Brief Introduction to Markov Chains

James Nolen

Math 340, Spring 2024

These notes give a summary of some basic ideas about Markov chains and branching processes. There are many good references about Markov Chains, for example:

- G. Grimmett and D. Stirzaker, *Probability and Random Processes*, Oxford University Press, 2001.
- G. Lawler, *Introduction to Stochastic Processes*, Chapman and Hall, 2006.
- J. R. Norris, *Markov Chains*, Cambridge University Press, 1998.
- R. Durrett, *Elementary Probability for Applications*, Cambridge University Press, 2009.

1 Discrete time chains

A **Markov chain** is a sequence of random variables $\{X_n\}_{n=0}^{\infty}$ which take values in some set $\mathcal{S}$, called the **state space**, and which satisfy the Markov property. For this course, we will always assume $\mathcal{S}$ is a countable set; sometimes we will assume that $\mathcal{S}$ is a finite set. Thus, the variables $\{X_n\}_{n\geq 0}$ will be *discrete* random variables. We usually think of the index n as a discrete "time" index; X_n is the state of the system at time n. The state of the system evolves in a random way.

The state space $\mathcal{S}$ and the "state of the system" at time n, represented by X_n could be many things, not necessarily numerical. For example, maybe X_n represents a simplistic description of the weather on day n, either sunny or cloudy. In that case, $\mathcal{S} = \{\text{cloudy}\,, \text{sunny}\}$. Or, in another example, X_n may represent the conformation state of a protein as it changes randomly; $\mathcal{S}$ is the set of all conformations. Or, maybe X_n is the size of a population, and $\mathcal{S}$ is the set of positive integers. In some early work (1913) on discrete-time random processes, Andrey Markov (for whom these processes are named) famously studied the distribution of vowels and consonants in the literary work of Alexander Pushkin.

To be a Markov chain, the sequence X_n must satisfy the **Markov Property**, which is the condition that

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n,\ X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = y \mid X_n = x_n) \tag{1.1}$$

holds for any choice of states $y, x_n, x_{n-1}, \ldots, x_0 \in \mathcal{S}$, and for any $n \geq 1$. The term on the left, $\mathbb{P}(X_{n+1} = y \mid X_n = x_n,\ X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)$, is the conditional probability that

$X_{n+1} = y$, given all the past states up to time n. The term on the right, $\mathbb{P}(X_{n+1} = y \mid X_n = x_n)$ is the conditional probability that $X_{n+1} = y$, given only the state at time n. So, roughly speaking, the Markov property may be interpreted to mean that if we know the state at time n (i.e. given $X_n = x_n$), we cannot improve our prediction of X_{n+1} by knowing any more information about the history of previous states. Thus, the "rule" for updating the system from time n to time $n+1$ may involve some randomness and it may depend on X_n; but given X_n, this update "rule" does not depend on other information from the past.

Remark 1.1 *An equivalent way to state the Markov property is that for any $n \geq 0$, and states $x, y \in \mathcal{S}$ and any subsets $A_k \subset \mathcal{S}$ for $k = 0, 1, \ldots, n-1$,*

$$\mathbb{P}(X_{n+1} = y \mid X_n = x,\ X_{n-1} \in A_{n-1}, \ldots, X_1 \in A_1,\ X_0 \in A_0) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$$

In particular, by choosing $A_k = \mathcal{S}$ for $k = 1, \ldots, n-1$ and $A_0 = \{x_0\}$, this implies statements like

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n,\ X_0 = x_0) = \mathbb{P}(X_{n+1} = y \mid X_n = x_n),$$

even though we don't explicitly condition on intermediate states as in (1.1).

Assuming the chain is time-homogeneous, the **transition probability matrix** is

$$P(x, y) = \mathbb{P}(X_1 = y \mid X_0 = x) = \mathbb{P}(X_{n+1} = y \mid X_n = x), \qquad x, y \in \mathcal{S}. \tag{1.2}$$

This means that $P(x, y)$ is the probability of moving from state x to state y in one step. We think of this as a (possibly infinite dimensional) matrix, indexed by states $x, y \in \mathcal{S}$. The dimensions of the transition matrix P are $|\mathcal{S}| \times |\mathcal{S}|$, where $|\mathcal{S}|$ is the number of possible states. Sometimes we use the simpler notation

$$P_{xy} = P(x, y)$$

for the transition matrix. The term **time-homogeneous** refers to the second inequality in (1.2): the one-step transition probabilities don't change with the time index n. Unless stated otherwise, we will always b talking about time-homogeneous Markov chains. Because of the axioms of probability, transition matrices must satisfy

$$0 \leq P(x, y) \leq 1, \quad \forall\ x, y \in \mathcal{S}; \qquad \sum_{y \in \mathcal{S}} P(x, y) = 1, \quad \forall\ x \in \mathcal{S}.$$

The transition matrix is essential for studying the chain X_n. Because of the Markov property, the distribution of X_n, the joint distribution of $(X_1, X_2, \ldots, X_n)$, and many other quantities of interest can be expressed in terms of this matrix P.

Example

One of the most famous examples of a Markov chain is a **random walk** on the integers. For this Markov chain, the state space is $\mathcal{S} = \mathbb{Z}$. At each step, toss a p-coin. If the coin lands heads, then X moves to the right (increases by one); if the coin lands tails, then X moves to the left (decreases by one). Thus,

$$X_n = X_0 + \#(\text{heads tossed up to time } n) - \#(\text{tails tossed up to time } n).$$

Also, for every n,

$$P(x,y) = \mathbb{P}(X_{n+1} = y \mid X_n = x) = \begin{cases} p, & \text{if } y = x+1 \\ 1-p, & \text{if } y = x-1 \\ 0, & \text{otherwise.} \end{cases}$$

If $p = 1/2$ (using a fair coin), then it is said to be a **unbiased** random walk or **simple** random walk. □

Example

The idea of a random walk on $\mathbb{Z}$ can be generalized to other discrete sets. Here is a description of a **simple random walk on a graph**. Suppose that $G = (V, E)$ is a finite graph consisting of vertices V and edges E which link the vertices. Suppose the edges are undirected and that any pair of vertices $(x, y) \in V \times V$ has at most one edge between them. Define the random walk on G as follows: Let $\mathcal{S} = V$. If $X_n = x$ (the walker is at vertex x at time n), then choose the next location uniformly at random from all the neighbors of x (i.e. from those vertices connected to x by an edge). That is, if the degree of the vertex x is k, then for each of the k vertices y connected to x,

$$P(x,y) = \frac{1}{k} = \frac{1}{\deg(x)}. \tag{1.3}$$

In this way, the walker hops randomly from vertex to vertex. If the graph is connected, then the walker can visit any vertex in the graph (although it could take many steps). □

Example

Suppose we have N marbles distributed among two urns (a red urn and a blue urn). At each step, choose a marble uniformly at random (from among all the marbles), remove it from its current urn, and put it in the other urn. So, for each urn, the number of marbles in that urn changes by ± 1 at each step. Let X_n the be the number of marbles in the red urn after the n^{th} step; the blue urn contains $N - X_n$ marbles. This defines a Markov chain on the state space $\mathcal{S} = \{0, 1, 2, \ldots, N\}$. For this chain the transition probabilities are:

$$P_{ij} = \begin{cases} \frac{i}{N}, & \text{if } j = i-1, \\ \frac{N-i}{N}, & \text{if } j = i+1, \\ 0, & \text{otherwise} \end{cases} \tag{1.4}$$

The ratio i/N is the probability that a marble in the red urn is chosen, while $(N-i)/N$ is the probability that a marble in the blue urn is chosen. Notice that when $X_n = N$, the next state is $N - 1$, with probability one; whenever $X_n = 0$, the next state must be $X_{n+1} = 1$. When X_n is large (more marbles in the red urn), then we are more likely to choose a marble from the red urn, making X_n decrease. When X_n is small (most marbles are in the blue urn), then we will more likely choose a marble from the blue urn, making X_n increase.□

Example

This example is a discrete-time version of what is called the Moran model, from population genetics. Consider a population of size N. Each individual is one of two types (say, red or blue). At each time step, the system evolves in the following way: First, one of the individuals is chosen uniformly at random to be eliminated from the population; and another individual is chosen uniformly at random to produce one offspring identical to itself. These two choices are made independently. So, if a red individual is chosen to reproduce, and a blue one is chosen for elimination, then the total number of red particles increases by one and the number of blue particles decreases by one. If a red is chosen for reproduction and a red is chosen for elimination, then there is no net change in the number of reds and blues. Let X_n be the number of red individuals at time n. The transition matrix for this chain is

$$P_{ij} = \begin{cases} \frac{i}{N}\left(\frac{N-i}{N}\right), & \text{if } j = i-1, \ \ i \neq 0 \\ \left(\frac{N-i}{N}\right)\frac{i}{N}, & \text{if } j = i+1, \ \ i \neq N \\ 1 - 2\left(\frac{N-i}{N}\right)\frac{i}{N}, & \text{if } j = i, \\ 0, & \text{otherwise} \end{cases} \tag{1.5}$$

Notice that $\frac{i}{N}$ is the probability that red is chosen for elimination, and $\frac{N-i}{N}$ is the probability that blue is chosen for reproduction. Notice also that if $i = 0$, then $P_{00} = 1$ and $P_{0j} = 0$ for all $j \neq 0$. This is because if $X_n = 0$, then there are no reds and there can never be any more reds after this. Similarly, $P_{NN} = 1$ and $P_{Nj} = 0$ if $j \neq N$. The states $X_n = 0$ and $X_n = N$ are called **absorbing states**, since the system cannot leave these states once those states are reached. In population genetics setting, this phenomenon is called **fixation**. The red and blue may represent different alleles in a population. □

Example

To illustrate the meaning of time-homogeneous, consider the following simple modification of the random walk described above: At the n^{th} step (for the transition $X_{n-1} \to X_n$) you toss a coin which is a p-coin with $p = 1/n$. If the coin lands heads, then X moves to the right (increases by one); if the coin lands tails, then X moves to the left (decreases by one). Thus, as n increases, it becomes more likely that the jumps will be to the left, rather than the right. The transition probabilities for this chain are

$$\mathbb{P}(X_{n+1} = y \mid X_n = x) = \begin{cases} \frac{1}{n+1}, & \text{if } \ y = x+1 \\ 1 - \frac{1}{n+1}, & \text{if } \ y = x-1 \\ 0, & \text{otherwise.} \end{cases}$$

In particular, $\mathbb{P}(X_{n+1} = 1 \mid X_n = 0) = 1/(n+1)$. So, the transition probabilities depend on n. This chain is **not** time-homogeneous; it is time-inhomogeneous.

