorem of Probability) For any X and Y

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

Comments

- Property 3 example: $\mathbb{E}[A^2|A] = A^2$. In other words, if you condition on A, treat the random variable like it is a constant.
- Property 4 example $X_1, X_2 \overset{\text{iid}}{\sim} \mathsf{Unif}(\{1, 2, 3, 4, 5, 6\}), S = X_1 + X_2$.

$$\begin{aligned}\mathbb{E}[S|X_2] &= \mathbb{E}[X_1 + X_2|X_2] \\ &= \mathbb{E}[X_1|X_2] + \mathbb{E}[X_2|X_2] \\ &= 3.5 + X_2.\end{aligned}$$

$3.5 + X_2$ is a function of X_2, the random variable that we are conditioning on.

More generally, $\mathbb{E}[X|A_1, A_2, \ldots, A_n]$ will be a function of $A_1, A_2, \ldots, A_n$. Recall functions don't always use all their inputs: $f(x, y, z) = xy$ doesn't use z, still a function of x, y, and z.

- Property 5 example:

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S|X_2]] = \mathbb{E}[3.5 + X_2] = 3.5 + \mathbb{E}[X_2] = 7$$

To solve QotD

Fact 15
For $M_0, M_1, \ldots$ a martingale and $t \geq 0$, $\mathbb{E}[M_t|M_0] = M_0$.

Proof. Proof by induction. For $t = 0$, $\mathbb{E}[M_0|M_0] = M_0$. Induction hypothesis: Suppose $\mathbb{E}[M_{n-1}|M_0] = M_0$.

$$\begin{aligned}\mathbb{E}[M_n|M_0] &= \mathbb{E}[\mathbb{E}[M_n|M_{n-1}, M_{n-2}, \ldots, M_0]|M_0]] \\ &= \mathbb{E}[M_{n-1}|M_0] = M_0\end{aligned}$$

by the induction hypothesis. □

For QotD:

$$\mathbb{E}[M_{12}|M_0 = 15] = 15.$$

What's coming...

- Suppose one step is fair...
- Then any fixed number of steps is fair.
- What about a random number of steps?
- Might not be, but often is.
- Optional Sampling Theorems tells when we can do that.
- One of our big theorems!
- Next time: stopping times

7 Encoding information

Question of the Day How can the information in a random variable be represented?

7.1 The σ-algebra generated by a random variable

Example

- I have a radiation badge that turns black when I receive more than 10 rems of radiation.
- Let outcome ω be amount of radiation that I receive.
- Let $X = \mathbb{1}(\omega \geq 10)$.
- What information does X give me about ω?
- Given X, can determine if $\omega \in [10, \infty)$.
- Given X, can determine if $\omega \in [0, 10)$.
- Given X, cannot determine if $\omega \in [5, 15]$.
- Let $\mathcal{F}$ be those A such that given X, can determine if $Y \in A$.
- Call $\mathcal{F}$ the *σ-algebra generated by* X, write $\mathcal{F} = \sigma(X)$.

$$\sigma(X) = \mathcal{F} = \{[0, \infty), \emptyset, [10, \infty), [0, 10)\}.$$

- Then $\mathcal{F}$ will be a σ-algebra.

Definition 25
Let X be a random variable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\Omega', \mathcal{F}')$. Then

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{F}'\}.$$

- Ex: $\Omega = \mathbb{R}$, $\mathcal{F}$ = Borel sets, $\mathbb{P}$ is unknown, $\Omega' = \{0, 1\}$, $\mathcal{F}' = \{\emptyset, \{0, 1\}, \{0\}, \{1\}\}$, $X(\omega) = \mathbb{1}(\omega \geq 10)$.
- Then

$$X^{-1}(\{1\}) = \{\omega : X(\omega) = 1\} = \{\omega \geq 10\} = [10, \infty).$$

7.2 X measurable with respect to $\mathcal{F}$

- Intuition: X is measurable with respect to $\mathcal{F}$ means that if you know YES/NO to is $\omega \in A$ for all $A \in \mathcal{F}$, then you can figure out the value of X.
- Note is rad badge ex, knowing yes or no for $[10, \infty)$ enough to pin down value of $\mathcal{F}$. Any σ-algebra that contains these sets determines X.

Definition 26
Say that X is **measurable with respect to $\mathcal{F}$** if $\sigma(X) \subseteq \mathcal{F}$.

Note that X is always measurable wrt $\sigma(X)$, since $\sigma(X) \subseteq \sigma(X)$.

Fact 16
If $X_1, \ldots, X_n$ are measurable with respect to $\mathcal{F}$, then so is any function of $X_1, \ldots, X_n$.

Fact 17
If $\mathcal{F} = \sigma(\mathcal{X})$, then

$$\mathbb{E}[Y|\mathcal{F}] = \mathbb{E}[Y|X].$$

Filtrations

- Another badge tells me if rems at least 20.
- Let $W = \mathbb{1}(Y \geq 20)$.

$$\sigma(X, W) = \{[0, \infty), \emptyset, [10, \infty), [0, 10), \\ [0, 20), [20, \infty), [0, 10) \cup [20, \infty), [10, 20)\}$$

- More information = larger σ-algebra.
- So $\sigma(X) \subseteq \sigma(X, W)$.
- For a process, $X_1, X_2, \ldots$,

$$\sigma(X_1) \subseteq \sigma(X_1, X_2) \subseteq \sigma(X_1, X_2, X_3) \subseteq \cdots .$$

Definition 27
A sequence of σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is a **filtration** if

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \cdots .$$

Definition 28
For random variables $X_1, X_2, \ldots$, the **natural** or *adapted filtration* is

$$\sigma(X_1), \sigma(X_1, X_2), \sigma(X_1, X_2, X_3), \ldots .$$

Now we can formally define what a martingale is!

Definition 29
A stochastic process $M_0, M_1, \ldots$ is a **martingale** with respect to a filtration $\mathcal{F}_n$ if for all n:

1: M_n is measurable with respect to $\mathcal{F}_n$.

2: $\mathbb{E}[|M_n|] < \infty$

3: $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n$.

Example

- Let $D_1, D_2, D_3, \ldots \overset{\text{iid}}{\sim} \mathsf{Unif}(\{-1, 1\})$.
- Let $\mathcal{F}_n = \sigma(D_1, D_2, \ldots, D_n)$.
- Let $M_n = \sum_{i=1}^{n} D_i$. (Empty sums equal 0, so $M_0 = 0$.)
- Is M_n a martingale? Yes!

Proof. Let $n > 0$. Then $D_1, \ldots, D_n$ are measurable with respect to $\mathcal{F}_n = \sigma(D_1, \ldots, D_n)$, and M_n is a function of $D_1, \ldots, D_n$, so it is $\mathcal{F}_n$ measurable.

Next $M_n \in [-n, n]$, so $|M_n| \leq n$ and $\mathbb{E}[|M_n|] \leq n < \infty$.

Also,

$$\begin{aligned} \mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \mathbb{E}[M_{n+1}|D_1, \ldots, D_n] \\ &= \mathbb{E}[D_{n+1} + D_1 + \cdots + D_n|D_1, \ldots, D_n] \\ &= D_1 + \cdots + D_n + \mathbb{E}[D_{n+1}] \\ &= M_n \end{aligned}$$

□

Conditional expectation One more rule for conditional expectation: if X is measurable with respect to $\mathcal{F}$, then

$$\mathbb{E}[XY|\mathcal{F}] = X\mathbb{E}[Y|\mathcal{F}].$$

So conditioning on $\mathcal{F}$ means treat X like a constant rather than a random variable.

7.3 Formal definition of conditional expectation

Use σ-algebras to formally define what conditional expectation:

Definition 30
The **conditional expectation of X with respect to $\mathcal{F}$** is any random variable $Y = \mathbb{E}[X|\mathcal{F}]$ that satisfies

1: Y is measurable with respect to $\mathcal{F}$.

2: For any $A \in \mathcal{F}$, $\mathbb{E}[Y\mathbb{1}(\omega \in A)] = \mathbb{E}[X\mathbb{1}(\omega \in A)]$.

Notes

- This definition is nonconstructive, it does not tell us **how** to find conditional expectations.
- To find $\mathbb{E}[X|Y]$, use rules from earlier.
- This definition solely for proving things about conditional expectation.

Example

- Let $\omega \sim \mathsf{Unif}([0, 100])$.
- Say $Y = \omega$, $X = \mathbb{1}(Y \geq 10)$.
- What is $\mathbb{E}[Y|X]$?
- Here $\sigma(X) = \mathcal{F} = \{[0, 100], \emptyset, [0, 10), [10, 100]\}$.
- Want $W = \mathbb{E}[Y|X]$ to be measurable with respect to $\mathcal{F}$.
- Want $\mathbb{E}[W\mathbb{1}(\omega \in A)] = \mathbb{E}[Y\mathbb{1}(\omega \in A)]$ for all A.

$$\begin{aligned}\mathbb{E}[Y\mathbb{1}(\omega \in [0, 10))] &= \int_{\omega=0}^{100} \omega\mathbb{1}(\omega \in [0, 10))(1/100)\ d\omega \\ &= \int_{\omega=0}^{10} \omega(1/100)\ d\omega \\ &= 1/2.\end{aligned}$$

- When $X = 1$, $Y \geq 10$, can show $[Y|X = 1] \sim \mathsf{Unif}([10, 100])$.
- When $X = 0$, $Y \geq 10$, can show $[Y|X = 0] \sim \mathsf{Unif}([0, 10))$.
- So $[Y|X] \sim \mathsf{Unif}([10X, 10 + 90X])$.
- So $\mathbb{E}[Y|X] = [10 + 90X + 10X]/2 = 5 + 50X$.
- Test this:

A	$\mathbb{E}[(5 + 50X)\mathbb{1}(\omega \in A)]$	$\mathbb{E}[Y\mathbb{1}(\omega \in A)]$
$[0, 100]$	$5 + 50(90/100) = 50$	$(0 + 100)/2 = 50$
$\emptyset$	0	0
$[0, 10)$	$\mathbb{E}[5\mathbb{1}(\omega \in A)] = 5(1/10) = 1/2$	$1/2$
$[10, 100]$	$(55)(9/10) = 49.5$	$50 - 1/2 = 49.5$

The point There is a formal definition of conditional expectation, but it is never used for calculation, only for proofs. (Much like the formal definition of expected value.)

8 Stopping Times

Question of the Day Suppose I play a fair game where I either win or lose a dollar each (independent) play with probability 1/2. If I start with 3 dollars, and quit when I hit 0 or 10, what is the chance that I walk away with 10?

8.1 What is a stopping time?

Stopping times

- Example sequence from qotd game: $3, 2, 3, 4, 3, 2, 1, 0$
- For QotD, have M_t dollars at time t:

$$\begin{aligned} T &= \text{first time the process hits 0 or 10} \\ &= \inf\{t : M_t = 0 \text{ or } M_t = 10\}, \end{aligned}$$

- Given $M_0, M_1, \ldots, M_{15}$, can tell if $T \leq 15$ or not.
- Recall: $\mathcal{F}_n$ is all the information up to time n.

Definition 31
T is a **stopping time** with respect to a filtration $\{\mathcal{F}_n\}$ if for all n, the event $\{T \leq n\}$ is in $\mathcal{F}_n$.

Intuitively: it must be possible to determine if T happens or not by time n given all the information up to time n. Cannot look into the future.

8.2 Using stopping times with martingales

Recall

- For any fixed t, $\mathbb{E}[M_t|M_0] = M_0 \Rightarrow \mathbb{E}[M_t] = \mathbb{E}[M_0]$
- Does $\mathbb{E}[M_T] = \mathbb{E}[M_0]$?

Qotd

- Suppose that in the QotD, $\mathbb{E}[M_T] = \mathbb{E}[M_0] = 3$, and that we know the stopping time is finite wp1, that is, $\mathbb{P}(T < \infty) = 1$.
- When $T < \infty$, we have $M_T \in \{0, 10\}$
- So
$$\mathbb{E}[M_T] = 0 \cdot \mathbb{P}(M_T = 0) + 10 \cdot \mathbb{P}(M_T = 10) = 10 \cdot \mathbb{P}(M_T = 10) = 3,$$
so $\mathbb{P}(M_T = 10) = 3/10 = \boxed{0.3000}$.

8.3 The stopped process

To show that $\mathbb{E}[M_T] = \mathbb{E}[M_0]$ for qotd, need the *stopped process*

Definition 32
If $\{M_t\}$ is a stochastic process with stopping time T, call

$$M'_t = M_{t \wedge T}$$

the **stopped process**.

(Recall $a \wedge b = \min\{a, b\}$.)

- Example: $M_0 = 3, M_1 = 4, 3, 2, 1, 0, -1, -2, -1, 0, 1, \ldots$
- $T = 5$. So $M_{3\wedge T} = M_3 = 2$. $M_{7\wedge T} = M_T = M_5 = 0$.
- $\{M_{t\wedge T}\} = 3, 4, 3, 2, 1, 0, 0, 0, 0, 0, 0$.

Fact 18
If M_t is a martingale and T is a stopping time wrt to the same filtration, the stopped process $M'_t = M_{t\wedge T}$ is also a martingale.

Proof. For M'_t to be a martingale, we must show three things for all $t \geq 0$:

1: M'_t is $\mathcal{F}_t$ measurable,

2: $\mathbb{E}[|M'_t|] < \infty$, and

3: $\mathbb{E}[M'_{t+1}|\mathcal{F}_t] = M'_t$.

Consider these in order.

1: Let $t \geq 0$. Then $M'_t = M_{t\wedge T} = M_t \cdot \mathbb{1}(t < T) + M_T \cdot \mathbb{1}(T \leq t)$. All the information up to time t is enough to determine M_t, $\mathbb{1}(T \leq t)$, and $\mathbb{1}(T > t)$. When $T \leq t$, M_T is one of $M_0, \ldots, M_t$, and so it determined as well by $\mathcal{F}_t$. Hence M'_t is $\mathcal{F}_t$ measurable.

2: Next,

$$\begin{aligned}\mathbb{E}[|M'_t|] &= \mathbb{E}[|M_t|\mathbb{1}(t < T)|] + \mathbb{E}[|M_T|\mathbb{1}(T \leq t)] \\ &\leq \mathbb{E}[|M_t|] + \mathbb{E}[|M_0| + |M_1| + |M_2| + |M_3| + \cdots + |M_t|]\end{aligned}$$

since if $T \leq t$ then M_T has to equal one of $M_0, \ldots, M_t$. But

$$\mathbb{E}[|M_t|] + \mathbb{E}[|M_0| + |M_1| + |M_2| + |M_3| + \cdots + |M_t|] \leq \mathbb{E}[|M_t|] + \sum_{i=0}^{t} \mathbb{E}[|M_i|],$$

and all of these are finite since M_t is a martingale, so their sum is finite as well.

3: Now to show $\mathbb{E}[M'_{t+1}|\mathcal{F}_t] = M'_t$. We'll use that

$$M'_{t+1} = M_{(t+1)\wedge T} = M_{t+1}\mathbb{1}(T \geq t+1) + M_T\mathbb{1}(T \leq t).$$

$$\mathbb{E}[M'_{t+1}|\mathcal{F}_t] = \mathbb{E}[M_{t+1}\mathbb{1}(T \geq t+1) + M_T\mathbb{1}(T \leq t)|\mathcal{F}_t]$$

Given the info up to time t, $\mathbb{1}(\{T \leq t\}$ is $\mathcal{F}_t$-measurable, which means $1 - \mathbb{1}(T \leq t) = \mathbb{1}(T \geq t+1)$ is also $\mathcal{F}_t$-measurable.

Also,

$$M_T\mathbb{1}(T \leq t) \in \{0, M_0, M_1, M_2, \ldots, M_t\},$$

so is also $\mathcal{F}_t$-measurable.

The only piece which isn't $\mathcal{F}_t$-measurable is M_{t+1}. So

$$\begin{aligned}\mathbb{E}[M'_{t+1}|\mathcal{F}_t] &= \mathbb{1}(T \geq t+1)\mathbb{E}[M_{t+1}|\mathcal{F}_t] + M_T\mathbb{1}(T \leq t) \\ &= M_t\mathbb{1}(T \geq t+1) + M_T\mathbb{1}(T \leq t) \\ &= M_{t\wedge T}\end{aligned}$$

□

Showing that $\mathbb{P}(T < \infty) = 1$

- We also needed in the QotD that $\mathbb{P}(T < \infty) = 1$. To show this, think of the martingale as an infinite sequence of $+1$ changes represented by $+$ and -1 represented by $-$.

$$+ + + + + + - - - + + - + - - - - - - + + \dots$$

- Break this stream of moves into sections of 10 moves. Note that if any group of ten moves is up, then $T < \infty$. Let A_k be the event that the kth group of moves are all up moves.

$$+ + + + - + + - - + | - - - - - + + - - - | - - + + - - + + - - | + + + + + + + + + + | \dots$$

$$A_1 = F \qquad A_2 = F \qquad A_3 = F \qquad A_4 = T$$

- If $A_k = T$ for any k, then $T < \infty$. So

$$\mathbb{P}(T = \infty) = \mathbb{P}(A_1^C A_2^C \cdots) = \prod_{i=1}^{\infty} \left[1 - \left(\frac{1}{2}\right)^{10}\right] = 0.$$

Solving the QotD

- Since $M_{t\wedge T}$ is a martingale, for all t:

$$\mathbb{E}[M_{t\wedge T} | M_0 = 3] = 3 \Rightarrow \lim_{t\to\infty} \mathbb{E}[M_{t\wedge T} | M_0 = 3] = 3$$

- Since $M_{t\wedge T} \in [0, 10]$, use dominated convergence theorem to take limit as $t \to \infty$ inside expression:

$$\mathbb{E}\left[\lim_{t\to\infty} M_{t\wedge T} | M_0 = 3\right] = 3.$$

- Since $\mathbb{P}(T < \infty) = 1$, $\lim_{t\to\infty} M_{t\wedge T} = M_T$.

$$\mathbb{E}[M_T | M_0 = 3] = 3$$

- Now just use regular method for evaluating M_T (sum outcomes times probabilities!)

$$\begin{aligned} \mathbb{E}[M_T | M_0 = 3] &= 10 \cdot \mathbb{P}(M_T = 10 | M_0 = 3) + 0 \cdot \mathbb{P}(M_T = 0 | M_0 = 3) \\ 3 &= 10\mathbb{P}(M_T = 10 | M_0 = 3) \\ \mathbb{P}(M_T = 10 | M_0 = 3) &= 3/10 = \boxed{0.3000}. \end{aligned}$$

Steps in this kind of problem

1: First show that $\mathbb{E}[M_T] = M_0$.

2: Then write $\mathbb{E}[M_T] = \sum_t \mathbb{P}(T = t)\mathbb{E}[M_t | T = t]$.

Does $\mathbb{E}[M_T]$ always equal M_0?

- NO!
- Let $T_0 = \inf\{t : M_t = 0\}$. Turns out, still have $\mathbb{P}(T < \infty) = 1$
- But $\mathbb{E}[M_T] = 0 \neq 3$!
- Recall earlier, used dominated convergence theorem since $M_{t\wedge T} \in [0, 10]$. But $M_{t\wedge T_0}$ here cannot be dominated by an integrable random variable.
- Cannot always bring limit as $t \to \infty$ inside an expectation.

9 Artificial martingales

Question of the Day Suppose I play an unfair game where I win a dollar with probability 45%, and lose with probability 55%. If I start with 3 dollars, and quit when I hit 0 or 10, what is the chance that I walk away with 10?

Stopping time

- Let M_t be money after t steps of game ($M_0 = 3$)
- Let $T = \inf\{t : M_t = 0 \text{ or } M_t = 10\}$
- Last time used the fact that M_t is a martingale.
- With only 45% chance of winning, M_t no longer a martingale.

$$\mathbb{E}[M_{t+1}|\mathcal{F}_t] = 0.45(M_t + 1) + 0.55(M_t - 1) = M_t - 0.1$$

- So $\mathbb{E}[M_T|M_0] \neq M_0$

9.1 Multiplicative martingales

Idea: create an artificial martingale

- Let $D_1, D_2, D_3, \ldots \overset{\text{iid}}{\sim} D$, $\mathbb{P}(D = 1) = 0.45$, $\mathbb{P}(D = -1) = 0.55$.
- Let $\mathcal{F}_t = \sigma(D_1, D_2, \ldots, D_t)$
- Then $M_t = M_0 + \sum_{i=1}^t D_i$ is not a martingale
- Let $N_t = (0.55/0.45)^{\sum_{i=1}^t D_i}$ is a martingale w.r.t $\mathcal{F}_t$

  1: N_t is a function of $D_1, \ldots, D_t$, so is measurable w.r.t $\mathcal{F}_t$.

  2: $|N_t| \in [0, 1]$, so $\mathbb{E}[|N_t|] < \infty$.

  3: For all $t > 0$,

$$\begin{aligned}\mathbb{E}[N_{t+1}|\mathcal{F}_t] &= 0.45[N_t(0.55/0.45)^1] + 0.55[N_t(0.55/0.45)^{-1}] \\ &= 0.55N_t + 0.45N_t = N_t.\end{aligned}$$

Solving the QotD

- First step, show that $\mathbb{P}(T < \infty) = 1$
- There is a $(1/2)^9$ chance next nine games are a loss.
- If nine games in a row lose, then $T \leq t$.
- So $\mathbb{P}(T > 9k) \leq (1 - (0.55)^9)^k$.
- $\mathbb{P}(T = \infty) = \lim_{k\to\infty}(1 - (0.55)^9)^k = 0$.
- Since N_t is a martingale, so is $N_{T\wedge t}$, so

$$\mathbb{E}[N_{t\wedge T}|N_0 = 1] = 1.$$

- Since $N_{t\wedge T} \in [(55/45)^{-10}, (55/45)^{10}]$, use DCT to take limit as $t \to \infty$ inside expression:

$$\mathbb{E}\left[\lim_{t\to\infty} N_{t\wedge T}|N_0 = 1\right] = 1.$$

- Since $\mathbb{P}(T < \infty) = 1$, $\lim_{t\to\infty} N_{t\wedge T} = N_T$.

$$\mathbb{E}[N_T | N_0 = 1] = 1.$$

- Now use basic method of evaluating N_T (sum outcomes times probabilities). Let $p = \mathbb{P}(M_T = 10 | M_0 = 3)$

$$1 = \mathbb{E}[N_T | N_0 = 1] = (0.55/0.45)^7 p + (0.55/0.45)^{-3}(1-p)$$
$$p = \frac{1 - (0.55/0.45)^{-3}}{(0.55/0.45)^7 - (0.55/0.45)^{-3}} \approx 0.1282\ldots$$

What if slightly worse game?

- Suppose chance of winning now 0.40.
- Then following the same path gives

$$\mathbb{P}(M_T = 10 | M_0 = 3) = \frac{1 - (0.60/0.40)^{-3}}{(0.60/0.40)^7 - (0.60/40)^{-3}} \approx 0.04191\ldots$$

- Small change in winning chance leads to about one third chance of winning!
- High sensitivity to chance of winning because of exponential.
- Sometimes called "Gambler's Ruin"

9.2 Additive artificial martingales

- Used multiplicative factors to make this a martingale.
- Can also use additive terms.

Example

- Back to 45% chance of winning, start with \$3, quit when reach \$0 or \$10.
- What is the expected number of steps before quitting?
- To solve, create artificial martingale with addition

$$R_t = M_t + 0.1t = M_0 + \sum_{i=1}^{t} D_t + 0.1t$$

- (The 0.1 comes from $-\mathbb{E}[D]$.)
- Is R_t a martingale?

  1: R_t is a function of $D_1, \ldots, D_t$, so is measurable w.r.t $\mathcal{F}_t$.

  2: $|R_t| \leq 1.1t$, so $\mathbb{E}[|R_t|] < \infty$.

3: For all $t > 0$,

$$\begin{aligned}\mathbb{E}[R_{t+1}|\mathcal{F}_t] &= \mathbb{E}\left[M_0 + \sum_{i=1}^{t+1} D_i + 0.1(t+1)|\mathcal{F}_t\right] \\ &= M_0 + \sum_{i=1}^{t} D_i + 0.1(t+1) + \mathbb{E}[D_{t+1}] \\ &= M_0 + \sum_{i=1}^{t} D_i + 0.1(t+1) + (0.45)(1) + (0.55)(-1) \\ &= M_0 + \sum_{i=1}^{t} D_i + 0.1(t+1) - 0.1 \\ &= M_0 + \sum_{i=1}^{t} D_i + 0.1t = R_t\end{aligned}$$

- So it is a martingale.
- Hence
$$\mathbb{E}[R_{t\wedge T}|R_0 = 3] = 3,$$
or equivalently
$$\mathbb{E}[M_{t\wedge T} + 0.1(t\wedge T)|R_0 = 3] = \mathbb{E}[M_{t\wedge T}|M_0 = 3] + 0.1\mathbb{E}[t\wedge T|R_0 = 3] = 3.$$
- Want to take limit as $t \to \infty$ inside mean.
- Since $M_{t\wedge T} \in [0, 10]$, can use bounded convergence theorem there.
- Since $t \wedge T$ is a monotonically increasing random variable, can use monotonic convergence theorem there.
- Using $\mathbb{P}(T < \infty) = 1$,
$$\lim_{t\to\infty} M_{t\wedge T} = M_T, \ \lim_{t\to\infty} t\wedge T = T,$$
so
$$\mathbb{E}[M_T|M_0 = 3] + 0.1\mathbb{E}[T|M_0 = 3] = 3.$$
- From earlier,
$$\begin{aligned}\mathbb{E}[M_T|M_0 = 3] = \frac{1-(0.55/0.45)^{-3}}{(0.55/0.45)^7-(0.55/0.45)^{-3}}(10)+ \\ 0\mathbb{P}(M_T = 0|M_0 = 3) + 0.1\mathbb{E}[T|M_0 = 3] = 3.\end{aligned}$$
$$\mathbb{E}[T|M_0 = 3] = (0.1)^{-1}\left[3 - 10\frac{1-(0.55/0.45)^{-3}}{(0.55/0.45)^7-(0.55/0.45)^{-3}}\right] \approx 17.17\ldots.$$

10 Uniform integrability

Question of the Day Is there a necessary and sufficient condition for

$$\mathbb{E}\left[\lim_{t\to\infty} X_t\right] = \lim_{t\to\infty} \mathbb{E}[X_t]?$$

Sufficient (but not necessary conditions)

- Two main theorems, (Lebesgue) dominated convergence theorem (DCT) (and it's special case, the bounded convergence theorem (BCT) and the monotonic convergence theorem.
- Both these give sufficient conditions for bringing the limit inside the expectation, but not necessary conditions.

Last time wanted $\mathbb{E}[M_T] = \mathbb{E}[M_0]$

- If $\mathbb{P}(T < \infty) = 1$, and $|M_{t\wedge T}| < B$ for all t:

$$\begin{aligned}
\mathbb{E}[M_{t\wedge T}] &= \mathbb{E}[M_0] \\
\lim_{t\to\infty} \mathbb{E}[M_{t\wedge T}] &= \mathbb{E}[M_0] \\
\mathbb{E}[\lim_{t\to\infty} \mathbb{E}[M_{t\wedge T}] &= \mathbb{E}[M_0] \quad \text{(by DCT)} \\
\mathbb{E}[M_T] &= \mathbb{E}[M_0] \text{ since } \mathbb{P}(T < \infty) = 1.
\end{aligned}$$

- Key was DCT allowed bringing limit inside.
- Under what conditions can you bring limits inside expectations?
- Remember our counterexample:

$$U \sim \mathsf{Unif}([0,1]), \quad X_n = n\mathbb{1}(U \le 1/n).$$

- Counterexample has small chance of being very large.

  1: DCT: $|X_i| \le Y$, $\mathbb{E}[Y] < \infty$, $X_i \to X$ w.p. 1.

  2: MCT: $0 \le X_1 \le X_2 \le \cdots$ w.p. 1.

- Recall $X_i \to X$ in probability means

$$(\forall \epsilon > 0)\left(\lim_{t\to\infty} \mathbb{P}(|X_t - X| > \epsilon) = 0\right).$$

The necessary and sufficient condition to bring limits inside expectation under convergence in probability we will call uniform integrability.

Theorem 5 (Uniformly integrable is equivalent to being able to swap limits and mean)
Suppose $X_t \to X$ in probability. Then $\lim_{t\to\infty} \mathbb{E}[X_t] = \mathbb{E}[X]$ if and only if the $\{X_t\}$ are uniformly integrable.

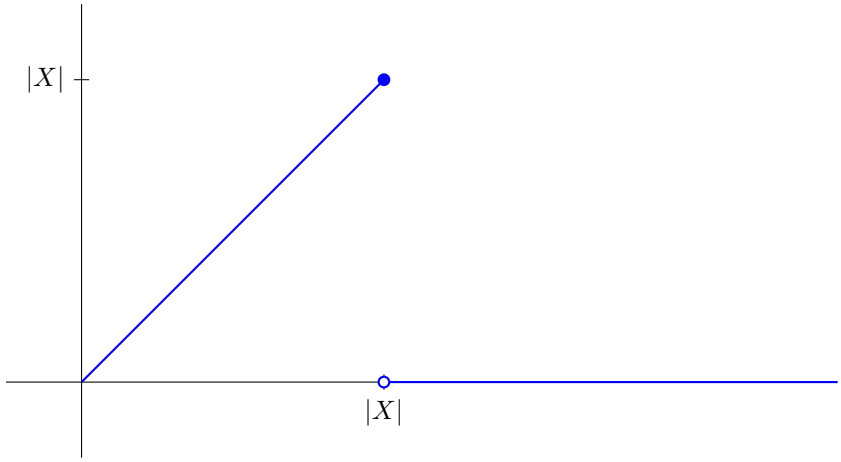
10.1 What is uniform integrability

Recall that a random variable X is integrable if $\mathbb{E}(|X|) < \infty$.

Fact 19
A random variable X is integrable if and only if

$$\lim_{B\to\infty} \mathbb{E}(|X|\mathbb{1}(|X| > B)) = 0.$$

Proof. ($\Rightarrow$) Assume $|X|$ is integrable. Then $\mathbb{P}(|X| < \infty) = 1$, so if I graph $f(B) = |X|\mathbb{1}(|X| > B)$, it looks like



So that means

$$\lim_{B\to\infty} |X|\mathbb{1}(|X| > B) = 0.$$

For all $B > 0$, $|X|\mathbb{1}(|X| > B) \leq |X|$ which is integrable, so we can use the DCT to say:

$$\lim_{B\to\infty} \mathbb{E}[|X|\mathbb{1}(|X| > B)] = \mathbb{E}\left[\lim_{B\to\infty} |X|\mathbb{1}(|X| > B)\right] = \mathbb{E}[0] = 0.$$

($\Leftarrow$) Suppose $\lim_{B\to\infty} \mathbb{E}(|X|\mathbb{1}(|X| > B) = 0$. Then

$$\mathbb{E}[|X|] = \mathbb{E}(|X|\mathbb{1}(|X| > B) + \mathbb{E}(|X|\mathbb{1}(|X| \leq B))$$

The first term on the RHS goes to 0 as $B \to \infty$, so for large enough B, it is at most 1. Note $\max_{|X|} |X|\mathbb{1}(|X| \leq B) = B$, so the RHS is at most $1 + B < \infty$, and the proof is complete. □

To move from a single random variable X to a stochastic process $\{X_t\}$, replace the stuff inside the limit with the supremum over all the variables in the process.

Definition 33
A stochastic process $\{X_t\}$ is **uniformly integrable** if

$$\lim_{B\to\infty} \left(\sup_t \mathbb{E}(|X_t|\mathbb{1}(|X_t| > B))\right) = 0.$$

10.2 Sufficient conditions for uniform integrability

Fact 20
Two conditions that imply $\{X_n\}$ is uniformly integrable are:

- Boundedness:
$$(\exists M)(\forall n)(|X_n| \leq M),$$
- Dominated by integrable r.v.:
$$(\exists Y : \mathbb{E}[|Y|] < \infty)(\forall n)(\mathbb{P}(X_n \leq Y) = 1).$$

Process is bounded

- Suppose $X_i \in [0, 10]$ for all i.
- Then $\sup_i \mathbb{E}[X_i \mathbb{1}(X_i > 11)] = \sup_i \mathbb{E}[0] = 0$.
- Generally, $|X_i|$ bounded above means $\{X_i\}$ are u.i.

Process is dominated by integrable random variable

- From the dominated convergence theorem, if $\lim_{n\to\infty} X_n = X$, and $|X_n| \leq Y$ with probability 1 where $\mathbb{E}[Y] < \infty$, then $\lim \mathbb{E}[X_n] = \mathbb{E}[\lim X_n]$. So the $\{X_n\}$ must be uniformly integrable.

A process that is not u.i

- X_i converges to X in probability means

$$\mathbb{E}[X_i] \to \mathbb{E}[X]$$

  if and only if X_i u.i.
- Now let $U \sim \mathsf{Unif}([0, 1])$, $Y_n = n\mathbb{1}(U < 1/n)$.
- So $Y_n \in \{0, n\}$.
- For $n > B$, $\mathbb{E}[Y_n \mathbb{1}(Y_n > B)] = 0(1 - 1/n) + n(1/n) = 1$.
- So $\mathbb{E}[Y_n] = 1$ for all n, so $\sup_n \mathbb{E}[Y_n] = 1 \neq 0$
- Does not converge to 0, so Y_n not u.i.
- Note $Y_n \to 0$ w.p. 1, and so also in probability.
- And $\lim_{n\to\infty} \mathbb{E}[Y_n] = 1$, but $\mathbb{E}[\lim_{n\to\infty} Y_n] = 0$.

10.3 Proof that uniform integrability allows swapping mean and limits

Proof u.i.+convergence in prob. allows swapping limits and mean. Suppose $X_t \to X$ in probability where the X_t are uniformly integrable. By properties of limits and expectations:

$$\lim \mathbb{E}(X_n) = \mathbb{E}(X) \Leftrightarrow \lim \mathbb{E}(X_n - X) = 0 \Leftrightarrow \lim \mathbb{E}(|X_n - X|) = 0.$$

Fix $\epsilon > 0$. Then there exists B such that for all $b \geq B$,

$$\sup_t \mathbb{E}[|X_t| \mathbb{1}(X_t > b)] < \min\{\epsilon/3, 1\}.$$

Let ϕ_M be the function that rounds x down to M if $x > M$, and up to $-M$ if $x < -M$. That is:

$$\phi_M(x) = M \cdot \mathbb{1}(x \geq M) + x \cdot \mathbb{1}(-M < x < M) - M \cdot \mathbb{1}(x \leq -M).$$

Then

$$X_n - X = (X_n - \phi_B(X_n)) + (\phi_B(X_n) - \phi_B(X)) + (\phi_B(X) - X).$$

Taking absolute values and means, and using the tri. ineq. gives...

$$\mathbb{E}|X_n - X| \leq \underbrace{\mathbb{E}|X_n - \phi_B(X_n)|}_{\text{term 1}} + \underbrace{\mathbb{E}|\phi_B(X_n) - \phi_B(X)|}_{\text{term 2}} + \underbrace{\mathbb{E}|\phi_B(X) - X|}_{\text{term 3}}.$$

The goal then becomes: show that each of these three terms is at most $\epsilon/3$ for large enough n and B.

The second term is easiest: since $\phi_B(X_n) - \phi_B(X)$ is between $-2M$ and $2M$, the bounded convergence theorem gives that this converges to 0 as $n \to \infty$. Hence there is an N such that $\mathbb{E}|\phi_B(X_n) - \phi_B(X)| \leq \epsilon/3$ for all $n \geq N$.

Uniform integrability also deals easily with the first term. When $X_n \in [-B, B]$, $\mathbb{1}(|X_n| > B) = 0$ and $X_n - \phi_B(X_n) = 0$. When $|X_n| > B$, then $|X_n - \phi_B(X_n)| = |X_n| - B \leq |X_n|$. Hence

$$|X_n - \phi_B(X_n)| \leq |X_n| \mathbb{1}(|X_n| > B).$$

So uniform integrability gives $\mathbb{E}|X_n - \phi_B(X_n)| \leq \epsilon/3$.

To show the third term converges to zero, the subgoal is to show that $\mathbb{E}[|X|] < \infty$, since $|\phi_B(X) - X| \leq |X| + B$, and so DCT could be used to bring the limit inside the mean.

Since $|X_n|$ and $|X| \geq 0$, Fatou's lemma says that

$$\liminf \mathbb{E}|X_n| \geq \mathbb{E}[\liminf |X_n|] = \mathbb{E}|X|.$$

So our new goal is to show $\liminf \mathbb{E}[|X_n|]$ is finite.

By u.i:

$$|X_n| = |X_n| \mathbb{1}(|X_n| > B) + |X_n| \mathbb{1}(|X_n| \leq B)$$

and $|X_n| \mathbb{1}(|X_n| \leq B) \leq B$, so

$$\sup_n \mathbb{E}|X_n| \leq \left[\sup_n \mathbb{E}(|X_n| \mathbb{1}(X_n > B) \right] + B \leq 1 + B.$$

So

$$\mathbb{E}[|X|] \leq \liminf \mathbb{E}|X_n| \leq \sup_n \mathbb{E}|X_n| \leq 1 + B.$$

Whew! So $\mathbb{E}|X| < \infty$, which means the dominated convergence theorem can be used to say $\mathbb{E}|\phi_B(X) - X| \to 0$ as $B \to \infty$.

Hence there is a $B \geq b$ such that for all $n \geq N$,

$$\mathbb{E}|X_n - X| \leq \epsilon/3 + \epsilon/3 + \epsilon/3 \leq \epsilon.$$

□

Proof convergence in prob.+limits and mean swap gives u.i. Suppose $X_n \to X$ in probability, and $\mathbb{E}[X_n] \to \mathbb{E}[X]$. Then $\mathbb{E}|X_n| \to \mathbb{E}|X|$ as well, and by Jensen's inequality:

$$|\mathbb{E}(X_n - X)| \leq \mathbb{E}|X_n - X| \leq \mathbb{E}|X_n| - \mathbb{E}|X| \to 0,$$

so $\mathbb{E}|X_n| \to \mathbb{E}|X|$.

So for some n large enough $|\mathbb{E}|X_n| - \mathbb{E}|X|| \leq \epsilon/3$.

