Lecture 2: Markov Chains (I)

Readings Strongly recommended:

- Grimmett and Stirzaker (2001) 6.1, 6.4-6.6

Optional:

- Hayes (2013) for a lively history and gentle introduction to Markov chains.
- Koralov and Sinai (2010) 5.1-5.5, pp.67-78 (more mathematical)

A canonical reference on Markov chains is Norris (1997).

We will begin by discussing *Markov chains*. In Lectures 2 & 3 we will discuss *discrete-time* Markov chains, and Lecture 4 will cover *continuous-time* Markov chains.

2.1 Setup and definitions

We consider a discrete-time, discrete space stochastic process which we write as $X(t) = X_t$, for $t = 0, 1, \ldots$. The state space S is discrete, i.e. finite or countable, so we can let it be a set of integers, as in $S = \{1, 2, \ldots, N\}$ or $S = \{1, 2, \ldots\}$.

Definition. The process $X(t) = X_0, X_1, X_2, \ldots$ is a *discrete-time Markov chain* if it satisfies the *Markov property*:

$$P(X_{n+1} = s | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = s | X_n = x_n). \tag{1}$$

The quantities $P(X_{n+1} = j | X_n = i)$ are called the *transition probabilities*. In general the transition probabilities are functions of i, j, n. It is convenient to write them as

$$p_{ij}(n) = P(X_{n+1} = j | X_n = i)\,. \tag{2}$$

Definition. The *transition matrix* at time n is the matrix $P(n) = (p_{ij}(n))$, i.e. the (i, j)th element of $P(n)$ is $p_{ij}(n)$.[1] The transition matrix satisfies:

(i) $p_{ij}(n) \geq 0 \quad \forall i, j$ (the entries are non-negative)

(ii) $\sum_j p_{ij}(n) = 1 \quad \forall i$ (the rows sum to 1)

Any matrix that satisfies (i), (ii) above is called a *stochastic matrix*. Hence, the transition matrix is a stochastic matrix.

***Exercise* 2.1.** Show that the transition probabilities satisfy (i), (ii) above.

***Exercise* 2.2.** Show that if $X(t)$ is a discrete-time Markov chain, then

$$P(X_n = s | X_0 = x_0, X_1 = x_1, \ldots, X_m = x_m) = P(X_n = s | X_m = x_m)\,,$$

for any $0 \leq m < n$. That is, the probabilities at the current time, depend only on the most recent known state in the past, even if it's not exactly one step before.

[1] We call it a matrix even if $|S| = \infty$.

Remark. Note that a "stochastic matrix" is *not* the same thing as a "random matrix"! Usually "random" can be substituted for "stochastic" but not here. A random matrix is a matrix whose entries are random. A stochastic matrix has completely deterministic entries. It probably gets its name because it is used to describe a stochastic phenomenon, but this is an unfortunate accident of history.

Definition. The Markov chain $X(t)$ is *time-homogeneous* if $P(X_{n+1} = j|X_n = i) = P(X_1 = j|X_0 = i)$, i.e. the transition probabilities do not depend on time n. If this is the case, we write $p_{ij} = P(X_1 = j|X_0 = i)$ for the probability to go from i to j in one step, and $P = (p_{ij})$ for the transition matrix.

We will only consider time-homogeneous Markov chains in this course, though we will occasionally remark on how some results may be generalized to the time-inhomogeneous case.

Examples

1. *Weather model* Let X_n be the state of the weather on day n in New York, which we assume is either *rainy* or *sunny*. We could use a Markov chain as a crude model for how the weather evolves day-by-day. The state space is $S = \{\text{rain}, \text{sun}\}$. One transition matrix might be

$$P = \begin{array}{c} \\ \text{sun} \\ \text{rain} \end{array} \begin{array}{c} \begin{array}{cc} \text{sun} & \text{rain} \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} \end{array}$$

This says that if it is sunny today, then the chance it will be sunny tomorrow is 0.8, whereas if it is rainy today, then the chance it will be sunny tomorrow is 0.4.

One question you might be interested in is: what is the long-run fraction of sunny days in New York?

2. *Coin flipping* Another two-state Markov chain is based on coin flips. Usually coin flips are used as the canonical example of independent Bernoulli trials. However, Diaconis et al. (2007) studied sequences of coin tosses empirically, and found that outcomes in a sequence of coin tosses are *not* independent. Rather, they are well-modelled by a Markov chain with the following transition probabilities:

$$P = \begin{array}{c} \\ \text{heads} \\ \text{tails} \end{array} \begin{array}{c} \begin{array}{cc} \text{heads} & \text{tails} \end{array} \\ \begin{pmatrix} 0.51 & 0.49 \\ 0.49 & 0.51 \end{pmatrix} \end{array}$$

This shows that if you throw a Heads on your first toss, there is a very slightly higher chance of throwing heads on your second, and similarly for Tails.

3. *Random walk on the line* Suppose we perform a walk on the integers, starting at 0. At each time we move right or left by one unit, with probability 1/2 each. This gives a Markov chain, which can be constructed explicitly as

$$X_n = \sum_{j=1}^{n} \xi_j, \qquad \xi_j = \pm 1 \quad \text{with probability } \frac{1}{2} \text{ each}, \qquad \xi_i \text{ i.i.d.}$$

The transition probabilities are

$$p_{i,i+1} = \frac{1}{2}, \quad p_{i,i-1} = \frac{1}{2}, \quad p_{i,j} = 0 \quad (j \neq i \pm 1).$$

The state space is $S = \{\ldots, -1, 0, 1, \ldots\}$, which is countably infinite.

There are various things we might want to know about this random walk – for example, what is the probability of ever reaching level 10, i.e. what is the probability that $X_n = 10$ for some n? And, what is the average number of steps it takes to do this? Or, what is the average distance to the origin, or squared distance to the origin, as a function of time? Or, how does the probability distribution of the walker evolve? We will see how to answer all of these questions.

4. *Gambler's ruin* This is a modification of a random walk on a line, designed to model certain gambling situations. A gambler plays a game where she either wins 1$ with probability p, or loses 1$ with probability 1-p. The gambler starts with $k$$, and the game stops when she either loses all her money, or reaches a total of $n$$.

   The state space of this Markov chain is $S = \{0, 1, \ldots, n\}$ and the transition matrix has entries

$$P_{ij} = \begin{cases} p & \text{if } j = i+1, \quad 0 < i < n, \\ 1-p & \text{if } j = i-1, \quad 0 < i < n, \\ 1 & \text{if } i = j = 0, \quad \text{or } i = j = n, \\ 0 & \text{otherwise.} \end{cases}$$

   For example, when $n = 5$ and $p = 0.4$ the transition matrix is

$$P = \begin{array}{c c} & \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \left(\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

   Questions of interest might be: what is the probability the gambler wins or loses? Or, after a given number of games m, what is the average amount of money the gambler has?

