Introduction to Stochastic Processes
Math 310-2

Introduction

In Math 310-1, you studied random variables and sequences of random variables that were *independent*. Because they were independent, they satisfied properties like the central limit theorem and the law of large numbers.

But, many worldly phenomena are not independent. In this class we will study a basic type of non-independent process—Markov chains.

Markov Chain

DEF

A sequence of random variables (X_n) is a ***Markov chain*** (or Markov for short) if for n and any sequence of events (A_i) having positive probability,

$$\mathbb{P}(X_n \in A_n \mid X_{n-1} \in A_{n-1}) = \mathbb{P}(X_n \in A_n \mid X_{n-1} \in A_{n-1}, X_{n-2} \in A_{n-2}, \ldots).$$

In other words, (X_n) is *Markov* if knowing the previous value in the sequence gives you the same amount of information as knowing all previous values. Since we will be looking at conditional probabilities a lot, to keep things less cluttered, we will sometimes avoid writing the events. That is,

$$\mathbb{P}(X_n \in A_n \mid X_{n-1}, X_{n-2}, \ldots) \equiv \mathbb{P}(X_n \in A_n | X_{n-1} \in A_{n-1}, X_{n-2} \in A_{n-2}, \ldots)$$

where (A_i) is a sequence of positive probability events. Further, $\mathbb{P}(X_n)$, will mean the distribution of X_n.

Let's look at some sequences and decide whether or not they are Markov.

1. You're walking around Chicago. At each street intersection, you roll a die to decide which street to go down next. (X_n) is the sequence of streets you walk down.

   Markov! The next street you decide to go down depends only on your current position, not on how you got there.

2. You're trying to hack a computer network. Every minute you have a 50% chance to crack another password. If you've cracked $k/2$ passwords in the last k minutes, your heart-rate gets high, otherwise it is normal. (X_i) is your heart-rate in the ith minute.

   Markov, but not *stationary*. For minute n, if you are told that X_{n-1} =caught, you know with probability 1 that X_n will be caught. Alternatively, if you're told X_{n-1} =not caught, you know that $X_{n-2}, \ldots$ =not caught, and so you get know extra information knowing all previous states.

3. You're trying to hack a computer network. Every minute you have a 50% chance to crack another password. If you crack k passwords, you will be discovered. (X_i) is how many passwords you've cracked by the ith minute.

   Markov! If you've cracked n passwords in the ith minute, you still have a 50/50 chance of cracking your $(n+1)$th password in the $(i+1)$th minute, regardless of whether you cracked n or $n-1$ in the $(i-1)$th minute.

4. (X_i) is an independent and identically distributed (iid) sequence of Bernoulli $(1/3, 2/3)$ random variables.

   Markov! We trivially have $\mathbb{P}(X_i \mid X_{i-1}) = \mathbb{P}(X_i) = \mathbb{P}(X_i \mid X_{i-1}, X_{i-2}, \ldots)$ by independence.

5. You repeatedly take a test to become part of your favorite fan club. (X_i) is your average score on the first i tests.

   Markov, but not *stationary*. This one is subtle. Suppose that your score on the ith test is given by the independent sequence S_i. Then

   $$X_i = \frac{1}{n}\sum_{n=1}^{i} S_i.$$

   We then have $\mathbb{P}(X_i = a \mid X_{i-1} = b) = \mathbb{P}\left(\frac{S_i+(i-1)b}{i} = a\right)$. Since S_i is an independent sequence, knowing X_{i-2} doesn't change this probability.

Day 2

Notation

Let (X_i) have the Markov property. We call the codomain$(X_i) = X^{(i)}$ the *state space*, and we will assume $X^{(i)} = X^{(j)}$.

$\mathbb{P}(X_i = b \mid X_{i-1} = a)$ is called the *transition probability* from state a to state b at time i. If

$$\mathbb{P}(X_i = b \mid X_{i-1} = a) = \mathbb{P}(X_j = b \mid X_{j-1} = a) \qquad \text{for all } i \text{ and } j,$$

we call (X_i) *stationary*. We will only consider stationary processes.

For a stationary process (X_i),

$$P(a,b) \equiv \mathbb{P}(X_i = b \mid X_{i-1} = a) \qquad \text{and} \qquad P^n(a,b) \equiv \mathbb{P}(X_i = b \mid X_{i-n} = a)$$

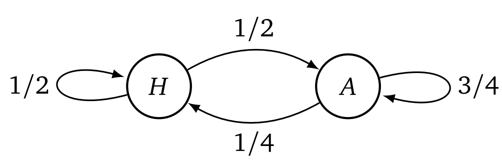
are called the *one-step transition function* and the *multi-step transition function*.

A Simple Example

Your very conservative aunt decided to go to a Chicago comedy club. Each joke leaves her happy (H) or angry (A). If she is happy, the next joke has a 50/50 chance of leaving her happy or angry. If she is angry, the next joke has a 1/4th chance of making her happy and a 3/4ths chance of making her angry.

Day 3

We can model this situation with a directed graph:



DEF

The ***transition matrix*** associated with a stationary Markov chain (X_i) having n states $\{1, 2, \ldots, n\}$ is the matrix $T = [t_{ij}]$ where $t_{ij} = P(i,j)$.

For this example, letting happy be the first state and angry be the second state, we have

$$T = \begin{array}{c c} & \begin{array}{cc} H & A \end{array} \\ \begin{array}{c} H \\ A \end{array} & \begin{bmatrix} \mathbb{P}(H \to H) & \mathbb{P}(H \to A) \\ \mathbb{P}(A \to H) & \mathbb{P}(A \to A) \end{bmatrix} \end{array} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}.$$

Let's compute $P^2(A,A) = \mathbb{P}(X_i = A \mid X_{i-2} = A)$. There are only two ways for this to happen. Either the states went

$$A \to A \to A \qquad \text{or} \qquad A \to H \to A.$$

$\mathbb{P}(A \to A \to A) = \frac{3}{4}\cdot\frac{3}{4} = \frac{9}{16}$ and $\mathbb{P}(A \to H \to A) = \frac{1}{4}\cdot\frac{1}{2} = \frac{2}{16}$, and so

$$P^2(A,A) = \mathbb{P}(A \to A \to A) + \mathbb{P}(A \to H \to A) = \frac{11}{16}.$$

The P^2 transition matrix now looks like $\begin{bmatrix} ? & ? \\ ? & \frac{11}{16}\end{bmatrix}$. Computing every entry we get $\begin{bmatrix} \frac{6}{16} & \frac{10}{16} \\ \frac{5}{16} & \frac{11}{16}\end{bmatrix}$.

Let's look at what happened from a matrix perspective: To compute $P^2(A,A)$, we need to compute the quantity $Q = \mathbb{P}(A \to A \to A) + \mathbb{P}(A \to H \to A)$, and by the Markov property,

$$\mathbb{P}(A \to A \to A) = \mathbb{P}(A \to A)\mathbb{P}(A \to A) \qquad \text{and} \qquad \mathbb{P}(A \to H \to A) = \mathbb{P}(A \to H)\mathbb{P}(H \to A).$$

If we squint, we see

$$Q = \begin{bmatrix}\mathbb{P}(H \to A)\\ \mathbb{P}(A \to A)\end{bmatrix}\cdot\begin{bmatrix}\mathbb{P}(A \to H)\\ \mathbb{P}(A \to A)\end{bmatrix} = \begin{bmatrix}\mathbb{P}(H \to A) & \mathbb{P}(A \to A)\end{bmatrix}\begin{bmatrix}\mathbb{P}(A \to H)\\ \mathbb{P}(A \to A)\end{bmatrix} = (\begin{bmatrix}0 & 1\end{bmatrix}T)\left(T\begin{bmatrix}0\\1\end{bmatrix}\right) = \begin{bmatrix}0 & 1\end{bmatrix}T^2\begin{bmatrix}0\\1\end{bmatrix}.$$

In other words, the $(2,2)$ entry of the transition matrix for $P^2(x,y)$ is the $(2,2)$ entry of T^2. In fact, the transition function $P^2(x,y)$ is given by the matrix T^2!

Using induction, we can show that the transition function $P^n(x,y)$ is given by the matrix T^n for any n, but first let's prove a lemma we've already used once.

