Stochastic Processes: Markov chains

Math 606, Spring 2025

Luc Rey-Bellet
University of Massachusetts Amherst

2025-04-21

Table of contents

1 Markov chains basics

We introduce the basic definitions necessary to describe Markov chains and provide a first series of examples. For further reading we recommend the books Lawler (2006) and Levin et al. (2017).

1.1 Markov chains on a discrete state spaces

Definition 1.1 (Stochastic process) A *discrete time stochastic process* is a infinite sequence of random variables $X_0, X_1, X_2, \cdots$ where all X_n take values in some space S, called the *state space* of the process. We think of n as time and X_0 as the initial condition.

Formally we can think of a stochastic process as a probability measure on S^∞ (the joint distribution of all X_n) but it is not convenient to put your hands on this object directly. A famous result in measure theory called Kolmogorov extension theorem say that is enough to specify all the joint distribution of $X_{n_1}, X_{n_2}, \cdots, X_{n_k}$ for all choices of k and $n_1 < n_2 \cdots < n_k$.

To start, in this chapter, we assume that S is discrete (either finite or countable) and without loss of generality we can write $S = \{1, 2, \cdots, N\}$ with N finite or not (by relabeling the state as needed). In this context we only need to specify the finite dimensional distributions

$$P\left(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\right) = P\left(X_0 = i_0\right) P\left(X_1 = i_1 | X_0 = i_0\right) P\left(X_1 = i_2 | X_1 = i_1, X_0 = i_0\right) \cdots$$
$$\cdots P\left(X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\right)$$

which we have written in terms of conditional probabilities using the product rule.

The **Markov property** is an assumption on the structure of these conditional probabilities: the future depends only on the present and not on the past.

Definition 1.2 (Markov Chains) A *Markov chain* is a stochastic process with a discrete state space S such that, for all n, all states $i_0, \cdots i_n$, we have

$$P\left(X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\right) = P(X_n = i_n | X_{n-i} = i_{n-1})$$

that is the probability to move to state j at time n depends only on the current position at times $n-1$.

The Markov chain is *time-homogeneous* if $P(X_n = j | X_{n-1} = i)$ is independent of n, that is the probability to move from state i to state j does not depend on the time n of the move.

As a consequence for a Markov the joint pdf can be written as

$$P\left(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\right) \;=\; P\left(X_0 = i_0\right) P\left(X_1 = i_1 | X_0 = i_0\right) \cdots P\left(X_n = i_n | X_{n-1} = i_{n-1}\right)$$

1.2 Matrix notation

For time homogeneous Markov chains it is natural to use the (possibly infinite) vector/matrix notation:

Without loss of generality, we can relabel the states so that $S = \{1, 2, 3, \cdots, N\}$ (with $N \le \infty$). It will be convenient to set

$$\mu = (\mu(1), \mu(2), \cdots, \mu(N))$$

that is μ is a row vector whose entries are the initial distribution.

Also we will write P for the $N \times N$ matrix whose entries are $P(i, j)$

$$P = \begin{pmatrix} P(1,1) & P(1,2) & \cdots & P(1,N) \\ P(2,1) & P(2,2) & \cdots & P(2,N) \\ \vdots & \vdots & & \vdots \\ P(N,1) & P(N,2) & \cdots & P(N,N) \end{pmatrix}$$

We denotethen μP the row vector with entries

$$\mu P(i) = \sum_j \mu(j) P(j, i)$$

1.3 Kolmogorov equations

Proposition 1.1 (Kolmogorov equations) For a Markov process with initial distribution μ and transition probabilities P:

1.The n-step transition probabilities are given by

$$P(X_n = j|X_0 = i) \; = \; P^n(i,j)$$

where P^n is the matrix product $\underbrace{P \cdots P}_{\text{n times}}$.

2. If $\mu(i) = P(X_0 = i)$ is the distribution of X_0 then

$$P(X_n = i) \; = \; \mu P^n(i)$$

is the distribution of X_n.

3. If $f = (f(1), \cdots, f(n))^T$ is a column vector (you may think of f as a function $f : S \to \mathbb{R}$) then we have

$$P^n f(i) \; = \; E\left[f(X_n)|X_0 = i\right].$$

Proof. For 1. we use induction and assume the formula is true for $n-1$. We condition on the state at time $n-1$, use the formula $P(AB|C) = P(A|BC)P(B|C)$, the Markov property, to find

$$
\begin{aligned}
P\left(X_n=j|X_0=i\right) &= \sum_{k\in S} P\left(X_n=j\,,X_{n-1}=k|X_0=i\right)\\
&= \sum_{k\in S} P\left(X_n=j|X_{n-1}=k\,,X_0=i\right) P\left(X_{n-1}=k|X_0=i\right)\\
&= \sum_{k\in S} P\left(X_n=j|X_{n-1}=k\right) P\left(X_{n-1}=k|X_0=i\right)\\
&= \sum_{k\in S} P^{n-1}(i,k)P(k,j) \;=\; P^n(i,j)\,.
\end{aligned}
$$

For 2. note that μP is a probability vector since

$$
\sum_i \mu P(i) \;=\; \sum_i \sum_j \mu(j)P(j,i) \;=\; \sum_j \mu(j) \sum_i P(j,i) \;=\; \sum_j \mu(j)\,.
$$

Furthermore by the formula for conditional probabilities part 1.

$$
P(X_n=j) \;=\; \sum_{k\in S} P(X_n=j|X_0=k)P(X_0=k) \;=\; \sum_k \mu(k)P^n(k,j) \;=\; \mu P^n(j)\,.
$$

For 3. we have

$$
P^n f(i) \;=\; \sum_k P^n(i,k)f(k) \;=\; \sum_k f(k)P(X_n=k|X_0=i) \;=\; E[f(X_n)\,|\,X_0=i]\,.
$$

1.4 Memoryless property

Markov chains forgets their past. For example if we observe that $X_m = i$ for some $i \in S$ then the Markov chain starts anew at i: conditional on the event $\{X_m = i\}$, the stochastic process $Y_n = X_{m+n}, n = 0, 1, 2, \cdots$ is a Markov chain with transition matrix P and initial condition i.
Indeed we have

$$P(X_{m+1} = i_{m+1} \cdots X_{m+n} = i_{m+n} | X_m = i) = \frac{P(X_m = i, X_{m+1} = i_{m+1} \cdots X_{m+n} = i_{m+n})}{P(X_m = i)}$$
$$= \frac{\mu P^m(i) P(i, i_{m+1}) \cdots P(i_{m+n-1}, i_{m+n})}{\mu P^m(i)} = P(i, i_{m+1}) \cdots P(i_{m+n-1}, i_m)$$

Actually a stronger statement holds and shows that the Markov chain after time n is independent of the past!

Theorem 1.1 Suppose the Markov chain is time homogeneous and A is any event which depends only $X_0, \cdots, X_{n-1}$ (the past) then we have

$$P(\{X_{m+1} = i_{m+1} \cdots X_{m+n} = i_{m+n}\} \cap A | X_n = i) = P(X_1 = i_{m+1} \cdots X_n = i_{m+n} | X_0 = i) P(A | X_n = i)$$

Proof. See Homework Exercise 1.3. Any such even A can be written as a union of event of the form $\{X_0 = i_0, \cdots, X_{m-1} = i_{m-1}\}$.

Using the language of measure theory A belong to the σ-algebra $\sigma(X_0, \cdots, X_{m-1})$.

1.5 Stationary and limiting distributions

Basic question: For Markov chain understand the distribution of X_n for large n, for example we may want to know whether the limit

$$\lim_{n\to\infty} P(X_n = i) = \lim_{n\to\infty} \mu P^n(i)$$

exists or not, whether it depends on the choice of initial distribution μ and how to compute it.