Now create a function ϕ that is the identity on $[0, M-1]$, 0 on $[M, \infty)$, and linearly interpolates in between on $(M-1, M)$. So

$$\psi_M(x) = x \mathbb{1}(x \in [0, M-1]) + ((M-1)(M-x)) \mathbb{1}(x \in (M-1, M)).$$

Then as $M \to \infty$, $\psi_M(|X|) \to |X|$, so by DCT $\mathbb{E}[\psi_M(|X|)] \to \mathbb{E}[|X|]$, and for some M large enough

$$\mathbb{E}|X| - \mathbb{E}\psi_M(|X|) \leq \epsilon/3.$$

For any fixed M, the bounded convergence theorem gives

$$\mathbb{E}\psi_M(|X_n|) \to \mathbb{E}\psi_M(|X|),$$

so for some n large enough

$$|\mathbb{E}\psi(|X|) - \mathbb{E}\psi_M(|X_n|)| \leq \epsilon/2.$$

Putting this together gives

$$\begin{aligned}
\mathbb{E}[|X_n| \mathbb{1}(|X_n| > M)] &\leq \mathbb{E}[|X_n| - \psi_M(|X_n|)] \\
&= \mathbb{E}|X_n| - \mathbb{E}|\psi_M(|X_n|)| \\
&\leq \mathbb{E}|X| + \epsilon/3 - \mathbb{E}|\psi_M(|X|)| + \epsilon/3 \\
&\leq \epsilon.
\end{aligned}$$

□

11 The Martingale Convergence Theorem

Last time we learned that a random variable X is integrable if and only if

$$\lim_{B\to\infty} \mathbb{E}[|X|\mathbb{1}(|X| > B)] = 0.$$

This idea gave us a way to characterize when a collection of random variables $\{X_\alpha\}$ are all integrable together, which we called uniform integrability:

$$\lim_{B\to\infty} \sup_\alpha \mathbb{E}[|X_\alpha|\mathbb{1}(|X_\alpha| > B)] = 0.$$

Question of the Day For a martingale $\{M_t\}$, when does $\lim_{t\to\infty} M_t$ exist?

Last time

- A single random variable X is integrable iff

$$\lim_{B\to\infty} \mathbb{E}[|X|\mathbb{1}(|X| > B)] = 0.$$

- A collection of random variables $\{X_\alpha\}$ is uniformly integrable iff

$$\lim_{B\to\infty} \sup_\alpha \mathbb{E}[|X_\alpha|\mathbb{1}(|X_\alpha| > B)] = 0.$$

Today

- A sufficient condition for $\lim_{t\to\infty} M_t$ to exist is that the martingale is uniformly integrable.

A martingale that does not converge

- Simple symmetric walk on the integers. Let $D_1, D_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Unif}(\{-1, 1\})$ and

$$X_t = \sum_{i=1}^{t} D_i$$

  (So X_t goes up one with probability 1/2, down 1 with probability 1/2.)

- Then X_t is a martingale.
- Note that X_t never converges as t goes to infinity!
- Always jumps by 1 at each step.

A martingale that does converge

- Now let

$$T = \inf\{t : X_t = -10 \text{ or } X_t = 5\}.$$

- Then $X_{t\wedge T}$ does converge as t goes to infinity.
- Let $X_\infty = \lim_{t\to\infty} X_t$.
- Then $X_\infty = -10$ with probability 1/3, 5 w/ prob. 2/3
- So stopping martingales converge to a random variable!
- What about a martingale that isn't a stopping martingale?

11.1 Polya's Urn

A martingales that converges in a more interesting way

- Polya's Urn
- Start with one red and one blue marble in an urn
- At each step, pick a marble uniformly from the urn
- Replace that marble, and add another marble of same color
- Example: red, red, blue, red results in 4 red marbles and 2 blue marbles
- Let M_n be the percentage of red marbles after n draws.
- So $M_0 = 1/2$ (one red out of two at start).
- $n+2$ marbles in urn after n draws.
- On $n+1$ draw, M_n chance of picking a red marble...

$$\begin{aligned}\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= M_n \frac{M_n(n+2)+1}{n+3} + (1-M_n)\frac{M_n(n+2)}{n+3} \\ &= \frac{M_n}{n+3} + \frac{M_n(n+2)}{n+3} = \frac{M_n(n+3)}{n+3} = M_n\end{aligned}$$

- (M_n bounded, measurable w.r.t $\mathcal{F}_n$) so M_n is a martingale!
- Does M_n converge to anything?

Distribution of M_n in Polya's Urn

- Let $N_n = M_n(n+2)$ be the # of red marbles after n draws
- $N_0 = 1$
- N_1 is 1 (w/ prob. 1/2) or 2 (w/ prob. 1/2)
- N_2 is 1 (w/ prob. $(1/2)(2/3) = 1/3$...
- ...or 2 (w/ prob. $(1/2)(1/3) + (1/2)(1/3) = 1/3$)...
- ...or 3 (w /prob. $(1/2)(2/3) = 1/3$)
- So $N_0 \sim \mathsf{Unif}(\{1\})$, $N_1 \sim \mathsf{Unif}(\{1,2\})$, $N_2 \sim \mathsf{Unif}(\{1,2,3\})$.
- Use induction to show $N_n \sim \mathsf{Unif}(\{1,2,\ldots,n+1\})$.
- As n goes to infinity,

$$\lim_{n\to\infty} N_n = N_\infty \sim \mathsf{Unif}([0,1]).$$

Theorem 6 (Martingale Convergence Theorem)
Let M_n be a uniformly integrable martingale. Then

$$M_\infty = \lim_{n\to\infty} M_n$$

exists with probability 1, and $\mathbb{E}[M_\infty|M_0] = M_0$.

Example application of Mart. Conv. Thm.

- Let X_t be simple symmetric random walk on the integers w/ $X_0 = 0$.
- Let $T = \inf\{t : X_t = -10 \text{ or } X_t = 5\}$.
- Then $\mathbb{P}(T < \infty) = 1$.
  - Since $X_{t\wedge T} \in \{-5, \ldots, 10\}$, $X_{t\wedge T}$ is u.i.
  - So $\lim_{t\to\infty} X_{t\wedge T}$ exists.
  - But when $t < T$, $X_{t\wedge T}$ changes by 1 at each step, so for $X_{t\wedge T}$ to converge, must have T finite with probability 1.

11.2 Proof of Martingale Convergence Theorem

Outline

- Show that if M_t is u.i, then for any interval $[a, b]$, the number of times that M_t crosses from at most a to at least b (called an upcrossing) is finite.
- If $\lim a_n$ does not exist, then $\liminf a_n < \limsup a_n$ and the sequence a_n has an infinite number of upcrossings from $\liminf a_n$ to $\limsup a_n$.

Lemma 1 (Upcrossing inequality)
Let $a < b$ be rational numbers, and M_t a u.i. martingale. Set

$$T_a = \inf\{t : M_t \le a\},\ T_b = \inf\{t > T_a : M_t \ge b\}.$$

(Moving from below a to above b is called an *upcrossing*.) Then $\mathbb{P}(T_b < \infty) < 1$.

Note that the set of pairs of rational numbers $a < b$ is countable, so with probability 1, a u.i. martingale M_t has a finite number of upcrossings for all rational numbers $a < b$.

Fact 21 (From real analysis)
Let $x_0, x_1, x_2, \ldots$ be a sequence of real numbers that does not converge to a real number or ∞ or $-\infty$. Then there exists rational numbers $a < b$ such that x_i upcrosses (a, b) infinitely often.

Proof. Some definitions from real analysis.

$$\limsup x_i = \lim_{n\to\infty} \sup_{i\ge n} x_i$$
$$\liminf x_i = \lim_{n\to\infty} \inf_{i\ge n} x_i$$

A useful real analysis fact is that $\lim x_i$ exists if and only if $\limsup x_i = \liminf x_i$.

Suppose $\limsup x_i < \liminf x_i$, then there must exist rational numbers a and b such that

$$\liminf x_i < a < b < \limsup x_i.$$

Since a and b are strictly inside $(\liminf x_i, \limsup x_i)$, there are an infinite number of x_i that are at least b, and an infinite number of x_i that are at most a. Hence x_i upcrosses (a, b) infinitely often. □

Proof of the Martingale Convergence Theorem. The last two lemmas give

$$\lim_{n\to\infty} M_n = M_\infty$$

exists with probability 1. Since M_t is a u.i. martingale

$$\mathbb{E}[M_\infty | M_0] = \mathbb{E}\left[\lim_{n\to\infty} M_n | M_0\right] = \lim_{n\to\infty} \mathbb{E}[M_n | M_0] = M_0.$$

□

12 The Optional Sampling Theorem

Question of the Day A bet of x dollars on red in American Roulette returns 2x dollars with probability 18/38, and 0 dollars with probability 20/38. Is there a betting scheme that guarantees I will win \$1 with probability 1?

Surprising answer

- YES
- The problem: you need an infinite amount of money to do it.
- (So u.i. does not apply!)

Martingale betting scheme

- Start by betting one dollar
- If you win, quit, you've won one dollar!
- Otherwise, double bet, and play again. Repeat until you win.

Analysis

- Suppose you win on fifth game
$$-1-2-4-8+16=1.$$
- Suppose you win on ith game
$$2^j-\sum_{j=1}^{i-1}2^{j-1}=2^j-\frac{2^j-2^0}{2-1}=1.$$
- If M_t is the money after t steps of betting, and T is the first time you win a game,
$$M_T=M_0+1 \text{ so } \mathbb{E}[M_T|M_0]\neq M_0.$$

Theorem 7 (Optional Sampling Theorem)
Suppose that $M_0, M_1, \ldots$ is a martingale and T is a stopping time with respect to $\{\mathcal{F}_n\}$. If $M_{T\wedge t}$ is uniformly integrable, then
$$\mathbb{E}[M_T|M_0]=M_0.$$

Proof. The u.i. of $M_{t\wedge T|}$ gives
$$M_0=\lim_{t\to\infty}\mathbb{E}[M_{T\wedge t}|M_0]=\mathbb{E}\left[\lim_{t\to\infty}M_{T\wedge t}|M_0\right]=\mathbb{E}[M_\infty|M_0].$$
where the last step comes from the Martingale Convergence Theorem.

If $T<\infty$, then $M_\infty=M_T$. If $T=\infty$, then $M_T=M_\infty$. Either way, $M_0=\mathbb{E}[M_T|M_0]$, and we are done. □

Supermartingales and submartingale

- So far we've focused on martingales, fair games.
- Some games are biased upwards or downwards.

Definition 34
A stochastic process $M_0, M_1, \ldots$ is a **submartingale** with respect to a filtration $\mathcal{F}_n$ if for all n:

1: M_n is measurable with respect to $\mathcal{F}_n$.

2: $\mathbb{E}[|M_n|] < \infty$

3: for all $n > 0$, $M_n \leq \mathbb{E}[M_{n+1}|\mathcal{F}_n]$.

Definition 35
A stochastic process $M_0, M_1, \ldots$ is a **supermartingale** with respect to a filtration $\mathcal{F}_n$ if for all n:

1: M_n is measurable with respect to $\mathcal{F}_n$.

2: $\mathbb{E}[|M_n|] < \infty$

3: for all $n > 0$, $M_n \geq \mathbb{E}[M_{n+1}|\mathcal{F}_n]$.

- Note: if M_t is a submartingale, $-M_t$ is a supermartingale. If M_t is both a submartingale and a supermartingale, it is just a martingale.
- Our two main theorems (Martingale Convergence Theorem and Optional Sampling Theorem) hold for sub and supermartingales with appropriate $\leq$ sign.

Theorem 8 (Martingale Convergence Theorem)
Let $\{M_n\}$ be a uniformly integrable [sub][super]martingale. Then $M_\infty = \lim_{n\to\infty} M_n$ exists with probability 1, and

$$\mathbb{E}[M_\infty|M_0] \begin{cases} \geq M_0 & \text{for submartingales} \\ = M_0 & \text{for martingales} \\ \leq M_0 & \text{for supermartingales} \end{cases}$$

Theorem 9 (Optional Sampling Theorem)
Suppose that $M_0, M_1, \ldots$ is a [sub][super]martingale and T is a stopping time with respect to $\{\mathcal{F}_n\}$. If $M_{T\wedge t}$ is uniformly integrable, then

$$\mathbb{E}[M_T|M_0] \begin{cases} \geq M_0 & \text{for submartingales} \\ = M_0 & \text{for martingales} \\ \leq M_0 & \text{for supermartingales} \end{cases}$$

An application

- Let M_t be the amount of money playing Roulette, betting on red.
- Given history of first t spins $\mathcal{F}_t$, let $f(\mathcal{F}_t)$ be the amount of money bet on $(t+1)$st spin.
- Then M_t is a supermartingale, since

$$\mathbb{E}[M_{t+1}|\mathcal{F}_t] = M_t + f(\mathcal{F}_n)(18/38) - (20/38)f(\mathcal{F}_n) \leq M_t.$$

- For $a < M_0 < b$, let

$$T = \inf\{t : M_t < a \text{ or } M_t > b\}.$$

- Then $M_{t\wedge T}$ is a u.i. supermartingale.
- So

$$\mathbb{E}[M_T|M_0] \leq M_0.$$

- No matter what betting scheme you use, there is no way to make money as long as you quit when you fall below a certain number or when you rise above a certain number.

Example Can you use the OST to show that simple symmetric random walk, which is $X_t = \sum_{i=1}^{t} D_i$ where $D_i \overset{\text{iid}}{\sim} \mathsf{Unif}(\{-1, 1\})$, reaches 1 in finite time with probability 1?

To answer this, consider $T_{a,1} = \inf\{t : X_t \in \{1, a\}\}$, and $T_1 = \inf\{t : X_t = 1\}$. The reason to use $T_{a,1}$ is that $X_{t \wedge T_{a,1}}$ is a bounded martingale, and so is uniformly integrable. That means that the Martingale Convergence Theorem can be used to say that $\lim_{t\to\infty} X_{t\wedge T_{a,1}} = X_\infty$ exists with probability 1, which is equivalent to saying that $T_{a,1}$ is finite with probability 1.

Hence $X_\infty = X_{T_{a,1}}$, and

$$\mathbb{E}[X_{T_{a,1}}] = \mathbb{E}[X_0] = 0,$$

where

$$\mathbb{E}[X_{T_{a,1}}] = (1 - \mathbb{P}(X_{T_{a,1}} = a)) \cdot 1 + \mathbb{P}(X_{T_{a,1}} = a)a.$$

Solving gives

$$\mathbb{P}(X_{T_{a,1}} = a) = 1/(1 - a).$$

Now, if $T_1 = \infty$, then $X_{T_{a,1}} = a$ since it must have reached a before reaching 1. So

$$(\forall a \in \{-1, -2, \ldots\})(\mathbb{P}(T_1 = \infty) \leq \mathbb{P}(X_{T_{a,1}} = a) = 1/(1 - a)).$$

The only number in $[0, 1]$ less than $1/(1 - a)$ for all negative integers a is 0, so $\mathbb{P}(T_1 = \infty) = 0$.

13 Markov chains and First step analysis

Question of the Day In a model for a queue, the queue length starts at 0. At each step, with probability 40% someone arrives to the queue, otherwise with probability 60% someone leaves the queue. If the queue already has 4 people in it, an arrival does not join the queue. Given the queue is empty, what is the expected time needed for the next time the queue is empty?

Main Idea A Markov chain is a stochastic process that has the memoryless property: given the entirety of the process $X_0, X_1, \ldots, X_t$, the distribution of X_{t+1} only depends on X_t, and not on the rest of the process.

Markov chains

- This model of a queue is an example of a *Markov chain*, a process $X_0, X_1, X_2, \ldots$ where the distribution of X_{t+1} only depends on X_t, and not on earlier values.
- The queue length at the next step only depends on the current queue length, not on past history.

Definition 36
A stochastic process $X_0, X_1, X_2, \ldots$ is a **Markov chain** with respect to filtration $\mathcal{F}_t$ if for all t,

1: X_t is $\mathcal{F}_t$ measurable and

2: $[X_{t+1}|\mathcal{F}_t] \sim [X_{t+1}|X_t]$.

Definition 37
A Markov chain $X_0, X_1, \ldots$ is **time-homogeneous** if for all t,

$$[X_{t+1}|X_t] \sim [X_1|X_0].$$

- Markov chains are often called *memoryless* processes because they can only remember the last state X_t, and forget $X_1, \ldots, X_{t-1}$ when generating X_{t+1}.
- For state space Ω, a Markov chain has the form:

$$X_{t+1} = f(X_t, U_{t+1}),$$

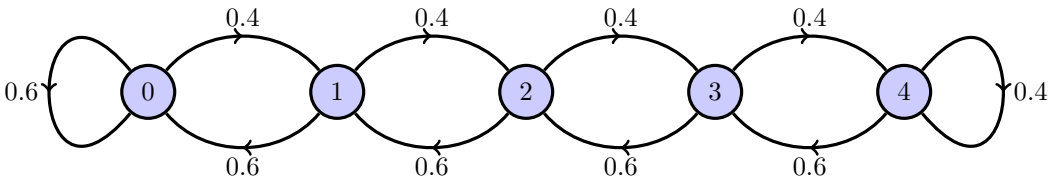
  where $f : \Omega \times [0, 1] \to \Omega$ is a deterministic function.
- Unless indicated otherwise, all Markov chains in this course will be time-homogeneous.

For the QotD

- The sequence of queue lengths could be:
  - 0,1,0
  - 0,0
  - 0,1,2,3,4,4,4,3,2,3,2,1,0
- Let $T = \inf\{t \geq 1 : X_t = 0\}$.
- What is $\mathbb{E}[T]$?

Representing Markov chains

- There are multiple ways to represent Markov chains.
- One way is graphically.
- (Directed) graphs consist of nodes (vertices) and edges (arcs) that run from a head node to a tail node.
- For a Markov chain, nodes represent states of the chain, edges marked with the probability of moving from one state to another.
- This is called a *transition graph*.
- For the QotD MC:



Back to the Question of the Day

- Need more variables to get to T.
- For $a \in \{0, 1, 2, 3, 4\}$, let
$$T_a = \inf\{t \geq 0 : X_t = 0 | X_0 = a\}.$$
- So $\mathbb{E}[T_3]$ is the number of steps needed to return to 0 starting at state 3.
- To understand T_3, do what is called first step analysis.
- Consider what happens at the first step of the chain.

$$\begin{aligned}\mathbb{E}[T_3] &= \mathbb{E}[\mathbb{E}[T_3|X_1]] \\ &= \mathbb{E}[T_3|X_1 = 2]\mathbb{P}(X_1 = 2) + \mathbb{E}[T_3|X_1 = 4]\mathbb{P}(X_1 = 4) \\ &= (1 + \mathbb{E}[T_2])(0.6) + (1 + \mathbb{E}[T_4])(0.4)\end{aligned}$$

- This gives us an equation for $\mathbb{E}[T_0], \ldots, \mathbb{E}[T_4]$.

$$\begin{aligned}\mathbb{E}[T_0] &= 0 \\ \mathbb{E}[T_1] &= (1 + \mathbb{E}[T_0])(0.6) + (1 + \mathbb{E}[T_2])(0.4) \\ \mathbb{E}[T_2] &= (1 + \mathbb{E}[T_1])(0.6) + (1 + \mathbb{E}[T_3])(0.4) \\ \mathbb{E}[T_3] &= (1 + \mathbb{E}[T_2])(0.6) + (1 + \mathbb{E}[T_4])(0.4) \\ \mathbb{E}[T_4] &= (1 + \mathbb{E}[T_3])(0.6) + (1 + \mathbb{E}[T_4])(0.4)\end{aligned}$$

- This can be written as a linear system. Let $w_i = \mathbb{E}[T_i]$,

$$\begin{aligned}w_0 &= 0 \\ w_1 &= 1 + 0.6w_0 + 0.4w_2 \\ w_2 &= 1 + 0.6w_1 + 0.4w_3 \\ w_3 &= 1 + 0.6w_2 + 0.4w_4 \\ w_4 &= 1 + 0.6w_3 + 0.4w_4,\end{aligned}$$

so

$$\begin{pmatrix}0\\1\\1\\1\\1\end{pmatrix} = \begin{pmatrix}1 & 0 & 0 & 0 & 0\\ -0.6 & 1 & -0.4 & 0 & 0\\ 0 & -0.6 & 1 & -0.4 & 0\\ 0 & 0 & -0.6 & 1 & -0.4\\ 0 & 0 & 0 & -0.6 & 0.6\end{pmatrix}\vec{w}$$

- Can use MATLAB/Mathematica/R/TI to solve systems of equations.
- Matrices in Wolfram Alpha/Mathematica
- To solve in Wolfram Alpha...

```
inverse{{1,0,0,0,0},{-0.6,1,-0.4,0,0},{0,-0.6,1,-0.4,0},
  {0,0,-0.6,1,-0.4},{0,0,0,-0.6,0.6}}*
  {{0},{1},{1},{1},{1}}
```

- Result (to four sig figs)

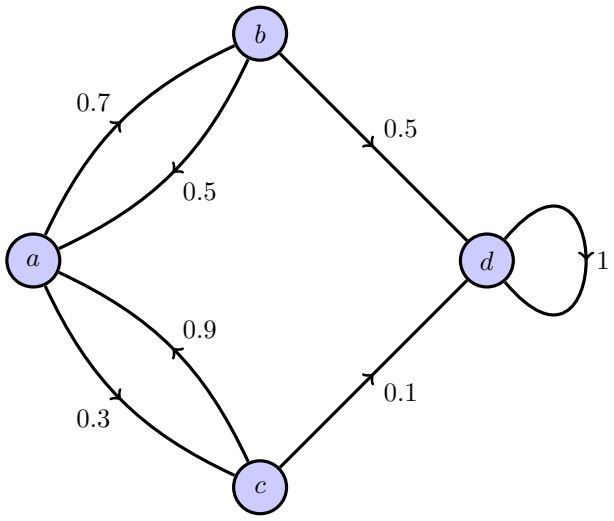
$$\begin{pmatrix}0\\4.012\\7.530\\10.30\\11.97\end{pmatrix}$$

Finishing the QotD

- So what is $\mathbb{E}[T]$?
- Use first step analysis...

$$\begin{aligned}\mathbb{E}[T] &= \mathbb{E}[\mathbb{E}[T|X_1]]\\ &= \mathbb{E}[T|X_1=1](0.4) + \mathbb{E}[T|X_1=0](0.6)\\ &= 0.4(1+w_1) + 0.6(1) = 2.60494 = \boxed{2.60494\ldots}.\end{aligned}$$

Example Consider the following 4-state Markov chain.



- Starting at a, what is the expected number of steps needed to get to d?
- First step analysis to the rescue!

- Let w_i be expected number of steps to reach d starting from i.

$$\begin{aligned} w_a &= 0.7(1 + w_b) + 0.3(1 + w_c) \\ w_b &= 0.5(1 + w_a) + 0.5(1 + w_d) \\ w_c &= 0.9(1 + w_a) + 0.1(1 + w_d) \\ w_d &= 0 \end{aligned}$$

```
solve a = 0.7(1 + b) + 0.3(1 + c) and b = 0.5(1 + a)
  + 0.5(1 + d) and c = 0.9(1 + a) + 0.1(1 + d) and
  d = 0
```

- Then $a = 100/19 = \boxed{5.263\ldots}$.

13.1 Example of a stochastic process that is not a Markov chain

Recall the Martingale betting scheme, where if we lose we double the bet, and if we win the bet returns to 1. So if the sequence of bets was

$$1, 1, 1, 2, 4, 8, 1, 1, 1, \ldots$$

then we know that we won the first two bets, lost the next three, then won the next three. (The 9th bet we wouldn't know if we won or lost until the next bet was revealed.) Suppose we started with $M_0 = 3$ dollars, then the $\{M_t\}$ sequence would be

$$3, 4, 5, 4, 2, -2, 6, 7, 8,$$

This is definitely not a Markov chain! For instance, if the sequence was

$$3, 2, 0, 4, 5,$$

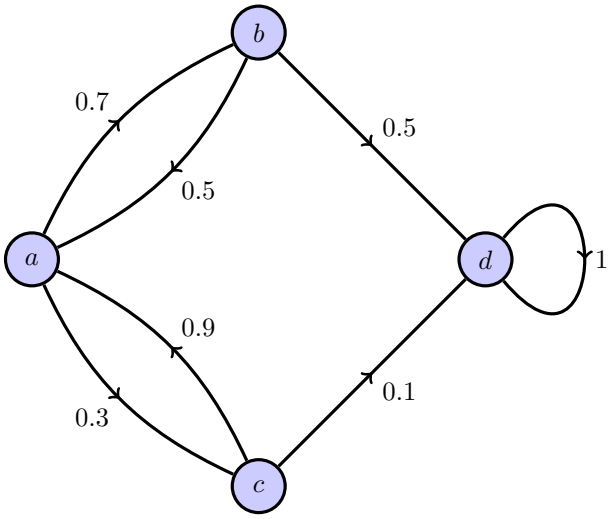
then we won the last bet, so the next bet would be 1 and $M_5 \in \{4, 6\}$. However, if the sequence was

$$3, 4, 5, 6, 5,$$

then we lost the last bet, but won the one before, so the next bet would be 2 and $M_t \in \{3, 7\}$. Hence the final state is not enough information to determine the distibution of the next state, so this is not a Markov chain.

14 Transition matrices and update functions

Question of the Day Consider the following Markov chain:



If it starts in state a, what is the chance after 5 steps that we are in state d?

Transitions

- Since the graph tells us the probability of transitioning from state to state, call this a *transition graph*.
- Can also create a *transition matrix*

Definition 38
Given a Markov chain with finite state space Ω, the **transition matrix** has entry in the ith row and jth column

$$a_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i).$$

For the transition graph in the Question of the Day:

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 0.7 & 0.3 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Taking one step in the Markov chain

- Suppose that $X_0 \sim \mathsf{Unif}(\{a, b, c, d\})$.
- What is the chance that $X_1 = d$?
- Can describe distribution of X_0 with a probability vector:

$$p_0 = \begin{array}{c} a \\ b \\ c \\ d \end{array} \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$$

- We can break down the event X_1 into different possibilities based on X_0:

$$\begin{aligned}\mathbb{P}(X_1 = a) &= \mathbb{P}(X_1 = a, X_0 = a) + \mathbb{P}(X_1 = a, X_0 = b) + \\ &\quad \mathbb{P}(X_1 = a, X_0 = c) + \mathbb{P}(X_1 = a, X_0 = d) \\ &= \mathbb{P}(X_1 = a|X_0 = a)\mathbb{P}(X_0 = a) + \mathbb{P}(X_1 = a|X_0 = b)\mathbb{P}(X_0 = b) + \\ &\quad \mathbb{P}(X_1 = a|X_0 = c)\mathbb{P}(X_0 = c) + \mathbb{P}(X_1 = a|X_0 = d)\mathbb{P}(X_0 = d) \\ &= (0)(1/4) + (0.5)(1/4) + (0.9)(1/4) + (0)(1/4) \\ &= 0.35\end{aligned}$$

- Of course we could also could write this same process using vector multiplication:

$$\mathbb{P}(X_1 = a) = \underbrace{\begin{pmatrix}1/4 & 1/4 & 1/4 & 1/4\end{pmatrix}}_{p_0} \quad \underbrace{\begin{pmatrix}0 \\ 0.5 \\ 0.9 \\ 0\end{pmatrix}}_{\text{first col of transition matrix}}$$

- That gives $\mathbb{P}(X_1 = a)$. To get $\mathbb{P}(X_1 = b), \mathbb{P}(X_1 = c), \mathbb{P}(X_1 = d)$, multiply by appropriate columns of transition matrix

- Or do it all at once with matrix multiplication!

$$\begin{pmatrix}1/4 & 1/4 & 1/4 & 1/4\end{pmatrix} \begin{pmatrix}0 & 0.7 & 0.3 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}0.35 & 0.175 & 0.075 & 0.4\end{pmatrix}$$

- So X_0 has prob. vector p_0, and X_1 has prob. vector p_1:

$$\underbrace{\begin{pmatrix}1/4 & 1/4 & 1/4 & 1/4\end{pmatrix}}_{p_0}, \quad \underbrace{\begin{pmatrix}0.35 & 0.175 & 0.0175 & 0.4\end{pmatrix}}_{p_1}$$

- Let A denote the transition matrix, then

$$p_1 = p_0 A.$$

- Note: multiplication on the left very important!

Taking multiple steps

- There was nothing special about X_0 to X_1 in last example.

- To move from prob. vector for X_5 to X_6, do something similar:

$$p_6 = p_5 A.$$

- Now consider X_0 to X_2. First go to X_1, then go to X_2:

$$p_2 = p_1 A = (p_0 A)A = p_0(A \cdot A) = p_0 A^2.$$

- Matrix multiplication is not commutative, it is associative

- Can't change order, can change parenthesis around.

QotD

- Want to take five steps in Markov chain:

$$p_5 = p_4 A = p_3 A^2 = p_2 A^3 = p_1 A^4 = p_0 A^5.$$

- For QotD, start in state a, so prob. vector is

$$p_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix},$$

and

$$p_0 A^5 = \begin{pmatrix} 0 & 0.26908 & 0.11532 & 0.6156 \end{pmatrix}.$$

- There is a $\boxed{61.56\%}$ chance of being in state d five steps after starting in state a.

Fact 22
If $X_0 \sim p$ for a Markov chain with transition matrix A, then for $t \geq 0$,

$$X_t \sim pA^t.$$

Taking many steps

- Another way to view this: If A is the transition matrix for 1 step, then A^t is the transition matrix for t steps.
- Back to A from QotD:

$$A^{10} \approx \begin{pmatrix} 0.09 & 0 & 0 & 0.9 \\ 0 & 0.5 & 0.02 & 0.92 \\ 0 & 0.09 & 0.03 & 0.86 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

last line means if you start in d, stay in d after 10 steps.

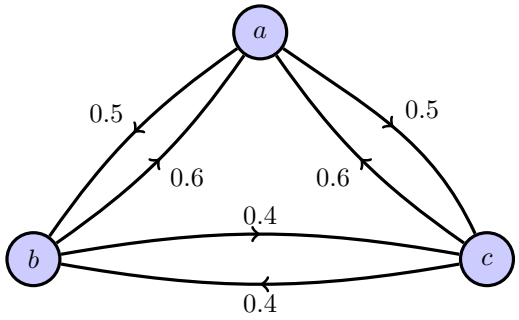
- After 100 steps:

$$A^{100} \approx \begin{pmatrix} 4 \cdot 10^{-11} & 0 & 0 & 1 \\ 0 & 2 \cdot 10^{-11} & 1 \cdot 10^{-11} & 1 \\ 0 & 4 \cdot 10^{-11} & 2 \cdot 10^{-11} & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- So with probability very close to 1, now in state d, no matter where you start.
- Call a, b, c *transient* states because they only occur a finite number of times, whereas d is *recurrent* because it occurs infinitely often (wp1)

Another example of taking multiple steps in the Markov chain

- Suppose that we have a 3-state Markov chain:



$$A = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.6 & 0 & 0.4 \\ 0.6 & 0.4 & 0 \end{pmatrix}$$

- Then after 100 steps

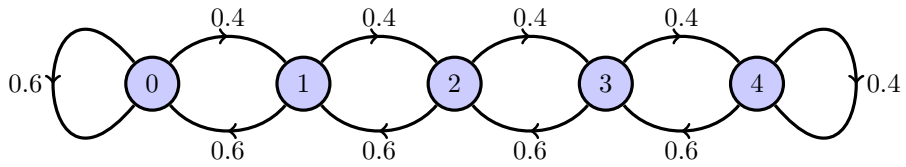
$$A^{100} \approx \begin{pmatrix} 0.375 & 0.3125 & 0.3125 \\ 0.375 & 0.3125 & 0.3125 \\ 0.375 & 0.3125 & 0.3125 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} A^{100} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} A^{100} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} A^{100} = \begin{pmatrix} 0.375 & 0.3125 & 0.3125 \end{pmatrix}$$

- So $p_0 A^{100} = \begin{pmatrix} 0.375 & 0.3125 & 0.3125 \end{pmatrix}$ for all prob. vectors p_0!
- This behavior is called *ergodicity* and we will discuss it in much more detail in the next section.

15 Limiting and Stationary distributions

Question of the Day For the queue model



what is the long term average # waiting in the queue?

Start with transition matrix

$$A = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 & 0 \\ 0 & 0.6 & 0 & 0.4 & 0 \\ 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0.6 & 0.4 \end{pmatrix}$$

(In Mathematica/WolframAlpha

```
{{0.6,0.4,0,0,0},{0.6,0,0.4,0,0},{0,0.6,0,0.4,0},
  {0,0,0.6,0,0.4},{0,0,0,0.6,0.4}}
```

So

$$A^{100} = \begin{pmatrix} 0.383886 & 0.255924 & 0.170616 & 0.113744 & 0.0758294 \\ 0.383886 & 0.255924 & 0.170616 & 0.113744 & 0.0758294 \\ 0.383886 & 0.255924 & 0.170616 & 0.113744 & 0.0758294 \\ 0.383886 & 0.255924 & 0.170616 & 0.113744 & 0.0758294 \\ 0.383886 & 0.255924 & 0.170616 & 0.113744 & 0.0758294 \end{pmatrix}$$

Note: rows of A^t always add to 1, since they are probabilities.

Definition 39
A matrix is **stochastic** (or more specifically *row stochastic*) if the sum of the entries of each row is 1.

Fact 23
Transition matrices and their powers are stochastic matrices.

15.1 Limiting distribution

We say that a Markov chain has a *limiting distribution* if after a large number of steps, the Markov chain "forgets" where it started at. Recall: the (i, j)th entry of A^{100} is

$$\mathbb{P}(X_{100} = j | X_0 = i).$$

Definition 40
π is a **limiting distribution** for a Markov chain if for all $x_0 \in \Omega$, and $A \in \mathcal{F}$,

$$\lim_{t\to\infty} \mathbb{P}(X_t \in A | X_0 = x_0) = \pi(A).$$

Note that if Ω is a discrete set $\{i_1, i_2, \ldots\}$, then this is the same as saying that for all $n, m \in \{1, 2, 3, \ldots\}$,

$$\lim_{t\to\infty} \mathbb{P}(X_t = i_n | X_0 = i_m) = \pi(\{i_n\}).$$

Eigenvalues and eigenvector

- Recall that if for matrix A and vector v,
$$vA = \lambda v$$
- Then λ is an eigenvalue (the German prefix eigen means "the same" here) and v is a left eigenvector because it is multiplying the matrix on the left.
- Note that left eigenvectors of A are right eigenvectors of A^T (read A transpose).
- Using

  ```
  eigenvector of transpose{{0.6,0.4,0,0,0},{0.6,0,0.4,0,0},
  {0,0.6,0,0.4,0},{0,0,0.6,0,0.4},{0,0,0,0.6,0.4}}
  ```

  gives the first eigenvalue of 1 with eigenvector
$$v_1 \approx \begin{pmatrix} -0.75194 & -0.501269 & -0.33418 & -0.222786 & -0.148524 \end{pmatrix}$$
- Linear algebra fact: if v is an eigenvector, so is Cv for any constant C:
$$CvA = C(vA) = C(\lambda v) = \lambda(Cv).$$
- So I can normalize v_1 by dividing by its sum to get:
$$\begin{aligned} v_1' &= \frac{v_1}{-0.751904 - 0.501269 - 0.33418 - 0.222786 - 0.148524} \\ &= \begin{pmatrix} 0.383886, 0.255924, 0.170616, 0.113744, 0.0758293 \end{pmatrix} \end{aligned}$$
- This should look familiar: it's the same as the limiting distribution!
- The eigenvalue is 1, so
$$v_1'A = v_1'$$
- Hence if $X_t \sim v_1'$, then $X_{t+1} \sim v_1'$.

15.2 stationary distributions

Formally, when you start in a distribution π, and after one step of the Markov chain you are still in distribution π, then we call the distribution *stationary*. Note that the state of the chain can be changing from step to step, it is only the distribution of the state that is unchanging.

Definition 41
Distribution π is a **stationary distribution** for a Markov chain if $X_t \sim \pi$ implies that $X_{t+1} \sim \pi$. (When π is given in the form of a probability vector, say that π is a stationary probability vector.)

Another example Suppose
$$A = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0 & 0.4 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$
Then
$$A^{100} = \begin{pmatrix} 0.313725 & 0.254902 & 0.431373 \\ 0.313725 & 0.254902 & 0.431373 \\ 0.313725 & 0.254902 & 0.431373 \end{pmatrix}$$
So $v = \begin{pmatrix} 0.313725 & 0.254902 & 0.431373 \end{pmatrix}$ is a limiting distribution.
Also
$$vA = v$$
So this is a stationary distribution as well!

Limiting distributions are always stationary distributions

Fact 24
Suppose for transition matrix A there is a probability vector π such that for all probability vectors v,

$$\lim_{t\to\infty} vA^t = \pi.$$

Then π is a stationary distribution probability vector.

Proof. Recall that you can bring linear operators in and out of limits:

$$\lim_{n\to\infty}(c_1 a_n + c_2 b_n) = c_1 \lim a_n + c_2 \lim b_n.$$

Hence

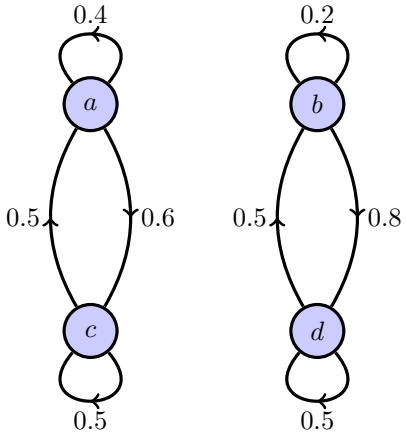
$$\begin{aligned}\pi A &= \left[\lim_{t\to\infty} vA^t\right] A = \lim_{t\to\infty} vA^t A \\ &= \lim_{t\to\infty} vA^{t+1} \\ &= \pi,\end{aligned}$$

making π a stationary probability vector. □

Sadly, the reverse is not true.

An example where a stationary distribution is not a limiting distribution

$$A = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0.4 & 0 & 0.6 & 0 \\ 0 & 0.2 & 0 & 0.8 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \end{pmatrix} \end{array}$$



- Start in a, stay in $\{a, c\}$
- Start in b, stay in $\{b, d\}$

$$A^{1000} = \begin{pmatrix} 0.4545 & 0 & 0.5455 & 0 \\ 0 & 0.3846 & 0 & 0.6154 \\ 0.4545 & 0 & 0.5455 & 0 \\ 0 & 0.3846 & 0 & 0.6154 \end{pmatrix}$$

- Restricted to $\{a, c\}$ there is a limiting distribution
- Restricted to $\{b, d\}$ there is a limiting distribution
- But no overall limiting distribution
- There are an infinite # of stationary distributions!

$$\begin{aligned}\pi_1 &= \begin{pmatrix} 5/11 & 0 & 6/11 & 0 \end{pmatrix} \\ \pi_2 &= \begin{pmatrix} 0 & 5/13 & 0 & 8/13 \end{pmatrix}\end{aligned}$$

- For any $\lambda \in [0, 1]$,

$$\pi_\lambda = \underbrace{\lambda\pi_1 + (1-\lambda)\pi_2}_{\text{convex linear combination}}$$

  is also stationary.