LEM

If (X_i) is a Markov chain and (s_i) is a sequence of states of (X_i), then

$$\mathbb{P}(s_1 \to s_2 \to \cdots \to s_n) = \prod_{i=1}^{n-1}\mathbb{P}(s_i \to s_{i-1}).$$

We used this lemma already when we said $\mathbb{P}(A \to H \to A) = \mathbb{P}(A \to H)\mathbb{P}(H \to A)$, and it seems quite intuitive given the definition of Markov.

Proof. We will proceed by induction. Trivially, $\mathbb{P}(s_1 \to s_2) = \mathbb{P}(s_1 \to s_2)$. Suppose $\mathbb{P}(s_1 \rightsquigarrow s_n) = \prod \mathbb{P}(s_i \to s_{i+1}$, and consider $\mathbb{P}(s_1 \rightsquigarrow s_n \to s_{n+1})$. By definition

$$\mathbb{P}(s_n \to s_{n+1}) = \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n)$$
$$= \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, X_{n-1} = s_{n-1}, \ldots)$$

with the last equality following from the Markov property. Now by the definition of conditional probability we have

$$\mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, X_{n-1} = s_{n-1}, \ldots) = \frac{\mathbb{P}(X_{n+1} = s_{n+1}, X_n = s_n, X_{n-1} = s_{n-1}, \ldots)}{\mathbb{P}(X_n = s_n, X_{n-1} = s_{n-1}, \ldots)} = \frac{\mathbb{P}(s_1 \rightsquigarrow s_n \to s_{n+1})}{\mathbb{P}(s_1 \rightsquigarrow s_n)}.$$

Multiplying both sides by $\mathbb{P}(s_1 \rightsquigarrow s_n)$ gives us

$$\mathbb{P}(s_1 \rightsquigarrow s_n \to s_{n+1}) = \mathbb{P}(s_1 \rightsquigarrow s_n)\mathbb{P}(s_n \to s_{n+1}).$$

Since we assume that $\mathbb{P}(s_1 \rightsquigarrow s_n)$ could be written as a product, we've completed the proof. □

Day 4

We're almost ready to tackle our first theorem, but let's first prove two more easy facts.

LEM

If (X_i) is a Markov chain with finite state space S, then

$$\mathbb{P}(X_n = b) = \sum_{s\in S}\mathbb{P}(X_n = b \mid X_{n-1} = s)\mathbb{P}(X_{n-1} = s).$$

Proof. Since $\mathbb{P}(X_i \in S) = 1$, we have

$$\mathbb{P}(X_n = b) = \mathbb{P}(X_n = b \text{ and } X_{n-1} \in S) = \mathbb{P}\left(\bigcup_{s\in S} X_n = b \text{ and } X_{n-1} = s\right).$$

Since the events ($X_n = b$ and $X_{n-1} = s$) and ($X_n = b$ and $X_{n-1} = s'$) are disjoint if $s \neq s'$, we have

$$\mathbb{P}\left(\bigcup_{s\in S} X_n = b \text{ and } X_{n-1} = s\right) = \sum_{s\in S} \mathbb{P}(X_n = b \text{ and } X_{n-1} = s).$$

Noticing the summand is just $\mathbb{P}(X_n = b \mid X_{n-1} = s)\mathbb{P}(X_{n-1} = s)$ completes the proof. □

LEM If (X_i) is a Markov chain with finite state space S, then $\mathbb{P}(X_n = b \mid X_1 = a)$ is the sum of $\mathbb{P}(a \to s_2 \to s_3 \to \cdots \to s_{n-1} \to b)$ for every possible combination of $s_2, \ldots, s_{n-1} \in S$.

Proof. Like the previous proof, this follows from manipulations of the basic definitions.

First note, if $s_2, \ldots, s_{n-1}$ and $s_2', \ldots, s_{n-1}'$ are not the exact same sequence, then the events $a \to s_2 \to \cdots \to s_{n-1} \to b$ and $a \to s_2' \to \cdots \to s_{n-1}' \to b$ are disjoint. Thus

$$\mathbb{P}\left(\bigcup_{s_2,\ldots,s_{n-1}\in S} a \to s_2 \to \cdots \to s_{n-1} \to b\right) = \sum_{s_2,\ldots,s_{n-1}\in S} \mathbb{P}(a \to s_2 \to \cdots \to s_{n-1} \to b).$$

Now, by the definition of conditional probability,

$$\mathbb{P}(a \to s_2 \to \cdots \to s_{n-1} \to b) = \frac{\mathbb{P}(X_1 = a, X_2 = s_2, \ldots, X_{n-1} = s_{n-1}, X_n = b)}{\mathbb{P}(X_1 = a)},$$

and so

$$\mathbb{P}\left(\bigcup_{s_2,\ldots,s_{n-1}\in S} a \to s_2 \to \cdots \to s_{n-1} \to b\right) = \frac{\mathbb{P}(X_1 = a, X_2 \in S, \ldots, X_{n-1} \in S, X_n = b)}{\mathbb{P}(X_1 = a)} = \frac{\mathbb{P}(X_1 = a, X_n = b)}{\mathbb{P}(X_1 = a)},$$

with the last equality following because $X_i \in S$ always. The right side of this equation is just $\mathbb{P}(X_n = b \mid X_1 = a)$, and so combining all of our equations proves the result. □

Now we're ready to tackle our theorem linking matrix multiplication and multi-step transition functions.

THM For a Markov chain (X_i) having states $\{1, \ldots d\}$ and transition function $P(x, y)$, let T be the corresponding transition matrix. Then, the transition matrix corresponding to the multi-step transition function $P^n(x, y)$ is T^n.

Proof. We will proceed by induction. By definition, T gives the transition probabilities for $P(x, y)$. Suppose that T^{n-1} gives the transition probabilities for $P^{n-1}(x, y)$. Since (X_i) is a Markov processes,

$$\mathbb{P}(s_1 \rightsquigarrow s_{n-1} \to s_n) = \mathbb{P}(s_1 \rightsquigarrow s_{n-1})\mathbb{P}(s_{n-1} \to s_n),$$

where $s_1, \ldots, s_n$ are states. Further, from our previous lemmas,

$$\mathbb{P}(X_n = s_n \mid X_1 = s_1) = \sum_{s\in S} \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1})\mathbb{P}(X_{n-1} = s_{n-1} \mid X_1 = s_1)$$

$$= \sum_{s\in S} P(s, s_n)P^{n-1}(s_1, s) = \sum_{s\in S} P^{n-1}(s_1, s)P(s, s_n).$$

However, the right hand side is the same as the dot product of the s_n column of T with the s_1 row of T^{n-1}, which is the definition of the (s_1, s_n) entry of $TT^{n-1} = T^n$. □

Now we know how to analyze transition probabilities for any finite-state Markov chain using matrices!

Let's see how our probability questions become linear algebra questions. Suppose we'd like to know the distribution after one joke supposing that your aunt started happy. That is, we'd like to know $f(x) = P(H, x)$. This is just the first row of T, and so the distribution is given by $\begin{bmatrix} 1 & 0 \end{bmatrix} T$. The distribution after 4 steps supposing your aunt started off angry is $\begin{bmatrix} 0 & 1 \end{bmatrix} T^7$. And $P^7(A, H)$ is $\begin{bmatrix} 0 & 1 \end{bmatrix} T^7 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

But we can do more. Suppose we wanted to know the distribution after one joke supposing that there was a 50/50 chance your aunt came into the club happy or angry. That's just $\begin{bmatrix} 1/2 & 1/2 \end{bmatrix} T$. In general, if $\vec{d}$ is a *row vector* specifying an initial distribution, then $\vec{d}T^n$ gives the resulting distribution after n steps.

Day 5

Consider now the following questions: After many many jokes, what is the probability that your aunt leaves the club happy? What's the probability she leaves angry? Does the state she comes into the club with affect these probabilities? How does the number of jokes told affect these probabilities?

We can answer all of these questions with linear algebra! Since we're going to be computing large powers of T, let's diagonalize it.

$$T = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix} = SDS^{-1}.$$

Since $(1/4)^n \to 0$ as $n \to \infty$, we can easily compute

$$T^\infty \equiv \lim_{n\to\infty} T^n = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix} = \begin{bmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{bmatrix}.$$

From this we can see $\begin{bmatrix} 1 & 0 \end{bmatrix} T^\infty = \begin{bmatrix} 0 & 1 \end{bmatrix} T^\infty$ and so her chances of being happy or angry upon leaving the club after many jokes doesn't depend on her initial state. In either case, she leaves happy with a 1/3 chance and angry with a 2/3 chance. We can also conclude that the number of jokes told doesn't really matter as long as the number is large. Indeed

$$T^n - T^m = S(D^n - D^m)S^{-1} = S \begin{bmatrix} \frac{1}{4^n} - \frac{1}{4^m} & 0 \\ 0 & 0 \end{bmatrix} S^{-1}.$$

As long as n and m are large, $\frac{1}{4^n} - \frac{1}{4^m} \approx 0$, and so the difference is approximately the zero matrix.