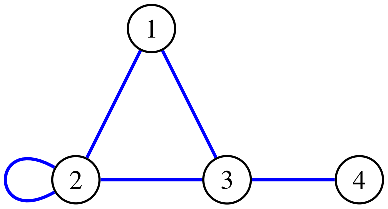
5. *Independent, identically distribute (i.i.d.) random variables* A sequence of i.i.d. random variables is a Markov chain, albeit a somewhat trivial one. Suppose we have a discrete random variable X taking values in $S = \{1, 2, \ldots, k\}$ with probability $P(X = i) = p_i$. If we generate an i.i.d. sequence $X_0, X_1, \ldots$ of random variables with this probability mass function, then it is a Markov chain with transition matrix

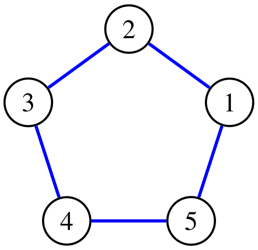
$$P = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & \cdots & k \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ k \end{array} & \left(\begin{array}{cccc} p_1 & p_2 & \cdots & p_k \\ p_1 & p_2 & \cdots & p_k \\ \vdots & \vdots & & \vdots \\ p_1 & p_2 & \cdots & p_k \end{array}\right) \end{array}$$

6. *Random walk on a graph (undirected, unweighted)* Suppose we have a graph (a set of vertices and edge connecting them.) We can perform a random walk on the graph as follows: if we are at node i,

choose an edge uniformly at random from the set of edges leading out of the node, and move along the edge to the node at the edge. Then repeat. If there are N nodes labelled by consecutive integers then this is a Markov chain on state space $S = \{1, 2, \dots, N\}$.

Here is are a couple of examples:



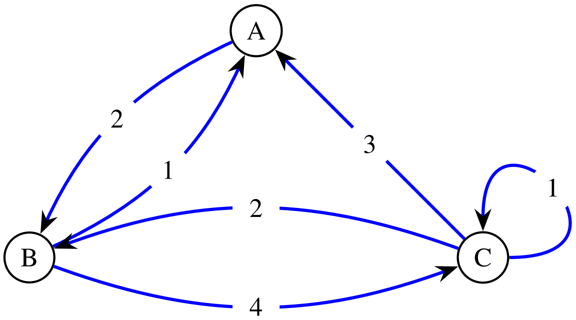


The corresponding transition matrices are:

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \qquad P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix} \end{array}$$

7. *Random walk on a graph (weighted, directed)*

   Every Markov chain can be represented as a random walk on a weighted, directed graph. A weighted graph is one where each edge has a positive real number assigned to it, its "weight," and the random walker chooses an edge from the set of available edges, in proportion to each edge's weight. In a directed graph each edge also has a direction, and a walker can only move in that direction. Here is an example:

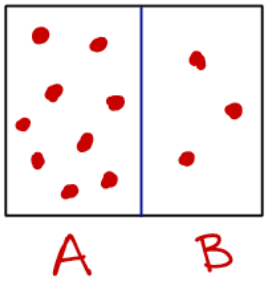


The corresponding transition matrix is:

$$P = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{5} & 0 & \frac{4}{5} \\ \frac{3}{6} & \frac{2}{6} & \frac{1}{6} \end{pmatrix} \end{array}$$

In fact, such a directed graph forms the foundation for Google's Page Rank algorithm, which has revolutionized internet searches. The earliest and best-known version of Page Rank constructs a directed graph of the internet, where nodes are webpages and there is a directed edge from webpage A to webpage B if A contains a link to B. Page Rank assumes an internet surfer clicks follows links at random, and ranks pages according to the long-time average fraction of time that the surfer spends on each page.

8. *Ehrenfest model of diffusion* Consider a container with a membrane in the middle, and m particles distributed in some way between the left and right sides. At each step, pick one particle at random and move it to the other side. Let X_n = # of particles in the left side at time n. Then X_n is a Markov chain, with transition probabilities $p_{i,i+1} = 1 - \frac{i}{m}$, $p_{i,i-1} = \frac{i}{m}$.



9. *Card shuffling* Shuffling a pack of cards can be modeled as a Markov chain. The state space S is the set of permutations of $\{1,2,\ldots,52\}$. A shuffle takes one permutation $\sigma \in S$, and outputs another permutation $\sigma' \in S$ with a certain probability.

   Perhaps the simplest model is the top-to-random shuffle: at each step, take a card from the top of the deck, and put it back in at a random location. The transition matrix has elements

$$P(X_1 = \sigma' | X_0 = \sigma) = \begin{cases} \frac{1}{52} & \text{if } \sigma' \text{ is obtained by taking an item in } \sigma \\ & \text{and moving it to the top,} \\ 0 & \text{otherwise.} \end{cases}$$

   One can also model more complicated shuffles, such as the riffle shuffle. While the state space is enormous ($|S| = 52!$) so you would not want to write down the whole transition matrix, one can still analyze these models using other techniques, from analysis and probability theory. Various authors have proven results about the number of shuffles needed to make the deck "close to random". For example, it takes seven riffle shuffles to get close to random, but it takes 11 or 12 to get so close that a gambler in a casino cannot exploit the deviations from randomness to win a typical game. See the online essay Austin (line) for an accessible introduction to these ideas, and Aldous and Diaconis (1986) for the mathematical proofs. (I first learned about this phenomenon in the beautiful *Proofs from the Book*, by Aigner and Ziegler.)

10. *Autoregressive model of order k (AR(k))* Given constants $a_1,\ldots,a_k \in \mathbb{R}$, let $Y_n = a_1 Y_{n-1} + a_2 Y_{n-2} + \ldots + a_k Y_{n-k} + W_n$, where W_n are i.i.d. random variables.

This process is *not* Markov, because it depends on the past k timesteps. However, we can form a Markov process by defining $X_n = (Y_n, Y_{n-1}, \dots, Y_{n-k+1})^T$. Then

$$X_n = AX_{n-1} + \underline{W}_n,$$

where $A = \begin{pmatrix} a_1 & a_2 & \cdots & & a_k \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & \cdots & & 0 \\ \cdots & \cdots & \cdots & 1 & 0 \end{pmatrix}$, and $\underline{W}_n = (W_n, 0, \dots, 0)^T$.

