1. LECTURE 1: APRIL 3, 2012

1.1. **Motivating Remarks: Differential Equations.** In the deterministic world, a standard tool used for modeling the evolution of a system is a **differential equation**. Such an equation relates the rate of change of a measured quantity, at a given time t, to the value of that quantity at time t. A typical first order ordinary differential equation takes the form

$$x'(t) = F(t, x(t))$$

for some (hopefully nice) function F. Note that the change in x does not depend on the value of x at *earlier* times in this equation. An example of such an equation would be

$$x'(t) = x(t - t_0)$$

for some fixed time-difference t_0. This is called a **delay equation**. Delay equations are extremely difficult to solve, or even qualitatively analyze; their behavior can depend on the parameter t_0 in complicated ways. For example, if $t_0 = \frac{\pi}{2}$ then the functions $x(t) = \cos t$ and $x(t) = \sin t$ are solutions of the above delay equation. But for other values of t_0 (even close to $\frac{\pi}{2}$), solutions cannot be described in terms of elementary functions, and are not periodic.

This is a course in **stochastic processes**, meaning processes that are inherently random. In the same manner as above, such processes are vastly easier to analyze if their future behavior can be predicted knowing only their *current* behavior, without reference to the past. We will begin this course with a study of the discrete-time version of this property.

1.2. **Markov Chains.** Before we define *Markov process*, we must define stochastic processes.

Definition 1.1. *Let S be a fixed state space. A* **(discrete time) stochastic process** $(X_n)_{n\in\mathbb{N}}$ *is a sequence of random variables all defined on a fixed probability space $(\Omega, \mathbb{P})$ all taking values in S. A* **(continuous time) stochastic process** $(X_t)_{t\geq 0}$ *is a collection of random variables X_t indexed by $t \in [0, \infty)$, all defined on a fixed probability space $(\Omega, \mathbb{P})$ all taking values in S.*

We will discuss exclusively discrete time processes for the first half of the course.

Example 1.2. Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables; then they form a discrete time stochastic process with state space $\mathbb{R}$. Similarly, the process $S_n = X_1 + \cdots + X_n$ is a discrete time stochastic process with state space $\mathbb{R}$.

Example 1.3. As a special case of the previous example, sample the i.i.d. variables from the symmetric Bernoulli distribution: $\mathbb{P}(X_n = \pm 1) = \frac{1}{2}$. Then The state space for the process (X_n) is just $\{\pm 1\}$. The corresponding process S_n has state space $\mathbb{Z}$. The process S_n is called the **symmetric random walk** on $\mathbb{Z}$.

Example 1.4. Fix $N > 0$ in $\mathbb{N}$. We will soon describe stochastic processes X_n which behaves as the symmetric random walk S_n does from Example 1.3 when $0 < X_n < N$, but have different behavior when X_n reaches $0, N$. These will be known as **reflected**, **partially reflected**, and **absorbing** random walks. To describe them, we need the technology of Markov processes, which we introduce next.

Given a collection $X_1, \ldots, X_n$ of random variables with discrete state space S, their joint distribution is completely described by the collection of probabilities

$$\{\mathbb{P}(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \colon i_1, \ldots, i_n \in S\}.$$

Thinking of the index as a time parameter, it is useful to express these numbers, using the definition of conditional probability, as

$$\begin{aligned}&\mathbb{P}(X_1 = i_1, \ldots, X_n = i_n)\\ &= \mathbb{P}(X_1 = i_1)\mathbb{P}(X_2 = i_2 | X_1 = i_1)\mathbb{P}(X_3 = i_3 | X_1 = i_1, X_2 = i_2) \cdots \mathbb{P}(X_n = i_n | X_1 = i_1, \ldots, X_{n-1} = i_{n-1}).\end{aligned}$$

Definition 1.5. *Let X_n be a discrete time stochastic process, with state space S that is finite or countably infinite. Then X_n is called a* **(discrete time) Markov chain** *if, for each $n \in \mathbb{N}$ and any $i_1, \ldots, i_n \in S$,*

$$\mathbb{P}(X_n = i_n | X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) = \mathbb{P}(X_n = i_1 | X_{n-1} = i_{n-1}).$$

A Markov chain (sometimes called a Markov process, though that is usuall reserved for the case of an uncountable state space) is a stochastic process whose evolution to the next state depends only on the current state, not the past behavior. Let's reconsider the above examples:

- In Example 1.2 (making the additonal assumption that the state space is a countable or finite subset of $\mathbb{R}$ to conform to the definition), the independence assumption shows that

  $$\mathbb{P}(X_1 = i_1, \ldots, X_n = i_n) = \mathbb{P}(X_1 = i_1) \cdots \mathbb{P}(X_n = i_n).$$

  In particular, this means that $\mathbb{P}(X_n = i_n | X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) = \mathbb{P}(X_n = i_n) = \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1})$. Hence, X_n is a Markov chain. For the process S_n of sums,

  $$\mathbb{P}(S_1 = i_1, S_2 = i_2, \ldots, S_n = i_n) = \mathbb{P}(X_1 = i_1, X_2 = i_2 - i_1, \ldots, X_n = i_n - i_{n-1}).$$

  Thus

  $$\begin{aligned}\mathbb{P}(S_n = i_n | S_1 = i_1, \ldots, S_{n-1} = i_{n-1}) &= \frac{\mathbb{P}(S_1 = i_1, \ldots, S_n = i_n)}{\mathbb{P}(S_1 = i_1, \ldots, S_{n-1} = i_{n-1})}\\ &= \frac{\mathbb{P}(X_1 = i_1, X_2 = i_2 - i_1, \ldots, X_n = i_n - i_{n-1})}{\mathbb{P}(X_1 = i_1, X_2 = i_2 - i_1, \ldots, X_{n-1} = i_{n-1} - i_{n-2})}\\ &= \mathbb{P}(X_n = i_n - i_{n-1})\end{aligned}$$

  while

  $$\mathbb{P}(S_n = i_n | S_{n-1} = i_{n-1}) = \frac{\mathbb{P}(S_{n-1} = i_{n-1}, S_n = i_n)}{\mathbb{P}(S_{n-1} = i_{n-1})}.$$

  The numerator can be rewritten as

  $$\begin{aligned}\mathbb{P}(S_n - S_{n-1} = i_n - i_{n-1}, S_{n-1} = i_{n-1}) &= \mathbb{P}(X_n = i_n - i_{n-1}, S_{n-1} = i_{n-1})\\ &= \mathbb{P}(X_n = i_n - i_{n-1})\mathbb{P}(S_{n-1} = i_{n-1}),\end{aligned}$$

  since X_n is independent from $X_1, \ldots, X_{n-1}$ and hence from S_{n-1}. Dividing, we achieve the desired equality. So S_n is a Markov chain.

- In the special case of Example 1.3, we can calculate the probabilities $\mathbb{P}(S_n = i_n | S_{n-1} = i_{n-1}) = \mathbb{P}(X_n = i_n - i_{n-1})$. Since $X_n \in \{\pm 1\}$ with equal probabilities, we must have $i_n - i_{n-1} \in \pm 1$ to have positive probability, and in this case each has value $\frac{1}{2}$. In other words:

$$\mathbb{P}(S_n = i \pm 1 | S_{n-1} = i) = \frac{1}{2},$$

  and all other "transition probabilities" are 0.

In general, for a Markov chain $(X_n)_{n\in\mathbb{N}}$, the probabilities $\mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1})$ are called the **transition probabilities**. The last example shows that the transition probabilities of some Markov chains are even more uniform than the definition insists.

Definition 1.6. *A Markov chain is called* **time-homogeneous** *if, for any $i, j \in S$, the transition probability $\mathbb{P}(X_n = j | X_{n-1} = i)$ is independent of n. That is: there is a fixed function $p\colon S \times S \to [0, 1]$ with the property that, for all n,*

$$\mathbb{P}(X_n = j | X_{n-1} = i) = p(i, j).$$

The symmetric random walk above is a homogeneous-time Markov chain, as is S_n for any independent sum of i.i.d. random variables. Note that

$$\sum_{j\in S} p(i, j) = \sum_{j\in S} \mathbb{P}(X_n = j | X_{n-1} = i) = \sum_{j\in S} \frac{\mathbb{P}(X_n = j, X_{n-1} = i)}{\mathbb{P}(X_{n-1} = i)} = 1.$$

for each $i \in S$, by the law of total probability. If we let $\mathbf{P}$ be the $|S| \times |S|$ matrix with i, j entry is $p(i, j)$, this says that the rows of $\mathbf{P}$ all sum to 1; $\mathbf{P}$ is a **stochastic matrix**.

Now let's flesh out Example 1.4.

Example 1.7. Consider the following Markov chain. The state space is $\{0, \dots, N\}$. As with the full random walk, we have transition probabilities $p(i, j) = \frac{1}{2}$ if $0 < i, j < N$ and $j = i \pm 1$, and $p(i, j) = 0$ otherwise. We have yet to specify the transition probabilities involving the states $0, N$. As usual $p(0, i) = p(i, 0) = p(N, j) = p(j, N) = 0$ if $i \neq 0, 1$ and $j \neq N - 1, N$. But we impose different boundary conditions for different kinds of random walks.

- **Reflecting Random Walk.** Here $p(0, 1) = 1$ and $p(N, N - 1) = 1$ (and so $p(0, 0) = 0 = p(N, N)$). That is: the random walker bounces off the walls surely when reaching them.
- **Absorbing Random Walk.** Here $p(0, 0) = p(N, N) = 1$; so when the walker reaches the boundary, he says there forever.
- **Partially Reflecting Walk.** In general, we can set $p(0, 0) = p$, $p(0, 1) = 1 - p$, $p(N, N) = q$, and $p(N, N - 1) = 1 - q$ for some arbitrary probabilities $0 \leq p, q \leq 1$.

As we will see in the next weeks, the long-time behavior of the random walk is very depenedent on the parameters p, q in the partially reflecting walk.

2. Lecture 2: April 4, 2012

2.1. Examples.

Example 2.1. Ehrenfest urn. There are N balls and 2 urns. At each time step, choose a ball from one of the two urns randomly and move it to the opposite urn. Let X_n denote the number of balls in the first urn after n time steps. Then $(X_n)_{n\in\mathbb{N}}$ is a Markov chain with state space $\{0,\dots,N\}$ and transition probabilities $p(i,i-1)=\frac{i}{N}$ and $p(i,i+1)=\frac{N-i}{N}$ (and all other $p(i,j)=0$). [This is a statistical mechanical model of the exchange of gas between two chambers connecting by a small hole.]

Example 2.2. Wright-Fisher model. A population of fixed size N can have genes of two types: A and a. The gene pool in generation $n+1$ is obtained by sampling with replacement from the population in generation n. Let X_n denote the number of A genes in the population in generation n. Then X_n is a Markov chain with state space $\{0,1,\dots,N\}$ and transition probabilities

$$p(i,j)=\binom{N}{j}\left(\frac{i}{N}\right)^j\left(\frac{N-i}{N}\right)^{N-j}.$$

I.e. the model is that, if i A genes are present at time n, then in generation $n+1$ each individual has probability $\frac{i}{N}$ of getting A and probability $1-\frac{i}{N}$ of getting a; so the number of A is given by the binomial distribution. [This is a model of genetic drift in diploid populations with non-overlapping generations; for example annual plants.]

2.2. n-Step Transition Probabilities.

The array $p(i,j)$ gives the one-step transition probabilities of the Markov chain. But the typical question is: given an initial probability distribution (for X_0), what is the probability that $X_n=i$ for some large n and any $i\in S$? To answer that question, we will have to delve into linear algebra, because:

Lemma 2.3. *Let $(X_n)_{n\in\mathbb{N}}$ be a homogeneous-time Markov chain with state space S, with transition probabilities $p\colon S\times S\to[0,1]$, and associated transition matrix $\mathbf{P}$. Then*

$$\mathbb{P}(X_n=j|X_0=i)=[\mathbf{P}^n]_{ij}.$$

That is: the n-step transition probabilities are given by the entries of $\mathbf{P}^n$.

Proof. Let $p_n(i,j)=\mathbb{P}(X_n=j|X_0=i)$. By definition, $p_1(i,j)=p(i,j)$ (by time-homogeneity). We proceed by induction. By the law of total probability

$$\begin{aligned}\mathbb{P}(X_{n+1}=j|X_0=i)&=\sum_{k\in S}\mathbb{P}(X_{n+1}=j,X_n=k|X_0=i)\\&=\sum_{k\in S}\mathbb{P}(X_{n+1}=j|X_n=k,X_0=i)\mathbb{P}(X_n=k|X_0=i).\end{aligned}$$

Since X_n is a Markov process, $\mathbb{P}(X_{n+1}=j|X_n=k,X_0=i)=\mathbb{P}(X_{n+1}=j|X_n=k)=p(k,j)$. Hence, we have

$$p_{n+1}(i,j)=\sum_{k\in S}\mathbb{P}(X_n=k|X_0=i)p(k,j)=\sum_{k\in S}p_n(i,k)p(k,j).$$

By the inductive hypothesis, $p_n(i,k)=[\mathbf{P}^n]_{ik}$, and so we have

$$\mathbb{P}(X_{n+1}=j|X_0=i)=\sum_{k\in S}[\mathbf{P}^n]_{ik}[\mathbf{P}]_{kj}=[\mathbf{P}^{n+1}]_{ij}.$$

This proves the claim. □

We can generalize this slightly, as follows.

Corollary 2.4. *If $(X_n)_{n\in\mathbb{N}}$ is a time-homogeneous Markov chain with transition matrix $\mathbf{P}$ with entries $[\mathbf{P}]_{ij} = p(i,j) = \mathbb{P}(X_n = j | X_{n-1} = i)$, then the n-step transition probabilities $p_n(i,j) = \mathbb{P}(X_n = j | X_0 = i)$ satisfy the* **Chapman-Kolmogorov Equations***:*

$$p_{m+n}(i,j) = \sum_{k\in S} p_m(i,k) p_n(k,j).$$

Proof. In light of Lemma 2.3, this is just the restatement of the fact that powers of $\mathbf{P}$ commute under matrix multiplication:

$$p_{m+n}(i,j) = [\mathbf{P}^{m+n}]_{ij} = [\mathbf{P}^m \mathbf{P}^n]_{ij} = \sum_{k\in S} [\mathbf{P}^m]_{ik} [\mathbf{P}^n]_{kj}$$

which yields the stated sum. □

2.3. **The Markov Property.** One way to state the definition of a Markov chain is that, conditioned on the present, the future is independent from the past. This can be made precise and slightly stronger, in the following proposition which may be taken as the (recursive) definition of a homogeneous time Markov chain.

Proposition 2.5. *Let $(X_n)_{n\in\mathbb{N}}$ be a time-homogeneous Markov chain with discrete state space S and transition probabilities $p(i,j)$. Fix $m \in \mathbb{N}$ and $i \in S$, and suppose that $\mathbb{P}(X_m = i) > 0$. Then conditional on $X_m = i$, the process $(X_{m+n})_{n\in\mathbb{N}}$ is a time-homogeneous Markov chain with the same transition probabilities $p(i,j)$, and* independent *from $X_0, X_1, \ldots, X_m$.*

In other words: if A is an event determined only by $X_0, X_1, \ldots, X_m$ with $\mathbb{P}(A \cap \{X_m = i\}) > 0$, then for all $n \geq 0$

$$\mathbb{P}(X_{m+1} = i_{m+1}, \ldots, X_{m+n} = i_{m+n} | A \cap \{X_m = i\}) = p(i, i_{m+1}) p(i_{m+1}, i_{m+2}) \cdots p(i_{m+n-1}, i_{m+n}).$$

Proof. From the definition of conditional probability, what we need to prove is that

$$\begin{aligned} &\mathbb{P}(\{X_{m+1} = i_{m+1}, \ldots, X_{m+n} = i_{m+n}\} \cap A) \\ &= \mathbb{P}(\{X_m = i\} \cap A) p(i, i_{m+1}) p(i_{m+1}, i_{m+2}) \cdots p(i_{m+n-1}, i_{m+n}). \end{aligned}$$

If A takes the form $A = \{X_0 = i_0, X_1 = i_1, \ldots, X_m = i_m\}$ (and so $i_m = i$ since we made the assumption that $\mathbb{P}(A \cap \{X_m = i\}) = 0$), the left hand side is just

$$\mathbb{P}(X_0 = i_0, \ldots, X_{m+n} = i_{m+n}) = \mathbb{P}(X_0 = i_0) p(i_0, i_1) \cdots p(i_{m-1}, i) p(i, i_{m+1}) \cdots p(i_{m+n-1}, i_{m+n})$$

because $(X_n)_{n\geq 0}$ is a Markov chain. But we also have

$$\mathbb{P}(\{X_m = i\} \cap A) = \mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_m = i) = \mathbb{P}(X_0 = i_0) p(i_0, i_1) \cdots p(i_{m-1}, i)$$

for the same reason. Comparing, this proves the desired equality for such special A. But we assumed that A depends only on the $X_0, \ldots, X_m$; thus, A must be a disjoint union of events of the assumed form, and the result follows by summing. □

2.4. **Hitting Times.** One important question about a stochastic process is: when is the first time it enters a certain set? Let A be an event, and define the random time τ_A by

$$\tau_A = \min\{n \in \mathbb{N} \colon X_n \in A\}.$$

(Here we make the convention that $\min \varnothing = \infty$.) For example, we may wish to know (the distribution of) the first time the symmetric random walk gets distance N away from the origin. Better yet: in the random walk with absorbing boundaries $0, N$, we may want to know the probability that the walk reaches N before it reachs 0, given an initial distribution. This can be formalized as a general problem of this form:

Let $A, B \subset S$ with $A \cap B = \varnothing$. For $i \in S$, let $h(i) = \mathbb{P}(\tau_A < \tau_B | X_0 = i)$. Calculate the function h.

When $(X_n)_{n\in\mathbb{N}}$ is a Markov chain, this becomes a linear algebra problem. First note that it is always true that

$$h(i) = 1 \text{ if } i \in A, \quad h(i) = 0 \text{ if } i \in B.$$

Now, take $i \notin A \cup B$. Conditioning on the first step of the process, we have

$$\mathbb{P}(\tau_A < \tau_B | X_0 = i) = \sum_{j\in S} \mathbb{P}(\tau_A < \tau_B | X_0 = i, X_1 = j)\mathbb{P}(X_1 = j | X_0 = i).$$

By the Markov property, conditioning on $X_1 = j$ gives a new Markov process with the same transition probabilities, hence the same joint distributions. The probability of the event $\{\tau_A < \tau_B\} = \{\tau_A - 1 < \tau_B - 1\}$ depends only on the joint distributions of the X_ns, and so it follows that

$$\mathbb{P}(\tau_A < \tau_B | X_0 = i, X_1 = j) = \mathbb{P}(\tau_A < \tau_B | X_1 = j) = \mathbb{P}(\tau_A - 1 < \tau_B - 1 | X_0 = j) = h(j).$$

Thus, we have the equations

$$h(i) = \sum_{j\in S} p(i, j)h(j). \tag{2.1}$$

If the state space is finite, then we can think of h as a vector $\mathbf{h} \in \mathbb{R}^{|S|}$ (with $\mathbf{h}_i = h(i)$), and these equations say that $\mathbf{h} = \mathbf{P}\mathbf{h}$. That is, $\mathbf{h}$ is an eigenvector of $\mathbf{P}$ with eigenvalue 1. This is not typically enough to nail down $\mathbf{h}$ (it could be scaled, or 1 may not be a simple eigenvalue), but the boundary conditions $h(i) = 1$ for $i \in A$ and $h(i) = 0$ for $i \in B$ are often enough to uniquely determine h. Let's consider an example.

Example 2.6. Let $(X_n)_{n\in\mathbb{N}}$ be random walk on $\mathbb{Z}$, but not necessarily symmetric: we define $p(i, i + 1) = q$ and $p(i, i - 1) = 1 - q$ for some $q \in [0, 1]$ and all $i \in \mathbb{Z}$. If $q = \frac{1}{2}$ this is the symmetric random walk of Example 1.3. Let's consider the probability that the random walk reaches N before 0; with the language above, $A = \{N\}$ and $B = \{0\}$. Thus $h(N) = 1$ and $h(0) = 0$. Equation 2.1 in this case says

$$h(i) = (1 - q)h(i - 1) + qh(i + 1), \quad 0 < i < N.$$

It is convenient to rewrite this in the form

$$(1 - q)[h(i) - h(i - 1)] = q[h(i + 1) - h(i)].$$

If $q = 0$ this gives $h(i)-h(i-1) = 0$ for $0 < i < N$ and since $h(0) = 0$ it follows that $h(i) = 0$ for $i < N$. (Indeed, here the walk always moves left.) Otherwise, let $c = h(1) - h(0)$, and

let $\theta = \frac{1-q}{q}$. Define $\Delta h(i) = h(i) - h(i-1)$ for $i \geq 1$. Thus, we have a geometric progression

$$\Delta h(i+1) = \theta \Delta h(i), \quad 1 \leq i \leq N-1$$

with initial condition $\Delta h(1) = c$. The solution is

$$\Delta h(i+1) = \theta^i c, \quad 1 \leq i \leq N-1.$$

Therefore, the telescoping sum

$$h(i) = h(i) - h(i-1) + h(i-1) - h(i-2) + \cdots + h(1) - h(0) + h(0)$$

yields (since $h(0) = 0$)

$$h(i) = \Delta h(1) + \cdots + \Delta h(i) = c + \theta c + \cdots + \theta^{i-1} c = c \sum_{j=0}^{i-1} \theta^j$$

for $2 \leq i \leq N$.

- **Symmetrix RW.** Here $q = \frac{1}{2}$ so $\theta = 1$. This gives $h(i) = ci$ for $2 \leq i \leq N$. In particular $h(N) = cN$. But also $h(N) = 1$, so $c = \frac{1}{N}$, and we have $h(i) = \frac{i}{N}$.

  In terms of gambling: suppose we bet on coin tosses. If heads comes up, you win \$1; if tails, you lose \$. You begin with \$i, and will cash out either when you go bust or win \$N. If the coin is fair, the probability of cashing out rather than going bust is i/N – it increases linearly th closer your intial capital is to the cash-out.

- **Biased RW.** For $q \neq \frac{1}{2}$, we have

$$h(i) = c \sum_{j=0}^{i-1} \theta^j = c \frac{1-\theta^i}{1-\theta}, \quad 2 \leq i \leq N.$$

  In particular $1 = h(N) = c\frac{1-\theta^N}{1-\theta}$, and so $c = \frac{1-\theta}{1-\theta^N}$. Combining, this yields

$$h(i) = \frac{1-\theta^i}{1-\theta^N}, \quad 1 \leq i \leq N.$$

  This highlights the *gambler's ruin* phenomenon. In the gambling game above, we suppose the coin is biased against you (perhaps only slightly). Suppose, for example, that $q = 0.49$. If you start with \$100 and set the cash-out at \$200, then $h(100) = 0.018$, as oppoed to 0.5 in the $q = 0.5$ case.

3. LECTURE 3: APRIL 6, 2012

Last time we saw how to calculate $\mathbb{P}_i(\tau_A < \tau_B)$, the probability that a Markov chain hits a set A before a set B (starting in a given state i). Today, we will consider the closely related question: what is $\mathbb{E}_i(\tau_A)$, the *expected hitting time* of a set A?

3.1. **Expected Hitting Times.** Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with state space S and transition probabilities $p(i,j)$. Let $A \subseteq S$, and as usual set $\tau_A = \min\{n \geq 0 \colon X_n \in A\}$. Let us define the function $g \colon S \to [0,\infty]$ by

$$g(i) = \mathbb{E}_i(\tau_A) = \mathbb{E}[\tau_A | X_0 = i].$$

Note: the *conditional expectation* here is the expectation with respect to the conditional probability $\mathbb{P}_i$,

$$\mathbb{E}_i(Y) = \sum_{n=1}^{\infty} n\mathbb{P}_i(Y = n) = \sum_{n=1}^{\infty} n\frac{\mathbb{P}(Y = n, X_0 = i)}{\mathbb{P}(X_0 = i)}.$$

Of course, $g(i) = 0$ if $i \in A$. For $i \notin A$, we condition on the first step of the trajectory and use the Markov property:

$$g(i) = \mathbb{E}_i(\tau_A) = \sum_{j\in S} \mathbb{E}_i[\tau_A | X_1 = j]\mathbb{P}_i(X_1 = j).$$

Now,

$$\begin{aligned}\mathbb{P}_i(\tau_A = n | X_1 = j) &= \mathbb{P}(X_0 \notin A, X_1 \notin A, \ldots, X_{n-1} \notin A, X_n \in A | X_0 = i, X_1 = j) \\ &= \mathbb{P}(X_0 \notin A, \ldots, X_{n-2} \notin A, X_{n-1} \in A | X_0 = j) = \mathbb{P}_j(\tau_A = n-1)\end{aligned}$$

by the Markov property. Hence

$$\begin{aligned}\mathbb{E}_i[\tau_A | X_1 = j] = \sum_{n=1}^{\infty} n\mathbb{P}_i(\tau_A = n | X_1 = j) &= \sum_{n=1}^{\infty} n\mathbb{P}_j(\tau_A = n-1) \\ &= \sum_{m=0}^{\infty} (m+1)\mathbb{P}_j(\tau_A = m) \\ &= \sum_{m=0}^{\infty} m\mathbb{P}_j(\tau_A = m) + \sum_{m=0}^{\infty} \mathbb{P}_j(\tau_A = m).\end{aligned}$$

The first term is just $\mathbb{E}_j(\tau_A)$, and since τ_A is $\mathbb{N} \cup \infty$-valued, the latter sum is 1. Hence we have

$$g(i) = \sum_{j\in S} [\mathbb{E}_j(\tau_j) + 1]\mathbb{P}_i(X_1 = j) = \sum_{j\in S} [g(j) + 1]p(i,j) = 1 + \sum_{j\in S} p(i,j)g(j). \tag{3.1}$$

This system of equations may not have a unique solution, since $g(i) = \infty$ for $i \notin A$ is always a solution. But if it is known a priori that the expected hitting time is finite, it may often be used to calculate the function g.

Remark 3.1. Let $\mathbf{1}$ denote the column vector with all entries equal to 1. If we interpret the function g as a vector $\mathbf{g} \in \mathbb{R}^{|S|}$, then Equation 3.1 can be written as

$$\mathbf{g} = \mathbf{1} + \mathbf{P}\mathbf{g}$$

where $\mathbf{P}$ is the transition matrix for the Markov chain. This is an *affine* equation, rather than a linear equation; it is less amenable to the techniques of linear algebra. But in many examples, it can be solved iteratively.

Example 3.2. *On average, how many times do we need to toss a coin to get two consecutive heads?* To answer this question, we need to construct an appropriate Markov chain. There are different ways to approach this; probably the simplest is to let X_n denote the number of consecutive heads one has immediately seen immediately after the n^{th} toss; we stop the chain once we see two consecutive heads. This means X_n is a Markov chain with state space $\{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}.$$

As above, set $g(i) = \mathbb{E}_i(\tau_{\{2\}})$. Then $g(2) = 0$. Then Equation 3.1 says

$$g(0) = 1 + p(0,0)g(0) + p(0,1)g(1) + p(0,2)g(2) = 1 + \tfrac{1}{2}g(0) + \tfrac{1}{2}g(1)$$
$$g(1) = 1 + p(1,0)g(0) + p(1,1)g(1) + p(1,2)g(2) = 1 + \tfrac{1}{2}g(0) + \tfrac{1}{2}g(2) = 1 + \tfrac{1}{2}g(0)$$

(and the third equation gives no information since it says $g(2) = g(2)$). The first equation simplifies to $\frac{1}{2}g(0) - \frac{1}{2}g(1) = 1$ or $g(0) - g(1) = 2$, and the seconq equation says $g(0) = 2g(1) - 2$. Thus $(2g(1) - 2) - g(1) = 2$ and so $g(1) = 4$; therefore $g(0) = g(1) + 2 = 6$. So, starting having seen no heads (of course), on average it takes about 6 tosses until 2 consecutive heads come up.

In making the above calculations, the key tool was the Markov property: conditioned on $X_1 = j$, the chain behaves after time 1 just like another Markov chain started from j. To make other calculations, we will need something even stronger: we will want to apply this reasoning not only to shifts by deterministic times, but also by (appropriate) *random times*. I.e. we will see that, if T is an appropraite $\mathbb{N}$-valued random variable, then the shifted process X_{T+n} is still a Markov chain with the same transition probabilities.

3.2. **Stopping Times.** The appropriate kind of random time alluded to in the above paragraph is a *stopping time*.

Definition 3.3. *Let $(X_n)_{n\in\mathbb{N}}$ be a discrete-time stochastic process. A* **stopping time** *is a random variable T with state space $\mathbb{N} \cup \{\infty\}$ such that, for each n, the event $\{T = n\}$ depends only on $X_0, X_1, \ldots, X_n$.*

It is typical to think of stopping times in terms of betting strategies. If X_n represents some statistic associated to a gambling game, one has to decide when to bet (or fold) based only on information up to the present time; but that information depends on the random occurrences in the game up to that time. For example: one could decide to sell a stock once it reaches \$200; this is a stopping time. One would prefer to sell it the day before it drops; this is *not* a stopping time.

Example 3.4. Fix $i \in S$. The **return time** T_i is the first time that X_n returns to state i after time $n = 0$ (where we condition on $X_0 = i$). That is:

$$T_i = \min\{n \geq 1 \colon T_n = i\}.$$

For any given n, the event $\{T_i = n\}$ can be written as $\{X_1 \neq i, X_2 \neq i, X_{n-1} \neq i, X_n = i\}$. Thus, the event $\{T_i = n\}$ depends only on the values of $X_0, \ldots, X_n$, and hence T_i is a stopping time. Note: the hitting time $\tau_{\{i\}} = \min\{n \geq 0 \colon X_n = i\}$ is generally different from T_i (they differ if $X_0 = i$ which we generally condition on).

Example 3.5. If $T = \min\{n \geq 0 \colon X_{n+1} = i\}$, then $\{T = n\} = \{X_0 \neq i, X_1 \neq i, \ldots, X_n \neq i, X_{n+1} = i\}$ depends on the future value of X_{n+1}, and so T is not a stopping time.

3.3. **The Strong Markov Property.** Restarting a Markov chain at a stopping time produces a new Markov chain with the same transition probabilities.

Proposition 3.6. *Let $(X_n)_{n\in\mathbb{N}}$ be a time-homogeneous Markov chain with state space S and transition probabilities $p(i,j)$. Let T be a stopping time. Suppose $i \in S$ and $\mathbb{P}(X_T = i) > 0$. Then, conditional on $X_T = i$, $(X_{T+n})_{n\in\mathbb{N}}$ is a time-homogeneous Markov chain with transition probabilities $p(i,j)$, independent of $X_0, X_1, \ldots, X_T$.*

In other words: if A is an event that depends only on $X_0, X_1, \ldots, X_T$ and $\mathbb{P}(A \cap \{X_T = i\}) > 0$, then for all $n \geq 0$ and all $i_1, \ldots, i_n \in S$,

$$\mathbb{P}(X_{T+1} = i_1, X_{T+2} = i_2, \ldots, X_{T+n} = i_n | A \cap \{X_T = i\}) = p(i, i_1)p(i_1, i_2) \cdots p(i_{n-1}, i_n).$$

Proof. What we need to show is that

$$\mathbb{P}(A \cap \{X_T = i, X_{T+1} = i_1, \ldots, X_{T+n} = i_n\}) = \mathbb{P}(A \cap \{X_T = i\})p(i, i_1) \cdots p(i_{n-1}, i_n).$$

Because A depends only on $X_0, \ldots, X_T$, the event $A \cap \{T = m\}$ depends only on $X_0, \ldots, X_m$. We therefore decompose according to the value m of T and apply the Markov property:

$$\begin{aligned}
&\mathbb{P}(A \cap \{X_T = i, X_{T+1} = i_1, \ldots, X_{T+n} = i_n\}) \\
&= \sum_{m=0}^{\infty} \mathbb{P}(\{T = m\} \cap A \cap \{X_m = i, X_{m+1} = i_1, \ldots, X_{m+n} = i_n\}) \\
&= \sum_{m=0}^{\infty} \mathbb{P}(\{T = m\} \cap A \cap X_m = i)p(i, i_1) \cdots p(i_{n-1}, i_n) \\
&= \mathbb{P}(A \cap \{X_T = i\})p(i, i_1) \cdots p(i_{n-1}, i_n).
\end{aligned}$$

□

4. LECTURE 4: APRIL 9, 2012

We now begin our discussion of the limiting (long-term) behavior of Markov chains. A quick reminder: we are utilizing the brief notation:

$$\mathbb{P}_i(X_n = j) = \mathbb{P}(X_n = j | X_0 = i).$$

4.1. **Transience and Recurrence.** Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with state space S. There are only two eventual behaviors a state can have under the action X_n.

Definition 4.1. *A state $i \in S$ is called* **recurrent** *if $\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 1$. A state $i \in S$ is called* **transient** *of $\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 0$.*

We will use the strong Markov property to show that every state is either recurrent or transient. The idea is to use the following collection of stopping times. Fixing $i \in S$, and conditioning on the event $X_0 = i$, we set

$$T_{i,1} = 0, \quad T_{i,k} = \min\{n > T_{i,k} \colon X_n = i\} \text{ for } k \geq 2.$$

So $T_{i,k}$ is the time of the k^{th} visit to i. Note that, for each $m \in \mathbb{N}$,

$$\{T_{i,k} = m\} = \{X_{T_{i,k-1}+1} \neq i, X_{T_{i,k-1}+2} \neq i, \ldots, X_{m-1} \neq i, X_m = i\}$$

depends only on $X_0, \ldots, X_m$ (inductively verified, since $T_{i,k-1}$ is involved). These generalize the first return time $T_i = T_{i,2} = \min\{n \geq 1 \colon X_n = i\}$, which you saw was a stopping time in the previous lecture. Now, define

$$r_i = \mathbb{P}_i(T_i < \infty).$$

Assume, for the moment, that, for fixed $k \geq 1$, $T_{i,k} < \infty$. So-conditioned, the strong Markov property yields that the process $X_{T_{i,k}+n}$ is a Markov chain with the same transition probabilities as (X_n). It follows that

$$\mathbb{P}(T_{i,k+1} < \infty | T_{i,k} < \infty) = \mathbb{P}_i(T_i < \infty) = r_i.$$

By induction, then, we have

$$\mathbb{P}(T_{i,k} < \infty) = r_i^{k-1}.$$

This leads to the desired dichotomy.

Theorem 4.2. *Let $i \in S$. Then $r_i = 1$ iff $\sum_{n=0}^{\infty} p_n(i,i) = \infty$ iff i is recurrent; and $r_i < 1$ iff $\sum_{n=0}^{\infty} p_n(i,i) < \infty$ iff i is transient.*

Remark 4.3. The statement of Theorem 4.2 shows that one can verify recurrence/transience for the state i either by calculating (or estimating) the probability r_i, or by calculating (or estimating) the sum $\sum p_n(i,i)$; both are useful in actual examples.

Remark 4.4. The intuition is pretty clear here. If $r_i = 1$ then we're guaranteed to return to i, and then (since the process is Markov) we start over and are guaranteed to return again, eventually visiting i infinitely often. If, onthe other hand, $r_i < 1$, there is a chance we never return the first time; in the (uncertain) case we do, there is the same chance we won't visit a second time, etc.; these probabilities compound and eventually we stop returning to i. The proof below makes this intuition rigorous.

In order to prove the theorem, we need to recall a useful fact from undergraduate probability. Let T be an $\mathbb{N}$-valued random variable. Then one way to calculate its expectation is

$$\mathbb{E}(T) = \sum_{j=1}^{\infty} j\mathbb{P}(T=j) = \sum_{j=1}^{\infty}\sum_{k=1}^{j} \mathbb{P}(T=j) = \sum_{k=1}^{\infty}\sum_{j=k}^{\infty} \mathbb{P}(T=j) = \sum_{k=1}^{\infty} \mathbb{P}(T \geq k).$$

(This is sometimes called the "layer cake representation" of T.)

Proof. Define N_i to be the number of visits to i:

$$N_i = \sum_{n=0}^{\infty} \mathbb{1}_{X_n=i}.$$

Then $\{N_i \geq k\} = \{T_{i,k} < \infty\}$, and so $\mathbb{P}_i(N_i \geq k) = \mathbb{P}(T_{i,k} < \infty) = r_i^{k-1}$. We also have

$$\mathbb{E}_i[N_i] = \sum_{n=0}^{\infty} \mathbb{E}_i[\mathbb{1}_{X_n=i}] = \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = i) = \sum_{n=0}^{\infty} p_n(i,i).$$

Now consider the two cases for r_i.

- If $r_i = 1$ then $\mathbb{P}_i(N_i \geq k) = 1^{k-1} = 1$ for all k, so $\mathbb{P}_i(N_i = \infty) = 1$, which is precisely the statement that i is recurrent. Moreover, we have
$$\sum_{i=0}^{\infty} p_n(i,i) = \mathbb{E}_i[N_i] = \infty$$
as claimed.
- If $r_i < 1$ then $\mathbb{P}_i(N_i \geq k) = r_i^{k-1}$ is exponentially small and so $\mathbb{P}_i(N_i = \infty) = 0$, which his precisely the statement that i is transient. Moreover, we have
$$\sum_{i=0}^{\infty} p_n(i,i) = \mathbb{E}_i[N_i] = \sum_{k=1}^{\infty} \mathbb{P}_i(N_i \geq k) = \sum_{k=1}^{\infty} r_i^{k-1} = \frac{1}{1-r_i} < \infty.$$

□

Example 4.5. Consider simple random walk on $\mathbb{Z}$, with $p(i, i+1) = p$ and $p(i, i-1) = 1-p$. For the n-step transition probabilities, we have $p_n(i,i) = 0$ unless n is even. So we calculated $p_{2n}(i,i)$. Consider the trajectory of the path; in order for $X_{2n} = i$ given $X_0 = i$, it must be that n of the steps are left and n are right. There are $\binom{2n}{n}$ such paths, and because of the independence of the steps, each has probability $p^n(1-p)^n$ of having occurred. Thus, we have

$$p_{2n}(i,i) = \binom{2n}{n} p^n (1-p)^n.$$

So the recurrence/transience is decided by the summability of

$$\sum_{n=0}^{\infty} p_n(i,i) = \sum_{n=0}^{\infty} \binom{2n}{n} p^n (1-p)^n.$$

Of course this is finite (in fact 0) of $p \in \{0,1\}$, where the walk moves deterministically either right or left. In general, we can make the blunt upper-bound $\binom{2n}{n} \leq 4^n$: this follows from the overcount of paths, since $\binom{2n}{n}$ counts the number of length-$2n$ paths that return

to their starting point, while $4^n = 2^{2n}$ counts the number of all length-$2n$ paths. Then we have

$$\sum_{n=0}^{\infty} p_n(i,i) \leq \sum_{n=0}^{\infty} 4^n p^n (1-p)^n = \sum_{n=0}^{\infty} (4p(1-p))^n.$$

The function $p \mapsto p(1-p)$ takes maximum value $\frac{1}{4}$ at $p = \frac{1}{2}$, and is strictly less than $\frac{1}{4}$ elsewhere in $[0,1]$; hence the series is summable when $p \neq \frac{1}{2}$ and all such non-symmetric random walks on $\mathbb{Z}$ are *transient*.

When $p = \frac{1}{2}$ this gives ∞ as an upper-bound for $\sum_{n=0}^{\infty} p_n(i,i)$ which is not useful; we need a lower-bound. Here we use Stirling's approximation:

$$n! \sim \sqrt{2\pi n}(n/e)^n.$$

From this it follows that

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \sim \frac{\sqrt{4\pi n}(2n/e)^{2n}}{2\pi n (n/e)^{2n}} = \frac{2^{2n}}{\sqrt{\pi n}}.$$

Thus, in the case $p = \frac{1}{2}$, we then have

$$\sum_{n=1}^{\infty} p_n(i,i) \sum \sum_{n=0}^{\infty} \frac{2^{2n}}{\sqrt{\pi n}} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n = \sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}} = \infty.$$

Hence, the *symmetric* random walk on $\mathbb{Z}$ is *recurrent*.

4.2. **Irreducible Markov chains.** Transience and recurrence are properties of a state of a Markov chain. Example 4.5 shows that all states of any simple RW on $\mathbb{Z}$ have the same behavior: either all recurrent (when $p = \frac{1}{2}$) or all transient (when $p \neq \frac{1}{2}$). But it is perfectly possible for different states to have different asymptotic properties in a Markov chain.

Example 4.6. Consider the stupid Markov chain with state space $\{1,2\}$ where $p(1,2) = 1$ and $p(2,2) = 1$. Then 1 is transient and 2 is recurrent.

The issue that makes Example 4.6 stupid is that 1 is transient for the silly reason that you can never get back to it once you leave. If we avoid this kind of thing, then we will always have universal asymptotic behavior for the states.

Definition 4.7. *a Markov chain is* **irreducible** *if, for any two states i, j there exists $n \in \mathbb{N}$ so that $p_n(i,j) > 0$.*

For example, in a simple random walk on $\mathbb{Z}$, as soon as $n > |i-j|$ and has the same parity as $i-j$, $p_n(i,j) > 0$. Hence, Example 4.5 highlights the following proposition.

Proposition 4.8. *If $(X_n)_{n\in\mathbb{N}}$ is an irreducible Markov chain, then either all states are recurrent or all states are transient.*

Proof. It suffices to show that if even one state is transient then all others must be transient. Suppose $i \in S$ is a transient state, and let $j \in S$. By assumption, there exists $n \in \mathbb{N}$ with $p_n(i,j) > 0$, and also there exists $m \in \mathbb{N}$ with $p_m(i,j) > 0$. Then we have for any $\ell \geq 0$

$$p_{\ell+m+n}(i,i) \geq p_m(i,j) p_\ell(j,j) p_n(j,i).$$

(I.e. one way to get from i back to i in $\ell+m+n$ steps is to move from i to j in m steps, that return to j in ℓ steps, and then move from j to i in n steps.) Thus, we have the estimate

$$\sum_{\ell=0}^{\infty} p_\ell(j,j) \leq \sum_{\ell=0}^{\infty} \frac{p_{\ell+m+n}(i,i)}{p_m(i,j)p_n(j,i)} = \frac{1}{p_m(i,j)p_n(j,i)} \sum_{k=m+n}^{\infty} p_k(i,i) < \infty$$

by the assumption that i is transient. Hence j is transient as well. $\square$

5. LECTURE 5: APRIL 11, 2012

5.1. **Random Walks on Graphs.** One of the most important class of examples of Markov chains consists of *simple random walks on graphs*.

Definition 5.1. *A* **graph** $G = (V, E)$ *is a collection of vertices* V *and a relation* G *on* $V \times V$ *(the set of edges). If* $x, y \in V$ *we write* $x \sim y$ *to mean* $(x, y) \in E$*. The relation* E *is assumed to be anti-reflexive (i.e. no loops,* $x \not\sim x$ *for any* x*) and symmetric (i.e. undirected,* $x \sim y$ *means the same as* $y \sim x$*).*

The **valence** *of a vertex* $x \in V$ *in graph is* $v_x = \#\{y \in V : x \sim y\}$*. We will generally work with graphs for which every vertex has finite, non-zero valence. (If* $v_x = 0$ *then* x *is an isolated point disconnected from the rest of the graph, and we may as well ignore it.*

Definition 5.2. *The* **simple random walk** *on a graph* $G = (V, E)$ *is the Markov chain* X_n *with state space* V*, and with transition probabilities given by* $p(x, y) = \frac{1}{v_x}$ *if* $x \sim y$ *and* $p(x, y) = 0$ *if* $x \not\sim y$*.*

Example 5.3. The integers $\mathbb{Z}$ form a graph: the vertices are the integers $\mathbb{Z}$ themselves, and the edges are all adjacent pairs $(i, i+1)$ for $i \in \mathbb{Z}$. The simple random walk on this graph is precisely the symmetric random walk we studied in Example 4.5.

Example 5.4. More generally, we consider the graphs $\mathbb{Z}^d$ for $d \geq 1$. Here the vertices are all d-tuples $(i_1, \ldots, i_d) \in \mathbb{Z}^d$, and we declare $(i_1, \ldots, i_d) \sim (j_1, \ldots, j_d)$ if there is an $m \in \{1, \ldots, d\}$ with $|i_m - j_m| = 1$ and $i_k = j_k$ for all $k \neq m$. I.e. two points are adjacent if they are exactly one step apart along one coordinate axis.

Thus, the simple random walk on $\mathbb{Z}^d$ has transition probabilities

$$p\left((i_1, \ldots, i_d), (i_1, \ldots, i_m \pm 1, \ldots, i_d)\right) = \tfrac{1}{2d}, \qquad m \in \{1, \ldots, d\}$$

and all other transition probabilities 0.

Let $\mathbf{i}$ and $\mathbf{j}$ be any two vertices in $\mathbb{Z}^d$. There is, of course, a path between them along coordinate-axis single steps. If n is the minimal length of such a path, then $p_n(\mathbf{i}, \mathbf{j}) > 0$ while $p_k(\mathbf{i}, \mathbf{j}) = 0$ for $k < n$. In particular, it follows that $\mathbb{Z}^d$ is irreducible, and so all states have the same asymptotic behavior. We will therefore restrict the discussion to the state $\mathbf{0} = (0, \ldots, 0)$, the origin, as a representative state for all others.

5.2. **SRW in $\mathbb{Z}^d$.** We have already seen (Example 4.5) that the SRW on $\mathbb{Z}$ is recurrent. Now consider $\mathbb{Z}^2$.

Example 5.5. SRW on $\mathbb{Z}^2$. As usual, $p_n(\mathbf{0}, \mathbf{0}) = 0$ if n is odd. Now, there are trajectories of length $2n$ that return to the origin; since each step has transition probability $\frac{1}{2d} = \frac{1}{4}$, any particular length $2n$ trajectory will have probability 4^{-2n} of occurring. We have to count the number of trajectories.