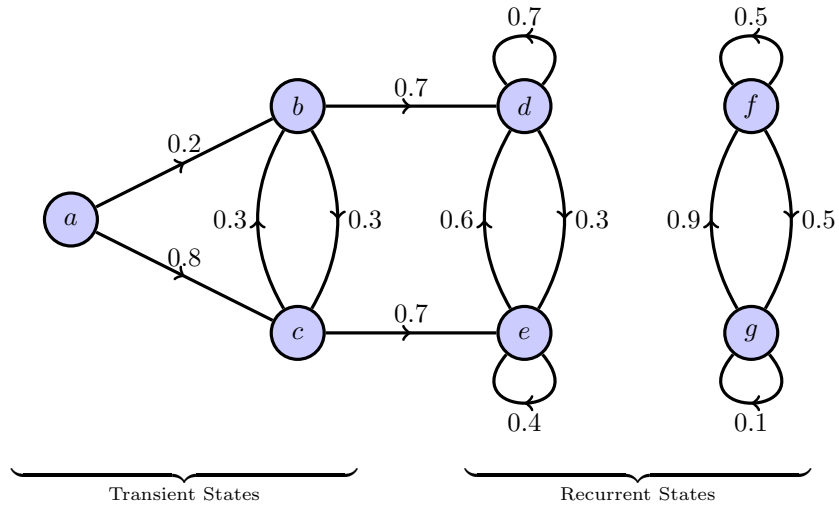
- Example:

$$\pi_{0.2} = \frac{1}{1430}\begin{pmatrix}130 & 440 & 156 & 704\end{pmatrix}$$
$$\pi_{0.5} = \frac{1}{1430}\begin{pmatrix}325 & 275 & 390 & 440\end{pmatrix}$$
$$\pi_{0.7} = \frac{1}{1430}\begin{pmatrix}455 & 165 & 546 & 264\end{pmatrix}$$

- A Markov chain with a transition graph with two or more strongly connected components is *reducible*.

16 Recurrent and Transient States

Question of the Day Consider the following Markov chain. For which states is the probability of returning to that state equal to 1?



A Markov chain consists of two types of states: recurrent and transient. A state is recurrent if the chain keeps returning to the state over and over again with probability 1. If a state is not recurrent, then it is transient.

Definition 42
Let $R_x = \inf\{t > 0 : X_t = x | X_0 = x\}$ be the **return time** for state x.

For example:

- Sequence 01210 has return time 4
- Sequence 00 has return time 1

Note that the smallest a return time can be is 1!

Definition 43
A state x is **transient** if $\mathbb{P}(R_x < \infty) < 1$. A state x is **recurrent** if $\mathbb{P}(R_x < \infty) = 1$.

Example

$$\begin{aligned}
&\mathbb{P}(R_b < \infty) < 0.3 \text{ if goes to } d, \text{ cannot return to } b \\
&\mathbb{P}(R_d < \infty) = \mathbb{P}(R_b = 1) + \mathbb{P}(R_b = 2) + \cdots \\
&\qquad = (0.7) + (0.3)(0.6) + (0.3)(0.4)(0.6) + (0.3)(0.4)(0.4)(0.6) + \cdots \\
&\qquad = (0.7) + (0.3)(0.6)\left[1 + (0.4) + (0.4)^2 + (0.4)^3 + \cdots\right] \\
&\qquad = (0.7) + (0.3)(0.6)\frac{1}{1 - 0.4} \\
&\qquad = (0.7) + (0.3) = 1.
\end{aligned}$$

Definition 44
States x and y of a Markov chain **communicate** if $\exists n, m \in \{0, 1, 2 \ldots\}$ such that $\mathbb{P}(X_n = y | X_0 = x) > 0$ and $\mathbb{P}(X_m = x | X_0 = y) > 0$. Write $x \leftrightarrow y$.

So $x \leftrightarrow y$ if there is a directed path (possibly of zero length) both from x to y and from y to x in the transition graph using positive probability edges.

Fact 25
Communication ($\leftrightarrow$) is an equivalence relation.

1: Reflexive: $x \leftrightarrow x$.

2: Symmetric: $x \leftrightarrow y$ implies $y \leftrightarrow x$.

3: Transitive: $x \leftrightarrow y$ and $y \leftrightarrow z$ implies $x \leftrightarrow z$.

Definition 45
If $x \leftrightarrow y$, say that x and y are in the same **communication class**.

The next fact gives us a simpler way to establish if a state is recurrent or transient.

Fact 26
If state i is in communication class C, the state is recurrent if and only if no edges with positive probability leave C.

To simplify writing probabilities, let

$$p(i,j) = \mathbb{P}(X_{t+1} = j | X_t = i)$$

be the chance of moving from i to j in one step of the Markov chain.

Proof. Saying that a recurrent communication class has no outgoing edges is equivalent to contrapositive statement that a communication class with an outgoing edge is transient.

Let C be a communication class with edge (i,j) where $i \in C$, $j \notin C$, $p(i,j) > 0$. Since i can reach j, but j isn't in the same communication class, so j cannot reach i. Hence $\mathbb{P}(R_i < \infty) \leq 1 - p(i,j) < 1$. So the communication class is not recurrent.

Now suppose that state i is in a communication class C with no outgoing edges. Let $n = \#C$.

Let j be any state in C reachable from i. Then since C is a communication class, there is an integer m_j such that $\mathbb{P}(X_{m_j} = i | X_0 = j) > 0$. Let $M = \max\{m_j\}$.

So no consider first taking a step in the chain. If we are back at i, then $R_i = 1$. Otherwise, after M steps there is a positive chance that we will have returned to i at least once. Let

$$\alpha = \min\{\mathbb{P}(i \in \{X_1, X_2, \ldots, X_M\} | X_0 = j)\}.$$

Then the probability we haven't returned to i in $1+M$ steps is at most $1-\alpha$. The probability we haven't returned to i in $1+2M$ steps is at most $(1-\alpha)^2$. And in general, the probability we haven't returned to i in $1+kM$ steps is $(1-\alpha)^k$. Since that goes to 0 as $k \to \infty$, $\mathbb{P}(R_i < \infty) = 1$. □

Using this fact allows us to classify the states as follows:

1: First write down the communication classes.

2: Elements of classes with no outgoing edges are recurrent, the rest are transient.

In the QotD chain:

$$\text{Communication classes} = \{\{a\}, \{b,c\}, \{d,e\}, \{f,g\}\}$$
$$\text{transient} = \{a,b,c\}, \quad \text{recurrent} = \{d,e,f,g\}.$$

Fact 27
If any element of a class is recurrent, they all are. So call classes *recurrent* if they have a recurrent element, and transient otherwise.

Proof. Follows directly from the previous fact: When x is recurrent, the class it is in has no outgoing edges, so every element in the class is in a class with no outgoing edges, and so is recurrent. □

In QotD chain

Classes	Type
$\{1\}$	transient
$\{2,3\}$	transient
$\{4,5\}$	recurrent
$\{6,7\}$	recurrent

Fact 28
If state k is recurrent in a finite state Markov chain, then $\mathbb{E}[R_k] < \infty$.

Proof. Since k is recurrent, there are no outgoing edges from its communication class, which is of size at most $\#(\Omega)$. Hence after $\#(\Omega)$ steps, there is an $\alpha > 0$ chance of returning to k at least once.

Let $n = \#(\Omega)$. Then $R_k \in \{0, 1, 2, 3, \ldots\}$, so can use the tail sum formula:

$$\begin{aligned}\mathbb{E}[R_k] &= \sum_{i=0}^{\infty} \mathbb{P}(R_k > i) \\ &= \mathbb{P}(R_k > 0) + \cdots + \mathbb{P}(R_k > n-1) + \\ &\quad \mathbb{P}(R_k > n) + \cdots + \mathbb{P}(R_k > 2n-1) + \\ &\quad \mathbb{P}(R_k > 2n) + \cdots + \mathbb{P}(R_k > 3n-1) + \cdots\end{aligned}$$

Now $\mathbb{P}(R_k > n) \leq 1 - \alpha$, $\mathbb{P}(R_k > 2n) \leq (1-\alpha)^2$ and so on, which gives:

$$\begin{aligned}\mathbb{E}[R_k] &\leq 1 + 1 + \cdots + 1 + \\ &\quad (1-\alpha) + (1-\alpha) + \cdots + (1-\alpha) + \\ &\quad (1-\alpha)^2 + (1-\alpha)^2 + \cdots + (1-\alpha)^2 + \cdots \\ &= n + n(1-\alpha) + n(1-\alpha)^2 + \cdots \\ &= \frac{n}{1-(1-\alpha)} = n\alpha^{-1} < \infty.\end{aligned}$$

□

For QotD, can find these exactly:

$$\begin{aligned}\mathbb{E}(R_d) &= \mathbb{E}(\mathbb{E}(R_d|X_1)) \\ &= \mathbb{E}(R_d|X_1 = d)\mathbb{P}(X_1 = d) + \mathbb{E}(R_d|X_1 = e)\mathbb{P}(X_1 = e) \\ &= (1)(0.7) + (1 + (1/0.6))(0.3) = \boxed{1.500},\end{aligned}$$

where $1/0.6$ is the mean of a geometric random variable with parameter 0.6.

Fact 29
For a transient state y, for all starting states x, $\lim_{t\to\infty} \mathbb{P}(X_t = y|X_0 = x) = 0$.

Proof. Fix y, and call its transient communication class C. Note that once the state leaves C, it can never return. Since C is transient, it has at least one outgoing edge, call it (a, b).

When the state is in C, after $\#(\Omega)$ steps, there is a positive chance α of moving across the edge (a, b), never to return. Let k be a positive integer, then

$$\mathbb{P}(X_{k\#(\Omega)} \in C|X_0 \in C) \leq (1-\alpha)^k.$$

Therefore in the limit as the number of steps goes to infinity, the chance of staying in C goes to 0. □

Fact 30
Every finite state Markov chain has at least one recurrent state.

Proof. Fix $x \in \Omega$. Then

$$1 = \lim_{t\to\infty} \mathbb{P}(X_t \in \Omega | X_0 = x) = \sum_{y\in\Omega} \lim_{t\to\infty} \mathbb{P}(X_t = y | X_0 = x),$$

so there must be at least one $y \in \Omega$ with $\lim_{t\to\infty} \mathbb{P}(X_t = y|X_0 = x) > 0$. That state y is recurrent. □

17 Building a stationary measure for a Markov chain

Question of the Day Do all finite state Markov chains have stationary distributions?

Today

- Learn about stationary measures.
- Give stationary measures for recurrent Markov chains using return times
- Use this to get stat. dist. for all finite state Markov chains

Fact 31
All finite state Markov chains have at least one stationary distribution.

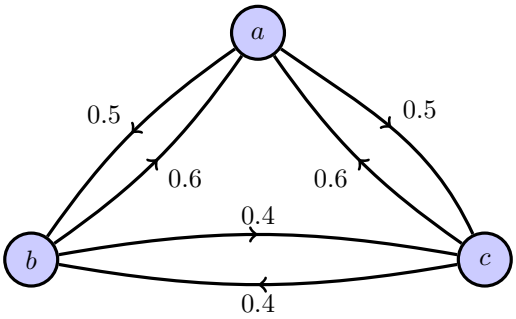
[Recall: π stationary means $X_t \sim \pi \Rightarrow X_{t+1} \sim \pi$.]

17.1 The stationary measure

Definition 46
Say that μ is a **stationary measure** for a discrete Markov chain if for all i

$$\mu(i) = \sum_{j \in \Omega} \mu(j) \mathbb{P}(X_1 = i | X_0 = j).$$

Note: it is not possible to use matrix notation anymore because Ω might be countably infinite!
Example:



$$A = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.6 & 0 & 0.4 \\ 0.6 & 0.4 & 0 \end{pmatrix}$$

- Suppose each node a, b, c has one unit of clay on it.

$$\mu_0(a) = 1, \ \mu_0(b) = 1, \ \mu_0(c) = 1.$$

- Unit of clay at node b sends 60% to node a and 40% to node c.
- Node a receives 0.6 clay from node b plus 0.6 clay from node c.
- New measure:

$$\mu_0(a) = 1.2, \ \mu_0(b) = 0.9, \ \mu_0(c) = 0.9.$$

Definition 47
Given $X_0 = x$, let N_y be the number of visits of the Markov chain to y in times $\{0, 1, 2, \ldots, R_x - 1\}$. So

$$N_y = \sum_{i=0}^{R_x} \mathbb{1}(X_i = y | X_0 = x).$$

Having a random variable R_x in the limit of the summation is difficult to deal with. Fortunately there is a trick that we can use to get rid of it using indicator functions.

$$N_y = \sum_{i=0}^{\infty} \mathbb{1}(X_i = y, i < R_x | X_0 = x).$$

In turns out that the expected value of the N_y form a stationary measure!

Fact 32
Fix x a recurrent state in the Markov chain. For all y, set

$$\mu(y) = \mathbb{E}_x(N_y).$$

Then μ is a stationary measure.

Proof. Fix $y \in \Omega$. Then $\mu(y)$ is the expected number of visits to y in $\{0, 1, \ldots, R_x - 1\}$. It does not matter if we count the visits in $\{1, 2, \ldots, R_x\}$ since either way there is exactly one visit to x counted, and the visits in $\{1, 2, \ldots, R_x - 1\}$ are the same. So

$$\mu(y) = \sum_{i=0}^{\infty} \mathbb{P}(X_t = y, i < R_x | X_0 = x) = \sum_{i=1}^{\infty} \mathbb{P}(X_t = y, i \leq R_x | X_0 = x),$$

where the last equality follows from $X_0 = X_{R_x} = x$.

Start with the case $y \neq x$. Then to show stationarity, consider

$$\begin{aligned}\sum_z \mu(z)\mathbb{P}(X_1 = y | X_0 = z) &= \sum_z \left(\sum_{j=0}^{\infty} \mathbb{P}(X_j = z, j < R_x | X_0 = x)\ \mathbb{P}(X_1 = y | X_0 = z) \right) \\ &= \sum_z \left(\sum_{j=0}^{\infty} \mathbb{P}(X_j = z, j < R_x | X_0 = x)\ \mathbb{P}(X_{j+1} = y | X_j = z) \right)\end{aligned}$$

since $\mathbb{P}(X_1 = y | X_0 = z) = p(z, y) = \mathbb{P}(X_{j+1} = y | X_j = z)$ by the time homogeneity of the Markov chain. Now use a change of variable $i = j + 1$ to get

$$\begin{aligned}\sum_z \mu(z)\mathbb{P}(X_1 = y | X_0 = z) &= \sum_z \sum_{i=1}^{\infty} \mathbb{P}(X_{i-1} = z, i \leq R_x | X_0 = x)\ \mathbb{P}(X_i = y | X_{i-1} = z) \\ &= \sum_{i=1}^{\infty} \sum_z \mathbb{P}(X_i = y, X_{i-1} = z, i \leq R_x | X_0 = x)\end{aligned}$$

Where we can swap the summations because the sum over z is a finite sum. But now the interior sum is just saying we have to pass through some state z on our way from x to y, and so

$$\begin{aligned}\sum_z \mu(z)\mathbb{P}(X_1 = y | X_0 = z) &= \sum_{i=1}^{\infty} \mathbb{P}(X_i = y, i \leq R_x | X_0 = x) \\ &= \mu(y),\end{aligned}$$

and the measure is stationary for y.

□

Note that we can use any recurrent state x to get a stationary measure. Starting at different states (if they are in different recurrent communication classes) will give you different stationary measures.

17.2 The stationary distribution

Fact 33
For a finite state Markov chain with a recurrent class, there is always a stationary distribution.

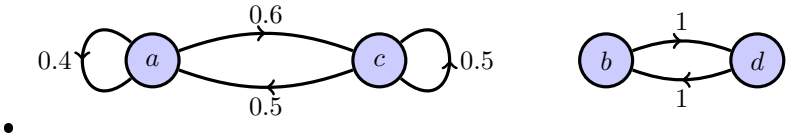
Proof. Start with the stationary measure $\mu(y) = \mathbb{E}_x(N_y)$. Then by definition $\sum_{y\in\Omega} N_y = R_x$, so they have the same mean. Since Ω is finite, linearly of expectation gives:

$$\sum_{y\in\Omega} \mathbb{E}[N_y] = \mathbb{E}[R_x] < \infty,$$

where the last inequality is true for finite state Markov chains. So finite state Markov chains always have at least one stationary distribution. □

- By definition, $N_x = 1$, so $\mathbb{E}[N_x] = 1$. So the stationary distribution associated with a recurrent state x is
$$\pi(x) = \frac{1}{\mathbb{E}[R_x]}.$$

17.3 Example of stationary measure



- Start with $x = b$.
  - Expected number of visits to d before return to a is 1. (Always goes b, d, b.)
  - Expected number of visits to b before return is always 1. (In general $N_x = 1$ given $X_0 = x$.)
  - Expected number of visits to a or c are both 0.
  - So stationary measure $(0, 1, 0, 1)$
  - Can normalize to stationary distribution $0 + 1 + 0 + 1 = 2$, so
  $$\frac{1}{2}(0, 1, 0, 1) = (0, 1/2, 0, 1/2)$$
  is stationary dist.
- Now do $x = a$.
  - As before, $N_a = 1$.
  - With 40% chance, $N_c = 0$, otherwise $N_c \sim \mathsf{Geo}(1/2)$.
  - So $\mathbb{E}[N_c] = (0.4)(0) + (0.6)(1/(1/2)) = 1.2$.
  - Final stationary measure:
  $$(1, 0, 1.2, 0).$$
  - For stat. dist.:
  $$\frac{1}{1 + 0 + 1.2 + 0}(1, 0, 1.2, 0) = (5/11, 0, 6/11, 0).$$
- Check answer. For μ_b:
$$\begin{pmatrix} 0 & 1/2 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 0.4 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$

- Check answer: For μ_a:

$$\begin{pmatrix} 5/11 & 0 & 6/11 & 0 \end{pmatrix} \begin{pmatrix} 0.4 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 5/11 & 0 & 6/11 & 0 \end{pmatrix}$$

18 Stationary distributions

Question of the Day What are the stationary distributions of a Markov chain?

Today

- Turning recurrent communication classes into stationary distributions.

Last time...

- μ is a stationary measure if
$$\mu(z) = \sum_{y \in \Omega} \mu(y) \mathbb{P}(X_{t+1} = z | X_t = y).$$
- Let N_y be # of visits to y in between visits to x. Then $\mu(y) = \mathbb{E}[N_y]$ is a stationary measure.
- Also, $\sum_{y \in \Omega} \mathbb{E}[N_y] = \sum_{y \in \Omega} \mu(y) = \mathbb{E}[R_x]$ Note that we can bring the expectation inside the sum when Ω is finite by linearity of expectation, and when Ω is a countable set by the Monotonic Convergence Theorem.
- If $\mathbb{E}[R_x] < \infty$, then $\mu(z)/\mathbb{E}[R_x]$ is a stationary distribution.

Definition 48
A recurrent state x is **positive recurrent** if
$$\mathbb{E}[R_x] < \infty.$$

We showed earlier the following useful fact:

Fact 34
All recurrent states in a finite state Markov chain are positive recurrent.

The next thing is to show that each recurrent communication class only has a single stationary distribution associated with it. The first step is to show that if $\pi(x) > 0$ for some x in a recurrent comm class, then $\pi(y) > 0$ for all y in that class.

Fact 35
Let $C \subseteq \Omega$ be a recurrent communication class and π be a stationary distribution. If there exists $x \in C$ with $\pi(x) > 0$, then for all $y \in C$ it holds that $\pi(y) > 0$.

Proof. Suppose $x \in C$ has $\pi(x) > 0$. Let $y \in C$. Then there exists n such that $\mathbb{P}(X_n = y | X_0 = 0) > 0$. Recall if $X_0 \sim \pi$, then $X_n \sim \pi$. Hence $\pi(y) \geq \pi(x)\mathbb{P}(X_n = y | X_0 = 0) > 0$. □

Fact 36
Let C be a recurrent communication class. There is a unique stationary distribution such that $\pi(x) > 0$ for all $x \in C$.

Actually it is better to show an equivalent statement.

Fact 37
Let C be a recurrent communication class. Then the set of stationary measures ν such that $\nu(C^C) = 0$ is unique up to constant multiples.

Notation: for $x, y \in \Omega$, let
$$p(x, y) = \mathbb{P}(X_1 = y | X_0 = x).$$

Proof. Let ν be a stationary measure with $\nu(C^C) = 0$ and $c \in C$. Then by definition, for any state $z \in C$,

$$\nu(z) = \sum_y \nu(y)p(y,z) = \nu(c)p(c,z) + \sum_{y \neq c} \nu(y)p(y,z).$$

We could write this same equation using different dummy variables as

$$\nu(y) = \nu(c)p(c,y) + \sum_{x \neq c} \nu(x)p(x,y).$$

Now use this identity recursively for $\nu(y)$ to get:

$$\begin{aligned}
\nu(z) &= \nu(c)p(c,z) + \sum_{y \neq c} \left[\nu(c)p(c,y) + \sum_{x \neq c} \nu(x)p(x,y) \right] p(y,z) \\
&= \nu(c)p(c,z) + \sum_{y \neq c} \nu(c)p(c,y)p(y,z) + \sum_{y \neq c} \sum_{x \neq c} \nu(x)p(x,y)p(y,z) \\
&= \nu(c) \left[\mathbb{P}_c(X_1 = z) + \mathbb{P}_c(X_1 \neq c, X_2 = z) \right] + \mathbb{P}_\nu(X_0 \neq c, X_1 \neq c, X_2 = z)
\end{aligned}$$

I can repeat this recursion n times (formally you would use an induction) to get:

$$\nu(z) = \nu(c) \sum_{m=1}^{n} \mathbb{P}_x(X_k \neq c \, \forall k \in \{1, \ldots, m\}, X_m = z) + \sum_{x_0 \neq c} \nu(x_0) \mathbb{P}(X_0 = x_0, X_1 \neq c, X_2 \neq c, \ldots, X_{n-1} \neq c, X_n = z).$$

Now take the limit as n goes to infinity. Since the probability of an event is the expected value of the indicator function of the event, and indicator functions are bounded by 1 in absolute value, the limit can be brought inside the probability. In the last term, no matter what x_0 is, it bounded above by the probability that the first $n-1$ steps never hit c, and this goes to 0 as n goes to infinity since c and x_0 have to be in the same communication class for $\nu(x_0) > 0$.

The sum in the first term converges by the monotone convergence theorem to

$$\sum_{m=1}^{\infty} \mathbb{P}_c(X_k \neq c \, \forall k \in \{1, \ldots, m\}, X_m = z) = \mathbb{E}[N_z].$$

Hence

$$\nu(z) = \nu(c)\mathbb{E}[N_z].$$

Note $\nu(x)$ is a fixed constant independent of z. So what this says is that ν must be a multiple of the stationary distribution we already know about, $\mathbb{E}[N_z]$, which is 0 for any z outside of the communication class. □

Fact 38
Let C be a recurrent communication class. Then the set of stationary distributions $\{\pi : \pi(C^C) = 0\}$ contains exactly one distribution:

$$\pi(x) = \frac{1}{\mathbb{E}[R_x]}.$$

Proof. From earlier if two stationary measures π and π' exist, then $\pi = c\pi'$ for a constant c. But $\sum_{x \in C} \pi(c) = \sum_{x \in C} \pi'(c) = 1$, so $c = 1$ and $\pi = \pi'$.

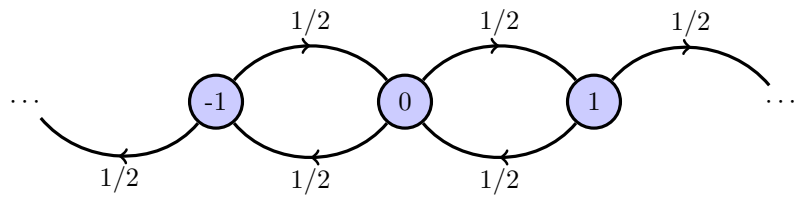
Let x be any element of C. Then $\pi(y) = \mathbb{E}[N_y]/\mathbb{E}[R_x]$ is the unique stationary distribution, and $N_x = 1$ since x is always visited exactly once in $\{0, 1, \ldots, R_x - 1\}$. So $\pi(x) = 1/\mathbb{E}[R_x]$. □

Intuition behind this result

- Suppose that on average it takes 5 steps starting from x to return to x.
- So in the long run, about 1/5 of the time the state will be state x.

18.1 Example of countably infinite chain with no stationary distribution

Simple symmetric random walk on $\mathbb{Z}$ is a Markov chain with no stationary distribution.



Proof that no stationary distribution exists. Suppose $\mu(i) \geq 0$ exists such that

$$\mu(i) = (1/2)\mu(i-1) + (1/2)\mu(i+1),$$

so that

$$\mu(i+1) = 2\mu(i) - \mu(i-1)$$

or

$$\mu(i+2) = 2\mu(i+1) - \mu(i).$$

This is a recurrence relation. If $\mu(0) = \mu(1)$, then this equation gives $\mu(2) = 2\mu(0) - \mu(0) = \mu(0)$, and

i	0	1	2	3	$\cdots$
$\mu(i)$	$\mu(0)$	$\mu(0)$	$\mu(0)$	$\mu(0)$	$\cdots$

An easy induction proof gives $\mu(i) = \mu(0)$ for all i. When $\mu(0) > 0$ then $\sum_i \mu(i) \neq 1$, and when $\mu(0) = 0$, $\sum_i \mu(i) = 0$. Either way, μ is not a probability distribution!

Suppose $\mu(1) \neq \mu(0)$. Then

$$\mu(2) = 2\mu(1) - \mu(0) = \mu(1) + (\mu(1) - \mu(0)).$$

Similarly

$$\mu(3) = 2\mu(2) - \mu(1) = \mu(2) + (\mu(2) - \mu(1)) = \mu(2) + (\mu(1) - \mu(0)),$$

and in general

$$\mu(i+1) = \mu(i) + (\mu(1) - \mu(0)).$$

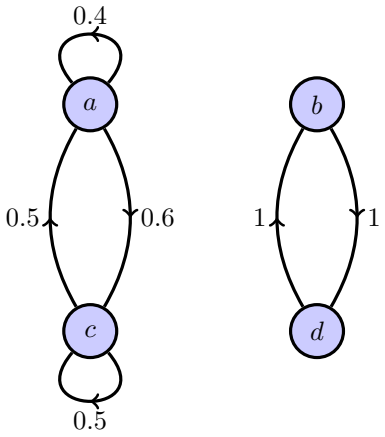
Again an easy induction gives $\mu(i+1) = \mu(0) + i(\mu(1) - \mu(0))$. when $\mu(1) - \mu(0) > 0$ then when i is large enough $\mu(i) > 1$, and when $\mu(1) - \mu(0) < 0$ then for i large enough $\mu(i) < 0$. Either way, μ is not a probability distribution! □

19 The ergodic theorem for finite state Markov chains

Question of the Day Finite state Markov chains always have a stationary distribution π. When is π also a limiting distribution?

Our story so far

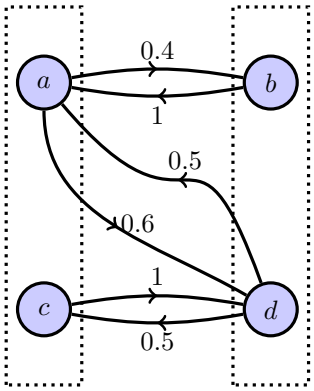
- Every finite state Markov chain has at least one stationary distribution.
- For applications (shuffling cards, Markov chain Monte Carlo (MCMC)), want stationary distribution to equal limiting distribution.
- What can go wrong?
- Two things:
  - First, more than one stationary distribution (can't converge to both!)



$\mu_1 = (0, 1/2, 0, 1/2)$

$\mu_2 = (5/11, 0, 6/11, 0)$

  - Second, periodicity.



Alternates $\{a, c\}$ and $\{b, d\}$

- It turns out, these are the only two problems!
- To solve first problem: only one recurrent communication class.
- To solve second problem: need aperiodicity.

Definition 49
For a recurrent communication class C, let k be the largest integer such that there is a partition C into $P_0, P_1, \ldots, P_{k-1}$ that satisfy

$$(\forall i \in \{0, 1, \ldots, k-1\})(\forall x \in P_i)(\mathbb{P}(X_{t+1} \in P_{i+1} | X_t = x) = 1)$$

where $P_k = P_0$. Call k the **period** of C.

Example

- In the example above $P_0 = P_2 = \{a, c\}$ and $P_1 = \{b, d\}$.
- From any state in P_0 you land in P_1 with probability 1.
- From any state in P_1 you land in $P_2 = P_0$ with probability 1.
- The period is 2.

Definition 50
A recurrent communication class C is **aperiodic** if it has period 1.

Theorem 10 (Ergodic Theorem for finite state Markov chains)
For finite state Markov chains

1: There is at least one stationary distribution.

2: The stationary distribution π is unique if and only if there is exactly one recurrent communication class.

3: If C is a recurrent communication class, then setting for all $x \in C$,

$$\pi(x) = \frac{1}{\mathbb{E}[R_x]}.$$

and for all $y \notin C$, $\pi(y) = 0$, π is a stationary distribution.

4: For one recurrent aperiodic communication class,

$$(\forall x \in \Omega)(\forall A \in \mathcal{F}) \left(\lim_{t \to \infty} \mathbb{P}(X_t \in A | X_0 = x) = \pi(A) \right)$$

Definition 51
A chain is **irreducible** if it consists of exactly one recurrent communication class. If that class is also aperiodic, then the chain is **ergodic**.

The short version of ergodic theorem:

> Limiting distribution equals unique stationary distribution if a finite state Markov chain is irreducible and aperiodic.

Slightly longer version of ergodic theorem:

> Limiting distribution equals unique stationary distribution iff a finite state Markov chain has one recurrent communication class which is aperiodic.

Another way to describe limiting distribution is by using a distance between probability measures.

Definition 52
The **total variation distance** between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ is

$$d_{\text{TV}}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{A \in \mathcal{F}} |\mathbb{P}_1(A) - \mathbb{P}_2(A)|.$$

[Recall: a distance (aka a metric) d is a function that takes pairs of states and returns a real number that satisfies for all states x, y and z:

1: $d(x, y) \geq 0$ (nonnegativity)

2: $d(x, y) = 0 \Leftrightarrow x = y$ (identity)

3: $d(x, y) = d(y, x)$ (symmetry)

4: $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality)

Straightforward to show that d_{TV} is a distance.]

Recall: The probability distribution associated with r.v. X is

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A).$$

So for instance, the total variation distance from X to π is:

$$d_{\text{TV}}(X, \pi) = \sup_{A} |\mathbb{P}(X \in A) - \pi(A)|.$$

With this notation, the ergodic theorem is:

Finite state Markov chains that are irreducible and aperiodic with stationary distribution π have

$$(\forall x \in \Omega) \left(\lim_{t \to \infty} d_{\text{TV}}([X_t | X_0 = x], \pi) = 0 \right).$$

Aperiodicity It turns out to be easy to force a Markov chain to be aperiodic.

Fact 39
Suppose x is in a recurrent communication class and $\mathbb{P}(X_1 = x | X_0 = x) > 0$. Then the class is aperiodic.

Proof. Let C_1, C_2 be any partition of the state space where $x \in C_1$. Then $\mathbb{P}(X_2 \in C_2 | X_1 = x) \leq 1 - \mathbb{P}(X_2 = x | X_1 = x) < 1$. Hence the chain is aperiodic. □

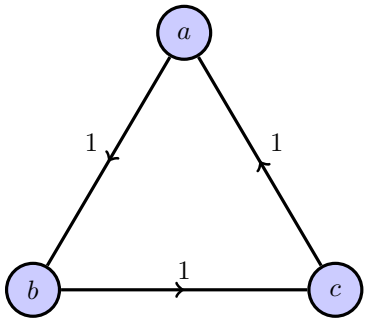
Using holding to give aperiodicity

- Want to give states a holding probability without changing the stationary distribution π.
- Idea: at each step, flip a fair coin. If heads, stay where you are, otherwise move according to the Markov chain.
- Now every state has a holding probability. (So in particular the recurrent states do.)
- Old transition matrix A, new transition matrix $(1/2)A + (1/2)I$.

$$\pi A = \pi \Rightarrow \pi[(1/2)A + (1/2)I] = (1/2)\pi + (1/2)\pi = \pi.$$

Eigenvalues and aperiodicity

- Period k chains cycle through a partition $P_1, P_2, \ldots, P_k$.
- After k steps, back in P_1.
- So like raising transition matrix to k power.
- If A^k is a Markov chain on P_1 with eigenvalue 1...
- A must have had an eigenvalue that is the kth root of 1.
- Example:



Eigenvalues $1, -1/2 \pm (\sqrt{3}/2)i$, cube roots of unity

19.1 Periodicity through greatest common divisors

Another way to view periodicity is through the lens of greatest common divisors. This is a more number theory way of looking at periodicity. First we need the notion of when an integer divides another integer.

Definition 53
A positive integer k divides a positive integer n if there exists an integer ℓ such that $k\ell = n$. Write $k|n$.

For example: 2 divides 6, 14 divides 14, and 1 divides every positive integer.

Definition 54
Let A be a set of positive integers. Then the **greatest common divisor** (or gcd) of A is

$$\max\{k \in \mathbb{Z}^+ | (\forall a \in A)(k|a).$$

For example, $\gcd(\{3,4\}) = 1$, $\gcd(\{2,4,6,\ldots\}) = 2$. Now we can find out the periodicity as a gcd.

Fact 40
Let a be any state in a recurrent communication class C. Let $M = \{m : \mathbb{P}(X_m = a | X_0 = a)\}$. Then the period of C equals $\gcd(M)$.

Proof. Let a be a state in recurrent communication class C. Suppose that the period of C is k. Then there exists a partition $(P_0, P_1, \ldots, P_{k-1})$ of C such that

$$(\forall i \in \{0,1,2,\ldots,k-1\})(\mathbb{P}(X_1 \in P_{i+1} | X_0 \in P_i) = 1).$$

where $P_k = P_0$. In fact for $d \in \{0,1,\ldots\}$ and $i \in \{0,1,\ldots,k-1\}$) set $P_{kd+i} = P_i$. Then it is straightforward to show

$$(\forall i \in \{0,1,2,\ldots\})(\mathbb{P}(X_1 \in P_{i+1} | X_0 \in P_i) = 1).$$

Now let m be a positive integer such that $\mathbb{P}(X_m = a | X_0 = a) > 0$. Then since $X_0 = a$, $a \in P_0$ in the partition. Then it must be true that $a \in P_m$, so $P_m = P_0$. Hence $k|m$, and k is a divisor of every element of M.

This is true for any partition of the form

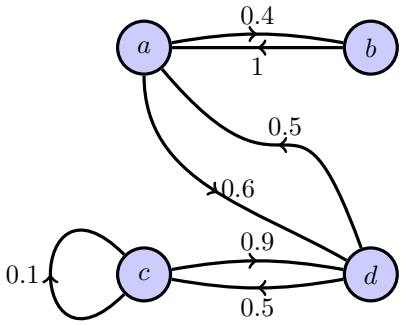
$$(\forall i \in \{0, 1, 2, \ldots, k-1\})(\mathbb{P}(X_1 \in P_{i+1} | X_0 \in P_i) = 1).$$

But the period is the greatest such k where a partition exists, and so it is the greatest common divisor of the integers in M. $\square$

20 Using the Ergodic Theorem to calculate expected travel times

Question of the Day Consider simple symmetric random walk with partially reflecting boundaries on $\{0, 1, \ldots, n\}$. What is the expected number of steps needed to return to 0 starting from 0?

Another example



What is the expected time to return to state a starting at state a?

- There is one recurrent communication class because the graph is connected.
- The chain is aperiodic because $\mathbb{P}(X_1 = c | X_0 = c) = 0.1 > 0$.
- So there is a unique stationary distribution. To find, solve

$$\begin{pmatrix} a & b & c & d \end{pmatrix} \begin{pmatrix} 0 & 0.4 & 0 & 0.6 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0.5 & 0 & 0.5 & 0 \end{pmatrix} = \begin{pmatrix} a & b & c & d \end{pmatrix}$$

- Unfortunately multiple solutions (if $\vec{v}$ solves, so does $c\vec{v}$ for any constant c.)
- We know one extra thing: $a + b + c + d = 1$.

$$\begin{pmatrix} a & b & c & d \end{pmatrix} \begin{pmatrix} 0 & 0.4 & 0 & 0.6 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.1 & 0.9 & 1 \\ 0.5 & 0 & 0.5 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b & c & d & 1 \end{pmatrix}$$

In Wolfram Alpha

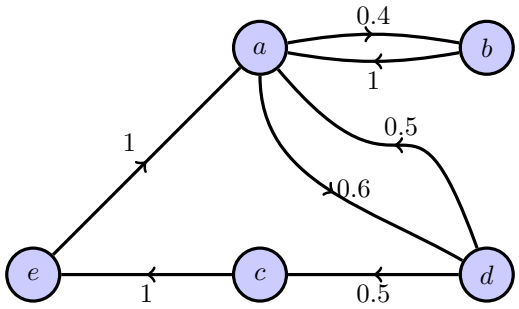
```
solve {{a,b,c,d}}{{0,0.4,0,0.6,1},{1,0,0,0,1},
{0,0,0.1,0.9,1},{0.5,0,0.5,0,1}}={{a,b,c,d,1}}
```

gives $\pi(a) \approx 0.306122$.
By the Ergodic Theorem

$$\mathbb{E}[R_a] = \frac{1}{\pi(a)} = \frac{1}{0.3061\ldots} \approx \boxed{3.266}.$$

Now continue the example. What is the expected time to travel from a to c?

- Idea: alter the Markov chain somewhat.
- First, add a dummy node e.
- Second, always move from c to e.
- Third, always move from e to a.



- To get from e to itself, have to move from a to c!
- Let $T_{a,c} = \inf\{t : X_t = c | X_0 = a\}$.
- Then $T_{a,c} = R_e - 2$. (So $\mathbb{E}[T_{a,c}] = \mathbb{E}[R_e] - 2$.)
- Find $\mathbb{E}[R_e]$ using the ergodic theorem.

$$\begin{pmatrix} a & b & c & d & e \end{pmatrix} \begin{pmatrix} 0 & 0.4 & 0 & 0.6 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0.5 & 0 & 0.5 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b & c & d & e & 1 \end{pmatrix}$$

- Solution

$$\mathbb{E}[T_{a,c}] = \frac{1}{0.11538\ldots} - 2 \approx \boxed{6.666}.$$

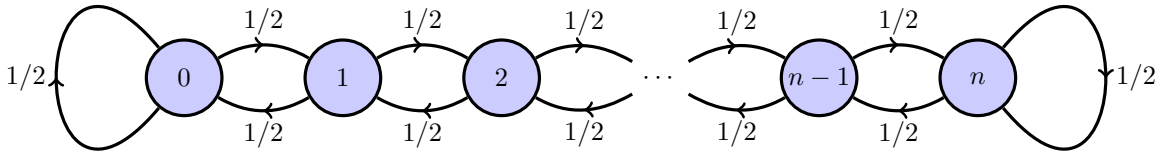
First step analysis

- Recall that we already had a method for finding $\mathbb{E}[T_{a,b}]$, first step analysis!
- Let $w_i = \mathbb{E}[T_{i,c}]$.
- Then

$$\begin{aligned} w_a &= 1 + 0.4w_b + 0.6w_d \\ w_b &= 1 + w_a \\ w_c &= 0 \\ w_d &= 1 + 0.5w_a + 0.5w_c \end{aligned}$$

- This has solution $w_a = 20/3$, so $\boxed{6.666}$ steps are needed on average.
- Obvious but worth noting: you should get the same answer!