Definition 1.3 (Stationary and limiting distributions)

- A probability vector π is called a *limiting distribution* if the limit $\lim_{n\to\infty} \mu P^n = \pi$ exists.
- A probability vector π is called a *stationary distribution* if $\pi P = \pi$.

Remark: Limiting distributions are always stationary distributions: If $\lim_{n\to\infty} \mu P^n = \pi$ then

$$\pi P = (\lim_{n\to\infty} \mu P^n)P = \lim_{n\to\infty} \mu P^{n+1} = \lim_{n\to\infty} \mu P^n = \pi\,.$$

Remark: If π is stationary and is the initial condition then X_n is a sequence of identically distributed random variables.

1.6 Examples

Example 1.1 (2-state Markov chain) Suppose $S = \{1, 2\}$ with transtion matrix $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$.

The equation for the stationary distribution, $\pi P = \pi$ gives

$$\pi(1)(1-p) + \pi(2)q = \pi(1) \qquad \pi(1)p + \pi(2)(1-q) = \pi(2)$$

that is $p\pi(1) = q\pi(2)$. Normalizing to a probability vector gives $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$.

We show π is a limiting distribution. Set $\mu_n \equiv \mu P^n$ and consider the difference $\mu_n - \pi$. Using $\mu_n(2) = 1 - \mu_n(1)$ we get the equation

$$\begin{aligned} \mu_n(1) - \pi(1) = \mu_{n-1}P(1) - \pi(1) &= \mu_{n-1}(1)(1-p) + (1-\mu_{n-1}(1))q - \frac{q}{p+q} \\ &= \mu_{n-1}(1)(1-p-q) - \frac{q}{p+q}(1-p-q) = (1-p-q)(\mu_{n-1}(1) - \pi(1)) \end{aligned}$$

By induction $\mu_n(1) - \pi(1) = (1-p-q)^n(\mu_0(1) - \pi(1))$. If either $p > 0$ or $q > 0$ then $\lim_{n\to\infty} \mu_n(1) = \pi(1)$ and this implies that $\lim_{n\to\infty} \mu_n(2) = \pi(2)$ as well.

If either p or q does not vanish then $\mu_n = \mu P^n$ converges to a stationary distribution for an arbitrary choice of the initial distribution μ.

Example 1.2 (Coupon collecting) A company offers toys in breakfast cereal boxes. There are N different toys available and each toy is equally likelky to be found in any cereal box.

Let X_n be the number of distinct toys that you collect after buying n boxes and is natural to set $X_0 = 0$. Then X_n is a Markov chain with state space $\{0, 1, 2, \cdots, N\}$ and it has a simple structure since X_n either stays the same of increase by 1 unit.

The transition probabilities are

$$
\begin{aligned}
P(j, j+1) &= P\{ \text{ new toy} \mid \text{already j toys}\} = \frac{N-j}{N} \\
P(j, j) &= P\{ \text{ no new toy} \mid \text{already j toys}\} = \frac{j}{N}
\end{aligned}
$$

The Markov chain X_n will eventually reach the state N and stays there forever (N is called an absorbing state). Let us denote by τ the (random) finite time τ it takes to reach the state N. To compute its expectation, $E[\tau]$, let us write

$$
\tau = T_1 + \cdots + T_N \,,
$$

where T_i is the time needed to get a new toy after you have gotten your $(i-1)^{th}$ toy. The T_i's are independent and have T_i has a geometric distribution with $p_i = (N-i)/N$. Thus

$$
E[\tau] = \sum_{i=1}^{N} E[T_i] = \sum_{i=1}^{N} \frac{N}{N-i} = N \sum_{i=1}^{N} \frac{1}{i} \approx N \ln(N) \,.
$$

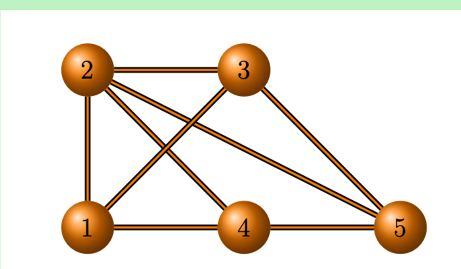
Example 1.3 (Random walk on a graph) Consider an *undirected graph* G consists with *vertex set* V and *edge set* E (each edge $e = \{v, w\}$ is an (unordered) pair of vertices). We say that the vertex v is a *neighbor* of the vertex w, and write $v \sim w$, if $\{v, w\}$ is an edge. The *degree* of a vertex v, denoted $\deg(v)$, is the number of neighbor of v.

Given such a graph $G = (V, E)$ the *simple random walk* on G is the Markov chain with state space V and transition matrix

$$P(v,w) = \begin{cases} \frac{1}{\deg(v)} & \text{if} w \sim v \\ 0 & \text{otherwise} \end{cases} .$$

The invariant distribution for the random walk on graph is given by $\pi(v) = \frac{\deg(v)}{2|E|}$ where is the number of edges. First note that $\sum_v \pi(v) = 1$ since each edge connects two vertices. To show invariance note that

$$\pi P(v) = \sum_w \pi(w) P(w,v) = \sum_{w; w \sim v} \frac{\deg(w)}{2|E|} \frac{1}{\deg(w)} = \frac{1}{2|E|} \sum_{w; w \sim v} 1 = \pi(v) .$$



$$P = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix} \qquad \pi = \left(\frac{3}{16}, \frac{4}{16}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}\right)$$

Example 1.4 (Random walk on the hypercube)

The d-dimensional hypercube graph Q_d has for vertices the binary d-tuples

$$\mathbf{x} = (x_1, \cdots, x_d) \quad \text{with } x_k \in \{0, 1\}$$

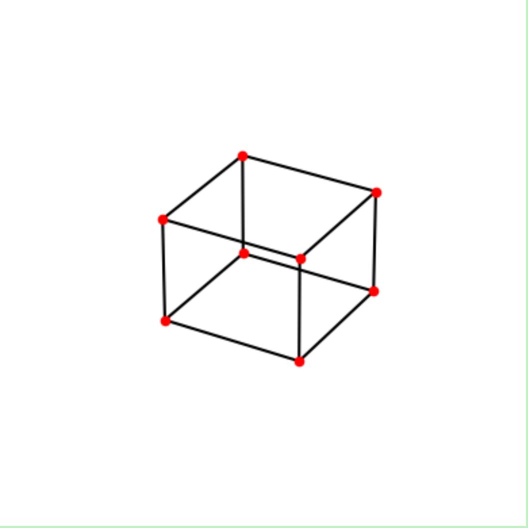
Two vertices are connected by an edge when they differ in exactly one coordinate (flipping a 0 into 1 or vice-versa).

The simple random walk on Q_d moves from one vertex $x = (x_1, \cdots, x_d)$ by choosing a coordinate $j \in \{1, 2, \cdots, d\}$ uniformly at random and setting $x_j \to (1 - x_j)$.

The degree of each vertex is d, the number of vertices is 2^d and the number of edges is $2^d \frac{d}{2}$.

The stationary distribution is $\pi(\mathbf{x}) = \frac{1}{2^d}$,the uniform distribution on Q_d.