Let m be the number of $(1, 0)$ steps; to return to the origin, there must be m $(-1, 0)$ steps. Since there are $2n$ steps total, this means $2m \leq 2n$. The same analysis for $(0, \pm 1)$ steps shows there must be k of each with $0 \leq k \leq n$. But there are exactly $2n$ steps so $2k + 2m = 2n$ so $k = n - m$. Hence, the total number of length $2n$ trajectories from $\mathbf{0}$ to $\mathbf{0}$ can be enumerated as

$$\sum_{m=0}^{n} \#\{\text{paths with } m \ (\pm 1, 0) \text{ steps and } n - m \ (0, \pm 1) \text{ steps}\}.$$

This summand is a multinomial coefficient: from $2n$, choose m $(1,0)$ steps, m $(-1,0)$ steps, $n-m$ $(0,1)$ steps, and $n-m$ $(0,1)$ steps. This is given by

$$\binom{2n}{m \;\; m \;\; n-m \;\; n-m} = \frac{(2n)!}{(m!)^2((n-m)!)^2}$$

and so we have the exact calculation

$$p_{2n}(\mathbf{0},\mathbf{0}) = 4^{-2n}\sum_{m=0}^{n}\frac{(2n)!}{(m!)^2((n-m)!)^2}.$$

This can be simplified as

$$4^{-2n}\sum_{m=0}^{n}\frac{(2n)!}{(m!)^2((n-m)!)^2} = 4^{-2n}\frac{(2n)!}{(n!)^2}\sum_{m=0}^{n}\left[\frac{n!}{m!(n-m)!}\right]^2.$$

The squared term in the sum is just $\binom{n}{m}$, and since $\binom{n}{m} = \binom{n}{n-m}$, we can write this as

$$4^{-2n}\binom{2n}{n}\sum_{m=0}^{n}\binom{n}{m}\binom{n}{n-m}.$$

A convenient binomial relationship is that

$$\sum_{m=0}^{n}\binom{n}{m}\binom{n}{n-m} = \binom{2n}{n}.$$

(Think of it this way: to choose n objects from $2n$, one could start by artificially dividing the $2n$ into two groups of n, then selecting m from the first group and $n-m$ from the second; summing over all allotments of $0 \le m \le n$ from the first and $n-m$ from the second, we achieve all ways of selecting n from the original $2n$.)

Hence, we have shown that

$$p_{2n}(\mathbf{0},\mathbf{0}) = 4^{-2n}\binom{2n}{n}^2 = \left(2^{-2n}\binom{2n}{n}\right)^2.$$

Looking back at Example 4.5, we see that this is exactly the square of $p_2 n(0,0)$ for the SRW on $\mathbb{Z}$. In particular, we already calculated that

$$2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}.$$

Hence, for SRW on $\mathbb{Z}^2$,

$$\sum_{n=0}^{\infty}p_n(\mathbf{0},\mathbf{0}) = \sum_{n=0}^{\infty}p_{2n}(\mathbf{0},\mathbf{0}) = \sum_{n=0}^{\infty}\frac{1}{\pi n} = \infty.$$

Thus, SRW on $\mathbb{Z}^2$ is recurrent.

It turns out the situation is different for SRW in $\mathbb{Z}^d$ for $d \ge 3$.

Example 5.6. SRW on $\mathbb{Z}^3$. Arguing exactly as we did above to enumerate the trajectories of length $2n$, now noting that the adjacent-step transition probabilities are $\frac{1}{6}$, we have

$$p_{2n}(\mathbf{0},\mathbf{0}) = 6^{-2n} \sum_{\substack{i,j,k\geq 0\\ i+j+k=n}} \frac{(2n)!}{(i!j!k!)^2} = 2^{-2n}\binom{2n}{n} \sum_{\substack{i,j,k\geq 0\\ i+j+k=n}} 3^{-2n}\left[\frac{n!}{i!j!k!}\right]^2.$$

You might hope this works out to exactly $p_{2n}(0,0)^3$ from SRW on $\mathbb{Z}$, but it does not. However, it turns out to be asymptotically a constant times this, as we will now see. Here are some facts that help in the calculation:

- The number of ways to put n balls into 3 boxes is, of course, 3^n; but by considering all ways of subdividing them individually, we have
$$3^n = \sum_{\substack{i,j,k\geq 0\\ i+j+k=n}} \binom{n}{i\ \ j\ \ k} = \sum_{\substack{i,j,k\geq 0\\ i+j+k=n}} \frac{n!}{i!j!k!}.$$
- If n is divisible by 3, say $n=3m$, and $i+j+k=n=3m$, then $i!j!k! \geq (m!)^3$ and so $\frac{n!}{i!j!k!} \leq \frac{n!}{(m!)^3}$. (I.e. the "middle" multinomial coefficient is largest; this can be checked easily by induction.)
- If $a_{\mathbf{i}} \geq 0$ are non-negative with sum $\sum_{\mathbf{i}} a_{\mathbf{i}} = 1$, then
$$\sum_{\mathbf{i}} a_{\mathbf{i}}^2 \leq \sum_{\mathbf{i}} a_{\mathbf{i}} \cdot \max_{\mathbf{j}} a_{\mathbf{j}} = \max_{\mathbf{j}} a_{\mathbf{j}}.$$

Now, define $a^n_{\mathbf{i}} = a^n_{i,j,k} = 3^{-n}\frac{n!}{i!j!k!}$. The first fact above asserts that $\sum_{\mathbf{i}} a^n_{\mathbf{i}} = 1$, where the sum is over those $\mathbf{i} = (i,j,k)$ in $\mathbb{N}$ with $i+j+k=n$. We calculated above that

$$p_{2n}(\mathbf{0},\mathbf{0}) = 2^{-2n}\binom{2n}{n}\sum_{\mathbf{i}}(a^n_{\mathbf{i}})^2.$$

So, the third fact above shows us that

$$p_{2n}(\mathbf{0},\mathbf{0}) \leq 2^{-2n}\binom{2n}{n}\max_{\mathbf{j}} a^n_{\mathbf{j}}.$$

For the moment, take $n=3m$. Then the second fact above gives

$$\max_{\mathbf{j}} a^n_{\mathbf{j}} = \max_{\substack{i,j,k\geq 0\\ i+j+k=n}} 3^{-n}\frac{n!}{i!j!k!} \leq 3^{-n}\frac{n!}{(m!)^3}.$$

Hence, we have

$$p_{6m}(\mathbf{0},\mathbf{0}) \leq 2^{-2n}\binom{2n}{n}\cdot 3^{-n}\frac{n!}{(m!)^3} \sim \frac{1}{\sqrt{\pi n}}\cdot\frac{3^{-n}\sqrt{2\pi n}(n/e)^n}{(2\pi m)^{3/2}(m/e)^{3m}} = \frac{1}{2\pi^{3/2}m^{3/2}}$$

using Stirling's formula again.

This only gives us an upper bound on p_{6m}; to get p_{2n} for all n, we also need p_{6m-2} and p_{6m-4}. But note: if there is a length $6m-2$ path from $\mathbf{0}$ to $\mathbf{0}$, then by traveling one step forward and back along any of the 3 axes, we produce a path of length $6m$; since each step

has transition probability $\frac{1}{6}$, we then have the lower bound $p_{2m}(\mathbf{0},\mathbf{0}) \geq (\frac{1}{6})^2 p_{6m-2}(\mathbf{0},\mathbf{0})$. Similarly, we have $p_{2m}(\mathbf{0},\mathbf{0}) \geq (\frac{1}{6})^4 p_{6m-4}(\mathbf{0},\mathbf{0})$. This gives the estimate

$$\sum_{n=1}^{\infty} p_{2n}(\mathbf{0},\mathbf{0}) = \sum_{m=1}^{\infty} [p_{6m-4}(\mathbf{0},\mathbf{0}) + p_{6m-2}(\mathbf{0},\mathbf{0}) + p_{6m}(\mathbf{0},\mathbf{0})] \leq 6^4 \sum_{m=1}^{\infty} p_{6m}(\mathbf{0},\mathbf{0})$$

and thus

$$\sum_{n=1}^{\infty} p_{2n}(\mathbf{0},\mathbf{0}) \leq \frac{6^4}{2\pi^{3/2}} \sum_{m=1}^{\infty} \frac{1}{m^{3/2}} < \infty.$$

Ergo, SRW on $\mathbb{Z}^3$ is transient.

In particular, by Theorem 4.2, this shows that $r_{\mathbf{0}}^3 = \mathbb{P}_{\mathbf{0}}(T_{\mathbf{0}}^3 < \infty)$, the probability of return to $\mathbf{0}$ in $\mathbb{Z}^3$, is < 1. Now, in $\mathbb{Z}^d$, in order for the process to return to $\mathbf{0}$, all components must return to $\mathbf{0}$; hence we have $\{T_{\mathbf{0}}^{d-1} < \infty\} \supseteq \{T_{\mathbf{0}}^d < \infty\}$ for all d, and hence the probability $r_{\mathbf{0}}^d$ is non-increasing with dimension. It therefore follows that SRW on $\mathbb{Z}^d$ is transient for all $d \geq 3$.

Remark 5.7. It can be calculated, using generating-function techniques we do not discuss here, that the return-probability in $\mathbb{Z}^3$ is $r_{\mathbf{0}}^3 \approx 0.34$.

6. LECTURE 6: APRIL 13, 2012

Let $(X_n)_{n\geq 0}$ be a Markov chain with state space S. For the moment, we will not condition on a particular initial state. Instead, suppose the initial probability distribution is π_0; i.e. $\mathbb{P}(X_0 = i) = \pi_0(i)$ for $i \in S$. The Markov property then allows us to compute the distribution of X_n for any n:

$$\mathbb{P}(X_n = j) = \sum_{i\in S} \mathbb{P}(X_n = j, X_0 = i) = \sum_{i\in S} \mathbb{P}(X_0 = i)\mathbb{P}(X_n = j|X_0 = i) = \sum_{i\in S} \pi_0(i)p_n(i,j).$$

As we established in Lecture 2 (cf. the Chapman-Kolmogorov Equations), $p_n(i,j)$ is the (i,j)-entry of the matrix $\mathbf{P}^n$. So we have for $n \geq 0$, the distribution $\pi_n(j) = \mathbb{P}(X_n = j)$ can be computed as

$$\pi_n(j) = \sum_{i\in S} \pi_0(i)p_n(i,j) = [\pi_0\mathbf{P}^n]_j.$$

I.e. considering π_0 and π_n as row-vectors, we have $\pi_n = \pi_0\mathbf{P}^n$.

So, the question of what happens for large n amounts to understanding the behavior of high powers of stochastic matrices.

6.1. Stationarity.

Suppose that π is a probability vector with the property that $\pi = \mathbf{P}\pi$; i.e. π is a right eigenvector of $\mathbf{P}$ with eigenvalue 1. Then, of course, $\pi = \pi\mathbf{P}^n$ for all n; in other words, if π is ever the distribution of the process at some time, it will remain *stationary* in that distribution for all time. So we call such a π a *stationary distribution*. We can be a little forgiving with the requirement that it be a distribution.

Definition 6.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space S transition matrix $\mathbf{P}$. A vector $\pi = [\pi(i)\colon i \in S]$ is called a* **stationary measure** *if $\pi(i) \geq 0$ for all $i \in S$ and $\pi\mathbf{P} = \pi$. If, in addtion, $\sum_{i\in S} \pi(i) = 1$, we call π a* **stationary distribution**.

Note: at least in the case of a finite-state Markov chain, if there is a stationary measure π with *at least one* non-zero entry, then we can normalize the vector to produce a stationary distribution $\hat{\pi}$: just take $\hat{\pi} = (\sum_{i\in S} \pi(i))^{-1}\pi$.

Example 6.2. Suppose a 2-state Markov chain has transition matrix

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}.$$

If $\pi = [x, y]$ is a stationary measure, then we have the equations

$$[x, y] = [x, y] \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = [\tfrac{1}{2}x + \tfrac{1}{4}y, \tfrac{1}{2}x + \tfrac{3}{4}y].$$

These two equations actually amount to the same thing: $\frac{1}{2}x = \frac{1}{4}y$, so $y = 2x$. Hence there is a one-parameter family of stationary measures; with the additional constraint that $x + y = 1$, we see the unique stationary distribution is

$$\pi = [1/3, 2/3].$$

(If there exists a stationary measure, it is never unique: one can always scale an eigenvector. The question is whether 1 is an eigenvalue.)

Now, let us look at the powers of the matrix $\mathbf{P}$. Accurate to 3 digits,

$$\mathbf{P}^2 = \begin{bmatrix} 0.375 & 0.625 \\ 0.313 & 0.688 \end{bmatrix}, \qquad \mathbf{P}^4 = \begin{bmatrix} 0.336 & 0.664 \\ 0.332 & 0.668 \end{bmatrix}, \qquad \mathbf{P}^7 = \begin{bmatrix} 0.333 & 0.667 \\ 0.333 & 0.667 \end{bmatrix}.$$

It appears (and is easily verified) that

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{bmatrix}.$$

I.e. all the rows of the limiting transition matrix are equal to the stationary distribution π. In particular, if $\nu = [x, y]$ is any distribution (so $x + y = 1$), it then follows that $\lim_{n\to\infty} \nu\mathbf{P}^n = \pi$; i.e. no matter what is the starting distribution, the limit distribution of the chain is the stationary distribution.

Example 6.2 leaves us with the following general questions.

(Q1) When does a Markov chain possess a stationary distribution?
(Q2) If it exists, when is the stationary distribution unique?
(Q3) When does the distirbution of X_n converge to the stationary distribution?

First let's consider some examples to see that all three of these can go wrong.

Example 6.3. (1) Let $S = \mathbb{Z}$ and set $p(i, i+1) = 1$ for $i \in \mathbb{Z}$. This chain is deterministic: it always moves right. So all states are transient. That is: regardless of the starting distribution, $\lim_{n\to\infty} \mathbb{P}(X_n = i) = 0$ for any $i \in \mathbb{Z}$. This means there is is no stationary distribution. (Indeed: if there were a stationary distribution, then starting in it would preserve the distribution for all time and so some state, with stationary distribution having positive probability, would have fixed positive probability of being visited each turn – it would therefore be recurrent.)

(2) Let $S = \{1, 2, 3, 4\}$ and suppose $\mathbf{P}$ has the block form

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}.$$

Note, then, that all powers of the matrix have this same block form; in particular $p_n(i, j) = 0$ for all n if $i \in \{1, 2\}$ and $j \in \{3, 4\}$. It is also easy to check that the two vectors $[1/2, 1/2, 0, 0]$ and $[0, 0, 1/2, 1/2]$ are both stationary states, so this chain possesses stationary states but they are not unique.

(3) Consider SRW on the square (vertices $1, 2, 3, 4$ with the four edges $(1, 2)$, $(2, 3)$, $(3, 4)$, $(4, 1)$). If $X_0 = 0$, then X_n is even if n is even and odd if n is odd. So for any $j \in \{1, 2, 3, 4\}$, $\lim_{n\to\infty} \mathbb{P}(X_n = j)$ cannot exist; this chain is *periodic* and cannot have a stationary distribution. (We'll come back to periodicity later.)

As we will see, the three examples in 6.3 show all the possible ways Questions 1–3 can fail to have positive answers.

6.2. **General 2-state Markov chain.** Let's take a moment to completely classify the behavior of all 2-state Markov chains to gain intuition. Any 2-state Markov chain has a transition matrix of the form

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

for some $0 \leq p, q \leq 1$. To compute all powers of this matrix, we diagonalize. The characteristic polynomial is $(\lambda - (1-p))(\lambda - (1-q)) - pq = \lambda^2 + (p+q-2)\lambda + 1 - p - q$ whose roots are $\lambda \in \{1, 1-p-q\}$.

First, suppose that $\{p, q\} \in \{0, 1\}$. That is, consider separately the four special cases:

$$\mathbf{P} \in \left\{ \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right\}.$$

In the first two cases, one can easily check that the matrix is already in the form $\begin{bmatrix} \pi \\ \pi \end{bmatrix}$ for π the stationary distribution, as desired. The latter two are different, however. For the identity matrix, *every* vector is a left-eigenvector with eigenvalue 1, and so although it is true that $\pi I^n = \pi$ for all n (and so too in the limit), the stationary distirbution is not unique, and the limit distribution depends on the initial distribution. Worse yet, for the last matrix: the powers do not stablize, but cycle periodically between the matrix and I. The eigenvalues are ± 1, and the unique stationary distribution is $[\frac{1}{2}, \frac{1}{2}]$, but the limit $\lim_{n\to\infty} \mathbb{P}(X_n = j)$ does not exists, regardless of the initial distribution.

Now, suppose that at least one of p, q is strictly in $(0, 1)$. Then the eigenvalues are distinct and we can find two linearly independent eigenvectors. The result is that we can write $\mathbf{P}$ in the form $\mathbf{P} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}$ where

$$\mathbf{Q} = \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix}, \qquad \mathbf{Q}^{-1} = \begin{bmatrix} q/(p+q) & p/(p+q) \\ -1/(p+q) & 1/(p+q) \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} 1 & 0 \\ 0 & 1-p-q \end{bmatrix}.$$

Note: writing it this way means $\mathbf{P}\mathbf{Q} = \mathbf{Q}\mathbf{D}$ or $\mathbf{Q}^{-1}\mathbf{P} = \mathbf{D}\mathbf{Q}^{-1}$ and so the columns of $\mathbf{Q}$ are *right*-eigenvectors for $\mathbf{P}$, and the *rows* of $\mathbf{Q}^{-1}$ are *left*-eigenvectors for $\mathbf{P}$. The normalized left-eigenvector for eigenvalue 1 is $[q/(p+q), p/(p+q)]$, so this is the stationary distribution π. Now, since $0 \leq p, q \leq 1$ and one is strictly in $(0, 1)$, the eigenvalue $1 - p - q$ has $|1 - p - q| < 1$. Thus

$$\mathbf{D}^n = \begin{bmatrix} 1 & 0 \\ 0 & (1-p-q)^n \end{bmatrix} \to \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{as } n \to \infty$$

and so

$$\begin{aligned} \mathbf{P}^n = \mathbf{Q}^{-1}\mathbf{D}^n\mathbf{Q} &\to \begin{bmatrix} q/(p+q) & p/(p+q) \\ -1/(p+q) & 1/(p+q) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} \\ &= \begin{bmatrix} q/(p+q) & p/(p+q) \\ q/(p+q) & p/(p+q) \end{bmatrix} = \begin{bmatrix} \pi \\ \pi \end{bmatrix}, \quad \text{as } n \to \infty. \end{aligned}$$

Hence, the chain converges to the stationary distribution regardless of initial distribution, provided at least one of p, q is in $(0, 1)$.

Remark 6.4. If both p, q are in $(0, 1)$, then all entries of $\mathbf{P}$ are strictly positive. This turns out to be enough in general to guarantee the existence of a unique stationary distribution and universal convergence to it. Now, if (say) $p \in (0, 1)$ but $q = 1$, then the matrix has the form

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix}, \text{ and so } \mathbf{P}^2 = \begin{bmatrix} 1-p+p^2 & p-p^2 \\ 1-p & p \end{bmatrix}$$

and so all entries of $\mathbf{P}^2$ are strictly positive, and so similar reasoning will allow us to say something about stationary distributions and limit theorems for such chains.

6.3. **Perron-Frobenius Theorem.** Let $\mathbf{P}$ be a (finite-dimensional) stochastic matrix. It is easy to check that the *column vector* $\mathbf{1}$ with all entries equal to 1 is a *left*-eigenvector for $\mathbf{P}$ with eigenvalue 1: $\mathbf{P1} = \mathbf{1}$. So 1 is always an eigenvalue of a stochastic matrix, and therefore there exists a left eigenvector for this eigenvalue. Suppose we could show that:

(1) 1 is a simple eigenvalue, so its (left-)eigenspace is 1-dimensional.
(2) There exists a left eigenvector with all non-negative entries.
(3) All non-1 eigenvalues have magnitude strictly < 1.

Even in this case, it may not be possible to diagonalize the matrix $\mathbf{P}$, but we don't need to. Using the *Jordan normal form* for $\mathbf{P}$, we can decompose $\mathbf{P} = \mathbf{QDQ}^{-1}$ where $\mathbf{D}$ is not diagonal (it may have entries on the super-diagonal as well), $\mathbf{Q}$ has columns that are generalized right-eigenvectors, and $\mathbf{Q}^{-1}$ has columns that are generalized left-eigenvectors. In particular, this means we will have

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & M & \\ 0 & & & \end{bmatrix}, \qquad \mathbf{Q} = \begin{bmatrix} 1 & \cdots \\ 1 & \cdots \\ \vdots & \ddots \\ 1 & \cdots \end{bmatrix}, \qquad \mathbf{Q}^{-1} = \begin{bmatrix} \pi \\ \vdots \end{bmatrix}$$

where π is a left-eigenvector with eigenvalue 1, and the submatrix M is bidiagonal matrix and easily seen to satisfy $M^n \to 0$ as $n \to \infty$ (by condition (3) above). (Consult a linear algebra textbook, or Wikipedia, regarding the Jordan normal form.) Hence

$$\lim_{n\to\infty} \mathbf{P}^n = \mathbf{Q} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{0} & \\ 0 & & & \end{bmatrix} \mathbf{Q}^{-1} = \begin{bmatrix} \pi \\ \vdots \\ \pi \end{bmatrix}.$$

Theorem 6.5 (Perron, Frobenius)**.** *Let $\mathbf{M}$ be an $N \times N$ matrix all of whose entries are strictly positive. Then there is an eigenvalue $r > 0$ such that all other eigenvalues λ satisfy $|\lambda| < r$. Moreover, r is a simple eigenvalue, and its one-dimensional eigenspace contains a vector with all strictly-positive eintres. Finally, r satisfies the bounds $\min_i \sum_j \mathbf{M}_{ij} \le r \le \max_i \sum_j \mathbf{M}_{ij}$.*

The Perron-Frobenius theorem is one of the most important results in linear algebra. Applied to a stochastic matrix (with strictly positive entries), we see that the Perron-Frobenius eigenvalue $r = 1$ (since it is bounded above by the maximal row sum and below by the minimal row sum, both of which are equal to 1 for a stochastic matrix). Hence, conditions (1), (2), and (3) above are satisfied if $\mathbf{P}$ has all strictly positive entries. Of course, there are plenty of Markov chains with some transition probabilities 0; but we can generalize a little bit.

Corollary 6.6. *Suppose $\mathbf{P}$ is a stochastic matrix with the property that, for some $n \in \mathbb{N}$, $\mathbf{P}^n$ has all entries strictly positive. Then $\mathbf{P}$ satisfies conditions (1), (2), and (3) above, and therefore there exists a unique stationary distribution π and $\lim_{n\to\infty} \nu\mathbf{P}^n = \pi$ for any initial distribution ν.*

Proof. If λ is an eigenvalue of $\mathbf{P}$ with eigenvector $\mathbf{v}$ then λ^n is an eigenvalue of $\mathbf{P}^n$ with eigenvector $\mathbf{v}$. By the Perron-Frobenius theorem, the non-1 eigenvalues of $\mathbf{P}^n$ are all < 1 in magnitude, and so the same holds true for their n^{th} roots; since 1 is a simple eigenvector for $\mathbf{P}^n$, the same must be true for $\mathbf{P}$, and the eigenvector π of 1 (for $\mathbf{P}^n$, and so also for $\mathbf{P}$) may be chosen with all positive entries. □

7. LECTURE 7: APRIL 16, 2012

We will soon classify all Markov chains whose transition matrices $\mathbf{P}$ have the property that $\mathbf{P}^n$ has all positive entries for some $n \geq 1$. In the meantime, we give a probabilistic interpretation to the (positive) entries of the stationary distribution.

Proposition 7.1. *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with finite state space S, satisying the conditions of Corollary 6.6. Let π denote the unique stationary distribution. Then for each $j \in S$, $\pi(j)$ is equal to the asymptotic expected fraction of time the chain spends in state j.*

Proof. To make the statement of the proposition clear, let $Y(j,n)$ denote the number of time steps the process is in state j up through time n:

$$Y(j,n) = \sum_{m=0}^{n} \mathbb{1}_{\{X_m=j\}}.$$

Then the precise statement we are proving is that, for any initial state $i \in S$,

$$\pi(j) = \lim_{n\to\infty} \frac{\mathbb{E}_i[Y(j,n)]}{n+1}.$$

To prove this, simply note that

$$\mathbb{E}_i[Y(j,n)] = \sum_{m=0}^{n} \mathbb{E}_i(\mathbb{1}_{\{X_m=j\}}) = \sum_{m=0}^{n} \mathbb{P}_i(X_m = j) = \sum_{m=0}^{n} [e_i\mathbf{P}^n]_j$$

where e_i is the vector with a 1 in the j-entry and 0s elsewhere. By assumption $\epsilon_i\mathbf{P}^n \to \pi$ as $n \to \infty$, ans the result follows. (I.e. this is the elementary fact that if a_n is any sequence of real numbers that converges to $\alpha \in \mathbb{R}$, then the "running average" $(a_1 + a_2 + \cdots + a_n)/n$ converges to α as well.) □

On Homework 2, you also explore the connection between the stationary distribution and the expected return times T_i for states $i \in S$. Here is a somewhat heuristic argument for the relation between them. As usual, let $T_{i,k}$ denote the time of the k^{th} return to the state i (so that, conditioned on $X_0 = i$, $T_i = T_{i,2}$). Write this time as

$$T_{i,k} = T_1 + T_2 + \cdots + T_k$$

where $T_\ell = T_{i,\ell} - T_{i,\ell-1}$ for $\ell \geq 1$. The Markov property shows that the T_ℓ are independent and identically distributed (with distribution $T_1 \sim T_i$). Now, the law of large numbers therefore gives

$$\frac{T_{i,k}}{k} = \frac{1}{k}(T_1 + \cdots + T_k) \approx \mathbb{E}(T_i).$$

I.e. there are approximately k visits to state i in the first $k\mathbb{E}(T_i)$ steps of the chain. But Proposition 7.1 states that in n steps there are about $n\pi(i)$ such visits to the state i. Taking $n = k\mathbb{E}(T_i)$, we find $k\mathbb{E}(T_i)\pi(i) \approx k$ for large k, and so we have loosely derived the fact

$$\mathbb{E}(T_i) = \frac{1}{\pi(i)}.$$

You will give a rigorous proof of this fact on Homework 2.

7.1. **Periodic Chains.** Let $\mathbf{P}$ be the transition matrix for an irreducible chain. The **period** of a state $i \in S$, denoted $d(i)$, is the greatest common divisor of the set of times J_i for which the chain can return from i to i:

$$d(i) = \gcd J_i, \qquad J_i = \{n \geq 0 \colon p_n(i,i) > 0\}.$$

Lemma 7.2. *If $\mathbf{P}$ is the transition matrix for an irreducible Markov chain, then $d(i) = d(j)$ for all states i, j.*

Proof. For fixed state i, note that if $n \in J_i$ and $m \in J_i$ then $n + m \in J_i$ (if there is a positive probability trajectory of length n and another of length m, then the conjunction of the two is a positive probability trajectory of length $n + m$). Hence, J_i is closed under addition. Now, if J is any subset of $\mathbb{N}$ closed under addition and $d = \gcd J$, then $J \subseteq \{0, d, 2d, \ldots\} = d\mathbb{N}$ (otherwise J would contain a number not divisible by d which contradicts the definition of d).

Now, let j be another state. The chain is irreducible, so there are $m, n \in \mathbb{N}$ so that $p_n(i,j) > 0$ and $p_m(j,i) > 0$; hence $m + n \in J_i$. Since $m + n \in J_i$ there is $k \in \mathbb{N}$ so that $m + n = kd(i)$. Now, suppose that $\ell \in J_j$. Since $p_\ell(j,j) > 0$, we have

$$p_{m+n+\ell}(i,i) \geq p_m(i,j)p_\ell(j,j)p_n(j,i) > 0$$

and so $m + n + \ell \in J_i$; hence $m + n + \ell$ is divisible by $d(i)$. Thus, ℓ is divisible by $d(i)$ as well. Thus, $d(i)$ is a common divisor of J_j, and it follows that $d(j)$ divides $d(i)$. The symmetric argument shows that $d(i)$ divides $d(j)$, so $d(i) = d(j)$. □

It therefore makes sense to talk about the **period of an irreducible Markov chain** or the **period of $\mathbf{P}$**.

Example 7.3. Let G be a finite, connected graph. Then the simple random walk on G is irreducible. Note, if i, j are two adjacent vertices, then $p_2(i,i) \geq p(i,j)p(j,i) > 0$, and so the period of i is ≤ 2; so simple random walks on graphs have period 1 or 2. It is not hard to see that the period is 2 if and only if the graph is *bipartite*: $G = (V_1 \sqcup V_2, E)$ where there are no edges between any two vertices in either V_1 or V_2, only between the two non-empty disjoint sets V_1 and V_2. For example, any cyclic graph (like the square, in Example 6.3(3) above) is bipartitie, partitioned by parity.

A Markov chain is called **aperiodic** of all of its states have period 1.

7.2. **Limit Theorem for Irreducible Aperiodic Chains.** The Perron-Frobenius approach we discussed in the last lecture works precisely for irreducible aperiodic Markov chains.

Theorem 7.4. *Let $\mathbf{P}$ be the transition matrix for a finite-state, irreducible, aperiodic Markov chain. Then there exists a unique stationary distribution π: $\pi\mathbf{P} = \pi$, and for any initial probability distribution ν,*

$$\lim_{n\to\infty} \nu\mathbf{P}^n = \pi.$$

For the proof, we need a slightly stronger number theoretic claim about additively-closed subsets of $\mathbb{N}$. If $J \subseteq \mathbb{N}$ is closed under addition, then $J \subseteq d\mathbb{N}$ where $d = \gcd J$ as shown above. What's more, it is true that J contains *all but finitely many multiples of d*. In other words, there is $M \in \mathbb{N}$ such that $md \in J$ for all $m \geq M$. (A proof of this fact is outlined in Exercise 1.21 on page 41 of Lawler.)

Proof. Using the Perron-Frobenius theorem as discussed in Corollary 6.6, it sufficies to show that there is an $n \in \mathbb{N}$ for which $\mathbf{P}^n$ has all strictly positive entries. Fix states i, j. Since the chain is irreducible, there is a time $n(i,j)$ such that $p_{n(i,j)}(i,j) > 0$. Since the chain is aperiodic, the state i has period 1 and so, as per the discussion preceding this proof, J_i contains all sufficiently large integers: there is $M(i) \in \mathbb{N}$ so that $p_m(i,i) > 0$ for all $m \geq M(i)$. Hence, if $m \geq M(i)$,

$$p_{m+n(i,j)}(i,j) \geq p_m(i,i)p_{n(i,j)}(i,j) > 0.$$

Now, since the state space is finite, $n = \max\{M(i) + n(i,j) \colon i,j \in S\}$ is finite. The above shows that $p_n(i,j) > 0$, concluding the proof. □

7.3. **Reducible Chains.** Irreducible Markov chains are the easiest to handle, in large part because of Proposition 4.8. Other Markov chains are called **reducible**, because they reduce (in some sense) to collections of irreducible chains. Following Definition 4.7 and the proof of Proposition 4.8, they key feature of irreducible chains is the ability of all states to eventually communicate with each other.

Definition 7.5. *Let (X_n) be a Markov chain with state space S. Say that two states $i, j \in S$ can* **communicate**, *written $i \leftrightarrow j$, if there exist $n, m \in \mathbb{N}$ so that $p_n(i,j) > 0$ and $p_m(j,i) > 0$.*

Lemma 7.6. *The relation $\leftrightarrow$ on the state space S (cf. Definition 7.5) is an equivalence relation.*

Proof. Since $p_0(i,i) = 1$ by definition, $i \leftrightarrow i$ and so the relation is reflexive. Symmetry is built into the definition. We need only verify transitivity. Suppose, then, that $p_m(i,j) > 0$ and $p_n(j,k) > 0$. Then

$$\begin{aligned} p_{m+n}(i,k) &= \mathbb{P}_i(X_{m+n} = k) \\ &\geq \mathbb{P}_i(X_{m+n} = k, X_m = j) \\ &= \mathbb{P}_i(X_m = j)\mathbb{P}_i(X_{m+n} = k | X_m = j) \\ &= \mathbb{P}_i(X_m = j)\mathbb{P}_j(X_n = k) \\ &= p_m(i,j)p_n(j,k) > 0 \end{aligned}$$

where the penultimate equality is the Markov property. A similar argument applied to (k,j) proves transitivity. □

The equivalence classes under $\leftrightarrow$ partition the state space S into **communication classes**. An irreducible Markov chain is one with exactly one communication class. Examining the proof of Proposition 4.8 shows that:

Corollary 7.7. *Let (X_n) be any Markov chain. Then all states within any one of its communication classes have the same asymptotic behavior: either all recurrent or all transient.*

Remark 7.8. Similarly, the proof of Lemma 7.2 shows that any two states in a given communication class have the same period.

Based on Corollary 7.7, we refer to communication classes either as **recurrent classes** or **transient classes**. If S is a *finite state space*, at least one class must be recurrent (since at least one state must be visited infnitely often).

Let i, j be states in two distinct classes. If $p(i,j) > 0$ then it must be true that $p_n(j,i) = 0$ for all n (or else they would be in the same class). It follows that $\mathbb{P}_j(X_n = i) = 0$ for all n, and so i is transient. Hence, if i, j are *recurrent* states in distinct communication classes, we must have $p(i,j) = 0$ (and also $p(j,i) = 0$ by the symmetric argument).

Let us agree to order the states by communication class, with recurrent classes $R_1, \ldots, R_r$ first. The above argument shows that the Markov chain "reduces" the subsets R_ℓ: once in a state in R_ℓ, the chain stays in R_ℓ forever; we can consider this as its own Markov chain with transition matrix $\mathbf{P}_\ell$. The transition matrix for the full chain has the form

$$\mathbf{P} = \left[\begin{array}{cccc|c} \mathbf{P}_1 & & & & \\ & \mathbf{P}_2 & & \mathbf{0} & \\ & & \mathbf{P}_3 & & \mathbf{0} \\ & \mathbf{0} & & \ddots & \\ & & & & \mathbf{P}_r \\ \hline & & \mathbf{S} & & \mathbf{Q} \end{array}\right]$$

where the bottom rows $[\ \mathbf{S}\ |\mathbf{Q}]$ give the transition probabilities starting in the transient classes. Now, this block diagonal form is preserved under matrix multiplication, so we have

$$\mathbf{P}^n = \left[\begin{array}{cccc|c} \mathbf{P}_1^n & & & & \\ & \mathbf{P}_2^n & & \mathbf{0} & \\ & & \mathbf{P}_3^n & & \mathbf{0} \\ & \mathbf{0} & & \ddots & \\ & & & & \mathbf{P}_r^n \\ \hline & & \mathbf{S}_n & & \mathbf{Q}^n \end{array}\right]$$

for some matrix $\mathbf{S}_n$. Note that the row sums of $[\ \mathbf{S}\ |\mathbf{Q}]$ are all 1 and so the same remains true for $[\ \mathbf{S}_n\ |\mathbf{Q}^n]$ since $\mathbf{P}^n$ is a stochastic matrix. So $\mathbf{Q}^n$ is *substochastic* for all n (its row-sums are ≤ 1). This will be important later.

8. LECTURE 8: APRIL 18, 2012

8.1. **Transient States.** We saw above that, at least in the case of an irreducible chain, if a stationary distribution π exists then $\pi(j)$ represents the average amount of time the chain spends in state j. If j is a transient state, however, this average ought to be 0. Of course, in a finite-state irreducible Markov chain, all states have the same asymptotic behavior, and so since there must be at least one recurrent state, all states are recurrent. But in an infinite-state Markov chain, or a reducible chain, this highlights one obstacle to the existence of a stationary distribution.

The question remains: how can we calculate or estimate the (finite) number of visits that will eventually be made to a transient state j? That is: fix a transient state j, and define

$$Y_j = \sum_{n=0}^{\infty} \mathbb{1}_{\{X_n=j\}}$$

(which, by assumption, is almost surely finite). We would like to calculate $\mathbb{E}_i[Y_j]$ for some starting state i.

$$\mathbb{E}_i[Y_j] = \mathbb{E}_i\left[\sum_{n=0}^{\infty} \mathbb{1}_{\{X_n=j\}}\right] = \sum_{n=0}^{\infty} \mathbb{E}_i[\mathbb{1}_{\{X_n=j\}}] = \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = j) = \sum_{n=0}^{\infty} p_n(i,j).$$

This last sum is the (i,j)-entry of the matrix $\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \cdots$ which, a priori, may not converge at all. However, consider the block-diagonal form above for the matrix $\mathbf{P}$. Let us assume that *both i and j are transient states*, which means that the (i,j)-entry of $\mathbf{P}^n$ is the (i,j)-entry of $\mathbf{Q}^n$, where $\mathbf{Q}$ is the substochastic matrix for the transitions of transient states to transient states. By definition of transience, for each i,j there is N so that $[\mathbf{Q}^n]_{ij} = 0$ for all $n \geq N$. Hence, the sum

$$[\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \cdots]_{ij} = [\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots]_{ij}$$

is actually a finitely-terminating series; so, in particular, it converges. In terms of calculating it, however, we use the fact that it converges to use the geometric series to express it as

$$\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots = (\mathbf{I} - \mathbf{Q})^{-1}.$$

Example 8.1. Consider the random walk on $\{0,1,2,3,4\}$ with absorbing boundaries (cf. Example 1.7). The states $0,4$ are recurrent, so we should order the states $\{0,4,1,2,3\}$; then the transition matrix becomes

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}, \qquad \mathbf{Q} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

So we can easily and quickly calculate

$$(\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{bmatrix}.$$

Thus, starting in state 1, the expected number of visits to state 1 is $\frac{3}{2}$, to state 2 is 1, and to state 3 is $\frac{1}{2}$. Hence, the total expected number of steps before absorption at the boundary, starting in state 1 is $\frac{3}{2} + 1 + \frac{1}{2} = 3$. The same expected number of steps holds starting in state 3, while starting in state 2 the expected number of steps is $1 + 2 + 1 = 4$.

8.2. **Eventual Recurrence Probabilities.** As another application of the matrix $(\mathbf{I} - \mathbf{Q})^{-1}$, consider a Markov chain with at least two distinct recurrent classes. What is the probability, when starting in a given initial state, of eventually ending up in a particular recurrent class? To answer this question, it suffices to consider a reduced Markov chain: collapse each recurrent class to a single point. The reduced Markov chain then has transition matrix of the form

$$\mathbf{P} = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{S} & \mathbf{Q} \end{array}\right]$$

(i.e. each $\mathbf{P}_k$ has been collapsed to 1). Denote the states of this Markov chain as either r_i (for the reduced recurrent states) or t_j (for the original transient states). Let

$$\alpha(t_i, r_j) = \mathbb{P}_{t_i}(X_n = r_j \text{ eventually}).$$

Of course we also have $\alpha(r_j, r_k) = \delta_{jk}$ and $\alpha(x, t_j) = 0$ for all states x and all transient states t_j. Let $\mathbf{A}$ be the square matrix with $\alpha(x, y)$ as entries, for all $x, y \in \{r_1, r_2, \ldots, r_k, t_1, \ldots, t_m\}$. To calculate these probabilities, we use the Markov property as usual to assert that

$$\begin{aligned}
\alpha(t_i, r_j) &= \mathbb{P}_{t_i}(X_n = r_j \text{ eventually}) \\
&= \sum_{x \in S} \mathbb{P}_{t_i}(X_n = r_j \text{ eventually}, X_1 = x) \\
&= \sum_{x \in S} \mathbb{P}_{t_i}(X_1 = x)\mathbb{P}_{t_i}(X_n = r_j \text{ eventually} | X_1 = x) \\
&= \sum_{x \in S} p(t_i, x)\mathbb{P}_x(X_n = r_j \text{ eventually}) \\
&= \sum_{x \in S} p(t_i, x)\alpha(x, r_j).
\end{aligned}$$

In other words, $\mathbf{A}$ satisfies the equations $\mathbf{A} = \mathbf{PA}$, at least for entries of the form (t_i, r_j). If we write $\mathbf{A}$ in block form as such, we have

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{A}_{tr} & \mathbf{0} \end{array}\right]$$

where $\mathbf{A}_{tr}$ are the entries calculated above. The matrix equation $\mathbf{A} = \mathbf{PA}$ then yields the matrix equation $\mathbf{A}_{tr} = \mathbf{S} + \mathbf{QA}_{tr}$, or

$$\mathbf{A}_{tr} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{S}.$$

Example 8.2. Again, consider the random walk on $\{0, 1, 2, 3, 4\}$ with absorbing boundaries, cf. Example 7.3 above. Then we have

$$\mathbf{S} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \qquad \mathbf{Q} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix}, \qquad (\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{bmatrix}.$$

Hence

$$\mathbf{A}_{tr} = (\mathbf{I}-\mathbf{Q})^{-1}\mathbf{S} = \begin{bmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}.$$

Thus, starting at state 1, the walk is eventually absorbed at 0 with probability $\frac{3}{4}$; starting at state 2 its an even split for where it ends up; starting at state 3 the walk is absorbed at state 1 with probability $\frac{1}{4}$.

8.3. **Transit Times.** Techniques like those above can be used to determine expected transit times from one state to another, in an irreducible chain. Fix a state i, and let $T_i = \min\{n \geq 1 \colon X_n = i\}$ as usual. We saw (in Lecture 3) a technique for calculating $\mathbb{E}_i[T_i]$, but what about the more general question of $\mathbb{E}_j[T_i]$ for different states i, j?

One approach is to define a new Markov chain from the original which converts the coveted state i into an absorbing state. First, write the transition matrix for the original chain with the stat i ordered first:

$$\mathbf{P} = \left[\begin{array}{c|c} p(i,i) & \mathbf{r} \\ \hline \mathbf{s} & \mathbf{Q} \end{array}\right].$$

Then the new chain X'_n has transition matrix

$$\mathbf{P}' = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{s} & \mathbf{Q} \end{array}\right].$$

Now, for any state $k \neq i$ let Y_i^k denote the number of visits to k before reaching i. Then

$$\mathbb{E}_j[T_i] = \mathbb{E}_j\left[\sum_{k\neq i} Y_i^k\right] = \sum_{k\neq i} \mathbb{E}_j[Y_i^k].$$

Now, because we have changed the chain to make i an absorbing state, if $X'_n = i$ then it will never reach a state $k \neq i$ at any time $\geq n$. Thus, we can calculate the expectation of Y_i^k simply as

$$\mathbb{E}_j[Y_i^k] = \mathbb{E}_j\left[\sum_{n=0}^{\infty} \mathbb{1}_{\{X'_n = k\}}\right] = \sum_{n=0}^{\infty} \mathbb{E}_j[\mathbb{1}_{\{X'_n=k\}}] = \sum_{n=0}^{\infty} \mathbb{P}_j(X'_n = k) = \sum_{n=0}^{\infty} p'_n(j,k).$$

As above, $p'_n(j,k) = [(\mathbf{P}')^n]_{jk}$, and so long as $j, k \neq i$, this means that

$$\mathbb{E}_j[Y_i^k] = \sum_{n=0}^{\infty} [\mathbf{Q}^n]_{jk} = [(\mathbf{I}-\mathbf{Q})^{-1}]_{jk}.$$

Thus, for $j \neq i$,

$$\mathbb{E}_j[T_i] = \sum_{k\neq i} [(\mathbf{I}-\mathbf{Q})^{-1}]_{jk} = [(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{1}]_j.$$

Here $\mathbf{1}$ is the column-vector with all entries equal to 1.

Example 8.3. Consider the random walk on $\{0, 1, 2, 3, 4\}$ with reflecting boundary:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Taking the state $i = 0$ as the target, we use the preceding technique to calculated the expected transit time to 0. The submatrix $\mathbf{Q}$ and the associated matrix $(\mathbf{I} - \mathbf{Q})^{-1}$ are

$$\mathbf{Q} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad (\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} 2 & 2 & 2 & 1 \\ 2 & 4 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 2 & 4 & 6 & 4 \end{bmatrix}.$$

Thus

$$(\mathbf{I} - \mathbf{Q})^{-1}\mathbf{1} = \begin{bmatrix} 2 & 2 & 2 & 1 \\ 2 & 4 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 2 & 4 & 6 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 12 \\ 15 \\ 16 \end{bmatrix}.$$

So, starting at 1, it takes about 7 steps to get to 0; starting at 4 it takes about 16 steps.

9. LECTURE 9: APRIL 20, 2012

We now turn our attention to issues that arise in the world of discrete-time Markov chains with countably infinitely many states. To begin, we consider a class of examples called *birth-and-death chains*.

9.1. **Birth-and-Death chains.** Consider a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathbb{N} = \{0, 1, 2, \ldots\}$. For $i \geq 1$, we will have $p(i, j) = 0$ unless $|i - j| = 1$, as with simple random walk. But we generally allow the transition probabilities to vary with location:

$$p(i, i+1) = p_i,\ p(i, i-1) = q_i = 1 - p_i,\ \ i \geq 1.$$

As for $p(0, 0) = 1 - p(0, 1)$, we can choose this parameter to be 1 (absorbing), 0 (reflecting), or anything in between for mixed boundary conditions. Such a chain can serve as a general model for population growth, for example. As such, the important questions are $\mathbb{P}_i(\exists n \geq 0\ X_n = 0)$ (i.e. starting with population i, what is the probability the population dies out) and $\mathbb{P}_i(X_n \to \infty \text{ as } n \to \infty)$ (i.e. starting with population i, what is the probability the population explodes).

As usual, let $h(i) = \mathbb{P}_i(\exists n \geq 0\ X_n = 0)$. Then standard first step analysis shows that

$$h(i) = \sum_{j\geq 0} \mathbb{P}_i(\exists n \geq 0\ X_n = 0 | X_1 = j)\mathbb{P}_i(X_1 = k) = \sum_{j\geq 0} \mathbb{P}_k(\exists n \geq 0\ X_n = 0)p(i, j) = \sum_{j\geq 0} p(i, j)h(j).$$

For this kind of chain, with $i \geq 1$, this says $h(i) = p_i h(i + 1) + q_i h(i - 1)$. This is a 2-step recurrence relation, but it has variable coefficients, so it is a little trickier to handle, but not too much trickier. Let $u(i) = h(i-1) - h(i)$ for $i \geq 1$. Then $h(i) = p_i h(i+1) + q_i h(i-1)$ can be rewritten as $p_i u(i + 1) = q_i u(i)$. Assuming $p_i \neq 0$ for any i, this gives us

$$u(i + 1) = \frac{q_i}{p_i} u(i) = \frac{q_i q_{i-1} \cdots q_1}{p_i p_{i-1} \cdots p_1} u(1)$$

by induction. Define

$$\rho_i = \frac{q_i q_{i-1} \cdots q_1}{p_i p_{i-1} \cdots p_1}.$$

Then the telescoping sum gives us $u(1) + u(2) + \cdots + u(i) = h(0) - h(i)$, and so

$$h(0) - h(i) = u_1(1 + \rho_1 + \cdots + \rho_{i-1}). \tag{9.1}$$

The value of $h(0)$ will be determined by the boundary conditions; for example if $p(0, 0) = 1$ then $h(0) = 1$; if $p(0, 1) = 1$ then $h(0) = 0$. The first difference $u(1)$ remains undetermined. We consider two cases.