Back to the Question of the Day Simple symmetric random walk with partially reflecting boundaries:



In order to use the Ergodic Theorem, we need to have a stationary distribution. The detailed balance equations are:

$$\mathbb{P}(X_{t+1} = j) = \sum_i \mathbb{P}(X_t = i)\mathbb{P}(X_{t+1} = j | X_t = i).$$

$$\begin{aligned}
\pi(0) &= (1/2)\pi(0) + (1/2)\pi(1) \\
\pi(1) &= (1/2)\pi(0) + (1/2)\pi(2) \\
\vdots &= \vdots \\
\pi(i) &= (1/2)\pi(i-1) + (1/2)\pi(i+1) \\
\vdots &= \vdots \\
\pi(n-1) &= (1/2)\pi(n-2) + (1/2)\pi(n) \\
\pi(n) &= (1/2)\pi(n-1) + (1/2)\pi(n).
\end{aligned}$$

Solving these equations one by one gives:

$$\begin{aligned}
\pi(0) &= \pi(1) \\
\pi(1) &= \pi(2) \\
\pi(2) &= \pi(3) \\
&\vdots
\end{aligned}$$

In fact, $\pi(i)$ all equal! Since must sum to 1,

$$\pi(i) = \frac{1}{n+1}.$$

So time to return to 0 is $\boxed{\text{n + 1}}$.
Time to return to 5 starting from 5, also $n + 1$!
The time to return to any state from that state in this chain: $n + 1$.

21 Coupling

Question of the Day How can we prove the ergodic theorem?

Today's lecture

- Notion of coupling random variables for proof of Ergodic Thm.
- Show a number theory way of looking at the period of a state.

Definition 55
Random variables X and Y are said to have **coupled** if $X = Y$.

Definition 56
A **coupling** of random variables X and Y is just another name for the bivariate distribution.

Example

- Suppose $U \sim \mathsf{Unif}([0,1])$, $X = U$, $Y = U$. Then X and Y are always coupled.
- Suppose $U \sim \mathsf{Unif}([0,1])$, $X = U$, $Y = 1 - U$. Then X and Y are uncoupled with probability 1.
- Suppose $U \sim \mathsf{Unif}([0,1])$, $X = \mathbb{1}(U \leq 0.3)$, $Y = \mathbb{1}(U \leq 0.6)$. Then

$$\mathbb{P}(X = Y) = \mathbb{P}(X = Y = 0) + \mathbb{P}(X = Y = 1) = 0.3 + 0.4 = 0.7.$$

  Note that in this example $\text{dist}_{TV}(X, Y) = \mathbb{P}(X \neq Y)$.

Rather than just coupling a single pair of variables, we could couple two entire stochastic processes.

Definition 57
A **coupling** of two stochastic processes X_t and Y_t is a sequence $\{(X_t, Y_t)\}$ such that $\{X_t\}$ and $\{Y_t\}$ viewed by themselves have the correct distribution for their process.

Example

- Suppose $D_1, D_2, D_3, \ldots \overset{\text{iid}}{\sim} \mathsf{Unif}(\{-1, 1\})$.
- $X_0 = Y_0 = 0$, $X_{t+1} = X_t + D_{t+1}$, $Y_{t+1} = Y_t + D_{t+1}$.
- Then X_t and Y_t both move according to the same Markov chain.
- Now suppose $W_0 = 0$, $W_{t+1} = W_t - D_{t+1}$.
- Since $-D_i \sim D_i$, W_t has the same transition probabilities...
- ...but it moves differently than X_t or Y_t.

What does coupling have to do with total variation distance?

Fact 41 (Doeblin 1933, Coupling Lemma)
For random variables X and Y,

$$d_{\text{TV}}(X, Y) \leq \mathbb{P}(X \neq Y).$$

Proof. Let $A \in \mathcal{F}$. Then

$$\begin{aligned}
|\mathbb{P}_X(A) - \mathbb{P}_Y(A)| &= |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\
&= |\mathbb{P}(X \in A, X = Y) + \mathbb{P}(X \in A, X \neq Y) - \\
&\quad (\underbrace{\mathbb{P}(Y \in A, X = Y)}_{\mathbb{P}(X \in A, X = Y)} + \mathbb{P}(Y \in A, X \neq Y))| \\
&= |\underbrace{\mathbb{P}(X \in A, X \neq Y)}_{\text{in } [0, \mathbb{P}(X \neq Y)]} \underbrace{- \mathbb{P}(Y \in A, X \neq Y)}_{\text{in } [-\mathbb{P}(X \neq Y), 0]}| \\
&\leq \mathbb{P}(X \neq Y).
\end{aligned}$$

□

21.1 Using coupling to show the ergodic theorem

Suppose that $Y_0 \sim \pi$, where π is a stationary distribution of a Markov chain. Then $Y_1 \sim \pi$, $Y_2 \sim \pi$, and so on. Now let X_0 be any state $x_0 \in \Omega$. Then if the Markov chains $\{X_t\}$ and $\{Y_t\}$ are coupled, then

$$\text{dist}_{\text{TV}}(X_t, Y_t) = \text{dist}_{\text{TV}}(X_t, \pi) \leq \mathbb{P}(X_t \neq Y_t).$$

In words, the distance that X_t is from the stationary distribution is equal to the probability that the X_t process has not yet run into the stationary process Y_t. Suppose that the two processes are simulated independently. Then the following fact will be useful.

Fact 42
Let $\{X_t\}$ be an irreducible, aperiodic finite state Markov chain over Ω. For any two states x and y in Ω, there is a time t such that $\mathbb{P}(X_t = x | X_0 = x) > 0$ and $\mathbb{P}(X_t = x | X_0 = y) > 0$.

A fact from number theory is needed to prove this.

Fact 43
Let $A \subseteq \mathbb{Z}^+$ with $\gcd(A) = 1$ satisfying $(\forall a_1, a_2 \in A)(a_1 + a_2 \in A)$. Then

$$(\exists n)(\forall N \geq n)(N \in A).$$

Proof of Fact 42. Since y and x communicate, let c be the length of a path from y to x. Let A be the set of times t' such that $\mathbb{P}(X_{t'} = x | X_0 = x) > 0$. If $t_1, t_2 \in A$, then $\mathbb{P}(X_{t_1+t_2} = x | X_0 = x) \geq \mathbb{P}(X_{t_1} = x | X_0 = x) \cdot \mathbb{P}(X_{t_2} = x | X_0 = x) > 0$, so $t_1 + t_2 \in A$.

Then since the chain is aperiodic, $\gcd(A) = 1$, and so satisfies the condition of the number theory fact. Hence there is a time t' such tht $t = t' + c$ is also in the set, which gives the desired time. □

Now to prove the ergodic theorem.

Proof there is at least one stationary distribution. Since the chain is finite there exists a recurrent state z (and class). The cycle trick proof shows that there is a stationary measure, and since $\mathbb{E}[R_z] < \infty$, it can be normalized to give a stationary distribution. □

Proof the stationary distribution π is unique if and only if there is exactly one recurrent communication class. If there are at least two recurrent communication class, then the stationary measure from the cycle trick gives two stationary measures, each of which can be converted to a stationary distribution, and which give probability 1 to disjoint states. Hence the stationary distribution is not unique.

If there is one recurrent communication class. □

Proof that π is the inverse average return time to recurrent states. Let x be a recurrent state. The stationary measure cycle trick normalized to a stationary distribution gives π with $\pi(x) = 1/\mathbb{E}[R_x]$. For any y that is transient, this cannot be reached from x and so the same stationary distribution has $\pi(y) = 0$. By the previous lemma this stationary distribution is unique, so this holds for any recurrent x and transient y. □

For one recurrent aperiodic communication class, the stationary distribution is limiting. Fix $x \in \Omega$ and let $X_0 = x$. Let $Y_0 \sim \pi$, where π is stationary. Then $Y_t \sim \pi$ for all t.

Advance the X_t and Y_t chains independently if $X_t \neq Y_t$, otherwise just advance X_t to X_{t+1} and set $Y_{t+1} = X_{t+1}$. With this coupling both $\{X_t\}$ and $\{Y_t\}$ are following the transition probabilities for the Markov chain. Hence

$$\mathrm{dist}_{\mathrm{TV}}(X_t, \pi) = \mathrm{dist}_{\mathrm{TV}}(X_t, Y_t) \leq \mathbb{P}(X_t \neq Y_t).$$

Now, for every pair of states $x, y \in \Omega$, there is a time $t_{x,y}$

$$\alpha_{x,y} = \min\{\mathbb{P}(X_{t_{x,y}} = z | X_t = x), \mathbb{P}(X_{t_{x,y}} = z | X_t = y)\} > 0.$$

So after $t_{x,y}$ steps, there is at most a $1 - \alpha_{x,y}$ chance that $X_{t_{x,y}} \neq Y_{t_{x,y}}$. Let $\alpha = \min_{x,y} \alpha_{x,y}$ and $t = \max_{x,y} t_{x,y}$.

Then for any $k \in \{1, 2, 3, \ldots\}$, $\mathbb{P}(X_{kt} \neq Y_{kt}) \leq (1 - \alpha)^k \to 0$. □

22 Using coupling to show mixing time

Question of the Day For simple symmetric random walk on the integers $\{0, 1, 2, \ldots, n-1\}$, find the mixing time.

Today

- Mixing time of the Markov chain.
- Using coupling to bound mixing times

Recall that for irreducible, aperiodic Markov chains, the limiting distribution equals the unique stationary distribution. Roughly speaking, the mixing time of a Markov chain is how many steps must be taking before the chain forgets the state that it is currently in. Or in other words, it is how many steps must be taken before the distribution of the state is close to stationarity.

Definition 58
For a Markov chain with unique stationary distribution π, let

$$\tau_{x,\epsilon} = \inf\{t : \text{dist}_{\text{TV}}(X_t | X_0 = x, \pi) \leq \epsilon\}.$$

Call

$$\tau_\epsilon = \sup_{x \in \Omega} \tau_{x,\epsilon}$$

the **mixing time** of the Markov chain.

Recall the coupling lemma says that $\text{dist}_{\text{TV}}(X_t | X_0 = x, \pi) \leq \mathbb{P}(X_t \neq Y_t | X_0 = x, Y_0 \sim \pi)$.
One way to build a coupling for two Markov chains is to create what is called an update function.

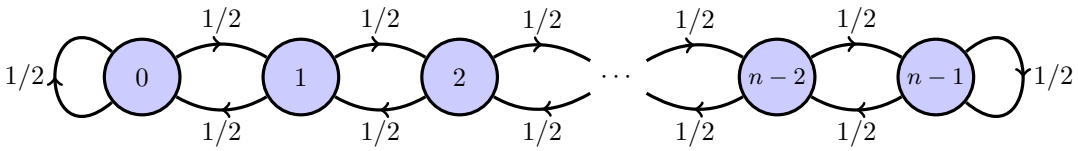
Definition 59
Let $R \in \Omega_R$ be a source of randomness. Then a Markov chain over Ω has an **update function**

$$f : \Omega \times \Omega_R \to \Omega$$

if f is a computable function that satisfies

$$(\forall A \in \mathcal{F})(\mathbb{P}(f(x, R) \in A) = \mathbb{P}(X_t \in A | X_{t-1} = x)).$$

In other words, an update function takes the current state plus some random choices, and returns the next state in the Markov chain. For the simple symmetric random walk on $\{0, \ldots, n-1\}$,



a simple update function is

$$f(x, U) = x + D \cdot \mathbb{1}(x + D \in \{0, \ldots, n-1\})$$

where $D \sim \text{Unif}\{-1, 1\}$. The indicator function makes the partially reflecting boundaries: if we try to move outside of the state space, then the indicator function forces us to stay where we are instead.

Now we can couple (X_t, Y_t) by saying

$$X_{t+1} = f(X_t, D_{t+1}), \; Y_{t+1} = f(Y_t, D_{t+1}),$$

so we are using the same random choice to move For example, say $n = 4$ so $\Omega = \{0, 1, 2, 3\}$.

t	D_t	(X_t, Y_t)
0		$(0, 2)$
1	+1	$(1, 3)$
2	+1	$(2, 3)$
3	-1	$(1, 2)$
4	+1	$(2, 3)$
5	+1	$(3, 3)$

so coupling occurred at $t = 3$.

Notice that for this chain, $X_t \leq Y_t$ implies $X_{t+1} \leq Y_{t+1}$. So if $X_t = n - 1$, then $Y_t = n - 1$, and $T_C \leq t$ where

$$T_C = \inf\{t : X_t = Y_t\}$$

we have for the stopping time

$$T = \inf\{t : X_t = n - 1\},$$

that $T_C \leq T$.

To bound $\mathbb{E}[T]$ (and hence $\mathbb{E}[T_C]$), we consider a third process whose update function is

$$W_{t+1} = W_t + D_{t+1}\mathbb{1}(W_t + D_{t+1} \geq 0),$$

so it is like the X_t process but can keep going up forever. Still, $W_{t\wedge T} = X_{t\wedge T}$, so studying the W_t process can help us understand the X_t process. In fact,

$$\mathbb{E}[W_{t+1}^2 | W_t] \geq \frac{1}{2}(W_t + 1)^2 + \frac{1}{2}(W_t - 1)^2 = W_t^2 + 1.$$

Hence $W_{t\wedge T}^2 - t$ is a martingale with respect to the natural filtration, and $\mathbb{E}[W_{t\wedge T}^2 - T] = 0$ for all t. Then by the ergodic theorem $\mathbb{P}(T < \infty) = 1$, and

$$\lim_{t\to\infty} \mathbb{E}[W_{t\wedge T}^2] - \lim_{t\to\infty} \mathbb{E}[t \wedge T] \geq 0.$$

Since $W_{t\wedge T}^2 \in [0, (n-1)^2]$, the limit can be brought inside by the bounded convergence theorem, and since $t \wedge T$ is increasing in t the limit can be brought inside by the monotonic convergence theorem. Using $T < \infty$ with probability 1 gives:

$$\mathbb{E}[W_T^2] - \mathbb{E}[T] \geq 0 \Rightarrow \mathbb{E}[T] \leq (n-1)^2.$$

So that tells us the average number of steps needed to couple. Can we bound the mixing time? Recall Markov's inequality:

Fact 44 (Markov's Inequality)
For an integrable random variable X,

$$\mathbb{P}(|X| \geq a) \leq \mathbb{E}[X]/a.$$

So $\mathbb{P}(T > 2(n-1)^2) \leq (n-1)^2/[2(n-1)^2] = 1/2$. In other words, after every $2(n-1)^2$ steps, we have at least a $1/2$ chance of coupling. Hence

$$\mathbb{P}(T_C > k \cdot 2(n-1)^2) \leq \mathbb{P}(T_C \leq \lfloor k \rfloor 2(n-1)^2) \leq (1/2)^{\lfloor k \rfloor - 1} \leq (1/2)^{k-2}.$$

for all $k > 0$. Then to make this at most ϵ,

$$k = \ln(1/\epsilon)/\ln(2) + 2,$$

so the mixing time is

$$\boxed{\tau_\epsilon \leq 2(n-1)^2[2 + \ln(1/\epsilon)/\ln(2)].}$$

23 Countable state spaces

Question of the Day Does the ergodic theorem hold for countably infinite state spaces?

Today

- Ergodicity for countable state spaces
- Still have recurrence and aperiodicity, need extra ingredient of positive recurrence.

Example: Simple symmetric random walk on $\mathbb{Z}$

- Transitions:
$$\mathbb{P}(X_{t+1} = x+1 | X_t = x) = \mathbb{P}(X_{t+1} = x-1 | X_t = x) = 1/2.$$
- This chain is recurrent: $\mathbb{P}(R_0 < \infty) = 1$ (can be shown using martingale techniques).
- This chain has a stationary measure: $\mu(\{i\}) = 1$ for all i.
- Because the chain is recurrent, the stationary measure is unique.
- This measure cannot be normalized (sums to infinity), by uniqueness that means that no stationary distribution can exist.

Countable versus finite state spaces

- Many definitions and facts are the same:
  - State i is **recurrent** if $\mathbb{P}(R_i < \infty) = 1$
  - Recurrence is a class property: If i is recurrent and i communicates with j, then j is recurrent.
  - A recurrent communication class C has **period** k if there exists a partition $P_0, P_1, \ldots, P_{k-1}$ of C such that
$$(\forall i \in \{0, 1, \ldots, k-1\})(\forall x \in P_i)(\mathbb{P}(X_{t+1} \in P_{i+1} | X_t = x) = 1)$$
    where $P_k = P_0$.
  - If $\mathbb{E}[R_i] < \infty$, then call i **positive recurrent**.
  - Positive recurrence is a class property.

Fact 45
If $\mathbb{E}[R_i] < \infty$ and i communicates with j, then $\mathbb{E}[R_j] < \infty$.

Since the state space is countable, we no longer have a matrix for the transitions:

Definition 60
A distribution π is **stationary** if
$$(\forall i \in \Omega)(\pi(i) = \sum_j \pi(j)\mathbb{P}(X_1 = i | X_0 = j)).$$
Call these the **balance equations**

It turns out that positive recurrence is exactly what is needed to make the ergodic theorem work.

Theorem 11 (Ergodic theorem for countable state space Markov chains)
For countable state space Markov chains

1: If C is a positive recurrent communication class, then setting for all $x \in C$,

$$\pi(x) = \frac{1}{\mathbb{E}[R_x]}.$$

and for all $y \notin C$, $\pi(y) = 0$, π is a stationary distribution.

2: Let π be a stationary distribution where $\pi(C) = 1$ for a communication class C. Then C is a positive recurrent communication class, and for all $x \in C$, $\pi(x) = 1/\mathbb{E}[R_x]$.

3: For one positive recurrent aperiodic communication class,

$$(\forall x \in \Omega)(\forall A \in \mathcal{F}) \left(\lim_{t \to \infty} \mathbb{P}(X_t \in A | X_0 = x) = \pi(A) \right)$$

Outline of proof of countable ergodic theorem. The theorem can be shown in the following way.

1: If x is positive recurrent, then the cycle trick gets a stationary measure $\mu(y) = \mathbb{E}_x[N_y]$ for all y in the recurrent communication class. Since $\sum_x [N_y] = \mathbb{E}[R_x]$ (by the MCT), this can be normalized by dividing by $\mathbb{E}[R_x] < \infty$ to get a stationary probability measure. This measure is unique by the same argument as Fact 37. Using $\mathbb{E}_y[N_y] = 1$ finishes the result.

2: This is actually the hardest to prove, since it is working in the opposite direction from the finite state Markov chain ergodic theorem.

3: The coupling proof applies in the same fashion as for the finite state problem.

□

Example

- Consider the asymmetric random walk on $\{0, 1, \ldots\}$ where for $i \geq 0$, $\mathbb{P}(X_1 = i+1 | X_0 = i) = 1/3$ and $\mathbb{P}(X_1 = i | X_0 = i+1) = 2/3$, and $\mathbb{P}(X_1 = 0 | X_0 = 0) = 2/3$. Find $\mathbb{E}[R_2]$.
- First, find a stationary distribution. The balance equations are:

$$\begin{aligned}
\pi(0) &= (2/3)\pi(0) + (2/3)\pi(1), \\
\pi(1) &= (2/3)\pi(2) + (1/3)\pi(0), \\
\pi(2) &= (2/3)\pi(3) + (1/3)\pi(1), \\
&\vdots = \vdots
\end{aligned}$$

Solving these gives

$$\begin{aligned}
\pi(0) &= 2\pi(1) \\
\pi(1) &= 2\pi(2) \\
\pi(2) &= 2\pi(3) \\
&\vdots = \vdots
\end{aligned}$$

Or

$$\pi(1) = (1/2)\pi(0), \ \pi(2) = (1/4)\pi(0), \ \pi(3) = (1/8)\pi(0), \ldots$$

and

$$\sum_{i=0}^{\infty} \pi(i) = \pi(0)[1 + (1/2) + (1/4) + \cdots] = 1,$$

which gives $\pi(0) = 1/2$ and $\pi(i) = (1/2)^{i+1}$.

- Next, since this is the stationary distribution, the chain must be positive recurrent, and $\mathbb{E}[R_2] = 1/(1/8) = \boxed{8}$.

24 General state spaces

Question of the Day Does the ergodic theorem hold for more general state spaces? Countably infinite? Uncountable?

Today

- Introduce notion of a *Harris chain*.
- Allows us to prove ergodic theorem in very similar way.
- Continuous state spaces do not return to exactly the same state.

In real world modeling, the state space of a Markov chain is often a continuous set, like $\mathbb{R}^n$. How can we extend the ergodic theorem to more general spaces?

24.1 Harris chain

A Harris chain is a Markov chain with two properties.

1: There is a special set A such that from any starting state, there is a positive chance of moving to A within a finite number of steps.

2: If the state is in A, then there is an ϵ chance that the chain can "forget" the exact location of the state within A in deciding the next move.

Definition 61
A Markov chain $\{X_t\}$ is a **Harris chain** if there exist a measurable set $A \subseteq \Omega$, $\epsilon > 0$, and a probability measure ρ where

1: For $T_A = \inf\{t \geq 0 : X_t \in A\}$,

$$(\forall z \in \Omega)(\mathbb{P}(T_A < \infty | X_0 = z) > 0).$$

2: For all $x \in A$ there is a distribution ν_x such that

$$[X_1 | X_0 = x] = \epsilon\rho + (1 - \epsilon)\nu_x.$$

The Harris chain definition encompasses two parts:

1: The first condition is the continuous equivalent of the chain consists of one communication class.

2: The second condition will allow the chain to couple effectively. It says that the distribution of $X_1|X_0 = x$ has an ϵ chance of being ρ, which does not depend on the value of x! With probability $1 - \epsilon$, the distribution of $X_1|X_0 = x$ is ν_x, which does depend on x.

Fact 46
Any countable state space chain with one communication class is a Harris chain.

Proof. Let A be any state x in the chain. Let $B = \{y : \mathbb{P}(X_1 = y | X_0 = x) > 0\}$, $\rho(C) = \mathbb{P}(X_1 \in C | X_0 = x)$, and $\epsilon = 1$.

Since all states communicate with x, $\mathbb{P}(T_A < \infty | X_0 = z) > 0)$ for all states z. Also, for $C \subseteq B$,

$$\mathbb{P}(X_1 \in C | X_0 = x) = (1)\rho(C),$$

so the choice of ϵ and ρ works as well. □

Random walk on $\mathbb{R}$

- Let $R_1, R_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Unif}([-1,1])$.
- Let $X_0 = 0$, $X_{t+1} = X_t + R_{t+1}$.
- Now it's a bad idea to set $A = \{0\}$, since chain returns to exactly 0 with probability 0!
- Let $A = [0,1]$.
- If $X_t = z > 0$, then suppose that

$$R_t, R_{t+1}, \ldots, R_{t+\lfloor z \rfloor} \in [-1, -(z-1)/\lfloor z \rfloor].$$

  Then

$$X_{t+\lfloor z \rfloor} = X_t + R_{t+1} + \cdots + R_{t+\lfloor z \rfloor} \in [z - \lfloor z \rfloor, z - (z-1)] \in [0,1].$$

  So

$$\mathbb{P}(T_A < \infty) \geq [\mathbb{P}(R_t \in [-1, (-(z-1)/\lfloor z \rfloor]]^{\lfloor z \rfloor} > 0.$$

- Notice from any point in A, there is at least a $1/2$ chance that the next point is also in $[0,1]$. So make $B = [0,1]$, and $\epsilon = 1/2$. Let $\rho(B) = \mathsf{Unif}(B)$. (Usually this is a good choice for ρ.)
- Let $C \subseteq B$. What is $\mathbb{P}(X_1 \in C | X_0 = x)$, where $x \in A$?
- If $x = 0$, $X_1 \sim \mathsf{Unif}([-1,1])$.
- If $x = 1$, $X_1 \sim \mathsf{Unif}([0,2]$.
- For any $x \in [0,1]$, there is a $1/2$ chance that $X_1 \in [0,1]$, so a $1/2$ chance that X_1 is uniform over $[0,1]$.
- So $\mathbb{P}(X_1 \in C | X_0 = x) = (1/2)\rho(C)$ for all $x \in [0,1]$.
- This is a Harris chain.

Why this definition?

- This is set up to allow us to use the same proof as before for the ergodic theorem!
- When $X_t \in A$, there is an ϵ chance that the next state comes from ρ, and is independent of the current state!
- Still need recurrence and transience, though.
- To make coupling happen, simulate chain as follows.
- First, some notation.

$$\tau_x(D) = \mathbb{P}(X_1 \in D | X_0 = x)$$
$$\psi_x(D) = (\tau_x(D) - \epsilon\rho(D))/(1-\epsilon)$$

  In words, $\tau_X(D)$ is the distribution of the next state of the chain given that the current state is x. ϕ_x is the distribution of the next state of the chain given that we started in A, but did not forget where in A we were.

  This is how to advance the original Harris chain to the next state.

Advancing X_t	
1)	If $X_t \notin A$
2)	Draw $X_{t+1} \leftarrow \tau_{X_t}$
3)	Else
4)	Draw $B \leftarrow \mathsf{Bern}(\epsilon)$
5)	If $B = 1$
6)	Draw $X_{t+1} \leftarrow \rho$
7)	Else
8)	Draw $X_{t+1} \leftarrow \psi_{X_t}$

- Note that for any measurable set D,
$$\tau_x(D) = \epsilon\rho(D) + \psi_x(D),$$
so this way of updating the chain value is valid.

Coupling X_t

- Suppose X_t and Y_t are coupled copies of the Markov chain.
- Then if $X_t \in A$ and $Y_t \in A$, and $B = 1$, then at the next time step X_{t+1} and Y_{t+1} can both be chosen according to ρ.
- When this happens $X_{t+1} = Y_{t+1}$ and coupling has occurred.

Definition 62
Let $R = \inf\{n > 0 : X_n \in A\}$. A Harris chain is **recurrent** if for all $x \in A$, $\mathbb{P}(R < \infty | X_0 = x) = 1$. A Harris chain that is not recurrent is transient.

Definition 63
A recurrent Harris chain is **aperiodic** if for all $x \in \Omega$, there exists n such that for all $n' \geq n$,
$$\mathbb{P}(X_{n'} \in A | X_0 = x) > 0.$$

As with finite state Markov chains, the easiest way to get aperiodicity is for every state in A to have a positive probability of holding.

Fact 47
Suppose for all $a \in A$ (from the Harris chain definition), $\mathbb{P}(X_1 \in A | X_0 = a) > 0$. Then X_t is aperiodic.

Proof. Fix $x \in \Omega$. Then for some n, $\mathbb{P}(X_n \in A | X_0 = x) > 0$. Then no matter where $X_n \in A$ is, there is a positive chance of landing in A at the next step. An induction yields
$$\mathbb{P}(X_{n'} \in A | X_0 = x) > 0$$
for all $n' \geq n$. □

The bad news is that even with aperiodicity and recurrence, we do not get everything in the ergodic theorem with Harris chains that we got in the finite or countable state space case. However, we do get the most important thing, which is that if we have a stationary distribution, then it will also be the limiting distribution.

Theorem 12 (Ergodic Theorem for Harris chains)
Let X_n be an aperiodic recurrent Harris chain with stationary distribution π. If $\mathbb{P}(R < \infty | X_0 = x) = 1$ for all x, then as $t \to \infty$,
$$d_{\text{TV}}([X_t | X_0 = x], \pi) \to 0.$$

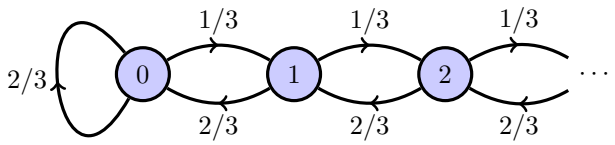
Notes

- In Ergodic Theorem for finite Markov chains
  - Require one recurrent communication class, aperiodicity.
  - Get existence of π for free.
  - Get limiting distribution equals π.
- In Ergodic Theorem for Harris chains
  - Require recurrent, aperiodic Harris chain.
  - Have to show that π exists.
  - Get limiting distribution equals π from any state that reaches A with probability 1.

25 Applying the Ergodic theorem for Harris chains

Question of the Day Let $B_1, B_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Bern}(1/3)$. Suppose $X_0 = 0$ and

$$X_{t+1} = \max\{X_t - 1 + 2B_{t_1}, 0\}.$$



What is the limiting distribution of $[X_t | X_0 = x]$?

To use the ergodic theorem for Harris chains:

- Need X_t is a Harris chain.
- Need X_t is aperiodic.
- Need $\mathbb{P}(R < \infty | X_0 = x) = 1$. (Note this also gives that the chain is recurrent.)
- Need a stationary distribution π.

X_t is a Harris chain.

- All states communicate with each and this is a countable state space, so it must be a Harris chain!

X_t is aperiodic.

- From the state a, there is a $(2/3)$ chance of moving back to a at the next step, so the chain is aperiodic.

$\mathbb{P}(R < \infty | X_0 = x) = 1$

- Each time we reach 0, there is a $2/3$ chance of moving to a. Therefore it suffices to show that $\mathbb{P}(T < \infty | X_0 = x)$, where $T = \inf\{t : X_t = 0 | X_0 = x\}$.
- Now $X_{t \wedge T}$ is a supermartingale, and $T_n = \inf\{t : X_t = 0 \text{ or } X_t = n | X_0 = x\}$ is a stopping time for this martingale.
- For $n > x$, $X_{t \wedge T_n}$ is a bounded (and hence u.i.) martingale, so the OST applies and

  $$\mathbb{E}[X_{T_n}] = \mathbb{E}[X_0] = x.$$

  Note $\mathbb{E}[X_{T_n}] = (0)\mathbb{P}(X_{T_n} = 0) + (n)\mathbb{P}(X_{T_n} = n)$, so $\mathbb{P}(X_{T_n} = n) = \mathbb{P}(T_n < T) = x/n$.

  For T to be infinite means that $T_n < T$ for all n. Hence $\mathbb{P}(T = \infty) \leq \mathbb{P}(T_n < T)$ for all n. But the only number less than x/n for all n is 0. So $\mathbb{P}(T < \infty) = 1$.

$\pi(i) = (1/2)^{i+1}$ is stationary

- How did I come up with that?
- Start writing down the equations that characterize stationarity

$$\mathbb{P}(X_{t+1} = y) = \sum_x \mathbb{P}(X_{t+1} = y | X_t = x)\mathbb{P}(X_t = x)$$
$$\pi(y) = \sum_x \pi(x)\mathbb{P}(X_{t+1} = y | X_t = x)$$

  So for $y \in \{0, 1, 2, \ldots\}$ these give:

$$\begin{aligned}
\pi(0) &= (2/3)\pi(0) + (2/3)\pi(1) \\
\pi(1) &= (1/3)\pi(0) + (2/3)\pi(2) \\
\pi(2) &= (1/3)\pi(1) + (2/3)\pi(3) \\
\pi(3) &= (1/3)\pi(2) + (2/3)\pi(4) \\
&\vdots = \vdots
\end{aligned}$$

- Solving the equations one by one gives:

$$\begin{aligned}
\pi(0) &= 2\pi(1) \\
\pi(1) &= 2\pi(2) \\
\pi(2) &= 2\pi(3) \\
&\vdots = \vdots
\end{aligned}$$

  So guess that the solution is

$$\pi(i) = C(1/2)^i$$

  and use $\sum_{i=0}^{\infty} C(1/2)^i = 2C = 1$ to get $C = 1/2$.

- Alternately, just verify that $\pi(i) = (1/2)^{i+1}$ satisfies the equations above:

$$\begin{aligned}
i = 0: &\quad 1 = (2/3)(1) + (2/3)(1/2) \text{ (yes)} \\
i \geq 1: &\quad (1/2)^{i+1} = (1/3)(1/2)^i + (2/3)(1/2)^{i+1} \text{ (yes)}
\end{aligned}$$

Result

- Can apply the Ergodic theorem for Harris chains to say that

$$\lim_{t \to \infty} \mathbb{P}(X_t = i | X_0 = x) = (1/2)^{i+1}$$

  for all $i \geq 0$.

Proving the Ergodic Theorem for Harris chains

Encode the advancing X_t algorithm as a stochastic process

- Recall

$$\tau_x(D) = \mathbb{P}(X_1 \in D | X_0 = x)$$
$$\psi_x(D) = (\tau_x(D) - \epsilon\rho(D))/(1 - \epsilon)$$

- Given $X_t = x$, let

$$W_t \leftarrow \tau_x$$
$$Y_t \leftarrow \rho$$
$$Z_3 \leftarrow \psi_x$$
$$B_t \leftarrow \mathsf{Bern}(\epsilon)$$

Then

$$X_{t+1} = W_{t+1}\mathbb{1}(X_t \notin A) + \mathbb{1}(X_{t+1} \in A)[Y_{t+1}B_{t+1} + Z_{t+1}(1 - B_{t+1})].$$

Proof of Ergodic Theorem for Harris chains. Fix x. Let $X_0 = x$ and $Y_0 \sim \pi$. Then the probability that X_t and Y_t hits A in finite time is 1. Since the chain is recurrent, Y_t will hit A infinitely often with probability 1 after that. After the first time X_t hits A, there exists n such that for $n' \geq n$, $\mathbb{P}(X'_n \in A) > 0$.

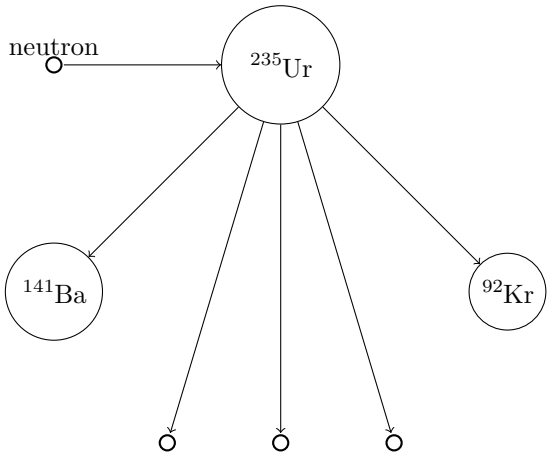
Taken together, that means there is a time t such that $\mathbb{P}(Y_t \in A, X_t \in A) = c > 0$. If $B_{t+1} = 1$ (which happens with probability ϵ), then that $X_{t+1} = Y_{t+1}$. After m such encounters, the chance that $X_t \neq Y_t$ is at most $(1 - c)^m$. Since this goes to 0 as m goes to infinity, as t goes to infinity the chance that X_t and Y_t have not met goes to 0. □

26 Branching processes and fission bombs

Question of the Day How much U-235 do you need to create a fission bomb?

How fission bombs work

- Need a collection of atoms that are prone to splitting in half.
- Uranium-235 (aka U-235 or ^{235}U) has 92 protons and 143 neutrons.
- When an extra neutron is added, you get extremely unstable U-236
- It quickly splits into Barium-141 and Krypton-92 and three more neutrons



- Hopefully those neutrons go on to split more U-235.
- Called a *chain reaction*.
- If chain reaction dies out, the bomb is fizzles.
- Otherwise, BOOM!

Branching process

- Given a population of individuals:

  1: Each individual has a random # of children.

  2: Each individual has children independently.

- Neutrons are an example of a branching process.
- Each neutron has 0 or 3 children.
- Each of those goes on to have a random number of children.
- Suppose there are six neutrons at stage t ($X_t = 6$).
- Then number of neutrons at stage $t+1$:

$$X_{t+1} = \sum_{i=1}^{6} D_i,$$

  where D_i are iid and in $\{0, 3\}$.

Definition 64
A **branching process** is a special kind of Markov chain with state space $\{0, 1, 2, \ldots\}$ where for $t, i \geq 0$, $Y_{t,i}$ are iid and

$$X_{t+1} = \sum_{i=1}^{X_t} Y_{t+1,i}.$$

- Let $\mu = \mathbb{E}[Y_{t,i}]$ denote the average # of children an individual has.
- Then
$$\mathbb{E}[X_t] = \mathbb{E}[\mathbb{E}[X_t|X_{t-1}]] = \mathbb{E}[X_{t-1}\mu] = \mu\mathbb{E}[X_{t-1}].$$
- This forms the basis of an induction proof that
$$\mathbb{E}[X_n] = \mu^n \mathbb{E}[X_0].$$

Fact 48
If $\mu < 1$, then as $n \to \infty$, $\mathbb{E}[X_n] \to 0$, and $\mathbb{P}(X_n > 0) \to 0$.

Proof. Since $X_n \geq 0$, Markov's inequality applies: $\mathbb{P}(X_n \geq 1) \leq \mathbb{E}[X_n]/1$. Since $\mu < 1$, as $n \to \infty$,

$$\mu^n \to 0 \Rightarrow \mathbb{E}[X_n] \to 0 \Rightarrow \mathbb{P}(X_n \geq 1) \to 0.$$

□

Definition 65
A branching process goes **extinct** if there is some n for which $X_n = 0$.

When $\mu < 1$, the processes goes extinct with probability 1. What about when $\mu = 1$? $\mu > 1$? Notation:

$$\begin{aligned} a_n(k) &= \mathbb{P}(X_n = 0|X_0 = k) \\ a(k) &= \lim_{n\to\infty} a_n(k) \\ p(i) &= \mathbb{P}(X_1 = i|X_0 = 1). \end{aligned}$$

- So $a(k)$ is the probability that you go extinct starting with k people.
- Recall that individuals reproduce independently.
- So for k people to go extinct, each individuals line must fail.
- Hence $a(k) = a(1)^k$.
- If $p(i) = 0$, then never go extinct (everyone has at least one child).
- From now on, assume $p(0) > 0$.

Definition 66
The probability the stochastic process goes extinct is the **extinction probability**.

[Let $a = a(1)$ be the extinction probability.]
Let's do a little first step analysis!

$$\begin{aligned} a &= \mathbb{P}(\exists n : X_n = 0|X_0 = 1) \\ &= \sum_i \mathbb{P}(\exists n : X_n = 0|X_1 = i)\mathbb{P}(X_1 = i) \\ &= \sum_i a(i)p(i) = \sum_i a^i p(i) = \mathbb{E}[a^{Y_{1,1}}]. \end{aligned}$$

Definition 67
For $a \neq 0$, $\text{gf}_X(a) = \mathbb{E}[a^X]$ is the **generating function** of the random variable X. For $a = 0$, $\text{gf}_X(0) = \mathbb{P}(X = 0)$.

Notes

- $\text{gf}_X(0) = \mathbb{P}(X = 0)$ makes generating function continuous where it exists.
- Recall the moment generating function is $\text{mgf}_X(t) = \mathbb{E}[e^{tX}]$.
- So generating function is mgf with $a = e^t$.

Notation: let $\phi(a) = \mathbb{E}[a^{Y_{1,1}}]$.

Fact 49 **1:** $\phi(0) = p(0)$.

2: $\phi'(1) = \mu$.

3: $\phi(1) = 1$.

4: $\phi''(a) \geq 0$.

Proof. $\phi(0) = p(0)$ is the convention for generating functions at 0.