Variation on this random walk have many applications, you can intepret the vector $\mathbf{x}$ as describing which one of d objects is on a list (see the section on Monte-Carlo Markov chains)



Example 1.5 (Assymetric random walks on {0,1, , N}) State space $S = \{0, 1, \cdots, N\}$ and the Markov chain goes up by 1 with probability p and down by 1 with probaility $1-p$.

$$P(j, j+1) = p,\ P(j, j-1) = 1-p\,, \quad \text{for } j = 1, \cdots, N-1$$

We can pick different boundary conditions (BC) at 0 and N:

- Absorbing BC: $P(0,0) = 1, P(N,N) = 1$
- Reflecting BC: $P(0,1) = 1, P(N, N-1) = 1$
- Partially reflecting BC: $P(0,0) = (1-p)\,, P(0,1) = pP(N, N-1) = (1-p)\,, P(N,N) = p$
- Periodic BC: $P(0,1) = p\,, P(0,N) = (1-p)\,, \quad P(N,0) = p\,, P(N,N-1) = (1-p)$

Example 1.6 (Ehrenfest urn model) Suppose d balls are distributed among two urns, urn A and urn B. At each move one ball is selected uniformly at random among the d balls and is transferred from its current urn to the other urn. If X_n is the number of balls in urn A then the state space is $S = \{0, 1, \cdots, d\}$ and the transition probabilities

$$P(j, j+1) \;=\; \frac{d-j}{d}\,, \quad P(j, j-1) \;=\; \frac{j}{d}\,.$$

We show that the invariant distribution is binomial with paramters $(d, \frac{1}{2})$, that is $\pi(j) \;=\; \binom{d}{j}\frac{1}{2^d}$.

$$\begin{aligned}\pi P(j) &= \sum_k \pi(k)P(k,j) = \pi(j-1)P(j-1,j) + \pi(j+1)P(j+1,j) \\ &= \frac{1}{2^d}\left[\binom{d}{j-1}\frac{d-(j-1)}{d} + \binom{d}{j+1}\frac{j+1}{d}\right] = \binom{d}{j}\frac{1}{2^d}\,.\end{aligned}$$

This Markov chain is closely related to the simple random walk Y_n on the hypercube Q_d. Indeed selecting randomly one of the d balls and moving it the other urn is equivalent to selecting a random coordinate y_k of $\mathbf{y} = (y_1, \cdots, y_d)$ and changing it to $1 - y_k$. If we denote by $j = |\mathbf{y}| = y_1 + \cdots + y_d$ to be the number of 1s in $\mathbf{y}$. Then

$$P(X_n = j+1 | X_n = j) = P(\text{ choose } k \text{ such that } y_k = 0) = \frac{d-j}{d}$$

1.7 Simulation of Markov chains

- It is relatively easy to compute the distribution of X_n by matrix multiplication, that is μP^n if P is not too large.
- It is also not difficult to generate the paths $X_0, X_1, X_2, \cdots$ of the Markov chain.
  - Set $X_0 = i$ (or generate the random variable X_0 with distribution μ)
  - Generate the RV Z with probability distribution $(P(i,1), \cdots P(i,n))$ and set $X_1 = Z$.
  - and so on.
- To generate a discrete RV with finite state space do

```
import numpy as np

# Create an array to describe the state space
elements = np.array([1, 2, 3, 4, 5])

# Define nonuniform probabilities for each element
probabilities = np.array([0.1, 0.2, 0.3, 0.2, 0.2])

# Use random.choice with nonuniform probabilities
random_element = np.random.choice(elements, p=probabilities)
```

1.8 Exercises

Exercise 1.1 A standard die is rolled repeatedly. Which of the following are Markov chains? For those that are, supply the transition matrix.

1. The largest number X_n shown up to the nth roll.
2. The number N_n of sixes in n rolls.
3. At time r the time C_r, since the most recent six.
4. At time r, the time B_r until the next six.

▶ Solution

Exercise 1.2 Suppose X_n is a Markov chain on the state space $\{1, 2, 3, 4, 5, 6\}$. Is it true that

$$P(X_2 = 6|X_1 \in \{3, 4\}, X_0 = 2) = P(X_2 = 6|X_1 \in \{3, 4\})?$$

$$P(X_2 = 6|X_1 = 3, X_0\{2, 5\}) = P(X_2 = 6|X_1 = 3)?$$

Prove or disprove.

Exercise 1.3 (More on the Markov property) We have introduced the Markov property as meaning that the future depends on the present but not on the past, see Definition 1.2. Show that for a Markov chain the following two condition hold

1. The past depends only on the present but not on the future:

$$P(X_0 = i_0 \,|\, X_1 = i_1, \cdots X_n = i_n) = P(X_0 = i_0 \,|\, X_1 = i_1)\,.$$

2. Conditioned on the present, the past and the future are independent,:

$$P(X_{n+1} = i_{n+1}, X_{n-1} = i_{n-1} | X_n = i_n) = P(X_{n+1} = i_{n+1} | X_n = i_n) P(X_{n-1} = i_{n-1} | X_n = i_n)$$

3. Generalize part b. and show that for any event A which depends only on $\{X_0, \cdots, X_{n-1}\}$ we have

$$\begin{aligned} &P(\{X_{n+1} = i_{n+1}, \cdots X_{n+m} = i_{n+m}\} \cap A | X_n = i) \\ &= P(X_{n+1} = i_{n+1}, \cdots X_{n+m} = i_{n+m} | X_n = i) P(A | X_n = i) \end{aligned}$$

▶ Solution

Exercise 1.4 (2-steps Markov chain) Suppose that X_n is a Markov chain with state space S, transition probabilities $P(i,j)$ and stationary distribution $\pi(i)$. Show that

$$Z_n = (X_n, X_{n+1})$$

is a Markov chain. What are (a) the state space, (b) the transition probabilities, and (c) the stationary distribution?

▶ Solution

Exercise 1.5 (Markov chain finite memory) Instead of the Markov property let us assume X_n depends on the previous two steps: i.e we have for all n and states

$$P\{X_n = i_n \,|\, X_{n-1} = i_{n-1}, \cdots X_0 = i_0\} = P\{X_n = i_n \,|\, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}\}\,.$$

This is called a 2**-Markov chain** and if it is time-homogeneous it is specified by the numbers $Q_{i,j,k} = P\{X_n = k \,|\, X_{n-1} = j, X_{n-2} = i\}$. Show that

$$Z_n = (X_n, X_{n+1})$$

is a Markov chain. Describe its state space and transition probabilities.

▶ Solution

Exercise 1.6 Suppose $\{X_n\}$ are independent indentically distributed random variables taking values in $\mathbb{Z}$. Determine which of the following are Markov chains.

1. X_n
2. $S_n = X_1 + \cdots + X_n$ (with $S_0 = 0$)
3. $Y_n = X_n + X_{n-1}$ (with $X_{-1} = 0$ so $Y_0 = X_0$)
4. $Z_n = \sum_{k=0}^{n} S_k$
5. (S_n, Z_n)

▶ Solution

Exercise 1.7 (Umbrella Markov chain) Jane possesses r umbrellas which she uses going from her home to her office in the morning and vice versa in the evening. If it rains in the morning or in the evening she will take an umbrella with her provided there is one available. Assume that independent of the past it will rain in the morning or evening with probability p. Let X_n denote the number of umbrellas at her home before she gets to work.

1. Give the state space and the transition probabilities describing the Markov chain X_n.
2. Find the stationary distribution $\pi(j)$.
3. Estimate the number of times in a year where Jane gets wet.

▶ Solution

Exercise 1.8 (Simulation of Markov chains) Write codes which simulate the umbrella Markov chain in the previous exercise (or other Markov chains where the transition matrix is not too big). The input should be the initial distribution, or initial state, and the transition probability matrix. Your code should return as outputs

- P^n for suffcently large n.
- One suffcently long path $X_0, X_1, X_2, \cdots, X_n$ of the Markov chain using the simulation algorithm in Section 1.7. Use this path to extract the path statistics: find the proportion of time spent in every state (for example do an histogram).
- What do you observe for the umbrella Markov chains?

Exercise 1.9 Suppose X_n is a Markov chain with state space S and $h : S \to T$ is a function.

- Show that if h is one-to-one then $Y_n = h(X_n)$ is a Markov chain.
- Show that if h is not one-to-one then $Y_n = h(X_n)$ is not a Markov chain in general.
- If h is not one-to-one and Y_n is a Markov chain then the Markov chain X_n is called lumpable. Show that the random walk on the hypercube is lumpable if $h(x_1, \cdots, x_d) = x_1 + \cdots x_d$.