- Suppose $\sum_{i=0}^{\infty} \rho_i = \infty$. Since $h(i)$ is a probability, $h(i) \in [0, 1]$. Thus the sequence $h(0) - h(i)$ must remain bounded. Taking $i \to \infty$ then shows that we must have $u(1) = 0$ here, and so $h(i) = h(0)$ for all i (boring case).
- If, on the other hand, $\sum_{i=0}^{\infty} \rho_i < \infty$, then $u(1)$ can be taken > 0 provided that

  $$h(i) = h(0) - u(1)(1 + \rho_1 + \cdots + \rho_i) \geq 0 \quad \text{for all } i \geq 1.$$

  This suggests how to find a solution: the minimal possible solution will be achieved when $\lim_{i\to\infty}[h(0) - u(1)(1 + \rho_1 + \cdots + \rho_i)] = 0$. So we take

  $$u(1) = h(0) \left(\sum_{j=0}^{\infty} \rho_j \right)^{-1}.$$

Thus, the solution is

$$h(i) = h(0) - h(0)\left(\sum_{j=0}^{\infty}\rho_i\right)^{-1}\sum_{j=0}^{i}\rho_i = h(0)\frac{\sum_{j=i}^{\infty}\rho_j}{\sum_{j=0}^{\infty}\rho_j}.$$

Example 9.1. In a population growth model we certainly must have $p(0,0) = 1$, giving boundary condition $h(0) = 1$. So long as $p_i > 0$ for each i, this implies that $\rho_i > 0$, and hence we either have $\sum \rho_i = \infty$ in which case $h(i) = 1$ for all i – i.e. guaranteed extinction – or $\sum \rho_i < \infty$ in which case $h(i) < 1$ and $h(i) \to 1$ as $i \to \infty$ – i.e. there is positive probability of survival, and this probability increased with larger starting population (as expected).

9.2. **Positive and Null Recurrence.** In general, we can refine the definition of *recurrence* further (although, as we'll see, this really only makes sense for infinite-state Markov chains).

Definition 9.2. *Let i be a recurrent state for a Markov chain. Denote $T_i = \min\{n \geq 1 \colon X_n = 1\}$ as usual. If $\mathbb{E}_i[T_i] < \infty$, we call i* **positive recurrent***. If, on the other hand, $\mathbb{E}_i[T_i] = \infty$, then we call i* **null recurrent***.*

Remark 9.3. Note, it is certainly possible for a finite-valued random variable T to have infinite expectation. For example, suppose that $\mathbb{P}(T = 2^k) = 2^{-k}$ for each $k \geq 1$. Then T takes only the values $\{2, 4, 8, 16, \ldots\}$ and is never infinite, but $\mathbb{E}[T] = \sum_k 2^k \cdot \mathbb{P}(T = 2^k) = \sum_{k=1}^{\infty} 1 = \infty$.

Proposition 9.4. *In a finite-state, irreducible Markov chain, all states are positive recurrent.*

Proof. Fix states $i, j \in S$. By the irreducible assumption, there is a finite time $n(i,j)$ so that $p_{n(i,j)}(i,j) > 0$. Let $N = \max\{n(i,j) \colon i,j \in S\}$ and let $q = \min\{p_{n(i,j)}(i,j) \colon i \in S\}$. Thus, regardless of the current state i, there is a positive probability $\geq q$ of reaching state j in the next $\leq N$ steps; i.e. $\mathbb{P}_i(T_j \leq N) \geq q$, and hence $\mathbb{P}_i(T_j > N) \leq 1 - q$. We now apply the Markov property: for any states $i, j \in S$ and any $k \in \mathbb{N}$,

$$\mathbb{P}_j(T_j > (k+1)N | T_j > kN, X_{kN} = i) = \mathbb{P}_i(T_j > N) \leq 1 - q.$$

Of course, conditioned on $T_j > kN$ the event $T_j > (k+1)N$ is independent of X_{kN} by the Markov property, and so in fact we have $\mathbb{P}_j(T_j > (k+1)N | T_j > kN) \leq 1 - q$. By induction it follows that $\mathbb{P}_j(T_j > kN) \leq (1-q)^k$ for all $k \in \mathbb{N}$. Thus:

$$\mathbb{E}_j[T_j] = \sum_{n=1}^{\infty}\mathbb{P}_j(T_j \geq n) = \sum_{k=0}^{\infty}\sum_{n=kN+1}^{(k+1)N}\mathbb{P}_j(T_j \geq n).$$

We can bound the internal sum by

$$\sum_{n=kN+1}^{(k+1)N}\mathbb{P}_k(T_j \geq n) \leq \sum_{n=kN+1}^{(k+1)N}\mathbb{P}_j(T_j > kN) = N\mathbb{P}_j(T_j > kN)$$

because $\mathbb{P}_j(T_j \geq n)$ is non-increasing with n. Thus

$$\mathbb{E}_j[T_j] \leq \sum_{k=0}^{\infty} N\mathbb{P}_j(T_j > kN) \leq N\sum_{k=0}^{\infty}(1-q)^k < \infty.$$

□

Thus, null recurrence can only occur if there are infinitely many states (at least in the case of an irreducible chain). For infinite state Markov chains, even irreducible ones, it can certainly happen that the states are null recurrent. Indeed, simple random walk on $\mathbb{Z}$ gives an example.

Example 9.5. Let $(X_n)_{n\geq 0}$ denote SRW on $\mathbb{Z}$. Let us calculate $\mathbb{E}_0[T_0]$, the expected return time to 0 starting at 0 (of course this will be the same as $\mathbb{E}_i[T_i]$ for any state $i \in \mathbb{Z}$).

$$\mathbb{E}_0[T_0] = \sum_{n\geq 1} \mathbb{P}_0(T_0 \geq n) = \mathbb{P}_0(X_1 \neq 0, X_2 \neq 0, \ldots, X_n \neq 0).$$

We can evaluate this probability by summing over all non-0 states:

$$\begin{aligned}\mathbb{P}_0(X_1 \neq 0, \ldots, X_n \neq 0) &= \sum_{i_1,\ldots,i_n \neq 0} \mathbb{P}_0(X_1 = i_1, \ldots, X_n = i_n) \\ &= \sum_{i_1,\ldots,i_n \neq 0} p(0,i_1)p(i_1,i_2)\cdots p(i_{n-1},i_n).\end{aligned}$$

Given the form of the transition probabilities, there are only finitely many $(i_1, \ldots, i_n)$ for which this product is non-zero: namely those that satisfy $i_{k+1} = i_k \pm 1$ for each k. So $i_1 = \pm 1$, and so forth. Thus $(i_1, \ldots, i_n)$ forms a lattice path. The restriction that $i_1, \ldots, i_n \neq 0$ means that, after the first step $i_1 = \pm 1$, the path must stay either strictly above 0 (if $i_1 = 1$) or strictly below 0 (if $i_1 = -1$). By symmetry, we consider only the first case, so we have $i_1 = 1$, $i_{k+1} = i_k \pm 1$, and $i_k \geq 1$ for all k; call such paths P_n^+. Along any such path, all the probabilities $p(i_k, i_{k+1}) = \frac{1}{2}$. Thus, we have

$$\mathbb{P}(X_1 \neq 0, \ldots, X_n \neq 0) = \sum_{(1,i_2,\ldots,i_n)\in P_n^+} \frac{1}{2^n} + \sum_{(-1,i_2,\ldots,i_n)\in -P_n^+} \frac{1}{2^n} = 2^{-n+1}|P_n^+|.$$

So we would like to count the number of paths in P_n^+. This turns out to be tricky. But we can bound the size from below as follows. Let $P_n^{+,m}$ denote the subset of P_n^+ of those paths that end at height m (so $1 + i_2 + \cdots + i_n = m$). Consider a pair π_1, π_2 of paths in P_n^+. By reversing π_2 (i.e. if $\pi_2 = (1, i_2, \ldots, i_n)$ then $\pi_2^{-1} = (i_n, i_{n-1}, \ldots, i_2, 1)$), the path $\pi_1\pi_2^{-1}$ is a lattice path of length $2n$ that starts and ends at height 1, and never drops below height 1. By decreasing all heights by 1, we see this path is a **Dyck path** of length $2n$. What's more, any Dyck path of length $2n$, whose middle-height is $m-1$, can be identified with the pair π_1, π_2 as above.

So, if $D_{n,m-1}$ denotes the set of Dyck paths of length $2n$ whose middle height is $m-1$, then we have a bijection $P_n^{+,m} \times P_n^{+,m} \cong D_{n,m-1}$ for all $m \geq 1$. (These sets are nonempty only if $m \leq n+1$.) So we have $|D_{n,m-1}| = |P_n^{+,m}|^2$, and so

$$|P_n^+| = \sum_{m=1}^{n+1} |P_n^{+,m}| = \sum_{m=1}^{n+1} |D_{n,m-1}|^{1/2}.$$

It follows that

$$|P_n^+|^2 = \left(\sum_{m=1}^{n+1} |D_{n,m-1}|^{1/2}\right)^2 \geq \sum_{m=1}^{n+1} |D_{n,m-1}| = |D_n|$$

where D_n is the set of all Dyck paths. It is a well known result that D_n is counted by the Catalan number $|D_n| = \frac{1}{n+1}\binom{2n}{n}$. Thus, we have

$$|P_n^+| \geq \sqrt{\frac{1}{n+1}\binom{2n}{n}} \sim \sqrt{\frac{1}{n+1}\frac{2^{2n}}{\sqrt{\pi n}}} \sim \frac{2^n}{\sqrt{\pi}}\frac{1}{n^{3/4}}.$$

Hence, we have

$$\mathbb{P}_0(X_1 \neq 0, \ldots, X_n \neq 0) \gtrsim 2^{-n+1} \cdot \frac{2^n}{\sqrt{\pi}}\frac{1}{n^{3/4}} = \frac{2}{\sqrt{\pi}}\frac{1}{n^{3/4}}.$$

In particular, this means that

$$\mathbb{E}_0[T_0] = \sum_{n\geq 1} \mathbb{P}_0(X_1 \neq 0, \ldots, X_n \neq 0) \gtrsim \sum_{n\geq 1} \frac{2}{\sqrt{\pi}}\frac{1}{n^{3/4}} = \infty.$$

We showed in Example 4.5 that all states of the SRW on $\mathbb{Z}$ are recurrent. So the state 0 is null recurrent.

9.3. **Positive Recurrence and Stationary Distributions.** Proposition 9.4 shows that finite-state (irreducible) Markov chains have only positive recurrent states. As the next theorem shows, this is the real reason that stationary distributions exists for such chains.

Theorem 9.6. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space S that is countable (but not necessarily finite). Suppose there exists a positive recurrent state $i \in S$: i.e. $\mathbb{E}_i[T_i] < \infty$, where, as usual, $T_i = \min\{n \geq 1 \colon X_n = i\}$. For each state $j \in S$, define*

$$\gamma(i,j) = \mathbb{E}_i\left[\sum_{n=0}^{T_i-1} \mathbb{1}_{\{X_n=j\}}\right]$$

be the expected number of visits to j before returning to i. Then the function $\pi \colon S \to \mathbb{R}_+$

$$\pi(j) = \frac{\gamma(i,j)}{\mathbb{E}_i[T_i]}$$

is a stationary distribution for $(X_n)_{\{n\geq 0\}}$.

Remark 9.7. Note: Theorem 9.6 gives a much simpler proof that no state can be positive recurrent for the SRW on $\mathbb{Z}$. Indeed, if such a state existed, then there would necessarily exist a stationary distribution π for the SRW. From the form of the transition probabilities, this would mean that

$$\pi(j) = \frac{1}{2}\pi(j-1) + \frac{1}{2}\pi(j+1).$$

for all j. Rearranging this (by writing $\pi(j) = \frac{1}{2}\pi(j) + \frac{1}{2}\pi(j)$, subtracting, and multiplying through by 2), we have

$$\pi(j+1) - \pi(j) = \pi(j) - \pi(j-1).$$

Thus, if there is ever a j such that $\pi(j+1) \neq \pi(j)$, we see that the sequence π is an arithmetic progression and is both unbounded and not always positive, which means it is not a stationary distribution. If, on the other hand, $\pi(j) = \pi(j+1)$ for all j, then π is constant, and is either 0 so is not a distribution or is not summable so is not a distribution.

Thus, the SRW on $\mathbb{Z}$ does not possess an invariant distribution, and so by Theorem 9.6, it cannot have any positive recurrent states.

10. LECTURE 10: APRIL 23, 2012

We begin with the proof of Theorem 9.6 from the previous lecture.

Proof of Theorem 9.6. First note that $\mathbb{E}_i[T_i]$ can be calculated as $1+$ the sum over all states $j \neq i$ of the expected number of visits to j before returning to i:

$$\mathbb{E}_i[T_i] = 1 + \sum_{j \neq i} \gamma(i,j) = \sum_{j \in S} \gamma(i,j)$$

where the second inequality follows from the fact that $\gamma(i,i) = 1$ by definition. It follows that

$$\sum_{j \in S} \pi(j) = \frac{\sum_{j \in S} \gamma(i,j)}{\mathbb{E}_i[T_i]} = 1.$$

Since $\gamma(i,j)$ and $\mathbb{E}_i[T_i]$ are ≥ 0, the components of π are ≥ 0; so π is a probability distribution. It remains to show that it is stationary: that $\pi(j) = \sum_{k \in S} \pi(k) p(k,j)$ for all $j \in S$. By multiplying through both sides by $\mathbb{E}_i[T_i]$, what we wish to show is that $\gamma(i,j) = \sum_{k \in S} \gamma(i,k) p(k,j)$ for all $j \in S$. To that end, we calculate: for $j \in S$,

$$\gamma(i,j) = \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \mathbb{1}_{\{X_n = j\}} \right] = \mathbb{E}_i \left[\sum_{n=0}^{\infty} \mathbb{1}_{\{X_n = j, T_i > n\}} \right] = \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = j, T_i > n).$$

On the other hand, so long as $j \neq i$ we have $\mathbb{E}_i[\mathbb{1}_{\{X_{T_i} = j\}}] = 0$ and so

$$\mathbb{E}_i \left[\sum_{n=0}^{T_i} \mathbb{1}_{\{X_n = j\}} \right] = \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \mathbb{1}_{\{X_n = j\}} \right] = \gamma(i,j).$$

Thus, in the case $j \neq i$, we have

$$\gamma(i,j) = \mathbb{E}_i \left[\sum_{n=0}^{T_i} \mathbb{1}_{\{X_n = j\}} \right] = \mathbb{E}_i \left[\sum_{n=1}^{T_i} \mathbb{1}_{\{X_n = j\}} \right] = \mathbb{E}_i \left[\sum_{m=0}^{T_i - 1} \mathbb{1}_{\{X_{m+1} = j\}} \right]$$

where the second equality follows from the fact that $\mathbb{E}_i[\mathbb{1}_{X_0 = j}] = 0$. On the other hand, when $j = i$ $\gamma(i,i) = 1$ and the right-hand-side is

$$\mathbb{E}_i \left[\sum_{m=0}^{T_i - 1} \mathbb{1}_{\{X_{m+1} = i\}} \right] = \sum_{m=0}^{T_i - 1} \mathbb{P}(X_{m+1} = i) = \sum_{n=1}^{T_i} \mathbb{P}(X_n = i) = 1$$

by definition of T_i. So for all states j we have

$$\gamma(i,j) = \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \mathbb{1}_{\{X_{n+1} = j\}} \right] = \sum_{n=0}^{\infty} \mathbb{P}_i(X_{n+1} = j, T_i > n).$$

All told, we have the two seemingly competing equalities:

$$\gamma(i,j) = \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = j, T_i > n), \quad \text{and}$$

$$\gamma(i,j) = \sum_{n=0}^{\infty} \mathbb{P}_i(X_{n+1} = j, T_i > n).$$

Therefore, applying the Markov property, we have

$$\begin{aligned}\sum_{k\in S}\gamma(i,k)p(k,j) &= \sum_{k\in S}\sum_{n=0}^{\infty}\mathbb{P}_i(X_n=k,T_i>n)p(k,j)\\ &= \sum_{n=0}^{\infty}\sum_{k\in S}\mathbb{P}_i(X_n=k,X_{n+1}=j,T_i>n)\\ &= \sum_{n=0}^{\infty}\mathbb{P}_i(X_{n+1}=j,T_i>n)=\gamma(i,j),\end{aligned}$$

as claimed.

□

Corollary 10.1. *If i is a positive recurrent state, then the stationary distribution π defined in Theorem 9.6 satisfies*

$$\pi(i)=\frac{1}{\mathbb{E}_i[T_i]}.$$

Proof. This follows immediately from the definition of π, since $\gamma(i,i)=1$. □

Corollary 10.1 gives a more succinct interpretation of the components of the stationary distribution, should it exist; it also points out the necessity of positive recurrence for the existence of a stationary distribution with positive entries. The next result makes this fact clearer.

Proposition 10.2. *Suppose $(X_n)_{n\geq 0}$ is a time-homogeneous Markov chain with state space S, and suppose the chain possesses a stationary distribution π.*

(1) *If $(X_n)_{n\geq 0}$ is irreducible, then $\pi(j)>0$ for all $j\in S$.*
(2) *In general, if $\pi(j)>0$, then j is positive recurrent.*

Proof. (1) Since π is a stationary distribution, $\pi=\pi\mathbf{P}$ and hence by induction $\pi=\pi\mathbf{P}^n$ for all n. Thus, $\pi(j)=\sum_{i\in S}\pi(i)p_n(i,j)$. There must exists some i with $\pi(i)>0$ (since $\pi(i)\geq 0$ for all i and $\sum_i\pi(i)=1>0$). By irreducibility, for this choice of i, there is some n with $p_n(i,j)>0$. Hence $\pi(j)\geq\pi(i)p_n(i,j)>0$.

(2) Suppose that j is not positive recurrent. Start the chain in distribution π: so $\mathbb{P}(X_0=i)=\pi(i)$ for all i. Denote by $V_n=V_n(j)$ the number of visits (after time 0) to j up to time n:

$$V_n(j)=\sum_{m=1}^{n}\mathbb{1}_{\{X_m=j\}}.$$

Since π is a stationary distribution, we have

$$\mathbb{E}_\pi[V_n(j)]=\sum_{m=1}^{n}\mathbb{E}_\pi[\mathbb{1}_{\{X_m=j\}}]=\sum_{m=1}^{n}\mathbb{P}_\pi(X_m=j)=\sum_{m=1}^{n}\pi(j)=n\pi(j).$$

Now, let T_j^k be the time of the kth visit to state j: $T_j^k=\min\{n\geq 0\colon V_n(j)=k\}$. We claim that

$$\mathbb{P}_\pi\left(\lim_{k\to\infty}\frac{T_j^k}{k}=\infty\right)=1. \tag{10.1}$$

Assume, for the moment, that we have proved Equation 10.1. It follows that, for fixed $M > 0$, the probability that $T_j^k/k \leq M$ is small for large k. In particular, there must exist N so that $\mathbb{P}(T_j^N/N \leq M) \leq \frac{1}{M}$. Note that

$$T_j^N/N \leq M \iff \min\{n \colon V_n(j) = N\} \leq NM.$$

Now, if $V_{NM}(j) \geq N$ (there are at least N visits to j in the first NM time steps) then the minimal time n by which there are N visits to j is $\leq NM$. In other words

$$\{V_{NM}(j) \geq N\} \subset \{T_j^N \leq NM\}$$

and so

$$\mathbb{P}_\pi(V_{NM}(j) \geq N) \leq \mathbb{P}_\pi(T_j^N \leq NM) \leq \frac{1}{M}.$$

But this implies that

$$MN\pi(j) = \mathbb{E}_\pi[V_{MN}] = \sum_{k=1}^{NM} \mathbb{P}_\pi(V_{NM} \geq k)$$

since $V_{NM} \leq NM$. We may estimate this sum as follows:

$$\sum_{k=1}^{NM} \mathbb{P}_\pi(V_{NM} \geq k) = \sum_{k=1}^{N} \mathbb{P}_\pi(V_{NM} \geq k) + \sum_{k=N+1}^{NM} \mathbb{P}_\pi(V_{NM} \geq k).$$

The first sum has N terms each ≤ 1, so is $\leq N$. Since the events $\{V_{NM} \geq k\}$ are decreasing as k increases, the second sum may be bounded by

$$\sum_{k=N+1}^{NM} \mathbb{P}_\pi(V_{NM} \geq k) \leq \sum_{k=N+1}^{NM} \mathbb{P}_\pi(V_{NM} \geq N) \leq \sum_{k=N+1}^{NM} \frac{1}{M} = \frac{NM - N}{M} < N.$$

Hence, we have $NM\pi(j) < 2N$, and so $\pi(j) < \frac{2}{M}$ for each $M > 0$, which contradicts the assumption that $\pi(j) > 0$.

It remains to verify Equation 10.1. By assumption j is not positive recurrent. If j is transient, then $T_j^k = \infty$ for sufficiently large k, proving Equation 10.1 trivially in this case. So, suppose j is null recurrent. Define

$$\tau_j^k = T_j^{k+1} - T_j^k$$

be the duration of the kth excursion away from state j. Then we can write $T_j^2 = T_j^1 + \tau_j^1$, $T_j^3 = T_j^2 + \tau_j^2$, and so forth. Clearly τ_j^k is a stopping time. Hence, by the strong Markov property, $\tau_j^1, \tau_j^2, \ldots$ are independent and identically-distributed. Moreover, the distirbution is the same as that of $T_j^1 - T_j^0 = T_j$; hence $\mathbb{E}(\tau_j^k) = \infty$ by the assumption of null recurrence. Now

$$\frac{T_j^k}{k} = \frac{T_j^1 + \tau_j^1 + \cdots + \tau_j^{k-1}}{k} = \frac{T_j^1}{k} + \frac{k-1}{k} \frac{\tau_j^1 + \cdots + \tau_j^{k-1}}{k-1}.$$

The first term tends to 0 since $T_j^1 < \infty$ (as the state j is recurrent). By the Strong Law of Large Numbers, the second term tends to $\mathbb{E}(T_j) = \infty$ with probability 1. This proves Equation 10.1. $\square$

11. Lecture 11: April 25, 2012

Proposition 10.2 allows us to conclude that positive recurrence is a class property.

Corollary 11.1. *For an irreducible Markov chain, the following are equivalent:*

(1) *There exists a stationary distribution with all entries* > 0.
(2) *There exists a stationary distribution.*
(3) *There exists a positive recurrent state.*
(4) *All states are positive recurrent.*

Proof. The implication (4) $\implies$ (3) is trivial. The implication (3) $\implies$ (2) is the statement of Theorem 9.6. Implication (2) $\implies$ (1) is statement (1) of Proposition 10.2, while implication (1) $\implies$ (4) is statement (2) of Proposition 10.2. □

Example 11.2. Consider a birth and death chain with homogeneous probabilities: $S = \{0, 1, 2, \ldots\}$ and $p(i, i+1) = q \in (0, 1)$ for $i \in S$ while $p(i, i-1) = 1 - q$ for $i \geq 1$ and $p(0, 0) = 1 - q$. Since $0 < q < 1$ it is easy to see that the chain is irreducible. We wish to determine if there exists a stationary distribution. If π is stationary, then $\pi(0) = \pi(0)(1-q) + \pi(1)(1-q)$, which we can rearrange as

$$\pi(1) = \beta\pi(0), \qquad \beta = \frac{q}{1-q}.$$

For $i \geq 1$, we have $\pi(i) = \pi(i-1)q + \pi(i+1)(1-q)$, which we can rearrange in the standard way as

$$\pi(i+1) - \pi(i) = \beta[\pi(i) - \pi(i-1)].$$

It is easy to verify that both these conditions are true of $\pi(i) = c\beta^i$ for some constant c. Of course, we want π to be a distribution, so $\pi(i) \geq 0$ and $\sum \pi(i) = 1$. The latter is only possible if $\beta < 1$, in which case

$$\sum_{i=0}^{\infty} c\beta^i = \frac{c}{1-\beta} = \frac{c}{1 - \frac{q}{1-q}} = c\frac{1-q}{1-2q}.$$

Thus, when $0 < q < \frac{1}{2}$ (so that $0 < \beta < 1$, if we take $c = \frac{1-2q}{1-q}$, we find that $\pi(i) = \frac{1-2q}{1-q}\left(\frac{q}{1-q}\right)^i$ is a stationary distribution for the chain. By Corollary 11.1, it follows that all states are positive recurrent in this case.

Now, when $q = \frac{1}{2}$, the situation is very similar to the SRW on $\mathbb{Z}$: we have $\pi(i+1) - \pi(i) = \pi(i) - \pi(i-1)$ for all $i \geq 1$, and so unless $\pi(1) - \pi(0) = 0$ the sequence π grows without bound and cannot be a distribution; if $\pi(1) = \pi(0)$, it follows that $\pi(i) = \pi(0)$ for all i and again π cannot be summable unless it is constantly 0 and so not a distribution. Thus the chain does not possess an invariant distirbution when $q = \frac{1}{2}$, and so no states are positive recurrent. On the other hand, conditioned on the event $X_1 = 1$ the chain behaves exactly as a SRW on $\mathbb{Z}$ until time T_0, and since SRW is recurrent, it follows that this chain is recurrent. Hence, all its states are null recurrent.

Finally, suppose $q < \frac{1}{2}$. The same first step analysis we did in Example 2.6 (biased SRW) shows that

$$\mathbb{P}_i(T_N < T_0) = \frac{1 - \left(\frac{1-q}{q}\right)}{1 - \left(\frac{1-q}{q}\right)^N}$$

which shows that $\lim_{N\to\infty} \mathbb{P}_1(T_N < T_0) = 1 - \frac{1-q}{q} = \frac{2q-1}{q} > 0$ in this case. Hence, there is a positive probability that the chain visits all sufficiently large N before visiting 0, so 0 is (and hence all states are) transient.

11.1. **Average Convergence to the Stationary Distribution: The Ergodic Theorem.** We have now seen that (in the irreducible case) the existence of a stationary distribution is equivalent to the existence of a positive recurrent state (which is the same as all states being positive recurrent). Even in the finite-state case, however, we saw examples (cf. Example 6.3(3), SRW on an even-length cycle) where a stationary distribution exists, but the chain does not converge to it; indeed, periodicity plays a role too. Nevertheless, there is still always a sense in which a chain posessing a stationary distribution converges to it: in terms of **running average**.

Theorem 11.3 (Ergodic Theorem)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible recurrent Markov chain with state space S. Let $j \in S$ be fixed, and define*

$$\pi(j) = \frac{1}{\mathbb{E}_j[T_j]}$$

(with the understanding that $\pi(j) = 0$ if j is null recurrent, i.e. if $\mathbb{E}_j[T_j] = \infty$). Let $V_n(j) = \sum_{m=1}^{n} \mathbb{1}_{\{X_m=j\}}$ be the number of visits to state j up to time n. Then for any state i,

$$\mathbb{P}_i\left(\lim_{n\to\infty} \frac{V_n(j)}{n} = \pi(j)\right) = 1$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} p_m(i,j) = \pi(j).$$

Remark 11.4. In general, the claim $\frac{1}{n}\sum_{m=1}^{n} p_m \to p$ is weaker than the claim $p_n \to p$. The exception is when $p_n \geq 0$ and $p = 0$; in this case, the average converging to 0 implies that the sequence converges to 0 as the reader should check. Hence, the theorem has the interesting corollary that, in a null recurrent irreducible chain, $p_n(i,j) \to 0$ for all i, j. This is born out, for example, in SRW on $\mathbb{Z}^1$ and $\mathbb{Z}^2$.

Proof. The proof is quite similar to the proof of Proposition 10.2(2). Since j is a recurrent state, we know that $V_n(j) \to \infty$ as $n \to \infty$ with $\mathbb{P}_i$-probability 1. It follows that $\frac{V_n(j)+1}{V_n(j)} \to 1$ with $\mathbb{P}_i$-probability 1. Now, since there were $V_n(j)$ visits to j by time n, the $V_n(j)$th visit to j occured at a time $\leq n$. Using the times T_j^k from the proof of Proposition 10.2, this means

$$T_j^{V_n(j)} \leq n$$

with $\mathbb{P}_i$-probability 1. On the other hand, the $(V_n(j)+1)$st visit to j is $\geq n$, meaning that

$$T_j^{V_n(j)+1} \geq n$$

with $\mathbb{P}_i$-probability 1. Hence $T_j^{V_n(j)} \leq n \leq T_j^{V_n(j)+1}$. Dividing through by $V_n(j)$, we have

$$\frac{T_j^{V_n(j)}}{V_n(j)} \leq \frac{n}{V_n(j)} \leq \frac{T_j^{V_n(j)+1}}{V_n(j)+1} \frac{V_n(j)+1}{V_n(j)}$$

In the proof of Proposition 10.2, we shows that $T_j^k/k \to \mathbb{E}_i[T_j]$ with $\mathbb{P}_i$-probability 1 (using the Strong Law of Large Numbers). Hence, by the Squeeze Theorem, $\frac{n}{V_n(j)} \to \mathbb{E}[T_j]$ as $n \to \infty$ with $\mathbb{P}_i$-probability 1. Taking reciprocals proves the first equation.

For the second equality, note that since $\lim_{n\to\infty} \frac{V_n(j)}{n} = \pi(j)$ with $\mathbb{P}_i$-probability 1, it follows that $\mathbb{E}_i[\frac{V_n(j)}{n}] \to \mathbb{E}_i[\pi(j)] = \pi(j)$ as $n \to \infty$. The proof is then concluded by noting that

$$\mathbb{E}_i\left[\frac{V_n(j)}{n}\right] = \frac{1}{n}\mathbb{E}_i[\sum_{m=1}^{n} \mathbb{1}_{\{X_m=j\}}] = \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}_i[1_{\{X_m=j\}}] = \frac{1}{n}\sum_{m=1}^{n} \mathbb{P}_i(X_m = j) = \frac{1}{n}\sum_{m=1}^{n} p_m(i,j).$$

□

12. LECTURE 12: APRIL 27, 2012

12.1. **The Convergence Theorem.** In order to have genuine convergence to the stationary distribution, we know (from the finite-state case, cf. Theorem 7.4) that aperiodicity is also required. As in the finite-state case, it is also sufficient.

Theorem 12.1. *Let $(X_n)_{n\geq 0}$ be an irreducible, aperiodic Markov chain possessing a stationary distribution π. Then for any states i, j, $\lim_{n\to\infty} p_n(i,j) = \pi(j)$.*

Remark 12.2. (1) Since limits of real number sequences are unique, it follows that an irreducible, aperiodic Markov chain has *at most one* stationary distribution (i.e. if a stationary distribution exists, then it is unique). Moreover, from Theorem 11.3 (in fact from Corollary 10.1, it follows that this stationary distribution is given by $\pi(i) = \frac{1}{\mathbb{E}_i[T_i]}$.

(2) It is a fact, which we won't prove however, that the uniqueness of the stationary distribution does not require aperiodicity. Any irreducible Markov chain possesses at most one stationary distribution; aperiodicity is only required to guarantee convergence toward the stationary distribution.

Example 12.3. Consider SRW on a 5-cycle. This graph is connected so the SRW is irreducible. Also, because there are an odd number of vertices, there are path of odd length from any state i to itself (going all around the cycle), and so $d(i) = 1$. Hence, this is an irreducible, aperiodic, finite-state Markov chain, and so possesses an invariant distribution to which the process converges in distribution. It is easy to check that the invariant distribution is $\pi(i) = \frac{1}{5}$ for each i. This immediately shows us that $\mathbb{E}_i[T_i] = 5$ for each i. Even more, we note from Theorem 9.6 that $\pi(j) = \frac{\gamma(i,j)}{\mathbb{E}_i[T_i]} = \frac{1}{5}\gamma(i,j)$ for all pairs i, j. So we can quickly calculate that the expected number of visits to any state j before returning to i satisfies $\frac{1}{5} = \frac{1}{5}\gamma(i,j)$; in other words, $\gamma(i,j) = 1$.

Proof of Theorem 12.1. Consider the conditioned Markov chain $(X_n^i)_{n\geq 0}$ where $X_0^i = i$, and $(X_n^i)_{n\geq 0}$ has the same transition probabilities $p(i,j)$ as X. Then $p_n(i,j) = \mathbb{P}_i(X_n = j) = \mathbb{P}(X_n^i = j)$. Similarly, let $(X_n^\pi)_{n\geq 0}$ be a Markov chain **independent** of $(X_n^i)_{n\geq 0}$, which has initial distribution π: $\mathbb{P}(X_0 = i) = \pi(i)$, and also has transition probabilities $p(i,j)$. Since π is stationary, it follows that $\mathbb{P}(X_n^\pi = i) = \pi(i)$ for all n.

Let $T = \min\{n \in \mathbb{N}\colon X_n^i = X_n^\pi\}$ (clearly T is a stopping time). Define a new process $(Y_n)_{n\geq 0}$ by

$$Y_n = \begin{cases} X_n^i, & n \leq T \\ X_n^\pi, & n > T \end{cases}$$

Since T is a stopping time, (Y_n) is a Markov chain with transition probabilities $p(i,j)$, which satisfies $Y_0 = i$; thus, $\mathbb{P}(Y_n = j) = p_n(i,j)$. Now, for any j, if $Y_n = j$ then either $X_n^\pi = j$ or $X_n^\pi \neq Y_n$; in other words $\{Y_n = j\} \subseteq \{X_n^\pi = j\} \cup \{X_n^\pi \neq Y_n\}$, and so

$$\mathbb{P}(Y_n = j) \leq \mathbb{P}(X_n^\pi = j) + \mathbb{P}(X_n^\pi \neq Y_n).$$

Similarly, $\{X_n^\pi = j\} \subseteq \{Y_n = j\} \cup \{X_n^\pi \neq Y_n\}$ and so

$$\mathbb{P}(X_n^\pi = j) \leq \mathbb{P}(Y_n = j) + \mathbb{P}(X_n^\pi \neq Y_n).$$

Together, these say that $|\mathbb{P}(Y_n = j) - \mathbb{P}(X_n^\pi = j)| \leq \mathbb{P}(Y_n \neq X_n^\pi)$. It follows that we can estimate

$$|p_n(i,j) - \pi(i)| = |\mathbb{P}(Y_n = j) - \mathbb{P}(X_n^\pi = j)| \leq \mathbb{P}(Y_n \neq X_n^\pi) = \mathbb{P}(T > n).$$

Hence, to prove the theorem, it suffices to prove that $\lim_{n\to\infty} \mathbb{P}(T > n) = 0$.

Consider the Markov chain $\mathbf{Z}_n = (X_n^i, X_n^\pi)$ (whose state space is $S \times S$).

Claim. For any state $j \in S$, $\mathbb{P}(\mathbf{Z}_n = (j, j) \text{ for some } n) = 1$.

Once the Claim is proved, it follows immediately that $\mathbb{P}(T < \infty) = 1$, which implies $\mathbb{P}(T > n)$ tends to 0 as $n \to \infty$, as required to prove the theorem. To prove the Claim, we refer to Exercise 1 on Homework 3, which states that if $(Z_n)_{n=0}^\infty$ is an irreducible, recurrent Markov chain, then $\mathbb{P}_i(Z_n = j \text{ for some } n) = 1$ for all states i, j. As such, it suffices to show that $\mathbf{Z}_n$ is irreducible and recurrent. In fact, we will show it is irreducible and positive recurrent, by demonstrating (for the latter) that it possesses an invariant distribution.

The key observation is that, since X_n^i and X_n^π are independent by construction, it follows that the transitions probabilities $q\left((i_1, i_2), (j_1, j_2)\right)$ for the chain $\mathbf{Z}_n$ are

$$q\left((i_1, i_2), (j_1, j_2)\right) = p(i_1, j_1)p(i_2, j_2).$$

- $\mathbf{Z}_n$ *is irreducible*. Let (i_1, i_2) and (j_1, j_2) be any two states in $S \times S$. By assumption, both X_n is irreducible, and so there are times m_1 and m_2 where $p_{m_1}(i_1, j_1) > 0$ and $p_{m_2}(i_2, j_2) > 0$. Since X_n is aperiodic, we know (cf. the remark preceeding the proof of Theorem 7.4) we know $p_n(j_1, j_1) > 0$ and $p_n(j_2, j_2) > 0$ for all sufficiently large n; in particular, for all sufficiently large k we have $p_{m_1+k}(j_1, j_1) > 0$ and $p_{m_2+k}(j_2, j_2) > 0$. Thus

$$\begin{aligned} q_{m_1+m_2+k}\left((i_1, i_2), (j_1, j_2)\right) &= p_{m_1+m_2+k}(i_1, j_1)p_{m_1+m_2+k}(i_2, j_2) \\ &\geq [p_{m_1}(i_1, j_1)p_{m_2+k}(j_1, j_1)][p_{m_2}(i_2, j_2)p_{m_1+k}(j_2, j_2)] > 0. \end{aligned}$$

- $\mathbf{Z}_n$ *positive recurrent*. By Corollary 11.1, it suffices to show there exists a stationary distribution for $\mathbf{Z}_n$. Define $\theta(i, j) = \pi(i)\pi(j)$ for $(i, j) \in S \times S$. Then

$$\begin{aligned} \sum_{(k,\ell)\in S\times S} \theta(k, \ell) \cdot q\left((k, \ell), (i, j)\right) &= \sum_{(k,\ell)\in S\times S} \pi(k)\pi(\ell)p(k, i)p(\ell, j) \\ &= \left[\sum_{k\in S} \pi(k)p(k, i)\right]\left[\sum_{\ell\in S} \pi(\ell)p(\ell, j)\right] \\ &= \pi(i)\pi(j) = \theta(i, j) \end{aligned}$$

  because π is stationary for X_n. Thus, θ is stationary for $\mathbf{Z}_n$.

□

13. Lecture 13: April 30, 2012

13.1. **Time Reversal.** Let $(X_n)_{n\geq 0}$ be a Markov chain. For a fixed time N, by definition X_{N+1} is conditionally independent of $X_0, X_1, \ldots, X_{N-1}$, conditioned on X_N. In other words, the past and the future are independent, conditioned on the present. Stated in this symmetric manner, it is natural to think of a Markov chain running *backward* in time, and expect this process to be another Markov chain. In fact, it is sometimes the *same* Markov chain.

Example 13.1. Let $(X_n)_{n\geq 0}$ be SRW on $\mathbb{Z}$. Then for any trajectory $(i_0, i_1, \ldots, i_N)$ for the first N steps, we have $\mathbb{P}(X_0 = i_0, \ldots, X_N = i_N) = 0$ unless $i_{n+1} = i_n \pm 1$ for each n; in this case, each such trajectory has probability 2^{-N}.

Now, let $Y_n = X_{N-n}$ for $0 \leq n \leq N$. Then for any trajectory $j_0, \ldots, j_N$,

$$\mathbb{P}(Y_0 = j_0, \ldots, Y_N = j_N) = \mathbb{P}(X_0 = j_N, X_1 = j_{N_1}, \ldots, X_N = j_0).$$

The condition $j_{n+1} = j_n \pm 1$ is equivalent to $j_{n-1} = j_n \pm 1$, and so we once again have the joint distribution of $(Y_0, \ldots, Y_N)$ is supported on the same trajectories as that of $(X_0, \ldots, X_N)$, and each such trajectory has probability 2^{-N}, the same as for $(X_0, \ldots, X_N)$.

Example 13.1 gives an appropriate sense we can give to the statement "the chain looks the same forwards and backwards": we consider the joint distirbution of $(X_0, \ldots, X_N)$ and $(X_N, X_{N-1}, \ldots, X_0)$. In this case, they are both the same. This will not always occur; but what is true is that, started/ended in the right distirbution, the backward process will still be a Markov chain. We have to be careful, though. If we (as is typical) condition the process $(X_n)_{n\geq 0}$ to start in a particular state, $X_0 = i$, we cannot expect the reversed process Y_n to be Markovian; for example, if X_n is irreducible and positive recurrent, then X_N will be close to stationarity for large N, not concentrated near the state i.

The best situation where we can hope for the reversed process to exhibit Markovian behavior is when it starts (and therefore stays) in its stationary state.

Theorem 13.2. *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain possessing a stationary distribution π. Let $N \in \mathbb{N}$, and for $0 \leq n \leq N$ define $Y_n = X_{N-n}$. Then $(Y_n)_{0\leq n\leq N}$ is an irreducible Markov chain with the same stationary distribution, and transition probabilities $q(i,j)$ given by*

$$q(i,j) = \frac{\pi(j)}{\pi(i)} p(j,i). \tag{13.1}$$

Remark 13.3. (1) Recall from Corollary 10.2 that the stationary distribution has strictly positive entries; hence there is no trouble dividing by $\pi(i)$ in Equation 13.1.

(2) Note that $q(i,j) \geq 0$ from Equation 13.1, and for fixed i

$$\sum_j q(i,j) = \frac{1}{\pi(i)} \sum_j \pi(j)p(j,i) = \frac{1}{\pi(i)} \cdot \pi(i) = 1$$

where the penultimate equality follows from the stationarity of π for p. Hence q is still stochastic, and so indeed gives the transition probabilities of *some* Markov chain.

Proof. First let's check that π is still stationary for q:

$$\sum_i \pi(i)q(i,j) = \sum_i \pi(i) \cdot \frac{\pi(j)}{\pi(i)} p(j,i) = \sum_i \pi(j)p(j,i) = \pi(j) \sum_i p(j,i) = \pi(i) \cdot 1 = \pi(j)$$

where the penultimate equality is just the statement that p is stochastic. Thus, π is q-stationary as well as p-stationary.

Now, for a given trajectory $i_0, \ldots, i_N$, we have

$$\begin{aligned}\mathbb{P}(Y_0 = i_0, Y_1 = i_1, \ldots, Y_N = i_N) &= \mathbb{P}(X_0 = i_N, X_1 = i_{N-1}, \ldots, X_N = i_0) \\ &= \mathbb{P}(X_0 = i_N)p(i_N, i_{N-1}) \cdots p(i_2, i_1)p(i_1, i_0) \\ &= \pi(i_N)p(i_N, i_{N-1}) \cdots p(i_2, i_1)p(i_1, i_0)\end{aligned}$$

by the Markov property and the assumption that X_n is in the stationary distribution. If we rewrite Equation 13.1 in the form

$$\pi(j)p(j, i) = q(i, j)\pi(i)$$

then we can proceed by induction and write

$$\begin{aligned}&\pi(i_N)p(i_N, i_{N-1})p(i_{N-1}, i_{N-2}) \cdots p(i_2, i_1)p(i_1, i_0) \\ &= q(i_{N-1}, i_N)\pi(i_{N-1})p(i_{N-1}, i_{N-2}) \cdots p(i_2, i_1)p(i_1, i_0) \\ &= q(i_{N-1}, i_N)q(i_{N-2}, i_{N-1})\pi(i_{N-2}) \cdots p(i_2, i_1)p(i_1, i_0) \\ &= q(i_{N-1}, i_N)q(i_{N-2}, i_{N-1}) \cdots q(i_1, i_2)\pi(i_1)p(i_1, i_0) \\ &= q(i_{N-1}, i_N)q(i_{N-2}, i_{N-1}) \cdots q(i_1, i_2)q(i_0, i_1)\pi(i_0).\end{aligned}$$

Thus, we have

$$\mathbb{P}(Y_0 = i_0, Y_1 = i_1, \ldots, Y_N = i_N) = \pi(i_0)q(i_0, i_1)q(i_1, i_2) \cdots q(i_{N-1}, i_N)$$

which is precisely to say that $(Y_n)_{0 \leq n \leq N}$ is a Markov chain.

Finally, since p is irreducible, given states i, j there is some finite n so that $p_n(i, j) > 0$. By the definition of matrix multiplication,

$$p_n(i, j) = \sum_{i_1, \ldots, i_{n-1}} p(i, i_1)p(i_1, i_2) \cdots p(i_{n-1}, j)$$

and since all these terms are ≥ 0 and $p_n(i, j) > 0$ it must be that there is at least one trajectory $i, i_1, \ldots, i_{n-1}, j$ such that $p(i, i_1)p(i_1, j_2) \cdots p(i_{n-1}, j) > 0$. But then

$$\begin{aligned}q(j, i_{n-1})q(i_{n-1}, i_{n-2}) \cdots q(i_1, i) &= \frac{\pi(i_{n-1})}{\pi(j)}p(i_{n-1}, j) \cdot \frac{\pi(i_{n-2})}{\pi(i_{n-1})}p(i_{n-2}, i_{n-1}) \cdots \frac{\pi(i)}{\pi(i_1)}p(i, i_1) \\ &= \frac{\pi(i)}{\pi(j)}p(i, i_1)p(i_1, i_2) \cdots p(i_{n-1}, j) > 0.\end{aligned}$$

Thus, at least one term in the matrix multiplication sum for $q_n(i, j)$ is > 0, so $q_n(i, j) > 0$. The symmetric argument for $q_n(j, i)$ shows that q is irreducible, as claimed. □

Remark 13.4. The final step in the above proof highlights a(n obvious) alternate definition of irreducibility of a Markov chain, or more precisely of communication classes. Say j *can be reached from* i if there is a finite trajectory $i_1, i_2, \ldots, i_{n-1}$ so that $p(i, i_1)$, $p(i_1, i_2)$, …, $p(i_{n-1}, j)$ are all > 0. Then say i and j can communicate if j can be reached from i and i can be reached from j. The above proof shows that this is equivalent to the usual definition of communication classes.

13.2. **Reversible Markov Chains.** Theorem 13.2 shows that a time reversed Markov chain (in stationary distribution) is still a Markov chain, but with a potentially different transition matrix. It is natural to ask when it happens that the reversed chain is actually the same as the original chain.