Day 6

Hitting Times

Consider a Markov chain for a gambler. The state space would be $\$d \geq 0$. If the gambler has any money, he can bet it and transition to a state with more money or a state with less money. But, if the gambler ever has \$0, he cannot bet any further.

The *Gambler's Ruin* problem asks, how long does it take to end up in state \$0?

Hitting Time

DEF

Let (X_i) be a Markov chain. A ***hitting time*** for an event B is the random variable

$$T_B = \min\{i \geq 0 : X_i \in B\}.$$

We will not compute hitting times right now, but hitting times give us another way to think about Markov chains. For example,

$$P^n(x,y) = \sum_{m\leq n} P(T_{\{y\}} = m \mid X_1 = x)P^{n-m}(y,y).$$

(Why is this formula intuitive?)

Different Types of States

Consider the gambler's ruin problem. If $a, b > 0$, there's a non-zero chance of starting at state a and ending at state b. (Is the correct way to mathematical quantify this statement $P(a,b) > 0$?) However, $P^n(0,b) = 0$ unless $b = 0$. That is, you cannot transition out of state \$0.

Absorbing

DEF

A state a of a Markov chain is called ***absorbing*** if $P(a,b) = 0$ whenever $b \neq a$.

In plain language, a state is absorbing if once you enter it, you can never leave.

For a Markov chain, define

$$\rho_{xy} = \mathbb{P}(T_{\{y\}} < \infty \mid X_1 = x).$$

That is, ρ_{xy} is the probability you end up at state y at some time in the future having started at state x.

DEF

Recurrent & Transient

For a Markov chain (X_i), a state a is called ***recurrent*** if $\rho_{aa} = 1$, otherwise it is called ***transient*** .

What are the transient and recurrent states for the gambler's ruin problem? Well, the state \$0 is absorbing, and if the odds are in the house's favor, $\rho_{x0} = 1$. That is, you will always eventually end up at \$0. Thus, the state \$0 is recurrent. However, if you start at state $\$d > 0$, there is a positive probability you end up at state \$0, and so $\rho_{dd} < 1$. This means that every state $\$d > 0$ is transient.

Now we get to our first big theorem.

THEOREM

Let (X_i) be a Markov chain, let $N(x)$ be the number of times state x is visited, and let $G(y,x) = \mathbb{E}(N(x) \mid X_0 = y)$. Then,

(i) if x is recurrent, $\mathbb{P}(N(x) = \infty \mid X_1 = x) = 1$ and $G(x,x) = \infty$, and

(i) if x is transient, $\mathbb{P}(N(x) < \infty \mid X_1 = y) = 1$ and

$$G(y,x) = \frac{\rho_{yx}}{1-\rho_{xx}}.$$

In plain words, the theorem says that recurrent states are visited infinitely often and transient states are visited finitely often. Also, we can precisely compute the expectation of the number of times a transient state is visited in terms of ρ.

Day 7

Proof. (i). Suppose x is a recurrent state. Let $T^{(n)}_{\{x\}} = \min\{i > n : X_i = x\}$ be the first return time to $\{x\}$ after time n. Since X_i is a stationary process,

$$\mathbb{P}(T_{\{x\}} < \infty \mid X_1 = x) = \mathbb{P}(T^{(n)}_{\{x\}} < \infty \mid X_n = x).$$

Since x is recurrent, $\mathbb{P}(T_{\{x\}} < \infty \mid X_1 = x) = 1$, and so $\mathbb{P}(T^{(n)}_{\{x\}} < \infty \mid X_n = x) = 1$. Thus if we are ever in state x we must again be in state x at some point in the future, and so $N(x) = \infty$. It immediately follows that $G(x,x) = \mathbb{E}(\infty) = \infty$.

(ii). Suppose x is transient. Let p_n be the probability that, starting at y, we hit x at least n times. Computing with the Markov property, we see

$$p_n = \rho_{yx}\rho_{xx}^{n-1} = \mathbb{P}(N(x) \geq n \mid X_1 = y).$$

We'd like to compute $\mathbb{P}(N(x) < \infty \mid X_1 = y)$. Applying an easy lower bound, we see

$$\begin{aligned}\mathbb{P}(N(x) < \infty \mid X_1 = y) &= \sum_{i<\infty} \mathbb{P}(N(x) = i \mid X_1 = y)\\ &\geq \sum_{i\leq n} \mathbb{P}(N(x) = i \mid X_1 = y)\\ &= 1 - \mathbb{P}(N(x) \geq n \mid X_1 = y) = 1 - \rho_{yx}\rho_{xx}^{n-1},\end{aligned}$$

and so $\mathbb{P}(N(x) < \infty \mid X_1 = y) = 1$.

Now, let $d(n) = \mathbb{P}(N(x) = n \mid X_1 = y)$ be the distribution of $N(x)$. We can explicitly write d as

$$\begin{aligned}d(n) &= \mathbb{P}(N(x) \geq n \mid X_1 = y) - \mathbb{P}(N(x) \geq n+1 \mid X_1 = y)\\ &= \rho_{yx}\rho_{xx}^{n-1} - \rho_{yx}\rho_{xx}^{n} = \rho_{yx}\rho_{xx}^{n-1}(1-\rho_{xx}),\end{aligned}$$

and so

$$G(y,x) = \mathbb{E}(N(x) \mid X_1 = y) = \sum nd(n) = \sum n\rho_{yx}\rho_{xx}^{n-1}(1-\rho_{xx}) = \frac{\rho_{yx}}{1-\rho_{xx}}.$$

$\square$

In light of the previous theorem, we can think of recurrent and transient states as synonymous with being visited a finite number of times.

To ease notation, since we will often talk about $\mathbb{P}(E \mid X_0 = x)$ for events E and states x, we will use the shorthand notation $\mathbb{P}_x(E)$.

Recurrent & Transient Chains

DEF

A Markov chain is called a ***recurrent chain*** if every state is recurrent and it is called a ***transient chain*** if every state is transient.

Note that, unlike states, there can be Markov chains that are neither recurrent nor transient.

Closed

DEF

Let (X_i) be a Markov chain with state space $\mathcal{S}$. A subset $\mathcal{A} \subseteq \mathcal{S}$ is called ***closed*** if $P(a, b) = 0$ whenever $a \in \mathcal{C}$ and $b \in \mathcal{A}^C$, where $\mathcal{A}^C$ is the complement of $\mathcal{C}$.

In other words, a set $\mathcal{A}$ is closed if $\mathcal{A}$ as a whole acts like an absorbing state.

LEM

Let $\mathcal{A}$ be a closed set for a Markov chain and suppose $a \in \mathcal{A}$ and $b \notin \mathcal{A}$. Then $\rho_{ab} = 0$.

Proof. Suppose $\rho_{ab} > 0$. Then there exists some sequence of states $a = x_0 \to x_1 \to \cdots \to x_{n-1} \to x_n = b$ such that $\mathbb{P}(x_0 \to x_1 \to \cdots \to x_{n-1} \to x_n) > 0$. For each i, $x_i \in \mathcal{A}$ or $x_i \notin \mathcal{A}$. Let k be the smallest index such that $x_k \notin \mathcal{A}$. Since $x_0 \in \mathcal{A}$, $k \geq 1$.

Now, since $\mathcal{A}$ is closed and $x_{k-1} \in \mathcal{A}$ and $x_k \notin \mathcal{A}$, we have that $P(x_{k-1}, x_k) = 0$. But this means, $\mathbb{P}(x_0 \to x_1 \to \cdots \to x_{n-1} \to x_n) = \prod_{i \geq 1}^{n} \mathbb{P}(x_{i-1} \to x_i) = 0$, which is a contradiction. □

Leads To

DEF

For a Markov chain, we say a state a ***leads to*** a state b if $\rho_{ab} > 0$, and we notate it $a \underset{\text{leads to}}{\rightsquigarrow} b$.

Irreducible

DEF

Let (X_i) be a Markov chain with state space $\mathcal{S}$. A subset $\mathcal{A} \subseteq \mathcal{S}$ is called ***irreducible*** if $\rho_{ab} > 0$ for any $a, b \in \mathcal{A}$.