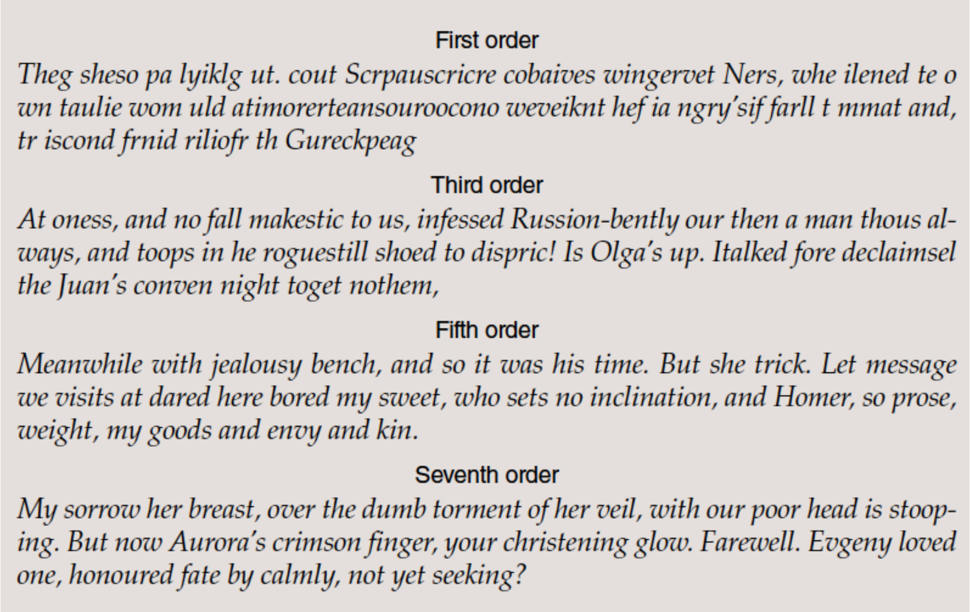
11. *Language, and history of the Markov chain* Markov chains were first invented by Andrei Markov to analyze the distribution of letters in Russian poetry (Hayes (2013)).[2] He meticulously constructed a list of the frequencies of vowel↔consonant pairs in the first 20,000 letters of Pushkin's poem-novel *Eugene Onegin*, and constructed a transition matrix from this data. His transition matrix was:

$$P = \begin{array}{r} \\ \text{vowel} \\ \text{consonant} \end{array} \begin{array}{c} \begin{array}{cc} \text{vowel} & \text{consonant} \end{array} \\ \begin{pmatrix} 0.175 & 0.825 \\ 0.526 & 0.474 \end{pmatrix} \end{array}$$

He showed that from this matrix one can calculate the average number of vowels and consonants. When he realized how powerful this idea was, he spent several years developing tools to analyze the properties of such random processes with memory.

Just for fun, here's an example (from Hayes (2013)) based on Markov's original example, to show how Markov chains can be used to generate realistic-looking text. In each of these excerpts, a Markov chain was constructed by considering the frequencies of strings of k letters from the English translation of the novel *Eugene Onegin* by Pushkin, for $k = 1, 3, 5, 7$, and was then run from a randomly-generated initial condition. You can see that when $k = 3$, there are English-looking syllables, when $k = 5$ there are English-looking words, and when $k = 7$ the words themselves almost fit together coherently.

[2] Actually, he invented Markov chains to disprove a colleague's statement that the Law of Large Numbers can only hold for independent sequences of random variables, and he illustrated his new ideas on this vowel/consonant example.



First order

Theg sheso pa lyiklg ut. cout Scrpauscricre cobaives wingervet Ners, whe ilened te o wn taulie wom uld atimorerteansouroocono weveiknt hef ia ngry'sif farll t mmat and, tr iscond frnid riliofr th Gureckpeag

Third order

At oness, and no fall makestic to us, infessed Russion-bently our then a man thous always, and toops in he roguestill shoed to dispric! Is Olga's up. Italked fore declaimsel the Juan's conven night toget nothem,

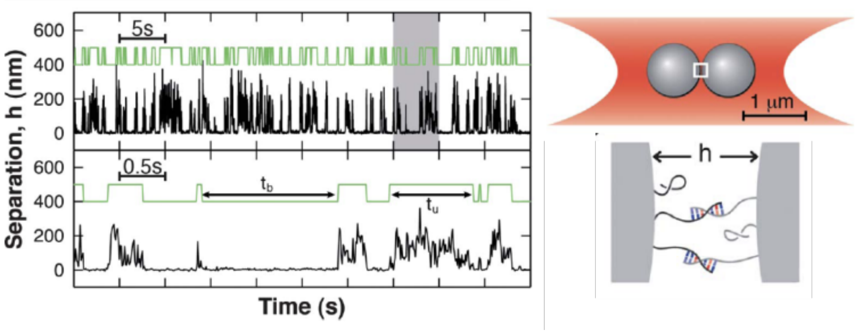
Fifth order

Meanwhile with jealousy bench, and so it was his time. But she trick. Let message we visits at dared here bored my sweet, who sets no inclination, and Homer, so prose, weight, my goods and envy and kin.

Seventh order

My sorrow her breast, over the dumb torment of her veil, with our poor head is stooping. But now Aurora's crimson finger, your christening glow. Farewell. Evgeny loved one, honoured fate by calmly, not yet seeking?

12. *Markov chains in applications.* Markov chains arise in a great many modern applications. Here is an example, from Rogers et al. (2013), where the configuration space of two DNA-coated colloids was modelled as a two-state Markov chain, with states "bound" and "unbound," depending on whether the distance between the particles was small or large:



Some other examples of applications that use Markov chains include:

- models of physical processes
  - rainfall from day-to-day
  - neural networks
  - population dynamics
  - lineups, e.g. in grocery stores, computer servers, telephone call centers, etc.
  - chemical reactions
  - protein folding
  - baseball statistics
- discretize a continuous system

- sampling from high-dimensional systems, e.g. Markov-Chain Monte-Carlo
- data/network analysis
  - clustering
  - speech recognition
  - PageRank algorithm in Google's search engine.

2.2 Evolution of probability

Given a Markov chain with transition probabilities P and initial condition $X_0 = i$, we know how to calculate the probability distribution of X_1; indeed, this is given directly from the transition probabilities. The natural question to ask next is: what is the distribution at later times? That is, we would like to know the n-step transition probabilities $P^{(n)}$, defined by

$$P_{ij}^{(n)} = P(X_n = j|X_0 = i). \tag{3}$$

For example, for $n = 2$, we have that

$$\begin{aligned} P(X_2 = j|X_0 = i) &= \sum_k P(X_2 = j|X_1 = k, X_0 = i)P(X_1 = k|X_0 = i) && \text{Law of Total Probability} \\ &= \sum_k P(X_2 = j|X_1 = k)P(X_1 = k|X_0 = i) && \text{Markov Property} \\ &= \sum_k P_{kj}P_{ik} && \text{time-homogeneity} \\ &= (P^2)_{ij} \end{aligned}$$

That is, the two-step transition matrix is $P^{(2)} = P^2$.

This generalizes:

Theorem. *Let $X_0, X_1, \ldots$ be a time-homogeneous Markov chain with transition probabilities P. The n-step transition probabilities are $P^{(n)} = P^n$, i.e.*

$$P(X_n = j|X_0 = i) = (P^n)_{ij}. \tag{4}$$

To make the notation cleaner we will write $(P^n)_{ij} = P_{ij}^n$. Note that this does *not* equal $(P_{ij})^n$.