Definition 13.5. *Let $(X_n)_{n\geq 0}$ be a (time-homogeneous) Markov chain, with countable state space S and transition probabilities $p(i,j)$ for $i,j \in S$. Call the chain* **reversible** *if there is a function $\pi\colon S \to [0,\infty)$, not identically 0, which verifies the* detailed balance condition*:*

$$\pi(i)p(i,j) = \pi(j)p(j,i), \quad \forall\, i,j \in S. \tag{13.2}$$

The detailed balance condition is invariant under scaling – that is, if $\pi(i)p(i,j) = \pi(j)p(j,i)$ then $[\lambda\pi(i)]p(i,j) = [\lambda\pi(j)]p(j,i)$ as well for any $\lambda > 0$. Thus, if the function π in Definition 13.5 is summable, then we can scale it to be a distribution. It is not hard to see that

Corollary 13.6. *If $(X_n)_{n\geq 0}$ is a reversible Markov chain that possesses a unique stationary distribution (i.e. if it is irreducible and positive recurrent), then that stationary distribution witnesses the local balance condition.*

Proof. We need only show that π is stationary for p; this follows from the same argument in the proof of Theorem 13.2:

$$\sum_{i\in S} \pi(i)p(i,j) = \sum_{i\in S} \pi(j)p(j,i) = \pi(j)\sum_{i\in S} p(j,i) = \pi(j)\cdot 1 = \pi(j).$$

□

It is possible, however, for a π verifying the detailed balance condition to exist even if a stationary distribution does not exist (cf. Example 13.1).

Example 13.7. Suppose the Markov chain is *symmetric*: the transition probabilities satisfy $p(i,j) = p(j,i)$. In this case, trivially the constant vector $\pi = (c,c,\ldots,c)$ witnesses the local balance condition for any c. If there are a finite number of states, N, it follows that $\pi(i) = \frac{1}{N}$ for each N is a stationary distribution, and the chain is reversible. If there are infinitely many states, there may not be any distribution that witnesses the local balance condition. For example, take SRW on $\mathbb{Z}$, whose transition probabilities are symmetric: $p(i,j) = p(j,i)$ has value $\frac{1}{2}$ if $|i-j| = 1$ and 0 otherwise. Here any constant $\pi = (c,c,c,\ldots)$ witnesses the local banace condition, and it is easy to see that *only* constants do. (Indeed, we know there is no stationary distribution.)

Perhaps the most important reason reversible processes are nice is that the detailed balance condition provides a much easier way to find the stationary distribution than the definition $\pi = \pi\mathbf{P}$. Here is an important class of examples, cf. Definition 5.2.

Example 13.8. Let G be a countable graph with no isolated vertices. The SRW on G has transition probabilities $p(i,j) = \frac{1}{v_i}$ where v_i is the valence of i (which is ≥ 1 since i is not isolated) if $i \sim j$, and $p(i,j) = 0$ if $i \not\sim j$. Then we can check that $\pi(i) = v_i$ satisfies

$$\pi(i)p(i,j) = \pi(j)p(j,i) = \begin{cases} 1, & i \sim j \\ 0, & i \not\sim j \end{cases}$$

so the "valence vector" π satisfies the local balance condition and SRW on graphs are reversible. In particular, this shows that if the graph is finite, there exists a stationary distribution: just normalize $\pi(i) = v_i / \sum_j v_j$.

Example 13.9. Capitalizing on the previous example, consider the following problem. *A Kinght starts at a corner of a standard 8×8 chessboard, and on each step moves at random. How long, on average, does it take to return to its starting position?*

If i is a corner, the question is what is $\mathbb{E}_i[T_i]$, where the Markov process in question is the one desribed in the problem: it is SRW on a graph whose vertices are the 64 squares of the chessboard, and two squares are connected in the graph if a Knight can move from one to the other. (Note: if the Knight can move from i to j, it can do so in exactly two ways: 1 up/down and 2 left/right, or 2 left/right and 1 up/down. So choosing uniformly among *moves* or among *positions* amounts to the same thing.) This graph is connected (as a little thought will show), and so the SRW is irreducible; therefore there is a unique stationary distribution. By Example 13.7, the stationary distribution π is given by $\pi(i) = v_i / \sum_j v_j$, and so by the Ergodic Theorem, $\mathbb{E}_i[T_i] = \frac{1}{\pi(i)} = \sum_j v_j / v_i$.

A Knight moves 2 units on one direction and 1 unit in the other direction. Starting at a corner $(1, 1)$, the Knight can only more to the positions $(3, 2)$ and $(2, 3)$, so $v_i = 2$. To solve the problem, we need to calculate v_j for all starting positions j on the board. By square symmetry, we need only calculate the numbers for the upper 4×4 grid. An easy calculation then gives these valences as

$$\begin{bmatrix} 2 & 3 & 4 & 4 \\ 3 & 4 & 6 & 6 \\ 4 & 6 & 8 & 8 \\ 4 & 6 & 8 & 8 \end{bmatrix}$$

The sum is 84, and so the sum over the full chess board is $4 \cdot 84 = 336$. Thus, the number of expected steps for the Knight to return to the corner is $\frac{1}{2} \cdot 336 = 168$.

14. LECTURE 14: MAY 2, 2012

One of the main uses of Markov chains is sampling (approximately) from probability distributions that are computationally intractable.

Example 14.1. A **Hard Core Configuration** on an $N \times N$ grid, HCC_N, is a function

$$c\colon \{1,\dots,N\}^2 \to \{0,1\}$$

such that if $c(i,j) = 1$ then the four values $c(i \pm 1, j \pm 1)$ are 0. (I.e. there are no adjacent vertices in the square lattice both having value 1.) This is a simple model of the configuration of an ideal gas (the 1s represent gas molecules, the 0s empty space). We would like to pick a random hard-core configuration – meaning sample from the uniform distribution on HCC_N, so each c is chosen with weight $1/|HCC_N|$. The problem is that $|HCC_N|$ is unknown, even to exponential order, for large N. A simple upper-bound is 2^{N^2} (which counts the set of all configurations, functions $f\colon \{1,\dots,N\}^2 \to \{0,1\}$ without constraint). It is conjectured that $|HCC_N| \sim \beta^{N^2}$ for some $\beta \in (1,2)$, but no one even has a good guess as to what β is.

So, how do we sample from a distribution we cannot even approximate? By using the following Markov chain $(X_n)_{n\geq 0}$. The state space is $S = HCC_N$, and the transition probabilities for $c \neq c'$ are:

$$p(c,c') = \begin{cases} 1/N^2, & \text{if } c, c' \text{ differ at exactly one point in } \{1,\dots,N\}^2, \\ 0, & \text{otherwise.} \end{cases}$$

For any fixed hard-core configuration c, there are at most N^2 hard-core configurations that differ from it in 1 position, so $\sum_{c'\neq c} p(c,c') \leq 1$. We therefore define $p(c,c) = 1 - \sum_{c'\neq c} p(c,c')$, and this gives a valid Markov transition kernel.

It is actually easy to implement this chain computationally: at each step, we choose one of the N^2 points in the grid uniformly at random, and change the value of the configuration at that point if we can; if the change would violate the HCC condition, then we leave the configuration unchanged for the next step. To be precise: if $X_n = c$, choose a point (i,j) in the grid uniformly at random from all N^2; if $c(i,j) = 1$, then let $X_{n+1} = c'$ where $c'(i,j) = 0$ and $c'(k,\ell) = c(k,\ell)$ for $(k,\ell) \neq (i,j)$; if $c(i,j) = 0$ and $c(i \pm 1, j \pm 1) = 0$ then similarly set $X_{n+1} = c'$ where $c'(i,j) = 1$ and $c'(k,\ell) = c(k,\ell)$ for $(k,\ell) \neq (i,j)$; if, on the other hand, $c(i,j) = 0$ but a neighboring point has $c = 1$, then set $X_{n+1} = X_n = c$. It is very easy to see that this algorithm produces a Markov chain with the above transition kernel $p(c,c')$.

Note that $p(c,c') = p(c',c)$, so the transition kernel is symmetric. It follows that the uniform distribution on the state space, π, satisfies the detailed balance condition with $p(c,c')$, and hence the uniform distirbution π is the stationary distribution for $(X_n)_{n\geq 0}$. (It is easy to see that $(X_n)_{n\geq 0}$ is irreducible; for example, to get from any c to any c', one can begin by eliminating the 1s in c one by one to get to the 0 configuration, then add new 1s one by 1 to reach c'.)

So, how do we select a uniformly random hard core configuration? Start at any configuration and run the Markov chain above. For large n, the distribution of X_n is approximately π, by the Convergence Theorem; so, for large n, X_n will be a (nearly) uniformly random hard-core configuration.

Example 14.2. Consider the problem of **graph colorings**. Let $G = (V, E)$ be a finite graph. A q**-coloring** of G (with $q \in \mathbb{N}$) is a function $f\colon V \to \{1, 2, \ldots, q\}$ with the property that, if $u, v \in V$ with $u \sim v$, then $f(u) \neq f(v)$. The set of q-colorings is very hard to count (especially for small q), so again we cannot directly sample a uniformly random q-coloring. Instead, define a Markov chain on the state space of all q-colorings f, where for $f \neq g$

$$p(f, g) = \begin{cases} \frac{1}{q|V|}, & \text{if } f \text{ and } g \text{ differ at exactly one vertex,} \\ 0, & \text{otherwise.} \end{cases}$$

Again: since there are at most q different q-colorings of G that differ at a given vertex, and there are $|V|$ vertives, we have $\sum_{f \neq g} p(f, g) \leq 1$, and so setting $p(f, f) = \sum_{g \neq f} p(f, g)$ yields a stochastic matrix p which is the transition kernel of a Markov chain. It is evidently symmetric, and by considerations like those in Example 14.1 the chain is irreducible and aperiodic (so long as q is large enough for any q-colorings to exist!); hence, the stationary distribution is uniform, and this Markov chain converges to the uniform distribution.

To simulate the corresponding Markov chain: given $X_n = f$, choose one of the $|V|$ vertices, v, uniformly at random and one of the q colors, k, uniformly at random. If the new function g defined by $g(v) = k$ and $g(w) = f(w)$ for $w \neq v$ is a q-coloring of the graph, set $X_{n+1} = g$; otherwise, keep $X_{n+1} = X_n = f$. This process has the transition probabilities listed above, and so running it for large n gives an approximately uniformly random q-coloring.

In Examples 14.1 and 14.2, we constructed Markov chains with symmetric kernels, therefore having the uniform distribution as stationary, and converging to it. This is often a good way to simulate uniform samples. But how can we sample from a non-uniform distribution?

14.1. **The Metropolis-Hastings Algorithm.** We can generalize Examples 14.1 and 14.2 to give a method for sampling *any* (strictly positive) distribution π, by building a 2-step Markov chain: one to propose moves, a second to accept/reject them. Here is the algorithm.

Let S be a finite set, and let π be a (strictly positive) distribution on S. First, construct an irreducible Markov chain on S which has transition probabilities $q(i, j) = q(j, i)$, symmetric. This chain is thus reversible, and has the uniform distribution as its stationary state; but π is not necessarily uniform. Instead, we construct a *new* Markov chain with transition kernel

$$p(i, j) = q(i, j) \min\left\{1, \frac{\pi(j)}{\pi(i)}\right\} \text{ for } i \neq j, \quad p(i, i) = 1 - \sum_{j \neq i} p(i, j).$$

Once we have constructed this chain, we run it for long time to sample from the distribution π: it is clear from the definition that p is irreducible (since q is), and

$$\pi(i)p(i, j) = q(i, j) \min\{\pi(i), \pi(j)\} = q(j, i) \min\{\pi(j), \pi(i)\} = \pi(j)p(j, i)$$

so π satisfies the detailed balance condition for p, and thus π is the stationary distribution for p. Hence, by the Convergence Theorem, the p-Markov chain converges to π in distribution for large time.

It remains to be seen how to actually simulate the Markov chain with kernel p (given that we know how to simluate the chain with kernel q). Here is the algorithm:

- Suppose we've reached $X_n = i$. For each $j \neq i$, propose the move $X_{n+1} = j$ according to the original chain; i.e. propose the move $i \to j$ with probability $q(i,j)$.
- Once the first move has been proposed (i.e. the proposed next state j has been chosen), accept this move with probability $\min\{1, \frac{\pi(j)}{\pi(i)}\}$; otherwise reject it and stay in state i. In other words: if $\pi(j) \geq \pi(j)$, set $X_{n+1} = j$; if $\pi(j) = \alpha\pi(i)$ with $0 < \alpha < 1$, then set $X_{n+1} = j$ with probability α and $X_{n+1} = i$ with probability $1 - \alpha$.

It is an easy exercise to show that the transition kernel p above is the result of this two-step Markov chain algorithm. Hence, it may be used to simluate any distribution π (approximately, for large run time). Note: the indeterminacy of q is a *benefit* here. It means we are free to choose the initial (symmetric) chain to fit the symmetries of the given state space.

There are many variations on the Metropolis-Hastings algorithm, all with the same basic ingredients and goal: to use a (collection of correlated) Markov chain(s), designed to have a given stationary distribution π, to sample from π approximately. They are collectively known as **Markov chain Monte Carlo** simulation methods. One place they are especially effective is in calculating high-dimensional integrals (i.e. expectations), a common problem in multivariate statistics.

14.2. **A word on convergence rates.** In order for the Metropolis-Hastings algorithm (or other MCMC methods) to be effective, the Markov chain designed to converge to the given distribution must converge reasonably fast. The term used to measure this rate is the **mixing time** of the chain. (To make this precise, we need a precise measure of closeness to the stationary distribution; then we can ask how long before the distance to stationarity is less than $\epsilon > 0$.) This is a topic of a great deal of current research, with literally *thousands* of papers published. We will not even scratch the surface here. Let's just point out the connection to eigenvalues.

Suppose our Markov chain has N states, is irreducible and aperiodic (as Homework 3, Problem 3 points out, every finite-state Markov chain is arbitrarily close to one). Let's suppose further that the chain is *symmetric*; i.e. the transition matrix $\mathbf{P} = \mathbf{P}^\top$. Then by the spectral theorem from linear algebra,

$$\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

where $\mathbf{U}$, the eigenvector matrix (columns are normalized eigenvalues) is an *orthogonal* matrix (the columns are orthogonal to each other), and

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{bmatrix}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ are the eigenvalues of $\mathbf{P}$. We also know, from the Perron-Frobenius theorem (Corollary 6.6) and the assumption of irreducibility and aperiodicity,

$\lambda_1 = 1$ and for $j \geq 2$ $|\lambda_j| < 1$. (I.e. the largest magnitude eigenvalue is 1 and all others are strictly smaller.) So $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N > -1$. Hence

$$\lim_{n\to\infty} \mathbf{P}^n = \lim_{n\to\infty} \mathbf{U}\mathbf{D}^n\mathbf{U}^\top = \mathbf{U} \lim_{n\to\infty} \begin{bmatrix} 1 & & & \\ & \lambda_2^n & & \\ & & \ddots & \\ & & & \lambda_N^n \end{bmatrix} \mathbf{U}^\top = \mathbf{U} \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \mathbf{U}^\top.$$

Moreover, we see that this convergence is exponentially-fast: setting $\mathbf{P}^\infty$ equal to the limit, we have

$$\|\mathbf{P}^n - \mathbf{P}^\infty\| = \left\| \begin{bmatrix} 1 & & & \\ & \lambda_2^n & & \\ & & \ddots & \\ & & & \lambda_N^n \end{bmatrix} - \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \right\| = \max_{2\leq j\leq N} (1-\lambda_j^n) = \max\{1-\lambda_2^n, 1-\lambda_N^n\}.$$

So, measuring in terms of matrix norm at least, the mixing rate is controlled by the size of $1 - \max\{\lambda_2, |\lambda_N|\}$; if this is big, then mixing is fast; if it is small, mixing is slow. For example:

$$\begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix} \qquad \lambda_1 = 1, \lambda_2 = 1-2\epsilon, \quad \textit{slow mixing}$$

$$\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \qquad \lambda_1 = 1, \lambda_2 = 0, \quad \textit{fast mixing}.$$

Much of the techonology for studying mixing times thus amounts to estimating the largest magnitude eigenvalues of (*huge*) stochastic matrices.

15. LECTURE 15: MAY 4, 2012

Today we consider a class of countable state Markov chains where the transition probabilities, although theoretically computable, are somewhat complicated to handle. We introduce alternate methods (generating functions) to compute the desired statistics.

15.1. **Galton-Watson branching process.** Consider an evolving popluation of asexual reproducers. For example: bacteria. One can also think of the propagation (and extinction) of family names of male offspring, at least under the (traditional, unrealistic) assumption that family name passes from father to son. Relatedly, one can think about the evolution of a Y-chromosome DNA haplogroups (genes).

Let X_n denote the number of individuals in the given population in the nth generation. We make the following assumptions:

- There is a fixed distribution $p_0, p_1, p_2, \ldots$ on $\mathbb{N}$ so that each individual produces k offspring with probability p_k, and then dies.
- All individuals reproduce independently.

It is clear, then, that the number of individuals in the nth generation is determined by the individuals present in generation $n-1$ only; that is, $(X_n)_{n\geq 0}$ is a Markov chain. This chain is called a (Galton-Watson) **branching process**, since one can represent it by a "family tree" diagram. Note that if the population ever reaches 0 then it remains 0 from then on – the population has gone extinct.

We could explicitly calculate the transition probabilities of the chain as follows. Let $Y_1, Y_2, Y_3, \ldots$ be a sequence of i.i.d. random variables each with distribution $\mathbb{P}(Y_i = k) = p_k$ for $k \in \mathbb{N}$. Then

$$p(i,j) = \mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(Y_1 + \cdots + Y_i = j).$$

Since the Y_i are independent, this probability can be written as an i-fold convolution of the distribution of Y_1 evaluated at j, and written down explicitly; but the formula is quite complicated and difficult to work with. We will avoid the exact form of the transition kernel, and use other tools to analyze the branching process.

The fundamental question we want to answer is: *what is the probability that the population never goes extinct*? That is, we want to calculate (or estimate)

$$q(i) = \mathbb{P}_i(X_n = 0 \text{ for some } n)$$

where i is the initial population size. It turns out this probability is largely controlled by the mean of the offspring distribution:

$$\mu = \sum_{k=0}^{\infty} k p_k.$$

To see why, consider the expected size of the population in generation n: $\mathbb{E}_i[X_n]$. We compute it by conditioning on the population in the previous generation:

$$\mathbb{E}_i[X_n] = \sum_{k=0}^{\infty} \mathbb{P}_i(X_{n-1} = k)\mathbb{E}[X_n | X_{n-1} = k].$$

But, using the i.i.d. Y_i variables above, we have

$$\mathbb{E}_i[X_n | X_{n-1} = k] = \mathbb{E}_k[X_1] = \mathbb{E}[Y_1 + \cdots + Y_k] = k\mathbb{E}[Y_1] = k\mu.$$

Thus

$$\mathbb{E}_i[X_n] = \sum_{k=0}^{\infty} \mathbb{P}_i(X_{n-1} = k) k\mu = \mu \sum_{k=0}^{\infty} k\mathbb{P}_i(X_{n-1} = k) = \mu \mathbb{E}_i[X_{n-1}].$$

By induction, it follows that

$$\mathbb{E}_i[X_n] = \mu^n \mathbb{E}_i[X_0] = i\mu^n.$$

Now, from the following simple estimate

$$i\mu^n = \mathbb{E}_i[X_n] = \sum_{k=0}^{\infty} \mathbb{P}_i(X_n \geq k) \geq \mathbb{P}_i(X_n \geq 1)$$

we see that if $\mu < 1$ then $\lim_{n\to\infty} \mathbb{P}_i(X_n \geq 1) = 0$. It follows that in this case $q(i) = \mathbb{P}(X_n = 0 \text{ eventually}) = 1$, for any i. This motivated (part of) the following definition.

Definition 15.1. *Let $(X_n)_{n\geq 0}$ be a branching process, with offspring distirbution $p_0, p_1, p_2, \ldots$ and mean $\mu = \sum_{k=0}^{\infty} kp_k$. Call the process* **subcritical** *of $\mu < 1$; **critical** if $\mu = 1$; and **supercritical** if $\mu > 1$.*

The calculation preceding Definition 15.1 shows that a subcritical branching process has $q(i) = 1$ for all i: no matter how large the initial population, the population eventually goes extinct. For critical and supercritical branching processes, we need another tool to analyze $q(i)$. Before we proceed, let us note that, since the individuals in each generation reproduce indpendently, the branches of the process are all independent; it follows that if $q_n(i) = \mathbb{P}_i(X_n = 0)$, then $q_n(i) = [q_n(1)]^i$. Since $q(i) = \lim_{n\to\infty} q_n(i)$, it therefore suffices to calculate $q(1) \equiv q$, which we define to be the **extinction probability**. (I.e. q is the probability the population eventually goes extinct, conditioned on starting with a single individual.)

15.2. **Probability generating function.** Let X be a $\mathbb{N}$-valued random variable. Define its **probability generating function** φ_X as

$$\varphi_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} s^k \mathbb{P}(X = k).$$

This is a variant of the characteristic function (a.k.a. Fourier transform) $\chi_X(t) = \mathbb{E}[e^{itX}]$ which is more commonly used for continuous distributions. Let us point out a few facts about φ_X.

- From the power-series, since $\mathbb{P}(X = k) \leq 1$ for all k, it follows that φ_X is analytic on $(-1, 1)$. We have $\varphi_X(1) = 1$, and $\varphi_X^{(n)}(0) = n!\,\mathbb{P}(X = n)$.
- For $|s| < 1$, we have $\varphi_X'(s) = \sum_{k=1}^{\infty} ks^{k-1}\mathbb{P}(X = k)$. If $\mathbb{E}[X] < \infty$, it follows that this series converges at $s = 1$, and $\varphi_X'(1) = \mathbb{E}[X]$.
- Because $\mathbb{P}(X = k) \geq 0$, the function $s \mapsto \varphi_X'(s)$ is nondecreasing; if $\mathbb{P}(X \geq 2) > 0$, it is strictly increasing.

Theorem 15.2. *Let $(X_n)_{n\geq 0}$ be a branching process with offspring distribution $p_0, p_1, p_2, \ldots$ Let φ be the probability generating function of this distribution $\varphi(s) = \sum_{k=0}^{\infty} p_k s^k$. Then the extinction probability q is given by*

$$q = \min\{s \in [0, 1] \colon \varphi(s) = s\}.$$

Proof. As noted above, $q(i) = q(1)^i = q^i$. Then by first step conditioning,

$$\begin{aligned} q = \mathbb{P}_1(X_n = 0 \text{ eventually}) &= \sum_{i=0}^{\infty} \mathbb{P}_1(X_n = 0 \text{ eventually} | X_1 = i)\mathbb{P}_1(X_1 = i) \\ &= \sum_{k=0}^{\infty} \mathbb{P}_i(X_n = 0 \text{ eventually}) p_i \\ &= \sum_{i=0}^{\infty} p_i q^i = \varphi(q) \end{aligned}$$

so q is a fixed point of φ. Of course, 1 is always a fixed point of φ; it remains to show that *if* there is a smaller fixed point, q takes it as its value.

Define $\hat{q} = \min\{s \in [0,1] \colon \varphi(s) = s\}$. Since $q_n \equiv q_n(1) = \mathbb{P}_1(X_n = 0)$ have limit $\lim_{n\to\infty} q_n = q$, it suffices to show that $q_n \leq \hat{q}$ for all n (then the limit q must be $\leq \hat{q}$, and therefore equal to $\hat{q}$ since q is a fixed point in $[0,1]$). Clearly $q_0 = 0 \leq \hat{q}$. Suppose we have shown that $q_{n-1} \leq \hat{q}$. Then

$$q_n = \mathbb{P}_1(X_n = 0) = \sum_{i=0}^{\infty} \mathbb{P}_1(X_n = 0 | X_1 = i)\mathbb{P}(X_1 = i) = \sum_{k=0}^{\infty} p_i q_{n-1}^i$$

by the independence of the i offspring in gneeration $n-1$. Thus, by the inductive hypothesis,

$$q_n = \sum_{k=0}^{\infty} p_i q_{n-1}^i \leq \sum_{k=0}^{\infty} p_i \hat{q}^i = \varphi(\hat{q}) = \hat{q},$$

concluding the proof. □

Corollary 15.3. *Suppose $p_1 \neq 1$. Then $q = 1$ if the process is critical or subcritical, and $q < 1$ if the process is supercritical.*

Note: if $p_1 = 1$ (so that $p_k = 0$ for $k \neq 1$) the population remains 1 for all time, the boring case. Excluding this deterministic situation, we know exactly when there is a chance that the population will survive in the limit: precisely when it is supercritical.

Proof. We have already seen that, in the subcritical case, $q = 1$. If the process is supercritical, then $1 < \mu = \varphi'(1)$. On the other hand we know that $\varphi'(0) = \mathbb{P}(Y = 1) = p_1 < 1$. Now, consider the function $f(s) = \varphi(s) - s$. Then $f(1) = 0$ and $f(0) = p_0 \geq 0$; we also have $f'(s) = \varphi'(s) - 1$ so $f'(1) > 0$. Since f is smooth, it follows that f is strictly increasing near 1, so $f(s) < 0$ for s near 1. Since f is continuous and $f(0) \geq 0$, there must be some point $s < 1$ so that $f(s) = 0$. Hence the minimal fixed point $\hat{q}$ of φ is in $(0,1)$, and so $q < 1$.

Now, suppose the process is critical: so $\mu = 1$ (in particular $p_0 \neq 1$). Since $p_1 \neq 0$, we have $\sum_{k=2}^{\infty} p_k > 0$. Thus $\varphi'(1) = 1$ and $\varphi'(t) < 1$ for all $t < 1$. Thus, if $s < 1$ then

$$1 - \varphi(s) = \varphi(1) - \varphi(s) = \int_s^1 \varphi'(t)\, dt < 1 - s$$

so $\varphi(s) > s$ for all $s < 1$. It follows that $q = \hat{q} = 1$. □

16. LECTURE 16: MAY 7, 2012

We are about to discuss a class of stochastic models known as **Hidden Markov models**. To begin, consider the following motivating example.

Example 16.1 (Occasionally Dishonest Casino)**.** In a casino game, a die is thrown. Usually a fair die is used, but sometimes the casino swaps it out for a loaded die. To be precise: there are two dice, F (fair) and L (loaded). For the fair die F, the probability of rolling i is $\frac{1}{6}$ for $i = 1, 2, \ldots, 6$. For the loaded die, however, 1 is rolled with probability 0.5 while 2 through 5 are each rolled with probability 0.1. Each roll, the casino randomly switches out F for L with probability 0.05; if L is in use, they switch back to F next roll with probability 0.9.

You notice that 1 is rolled 6 times in a row. How likely is it that the fair die was in use for those rolls (given the above information)? How likely is it the loaded die was used in rolls 2, 3, and 5? More generally, what is the *most likely* sequence of dice that was used to generate this sequence of rolls?

16.1. **Hidden Markov Models.** Example 16.1 is a very typical situation that occurs in many real-world problems. To model it, we consider it as two processes: there is a process of observations (the rolls of the die), and an underlying *hidden* process (which die is chosen). We would like to "decode" the hidden process, but all the information we have are the conditional probabilities of the observed process given the hidden process.

Definition 16.2. *A* Hidden Markov Model *(HMM) is a pair of stochastic processes, $(X_n, Y_n)_{n\geq 0}$, where $(Y_n)_{n\geq 0}$ is a Markov chain with state space S, and $(X_n)_{n\geq 0}$ has a possibly different state space R, and the vector-valued process $\mathbf{Z}_n = (X_n, Y_n)$ is a Markov chain. For $y \in S$ and $x \in R$, the conditional probabilities*

$$e_y(x) = \mathbb{P}(X_n = x | Y_n = y)$$

are called the **emission probabilities** *of the model. Let $p\colon S \times S \to [0, 1]$ be the transition kernel for $(Y_n)_{n\geq 0}$. It is taken as an assumption that the transition kernel for $(\mathbf{Z}_n)_{n\geq 0}$ is*

$$\mathbb{P}(\mathbf{Z}_{n+1} = (x', y') | \mathbf{Z}_n = (x, y)) = p(y, y')e_{y'}(x'). \tag{16.1}$$

It is easy to verify that Equation 16.1 defined a valid transition kernel on $R \times S$ (the reader should take a moment to check this). Note: it is not assumed that $(X_n)_{n\geq 0}$ alone is a Markov chain. The form of the transition kernel of $(\mathbf{Z}_n)_{n\geq 0}$ is special; it is not generally true that a vector-valued Markov chain's transition probabilities will only depend on the new state of one of the coordinates. This is the defining characteristic of Hidden Markov models.

By induction and the Markov property, it follows that

$$\begin{aligned} &\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_0 = y_0, \ldots, Y_n = y_n) \\ &= \mathbb{P}(X_0 = x_0, Y_0 = y_0)p(y_0, y_1)e_{y_1}(x_1) \cdots p(y_{n-1}, y_n)e_{y_n}(x_n) \\ &= \mathbb{P}(Y_0 = y_0)e_{y_0}(x_0)p(y_0, y_1)e_{y_1}(x_1) \cdots p(y_{n-1}, y_n)e_{y_n}(x_n). \end{aligned} \tag{16.2}$$

On the other hand, we could condition as

$$\begin{aligned} &\mathbb{P}(X_0 = x_0, Y_0 = y_0, \ldots, X_n = x_n, Y_n = y_n) \\ &= \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n)\mathbb{P}(X_0 = x_0, \ldots, X_n = y_n | Y_0 = y_0, \ldots, Y_n = y_n). \end{aligned}$$

The former probability is $\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n) = \mathbb{P}(Y_0 = y_0)p(y_0, y_1) \cdots p(y_{n-1}, y_n)$ since $(Y_n)_{n\geq 0}$ is a Markov chain, and so combining Equations 16.2 and 16.3 we have

$$\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n | Y_0 = y_0, \ldots, Y_n = y_n) = e_{y_0}(x_0) \cdots e_{y_n}(x_n). \tag{16.3}$$

By reversing the above calculations, we see that Equation 16.3 implies that $(\mathbf{Z}_n)_{n\geq 0}$ is a Markov chain. So, we could alternatively take, as the definition of a HMM, a pair (X_n, Y_n) such that Y_n is a Markov chain and Equation 16.3 holds for the joint distirbution.

The factorization of Equation 16.3 is a kind of independence property (though not particularly eary to formalize). The intuition is that the underlying process $(Y_n)_{n\geq 0}$ is running, and then at each time k the observed signal X_k is generated from the present value of Y_k, independently of the previous observations. Problems 1 and 2 on Homework 5 show how to construct Hidden Markov models.

The language is suggestive of the chain $(Y_n)_{n\geq 0}$ *generating signals* which we observe as $(X_n)_{n\geq 0}$, possessing some random noise. Indeed, one of the main application of HMMs is in *machine speech recognition*. Here we have $S = \{\text{words}\}$ and $R = \{\text{wave forms}\}$. We want to develop a model to decode which wave forms correspond to which words (i.e. determine the emission probabilities).

Example 16.3 (Occasionally Dishonest Casino, Continued)**.** Example 16.1 can be described as a HMM: the underlying Markov chain $(Y_n)_{n\geq 0}$ (which we cannot observe directly) is the switching between fair and loaded dice: $S = \{F, L\}$. The model presented is that $p(F, L) = 0.05$ and $p(L, F) = 0.9$ (so $p(F, F) = 0.95$ and $p(L, L) = 0.1$). The observed process $(X_n)_{n\geq 0}$ is the roll of the die, so $R = \{1, 2, 3, 4, 5, 6\}$. In this case, we know a priori that $e_F(i) = \frac{1}{6}$ for $i = 1, \ldots, 6$, while $e_L(1) = 0.5$ and $e_L(2) = \cdots = e_L(6) = 0.1$. Since all the rolls (of any dice) are independent of each other, the product formula of Equation 16.3 follows.

Example 16.4. Let $(Y_n)_{n\geq 0}$ be a Markov chain with state space S. Let R be a set, and let $f\colon S \to R$ be a function. Then $X_n = f(Y_n)$ is a Hidden Markov Model, with $e_i(x) = \mathbb{P}(X_n = x | Y_n = i) = \mathbb{P}(f(Y_n) = x | Y_n = i) = \mathbb{1}_{\{f(i)=x\}}$. (As Homework 5, Problem 2 shows, this is a special case of a general construction for HMMs: we can always represent $X_n = F(Y_n, V_n)$ where $F\colon S \times [0, 1] \to R$ is a fixed function and V_n are i.i.d. uniform random variables.)

16.2. **The Forward Algorithm.** Let (X_n, Y_n) be a HMM (with known parameters – i.e. known transition kernel for Y_n and known emission probabilities). As per Example 16.1, the kinds of questions we're interested in are: what is the probability of a given sequence of states $y_0, y_1, \ldots, y_N$ of Y_n occurring, given an observed sequence of symbols $x_0, x_1, \ldots, x_N$?

Fix an **observation sequence** $\mathbf{x} = (x_0, x_1, \ldots, x_N)$ and a **state sequence** $\mathbf{y} = (y_0, y_1, \ldots, y_N)$. We use the notation:

$$\mathbb{P}(\mathbf{x}) = \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_N = x_n)$$
$$\mathbb{P}(\mathbf{x}, \mathbf{y}) = \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_N = x_N, Y_0 = y_0, Y_1 = y_1, \ldots, Y_N = y_N)$$

Then what we seek to calculate is

$$\mathbb{P}(\mathbf{y}|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}, \mathbf{y})}{\mathbb{P}(\mathbf{x})}.$$

In this language, Equation 16.2 takes the form

$$\mathbb{P}(\mathbf{x},\mathbf{y}) = \mathbb{P}(Y_0 = y_0)p(y_0,y_1)e_{y_0}(x_0)p(y_1,y_2)e_{y_1}(x_1)\cdots p(y_{N-1},y_N)e_{y_N}(x_N). \tag{16.4}$$

We could then calculate $\mathbb{P}(\mathbf{x}) = \sum_{\mathbf{y}\in S^N}\mathbb{P}(\mathbf{x},\mathbf{y})$ to determine the conditional probability $\mathbb{P}(\mathbf{x}|\mathbf{y})$ in question. But this approach is computationally infeasible. For each $\mathbf{y} \in S^N$ there are $2N$ multiplications to be done to compute the term $\mathbb{P}(\mathbf{x},\mathbf{y})$ in Equation 16.4, and then the sum is over $|S|^N$ such terms: so there are $2N|S|^N$ computations. Even if $|S|$ is very small, as N grows moderatley this becomes intractible for any computer.

The *forward algorithm* is a recursive procedure to compute the conditional probability in question in polynomial time. Here is how it works. Fix an observation sequence $\mathbf{x} = (x_0, x_1, \ldots, x_N)$. For any state $y \in S$ and any time $0 \leq n < N$, set

$$\alpha_n(y) = \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n, Y_n = y)$$

(I.e. the probability that the first n observations have occurred and the hidden state at the present time is y). Then we have

$$\begin{aligned}\alpha_{n+1}(y') &= \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1}, Y_{n+1} = y') \\ &= \sum_{y\in S}\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1}, Y_n = y, Y_{n+1} = y').\end{aligned}$$

For each such term,

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1}, Y_n = y, Y_{n+1} = y') \\ &= \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_n = y)\mathbb{P}(X_{n+1} = x_{n+1}, Y_{n+1} = y'|X_0 = x_0, \ldots, X_n = x_n, Y_n = y) \\ &= \alpha_n(z)\mathbb{P}(X_{n+1} = x_{n+1}, Y_{n+1} = y'|X_0 = x_0, \ldots, X_n = x_n, Y_n = y)\end{aligned}$$

This last term is handled easily by the Markov property for $\mathbf{Z}_n$.

Lemma 16.5. *Let $\mathbf{Z}_n = (X_n, Y_n)$ be a Markov chain with state space $R \times S$. For any states $x_0, \ldots, x_{n+1} \in R$ and $y, y' \in S$,*

$$\begin{aligned}&\mathbb{P}(\mathbf{Z}_{n+1} = (x_{n+1}, y')|X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n, Y_n = y) \\ &= \mathbb{P}(\mathbf{Z}_{n+1} = (x_{n+1}, y')|\mathbf{Z}_n = (x_n, y)).\end{aligned}$$

Remark 16.6. Below we give an elementary but cumbersome discrete probability proof of this fact. The intuition is quite clear. Let's denote the information in all the random variables $\mathbf{Z}_0, \ldots, \mathbf{Z}_n$ symbolically as $\mathscr{F}_{\leq n}$, and the information contained in just $\mathbf{Z}_n$ as $\mathscr{F}_n$. The Markov property then says that

$$\mathbb{P}(\mathbf{Z}_n = \mathbf{z}|\mathscr{F}_{\leq n}) = \mathbb{P}(\mathbf{Z}_n = \mathbf{z}|\mathscr{F}_n).$$

Of coursr $\mathscr{F}_n \subseteq \mathscr{F}_{\leq n}$. Now, denote by $\mathscr{G}$ the information contained in the random variables $X_0, \ldots, X_n, Y_n$. This list is strictly contained in the components of $\mathbf{Z}_0, \ldots, \mathbf{Z}_n$, and it contains the components of $\mathbf{Z}_n$, so we have

$$\mathscr{F}_n \subseteq \mathscr{G} \subseteq \mathscr{F}_{\leq n}.$$

Since the conditional distribution of $\mathbf{Z}_{n+1}$, conditioned on the *big* set $\mathscr{F}_{\leq n}$ is the same as conditioned on the *small* set $\mathscr{F}_n$, it stands to reason it should also be the same when conditioned on the *intermediate* set $\mathscr{G}$. This intuition can be formalized and forms the basis of the theory of **conditional expectation** and **Martingales**, that we will study later in this course.

Proof. By definition

$$\begin{aligned}&\mathbb{P}(\mathbf{Z}_{n+1} = (x_{n+1}, y') | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n, Y_n = y)\\ &= \frac{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1}, Y_n = y, Y_{n+1} = y')}{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_n = y)}.\end{aligned}$$

The numerator can be expanded by summing over the probabilities of all possible values taken by $Y_0, \ldots, Y_{n-1}$:

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1}, Y_n = y, Y_{n+1} = y')\\ &= \sum_{y_0, \ldots, y_{n-1}} \mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y), \mathbf{Z}_{n+1} = (x_{n+1}, y')\right)\end{aligned}$$

We can then express

$$\begin{aligned}&\mathbb{P}(\mathbf{Z}_{n+1} = (x_{n+1}, y') | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n, Y_n = y)\\ &= \sum_{y_0, \ldots, y_{n-1}} \frac{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y), \mathbf{Z}_{n+1} = (x_{n+1}, y')\right)}{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_n = y)}\\ &\qquad \cdot \frac{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y)\right)}{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y)\right)}\\ &= \sum_{y_0, \ldots, y_1} \frac{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y), \mathbf{Z}_{n+1} = (x_{n+1}, y')\right)}{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y)\right)}\\ &\qquad \cdot \frac{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y)\right)}{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_n = y)}\\ &= \sum_{y_0, \ldots, y_{n-1}} \mathbb{P}(\mathbf{Z}_{n+1} = (x_{n+1}, y') | \mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y))\\ &\qquad \cdot \frac{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y)\right)}{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_n = y)}.\end{aligned}$$

Since $\mathbf{Z}_n$ is a Markov chain, the conditional probability is simply equal to the one step transition probability $\mathbb{P}(\mathbf{Z}_{n+1} = (x_{n+1}, y') | \mathbf{Z}_n = (x_n, y))$, which does not depend on the summation variables. Hence we have

$$\mathbb{P}(\mathbf{Z}_{n+1} = (x_{n+1}, y') | \mathbf{Z}_n = (x_n, y)) \sum_{y_0, \ldots, y_{n-1}} \frac{\mathbb{P}\left(\mathbf{Z}_0 = (x_0, y_0), \ldots, \mathbf{Z}_{n-1} = (x_{n-1}, y_{n-1}), \mathbf{Z}_n = (x_n, y)\right)}{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_n = y)}.$$

The numerator is just the expansion of the denominator over all possible values of $Y_0, \ldots, Y_{n-1}$, so the ratio is 1, proving the lemma. □

We therefore have

$$\begin{aligned}\alpha_{n+1}(y') &= \sum_{y \in S} \alpha_n(y) \mathbb{P}\left(\mathbf{Z}_{n+1} = (x_{n+1}, y') | \mathbf{Z}_n = (x_n, y)\right)\\ &= \sum_{y \in S} \alpha_n(y) p(y, y') e_{y'}(x_{n+1})\end{aligned}$$

by Equation 16.1. We can now factor out the term $e_{y'}(x_{n+1})$ which is independent of the summation variable y, yo achieve

$$\alpha_{n+1}(y') = e_{y'}(x_{n+1}) \sum_{y \in S} \alpha_n(y) p(y, y'). \tag{16.5}$$

Then, to compute $\mathbb{P}(\mathbf{x}) = \mathbb{P}(X_0 = x_0, \ldots, X_N = x_N)$, we simply sum

$$\mathbb{P}(\mathbf{x}) = \sum_{y \in S} \mathbb{P}(X_0 = x_0, \ldots, X_N = x_N, Y_N = y) = \sum_{y \in S} \alpha_N(y).$$

Let us now analyze the Forward Algorithm for complexity. Let $0 < n < N$, and suppose we have calculated all the quantities $\{\alpha_n(y) \colon y \in S\}$. For each $y' \in S$, Equation 16.5 allows us to compute $\alpha_{n+1}(y')$ with $|S|$ multiplications, then the sum of those $|S|$ terms, and finally the product with the one remaining term, for a total of $2|S| + 1$ operations. We must compute all $|S|$ terms $\{\alpha_{n+1}(y') \colon y' \in S\}$, which means repeating this $|S|$ times for a total of $|S|(2|S| + 1)$ operations.

At time 0, we initialize with the known distribution $\alpha_0(y) = \mathbb{P}(X_0 = x_0, Y_0 = y) = \mathbb{P}(Y_0 = y) e_y(x_0)$, requiring one computation for each state, so $|S|$ computations in total. Then we need $N \cdot |S|(2|S|+1)$ operations to compute all the terms $\{\alpha_N(y) \colon y \in S\}$. Finally, to compute $\mathbb{P}(\mathbf{x})$ we must sum up these $|S|$ terms, for a total of $2|S| + N|S|(2|S| + 1) = O(N|S|^2)$ computations.

Example 16.7. To reiterate, if $(X_n, Y_n)_{n \geq 0}$ is a HMM and $\mathbf{x} = (x_0, \ldots, x_N)$ is a fixed observation sequence, computing $\mathbb{P}(\mathbf{x})$ by

- the straightforward method $\sum_{\mathbf{y} \in S^N} \mathbb{P}(\mathbf{x}, \mathbf{y})$ requires $2N|S|^N$ operations.
- the Forward algorithm requires $2N|S|^2 + (N + 2)|S|$ operations.

For example, suppose $|S| = 2$ (as in the occasionally dishonest casino example) and $N = 100$. Then the Forward Algorithm takes 1004 computations, while the straightforward method takes 2.54×10^{32} computations. Note, the world's fastest supercomputer (as of November, 2011) operates at about 10.5 PFLOPS, or 1.05×10^{16} flops (floating-point operations per second). On this computer, the straightforward computation above would take more than 750 million years. The computation by the forward algorithm, by comparison, could be implemented on my iPhone in 1 microsecond.

17. LECTURE 16: MAY 9, 2012

To recap from last time: a Hidden Markov Model (HMM) is a pair $(X_n, Y_n)_{n\geq 0}$ of stochastic processes, where $(Y_n)_{n\epsilon 0}$ is a Markov chain with transition kernel $p(\cdot,\cdot)$, and the joint distribution $\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_0 = y_0, \ldots, Y_n = y_n)$ is given by

$$\mathbb{P}(Y_0 = y_0)e_{y_0}(x_0)p(y_0, y_1)e_{y_1}(x_1)\cdots p(y_{n-1}, y_n)e_{y_n}(x_n).$$

Here $e_y(x) = \mathbb{P}(X_n = x|Y_n = y)$ are the *emission probabilities* (and are time-homogeneous). Equivalently, the model is that $(Y_n)_{n\geq 0}$ *and the pair* $(X_n, Y_n)_{n\geq 0}$ are Markov chains, where the transition probabilities for the pair chain are

$$\mathbb{P}\left((X_{n+1}, Y_{n+1}) = (x', y')|(X_n, Y_n) = (x, y)\right) = p(y, y')e_{y'}(x').$$

This models a scenario where there is an underlying driving Markov process Y_n which is "hidden", but it generates a process X_n which can be observed (as signals with some random noise). The emission probabilities are known, and we would like to determine from a given signal what the hidden state of the underlying process is.

One question we can answer (in a computationally tractable way) is: given an observed output $\mathbf{x} = (x_0, x_1, \ldots, x_N)$, what is the probability that the hidden chain followed the trajectory $\mathbf{y} = (y_0, y_1, \ldots, y_N)$? That is, what is $\mathbb{P}(\mathbf{y}|\mathbf{x})$? By definition is it $\mathbb{P}(\mathbf{x}, \mathbf{y})/\mathbb{P}(\mathbf{x})$. The numerator is calculated (as above) as the product of transition probabilities and emission probabilities (and the probability of the initial state). To calculate the denominator $\mathbb{P}(\mathbf{x})$, we use the **Forward Algorithm**: for fixed observation sequence $\mathbf{x}$ as above

Initializiation: For $y \in S$, set $\alpha_0(y) = \mathbb{P}(X_0 = x_0, Y_0 = y_0) = \mathbb{P}(Y_0 = y_0)e_{y_0}(x_0)$.

Recursion: For $y' \in S$ and $0 < n < N$, set $\alpha_{n+1}(y') = e_{y'}(x_{n+1}) \sum_{y\in S} \alpha_n(y)p(y, y')$.

Termination: $\mathbb{P}(\mathbf{x}) = \sum_{y\in S} \alpha_N(y)$.

If S is the state space of the hidden chain $(Y_n)_{n\geq 0}$, this algorithm is $O(N|S|^2)$ and so quickly implemented on a computer. Homework 5 Problem 1 asks you to do this computation for a Crooked Casino style HMM.

17.1. **Most Likely Trajectory.** HMMs are used extensively in signal processessing applications: notably speach recognition and error correcting codes for wireless communications. The observed process X_n is interpretted as a signal with random noise, and the goal is to decode the true source Y_n. This is, of course, impossible, even if all the emission probabilities for X_n given Y_n and all the transition probabilities for Y_n are known exactly (which is usually not true anyhow). The question is, then, what is our best guess about the values of $y_0, y_1, \ldots, y_N$ given an observed sequence $x_0, x_1, \ldots, x_N$?