Eventually we will learn how to decompose a Markov chain into closed and irreducible components.

THM

Let (X_i) be a Markov chain with state space $\mathcal{S}$. For $x, y \in \mathcal{S}$, suppose that x is recurrent and $x \underset{\text{leads to}}{\rightsquigarrow} y$. Then y is recurrent and $\rho_{xy} = \rho_{yx} = 1$.

Proof. We will first prove the result for the one-step process. Fix $x, y \in \mathcal{S}$ such that x is recurrent and $\mathbb{P}(x \to y) > 0$. Since x is recurrent, $\mathbb{P}_x(T_x = \infty) = 0$, so

$$0 = \mathbb{P}_x(T_x = \infty) \geq \mathbb{P}(x \to y)(1 - \rho_{yx}).$$

But, we assumed $\mathbb{P}(x \to y) > 0$, so we conclude $\rho_{yx} = 1$ (because $1 - \rho_{yx} = 0$).

Let $N^j(a) = \#\{X_i = a : i \leq j\}$. Now, since x is recurrent, for any $\varepsilon > 0$ and any k, we may find an n such that $\mathbb{P}_x(N^n(x) \geq k) > 1 - \varepsilon$. Let $C^n_w(a) = \{\text{chains of length } n \text{ with } w \text{ occurrences of } a\}$. Now, we have

$$\mathbb{P}_x(N^n(x) \geq k) = \sum_{w=k}^{n} \sum_{C \in C^n_w(x)} \mathbb{P}_x(C) > 1 - \varepsilon,$$

and so in particular

$$\sum_{w=0}^{k} \sum_{C \in C^n_w(x)} \mathbb{P}_x(C) \leq \varepsilon.$$

Consider a chain $C \in C^n_w(x) \cap C^n_0(y)$. Since the chain is Markov, we know $\mathbb{P}(\cdots \to x \to s \to \ldots) = \mathbb{P}(\cdots \to x)\mathbb{P}(x \to s)\mathbb{P}(s \to \ldots)$. Since there are no occurrences of y, $\mathbb{P}(x \to s) = 1 - P(x, y)$. Since there are w occurrences of x in C, $\mathbb{P}_x(C) \leq (1 - P(x, y))^k$.

We are now ready to compute

$$\mathbb{P}_x(N(y)=0) \le \mathbb{P}_x(N^n(y)=0) = \sum_{w=0}^{n} \sum_{C \in C_w^n(x) \cap C_0^n(y)} \mathbb{P}_x(C)$$

$$= \sum_{w=k+1}^{n} \sum_{C \in C_w^n(x) \cap C_0^n(y)} \mathbb{P}_x(C) + \sum_{w=0}^{k} \sum_{C \in C_w^n(x) \cap C_0^n(y)} \mathbb{P}_x(C)$$

$$\le (1-P(x,y))^k + \varepsilon.$$

Since $P(x,y) > 0$ and this holds for all ε and k, we conclude that, we conclude that $\mathbb{P}_x(N(y) \ge 1) = 1$ and so $\rho_{xy} = 1$.

Since

$$\rho_{yy} \ge \rho_{yx}\rho_{xy} = 1,$$

y is recurrent.

We have proven that if y is one step from x and $\rho_{xy} > 0$ and x is recurrent, then y is recurrent.

Now, suppose $x \underset{\text{leads to}}{\rightsquigarrow} y$ and x is recurrent. This means there exists some chain $x = s_0 \to s_1 \to \cdots \to s_{n-1} \to s_n = y$ such that $\mathbb{P}(s_0 \to \ldots \to s_n) > 0$. Suppose one of the s_i is not recurrent, and let k be the minimum index of such an s_i. Since $s_0 = x$ is recurrent, $k \ge 1$. We now have s_{k-1} is recurrent and $\mathbb{P}(s_{k-1} \to s_k) > 0$, so s_k must be recurrent, a contradition which completes the proof. □

Day 10

COR

Suppose $\mathcal{S}$ is the state space of a Markov chain and $\mathcal{B} \subseteq \mathcal{S}$ is a closed and irreducible set. Then, if $\mathcal{B}$ contains a recurrent state, $\rho_{xy} = 1$ for all $x, y \in \mathcal{B}$.

THM

If $\mathcal{B} \subseteq \mathcal{S}$ is a finite, closed, and irreducible subset of the state space $\mathcal{S}$ for a Markov chain, then every state in $\mathcal{B}$ is recurrent.

Proof. Fix $x \in \mathcal{B}$. By the pidgeon hole principle, there must be some state $y \in \mathcal{B}$ so that $\mathbb{E}(N(y) \mid X_0 = x) = \infty$. By our first theorem, this means that y is recurrent. Since $\mathcal{B}$ is irreducible, $\rho_{yt} > 0$ for any $t \in \mathcal{B}$, and so by our second theorem, $\rho_{yt} = 1$. □

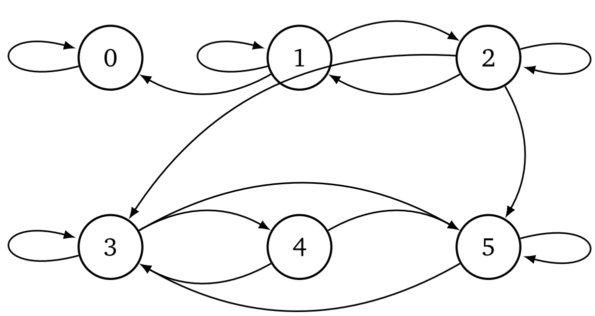
We've stated a lot of theorems. Let's look at a simple example.

Suppose the transition matrix for a Markov chain is

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/5 & 2/5 & 1/5 & 0 & 1/5 \\ 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/4 & 0 & 3/4 \end{bmatrix} \end{array}.$$

Can we identify the closed, irreducible components? Can we identify the recurrent states?

Well, the transition graph for this chain looks like



Subsets of the graph with only incoming edges are closed. Subsets of the graph where there is a path going through every element are irreducible. We therefore see the closed, irreducible subsets of the state space are $C_0 = \{0\}$ and $C_1 = \{3,4,5\}$. Since these are finite, they must consist only of recurrent points. We can also see that states 1 and 2 are not recurrent since there is a positive probability of transitioning into C_0 or C_1.

Day 11

THM Let $\mathcal{S}_R \subseteq \mathcal{S}$ be the set of all recurrent states for a Markov chain with state space $\mathcal{S}$. Then $\mathcal{S}_R = \bigcup C_i$ where $\{C_i\}$ are disjoint, closed, and irreducible.

Proof. Suppose C is a closed set of recurrent points. Then $\underset{\text{leads to}}{\rightsquigarrow}$ is an equivalence relation on C.

It is clear that since for any two states x, y, either $\rho_{xy} > 0$ or $\rho_{xy} = 0$, $\underset{\text{leads to}}{\rightsquigarrow}$ is a relation on the set of states. We will now show it satisfies the properties of an equivalence relation.

(reflexive) If $x \in C$, then $x \underset{\text{leads to}}{\rightsquigarrow} x$ because x is recurrent.

(symmetric) If $x, y \in C$ and $x \underset{\text{leads to}}{\rightsquigarrow} y$, then $y \underset{\text{leads to}}{\rightsquigarrow} x$ by our second theorem.

(transitive) If $x, y, z \in C$ and $x \underset{\text{leads to}}{\rightsquigarrow} y$ and $y \underset{\text{leads to}}{\rightsquigarrow} z$, then $\rho_{xz} \geq \rho_{xy}\rho_{yz} > 0$, so $x \underset{\text{leads to}}{\rightsquigarrow} z$.

Now, notice that $\mathcal{S}_R$ is closed. Suppose $y \notin \mathcal{S}_R$. Then,

$$1 = \rho_{xx} \leq 1 - \rho_{xy},$$

and so $\rho_{xy} = 0$ which is equivalent to $\mathcal{S}_R$ beging closed. Now, let $\{C_i\}$ be the partition coming from the equivalence relation $\underset{\text{leads to}}{\rightsquigarrow}$. $\square$

DEF For a set of states A, let $\rho_{xA} = \mathbb{P}_x(T_A < \infty)$ be the probability of transitioning from the state x to any state in A in finite time.

If A is a closed set of states, we can think of ρ_{xA} as the probability that starting at x, you will be absorbed by the set A. In light of the previous theorem, we now have a strategy for studying Markov chains.