Questions

There are lots of interesting questions that one might ask about Markov chains. Here are a few basic questions:

- If we know the initial state X_0, what will be the distribution of X_n?
- Will X ever reach a given state y?
- If so, how long will it take before X reaches a given state y?
- What is the probability that the system reaches state y before it reaches state z?
- In the long run, what fraction of time will the system spend in state y?

If $X_0 = x$, the probability that $X_1 = y$ is $P(x, y)$. What is the probability that $X_2 = y$? To compute this, we just sum over all possible ways to go from x to y in 2 steps. Using the rules of conditional probability (partition rule), we compute:

$$\begin{aligned}
\mathbb{P}(X_2 = y \mid X_0 = x) &= \sum_{z \in \mathcal{S}} \mathbb{P}(X_2 = y,\ X_1 = z \mid X_0 = x) \\
&= \sum_{z \in \mathcal{S}} \mathbb{P}(X_2 = y \mid X_1 = z\ ,\ X_0 = x)\mathbb{P}(\ X_1 = z\ ,\ X_0 = x) \\
&= \sum_{z \in \mathcal{S}} \mathbb{P}(X_2 = y \mid X_1 = z)\mathbb{P}(\ X_1 = z\ ,\ X_0 = x), \quad \text{(using the Markov property)} \\
&= \sum_{z \in \mathcal{S}} P(x, z)P(z, y) \\
&= P^{(2)}(x, y) \qquad (1.6)
\end{aligned}$$

where $P^{(2)}$ denote the square of the matrix P. Notice that in the last sum, the product $P(x, z)P(z, y)$ is the probability of going from $x \mapsto z$ followed by $z \mapsto y$. Using the Markov property in the same way, one can show that the n-step transition probabilities are

$$\mathbb{P}(X_{k+n} = y \mid X_k = x) = P^{(n)}(x, y)$$

where $P^{(n)}(x, y)$ denotes the (x, y) entry of the n^{th} power of the matrix P (e.g. $P^{(3)} = P \cdot P \cdot P$). In particular, $P^{(n)}(x, y)$ is **not** the same as $(P(x, y))^n$. By a similar computation, the joint distribution of $(X_1, X_2, \ldots, X_n)$ can be expressed in terms of the matrix P. For example,

$$\mathbb{P}(X_2 = x_2, X_2 = x_1 \mid X_0 = x_0) = P(x_0, x_1)\mathbb{P}(x_1, x_2)$$

and

$$\mathbb{P}(X_{j+k} = x_{j+k},\ X_j = x_j \mid X_0 = x_0) = P^{(j)}(x_0, x_j)P^{(k)}(x_j, x_{j+k}).$$

Instead of a fixed initial condition for X_0, the initial state could be random. If a vector ν (indexed by states $x \in \mathcal{S}$) satisfies

$$0 \le \nu(x) \le 1, \quad \forall\ x \in \mathcal{S}, \quad \text{and} \qquad \sum_{x \in \mathcal{S}} \nu(x) = 1,$$

then it defines a probability distribution on $\mathcal{S}$ ($x \mapsto \nu(x)$ is a probability mass function). We say that ν is the **initial distribution** for the chain if $\mathbb{P}(X_0 = x) = \nu(x)$; in this case

we write $X_0 \sim \nu$. If ν is the initial distribution, then the distribution of X_n is given by the vector $\nu P^{(n)}$; this is vector-matrix (left) multiplication where we regard ν as a row vector. That is,

$$\mathbb{P}(X_n = y \mid X_0 \sim \nu) = \sum_{x \in \mathcal{S}} \nu(x) P^{(n)}(x, y).$$

An **invariant distribution** (or **stationary distribution**) π is a probability distribution on $\mathcal{S}$ such that $\pi P = \pi$. Hence $\pi P^{(n)} = \pi$ holds for all n, as well. Thus, an invariant distribution is a left eigenvector of the matrix P with eigenvalue 1. If π is an invariant distribution for the chain, and $X_0 \sim \pi$, then the distribution of X_n does not change with n – it is invariant. This does *not* mean that X_n is contant; rather, it means that the distribution of X_n is not changing.

Notice that although X_n is a random variable, it may not make sense to talk about $\mathbb{E}[X_n]$ since X takes values in $\mathcal{S}$, which may not be a linear space. Adding or averaging elements of $\mathcal{S}$ may have no meaning. On the other hand, it is often useful to consider quantities like $\mathbb{E}[f(X_n)]$ for some function $f : \mathcal{S} \to \mathbb{R}$. Considering the definitions above, observe that

$$\mathbb{E}[f(X_n) \mid X_0 = x] = \sum_{y \in \mathcal{S}} f(y) \mathbb{P}(X_n = y \mid X_0 = x) = \sum_{y \in \mathcal{S}} P^{(n)}(x, y) f(y).$$

If we think of $\{f(y)\}_{y \in S}$ as a column vector, this expression has the form of a matrix-vector multiplication, $(P^{(n)} f)(x)$. Observe that $f \mapsto \mathbb{E}[f(X_n) \mid X_0 = x] = (P^{(n)} f)(x)$ is a linear operation on such functions f.

Example

Consider the simple random walk on a graph $G = (V, E)$, as defined above. Define

$$\pi(x) = \frac{d(x)}{D}, \quad x \in V \tag{1.7}$$

where $d(x)$ is the degree of vertex x (the number of vertices connected to x by an edge, possibly including the vertex x itself), and the constant

$$D = \sum_{x \in V} d(x)$$

is the sum of all the degrees. Then π defines a probability distribution on V. In fact, one can check that this π is an invariant distribution for the simple random walk on the graph! Indeed, recalling P defined by (1.3), we see that for each $y \in V$,

$$(\pi P)_y = \sum_{x \in D} \pi(x) P(x, y) = \sum_{x \sim y} \frac{d(x)}{D} \frac{1}{d(x)} = \frac{1}{D} \sum_{x \sim y} 1 = \frac{d(y)}{D} = \pi(y). \tag{1.8}$$

That is, $\pi P = \pi$. $\square$

Example

Perhaps the simplest non-trivial Markov chain is a chain on two states, which we will label L and R ("left" and "right"), so $\mathcal{S} = \{L, R\}$. Consider a chain on this state space with transition probability matrix

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}. \tag{1.9}$$

That is, $P(L,L) = 1-a$, $P(L,R) = a$, $P(R,L) = b$, $P(R,R) = 1-b$. It is easy to see that the vector

$$\pi = (\pi(L), \pi(R)) = \left(\frac{b}{a+b}\,,\ \frac{a}{a+b} \right)$$

is a stationary probability distribution for this chain, assuming $a + b > 0$. Notice that if $a = b = 0$, then this definition is ill-defined. Nevertheless, if $a = b = 0$, then *any* probability distribution is invariant, since P is just the identity matrix in this case.

I claim that if $|1 - a - b| < 1$, then for any initial distribution ν_0, we have $\nu P^{(n)} \to \pi$ as $n \to \infty$. In particular, the stationary distribution is unique. Given an initial distribution ν_0, let $\nu_n = \nu_0 P^{(n)}$ denote the distribution at time n. Then $\nu_{n+1} = \nu_n P$, so

$$\begin{aligned} \nu_{n+1}(L) &= \nu_n(L)P(L,L) + \nu_n(R)P(R,L) \\ &= \nu_n(L)(1-a) + (1-\nu_n(L))b \\ &= \nu_n(L)(1-a-b) + b. \end{aligned}$$

Therefore,

$$\begin{aligned} \nu_{n+1}(L) - \frac{b}{a+b} &= \nu_n(L)(1-a-b) + b - \frac{b}{a+b} \\ &= \left(\nu_n(L) - \frac{b}{a+b} \right)(1-a-b). \end{aligned} \tag{1.10}$$

In particular,

$$|\nu_{n+1}(L) - \pi(L)| \le |1-a-b||\nu_n(L) - \pi(L)|.$$

If $r = |1-a-b| < 1$, we can iterate this bound to obtain

$$|\nu_n(L) - \pi(L)| \le r^n |\nu_0(L) - \pi(L)|, \quad n \ge 1.$$

The same bound holds for $|\nu_n(R) - \pi(R)|$. Notice that $r^n \to 0$ as $n \to \infty$, since $0 < r < 1$. □

We say that a state $x \in \mathcal{S}$ is **recurrent** if

$$\mathbb{P}(X_n = x \text{ for some } n \ge 1 \mid X_0 = x) = 1.$$

This means that if the initial state is x, the chain is sure to return to x at some later time. If a state is not recurrent (i.e. this probability is less than 1), then we say the state is

transient. So, if x is transient, there is some positive probability that the chain will never return to x. In the urn model example defined by (1.4), every state is recurrent and there are no transient states. In the Moran model example defined by (1.5), the states $X = 0$ and $X = N$ are recurrent; every other state is transient.

We say that two states $x, y \in \mathcal{S}$ **communicate** if there are positive integers n and m such that

$$P^{(n)}(x, y) > 0 \quad \text{and} \quad P^{(m)}(x, y) > 0.$$

This means that it is possible (positive probability) that the chain can go from x to y in some number of steps and from y to x in some number of steps. We also write $x \leftrightarrow y$ to say that x and y communicate (or intercommunicate). If all pairs $x, y \in \mathcal{S}$ communicate, then we say that the chain is **irreducible**. Otherwise, the chain is **reducible**. This notion of communication ($\leftrightarrow$) between two states determines equivalence classes: if $x \leftrightarrow y$ and $y \leftrightarrow z$, then $x \leftrightarrow z$. For any chain, the state space can be partitioned uniquely according to

$$\mathcal{S} = T \cup C_1 \cup C_2 \cup \dots$$

where T is the set of all **transient states**, and the sets C_k, for $k = 1, 2, \dots$ are **closed communication classes** of recurrent states. Saying that C_k is a closed communication class means that

(i) For all $x, y \in C_k$, we have $x \leftrightarrow y$, and

(ii) $P(x, z) = 0$ whenever $x \in C_k$ but $z \notin C_k$.