Referring to the previous lecture and the forward algorithm, we have an efficient way to calculate $\mathbb{P}(\mathbf{y}|\mathbf{x})$. The usual answer for the best guess $\mathbf{y}$ for the true value of $\mathbf{y}$ is then

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}} \mathbb{P}(\mathbf{y}|\mathbf{x}).$$

The operation argmax is Engineering Speak: the meaning is that $\mathbf{y}^*$ is chosen so that

$$\mathbb{P}(\mathbf{y}^*|\mathbf{x}) = \max_{\mathbf{y}} \mathbb{P}(\mathbf{y}|\mathbf{x}).$$

Such a $\mathbf{y}^*$ surely exists (since we are dealing with finite-state processes here). A more subtle theoretical question is whether the maximum is unique (the answer is *no* typically), so we can weaken our goal to find *a* maximizer. The intuition is simple: we observe a given $\mathbf{x}$. We check all possible strings $\mathbf{y}$, and the one that has the highest probability of generating $\mathbf{x}$ is (probably close to) the true string.

We now come to the question of computational feasibility. We have an algorithm (the Forward Algorithm) to compute $\mathbb{P}(\mathbf{y}|\mathbf{x})$ for fixed $\mathbf{x}$ and $\mathbf{y}$, and it has complexity $O(N|S|^3)$ where S is the state space of the hidden chain. The most obvious way to find $\operatorname{argmax}_{\mathbf{y}}\mathbb{P}(\mathbf{y}|\mathbf{x})$ is to compute all of the probabilities $\{\mathbb{P}(\mathbf{y}|\mathbf{x})\colon \mathbf{y} \in S^N\}$, and then select $\mathbf{y}^*$ as (one of) the maximizer(s). Of course, this procedure will exponentially complexify the calculation, yielding $O(N|S|^{N+3})$. This will never be computationally possible for moderately large N.

Enter **Viterbi**. In 1967, he was an EE professor here at UCLA, when he invented the algorithm (below) for recursively effectively computing the most likely trajectory in a HMM. He quickly realized the enormous practical applications, and formed a company (called Linkabit) to market it (initially only to the military and NASA). In 1973, he resigned from UCLA and moved to San Diego (where Linkabit was based) to run the company full-time. In 1980, Linkabit was acquired by a larger company called Microwave Associates Communications. A change in management caused Viterbi (and his original Linkabit team) to resign en masse in 1985, at which time they formed a new startup company that they called QUALCOMM (which is, of course, now an industry giant in communications technology). While at QUALCOMM, Viterbi more or less single-handedly invented the CDMA algorithm used by all modern cellular providers; the Viterbi algorithm is used at the heart of it (for error-correcting codes). Viterbi retired from QUALCOMM in 2000 (at age 65), and became an EE professor here at UCSD, where he is still on the faculty.

17.2. **The Viterbi Algorithm.** Fix an observation sequence $\mathbf{x} = (x_0, x_1, \ldots, x_N)$. First note that $\mathbb{P}(\mathbf{y}|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x},\mathbf{y})}{\mathbb{P}(\mathbf{x})}$, so

$$\max_{\mathbf{y}} \mathbb{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathbb{P}(\mathbf{x})} \max_{\mathbf{y}} \mathbb{P}(\mathbf{x},\mathbf{y}).$$

It therefore suffices to maximize $\mathbb{P}(\mathbf{x},\mathbf{y})$ over $\mathbf{y}$; a maximizer $\mathbf{y}^*$ for this joint probability is also a maximizer for the conditional probability.

We build up recursively the maximum conditional probability of a state sequence given the observed sequence. For $0 < n \leq N$, define, for any state y,

$$V_n(y) = \max_{y_0,\ldots,y_{n-1}\in S} \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_0 = y_0, \ldots, Y_{n-1} = y_{n-1}, Y_n = y).$$

Then $\max_{y\in S} V_n(y) = \max_{\mathbf{y}\in S^N} \mathbb{P}(\mathbf{x},\mathbf{y})$, whose argmax we wish to compute. On its face, it seems that computing $V_n(y)$ requires a brute force approach with $|S|^{n-1}$ computations of probabilities; however, note that the Hidden Markov Model provides a simple product structure for the probabilities in question. From Equation 16.4, we have for any states $y_0, \ldots, y_{n-1}, y, y'$ that

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, \ldots, X_{n+1} = x_{n+1}, Y_0 = y_0, \ldots, Y_n = y, Y_{n+1} = y')\\ &= \mathbb{P}(Y_0 = y_0)e_{y_0}(x_0)p(y_0,y_1)e_{y_1}(x_1)\cdots p(y_{n-1},y)e_y(x_n)p(y,y')e_{y'}(x_{n+1})\end{aligned}$$

and so

$$V_{n+1}(y') = \max_{y_0,\dots,y_{n-1},y} \Big(\mathbb{P}(Y_0 = y_0)p(y_0,y_1)e_{y_0}(x_0)\cdots p(y_{n-1},y)e_{y_n}(x_n)p(y,y')e_y(x_{n+1})\Big)$$
$$= \max_{y_0,\dots,y_{n-1},y} \Big(\mathbb{P}(Y_0 = y_0)p(y_0,y_1)e_{y_0}(x_0)\cdots p(y_{n-1},y_n)e_y(x_n)p(y,y')\Big)e_{y'}(x_{n+1}).$$

The quantity inside the large parentheses is $\mathbb{P}(X_0 = x_0,\dots,X_n = x_n, Y_0 = y_0,\dots,Y_{n-1} = y_{n-1}, Y_n = y)\cdot p(y,y')$, again by Equation 16.4. Thus, we have

$$V_{n+1}(y') = \max_{y_0,\dots,y_{n-1},y} \Big(\mathbb{P}(X_0 = x_0,\dots,X_n = x_n, Y_0 = y_0,\dots,Y_{n-1} = y, Y_n = y')p(y,y')\Big)e_{y'}(x_{n+1}).$$

The maximum can be rewritten as

$$\max_y \Big(p(y,y') \max_{y_0,\dots,y_{n-1}} \mathbb{P}(X_0 = x_0,\dots,X_n = x_n, Y_0 = y_0,\dots,Y_n = y)\Big)$$
$$= \max_y \Big(p(y,y')V_n(y)\Big).$$

Altogether, then, we have the recursive formula

$$V_{n+1}(y') = \max_{y\in S} V_n(y)p(y,y')e_{y'}(x_{n+1}). \tag{17.1}$$

Actually computing this maximum means computing the quantitites

$$W_{n+1}(y;y') = V_n(y)p(y,y')e_{y'}(x_{n+1}) \tag{17.2}$$

for each pair y, y'. Then define, for each state $y' \in s$, a function $\psi_n^*\colon S \to S$, by

$$\psi_n^*(y') = \operatorname*{argmax}_y W_{n+1}(y;y'). \tag{17.3}$$

Then, by definition, we have

$$V_{n+1}(y') = W_{n+1}(\psi_n^*(y'), y'). \tag{17.4}$$

Once we have computed $W_k(y;y')$ and $V_k(y')$ for all $0 < k \le N$, we can then do the final maximization: since $\max_{\mathbf{y}} \mathbb{P}(\mathbf{x},\mathbf{y}) = \max_y V_N(y)$, we set

$$y_N^* = \operatorname*{argmax}_y V_N(y); \quad \therefore \max_{\mathbf{y}} \mathbb{P}(\mathbf{x},\mathbf{y}) = V_N(y_N^*).$$

Now, to compute the maximizing sequence of states, we back-track: we have already computed y_N^*. In the process of computing $V_N(y')$, we computed $\psi_{N-1}^*(y')$ for each y', i.e. the state $y = \psi_{N-1}^*(y')$ which maximizes $W_N(y;y')$. We now know that the state y' which maximizes the end result is $y' = y_N^*$, and so we define $y_{N-1}^* = \psi_{N-1}^*(y_N^*)$. Continuing recursively, we define

$$y_k^* = \psi_k^*(y_{k+1}), \quad 0 \le k < N-1.$$

This produces the maximizing sequence $\mathbf{y}^* = (y_0^*, y_1^*, \dots, y_N^*)$ that we sought.

To summarize, here is the laid out **Viterbi algorithm**.

Initializiation: For $y \in S$, set $V_0(y) = \mathbb{P}(X_0 = x_0, Y_0 = y_0) = \mathbb{P}(Y_0 = y_0)e_{y_0}(x_0)$.

Recursion: For $y, y' \in S$ and $0 < n < N$, set $W_{n+1}(y; y') = V_n(y)p(y, y')e_{y'}(x_{n+1})$.

Then compute $\psi_n^*(y') = \operatorname{argmax}_y W_{n+1}(y; y')$.

Set $V_{n+1}(y') = W_{n+1}(\psi_n^*(y'); y')$.

Termination: $\max_{\mathbf{y}} \mathbb{P}(\mathbf{x}, \mathbf{y}) = \max_{y \in S} V_N(y)$, so define $y_N^* = \operatorname{argmax}_y V_N(y)$.

Back-tracking: For $0 \leq k < N$, set $y_k^* = \psi_k^*(y_{k+1}^*)$.

Note: at the $n+1$st recursion step we compute $|S|$ maximizers $\psi_n^*(y')$ for all states $y' \in S$. It could well be that more than one state y maximizes $W_{n+1}(y; y')$, so there is some indeterminacy. In practice, we just define $\psi_n^*(y')$ to be the *first* y to maximize $W_{n+1}(y; y')$ (i.e. we order the states of S to begin with). The same goes for the final mazimization of $V_n(y)$ over $y \in S$. So the order we choose for S determines which maximum we find.

17.3. **Complexity Analysis.** The Viterbi algorithm produces a maximizing state sequence $\mathbf{y}^* = (y_0^*, \ldots, y_N^*)$ for a given observation sequences $\mathbf{x} = (x_0, \ldots, x_N)$; that is, $\mathbb{P}(\mathbf{y}^*|\mathbf{x}) = \max_{\mathbf{y}} \mathbb{P}(\mathbf{y}|\mathbf{x})$. The reason it is so useful is that, unlike the brute force aprroach discussed above that is exponential in time N, this algorithm is *linear* in time N. Let's run through the algorithm to see why.

- For each y, computing $V_0(y)$ requires one multiplication; so computing the quantities $\{V_0(y) \colon y \in S\}$ requires $|S|$ operations.
- Once we have computed all the quantities $\{V_n(y) \colon y \in S\}$, for any fixed state y' we need two multiplications to compute $W_{n+1}(y; y') = V_n(y)p(y, y')e_{y'}(x_{n+1})$ for each pair $y, y' \in S$. So there are $2|S|^2$ operations to compute the quantities $\{W_{n+1}(y; y') \colon y, y' \in S\}$; in computing them, we run a pointer to identify the (first) y that maximizes $W_{n+1}(y; y')$ for each given y', so we compute $\psi_n^*(y')$ and $V_{n+1}(y')$ in the process $|S|$ additional operations (specifying the pointer $\psi_n^*(y')$ for each y').
- At termination, we argument-maximize $V_N(y)$ over y to find y_N^*, and backtrack to compute the maximizing sequence $y_k^* = \psi_k^*(y_{k+1}^*)$; this is all a matter of calling the pointers we specified through the process, so no additional operations are needed.

Thus, the total number of operations is: $|S|$ for intialization, and $N \cdot (2|S|^2 + |S|)$ to compute $V_1, V_2, \ldots, V_N$ and specify the relevant pointers. The total complexity of the Viterbi algorithm is then $|S| + N(2|S|^2 + |S|) = 2N|S|^2 + (N+1)|S| = O(N|S|^2)$.

18. Lecture 18: Continuous Time Markov Chains

Let S be a finite or countable state space. A stochastic process $(X_t)_{t\geq 0}$ with state space S, indexed by non-negative reals t (in the half-infinite interval $[0,\infty)$, or perhaps in a subinterval $[a,b]$) is called a **continuous-time Markov chain** if the following two properties hold.

(1) (Markov property) Let $0 \leq t_0 < t_1 < \cdots < t_n < \infty$ be a sequence of times, and let $i_0, i_1, \ldots, i_n \in S$ be a sequence of states such that $\mathbb{P}(X_{t_0} = i_0, X_{t_1} = i_1, \ldots, X_{t_{n-1}} = i_{n-1}) > 0$. Then

$$\mathbb{P}(X_{t_n} = i_n | X_{t_0} = i_0, X_{t_1} = i_1, \ldots, X_{t_{n-1}} = i_{n-1}) = \mathbb{P}(X_{t_n} = i_n | X_{t_{n-1}} = i_{n-1}).$$

(2) (Right-continuity) For $t \geq 0$ and $i \in S$, if $X_t = i$ then there is $\epsilon > 0$ such that $X_s = i$ for $t \leq s \leq t + \epsilon$.

The Markov property is fundamentally the same as the the one in Definition 1.5, except that we now must allow the times to be chosen from a larger set (all positive reals). The intuition remains the same: the distribution of the process at the present time t_n, conditioned on *any* past information, is the same as conditioned on only the *immediate* past. (I.e. for any time t earlier than t_n, conditioning on times t and earlier is the same as conditioning only on time t.) Condition (2) (right-continuity) is a technical condition to guarantee that very wild small-time behavior is ruled out.

As with the discrete time case, we immediately restrict our attention to **time-homogeneous** chains:

(3) For any times $0 \leq s < t < \infty$ and states $i, j \in S$,

$$\mathbb{P}(X_t = j | X_s = i) = \mathbb{P}(X_{t-s} = j | X_0 = i).$$

For discrete time chains, it was really enough to understand the one-step transition matrix $\mathbb{P}(X_1 = j | X_0 = i)$, since there were no times in between 0 and 1. For continuous-time chains, however, we really need all the information in the **transition kernel**

$$p_t(i,j) = \mathbb{P}(X_t = j | X_0 = i), \quad t > 0.$$

This is a lot more information to process, so we will need a different approach to analyze such chains.

18.1. **Jump-Times.** Because of the right-continuity assumption, if $X_t = i$ then $X_{t+\epsilon} = i$ for all sufficiently small $\epsilon > 0$. Hence, if we define the **jump-time**

$$J_1 = \min\{t \geq 0 \colon X_t \neq X_0\}$$

then $J_1 > 0$ with probability 1. Now, suppose $J_1 > s$ – so we have already waited time s in state $X_0 = i$. What is the probability that $J_1 > s + t$ – i.e. how likely is it wait at least t longer in state i?

Proposition 18.1. *For $s, t > 0$ and $i \in S$,*

$$\mathbb{P}(J_1 > s + t | J_1 > s) = \mathbb{P}(J_1 > t).$$

Proof. Let $(X_t)_{t\geq 0}$ be a time-homogeneous Markov chain, with jump time J_1 from the initial state. Conditioned on $X_0 = i$, he event $\{J_1 > s\}$ is equal to the event $\{X_u = i,\ \forall u \geq s\}$. Let's consider a discrete version of this: fix a sequence of times

$$0 \leq u_0 < u_1 < \cdots < u_n = s < u_{n+1} < \cdots < u_{n+m} = s + t.$$

Then $\mathbb{P}_i(J_1 > s + t | J_1 > s)$ is approximated by

$$\mathbb{P}(X_{u_0} = X_{u_1} = \cdots = X_{u_{n+m}} = i | X_{u_0} = X_{u_1} = \cdots = X_{u_n} = i).$$

Abbreviate the event $\{X_{u_0} = X_{u_1} = \cdots = X_{u_n} = i\}$ as $A_n(i)$; then we have the usual decomposition

$$\begin{aligned}
&\mathbb{P}(X_{u_0} = X_{u_1} = \cdots = X_{u_{n+m}} = i | A_n(i)) \\
&= \mathbb{P}(X_{u_{n+m}} = \cdots = X_{u_{n+1}} = i | A_n(i)) \\
&= \mathbb{P}(X_{u_{n+1}} = i | A_n(i)) \mathbb{P}(X_{u_{n+2}} = i | X_{u_n}, A_n(i)) \cdots \mathbb{P}(X_{u_{n+m}} = i | X_{u_{n+m-1}} = \cdots = X_{u_{n+1}} = i, A_n(i)).
\end{aligned}$$

By the Markov property, this is simply the product

$$\mathbb{P}(X_{u_{n+1}} = i | X_{u_n} = i) \mathbb{P}(X_{u_{n+2}} = i | X_{u_{n+1}} = i) \cdots \mathbb{P}(X_{u_{n+m}} = i | X_{u_{n+m-1}} = i).$$

Now lets choose the partition points to evenly subdivide the interval $[s, s+t]$ into m parts – that is, $u_{n+j+1} = u_{n+j} + \frac{t}{m}$. Then by time-homogeneity, each of the above terms in the product is equal to

$$\mathbb{P}(X_{t/m} = i | X_0 = i).$$

By letting the partition witdth tend to 0, we recover the desired probability, so

$$\mathbb{P}_i(J_1 > s + t | J_1 > s) = \lim_{m\to\infty} \left[\mathbb{P}(X_{t/m} = i | X_0 = i)\right]^m. \tag{18.1}$$

As for $\mathbb{P}_i(J_1 > t)$, we can approximate this probability similar to above by choosing a partition $0 = v_1 < v_2 < \cdots < v_m = t$ and computing

$$\begin{aligned}
&\mathbb{P}_i(X_{v_1} = \cdots = X_{v_m} = i) \\
&= \mathbb{P}(X_{v_1} = i | X_0 = i) \mathbb{P}(X_{v_2} = i | X_{v_1} = i) \cdots \mathbb{P}(X_{v_m} = i | X_{v_0} = \cdots = X_{v_{m-1}} = i) \\
&= \mathbb{P}(X_{v_1 - v_0} = i) \mathbb{P}(X_{v_2 - v_1} = i) \cdots \mathbb{P}(X_{v_m - v_{m-1}} = i)
\end{aligned}$$

where we have used the Markov property and time-homogeneity in the last equality. If we again choose the even partition with $v_{j+1} - v_j = t/m$, and take limits, this gives

$$\mathbb{P}_i(J_1 > t) = \lim_{m\to\infty} \left[\mathbb{P}(X_{t/m} = i)\right]^m. \tag{18.2}$$

Equations 18.1 and 18.2 prove that $\mathbb{P}_i(J_1 > s + t | J_1 > s) = \mathbb{P}_i(J_1 > t)$ for all states i. This proves the proposition. □

18.2. **The Exponential Distribution.** The formula in Equation 18.2 for the (reverse) cumulative probability distribution function of J_1 has the feel of an exponential to it. This turns out to indeed be true. The result of Proposition 18.1 is a "memoryless" property: it says that the distribution of jump times after already waiting some fixed length of time is the same as it was at the beginning. In fact, there is only a one-parameter family of probability distributions with this property: the **exponential distribution**s.

Proposition 18.2. *If T is a random variable taking values in $(0,\infty)$, and if T has the memoryless property $\mathbb{P}(T > s+t|T > s) = \mathbb{P}(T > t)$ for all $s,t > 0$, then T is an exponential random variable with some intensity $q > 0$:*

$$\mathbb{P}(T > t) = e^{-qt}, \quad t \geq 0.$$

Notation: we write $T \sim \mathrm{Exp}(q)$.

Proof. Let $G_T(t) = \mathbb{P}(T > t)$. Then for $s,t > 0$

$$G_T(s+t) = \mathbb{P}(T > s+t) = \mathbb{P}(T > s)\mathbb{P}(T > s+t|T > s) = \mathbb{P}(T > s)\mathbb{P}(T > t) = G_T(s)G_T(t).$$

It follows from this functional equation that $G_T(t) = e^{-qt}$ for some $q > 0$: by induction it follows that $G_T(nt) = G_T(t)^n$ for all natural numbers n and $t > 0$. In particular, $G_T(1/n)^n = G_T(1) \in (0,1)$ and so there is $q > 0$ with $G_T(1) = e^{-q}$. Thus, for $m,n \in \mathbb{N}$, $G_T(m/n)^n = G_T(1)^m = e^{-qm}$ so that $G_T(m/n) = e^{-qm/n}$. So for rational t we have proved the result. The function G_T is decreasing, and so since it is equal to the function $e^{-q(\cdot)}$ on the dense set of positive rationals, it is equal to it everywhere. □

Here are a few useful properties of the exponential distribution.

Proposition 18.3. *Let $T_1, T_2, \ldots, T_n$ be independent with $T_j \sim \mathrm{Exp}(q_j)$.*

(a) *The density of T_j is $f_{T_j}(t) = q_j e^{-q_j t}\mathbb{1}_{\{t\geq 0\}}$, and $\mathbb{E}[T_j] = \frac{1}{q_j}$ and $\mathrm{Var}[T_j] = \frac{1}{q_j^2}$.*

(b) *For $s,t > 0$, $\mathbb{P}(T_j \geq s+t|T_j > s) = \mathbb{P}(T_j > t) = e^{-q_t t}$.*

(c) *$T = \min_j T_j$ is also exponential, with $T \sim \mathrm{Exp}(q_1 + \cdots + q_n)$. Moreover, $\mathbb{P}(T_j = T) = \frac{q_j}{q_1+\cdots+q_n}$.*

Proof. Parts (a) and (b) follow from easy calculation. For part (c), simply note from independence that

$$\mathbb{P}(T \geq t) = \mathbb{P}(T_1 \geq t, \ldots, T_n \geq t) = \mathbb{P}(T_1 \geq t)\cdots\mathbb{P}(T_n \geq t) = e^{-q_1 t}\cdots e^{-q_n t} = e^{-t(q_1+\cdots+q_n)}.$$

Moreover,

$$\begin{aligned}
&\mathbb{P}(T_1 = T)\\
&= \mathbb{P}(T_2 > T_1, \ldots, T_n > T_1)\\
&= \int_{t_2>t_1,\ldots,t_n>t_1} f_{T_1,\ldots,T_n}(t_1,\ldots,t_n)\,dt_1\cdots dt_n\\
&= \int_0^\infty dt \int_t^\infty \cdots \int_t^\infty f_{T_1,\ldots,T_n}(t,t_2,\ldots,t_n)\,dt_2\cdots dt_n\\
&= \int_0^\infty dt \int_t^\infty \cdots \int_t^\infty q_1 e^{-q_1 t} q_2 e^{-q_2 t_2}\cdots q_n e^{-q_n t}\,dt_2\cdots dt_n\\
&= \int_0^\infty q_1 e^{-q_1 t}\,dt \prod_{i=2}^n \int_t^\infty q_i e^{-q_i t_i}\,dt_i\\
&= \int_0^\infty q_1 e^{-q_1 t} \prod_{i=2}^n e^{-q_i t} = q_1 \int_0^\infty e^{-(q_1+\cdots+q_n)t}\,dt = \frac{q_1}{q_1+\cdots+q_n}.
\end{aligned}$$

□

Remark 18.4. In fact, one can go further. Let M be the $\{1, 2, \ldots, n\}$-valued random variable defined by $M = j$ iff $\min\{T_1, \ldots, T_n\} = T_j$. We just computed that the distribution of M is $\mathbb{P}(M = j) = \frac{q_j}{q_1+\cdots+q_j}$. Similar calculations show that M is independent of $T_1, \ldots, T_n$. This is quite remarkable. For example, suppose we have two processes, one $\mathrm{Exp}(1)$ and the other $\mathrm{Exp}(100)$. If we sample one of the two at random without knowing which one, even if we get value 0.9, we do not have *any* information about which one we actually sampled. This is a very strong form of the memoryless property.

18.3. **Transition Rates.** Proposition 18.2 shows that the first jump time J_1 of a Markov chain has an exponential distribution with some intensity $q > 0$, which may differ depending on the initial state. That is: for each $i \in S$, there is a $q(i) > 0$ so that the conditional distribution of J_1, conditioned on $X_0 = i$, is $\mathrm{Exp}(q(i))$. This gives a convenient way to think about any continuous-time Markov chain: at each state, there is an "exponential clock", which will ring at a random time whose distribution is exponential with some rate parameter $q(i)$. Once arriving at state i, the clock is started; it rings at time J_1, and the process moves to another state j, chosen with probabilities

$$p(i, j) = \mathbb{P}(X_{J_1} = j | X_0 = i).$$

Note that, by definition of the jump-time J_1, $X_{J_1} \neq X_0$, so $p(i, i) = 0$.

There is a slightly different interpretation which is somewhat more convenient for calculations. Define

$$q(i, j) = q(i)p(i, j) = q(i)\mathbb{P}(X_{J_1} = j | X_0 = i). \tag{18.3}$$

The numbers $q(i, j)$ for $i, j \in S$ are called the **transition rates** of the Markov chain. They satisfy

$$q(i, j) \geq 0, \quad q(i, i) = 0, \quad i, j \in S$$
$$\sum_{j \in S} q(i, j) = q(i) \sum_{j \in S} p(i, j) = q(i) < \infty.$$

The transition rates fully determine the Markov chain, so we can specify the chain simply by specifying the matrix $\mathbf{Q} = [q(i, j)]_{i,j \in S}$. Their interpretation in terms of the process is thus: when we arrive at state i, all other states start exponential clocks – state j's clock as rate $q(i, j)$. When a clock goes off, the chain moves to the state whose clock rang first. All of the clocks are independent, so the time that the first clock will go off is the minimum of all the independent exponential clocks; by Proposition 18.3(c), this is exponential with rate $\sum_j q(i, j) = q(i)$, as required. The probability that the clock at j goes off first is, again by Proposition 18.3(c), $\frac{q(i,j)}{\sum_k q(i,k)} = \frac{q(i,j)}{q(i)} = p(i, j)$.

18.4. **The Poisson Process.** The simplest non-trivial continuous time Markov chain is a **Poisson Process**, with rate $\lambda > 0$. This is a Markov chain with state space $S = \{0, 1, 2, \ldots\}$ and transition rates $q(i, i + 1) = \lambda$ for all $i \geq 0$, and $q(i, j) = 0$ when $j \neq i + 1$. So the process is non-decreasing; the moment that it arrives at state i, it begins an $\mathrm{Exp}(\lambda)$ clock, and jumps to state $i + 1$ when it rings. If $(X_t)_{t \geq 0}$ is a Poisson process, we can think of the random variable X_t as counting the number of events that have occurred up to time t. It is a very good model for radioactive decay, telephone calls coming into a call-center, the number of requests for a particular page coming into a web server, etc.

Why do we call it a Poisson process?

Proposition 18.5. *Let $(X_t)_{t\geq 0}$ be a Poisson process with rate $\lambda > 0$. Then for each $t > 0$, conditioned on $X_0 = 0$, the random variable X_t has a Poisson distribution with rate λt:*

$$\mathbb{P}_1(X_t = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k \in \mathbb{N}.$$

We will prove Proposition 18.5 in the next lecture, using the infinitesimal description of continuous-time processes and the forward and backward Kolmogorov equations.

19. LECTURE 19: MAY 14, 2012

19.1. **The Strong Markov Property.** In complete analogy with the discrete-time case, we have continuous-time **stopping times**. Given a Markov chain $(X_t)_{t\geq 0}$, a **stopping time** T is a random variable taking values in $[0,\infty]$ with the property that:

$$\text{for } t \geq 0, \text{ the event } \{T \leq t\} \text{ depends only on } \{X_s \colon s \leq t\}.$$

For example: the jump-time J_1 of a Markov chain, $J_1 = \min\{t \geq 0 \colon X_t \neq X_0\}$ is a stopping time: the event $\{J_1 \leq t\} = \{\exists s \leq t \; X_s \neq X_0\}$ manifestly depends only on X_s for $s \leq t$.

Theorem 19.1 (Strong Markov property)**.** *Let $(X_t)_{t\geq 0}$ be a continuous-time Markov chain with state space S and transition rates $[q(i,j)]_{i,j\in S}$. Let T be a stopping time. For some $i \in S$, suppose that $\mathbb{P}(X_T = i) > 0$. Then conditioned on $X_T = i$, $(X_{T+t})_{t\geq 0}$ is a Markov chain with the same transition rates $[q(i,j)]_{i,j\in S}$, independent from $(X_t)_{0\leq t\leq T}$.*

The proof follows the same idea as the proof of the discrete-time Strong Markov property, Proposition 3.6. To make the terminology fully rigorously precise (i.e. in particular what it means for a stopping time to depend only on an uncountable collection like $\{X_s \colon s \leq t\}$) really requires measure theory.

We can use the Strong Markov property to develop tools for "first-step" computations with continuous time Markov chains, in complete analogy with the techniques in Lectures 2 and 3. For example, for a subset $A \subseteq S$ of the state space, as usual define the **hitting time** $\tau_A = \min\{t \geq 0 \colon X_t \in A\}$. How likely is it that X_t hits A before B, for two disjoint subsets $A, B \subset S$? To answer, for each state i set $h(i) = \mathbb{P}_i(\tau_A < \tau_B)$. Then $h(i) = 1$ for $i \in A$ and $h(i) = 0$ for $i \in B$. For $i \in A \cup B$, by the Strong Markov property:

$$h(i) = \sum_{j\in S} \mathbb{P}_i(X_{J_1=j})\mathbb{P}_i(\tau_A < \tau_B | X_{J_1} = j) = \sum_{j\in S} p(i,j)h(j) = \sum_{j\in S} \frac{q(i,j)}{q(i)} h(j). \tag{19.1}$$

(That is: $\mathbb{P}_i(\tau_A < \tau_B | X_{J_1} = j) = \mathbb{P}_j(\tau_A - J_1 < \tau_B - J_1) = \mathbb{P}_j(\tau_A < \tau_B)$ by the Strong Markov property with stopping time equal to the jump time J_1.) Similarly, we can compute the expected hitting time $g(i) = \mathbb{E}_i[\tau_A]$ for $i \notin A$ by

$$g(i) = \sum_{j\in S} \mathbb{P}_i(X_{J_1} = j)\mathbb{E}_i[\tau_A | X_{J_1} = j] = \sum_{j\in S} \frac{q(i,j)}{q(i)} \mathbb{E}_i[\tau_A | X_{J_1} = j].$$

Now, define $Y_t = X_{J_1+t}$. Then $\tau_A = \min\{t \geq 0 \colon X_t \in A\} = \min\{t + J_1 \colon X_{t+J_1} \in A\} = \min\{t + J_1 \colon Y_t \in A\} = J_1 + \tau'_A$, where τ'_A is the hitting time of A for the process $(Y_t)_{t\geq 0}$. By the Strong Markov property

$$\mathbb{E}_i[\tau_A | X_{J_1=j}] = \mathbb{E}[\tau'_A + J_1 | X_0 = i, X_{J_1} = j] = \mathbb{E}_j[\tau_A] + \mathbb{E}_i[J_1].$$

Conditoned on $X_0 = i$, J_1 is $\mathrm{Exp}(q(i))$, so its expectation is $\frac{1}{q(i)}$. Altogether, then, we have

$$g(i) = \sum_{j\in S} \frac{q(i,j)}{q(i)} \left(g(j) + \frac{1}{q(i)} \right).$$

Multiplying through by $q(i)$, and using the fact that $\sum_{j\in S} \frac{q(i,j)}{q(i)} = 1$, this gives us the equations

$$q(i)g(i) = 1 + \sum_{j\in S} q(i,j)g(j). \tag{19.2}$$

Equations 19.1 and 19.2 can be used to iteratively calculate $\mathbb{P}(\tau_A < \tau_B)$ and $\mathbb{E}[\tau_A]$ just as in the the discrete-time setting.

19.2. **Embedded Jump Chain.** Given a continuous-time Markov chain, we have already defined the first jump-time $J_1 = \min\{t \geq 0 \colon X_t \neq X_0\}$. This is part of a whole sequence of jump times. We can inductively define

$$J_0 = 0, \quad J_{n+1} = \min\{t \geq J_n \colon X_t \neq X_{J_n}\}.$$

By the strong Markov property, for any states $i_0, i_1, \ldots, i_n \in S$,

$$\begin{aligned}
&\mathbb{P}(X_{J_n} = i_n | X_{J_{n-1}} = i_{n-1}, \ldots, X_{J_1} = i_1, X_0 = i_0) \\
&= \mathbb{P}(X_{J_n - J_1} = i_n | X_{J_{n-1} - J_1} = i_{n-1}, \ldots, X_{J_2 - J_1} = i_2, X_0 = i_1) \\
&= \mathbb{P}(X_{J_n - J_2} = i_n | X_{J_{n-1} - J_2} = i_{n-1}, \ldots, X_{J_3 - J_2} = i_3, X_0 = i_2) \\
&\vdots \\
&= \mathbb{P}(X_{J_n - J_{n-1}} = i_n | X_0 = i_{n-1}) \\
&= \mathbb{P}(X_{J_n} = i_n | X_{J_{n-1}} = i_{n-1})
\end{aligned}$$

where the final step is from time-homogeneity, and the intermediate kth step follows by restarting the process at time J_k (note that $(J_n - J_{k-1}) - (J_k - J_{k-1}) = J_n - J_k$, so the differences telescope). This shows that the discrete-time stochastic process

$$Y_n = X_{J_n},$$

called the **embedded jump chain** of $(X_t)_{t \geq 0}$, is a Markov chain. Moreover, its transition probabilities are $\mathbb{P}(Y_1 = j | Y_0 = i) = \mathbb{P}(X_{J_1} = j | X_0 = i) = p(i,j) = \frac{q(i,j)}{q(i)}$.

Now, let us condition on a particular trajectory for the jump chain: let $A = \{Y_0 = i_0, \ldots, Y_n = i_n\}$. Let us consider the so-called **sojourn times** $S_k = J_k - J_{k-1}$ for $k \geq 1$. Then we have for any $0 < s_1 < \cdots < s_n < \infty$

$$\begin{aligned}
&= \mathbb{P}(S_1 > s_1, \ldots, S_n > s_n | A) \\
&= \mathbb{P}(J_1 > s_1, J_2 - J_1 > s_2, \ldots, J_n - J_{n-1} > s_n | A) \\
&= \mathbb{P}(J_1 > s_1, J_2 > s_2 + J_1, \ldots, J_n > s_n + J_{n-1} | A) \\
&= \mathbb{P}_A(J_1 > s_1) \mathbb{P}_A(J_2 > s_2 + J_1 | J_1 > s_1) \cdots \mathbb{P}_A(J_n > s_n + J_{n-1} | J_1 > s_1, \ldots, J_{n-1} > s_{n-1}).
\end{aligned}$$

By the Strong Markov property,

$$\begin{aligned}
\mathbb{P}_A(J_k > s_k + J_{k-1} | J_1 > s_1, \ldots, J_{k-1} > s_{k-1}) &= \mathbb{P}_A(J_k > s_k + J_{k-1} | J_{k-1} > s_{k-1}) \\
&= \mathbb{P}(J_1 > s_k | X_0 = i_{k-1}) = e^{-q(i_{k-1}) s_k}.
\end{aligned}$$

(I.e. restarting the process at time J_{k-1}, the next jump time J_k becomes the first jump-time of the shifted process; the conditioning $X_{J_{k-1}} = Y_{k-1} = i_{k-1}$ translates to $X_0 = i_{k-1}$ for the shifted process.) So, we have shown that

$$\mathbb{P}(S_1 > s_1, \ldots, S_n > s_n | A) = e^{-q(i_0) s_1} \cdots e^{-q(i_{n-1}) s_n}$$

which is precisely the joint distribution of independent exponential random with parameters $q(i_0), \ldots, q(i_{n-1})$. In particular, this shows that

Proposition 19.2. *Conditioned on $Y_0, \ldots, Y_{n-1}$, the sojourn times $S_1, \ldots, S_n$ are independent exponential random variables with $S_k \sim \mathrm{Exp}(q(Y_{k-1}))$.*

19.3. **Infinitesimal Description.** We know that the transition rates $q(i,j)$ completely determine the Markov chain, and give a convenient description in terms of exponential jumps. Since they give a complete description, they must therefore encode the transition probaiblities $p_t(i,j) = \mathbb{P}_i(X_t = j)$ for all $t \geq 0$. One way to see how $p_t(i,j)$ is determined from $q(i,j)$ is through the following so-called *infinitesimal description* of the chain.

Theorem 19.3. *Let $(X_t)_{t\geq 0}$ be a Markov chain with state space S and transition rates $[q(i,j)]_{i,j\in S}$. Then the transition probabilities $p_t(i,j) = \mathbb{P}_i(X_t = j)$ satisfy*

$$p_t(i,i) = 1 - q(i)t + o(t) \qquad \textit{for } i \in S$$
$$p_t(i,j) = q(i,j)t + o(t) \qquad \textit{for } i \neq j \in S.$$

In the statement of Theorem 19.3, we have (as usual) $q(i) = \sum_{j\in S} q(i,j)$. The notation $f(t) = g(t) + o(t)$ is short-hand for

$$\lim_{t\downarrow 0} \frac{f(t) - g(t)}{t} = 0.$$

That is, to first order f and g agree near 0. So the statement is that, for small $t > 0$, $p_t(i,i)$ is very close to $1 - q(i)t$, and $p_t(i,j)$ is very close to $q(i,j)t$ for $i \neq j$.

Proof. First, for the $i = j$ case, we have $p_t(i,i) = \mathbb{P}_i(X_t = i)$. The event $\{X_t = i\}$ contains the event $\{\forall s \leq t\ X_s = i\}$ which (conditioned on $X_0 = i$) is the event $\{J_1 > t\}$. Thus

$$p_t(i,i) = \mathbb{P}_i(X_t = i) \geq \mathbb{P}_i(J_1 > t) = e^{-q(i)t} = 1 - q(i)t + o(t). \tag{19.3}$$

For $j \neq i$, we can similarly estimate as follows: if $J_1 \leq t$, the first jump state is $Y_1 = j$, and the following sojurn time $S_2 > t$, then the process jumps to j before time t and doesn't leave it until time $J_2 = J_1 + S_2 > t$, thus $X_t = j$. Hence

$$p_t(i,j) = \mathbb{P}_i(X_t = j) \geq \mathbb{P}_i(J_1 \leq t, Y_1 = j, S_2 > t) = \mathbb{P}_i(Y_1 = j)\mathbb{P}_i(J_1 \leq t, S_2 > t | Y_1 = j).$$

By definition $\mathbb{P}_i(Y_1 = j) = \mathbb{P}_i(X_{J_1} = j) = p(i,j)$. By Proposition 19.2, since $J_1 = S_1$, we have

$$\mathbb{P}_i(J_1 \leq t, S_2 > t | Y_1 = j) = \mathbb{P}_i(J_1 \leq t | Y_1 = j)\mathbb{P}_i(S_2 > t | Y_1 = j) = (1 - e^{-q(i)})e^{-q(j)t}.$$

Thus, we have the inequality

$$p_t(i,j) \geq p(i,j)(1-e^{-q(i)t})e^{-q(j)t} = p(i,j)(q(i)t+o(t))(1-q(j)t+o(t)) = q(i,j)t+o(t). \tag{19.4}$$

(The final equality follows from the definition $q(i,j) = q(i)p(i,j)$.) Now, fixing i and summing over all states j, we have from Inequalities 19.3 and 19.4

$$\sum_{j\in S} p_t(i,j) \geq 1 - q(i)t + o(t) + \sum_{j\neq i} q(i,j)t + o(t).$$

Since $q(i) = \sum_j q(i,j)$ and $q(i,i) = 0$, this shows that

$$\sum_{j\in S} p_t(i,j) \geq 1 + o(t).$$

Hence, if any one of the inequailities in 19.3 or 19.4 were strict, we would have the strict inequality $\sum_{j\in S} p_t(i,j) > 1 + o(t)$ and so for small enough t the sum would exceed 1. This contradicts the stochastic condition $\sum_{j\in S} p_t(i,j) = 1$, and so the inequalites must in fact be equalities. $\square$

The infinitesimal description is a very handy way to identify a Markov chain. Indeed, the chain is completely determined by the parameters $q(i)$ and $q(i,j)$ for $i,j \in S$, so to fully understand the chain one need only understand $p_t(i,j)$ to first order in t. This can be used to identify a process quickly.

Example 19.4. Let $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ be independent Poisson processes, with parameters λ_1 and λ_2. The infinitesimal description of Theorem 19.3 can be used to prove that $X_t + Y_t$ is a Poisson process with paramter $\lambda_1 + \lambda_2$; this is an exercise on Homework 5.

20. LECTURE 20: MAY 16, 2012

20.1. **Kolmogorov's Forward and Backward Equations.** The infinitesimal decription gives rise to a system of differential equations for $p_t(i,j)$. It can be described elegantly in terms of the following matrix.

Definition 20.1. *Let $(X_t)_{t\geq 0}$ be a continuous-time Markov chain with state space S and transition rates $[q(i,j)]_{i,j\in S}$. Define the* **infinitesimal generator** *matrix $\mathbf{A}$ of the chain as follows: for $i \neq j$, $[\mathbf{A}]_{ij} = q(i,j)$, and $[\mathbf{A}]_{ii} = -q(i) = -\sum_{j\in S} q(i,j)$.*

So we can specify a Markov chain by specifying its infinitesimal generator.

Theorem 20.2 (Kolmogorov's Forward and Backward Equations)**.** *Le $(X_t)_{t\geq 0}$ be a continuous-time Markov chain with state space S and infinitesimal generator $\mathbf{A}$. Let $\mathbf{P}_t$ denote the matrix with entries $[\mathbf{P}_t]_{ij} = p_t(i,j)$ for $i,j \in S$. Then*

$$\frac{d}{dt}\mathbf{P}_t = \mathbf{P}_t\mathbf{A} = \mathbf{A}\mathbf{P}_t. \tag{20.1}$$

The first equality is called the **Kolomogorov Forward Equation**; the **Kolmogorov Backward Equation** is the equality $\frac{d}{dt}\mathbf{P}_t = \mathbf{A}\mathbf{P}_t$. At least in the case of a finite state-space, the second equality follows from the first, as will be evident in the discussion following the proof.

Proof. Fix $t \geq 0$ and $h > 0$. By the Markov property,

$$\begin{aligned} p_{t+h}(i,j) = \mathbb{P}_i(X_{t+h} = j) &= \sum_{k\in S} \mathbb{P}_i(X_t = k)\mathbb{P}_i(X_{t+h} = j | X_t = k) \\ &= \sum_{k\in S} p_t(i,k)\mathbb{P}_k(X_h = j) \\ &= \sum_{k\in S} p_t(i,k)p_h(k,j). \end{aligned}$$

From the infinitesimal description, we have $p_h(j,j) = 1 - q(j)h + o(h)$ and $p_h(k,j) = q(k,j)h + o(h)$ when $k \neq j$. Thus

$$p_{t+h}(i,j) = p_t(i,j)(1 - q(j)h + o(h)) + \sum_{k\neq j} p_t(i,k)(q(k,j)h + o(h))$$

which says

$$p_{t+h}(i,j) - p_t(i,j) = -p_t(i,j)q(j)h + \sum_{k\neq j} p_t(i,k)q(k,j)h + o(h).$$

Dividing both sides by h and taking the limit as $h \to 0$ gives

$$\frac{d}{dt}[\mathbf{P}_t]_{ij} = \frac{d}{dt}p_t(i,j) = -p_t(i,j)q(j) + \sum_{k\neq j} p_t(i,k)q(k,j) = \sum_{k\in S}[\mathbf{P}]_{ik}[\mathbf{A}]_{kj}$$

which proves the Forward Equation. The Backward Equation is proved similarly, in that case conditioning on $X_h = k$ rather than $X_t = k$. □

From the elementary theory of ODEs, we can now easily solve the Kolomogorov Forward and Backward Equations, at least in the case of a finite state space.

Corollary 20.3. *Let $(\mathbf{X}_t)_{t\geq 0}$ be a finite-state continuous time Markov chain with infinitesimal generator $\mathbf{A}$. Then the transition probabilities are given by*

$$\mathbf{P}_t = e^{t\mathbf{A}}.$$

The exponential of a matrix can be defined, for example, by the power-series $e^{t\mathbf{A}} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{A}^k$. The essential point here is that $\frac{d}{dt}e^{t\mathbf{A}} = \mathbf{A}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A}$, which shows that $\mathbf{P}_t = e^{t\mathbf{A}}$ satisfies the Kolmogorov Forward and Backward equations. In addition, we have $\epsilon^{0\mathbf{A}} = \mathbf{I}$, the identity matrix, and indeed $[\mathbf{P}_0]_{ij} = \mathbb{P}(X_0 = j|X_0 = i) = \delta_{ij} = [\mathbf{I}]_{ij}$; hence, uniqueness of the solution to an ODE with given initial conditions proves the corollary. It is worth noting that, when there are infinitely-many states, the matrix exponential may not even make sense, and there can in fact be more than one solution to the equations. In this case, at can be proved (with probabilistic techniques) that the Kolmogorov Backward equation *possesses a unique minimal non-negative solution*, and this solution gives the transition probabilities $p_t(i, j)$.

Example 20.4. Consider an arbitrary two-state chain. The infinitesimal generator $\mathbf{A}$ then has the form

$$\mathbf{A} = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}$$

where $\alpha = q(1, 2)$ and $\beta = q(2, 1)$. We can calculate that

$$\mathbf{A}^2 = \begin{bmatrix} \alpha^2 + \alpha\beta & -\alpha^2 - \alpha\beta \\ -\alpha\beta - \beta^2 & \alpha\beta + \beta^2 \end{bmatrix} = -(\alpha + \beta)\begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix} = -(\alpha + \beta)\mathbf{A}.$$

Thus $\mathbf{A}^n = [-(\alpha + \beta)]^{n-1}\mathbf{A}$, and so we have

$$e^{t\mathbf{A}} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{A}^k = \mathbf{I} + \sum_{k=1}^{\infty} \frac{t^k[-(\alpha + \beta)]^{k-1}}{k!}\mathbf{A} = \mathbf{I} - \frac{1}{\alpha + \beta}\mathbf{A}\sum_{k=1}^{\infty} \frac{[-(\alpha + \beta)t]^k}{k!}$$

which we can write as

$$\mathbf{P}_t = e^{t\mathbf{A}} = \mathbf{I} - \frac{1}{\alpha + \beta}\mathbf{A}\left(e^{-(\alpha+\beta)t} - 1\right) = \mathbf{I} + \frac{1 - e^{-(\alpha+\beta)t}}{\alpha + \beta}\mathbf{A}.$$

So we have a closed-form expression for the transition probabilities of any 2-state continuous-time Markov chain.

Example 20.5. Let $(X_t)_{t\geq 0}$ be a Poisson process with rate $\lambda > 0$. So the transition rates are $q(i, j) = \lambda\delta_{j,i+1}$. It follows that $q(i) = \sum_{j\neq i} q(i, j) = \lambda$. Hence, the Kolmogorov Forward Equations are

$$\frac{d}{dt}p_t(i, 0) = -p_t(i, 0)q(0) + \sum_{k\neq 0} p_t(i, k)q(k, 0) = -\lambda p_t(i, 0)$$

$$\frac{d}{dt}p_t(i, j) = -p_t(i, j)q(j) + \sum_{k\neq j} p_t(i, j)q(k, j) = -\lambda p_t(i, j) + \lambda p_t(i, j - 1), \quad j \geq 1.$$

Since this is an infinite system of ODEs, we cannot simply solve it by taking a matrix exponential; we do not even have a general theorem to tell us that there is a unique solution.