(i) Find the closed, irreducible subsets of the state space $\{C_i\}$.

(ii) For a transient state x, compute ρ_{xC_i}.

(iii) Study irreducible Markov chains.

Day 12

For a moment, let's consider (ii). Suppose $\{C_i\}$ is the collection of closed, irreducible subsets of the state space. Then, if $x \in C_i$,

$$\rho_{xC_i} = 1 \qquad \text{and} \qquad \rho_{xC_j} = 0 \text{ for } i \neq j.$$

If x is transient, things are a little harder. But, we do know if x transitions to C_i, it does so in one step or more than one step. Thus,

$$\rho_{xC_i} = \sum_{y \in C_i} P(x,y) + \sum_{y \notin C_i} P(x,y)\rho_{yC_i} = \sum_{y \in C_i} P(x,y) + \sum_{y \in \mathcal{S}_T} P(x,y)\rho_{yC_i}, \tag{1}$$

where $\mathcal{S}_T$ is the set of all transient points.

If we squint at equation (1), we see that taking the collection of $\{\rho_{zC_i}\}$ as unknowns, equation (1) gives us a system of linear equations for $\{\rho_{zC_i}\}$. Thus, finding ρ_{xC_i} should just be a little linear algebra (if the system actually has a solution).

Suppose now that there are only a finite number of transient states $\{t_1, \ldots, t_n\}$. Then, we can rewrite the system corresponding to (1) as a matrix equation of the form

$$\vec{\rho}_{C_i} = P_T \vec{\rho}_{C_i} + \vec{p}_{C_i}, \tag{2}$$

where

$$\vec{\rho}_{C_i} = \begin{bmatrix} \rho_{t_1 C_i} \\ \rho_{t_2 C_i} \\ \vdots \end{bmatrix} \qquad \vec{p}_{C_i} = \begin{bmatrix} P(t_1, C_i) \\ P(t_2, C_i) \\ \vdots \end{bmatrix}$$

and P_T is the matrix with j,k entry $P(t_j, t_k)$.

Equation (2) has a solution if and only if $(I - P_T)$ is invertible. As it turns out, it always is.

Day 13

THM

If a Markov chain has a finite number of transient states and P_T is the transition matrix between transient states, then $(I - P_T)$ is invertible.

Proof. Showing that $(I - P_T)$ is invertible is equivalent to showing that 1 is not an eigenvalue of P_T.

Suppose that 1 were an eigenvalue for P_T and $\vec{v}$ a corresponding eigenvector. Then

$$\lim_{n\to\infty} P_T^n \vec{v} = \vec{v}.$$

However, the i,j entry of P_T^n is the transition probability $P^n(t_i, t_j)$. Since t_j is transient, it is only hit a finite number of times and so

$$\lim_{n\to\infty} P^n(i,j) = 0.$$

Thus we have that

$$\lim_{n\to\infty} P_T^n \vec{v} = 0\vec{v} = \vec{0},$$

which contradicts $\vec{v}$ being an eigenvector. □

A Gambling Problem

You're gambling. Every bet of one dollar that you make, you have a p chance of winning (and gaining a dollar) and a $1-p$ chance of losing your dollar. You will stop if either: (a) you won \$25 or (b) you lost \$10.

What is the probability that you will win vs. lose?

This question could be reformulated as: How do we compute the quantity

$$\mathbb{P}_x(T_{x-10} < T_{x+25})?$$

Instead of tackling this question directly, we'll analyze a more general process.

Day 14

Birth and Death Chains

A birth and death chain is roughly inspired by the idea that whatever is alive has some chance of breeding. We model this as a Markov chain with state space $\mathbb{N}$ and transition probabilities

$$P(x,y) = \begin{cases} p_x & \text{if } y = x+1 \\ r_x & \text{if } y = x \\ q_x & \text{if } y = x-1 \\ 0 & \text{otherwise} \end{cases},$$

where $p_x + r_x + q_x = 1$ and may depend on x. Further, to keep this chain in the sate space, $q_0 = 0$.

Now, for a fixed $a < b$, we can consider

$$u_{ab}(x) = \mathbb{P}_x(T_a < T_b).$$

We trivially define $u_{ab}(x) = 1$ if $x \le a$ and $u_{ab}(x) = 0$ if $x \ge b$.

Our goal is to compute $u_{ab}(x)$. Since we have the one-step transition probabilities, we may expand

$$u_{ab}(x) = p_x u_{ab}(x+1) + r_x u_{ab}(x) + q_x u_{ab}(x-1),$$

which is valid *only when* $x \neq a$ and $x \neq b$. Now, since $p_x + r_x + q_x = 1$, we have $u_{ab}(x) = (p_x + r_x + q_x)u_{ab}(x)$, and so after some rearranging of terms we have

$$u_{ab}(x+1) - u_{ab}(x) = \frac{q_x}{p_x}\Big(u_{ab}(x) - u_{ab}(x-1)\Big),$$

which again is only valid if $x \neq a$ and $x \neq b$. But, this is a recursive formula! So, if $x > a$, we may substitute down to $x = a+1$ to get

$$u_{ab}(x+1) - u_{ab}(x) = \frac{q_x q_{x-1} \cdots q_{a+1}}{p_x p_{x-1} \cdots p_{a+1}}\Big(u_{ab}(a+1) - u_{ab}(a)\Big).$$

Let

$$\gamma_x = \frac{q_x q_{x-1} \cdots q_1}{p_x p_{x-1} \cdots p_1}.$$

We may now express $u_{ab}(x+1) - u_{ab}(x)$ as

$$u_{ab}(x+1) - u_{ab}(x) = \frac{\gamma_x}{\gamma_a}\Big(u_{ab}(a+1) - u_{ab}(a)\Big).$$

Notice that this formula, which was only valid for $x > a$ and $x \neq b$ is now trivially valid for $x = a$, and so it holds for all $a \leq x < b$. And, for any such x, doing a telescoping sum, we see

$$u_{ab}(x+1) - u_{ab}(a) = \sum_{y=a}^{x}\Big(u_{ab}(y+1) - u_{ab}(y)\Big) = \sum_{y=a}^{x}\frac{\gamma_y}{\gamma_a}\Big(u_{ab}(a+1) - u_{ab}(a)\Big).$$

Since $u_{ab}(a) = 1$, if we could somehow compute $\frac{u_{ab}(a+1)-u_{ab}(a)}{\gamma_a}$, we could compute $u_{ab}(x)$ for any x!

Notice the previous formula is valid for all $a \leq x < b$ and so letting $x = b-1$ we see,

$$\sum_{y=a}^{b-1}\frac{\gamma_y}{\gamma_a}\Big(u_{ab}(a+1) - u_{ab}(a)\Big) = u_{ab}(b) - u_{ab}(a) = 0 - 1 = -1.$$

Day 15

Thus,

$$\frac{u_{ab}(a+1) - u_{ab}(a)}{\gamma_a} = \frac{-1}{\sum_{y=a}^{b-1}\gamma_y}.$$

Putting everything together, we finally have a formula,

$$u_{ab}(x) - u_{ab}(a) = \mathbb{P}_x(T_a < T_b) - 1 = -\frac{\sum_{y=a}^{x-1}\gamma_y}{\sum_{y=a}^{b-1}\gamma_y},$$

and so

$$u_{ab}(x) = 1 - \frac{\sum_{y=a}^{x-1}\gamma_y}{\sum_{y=a}^{b-1}\gamma_y}.$$

We may now answer our gamblers problem. The gamblers problem is equivalent to a birth and death process with $p_x = p$, $r_x = 0$, and $q_x = 1-p$. We may then ask, what is

$$\mathbb{P}_{10}(T_0 < T_{35}).$$

Here $\gamma_y = (1-p)^y/p^y$. Letting $\alpha = (1-p)/p$,

$$\sum_{y=0}^{k}\gamma_y = \frac{\alpha^{k+1}-1}{\alpha - 1}.$$

This means,

$$\mathbb{P}_{10}(T_0 < T_{35}) = 1 - \frac{\frac{\alpha^{10}-1}{\alpha-1}}{\frac{\alpha^{35}-1}{\alpha-1}} = 1 - \frac{\alpha^{10}-1}{\alpha^{35}-1}.$$

Day 16

Branching Chains and Ninjas

Cultural preservation groups have been closely watching ninja dojos in the mountains of Japan. They've noticed that every dojo trains 3 warriors and each warrior has a 1/2 chance of starting a new dojo and a 1/2 chance of moving to the city and giving up the ninja life.