The second fact needs that you can differentiate power series term by term inside their radius of convergence. Since $\phi(1) = \mathbb{E}[1] = 1$, the radius of convergence is at least 1. Hence for all $a < 1$,

$$\begin{aligned}\phi'(a) &= (d/da)[p_0 + \sum_{i=1}^{\infty} a^i p_i \\ &= \sum_{i=1}^{\infty} i a^{i-1} p_i = \mathbb{E}[X \cdot a^{\max(X-1,0)}]\end{aligned}$$

For $a < 1$ this is bounded above by $\mathbb{E}[X] = \mu$, so the dominated convergence theorem applies and taking the limit as $a \to 1^-$ gives $\phi'(1) = \mu$.

The third fact is just

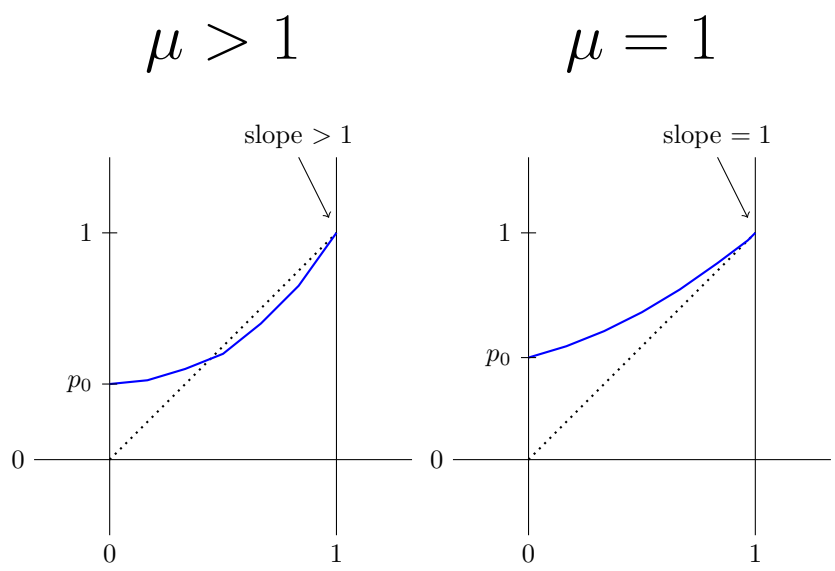
$$\phi(1) = \mathbb{E}[1^X] = 1,$$

and the last fact differentiates term by term again to get:

$$\phi''(a) = \sum_{i=2}^{\infty} (i)(i-1) a^{i-2} p_i \geq 0.$$

□

So the picture looks like:



When $\mu > 1$, $a = \phi(a)$ at $a = 1$, and once for $a \in (0, 1)$. When $\mu = 1$, $a = \phi(a)$ only at $a = 1$.

What does this mean for the bomb?

- Every neutron that fires has a positive chance of never going extinct.
- So positive chance of causing a chain reaction. Boom!
- Need greater than $1/3$ chance of causing a fission event to have positive probability of chain reaction.

27 Calculating with generating functions

Question of the Day A branching process has $p_0 = 1/2$, $p_1 = 1/3$, $p_2 = 1/6$. What is the chance a population of 1 is extinct after 5 generations?

Today

- Generating functions for random sum of random variables.

Summing two fair six sided dice

- Let $X, Y \sim \mathsf{Unif}(\{1,2,3,4,5,6\})$ be independent.
- What is $\mathrm{gf}_{X+Y}(a)$?

$$\mathrm{gf}_{X+Y}(a) = \mathbb{E}[a^{X+Y}] = \mathbb{E}[a^X a^Y] = \mathbb{E}[a^X]\mathbb{E}[a^Y] = \mathrm{gf}_X(a)\,\mathrm{gf}_Y(a).$$

$$\begin{aligned}&[(1/6)(a + a^2 + a^3 + a^4 + a^5 + a^6)]^2\\ &\quad = (1/36)(a^2 + 2a^3 + 3a^4 + 4a^5 + 5a^6 + 6a^7 + 5a^8 +\\ &\qquad 4a^9 + 3a^{10} + 2a^{11} + a^{12}),\end{aligned}$$

so $\mathbb{P}(X + Y = 5) = 4/36 = 0.1111\ldots$.

Summing a random number of six sided dice

- Let $X_1, X_2, X_3 \overset{\text{iid}}{\sim} X \sim \mathsf{Unif}(\{1, \ldots, 6\})$.
- Let $N \sim \mathsf{Unif}(\{1,2,3\})$.
- What is the generating function of $S = \sum_{i=1}^{N} X_i$?

$$\begin{aligned}\mathrm{gf}_S(a) &= \mathbb{E}[a^S]\\ &= \mathbb{E}[\mathbb{E}[a^S|S]]\\ &= \mathbb{E}[a^S|S=1]\mathbb{P}(S=1) + \mathbb{E}[a^S|S=2]\mathbb{P}(S=2) +\\ &\quad \mathbb{E}[a^S|S=3]\mathbb{P}(S=3)\\ &= \mathrm{gf}_X(a)^1(1/3) + \mathrm{gf}_X(a)^2(1/3) + \mathrm{gf}_X(a)^3(1/3)\\ &= \mathrm{gf}_N(\mathrm{gf}_X(a)).\end{aligned}$$

Fact 50
Suppose $X_1, X_2, \ldots$ are iid with distribution the same as X. Then for any nonnegative integer valued N,

$$\mathrm{gf}_{X_1+\cdots+X_N}(a) = \mathrm{gf}_N(\mathrm{gf}_X(a)).$$

Applying to branching processes

- Recall: $X_n = Y_1 + \cdots + Y_{X_{n-1}}$, where $Y_i \overset{\text{iid}}{\sim} Y$.
- So

$$\begin{aligned}\mathrm{gf}_{X_n} &= \mathrm{gf}_{X_{n-1}} \circ \mathrm{gf}_Y\\ &= (\mathrm{gf}_{X_{n-2}} \circ \mathrm{gf}_Y) \circ \mathrm{gf}_Y\\ &= \vdots\\ &= \mathrm{gf}_Y \circ \mathrm{gf}_Y \circ \cdots \circ \mathrm{gf}_Y\end{aligned}$$

- Recall: for a r.v. X, $\mathrm{gf}_X(0) = \mathbb{P}(X = 0)$.
- So $\mathbb{P}(X_n = 0) = [\mathrm{gf}_Y \circ \mathrm{gf}_Y \circ \cdots \circ \mathrm{gf}_Y](0)$.

Question of the day

- $\text{gf}_Y = 1/2 + (1/3)a + (1/6)a^2$.
- So $\text{gf}_{X_1}(0) = \text{gf}_Y(0) = 1/2$.
- There is a $1/2$ chance that $X_1 = 0$.
- Now it gets interesting...

$$\begin{aligned}\text{gf}_{X_2}(0) &= \text{gf}_Y(\text{gf}_Y(0)) = \text{gf}_Y(1/2) = (1/2) + (1/3)(1/2) + (1/6)(1/2)^2 \\ &= (12 + 4 + 1)/24 = 17/24.\end{aligned}$$

- Can keep going and make a table:

n	0	1	2	3	4	5
$\mathbb{P}(X_n = 0)$	0	0.5	0.7083	0.8197	0.8852	0.9256

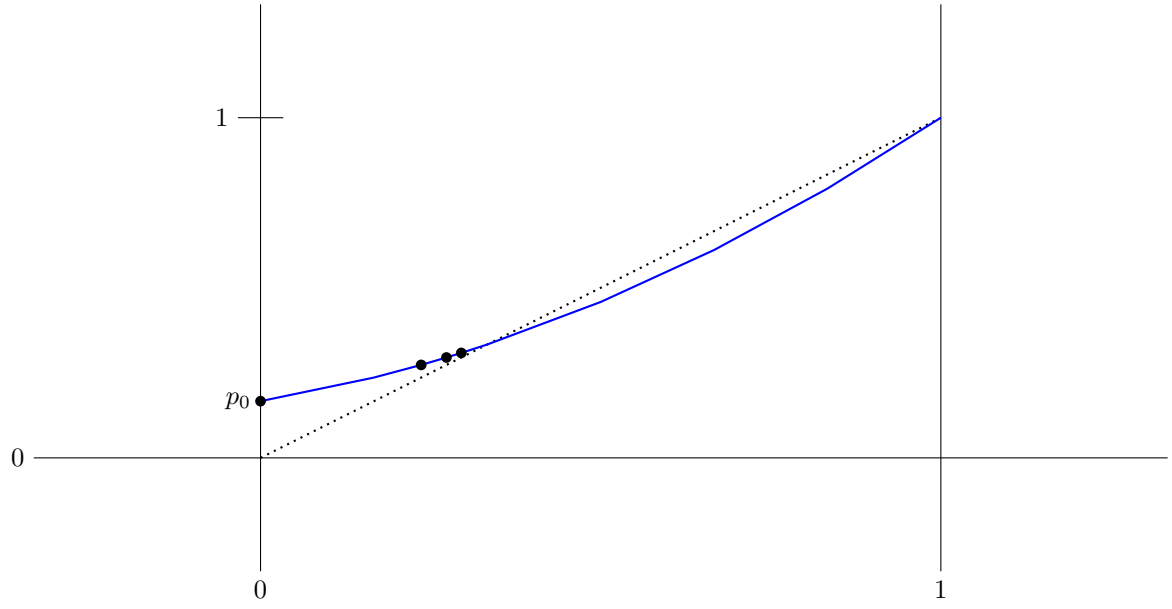
- After 5 generations, there is $\boxed{92.56\%}$ chance of being extinct.
- Here $\mu = (1/3)(1) + (1/6)(2) < 1$, so $a = 1$.

Finding the extinction probability

- Suppose $\text{gf}_Y(a) = (1/6) + (1/3)a + (1/2)a^2$.
- Using the same procedure as before:

n	0	1	2	3	4	5
$\mathbb{P}(X_n = 0)$	0	0.1666	0.2361	0.2732	0.2950	0.3085

- Looking at the $\text{gf}_Y(a)$ curve:



- As long as you start out below point where $a = \text{gf}_Y(a)$, will converge to that point.
- This gives the following fact.

Fact 51
For a branching process with:

condition	extinction probability
$p_0 = 0$	$a = 0$
$p_0 \in [0,1),\ \mu \leq 1$	$a = 1$
$p_0 \in [0,1),\ \mu > 1$	unique solution to $a = \mathrm{gf}_Y(a)$ in $(0,1)$.

Notes

- For general Y, easiest to just use convergence.
- When Y small polynomial, can solve exactly:

$$\begin{aligned} a &= (1/6) + (1/3)a + (1/2)a^2 \\ 0 &= (1/6) - (2/3)a + (1/2)a^2 \\ 0 &= 1 - 4a + 3a^2 \\ 0 &= (a-1)(3a-1) \text{ (since } (a-1) \text{ always a factor)} \\ a &= 1/3. \end{aligned}$$

What happened with the German bomb?

- Werner Heisenberg (of uncertainty principle fame) in charge of project.
- Made enormous mistake in calculation.
- Told German high command needed several tons of U-235 to make it work.
- With proper branching process analysis: 60 kg.
- Later told the British that he lied on purpose so Nazi's wouldn't get the weapon.

28 Brownian Motion

Question of the Day What happens to simple random walk as the time step goes to 0?

Today

- Brownian Motion
- Scaling limit of simple random walk.

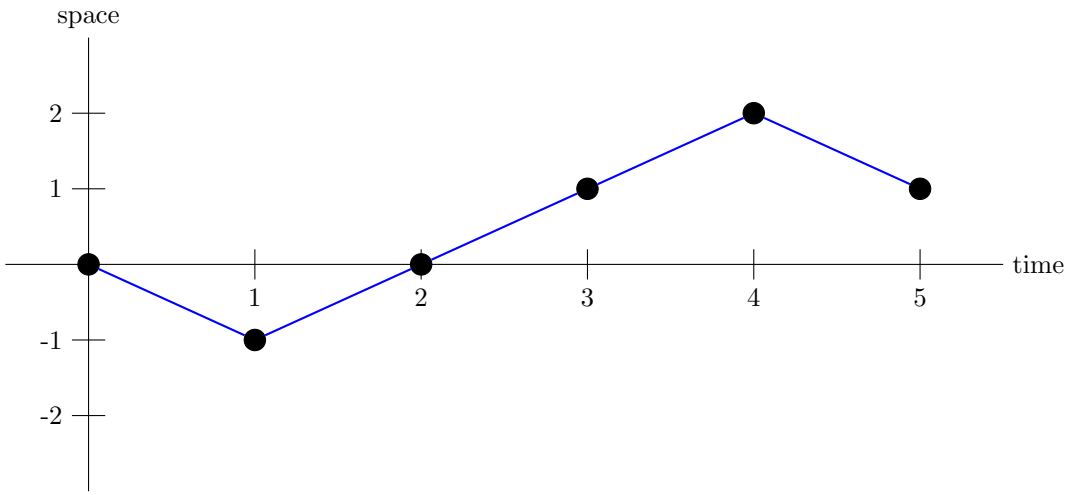
History

- In 1595, microscope was invented.
- In 1696, Gray noticed small grains suspended in fluids were moving.
- In 1827, Robert Brown saw pollen grains moving around.
- Idea then was that only organic material moved on its own.
- Brown also saw movement from inorganic material!
- Late 19th century: is matter continuous or discontinuous?
- (Thought that Newton had settled question on continuous side.)
- In 1900, Bachelier modeled stock prices with Brownian motion.
- First mathematical treatment of phenomenon.
- In 1905, Einstein has his "Miracle Year"
  - **1:** Photoelectric Effect (1921 Nobel Prize)
  - **2:** Brownian Motion displacement prediction
  - **3:** Special Relativity
  - **4:** Mass-Energy Equivalence ($E = mc^2$).
- Perrin actually measured displacement in 1908 (1926 Nobel Prize).

Simple symmetric random walk on $\mathbb{Z}$

$$Z_{t+1} = \begin{cases} Z_t + 1 & \text{with probability } 1/2 \\ Z_t - 1 & \text{with probability } 1/2 \end{cases}$$

A picture might look something like this:



Another description:

$$D_1, D_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Unif}(\{-1, 1\})$$
$$X_t = \sum_{i=1}^{t} D_i$$

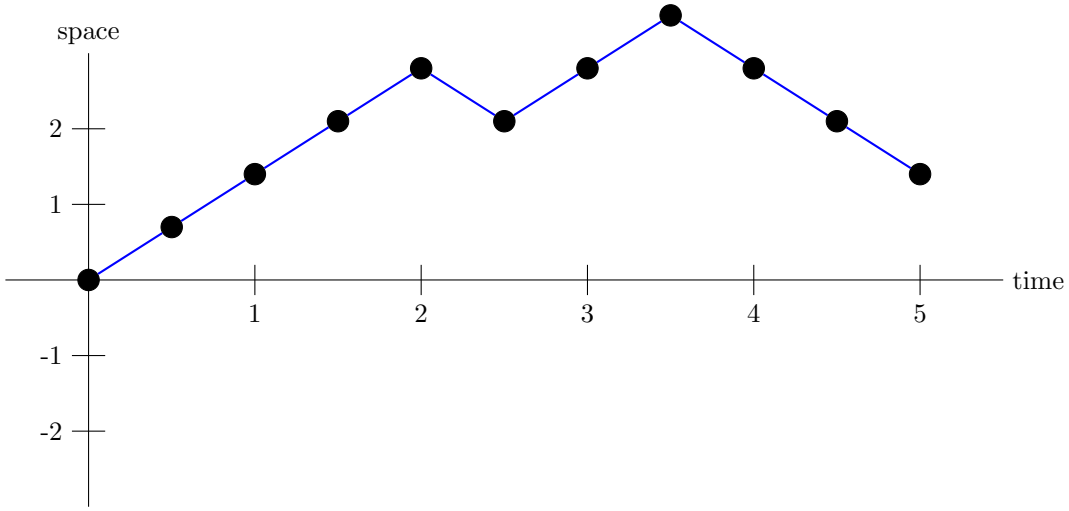
That allows us to calculate mean and variance:

$$\mathbb{E}[X_t] = \mathbb{E}(\sum D_i) = \mathbb{E}[D_1] + \cdots + \mathbb{E}[D_t] = t\mathbb{E}[D_1] = 0$$
$$\mathbb{V}[X_t] = \mathbb{V}(\sum D_i) = \mathbb{V}[D_1] + \cdots + \mathbb{V}[D_t] = t\mathbb{V}[D_1].$$

Now $\mathbb{V}[D_1] = \mathbb{E}[D_1^2] - \mathbb{E}[D_1]^2 = 1$, and standard deviation is square root of variance, so

$$\mathrm{SD}(X_t) = \sqrt{t}\,\mathrm{SD}(D_1) = \sqrt{t}.$$

Now scale it so using time steps of length $1/2$:



So

$$H_{1/2}, H_1, H_{3/2}, \ldots \overset{\text{iid}}{\sim} ?$$
$$X_t = \sum_{i \le t} H_i$$

Adding $2t$ of the H_i together to get X_t.

$$\mathrm{SD}(X_t) = \sqrt{2t}\,\mathrm{SD}(H_i),$$

so to make $\mathrm{SD}(X_t) = \sqrt{t}$ as before, $\mathrm{SD}(H_i) = 1/2$. Simple way:

$$H_i \sim \mathsf{Unif}(\{-\sqrt{1/2}, \sqrt{1/2}\}).$$

So still have

$$\mathbb{E}[X_t] = 0,\ \mathrm{SD}(X_t) = \sqrt{t}.$$

Make time step h. Add t/h of H_i together to get X_t,

$$\mathrm{SD}(X_t) = \sqrt{t/h}\,\mathrm{SD}(H_i),$$

so make $\mathrm{SD}(H_i) = \sqrt{h}$. Simple way:

$$H_i \sim \mathsf{Unif}(\{-\sqrt{h}, \sqrt{h}\}).$$

Check

$$\mathbb{V}(H_i) = \mathbb{E}[H_i^2] - \mathbb{E}[H_i]^2 = h - 0 = h.$$

Idea: Brownian motion is the limit as $h \to 0$

- Note that X_t is sum of iid random variables.
- Central limit theorem: X_t approximately normal, $N(\mathbb{E}(X_t), \mathrm{SD}(X_t)^2)$.
- This gives rise to the mathematical notion of Brownian Motion.

Definition 68
Say that a stochastic process B_t is **standard Brownian Motion** if

1: Centered: $B_0 = 0$.

2: Independent increments: for all $a < b \leq c < d$, $B_d - B_c$ and $B_b - B_a$ are independent random variables.

3: Normal increments: for all $a < b$,

$$B_b - B_a \sim \mathsf{N}(0, b-a).$$

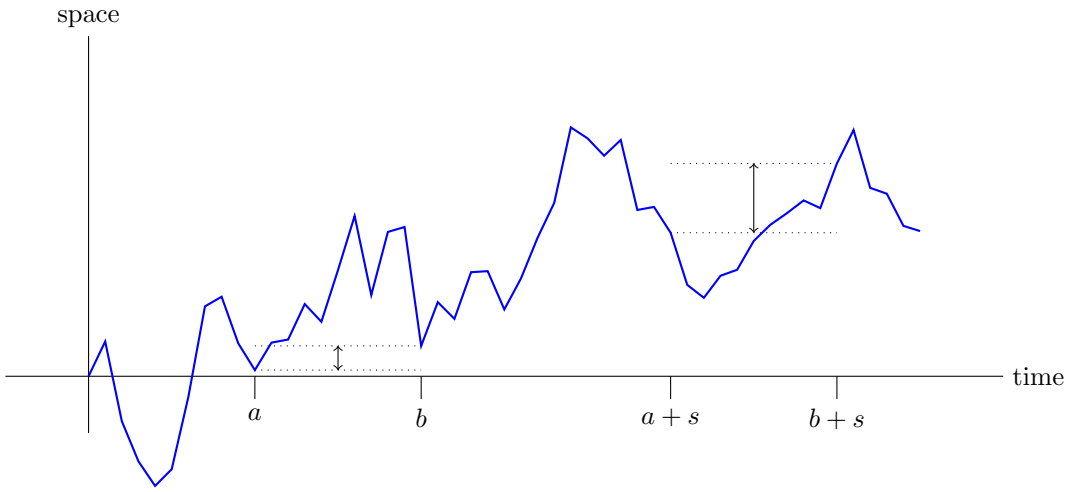
4: Continuity: the process B_t is continuous with probability 1.

Notes

- Haven't actually proved that Brownian motion exists.
- Definition is just "something with these properties" is Brownian Motion.
- Existence beyond the scope of this course.
- With this definition, can prove facts about Brownian motion consistent with its use in models.

Definition 69
A stochastic process is **stationary** (more precisely, has *stationary increments* if $\forall a < b$ and $s > 0$,

$$X_b - X_a \sim X_{b+s} - X_{a+s}$$



Change from a to b same distribution as change from $a + s$ to $b + s$

Fact 52
Standard Brownian motion is stationary.

Proof. Let $a < b$ and $s > 0$. Then $X_b - X_a \sim \mathsf{N}(0, b-a)$ and $X_{b+s} - X_{a+s} \sim \mathsf{N}(0, (b+s)-(a+s))$, which is the same distribution. □

Definition 70
A **Lévy Process** is a stochastic process with independent and stationary increments.

Although Brownian motion is continuous, it is not differentiable!

Fact 53
With probability 1, Brownian motion is not differentiable anywhere.

Proof idea: A formal proof is beyond this course, but here's the intuition. Recall that the derivative of X_t is

$$\lim_{h \to 0} \frac{X_{t+h} - X_t}{h}.$$

For Brownian Motion, $X_{t+h} - X_t \sim \mathsf{N}(0, h)$, and so has standard deviation of $\sqrt{h}$. A normal random variable has at least a 31% chance of being at least one standard deviation away from the mean. So if $|X_{t+h} - X_t| \geq \sqrt{h}$, $|X_{t+h} - X_t|/h = 1/\sqrt{h}$ which goes to infinity as $h \to 0$! This prevents the derivative from existing. □

29 Simulation of Brownian Motion

Question of the Day Suppose $Z_1, Z_2, Z_3 \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$. How can I create Brownian motion at times $t \in \{0,2,4,6\}$ as a function of Z_1, Z_2, Z_3?

Today

- Markovian processes.
- Simulating Brownian motion from one endpoint.
- Interpolating Brownian motion.

For discrete time

- Markov chains have for integer n and measurable A

$$\mathbb{P}(X_{n+1} \in A | \mathcal{F}_n) = \mathbb{P}(X_{n+1} \in A | X_n).$$

- No memory.
- Brownian motion also memoryless, but now in continuous time.

Definition 71
A stochastic process $\{X_t\}_{t \geq 0}$ has the **Markov property** if for all $s \leq t$ and measurable A,

$$\mathbb{P}(X_t \in A | \mathcal{F}_s) = \mathbb{P}(X_t \in A | X_s).$$

Fact 54
Brownian motion has the Markov property.

This allows us to simulate Brownian motion!

Simulation=for a finite set of times $\{t_1, \ldots, t_n\}$, draw

$$B_{t_1}, B_{t_2}, \ldots, B_{t_n}.$$

For $t \in \{0,2,4,6\}$, B_0, $B_2 - B_0$, $B_4 - B_2$, $B_6 - B_2$ are independent

$B_0 = 0$, $B_2 - B_0 \sim \mathsf{N}(0,2)$, $B_4 - B_2 \sim \mathsf{N}(0,2)$, $B_6 - B_4 \sim \mathsf{N}(0,2)$.

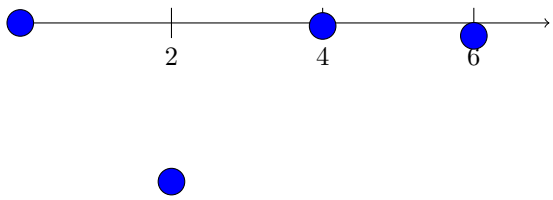
Example

- Suppose

$$Z_1 = -1.484, \; Z_2 = 1.456, \; Z_3 = -0.09262.$$

- Recall for $c \in \mathbb{R}$, $\mathbb{V}(cX) = c^2 \mathbb{V}(X)$.
- So $Z_i \sim \mathsf{N}(0,1) \Rightarrow \sqrt{2}Z \sim \mathsf{N}(0,2)$.

$$\begin{aligned}
B_0 &= 0 \\
B_2 &= B_0 + (B_2 - B_0) = 0 + \sqrt{2}(-1.484) = -2.099 \\
B_4 &= B_2 + (B_4 - B_2) = -2.099 + \sqrt{2}(1.456) = -0.03991 \\
B_6 &= B_4 + (B_6 - B_4) = -0.03991 + \sqrt{2}(-0.09262) = -0.1706
\end{aligned}$$



- That is only 4 out of an infinite number of values for B_t!

Filling in times

- Now suppose I also want to know B_1,
  given $B_0 = 0$ and $B_2 = -2.099$. Note:

$$\begin{aligned}[B_1|B_2 = b] &\sim [B_1 - B_0|(B_2 - B_1) + (B_1 - B_0) = b] \\ &\sim [Z_1|Z_1 + Z_2 = b],\end{aligned}$$

  where $Z_1, Z_2 \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$.

Fact 55
For X and Y continuous with densities f_X and f_y with respect to Lebesgue measure,

$$f_{X|X+Y=s}(x) \propto f_X(x) f_Y(s - x)$$

Adding normals

- Suppose $t \in [t_1, t_2]$, so $t = \lambda t_1 + (1 - \lambda)t_2$.
- Suppose $X \sim \mathsf{N}(0, \lambda(t_2 - t_1))$
- $Y \sim \mathsf{N}(0, (1 - \lambda)(t_2 - t_1))$
- (Then $X + Y \sim \mathsf{N}(0, t_2 - t_1) \sim B_{t_2} - B_{t_1}$)

$$\begin{aligned}f_{X|X+Y=s}(x) &\propto \exp\left(-\frac{x^2}{2\lambda(t_2 - t_1)}\right) \exp\left(-\frac{(s-x)^2}{2(1-\lambda)(t_2 - t_1)}\right) \\ &= \exp\left(-\frac{(1-\lambda)x^2 + \lambda(s-x)^2}{2\lambda(1-\lambda)(t_2 - t_1)}\right) \\ &= \exp(\left(-\frac{x^2 - 2sx\lambda - \lambda s^2 + (\lambda s)^2 - (\lambda s)^2}{2\lambda(1-\lambda)(t_2 - t_1)}\right) \\ &\propto \exp\left(-\frac{(x - s\lambda)^2}{2\lambda(1-\lambda)(t_2 - t_1)}\right)\end{aligned}$$

  which means

$$[X|X + Y = s] \sim \mathsf{N}(\lambda s, \lambda(1 - \lambda)(t_2 - t_1)).$$

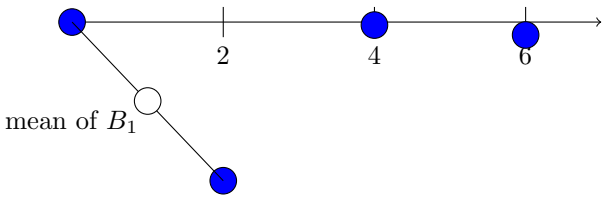
- Recall $X = B_t - B_{t_1}$ and $Y = B_{t_2} - B_t$, so

$$[B_t - B_{t_1}|B_{t_1}, B_{t_2}] \sim \mathsf{N}(\lambda s, \lambda(1 - \lambda)(t_2 - t_1))$$

- This gives the following fact!

Fact 56
Suppose $t_1 < t_2$. For $\lambda \in [0,1]$ and $t = (1-\lambda)t_1 + \lambda t_2$:

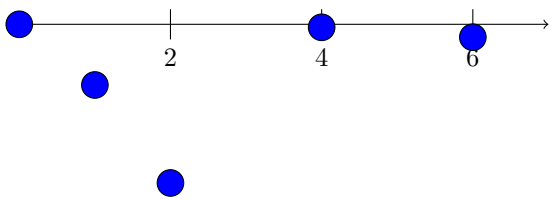
$$[B_t | B_{t_1}, B_{t_2}] \sim \mathsf{N}(B_{t_1} + \lambda(B_{t_2} - B_{t_1}), \lambda(1-\lambda)(t_2 - t_1)).$$



Here is the intuition behind This result. Pretend that a Hooke's law spring is connecting B_t to B_{t-1}. Given B_{t_1}, B_{t_2}, one spring attaches B_t to B_{t_1}, and another to B_{t_2}. Each is pulling B_t towards it, and so the overall give in the spring is lessened.

Going back to the earlier problem

- Suppose $Z \sim \mathsf{N}(0,1)$ is $Z = 0.349643$.
- Use this to find $B_1 | B_0 = 0, B_2 = -2.099779$.
- Here $\lambda = (1-0)/(2-0) = 1/2$.
- So $B_1 - B_0 \sim \mathsf{N}(1/2(0) + (1/2)(-2.099779), (1/2)(1/2)(2))$.
- So $B_1 = -2.099779/2 + (1/2)^{1/2} Z = -0.8026545$.



Summary Given the Brownian Motion at times $\{t_1, \ldots, t_n\}$, $Z \sim \mathsf{N}(0,1)$, and a new time t, find B_t as follows:

1: When $t > t_n$,

$$B_{t_n} + (t - t_n)^{1/2} Z.$$

2: When $t = (1-\lambda)t_i + \lambda t_{i+1}$ for some $\lambda \in [0,1]$

$$(1-\lambda)B_{t_i} + \lambda B_{t_{i+1}} + [\lambda(1-\lambda)(t_{i+1} - t_i)]^{1/2} Z.$$

3: When $t < t_1$:

$$B_{t_1} + (t_1 - t)^{1/2} Z.$$

[Note: Brownian motion looks the same when run forward in time or backwards in time. Called a *reversible* process.]

Proof of Fact 55 [With the added assumption that the partial derivatives of the densities are continuous.] For any continuous random variables X and S with joint pdf $f_{X,S}(x,s)$:

$$f_{X|S=s}(x) = \frac{f_{X,S}(x,s)}{f_S(s)} \propto f_{X,S}(x,s).$$

So the key is finding the joint density of X and S when $S = X + Y$. The joint pdf is perhaps easier:

$$\begin{aligned}\mathbb{P}(X \le a, S \le b) &= \mathbb{P}(X \le a, X + Y \le b) \\ &= \int_{(x,y):x\le a \wedge x+y\le b} f_{X,Y}(x,y)\ d\mathbb{R}^2 \\ &= \int_{-\infty}^{a}\int_{-\infty}^{b-a} f_{X,Y}(x,y)\ dx\ dy\end{aligned}$$

We can make this an iterated integral by either Tonelli (since nonnegative) or Fubini (since the absolute integral is at most 1.)

With the assumption that the partial derivatives of the integrands are continuous everywhere, the partial derivatives with respect to b can be placed inside the integral to give

$$\frac{\partial}{\partial b}\mathbb{P}(X \le a, S \le b) = \int_{-\infty}^{a} f_{X,Y}(x, b-a)\ dx,$$

and then

$$\frac{\partial^2}{\partial a \partial b}\mathbb{P}(X \le a, S \le b) = f_{X,Y}(a, b-a)\ dx.$$

Hence $f_{X,S}(a,b) = f_{X,Y}(a, b-a)$, and we are done.

30 Poisson point processes

Question of the Day A concrete slab one meter in length has defects scattered as a Poisson point process with rate 2 per meter. What is the chance that the slab has no defects whatsoever?

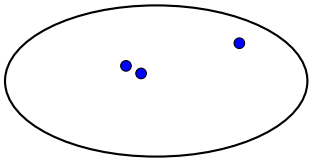
Today

- Poisson point processes (PPP)
- Poisson process
- Arrival and Interarrival times

30.1 Poisson point process on $\mathbb{R}^n$

What is a Poisson point process

- A set of points in $\mathbb{R}^d$



What does *rate* mean?

- Controls frequency of defects.
- Typically use λ for the rate.
- High λ = many defects likely.
- Low λ = few defects likely.

Definition 72
A random collection P of points in $\mathbb{R}^n$ is a **Poisson point process** with measure μ if

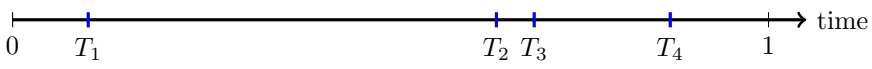
1: For any measurable A

$$\mathbb{E}[\#\{P \cap A\}] = \mu(A)$$

2: For any measurable A, B with $A \cap B = \emptyset$, $\#(P \cap A)$ and $\#(P \cap B)$ are independent.

When is PPP a good model?

- Material defects
- Times of radioactive decay
- Misprints in a book
- Arrival times of people to a queue.



- $T_1, T_2, \ldots$ called the *arrival times*.
- The times between arrivals, $T_1 - 0, T_2 - T_1, T_3 - T_2, \ldots$ are called *interarrival times*.
- [Interstates travel between states...interarrival times are between arrival times.]

Mean and variance grow linearly with $\mu(A)$

- For $A \cap B = \emptyset$,

$$\mathbb{E}[\#(P \cap (A \cup B))] = \mathbb{E}[\#(P \cap A)] + \mathbb{E}[\#(P \cap B)]$$
$$\mathbb{V}[\#(P \cap (A \cup B))] = \mathbb{V}[\#(P \cap A)] + \mathbb{V}[\#(P \cap B)]$$

- Only one distribution on $\{0, 1, 2, \ldots\}$ has mean and variance growing the same way...

Definition 73
X has a **Poisson distribution with mean μ** (written $X \sim \mathsf{Pois}(\mu)$) if for all $i \in \{0, 1, 2 \ldots\}$

$$\mathbb{P}(X = i) = \exp(-\mu)\frac{\mu^i}{i!}.$$

Fact 57
For $N \sim \mathsf{Pois}(\mu)$, $\mathbb{E}[N] = \mu$, $\mathbb{V}(N) = \mu$. If $X \sim \mathsf{Pois}(\mu_X)$ and $Y \sim \mathsf{Pois}(\mu_Y)$ are independent, then $X + Y \sim \mathsf{Pois}(\mu_X + \mu_Y)$.

Fact 58
Let P be a Poisson point process of rate λ over $\mathbb{R}^n$. Then $\#(P \cap A) \sim \mathsf{Pois}(\lambda \cdot \text{Lebesgue}(A))$.

QotD

- Rate $= 2/\text{m}$, measure $= 1$ m, so $\mu(A) = (2/\text{m})(1\text{ m}) = 2$
- So $\mathbb{P}(\#(P \cap [0, 1]) = 0) = \exp(-2) \approx 0.1353$.

Conditioning on number of points

- Let A and B be disjoint with equal measure. Then for $P = \{X_1, \ldots, X_n\}$, X_i equally likely to be in A or B.

Fact 59
Let $n = \#P$ and $P = \{X_1, \ldots, X_n\}$. Then the X_i are iid $\mathsf{Unif}(A)$.

Example

- Q: Suppose the marble is known to have 3 defects. What is the chance that all the defects lie in $[1/2, 1]$?
- A: Each of the three defects (independently) has a $1/2$ chance of falling into $[1/2, 1]$, so the chance that all three do is $(1/2)^3 = 1/8 = 0.1250$.

Fact 60
Let $A \subseteq B$ be measurable. Then for P a PPP of measure μ,

$$[\#(P \cap A) | \#(P \cap B) = n] \sim \mathsf{Bin}(n, \mu(A)/\mu(B)).$$

30.2 Poisson processes on $[0, \infty)$

Sorting the points

- For points on $[0, \infty)$, they can be sorted:

$$P = \{T_1, T_2, T_3, \ldots\},$$

where $T_i < T_j$ for $i \leq j$.

What is the distribution of T_1?

- Let $a > 0$. Then

$$\begin{aligned}\mathbb{P}(T_1 \le a) &= \mathbb{P}(\#(P \cap [0,a]) > 0) = 1 - \mathbb{P}(\#(P \cap [0,a]) = 0) \\ &= 1 - \exp(-\lambda a).\end{aligned}$$

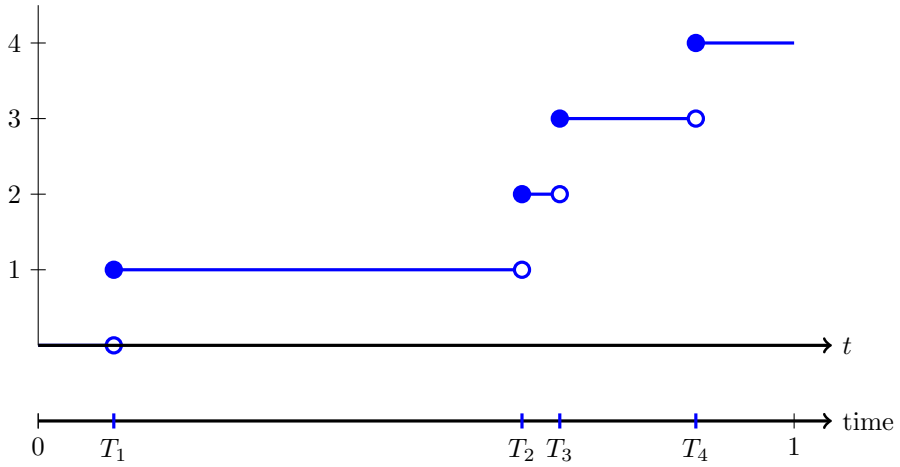
- This is cdf of $\mathsf{Exp}(\lambda)$.

Fact 61
The distribution of $T_{i+1} - T_i \sim \mathsf{Exp}(\lambda)$. The distribution of T_i is $\mathsf{Gamma}(i, \lambda)$, or $\mathsf{Erlang}(i, \lambda)$.

Definition 74
Let $T_1 < T_2 < \cdots$ be a PPP on $[0, \infty)$. Then let $T_0 = 0$, and $A_i = T_i - T_{i-1}$ form the **interarrival times** of the process.

Definition 75
For $t > 0$ and P a PPP, let $N_t = \#(P \cap [0,t])$. Then $\{N_t\}$ is a **Poisson process**

The following picture shows a Poisson point process on the bottom, and the associated Poisson process on the top:



$$P = \{0.10\ldots, 0.64\ldots, 0.69\ldots, 8.7\ldots, \ldots\}$$

Example A web page receives hits at rate 4 per minute. Suppose that five hits are received in the first two minutes. What is the chance that exactly three of them arrived in the first minute?

- $A = [0,1]$, $B = [0,2]$, $\mu(A)/\mu(B) = 1/2$,

$$\mathbb{P}(N_1 = 3 | N_2 = 5) = \binom{5}{3}(1/2)^3(1/2)^2 = 5/16 \approx \boxed{0.3125}.$$

Note, points in $[0,t]$ and (t, ∞) are independent of each other...

Fact 62
Poisson processes $\{N_t\}$ are Markovian.