But we can, in fact, solve the system and prove it has a unique solution, as follows. Fix i, and consider the probability generating function

$$\varphi_i(t,z) = \sum_{j=0}^{\infty} p_t(i,j) z^j.$$

All the coefficients $p_t(i,j)$ are ≤ 1 in modulus, so the series converges (uniformly) for $|z| < 1$. Taking the time derivative can then be done term-by-term:

$$\frac{\partial}{\partial t}\varphi_i(t,z) = \sum_{j=0}^{\infty} \left[\frac{d}{dt} p_t(i,j)\right] z^j = \frac{d}{dt} p_t(i,0) + \sum_{j=1}^{\infty} \left[\frac{d}{dt} p_t(i,j)\right] z^j.$$

Kolmogorov's Forward Equations then give

$$\begin{aligned}
\frac{\partial}{\partial t}\varphi_i(t,z) &= -\lambda p_t(i,0) + \sum_{j=1}^{\infty} \left[-\lambda p_t(i,j) + \lambda p_t(i,j-1)\right] z^j \\
&= -\lambda \sum_{j=0}^{\infty} p_t(i,j) z^j + \lambda \sum_{j=1}^{\infty} p_t(i,j-1) z^j \\
&= -\lambda \varphi_i(t,z) + \lambda \sum_{k=0}^{\infty} p_t(i,k) z^{k+1} \\
&= -\lambda \varphi_i(t,z) + \lambda z \varphi_i(t,z).
\end{aligned}$$

In other words, the function $\varphi_i(t,z)$ satisfies the partial differential equation $\partial_t \varphi_i(t,z) = \lambda(z-1)\varphi_i(t,z)$. Dividing through by $\varphi_i(t,z)$ gives

$$\frac{\partial}{\partial t} \ln \varphi_i(t,z) = \frac{1}{\varphi_i(t,z)} \frac{\partial}{\partial t} \varphi_i(t,z) = \lambda(z-1).$$

It follows immediately that the solution is

$$\varphi_i(t,z) = \varphi_i(0,z) e^{\lambda t(z-1)} = z^i e^{\lambda t(z-1)} = e^{-\lambda t} z^i e^{\lambda t z}$$

where the constant term $\varphi_i(0,z) = z^i$ comes from the fact that $p_i(0,j) = \delta_{ij}$. We can then recover the desired coefficients $p_t(i,j)$ by expanding the power-series

$$\sum_{j=0}^{\infty} p_t(i,j) z^j = \varphi_i(t,z) = e^{-\lambda t} z^i \sum_{k=0}^{\infty} \frac{1}{k!} (\lambda t z)^k = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} z^{k+i}.$$

Comparing coefficients, we see that $p_t(i,j) = 0$ for $j < i$, while for $j \geq i$

$$p_t(i,j) = e^{-\lambda t} \frac{(\lambda t)^{j-i}}{(j-i)!}.$$

In particular, when $i = 0$, we have

$$\mathbb{P}(X_t = j | X_0 = 0) = p_t(0,j) = e^{-\lambda t} \frac{(\lambda t)^j}{j!},$$

confirming that X_t is a Poisson random variable with rate λt.

20.2. **More in Poisson Processes.** The Poisson process with rate λ has a very special property that is not shared by all Markov processes. To get a hint of what it is, consider the embedded jump chain $(Y_n)_{n\geq 0}$ of such a Poisson process $(X_t)_{t\geq 0}$. Since $q(i,j) = 0$ unless $j = i+1$, we have $X_{J_{n+1}} = X_{J_n} + 1$. It follows that $Y_n = Y_0 + n$; the jump chain is just a deterministic one-step increase chain. Now, by Proposition 19.2, the sojourn times are independent exponentials conditioned on the trajectory of (Y_n); but since there is only one trajectory of (Y_n), and the jump rates $q(i,i+1) = \lambda$ do not depend on i, it follows that the sojourn times are just plan i.i.d. Indeed, we can compute $\mathbb{P}(S_1 > s_1, \dots, S_n > s_n)$ as the conditional sum

$$\sum_{i_0,\dots,i_n} \mathbb{P}(S_1 > s_1, \dots, S_n > s_n | Y_0 = i_0, \dots, Y_n = i_n)\mathbb{P}(Y_0 = i_0, \dots, Y_n = i_n).$$

All of the terms $\mathbb{P}(Y_0 = i_0, \dots, Y_n = i_n)$ are 0 except for those for which $i_k = i_0 + k$ for $1 \leq k \leq n$. So we have just the single sum

$$\sum_i \mathbb{P}(S_1 > s_1, \dots, S_n > s_n | Y_0 = i, \dots, Y_n = i+n)\mathbb{P}(Y_0 = i, \dots, Y_n = i+n).$$

Because the jump chain is deterministic, $\mathbb{P}(Y_0 = i, \dots, Y_n = i+n) = \mathbb{P}(Y_0 = i)$. By Proposition 19.2,

$$\begin{aligned}\mathbb{P}(S_1 > s_1, \dots, S_n > s_n | Y_0 = i, \dots, Y_n = i+n) &= e^{-q(i)s_1} \cdots e^{-q(i+n-1)s_n} \\ &= e^{-\lambda(s_1+\cdots+s_n)}.\end{aligned}$$

So in total we have

$$\mathbb{P}(S_1 > s_1, \dots, S_n > s_n) = \sum_i e^{-\lambda(s_1+\cdots+s_n)}\mathbb{P}(Y_0 = i) = e^{-\lambda(s_1+\cdots+s_n)}.$$

That is:

Proposition 20.6. *If $(X_t)_{t\geq 0}$ is a Poisson process with rate λ, then its sojourn times are i.i.d.* $\mathrm{Exp}(\lambda)$ *random variables.*

This gives us a slightly different way to describe/construct a Poisson process. We can begin with a sequence $S_1, S_2, \dots$ of i.i.d. $\mathrm{Exp}(\lambda)$ random variables, define $J_n = S_1 + \cdots + S_n$, and then define $X_t = n$ for $J_n \leq t < J_{n+1}$; then $(X_t)_{t\geq 0}$ is a Poisson process of rate λ.

Theorem 20.7. *Let $(X_t)_{t\geq 0}$ be a Poisson process of rate λ (with $X_0 = 0$). Then for any $s \geq 0$, the process $\widetilde{X}_t = X_{t+s} - X_s$ is a Poisson process of rate λ, independent of $\{X_u \colon 0 \leq u \leq s\}$.*

Proof. It suffices to prove the claim under the conditioning $X_s = i$ for an arbitrary fixed state i. Indeed, if B is any event determined by the process $\widetilde{X}_t$ for $t \geq 0$, and A is any event determined by $\{X_u \colon 0 \leq u \leq s\}$, if $A$and B are independent conditioned on $\{X_s = i\}$ then

$$\mathbb{P}(A, B) = \sum_i \mathbb{P}(A, B | X_s = i)\mathbb{P}(X_s = i) = \sum_i \mathbb{P}(A | X_s = i)\mathbb{P}(B | X_s = i)\mathbb{P}(X_s = i).$$

Note that, by the conditional independence claim, we have B and $\{X_s = i\}$ are independent conditioned on $\{X_s = i\}$, which means simply that B and $\{X_s = i\}$ are independent. Thus $\mathbb{P}(B | X_s = i) = \mathbb{P}(B)$, and so

$$\mathbb{P}(A, B) = \sum_i \mathbb{P}(A | X_s = i)\mathbb{P}(B)\mathbb{P}(X_s = i) = \mathbb{P}(B)\sum_i \mathbb{P}(A | X_s = i)\mathbb{P}(X_s = i) = \mathbb{P}(B)\mathbb{P}(A)$$

as required.

Let $J_1, J_2, \ldots$ be the jump times and $S_1, S_2, \ldots$ the sojourn times. The event $\{X_s = i\}$ can be written exclusively in terms of these variables: $X_s = i$ means that there have been exactly i jumps (and no more) by time s. Hence

$$\{X_s = i\} = \{J_i \leq s < J_{i+1}\} = \{J_i \leq s\} \cap \{S_{i+1} > s - J_i\}.$$

By definition $\widetilde{X}_0 = X_s - X_0 = X_s = i$. The first jump time $\widetilde{J}_1$ of $\widetilde{X}_t$ is the portion of S_{i+1} to the future of s; since S_{i+1} is the interval between J_i and J_{i+1}, this means that

$$\widetilde{J}_1 = \widetilde{S}_1 = S_{i+1} - (s - J_i) = J_{i+1} - s.$$

For further sojourn times, we simply have a time shift

$$\widetilde{S}_n = S_{i+n}, \qquad n \geq 2.$$

We can then calculate the distirbution of $\widetilde{S}_1, \ldots, \widetilde{S}_n$ conditioned on $S_1, \ldots, S_i$ and the event $\{X_s = i\} = \{J_i \leq s\} \cap \{J_{i+1} > s\}$.

$$\begin{aligned}
&\mathbb{P}(\widetilde{S}_1 > \tilde{s}_1, \ldots, \widetilde{S}_n > \tilde{s}_n | S_1 > s_1, \ldots, S_i > s_i, X_s = i) \\
&= \mathbb{P}(S_{i+n} > \tilde{s}_n, \ldots, S_{i+2} > \tilde{s}_2, J_{i+1} > \tilde{s}_1 + s | J_{i+1} > s, J_i \leq s, S_i > s_i, \ldots, S_1 > s_1) \\
&= \frac{\mathbb{P}(S_{i+n} > \tilde{s}_n, \ldots, S_{i+2} > \tilde{s}_2, J_{i+1} > \tilde{s}_1 + s, J_i \leq s, S_i > s_i, \ldots, S_1 > s_1)}{\mathbb{P}(J_{i+1} > s, J_i \leq s, S_i > s_i, \ldots, S_1 > s_1)}. \qquad (20.2)
\end{aligned}$$

Both numerator and denominator can be computed as multiple integrals. Recalling that $J_i = S_1 + \cdots + S_i$, and that $S_1, S_2, \ldots$ are i.i.d. $\mathrm{Exp}(\lambda)$ random variables, the denominator of Equation 20.2 is

$$\begin{aligned}
&\mathbb{P}(S_1 + \cdots + S_{i+1} > s, S_1 + \cdots + S_i \leq s, S_i > s_i, \ldots, S_1 > s_1) \\
&= \int_{\substack{u_1 > s_1, \ldots, u_i > s_i \\ u_1 + \cdots + u_i \leq s}} \int_{u_{i+1} \geq 0} \mathbb{1}_{\{u_1 + \cdots + u_{i+1} > s\}} \lambda^{i+1} e^{-\lambda(u_1 + \cdots + u_{i+1})} \, du_1 \cdots du_{i+1} \\
&= \int_{\substack{u_1 > s_1, \ldots, u_i > s_i \\ u_1 + \cdots + u_i \leq s}} \int_{u_{i+1} \geq s - (u_1 + \cdots + u_i)} \lambda^{i+1} e^{-\lambda(u_1 + \cdots + u_{i+1})} \, du_1 \cdots du_{i+1}.
\end{aligned}$$

The inside integral can be computed as

$$\begin{aligned}
\lambda^i e^{-\lambda(u_1 + \cdots + u_i)} \int_{s - (u_1 + \cdots + u_n)}^{\infty} \lambda e^{-\lambda u_{i+1}} \, du_{i+1} &= \lambda^i e^{-\lambda(u_1 + \cdots + u_i)} e^{-\lambda(s - (u_1 + \cdots + u_i))} \\
&= \lambda^i e^{-\lambda s}.
\end{aligned}$$

Hence, the denominator of Equation 20.2 is simply equal to

$$\lambda^i e^{-\lambda s} \int_{\substack{u_1 > s_1, \ldots, u_i > s_i \\ u_1 + \cdots + u_i \leq s}} du_1 \cdots du_i.$$

A totally analogous argument applied to the numerator of Equation 20.2, integrating out all the variables $u_{i+1}, \ldots, u_{i+n}$, yields

$$\begin{aligned}
&\mathbb{P}(S_{i+n} > \tilde{s}_n, \ldots, S_{i+2} > \tilde{s}_2, J_{i+1} > \tilde{s}_1 + s, J_i \leq s, S_i > s_i, \ldots, S_1 > s_1) \\
&\qquad = \; e^{-\lambda \tilde{s}_n} \cdots e^{-\lambda \tilde{s}_2} \cdot \lambda^i e^{-\lambda(s + \tilde{s}_1)} \int_{\substack{u_1 > s_1, \ldots, u_i > s_i \\ u_1 + \cdots + u_i \leq s}} du_1 \cdots du_i.
\end{aligned}$$

Taking the ratio then yields

$$\mathbb{P}(\widetilde{S}_1 > \tilde{s}_1, \dots, \widetilde{S}_n > \tilde{s}_n | S_1 > s_1, \dots, S_i > s_i, X_s = i) = e^{-\lambda \tilde{s}_1} \cdots e^{-\lambda \tilde{s}_n}.$$

This shows that, conditioned on $X_s = i$, $\widetilde{S}_1, \dots, \widetilde{S}_n$ are independent $\mathrm{Exp}(\lambda)$ random variables, all independent from $S_1, \dots, S_i$. Since it is clear that $\widetilde{X}_t = X_{s+t} - X_s$ also has embedded jump chain $\widetilde{Y}_n = n$, we see that $\widetilde{X}_t$ is a Poisson process of rate λ.

The brief discussion following Proposition 20.6 shows that the process X_t is completely determined by its Sojourn times. Indeed, conditioned on $X_s = i$, we can write X_r for $r \leq s$ as

$$X_r = \sum_{j=1}^{i} \mathbb{1}_{\{S_j \leq r\}}. \tag{20.3}$$

Similarly, for any state k the event $\widetilde{X}_t = k$ is equal to $\{\widetilde{S}_1 + \cdots + \widetilde{S}_k \leq t < \widetilde{S}_1 + \cdots + \widetilde{S}_k + \widetilde{S}_{k+1}\}$, and by the previous paragraph this event is independent from $S_1, \dots, S_i$ and therefore from X_r for $r \leq s$. This concludes the proof. □

20.3. **Independent Increments.** Given a stochastic process $(X_t)_{t \geq 0}$, its **increments** are the random variables $\{X_t - X_s \colon 0 \leq s < t < \infty\}$. If $(X_t)_{t \geq 0}$ is a Poisson process (or any counting process, so that X_t represents the number of events that have occurred by time t), then the increment $X_t - X_s$ is the number of events that have occurred in the interval $(s, t]$. It is because of the following corollary to Theorem 20.7 that the Poisson process is so useful in applications.

Corollary 20.8. *If $(X_t)_{t \geq 0}$ is a Poisson process with rate λ, then for any sequence $0 \leq t_0 < t_1 < \cdots < t_n$ the increments $X_{t_n} - X_{t_{n-1}}, \dots, X_{t_1} - X_{t_0}$ are independent, and each increment $X_t - X_s$ is a Poisson random variable with rate $\lambda(t - s)$. Moreover, these properties uniquely characterize the Poisson process of rate λ.*

Proof. First, write any increment $X_t - X_s = X_{(t-s)+s} - X_s$. By Theorem 20.7, $\widetilde{X}_u = X_{u+s} - X_s$ is a Poisson process with rate λ, and hence $X_t - X_s = \widetilde{X}_{t-s}$ is $\mathrm{Poisson}(\lambda(t - s))$ as claimed. For independence, proceed by induction. By Theorem 20.7, $\widetilde{X}_t = X_{t+t_n} - X_{t_n}$ is independent of X_s for $s \leq t_n$; in particular, it is independent of $X_{t_n} - X_{t_{n-1}}, \dots, X_{t_1} - X_{t_0}$, which are themselves all independent by the inductive hypothesis. It follows that for any $t_{n+1} = t_n + t > t_n$, all the increments $X_{t_{n+1}} - X_{t_n}, X_{t_n} - X_{t_{n-1}}, \dots, X_{t_1} - X_{t_0}$ are independent.

For the converse, note that independent increments with Poisson distributions completely determine the joint distributions of $(X_{t_0}, \dots, X_{t_n})$ for any choice of times $0 \leq t_0 < \cdots < t_n$. Hence, there is a unique process with these properties. The above proof shows that the Poisson process of rate λ is such a process, hence it is the only one.

□

Not all Markov chains have independent increments, although the converse is true (and easy to prove): if a stochastic process (with a countable state space and right-continuous trajectories) has independent increments, then it is a Markov chain; if the increments have the property that the distirbution of $X_t - X_s$ only depends on $t - s$ (we say the process has *stationary increments*), then $(X_t)_{t \geq 0}$ is time-homogenous as well. The class of such stochastic processes is called the **Lévy processes**. Most Markov chains are not Lévy processes; the Poisson process is almost alone in this regard (with countable state space).

21. LECTURE 21: MAY 18, 2012

21.1. **Birth and Death Chains.** A continuous time Markov chain is called a **birth and death chain** if

- the state space is $S = \{0, 1, 2, \ldots\}$
- the transition rates are $q(i, i+1) = \lambda_i \geq 0$ for $i \geq 0$, $q(i, i-1) = \mu_i \geq 0$ for $i \geq 1$, and $q(i, j) = 0$ if $j \neq i \pm 1$.

If all the $\mu_i = 0$, it is a **pure birth process**; if all the $\lambda_i = 0$ it is a **pure death process**. The Poisson process with rate λ is a pure birth process with $\lambda_i = \lambda$ for all i.

Example 21.1 (Kingman's Coalescent Process)**.** Consider the pure death process $(X_t)_{t\geq 0}$ with $\mu_1 = 0$ and $\mu_k = \binom{k}{2}$ for $k \geq 2$. This is a model of genetic coalescence: tracking ancestor lines backward in time. It models the following: suppose we take a sample (size N) from a larrge population (of genes, for example), and track their ancestors back in time. As soon as we find a common ancestor, that shrinks the sample size by 1; the set of pairs of members of the population has size $\binom{N}{2}$, so there are this many possible coalescent points. Once one is reached, the process starts over with the new coalesced population of size $N - 1$.

What we are most interested in is $T = \min\{t \colon X_t = 1\}$, the *Time to Most Recent Common Ancestor* (TMRCA). Conditioned on $X_0 = N$, this time is

$$T = S_1 + S_2 + \cdots + S_{N-1}$$

(here the sojourn times are given by $S_1 =$ amount of time spend in state N, $S_2 =$ amount of time spent in state $N-1$, etc.) Conditioned on the trajectory of the jump chain (which is then guaranteed to be $Y_j = N - j$), these times are independent and $S_j \sim \mathrm{Exp}(q(Y_{j-1})) = \mathrm{Exp}(\binom{N-j+1}{2})$. Therefore we have

$$\mathbb{E}[S_{N-k}] = \frac{1}{\binom{k+1}{2}} = \frac{2}{(k+1)k} = 2\left(\frac{1}{k} - \frac{1}{k+1}\right).$$

Hence

$$\mathbb{E}[T] = \sum_{j=1}^{N-1} \mathbb{E}[S_j] = \sum_{k=1}^{N-1} 2\left(\frac{1}{k} - \frac{1}{k+1}\right) = 2\left(1 - \frac{1}{N}\right).$$

Thus, in this model, as N grows large, the TMRCA approaches 2.

Another interesting statistic is the sum of the branch lengths in the ancestral tree, L. This gives a measure of the number of mutations in the population: if we assume mutations happen randomly along the branches with a constant rate, then the number of mutations is proportional to L. Conditioned on $X_0 = N$, the amount of time spent with population size k is S_{N-k+1}, so this branch length is given by $L = \sum_{k=2}^{N} kS_{N-k+1}$. So we have

$$\mathbb{E}[L] = \sum_{k=2}^{N} k\mathbb{E}[S_{N-k+1}] = \sum_{k=2}^{N} k\frac{2}{k(k-1)} = \sum_{k=2}^{N} \frac{1}{k-1} \approx \ln N.$$

Hence, the number of mutations in a sample of size N is proportional to $\ln N$.

Example 21.2 (Explosion)**.** Let $(X_t)_{t\geq 0}$ be a pure birth process with $\lambda_i = i^2$ for $i \in \mathbb{N}$. Condition on $X_0 = 1$. If T_N is the time to reach state N, then $T_{N^2} = S_1 + \cdots + S_N$, and so

$$\mathbb{E}[T_N] = \sum_{i=1}^{N} \mathbb{E}[S_i] = \sum_{i=1}^{N} \frac{1}{i^2}.$$

In particular, this means that $\lim_{N\to\infty} \mathbb{E}[T_N] = \frac{\pi^2}{6} < \infty$. Since all the random variables are positive, we can exchange the limit and the sum, and if we set $T = \sum_{i=1}^{\infty} S_i$ (the time to reach ∞), then $\mathbb{E}[T] = \frac{\pi^2}{6} < \infty$. In particular, this means $\mathbb{P}(T < \infty) = 1$.

The situation in Example 21.2 is called **explosion**. It presents a subtlety that we have thus far ignored: what happens after $T = \sum_{i=1}^{\infty} S_i$? The jump-chain, sojourn-times description is insufficient to answer this. Indeed, after this time, the process may restart in any state, and the Markov property is still obeyed. It is for this reason that the Kolmogorov Forward- and Backward- equations can have multiple solutions (despite always having the same initial date $P_0 = I$). If explosion occurs, then the very same transition rates describe different chains: one that stays at ∞ ever after the explosion, and others than restart in other states. Such chains are called *non-minimal*. We will not discuss such chains. But it is important to realize that the jump-chain sojourn-time description is only valid up to the time of explosion (if it occurs).

21.2. **Recurrence and Transience.** The same asymptotic notions for the behavior of a Markov chain apply in the continuous-time setting as the discrete-time setting.

Definition 21.3. *Let $(X_t)_{t\geq 0}$ be a continuous-time Markov chain with state space S, and let $i \in S$. Let $T = \min\{t > 0 \colon X_t = i\}$. The state i is called* **transient** *if $\mathbb{P}_i(T_i < \infty) = 0$; it is called* **recurrent** *if $\mathbb{P}_i(T_i < \infty) = 1$; it is called* **positive recurrent** *if $\mathbb{E}_i[T_i] < \infty$.*

Note that if $X_t = i$ then there is k with $J_k \leq t < J_{k+1}$, and so the the jump chain also satisfies $Y_k = i$. Since $Y_k = X_{J_k}$, it follows that Y_n returns to i iff X_t returns to i, and so the state i is recurrent (resp. transient) for the chain X_t iff it is recurrent (resp. transient) for the jump chain Y_n. We therefore need no new technology to analyze a continuous-time chain for recurrence/transience; we need only look at its jump chain. (In particular, as before, if the chain is irreducible – meaning for each $i, j \in S$ $p_t(i, j) > 0$ for some $t > 0$ – then either all states are recurrent or all states are transient.)

However, positive recurrence is determined not only by the property of return to a state, but how long it takes to get there, which is different for the two chains. Hence, positive recurrence is a different matter for continuous-time chains than discrete-time.

Example 21.4. Let $(X_t)_{t\geq 0}$ be a birth and death chain with parameters $\lambda_i = q(i, i+1) > 0$ for $i \geq 0$ and $\mu_i = q(i, i-1) > 0$ for $i \geq 1$. Since all the parameters are > 0, the chain is irreducible, so we may test any state (by default 0) for transience/recurrence of the chain. We can analyze such a chain in exactly the same manner we did for discrete-time birth and death chains in Section 9.1.

Let $h(i) = \mathbb{P}_i(\exists t \geq 0\ X_t = 0)$; then $h(0) = 1$. For $i \geq 1$, we do a first-jump analysis, we have

$$h(i) = \sum_{j\geq 0} \mathbb{P}_i(\exists t \geq 0\ X_t = 0 | X_{J_1} = j)\mathbb{P}_i(X_{J_1} = j).$$

By the strong Markov, we can restart the process at time J_1. Since $i \neq 0$, $X_t \neq 0$ for any $t < J_1$, and so $\mathbb{P}_i(\exists t \geq 0\ X_t = 0 | X_{J_1} = j) = \mathbb{P}_j(\exists t \geq 0\ X_t = 0) = h(j)$. Hence we have the usual recursion

$$h(i) = \sum_{j \geq 0} p(i,j)h(j), \quad i \geq 1.$$

For this chain $p(i,j) = q(i,j)/q(i) = 0$ unless $j = i \pm 1$. So we have $q(i) = \lambda_i + \mu_i$, and the recursion becomes

$$h(i) = \frac{\lambda_i}{\lambda_i + \mu_i} h(i+1) + \frac{\mu_i}{\lambda_i + \mu_i} h(i-1).$$

Rewriting this as

$$(\lambda_i + \mu_i)h(i) = \lambda_i h(i+1) + \mu_i h(i-1)$$

we can subtract to get

$$0 = \lambda_i h(i+1) - \lambda_i h(i) - \mu_i h(i) + \mu_i h(i-1) = \lambda_i [h(i+1) - h(i)] - \mu_i [h(i) - h(i-1)].$$

Hence, we have

$$h(i+1) - h(i) = \frac{\mu_i}{\lambda_i}[h(i) - h(i-1)], \quad i \geq 1.$$

By induction, then, we have

$$h(i+1) - h(i) = \frac{\mu_i \mu_{i-1} \cdots \mu_1}{\lambda_i \lambda_{i-1} \cdots \lambda_1}[h(1) - h(0)] \equiv \rho_i [h(1) - h(0)].$$

$$h(n) - h(0) = \sum_{i=1}^{n-1} [h(i+1) - h(i)] = [h(1) - h(0)] \sum_{i=1}^{n-1} \rho_i. \tag{21.1}$$

Now, we have two cases. Suppose that $\sum_{i=1}^{n-1} \rho_i \to \infty$ as $n \to \infty$. Since the left-hand-side $h(n) - h(0)$ is a difference of two probabilities, it stays bounded in $[0,1]$ for all n; the right-hand-side must as well, which means we must have $h(1) - h(0) = 0$. Since $h(0) = 1$, this means $1 = h(1) = \mathbb{P}_1(\exists t \geq 0\ X_t = 0)$. It then similarly follows that the left-hand-side is always 0, so $1 = h(n) = \mathbb{P}_n(\exists t \geq 0\ X_t = 0)$ for all n. This means (by Homework 4, Problem 1) that the chain is recurrent.

Thus, if $\sum_{i=1}^{\infty} \rho_i = \infty$ then the chain is recurrent. Suppose, on the hand, that this sum is finite. Referring to Equation 21.1 once more, set $u(n) = h(0) - h(n) = 1 - h(n)$. Since $h(n)$ is a probability, $u(n) \in [0,1]$. Of course $u(n) = 0$ for all n is a solution, as in the case that $\sum_{i=1}^{\infty} \rho_i = \infty$. But there are strictly positive solutions as well. Indeed, we can choose the value of $u(1)$ as we like, provided that $u(n) \in [0,1]$ for all n. Since $\rho_i > 0$, the sequence $u(1) \sum_{i=1}^{n-1} \rho_i$ is increasing, and so we need

$$u(1) \sum_{i=1}^{\infty} \rho_i = \lim_{n \to \infty} u(1) \sum_{i=1}^{n-1} \rho_i = \lim_{n \to \infty} u(n) \leq 1.$$

The minimal solution $h(n)$, meaning the maximal solution $u(n)$, is then achieved when $u(1) = \left(\sum_{i=1}^{\infty} \rho_i\right)^{-1}$. So we have

$$1 - h(n) = u(1) \sum_{i=1}^{n-1} \rho_i = \frac{\sum_{i=1}^{n-1} \rho_i}{\sum_{i=1}^{\infty} \rho_i}$$

meaning that

$$h(n) = 1 - \frac{\sum_{i=1}^{n-1} \rho_i}{\sum_{i=1}^{\infty} \rho_i} = \frac{\sum_{i=n}^{\infty} \rho_i}{\sum_{i=1}^{\infty} \rho_i}.$$

As $n \to \infty$, $h(n) \to 0$, which means that, for some n, there is a positive probability that the chain started at n never reaches 0. This means 0 is a transient state, and so the chain is transient.

Hence, recurrence is characterized by the summability of $\sum_{i=1}^{\infty} \rho_i$: recurrent if the sum is ∞, transient if it is $< \infty$.

Example 21.5 (M/M/1 queueing system)**.** Consider a birth and death chain with $\lambda_i = \lambda$ and $\mu_i = \mu$ both constant and non-zero. This is called an **M/M/1 queue**. It is a model for a system where jobs (or customers) arrive at Poissonian times (at rate λ), queue up, and are served in the order they arrived at rate μ. The process X_t is the number of jobs in the queue at time t. There are much more general kinds of queueing models; the M/M stands for Markov/Markov (meaning that the arrivals and service times are both Markovian), and the index 1 means that only 1 job is served at a time.

Following example 21.4, here we simply have $\rho_i = \frac{\mu^{i-1}}{\lambda^{i-1}} = (\mu/\lambda)^{i-1}$. Hence $\sum_{i=1}^{\infty} \rho_i = \sum_{j=0}^{\infty} (\mu/\lambda)^j$ is infinite if $\mu \geq \lambda$, and is finite and equal to $(1 - \mu/\lambda)^{-1}$ if $\mu < \lambda$. So X_t is recurrent if $\mu \geq \lambda$, and transient of $\mu < \lambda$. Indeed: if the service rate is at least the arrival rate, we expect the queue to keep emptying out over time; if the service rate is less than the arrival rate, then the line eventually grows without bound and never empties again (i.e. 0 is never revisited).

22. LECTURE 22: MAY 23

22.1. **Stationary Distribution.**

Definition 22.1. *Let $(X_t)_{t\geq 0}$ be a continuous time Markov chain with transition rates $q(i,j)$. A probability distribution π is called* **stationary** *or* **invariant** *if, for each state j,*

$$q(j)\pi(j) = \sum_i \pi(i)q(i,j).$$

In terms of the infinitesimal generator $\mathbf{A}$, this says $\pi\mathbf{A} = \mathbf{0}$.

Remark 22.2. One might think that the stationary distribution for X_t should be the same as the stationary distribution ϖ for its jump chain Y_n. However, the distribution ϖ satisfies $\varpi(j) = \sum_i \varpi(i)p(i,j) = \sum_i \varpi(i)\frac{q(i,j)}{q(i)}$. Therefore, if π is invariant for $(X_t)_{t\geq 0}$ then $\varpi(i) = q(i)\pi(i)$ is invariant for $(Y_n)_{n\geq 0}$. Note that, even if π is a probability distribtion $\sum_i \pi(i) = 1$, it may be that ϖ is not summable and so there is no invariant distribution for Y_n. The converse holds similarly: $(Y_n)_{n\geq 0}$ may have an invariant distribution while $(X_t)_{t\geq 0}$ fails to. This contributes to the fact that positive recurrence can differ between the process and its jump chain.

Proposition 22.3. *Let $(X_t)_{t\geq 0}$ be a continuous time non-explosive Markov chain, and suppose that π is a stationary distribution for $(X_t)_{t\geq 0}$. If $\mathbb{P}(X_0 = j) = \pi(j)$ for all states i, then $\mathbb{P}(X_t = j) = \pi(j)$ for all $t > 0$ and all states j.*

Remark 22.4. In the following proof, we will basically assume there are finitely-many states (this is required to interchange the derivatve and the sum). A different proof is needed to show the result holds true for infinitely-many states. Since Kolmogorov's equations actuallyonly determine the process up to the time of explosion, however, the fact that $\frac{d}{dt}\mathbb{P}(X_t = j) = 0$ only shows that $\mathbb{P}(X_t = j)$ is constant up to explosion time. If the process explodes, it can well start in and stay in the stationary distribution up to explosion time, and then be in a(n arbitrary) new distribution after explosion.

Proof. Fix a state j. Then

$$\frac{d}{dt}\mathbb{P}(X_t = j) = \frac{d}{dt}\sum_i \mathbb{P}(X_0 = i)\mathbb{P}(X_t = j|X_0 = i) = \frac{d}{dt}\sum_i \pi(i)p_t(i,j) = \sum_i \pi(i)\frac{d}{dt}p_t(i,j).$$

By Kolmogorov's backward equation,

$$\frac{d}{dt}p_t(i,j) = \sum_{k\neq i} q(i,k)p_t(k,j) - q(i)p_t(i,j)$$

and so

$$\sum_i \pi(i)\frac{d}{dt}p_t(i,j) = \sum_i \pi(i)\sum_{k\neq i} q(i,k)p_t(k,j) - \sum_i \pi(i)q(i)p_t(i,j).$$

In the first term, exchanging the order of the summation yields

$$\sum_k p_t(k,j)\sum_{i\neq k} \pi(i)q(i,k) = \sum_k p_t(k,j)\sum_i \pi(i)q(i,k) = \sum_k p_t(k,j)q(k)\pi(k)$$

where we have used the fact that $q(k,k) = 0$ and the definition of invariance of π. Hence we have

$$\frac{d}{dt}p_t(i,j) = \sum_k p_t(k,j)q(k)\pi(k) - \sum_i \pi(i)q(i)p_t(i,j) = 0.$$

It follows that $t \mapsto p_t(i,j)$ is constant, as claimed. □

Hence, it makes sense to call such a distribution stationary or invariant.

Example 22.5. Let $(X_t)_{t\geq 0}$ be an irreducible birth-and-death chain. (Irreducible here simply means that $\lambda_i > 0$ and $\mu_i > 0$ for all i.) Then $q(0) = \lambda_0$, $q(j) = \lambda_j + \mu_j$ for $j \geq 1$, and the equations for a stationary distribution are

$$\lambda_0\pi(0) = \mu_1\pi(1), \qquad (\lambda_j + \mu_j)\pi(j) = \mu_{j+1}\pi(j+1) + \lambda_{j-1}\pi(j-1), \quad j \geq 1.$$

The equations for $j \geq 1$ can be written in the form

$$\mu_{j+1}\pi(j+1) - \mu_j\pi(j) = \lambda_j\pi(j) - \lambda_{j-1}\pi(j-1).$$

Thus, for example, $\mu_2\pi(2) - \mu_1\pi(1) = \lambda_1\pi(1) - \lambda_0\pi(0)$; but $\lambda_0\pi(0) = \mu_1\pi(1)$ and so this simplifiles to $\mu_2\pi(2) = \lambda_1\pi(1)$. By induction, it follows that

$$\mu_{j+1}\pi(j+1) = \lambda_j\pi(j), \quad j \geq 0$$

and hence (since $\mu_j > 0$ for all j), again by induction, we have

$$\pi(j) = \frac{\lambda_{j-1}}{\mu_j}\pi(j-1) = \frac{\lambda_{j-1}\lambda_{j-2}}{\mu_j\mu_{j-1}}\pi(j-2) = \cdots = \frac{\lambda_0\cdots\lambda_{j-1}}{\mu_1\cdots\mu_j}\pi(0).$$

Define this new coefficient as θ_j; in terms of the previous notation ρ_j, we have

$$\theta_j = \frac{\lambda_0\cdots\lambda_{j-1}}{\mu_1\cdots\mu_j} = \frac{\lambda_0}{\mu_j}\frac{1}{\rho_j}.$$

By convention, we set $\theta_0 = 1$. Thus, normalizing $\pi(0)$ as we like, we have that $\pi(j) \sim \theta_j$ is an invariant measure; in order for it to be a stationary distribution, we need it to be summable. Hence, an invariant birth-and-death chain has a stationary distribution if and only if $\sum_{i=0}^{\infty}\theta_i < \infty$, in which case the stationary distribution is $\pi(j) = (\sum_{i=0}^{\infty}\theta_i)^{-1}\theta_j$.

Example 22.6. Consider the M/M/1 queue of Example 21.5. This is a birth-and-death chain with $\mu_i = \mu$ and $\lambda_i = \lambda$ for all i. Hence, we have

$$\theta_j = \frac{\lambda_0\cdots\lambda_{j-1}}{\mu_1\cdots\mu_j} = \frac{\lambda^j}{\mu^j} = \left(\frac{\lambda}{\mu}\right)^j.$$

So $\sum_{j=0}^{\infty}\theta_j$ is summable if and only if $\lambda < \mu$, in which case the sum is

$$\sum_{j=0}^{\infty}\theta_j = \sum_{j=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^j = \frac{1}{1-\lambda/\mu}.$$

Hence, the stationary distribution for this chain is

$$\pi(j) = \left(\sum_{i=0}^{\infty}\theta_i\right)^{-1}\theta_j = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^j$$

which is a (shifted) geometric distribution.

Example 22.7 (M/M/∞ queue)**.** Consider a queueing system with infinitely many servers. Jobs arrive according to a Poisson process with rate μ, and are served independently with independent service times at rate μ, where all jobs currently in the queue are served simultaneously. We model this as a birth-and-death-chain with $\lambda_j = \lambda$ for all $j \geq 0$, and $\mu_j = j\mu$ for all $j \geq 1$. Thus, we have

$$\theta_j = \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} = \frac{\lambda^j}{\mu \cdot 2\mu \cdots j\mu} = \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j$$

and so

$$\sum_{j=0}^{\infty} \theta_j = \sum_{j=0}^{\infty} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j = e^{\lambda/\mu}$$

which is finite for any $\lambda, \mu > 0$. Hence the stationary distribution always exists, and is given by

$$\pi(j) = \frac{e^{-\lambda/\mu}}{j!}\left(\frac{\lambda}{\mu}\right)^j.$$

That is, the stationary distribution is Poissonian with mean λ/μ. In particular, if the queue is at equilibrium, i.e. in the stationary distribution, then the mean number of customers in the queue is the mean of $\mathrm{Poisson}(\lambda/\mu)$, which is λ/μ.

22.2. **Convergence to the Stationary Distribution.** The exact analog of the basic convergence theory for discrete Markov chains (cf. Corollary 11.1, Theorem 11.3, and Theorem 12.1) holds for continuous time chains as well, with the proviso that the theory is simpler: there is no issue with periodicity. Indeed, for any continuous-time Markov chain, if $p_t(i,j) > 0$ for some time $t > 0$, then in fact $p_t(i,j) > 0$ for *all time* $t > 0$ (this is Problem 2 on Homework 6).

Theorem 22.8. *Let $(X_t)_{t\geq 0}$ be an irreducible, continuous-time Markov chain, with transition kernel $p_t(i,j) = \mathbb{P}_i(X_t = j)$ for states i, j, and transition rates $q(i,j) > 0$ and $q(i) = \sum_{j\neq i} q(i,j)$. The following are equivalent.*

(1) *All states are positive recurrent.*
(2) *Some state is positive recurrent.*
(3) *The chain is non-explosive, and there exists a stationary distribution π.*

Moreover, when these conditions hold, the stationary distribution is given by

$$\pi(j) = \frac{1}{\mathbb{E}_j[T_j]}$$

where T_j is the return time to state j. Finally, under these conditions, we have $p_t(i,j) \to \pi(j)$ as $t \to \infty$ for any states i, j.

Hence, Example 22.6 shows that the M/M/1 queue is, in fact, positive recurrent if and only if $\lambda < \mu$, and that it converges to equilibrium in this case. We already saw, in Example 21.5, that the M/M/1 queue is recurrent when $\lambda \leq \mu$; we conclude that in the critical case $\lambda = \mu$, the chain is *null recurrent*. As usual, this means that $p_t(i,j) \to 0$ for each fixed i, j in this case. Example 22.7 shows that the M/M/∞ queue is always positive recurrent, and always converges to its stationary (Poissonian) distribution.

The proofs of all parts of Theorem 22.8 are very similar to the proofs of Corollary 11.1 and Theorem 12.1, and so we will not touch them here. It is important to note that the non-explosion clause (3) in the theorem is absolutely necessary. For example, for convergence, we know that an exploding chain can start in the stationary distribution and hence stay in it until it explodes; afterward it can take any distribution at all, and so may not converge back to the stationary distribution.

Example 22.9. Consider a birth-and-death chain with $\lambda_j = \lambda \cdot 2^j$ and $\mu_j = \mu \cdot 2^j$ for some parameters λ, μ chosen so that $1 < \lambda/\mu < 2$. Then

$$\theta_j = \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} = \frac{\lambda \cdot 2\lambda \cdots 2^{j-1}\lambda}{2\mu \cdot 4\mu \cdots 2^j \mu} = \left(\frac{\lambda}{\mu}\right)^j \frac{1}{2^j}$$

and so $\sum_{j=0}^{\infty} \theta_j < \infty$, so an invariant distribution exists. But because the rates increase exponentially up the chain, explosion does occur, so the limiting behavior of the chain is not determined by rate paramters at all.

23. LECTURE 23: MAY 23, 2012

We are about to discuss a new kind of stochastic process, called **martingales**, that are not necessarily Markov processes (but are associated to them in important ways). We begin with a motivating example that we have already studied quite a bit from a different perspective.

Example 23.1. Let $(X_n)_{n\geq 0}$ be SRW on $\mathbb{Z}$. This is a Markov chain, of course, but it is even more "memoryless" than your average Markov chain. To see how, let's think about it in terms of a game: recall that X_n can be thought of as the sum of n independent Bernoulli trials. As such, we think of it as a kind of gambling game: you and your opponent each wager \$1 each toss; if heads comes up you win your dollar and your opponent's, while if tails you lose your dollar. Hence X_n represents your winnings after n tosses. In this context, it is natural ask about the *expectation* of the process: what is your expected winnings after n tosses? In this simple game, we have

$$\mathbb{E}[X_n] = \mathbb{E}[B_1 + \cdots + B_n] = \sum_{j=1}^{n} \mathbb{E}[B_j] = 0$$

where B_j are the independent Bernoulli trials that all have expectation 0. So the expectation is constant(ly equal to 0). This is inevitable for a "fair game" – if your expected winnings were ever positive, your opponent's would be negative, which means the game would be slanted in your favor.

Actually, there is a stronger "fairness" property that holds for this game. Suppose m tosses have already occurred, and your fortune has varied according to the trajectory $i_0, i_1, \ldots, i_m$ over that time. After some larger number of tosses n, what is your expected winnings? Since we already know $X_m = i_m$, the interesting quantity is $X_n - X_m$: your net winnings (or loss) since time m. So what we want to know is

$$\mathbb{E}[X_n - X_m | X_0 = i_0, \ldots, X_m = i_m].$$

Expressing this as a telescoping sum and using the linearity of expectation, we have

$$\mathbb{E}[X_n - X_m | X_0 = i_0, \ldots, X_m = i_m] = \sum_{k=m}^{n-1} \mathbb{E}[X_{k+1} - X_k | X_0 = i_0, \ldots, X_m = i_m].$$

Now further conditioning up to time k, we have $\mathbb{E}[X_{k+1} - X_k | X_0 = i_0, \ldots, X_m = i_m]$ is equal to

$$\sum_{i_{m+1}, \ldots, i_k} \mathbb{E}[X_{k+1} - X_k | X_0 = i_0, \ldots, X_m = i_m, X_{m+1} = i_{m+1}, \ldots, X_k = i_k] \cdot \mathbb{P}(X_{m+1} = i_{m+1}, \ldots, X_k = i_k | X_0 = i_0, \ldots, X_m = i_m).$$

The one-step increment conditional probabilities are easy to calculate. From the Markov property,

$$\begin{aligned} &\mathbb{E}[X_{k+1} - X_k | X_0 = i_0, \ldots, X_m = i_m, X_{m+1} = i_{m+1}, \ldots, X_k = i_k] \\ &= \mathbb{E}[X_{k+1} | X_0 = i_0, \ldots, X_m = i_m, X_{m+1} = i_{m+1}, \ldots, X_k = i_k] - i_k \\ &= \mathbb{E}[X_{k+1} | X_k = i_k] - i_k \\ &= \mathbb{E}[X_1 | X_0 = i_k] - i_k. \end{aligned}$$

Since $X_1 = X_0 + B_1$ where B_1 is a symmetric Bernoulli trial independent from X_0, and since $\mathbb{E}[B_1] = 0$, it follows that

$$\mathbb{E}[X_1|X_0 = i_k] = \mathbb{E}[X_0 + B_1|X_0 = i_k] = i_k + \mathbb{E}[B_1|X_0 = i_k] = i_k + \mathbb{E}[B_1] = i_k.$$

Hence all the terms in the above sum are 0, and we have the strong averaging property

$$\mathbb{E}[X_n - X_m|X_0 = i_0, \ldots, X_m = i_m] = 0 \tag{23.1}$$

for any initial trajectory $i_0, \ldots, i_m$ with $m \leq n$. In other words, no matter what has happened to the player's fortune so far, the expected net win or loss for any future time is always 0.

A process which satisfies Equation 23.1 is called a (discrete-time) **martingale**. So SRW is a martingale. It is also a Markov chain, and in fact we used the Markov property to deduce the martingale property – but not alone. Indeed, Markov chains need not take real values, so the expectation is not in general even defined. Even when it is, in general a Markov chain X_n need not have constant expectation – $\mathbb{E}[X_n]$ may well vary with n. For example, consider a non-symmetric random walk on $\mathbb{Z}$, with $p(i, i+1) = p$ and $p(i, i-1) = 1-p$ with $p \neq \frac{1}{2}$, cf. Example 4.5. This random walk can be realized as $X_n = B_1 + \cdots + B_n$ where B_j are i.i.d. biased Bernoulli random variables with $\mathbb{P}(B_j = 1) = p$ and $\mathbb{P}(B_j = -1) = 1 - p$. Then $\mathbb{E}[B_j] = 1 \cdot p + (-1) \cdot (1-p) = 2p - 1$, and so $\mathbb{E}[X_n] = (2p-1)n$ is only constant when $p = \frac{1}{2}$. Most Markov chains are not martingales. The converse is also true, but before we see examples, it will be useful to improve our understanding of conditional expectation.

23.1. **Conditional Expectation.** We are accustomed to thinking of conditional expectation as a number: if X is a random variable with (discrete) state space $S \subset \mathbb{R}$ and B is an event, then $\mathbb{E}[X|B] = \sum_{x \in S} x \cdot \mathbb{P}(X = x|B)$. The kinds of events we typically condition on are the outcomes of (other) random variables: for example $B = \{Y_1 = i_1, \ldots, Y_n = i_n\}$ for some random variables $Y_1, \ldots, Y_n$ and some states $i_1, \ldots, i_n$. As a matter of convenience, we can then group all of the conditional expectactions together into a **new random variable** which we call $\mathbb{E}[X|Y_1, \ldots, Y_n]$. It simply takes the value $\mathbb{E}[X|Y_1 = i_1, \ldots, Y_n = i_n]$ in the event $\{Y_1 = i_1, \ldots, Y_n = i_n\}$. That is:

$$\mathbb{E}[X|Y_1, \ldots, Y_n] = \sum_{i_1, \ldots, i_n} \mathbb{E}[X|Y_1 = i_1, \ldots, Y_n = i_n] \mathbb{1}_{\{Y_1 = i_1, \ldots, Y_n = i_n\}}.$$

This may look like simple book-keeping, but it brings new intuition about conditioning. We start with a random variable X. We are then given some *information*, in the form of some other random variables $Y_1, \ldots, Y_n$ that we may observe. Given these observations, what is our best guess for the value of X? In the (random event) that $Y_1 = i_1, \ldots, Y_n = i_n$, our best guess is the conditional expectation $\mathbb{E}[X|Y_1 = i_1, \ldots, Y_n = i_n]$. Hence, the new random variable $\mathbb{E}[X|Y_1, \ldots, Y_n]$ is our best (random) guess about the value of X given any observations of $Y_1, \ldots, Y_n$.