Supposing it all started with one dojo, what are the chances the ninja tradition eventually goes extinct?

Let X_i be the number of dojos at the ith stage. The Markov chain (X_i) is called a *branching chain* because every dojo has the possibility to spawn new dojos (to branch). There's also a possibility a dojo will never produce new dojos, in which case the dojo's linear will go extinct.

To answer the extinction question, we'd like to compute ρ_{10} for this Markov chain. (0 is an absorbing state for branching processes). Since dojos act independently, we see

$$\rho_{y0} = \rho_{10}^{y}.$$

That is, the probability of y dojos going extinct is the y-fold product of one dojo going extinct.

Directly computing, we get

$$\rho_{10} = P(1,0) + \sum_{y>0} P(1,y)\rho_{y0}.$$

Let $f(y) = P(1,y)$ be the probability mass function for the one-step process starting in state 1. Now, using the convention that $a^0 = 1$ for all $a \in \mathbb{R}$, we may simply write

$$\rho_{10} = \sum_{y\geq 0} P(1,y)\rho_{10}^{y} = \sum_{y\geq 0} f(y)\rho_{10}^{y} = \Phi(\rho_{10}), \tag{3}$$

where Φ is the probability generating function for the one-step process. What equation (3) says is that ρ_{10}, whatever it is, must be a fixed point of Φ.

Let's compute some properties of Φ. First, note, $\Phi(1) = 1$, so Φ always has at least one fixed point. Now,

$$\Phi'(t) = \sum_{y\geq 1} y f(y) t^{y-1} \geq 0,$$

so Φ is non-decreasing. In fact, if $\mu = \sum_{y\geq 1} y f(y)$ is the expectation of the one-step process and $\mu > 0$, then $\Phi'(t) > 0$ for $t > 0$. In particular, this means if $a < b$, then $\Phi(a) < \Phi(b)$.

Let's now unravel ρ_{10}. From the definition of ρ_{10},

$$\rho_{10} = \lim_{n\to\infty} \mathbb{P}_1(T_0 \leq n).$$

Now,

$$\mathbb{P}_1(T_0 \leq n+1) = P(1,0) + \sum_{y>0} P(1,y)\mathbb{P}_y(T_0 \leq n).$$

But, again, $\mathbb{P}_y(T_0 \leq n) = \mathbb{P}_1(T_0 \leq n)^y$ since every dojo independently trains people. Thus

$$\mathbb{P}_1(T_0 \leq n+1) = \sum_{y\geq 0} P(1,y)\mathbb{P}_1(T_0 \leq n)^y = \Phi(\mathbb{P}_1(T_0 \leq n)).$$

Recursively applying this property,

$$\mathbb{P}_1(T_0 \leq k) = \Phi^{k+1}(\mathbb{P}_1(T_0 \leq 0)) = \Phi^{k+1}(0),$$

since $\mathbb{P}_1(T_0 \leq 0) = 0$.

Suppose that $f_0 \in [0,1]$ is the smallest fixed point of Φ. That is, $\Phi(f_0) = f_0$ and $\Phi(x) > x$ for all $x \in [0,1]$ satisfying $x < f_0$. Since Φ is non-decreasing, we trivially have the following string of inequalities,

$$0 \le f_0$$

$$0 \le \Phi(0) \le \Phi(f_0) = f_0$$
$$0 \le \Phi^2(0) \le \Phi^2(f_0) = f_0$$
$$\vdots$$

Thus, $\lim_{n\to\infty} \Phi^n(0) = \alpha \le f_0$. But $\alpha \ge 0$ is a fixed point and so $\alpha = f_0$. Thus,

$$\lim_{n\to\infty} \mathbb{P}_1(T_0 \le n) = \rho_{10} = f_0.$$

What remains is to find the smallest fixed point of Φ.

THEOREM

Let f be the probability mass function for a random variable X with sate space $\mathbb{N}$, let μ be the mean of X, and let

$$\Phi(t) = \sum_{y\ge 0} f(y)t^y.$$

Then, if $0 < \mu \le 1$ and $f(1) < 1$, then $\Phi(t) = t$ and $t \in [0,1]$ implies $t = 1$.

If $\mu > 1$, then $\Phi(t) = t$ where $t \in [0,1]$ has exactly two solutions, $t = 1$ and $t = t_0 \in [0,1)$.

Proof. Notice that

$$\mu = \sum_{y\ge 1} y f(y)$$

and

$$\Phi'(t) = \sum_{y\ge 1} y f(y) t^{y-1}.$$

In particular, $\Phi'(1) = \mu$ and $\Phi'(t) \le \mu$ for $t \in [0,1)$ with a strict inequality if $f(0), f(1) < 1$.

Suppose $0 < \mu \le 1$ and $f(1) < 1$, and let $h(t) = \Phi(t) - t$. It is clear that $h(1) = 0$, and our goal is to show this is the only zero. By our computations above,

$$h'(t) = \Phi'(t) - 1 \le \mu - 1 \le 0$$

and so h is non-increasing on $[0,1]$. Since $\mu > 0$, we deduce $f(0) < 1$. By assumption we have $f(1) < 1$, and so in fact

$$h'(t) < 0.$$

for $t \in [0,1]$. A strictly decreasing function on any interval can have at most one root, and so $h(1) = 1$ must be the only root.

Now suppose $\mu > 1$ and notice that in this case we must have $f(0) + f(1) < 1$ since

$$\mu = \sum_{y\ge 1} y f(y) = 0f(0) + 1f(1) + \sum_{y\ge 2} y f(y) > 1$$

implies $f(y) > 0$ for at least one $y \ge 2$.

Now, as before, $h(1) = 0$ is a root, and

$$h'(1) = \Phi'(1) - 1 = \mu - 1 > 0.$$

Further, $h(0) = \Phi(0) = f(0) \ge 0$. If h starts above zero and ends at zero with a positive slope, at some point $t \in [0,1)$, $h(t) < 0$ (you can formally prove this with the mean value theorem). Therefore, by the intermediate value theorem, there is some $t_0 \in [0,1)$ so that $h(t_0) = 0$.

Finally we will show that t_0 is the unique zero in $[0,1)$. Consider

$$\Phi''(t) = \sum_{y\ge 2} y(y-1) f(y) t^{y-2} > 0$$

for $t > 0$ since $f(y) > 0$ for at least one $y \geq 2$. Therefore, Φ does not change concavity on the interval $[0,1]$ and so Φ has at most two roots on $[0,1]$. In particular, those roots must be t_0 and 1. $\square$

Day 18

Now we can finally answer the dojo extinction question. For the one-step process, since each of the three trained ninjas have a $1/2$ chance of starting a new dojo, the distribution of dojos will be

$$f(0) = 1/8 \qquad f(1) = 3/8 \qquad f(2) = 3/8 \qquad f(3) = 1/8.$$

Thus,

$$\Phi(t) = \tfrac{1}{8} + \tfrac{3}{8}t + \tfrac{3}{8}t^2 + \tfrac{1}{8}t^3.$$

Letting $h(t) = \Phi(t) - t$ and noting that $h(t) = 0$ if and only if $8h(t) = 0$, we see

$$8h(t) = 1 + 3t + 3t^2 + t^3 - 8t = (t-1)(t^2 + 4t - 1) = 0$$

at $t = 1$ and $t = -2 \pm \sqrt{5}$. We already know that ρ_{10} must be the smallest zero of h in $[0,1]$, so

$$\rho_{10} = -2 + \sqrt{5} \approx 0.236.$$

Stationary Distributions

If (X_i) is a sequence of independent and identically distributed random variables with distribution f and and $\mu = \text{mean}(f)$, the law of large numbers tells us that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i<k} X_i = \mu.$$

In fact, we can deduce that

$$\lim_{k\to\infty} \frac{1}{k} \#\{i < k : X_i = y\} = f(y).$$

In other words, we can recover the distribution f from the *statistics* of a sequence of i.i.d. "copies" of a random variable (in this case the "copies" are the X_i).

Markov chains are not independent (most of the time), but we could still collect statistics. For a Markov chain (X_i), let

$$N_k(y) = \#\{i < k : X_i = y\}$$

be the number of times the chain hits y in at most k steps.