Thus, for all $x, y \notin T$, x and y intercommunicate ($x \leftrightarrow y$) if and only if x and y are in the same class C_k. Moreover, once the chain reaches one of the sets C_k, it cannot leave C_k. Strange as it may seem, it is possible that all states are transient (i.e. $T = \mathcal{S}$ and there are no recurrent communications classes C_k's), but this can only happen in cases where $|\mathcal{S}| = \infty$.

Lemma 1.1 *If $|\mathcal{S}| < \infty$, then there is at least one recurrent state.*

Lemma 1.2 *If the chain is irreducible, then either (i) all states are recurrent, or (ii) all states are transient.*

For any state $x \in \mathcal{S}$, we define the **period** of x to be

$$d(x) = \gcd\{n \geq 1 \mid P^{(n)}(x, x) > 0\}$$

If two states x and y communicate, then they must have the same period: $d(x) = d(y)$; if the chain is irreducible, then all states must have the same period. Therefore, if the chain is irreducible, we can define the period of the chain to be $d(x)$ (which is the same for all $x \in \mathcal{S}$). If an irreducible chain has period 1, we say the chain is **aperiodic**; otherwise the chain is **periodic** with period $d > 1$.

Theorem 1.1 *Suppose $|\mathcal{S}| < \infty$. If the chain is irreducible, then there is a unique invariant probability distribution π. If the chain is also aperiodic, then for any initial probability distribution ν,*

$$\lim_{n \to \infty} \nu P^{(n)} = \pi. \tag{1.11}$$

Hence

$$\lim_{n\to\infty} P^{(n)}(x,y) = \pi(y)$$

for all $x, y \in \mathcal{S}$. Furthermore, for any function $F : \mathcal{S} \to \mathbb{R}$, the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} F(X_n) = \sum_{x\in\mathcal{S}} F(x)\pi(x) = \mathbb{E}_\pi[F(x)] \tag{1.12}$$

holds with probability one. In particular, the limit does not depend on the initial distribution.

If P is a transition probability matrix such that for some fixed $n > 0$, the condition $P^{(n)}(x,y) > 0$ holds for all $x, y \in \mathcal{S}$, then the chain must be both irreducible and aperiodic.

Corollary 1.1 *If $|\mathcal{S}| < \infty$ and for some fixed $n > 0$, the condition $P^{(n)}(x,y) > 0$ holds for all $x, y \in \mathcal{S}$, then the conclusions of Theorem 1.1 must hold.*

Theorem 1.1 implies that a stationary distribution $\pi(x)$ may be interpreted as the long-run fraction of time that the chain spends in state x. To see this, suppose that $|\mathcal{S}| < \infty$ and the chain is irreducible and aperiodic. Then, if for some fixed $x \in \mathcal{S}$ we take $F : \mathcal{S} \to \mathbb{R}$ to be $F(y) = 1$ if $y = x$, $F(y) = 0$ if $y \neq x$. That is, F is the indicator function of state x, $F(y) = \mathbb{I}(y = x)$. Then (1.12) implies that

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \#\{n \in \{1,\ldots,N\} \mid X_n = x\} &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1} \mathbb{I}(X_n = x) \\ &= \sum_{y\in\mathcal{S}} F(y)\pi(y) = \pi(x). \end{aligned} \tag{1.13}$$

Example

Recall the distribution π defined at (1.7), which is an invariant distribution for the simple random walk on a graph. If the graph is connected, then the chain is irreducible (why?). Consequently, this π is the unique invariant distribution (by Theorem 1.1). In view of the preceding comments, $\pi(x)$ is the long-run average time that the chain spends at vertex x. Thus, the long-run average time is proportional to the degree $\deg(x)$: on average, the chain will spend more time at vertices with lots of neighbors. □

For each $x \in \mathcal{S}$, define the **first visit** to x by

$$T_x = \min\{n \geq 1 \mid X_n = x\}$$

(note: $n \geq 1$ here excludes the initial time $n = 0$.) This T_x is an integer-valued random variable. We say $T_x = +\infty$ if X_n never reaches x. Then we define the **mean return time** to x by

$$\mu_x = \mathbb{E}[T_x \mid X_0 = x].$$

If x is transient, then $\mu_x = +\infty$, since there is positive probability that $T_x = +\infty$. It is possible that $\mu_x = +\infty$ even if x is recurrent; that is, even if X_n is sure to return to x,

the return time might have infinite expectation. In this case where x is recurrent while $\mu_x = +\infty$, we say that x is **null recurrent**. If x is recurrent and $\mu_x < \infty$, we say that x is **positive recurrent**.

Theorem 1.2 *An irreducible chain has a stationary probability distribution π if and only if all states are positive recurrent. If the chain is irreducible and all states are positive recurrent, then $\pi(x) = (\mu_x)^{-1}$ for all $x \in \mathcal{S}$; in particular, π is unique.*

If the state space is finite ($|\mathcal{S}| < \infty$) and the chain is irreducible, then all states are positive recurrent and there is a unique stationary distribution, whether or not the chain is aperiodic (but the convergence described in (1.11) may not hold if the chain is periodic). Null-recurrence for an irreducible chain can happen only if the state space is infinite. For example, for the simple random walk on the integers $\mathbb{Z}$, one can show that all states are null-recurrent.

1.1 Some calculations with Markov chains

Exit probabilities

Suppose a chain is finite and irreducible. Let $a, b \in \mathcal{S}$ be given states. Let $h(x)$ be the probability of hitting b before hitting a, starting from x:

$$h(x) = \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_0 = x). \tag{1.14}$$

Clearly $h(b) = 1$ and $h(a) = 0$. By conditioning on the first jump out of x, we also have

$$\begin{aligned} h(x) &= \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_0 = x) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y,\, X_0 = x)\mathbb{P}(X_1 = y \mid X_0 = x) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y,\, X_0 = x)P(x,y) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y)P(x,y) \\ &= \sum_y h(y)P(x,y) \end{aligned} \tag{1.15}$$

The sum is over all $y \in \mathcal{S}$ for which $P(x,y) \neq 0$. This gives us a linear system of equations to solve for h:

$$\begin{aligned} h(x) &= \sum_y P(x,y)h(y), \quad \forall\, x \in \mathcal{S} \setminus \{a,b\}, \\ h(b) &= 1, \\ h(a) &= 0. \end{aligned} \tag{1.16}$$

Exit prize

Here is a useful generalization of the above computation. Let $B \subset \mathcal{S}$ be some subset of the state space, and let $g : B \to \mathbb{R}$ be some function. Consider the function

$$h(x) = \mathbb{E}\left[g(X_\tau) \mid X_0 = x\right] \tag{1.17}$$

where $\tau = \min\{n \geq 0 \mid X_n \in B\}$ is the first time that the chain reaches some state in the set B (this time is random). Think of $g(y)$ as a "prize" that is awarded if the chain first reaches B through state y. Then $h(x)$ defined by (1.17) is the expected prize, given that $X_0 =$. If $x \in B$, then $\tau = 0$, so $h(x) = g(x)$ in this case. However, if $x \notin B$, then it is not so simple to compute $h(x)$. By the same argument as above (contioning on the first jump out of x), one can show that h satisfies the linear system of equations:

$$\begin{aligned} h(x) &= \sum_y P(x,y)h(y), \quad \forall\, x \in \mathcal{S} \setminus B, \\ h(x) &= g(x), \quad x \in B. \end{aligned} \tag{1.18}$$

Notice that (1.14) is a special case of (1.17), with $B = \{a, b\}$ and g defined by $g(a) = 0$, $g(b) = 1$; thus, prize of 1 is awarded if the chain hits b before a, while the prize of 0 is awarded if the chain hits a before b.

Example

Here we compute (1.14) in the context of a specific chain. Consider a birth-death chain on integers $\{0, 1, 2, \ldots, L\}$ such that at each jump time, the process either moves left one step, right one step, or stays put. That is,

$$P(k,k) = 1 - P(k,k+1) - P(k,k-1), \quad k = 1, 2, \ldots, L-1.$$

Suppose $X_0 = x$. What is the probability that the system reaches state L before state 0? This is a case of (1.14) with $b = L$ and $a = 0$. In this case, the system (1.18) takes the form

$$h(x) = \alpha_x h(x-1) + (1 - \alpha_x - \beta_x)h(x) + \beta_x h(x+1) \tag{1.19}$$

where $\alpha_x = P(x, x-1)$ and $\beta_x = P(x, x+1)$. Also, $h(L) = 1$, $h(0) = 0$. We can solve this system explicitly. Rearranging (1.19), we obtain the relation

$$h(x+1) - h(x) = \frac{\alpha_x}{\beta_x}\left(h(x) - h(x-1)\right)$$

Let $d(x) = h(x) - h(x-1)$. Thus, $d(x+1) = \frac{\alpha_x}{\beta_x} d(x)$ for all $x \in \{1, 2, \ldots, L-1\}$. This implies that

$$d(x) = d(1) \prod_{j=1}^{x-1} \frac{\alpha_j}{\beta_j}, \quad x = 2, \ldots, L.$$

(Note: for $x = 1$, we define $\prod_{j=1}^{0} \frac{\alpha_j}{\beta_j} = 1$.) We also know that h can be written as a sum of these differences:

$$h(x) = h(x) - h(0) = \sum_{k=1}^{x} d(k) = d(1) \sum_{k=1}^{x} \prod_{j=1}^{k-1} \frac{\alpha_j}{\beta_j}$$

In particular, since $h(L) = 1$, this formula implies that

$$d(1) = \left(\sum_{k=1}^{L} \prod_{j=1}^{k-1} \frac{\alpha_j}{\beta_j} \right)^{-1}.$$

Therefore,

$$h(x) = \frac{\sum_{k=1}^{x} \prod_{j=1}^{k-1} \frac{\alpha_j}{\beta_j}}{\sum_{k=1}^{L} \prod_{j=1}^{k-1} \frac{\alpha_j}{\beta_j}}$$