Example 23.2. Suppose that X is a function of $Y_1, \ldots, Y_n$, $X = F(Y_1, \ldots, Y_n)$ for some fixed (deterministic F). Then complete information about $Y_1, \ldots, Y_n$ yields complete information about X, so our "best guess" about X should be X itself. Indeed, in this case

we have

$$\begin{aligned}\mathbb{E}[X|Y_1,\dots,Y_n] &= \mathbb{E}[F(Y_1,\dots,Y_n)|Y_1,\dots,Y_n] \\ &= \sum_{i_1,\dots,i_n} \mathbb{E}[F(Y_1,\dots,Y_n)|Y_1=i_1,\dots,Y_n=i_n]\mathbb{1}_{\{Y_1=i_1,\dots,Y_n=i_n\}} \\ &= \sum_{i_1,\dots,i_n} F(i_1,\dots,i_n)\mathbb{1}_{\{Y_1=i_1,\dots,Y_n=i_n\}}.\end{aligned}$$

On the other hand, we also have

$$X = \sum_{i_1,\dots,i_n} X\mathbb{1}_{\{Y_1=i_1,\dots,Y_n=i_n\}}$$

and since $X = F(Y_1,\dots,Y_n)$, on the event $\{Y_1=i_1,\dots,Y_n=i_n\}$ $X = F(i_1,\dots,i_n)$. Thus

$$X = \sum_{i_1,\dots,i_n} X\mathbb{1}_{\{Y_1=i_1,\dots,Y_n=i_n\}} = \sum_{i_1,\dots,i_n} F(i_1,\dots,i_n)\mathbb{1}_{\{Y_1=i_1,\dots,Y_n=i_n\}} = \mathbb{E}[X|Y_1,\dots,Y_n].$$

When X is a function of $Y_1,\dots,Y_n$, we say X **is measurable with respect to** $Y_1,\dots,Y_n$. So Example 22.6 shows that, when X is measurable with respect to some random variables $Y_1,\dots,Y_n$, we have $\mathbb{E}[X|Y_1,\dots,Y_n] = X$ – our best guess for the value of X is X itself, since we have complete information.

Example 23.3. On the other hand, suppose that X and $\{Y_1,\dots,Y_n\}$ are independent. In this case, information about $Y_1,\dots,Y_n$ should be essentially useless in determining the value of X. What does this mean about $\mathbb{E}[X|Y_1,\dots,Y_n]$? For any states $x, i_1,\dots,i_n$ the events $\{X=k\}$ and $\{Y_1=i_1,\dots,Y_n=i_n\}$ are independent, so $\mathbb{P}(X=x|Y_1=i_1,\dots,Y_n=i_n) = \mathbb{P}(X=x)$, and therefore

$$\mathbb{E}[X|Y_1=i_1,\dots,Y_n=i_n] = \sum_x x\cdot\mathbb{P}(X=x|Y_1=i_1,\dots,Y_n=i_n) = \sum_x x\cdot\mathbb{P}(X=x) = \mathbb{E}[X].$$

Thus

$$\begin{aligned}\mathbb{E}[X|Y_1,\dots,Y_n] &= \sum_{i_1,\dots,i_n} \mathbb{E}[X|Y_1=i_1,\dots,Y_n=i_n]\mathbb{1}_{\{Y_1=i_1,\dots,Y_n=i_n\}} \\ &= \mathbb{E}[X]\sum_{i_1,\dots,i_n} 1_{\{Y_1=i_1,\dots,Y_n=i_n\}} = \mathbb{E}[X].\end{aligned}$$

Thus, when X is independent from $Y_1,\dots,Y_n$, the "best guess" $\mathbb{E}[X|Y_1,\dots,Y_n]$ is just the constant $\mathbb{E}[X]$. This is built into conditional expectation; the average value of X without conditioning is always known. If no further information is given, this constant is the conditional expectation as well.

Example 23.4. Consider the SRW X_n of Example 22.5 again. Using Equation 23.1, we have

$$\mathbb{E}[X_n - X_m|X_0,\dots,X_m] = \sum_{i_0,\dots,i_m} \mathbb{E}[X_n - X_m|X_0=i_0,\dots,X_m=i_m]\mathbb{1}_{\{X_0=i_0,\dots,X_m=i_m\}} = 0.$$

Now, X_m is certainly measurable with respect to $X_1,\dots,X_m$, and so we also have

$$\mathbb{E}[X_n - X_m|X_0,\dots,X_m] = \mathbb{E}[X_n|X_0,\dots,X_m] - \mathbb{E}[X_m|X_0,\dots,X_m] = \mathbb{E}[X_n|X_0,\dots,X_m] - X_m$$

by Example 22.6. Thus, for $n \geq m$, we have $\mathbb{E}[X_n|X_0, \ldots, X_m] = X_m$; the best guess about our future fortune at time n, given complete information up to time m, is our present fortune X_m. This is the essence of the "average fairness" property that defines a martingale.

Let's collect the properties of conditional expectation from the last examples (and a few more) in the following proposition.

Proposition 23.5. *Let X, X' be random variables, and $\mathbf{Y} = Y_1, \ldots, Y_n$ a collection of random variables. The following properties hold.*

(1) *For $a, b \in \mathbb{R}$, $\mathbb{E}[aX + bX'|\mathbf{Y}] = a\mathbb{E}[X|\mathbf{Y}] + b\mathbb{E}[X'|\mathbf{Y}]$.*
(2) *If X is $\mathbf{Y}$-measurable, then $\mathbb{E}[X|\mathbf{Y}] = X$.*
(3) *If X is independent of $\mathbf{Y}$, then $\mathbb{E}[X|\mathbf{Y}] = \mathbb{E}[X]$.*
(4) *(Tower Property) Let $\mathbf{Z}$ be another collection of random variables, and suppose $\mathbf{Y}$ are $\mathbf{Z}$-measurable, so $\mathbf{Y} = F(\mathbf{Z})$ for some function F. (A typical situation is when $\mathbf{Z} \supseteq \mathbf{Y}$). Then*
$$\mathbb{E}[\mathbb{E}[X|\mathbf{Z}]|\mathbf{Y}] = \mathbb{E}[X|\mathbf{Y}].$$
(5) *(Factoring) If Y is $\mathbf{Y}$-measurable then $\mathbb{E}[XY|\mathbf{Y}] = Y\mathbb{E}[X|\mathbf{Y}]$.*

Proof. Properties (2) and (3) were proved in Examples 23.2 and 23.3. Properties (1), (4), and (5) all follow from straightforward (but lengthy) calculations immediate from the definitions. □

The simplest case of the Tower Property (4) in Proposition 23.5 is worthy of its own statement.

Corollary 23.6. *If X and $\mathbf{Y} = Y_1, \ldots, Y_n$ are random variables, then $\mathbb{E}[\mathbb{E}[X|\mathbf{Y}]] = \mathbb{E}[X]$.*

Proof. The set $\varnothing$ of random variables is (vacuously) independent from all random variables, including X, and $\mathbf{Y} \supset \varnothing$ for any $\mathbf{Y}$. Thus, by the Tower Property (4) of Proposition 23.5, we have
$$\mathbb{E}[\mathbb{E}[X|\mathbf{Y}]|\varnothing] = \mathbb{E}[X|\varnothing] = \mathbb{E}[X]$$
where the second equality follows from (3) in Propostion 23.5. By the same argument, the random variable $\mathbb{E}[X|\mathbf{Y}]$ is independent from $\varnothing$, and so again by (3)
$$\mathbb{E}[\mathbb{E}[X|\mathbf{Y}]|\varnothing] = \mathbb{E}[\mathbb{E}[X|\mathbf{Y}]]$$
which proves the claim. □

24. LECTURE 24: MAY 23, 2012

24.1. **Martingales.** We have now seen several examples and equivalent formulations of the martingale concept. Let's give the full definition. To make all calculations valid, we need to ensure that expectations are finite.

Definition 24.1. *A (discrete-time)* **martingale** *is a stochastic process $(X_n)_{n\geq 0}$ which satisfies $\mathbb{E}[|X_n|] < \infty$ and*

$$\mathbb{E}[X_{n+1}|X_0, \ldots, X_n] = X_n$$

for all $n \geq 0$.

In Example 23.4, we showed that SRW is a martingale. (Indeed, since $|X_n| \leq n$, $\mathbb{E}[|X_n|] \leq n < \infty$ for all n.) In fact, we showed something apparently stronger: that $\mathbb{E}[X_n|X_0, \ldots, X_m] = X_m$ for all $m \leq n$. In fact, this nominally stronger property is equivalent to being a martingale.

Lemma 24.2. *If $(X_n)_{n\geq 0}$ is a martingale, then $\mathbb{E}[X_n|X_0, \ldots, X_m] = X_m$ for all $m \leq n$.*

Proof. The proof is by induction. The statement holds when $m = n$ since then X_n is $(X_1, \ldots, X_m)$-measurable, cf. Proposition 23.5(2). Suppose that we have shown that, for some $n \geq m$, $\mathbb{E}[X_n|X_0, \ldots, X_m] = X_m$. Then by the Tower Property of Proposition 23.5(4)

$$\mathbb{E}[X_{n+1}|X_0, \ldots, X_m] = \mathbb{E}[\mathbb{E}[X_{n+1}|X_0, \ldots, X_n]|X_0, \ldots, X_m].$$

We assumed that X_n is a martingale, meaning that the inside conditional expectation is $\mathbb{E}[X_{n+1}|X_0, \ldots, X_n] = X_n$. Hence, we have

$$\mathbb{E}[X_{n+1}|X_0, \ldots, X_m] = \mathbb{E}[X_n|X_0, \ldots, X_m] = X_m$$

where the second equality is the inductive hypothesis. □

Corollary 24.3. *Of $(X_n)_{n\geq 0}$ is a martingale, then it has constant expectation: $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ for all n.*

Proof. Here we simply use the Double Expectation property (Corollary 23.6). By Lemma 24.2 $X_0 = \mathbb{E}[X_n|X_0]$, and thus

$$\mathbb{E}[X_0] = \mathbb{E}[\mathbb{E}[X_n|X_0]] = \mathbb{E}[X_n].$$

□

Example 24.4 (Betting on Independent Coin Tosses)**.** We motivated the discussion of martingales with a simple betting strategy on coin tosses (i.e. SRW), where we wager \$1 on each toss regardless. In fact, it doesn't matter how we bet – as long as we have no future information, our winnings forms a martingale. To make this precise: let $X_1, X_2, \ldots$ be i.i.d. Bernoulli trials (so $\mathbb{P}(X_n = \pm 1) = \frac{1}{2}$, and thus $\mathbb{E}[X_n] = 0$). We start with some initial fortune W_0 (which is independent of the tosses X_n). Suppose we decide to bet an amount, B_n, on the nth coin toss. Thus, after the toss, we either increase our fortune by B_n if $X_n = 1$, or decrease it by B_n if $X_n = -1$ – that is, our winnings are

$$W_n = \sum_{i=1}^{n} X_i B_i.$$

The sequence $B_1, B_2, \ldots$ is our **betting strategy**, which we allow to be random. We insist that $\mathbb{E}[|B|_n] < \infty$ for each n – in a real betting game, we would be restricted to have

$B_n \leq \max\{W_n, 0\}$, and so long as we start with finite capital, by induction this means B_n is a bounded random variable so certainly has finite expectation. We also insist that the betting strategy at time n can only use information up to time n: in other words, we require B_n to be $(W_0, W_1, \ldots, W_{n-1})$-measurable.

Under these conditions, we have

$$\mathbb{E}[|W_n|] \leq \mathbb{E}\left[\sum_{i=1}^{n} |X_i B_i|\right] = \mathbb{E}\left[\sum_{i=1}^{n} |B_i|\right] = \sum_{i=1}^{n} \mathbb{E}[|B_i|] < \infty$$

for each n, as required. Now, the winnings at time $n+1$ are $W_{n+1} = W_n + X_{n+1}B_{n+1}$, and so

$$\mathbb{E}[W_{n+1}|W_0, \ldots, W_n] = \mathbb{E}[W_n|W_0, \ldots, W_n] + \mathbb{E}[X_{n+1}B_{n+1}|W_0, \ldots, W_n].$$

The first term is $\mathbb{E}[W_n|W_0, \ldots, W_n] = W_n$. For the second term: by assumption B_{n+1} is measurable with respect to $W_0, \ldots, W_n$, and so by the factoring property of Proposition 23.5(5), we have

$$\mathbb{E}[X_{n+1}B_{n+1}|W_0, \ldots, W_n] = B_{n+1}\mathbb{E}[X_{n+1}|W_0, \ldots, W_n].$$

Now, B_1 is W_0 measurable, and W_0 is independent of X_k for all k, so B_1 is as well. It follows that X_1B_1 is independent of X_{n+1}, and therefore so is $W_1 = W_0 + X_1B_1$. Then B_2 is (W_0, W_1)-measurable, and since both are independent of X_{n+1}, so is B_2. Thence so is X_2B_2, and so too $W_2 = W_1 + X_2B_2$. Continuing this way, we find that $W_0, \ldots, W_n$ are all independent of X_{n+1}. Hence, by Propositon 23.5(3),

$$\mathbb{E}[X_{n+1}|W_0, \ldots, W_n] = \mathbb{E}[X_{n+1}] = 0.$$

So, we have shown that $\mathbb{E}[W_{n+1}|W_0, \ldots, W_n] = W_n + 0 = W_n$, and so W_n is a martingale.

24.2. **Stopping Martingales.** Just as for a (discrete-time) Markov chain, or any stochastic process $(X_n)_{n\geq 0}$, a **stopping time** T is a random variable taking values in $\{0, 1, \ldots\} \cup \{\infty\}$, which has the property that the event $\{T \leq n\}$ depends only on $X_0, \ldots, X_n$. (To be precise: $\{T \leq n\}$ can be expressed in the form $\{(X_0, \ldots, X_n) \in U\}$ for some set of states U.)

In the context of Example 24.4, we would like to stop betting at some time. We could stop after a fixed number n of coin tosses. More likely, we would like to stop as soon as $W_n \geq F$ for some fixed fortune F. (Indeed, $T = \min\{n \geq 0 \colon W_n \geq F\}$ is a stopping time.) If we stop betting at this time, then no matter what happens to the coin, our fortune is equal to W_T from then on. In other words, we have replaced the original martingale W_n with the new process $W_{T\wedge n}$, where $T \wedge n = \min\{T, n\}$. In fact, this stopped martingale is still a martingale.

Proposition 24.5. *Let $(X_n)_{n\geq 0}$ be a martingale and let T be a stopping time for this martingale. Then $(X_{T\wedge n})_{n\geq 0}$ is a martingale.*

Proof. Let $Y_n = X_{T\wedge n}$. First note that $Y_n = M_{T\wedge n} \in \{X_0, \ldots, X_n\}$ for each n, so

$$|Y_n| \leq \max\{|X_0|, \ldots, |X_n|\} \leq |X_0| + \cdots + |X_n|,$$

hence $\mathbb{E}[|Y_n|] \leq \mathbb{E}[|X_0|] + \cdots + \mathbb{E}[|X_n|] < \infty$, as required. We now need to show that $\mathbb{E}[Y_{n+1}|Y_0, \ldots, Y_n] = Y_n$. We will first show that

$$\mathbb{E}[Y_{n+1}|X_0, \ldots, X_n] = Y_n. \tag{24.1}$$

To prove Equation 24.1, we can express $Y_{n+1} = X_{T\wedge(n+1)T}$ as follows. If $T \geq n+1$, then $T \wedge (n+1) = n+1$ and so $Y_{n+1} = X_{n+1}$ in this case. If, on the other hand, $T \leq n$, then $T \wedge (n+1) = T$ and $Y_{n+1} = T$. Thus

$$Y_{n+1} = X_T \mathbb{1}_{\{T\leq n\}} + X_{n+1}\mathbb{1}_{\{T>n\}}.$$

Since T is a stopping-time, $\{T \leq n\}$ only depends on $X_0, \ldots, X_n$; in other words $\mathbb{1}_{\{T\leq n\}}$ is $(X_0, \ldots, X_n)$-measurable. Of course, in the event $T \leq n$, X_T is $(X_0, \ldots, X_n)$-measurable as well, and so the product $X_T\mathbb{1}_{\{T\leq n\}}$ is $(X_0, \ldots, X_n)$-measurable. In addition, the function $\mathbb{1}_{\{T>n\}} = 1 - \mathbb{1}_{\{T\leq n\}}$ is $(X_0, \ldots, X_n)$-measurable. By Proposition 23.5, it follows that

$$\begin{aligned}\mathbb{E}[Y_{n+1}|X_0, \ldots, X_n] &= \mathbb{E}[X_T\mathbb{1}_{\{T\leq n\}}|X_0, \ldots, X_n] + \mathbb{E}[X_{n+1}\mathbb{1}_{\{T>n\}}|X_0, \ldots, X_n] \\ &= X_T\mathbb{1}_{\{T\leq n\}} + \mathbb{1}_{\{T>n\}}\mathbb{E}[X_{n+1}|X_0, \ldots, X_n].\end{aligned}$$

By assumption X_n is a martingale, and so we have

$$\mathbb{E}[Y_{n+1}|X_0, \ldots, X_n] = X_T\mathbb{1}_{\{T\leq n\}} + \mathbb{1}_{\{T>n\}}X_n.$$

Note that, in the case $T = n$, $X_T = X_n$, so we can rewrite this as

$$X_T\mathbb{1}_{\{T\leq n\}} + \mathbb{1}_{\{T>n\}}X_n = X_T\mathbb{1}_{\{T<n\}} + \mathbb{1}_{\{T\geq n\}}X_n = X_{T\wedge n} = Y_n.$$

This proves Equation 24.1.

Now, the collection $\mathbf{Y}_n = (Y_0, \ldots, Y_n)$ is $\mathbf{X}_n = (X_0, \ldots, X_n)$-measurable. Hence, we can use the Tower Property:

$$\mathbb{E}[Y_{n+1}|Y_0, \ldots, Y_n] = \mathbb{E}[Y_{n+1}|\mathbf{Y}_n] = \mathbb{E}[\mathbb{E}[Y_{n+1}|\mathbf{X}_n]|\mathbf{Y}_n].$$

Equation 24.1 says that $\mathbb{E}[Y_{n+1}|\mathbf{X}_n] = Y_n$, and so we have

$$\mathbb{E}[Y_{n+1}|Y_0, \ldots, Y_n] = \mathbb{E}[Y_n|\mathbf{Y}_n] = \mathbb{E}[Y_n|Y_0, \ldots, Y_n] = Y_n$$

concluding the proof.

□

As a corollary, we therefore have that $\mathbb{E}[X_{T\wedge n}] = \mathbb{E}[X_{T\wedge 0}] = \mathbb{E}[X_0]$ for all n. This is not *quite* the same as saying that, when we've stopped playing, our expected fortune is still unchanged. What we would like to see, for a truly "fair game", is that $\mathbb{E}[X_T] = \mathbb{E}[X_0]$. If this can be violated, then even though at each fixed time n we cannot expect any profit from any betting strategy, perhaps we can choose a clever enough random stopping time (based on events we've seen thus far in the game) to come out on top, on average. In fact, this *is* possible.

Example 24.6 (Martingale Betting Strategy)**.** Consider a betting strategy as in Example 24.4, where the strategy is fixed as $B_n = 2^n$ (double your bet each round). Example 24.4 shows that the winnings W_n form a martingale. Now, let $T = \min\{n\colon X_n = 1\}$, the first time heads comes up. If the initial capital is the fixed amount $W_0 = C$, then we have

$$\mathbb{E}[W_T] = \sum_{n=0}^{\infty}\mathbb{E}[W_T|T = n]\mathbb{P}(T = n).$$

The event $\{T = n\}$ is the event that $X_1 = \cdots = X_{n-1} = -1$ and $X_n = 1$, so it fixes a trajectory and thus $\mathbb{P}(T = n) = 2^{-n}$. So we have

$$\mathbb{E}[W_T] = \sum_{n=1} \mathbb{E}[W_T | T = n] \cdot 2^{-n} = \sum_{n=1}^{\infty} 2^{-n} \mathbb{E}[W_n | X_1 = \cdots = X_{n-1} = -1, X_n = 1].$$

By the betting strategy, conditioned on $X_1 = \cdots = X_{n-1} = -1$ and $X_n = 1$, we have

$$W_n = \sum_{i=1}^{n} X_i B_i = C - 2 - 4 - \cdots - 2^{n-1} + 2^{n+1} = C + 1$$

and so

$$\mathbb{E}[W_T] = \sum_{n=1}^{\infty} 2^{-n}(C + 1) = C + 1.$$

Hence, $\mathbb{E}[W_T] = C + 1 \neq C = W_0 = \mathbb{E}[W_0]$. (However, by Proposition 24.5, $\mathbb{E}[W_{T \wedge n}] = C$ for all n. The problem is that the stopping time can be arbitrarily large, so it is never true that $T \wedge n = T$ for any finite n.)

Remark 24.7. Example 24.6 is the reason for the terminology "martingale" for the processes we're studying (at least according to legend). This betting strategy was popular with 18th Century French gamblers. The term "martingale" comes from the shape of a particular horse harness; the reason it was used to name the betting strategy is lost to the mysts of time (though there are many interesting theories that can be found online).

This betting strategy appears to provide a sure-fire way to make money (which is what 18th Century gamblers thought). The problem is that the above model is not quite accurate; the real betting strategy is $B_n = 2^n \mathbb{1}_{\{W_{n-1} \geq 2^n\}}$, since we cannot bet money we do not have. The strategy therefore only really works if $C = \infty$, i.e. infinite capital. When $C < \infty$, the player will eventually go bust (by the law of large numbers).

Example 24.6 shows that "arbitrage" is possible: one can employ a betting strategy that is guaranteed to make money in a fair game (if one has infinite capital). But the winnings-martingale W_n wildly fluctuates in this case. In fact, if we don't allow this (or require that our stopping time must occur by a fixed time), then arbitrage is impossible.

Theorem 24.8 (Optional Sampling Theorem)**.** *Let $(X_n)_{n \geq 0}$ be a martingale, and let T be a finite stopping time. Suppose that either:*

(1) *T is bounded: there is some $N < \infty$ so that $\mathbb{P}(T \leq N) = 1$; or*
(2) *$(X_n)_{0 \leq n \leq T}$ is bounded: for some $B < \infty$, $\mathbb{P}(|X_n| \leq B \text{ for } n \leq T) = 1$.*

Then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

Proof. First, suppose (1) holds. By proposition 24.5, $X_{T \wedge n}$ is a martingale, and so $\mathbb{E}[X_{T \wedge n}] = \mathbb{E}[X_{T \wedge 0}] = \mathbb{E}[X_0]$ for all n. In particular, $\mathbb{E}[X_0] = \mathbb{E}[X_{T \wedge N}]$; but $T \wedge N = T$, so $\mathbb{E}[X_0] = \mathbb{E}[X_T]$ as claimed.

Now, suppose T is not necessarily bounded, but (2) holds. Then for any n, we can write

$$X_T = X_{T \wedge n} + (X_T - X_{T \wedge n}) = X_{T \wedge n} + (X_T - X_{T \wedge n}) \mathbb{1}_{\{T > n\}}.$$

(The second equality follows from the fact that, if $T \leq n$, then$X_T - X_{T\wedge n} = X_T - X_T = 0$.) So we have

$$\mathbb{E}[X_T] = \mathbb{E}[X_{T\wedge n}] + \mathbb{E}[(X_T - X_{T\wedge n})\mathbb{1}_{\{T>n\}}].$$

Now, $T \wedge n$ bounded by n, and it is easy to see that it is a stopping time; so by part (1), it follows that $\mathbb{E}[X_{T\wedge n}] = \mathbb{E}[X_0]$. For the second term, we have

$$|\mathbb{E}[(X_T - X_{T\wedge n})\mathbb{1}_{\{T>n\}}]| \leq \mathbb{E}[|X_T - X_{T\wedge n}|\mathbb{1}_{\{T>n\}}] \leq 2B\mathbb{E}[\mathbb{1}_{\{T>n\}}] = 2B\mathbb{P}(T > n).$$

Since $\mathbb{P}(T < \infty) = 1$, $\mathbb{P}(T > n) \to 0$ as $n \to \infty$. Hence, we have

$$|\mathbb{E}[X_T] - \mathbb{E}[X_0]| = |\mathbb{E}[X_T] - \mathbb{E}[X_{T\wedge n}]| \leq 2B\mathbb{P}(T > n) \to 0.$$

Since the left-hand-side does not depend on n, it follows that $\mathbb{E}[X_T] = \mathbb{E}[X_0]$, as required. □

25. LECTURE 25: MAY 30, 2012

25.1. **Applications of the Optional Sampling Theorem.**

Example 25.1. Let $(X_n)_{n\geq 0}$ be SRW on $\mathbb{Z}$, conditioned to start at $X_0 = j$ for some integer $j \in \{0, 1, \ldots, N\}$. This is a martingale (indeed it was our motivating example of a martingale). Let $\tau_k = \min\{n\colon X_n = k\}$ and set $T = \min\{\tau_0, \tau_N\}$. Using the first-step analysis iterative scheme from Section 2.4, we know how to calculate $h(j) = \mathbb{P}_j(\tau_0 < \tau_N)$. Martingale theory provides a different, much simplere method. Note that, by definition, $|X_n| \leq N$ for $n \leq T$, hence the optional sampling theorem applies to $(X_n)_{0\leq n\leq T}$. Therefore $\mathbb{E}[X_0] = \mathbb{E}[X_T]$. Since we conditioned $X_0 = j$, $\mathbb{E}[X_0] = j$. On the other end, the random variable X_T can take only two values: $X_T = 0$ or $X_T = N$. Thus

$$j = \mathbb{E}[X_0] = \mathbb{E}[X_T] = 0 \cdot \mathbb{P}(X_T = 0) + N \cdot \mathbb{P}(X_T = N).$$

It follows that $\mathbb{P}(X_T = N) = j/N$, and sp $\mathbb{P}(\tau_0 < \tau_N) = \mathbb{P}(X_T = 0) = 1 - j/N$.

Example 25.2. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common mean $\mathbb{E}[X_n] = \mu$, and let $S_n = X_1 + \cdots + X_n$. Define $M_n = S_n - n\mu$, and $M_0 = 0$. Then note that

$$\begin{aligned}
\mathbb{E}[M_{n+1}|M_0, M_1, \ldots, M_n] &= \mathbb{E}[S_{n+1} - (n+1)\mu|M_0, \ldots, M_n] \\
&= \mathbb{E}[S_n - n\mu + X_{n+1} - \mu|M_0, \ldots, M_n] \\
&= \mathbb{E}[M_n + X_{n+1} - \mu|M_0, \ldots, M_n] \\
&= M_n + \mathbb{E}[X_{n+1} - \mu|M_0, \ldots, M_n]
\end{aligned}$$

Since the X_n are independent, X_{n+1} is independent from S_k for all $k \leq n$, and hence also from M_k for $k \leq n$. Thus the last term is

$$\mathbb{E}[X_{n+1} - \mu|S_0, \ldots, S_n] = \mathbb{E}[X_{n+1} - \mu] = \mu - \mu = 0.$$

Also: $\mathbb{E}[M_1|M_0] = \mathbb{E}[M_1|0] = \mathbb{E}[M_1] = \mathbb{E}[X_1 - \mu] = 0 = M_0$. Thus $(M_n)_{n\geq 0}$ is a martingale.

Now, let T be a bounded stopping time for $(X_n)_{n\geq 0}$ (which is therefore also a bounded stopping time for $(M_n)_{n\geq 0}$. Then by the Optional Sampling Theorem,

$$0 = \mathbb{E}[M_0] = \mathbb{E}[M_T] = \mathbb{E}[S_T - T\mu] = \mathbb{E}[S_T] - \mu\mathbb{E}[T].$$

We have thus proved **Wald's equation**: if T is a bounded stopping time for an i.i.d. sequence with mean μ, then $\mathbb{E}[X_1 + X_2 + \cdots + X_T] = \mu\mathbb{E}[T]$. (In fact, this is true for any stopping time T with $\mathbb{E}[T] < \infty$, and it can be proved using the optional sampling theorem in this stronger form, but the proof is more involved.)

25.2. **Submartingales.**

A stochastic process $(X_n)_{n\geq 0}$ is called a **submartingale** (resp. **supermartingale**) if, for each n, $\mathbb{E}[X_{n+1}|X_0, \ldots, X_n] \geq X_n$. (resp. $\mathbb{E}[X_{n+1}|X_0, \ldots, X_n] \leq X_n$). For example, suppose we bet on coin-tosses with a betting strategy as in Example 24.4, but we additionally have a secret admirer who watches us bet. At each time n, the admirer adds some amount G_n to our winnings, where the amount given depends only on our betting and the coin tosses up to time $n-1$. That is: $W_{n+1} = W_n + B_{n+1}X_{n+1} + G_{n+1}$ where G_{n+1} is $(X_1, \ldots, X_n, B_1, \ldots, B_n)$-measurable. Then our total fortune W_n is a submartingale: $\mathbb{E}[W_{n+1}|W_n] = W_n + G_{n+1} \geq W_n$. (We might similarly consider a game where, at each time, we bet but may also buy some number of drinks, with the number depending only on our current and past fortune; this would give a supermartingale.)

Sub- and supermartingales are very useful technical tools in analyzing martingales and other stochastic processes. The main example we will consider comes from *maximal inequalities*. One of the most basic estimates useful in probability theory is *Markov's inequality*: $\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}$. There are more general versions (proved similarly) that give $\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[|X|^r]}{a^r}$ and $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[e^{bX}]}{e^{ba}}$ for any $b > 0$. If X_n is a martingale, stronger estimates are possible: the same bounds can be given not only for X_n, but for $\max\{X_0, \ldots, X_n\}$. The key is the following strengthening of the $r = 1$ Markov inequality for a positive submartingale.

Theorem 25.3 (Doob's Maximal Inequality)**.** *Let $(X_n)_{n\geq 0}$ be a non-negative submartingale. Then for any $a > 0$,*

$$\mathbb{P}(\max\{X_0, \ldots, X_n\} \geq a) \leq \frac{\mathbb{E}[X_n]}{a}.$$

Proof. Let $T = \min\{n \colon X_n \geq a\}$ (a stopping time). Consider the event

$$A_k = \{T = k\} = \{T \leq k\} \cap \{T \not\leq k-1\}.$$

Since T is a stopping time, $\{T \leq k\}$ is $(X_0, \ldots, X_k)$-measurable, and the event $\{T \leq k-1\}$ is $(X_0, \ldots, X_{k-1})$-measurable, so too is its complement, and therefore A_k is $(X_0, \ldots, X_k)$-measurable. Now, since $X_n \geq 0$, we have $X_n \geq X_n \mathbb{1}_{\{T \leq n\}}$, and hence

$$\mathbb{E}[X_n] \geq \mathbb{E}[X_n \mathbb{1}_{\{T\leq n\}}] = \mathbb{E}\left[X_n \sum_{k=0}^{n} \mathbb{1}_{A_k}\right] = \sum_{k=0}^{n} \mathbb{E}[X_n \mathbb{1}_{A_k}].$$

We now use Proposition 23.5 to simplify the summands:

$$\mathbb{E}[X_n \mathbb{1}_{A_k}] = \mathbb{E}[\mathbb{E}[X_n \mathbb{1}_{A_k} | X_0, \ldots, X_k]] = \mathbb{E}[\mathbb{1}_{A_k} \mathbb{E}[X_n | X_0, \ldots, X_k]].$$

Since X_n is a submartingale, it follows by induction that for $k \leq n$ $\mathbb{E}[X_n | X_0, \ldots, X_k] \geq X_k$. Thus, we have

$$\mathbb{E}[X_n \mathbb{1}_{A_k}] \geq \mathbb{E}[1_{A_k} X_k].$$

Now, A_k is the event that $\min\{n \colon X_n \geq a\} = k$; in particular, when A_k holds, $X_k \geq a$. Thus $\mathbb{1}_{A_k} X_k \geq a \mathbb{1}_{A_k}$, and so we have

$$\mathbb{E}[X_n] \geq \sum_{k=0}^{n} \mathbb{E}[1_{A_k} X_k] \geq \sum_{k=0}^{n} a\mathbb{E}[1_{A_k}] = a\mathbb{E}\left[\sum_{k=0}^{n} \mathbb{1}_{A_k}\right] = a\mathbb{P}(T \leq n).$$

Now, the event $\{T \leq n\}$ is the event that there exists a $k \leq n$ with $X_k \geq a$; i.e. $\{T \leq n\} = \{\max\{X_0, \ldots, X_n\} \geq a\}$. This proves the theorem. □

We will use Doob's maximal inequality directly for the special case of a non-negative martingale, in the next section. The more general result for *sub*martingales actually allows a much more powerful family of estimates for martingales. This is because of the following.

Lemma 25.4. *Let $(X_n)_{n\geq 0}$ be a martingale, and let $f \colon \mathbb{R} \to \mathbb{R}$ be a convex function such that $\mathbb{E}[|f(X_n)|] < \infty$ for all n. Then $Y_n = f(X_n)$ is a submartingale.*

The proof of Lemma 25.4 is an Exercise on Homework 7.

Corollary 25.5. *Let $(X_n)_{n\geq 0}$ be a martingale, and let $r \geq 1$ and $a, b > 0$. Then*

$$\mathbb{P}(\max\{X_0, \ldots, X_n\} \geq a) \leq \frac{\mathbb{E}[|X_n|^r]}{a^r} \tag{25.1}$$

$$\mathbb{P}(\max\{X_0, \ldots, X_n\} \geq a) \leq \frac{\mathbb{E}[e^{bX_n}]}{e^{ba}}. \tag{25.2}$$

Proof. The function $f(x) = |x|^r$ is convex for $r \geq 1$. Since $(X_n)_{n\geq 0}$ is a martingale, Lemma 25.4 shows that $(|X_n|^r)_{n\geq 0}$ is a (non-negative) submartingale. Now, for fixed a, if there exists a $k \in \{0, \ldots, n\}$ with $X_k \geq a$ then $|X_k|^r \geq a^r$. Thus $\{\max\{X_0, \ldots, X_n\} \geq a\} \subseteq \{\max\{|X_0|^r, \ldots, |X_n|^r\} \geq a^r\}$. Therefore, using Theorem 25.3 for $Y_n = |X_n|^r$, we have

$$\mathbb{P}(\max\{X_0, \ldots, X_n\} \geq a) \leq \mathbb{P}(\max\{Y_0, \ldots, Y_n\} \geq a^r) \leq \frac{\mathbb{E}[Y_n]}{a^r}$$

which proves Equation 25.1. The proof of Equation 25.2 is very similar. □

Example 25.6. Let $X_1, X_2, \ldots$ be i.i.d. symmetric Bernoulli random variables, so that $S_n = X_1 + \cdots + X_n$ (and $S_0 = 0$) forms a martingale. Taking $b = 1/\sqrt{n}$ in Equation 25.2, and setting $\alpha = a/\sqrt{n}$, we have the estimate

$$\mathbb{P}(\max\{S_1, \ldots, S_n\} \geq \alpha\sqrt{n}) \leq e^{-ba}\mathbb{E}[e^{bS_n}] = e^{-\alpha}\mathbb{E}[e^{S_n/\sqrt{n}}]. \tag{25.3}$$

Now, using independence and identical distribution of the X_k, we have

$$\mathbb{E}[e^{bS_n}] = \mathbb{E}[e^{b(X_1+\cdots+X_n)}] = \mathbb{E}[e^{bX_1}]^n.$$

Of coursr $\mathbb{E}[e^{bX_1}] = \frac{1}{2}e^b + \frac{1}{2}e^{-b}$, and so

$$\mathbb{E}[e^{S_n/\sqrt{n}}] = \left(\frac{e^{1/\sqrt{n}} + e^{-1/\sqrt{n}}}{2}\right)^n.$$

One can now employ elementary calculus (for example l'Hôpital's Rule) to show that

$$\lim_{n\to\infty} \mathbb{E}[e^{S_n/\sqrt{n}}] = \lim_{n\to\infty} \left(\frac{e^{1/\sqrt{n}} + e^{-1/\sqrt{n}}}{2}\right)^n = e^{1/2}.$$

Since this limit exists, we can conclude from the estimate of Equation 25.3 that there exists a constant $C > 0$ such that for $\alpha > 0$

$$\mathbb{P}(\max\{S_0, \ldots, S_n\} \geq \alpha\sqrt{n}) \leq Ce^{-\alpha}.$$

This is a very strong statement to the effect that, with high probability, S_n is not bigger than $\sqrt{n}$.

26. LECTURE 26: JUNE 1, 2012

26.1. **The Martingale Convergence Theorem.** Perhaps the central result about martingales is the following convergence theorem.

Theorem 26.1. *Let $(X_n)_{n\geq 0}$ be a martingale, and suppose there is a constant $C > 0$ so that $\mathbb{P}(X_n \geq -C) = 1$ for all n. Then there is a random variable X_∞ such that $\lim_{n\to\infty} X_n = X_\infty$ with probability 1.*

Remark 26.2. The condition of the theorem is often stated as requiring either $X_n \geq 0$ or $|X_n| \leq C$ (i.e. non-negative or bounded martingales). The standard proof does not require an upper-bound, only a lower-bound, so we state the theorem in this stronger form .

Proof. First, suppose we have proved the theorem in the special case $C = 0$: i.e. for $X_n \geq 0$. Then for any martingale $Y_n \geq -C$, set $X_n = Y_n + C$. It is easy to check that X_n is a martingale, and of course $X_n \geq 0$, so by the theorem $X_n \to X_\infty$ with probability 1. Hence, $Y_n \to X_\infty - C$ with probability 1, proving the theorem in general. Hence, we will assume that $X_n \geq 0$.

Since $X_n \geq 0$ is a (sub)martingale, Doob's Maximal Inequality shows that, for any fixed $N \in \mathbb{N}$,

$$\mathbb{P}\left(\max_{0\leq n\leq N} X_n \geq a\right) \leq \frac{\mathbb{E}[X_N]}{a} = \frac{\mathbb{E}[X_0]}{a}.$$

Taking $N \to \infty$, this shows that

$$\mathbb{P}\left(\max_{n\geq 0} X_n \geq a\right) \leq \frac{\mathbb{E}[X_0]}{a}.$$

It follows that $\mathbb{P}(\max_n X_n < \infty) = 1$, and so X_n is bounded with probability 1. In particular, this means that X_n has a convergent subsequence. This means there is a small interval $[a, b]$ that X_n keeps returning to infinitely often.

It is still possible for the sequence X_n to diverge, if X_n *oscillates*: i.e. if it also leaves the interval $[a, b]$ infinitely often. So we fix $0 \leq a < b$, and consider these potential oscillations. Let $S_0 = \min\{k\colon X_k \leq a\}$; for $n \geq 1$ let $T_n = \min\{k \geq S_{n-1}\colon X_k \geq b\}$, and $S_n = \min\{k \geq T_n\colon X_k \leq a\}$. So S_n are the *upcrossings* and T_n are the *downcrossings* through the interval $[a, b]$. The bounded sequence X_n fails to converge only if there is are some fixed $a < b$ such that there are infinitely many up-crossing and down-crossings through $[a, b]$. So define U_n to be the number of up-and-down crossings through $[a, b]$ by time n:

$$U_n = \max\{k\colon T_k \leq n\}.$$

Since U_n is non-decreasing, $U = \lim_{n\to\infty} U_n$ exists (though it may be ∞). Our goal is to show that $\mathbb{P}(U < \infty) = 1$.

We will prove this with a betting strategy akin to Example 24.4. Let $B_j = 1$ if $S_{k-1} < j \leq T_k$ for some k, and $B_j = 0$ if $T_k < j \leq S_{k+1}$ for some k. (So we bet 1 dollar every step during an up-crossing, and we don't bet during a down-crossing.) Then our winnings by time n are

$$W_n = \sum_{j=1}^{n} B_j(X_j - X_{j-1}).$$

In fact, $(W_n)_{n\geq 1}$ is a martingale; the proof is similar to the proof in Example 24.4. We have

$$\mathbb{E}[W_{n+1}|X_1,\dots,X_n] = \mathbb{E}[W_n + B_{n+1}(X_{n+1}-X_n)|X_1,\dots,X_n].$$

Now, the bet B_{n+1} is determined by $X_1,\dots,X_n$, and the winnings W_n are determined by $X_1,\dots,X_n$ and $B_1,\dots,B_n$ (hence just by $X_1,\dots,X_n$). Thus W_n and B_{n+1} are $(X_1,\dots,X_n)$-measurable, and so

$$\mathbb{E}[W_{n+1}|X_1,\dots,X_n] = W_n + B_{n+1}\mathbb{E}[X_{n+1}-X_n|X_1,\dots,x_n] = W_n$$

since X_n is a martingale. Now, since $W_1,\dots,W_n$ are $(X_1,\dots,X_n)$-measurable, we therefore have

$$W_n = \mathbb{E}[W_{n+1}|W_1,\dots,W_n] = \mathbb{E}[\mathbb{E}[W_{n+1}|X_1,\dots,X_n]|W_1,\dots,W_n] = \mathbb{E}[W_{n+1}|W_1,\dots,W_n]$$

and so $(W_n)_{n\geq 1}$ is a martingale. In particular $\mathbb{E}[W_n] = \mathbb{E}[W_1]$ for all n, and W_1 equals either $X_1 - X_0$ or 0, both if which have $\mathbb{E} = 0$. Thus $\mathbb{E}[W_n] = 0$ for all n.

Now, note that

$$\begin{aligned} W_n &= \sum_{k:\, T_k\leq n}\ \sum_{S_{k-1}<j\leq T_k} 1\cdot(X_j - X_{j-1}) \\ &= \sum_{k=0}^{U_n}\ \sum_{S_{k-1}<j\leq T_k}(X_j - X_{j-1}) + \sum_{j=S_{U_n}+1}^{n}(X_j - X_{j-1}). \end{aligned}$$

The first sum is over precisely the complete U_n up-crossings. In each up-crossing, the sum over j is a telescoping sum with value $X_{T_k} - X_{S_{k-1}}$; by definition $X_{T_k}\geq b$ and $X_{S_{k-1}}\leq a$. Hence, each of those terms is $\geq b-a$, and we have

$$W_n \geq \sum_{k=0}^{U_n}(b-a) + \sum_{j=S_{U_n}+1}^{n}(X_j - X_{j-1}).$$

The latter sum is also telescoping, with value $X_n - X_{S_{U_n}}$. By assumption $X_n\geq 0$, and $S_{U_n}\leq a$, so we conclude that

$$W_n \geq (b-a)U_n - a.$$

Hence $0 = \mathbb{E}[W_n]\geq (b-a)\mathbb{E}[U_n] - a$, and so we have $\mathbb{E}[U_n] l \frac{a}{b-a}$ for all n. It follows that the limit $U = \lim_{n\to\infty} U_n$ also has $\mathbb{E}[U]\leq \frac{a}{b-a} < \infty$. In particular, $\mathbb{P}(U<\infty) = 1$, which concludes the proof of the theorem.

□

Example 26.3. Let $(X_n)_{n\geq 0}$ be SRW on $\mathbb{Z}$, conditioned so $X_0 = 1$. Let $T = \min\{n\geq 0 : X_n = 0\}$. Then T is a stopping time, and so, by Proposition 24.5 $M_n = X_{T\wedge n}$ is a martingale. Before time T, $X_n\geq 1$, and so $M_n\geq 0$. Thus by the Martingale Convergence Theorem, there is a limit M_∞ so that $M_n\to M_\infty$ with probability 1.

What could this limit M_∞ possibly be? First, note that M_n is integer valued (since X_n is), so convergence means that M_n is eventually constant. Fix $k\in\mathbb{Z}$; then $\{M_\infty = k\} = \{M_n = k \text{ for all but finitely many } n\}$. If $k\neq 0$, then this is the same as the event $\{X_n = k \text{ for all but finitely many } n\}$; but this eventually constant trajectory has probability $\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdots = 0$. Thus, for $k\neq 0$, $\mathbb{P}(M_\infty = k) = 0$. It follows that $M_\infty = 0$ with probability 1.

So, start X_n as $X_0 = 0$ as usual. Its first step is to ± 1. In the case it is $+1$, the above analysis shows that $X_n = 0$ for some n; a similar analysis works for $X_1 = -1$. This gives an alternate proof the SRW is recurrent.

Note: in Example 26.3, conditioning $X_0 = 1$ means that $M_0 = X_{T\wedge 0} = X_0 = 1$, and so $\mathbb{E}[M_n] = \mathbb{E}[M_0] = 1$ for all n. But $M_\infty = 0$ so $\mathbb{E}[M_\infty] = 0$. We see that the constant expectation property of martingales does not generally extend to their limits. This is not hard to believe – it is always a tricky business exchanging a limit with a sum/integral/expectation.

Example 26.4 (Polya Urns)**.** An urn initially contains a red balls and b blue balls. At each time step, draw a ball at random from the urn, then return it along with another ball of the same color. Let X_n denote the number of red balls after n turns. Then $(X_n)_{n\geq 0}$ is a (time-inhomogeneous) Markov chain. Indeed, suppose at time n there are k red balls in the bin. The total number of balls is at time n is $n + a + b$, and so the probability that the ball we pick a red ball next is $\frac{k}{n+a+b}$. Then at time $n + 1$, we return this ball with another of the same color; if it was red, $X_{n+1} = k + 1$, and if it was blue, $X_{n+1} = k$ again. Hence, for $k \geq a$,

$$\mathbb{P}(X_{n+1} = k+1 | X_n = k) = \frac{k}{n+a+b} \quad \mathbb{P}(X_{n+1} = k | X_n = k) = 1 - \frac{k}{n+a+b}.$$

All our techniques for studying Markov chains required time homogeneity, so we are left in the lurch to analyze this process directly.