Now, we could define a function

$$f(y) := \lim_{k\to\infty} \frac{1}{k} N_k(y)$$

whenever the limit exists. Surprisingly, this limit exists for Markov chains, but this isn't obvious and we'll just accept this for now. The question we're interested in is whether f depends on X_0.

Day 19

We can come up with examples that go both ways.

1. Consider a Markov chain with transition matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Then

   $$f(y) = \begin{cases} 1 & \text{if } y = X_0 \\ 0 & \text{else} \end{cases}.$$

   Since this Markov chain remains in its initial state, f clearly depends on the initial state.

2. Consider a Markov chain with transition matrix $\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$. Then this Markov chain is actually independent and

   $$f(y) = 1/2$$

   regardless of y.

3. Consider a Markov chain with transition matrix $M = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$. In this case we can diagonalize

$$M = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}$$

and use this to compute

$$\lim_{k\to\infty} M^k = \begin{bmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{bmatrix}.$$

It takes some careful analysis to unpack the statement about $\lim_{k\to\infty} M^k$ and turn it into a statement about the statistics of our Markov chain, but, it turns out the straightforward thing happens. That is,

$$f(y) = \begin{cases} 1/3 & \text{if } y = 0 \\ 2/3 & \text{if } y = 1 \end{cases}.$$

Stationary Distribution

DEFINITION

Let (X_i) be a Markov chain with transition function P and state space $\mathcal{S}$. The distribution π is called a ***stationary distribution*** if

$$\pi(y) = \sum_{x\in\mathcal{S}} \pi(x)P(x, y).$$

Day 20

A stationary distribution is a distribution that equals its own *push forward* distribution. That is, $\mathrm{dist}(X_i \mid X_{i-1} \sim \pi) = \pi$ where $X_{i-1} \sim \pi$ means that X_{i-1} has probability distribution π.

If $\mathcal{S}$ is finite, we can phrase this in terms of the transition matrix. Let T be the transition matrix for a Markov chain. Then if $\vec{\pi}$ is a distribution represented as a probability vector, $\vec{\pi}$ would be stationary if and only if

$$\vec{\pi} = \vec{\pi}T.$$

In other words, $\vec{\pi}$ would be a left eigenvector of T with eigenvalue 1. Since T always has an eigenvalue of 1, we know there is always at least one stationary distribution for any Markov chain. If the eigenvector with eigenvalue 1 is unique (up to scalar multiples), we know that this stationary distribution cannot depend on the distribution of X_0. Since we'll see that the statics of a Markov chain will always converge to an invariant distribution, this means if there is only one invariant distribution, the statistics won't depend on the initial state of the Markov chain.

Stationary distributions are closely related to the statistically-derived f we've been looking at, and we will come to see that if f is a distribution and is independent of X_0, it will coincide with a stationary distribution.

For now, let's compute a simple example of a stationary distribution. Consider the Markov chain with transition matrix $M = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$. If $\vec{\pi}$ is a left eigenvector of M, then $\vec{\pi}M = \vec{\pi}$. Suppose $\vec{\pi}$ is a probability vector. We may now compute

$$\lim_{n\to\infty} \vec{\pi}M^n = \vec{\pi}\left(\lim_{n\to\infty} M^n\right) = \vec{\pi}\begin{bmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{bmatrix} = \begin{bmatrix} 1/3 & 2/3 \end{bmatrix},$$

so if $\vec{\pi}$ is an eigenvector with eigenvalue 1 for M, it must be the vector $\begin{bmatrix} 1/3 & 2/3 \end{bmatrix}$. Verifying, this is indeed an eigenvector, and so we found it by luck. We of course can systematically find the eigenvectors of any finite matrix.

If the Markov chain has an infinite state space, things can get a little more tricky.

Stationary Measure of a Birth and Death Chain

Suppose (X_i) is a birth and death chain with transition function

$$P(x, y) = \begin{cases} p_x > 0 & \text{if } y = x + 1 \\ r_x & \text{if } y = x \\ q_x > 0 & \text{if } y = x - 1 \end{cases}$$

and $q_0 = 0$. Suppose π is a stationary distribution for (X_i). This means

$$\pi(y) = \sum_{x\in\mathbb{N}} \pi(x)P(x,y),$$

and so in particular

$$\pi(0) = r_0\pi(0) + q_1\pi(1)$$

and

$$\pi(y) = p_{y-1}\pi(y-1) + r_y\pi(y) + q_{y+1}\pi(y+1)$$

if $y > 0$.

From the first equation, since $r_0 = 1 - p_0$ (because $q_0 = 0$) we see

$$q_1\pi(1) - p_0\pi(0) = 0.$$

Now, in the second equation, $r_y = 1 - p_y - q_y$ and so, subtracting $\pi(y)$ from both sides,

$$0 = p_{y-1}\pi(y-1) - q_y\pi(y) + q_{y+1}\pi(y+1) - p_y\pi(y).$$

Applying this to the case where $y = 1$, we see

$$0 = q_2\pi(2) - p_1\pi(1).$$

Recursively applying this substitution gives

$$0 = q_{y+1}\pi(y+1) - p_y\pi(y)$$

for any $y \geq 0$. And so we're left with a beautiful recursive formula

$$\pi(y+1) = \frac{p_y}{q_{y+1}}\pi(y) = \frac{p_y p_{y-1}}{q_{y+1}q_y}\pi(y-1) = \cdots = \frac{p_y p_{y-1}\cdots p_0}{q_{y+1}q_y\cdots q_1}\pi(0).$$

Since we assumed π to be a distribution,

$$1 = \sum_y \pi(y) = \pi(0)\left(1 + \sum_{y\geq 1}\frac{p_y p_{y-1}\cdots p_0}{q_{y+1}q_y\cdots q_1}\right),$$

and so

$$\pi(0) = \frac{1}{1 + \sum_{y\geq 1}\frac{p_y p_{y-1}\cdots p_0}{q_{y+1}q_y\cdots q_1}},$$

from which we can derive $\pi(y)$ for any $y > 0$.

We've shown that assuming π is a stationary distribution, it is determined uniquely by the p_x's and q_x's. However, if $\sum_{y\geq 1}\frac{p_y p_{y-1}\cdots p_0}{q_{y+1}q_y\cdots q_1} = \infty$, then $\pi(0) = 0 = \pi(y)$ for all y, which is clearly not a distribution. Therefore, we conclude that a birth and death chain has a stationary distribution if and only if

$$\sum_{y\geq 1}\frac{p_y p_{y-1}\cdots p_0}{q_{y+1}q_y\cdots q_1} < \infty.$$

Day 21

Mean Waiting Time

One of our goals will to be to show that if $f(y) := \lim_{k\to\infty}\frac{1}{k}N_k(y)$ is a distribution, then it is a stationary distribution. And, though it may seem intuitive, actually showing this will take some machinery. Here, we'll establish that machinery, piggybacking on the *strong law of large numbers*.

Strong Law of Large Numbers

THEOREM

Let X_i be an independent and identically distributed sequence of random variables and let $\mu \in \mathbb{R}\cup\{\pm\infty\}$. Then

$$\lim_{n\to\infty}\frac{X_1 + \cdots + X_n}{n} = \mu$$

with probability 1.

The strong law of large numbers is fairly technical to prove, so we will just use it as a black box for now.

Indicator Function

DEF

For an event X, the ***indicator function*** of X is

$$\mathbf{1}_X(\omega) = \begin{cases} 1 & \text{if } \omega \in X \\ 0 & \text{else} \end{cases}.$$

Indicator functions can be used to create random variables whose expectations are probabilities. That is, if X is an event,

$$\mathbb{P}(X) = \mathbb{E}(\mathbf{1}_X).$$

Mean Waiting Time

DEF

For a Markov chain (X_i), the ***mean waiting time*** for the state y is

$$m_y = \mathbb{E}(T_y) = \mathbb{E}(\min\{i > 0 : X_i = y \mid X_0 = y\}).$$

Recall that

$$N_k(y) = \#\{i < k : X_i = y\},$$

and define

$$G_k(x, y) = \mathbb{E}(N_k(y) \mid X_0 = x).$$

Finally, let $\{T_y < \infty\}$ be the event that y occurs. With this last bit of notation we can formally state a seemingly-intuitive relationship between $N_k(y)$ and m_y.