Recall that X_t is a Lévy process if

1: Increments are stationary: for $a < b$, $c < d$, and $d - c = b - a$:

$$X_b - X_a \sim X_d - X_c.$$

2: Increments are independent: for $a \leq b \leq c \leq d$:

$$X_b - X_a \text{ is independent of } X_d - X_c.$$

Fact 63
A Poisson process of rate λ is a Lévy process where for $a < b$,

$$N_b - N_a \sim \mathsf{Pois}(\lambda(b-a)).$$

Problems

30.1: Suppose that the arrivals of airport shuttles at a particular stop follow a Poisson process of rate 1/[10 min].

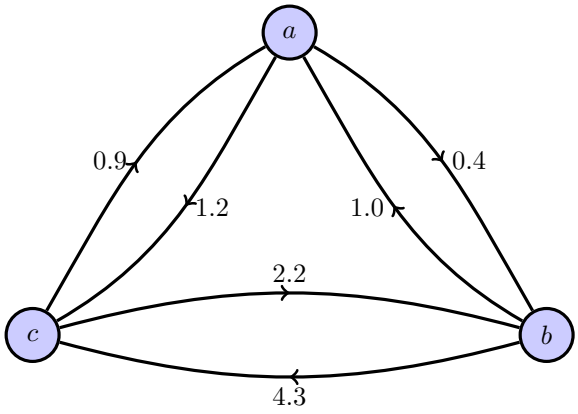
(a) On average, how many shuttles will arive in an hour?

(b) What is the chance that there are no shuttles in the first 20 minutes?

(c) What is the chance that the second shuttle arrives somewhere in $[15, 25]$ minutes?

31 Continuous time Markov chains

Question of the Day For the following continuous time Markov chain:



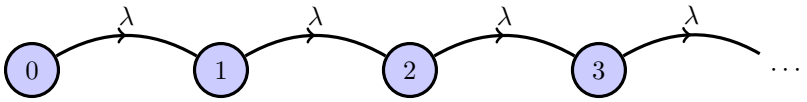
find $\mathbb{P}(X_{2.5} = c | X_0 = a)$.

Today

- Infinitesimal generator = transition rate matrix (CTMC version of transition matrix)
- A Poisson point process is a collection of points.
- A Poisson process is a random function (like Brownian Motion).

Continuous time Markov chain

- For all t, $N_t \in \{0, 1, 2, \ldots\}$.
- At arrival times, N_t "jumps" from current state to different state.
- Is an example of a *jump process*.
- Can be represented graphically. Here is a Poisson process



- This means that the time to jump from 0 to 1 has a distribution that is $\mathsf{Exp}(\lambda)$.
- Consider the question of the day.
  - The time to jump from a to b has distribution $\mathsf{Exp}(0.4)$.
  - The time to jump from a to c has distribution $\mathsf{Exp}(1.2)$.
  - Whichever jump occurs first is what actually occurs.
  - For $X \sim \mathsf{Exp}(0.4)$, $Y \sim \mathsf{Exp}(1.2)$, $\mathbb{P}(X < Y) = 0.4/(0.4 + 1.2)$.

Fact 64
Let $X_1, X_2, \ldots, X_k$ be independent random variables where $X_i \sim \mathsf{Exp}(\lambda_i)$. Then

$$\min(X_1, X_2, \ldots, X_k) \sim \mathsf{Exp}(\lambda_1 + \lambda_2 + \cdots + \lambda_k).$$

and

$$\mathbb{P}(\min(X_1, \ldots, X_k) = X_i) = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_k}.$$

Intuition: type I problems happen at rate 1.2 per day, type II problems happen at rate 3.2 per day. Type I or type II problems happen at rate $1.2 + 3.2 = 4.4$ per day.

Proof. Note: $1 - F_X(a) = \mathbb{P}(X > a)$ is called the *survival function* of X. Two r.v.'s with the same survival function have the same cdf and so the same distribution. Recall the survival function for $A \sim \mathsf{Exp}(\lambda)$ is $\exp(-\lambda a)\mathbb{1}(a \geq 0)$.

Let $X = \min_i X_i$. Then for $a > 0$,

$$\begin{aligned}\mathbb{P}(X > a) &= \mathbb{P}(X_1 > a, X_2 > a, \ldots, X_k > a) \\ &= \prod_i \mathbb{P}(X_i > a) = \prod \exp(-a\lambda_i) = \exp(-a \sum_i \lambda_i).\end{aligned}$$

So X has the survival function of an exponential with rate $\lambda_1 + \cdots + \lambda_k$. □

In example, rate at which leave state a is $1.2 + 0.4 = 1.6$.

31.1 The infinitesimal generator

Start with $X_0 = a$

- Now $\mathbb{P}(X_t = c | X_0 = a)$ changes continuously with c.
- Solve $(d/dt)\mathbb{P}(X_t = c | X_0 = a)$.
- Find $\mathbb{P}(X_h = c | X_0 = a)$ for small h. If h small, usually only one jump in time $[0, h]$. Specifically

  $$\mathbb{P}(X_h = a | X_0 = a) = \mathbb{P}(T_1 > h) + O(h^2)$$

  where $T_1 \sim \mathsf{Exp}(1.6)$ is time of first jump.

  $$\mathbb{P}(T_1 \leq h) = \exp(-1.6h) = 1 - 1.6h + O(h^2)$$

  using linear approximation for $\exp(x)$.
- Consider, what is $\mathbb{P}_a(X_h = b)$?
- Must have $T_1 \leq h$ and jump to b occurs before jump to c

  $$1.6h \cdot \frac{0.4}{1.6} + O(h^2) = 0.4h + O(h^2).$$
- Similarly, $\mathbb{P}_a(X_h = c) = 1.2h + O(h^2)$.

Repeat for $X_0 = b$ and $X_0 = c$

- Collecting all the results together gives:

  $$p_h = p_0 \begin{pmatrix} 1 - 1.6h & 0.4h & 1.2h \\ 1.0h & 1 - 5.3h & 4.3h \\ 0.9h & 2.2h & 1 - 3.1h \end{pmatrix} + O(h^2) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

  where $p_h(i) = \mathbb{P}(X_h = i)$.
- So that means:

  $$\begin{aligned} p'_{t=0} &= \lim_{h \to 0^+} \frac{p_h - p_{t=0}}{h} \\ &= \lim_{h \to 0^+} p_0 \begin{pmatrix} -1.6 & 0.4 & 1.2 \\ 1.0 & -5.3 & 4.3 \\ 0.9 & 2.2 & -3.1 \end{pmatrix} + wO(h) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \end{aligned}$$

- The end result:

$$p_t' = p_t \begin{pmatrix} -1.6 & 0.4 & 1.2 \\ 1.0 & -5.3 & 4.3 \\ 0.9 & 2.2 & -3.1 \end{pmatrix}$$

- Call this matrix multiplying p_t the *infinitesimal generator* or *transition rate matrix*.

Definition 76
The **infinitesimal generator** of a continuous time Markov chain, has $A(i,j)$ equal the rate at which the chain moves from state i to state j when $i \neq j$. The value of $A(i,i)$ is set as:

$$A(i,i) = -\sum_{j \neq i} A(i,j).$$

Fact 65
For continuous time Markov chain with infinitesimal generators A,

$$p_t' = p_t A,$$

which has solution

$$p_t = p_0 \exp(tA).$$

Compare to the discrete time case, where for transition matrix T,

$$p_{t+1} = p_t T.$$

Differential Equations

- Note $p_t' = p_t A$ is a system of differential equations.
- Fortunately, easy to solve!
- Recall: solution to $y' = ky$ is $y_t = y_0 e^{-kt}$.
- Same for matrices!
- Solution to $p_t' = p_t A$ is

$$p_t = p_0 e^{At}.$$

Definition 77
For a square matrix At, the **matrix exponential** is

$$e^A = I + A + A^2/2! + A^3/3! + \cdots$$

when this series converges.

Mathematica/WolframAlpha	MatrixExp[matrix]
Matlab	expm(matrix)
[For R must install and load package expm, then expm(matrix)]	

QotD

```
{{1,0,0}}*MatrixExp[2.5*{{-1.6,0.4,1.2},{1.0,-5.3,4.3},
  {0.9,2.2,-3.1}}]
```

gives

$$\mathbb{P}(X_{2.5} = c | X_0 = a) \approx \boxed{0.426125}.$$

Stationary Distributions

Definition 78
For a continuous time process X_t, π is **stationary** if $X_t \sim \pi \Rightarrow X_{t+s} \sim \pi$ for all $s \geq 0$.

- Then π will be stationary if for $p_t = \pi$, $p_t' = 0$.
- In other words
$$\pi A = 0,$$
where A is the infinitesimal generator.
- Because the rows of the A sum to 0, always a 0 eigenvalue.
- Similar argument to discrete time case show that left eigenvector for 0 is positive.
- For finite state space can always be normalized to give stationary distribution.
- Solve $\pi A = 0$ plus $\sum_{x \in \Omega} \pi(x) = 1$ to give solution.
- Continuous time chains always aperiodic!

Theorem 13 (Ergodic Theorem for finite state continuous time Markov chains)
For finite state Markov chains

1: There is at least one stationary distribution.

2: The stationary distribution π is unique if and only if there is exactly one recurrent communication class.

3: For one recurrent communication class,
$$(\forall x \text{ recurrent})(\forall A \in \mathcal{F}) \left(\lim_{t \to \infty} \mathbb{P}(X_t \in A | X_0 = x) = \pi(A) \right),$$
where π is the unique stationary distribution.

32 Stationary and Limiting distributions for CTMC

Question of the Day For the continuous time Markov chain with infinitesimal generator

$$\begin{pmatrix} -1.6 & 0.4 & 1.2 \\ 1.0 & -5.3 & 4.3 \\ 0.9 & 2.2 & -3.1 \end{pmatrix}$$

Is there a unique stationary distribution? What is it? Is it also the a limiting distribution?

Today

- Stationary distributions for CTMC
- Ergodic theorem for CTMC

Key ingredients for discrete time Harris chain ergodic theorem

- Aperiodicity.
- Existence of π.

The good news

- All continuous time Markov chains "aperiodic".
- Time to move any real number, so period becomes meaningless.
- So only need existence of π!
- Recall π stationary means:
$$X_t \sim \pi \Rightarrow X_s \sim \pi \text{ for all } s > t.$$
- No change in distribution at each step.
- For discrete time
$$p_{t+1} - p_t = 0 \Leftrightarrow \pi(A - I) = 0.$$
- For continuous time no change means derivative is 0:
$$p_t' = 0 \Leftrightarrow \pi A = 0.$$

QotD

- Eigenvalues:

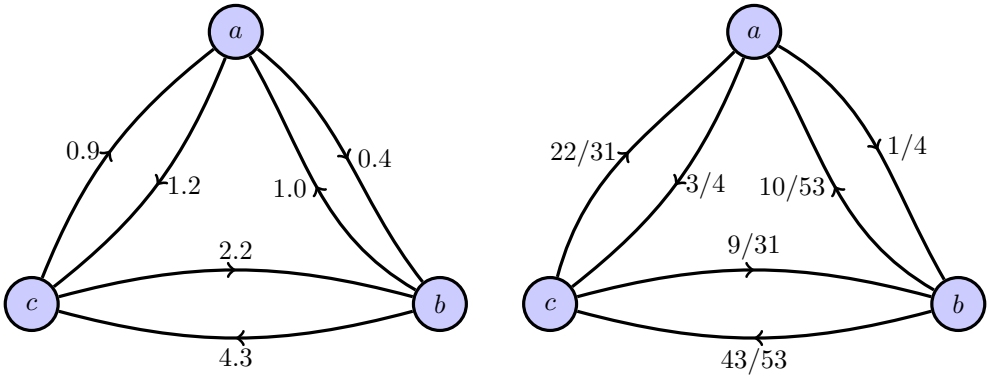
  ```
  eigenvalues {{-1.6,0.4,1.2},{1,-5.3,4.3},{0.9,2.2,-3.1}}
  ```