Instead, let $M_n = \frac{X_n}{n+a+b}$ be the *fraction* of red balls after n turns. Then $(M_n)_{n\geq 0}$ is a martingale: first, note that $0 \leq M_n \leq 1$, so $\mathbb{E}[|M_n|] \leq 1 < \infty$. Now, since X_n is a Markov chain, we know that for any states $i_0, i_1, \ldots, i_n, i_{n+1}$, $\mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n)$. It follows (from the summation definition of conditional expectation) that $\mathbb{E}[X_{n+1} | X_0, \ldots, X_n] = \mathbb{E}[X_{n+1} | X_n]$. This we can compute:

$$\begin{aligned}
\mathbb{E}[X_{n+1} | X_n = k] &= \sum_j j \mathbb{P}(X_{n+1} = j | X_n = k) \\
&= k\mathbb{P}(X_{n+1} = k | X_n = k) + (k+1)\mathbb{P}(X_{n+1} = k+1 | X_n = k) \\
&= k \cdot \left(1 - \frac{k}{n+a+b}\right) + (k+1) \cdot \frac{k}{n+a+b} = k + \frac{k}{n+a+b}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}[X_{n+1} | X_n] &= \sum_{k\geq a} \mathbb{E}[X_{n+1} = k | X_n = k] \mathbb{1}_{\{X_n = k\}} = \sum_{k \geq a} \left(k + \frac{k}{n+a+b}\right) \mathbb{1}_{\{X_n = k\}} \\
&= \sum_{k \geq a} \left(X_n + \frac{X_n}{n+a+b}\right) \mathbb{1}_{\{X_n = k\}} = \left(X_n + \frac{X_n}{n+a+b}\right) \sum_{k \geq a} \mathbb{1}_{\{X_n = k\}} \\
&= X_n + \frac{X_n}{n+a+b}.
\end{aligned}$$

(This makes good sense: our best guess for the number of red balls after the known outcome of the previous selection is the number of red balls previously plus the probability

of adding a red ball at the next step.) Since M_n is X_n-measurable, we therefore have

$$\mathbb{E}[M_{n+1}|M_n, \ldots, M_0] = \mathbb{E}[M_{n+1}|X_n] = \frac{1}{n+1+a+b}\mathbb{E}[X_{n+1}|X_n]$$

and this equals

$$\frac{1}{n+1+a+b}\left(X_n + \frac{X_n}{n+a+b}\right) = \frac{X_n}{n+a+b} = M_n$$

showing that $(M_n)_{n\geq 0}$ is a martingale.

This martingale is non-negative, hence by the Martingale Convergence Theorem, $M_n \to M_\infty$ with probability 1 for some random variable M_∞. It can be shown (though we are not equipped to show it) that M_∞ has a *beta distribution*: it is a continuous random variable on $[0, 1]$ with density

$$f_{M_\infty}(x) = \frac{(a+b-1)!}{(a-1)!(b-1)!}x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

There are good heuristics for why a beta should show up, so it is not too hard to justify that *if there is a limit* it ought to be beta. The martingale convergence theorem is the key to making the proof rigorous.

27. LECTURE 27: JUNE 4, 2012

Let's take a moment to recall the *Poisson process*: the simplest (continuous-time) birth-and-death process, with $\lambda_i = \lambda$ for all i and $\mu_i = 0$ for all i. As we showed in Lecture 20, there is an alternative description of this process: the Poisson process $(X_t)_{t\geq 0}$ with rate $\lambda > 0$ satisfies

(1) **Stationary Increments**: if $0 \leq s < t < \infty$, then $X_t - X_s$ has a Poisson distribution $\text{Poisson}(\lambda(t-s))$ that only depends on s and t through the difference $t - s$.
(2) **Independent Increments**: If $0 \leq t_0 < t_1 < \cdots < t_n < \infty$, then the increments of the process $X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent.
(3) **Right-continuous Trajectories**: For each $t \geq 0$, $\lim_{s \downarrow t} X_s = X_t$.

Properties (1) and (2) are essential for what we should call **random motion**. If we think of the trajectory $t \mapsto X_t$ as describing the motion of a particle, property (2) says that once the particle reaches position X_s, its change in position $X_t - X_s$ at a future time interval is independent of any previous motion. (This is a strong kind of Markov property.) Property (1) is a time-homogeneity condition: the distirbution of the random choice of "change in direction" also does not depend on your position or time. The Poisson process is not a good description of motion, however, since its paths are not continuous.

27.1. **Random Continuous Motion.** Call a stochastic process $(X_t)_{t\geq 0}$ a **random continuous motion** if it has:

(1) **Stationary Increments**: There is a one parameter family of probability distributions $\{\mu_t\}_{t\geq 0}$ so that, for $0 \leq s < t < \infty$, the distribution of $X_t - X_s$ is μ_{t-s}.
(2) **Independent Increments**: If $0 \leq t_0 < t_1 < \cdots < t_n < \infty$, then the increments of the process $X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent.
(3) **Continuous Trajectories**: The function $t \mapsto X_t$ is continuous with probability 1.

A Poisson process only misses the mark on item (3); but this is a serious change. Indeed, unless $(X_t)_{t\geq 0}$ is constant (the degenerate case), the condition that $t \mapsto X_t$ be continuous means that *the sample space is continuous* (i.e. it must contain an interval). Hence, a random continuous motion is a new kind of stochastic process beyond what we've studied: it is a **continuous-time, continuous-state** process.

One might expect this to open a can of worms on the kinds of processes that fit the bill; but the continuity of the paths (together with the other properties) actually pins down the measures: up to scale, there is only **one** random continuous motion! To see how the measure μ_t is determined, note that μ_t is the distribution of $X_t - X_0$. We can write this increment as a telescoping sum: let $t_j = \frac{j}{n}t$, so that $t_0 = t$ and $t_n = t$. Then

$$X_t - X_0 = [X_{t_n} - X_{t_{n-1}}] + [X_{t_{n-1}} - X_{t_{n-2}}] + \cdots + [X_{t_1} - X_{t_0}].$$

By conditions (1) and (2), this means $X_t - X_0$ is a sum of n independent random variables, each of which has distribution $\mu_{t_j - t_{j-1}} = \mu_{t/n}$. Probability distributions μ_t that can be written as the sum of n i.i.d. random variables for any n are called **infinitely-divisible**, and they form a small (but still infinite) family of probability distributions. Indeed, it we forget about the continuity, then there are many such processes, called **Lévy processes**, among which the Poisson process is an example. (As follows from Homework 5, Problem 3, a sum if independent Poisson random variables is Poisson, hence the Poisson distribution of any fixed rate is infinitely-divisible.)

Now we take the continuity condition (3) into account. Define

$$M_n = \max\{|X_{t_n} - X_{t_{n-1}}|, |X_{t_{n-1}} - X_{t_{n-2}}|, \ldots, |X_{t_1} - X_{t_0}|\}.$$

Since $t \mapsto X_t$ is continuous on the compact interval $[0, t]$, is is uniformly continuous there. So for any $\epsilon > 0$, there is $\delta > 0$ such that $|X_s - X_{s'}| < \epsilon$ whenever $|s - s'| < \delta$. For large enough n, all the partitions $|t_j - t_{j-1}|| < \delta$, and so for all large n $M_n < \epsilon$. It follows that $\lim_{n\to\infty} M_n = 0$. It turns out that this condition – that the maximum value of the increments tends to 0 as the partition-width tends to 0 – pins down μ_t: **the distribution must be a centered Gaussian**.

Definition 27.1. *A stochastic process $(B_t)_{t\geq 0}$ with state space $\mathbb{R}$ is called a* **Brownian Motion** *or* **Wiener process** *if:*

(1) *For $0 \leq s < t < \infty$, $B_t - B_s$ has a normal distribution $N(0, \sigma^2(t-s))$ for some $\sigma > 0$.*
(2) *The increments are independent.*
(3) *The trajectories are continuous with probability 1.*

If we additionally normalize $\sigma^2 = 1$ and condition $B_0 = 0$, then $(B_t)_{t\geq 0}$ is called a **standard Brownian motion**.

To be clear: condition (1) says that the increment $B_t - B_s$ is a continuous random variable with density

$$f_{B_t - B_s}(x) = \frac{1}{\sqrt{2\pi(t-s)}} e^{-x^2/2(t-s)}.$$

Our preceding discussion shows that **all random continuous motions are Brownian motions**. The process is named in honorof the early 19th Century botanist Robert Brown who noticed that pollen particles on the surface of apparently still water appeared, under a microscope, to execute "random continuous motion" – they jiggled around seemingly at random. The explanation for this motion (which is the result of zillions of collisions with water molecules all moving essentially randomly) in terms of the above definition of "random continuous motion" was provided by Einstein (and is one of his key papers that led to his Nobel Prize). However, an actually mathematically rigorous proof of the existence of this motion was first given by Norbert Wiener in the 1920s; for this reason, the process is often called the **Wiener process** and often denoted W_t rather than B_t.

Note that $\mathbb{E}[B_t] = 0$ since the normal distribution is centered. It is also interesting to consider the covariances. Let's take $s \leq t$; then

$$\begin{aligned}\mathrm{Cov}(B_s, B_t) &= \mathbb{E}[(B_s - \mathbb{E}[B_s])(B_t - \mathbb{E}[B_t])] \\ &= \mathbb{E}[B_s B_t] = \mathbb{E}[B_s(B_s + B_t - B_s)] = \mathbb{E}[B_s^2] + \mathbb{E}[B_s(B_t - B_s)].\end{aligned}$$

It is customary to take B_0 to be a constant; in general, we can allow it to be any random variable, as long as we assume B_0 is independent of all increments. Then we have the last term is

$$\begin{aligned}\mathbb{E}[(B_s - B_0 + B_0)(B_t - B_s)] &= \mathbb{E}[(B_s - B_0)(B_t - B_s)] + \mathbb{E}[B_0(B_t - B_s)] \\ &= \mathbb{E}[B_s - B_0]\mathbb{E}[B_t - B_s] + \mathbb{E}[B_0]\mathbb{E}[B_t - B_s] = 0.\end{aligned}$$

Hence

$$\mathrm{Cov}(B_s, B_t) = \mathbb{E}[B_s^2] = \mathrm{Var}(B_s) = \sigma^2 s, \quad s \leq t$$

since $\mathbb{E}[B_s]^2 = 0$, and the variance of $N(0, \sigma^2 s)$ is (by definition) $\sigma^2 s$. We can restate all of this as

$$\mathrm{Cov}(B_s, B_t) = \sigma^2 \min\{s, t\}. \tag{27.1}$$

On the subject of variances, note that we have

$$\mathrm{Var}(B_{t+h} - B_t) = \sigma^2 h.$$

In other words, the standard deviation of the increment $B_{t+h} - B_t$ is $\sigma\sqrt{h}$. For small h, this is $\gg h$. So, on average, we cannot expect $\frac{1}{h}[B_{t+h} - B_t]$ to have a limit as $h \downarrow 0$. Indeed, Brownian trajectories are **nowhere differentiable**.

Here is a very useful property of Brownian motion we will exploit in investigating its properties.

Lemma 27.2 (Scaling Property)**.** *Let $(B_t)_{t\geq 0}$ be a standard Brownian motion, and let $a > 0$. For $t \geq 0$, define $W_t = a^{-1/2} B_{at}$. Then $(W_t)_{t\geq 0}$ is a standard Brownian motion.*

Proof. Since $B_0 = 0$, $W_0 = 0$. Scaling does not affect continuity. So we need only check that the increments are independent and normally distributed. For fixed times $0 \leq t_0 < t_1 < \cdots < t_n$, we also have $0 \leq at_0 < at_1 < \cdots < at_n$, and so by assumption the increments $B_{at_1} - B_{at_0}, \ldots, B_{at_n} - B_{at_{n-1}}$ are independent. Scaling them by $a^{-1/2}$ preserves independence, and so the increments of W_t are independent. Finally, $W_t - W_s = a^{-1/2}(B_{at} - B_{as})$. By assumption the distirbution of $B_{at} - B_{as}$ is $N(0, a(t-s))$, and so scaling by $a^{-1/2}$ scales the variance by $(a^{-1/2})^2 = a^{-1}$, showing that $W_t - W_s$ has distribution $N(0, t-s)$. Thus $(W_t)_{t\geq 0}$ is a standard Brownian motion. □

27.2. **Gaussian Processes.**

Definition 27.3. *Let T denote a set of times (so $T \subseteq \mathbb{R}$, could be discrete or continuous). A stochastic process $(X_t)_{t\in T}$ is called a* **Gaussian process** *if its state space is $\mathbb{R}$, and if for all $t_1, \ldots, t_n \in T$ and all $a_1, \ldots, a_n \in \mathbb{R}$, $a_1 X_1 + \cdots + a_n X_n$ is a Gaussian random variable.*

For any stochastic process $(X_t)_{t\geq 0}$, its **marginal distributions** are the joint distributions of all random vectors $(X_{t_1}, \ldots, X_{t_n})$ for all choices of n and times $t_1, \ldots, t_n$. Definition 27.3 is a statement about the marginal distributions of the process. Properties (1) and (2) in the definition of Brownian motion are also statements exclusively about the marginal distributions of the process.

Example 27.4. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion. Fix $0 \leq t_1 < \cdots < t_n$, and set $G_j = B_{t_j} - B_{t_{j-1}}$. Then for any $a_1, \ldots, a_n \in \mathbb{R}$,

$$a_1 B_{t_1} + a_2 B_{t_2} + \cdots + a_n B_{t_n} = a_1 G_1 + a_2(G_1 + G_2) + \cdots + a_n(G_1 + \cdots + G_n).$$

We can rewrite this as

$$(a_1 + \cdots + a_n)G_1 + (a_2 + \cdots + a_n)G_2 + \cdots + (a_{n-1} + a_n)G_{n-1} + a_n G_n.$$

The increments G_j are independent Gaussians, and therefore this linear combination of them is also Gaussian. This shows that $(B_t)_{t\geq 0}$ is a Gaussian process.

It is very important to note that the marginal distributions do not encode fine properties of the trajectories, like continuity.

Example 27.5. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion, and let U be a uniform random variable (on $[0,1]$) independent from $(B_t)_{t\geq 0}$. Define

$$X_t = \begin{cases} B_t, & t \neq U \\ 0, & t = U \end{cases}.$$

In fact, X_t is a Gaussian process with the *same* marginal distributions as B_t. To see this, it suffices to show that for any times $t_1, \ldots, t_n$ and any coefficients $a_1, \ldots, a_n$, the two random variables $a_1 X_{t_1} + \cdots + a_n X_{t_n}$ and $a_1 B_{t_1} + \cdots + a_n B_{t_n}$ have the same distribution. This can be tested by looking at their *characteristic functions* $\chi_X(\xi) = \mathbb{E}[e^{i\xi X}]$. We have, for any $\xi \in \mathbb{R}$,

$$\mathbb{E}[e^{i\xi(a_1 X_{t_1} + \cdots + a_n X_{t_n})}] = \mathbb{E}[e^{i\xi(a_1 X_{t_1} + \cdots + a_n X_{t_n})} \mathbb{1}_{\{U \neq t_1, \ldots, t_n\}}]$$

because the event $\{U \neq t_1, \ldots, t_n\}$ has probability 1. But on this event $X_t = B_t$ for each t, so

$$\begin{aligned} \mathbb{E}[e^{i\xi(a_1 X_{t_1} + \cdots + a_n X_{t_n})}] &= \mathbb{E}[e^{i\xi(a_1 B_{t_1} + \cdots + a_n B_{t_n})} \mathbb{1}_{\{U \neq t_1, \ldots, t_n\}}] \\ &= \mathbb{E}[e^{i\xi(a_1 B_{t_1} + \cdots + a_n B_{t_n})}] \mathbb{E}[\mathbb{1}_{\{U \neq t_1, \ldots, t_n\}}] \\ &= \mathbb{E}[e^{i\xi(a_1 B_{t_1} + \cdots + a_n B_{t_n})}], \end{aligned}$$

where the second equality follows from the independence of U from $B_{t_1}, \ldots, B_{t_n}$, and the third is just the statement that $\mathbb{P}(U \neq t_1, \ldots, t_n) = 1$. So $a_1 X_{t_1} + \cdots + a_n X_{t_n}$ has the same characteristic function as $a_1 B_{t_1} + \cdots + a_n B_{t_n}$ for all times $t_1, \ldots, t_n$ and all coefficients $a_1, \ldots, a_n$. It follows that X_t and B_t have the same marginal distributions, and that X_t is a Gaussian process. But the function $t \mapsto X_t$ disagrees with $t \mapsto B_t$ at $t = U$ (unless $B_U = 0$, which is an event of probability 0), hence with probability 1 the trajectory $t \mapsto B_t$ has a discontinuity in $[0,1]$.

We therefore have to be careful to verify continuity properties separarely from marginal distribution properties. In particular, Brownian motion is a **continuous Gaussian process**; the "continuous" has to be added to the description, as it does not follow from a description of the marginal distributions.

28. LECTURE 28: JUNE 6, 2012

One of the reasons that the Gaussian distribution is so nice is that is is completely determined by its mean and variance. This fact implies that Gaussian processes are determined (in the sense of marginal distributions) by their mean and covariances.

Lemma 28.1. *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be Gaussian processes. Suppose that $\mathbb{E}[X_t] = \mathbb{E}[Y_t]$ and $\mathrm{Cov}(X_s, X_t) = \mathrm{Cov}(Y_s, Y_t)$ for all $s, t \in T$. Then $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ have the same marginal distributions.*

(The proof is somewhat involved, but not difficult.) We therefore use some notation: for a Gaussain process X_t, set

$$\mu(t) = \mathbb{E}[X_t], \qquad \Gamma(s,t) = \mathrm{Cov}(X_s, X_t).$$

Example 28.2. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion. Then B_t is a Gaussian process with $\mu(t) = 0$ and $\Gamma(s,t) = \min\{s,t\}$. It is a continuous process (meaning its trajectories are continuous with probability 1). The process described in Example 27.5 is also a Gaussian process with the same mean and covariance functions; but it is not a continuous process.

Example 28.3. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion, and let $X_t = B_t - tB_1$ for $0 \leq t \leq 1$. Then for any coefficients $a_1, \ldots, a_n$ and times $0 \leq t_1 < \cdots < t_n \leq 1$,

$$a_1 X_{t_1} + \cdots + a_n X_{t_n} = (a_1 B_{t_1} + \cdots + a_n B_{t_n}) - (a_1 t_1 + \cdots + a_n t_n) B_1.$$

If $t_n = 1$ we just combinate the last terms. In either case, this reduces to a linear combination of instances of B_t, and from the previous example we saw this yields a Gaussian. Hence X_t is a Guassian process. Note that it also has continuous trajectories. X_t is called a **Brownian bridge**.

Example 28.4. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion, and let $Z_t = e^{-t} B_{e^{2t}}$ for $t \geq 0$. Then a linear combination of Z_{t_j} is a linear combination of B_{s_j} for $s_j = e^{2t} t_j$, and hence is Gaussian; so Z_t is a Gaussian process. Again, it is clear that is has continuous trajectories. Z_t is called an **Ornstein-Uhlenbeck process**.

The Gaussian processes point-of-view gives us another useful invariance property of Brownian motion.

Lemma 28.5 (Time Inversion)**.** *Let $(B_t)_{t\geq 0}$ be a standard Brownian motion. Let $W_0 = 0$ and for $t > 0$ set $W_t = tB_{1/t}$. Then $(W_t)_{t\geq 0}$ is a standard Brownian motion.*

Proof. Let $t_1, \ldots, t_n \geq 0$. If any $= 0$, then since $W_t = 0$ we may ignore these terms in a sum, so wlog $t_j > 0$. Set $s_j = 1/t_j$. Then for any coefficients $a_1, \ldots, a_n$,

$$a_1 W_1 + \cdots + a_n W_n = a_1 t_1 B_{s_1} + \cdots + a_n b_n B_{s_n}$$

is a linear combination of Brownian points, and as shown in Example 27.4, it follows that $a_1 W_1 + \cdots + a_n W_n$ is Gaussian. Hence $(W_t)_{t\geq 0}$ is a Gaussian process. We can also calculate $\mu(t) = \mathbb{E}[W_t] = t\mathbb{E}[B_{1/t}] = 0$ for $t > 0$ and $\mu(0) = \mathbb{E}[W_0] = \mathbb{E}[0] = 0$. For covariances, if either $s = 0$ or $t = 0$ then $\mathrm{Cov}(W_s, W_t) = 0$, while if $0 < s \leq t$

$$\mathrm{Cov}(W_s, W_t) = \mathbb{E}[W_s W_t] = st\mathbb{E}[B_{1/s} B_{1/t}] = st \min\{1/s, 1/t\} = st \cdot 1/t = s.$$

So $\Gamma(s,t) = \mathrm{Cov}(W_s, W_t) = \min\{s,t\}$ as well, and we have shown that $(W_t)_{t\geq 0}$ has the same marginal distributions as $(B_t)_{t\geq 0}$.

It now suffices to show that $t \mapsto W_t$ is continuous with probability 1. For $t > 0$ this is clear, since $t \mapsto tB_{1/t}$ is continuous wherever $1/t$ is continuous. At $t = 0$, first we note that the above treatment shows that B_t and X_t have the same marginal distributions. In particular, their distributions agree on the dense set of rationals $\mathbb{Q} \cap (0, \infty)$. Since B_t has continuous paths and X_t is continuous on $[0, \infty)$, it follows (from the uniqueness of Brownian motion) that W_t is continuous at 0 as well. Hence $(W_t)_{t \geq 0}$ is a standard Brownian motion. □

Corollary 28.6. *If $(B_t)_{t \geq 0}$ is a standard Brownian motion, then $\lim_{t \to \infty} \frac{B_t}{t} = 0$.*

Remark 28.7. For integer $t = n$, we can write $B_n = B_n - B_0 = \sum_{k=1}^{n} [B_k - B_{k-1}]$ The increments $B_k - B_{k-1}$ are i.i.d. with mean 0, and hence by the strong law of large numbers $\frac{B_n}{n} \to 0$ with probability 1. Extending this argument to $t \notin \mathbb{N}$ directly takes a lot of work (owing to the very rough nature of Brownian motion: just because it is continuous doesn't mean there's a very nice way to bound B_t by B_n and B_{n+1} for $n \leq t < n + 1$). But time-inversion gives an immediate easy proof.

Proof. Let $W_t = tB_{1/t}$; then $W_{1/t} = \frac{B_t}{t}$. In Lemma 28.5, we saw that W_t is a standard Brownian motion; in particular this means $W_0 = 0$ and $t \mapsto W_t$ is continuous. It follows that $\lim_{t \to \infty} \frac{B_t}{t} = \lim_{t \to \infty} W_{1/t} = \lim_{s \downarrow 0} W_s = 0$. □

28.1. **Donsker's Scaling Limit.** Brownian motion is the pre-eminent Gaussian process. The Gaussian distribution arises as the limit distribution in the central limit theorem; similarly, Brownian motion arises as the "limit process" in a collection of limit theorems. The gist is that if one rescales a discrete-time random walk with space scaled as the square-root of time, then in the scaling limit the result is Brownian motion. Here is a precise statement.

Theorem 28.8 (Donsker's Invariance Principle)**.** *Let $(X_n)_{n \geq 0}$ be SRW on $\mathbb{Z}$. Let $c > 0$, and define a continuous-time, continuous process $X_t^{(c)}$ as follows: for $t \geq 0$ such that $ct \in \mathbb{N}$, set $X_t^{(c)} = c^{-1/2} X_{ct}$; for other times t, define $X_t^{(c)}$ by linear interpolation. Then for any times $0 \leq t_1 < \cdots < t_n$,*

$$(X_{t_1}^{(c)}, \ldots, X_{t_n}^{(c)}) \xrightarrow{\mathscr{D}} (B_{t_1}, \ldots, B_{t_n})$$

as $c \to \infty$, where $(B_t)_{t \geq 0}$ is a standard Brownian motion, and $\xrightarrow{\mathscr{D}}$ means convergence in distribution.

The proof, boiled down to its basic idea, is just repeated application of the central limit theorem. Indeed, one does not need to use the SRW; instead, one could take $X_n = Y_1 + \cdots + Y_n$ where $(Y_n)_{n \geq 0}$ is a sequence of i.i.d. random variables each with $\mathbb{E}[Y_j] = 0$ and $\mathrm{Var}[Y_j] = 1$. (If one takes $\mathrm{Var}[Y_j] = \sigma^2$ then the theorem holds with a Brownian motion of variance σ^2 in the limit.) Note also that the linear-interpolation definition of the scaled process $X_t^{(c)}$ is included only to make these processes continuous. One could just as well use (and indeed the proof uses)

$$\widetilde{X}_t^{(c)} = c^{-1/2} X_{\lfloor ct \rfloor}.$$

The sense of convergence stated in Theorem 28.8 is very weak; much stronger forms of convergence hold. For example, the random continuous path $(t \mapsto X_t^{(c)})_{0 \leq t \leq R}$ converge

to the Brownian path $(t \mapsto B_t)_{0 \le t \le R}$ uniformly, with probability 1, for any $R > 0$. (This is one way to prove the existence of Brownian motion – if one can show this uniform limit exists, then the limit is a uniform limit of continuous paths and therefore has continuous paths; that is satisfies the marginal distribution properties of Brownian motion is then verified via central limit theorem considerations as above.)

28.2. **Brownian Motion as a Markov Process.** It is difficult to formulate the Markov property in exact accordinace with the definition of continuous time Markov chains (with countable state space) in Lecture 18, without the benefit of more sophisticated technology (i.e. measure theory). But because of the indpendent increments, it is possible to state the Markov property in the following form.

Proposition 28.9 (Markov Property for Brownian Motion)**.** *Let $(B_t)_{t \ge 0}$ be a standard Brownian motion. Fix as time $s \ge 0$, and let $X_t = B_{t+s} - B_s$. Then $(X_t)_{t \ge 0}$ is a standard Brownian motion, independent from $(B_t)_{0 \le t \le s}$.*

Proof. Of course $X_0 = B_s - B_s = 0$, and the trajectories of X_t are continuous. For given $0 \le t_1 < \cdots < t_n < \infty$, the increments of X_t are $X_{t_j} - X_{t_{j-1}} = (B_{t_j+s} - B_s) - (B_{t_{j-1}+s} - B_s) = B_{t_j+s} - B_{t_{j-1}+s} = B_{s_j} - B_{s_{j-1}}$, where $s_j = t_j + s$ so $0 \le s_1 < \cdots < s_n < \infty$. Hence, they are shifted increments of the Brownian motion, and hence are independent. Finally, if $t_1 < t_2$ then $X_{t_2} - X_{t_1} = B_{t_2+s} - B_{t_1+s}$ which is a Brownian increment of length $(t_2 + s) - (t_1 + s) = t_2 - t_1$, and therefore has a $N(0, t_2 - t_1)$ distribution, as required. Thus, $(X_t)_{t \ge 0}$ is a standard Brownian motion.

Moreover, suppose $s > 0$ and let $0 \le t_1 < \cdots < t_n < s$. Now for $t \ge 0$, since $X_t = B_{t+s} - B_s$, we know that the increments $B_{t+s} - B_s, B_s - B_{t_n}, B_{t_n} - B_{t_{n-1}}, \ldots, B_{t_2} - B_{t_1}, B_{t_1} - 0$ ar independent. So, in particular, X_t is independent from $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}$. It is therefore also independent from the successive sums of these, namely $B_{t_2} = (B_{t_2} - B_{t_1}) + B_{t_1}$, and $B_{t_3} = (B_{t_3} - B_{t_2}) + B_{t_2}$, and so forth.

□

29. LECTURE 29: JUNE 9, 2012

Just as in the discrete state-space case, this implies a *Strong Markov property*. As usual, a **stopping time** (in this case for Brownian motion) is a random variable T taking values in $[0, \infty]$ with the property that, for each $t \geq 0$, the event $\{T \leq t\}$ depends only on $\{B_s\}_{0 \leq s \leq t}$.

Example 29.1. Fix $x \in \mathbb{R}$, and let $\tau_x = \min\{t \geq 0 \colon B_t = x\}$, the **first passage time** to x. Then

$$\{\tau_x \leq t\} = \{\exists s \leq t \colon B_s = x\}$$

which is an event determined by the random variables $(B_s)_{0 \leq s \leq t}$. Hence τ_x is a stopping time for $(B_t)_{t \geq 0}$.

Theorem 29.2 (Strong Markov property)**.** *Let $(B_t)_{t \geq 0}$ be a standard Brownian motion, and let T be a stopping time. For $t \geq 0$, set $X_t = B_{T+t} - B_T$. Then $(X_t)_{t \geq 0}$ is a standard Brownian motion, independent from $(B_t)_{0 \leq t \leq T}$.*

Example 29.3. Fix $x > 0$; we will determine the distribution of the first passage time τ_x. This can be done with the following trick. If we had a discrete random variable Y, then we could compute using the Law of Total Probability that, for any event A,

$$\mathbb{P}(A) = \sum_y \mathbb{P}(A, Y = y) = \sum_y \mathbb{P}(A > x | Y = y) \mathbb{P}(Y = y).$$

This formula does not make sense of Y is a continuous random variable with density f_Y, since the sum is over a continuous set, and since $\mathbb{P}(Y = y) = 0$. The analog of this formula which does hold is

$$\mathbb{P}(A) = \int \mathbb{P}(A | Y = y) f_Y(y) \, dy. \tag{29.1}$$

The range of the integral is all possible states that Y can take, in the event A. We apply this in the setting $A = \{B_t > x\}$ and $Y = \tau_x$. Note: when $B_t > x$, since $B_0 = 0$ and the Brownian path is continuous, by the intermediate value theorem there is a time $s \in (0, t)$ where $B_s = x$; hence $\tau_x \in (0, t)$. Thus

$$\mathbb{P}(B_t > x) = \int_0^t \mathbb{P}(B_t > x | \tau_x = s) f_{\tau_x}(s) \, ds. \tag{29.2}$$

We can deal with the integrand as follows. First, by definition $B_{\tau_x} = x$. Also, as we integrate over $0 < s < t$, we may write $t = s + (t - s)$. Hence

$$\mathbb{P}(B_t > x | \tau_x = s) = \mathbb{P}(B_{s+(t-s)} - B_{\tau_x} > 0 | \tau_x = s) = \mathbb{P}(B_{\tau_x + (t-s)} - B_{\tau_x} > 0 | \tau_x = s).$$

By the strong Markov property, $B_{\tau_x + t} - B_{\tau_x}$ is a standard Brownian motion, independent from $(B_s)_{0 \leq s \leq \tau_x}$. In particular, it is independent from τ_x, which is measurable with respect to $(B_s)_{0 \leq s \leq \tau_x}$, and so

$$\mathbb{P}(B_{\tau_x + (t-s)} - B_{\tau_x} > 0 | \tau_x = s) = \mathbb{P}(B_{\tau_x + (t-s)} - B_{\tau_x} > 0) = \mathbb{P}(B_{t-s} > 0) = \frac{1}{2}$$

where the last equality follows from the fact that B_{t-s} is a $N(0, t - s)$ random variable (hence symmetric). Plugging this into Equation 29.2 yields

$$\mathbb{P}(B_t > x) = \int_0^t \frac{1}{2} f_{\tau_x}(s) \, ds = \frac{1}{2} \mathbb{P}(\tau_x \in (0, t)) = \frac{1}{2} \mathbb{P}(\tau_x < t).$$

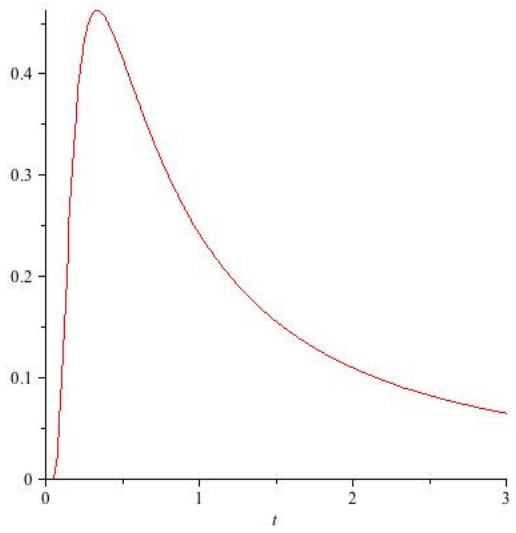

FIGURE 1. The density f_{τ_1} of the passage time to 1 for a standard Brownian motion.

Thus $\mathbb{P}(\tau_x < t) = 2\mathbb{P}(B_t > x)$, which is a probability we can calculate explicitly (in terms of the Gaussian kernel).

$$\mathbb{P}(\tau_x < t) = 2\mathbb{P}(B_t > x) = 2\int_x^\infty \frac{1}{\sqrt{2\pi t}} e^{-y^2/2t}\, dy.$$

To calculate the density f_{τ_x}, it is convenient to make the change of variables $y/\sqrt{t} = u$, so the integral becomes

$$\mathbb{P}(\tau_x < t) = 2\int_{x/\sqrt{t}}^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du.$$

This is convenient because the t only appears in the limits of integration. So we can differentiate easily:

$$f_{\tau_x}(t) = \frac{d}{dt}\mathbb{P}(\tau_x < t) = -\frac{2}{\sqrt{2\pi}} e^{-(x/\sqrt{t})^2/2} \frac{d}{dt}\frac{x}{\sqrt{t}} = \frac{x}{\sqrt{2\pi t^3}} e^{-x^2/2t}.$$

We assumed that $x > 0$ in all of the above. Note, however, that $(-B_t)_{t\geq 0}$ is also a standard Brownian motion (as can be very easily verified from the definition), and so by symmetry we will then have the more general formula

$$f_{\tau_x}(t) = \frac{|x|}{\sqrt{2\pi t^3}} e^{-x^2/2t}, \quad x \in \mathbb{R}.$$

Example 29.4. Consider the **Maximum process** for Brownian motion: $M_t = \max_{0\leq s\leq t} B_s$. First note that $M_t \geq 0$ for all t, since M_t is increasing and $M_0 = B_0 = 0$. We can calculate the distribution of the random variable M_t using the passage time. Notice that the event $\{M_t \geq x\}$ is equal to the event that there is some time $s \leq t$ for which $B_s \geq x$. Since Brownian paths are continuous, this is equal to the event that there is some (possibly different) time $s' \leq t$ where $B_{s'} = x$, and so we have $\{M_t \geq x\} = \{\tau_x \leq t\}$. Thus, by

Example 29.3, for $x \geq 0$ we have

$$\mathbb{P}(M_t \geq x) = \mathbb{P}(\tau_x \leq t) = 2\int_x^\infty \frac{1}{\sqrt{2\pi t}} e^{-y^2/2t}\, dy.$$

Thus, M_t has a density:

$$f_{M_t}(x) = \frac{d}{dx} 2\int_x^\infty \frac{1}{\sqrt{2\pi t}} e^{-y^2/2t}\, dy = \sqrt{\frac{2}{\pi t}} e^{-x^2/2t} \mathbb{1}_{x\geq 0}.$$

So the density of the maximum process is exactly 2 times the Gaussian density on the positive real axis. I.e. it is exactly twice as likely for M_t to exceed x as it is for B_t to exceed x. (This is a symmetry you may have expected, but it is not obvious.)

29.1. **The Zero Set of Brownian Motion.** Let $(B_t)_{t\geq 0}$ be a standard Brownian motion, and denote by $\mathscr{Z}$ the (random) set of zeroes of B_t:

$$\mathscr{Z} = \{t \geq 0 \colon B_t = 0\}.$$

Using techniques like the ones above, we can prove the following exact formula which is great asset in understand the zeroes, and general behavior, of Brownian motion.

Proposition 29.5 (Arcsine Law for Zeroes of BM)**.** *Let $0 < s < t$. Define*

$$\Theta(s,t) = \mathbb{P}(\mathscr{Z} \cap [s,t] \neq \varnothing).$$

Then

$$\Theta(s,t) = 1 - \frac{2}{\pi} \arcsin\sqrt{\frac{s}{t}}.$$

Proof. To begin, we will condition on the value of Brownian motion at time s. Using Equation 29.1,

$$\Theta(s,t) = \mathbb{P}(\mathscr{Z} \cap [s,t] \neq \varnothing) = \int_{-\infty}^\infty \mathbb{P}(\mathscr{Z} \cap [s,t] \neq \varnothing | B_s = x) f_{B_s}(x)\, dx.$$

The density f_{B_s} is a centered Gaussian of variance s. As for the conditional probability, we can use the Markov property: given $B_s = x$, we restart the process at time s: take $X_r = B_{s+r} - B_s = B_{s+r} - x$. Then $(X_r)_{r\geq 0}$ is a standard Brownian motion. Translating, the event that $B_\cdot$ has a zero in $[s,t]$ is the same as the event that $X_\cdot$ reaches height $-x$ in the interval $[0, t-s]$. Since $X_\cdot$ is a standard Brownian motion, this is the event that $\tau_{-x} \leq t-s$. Hence:

$$\Theta(s,t) = \int_{-\infty}^\infty \mathbb{P}(\tau_{-x} \leq t-s) \frac{1}{\sqrt{2\pi s}} e^{-x^2/2s}\, dx.$$

Referring to Example 29.3, the random variable τ_{-x} has density $f_{\tau_{-x}}(t) = \frac{|x|}{\sqrt{2\pi t^3}} e^{-x^2/2t}$. This is symmetric in x, as is the Gaussian density, so we have

$$\begin{aligned}\Theta(s,t) &= 2\int_0^\infty \int_0^{t-s} f_{\tau_{-x}}(r)\, dr\, \frac{1}{\sqrt{2\pi s}} e^{-x^2/2s}\, dx \\ &= \sqrt{\frac{2}{\pi s}} \int_0^\infty e^{-x^2/2s}\, dx \int_0^{t-s} \frac{x}{\sqrt{2\pi r^3}} e^{-x^2/2r}\, dr.\end{aligned}$$

We must now evaluate this double integral. To start, we exchange the order of integration (since everything in sight is positive, this follows from Fubini's theorem).

$$\Theta(s,t) = \frac{1}{\pi\sqrt{s}} \int_0^{t-s} r^{-3/2}\, dr \int_0^\infty x e^{-x^2/2r - x^2/2s}\, dx.$$

The inside integral can be explicitly evaluated, since the derivative $\frac{d}{dx} e^{-x^2/2r - x^2/2s}$ is equal to $-(1/r + 1/s) x e^{-x^2/2r - x^2/2s}$. Thus

$$\begin{aligned}\Theta(s,t) &= \frac{1}{\pi\sqrt{s}} \int_0^{t-s} r^{-3/2} \cdot \frac{-1}{1/r + 1/s} \left(e^{-x^2/2r - x^2/2s} \Big|_{x=0}^{x\to\infty} \right) dr \\ &= \frac{1}{\pi\sqrt{s}} \int_0^{t-s} r^{-3/2} \frac{rs}{r+s}\, dr = \frac{1}{\pi} \int_0^{t-s} \sqrt{\frac{s}{r}} \frac{dr}{r+s}.\end{aligned}$$

Now, we substitute $u = \sqrt{r/s}$. Then $r = su^2$, so $dr = 2su\, du$. The limits of integration become 0 up to $\sqrt{(t-s)/s}$, and we have

$$\Theta(s,t) = \frac{1}{\pi} \int_0^{\sqrt{(t-s)/s}} \frac{1}{u} \frac{2su\, du}{su^2 + s} = \frac{2}{\pi} \int_0^{\sqrt{(t-s)/s}} \frac{du}{1+u^2}.$$

The antiderivative of $\frac{1}{1+u^2}$ is $\arctan u$, so we can write this as

$$\Theta(s,t) = \frac{2}{\pi} \left[\arctan \sqrt{\frac{t-s}{s}} - \arctan 0 \right] = \frac{2}{\pi} \arctan \sqrt{\frac{t-s}{s}}.$$

This is a perfectly good final answer, but it is more common to write this as an arcsine. If we let $\theta = \arctan \sqrt{\frac{t-s}{s}}$, then θ is an angle in a right triangle whose opposite side has length $\sqrt{t-s}$ and adjacent side has length $\sqrt{s}$. The hypoteneuse therefore has length $\sqrt{s + (t-s)} = \sqrt{t}$. Hence $\cos\theta = \frac{\sqrt{s}}{\sqrt{t}}$. Using the relation $\sin(\frac{\pi}{2} - \theta) = \cos\theta$, we therefore have $\frac{\pi}{2} - \theta = \arcsin \frac{\sqrt{s}}{\sqrt{t}}$, and so

$$\Theta(s,t) = \frac{2}{\pi} \left[\frac{\pi}{2} - \arcsin \sqrt{\frac{s}{t}} \right] = 1 - \frac{2}{\pi} \arcsin \sqrt{\frac{s}{t}}.$$

□

Example 29.6. Let $L = \max\{t \le 1 \colon B_t = 0\}$ be the largest 0 of Brownian motion ≤ 1. Then for $t < 1$ the event $\{L \ge t\}$ is the event that $\mathscr{Z} \cap [t,1] \ne \varnothing$, and so

$$\mathbb{P}(L \ge t) = \mathbb{P}(\mathscr{Z} \cap [t,1] \ne \varnothing) = \Theta(t,1) = 1 - \frac{2}{\pi} \arcsin \sqrt{t}.$$

So we can calculate the density of L:

$$f_L(t) = \frac{d}{dt} \mathbb{P}(L < t) = \frac{d}{dt} \frac{2}{\pi} \arcsin \sqrt{t} = \frac{1}{\pi\sqrt{t(1-t)}}, \quad 0 \le t \le 1.$$

This density is symmetric about $\frac{1}{2}$, with a *minimum* there. It blows up at 0 and at 1; so the maximal 0 is most likely to be either very close to 0 or very close to 1.

We can use the arcsine law for the zeroes to show, for example, that Brownian motion (in one dimension) is recurrent.

Theorem 29.7 (Recurrence of Brownian Motion on $\mathbb{R}$)**.** *Let $(B_t)_{t\geq 0}$ be a standard Brownian motion on $\mathbb{R}$. Then the zero set $\mathscr{Z}$ is unbounded with probability 1. (I.e. the process returns to 0 at arbitrarily large times.)*

Proof. If the process is not recurrent, there is some time s after which there are no zeroes (with positive probability):

$$\mathbb{P}(\forall t \geq s\ B_t \neq 0) = \mathbb{P}(\mathscr{Z} \cap [s, \infty) = \varnothing).$$

The event $\{\mathscr{Z} \cap [t_0, \infty) = \varnothing\}$ is contained in the event $\{\mathscr{Z} \cap [s, t] = \varnothing\}$ for any $t > s$. Hence

$$\mathbb{P}(\mathscr{Z} \cap [s, \infty) = \varnothing) \leq \mathbb{P}(\{\mathscr{Z} \cap [s, t] = \varnothing) = 1 - \Theta(s, t) = \frac{2}{\pi} \arcsin \sqrt{\frac{s}{t}}.$$

As $t \to \infty$, $\arcsin \sqrt{\frac{s}{t}} \to 0$ for any fixed s, and hence $\mathbb{P}(\mathscr{Z} \cap [s, \infty) = \varnothing) = 0$ for any $s > 0$. Thus, there are arbitrarily large elements in $\mathscr{Z}$, with probability 1. □

Remark 29.8. Since we proved, early on, that SRW on $\mathbb{Z}$ is recurrent, it is natural that its scaling limit Brownian motion is recurrent as well – but this is highly non-obvious. Indeed, given space is scaled along with time, the kind of recurrence that is more natural to expect for Brownian motion is this: *given any $\epsilon > 0$, the set of $t \geq 0$ for which $|B_t| \leq \epsilon$ is unbounded with probability 1.* The fact that the Brownian motion actually returns to 0 infinitely often is a very fine property. Now, one can consider Brownian motion on $\mathbb{R}^d$: simply take $B_t^{(d)} = (B_t^1, B_t^2, \ldots, B_t^d)$ where the B_t^j are independent Brownian motions. Donsker's invariance theorem extends to this case, and $B_t^{(d)}$ is the scalaing limit of the SRW on $\mathbb{Z}^d$. Recall that SRW on $\mathbb{Z}^2$ is recurrent, while SRW on $\mathbb{Z}^d$ is transient for $d \geq 3$. If we interpret recurrence and transience in the weaker sense of return to a *neighborhood* of 0, then these translate to Brownian motion as well. But it is *not true* that $B_t^{(2)}$ is recurrent in the strong sense: the event that $\{B_t^{(2)} \neq (0, 0)$ for all large $t\}$ has positive probability. For $d \geq 3$, the results match up again: $\{|B_t^{(d)}| \to \infty\}$ has probability 1 for $d \geq 3$.

The recurrence of Brownian motion on $\mathbb{R}$, together with its time inversion property, gives one of the most graphic demonstration of just how rough Brownian paths are.

Corollary 29.9. *For any $\epsilon > 0$, the $(0, \epsilon) \cap \mathscr{Z} \neq \varnothing$ with probability 1; hence it contains infinitely many zeroes of $(B_t)_{t\geq 0}$.*

Proof. Consider $W_t = tB_{1/t}$. We have shown that $(W_t)_{t\geq 0}$ is a standard Brownian motion, and hence with probability 1 $W_t = 0$ for arbitrarily large times t. But $B_t = tW_{1/t}$, and therefore B_t has zeroes for t arbitrarily close to 0. □

Remark 29.10. By the strong Markov property, if T is a stopping time at which $B_T = 0$, then the same argument shows that $\mathscr{Z} \cap [T, T + \epsilon)$ contains infinitely-many zeroes of the Brownian motion. It is important to note that this only holds true for *stopping times T*, not just any random time that the Brownian motion has value 0. So, for example, if we fix a time t_0 and define $T_{t_0} = \min\{t \geq t_0 \colon B_t = 0\}$, then T_{t_0} is a stopping time, and indeed zeroes of the Brownian motion will accumulate to the right of T_{t_0}. On the other hand, there is a largest zero L smaller than 1 and, by Example 29.6, it has a positive probability of being less than 0.9. Indeed, from the arcine law for $\mathscr{Z}$, given any interval $[s, t]$ with $s < t$, the probability that there are *no zeroes* in $[s, t]$ is $\frac{2}{\pi} \arcsin \sqrt{\frac{s}{t}}$ which is strictly positive.

What can be proved, using the stopping times T_{t_0} above, is that $\mathscr{Z}$ is a *perfect set*: it is closed (since $t \mapsto B_t$ is continuous) and contains no isolated points. So, considering the largest zero L less than 1 again, there must be a sequence of zeroes *less than* L convering to L from below. In general, the zero set has the topological structure of a *Cantor set*.