Exercise 1.10 (Canonical representation of Markov chains)

1. A general way to construct a Markov chain is to use a noise model. Take $Z_1, Z_2, Z_3 \cdots$ to be IID random variables taking value in some arbitary space E. If $f : S \times E \to S$ is a function show that

$$X_{n+1} = f(X_n, Z_{n+1})$$

defines a Markov chain. For example you can think of $f(X_n) + Z_{n+1}$ as determinstic evolution plus noise.

2. Conversely show that any Markov chain can be represented in this form.
Hint: Pick Z_n to be independent uniform random variable on $[0, 1]$ and use the simulation algorithm in Section Section 1.7.

▶ Solution

2 Ergodic theory of finite state space Markov chains

We present the basic convergence theory for finite state Markov chains and introduce the concept of irreducubility and period of a Markov chain. This is very classical stuff, see e.g. Lawler (2006) and Levin et al. (2017).

2.1 Existence of stationary distributions

Stationary distributions always exist for finite state Markov chains. This will not be the case if the state space is countable.

Theorem 2.1 (Existence) Let X_n be a Markov chain on a finite state space S. Then X_n has at least one stationary distribution.

Proof. Boltzano Weierstrass theorem asserts that any bounded sequences x_k in $\mathbb{R}^d$ has a convergence subsequence.

Choose μ_0 to be an arbitrary initial distribution and let $\mu_n = \mu P^n$. Appling Boltzano Weierstrass directly to the sequence μ_n will not work but consider instead the time averages sequences

$$\nu_n = \frac{\mu + \mu P + \cdots \mu P^{n-1}}{n}$$

The sequence ν_n is bounded since μ_n and thus ν_n are probability vector. Therefore there exists a convergent subsequence ν_{n_k} with $lim_k \nu_{n_k} = \pi$ and π is a probability vector as well. We show that π is a stationary distribution. Note

$$\nu_n P - \nu_n = \frac{\mu P + \cdots \mu P^n - \mu - \cdots - \mu P^{n-1}}{n} = \frac{\mu P^n - \mu}{n} \quad \text{and thus} \quad |\nu_n P(j) - \nu_n(j)| \leq \frac{1}{n}$$

To conclude we note that

2.2 Uniqueness of stationary distributions

It is easy to build examples of Markov chain with multiple stationary distributions, for example

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & \frac{5}{6} \\ 0 & 0 & \frac{4}{5} & \frac{1}{5} \end{pmatrix} \qquad \pi_1 = \left(\frac{3}{5}, \frac{2}{5}, 0, 0\right) \text{ and } \pi_2 = \left(0, 0, \frac{25}{49}, \frac{24}{49}\right) \qquad \text{are stationary}$$

If the Markov chain starts in state 1 or 2 it will only visits states 1 and 2 in the future and so we can build a stationary distribution (by Theorem 2.1 using an initial μ_0 which vanishes on 3 and 4) which will vanish on the state 3 and 4.

Definition 2.1 (communication and irreducibility)

- We say that j **is accessible from** i, symbolically $i \rightsquigarrow j$ if there exists $n \geq 0$ such that $P^n(i,j) > 0$.
- We say that i **and** j **communicate**, symbolically $i \leftrightsquigarrow j$ if $i \rightsquigarrow j$ and $j \rightsquigarrow i$.
- A Markov chain X_n is **irreducible** if every state $i \in S$ communicate with every other state $j \in S$, that is, for any pair of states i, j in S there exists n such that $P^n(i,j) > 0$.

In an irreducible Markov chain, starting with any state the Markov chain will eventually visit any other state.

Theorem 2.2 (Uniqueness) Suppose X_n is an irreducible Markov chain with finite state space S.

1. If π is a stationary distribution then $\pi(j) > 0$ for all $j \in S$,
2. Suppose h is a column vector such that $Ph = h$ then $h = c(1, 1, \cdots, 1)^T$ is a constant vector.
3. The Markov chain X_n has a unique stationary distribution π.

Proof. For 1. choose i with $\pi(i) > 0$. If j is such that $i \rightsquigarrow j$ then $P^r(i,j) > 0$ for some r and thus

$$\pi(j) = \pi P^r(j) \; = \; \sum_k \pi(k) P^r(k,j) \; \geq \; \pi(i) P^r(i,j) > 0 \, .$$

Since X_n is irreducible, $\pi(j) > 0$ for any $j \in S$.

For 2. suppose that $Ph = h$ and i_0 such that $h(i_0) = \max_{i \in S} h(i) \equiv M$. If h is not a constant vector there exists j with $i_0 \rightsquigarrow j$ but $h(j) < M$ and $P^r(i_0, j) > 0$. Since $P^r h = h$,

$$M = h(i_0) \; = \; P^r h(i_0) \; = \; P^r(i_0, j) \underbrace{h(j)}_{<M} + \sum_{l \neq i_0} P^r(i_0, l) \underbrace{h(l)}_{\leq M} < M \sum_l P^r(i_0, l) = M \, ,$$

and this is a contradiction.

For 3. part 2. shows that the geometric multiplicity of the eigenvalue 1 matrix P is equal to 1 and thus so is the geometric multiplicity of the eigenvalue 1 for the adjoint P^T and thus P has at most one left eigenvector π.

2.3 Convergence to stationary distribution

- **Question**: If we have a unique stationary distribution is it correct that $\lim_{n\to\infty} \mu P^n = \pi$?

- Without further assumption the answer is NO, because of possible periodic behavior. Consider for example the random walk on $\{1, \cdots, 2N\}$ with, for example, periodic boundary conditions. The stationary distribution is uniform $\pi = \left(\frac{1}{2N}, \cdots, \frac{1}{2N}\right)$. But if X_0 is even then X_1 is odd and then alternating periodically between odd and even positions. For example for

$$P = \begin{pmatrix} 0 & 1/4 & 0 & 3/4 \\ 3/4 & 0 & 1/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \end{pmatrix} \qquad \pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \text{ is stationary}$$

and

$$P^{10} = \begin{pmatrix} 0.4995 & 0 & 0.5004 & 0 \\ 0 & 0.4995 & 0 & 0.5004 \\ 0.5004 & 0 & 0.4995 & 0 \\ 0 & 0.5004 & 0 & 0.4995 \end{pmatrix} \qquad P^{11} = \begin{pmatrix} 0 & 0.5002 & 0 & 0.4997 \\ 0.4997 & 0 & 0.5002 & 0 \\ 0 & 0.4997 & 0 & 0.5002 \\ 0.5002 & 0 & 0.4997 & 0 \end{pmatrix}$$

and $P^n(i,j)$ does not converge, oscillates, asymptotically between 0, and $1/2$.

Definition 2.2 A Markov chain X_n is called *irreducible and aperiodic* if there exists an integer n such that $P^n(i,j) > 0$ for all pair i, j in S. (The meaning of the terminology will become clearer later on.)

Theorem 2.3 (Doeblin's Theorem) Suppose X_n be an irreducible and aperiodic Markov chain on the finite state space S with stationary distribution π. There exists a constant $C > 0$ and number α with $0 \le \alpha < 1$ such that for any initial distribution μ we have

$$|\mu P^n(j) - \pi(j)| \le C\alpha^n \,, \tag{2.1}$$

i.e., the distribution of X_n converges, exponentially fast, to π.

Proof. Since the Markov chain is irreducible and aperiodic we can find an integer r such that P^r has strictly positive entries. Let Π be the stochastic matrix

$$\Pi = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(N) \\ \pi(1) & \pi(2) & \cdots & \pi(N) \\ \vdots & \vdots & & \vdots \\ \pi(1) & \pi(2) & \cdots & \pi(N) \end{pmatrix}$$

where every row is the stationary distribution π. Note that this corresponds to independent sampling from the stationary distribution.