***Exercise* 2.3.** Prove this theorem. Hint: use induction.

A more general equation relating the transition probabilities, that holds even in the *time-inhomogeneous* case, is:

Chapman-Kolmogorov Equation.

$$P(X_n = j|X_0 = i) = \sum_k P(X_n = j|X_m = k)P(X_m = k|X_0 = i). \tag{5}$$

Proof.

$$\begin{aligned}
P(X_n=j|X_0=i) &= \sum_k P(X_n=j,X_m=k|X_0=i) && \text{Law of Total Probability}\\
&= \sum_k P(X_n=j|X_m=k,X_0=i)P(X_m=k|X_0=i) && \because P(A\cap B|C)=P(A|B\cap C)P(B|C)\\
&= RHS && \text{by Markov property (see Ex.(2.2))}
\end{aligned}$$

□

Remark. In the time-homogeneous case, the CK equations may be written more compactly as

$$P^{m+n}=P^mP^n\,. \tag{6}$$

When $|S|=\infty$, we understand a power such as P^2 to mean the infinite sum $(P^2)_{ij}=\sum_k P_{kj}P_{ik}$.

Remark. All Markov Processes satisfy a form of Chapman-Kolmogorov equations, from which many other equations can be derived. However, not all processes which satisfy Chapman-Kolmogorov equations, are Markov processes. See Grimmett and Stirzaker (2001), p.218 Example 14 for an example.

Now suppose we don't know the initial condition of the Markov chain, but rather we know the probability distribution of the initial condition. That is, X_0 is a random variable with distribution $P(X_0=i)=a_i$. We will write this as

$$X_0\sim\alpha^{(0)}=(a_1,a_2,\ldots,).$$

Here $\alpha^{(0)}$ is a *row vector* representing the pmf of the initial condition. It is important that it is a row vector – this is the convention, which both simplifies notation and makes it easier to generalize to continuous-state Markov processes.

We would like to calculate the distribution of X_n, which we write as a row vector $\alpha^{(n)}$:

$$\alpha_i^{(n)}=P(X_n=i).$$

Let's calculate $\alpha^{(n+1)}$, assuming we know $\alpha^{(n)}$.

$$\begin{aligned}
\alpha_j^{(n+1)} &= \sum_i P(X_{n+1}=j|X_n=i)P(X_n=i) && \text{Law Of Total Probability}\\
&= \sum_i P_{ij}\alpha_i^{(n)} && \text{time-homogeneity and defn of } \alpha^{(n)}.
\end{aligned}$$

We obtain:

Forward Kolmogorov Equation. *(for a time-homogeneous, discrete-time Markov Chain)*

$$\alpha^{(n+1)}=\alpha^{(n)}P. \tag{7}$$

This equation shows how to evolve the probability in time. The solution is clearly $\alpha^{(n)}=\alpha^{(0)}P^n$, which you could also show directly from the n-step transition probabilities.

***Exercise* 2.4.** Do this! Show that (7) holds, directly from the formula for the n-step transition probabilities.

Therefore, if we know the initial probability distribution $\alpha^{(0)}$, then we can find the distribution at any later time using powers of the matrix P.

Now consider what happens if we ask for the expected value of some function of the state of the Markov chain, such as $\mathbb{E}X_n^2$, $\mathbb{E}X_n^3$, $\mathbb{E}|X_n|$, etc. Can we derive an evolution equation for this quantity?

Let $f : S \to \mathbb{R}$ be a function defined on state space, and let

$$u_i^{(n)} = \mathbb{E}_i f(X_n) = \mathbb{E}[f(X_n)|X_0 = i]. \tag{8}$$

You should think of $u^{(n)}$ as a *column vector*; again this is a convention whose convenience will become more transparent later in the course. Then $u^{(n)}$ evolves in time as:

Backward Kolmogorov Equation. *(for a time-homogeneous, discrete-time Markov Chain)*

$$u^{(n+1)} = Pu^{(n)}, \qquad u^{(0)}(i) = f(i) \quad \forall i \in S. \tag{9}$$

Proof. We have

$$\begin{aligned}
u^{(n+1)}(i) &= \sum_j f(j)P(X_{n+1}{=}j|X_0 = i) && \text{definition of expectation}\\
&= \sum_j \sum_k f(j)P(X_{n+1}{=}j|X_1{=}k, X_0{=}i)P(X_1{=}k|X_0{=}i) && \text{LoTP}\\
&= \sum_j \sum_k f(j)P(X_{n+1}{=}j|X_1{=}k)P(X_1{=}k|X_0{=}i) && \text{Markov property}\\
&= \sum_j \sum_k f(j)P^n_{kj}P_{ik} && \text{time-homogeneity}\\
&= \sum_k \sum_j f(j)P^n_{kj}P_{ik} && \text{switch order of summation}\\
&= \sum_k u^{(n)}(k)P_{ik} && \text{definition of } u^{(n)}\\
&= (Pu)_i
\end{aligned}$$

We can switch the order of summation above, provided we assume that $\mathbb{E}_i|f(X_n)| < \infty$ for each i and each n. □

This proof illustrates a technique sometimes known as first-step analysis, where one conditions on the first step of the Markov chain and uses the Law of Total Probability. Of course, you could also derive this equation more directly from the n-step transition probabilities.

***Exercise* 2.5.** Do this! Derive (9) directly from the formula for the n-step transition probabilities.

Remark. What is so backward about the backward equation? It gets its name from the fact that it can be used to describe how conditional expectations propagate backwards in time. To see this, suppose that instead of (8), which computes the expectation of a function after a certain number of steps has passed, we choose a fixed time T and compute the expectation at that time, given an earlier starting position. That is, for each $n \leq T$, define a column vector $u^{(n)}$ with components

$$u_i^{(n)} = \mathbb{E}[f(X_T)|X_n{=}i]\,. \tag{10}$$

Such a quantity is studied a lot in financial applications, where, say, X_n is the price of a stock at time n, f is a value function representing the value of an option to sell, T might be a time at which you decide (in advance) to sell a stock, and quantities of the form (10) above would represent your expected payout, conditional on being in state i at time n. Then, the vector $u^{(n)}$ evolves according to

$$u^{(n)} = Pu^{(n+1)}, \qquad u_i^{(T)} = f(i) \quad \forall i \in S. \tag{11}$$

Therefore you find $u^{(n)}$ by evolving it *backwards* in time – you are given a final condition at time T, and you can solve for u_n at all earlier times $n \leq T$.

Interestingly, (11) holds even when the chain is not time-homogeneous, provided that P in (11) is replaced by $P(n)$, the transition probabilities starting at time n. This same statement is not true for (9).