  gives $\lambda_1 \approx -7.4, \lambda_2 \approx -2.5, \lambda_3 = 0$, so unique solution (up to a constant factor) for $\pi A = 0$.
- ```
  solve -1.6a + 1b + 0.9c = 0 and 0.4a -5.3b + 2.2c = 0
    and 1.2a + 4.3b - 3.1c = 0 and a + b + c = 1
  ```
  gives $\pi = (697, 388, 808)/1893 \approx (0.3681, 0.2049, 0.4268)$.
- Is this unique stationary solution also a limiting distribution?
  ```

Underlying discrete chain

- Suppose $X_0 = a$.
- Jump to next state at time τ.
- What is $\mathbb{P}(X_\tau = b)$? $\mathbb{P}(X_\tau = c)$?

$$\mathbb{P}(X_\tau = b) = \frac{0.4}{0.4 + 1.2}, \ \mathbb{P}(X_\tau = c) = \frac{1.2}{0.4 + 1.2}$$



Original chain Underlying chain

Definition 79
A **continuous time Harris chain** is a Markov chain whose underlying discrete chain is Harris.

Definition 80
A continuous time Harris chain is **irreducible** if the underlying discrete chain is irreducible.

Theorem 14 (Ergodic Theorem for continuous time Harris chains)
Let X_n be a continuous time Harris chain with stationary distribution π. If $\mathbb{P}(R < \infty | X_0 = x) = 1$ for all x, then as $t \to \infty$,

$$d_{\text{TV}}([X_t | X_0 = x], \pi) \to 0.$$

Definition 81
For a countable state space, the stationary distribution satisfies the **balance equations**:

$$(\forall i) \left(\pi(i) \sum_{j \neq i} \lambda(i,j) = \sum_{j \neq i} \pi(j) \lambda(j,i) \right).$$

Matrix Exponentials

- What's happening as t goes to infinity?
- Recall that e^{tA} is defined as:

$$e^{tA} = I + tA + t^2 A^2/2! + t^3 A^3/3! + \cdots.$$

- Recall every matrix similar to its Jordan Canonical Form:

$$A = PJP^{-1},$$

where eigenvalues of A are on diagonal.

- For infinitesimal generators, eigenvalues must be real, otherwise probabilities would be complex!
- (More generally, Linear algebra: self-adjoint operators always have real eigenvalues.)
- When eigenvectors span space, J is a diagonal matrix D.
- For example:
$$\begin{pmatrix} -1.6 & 0.4 & 1.2 \\ 1.0 & -5.3 & 4.3 \\ 0.9 & 2.2 & -3.1 \end{pmatrix} = P \begin{pmatrix} -7.463 & 0 & 0 \\ 0 & -2.536 & 0 \\ 0 & 0 & 0 \end{pmatrix} P^{-1}$$
where
$$P = \begin{pmatrix} -0.03247 & 0.7745 & 0.5773 \\ -0.8897 & -0.4346 & 0.5773 \\ 0.4552 & -0.4594 & 0.5773 \end{pmatrix}$$
- Now consider A^3:
$$\begin{aligned} A^3 &= (PDP^{-1})(PDP^{-1})(PDP^{-1}) \\ &= PD(P^{-1}P)D(P^{-1}P)DP^{-1} \\ &= PD^3P^{-1}. \end{aligned}$$
To cube a diagonal matrix, just cube each of the entries.
- Back to the exponential:
$$e^{tA} = I + tPDP^{-1} + t^2PD^2P^{-1}/2! + \cdots,$$
and since this is all linear transformations, if the limit exists:
$$e^{tA} = P[I + tD + t^2D^2/2! + \cdots]P^{-1} = Pe^{tD}P^{-1}.$$
- A simple example: $A = \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix}$

`jordan form of {{-1,1},{2,-2}}`

$$S = \begin{pmatrix} -1 & 1 \\ 2 & 1 \end{pmatrix},\ D = \begin{pmatrix} -3 & 0 \\ 0 & 0 \end{pmatrix},\ S^{-1} = \begin{pmatrix} -1/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix},$$
[Note that D shows one eigenvalue 0, the other negative.]
$$\begin{aligned} e^{tA} &= Se^{tD}S^{-1} \\ &= \begin{pmatrix} -1 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} e^{-3t} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix} \\ &= \begin{pmatrix} (2/3) + (1/3)e^{-3t} & (1/3) - (1/3)e^{-3t} \\ (2/3) - (2/3)e^{-3t} & (1/3) + (2/3)e^{-3t} \end{pmatrix} \\ &= \begin{pmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix} + e^{-3t} \begin{pmatrix} 1/3 & -1/3 \\ -2/3 & 2/3 \end{pmatrix} \end{aligned}$$
- The rows $(2/3, 1/3)$ of the first part are the same, as $t \to \infty$ the second part goes to 0.
- Hence $(2/3, 1/3)$ is the limiting distribution.
- Faster to just say $(2/3, 1/3)$ is stationary:
$$\begin{pmatrix} 2/3 & 1/3 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}$$
and underlying discrete chain is irreducible, so Ergodic Theorem says the limiting distribution is the stationary distribution.

33 Differential equations

Question of the Day How can the classic predator prey model be implemented while keeping integers numbers of predators and prey?

Today

- Setting up continuous time Markov chains rather than differential equation models.

There are several reasons to use D.E.'s

- Continuous easier to solve exactly.
- Prepackaged numerical solvers give false sense of security.
- Results deterministic–same every time you run the model.

There are big drawbacks

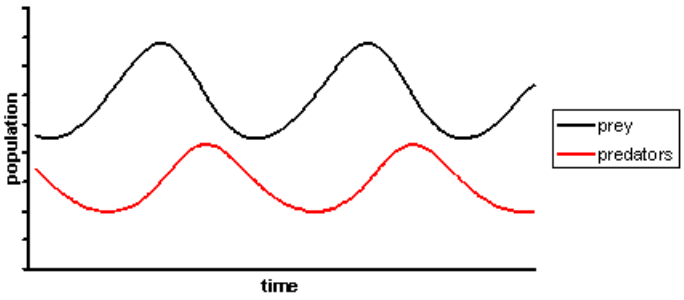
- Noninteger solutions.
- Failure to capture interesting phenomena.
- Results deterministic–same every time you run the model.

Classic predatory-prey model

- Lotka-Volterra model

$$x = \#\text{ of prey}$$
$$y = \#\text{ of predators}$$
$$a, b, c, d = \text{constants of model}$$

$$\frac{dx}{dt} = \underbrace{ax}_{\text{births}} - \underbrace{bxy}_{\text{deaths}}, \quad \frac{dy}{dt} = \underbrace{cxy}_{\text{births}} - \underbrace{dy}_{\text{deaths}}$$



Wikipedia. Retrieved 27 August, 2013 from http://en.wikipedia.org/wiki/Lotka

- This "solution" yields noninteger numbers of prey and predators.
- It also disallows the possibility that the predator or prey go extinct.

CTMC model

- Already have rates!
- State is (X_t, Y_t), both nonnegative integers.

$$X_t = \#\text{ of prey at time } t$$
$$Y_t = \#\text{ of predators at time } t$$

- Four events: prey birth, prey death, predator birth, predator death.
- Each corresponds to a move in the Markov chain.
- From current state (x, y):

new state	rate
$(x+1, y)$	ax
$(x-1, y)$	bxy
$(x, y+1)$	cxy
$(x, y-1)$	dy

- Steady state is $(0, 0)$ because you cannot leave this state.
- Will take a *very* long time to get there if large starting state.
- This type of systems called quasistable or quasistationary.

Conservation

- In some models, important to conserve quantities.
- Physics: matter
- Epidemiology: people

SEIR model

- Model of disease spread
- A person goes through four stages: Susceptible, Exposed, Infected, Recovered.
- What is conserved here is the total number of people:

$$S_t + E_t + I_t + R_t = N.$$

- Assume population is constant: so if someone dies from other causes (hit by car) immediately replaced by new baby (which is susceptible).
- First the D.E. model:

$$\begin{aligned}
\frac{dS}{dt} &= \underbrace{b(E+I+R)}_{\text{births}} - \underbrace{\beta SI}_{\text{exposure}} \\
\frac{dE}{dt} &= \underbrace{\beta SI}_{\text{exposure}} - \underbrace{\sigma E}_{\text{infection}} - \underbrace{bE}_{\text{deaths}} \\
\frac{dI}{dt} &= \underbrace{\sigma E}_{\text{infection}} - \underbrace{bI}_{\text{deaths}} - \underbrace{\gamma I}_{\text{recovery}} \\
\frac{dR}{dt} &= \underbrace{\gamma I}_{\text{recovery}} - \underbrace{bR}_{\text{deaths}}.
\end{aligned}$$

- Notice terms come in pairs: dS/dt has $-\beta SI$ term, while dE/dt has $+\beta SI$ term. Person moves from type S to type E.
- Reflected in the moves made by the chain.
- Let (s, e, i, r) be the current state (all nonnegative integers.)

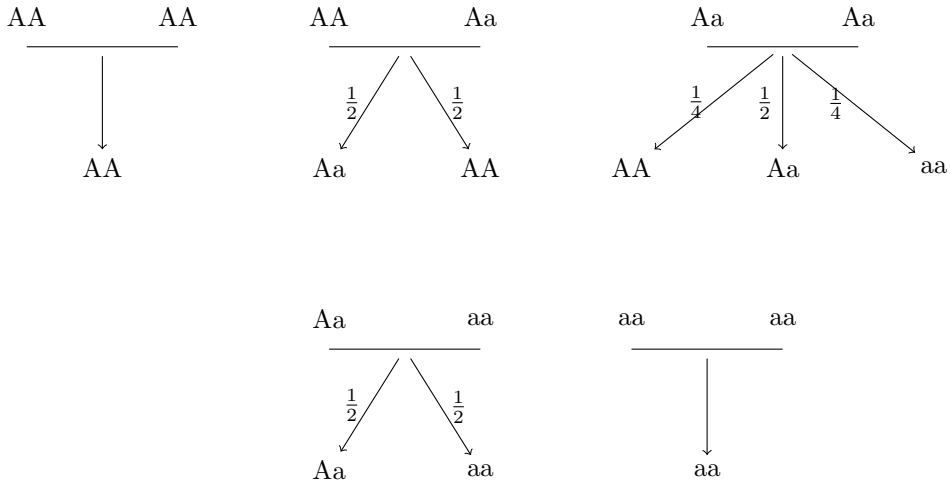
new state	rate
$(s-1, e+1, i, r)$	βsi
$(s+1, e, i, r-1)$	br
$(s+1, e, i-1, r)$	bi
$(s+1, e-1, i, r)$	be
$(s, e-1, i+1, r)$	σe
$(s, e, i-1, r+1)$	γi

Population genetics

- Single locus (site) for trait, diploid chromosome means two copies of each gene.
- Say gene comes in form A or a.
- So three genotypes: AA, Aa, aa. (Note Aa and aA result in same outcome.)

Definition 82
A population with random mating and no selection, no migration is in **Hardy-Weinberg Equilibrium**.

Flip coin for which gene from father, and flip coin for which gene comes from mother.



Gives the following D.E. model:

$$\frac{dy_{AA}}{dt} = b[y_{AA}^2 + (1/2)y_{AA}y_{Aa} + (1/2)y_{Aa}y_{AA} + (1/4)y_{Aa}y_{Aa}]/N - dy_{AA}$$
$$\frac{dy_{Aa}}{dt} = b[y_{AA}y_{Aa} + (1/2)y_{Aa}^2 + y_{Aa}y_{aa}]/N - dy_{Aa}$$
$$\frac{dy_{aa}}{dt} = b[(1/4)y_{Aa}^2 + (1/2)y_{Aa}y_{aa} + y_{aa}^2]/N - dy_{aa}$$

and adding them all together gives:

$$\frac{dN}{dt} = (b-d)N.$$

When $b \neq d$, the total population is not conserved, but always have

$$N = y_{AA} + y_{Aa} + y_{aa}.$$

If current state is $(n_{AA}, n_{Aa}, n_{aa}, n)$, where $n = n_{AA} + n_{Aa} + n_{aa}$, then our changes in state must preserve this final equation:

change in state	rate
$(1, 0, 0, 1)$	$b[n_{AA}^2 + (1/2)n_{AA}n_{Aa} + (1/2)n_{Aa}n_{AA} + (1/4)n_{Aa}^2]/n$
$(0, 1, 0, 1)$	$b[n_{AA}n_{Aa} + (1/2)n_{Aa}^2 + n_{Aa}n_{aa}]/n$
$(0, 0, 1, 1)$	$b[(1/4)n_{Aa}^2 + (1/2)n_{Aa}n_{aa} + n_{aa}^2]/n$
$(-1, 0, 0, -1)$	dn_{AA}
$(0, -1, 0, -1)$	dn_{Aa}
$(0, 0, -1, -1)$	dn_{aa}

Comments on rates

- Rates always positive.
- Negative terms in D.E. model becomes negative change at positive rate.

34 Uniform and square integrability

Question of the Day Is there an easier way to show uniform integrability?

Today

- Domination
- Square integrability

Recall

- $X_n \to X$ in probability means

$$(\forall \epsilon > 0) \left(\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0 \right).$$

- When $X_n \to X$ in probability $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$ if and only if the $\{X_n\}$ are uniformly integrable, which means

$$\lim_{B\to\infty} \left(\sup_n \mathbb{E}(|X_n| \mathbb{1}(|X_n| > B)) \right) = 0.$$

- Intuition: can't have small chance of being big, that throws off the expected value.
- Sadly, tough to apply directly in practice.
- Here are two simpler conditions that imply uniform integrability.

Theorem 15
Suppose that $\{X_\alpha\}$ satisfy:

1: *Dominated.* There exists Y with $\mathbb{E}[|Y|] < \infty$ where $|X_\alpha| \leq Y$ for all α. Or,

2: *Square integrable.* There exists M such that $\mathbb{E}[X_\alpha^2] \leq M$ for all α.

Then the $\{X_\alpha\}$ are uniformly integrable.

Example

- Suppose we have a Markov chain on the integers where we move left (subtract 1) with probability 2/3 and move right (add 1) with probability 1/3.
- Let $X_0 = x_0 > 0$, $T = \inf\{t : X_t = 0\}$.
- Question: Is $X_{t\wedge T}$ u.i.?
- Encode chain as follows: $X_{t+1} = X_t + D_{t+1}$, where $D_1, D_2, \ldots \overset{\text{iid}}{\sim} D$, $\mathbb{P}(D = -1) = 2/3$, $\mathbb{P}(D = 1) = 1/3$.
- $X_{t\wedge T}$ is square integrable.

  Proof. We will show by induction that $\mathbb{E}[X_{t\wedge T}^2] \leq \max\{5, x_0^2\}$.
  Base case: when $t = 0$, $X_0 = x_0$ so $\mathbb{E}[X_{t\wedge T}^2] = x_0^2$.
  Induction hypothesis: $\mathbb{E}[X_{t\wedge T}^2] \leq \max\{5, x_0^2\}$, consider $\mathbb{E}[X_{(t+1)\wedge T}^2]$. Use our standard trick:

$$\mathbb{E}[X_{(t+1)\wedge T}^2] = \mathbb{E}[\mathbb{E}[X_{(t+1)\wedge T}^2 | X_{t\wedge T}]].$$

Look at the inside expectation first. Things will be different if $T \leq t$ or $T > t$, so break this up using indicator functions:

$$\begin{aligned} X_{(t+1)\wedge T} &= X_{(t+1)\wedge T}\mathbb{1}(T \leq t) + X_{(t+1)\wedge T}\mathbb{1}(T > t) \\ &= X_{(t+1)\wedge T}\mathbb{1}(T > t) \end{aligned}$$

because if $T \leq t$ then $X_{t\wedge T} = X_{(t+1)\wedge T} = 0$. Now

$$X_{(t+1)\wedge T}\mathbb{1}(T > t) = [X_{t\wedge T} + D_{t+1}]\mathbb{1}(T > t).$$

Plug this into out expectation to get:

$$\begin{aligned} \mathbb{E}[X^2_{(t+1)\wedge T}|X_{t\wedge T}] &= \mathbb{E}[(X_{t\wedge T} + D_{t+1})^2]\mathbb{1}(T > t)^2|X_{t\wedge T}] \\ &= \mathbb{E}[(X^2_{t\wedge T} + 2X_{t\wedge T}D_{t+1} + D^2_{t+1})\mathbb{1}(T > t)|X_{t\wedge T}] \end{aligned}$$

First note: the square of an indicator function is just the indicator function since its value is 0 or 1. Second note: $D^2_{t+1} = 1$ always. Third note: D_{t+1} is independent of $X_{t\wedge T}$. Fourth note: $\mathbb{1}(T > t) = \mathbb{1}(X_{t\wedge T} > 0)$, so it acts as a constant in the conditional expectation. This gives us:

$$\mathbb{E}[X^2_{(t+1)\wedge T}|X_{t\wedge T}] = \mathbb{1}(T > t)[X^2_{t\wedge T} + 2X_{t\wedge T} \underbrace{\mathbb{E}[D_{t+1}]}_{(1/3)(1)+(2/3)(-1)=-1/3} + 1].$$

Now if $X_{t\wedge T}$ is at most 2, that says:

$$\mathbb{E}[X^2_{(t+1)\wedge T}|X_{t\wedge T}] = \mathbb{1}(T > t)[4 + 1] \leq 5$$

On the other hand, if $X_{t\wedge T}$ is at least 3, then

$$\begin{aligned} \mathbb{E}[X^2_{(t+1)\wedge T}|X_{t\wedge T}] &\leq \mathbb{1}(T > t)[X_{t\wedge T} + 2(3)(-1/3) + 1] \\ &\leq X_{t\wedge T} \leq \max\{5, x_0^2\} \end{aligned}$$

which completes the induction. □

Proof Dominated implies U.I.. Let's look at what domination gives us:

$$\mathbb{E}(|X_n|\mathbb{1}(|X_n| > B)) \leq \mathbb{E}(Y\mathbb{1}(Y > B)).$$

This holds for every n, so

$$\sup_n \mathbb{E}(|X_n|\mathbb{1}(|X_n| > B)) \leq \mathbb{E}(Y\mathbb{1}(Y > B)).$$

By the DCT:

$$\lim_{B\to\infty} \mathbb{E}(Y\mathbb{1}(Y > B)) = 0,$$

and we're done! □

Proof Square Integrable implies U.I. We are given that there exists M with $\mathbb{E}[X^2_\alpha] \leq M$. Our goal is to show $\mathbb{E}[|X_\alpha|\mathbb{1}(\cdot|X_\alpha|$ goes to 0. The key observation is

$$|X_\alpha|\mathbb{1}(|X_\alpha| > B) \leq |X_\alpha|^2\mathbb{1}(|X_\alpha| > B)/B.$$

Taking the mean of both sides:

$$\begin{aligned} \mathbb{E}[|X_\alpha|\mathbb{1}(|X_\alpha| > B)] &\leq \mathbb{E}[|X_\alpha|^2\mathbb{1}(|X_\alpha| > B)/B] \\ &\leq (1/B)\mathbb{E}[X^2_\alpha] \leq M/B. \end{aligned}$$

This holds for any α, so $\sup_\alpha \mathbb{E}[|X_\alpha|\mathbb{1}(|X_\alpha| > B)] \leq M/B$. This goes to 0 as B goes to infinity. □

Example continued

- Suppose I make a martingale out of X_t by setting:

$$M_t = X_t + (1/3)t.$$

$$\begin{aligned}\mathbb{E}[M_{t+1}|X_0,\ldots,X_t] &= \mathbb{E}[X_{t+1} + (1/3)(t+1)|X_0,\ldots,X_t] \\ &= \mathbb{E}[X_t + D_{t+1} + (1/3)t + (1/3)|X_0,\ldots,X_t] \\ &= X_t + (1/3)t + \mathbb{E}[D_{t+1} + 1/3] \\ &= M_t\end{aligned}$$

- As before, let $T = \inf\{t : X_t = 0\}$.
- Let's see if $M_{t\wedge T}$ is u.i. by checking for square integrability.

$$\begin{aligned}\mathbb{E}[M_{t\wedge T}^2] &= \mathbb{E}[\mathbb{E}[M_{t\wedge T}^2|M_{(t-1)\wedge T}]] \\ &= \mathbb{E}[\mathbb{E}[(M_{(t-1)\wedge T} + (D_t + 1/3)\mathbb{1}(T > t-1))^2|M_{(t-1)\wedge T}] \\ &= \mathbb{E}[\mathbb{E}[M_{(t-1)\wedge T}^2 + 2M_{(t-1)\wedge T}(D_t + 1/3)\mathbb{1}(T > t-1) + \\ &\quad (D_t + 1/3)^2\mathbb{1}(T > t-1)|M_{(t-1)\wedge T}]]\end{aligned}$$

  As before D_t and $\mathbb{1}(T > t-1)$ are independent, and $\mathbb{E}[D_t + 1/3] = 0$ so the middle term goes away leaving

$$\mathbb{E}[M_{t\wedge T}^2] = \mathbb{E}[M_{(t-1)\wedge T}^2 + \mathbb{1}(T > t-1)\mathbb{E}[(D_t + 1/3)^2]]$$

  Since $\mathbb{E}(D_t + 1/3)^2 = (4/3)^2(1/3) + (-2/3)^2(2/3) = 8/9$,

$$\begin{aligned}\mathbb{E}[M_{t\wedge T}^2] &= \mathbb{E}[M_{(t-1)\wedge T}^2 + \mathbb{1}(T > t-1)(8/9)] \\ &= \mathbb{E}[M_{(t-1)\wedge T}^2] + (8/9)\mathbb{P}(T > t-1)\end{aligned}$$

  An induction proof gives:

$$\mathbb{E}[M_{t\wedge T}^2] = x_0^2 + \mathbb{P}(T > 0) + \mathbb{P}(T > 1) + \cdot\mathbb{P}(T > t-1).$$

  Hence $M_{t\wedge T}^2$ is square integrable if and only if

$$\mathbb{P}(T > 0) + \mathbb{P}(T > 1) + \cdots = \mathbb{E}[T] < \infty.$$

- The last line comes from the tail sum formula for expected value.
- Next time we'll look at a strategy for showing that $\mathbb{E}[T] < \infty$ for these cases.

35 Wald's Equation

Question of the Day Suppose I roll a fair six sided die until the sum of the numbers is at least 100. Bound as best as possible the expected number of rolls needed.

Today

- Wald's Equation

Abraham Wald

- Born 1902 in Austria
- Fled in 1938 because he was Jewish, came to U.S.
- Best known for work on sequential experiments: when can you stop testing patients and declare a drug a success early?
- Wald's Equation came out of this work.
- Shortcut for many problems that use Optional Sampling Theorem
- Even better, it's one of those theorems that make intuitive sense.

Theorem 16 (Wald's Equation)
Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} X$ where $\mathbb{E}[X]$ is finite. Let T be a stopping time with respect to $X_1, X_2, \ldots$. If

1: $\mathbb{P}(X \geq 0) = 1$ or

2: $\mathbb{E}[T] < \infty$, then

$$\mathbb{E}\left(\sum_{n=1}^{T} X_n\right) = \mathbb{E}[T]\mathbb{E}[X_i].$$

Using Wald on the Qotd

- Say $X_i \overset{\text{iid}}{\sim} \mathsf{Unif}(\{1,2,3,4,5,6\})$.
- Let $T = \inf\{t : X_1 + \cdots + X_t \geq 100\}$.
- Then $T < 100$, so $\mathbb{E}[T] < 100$.
- Hence

$$\mathbb{E}\left(\sum_{i=1}^{T} X_i\right) = \mathbb{E}[T]\mathbb{E}[X_i].$$

- $\mathbb{E}[X_i] = (1+6)/2$.
- $\mathbb{E}\left(\sum_{i=1}^{T} X_i\right) \in \{100, 101, 102, 103, 104, 105\}$
- Hence

$$100 \leq \mathbb{E}[T](7/2) \leq 105,$$
$$\mathbb{E}[T] \in [200/7, 235/7] \approx [28.57, 30].$$

Proof when X is nonnegative. Note

$$\mathbb{E}\left(\sum_{i=1}^{T} X_i\right) = \mathbb{E}\left(\sum_{i=1}^{\infty} X_i \mathbb{1}(i \leq T)\right) = \sum_{i=1}^{\infty} \mathbb{E}[X_i \mathbb{1}(i \leq T)],$$

where the last step uses the Monotone Convergence Theorem and the fact that the summand is nonnegative.

To break apart $\mathbb{E}[X_i \mathbb{1}(i \leq T)]$, note that X_i and $\mathbb{1}(i \leq T)$ are actually independent. This is because knowing $X_1, \ldots, X_{i-1}$ are sufficient to determine if T is at least i, and these are independent of X_i, so

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{T} X_i\right) &= \sum_{i=1}^{\infty} \mathbb{E}[X_i]\mathbb{E}[\mathbb{1}(i \leq T)] \\
&= \sum_{i=1}^{\infty} \mathbb{E}[X_i]\mathbb{P}(i \geq T) \\
&= \mathbb{E}[X] \sum_{i=1}^{\infty} \mathbb{P}(T \geq i)
\end{aligned}$$

since $\mathbb{E}[X_i] = \mathbb{E}[X]$ for all i.

since the right hand side is just the tail sum formula for expected value of T, the equation is proved. □

Proof when T has finite expected value. Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} X$, and T be a stopping time w.r.t $X_1, X_2, \ldots$. Let $\mu = \mathbb{E}(X_i) < \infty$. Create a martingale out of the sums of X_i:

$$M_n = \sum_{i=1}^{n} (X_i - \mu).$$

Part 1: Check that M_n is a martingale w.r.t $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.

$$\mathbb{E}[|M_n|] \leq \mathbb{E}[|X_1 - \mu + X_2 - \mu + \cdots + X_n - \mu|] \leq \mathbb{E}[|X_1| + \cdots + |X_n| + n|\mu|].$$

Since X_i has finite expectation, $\mathbb{E}[|X_i|]$ is also finite, and $\mathbb{E}[|M_n|] \leq n(\mathbb{E}(|X_i|) + \mu) < \infty$.

Next

$$\begin{aligned}
\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \mathbb{E}[(X_1 - \mu) + \cdots + (X_{n+1} - \mu)|X_1, \ldots, X_n] \\
&= (X_1 - \mu) + \cdots + (X_n - \mu) + \mathbb{E}[X_{n+1} - \mu] \\
&= (X_1 - \mu) + \cdots + (X_n - \mu) \\
&= M_n.
\end{aligned}$$

So yes, it's a martingale.

Part 2: show that $M_{t \wedge T}$ is a uniformly integrable martingale. The approach is to show that the $M_{t \wedge T}$ are dominated by an integrable random variable. Let

$$Y = \sum_{i=1}^{T} |X_i - \mu| \geq |M_{t \wedge T}|.$$

Since $|X_i - \mu|$ are nonnegative, can use the nonnegative version of Wald's Equation to get

$$\mathbb{E}[Y] = \mathbb{E}[|X_i - \mu|]\mathbb{E}[T] < \infty.$$

Therefore $|M_{t \wedge T}| \leq Y$, where Y is an an integrable random variable.

Part 3: Hence the $M_{t \wedge T}$ is uniformly integrable, and we can use the OST to say

$$\mathbb{E}[M_T | M_0] = M_0,$$

or in this case

$$\mathbb{E}\left[\sum_{i=1}^{T} (X_i - \mu)\right] = \mathbb{E}\left[\sum_{i=1}^{T} X_i - \mu T\right] = 0.$$

Adding $\mathbb{E}(\mu T) = \mu \mathbb{E}[T]$ to both sides completes the proof. □

Roulette

- Odd-Even bets win with probability 18/38 on an American Roulette wheel.
- Betting \$1 a spin, what is the expected number of spins needed before I lose \$20?
- Let $\mathbb{P}(X = -1) = 20/38$ and $\mathbb{P}(X = 1) = 18/38$.
- So $\mu = \mathbb{E}[X] = (18/38) + (-1)(20/38) = -2/38$.
- Let $T = \inf\{t : X_1 + \cdots + X_t = -20\}$.
- Problem: X is not nonnegative, so is $\mathbb{E}[T] < \infty$?
- Solution: Let $T_a = \inf\{t : X_1 + \cdots + X_t \in \{-20, a\}\}$.
- Then for $S_t = X_1 + \cdots + X_t$, $S_{t \wedge T_a}$ is a finite state Markov chain, so $\mathbb{E}[T_a] < \infty$.
- Hence
$$\mathbb{E}\left[\sum_{i=1}^{T_a} X_i\right] = \mathbb{E}[T_a](-2/38).$$
- Since $-20 \leq \mathbb{E}[\sum_{i=1}^{T_a} X_i]$,
$$\mathbb{E}[T_a] \leq 20(38/2).$$
- Since T_a are nonnegative and increasing to T,
$$\mathbb{E}[T] \leq 20(38/2).$$
- That means we could have just applied Wald to T in the first place!
$$\mathbb{E}[\sum_{i=1}^{T} X_i] = \mathbb{E}[T](-2/38) \Rightarrow \mathbb{E}[T] = 20(38/2) = \boxed{380}.$$

36 Stochastic Integration

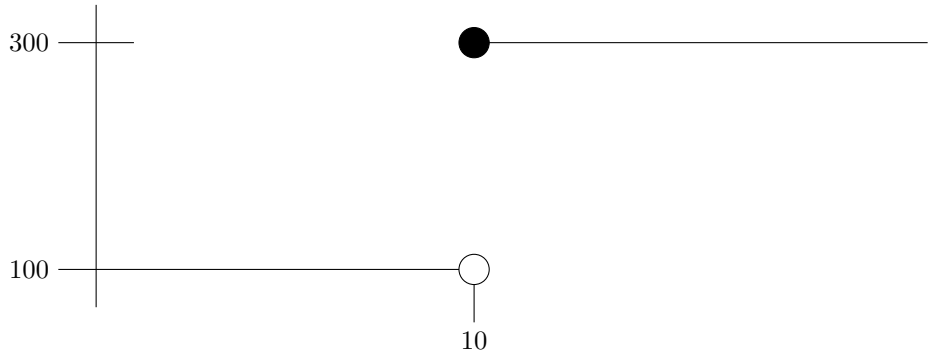
Question of the Day Can we integrate with respect to Brownian Motion?

Integrals as financial model

- Let S_t be the value of a stock.
- Let Y_t denote the number of shares of a stock I own.
- Intuition:
$$\int_0^T Y_t \, dS_t$$
represents the amount of profit made from time 0 to T.

First example

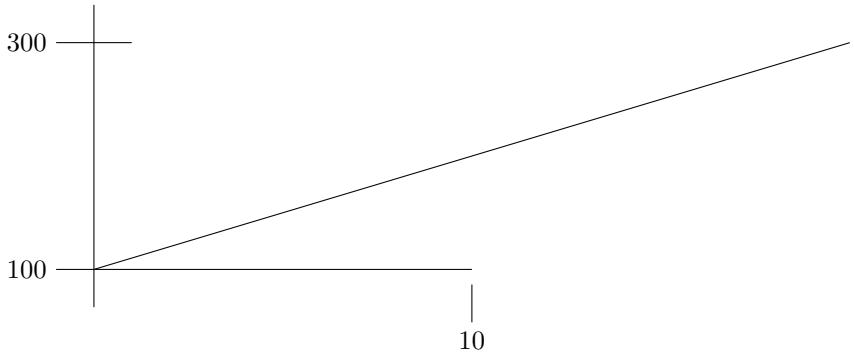
- $S_t = 100 + 200 \cdot \mathbb{1}(t \geq 10)$



- Own 3 shares for all $t \geq 0$ ($Y_t = 3$)
$$\int_0^{20} 3 \, dS_t = (3)(300) - (3)(100) = 600$$
- Own t shares of stock at time t ($Y_t = t$)
$$\int_0^{20} t \, dS_t = 10(300 - 100) = 2000$$
- Own t^2 shares of stock at time t ($Y_t = t^2$)
$$\int_0^{20} t^2 \, dS_t = 10^2(300 - 100) = 20\ 000.$$

Second example

- $S_t = 100 + 10t$.



- Now S_t changing continuously.
- Own 3 shares of stock:
$$\int_0^{20} 3\ d(100+10t) = 3(300) - 3(100) = 600$$
- Own t^2 shares of stock at time t:
$$\int_0^{20} t^2\ d(100+10t) = ?$$
- How much profit to I make over interval $[t, t+dt]$?:
$$t^2(100+10(t+dt)) - t^2(100+10t) = 10t^2\ dt$$
- So altogether:
$$\begin{aligned} &10(0)\ dt + 10(dt)^2\ dt + 10(2dt)^2\ dt + \cdots + 10(20-dt)^2\ dt \\ &= \sum_{i=0}^{20/dt} 10(i\ dt)^2\ dt \\ &= 10(1/3)(20/dt)(20/dt+1)(20/dt+1/2)\ dt^3 \\ &= (20)^3(10)/3. \end{aligned}$$
- Note regular Riemann integration:
$$\int_0^{20} 10t^2\ dt = 10t^3/3|_0^{20} = (10)20^3/3.$$

Results so far

- When S_t is differentiable,
$$\int_0^T Y_t\ dS_t = \int_0^T Y_t S_t'\ dt.$$
- When S_t has a single jump at time $T_{\text{jump}} \in [0,T]$ from a to b:
$$\int_0^T Y_t\ dS_t = Y_{T_{\text{jump}}}(b-a).$$
- Together give Stieltjies Integral (1894).
- The Fundamental Theorem of Calculus is when you only own one share of stock:
$$\int_0^T S_t'\ dt = S_t|_0^T = S_T - S_0.$$

Brownian Motion

- What if $S_t = B_t$?
- Brownian Motion is not differentiable!
- Brownian Motion is random.
- So amount of profit you make off of stock is random! That is,
$$\int_0^T Y_t\ dB_t$$
is itself a random variable!

Example

- Suppose $Y_t = 1$.
- Then
$$\int_0^1 Y_t \, dB_t = B_1 - B_0.$$
If $Y_t = 2$:
$$\int_{0.5}^{1.3} Y_t \, dB_t = 2(B_{1.3} - B_{0.5})$$

Definition 83
Y_t is a **simple strategy** if it can be written in the form
$$Y_t = y_1 \mathbb{1}(t \in [0, t_1)) + y_2 \mathbb{1}(t \in [t_1, t_2)) + \cdots + y_k \mathbb{1}([t_{k-1}, t_k)),$$
that is, Y_t changes value a finite number of times.

Definition 84
For $Y_t = y_1 \mathbb{1}(t \in [0, t_1)) + y_2 \mathbb{1}(t \in [t_1, t_2)) + \cdots + y_k \mathbb{1}([t_{k-1}, T])$, the Ito integral is defined as
$$\int_0^T Y_t \, dB_t = y_1(B_{t_1} - B_0) + y_2(B_{t_2} - B_{t_1}) + \cdots + y_k \mathbb{1}(B_T - B_{t_{k-1}}).$$

Now that have a definition, can prove properties of Ito integral:

Fact 66
The Ito integral satisfies:

1: Linearity:
$$\int_0^T (aX_t + bY_t) \, dB_t = a \int_0^T X_t \, dB_t + b \int_0^T Y_t \, dB_t.$$

2: Martingale: $Z_T = \int_0^T Y_t \, dB_t$ is a martingale.

3: Second moment:
$$\mathbb{E}\left(\int_0^T Y_t \, dB_t \right)^2 = \int_0^T \mathbb{E}[Y_t^2] \, dt.$$

What if Y_t is not simple?

- Things become a lot more complicated!
- Use approach similar to Riemann.
- Approximation Y_t over intervals of small width with an average.
- Take limit as width approaches 0.
- When Y_t is right-continuous with left limits (called **cadlag** for the French "continue à droite, limite à gauche"), this limit exists.

Expectations and integrals

- Recall that expected values are themselves really integrals.
- So you can swap expectation and integration as long as Fubini or Tonnelli holds.
- For example: does
$$\mathbb{E}\left[\int_0^T tB_t\ dt\right] = \int_0^T \mathbb{E}[tB_t]\ dt?$$
- Yes!
$$\begin{aligned}\left|\mathbb{E}\left[\int_0^T tB_t\ dt\right]\right| &\leq \mathbb{E}\left[\int_0^T |tB_t|\ dt\right] \\ &= \int_0^T \mathbb{E}|tB_t|\ dt && \text{by Tonnelli} \\ &= \int_0^T t^{3/2}\mathbb{E}|Z|\ dt && \text{by Tonnelli} \\ &= \sqrt{2/\pi}T^{5/2}/(5/2) < \infty,\end{aligned}$$
so by Fubini
$$\mathbb{E}\left[\int_0^T tB_t\ dt\right] = \int_0^T \mathbb{E}[tB_t]\ dt.$$

37 Ito's Formula

Question of the Day Write $\int_0^T Y_t \, de^{B_t+2t}$ as an integral with respect to Brownian Motion.

Today

- Ito's Formula
- The one nontrivial stochastic integral we can solve analytically.

Last time

- Geometry gets us Riemann integral.
- Not good enough for integrating over stochastic processes.
- Finance the best way to look at it.
- How much money we make on a stock price.
- Since we can simulate Brownian motion, can simulate stochastic integrals as well.

Next step

- Problem: B_t not a very good model of stock price.
- Better model: geometric Brownian Motion.
- For example, what if $S_t = e^{B_t+2t}$, what is

$$\int_0^T Y_t \, dS_t?$$

- This stock price is a function of Brownian motion:

$$f(t,x) = e^{x+2t}, \; S_t = f(t, B_t)$$

- Then can write question as

$$\int_0^T Y_t \, df(t, B_t) =?$$

- So what profit is made over interval $[t, t+h]$?

$$Y_t[f(t+h, B_{t+h}) - f(t, B_t)].$$

- Need to understand how changing two coordinates changes value.

Multivariate Taylor series expansion

- Start with one dimensional Taylor series expansion. The change in $f(t)$ when I move to $f(t+h)$ is

$$f(t+h) - f(t) = hf'(t) + h^2 f''(t)/2! + h^3 f'''(t)/3! + \cdots .$$

- But we have a function of two variables: $f(t,x)$.

- So Taylor series is more interesting. $t \to t + dt$ and $x \to x + dx$.

$$f(t+h, x+g) - f(t,x) = h\frac{\partial f}{\partial t} + g\frac{\partial f}{\partial x} + (1/2)\left[h^2\frac{\partial^2 f}{\partial t^2} + g^2\frac{\partial^2 f}{\partial x^2} + hg\frac{\partial}{\partial x}\frac{\partial f}{\partial t} + gh\frac{\partial}{\partial t}\frac{\partial f}{\partial x}\right] + \cdots$$

- When f has continuous second partial derivatives:

$$\frac{\partial^2 f}{\partial x \partial t} = \frac{\partial^2 f}{\partial t \partial x}.$$

So

$$f(t+h, x+g) - f(t,x) = h\frac{\partial f}{\partial t} + g\frac{\partial f}{\partial x} + \frac{1}{2}\left[h^2\frac{\partial^2 f}{\partial t^2} + 2hg\frac{\partial^2 f}{\partial x \partial t} + g^2\frac{\partial^2 f}{\partial x^2}\right] + \cdots$$

Multivariate Taylor with Brownian motion

- Time moves from t to $t+h$; g is random!
- $g = B_{t+h} - B_t \sim \mathsf{N}(0,h) = \sqrt{h}\mathsf{N}(0,1)$.
- $g^2 \sim h[\mathsf{N}(0,1)]^2 = h\chi^2(1)$. $\mathbb{E}[g^2] = 1$, $\mathbb{V}(g^2) = 2$.
- Let $Z_h \sim \mathsf{N}(0,1)$. Then

$$f(t+h, B_{t+h}) - f(t, B_t) = h\frac{\partial f}{\partial t} + \sqrt{h}Z\frac{\partial f}{\partial x} + \frac{1}{2}\left[h^2\frac{\partial^2 f}{\partial t^2} + 2h\sqrt{h}Z\frac{\partial^2 f}{\partial x \partial t} + hZ^2\frac{\partial^2 f}{\partial x^2}\right] + \cdots$$

- Now compared to h, h^2 term small. Compared to $\sqrt{h}Z$, $h\sqrt{h}Z$ term is small. So those become higher order terms:

$$f(t+h, B_{t+h}) - f(t, B_t) = h\frac{\partial f}{\partial t} + \sqrt{h}Z\frac{\partial f}{\partial x} + \frac{1}{2}hZ^2\frac{\partial^2 f}{\partial x^2} + \cdots$$

Towards Ito's Lemma

- What happens as $h \to 0$?
- $h \to dt$.
- $\sqrt{h}Z \to dB_t$.
- $hZ^2 \to dt$.
- Note: this is not in *any way* a formal derivation!
- But hopefully it makes the following result a little less mysterious.

Theorem 17 (Ito's Lemma)
Let $f(t,x)$ be a function with continuous second partial derivatives. Then

$$df(t, B_t) = \left[\frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\right]dt + \frac{\partial f}{\partial x}\, dB_t.$$

Question of the Day

- What is $\int_0^T Y_t de^{B_t+2t}$?
- Here $f(t,x) = e^{x+2t}$.
- So
$$\partial f/\partial t = 2e^{x+2t},\ \partial f/\partial x = e^{x+2t},\ \partial^2 f/\partial x^2 = e^{x+2t}.$$
- By Ito's Lemma:
$$de^{B_t+2t} = (2e^{B_t+2t} + (1/2)e^{B_t+2t})\ dt + e^{B_t+2t}\ dB_t.$$
Or in integral form:
$$\int_0^T Y_t\ de^{B_t+2t} = \int_0^T Y_t(2e^{B_t+2t} + (1/2)e^{B_t+2t})\ dt + \int_0^T Y_t e^{B_t+2t}\ dB_t.$$

A stochastic integral that can be solved analytically

- Many Riemann integrals have analytical solutions:
$$\int_0^T x\exp(x)\ dx = e^T(T-1)+1.$$
- Most Ito integrals do not.
- Simple one:
$$\int_0^T 1\ dB_t = B_T.$$
- A nontrivial one that I know how to solve is:
$$\int_0^T B_t\ dB_t.$$
- Consider $f(t,x) = (1/2)[x^2 - t]$. Then $\partial f/\partial t = -1/2$, $\partial f/\partial x = x$, $\partial^2 f/\partial x^2 = 1$. So
$$\int_0^T 1\ df(t,B_t) = \int_0^T (-1/2 + 1/2)\ dt + B_t\ dB_t = \int_0^T B_t\ dB_t.$$
- By first principles:
$$\int_0^T 1\ df(t,B_t) = f(T,B_T) - f(0,B_0) = \boxed{(1/2)[B_T^2 - T]}$$

Checking properties

- Recall Z_T should be a martingale, and $\mathbb{E}[Z_T^2] = \int_0^T \mathbb{E}[B_t^2]\ dt$.
- Is $Z_T = B_T^2 - T$ a martingale?
- Yes, for $S < T$,
$$\begin{aligned}
\mathbb{E}[Z_T|\mathcal{F}_S] &= \mathbb{E}[(1/2)[B_T^2 - T]|B_S] \\
&= (1/2)\mathbb{E}[(B_S + (B_T - B_S))^2 - T|B_S] \\
&= (1/2)\mathbb{E}[(B_S^2 + 2B_S(B_T - B_S) + (B_T - B_S)^2 - T|B_S] \\
&= (1/2)[B_S^2 + 2B_S(0) + (T-S) - T] \\
&= (1/2)[B_S^2 - T] \\
&= Z_S.
\end{aligned}$$

also

$$\begin{aligned}\mathbb{E}[Z_T^2] &= \mathbb{E}[[(1/2)(B_T^2 - T)]^2] \\ &= (1/4)\mathbb{E}[B_T^4 - 2B_T^2 T + T^2] \\ &= (1/4)[3T^2 - 2T^2 + T^2] = T^2/2.\end{aligned}$$

Since $\mathbb{E}[B_t^2] = t$,

$$\int_0^T t \; dt = T^2/2.$$

Weiner Process = Brownian Motion One more note: since Norbert Weiner put the whole endeavor on a firm theoretical foundation, Brownian Motion is also known as the Weiner Process, and W_t is commonly used as well as B_t.

38 Reversibility

Question of the Day Construct a Markov chain on $\{a, b, c\}$ with limiting distribution $\pi(a) = 0.5,\ \pi(b) = 0.3,\ \pi(c) = 0.2$.

Today

- Balance Equations
- Reversibility/Detailed Balance

From ergodic theorem for finite state Markov chains

- If irreducible and aperiodic, limiting distribution = stationary distribution.
- So need to construct an irreducible, aperiodic chain with

$$\pi A = \pi.$$

- Notation:

$$p(x, y) = \mathbb{P}(X_{t+1} = y, X_t = x).$$

- Writing out these equations for $\Omega = \{a, b, c\}$:

$$\begin{aligned}
\pi(a) &= \pi(a)p(a, a) + \pi(b)p(b, a) + \pi(c)p(c, a) \\
\pi(b) &= \pi(a)p(a, b) + \pi(b)p(b, b) + \pi(c)p(c, b) \\
\pi(c) &= \pi(a)p(a, c) + \pi(b)p(b, c) + \pi(c)p(c, c)
\end{aligned}$$

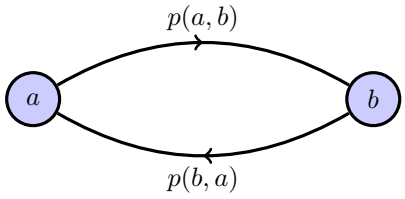
- Can think of this as "probability" flowing along edges of the transition graph:

$$\text{Flow out} = \text{Flow in}.$$

- These are called the *balance equations* because the probability flow out of a node is balanced by the probability flow into a node.

Detailed Balance

- Now just concentrate on two nodes:



- Suppose the flow from a to b matches that from b to a:

$$\pi(a)p(a, b) = \pi(b)p(b, a).$$

- This is a *detailed balance equation*.

Definition 85
A probability distribution π on a discrete state space Ω satisfies the **detailed balance equations** (also called *reversibility* for a discrete time Markov chain if for all $a, b \in \Omega$:

$$\pi(b)\mathbb{P}(X_{t+1} = a, X_t = b) = \pi(a)\mathbb{P}(X_{t+1} = b, X_t = a).$$

Fact 67
If π is reversible, then it is stationary.

Proof. Let $a \in \Omega$. Then

$$\sum_{i\in\Omega} \pi(i)p(i,a) = \sum_{i\in\Omega} \pi(a)p(a,i) = \pi(a).$$

So detailed balance implies balance. □

QotD

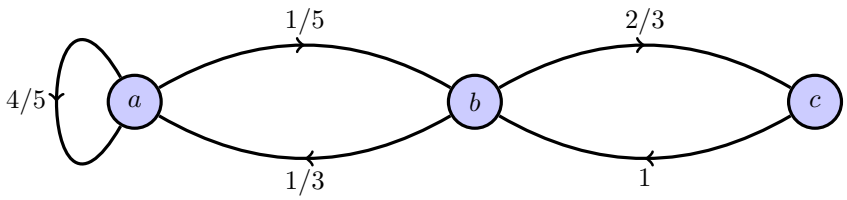
- Often easier to make a chain reversible than go for balance.
- $\pi(a) = 0.5, \pi(b) = 0.3,\ \pi(c) = 0.2.$
- For reversibility, $\pi(a)p(a,b) = \pi(b)p(b,a)$

$$p(a,b) = (3/5)p(b,a).$$

- Similarly, $\pi(b)p(b,c) = \pi(c)p(c,b)$.

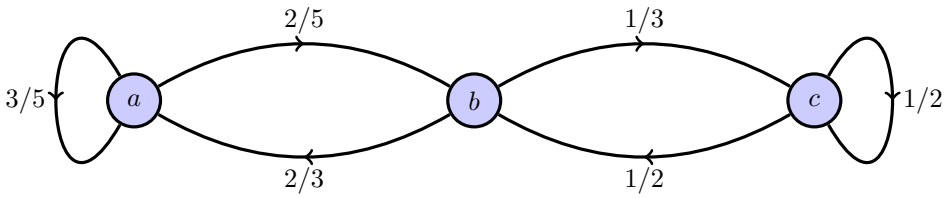
$$p(b,c) = (2/3)p(c,b).$$

- Don't have to worry about reversibility when $p(x,y) = p(y,x) = 0$.
- Now choose numbers that lie between 0 and 1. Let $p(c,b) = 1$. Then $p(b,c) = 2/3$. Let $p(b,a) = 1/3$, then $p(a,b) = 1/5$.



More than one chain has same stationary distribution

- More than one way to get π!
- Suppose $p(c,b) = 1/2$. Then $p(c,c) = 1/2$, $p(b,c) = 1/3$, $p(b,a) = 2/3$, $p(a,b) = 2/5$, $p(a,a) = 3/5$.



Symmetry

- Symmetric moves easy way to get uniform distribution.

Fact 68
Suppose $p(x, y) = p(y, x)$. Then the uniform distribution is stationary.

Proof. Let π be the uniform distribution. Then for all x and y in the state space, $\pi(x) = \pi(y)$, and $\pi(x)p(x, y) = \pi(y)p(y, x)$, so π is reversible, and hence stationary. □

Example

- Consider the Markov chain on permutations where two elements of the permutation are chosen uniformly at random and swapped (transposed).
- So if permutation is 43215 and elements 2 and 4 are swapped, new permutation is 41235.
- Consider two permutations x and y of $\{1, 2, \ldots, n\}$ connected by a transposition.
- Then $p(x, y) = p(y, x) = 1/n^2$.
- Hence the uniform distribution is stationary.
- Any sorting algorithm shows this chain is irreducible, so it is also the unique stationary distribution.
- Aperiodic since if you pick the same element twice to swap, you hold position.
- So also the limiting distribution.

Why "reversibility"?

- Consider two Markov chains on $\{1, 2, 3, 4\}$.
- Chain 1: $p(1, 2) = 1,\ p(2, 3) = 1,\ p(3, 4) = 1,\ p(4, 1) = 1$.
- Chain 2: $p(1, 2) = p(2, 1) = 1/2,\ p(2, 3) = p(3, 2) = 1/2,\ p(3, 4) = p(4, 3) = 1/2,\ p(4, 1) = p(1, 4) = 1/2$.
- A run of chain 1: $1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, \ldots$.
- If you reverse it, you immediately can tell its been reversed: $4, 3, 2, 1, 4, 3, 2, 1, \ldots$.
- A run of chain 2: $1, 4, 3, 4, 3, 4, 1, 2, 3, 4, 3, 4, 2, \ldots$.
- Looks the same run forward or backwards!

Fact 69
Suppose a Markov chain is reversible with respect to π and $X_0 \sim \pi$. Then

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = \mathbb{P}(X_0 = x_t, X_1 = x_{t-1}, \ldots, X_t = x_0).$$

That is, the distribution of $(X_0, X_1, \ldots, X_t)$ given $X_0 \sim \pi$ has the same distribution of $(X_t, X_{t-1}, \ldots, X_0)$ given $X_t \sim \pi$.

Proof. Consider:

$$\begin{aligned}
&\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) \\
&\qquad = \pi(x_0)p(x_0, x_1)p(x_1, x_2)\cdots p(x_{t-1}, x_t) \\
&\qquad = \pi(x_1)p(x_1, x_0)p(x_1, x_2)\cdots p(x_{t-1}, x_t) \\
&\qquad = \pi(x_2)p(x_1, x_0)p(x_2, x_1)p(x_2, x_3)\cdots p(x_{t-1}, x_t) \\
&\qquad = \qquad \vdots \\
&\qquad = \pi(x_t)p(x_t, x_{t-1})\cdots p(x_1, x_0) \\
&\qquad = \mathbb{P}(X_0 = x_t, X_1 = x_{t-1}, \ldots, X_t = x_0).
\end{aligned}$$

□

39 Metropolis-Hastings

Question of the Day Design a chain on $\{1, 2, 3, \ldots\}$ whose stationary distribution is

$$\pi(i) \propto 1/i^3.$$

Today

- Metropolis-Hastings method.

Normalizing constants

- Note the proportionality symbol $\propto$ in π.
- $\pi(i) = 1/i^3$ is not a probability distribution.

  $$\pi(i) = \frac{1/i^3}{\sum_{j=1}^{\infty} 1/j^3},$$

  is a probability distribution.
- Recall: call $\sum_{j=1}^{\infty} 1/j^3$ a normalizing constant of the distribution.
- More generally:

  $$\pi(i) = \frac{w(i)}{Z}, \quad Z = \sum_{x \in \Omega} w(x).$$

- Turns out: finding Z can be difficult in many cases.
- Often turn out to be $\#P$ complete problems.

Reversibility and normalizing constants

- Reversibility comes in when you can't find Z. Want:

  $$\pi(x)p(x, y) = \pi(y)p(y, x).$$

  All you need is

  $$w(x)p(x, y) = w(y)p(y, x).$$

- You don't need to know Z to build a Markov chain with π as the stationary distribution!
- Applications abound in statistics and estimation of Z.

Metropolis-Hastings

- Metropolis approach goes back to a 1953 paper with 5 authors.
- Nicholas Metropolis was project leader and first in alphabetical order, so his name was first on paper.
- Unclear whether he actually helped develop the algorithm.
- He did coin term: Monte Carlo method however.
- Authors were physicists, stuck to Boltzmann distributions.
- Later: Hastings gave general formulation for statisticians.
- Often called Metropolis-Hastings method.

Two steps in MH chain

1: Propose moving from $X_t = x$ to y using transitions $q(\cdot, \cdot)$.

2: With probability

$$\min\left\{\frac{w(y)q(y,x)}{w(x)q(x,y)}, 1\right\}$$

accept move and set $X_{t+1} = y$. Otherwise, $X_{t+1} = x$.

Using MH for QotD

- First need a proposal chain:

$$q(0,0) = 1/2, \quad \text{for all } i,\ q(i, i+1) = q(i+1, i) = 1/2.$$

  In words: fifty fifty chance of adding or subtracting 1, but don't go negative.

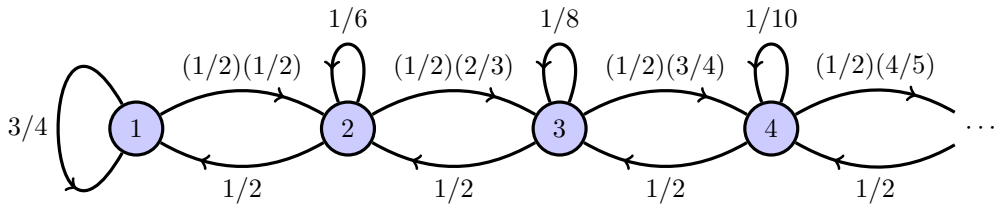
- If current state i, propose adding 1, accept move with probability:

$$\min\left\{\frac{(1/(i+1)^3)(1/2)}{(1/i^3)(1/2)}, 1\right\} = \left(\frac{i}{i+1}\right)^3.$$

- If current state $i > 0$, propose subtracting 1, accept move with probability:

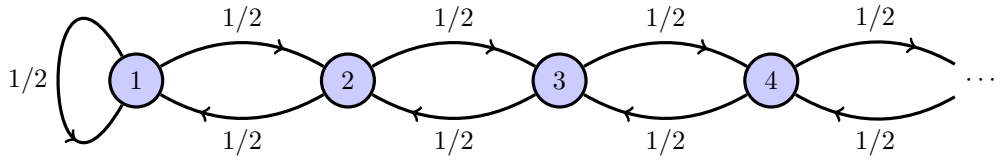
$$\min\left\{\frac{(1/i^3)(1/2)}{(1/(i+1)^3)(1/2)}, 1\right\} = 1.$$

- Then the final chain looks like:



Something interesting...

- Look at the original proposal chain again:



- This chain is recurrent and aperiodic, but does not have a stationary distribution.
- The point: You can use Metropolis-Hastings to turn a chain with no stationary distribution into one with the target distribution.

Fact 70
Suppose $q(x,y) > 0 \Rightarrow q(y,x) > 0$. Then the Metropolis-Hastings chain is reversible with respect to $\pi(i) = w(i)/Z$.

Proof. Let x and y be any two states in Ω with $w(x), w(y)$, $q(x,y)$ and $q(y,x)$ all positive. Without loss of generality $w(x)q(x,y) \leq w(y)q(y,x)$. Then

$$p(x,y) = q(x,y) \min\left\{\frac{w(y)q(y,x)}{w(x)q(x,y)}, 1\right\} = q(x,y)$$

and

$$p(y,x) = q(y,x) \min\left\{\frac{w(x)q(x,y)}{w(y)q(y,x)}, 1\right\} = \frac{w(x)q(x,y)}{w(y)}.$$

Hence

$$\pi(y)p(y,x) = \frac{w(y)}{Z}\frac{w(x)q(x,y)}{w(y)} = \frac{w(x)}{Z}q(x,y) = \pi(x)p(x,y).$$

□

Continuous Time Markov chains and reversibility

- MH needs the weird $\min\{\cdot, 1\}$ formulation to make probabilities between 0 and 1.
- Continuous time Markov chains just need rates nonnegative.
- Basic idea of reversibility:

$$\text{Flow from } x \text{ to } y = \text{Flow from } y \text{ to } x.$$

- In discrete time, this is $\pi(x)p(x,y) = \pi(y)p(y,x)$.
- For continuous time, this is $\pi(x)r(x,y) = \pi(y)r(y,x)$.

Definition 86
A probability distribution π on a discrete state space Ω satisfies the **detailed balance equations** (also called *reversibility* for a continuous time Markov chain if for all $a, b \in \Omega$:

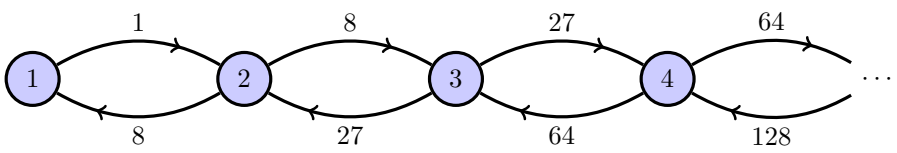
$$\pi(b)\lambda(b,a) = \pi(a)\lambda(a,b)$$

- [Here $r(x,y)$ is the rate of moving from x to y.]
- One easy way to accomplish when $\pi(x) = w(x)/Z$:

$$r(x,\cdot) = \frac{1}{w(x)}.$$

Back to the question of the day

- As a continuous chain, looks like:



Continuous state spaces

- Reversibility (and Metropolis-Hastings) also work with continuous state spaces.
- Recall X has density f if $\mathbb{P}(X \in dx) = f(x)\, dx$.

Definition 87
A probability distribution π on a continuous state space Ω satisfies the **detailed balance equations** (also called *reversibility* for a discrete time Markov chain if for all $a, b \in \Omega$:

$$\pi(y)\mathbb{P}(X_{t+1} \in dx, X_t = y) = \pi(x)\mathbb{P}(X_{t+1} \in dy, X_t = x).$$

40 Birth death chains

Question of the Day Are there chains for which stationary implies reversible?

Today

- Birth death chains

Reversibility and stationarity Here are the main ideas about reversiblility:

- If π is reversible for a Markov chain, it is also stationary for the chain.
- The inverse is not always true: one can have π stationary without it always being reversible.
- But for one particular kind of chain, a birth death chain, π is stationary if and only if it is also reversible.

Definition 88
A Markov chain on $\{0, 1, 2, \ldots\}$ is a **birth death chain** if the change in the state after one time step or jump is at most 1.

Using our earlier notation, if $p(x, y)$ or $r(x, y)$ is positive, then $|x - y| \leq 1$.

Fact 71
For birth death chains, if π is stationary, then π is reversible for the chain as well.