□

Occupation times, absorbing states

Suppose that a chain on a finite $\mathcal{S}$ is irreducible. Let $B \subset \mathcal{S}$ be some subset of states. Let $A = \mathcal{S} \setminus B$ be the other states. For $x \in A$, we ask: how many steps will the chain take before reaching a state in the set B? Define

$$\tau_B = \min\{n \geq 0 \mid X_n \in B\}$$

This is the first time that X is in B; it is an integer-valued random variable. We want to compute

$$h(x) = \mathbb{E}[\tau_B \mid X_0 = x].$$

Clearly $h(y) = 0$ for all $y \in B$. For $x \in A$, it takes at least one step to reach B. So, clearly $h(x) \geq 1$ for $x \in A$. To compute h, one must solve a linear system of equations. To derive this system, it is useful to condition on the first step from x. This leads to the system

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x, y)\mathbb{E}[\tau_B \mid X_1 = y], \quad \forall \ x \in A = \mathcal{S} \setminus B.$$

Since the chain is time-homogeneous, this means that

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x, y)h(y), \quad \forall \ x \in A = \mathcal{S} \setminus B.$$

Since $h(y) = 0$ for all $y \in B$, this is equivelent to

$$h(x) = 1 + \sum_{y \in A} P(x, y)h(y), \quad \forall \ x \in A = \mathcal{S} \setminus B,$$

where we are summing only over A. Let M denote the $|A| \times |A|$ submatrix of P obtained by keeping only the entries $P(x, y)$ with $x, y \in A$. So, the system can be written as,

$$h(x) = 1 + \sum_{y \in A} M(x, y)h(y), \quad \forall \ x \in A = \mathcal{S} \setminus B,$$

For small systems or when M has nice structure, it may be feasible to solve this easily by hand. For larger systems, you could use a computer. Using matrix-vector notation, this system is equivalent to

$$(I - M)h = \vec{1}$$

where $\vec{1} = (1, 1, 1, \ldots, 1)^T$ is a column vector of all $1's$. The solution vector is:

$$h = (I - M)^{-1}\vec{1}.$$

So, for a particular $x \in A$,

$$h(x) = \sum_{y \in A} (I - M)^{-1}(x, y). \tag{1.20}$$

Alternative perspective: Consider modifying the chain, by replacing P with the matrix $\tilde{P}$ defined by

$$\tilde{P}(x, y) = \begin{cases} P(x, y), & \text{if } x \in A, \ y \in \mathcal{S} \\ 1, & \text{if } x = y \in B \\ 0, & \text{otherwise} \end{cases}$$

This modification means that: (i) all transitions from a state $x \in A$ to any other state are preserved, and (ii) the only transitions from a state $x \in B$ are self-loops. In particular, all transitions from states $x \in B$ to states $y \in A$ are removed. Therefore, under this modified transition matrix, the states in B become **absorbing states** – once $\tilde{X}_n$ reaches a state $z \in B$ it never leaves that state.

Using the tail sum formula, we have

$$\mathbb{E}[\tau_B \mid X_0 = x] = \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k \mid X_0 = x). \tag{1.21}$$

Notice that before hitting a state in B, the original chain (with transition matrix P) and the modified chain (with transition matrix $\tilde{P}$) have the same transition rules. Therefore,

$$\mathbb{P}(\tau_B > k \mid X_0 = x) = \mathbb{P}(\tilde{X}_k \in A \mid X_0 = x) = \sum_{y \in A} \tilde{P}^{(k)}(x, y)$$

For $x, y \in A$, $P^{(k)}(x, y) = M^{(k)}(x, y)$, where M is the $|A| \times |A|$ sub-matrix described above. Therefore, putting this all together we have:

$$\begin{aligned} \mathbb{E}[\tau_B \mid X_0 = x] &= \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k \mid X_0 = x) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} \tilde{P}^{(k)}(x, y) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} M^{(k)}(x, y) \\ &= \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right)(x, y) \end{aligned} \tag{1.22}$$

Recall that if $|m| < 1$, then $\sum_{n=0}^{\infty} m^n = \frac{1}{1-m}$. A version of this statement is also true with square matrices: if all the eigenvalues of a $d \times d$ matrix M have modulus strictly less than one, then $I - M$ is invertible and

$$\sum_{k=0}^{\infty} M^{(k)} = (I - M)^{-1},$$

where I is the $d \times d$ identity matrix. If M is the $|A| \times |A|$ submatrix described above, one can show that M has this property and that $I - M$ is invertible. Hence,

$$\mathbb{E}[\tau_B \mid X_0 = x] = \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right) (x, y) = \sum_{y \in A} (I - M)^{-1}(x, y)$$

This last expression refers to the (x, y) entry of the matrix $(I - M)^{-1}$. Notice that this formula agrees with (1.20).

2 Markov Chain Monte Carlo Algorithms

There are many practical problems in which one needs to sample from a given probability distribution

$$\pi(x) = \frac{f(x)}{c}, \quad x \in \mathcal{S} \tag{2.23}$$

on a discrete space $\mathcal{S}$, where $c = \sum_{x \in \mathcal{S}} f(x) > 0$ is a normalizing constant. It is often the case that the constant c is unknown, and the state space S is so large that computing c directly is prohibitively expensive. (For example, this is often the case in the context of sampling from a Bayesian posterior distribution, or when sampling from a Gibbs measure in statistical physics.) Therefore, it is desirable to have an algorithm that generates samples from π *without knowing* c. One very useful approach to doing this to construct a Markov chain which has π as its stationary distribution; remarkably, this can be done without knowing c. This general idea (this type of algorithm) is known as **Markov Chain Monte Carlo** (MCMC).

The following MCMC-type algorithm is known as the **Metropolis-Hastings algorithm**. The reason this algorithm is so useful is that it does not require knowledge of the normalizing constant c in (2.23), since the algorithm only requires evaluations of ratios $\pi(x)/\pi(y) = f(x)/f(y)$. Here is the algorithm:

Given $X_n = x$, generate a new state X_{n+1} according by the following two-stage process:

- **Stage 1:** Propose a new state $y \in \mathcal{S}$ with probability $q(x, y)$. This q is some known transition probability matrix on S; q is the only ingredient needed for the algorithm.
- **Stage 2:** Decide whether to accept or reject the proposal. With probability

$$\min\left(1\ ,\ \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)$$

we accept the proposal and set $X_{n+1} = y$. Otherwise, the proposal is rejected and the new state is $X_{n+1} = x$ (the state does not change).

Proposition 2.1 *For the chain defined by this algorithm, the distribution π is stationary.*

Before proving this, let us note that the transition probability q is used to generate proposal steps in the chain, but it is *not* the transition probability for the chain X_n, because there is an additional random aspect to the transitions of X_n in Step 2. There is a lot of flexibility in choosing q, although the performance of the algorithm (how fast the distribution of X_n converges to the stationary distribution) will depend on this choice. The distribution π need not be stationary for q (typically it is not). To illustrate the idea, suppose $|S| < \infty$, and we choose $q(x, y) = 1/|S|$. This means that the proposal is chosen uniformly at random from all states in S. The stationary distribution for a chain with transition matrix q would be the uniform distribution on S. However, this algorithm (with that choice of q) will work for any probability distribution π on S.

Proof of Proposition 2.1: Let us define

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

First, observe that if $x \neq y$, the transition probability for the chain defined by the algorithm is

$$P(x, y) = q(x, y) \min\left(1\ ,\ \alpha(x, y)\right). \tag{2.24}$$

This is just the probability of proposing y times the probability of accepting y once it is proposed. Next, we claim that

$$\pi(x)P(x, y) = \pi(y)P(y, x), \quad \forall\ x, y \in \mathcal{S}. \tag{2.25}$$

Assuming $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \leq 1$, we see that

$$\pi(x)P(x, y) = \pi(x)q(x, y)\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \pi(y)q(y, x).$$

In this case, we also have $\alpha(y, x) = 1/\alpha(x, y) \geq 1$. So,

$$\pi(y)P(y, x) = \pi(y)q(y, x).$$

This shows that (2.25) holds. Now, sum (2.25) over x:

$$\sum_x \pi(x)P(x, y) = \sum_x \pi(y)P(y, x) = \pi(y)\sum_x P(y, x) = \pi(y).$$

Hence, π is stationary for this chain. □

The condition (2.25) is called **detailed balance.**

Here is another algorithm known as **Gibb's sampling**. This algorithm can be used in cases where the state space $\mathcal{S}$ has a particular structure. Let $A = \{a_1, \ldots, a_k\}$ be some finite set of k distinct elememts. Let $M > 1$ be an integer. Let $\mathcal{S} = A^M$. We think of a system which consists of M distinct components, and each component can be in one of k states: $a_1, \ldots, a_k$ are the different "states" or "labels" or "types" that a component can take. An element $x \in \mathcal{S}$ is a "configuration" of the system. Alternatively, you might think of $\mathcal{S}$ as the set of all distinct "words" of length M that can be made from letters in alphabet A.

The following alorithm generates a Markov chain on $\mathcal{S}$ with stationary distribution

$$\pi(x) = \frac{f(x_1, \ldots, x_M)}{c}, \quad x = (x_1, \ldots, x_M) \in \mathcal{S},$$

where $c > 0$ is a normalizing constant. However, the algorithm does not require knowledge of c. Observe that the total number of possible configurations of the system is $|\mathcal{S}| = k^M$, so computing c may be prohibitively expensive when M is large. The current state of the chain is denoted

$$X_n = (X_n^1, \ldots, X_n^M).$$

We think of X_n as the current "configuration" of the system, and X_n^j is the "state"' of the j^{th} component. Here is the algorithm:

Given $X_n = (x^1, \ldots, x^M)$, generate the next state X_{n+1} as follows:

Step 1: Pick a component index $i \in \{1, \ldots, M\}$ uniformly at random.

Step 2: Choose a new state $Y^i \in A$ for the i^{th} component randomly, according to distribution

$$\mathbb{P}(Y^i = a) = \frac{f(x^1, \ldots, x^{i-1}, a, x^{i+1}, \ldots, x^M)}{\sum_{j=1}^k f(x^1, \ldots, x^{i-1}, a_j, x^{i+1}, \ldots, x^M)}, \quad a \in \{a_1, \ldots, a_k\}. \tag{2.26}$$

Then set $X_{n+1} = (X_n^1, \ldots, X_n^{i-1}, Y^i, X_n^{i+1}, \ldots, X_n^M)$.

□

At each step, only one component of X_n is updated. Observe that (2.26), the distribution from which the component is updated, is equivalent to

$$\mathbb{P}(Y^i = a) = \frac{\pi(X_n^1, \ldots, X_n^{i-1}, a, X_n^{i+1}, \ldots, X_n^M)}{\sum_{j=1}^k \pi(X_n^1, \ldots, X_n^{i-1}, a_j, X_n^{i+1}, \ldots, X_n^M)}$$

which is the marginal distribution of the i^{th} component, given the values of the other components. Nevertheless, this does not require knowledge of the normalizing constant c, which cancels in the ratio (2.26).