***Exercise* 2.6.** Show (11), and argue it holds even when the Markov chain is not time-homogeneous.

2.2.1 Evolution of the full transition probabilities*

Another approach to the forward/backward equations is to define a function $P(j,t|i,s)$ to be the transition probability to be in state j at time t, given the system started in state i at time s, i.e.

$$P(j,t|i,s) = P(X_t = j|X_s = i). \tag{12}$$

One can then derive equations for how $P(j,t|i,s)$ evolves in t and s. For evolution in t (forward in time) we have, from the Chapman-Kolmogorov equations,

$$P(j,t+1|i,s) = \sum_k P(k,t|i,s)P(j,t+1|k,t). \tag{13}$$

For evolution in s (backward in time) we have, again from the Chapman-Kolmogorov equations,

$$P(j,t|i,s) = \sum_k P(j,t|k,s+1)P(k,s+1|i,s). \tag{14}$$

These are general versions of the forward and backward equations, respectively. They hold regardless of whether the chain is time-homogeneous or not. From them, we can derive the time-inhomogeneous versions of the forward and backward equations (7), (9).

To derive the time-inhomogeneous forward equation, notice that the probability distribution at time t, $\alpha^{(t)}$, has components $\alpha_j^{(t)} = \sum_i P(j,t|i,0)\alpha_i^{(0)}$. Therefore, multiplying (13) by $\alpha^{(0)}$ on the left (contracting it with index i) and letting $s = 0$, we obtain

$$\alpha_j^{(t+1)} = \sum_k \alpha_k^{(t)} P(j,t+1|k,t) \qquad \Leftrightarrow \qquad \alpha^{(t+1)} = \alpha^{(t)} P_{kj}^{(t)}\,. \tag{15}$$

To derive the time-inhomogeneous backward equation, let $f : S \to \mathbb{R}$, and let $u_i^{(s)} = \mathbb{E}[f(X_t)|X_s = i]$ (recall (8),(10).) Notice that $u_i^{(s)} = \sum_k f(k)P(k,t|i,s)$, so multiplying (14) by the column vector f on the right (contracting it with index j) gives

$$u_i^{(s)} = \sum_k P(k,s+1|i,s)u_k^{(s+1)} \qquad \Leftrightarrow \qquad u^{(s)} = P_{ij}^{(s)} u^{(s+1)}\,. \tag{16}$$

2.3 Long-time behaviour and stationary distribution

Suppose we take a Markov chain and let it run for a long time. What happens? Clearly the chain itself does not converge to anything, because it is continually jumping around, but the probability distribution might. Before getting into the theory, let's look at some examples.

Consider the two-state weather model from section 2.1, example 1. Suppose it is raining today. What is the probability distribution for the weather in the future? We calculate this from the n-step transition probabilities (4):

n	P(sun)	P(rain)
0	0	1
1	0.4000	0.6000
2	0.5600	0.4400
3	0.6240	0.3760
4	0.6496	0.3504
5	0.6598	0.3402
6	0.6639	0.3361
7	0.6656	0.3344
8	0.6662	0.3338
9	0.6665	0.3335
10	0.6666	0.3334
11	0.6666	0.3334
12	0.6667	0.3333
13	0.6667	0.3333
14	0.6667	0.3333

It seems like the probability distribution is converging to something. After 12 days, the distribution doesn't change, to 4 digits. You can check that the distribution it converges to does not depend on the initial condition. For example, if we start with a sunny day, $\alpha^{(0)} = (1,0)$, then $\alpha^{(10)} = (0.6666, 0.3334)$, $\alpha^{(11)} = (0.6666, 0.3334)$, $\alpha^{(12)} = (0.6667, 0.3333)$. In fact, this example is simple enough that you could work out the n-step transition probabilities for any initial condition analytically, and show they converge to $(2/3, 1/3)$.

***Exercise* 2.7.** Do this! (Hint: calculate eigenvalues of the transition matrix.)

Does the probability always converge? Let's look at another example. Consider a Markov chain on state space $\{0,1\}$ with transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Suppose the random walker starts at state 0. Its distribution at time n is:

n	P(0)	P(1)
0	1	0
1	0	1
2	1	0
3	0	1
4	1	0
5	0	1
6	1	0
$\vdots$	$\vdots$	$\vdots$

You can see the pattern. Clearly the distribution doesn't converge. Yet, if we start with initial distirbution $\alpha^{(0)} = (0.5, 0.5)$, then we obtain

n	P(0)	P(1)
0	0.5	0.5
1	0.5	0.5
2	0.5	0.5
$\vdots$	$\vdots$	$\vdots$

The distribution never changes!

2.3.1 Limiting and stationary distributions

In applications we are often interested in the long-term probability of visiting each state.

Definition. Consider a time-homogeneous Markov chain with transition matrix P. A row vector λ is a *limiting distribution* if $\lambda_i \geq 0$, $\sum_j \lambda_j = 1$ (so that λ is a probability distribution), and if, for every i,

$$\lim_{n\to\infty} (P^n)_{ij} = \lambda_j \qquad \forall j \in S.$$

In other words,

$$P^n \to \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \dots \\ \lambda_1 & \lambda_2 & \lambda_3 & \dots \\ \lambda_1 & \lambda_2 & \lambda_3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \qquad \text{as } n \to \infty.$$

***Exercise* 2.8.** Show that, if $|S| < \infty$, then λ is a limiting distribution if and only if the definition $\lim_{n\to\infty} \alpha P^n = \lambda$ for any initial probability distribution α.

As we saw in the earlier examples, a limiting distribution doesn't have to exist. If it exists, it must be unique. What happens if we start the chain in the limiting distribution? Let's calculate the distribution $\alpha^{(1)}$ at the next step of the chain, assuming initial distribution $\alpha^{(0)} = \lambda$. For simplicity, we will assume a finite state space, $|S| < \infty$, which lets us interchange the sum and the limit in the calculations below. Choose any i, and calculate, from (7), (writing $A_{i,\cdot}$ for the row vector corresponding to the ith row of matrix A):

$$\alpha^{(1)} = \lambda P = \left(\lim_{n\to\infty} P^n_{i,\cdot}\right) P = \left(\lim_{n\to\infty} P^{n+1}_{i,\cdot}\right) = \lambda.$$

Therefore if we start the chain in the limiting distribution, its distribution remains there forever. This motivates the following definition:

Definition. Given a Markov chain with transition matrix P, a *stationary distribution* is a probability distribution π which satisfies

$$\pi = \pi P \qquad \Longleftrightarrow \qquad \pi_j = \sum_i \pi_i P_{ij} \quad \forall\, j. \tag{17}$$

This says that that if we start with distribution π and run the Markov chain, the distribution will not change. That is why it is called "stationary." In other words, if $X_0 \sim \pi$, then $X_1 \sim \pi$, $X_2 \sim \pi$, etc.

Remark. Other synonyms you might hear for stationary distribution include *invariant measure*, *invariant distribution*, *steady-state probability*, *equilibrium probability* or *equilibrium distribution* (the latter two are from physics.).

In applications we want to know the limiting distribution, but it is usually far easier to calculate the stationary distribution, because it is obtained by solving a system of linear equations. Therefore we will restrict our focus to the stationary distribution. Some questions we might ask about π include:

(i) Does it exist?
(ii) Is it unique?
(iii) When is it a limiting distribution, i.e. when does an arbitrary distribution converge to it?