THEOREM

Let (X_i) be a Markov chain and y be a recurrent state. Then

$$\lim_{k\to\infty} \frac{N_k(y)}{n} = \frac{\mathbf{1}_{\{T_y<\infty\}}}{m_y}$$

with probability 1 and

$$\lim_{k\to\infty} \frac{G_k(x,y)}{n} = \frac{\rho_{xy}}{m_y}.$$

We need the $\mathbf{1}_{\{T_y<\infty\}}$ to cover the cases where state y is never reached, but if we condition on starting in state y, this theorem would say

$$\lim_{k\to\infty} \frac{(N_k(y) \mid X_0 = y)}{n} = \frac{1}{m_y}.$$

In other words,

$$\lim_{k\to\infty} \frac{n}{(N_k(y) \mid X_0 = y)} = m_y,$$

and so if we take a large number of steps and divide by the number of times we encounter y, we get the expected amount of time we need to wait for y. This is sounding a lot like the law of large numbers!

Proof. Notice that we can define

$$\mathbb{N}_k(y) = \sum_{i<k} \mathbf{1}_{\{X_i=y\}}.$$

If the $\mathbf{1}_{\{X_i=y\}}$ were independent, we could apply the law of large numbers to $\frac{1}{k}\sum_{i<k} \mathbf{1}_{\{X_i=y\}}$ and we'd be set. Unfortunately, we're not so lucky. But, we'll search for something that is independent.

Let

$$T_y^r = \min\{i > 0 : N_i(y) = r \mid X_0 = y\}$$

be the time until y is hit r times, and define

$$W_y^i = T_y^i - T_y^{i-1}$$

to be the time between the $(i-1)$st and ith occurrences of y. By the Markov property,

$$\mathbb{E}(W_y^i) = \mathbb{E}(W_y^j) = \mathbb{E}(T_y) = m_y$$

and the W_y^i are independent and identically distributed. Thus, by the strong law of large numbers,

$$\lim_{k\to\infty} \frac{W_y^1 + \cdots W_y^k}{k} = m_y$$

with probability 1. Now, $W_y^1 + \cdots W_y^k = T_y^k$ and

$$(N_{T_y^k}(y) \mid X_0 = y) = k.$$

So, we may write the previous limit as

$$\lim_{k\to\infty} \frac{W_y^1 + \cdots W_y^k}{k} = \lim_{k\to\infty} \frac{T_y^k}{(N_{T_y^k}(y) \mid X_0 = y)} = m_y.$$

Now, for r such that $T_y^k \le r < T_y^{k+1}$, by definition we have

$$\frac{T_y^k}{(N_{T_y^k}(y) \mid X_0 = y)} \le \frac{r}{(N_r(y) \mid X_0 = y)} < \frac{T_y^{k+1}}{(N_{T_y^{k+1}}(y) \mid X_0 = y) - 1}.$$

Since T_y^k is monotone increasing in k and $T_y^k \to \infty$ as $k \to \infty$ and because both the left and right sides of the above inequality converge to m_y, we deduce

$$\lim_{r\to\infty} \frac{r}{(N_r(y) \mid X_0 = y)} = m_y.$$

Inverting we have

$$\lim_{k\to\infty} \frac{(N_k(y) \mid X_0 = y)}{k} = \frac{1}{m_y}.$$

Now, notice that

$$(N_k(y) \mid X_0 = y) - n \le (N_k(y) \mid X_0 = y \text{ or } X_1 = y \text{ or } \ldots \text{ or } X_{n-1} = y) \le (N_k(y) \mid X_0 = y).$$

For any fixed n, $n/k \to 0$ as $k \to \infty$, and so

$$\lim_{k\to\infty} \frac{(N_k(y) \mid X_0 = y \text{ or } X_1 = y \text{ or } \ldots \text{ or } X_{n-1} = y)}{k} = \frac{1}{m_y},$$

but the event $\{T_y < \infty\}$ precisely corresponds to the existence of an $n < \infty$ so that $X_n = y$. And, if $T_y = \infty$, $N_k(y) = 0$ for all k. Thus, we conclude

$$\lim_{k\to\infty} \frac{N_k(y)}{k} = \frac{\mathbf{1}_{\{T_y<\infty\}}}{m_y}$$

with probability 1. Now, since $\mathbb{E}(\mathbf{1}_{\{T_y<\infty\}} \mid X_0 = x) = \rho_{xy}$, taking expectations of both sides (and using the bounded convergence theorem),

$$\mathbb{E}_x\left(\lim_{k\to\infty} \frac{N_k(y)}{k}\right) = \lim_{k\to\infty} \frac{\mathbb{E}_x(N_k(y))}{k} = \lim_{k\to\infty} \frac{G_k(x,y)}{k} = \frac{\mathbb{E}_x(\mathbf{1}_{\{T_y<\infty\}})}{m_y} = \frac{\rho_{xy}}{m_y}.$$

□

To complement this theorem, if y is transient, we know that with probability 1, $\lim_{k\to\infty} N_k(y) < \infty$, and so

$$\lim_{k\to\infty} \frac{N_k(y)}{k} = 0.$$

Positive and Null Recurrent

Not all recurrent states are created equally. Some, although you return with probability one, it may take an exceedingly long time to return. In fact, the mean return time to a recurrent state may be infinite!

DEF

Positive & Null Recurrent

A recurrent state y is called ***positive recurrent*** if $m_y < \infty$. Otherwise y is called ***null recurrent*** .

Some examples.

1. Consider a two state Markov chain whose transition matrix is $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Here $P(0,1) = 1$ and $P(1,0) = 1$ and so both states are recurrent with $m_0 = m_1 = 2$. Therefore, both states are positive recurrent.

2. Let R_i be a ring of length i with starting state r_i and ending state e. That is, from r_i, you transition to the next state with probability 1 and to the next, next state with probability 1 until you reach state e.

   Now, consider the Markov chain with state space $\bigcup R_{2^i}$ and transition probabilities $P(e, r_{2^i}) = 1/2^i$. Since every ring leads back to e, e is clearly a recurrent state. However, if you ever travel to ring R_t, you must wait t steps before returning to state e. Therefore,

   $$m_e = \mathbb{E}(T_e) = \sum 2^i \mathbb{P}(e, r_{2^i}) = \sum 1 = \infty,$$

   so the mean return time to e is infinite.

We now have an analog of a previous theorem.

THM

If y is positive recurrent and $y \underset{\text{leads to}}{\rightsquigarrow} x$, then x is positive recurrent.

Proof. We're going to leverage the previous theorem to make this theorem easier. As such, we need to better understand $G_n(x,y)$.

Rewriting the definition of $G_n(x,y)$ with indicator functions gives us some insight.

$$G_n(x,y) = \mathbb{E}_x(N_n(y)) = \mathbb{E}_x\left(\sum_{i<n} \mathbf{1}_{X_i=y}\right) = \sum_{i<n} \mathbb{E}_x(\mathbf{1}_{X_i=y})$$
$$= \sum_{i<n} \mathbb{P}_x(X_i = y) = \sum_{i<n} P^i(x,y).$$

Now, since $y \underset{\text{leads to}}{\rightsquigarrow} x$ implies $x \underset{\text{leads to}}{\rightsquigarrow} y$, there are $a, b > 0$ so that $P^a(x,y), P^b(y,x) > 0$. This means that for $k > a + b$,

$$P^k(x,x) \geq P^a(x,y)P^{k-a-b}(y,y)P^b(y,x).$$

Using the alternate formulation for G_n we see

$$G_n(x,x) = \sum_{i<n} P^n(x,x) \geq \sum_{a+b\leq i<n} P^n(x,x)$$
$$\geq \sum_{a+b\leq i<n} P^a(x,y)P^{k-a-b}(y,y)P^b(y,x)$$
$$= P^a(x,y)P^b(y,x) \sum_{i<n-a-b} P^i(y,y) = P^a(x,y)P^b(y,x)G_{n-a-b}(y,y).$$

Now, for any constant k,

$$\lim_{n\to\infty} \frac{G_n(y,y)}{n} = \lim_{n\to\infty} \frac{G_{n-k}(y,y)}{n} = \frac{1}{m_y}.$$

y is positive recurrent, so $m_y < \infty$ by definition. Now, since $\rho_{xx} = 1$,

$$\frac{1}{m_x} = \lim_{n\to\infty} \frac{G_n(x,x)}{n} \geq P^a(x,y)P^b(y,x) \lim_{n\to\infty} \frac{G_{n-a-b}(y,y)}{n} = \frac{P^a(x,y)P^b(y,x)}{m_y} > 0.$$

Thus, $m_x < \infty$ and so m_x is positive recurrent. □