Proposition 2.2 *For the chain defined by this algorithm, the distribution π is stationary.*

Proof: Verify that the detailed balance condition (2.25) holds.... It is important to note that $P(x, y) \neq 0$ if and only if x and y differ only in one coordinate. That is, $P(x, y) \neq 0$ if and only if there is $i \in \{1, 2, \ldots, M\}$ such that $x_k = y_k$ for all $k \neq i$.

□

3 Continuous time Markov processes

In a continuous time Markov chain X_t, the time parameter is continous, $t \geq 0$. As before, the system jumps randomly between states in $\mathcal{S}$, but now the jumps may occur at any time and they occur randomly. There are now two sources of randomness: *where* the system jumps and *when* the system jumps. That is, the time that passes between each jump is random.

One can formulate a Markov property in the continuous time case. The continuous time version of the Markov property says that for $s, t \geq 0$ and $y \in \mathcal{S}$,

$$\mathbb{P}(X_{t+s} = y \mid X_t) = \mathbb{P}(X_{t+s} = y \mid X_r,\ 0 \leq r \leq t)$$

In words: the conditional distribution of X_{t+s} given the history up to time t is the same as the conditional distribution of X_{t+s} given only X_t. Thus, if we know the current state X_t, then knowing more information about the past doesn't help us better predict the future state X_{t+s}.

3.0.1 Exponential random variables

It turns out that in order for the Markov property to hold, the times between jumps *must be exponentially distributed random variables*. Recall that a random variable T has the $\text{Exp}(\lambda)$ distribution if

$$\mathbb{P}(T > t) = e^{-\lambda t}.$$

The parameter $\lambda > 0$ is positive; this is sometimes called the **rate**. The mean of T is $\mathbb{E}[T] = \lambda^{-1}$ and the variance is $\text{Var}(T) = \lambda^{-2}$. This distribution has the density $\rho(t) = \lambda e^{-\lambda t}$ on $(0, \infty)$. Therefore,

$$\mathbb{E}[f(T)] = \int_0^\infty f(t)\lambda e^{-\lambda t}\, dt.$$

Exponential random variables have the important property that

$$\mathbb{P}(T > s + t \mid T > t) = \mathbb{P}(T > s), \quad \forall\ s, t \geq 0. \tag{3.27}$$

This is sometimes called the **memoryless property** of exponential random variables, and it is this property that makes exponential random variables so important for continuous-time Markov processes. The left side can be written as $\mathbb{P}(T - t > s \mid T > t)$. Think of T as a time we have to wait until a random alarm clock rings. The event $T > t$, means the alarm has not rung yet at time t, and $T - t$ is the additional time we must wait beyond time t for the clock to ring. So, this property says that, given the alarm has not rung by time t, the distribution of the additional time we must wait for the alarm has the same distribution as the original T. So, having waited a long time doesn't make it more likely that the alarm will

ring soon in the future. This property actually characterizes the exponential distribution in the following sense:

Theorem 3.1 *Let T be a non-negative, continuously distributed random variable. Then $T \sim Exp(\lambda)$ for some $\lambda > 0$ if and only if T satisfies property* (3.27)*.*

Another very important property of exponentials is that if $T_1, T_2, \ldots, T_n$ are independent exponential random variables with rates $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, then the random variable

$$T = \min(T_1, T_2, \ldots, T_n)$$

is exponential with rate $\lambda_1 + \lambda_2 + \cdots + \lambda_n$. Imagine that we have n alarm clocks, labeled $1, 2, \ldots, n$, which ring independently at random times. The clocks are all set simultaneously. Suppose that clock k will ring after T_k units of time have expired, where T_k is a random variable distributed as $\text{Exp}(\lambda_k)$. Then $T = \min(T_1, T_2, \ldots, T_n)$ is the time at which the first ring occurs – this is exponential with rate $\lambda_1 + \lambda_2 + \cdots + \lambda_n$. Moreover, one can show that

$$\mathbb{P}(T_k = \min(T_1, T_2, \ldots, T_n)) = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_n}$$

is the probability that clock k is the first alarm to ring.

Example

One of the simplest and most important continuous time Markov chains is the **Poisson arrival process**. The process has a single parameter, $\lambda > 0$, in its definition. The process is integer-valued. At each jump time, the process increases by 1. The time *between* jumps are independent, distributed as $\text{Exp}(\lambda)$.

Here's another way to express the same idea: Suppose $T_1, T_2, T_3, \ldots$ is a collection of independent $\text{Exp}(\lambda)$ random variables. Then define

$$X_t = \max\left\{n \geq 0 \mid T_1 + T_2 + \cdots + T_n \leq t\right\}.$$

Thus, X_t is the number of jumps that have occured up to time t. T_k is the time between jump $k-1$ and jump k. The sum $T_1 + T_2 + \cdots + T_n$ is the time at which the n^{th} jump occurs.

An alternative point of view is the following: suppose $X_t = n$. Set an alarm clock which rings a random time $T \sim \text{Exp}(\lambda)$. When the clock rings, X increases by one. Then the clock is reset and the process continues in this way.

The parameter λ is called the **rate** of the process. Notice that when λ is large, the arrivals occur more frequently than when λ is small, because the expected time between arrivals is $1/\lambda$. One can show that X_t, which is integer-valued, has the $\text{Poisson}(\lambda t)$ distribution. That is,

$$\mathbb{P}(X_t = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \ldots$$

In particular, $\mathbb{E}[X_t] = \lambda t$ and $\text{Var}(X_t) = \lambda t$. $\square$

Example: Jump rates

Suppose $\mathcal{S}$ is a finite set, whose elements we label $\{1, \ldots, n\}$, $n = |\mathcal{S}|$. For $j, k \in \mathcal{S}$, let $\alpha_{jk} \geq 0$. We now describe the continuous time Markov chain on $\mathcal{S}$ with jump rates α_{jk}.

There are two equivalent ways of contructing such a process. First, you may imagine that when X lands at state j, there are many exponential alarm clocks (one clock for each possible target location) which are all reset simultaneously, with rates α_{jk}. Thus, if T_{jk} denotes the time at which the (j, k) clock rings, then $T_{jk} \sim \text{Exp}(\alpha_{jk})$ and these are independent random variables. When the first of those clocks rings, the system makes a jump; if the (j, ℓ) clock is first to ring, then the system jumps to state ℓ.

Notice that the time of the jump is

$$T_j = \min\{T_{jk} \mid j \neq k\}$$

which is also exponentially distributed, $\text{Exp}(\bar{\alpha}_j)$, with rate

$$\bar{\alpha}_j = \sum_{k \neq j} \alpha_{jk}.$$

Moreover, the probability that the (j, k) clock will be the first to ring is $\alpha_{jk}/(\bar{\alpha}_j)$.

A second, equivalent way to think about this process is as follows: when X reaches state j, a single alarm clock is set. This clock rings at a random time which is distributed as $\text{Exp}(\bar{\alpha}_j)$. When the clock rings, the process jumps. To determine the location of the jump, dice are rolled, independent of the past, and X moves to site k with probability $\alpha_{jk}/(\bar{\alpha}_j)$.

3.0.2 Example: Birth-death process

A continuous-time **birth-death process** is a continuous-time Markov chain on the non-negative integers such that all jumps are of size 1. Thus, the rates satisfy

$$\alpha_{jk} = 0 \quad \text{if } \; k \neq j \pm 1.$$

The rate

$$\alpha_{j,j+1}$$

is the birth-rate when $X = j$, the rate of jumps from j to $j + 1$; while $\alpha_{j,j-1}$ is the death rate when $X = j$, the rate of jumps from j to $j - 1$.

The Poisson arrival process introduced earlier with rate λ is an example of a birth death process where are the birth rates are constant: $\alpha_{j,j+1} = \lambda$ for all $j \geq 0$, and there is no death, meaning that $\alpha_{j,j-1} = 0$ for all j.

3.0.3 Example: non-homogeneous birth process

Imagine the growth of a collection of cells which divide at random times, independently of each other. Suppose that there is no limit to the number of times a cell can divide, and that when a cell divides it does so instantly. We might model this by a birth process, as follows. Suppose that each cell divides at rate λ, so that the time before a cell's next division is $\text{Exp}(\lambda)$ random variable. If there are currently n cells, then the time T until the next

division is the minimum of n independent $\text{Exp}(\lambda)$ random variables. Thus, $T \sim \text{Exp}(\lambda n)$. If X_t is the number of cells at time t, then the transition rates for X are:

$$\alpha_{j,j+1} = \lambda j, \quad j \geq 1.$$

All other rates are zero (there are no deaths). Suppose the size of the system is large: $X_0 = N \gg 1$. One can show that as $N \to \infty$,

$$\frac{X_t}{X_0} = \frac{X_t}{N} \to Y(t)$$

where $Y'(t) = \lambda Y(t)$, $Y(0) = 1$. That is, $Y(t) = e^{\lambda t}$. Thus, the relative population size (relative to X_0) is approximated well by a solution to a differential equation. For finite N, however, there are stochastic effects.

3.0.4 Example: simple stochastic SIR model

Imagine a population of N individuals, who are classified as either susceptible, infected, or recovered. There are no births or deaths. Let $X_t = (S_t, I_t, R_t)$ be the state of the population at time t, which we model by a continuous time Markov chain. The state space is

$$\mathcal{S} = \{(S, I, R) \in \mathbb{Z}^3 \mid S + I + R = N, \quad S \geq 0,\ I \geq 0,\ R \geq 0\}$$

Given S, I, R at time t, the possible transitions are

$$\begin{aligned} (S, I, R) &\mapsto (S - 1, I + 1, R) \quad \text{at rate } \beta \frac{S}{N} I \\ (S, I, R) &\mapsto (S, I - 1, R + 1) \quad \text{at rate } \lambda I \end{aligned}$$

No other transitions are possible, in this simple model.