Let us set $\theta = 1 - \delta$ and by Equation 2.2 we define a stochastic matrix Q through the equation

$$P^r = (1-\theta)\Pi + \theta Q\,.$$

Note the following facts:

- Since $\pi P = \pi$ we have $\Pi P^n = \Pi$ for any $n \geq 1$.
- For any stochastic matrix M we have $M\Pi = \Pi$ since all rows of Π are identical.

Using these two properties we show, by induction, that any integer $k \geq 1$, $P^{kr} = (1-\theta^k)\Pi + \theta^k Q^k$. This is true for $k = 1$ and so let us assume this to be true for k. We have

$$\begin{aligned} P^{r(k+1)} = P^{rk}P^r &= \left[(1-\theta^k)\Pi + \theta^k Q^k\right] P^r = (1-\theta^k)\Pi P^r + \theta^k Q^k[(1-\theta)\Pi + \theta Q] \\ &= (1-\theta^k)\Pi + \theta^k(1-\theta)\Pi + \theta^{k+1}Q^{k+1} = (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1} \end{aligned}$$

and this concludes the induction step.

From this we conclude that $P^{rk} \to \Pi$ as $k \to \infty$. Write $n = kr + l$ where $0 \leq l < r$. We have then

$$P^n = P^{kr}P^l = [(1-\theta^k)\Pi + \theta^k Q^k]P^l = \Pi + \theta^k\left[Q^k P^l - \Pi\right]$$

and thus

2.4 The period of a Markov chain

As we have seen before Markov chain can exhibit periodic behavior and circle around various part of the state space.

Definition 2.3 The *period of state* j, $\mathrm{per}(j)$ is the greatest common divisor of the set

$$\mathcal{T}(j) = \{n \geq 1\,,\, P^n(j,j) > 0\}$$

that is $\mathcal{T}(j)$ is set of all times at which the chain can return to j with positive probability.

For example for random walks with periodic boundary conditions the period is 2 is the number of sites is even and 1 if the number of sites is odd.

Lemma 2.1 If $i \leftrightsquigarrow j$ then $\text{per}(i) = \text{per}(j)$.

Proof. Suppose $\text{per}(j) = d$ and $i \leftrightsquigarrow j$. There exist integers m and m such that $P^m(i,j) > 0$ and $P^n(j,i) > 0$ and thus

$$P^{m+n}(i,i) > 0 \text{ and } P^{m+n}(j,j) > 0$$

which means that $m + n \in \mathcal{T}(i) \cap \mathcal{T}(j)$ and so $m + n = kd$ for some integer k.

Suppose $l \in \mathcal{T}(i)$ then

$$P^{n+m+l}(j,j) \geq P^n(j,i)P^l(i,i)P^m(i,j) > 0$$

and $n + m + l \in \mathcal{T}(j)$. Therefore d divides l since it divides both $n + m + l$ and $n + m$ and this shows that $\text{per}(j) \leq \text{per}(i)$. Reversing the roles of i and j we find $\text{per}(j) = \text{per}(i)$. ■

Consequences:

- If X_n is irreducible then all states have the same period and we can define the period of the irreducible Markov chain X_n .
- If X_n is irreducible and has period 1 (also called aperiodic) and S is finite then $P^n(i,j) > 0$ for all suficiently large n (compare with our earlier definition of irreducible and paeriodic).

Decomposing the state space $S = G_1 \cup G_2 \cup \cdots G_d$:

- Start with a state $i \in S$ and let $i \in G_1$.
- For all j such that $P(i,j) > 0$ let $j \in G_2$.
- For all $j \in G_2$ and all k such that $P(j,k) > 0$ let $k \in G_3$.
- After assigning set to G_d assign the next states to G_1.
- Repeat until all states have been assigned. If $|S|$ is finite this takes at most $|S|$ steps.

By construction we get a partition of S into d distinct subsets $G_1, \cdots, G_d$ and

$$P(i,j) > 0 \implies i \in G_k \text{ and } j \in G_{k+1 (\text{mod} d)}$$

For example

$$P = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix} \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .2 & 0 & 0 & .8 & 0 \\ .3 & .7 & 0 & 0 & 0 & 0 & 0 \\ .2 & .8 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & .7 & .3 & 0 \end{pmatrix}$$

Iteratively we find

$$1 \rightsquigarrow 4 \rightsquigarrow 3,6 \rightsquigarrow 1,2 \rightsquigarrow 4,7 \rightsquigarrow 3,5,6 \rightsquigarrow 1,2 \rightsquigarrow 4,7$$

and thus the Markov chain is irreducible with period 3 and so we have

$$S = \{1,2\} \cup \{4,7\} \cup \{3,5,6\}$$

Relabeling the state the transition matrix P takes the following block matrix form

$$P = \begin{matrix} G_1 \\ G_2 \\ \vdots \\ G_{d-1} \\ G_d \end{matrix} \begin{pmatrix} 0 & P_{G_1 G_2} & 0 & \cdots & 0 \\ 0 & 0 & P_{G_2 G_3} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & P_{G_{d-1} G_d} \\ P_{G_d G_1} & 0 & \cdots & 0 & 0 \end{pmatrix}$$

where $P_{G_l G_{l+1}}$ describe the transition between states in G_l and G_{l+1}.

Putting the matrix P to the power d we find the block diagonal form

$$P^d = \begin{matrix} G_1 \\ G_2 \\ \vdots \\ G_d \end{matrix} \begin{pmatrix} Q_{G_1} & 0 & 0 & \cdots & 0 \\ 0 & Q_{G_2} & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & Q_{G_d} \end{pmatrix}$$

Theorem 2.4 If X_n is an irreducible Markov chain on a finite state space with stationary distribution π then we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k(i,j) = \pi(j) .$$

Proof. The matrix Q_{G_l}, restricted to state space G_l governs the evolution of an irreducible aperiodic Markov chain and thus, by Theorem 2.3, we have

$$\lim_{n\to\infty} P^{nd}(i,j) \;=\; \lim_{n\to\infty} Q_{G_l}^n(i,j) = \pi_{G_l}(j) \qquad \text{for } i \in G_l, j \in G_l$$

where π_{G_l} is the stationary distribution for Q_{G_l}. Morever $P^l(i,j) = 0$ if $l \neq nd$ for some n.

If $i \in G_{l-1}$ and $j \in G_l$ then

$$\lim_{n\to\infty} P^{nd+1}(i,j) \;=\; \lim_{n\to\infty} \sum_{k\in G_l} P(ik)P^{nd}(k,j) \;=\; \sum_{k\in G_l} P(ik)\pi_{G_l}(j) \;=\; \pi_{G_l}(j)\,,$$

and $P^m(i,j) = 0$ if $m \neq nd+1$ for some n. Similarly if $i \in G_{l-r}$ and $j \in G_l$ we have

$$\lim_{n\to\infty} P^{nd+r}(i,j) \;=\; \pi_{G_l}(j)\,.$$

and $P^m(i,j) = 0$ if $m \neq nd+r$ for some n.

The sequence $P^n(i,j)$ is asymptotically periodic where $d-1$ successive 0 alternates with a number converging to $\pi_{G_l}(j)$ and thus if we define $\pi \equiv \frac{1}{d}(\pi_{G_1}, \cdots, \pi_{G_d})$ then π is normalized, stationary and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} P^k(i,j) \;=\; \pi(k)$$

since the time spend in state j is asymptotically equal to $\frac{1}{d}\pi_{G_l}(j)$.

2.5 Hitting times, return times and the strong Markov property

We now have good understanding the asymptotic behavior of $P^n(i,j)$ or μP^n as $n \to \infty$. When simulating or observing a Markov chain we usually observe a single sequence

$$X_0, X_1, X_2, \cdots$$

and our next goal is to derive a **law large numbers** for this sequence of random variables. To do this it will be useful to track how often the Markov chain visits some state.