For (iii), we saw that a limiting distribution is a stationary distribution, but the converse is not always true. Indeed, in our second example, you can calculate that a stationary distribution is $\pi = (0.5, 0.5)$, but this is not a limiting distribution. What are the conditions that guarantee a stationary distribution is also the limiting distribution?

This is the subject of a rich body of work on the limiting behaviour of Markov chains. We will not go deeply into the results, but will briefly survey a couple of the major theorems.

2.3.2 A limit theorem or two

Definition. A matrix A is *positive* if it has all positive entries: $A_{ij} > 0$ for all i, j. In these notes we will write $A > 0$ when A is positive.

Remark. This is *not* the same as being positive-definite!

Definition. A stochastic matrix is *regular* if there exists some $s > 0$ such that P^s is positive, i.e. the s-step transition probabilities are positive for all i, j: $(P^s)_{ij} > 0 \ \forall i, j$.

Remark. Some books call such a matrix *primitive*. The text Koralov and Sinai (2010)) calls it *ergodic* (when the state space is finite), though usually this word is reserved for something slightly different.

This means that there is a time s such that, no matter where you start, there is a non-zero probability of being at any other state.

Theorem (Ergodic Theorem for Markov Chains, (one version))**.** *Assume a Markov Chain is regular and has a finite state space with size N. Then there exists a unique stationary probability distribution* $\pi = (\pi_1, \ldots, \pi_N)$*, with* $\pi_j > 0\, \forall j$*. The n-step transition probabilities converge to* π*: that is,* $\lim_{n\to\infty} P^n_{ij} = \pi_j$.

Remark. The name of this theorem comes from Koralov and Sinai (2010); it may not be universal. There are many meanings of the word "ergodic"; we will see several variants throughout this course.

Remark. For a Markov chain with an infinite state space, the Ergodic theorem holds provided the chain is irreducible (see below), aperiodic (see below), and all states have finite expected return times (ask instructor – this condition is always true for finite irreducible chains.) For a proof, see Dobrow (2016), section 3.10.

Proof. This is a sketch, see Koralov and Sinai (2010) p.72 for all the details.

- Define a metric $d(\mu',\mu'') = \frac{1}{2}\sum_{i=1}^{N}|\mu_i' - \mu_i''|$ on the space of probability distributions. It can be shown that d is a metric, and the space of distributions is complete.
- Show (*) $d(\mu'Q,\mu''Q) \leq (1-\alpha)d(\mu',\mu'')$, where $\alpha = \min_{i,j} Q_{ij}$ and Q is a stochastic matrix.
- Show that $\mu^{(0)}, \mu^{(0)}P, \mu^{(0)}P^2, \ldots$ is a Cauchy sequence. Therefore it converges, so let $\pi = \lim_{n\to\infty}\mu^{(n)}$.
- Show that π is unique: let π_1,π_2 be two distributions such that $\pi_i = \pi_1 P$. Then $d(\pi_1,\pi_2) = d(\pi_1 P^s, \pi_2 P^s) \leq (1-\alpha)d(\pi_1,\pi_2)$ by (*). Therefore $d(\pi_1,\pi_2) = 0$.
- Let $\mu^{(0)}$ be the probability distribution which is concentrated at point i. Then $\mu^{(0)}P^n$ is the probability distribution $(p_{ij}^{(n)})$. Therefore, $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$.

□

Remark. This shows that $d(\mu^{(0)}P^n,\pi) \leq (1-\alpha)^{-1}\beta^n$, with $\beta = (1-\alpha)^{1/s} < 1$. Therefore the rate of convergence to the stationary distribution is exponential.

There is also a version of the Law of Large Numbers.

Theorem (Law of large numbers for a regular Markov chain)**.** *Let π be the stationary distribution of a regular Markov chain, and let $\nu_i^{(n)}$ be the number of occurences of state i in after n time steps of the chain, i.e. among the values of $X_0, X_1, \ldots, X_n$. Let $\nu_{ij}^{(n)}$ be the number of values of k, $1 \leq k \leq n$, for which $X_{k-1} = i, X_k = j$. Then for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\left(\left|\frac{\nu_i^{(n)}}{n} - \pi_i\right| \geq \varepsilon\right) = 0$$

$$\lim_{n\to\infty} P\left(\left|\frac{\nu_{ij}^{(n)}}{n} - \pi_i p_{ij}\right| \geq \varepsilon\right) = 0$$

Proof. For a proof, see Koralov and Sinai (2010), p. 74. □

Remark. This implies that the long-time average of any function of a regular Markov chain, $\frac{1}{n}\sum_{i=1}^{n} f(X_n)$, approaches the average with respect to the stationary distribution, $\mathbb{E}_\pi f(X)$.

This theorem can be weakened slightly by allowing for Markov chains with some kind of periodicity. We need to consider a chain which can move between any two states (i,j), but not necessarily at a time s that is the same for all pairs.

Definition. A stochastic matrix is *irreducible* if, for every pair (i,j) there exists an $s > 0$ such that $(P^s)_{ij} > 0$.

There are limit theorems for irreducible chains, with slightly weaker conditions. Irreducible chains also have a unique stationary distribution – this follows from the Perron-Frobenius Theorem (see below.) However, it is not true that an arbitrary distribution converges to it; rather, we have that $\mu^{(0)}\bar{P}^{(n)} \to \pi$ as $n \to \infty$, where $\bar{P}^{(n)} = \frac{1}{n}\sum_{k=1}^{n} P^k$. This means that the average distribution converges. We need to form the average, because there may be a built-in periodicity, as in the chain in the second example. In this case $P^{2n} = I$, and $P^{2n+1} = P$, so α_n oscillates between two distributions, instead of converging to a fixed limit.

2.3.3 The linear algebra connection

Questions about the stationary and limiting distributions can also be addressed using linear algebra, by examining the eigenvalues of P. (We assume in this section that $|S| = N < \infty$.) Indeed, if π is a stationary distribution, then π is a left eigenvector of P corresponding to eigenvalue $\lambda = 1$.

We know that P *has* an eigenvalue $\lambda = 1$, since the rows of P sum to 1 so we have

$$P\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} = \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix},$$

and therefore $(1,1,\dots,1)^T$ is a right eigenvector. To ensure that the corresponding left eigenvector is a stationary distribution, we need to know that its entries are all nonnegative.