Proof. Assume π is stationary. For a continuous time Markov chain, the balance equations are

$$\begin{aligned}
\pi(0)r(0,1) &= \pi(1)r(1,0) \\
\pi(1)[r(1,0) + r(1,2)] &= \pi(0)r(0,1) + \pi(2)r(2,1) \\
\pi(2)[r(2,1) + r(2,3)] &= \pi(1)r(1,2) + \pi(3)r(3,2) \\
\vdots &= \vdots
\end{aligned}$$

The first equation is just reversibility between states 0 and 1. To get reversibility between states i and $i + 1$, just add the first $i + 1$ equations. For instance, the sum of the first three equations gives

$$\begin{aligned}
&\pi(0)r(0,1) + \pi(1)r(1,0) + \pi(1)r(1,2) + \pi(2)r(2,1) + \pi(2)r(2,3) \\
&\quad = \pi(1)r(1,0) + \pi(0)r(0,1) + \pi(2)r(2,1) + \pi(1)r(1,2) + \pi(3)r(3,2),
\end{aligned}$$

or after canceling,

$$\pi(2)r(2,3) = \pi(3)r(3,2).$$

More generally, for $j < i$, the term $\pi(j)r(j, j+1)$ appears on the left in equation $j-1$, and the term appears on the right in equation $j + 1$. Hence for $j < i$, these terms all cancel. Similarly, the term $\pi(j)r(j, j - 1)$ appears on the left in equation $j - 1$, and on the right in equation j.

However, $\pi(i)r(i, i + 1)$ only appears on the left (in equation $i + 1$) and $\pi(i + 1)r(i + 1, i)$ only appears on the right (in equation $i + 1$), so the only terms that do not cancel are

$$\pi(i)r(i, i + 1) = \pi(i + 1)r(i + 1, i),$$

which establishes reversibility. □

Even more generally, if π is stationary, then the probability flow across a cut in the graph (a subset of nodes) always balances the probability flow in the other direction across the cut.

However, in birth death chains, a cut of the form $\{0, 1, 2, \ldots, i\}$ only has edges from i to $i + 1$ and $i + 1$ to i crossing the cut. Hence reversibility is automatically met!

What's in a name?

- The reason for the name birth death chain, is that this can be thought of as a model for a population.
- $r(i, i+1)$ is the rate at which births occur when population is i.
- $r(i, i-1)$ is the rate at which deaths occur when population is i.
- Notation:
$$\lambda_i = r(i, i+1), \quad \mu_i = r(i, i-1).$$
- [λ stands for life, μ for mortis/death.]

Poisson Process

- Recall Poisson process keeps tracks of number of arrivals.
- Can be viewed as birth death chain...
$$\text{for all } i, \ \lambda_i = \lambda, \mu_i = 0.$$
- Of course, this chain has no stationary distribution.

Exponential growth

- Simple exponential growth of a population.
$$\text{for all } i, \ \lambda_i = \lambda i, \ \mu_i = 0.$$
- Also not stable...goes to infinity.
- Add deaths...
$$\text{for all } i, \ \lambda_i = \lambda i, \ \mu_i = \mu i.$$
- When $\lambda > \mu$ no stationary distribution, when $\mu > \lambda$, converges to 0 with probability 1.

Fact 72
A birth death process has a stationary distribution if and only if
$$\sum_{i=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_i}{\mu_1 \mu_2 \cdots \mu_{i+1}} < \infty.$$

Proof. A birth death process has stationary distribution π if and only if π is reversible for the chain. Reversibility here means:
$$\pi(i)\lambda_i = \pi(i+1)\mu_{i+1} \Leftrightarrow \pi(i+1) = \pi(i)\frac{\lambda_i}{\mu_{i+1}}.$$
A simple induction shows that for all i,
$$\pi(i) = \pi(0)\frac{\lambda_0 \lambda_1 \cdots \lambda_i}{\mu_1 \mu_2 \cdots \mu_{i+1}}.$$
This has solution
$$\pi(i) = \frac{\lambda_0 \lambda_1 \cdots \lambda_i}{\mu_1 \mu_2 \cdots \mu_{i+1}} \left[\sum_{j=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_j}{\mu_1 \mu_2 \cdots \mu_{j+1}} \right]^{-1}$$
which is a probability distribution if and only if the series in the brackets is finite. □

Example application

- Model of a queue (called M/M/1 queue in queuing theory).
- $\lambda_i = \lambda$, for $i \geq 0$, $\mu_i = \mu$ for $i \geq 1$, $\mu > \lambda$.
- λ is the rate at which arrivals come to the queue, μ is the rate at which people in the queue are served.
- What is the stationary distribution?
- In this case
$$\frac{\lambda_0 \cdots \lambda_i}{\mu_1 \cdots \mu_{i+1}} = \left(\frac{\lambda}{\mu}\right)^i,$$
- Since $\lambda < \mu$, this is a convergence geometric series, and the stationary distribution is:
$$\pi(i) = \frac{(\lambda/\mu)^i}{1 - \lambda/\mu}.$$

Yule process

- Suppose
$$\text{for all } i,\ \lambda_i = \lambda i^2,\ \mu_i = 0$$
- Expected time to go from 1 to 2 is $1/\lambda$,...
- ...from 2 to 3 is $\lambda^{-1}/2^2$,...
- ...time to get to infinity...
$$\frac{\lambda}{1^2} + \frac{\lambda}{2^2} + \frac{\lambda}{3^2} + \cdots = \frac{\lambda\pi^2}{6}.$$
- So with this level of growth, population reaches infinity in finite amount of time!
- Called a *population explosion.*

A Probability review

A.1 Elementary facts

Combinatorics The number of ways to arrange n objects in order is n factorial:

$$n! = n(n-1)(n-2)\cdots 1,$$

where $0! = 1$. The number of ways to choose r objects from n objects is:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

For $n_1 + n_2 + \ldots n_r = n$, the number of ways to choose n_1 objects of type 1, n_2 objects of type 2, up to n_r objects of type r, is

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

Definitions These are the basic definitions for talking about probability.

The set of outcomes is called the *sample space* or *outcome space*, and is usually denoted Ω.

An *event* is a subset E of Ω such that $\mathbb{P}(E)$ is defined (an event is also sometimes called a *measurable* subset). When A is an event, the complement of A is also an event. Also if $A_1, A_2, \ldots$ is a sequence of events, then $\cup_{i=1}^{\infty} A_i$ is also an event. (Any set of events with these properties is called a σ-algebra or σ-field.)

$\mathbb{P}$ is a function that given an event A, outputs the probability that the outcome lies in A.

The events A and B are *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$.

Measures A probability is a special type of measure that obeys the following four rules:

Rule 1: For event B, $0 \leq \mathbb{P}(B)$ (probabilities are nonnegative real numbers)

Rule 2: $\mathbb{P}(\emptyset) = 0$ (the probability nothing happens is zero).

Rule 3: For $B_1, B_2, \ldots$ disjoint events,

$$\mathbb{P}\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i).$$

Rule 4: $\mathbb{P}(\Omega) = 1$ (the probability that something occurs is 1).

Simple facts Some basic facts follow from these rules.

Prop: $0 \leq \mathbb{P}(A) \leq 1$.

Prop: $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$.

Prop: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$

Prop: $\mathbb{P}(\emptyset) = 0$.

A word about intersection For sets A and B, the intersection of A an B can be denoted $A \cap B$, AB, or A, B. All of these notations mean the same thing:

$$A \cap B := \{x : x \in A \text{ and } x \in B\}.$$

Conditional probabilities If $\mathbb{P}(B) > 0$, the conditional probability of A given B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

Bayes' Formula If $F_1, \ldots, F_n$ are disjoint and $\cup_{i=1}^{n} F_i = \Omega$, then

$$\mathbb{P}(F_i \mid A) = \frac{\mathbb{P}(A \mid F_i)\mathbb{P}(F_i)}{\mathbb{P}(A \mid F_1)\mathbb{P}(F_1) + \ldots \mathbb{P}(A \mid F_n)\mathbb{P}(F_n)}.$$

Random variables A *random variable* is a function of the outcome. The values the random variable can take on are called *states*, and lie in the *state space*. In other words, a random variable is a function from the sample space to the state space.

For a discrete random variable $X \in \{x_1, x_2, x_3, \ldots\}$, the expected value of X is

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i).$$

For a continuous random variable $X \in \mathbb{R}$ with density f_X, the expected value of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} s f_X(s)\, ds.$$

For any two random variables X and Y,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

For two random variables X and Y are *uncorrelated* if and only if

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Independent random variables (see below) are always uncorrelated, but uncorrelated random variables are not always independent!

Independence Two events A and B are *independent* if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) \Leftrightarrow \mathbb{P}(A \mid B) = \mathbb{P}(A).$$

Two random variables X and Y are independent if for any event $X \in A$ and $Y \in B$,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

A.2 A short guide to solving probability problems

Equally likely outcomes. If all outcomes are equally likely,

$$\mathbb{P}(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes}}.$$

Trick #1: Use complements. It is often easier to find $\mathbb{P}(A^C)$ then $\mathbb{P}(A)$, remember

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

Trick #2: Use independence to turn intersections into products. If we want the probability of the intersection of $A_1, \ldots, A_n$, then we can break it apart only when the events are independent:

$$\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \cdots \mathbb{P}(A_n).$$

Trick #3: Use disjointness to turn unions into sums. If the events $A_1, \ldots, A_n$ are disjoint,

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots \mathbb{P}(A_n).$$

Trick #4: Use Principle of In/Ex to deal with any union. We can *always* break apart unions of events $A_1 \ldots A_n$ using the Principle of Inclusion/Exclusion, which we use most often when $n = 2$:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 A_2).$$

Its easier to say the Principle of Inclusion/Exclusion in words than symbols: the probability of any event occurring is the sum of the probabilities that one event occurs minus the sum of the probabilities that 2 events occur plus the sum of the probabilities that 3 events occur etcetera until we reach the probability that all events occur.

Trick #5: Use De Morgan's Laws to covert unions and intersections. Convert back and forth between union and intersection using De Morgan's Laws:

$$(A_1 A_2 \cdots A_n)^C = A_1^C \cup A_2^C \cdots \cup A_n^C,$$

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^C = A_1^C A_2^C \cdots A_n^C.$$

Trick #6: Use Bayes' Formula to reverse conditional probabilities. If you know $\mathbb{P}(A \mid F_i)$ for all i as well as $\mathbb{P}(F_i)$, and want $\mathbb{P}(F_i \mid A)$, then use Bayes' Formula.

Trick #7: Acceptance/Rejection Method 1 Suppose that we perform a trial which if successful, has outcomes $A_1, \ldots, A_n$. If we fail, then we try again until one of A_1 through A_n occur. Then

$$\mathbb{P}(A_i \text{ occurs on final trial }) = \mathbb{P}(A_i \text{ on first trial}) \mid \text{first trial a success}) = \frac{\mathbb{P}(A_i \text{ on first trial})}{\mathbb{P}(\text{first trial a success})}.$$

Trick #8: Acceptance/Rejection Method 2 The other way to tackle acceptance rejection problem is using infinite series. Remember, when $|r| < 1$,

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}.$$

Common errors Some things to watch out for! Events use complements, unions, and intersections. A statement like $\mathbb{P}(A)^C$ doesn't make sense, since $\mathbb{P}(A)$ is a number. What was probably meant was $\mathbb{P}(A^C)$. Similarly, use +, - and times for numbers like probabilities, and never for sets. We haven't defined A + B, what was probably intended was $\mathbb{P}(A) + \mathbb{P}(B)$.

Steps to a problem: If you don't know how to get started on a problem, the following steps usually can get you going:

(1) Write down the sample space. Even if you can't write down the whole sample space, write down some of the outcomes. Make up symbols, like H for head or T for tails or W for win and L for a loss to make writing outcomes easier.

(2) Write down the events that you are given probabilities for, and the event that you are trying to find the probability of (the target event).

(3) See if you can express the target event in terms of union, intersection, or complements of the events that you are given (here is where the five tricks come into play).

Simple checks on an answer: Make sure that your final probabilities lie between 0 and 1. If you know that a set of probabilities must add to 1, then check by actually adding them. If you have a simple intuitive reason to believe that A is more likely than B, check that $\mathbb{P}(A) > \mathbb{P}(B)$.

A.3 A short guide to counting

Order matters When order matters, then there are $n!$ ways to order n objects.

Thinking about n choose k. There are several ways of thinking about $\binom{n}{k}$, all of which are equivalent.

(1) It's the number of the ways to choose a subset of size k from a set of size n.

(2) It's the number of ways to order a group of letters $A \ldots AB \ldots B$ where A appears k times and B appears $n-k$ times.

(3) Given n spaces, it's the number of ways to mark k of those spaces in some way.

(4) It's the number of ways of choosing k out of n trials to be successful.

Multichoosing Now $\binom{n}{n_1,\ldots,n_r}$ is similar, in that it generalizes $\binom{n}{k}$. This is because $\binom{n}{k} = \binom{n}{k,n-k}$. The number n multichoose $n_1, n_2, \ldots, n_r$ counts the following.

(1) It's the number of the ways to choose a partition of a set of size n where the first subset has size n_1, the second n_2, etcetera.

(2) It's the number of ways to order a group of letters $A_1 \ldots A_1 A_2 \ldots A_2 \ldots A_r \ldots A_r$ where A_i appears n_i times.

(3) Given n spaces, it's the number of ways to mark n_1 of those spaces with a 1, n_2 spaces with a 2, up to n_r spaces with n_r.

(4) Suppose each trial has r different outcomes. Then its the number of ways of choosing n_1 trials to have outcome 1, n_2 trials to have outcome 2, up to n_r trials having outcome r.

When all else fails. Almost any problem can be written as a problem with ordering. If you are uncomfortable with n choose r or can't figure out what should be ordered and what shouldn't then give everything in your problem a number and order everything.

For example, what's the probability of choosing a given five card hand from a set of 52 cards? One way: number of outcomes is 1, total number of outcomes is $\binom{52}{5}$, so

$$\mathbb{P}(\text{hand}) = \frac{1}{\binom{52}{5}}.$$

Another way: number all the cards $1, \ldots, 52$ and order them in any one of 52! ways. Then any outcome where the five cards we are interested in appear first in the ordering of cards works. There are 5! ways to order these cards and (52 - 5)! ways to order the remaining 47 cards, so the total number of outcomes is 5!(47!), so

$$\mathbb{P}(\text{hand}) = \frac{5!47!}{52!},$$

which is the same answer as the other way.

Another example: given a random ordering of the letters MIIIISSSSPP, what's the probability that it spells MISSISSIPPI? Think about numbering every symbol, so we are ordering $x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}$, where $x_1 = M$, x_2 through x_5 equal I, etc. Then the total number of outcomes is 11!. The number of outcomes that are successful? Well x_1 has to be in first position, x_2, x_3, x_4 and x_5 have to occupy positions $2, 5, 8$, and 10 (which they can do in 4! ways, there are 4! ways to order the x_i that equal S and 2! ways to order the x_i that equal P. So

$$\mathbb{P}(\text{MISSISSIPPI}) = \frac{1!4!4!2!}{11!}.$$

A.4 How to find $\mathbb{E}[X]$

Step 1 Find the values that X can take on with positive probability (this is called the *positive support* of X). If X is discrete, this will be either a finite number of values $\{x_1, \ldots, x_n\}$ or a countable number of values $\{x_1, x_2, \ldots\}$. If X is continuous, it could be an interval or union of intervals, like $(0, \infty)$ or $(3, 4) \cup [10, 15)$.

Step 2 Use the right formula. If X is discrete, then $\mathbb{E}[X]$ is the sum over all values of x such that $\mathbb{P}(X = x) > 0$ of the outcome times the probability. So if $X \in \{x_0, x_1, \ldots\}$, then

$$\mathbb{E}[X] = \sum_{x:p(x)>0} xp(x) = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i).$$

If X is continuous with density f_X then

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\, dx.$$

If $X \in \{0, 1, 2, 3, \ldots\}$, then the *Tail Sum Formula* gives an alternate way to find the expected value:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} \mathbb{P}(X > i).$$

If X is continuous and $\mathbb{P}(X \geq 0) = 1$, then the Tail Sum Formula is

$$\mathbb{E}[X] = \int_{x=0}^{\infty} \mathbb{P}(X > x)\ dx.$$

Conditional expectation To find $\mathbb{E}[A|B]$, treat B as a constant and calculate the probability in the exact same way as above. For all random variables A and B:

$$\mathbb{E}[\mathbb{E}[A|B]] = \mathbb{E}[A].$$

Note: If we wish to find $\mathbb{E}[g(X)]$ then use

$$\mathbb{E}[g(X)] = \sum_{x:\mathbb{P}(X=x)>0} g(x)\mathbb{P}(X = x) = \sum_{i=1}^{\infty} g(x_i)\mathbb{P}(X = x_i),$$

and

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(s)f_X(s)ds.$$

For uncorrelated random variables, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Independent random variables are uncorrelated, but uncorrelated random variables might not be independent.

Some properties of expected value:

- For any two random variables (correlated or uncorrelated) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

A.5 How to find $\mathbb{V}(X)$

Method 1: Use

$$\mathbb{V}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Method 2: Use

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Some properties

- For uncorrelated random variables, $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.
- For random variable X and constant $\alpha \in \mathbb{R}$, $\mathbb{V}(\alpha X) = \alpha^2\mathbb{V}(X)$, $\mathrm{SD}(\alpha X) = \alpha\,\mathrm{SD}(X)$.

A.6 Distributions

The *distribution* of a random variable is a complete listing of $\mathbb{P}(X \in A)$ for all sets A of interest. The distribution also referred to as the law of X, and denoted $\mathcal{L}(X)$. When X and Y have the same distribution, this is denoted

$$X \sim Y, \text{ or } \mathcal{L}(X) = \mathcal{L}(Y).$$

The *distribution function* of a random variable X (also known as the cumulative distribution function) is

$$F(a) = \mathbb{P}(X \leq a).$$

This is a function that is bounded, that is, it always lies between 0 and 1. It is also right continuous, that is if $a_1, a_2, a_3, \ldots$ decrease and their limit is a, then limit of $F(a_1), F(a_2), \ldots$ equals $F(a)$.

Because of a theorem from measure theory called the Carathéodory Extension Theorem, knowing F allows computation of $\mathbb{P}(X \in A)$ for any A of interest. In particular, if $A = (a, b]$, then $\mathbb{P}(X \in A) = F(b) - F(a)$. (Looks a bit like the fundamental theorem of calculus, which is one reason why F is always capitalized when used for the distribution function.)

More precisely, if F_X is the distribution function of X and F_Y is the distribution function of Y, then

$$\mathcal{L}(X) = \mathcal{L}(Y) \iff F_X(a) = F_Y(a)\ \forall a.$$

If X is discrete then the graph of $F(a)$ will have jumps, if X is continuous then $F(a)$ will be continuous. Some more formulas that come in handy:

$$\begin{aligned}
\mathbb{P}(a < X \leq b) &= F(b) - F(a) \\
\mathbb{P}(a < X < b) &= F(b) - F(a) - \mathbb{P}(X = b) \\
\mathbb{P}(a \leq X < b) &= F(b) - F(a) - \mathbb{P}(X = b) + \mathbb{P}(X = a) \\
\mathbb{P}(a \leq X \leq b) &= F(b) - F(a) + \mathbb{P}(X = a).
\end{aligned}$$

Remember that for continuous random variables $\mathbb{P}(X = s) = 0$ for any s, so the right hand side of these formula just becomes $F(b) - F(a)$. Also for continuous X,

$$f(a) = \frac{dF(a)}{da}$$

and

$$F(a) = \int_{-\infty}^{a} f(a)da,$$

where $f(x)$ is the *probability density function* (sometimes just called the density) of X.

Finally, say that $X_1, X_2, \ldots$ are independent identically distributed, or iid, if they are independent and all have the same distribution.

A.7 Discrete distributions

A random variable is *discrete* if it only takes on a finite or countably infinite number of values. The distribution of a discrete random variable is also called discrete in this instance.

Uniform Written: $\mathsf{Unif}(\{1, \ldots, n\})$. The story: roll a fair die with n sides.

$$\begin{aligned}
\mathbb{P}(X = i) &= \frac{1}{n}\mathbb{1}(i \in \{1, \ldots, n\}) \\
\mathbb{E}[X] &= \frac{n+1}{2} \\
\mathbb{V}(X) &= \frac{(n-1)(n+1)}{12}
\end{aligned}$$

Bernoulli Written: $\mathsf{Bern}(p)$. The story: flip a coin that comes up heads with probability p, and count the number of heads on the single coin flip. Also, the number of successes in a single trial where the trial is a success with probability p.

$$\begin{aligned}
\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) &= 1 - p \\
\mathbb{E}[X] &= p \\
\mathbb{V}(X) &= p(1-p).
\end{aligned}$$

Binomial Written: $\mathsf{Bin}(n, p)$. The story: flip iid coins n times where the probability of heads is p and count the number of heads. Also, the number of successes in a single trial where the trial is a success with probability p. Also if $X_1, \ldots, X_n$ are iid $\mathsf{Bern}(p)$, then $X = X_1 + X_2 + \ldots X_n \sim \mathsf{Bin}(n, p)$.

$$\begin{aligned}
\mathbb{P}(X = i) &= \binom{n}{i} p^i (1-p)^{n-i} \mathbb{1}(i \in \{\{0, \ldots, n\}) \\
\mathbb{E}[X] &= np \\
\mathbb{V}(X) &= np(1-p).
\end{aligned}$$

Geometric Written: $\mathsf{Geo}(p)$. The story: flip iid coins with probability p of heads and counting the number of flips needed for one head. Also, the number of trials needed for 1 success when the probability of success at each trial is p and each trial is independent.

$$\begin{aligned}\mathbb{P}(X = i) &= (1-p)^{i-1} p \mathbb{1}(\{0, 1, \ldots\}) \\ \mathbb{E}[X] &= \frac{1}{p} \\ \mathbb{V}(X) &= \frac{1-p}{p^2}.\end{aligned}$$

Negative Binomial Written: $\mathsf{NB}(r, p)$. The story: flipping iid coins with probability p of heads and counting the number of flips needed for r heads to arrive. Also, the number of trials needed for r successes when the probability of success at each trial is p and each trial is independent.

Also $X = X_1 + X_2 + \ldots X_r$, where X_i are iid and distributed as $\mathsf{Geo}(p)$.

$$\begin{aligned}\mathbb{P}(X = i) &= \binom{i-1}{r-1} p^r (1-p)^{i-r} \mathbb{1}(\{0, 1, \ldots\}) \\ \mathbb{E}[X] &= \frac{r}{p} \\ \mathbb{V}(X) &= r\frac{1-p}{p^2}.\end{aligned}$$

Hypergeometric Written: $\mathsf{HG}(N, m, n)$. The story: drawing n balls from an urn holding m green balls and $N - m$ blue balls and counting the number of green balls chosen.

$$\begin{aligned}\mathbb{P}(X = i) &= \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \mathbb{1}(\{0, 1, \ldots, n\}) \\ \mathbb{E}[X] &= \frac{nm}{N} \\ \mathbb{V}(X) &= \frac{N-n}{N-1} np(1-p).\end{aligned}$$

Zeta Written: $\mathsf{Zeta}(\alpha)$. A.k.a. Zipf or power law. The story: things like city sizes and incomes have Zeta distributions.

$$\begin{aligned}\mathbb{P}(X = i) &= \frac{C}{i^{\alpha+1}} \mathbb{1}(\{1, 2, \ldots\}) \\ \mathbb{E}[X] &= \text{no closed form} \\ \mathbb{V}(X) &= \text{no closed form.}\end{aligned}$$

Special notes: Except for special values of α like 1, we do not have a closed form solution for the value of C, the normalizing constant. Choose C so that $\sum_{i=1}^{\infty} \mathbb{P}(X = i) = 1$. Similarly, there are no closed form solutions for $\mathbb{E}[X]$ or $\mathbb{V}(X)$. These must be evaluated numerically. When $\alpha < 1$, $\mathbb{E}[X]$ does not exist (or is considered infinite). Similarly, when $\alpha < 2$, $\mathrm{Var}(X)$ does not exist (or can be considered infinite).

Poisson Written: $\mathsf{Pois}(\mu)$. The story: given that the chance of an arrival in time t to $t+dt$ is $\lambda\ dt$, and $\mu = \lambda T$, then this is the number of arrivals in the interval $[0, T]$. $X_1, X_2, \ldots$, it is

$$\max_i X_1 + X_2 + \ldots + X_i < 1.$$

$$\begin{aligned}
\mathbb{P}(X = i) &= e^{-\mu}\frac{\mu^i}{i!}\mathbb{1}(\{0, 1, \ldots\}) \\
\mathbb{E}[X] &= \mu \\
\mathbb{V}(X) &= \mu.
\end{aligned}$$

A.8 Continuous Distributions

A random variable is *continuous* if $\mathbb{P}(X = a) = 0$ for all a. The distribution of a continuous random variable is also called continuous.

Uniform (continuous) Written: $\mathsf{Unif}(A)$. The story: a point is uniform over A if for all $B \subseteq A$, the chance the point falls in B is the Lebesgue measure of B divided by the Lebesgue measure of A. For general A:

$$f(x) = \frac{1}{\text{Lebesgue measure of } A}\mathbb{1}(x \in A)$$

When $A = [a, b]$, more specifically:

$$\begin{aligned}
f(x) &= \frac{1}{b-a}\mathbb{1}(x \in (a, b]) \\
F(x) &= \frac{x-a}{b-a}\mathbb{1}(x \in [a, b]) + \mathbb{1}(x > b) \\
\mathbb{E}[X] &= \frac{b+a}{2} \\
\mathbb{V}(X) &= \frac{(b-a)^2}{12}
\end{aligned}$$

Normal Written: $\mathsf{N}(\mu, \sigma^2)$. The story: when you sum variables with finite mean and standard deviation together, they are well approximated by a normal distribution.

$$\begin{aligned}
f(x) &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\
F(x) &= \Phi\left(\frac{x-\mu}{\sigma}\right) \\
\mathbb{E}[X] &= \mu \\
\mathbb{V}(X) &= \sigma^2
\end{aligned}$$

Addition of normals. Adding independent normal random variables gives back another normal random variable. If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$, and $X = X_1 + X_2 + \ldots + X_n$, then

$$X \sim \mathsf{N}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right).$$

For X, Y independent $N(0, 1)$ random variables, the joint distribution of (X, Y) is rotationally invariant. Normal random variables are symmetric around μ, and so $\Phi(x) = 1 - \Phi(-x)$.

Exponential Written: $\mathsf{Exp}(\lambda)$. What it is: when events occur continuously over time at rate λ, this is the time you have to wait for the first event to occur.

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t} \mathbb{1}(t \in (0, \infty)) \\
F(t) &= \begin{cases} 1 - e^{-\lambda t} & a \geq 0 \\ 0 & a < 0 \end{cases} \\
\mathbb{E}[X] &= \frac{1}{\lambda} \\
\mathbb{V}(X) &= \frac{1}{\lambda^2}
\end{aligned}
$$

A.9 How to use the Central Limit Theorem (CLT)

The CLT says that if $X_1, X_2, \ldots$ are identically distributed random variables and $S_n = X_1 + \cdots + X_n$, then

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}(S_n)}} \leq a \right) = \Phi(a).$$

We use it as an approximation tool for $S = X_1 + \ldots X_n$:

$$\mathbb{P}\left(\frac{S - \mathbb{E}[S]}{\sqrt{\mathbb{V}(S)}} \leq a \right) \approx \Phi(a).$$

Often we are interested in approximating the probability of things like $\mathbb{P}(S \leq b)$ where $S = X_1 + \cdots + X_n$. This takes two steps.

Step 1 If S must be an integer, apply the half integer correction. So instead of $\mathbb{P}(S \leq i)$ we write $\mathbb{P}(S \leq i + 1/2)$.

Step 2 Subtract off $\mathbb{E}[S]$ and divide by the square root of $\mathrm{Var}(S)$. So

$$\mathbb{P}(S \leq b + 0.5) = \mathbb{P}\left(\frac{S - \mathbb{E}[S]}{\sqrt{\mathbb{V}(S)}} \leq \frac{b + 0.5 - \mathbb{E}[S]}{\sqrt{\mathbb{V}(S)}} \right).$$

Step 3 Apply the CLT and say

$$\mathbb{P}(S \leq b) \approx \Phi\left(\frac{b + 0.5 - \mathbb{E}[S]}{\sqrt{\mathbb{V}(S)}} \right).$$

A.10 Moment generating functions

Definition The *moment generating function* of a random variable X is $\mathrm{mgf}_X(t) = \mathbb{E}[\exp(tX)]$.

About the mgf While $\mathrm{mgf}_X(0) = 1$ for all X, it might be the case that the mgf_X does not exist for any t other than 0. All of the following results apply to values of t for which $\mathrm{mgf}_X(t)$ is finite.

Proposition If X and Y are independent, then $\mathrm{mgf}_{X+Y}(t) = \mathrm{mgf}_X(t) \cdot \mathrm{mgf}_X(t)$.

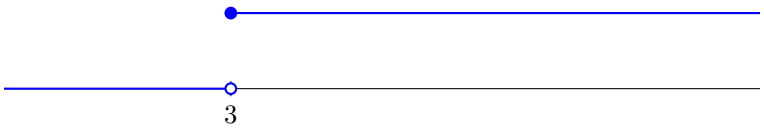
Proposition If $\mathrm{mgf}_X(t)$ exists for $tin(a, b)$ where $a < 0 < b$, and $[f]^{(i)}$ is the ith derivative of a function f, then $[\mathrm{mgf}_X(t)]^{(i)}|_{t=0} = \mathbb{E}[X^i]$.

B Definitions of functions

B.1 Indicator function

Definition 89
The **indicator function**, written $\mathbb{1}(\cdot)$, takes an argument that is either true or false, and returns 1 if the argument is true and 0 if it is false.

Examples: $\mathbb{1}(3 < 4) = 1$, $\mathbb{1}(3 > 4) = 0$. The graph of $\mathbb{1}(x \geq 3)$ looks like:



B.2 Minimum and maximum

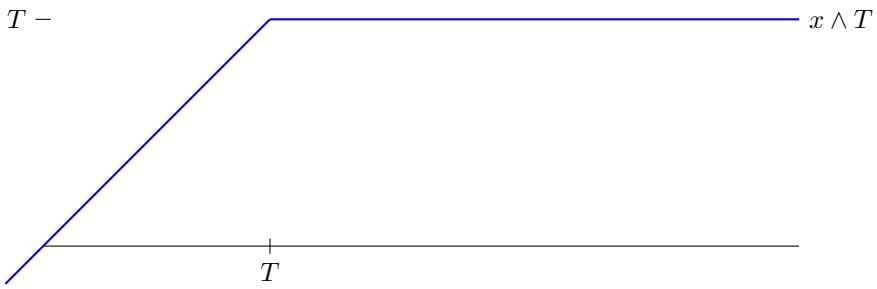
Definition 90
The **minimum** of a and b real numbers is

$$\min\{a, b\} = a \wedge b = a\mathbb{1}(a \leq b) + b\mathbb{1}(a > b).$$

Definition 91
The **maximum** of a and b real numbers is

$$\max\{a, b\} = a \vee b = a\mathbb{1}(a \geq b) + b\mathbb{1}(a < b).$$

Examples: $\max\{3, 4\} = 4$, $3 \wedge 4 = 3$. The graph of $f(x) = \min\{x, T\}$ looks like



B.3 Ceiling and floor

Definition 92
The **ceiling** of $x \in \mathbb{R}$, written $\lceil x \rceil$ is the smallest integer that is greater than or equal to x.

Examples: $\lceil 4.3 \rceil = 5$, $\lceil 4 \rceil = 4$, $\lceil -2.3 \rceil = -2$.

Definition 93
The **floor** of $x \in \mathbb{R}$, written $\lfloor x \rfloor$ is the greatest integer that is less than or equal to x.

Examples: $\lfloor 4.3 \rfloor = 4$, $\lfloor 4 \rfloor = 4$, $\lfloor -2.3 \rfloor = -3$.

C Vector Spaces

A set V together with another set S form a *vector space* if they have certain operations that obey nice properties.

Definition 94
Say that V is a **vector space** with **scalars** S if there exists vector addition, scalar multiplication, scalar addition, and vector-scalar multiplication such that

1: $(\forall v \in V)(\forall s \in S)(sv \in V)$

2: $(\forall v_1, v_2 \in V)(v_1 + v_2 \in V)$

3: $(\forall v_1, v_2, v_3 \in V)(v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3)$

4: $(\forall v_1, v_2 \in V)(v_1 + v_2 = v_2 + v_1)$

5: $(\exists 0 \in V)(\forall v \in V)(0 + v = v)$

6: $(\forall s_1, s_2 \in S)(\forall v \in V)(s_1(s_2 v) = (s_1 s_2)v)$

7: $(\exists 1 \in S)(\forall s \in S)(1s = s)$

8: $(\forall s \in S)(\forall v_1, v_2 \in V)(s(v_1 + v_2) = sv_1 + sv_2)$

9: $(\forall s_1, s_2 \in S)(\forall v \in V)((s_1 + s_2)v = s_1 v + s_2 v)$

For examples, for V equal to the set of real-valued random variables and S equal to the real numbers, (V, S) form a vector space.

D Proofs of theorems

D.1 Expectation is a linear operator

To show that expectation is a linear operator, first show for simple functions. This was done as Facts 2 and 3 in the text. Combining these facts gives the following.

Fact 73
For X and Y simple random variables with $a, b \in \mathbb{R}$,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Recall that the expected value of a nonnegative random variable X is the supremum of the expected value of simple random variables dominated by X. So first show the following facts about supremum.

Fact 74
Let $A \subseteq B$. Then $\sup A \geq \sup B$.

Proof. Since $\sup B$ is an upper bound on every element of B, it is also an upper bound on every element of A. Since the supremum is the least upper bound of A, and $\sup B$ is an upper bound, $\sup A \leq \sup B$. □

Fact 75
For $A \subset \mathbb{R}$, let $cA = \{b : b = ca \text{ for some } a \in A\}$. Then $\sup cA = c \sup A$.

Proof. Let $ca \in cA$. Then $\sup A \geq a$ and so $c \sup A \geq cA$. Since this holds for arbitrary $ca \in cA$, $c \sup A \geq \sup cA$.

For the other direction, let $a \in A$. Then $ca \in cA$, so $\sup cA \geq ca$ and $(\sup cA)/c \geq a$. Since a was an arbitrary element of A, $(\sup cA)/c \geq \sup A$, which gives $\sup cA \geq c \sup A$. □

Fact 76
Let A have finite supremum M. Then

$$(\forall \epsilon > 0)(\exists a \in A)(a > M - \epsilon)$$

Proof. The contrapostive is easier to show: if $(\exists \epsilon > 0)(\forall a \in A)(a \leq M - \epsilon)$, then M is not the supremum of A.

Let $\epsilon > 0$ satisfy for all $a \in A$, $a \leq M - \epsilon$. Then $\sup M \leq M - \epsilon < M$. So $\sup A \neq A$. □

Fact 77
Let A and B be subsets of $\mathbb{R}$ with finite supremum, then for

$$C = A + B = \{c : c = a + b, \text{ where } a \in A, b \in B\},$$

then $\sup C = \sup A + \sup B$.

Proof. Let $c = a + b$ be an arbitrary element of C, where $a \in A$ and $b \in B$. Then $a \leq \sup A$ and $b \leq \sup B$, so $c \leq \sup A + \sup B$.

Let $\epsilon > 0$. From the previous fact there exists $a \in A$ with $a \geq \sup A - \epsilon/2$. Similarly there exists $b \in B$ with $b \geq \sup B - \epsilon/2$. So there exists $c = a + b \in C$ such that $c \geq \sup A$. □

Fact 78
For any nonnegative integrable random variables X and Y, and real numbers a and b:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof. Let W be a simple function with $W \leq X$, and Z be a simple function with $Z \leq Y$. Then $W+Z$ is also a simple function dominated by $X+Y$ where $\mathbb{E}[W+Z] = \mathbb{E}[W] + \mathbb{E}[Z]$. Now if $W \leq X$, $Z \leq Y$, then $W+Z \leq X+Y$. So $\{A : A = W+Z, W \leq X, Z \leq Y\} \subseteq \{B : B \leq X+Y\}$. So

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sup_{A \leq X+Y} \mathbb{E}[A] \\
&\geq \sup_{W \leq X, Z \leq Y} \mathbb{E}[W+Z] \\
&= \sup_{W \leq X, Z \leq Y} \mathbb{E}[W] + \mathbb{E}[Z] \\
&= \sup_{W \leq X} \mathbb{E}[W] + \sup_{Z \leq Y} \mathbb{E}[Z] && \text{(by Fact 77)} \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

Now for the other direction, which is somewhat trickier. Let $\epsilon > 0$. Recall $\mathbb{E}[X+Y] = \sup \mathbb{E}[A]$ taken over A such that A is simple and $A \leq X+Y$. So from Fact 76, there exists a simple A dominated by $X+Y$ such that $\mathbb{E}[A] \geq \mathbb{E}[X+Y] - \epsilon/3$.

Now let $A_1 = \min\{A, X\}$ and $A_2 = \min\{A, Y\}$. Then by looking at the possible values for A_1 and A_2, it is easy to see that $A \leq A_1 + A_2 \leq X+Y$. The problem is that A_1 and A_2 might no longer be simple. However, since A was simple, A is bounded above by its maximum value, call in M.

Now let $f_\alpha(x)$ be the function that rounds x down to the nearest multiple of α. That is, $f_\alpha(x) = \alpha \lfloor x/\alpha \rfloor$ where $\lfloor \cdot \rfloor$ is the floor function that rounds its argument down to the nearest integer.

Note that $\lfloor x/\alpha \rfloor \geq x/\alpha - 1$, so $f_\alpha(x) \geq x - \alpha$.

Then $f_\alpha(A_1)$ and $f_\alpha(A_2)$ are simple, and $f_\alpha(A_i) \geq A_i - \alpha$. Setting $\alpha = \epsilon/3$, that gives $f_{\epsilon/3}(A_1) \leq X$ and is simple, $f_{\epsilon/3}(A_2) \leq Y$ and is simple. Moreover,

$$\begin{aligned}
\mathbb{E}[X] + \mathbb{E}[Y] &\geq \mathbb{E}[f_{\epsilon/3}(A_1)] + \mathbb{E}[f_{\epsilon/3}(A_2)] \\
&\geq \mathbb{E}[f_{\epsilon/3}(A_1) + f_{\epsilon/3}(A_2)] \\
&\geq \mathbb{E}[A_1 - \epsilon/3 + A_2 - \epsilon/3] \\
&\geq \mathbb{E}[A] - 2\epsilon/3 \\
&\geq \mathbb{E}[X+Y] - \epsilon/3 - 2\epsilon/3 = \mathbb{E}[X+Y] - \epsilon.
\end{aligned}$$

Since this was true for arbitrary $\epsilon > 0$, $\mathbb{E}[X] + \mathbb{E}[Y] \geq \mathbb{E}[X+Y]$. □

Fact 79
For any $a \geq 0$, and nonnegative random variable X, $\mathbb{E}[aX] = a\mathbb{E}[X]$.

Proof. This holds trivially if $a = 0$. Suppose $a > 0$. Then Y is a simple function dominated by X if and only if aY is a simple function dominated by aX.

Since scaling holds for simple functions

$$\begin{aligned}
\mathbb{E}[aX] &= \sup_{aY \text{:simple and } aY \leq aX} \mathbb{E}[aY] \\
&= \sup_{Y \text{:simple and } Y \leq X} a\mathbb{E}[Y] \\
&= a \sup_{Y \text{:simple and } Y \leq X} \mathbb{E}[Y] \\
&= a\mathbb{E}[X].
\end{aligned}$$

□

Fact 80
For integrable X and Y, $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Proof. Write $X = X^+ - X^-$ and $Y = Y^+ - Y^-$, where X^+, X^-, Y^+, and Y^- are nonnegative. Then $X^+ + Y^+$ and $X^- + Y^-$ are nonnegative, and $X + Y = (X^+ + Y^+) - (X^- + Y^-)$, so

$$\begin{aligned}\mathbb{E}[X+Y] &= \mathbb{E}[X^+ + Y^+] - \mathbb{E}[X^- + Y^-] \\ &= \mathbb{E}[X^+] + \mathbb{E}[Y^+] - (\mathbb{E}[X^-] + \mathbb{E}[Y^-]) \\ &= \mathbb{E}[X^+] - \mathbb{E}[X^-] + \mathbb{E}[Y^+] - \mathbb{E}[Y^-] \\ &= \mathbb{E}[X] - \mathbb{E}[Y].\end{aligned}$$

□

Fact 81
For $a \geq 0$, and integrable X, $\mathbb{E}[aX] = a\mathbb{E}[X]$.

Proof. If $a = 0$, this is just $0 = 0$, which is true.

Suppose $a > 0$. Let $X = X^+ - X^-$ where X^+ and X^- are nonnegative. Then $aX = aX^+ - aX^-$ where aX^+ and aX^- are nonnegative. So

$$\mathbb{E}[aX] = \mathbb{E}[aX^+] - \mathbb{E}[aX^-] = a\mathbb{E}[X^+] - a\mathbb{E}[X^-] = a(\mathbb{E}[X^+] - \mathbb{E}[X^-]) = a\mathbb{E}[X].$$

Now suppose $a < 0$. Then $aX = (-a)X^- - (-a)X^+$, where $(-a)X^-$ and $(-a)X^+$ are nonnegative. So

$$\mathbb{E}[aX] = \mathbb{E}[(-a)X^-] - \mathbb{E}[(-a)X^+] = (-a)\mathbb{E}[X^-] - (-a)\mathbb{E}[X^+] = a(\mathbb{E}[X^+] - \mathbb{E}[X^-]) = a\mathbb{E}[X].$$

□

Fact 82
For integrable X and Y and real a and b,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Proof. This follows direction from the previous two facts. □

D.2 Convergence with probability 1 implies convergence in probability

In this section we will prove that convergence wp 1 implies convergence in probability. First, a helpful result about limits.

Fact 83
Suppose we have a sequence of decreasing events: $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$. Then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\cap_{n=1}^{\infty} A_n\right).$$

Proof. Note that for any event $\mathbb{P}(A) = \mathbb{E}[\mathbb{1}(A)]$. Since $|\mathbb{1}(A)| \leq 1$, the bounded convergence theorem 10 allows us to say

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \mathbb{E}[\mathbb{1}(A_n)] = \mathbb{E}[\lim_{n\to\infty} \mathbb{1}(A_n)]$$

Since the A_n are decreasing, the only way $\lim_{n\to\infty} \mathbb{1}(A_n)$ exists is if $\mathbb{1}(A_n) = 0$ for all $n \geq N$ or $\mathbb{1}(A_n) = 1$ for all n. But that means $\lim_{n\to\infty} = \mathbb{1}(\cap A_n)$. So

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{E}[\mathbb{1}(\cap A_n)] = \mathbb{P}(\cap A_n),$$

and we are done. □

Fact 84
Suppose that $X_t \to X$ with probability 1. Then $X_t \to X$ in probability.

Proof. Suppose $X_t \to X$ wp 1. That is, the set of outcomes $A = \{\Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}$ satisfies $\mathbb{P}(\omega \in A) = 0$.

Fix $\epsilon > 0$. We want to show that $\mathbb{P}(|X_n - X| > \epsilon) \to 0$. To do that, consider the events

$$A_N = \cup_{n\geq N} |X_N - X| \geq \epsilon.$$

Note that $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$. Suppose that $\omega \in A_N$ for all N. Then that means that for every N, there exists some $n \geq N$ where $|X_n - X| > \epsilon$. That means that $\lim_{n\to\infty} X_n(\omega) \neq X(\omega)$. So

$$\cap_{N=1}^{\infty} A_N \subseteq A.$$

Hence

$$\mathbb{P}\left(\cap_{N=1}^{\infty} A_n\right) \leq \mathbb{P}(A) = 0 \Rightarrow \mathbb{P}\left(\cap_{N=1}^{\infty} A_n\right) = 0.$$

The previous fact then gives $\lim_{n\to\infty} P(A_n) = 0$. But that is exactly what is required for convergence in probability! □

E Problem Solutions

30.1: Suppose that the arrivals of airport shuttles at a particular stop follow a Poisson process of rate 1/[10 min].

(a) On average, how many shuttles will arive in an hour?

(b) What is the chance that there are no shuttles in the first 20 minutes?

(c) What is the chance that the second shuttle arrives somewhere in $[15, 25]$ minutes?

Solution

(a) The average number of shuttles in the first 1 hour (60 minutes) is $60/10 = \boxed{6}$.

(b) The number of shuttles in the first 20 minutes is Poisson distributed with mean $20/10 = 2$. Therefore the chance that $N_2 \sim \mathsf{Pois}(2)$ has $N_2 = 0$ is $\exp(-2)2^0/0! \approx 0.1353$.

(c) The time T_2 of the second shuttle arrival has an Erlang/gamma distribution with parameters 2 and 1/10. This has density

$$f_{T_2}(t) = \frac{(1/10)^2 t \exp(-t/10)}{(2-1)!},$$

so

$$\begin{aligned}
\mathbb{P}(T_2 \in [15, 25]) &= \int_{t=15}^{25} (1/100) t \exp(-t/10)\ dt \\
&= (1/100) \int_{t=15} t[-10\exp(-t/10)]'\ dt \\
&= (1/100) \int_{t=15} [-10t\exp(-t/10)]' - [t]'(-10\exp(-t/10))dt \\
&= (-t/10\exp(-t/10)|_{15}^{25} - \int_{t=15}^{25} -(1/10)\exp(-t/10)\ dt \\
&= 1.5\exp(-1.5) - 2.5\exp(-2.5) - \exp(-t/10)|_{15}^{25} \\
&= 2.5\exp(-1.5) - 3.5\exp(-2.5) \approx \boxed{0.2705}.
\end{aligned}$$

Index