Definition 2.4 (Hitting time and return times)

- The **Hitting time for the state** j, $\sigma(j)$, is the first time the Markov chain visits the state j.

$$\sigma(j) = \inf\{n \geq 0\,;\, X_n = j\} \quad \text{hitting time}$$

- The **return time for the state** j, $\tau(j)$ is the first time the Markov chain returns to the state j.

$$\tau(j) = \inf\{n \geq 1\,;\, X_n = j\} \quad \text{return time}$$

Hitting time and return times are closely related and are distinct only from the way they consider what happen at time 0 (both will be useful later on).

Hitting time and return times are example of stopping times and have the crucial property that to decide the events $\tau(j) = k$ we only need to know what happened in $X_0, \cdots X_k$.

Definition 2.5 (Stopping times) A random variable T taking values in the integer is called a stopping time for the Markov chain X_n if, for any k, the event $\{T = k\}$ depends only on $X_0, X_1, \cdots, X_k$, that is $\{T = k\}$ belong to the σ-algebra $\sigma(X_0, \cdots, X_k)$.

Intuitively stopping times are random times for the Markov chains which are decided only by looking at the past and the present and not the future, such as hitting time and return times.

Stopping times are very useful for Markov chain: if you run a Markov chain up to a stopping time T, then after T the Markov chain starts anew, that is
$Y_n = X_{T+n}$ is a Markov chain which is independent of the past. More precisely we have

Theorem 2.5 (Stong Markov property) Consider a stopping time T with $P(T < \infty) = 1$. Then conditional on $X_T = i, Y_n = X_{T+n}$ is a Markov chain with transition P and initial condition i.

Proof. If A is an event determined by $X_0, \cdots, X_T$ then $A \cap \{T = m\}$ is determined by $X_0, \cdots, X_m$. By the Markov property at time m we have

$$P(X_T = j_0, X_{T+1} = j_1, \cdots, X_{T+n} = j_n \cap A \cap T = m \cap X_T = i) = P(X_1 = j_1, \cdots, X_n = j_n | X_0 = i)P($$

2.6 Strong law of large numbers for Markov chains

Irreducibility means that starting from i the Markov chain will reach j and thus for any i we have

$$P\{\tau(j) < \infty | X_0 = i\} > 0$$

If the state space is finite then we have something much stronger. Not only we have $P\{\tau(j) < \infty | X_0 = i\} = 1$, the expectation of $\tau(j)$ is actually finite.

Lemma 2.2 For an irreducible Markov chain with finite state space S the expected return time $E[\sigma(j)|X_0 = i] < \infty$

Proof. By irreducibility the Markov chain can reach the state j from any state i in a finite time. Since S is finite this time is uniformly bounded. So there exist a time m and $\varepsilon > 0$ such that $P^l(i,j) \geq \epsilon$ for some $l \leq m$ uniformly in $i \in S$. Using the strong Markov property this implies that

$$P\{\sigma(j) > km | X_0 = i\} \leq (1-\varepsilon) P\{\sigma(j) > (k-1)m | X_0 = i\} \leq \cdots \leq (1-\varepsilon)^k$$

Therefore, using that $P\{\sigma(j) > n | X_0 = i\}$ is a decreasing function of n

$$E[\sigma(j)|X_0 = i] = \sum_{n \geq 0} P\{\sigma(j) > n | X_0 = i\} \leq m \sum_{k \geq 0} P\{\sigma(j) > km | X_0 = i\} \leq m \sum_{k \geq 0} (1-\varepsilon)^k < \infty \quad \blacksquare$$

Remark If S is infinite this is not necessarily true and this will lead to the concepts of transience, recurrence and positive recurrence.

We will also need the random variable $Y_n(j)$ which counts the number of visits to starte j up to time n that is

$$Y_n(j) \equiv \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k=j\}} \quad = \text{ the number of visits to the state } j \text{ up to time } n$$

Since $E\left[\mathbf{1}_{\{X_k=j\}}|X_0=i\right] = P^k(i,j)$ and we have

$$\frac{1}{n}E[Y_n(j)|X_0=i] = \frac{1}{n}\sum_{k=0}^{n-1} P^k(i,j)$$

and thus by Theorem 2.3 or Theorem 2.4

$$\pi(j) \;=\; \lim_{n\to\infty}\frac{1}{n}E\left[\sum_{k=1}^{n}\mathbf{1}_{\{X_k=j\}}\right] \quad = \text{ the proportion of time spent in state } j \text{ in the long run}$$

This is an important intuition

$$\pi(j) = \text{expected fraction of time that the Markov chain spends in state } j \text{ in the long run.}$$

The next results strengthen that intepretation.

Theorem 2.6 (Ergodic Theorem for Markov chain) Let X_n be an irreducible Markov chain with an arbitrary initial condition μ, then, with probability 1 we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_n=j\}} = \pi(j) \,,$$

this is the *ergodic theorem* for Markov chain.

Moreover if $\tau(j)$ is the first return time to j we have the *Kac's formula*

$$\pi(j) = \frac{1}{E[\tau(j)|X_0=j]} \,.$$

Proof. For a Markov chain starting in some state i consider the successive return to the state j. By the strong Markov property, once the Markov chain has reached j it starts a fresh. So the time when the Markov chain returns to state j for the k^{th} time is

$$T_k(j) = \tau_1(j) + \cdots + \tau_k(j) \,,$$

where $\tau_l(j)$ are independent copies of the return times $\tau(j)$. For $l \geq 2$ $\tau_l(j)$ is conditioned on starting at j while for $l=1$ it depends on the initial condition. Note that by the strong LLN for IID random variables we have

$$\lim_{k\to\infty} \frac{T_k(j)}{k} = \lim_{k\to\infty} \frac{1}{k} \left(\tau_1(j) + \cdots + \tau_k(j)\right) = E[\tau(j)|X_0=j] \,.$$

Note further that if the total number of visits to state j up to time n is k, it means that that we have returned exactyl k times and so

$$\{Y_n(j) = k\} = \{T_k(j) \le n < T_{k+1}(j)\}$$

So we have

$$\frac{T_{Y_n(j)}(j)}{Y_n(j)} < \frac{n}{Y_n(j)} \le \frac{T_{Y_n(j)+1}(j)}{Y_n(j)+1}\frac{Y_n(j)+1}{Y_n(j)}$$

As $n \to \infty$ $Y_n(j) \to \infty$ too and taking $n \to \infty$ both extremes of the inequality converge to $E[\tau(j)|X_0 = j]$ with probability 1 and thus we conclude that, with probability 1,

$$\lim_{n\to\infty} \frac{Y_n(j)}{n} = \frac{1}{E[\tau(j)|X_0 = j]} .$$

On the other hand we know that

$$\lim_{n\to\infty} \frac{E[Y_n(j)]}{n} = \pi(j) ,$$

and thus

$$\pi(j) = \frac{1}{E[\tau(j)|X_0 = j]} .$$

Putting all the pieces together gives the theorem. ■.

2.7 Exercises

Exercise 2.1 (Algorithm to compute the stationary distribution) Instead of solving $\pi P = \pi$ (that is computation of an eigenvector) we provide an alternative formula for π which is very easy to implement on a computer. We let I be the identity matrix and we let M be the matrix whose all entries are $M(i,j) = 1$.

- Prove that if X_n is an irreducible Markov chain with transition probabilities matrix P, then the unique stationary distribution π is given by

$$\pi \; = \; (1, 1, \cdots, 1)\,(I - P + M)^{-1} \; . \tag{2.3}$$

  Hint: Assuming first that the matrix $(I - P + M)$ is invertible show Equation 2.3. To that $(I - P + M)$ is invertible is equivalent to proving that $(I - P + M)x = 0 \implies x = 0$. To do this multiply $(I - P + M)x = 0$ by on the left by π and use that the only solutions of $Px = x$ are of the form $x = c(1 \cdots, 1)^T$.