Let's put this issue on hold for a moment, and just assume that the corresponding left eigenvector π is a stationary distribution. When is it also a limiting distribution? Suppose that P has a full set of eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_N$ which are distinct, with $\lambda_1 = 1$. Then there exists a matrix B such that

$$P = B^{-1}\Lambda B, \qquad \text{where} \quad \Lambda = \begin{pmatrix}\lambda_1 & 0 & 0 & \cdots & 0\\ 0 & \lambda_2 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots & 0\\ 0 & \cdots & 0 & 0 & \lambda_N\end{pmatrix}. \tag{18}$$

The rows of B are left eigenvectors of P, and the columns of B^{-1} are right eigenvectors. Therefore

$$P^n = B^{-1}\Lambda^n B, \qquad \text{where} \quad \Lambda = \begin{pmatrix}\lambda_1^n & 0 & 0 & \cdots & 0\\ 0 & \lambda_2^n & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots & 0\\ 0 & \cdots & 0 & 0 & \lambda_N^n\end{pmatrix}. \tag{19}$$

What happens as $n \to \infty$? For the first eigenvalue we have $\lambda_1^n = 1$. Any eigenvalue such that $|\lambda_i| < 1$ will converge to zero, $\lambda_i^n \to 0$. Therefore, there is a limiting distribution, only if $|\lambda_i| < 1$ for $i \geq 2$. In this case we have

$$\lim_{n\to\infty} P^n = B^{-1}\begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & 0\end{pmatrix}B.$$

We know that the right eigenvector associated with λ_1 is $\mathbf{v} = (1,1,\dots,1)^T$. By assumption, the left eigenvector is a stationary distribution π. Therefore we have

$$\lim_{n\to\infty} P^n = \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}\begin{pmatrix}\pi_1 & \pi_2 & \cdots & \pi_N\end{pmatrix} = \begin{pmatrix}\pi_1 & \pi_2 & \cdots & \pi_N\\ \pi_1 & \pi_2 & \cdots & \pi_N\\ \vdots & \vdots & \vdots & \vdots\\ \pi_1 & \pi_2 & \cdots & \pi_N\end{pmatrix},$$

so π is also a limiting distribution. (If P does not have a full set of distinct eigenvalues, then we can do a similar calculation using the Jordan canonical form of the matrix.)

We can justify the above calculations using some results from linear algebra.

Lemma. *The spectral radius of a stochastic matrix P is 1, i.e.* $\rho(P) = \max_\lambda |\lambda| = 1$, *where the max is over all eigenvalues.*

Proof. Let η be a left eigenvector with eigenvalue λ. Then $\lambda\eta_i = \sum_{j=1}^N \eta_j p_{ji}$,

$$|\lambda|\sum_{i=1}^N|\eta_i| = \sum_{i=1}^N|\sum_{j=1}^N \eta_i p_{ji}| \le \sum_{i,j=1}^N |\eta_j| p_{ji} = \sum_{j=1}^N |\eta_j|.$$

Therefore $|\lambda| \le 1$. □

Whew. This is good news – it shows that no eigenvalue of P has complex norm greater than 1 – but it still doesn't rule out the possibility that there are other eigenvalues with complex norm equal to 1. But, you may recall this theorem from linear algebra.

Theorem. (Perron-Frobenius Theorem, for aperiodic positive matrices.) *Let M be a positive*[3] $k\times k$ *matrix, with* $k < \infty$. *Then the following statements hold:*

(i) There is a positive real number λ_1 *which is an eigenvalue of M. All other eigenvalues* λ *of M satisfy* $|\lambda| < \lambda_1$.

(ii) The eigenspace of eigenvectors associated with λ_1 *is one-dimensional.*

(iii) There exists a positive right eigenvector $\mathbf{v}$ *and a positive left eigenvector* $\mathbf{w}$ *associated with* λ_1.

(iv) M has no other eigenvector with nonnegative entries.

For a proof, see an advanced linear algebra textbook, such as Lax (1997), Chapter 16. There is also a brief description of the proof in Strang (1988), section 5.3.

The Perron-Frobenius theorem implies that if P is positive, then it has a one-dimensional eigenspace associated with the eigenvalue $\lambda = 1$, and the corresponding left eigenvector π is positive. Therefore, π is the unique stationary distribution. The theorem also shows that all other eigenvalues have complex norm less than 1, so combined with the calculations above (or an enhanced version of the Perron-Frobenius theorem[4]), we have that π is a limiting distribution.

[3] $M_{ij} > 0$ for all i, j

[4] See for example Theorems 8.2.7, 8.2.8, in "Matrix Analysis" by Horn & Johnson.

With a little more work, one can show that the above statements hold for a transition matrix that is *regular* ($\exists\, s > 0$ such that P^s is positive.)

There is a more general version of the Perron-Frobenius theorem that works for finite, *irredicible* matrices (not only positive ones.) This allows a Markov chain to have some built-in periodicity. For curiosity's sake we will state the theorem here, but it will not be important for the course.

Definition. The *period* $d(i)$ of a state i is defined by $d(i) = \gcd\{n : P^n_{ii} > 0\}$, i.e. it is the greatest common divisor of all the possible number of steps it takes to return to the state, if you start at the state. An irreducible Markov chain is called *periodic with period* d if all states have period $d > 1$. An irreducible Markov chain is called *aperiodic* all states have period equal to 1.

We state the theorem for transition matrices, the way it appears in Grimmett and Stirzaker (2001), section 6.6.

Theorem. (Perron-Frobenius Theorem for irreducible matrices)

If P is the transition matrix of a finite irreducible chain with period d then:

(i) $\lambda_1 = 1$ *is an eigenvalue of P,*

(ii) the d complex roots of unity

$$\lambda_1 = \omega^0,\ \lambda_2 = \omega^1, \ldots, \lambda_d = \omega^{d-1} \quad \textit{where } \omega = e^{2\pi i/d},$$

are eigenvalues of P,

(iii) the remaining eigenvalues $\lambda_{d+1}, \ldots, \lambda_N$ satisfy $|\lambda_j| < 1$.

2.4 Mean first passage time

Sometimes we want to ask about how long it takes a Markov chain to do something: how long until the weather turns sunny again, how long will a gambler play a game if she stops when she has won a given amount of money or goes broke, and when she stops playing, is it because she won money, or went broke? Answering these questions requires asking about the probability distributions of random times, that depend on the realization of a Markov chain. We can't handle question about any kind of random time, but there is a certain class of random times that can be tractably handled using the ideas of Markov chains and the tools of linear algebra.

Definition. A *stopping time T* for a discrete-time Markov chain is a random variable taking values in $\{0,1,2,\ldots,\} \cup \{\infty\}$ with the property that the indicator function $1_{\{T=n\}}$ for the event $\{T = n\}$ is a function only of the variables $X_0, X_1, X_2, \ldots, X_n$.

That is, T is a stopping time if we can decide whether $T = n$ using only knowledge of the past and present states of the Markov chain, with no information about the future.[5]

A stopping time that we will be particularly interested in is the time it takes for a Markov chain to first hit a given set.

[5] The rigorous definition of a stopping time, is that, for all $n \geq 0$, the event $\{T = n\}$ is measurable with respect to the σ-algebra generated by $X_0, X_1, X_2, \ldots, X_n$.

Definition. The *first-passage time* or *first-hitting time* of a set $A \subset S$ is defined by

$$T_A = \min\{n \geq 0 : X_n \in A\}.$$

To show that T_A is a stopping time, observe that

$$\{T_A = n\} = \{X_0 \in A^c, X_1 \in A^c, \ldots, X_{n-1} \in A^c, X_n \in A\},$$

and the event on the right-hand side depends only on the random variables $X_0, \ldots, X_n$.