Here is an equivalent "alarm clock" description of this process: each infected person has two alarm clocks – an infection clock and a recovery clock. The clocks ring after a random delay, which is exponentially distributed. All the clocks are independent. The delay for each infection clock is exponentially distributed with rate $\beta > 0$. When an infection clock rings, that infected person chooses someone from the population at random with whom to make contact. If that person is susceptible, then the susceptible person becomes infected. Since there are fraction S/N of susceptible people, the rate at which a single infected person gets someone sick is $\beta S/N$. So, considering all infected people, the total rate at which new infections occur is $\beta(S/N)I$. The delay for each recovery clock is distributed as $\text{Exp}(\lambda)$. When one of those clocks rings, the infected person becomes recovered. Since their are I infected, there are I recovery clocks, and the total rate of recovery is λI.

Each time one of the clocks rings (either infection or recover), the state of the system changes, and all the clocks are reset according to the new parameters. Notice that if the current state is (S, I, R), then the time until the next clock rings (either an infection clock or a recovery clock) is distributed as $\text{Exp}(\beta(S/N)I + \lambda I)$. The probability that the next event is an infection event is

$$p_{inf} = \frac{\beta(S/N)I}{\beta(S/N)I + \lambda I}$$

while the probability that the next event is a recovery event is

$$p_{rec} = \frac{\lambda I}{\beta(S/N)I + \lambda I}$$

So, because $(S/N) \in [0,1]$, if $\mathcal{R}_0 = \beta/\lambda < 1$, then $p_{rec} > p_{inf}$. However, if $\mathcal{R}_0 = \beta/\lambda > 1$, then in the early stage of an epidemic where $(S/N) \approx 1$, we will have $p_{inf} > p_{rec}$.

3.0.5 Transition probabilities

Recall that for a discrete time chain with initial distribution ν, the distribution of X_n is $\nu P^{(n)}$ where P is the n^{th} power of the transition matrix. For a continuous time chain with jumps rates α_{jk} (as defined in the previous example), the probability distribution of X_t satisfies a system of ordinary differential equations (ODE). Specifically, suppose that $X_0 = x \in \mathcal{S}$. Then define

$$p_k(t) = \mathbb{P}(X_t = k \mid X_0 = x), \quad k \in \mathcal{S}$$

One can show that

$$\frac{dp_k}{dt} = -p_k\bar{\alpha}_k + \sum_{j\neq k} p_j\alpha_{jk}. \tag{3.28}$$

The first term, $-p_k\bar{\alpha}_k \leq 0$ involves the rate at which the system is jumping out of state k; the second term $\sum_{j\neq k} p_j\alpha_{jk}$ describes an increase in p_k due to jumps into k from other states. The system (3.28) is a system of linear, homogeneous differential equations. It could be written in the form

$$\frac{dp_k}{dt} = \sum_{j\in\mathcal{S}} p_j Q_{jk}, \qquad k \in \mathcal{S}. \tag{3.29}$$

where Q is the matrix

$$Q_{jk} = \alpha_{jk}, \quad \text{for } j \neq k, \qquad Q_{kk} = -\bar{\alpha}_k.$$

This matrix is sometimes called the rate matrix. Or, thinking of $\mathbf{p}(t) = (p_j(t))_{j\in\mathcal{S}}$ as a row vector, we might write (3.29) as

$$\frac{d\mathbf{p}}{dt} = \mathbf{p}Q \tag{3.30}$$

The initial condition for $\mathbf{p}$ is:

$$p_k(0) = 1, \quad \text{if } k = x, \qquad p_k(t) = 0, \quad \text{if } k \neq x.$$

Notice that the rate matrix Q has the property that all rows sum to zero: $\sum_j Q_{jk} = 0$, since $\bar{\alpha}_k = \sum_{j\neq k} \alpha_{jk}$.

Remark 3.1 *I'm writing this as $\mathbf{p}Q$ rather than $Q\mathbf{p}$, because I have used the convention that Q_{jk} is the rate from jumps from $j \to k$ (so that reading the indices left-to-right is consistent with the source-to-target idea). This is common practice in certain communities. The price to pay for this convention is that we have to think of $\mathbf{p}$ as a row vector and deal with "left multiplication" $\mathbf{p}Q$. Of course, we could have defined α_{jk} to be the rate of jumps from k to j, to avoid this notational issue.*

4 Branching processes

A **branching process** is a type of Markov chain modeling a population in which each individual produces a random number of children (possibly 0) and dies. The state space is $\mathcal{S} = \{0, 1, 2, \dots\}$, the non-negative integers. There is a discrete time version and a continuous time version of the chain. In the discrete case, the state is Z_n, the size of the population at time $n = 0, 1, 2, \dots$. In the continuous time case, the state is Z_t, for $t \geq 0$.

4.1 Discrete time case

In the discrete case, all of the Z_n individuals in the current generation branch at the same time (and immediately die). "Branching" means that these Z_n individuals gives birth to a random number of offspring. The births are independent and distributed according to the **offspring distribution** $\{p_k\}_{k=0}^{\infty}$. Specifically, if $Z_n = m$, then

$$Z_{n+1} = Y_1^n + Y_2^n + \cdots + Y_m^n$$

where Y_i^n are independent, identically distributed random variables such that $\mathbb{P}(Y_i^n = k) = p_k$ for $k = 0, 1, 2, \dots$. Think of Y_i^n as the number of children of the i^{th} individual in the n^{th} generation; p_k is the probability that a given parent has k children. If $p_0 \neq 0$, then it is possible that $Y_i^n = 0$ for all i (all parents produce no children) so that the population goes extinct forever. Let us suppose that the mean number of offspring of a single parent is finite:

$$\mu = \mathbb{E}[Y] = \sum_{k=0}^{\infty} k\mathbb{P}(Y = k) = \sum_{k=0}^{\infty} kp_k < \infty.$$

Recall that if Y_1 and Y_2 are two indepdendent, discrete random variables, then

$$\begin{aligned} \mathbb{P}(Y_1 + Y_2 = k) &= \sum_j \mathbb{P}(Y_1 + Y_2 = k \mid Y_2 = j)\mathbb{P}(Y_2 = j) \\ &= \sum_j \mathbb{P}(Y_1 + j = k \mid Y_2 = j)\mathbb{P}(Y_2 = j) \\ &= \sum_j \mathbb{P}(Y_1 = k - j)\mathbb{P}(Y_2 = j). \end{aligned} \tag{4.31}$$

So, if they are identically distributed with $\mathbb{P}(Y_i = k) = p_k$ for $k = 0, 1, 2, \dots$, then

$$\mathbb{P}(Y_1 + Y_2 = k) = \sum_{j=0}^{\infty} p_{k-j}p_j, \qquad k = 0, 1, 2, \dots$$

This expression on the right is called a convolution (of the sequence $\{p_k\}$ with itself). Similarly, if $Y_1, \dots, Y_m$ are independent with same distribution $\{p_k\}_{k=0}^{\infty}$, then their sum has the distribution

$$\mathbb{P}(Y_1 + \cdots + Y_m = k) = p_k^{*m}$$

where p^{*m} denotes the sequence which is obtained from m-fold convolution of the sequence $\{p_j\}$ with itself, and p_k^{*m} is the k^{th} term in this sequence. That is,

$$p_k^{*2} = \sum_{j=0}^{\infty} p_{k-j} p_j, \qquad p_k^{*n+1} = \sum_{j=0}^{\infty} p_{k-j} p_j^{*n}, \qquad \text{etc.}$$

Using this fact, we can write down transition probabilities for the markov chain Z_n:

$$\mathbb{P}(Z_{n+1} = k \mid Z_n = m) = \begin{cases} 0, & \text{if } m = 0, \\ p_k^{*m}, & \text{if } m \geq 1, \quad k \geq 0 \end{cases}$$

Thus, the branching process is determined by the distribution of Z_0 and the offspring distribution $\{p_k\}_{k=0}^{\infty}$.

One can show that

$$\mathbb{E}[Z_n \mid Z_0 = 1] = \mu^n$$

where μ is the mean of the offspring distribution. In particular, if $\mu > 1$, the mean of Z_n grows exponentially; if $\mu < 1$, then the mean of Z_n decreases exponentially as $n \to \infty$.

4.2 Continuous time case

A continuous time branching process Z_t has very similar structure to the discrete time branching process, except that the times between branch events are random variables. Instead of branching at fixed times, each individual branches independently after a time which is distributed as $\text{Exp}(\lambda)$ where the parameter $\lambda > 0$ is called the **branching rate**. It is as though each individual has an independent alarm clock which rings as a time that is $\text{Exp}(\lambda)$, independently of all other clocks. So, if there are currently N individuals, then the next alarm will ring at rate λN; that is, the time until the next ring is distributed as $\text{Exp}(N\lambda)$, since it is the minimum of N independent $\text{Exp}(\lambda)$ random variables. When an individual branches (clock rings), that individual produces a random number of offspring, according to the offspring distribution $\{p_k\}$, as before. So, a continuous time branching process has the same geneological structure as the discrete time process, but the times between branch events is randomized. Consequently, whether or not the process eventually goes extinct, depends only on the offspring distribution, not on the branching rate λ.

Let $m_1(t) = \mathbb{E}[Z_t]$ denote the expected population size at time t. One can show that $m_1(t)$ satisfies the ODE

$$\frac{d}{dt} m_1(t) = \lambda(\mu - 1) m_1(t)$$

where

$$\mu = \sum_{k=1}^{\infty} k p_k$$

is the mean of the offspring distribution. Therefore, $m_1(t) = e^{\lambda(\mu-1)t} m_1(0)$. If $\mu > 1$, the mean population size grows exponentially; if $\mu < 1$, the mean population size decreases exponentially.

Extinction probability, generating functions

The expression for the transition probabilities of Z_n (discrete case) is not so easy to work with. When doing calculations with branching processes, it is sometimes convenient to work with **generating functions**. The generating function for the offspring distribution is the function

$$G(s) = \sum_{k=0}^{\infty} p_k s^k = \mathbb{E}[s^Y]$$

where $Y \sim \{p_k\}$ is a random variable representing the number of children produced by a given individual. This function $G(s)$ is a power series. This function encodes information about the sequence $\{p_k\}_{k=0}^{\infty}$, about the offspring distribution. Using the fact that $\sum_{k=0}^{\infty} p_k = 1$ and $0 \leq p_k \leq 1$, one can check that $G(s)$ has the following properties:

- The radius of convergence of the power series $G(s)$ is at least 1. In particular, $G(s)$ defines a continuous function on $|s| \leq 1$.
- On the interval $[0, 1]$, $G(s)$ is increasing and convex. Under the condition that $p_0 + p_1 < 1$, then $G(s)$ is strictly convex for $s \in [0, 1]$.
- $G(0) = p_0$.
- $G(1) = 1$.
- $G'(1^-) = \mu$ is the expected number of offspring of a single individual.