- Add this to your simulation algorithm for Markov chain of Exercise 1.8

▶ Solution:

Exercise 2.2

1. A transition matrix for a Markov chain is called doubly stochastic if for any $j \in S$ we have

$$\sum_{i \in S} P(i,j) = 1$$

   Show that the uniform distribution $\pi(j) = \frac{1}{|S|}$ is a stationary distribution.

2. Let S_n be the sum of n independent rolls of a fair dice. Find $\lim_{n\to\infty} P(S_n \text{ is a multiple of } 8)$. *Hint*: Define an appropriate Markov chain and apply part 1.

▶ Solution

Exercise 2.3 Consider the Markov chain with transition matrix P.

$$P = \begin{pmatrix} 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 3/4 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

1. Is the chain irreducible?
2. What is the period of the chain?
3. Let $\tau^{(1)}$ be the first return time to state 1. Compute directly $E[\tau(1)|X_0 = 1]$ by computing the pdf of $\tau(1)$.
4. Compute the stationary distribution π.

▶ Solution

Exercise 2.4 (A Markov chain for card shuffling) Suppose you have a deck of 52 cards. You shuffle your deck of cards by picking the top card of the deck and insert it uniformly at random in the deck.

1. What is the state space?
2. Show that Markov chain defined in this way is irreducible and aperiodic.
3. What is the stationary distribution.
4. If you shuffle the deck of cards every second. What is the average time (in years) until the deck returns to the original order?
5. Starting with an arbitrary deck consider the card at the bottom of the deck. What happens to that particular card after one shuffle? after n shuffle?
6. Suppose that at time n there are k cards under the original bottom card. Show (by induction on n) that each the possible $k!$ ordering of the cards is equally likely.
7. Show that if τ is the random time at which the card originally at the bottom reach the top of the deck then at time $\sigma = \tau + 1$ the cards are uniformly distributed on the set of all permutations. That is the system reaches reaches the stationary distribution at the (random) time $\tau + 1$.
8. Compute $E[\sigma]$ and $\text{Var}(\sigma)$. *Hint*: Coupon collector problem.

▶ Solution

Exercise 2.5 (More on the Law of Large numbers) Suppose X_n is irreducible so that the strong Law of Large numbers holds.

1. Show that for any $f : S \to \mathbb{R}$ (think of f as a column vector) such that $\pi f = \sum_{j \in S} f(j)\pi(j) < \infty$ we have

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} f(X_k) = \pi f = \sum_{j \in S} \pi(j) f(j)$$

2. Suppose that you observe a sample of an (irreducible) Markov chain $X_0, X_2, \cdots, X_{N-1}$ of length N (where N sufficiently large). Build an estimator for $P(i,j)$ based on the sample $X_0, X_2, \cdots, X_{N-1}$.
   Hint: Use part 1. and the result in Exercise 1.4 in Homework 4 for the Markov chain (X_n, X_{n+1})

▶ Solution

Exercise 2.6 (Chess)

1. A chess piece performs a random walk on a chessboard; at each step it is equally likely to make any one of the available moves. What is the mean return time of a corner square if the piece is a: (a) king? (b) queen? (c) bishop? (d) knight? (e) rook?

2. A rook and a bishop perform independent random walks with synchronous steps on a 4 x 4 chessboard (16 squares). If they start together at a corner, show that the expected number of steps until they meet again at the same corner is 448/3.

▶ Solution

Exercise 2.7 (Cesaro limit) A sequence of numbers $\{a_n\}_{n=1}^{\infty}$ *converges in the sense of Cesaro to* a if the sequence $b_n = \frac{a_1+\cdots+a_n}{n}$ converges to a.

1. Prove that if a sequence a_n converges to a then a_n also converges to a in the sense of Cesaro.
2. Show that he converse is generally not true.

▶ Details

Exercise 2.8 (Simulation of Markov chains, continued) Improve your code of Exercise 1.8 such that given a sufficently long path $X_0, X_1, X_2, \cdots, X_n$ of a Markov chain it will returns estimates of both the stationary distribution $\pi(i)$ and the transition probabilities $P(i,j)$. You will need part 2 of Exercise 2.5.

Exercise 2.9 (Simulation of Markov chains, irreducibility and period)

- Write a code which determines whether a Markov chain is irreducible or not. The input is the transition matrix P.
- Write a code which computes the period of the states for a given Markov chain. The input is the transition matrix P.

3 Transient behavior of Markov chains

3.1 Decomposition of state space

We now drop the assumption of irreducibility.

- The communication relation $i \leftrightsquigarrow j$ is an **equivalence relation**. We use the convention $P^0 = I$ (the identity matrix) and then also that $i \leftrightsquigarrow i$ (i communicates with itself $P^0(i,i) = 1$).

  1. It is *reflexive* : $i \leftrightsquigarrow i$.
  2. It is *symmetric*: $i \leftrightsquigarrow j$ implies $j \leftrightsquigarrow i$.
  3. It is *transitive*: $i \leftrightsquigarrow j$ and $j \leftrightsquigarrow l$ implies $i \leftrightsquigarrow l$.

- Using this equivalence relation we can decompose the state space S into mutually disjoint communication classes

$$S = C_1 \cup C_2 \cup \cdots \cup C_M \,.$$

Definition 3.1 (transient and closed classes)

- A class C is called **transient** if there exists $i \in C$ and $j \in S \setminus C$ with $i \rightsquigarrow j$.
  It means it is possible to exit the class C and never come back.

- A class C is called **closed** classes if it is not transient that is, for any pair $i \in C$ and $j \in S \setminus C$ we have $i \not\rightsquigarrow j$.
  Clearly it is impossible to exit a closed class.

The next result state if you start in a finite transient class you will always eventually exit it (this may not be true any more S is infinite).

Lemma 3.1 Suppose S is finite and $X_0 \in C$ where C is a transient class. Then X_n exits C after a finite time with probability 1. As a consequence for $i, j \in C$ we have

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

Proof. By irreducibility X_n can exit C starting from any $i \in C$ after a finite time. Since C is finite, there exists k and $\theta < 1$ such that have

$$P\{X_k \in C | X_0 = i\} \le \theta \ \text{ for all } i \in C\,.$$

Using the (strong) Markov property this implies that $P\{X_{nk} \in C | X_0 = i\} \le \theta^n$ and so so the probability to stay in transient class goes to 0 as time goes by. If i and j both belong to the transient class C this can be re-expressed as

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

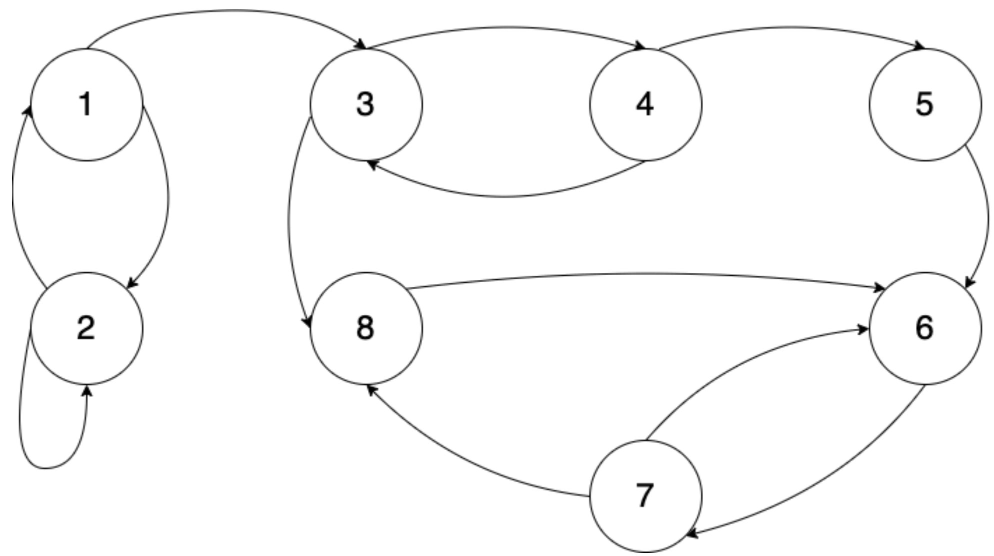
■.