Here are some other examples of stopping times:

1. $T = c$, where $c \in \mathbb{N}$ is a constant.
2. Given two stopping times S and T the random variables $U = S \wedge T$ (minimum of S, T) and $V = S \vee T$ (maximum of S, T) are stopping times.
3. Given two stopping times S and T the random variable $\tau = S + T$ is a stopping time.
4. $T = \min\{n \geq 0 : X_i > a \text{ for } i \in \{n-2, n-1, n\}\}$, where $a \in \mathbb{R}$ is some constant. That is, the first time the process has remained above a level for a sufficiently long amount of time.

***Exercise* 2.9.** Show that all of the examples above are stopping times.

An examples of a random times that is *not* a stopping time is the *last* visit to a set A, i.e. $T = \max\{n : X_n \in A\}$. The event $\{T = n\}$ depends on all future values $X_n, X_{n+1}, \ldots$ so it cannot be a stopping time.

Here are some other examples of random times that are *not* stopping times:

1. $T - 1$, where T is a stopping time.
2. $S - T$, where S, T are stopping times.
3. $\frac{1}{2}(S + T)$ where S, T are stopping times.
4. T = first time to reach $\max(X_0, X_1, \ldots)$ (such as, in gambling, the first time to reach the maximum amount of money you will ever reach.)

***Exercise* 2.10.** Argue why each of the above examples is not a stopping time.

We can answer many questions about stopping times, by solving linear equations. A common quantity of interest is the average time it takes to hit a set $A \subset S$.

Definition. The *mean first passage time* (mfpt) to set A starting at state i is

$$\tau_i = \mathbb{E}(T_A | X_0 = i). \tag{20}$$

Let's compute the mfpt τ_i, using a first-step analysis. Let's assume that $P(T_A < \infty | X_0 = i) = 1$ for all $i \in S$, and furthermore that $\tau_i < \infty$ for all $i \in S$.

For $i \in A$, we know that $T_A = 0$. Consider $i \notin A$. Then we have

$$\begin{aligned}
\tau_i &= \sum_{t=1}^{\infty} tP(T_A{=}t|X_0{=}i) \\
&= \sum_{t=1}^{\infty}\sum_{j=1}^{\infty} tP(T_A{=}t|X_0{=}i, X_1{=}j)P(X_1{=}j|X_0{=}i) && \text{LOTP} \\
&= \sum_{t=1}^{\infty}\sum_{j=1}^{\infty} tP(T_A{=}t|X_1{=}j)P(X_1{=}j|X_0{=}i) && \text{Markov property}
\end{aligned}$$

Because the chain is time-homogeneous, we expect that $P(T_A{=}t|X_1{=}j) = P(T_A{=}t-1|X_0{=}j)$. To show this explicitly, write

$$\begin{aligned}
P(T_A{=}t|X_1{=}j) &= P(X_2 \in A^c, \ldots, X_{t-1} \in A^c, X_t \in A|X_1 = j) && \text{by definition} \\
&= P(X_1 \in A^c, \ldots, X_{t-2} \in A^c, X_{t-1} \in C|X_0 = j) && \text{by time-homogeneity} \\
&= P(T_A{=}t-1|X_0{=}j).
\end{aligned}$$

Therefore, substituting into the above and changing the index $t \to t+1$, we have

$$\begin{aligned}
\tau_i &= \sum_{t=0}^{\infty}\sum_{j=1}^{\infty} (t+1)P(T_A{=}t|X_0{=}j)P_{ij} \\
&= \sum_{j=1}^{\infty}\sum_{t=0}^{\infty} tP(T_A{=}t|X_0{=}j)P_{ij} + \sum_{j=1}^{\infty}\sum_{t=0}^{\infty} P(T_A{=}t|X_0{=}j)P_{ij} \\
&= \sum_{j=1}^{\infty} \tau_j P_{ij} \quad + \ 1.
\end{aligned}$$

The second term is 1, because $\sum_{t=0}^{\infty} P(T_A{=}t|X_0{=}j) = 1$, since this sum is the probability that T_A takes any value (we are assuming that $P(T_A < \infty) = 1$.) Summing over j gives $\sum_{j=1}^{\infty} P_{ij} = 1$, which holds because we are simply summing the rows of P, which form a probability distribution. We can interchange the order of summation in the second step, because all the terms we are adding up are nonnegative, and we assume the sum exists since the mfpt exists.

We just showed the following:

Theorem. *Let $\tau = (\tau_1, \tau_2, \ldots)^T$ be a vector of mean first passage times from each state $i \in S$. Then τ solves the following system of equations:*

$$\tau_i = \begin{cases} 0 & i \in A \\ 1 + \sum_j P_{ij}\tau_j & i \notin A. \end{cases} \tag{21}$$

Remark. Another way to write (21) is

$$P'\tau' + \mathbf{1} = \tau', \tag{22}$$

where P' is P with the rows and columns corresponding to elements in A removed, and τ' is τ with the elements in A removed. Equation (22) can in turn be written as

$$(P' - I)\tau' = -1. \tag{23}$$

This form will make it easier to make the connection to continuous-time Markov chains and processes, later in the course.

Equation (21) gives a way to find the mean first passage time by solving a linear system of equations. Note that we can't find the mfpt from state i in isolation; we have to solve for the mfpt from all states i simultaneously. For systems that are not too large, this means we can use built-in linear algebra solvers to calculate mfpts. If the problem has some extra structure, we can sometimes even find analytical solutions.

***Exercise* 2.11.** Suppose you perform a random walk on the integers where at each step you jump left or right with equal probability, and let X_n be your position at time n. Calculate the mean first passage time τ_0 to leave the interval $(-6,6)$, starting at $X_0 = 0$.

References

Aldous, D. and Diaconis, P. (1986). Shuffling cards and stopping times. *American Mathematical Monthly*, 93:333–348.

Austin, D. (online). How many times do I have to shuffle this deck? `http://www.ams.org/samplings/feature-column/fcarc-shuffle`.

Diaconis, P., Holmes, S., and Montgomery, R. (2007). Dynamical Bias in the Coin Toss. *SIAM Rev.*, 49(2):211–235.

Dobrow, R. P. (2016). *Introduction to Stochastic Processes with R*. Wiley.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes*. Oxford University Press.

Hayes, B. (2013). First links in the Markov chain. *American Scientist*, 101.

Koralov, L. B. and Sinai, Y. G. (2010). *Theory of Probability and Random Processes*. Springer.

Lax, P. (1997). *Linear Algebra*. John Wiley & Sons.

Norris, J. R. (1997). *Markov Chains*. Cambridge University Press.

Rogers, W. B., Sinno, T., and Crocker, J. C. (2013). Kinetics and non-exponential binding of DNA-coated colloids. *Soft Matter*, 9(28):6412–6417.

Strang, G. (1988). *Linear Algebra and its Applications*. Brooks/Cole, 3rd edition.