Theorem 4.1 *Suppose that $Z_0 = 1$. Suppose that $p_0 + p_1 < 1$. Then*

$$\lim_{n\to\infty} \mathbb{P}(Z_n = 0) = \mathbb{P}(\textit{eventual extinction}) = t$$

where $t \in [0, 1]$ is the smallest non-negative root of the equation $t = G(t)$. If $\mu \leq 1$, then $t = 1$ (population eventually becomes extinct). If $\mu > 1$, then $t < 1$ (there is some probability that population never goes extinct).

Remark 4.1 *Note: this result applies to both the discrete time case, and the continuous time case. Whether or not the population goes extinct does not depend on λ. The λ effects the time at which exctintion occurs (if it occurs), but does not effect the probability that it occurs. However, the extinction probability certainly does depend on the offspring distribution.*

Let t be the probability that an individual's descendent family tree goes extinct. That is, $t = \mathbb{P}(Z_n = 0 \text{ for some } n \geq 1 \mid Z_0 = 1)$. To derive the equation $t = G(t)$, let us condition on the first generation, with Y_1 denoting the number of offspring of the (single) parent:

$$\begin{aligned} t &= \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1,\ Y_1 = k)\mathbb{P}(Y_1 = k \mid Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1,\ Y_1 = k)p_k \end{aligned} \tag{4.32}$$

Given that there are k children of the first individual, the probability that this first individual's descendent family tree will go extinct is equal to the probability that each of the k children's trees go extinct. These extinction events are independent. Therefore, $\mathbb{P}(\text{eventual extinction} \mid Z_0 = 1,\ \ Y_1 = k) = t^k$, hence

$$t = \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1,\ \ Y_1 = k) p_k = \sum_{k=0}^{\infty} t^k p_k = G(t).$$

Under the hypothesis that $p_0 + p_1 < 1$, then $G(s)$ is strictly convex on $[0, 1]$. Hence if $G'(1) = \mu \leq 1$, the smallest non-negative root of $t = G(t)$ must be $t = 1$, so extinction occurs with probability 1. On the other hand, if $G'(1) = \mu > 1$, then the smallest root of $t = G(t)$ occurs in the interval $[0, 1)$. If $p_0 = 0$, then clearly $t = 0$.

We can also prove this directly, by thinking about the distribution of Z_n. It will be conventient to consider the generating function for the random variable Z_n:

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

We need a few simple facts about generating functions. We say that a random variable X is a random **counting variable** if it takes values in $\{0, 1, 2, \dots\}$.

Lemma 4.1 *Let X and Y be two independent random counting variables, with generating functions $G_X(s) = \mathbb{E}[s^X]$ and $G_Y(s) = \mathbb{E}[s^Y]$. Then the generating function for the random variable $Z = X + Y$ is $G_Z(s) = G_X(s)G_Y(s)$.*

The proof is simple. Since X and Y are independent, we have:

$$G_Z(s) = \mathbb{E}[s^Z] = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X s^Y] = \mathbb{E}[s^X]\mathbb{E}[s^Y] = G_X(s)G_Y(s)$$

In particular, if X and Y are independent *and* identically distributed, then $G_Z(s) = G_X(s)^2$, since $G_X = G_Y$. Applying this argument iteratively, we obtain the following:

Lemma 4.2 *Let $N \geq 1$ be a fixed positive integer. $Y_1, \dots, Y_N$ be independent, identically distributed random counting variables with generating function $G_Y(s) = \mathbb{E}[s^Y]$. Then the generating function for the sum $Z = Y_1 + \cdots + Y_N$ is $G_Z(s) = G_Y(s)^N$.*

Now suppose N is actually a random variable: we wish to describe the distribution of the sum of a random number of random variables:

Lemma 4.3 *$Y_1, Y_2, Y_3, \dots$ be a collection of independent, identically distributed random variables with generating function $G_Y(s) = \mathbb{E}[s^Y]$. Let N be a random counting variable, independent of the Y_i. Let N have generating function $G_N(s)$. Then the generating function for $Z = Y_1 + \cdots + Y_N$ is $G_Z(s) = G_N(G_Y(s))$.*

proof: Just condition on $N = k$:

$$\begin{aligned} G_Z(s) = \mathbb{E}[s^Z] &= \sum_{k=0}^{\infty} \mathbb{E}[s^Z \mid N = k]\mathbb{P}(N = k) \\ &= \sum_{k=0}^{\infty} \mathbb{E}[s^{Y_1+\cdots+Y_k} \mid N = k]\mathbb{P}(N = k) \\ &= \sum_{k=0}^{\infty} G_Y(s)^k \mathbb{P}(N = k) = \mathbb{E}[G_Y(s)^N] = G_N(G_Y(s)). \end{aligned} \tag{4.33}$$

Now, back to the branching process Z_n.

Lemma 4.4 *Let $G(s)$ be the generating function for the offspring distribution $G(s) = \sum_{k=0} p_k s^k$. Suppose that $Z_0 = 1$ and let $G_n(s) = \mathbb{E}[s^{Z_n}]$ be the generating function for the random variable Z_n. Then*

$$G_{n+m}(s) = G_n(G_m(s)) = G_m(G_n(s)).$$

Hence,

$$G_n(s) = G(G(G(\cdots(G(s))\cdots))) \qquad n\text{-fold composition.}$$

proof: Each individual in generation $n + m$ has is a unique ancestor in generation m (but multiple individuals in generation $n + m$ could have the same ancestor). There are Z_m individuals in generation m. If $X_1, X_2, \ldots, X_{Z_m}$ denote the number of descendents in generation $n+m$ of those individuals in generation m, respectively, then $Z_{n+m} = X_1 + \cdots + X_{Z_m}$. Therefore, given $Z_m = k$, the distribution of Z_{n+m} is the distribution of a sum of k random variables each having same distribution as Z_n (since there are n generations between generation $n + m$ and m). That is, Z_{n+m} is a sum of a random number N of independent random variables, each having the same distribution as Z_n, while N has the distribution of Z_m. So by Lemma 4.3, we know that the generating function for Z_{n+m} is exactly $G_n(G_m(s))$.

Observe that

$$\mathbb{E}[Z_n] = \sum_{k=0}^{\infty} k\mathbb{P}(Z_n = k) = G_n'(1) = G'(1)G_{n-1}'(1) = \cdots = (G'(1))^n = \mu^n.$$

In particular, if $\mu \neq 1$ then the expectation $\mathbb{E}[Z_n]$ either grows or shrinks exponentially.

Let us show that $\mathbb{P}(Z_n = 0) \to t$ as $n \to \infty$, where t is the smallest non-negative root of $t = G(t)$. Let $t_n = \mathbb{P}(Z_n = 0) = G_n(0)$. Then Lemma 4.4 implies that $t_{n+1} = G_{n+1}(0) = G(G_n(0)) = G(t_n)$. Considering the properties of $G(s)$, we see that $t_n \to t$ must happen as $n \to \infty$.

Case I: Suppose $m < 1$. Then $0 \leq G'(s) \leq m < 1$ for all $s \in [0, 1]$, and $t = 1$. Hence

$$|t_{n+1} - t_n| = |G(t_n) - G(t_{n-1})| \leq m|t_n - t_{n-1}|$$

so that

$$|t_{n+1} - t_n| \leq m^n |t_1 - t_0|$$

and

$$|1 - t_n| = \left| \sum_{k=n}^{\infty} t_{k+1} - t_k \right| \leq \sum_{k=n}^{\infty} |t_{k+1} - t_k| \leq |t_1 - t_0| \sum_{k=n}^{\infty} m^k = |t_1 - t_0| \frac{m^n}{1-m} \leq \frac{m^n}{1-m}$$

So the convergence $t_n \to 1$ is exponentially fast as $n \to \infty$.

Case II: Suppose $m > 1$. This case is similar to the case of $m < 1$. One can show that the iterates converge $t_n \to t \in [0, 1)$ exponentially fast, using the fact that $G'(s) < 1$ in a neighborhood of t.

Case III: Suppose $m = 1$. In this case, $t = 1$, but the convergence $t_n \to t$ may be slower than exponential. Observe that

$$0 \leq 1 - t_{n+1} = 1 - G(t_n).$$

Taylor expanding at $s = 1$, we find that $G(s) - s \geq c|1 - s|^2$ for some positive $c \in (0, 1)$. Hence

$$0 \leq 1 - t_{n+1} = 1 - G(t_n) \leq 1 - t_n - c|1 - t_n|^2,$$

which means that the error $a_n = |1 - t_n|$ satisfies the relation

$$0 \leq a_{n+1} \leq a_n - c a_n^2 = a_n(1 - c a_n).$$

This implies that $a_{n+1} < a_n$, and $a_n \to 0$ as $t \to \infty$. To see why $a_n \to 0$, notice that

$$\begin{aligned} \frac{1}{a_k} - \frac{1}{a_0} &= \sum_{n=0}^{k-1} \left(\frac{1}{a_{n+1}} - \frac{1}{a_n} \right) \\ &= \sum_{n=0}^{k-1} \frac{a_n - a_{n+1}}{a_{n+1} a_n} \\ &\geq \sum_{n=0}^{k-1} \frac{c a_n^2}{a_{n+1} a_n} \\ &= \sum_{n=0}^{k-1} \frac{c a_n}{a_{n+1}} \\ &\geq \sum_{n=0}^{k-1} c = ck \end{aligned} \tag{4.34}$$

This shows that $a_k \leq \frac{1}{ck}$. □

Considering the properties of $G(s)$ we have

- If $s \in [0, t)$, then $G_n(s) \nearrow t$ as $n \to \infty$.

- If $s \in (t, 1)$, then $G_n(s) \searrow t$ as $n \to \infty$.

For any $\epsilon > 0$, the convergence $G_n(s) \to t$ is uniform on $[0, 1 - \epsilon)$.

Example: Suppose the offspring distribution is:

$$p_k = qp^k, \quad k \geq 0$$

for some $p \in (0, 1)$, where $q = 1 - p$. Thus, the number of children from a given parent is $Y = X - 1$ where $X \sim \text{Geom}(q)$. Then $m = \mathbb{E}[Y] = \frac{1}{q} - 1 = \frac{p}{q}$. One can compute:

$$G(s) = \frac{q}{1 - ps}.$$

and $t = \min(1, q/p)$.

4.3 A necessary and sufficient condition for transience

Having now learned a bit about generating functions, let us use this machinery to derive a necessary and sufficient condition for transience of a Markov chain. Suppose X_n is an irreducible markov chain on a discrete state space $\mathcal{S}$. If $|S| < \infty$ then every state must be recurrent (actually positive recurrent). If the state space is infinite, however, a state may be transient, even though the chain is irreducible.

Theorem 4.2 *A state $x \in \mathcal{S}$ is transient if and only if $\sum_{n=0}^{\infty} P^{(n)}(x, x) < \infty$. In other words, a state x is recurrent if an only if $\sum_{n=0}^{\infty} P^{(n)}(x, x) = +\infty$.*

Recall that $P^{(n)}(x, x)$ is the probability that the chain goes from x back to x in n steps (although this may not be the first return visit to x).

We can prove Theorem 4.2 using generating functions. To simplify notation, let $p_n = P^{(n)}(x, x)$, for $n = 1, 2, 3, \ldots$, and set $p_0 = 1$. Let the random variable T_x be the first return time to state x (we'll assume $X_0 = x$):

$$T_x = \min\{n \geq 1 \mid X_n = x\},$$

and let us say that $T_x = +\infty$ if $X_n \neq x$ for all $n \geq 1$. Thus,

$$\mathbb{P}(T_x = k \mid X_0 = x) = \mathbb{P}(X_1 \neq x,\ X_2 \neq x,\ \ldots\ ,\ X_{k-1} \neq x,\ X_k = x \mid X_0 = x).$$

For convenience let us define $r_k = \mathbb{P}(T_x = k \mid X_0 = x)$. Observe that

$$\sum_{k=1}^{\infty} r_k = \sum_{k=1}^{\infty} \mathbb{P}(T_x = k \mid X_0 = x) = \mathbb{P}(T_x < \infty \mid X_0 = x) = 1 - \mathbb{P}(T_x = +\infty \mid X_0 = x). \tag{4.35}$$

Thus, state x is transient if and only if $\sum_{k=1}^{\infty} r_k < 1$.

Although the sequences $\{p_k\}$ and $\{r_k\}$ may not define probability distributions on $\{0, 1, 2, \ldots\}$, it is still useful to consider the generating functions for these two sequences:

$$P(s) = \sum_{k=0}^{\infty} s^k p_k, \qquad R(s) = \sum_{k=1}^{\infty} s^k r_k.$$

Since $0 \le p_k, r_k \le 1$ for all k, the radii of convergence for these two power series are at least 1. I claim that

$$P(s) = 1 + P(s)R(s), \quad \text{for all } s \in [0,1). \tag{4.36}$$

To see why this is true, observe that for all $n \ge 1$,

$$\begin{aligned} p_n &= \mathbb{P}(X_n = x \mid X_0 = x) \\ &= \sum_{k=1}^{n} \mathbb{P}(X_n = x \mid T_x = k, \ \ X_0 = x)\mathbb{P}(T_x = k \mid X_0 = x) \\ &= \sum_{k=1}^{n} \mathbb{P}(X_n = x \mid T_x = k, \ \ X_0 = x) r_k. \end{aligned} \tag{4.37}$$

By the Markov property,

$$\mathbb{P}(X_n = x \mid T_x = k \ \ X_0 = x) = \mathbb{P}(X_n = x \mid X_k = x, \ X_{k-1} \neq x, \ \dots) = \mathbb{P}(X_n = x \mid X_k = x) = p_{n-k}$$

Therefore,

$$p_n = \sum_{k=1}^{n} p_{n-k} r_k \tag{4.38}$$

must hold for all $n \ge 1$. Now, multiply (4.38) by s^n and add the resulting equations for $n \ge 1$ to get an expression for $P(s)$:

$$\begin{aligned} P(s) &= 1 + sp_1 + s^2 p_2 + \dots \\ &= 1 + \sum_{n=1}^{\infty} s^n p_n \\ &= 1 + \sum_{n=1}^{\infty} \sum_{k=1}^{n} s^n p_{n-k} r_k \\ &= 1 + \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \mathbb{I}_{\{1 \le k \le n\}} s^{n-k} p_{n-k} s^k r_k \\ &= 1 + \sum_{k=1}^{\infty} s^k r_k \sum_{n=1}^{\infty} \mathbb{I}_{\{1 \le k \le n\}} s^{n-k} p_{n-k} \\ &= 1 + \sum_{k=1}^{\infty} s^k r_k \sum_{j=0}^{\infty} s^j p_j \\ &= 1 + R(s)P(s). \end{aligned} \tag{4.39}$$

This establishes the claim (4.36).

Using (4.36) we solve for $P(s)$:

$$P(s) = \frac{1}{1 - R(s)}, \quad s \in [0,1).$$

Therefore,

$$P(1^-) = +\infty \quad \text{if and only if} \quad R(1^-) = 1.$$

Since $P(1^-) = \sum_{n=0}^{\infty} p_n$ and $R(1^-) = \sum_{k=1}^{\infty} r_k = \mathbb{P}(T_x < \infty)$, by (4.35), this shows that

$$\sum_{n=0}^{\infty} P^{(n)}(x,x) = +\infty \quad \text{if and only if} \quad \mathbb{P}(T_x < \infty) = 1.$$

Thus we have proved Theorem 4.2.

As an example, consider a random walk on the integers $\mathbb{Z}$ with transition probability $P(x, x+1) = p$ and $P(x, x-1) = q = 1-p$. When $p > q$, there is a bias to the right; when $p < q$, the bias is to the left; when $p = q = 1/2$, this is an unbiased random walk. Let T_0 be the first return time to the origin $x = 0$, given $X_0 = 0$. By what we have proved above, the state $x = 0$ is transient if and only if $\sum_n P^{(n)}(0,0) < \infty$. It is easy to see that $P^{(n)}(0,0) = 0$ whenever n is odd, because the number of steps to the left must equal the number of steps to the right if the walk begins and ends at the same state. When n is even $P^{(n)}(0,0)$ will be positive. In fact, $P^{(2n)}(0,0)$ is the probability of tossing exactly n heads and n tails among $2n$ tosses of a coin:

$$P^{(2n)}(0,0) = \binom{2n}{n} p^n q^n, \quad n = 1, 2, 3, \ldots$$

Thus,

$$\sum_n P^{(n)}(0,0) = \sum_n \binom{2n}{n} p^n q^n$$

This also shows that the generating function defined above is:

$$P(s) = \sum_{n=0} \binom{2n}{n} p^n q^n s^{2n} = \sum_{n=0} \frac{(2n)!}{n!n!} p^n q^n s^{2n}.$$

In calculus, we learned how to compute Taylor series. It is easy to check that the Taylor series for the function $f(t) = (1-t)^{-1/2}$ about the point $t = 0$ is:

$$(1-t)^{-1/2} = \sum_{k=0} \frac{(2k)!}{k! 4^k} \frac{t^k}{k!}, \qquad |t| < 1,$$

so that $P(s)$ is exactly

$$P(s) = (1 - 4pqs^2)^{-1/2}.$$

Therefore,

$$R(s) = 1 - \frac{1}{P(s)} = 1 - (1 - 4pqs^2)^{1/2}.$$

In particular,

$$\mathbb{P}(T_x < \infty) = R(1^-) = 1 - (1 - 4pq)^{1/2} = 1 - (1 - 4p + 4p^2)^{1/2} = 1 - |2p - 1| = 1 - |p - q|.$$

So, if $p \neq q$, then $\mathbb{P}(T_x < \infty) = 1 - |p-q| < 1$. In this case, the state $x = 0$ is transient (and so is every other state); even though $X_0 = 0$, the random walk may never come home to 0 (how sad!). On the other hand, in the balanced case, where $p = q = 1/2$, then $\mathbb{P}(T_x < \infty) = 1$ and the state $x = 0$ is recurrent (and so is every other state). More precisely, we can show

that all states are **null recurrent**. If $p = q = 1/2$, then $R(s)$ is the generating function for the random variable T_0, since $\sum_k r_k = 1$ in this case. Hence, $\mathbb{E}[T_0] = R'(1^-)$. Using our explicit formula computed above, with $p = q = 1/2$,

$$R'(s) = s(1 - s^2)^{-1/2}, \quad R'(1^-) = +\infty.$$

So, $\mathbb{E}[T_0] = +\infty$, even though $\mathbb{P}(T_0 < \infty) = 1$: the chain is null-recurrent. In particular, the chain has no invariant probability distribution. It turns out that a simple random walk in $\mathbb{Z}^2$ (i.e. an unbiased walk) is also null-recurrent. For dimension $d \geq 3$, however, the unbiased random walk on $\mathbb{Z}^d$ is transient!

In the computation above, we evaluated $P(s)$ by recognizing that it was related to the power series for the function $f(t) = (1 - t)^{-1/2}$. We could also show directly that $P(1^-) = +\infty$. To do this, we could invoke Stirling's approximation for factorials:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \quad \text{as } n \to \infty,$$

which means that

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1.$$

Using this we obtain (with $p = q = 1/2$)

$$P^{(2n)}(0,0) = \binom{2n}{n} p^n q^n = \frac{(2n)!}{(n!)^2}\frac{1}{4^n} \sim \frac{\sqrt{2\pi 2n}\left(\frac{2n}{e}\right)^{2n}}{2\pi n\left(\frac{n}{e}\right)^{2n}}\frac{1}{4^n} = \frac{1}{\sqrt{\pi n}}$$

as $n \to \infty$. This means that

$$\lim_{n\to\infty} P^{(2n)}(0,0)\sqrt{\pi n} = 1.$$

So, by the limit comparison test for series, the series $\sum_n P^{(2n)}(0,0)$ must diverge, since $\sum_n \frac{1}{\sqrt{\pi n}}$ diverges. The fact that the unbiased walk is recurrent for $d = 2$ and transient for $d \geq 3$ is related to the fact that the series $\sum \frac{1}{n^{d/2}}$ diverges if $d = 1, 2$ and converges if $d \geq 3$.