3.2 Communication diagram

To figure out the class structure it is convenient to build a directed graph where the vertices are the state and each directed edges corespond to a pair (i,j) with $P(i,j)>0$.

For example

$$P = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{pmatrix} 0 & .4 & .6 & 0 & 0 & 0 & 0 & 0 \\ .3 & .7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 \\ 0 & 0 & .2 & 0 & .8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & .3 & 0 & .7 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$



There are 4 classes: $T_1=\{1,2\}, T_2=\{3,4\}, T_3=\{5\}$ are all transient and $C=\{6,7,8\}$ which is closed.

3.3 Decomposition of reducible Markov chains

Markov chain with closed classes $R_1, \cdots R_L$ and transient classes denoted by $T_1, \cdots T_K$ (set $T = T_1 \cup \cdots T_K$). After reordering the states the transition matrix has the block form

$$P = \begin{array}{c} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_L \\ T \end{array} \begin{pmatrix} P_1 & & & & & \\ & P_2 & & 0 & & \\ & & P_3 & & & \\ & 0 & & \ddots & & \\ & & & & P_L & \\ & & S & & & Q \end{pmatrix} \tag{3.1}$$

where P_l gives the transition probabilities within the class R_l, Q the transition within the transient classes and $S \neq 0$ the transition from the transient classes into the closed classes.

It is easy to see that P^n has the form

$$P^n = \begin{array}{c} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_L \\ T \end{array} \begin{pmatrix} P_1^n & & & & & \\ & P_2^n & & 0 & & \\ & & P_3^n & & & \\ & 0 & & \ddots & & \\ & & & & P_L^n & \\ & & S_n & & & Q^n \end{pmatrix}$$

for some matrix S_n.

Example: Consider the random walk with absorbing boundary conditions (see Example 1.5). There are three classes, 2 closed ones
$\{0\}$, $\{N\}$ and 1 transient $\{1, \cdots, N-1\}$ for example with $N=5$ we have

$$P = \begin{matrix} 0 \\ 5 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix}$$

To get a feel of what's going on we find (rounded to 4 digit precision)

$$P^{20} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.7942 & 0.1953 & 0.004 & 0 & 0.0065 & 0 \\ 0.5924 & 0.3907 & 0 & 0.0104 & 0 & 0.0065 \\ 0.3907 & 0.5924 & 0.0065 & 0 & 0.0104 & 0 \\ 0.1953 & 0.7942 & 0 & 0.0065 & 0 & 0.004 \end{pmatrix}, \quad P^{50} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.8 & 0.2 & 0 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 & 0 & 0 \end{pmatrix}$$

As expected starting in transient class the Markoveventually exit it to reach here of one the two closed class. From P^{50} one can read that starting in, say state 2 with probability $.6$ the Markov chain will end up at 0 and with probability $.4$ the Markov chain will end up in 5. We will learn how to compute these probability later on but we study first how long it takes to exit the closed class.

3.4 Absorption time

- Starting in a closed class R_l the Markov chain stays in R_l forever and its behavior is dictated entirely by the irreducible matrix P_l with state space R_l.
- Starting from a transient class the Markov chain will eventually exit the class T and is absorbed into some closed class. One basic question is: **What is the expected time until absorption**?
- We define the absorption time

$$\tau_{\text{abs}} = \min\{n \geq 0\,;\, X_n \notin T\}$$

Theorem 3.1 (Expected time until absorption) Let j be a transient state and let τ_{abs} to be the time until the Markov chain reaches some closed class. Then we have

$$E[\tau_{abs}|X_0 = j] = \sum_{i \in T} M(j,i)\,.$$

where

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

and Q is from the decomposition Equation 3.1.

Proof. From Lemma 3.1 we know that $Q^n(i,j) \to 0$ and so all eigenvalues of Q have absolute values strictly less than 1. Therefore $I - Q$ is an invertible matrix and we can define

$$M = (I-Q)^{-1} = I + Q + Q^2 + Q^3 + \cdots$$

The second equality is the geometric series for matrices which follows from the identity

$$(I + Q + \cdots Q^n)(1-Q) = I - Q^{n+1} \,.$$

To give a probabilistic interpretation to the matrix M we introduce the random variable

$$Y(i) = \sum_{n=0}^{\infty} 1_{\{X_n = i\}} \quad = \text{ total number of visits to state } i$$

If i is transient by Lemma 3.1 $Y(i) < \infty$ with probability 1. If j is another transient state we have

$$E[Y(i) \,|\, X_0 = j] = E\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n = i\}} \,|\, X_0 = j\right] = \sum_{n=0}^{\infty} P\left\{X_n = i \,|\, X_0 = j\right\} = \sum_{n=0}^{\infty} Q^n(i,j) = M(i,j)$$

That is $M(j,i)$ is simply the expected number of visits to i if $X_0 = j$ and thus, summing over all the transient states we obtain

$$E[\tau_{abs} | X_0 = j] = \sum_{i \in T} M(j,i) \,. \quad \blacksquare$$

3.5 Hitting time in irreducible Markov chain

- Suppose X_n is irreducible and we are interested in computing the expected time to reach one state from another state that is $E[\tau(i)|X_0 = j]$ for $j \neq i$.
- If $i = j$ then from Theorem 2.6 that $E[\tau(i)|X_0 = i] = \pi(i)^{-1}$ where π is the stationary distribution.
- If $i \neq j$ we use the following idea. Relabel the states so that the first state is i and turn i into an absorbing state

$$P = \begin{pmatrix} P(i,i) & R \\ S & Q \end{pmatrix} \quad \longrightarrow \quad \widetilde{P} = \begin{pmatrix} 1 & 0 \\ S & Q \end{pmatrix} \tag{3.2}$$

The fact that P is irreducible implies that $S \setminus \{i\}$ is a transient class for the modified transition matrix $\widetilde{P}$. The return time $\tau(i)$ starting from $i \neq j$ for the Markov chain with transition P is exactly the same as the absorption time for the Markov chain with transtion $\widetilde{P}$. Indeed the hitting time does not depend on the submatrix R. Therefore from Theorem 3.1 we obtain immediately

Theorem 3.2 (Expected hitting time in irreducible Markov chain) Let X_n be an irreducible Markov chain. For $i \neq j$ we

$$E[\tau(i)|X_0 = j] = \sum_{l \in T} M(j,l)\,.$$

where $M = (I - Q)^{-1}$ and Q is given in Equation 3.2 and is obtained by deleting the i^{th} row and i^{th} column from P.

3.6 Examples

Example(continued): Random walk with absorbing boundary conditions on $\{0, 1, \cdots, 5\}$. We get

$$Q = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 \end{pmatrix} \quad \Longrightarrow \quad M = (I-Q)^{-1} = \begin{pmatrix} 1.6 & 1.2 & 0.8 & 0.4 \\ 1.2 & 2.4 & 1.6 & 0.8 \\ 0.8 & 1.6 & 2.4 & 1.2 \\ 0.4 & .8 & 1.2 & 1.6 \end{pmatrix}$$

and thus the expected times until absorption are 4 for states 1 and 4 and 6 for states 2 and 3.

Example: Random walk with reflecting boundary conditions (see Example 1.5) with $N = 5$ and we compute $E[\tau(0)|X_0 = i]$. The stationary distribution is $\pi = \left(\frac{1}{10}, \frac{2}{10}, \frac{2}{10}, \frac{2}{10}, \frac{2}{10}, \frac{1}{10}\right)$ so $E[\tau(0)|X_0 = 0] = 10$. For $i \neq 0$ we delete the first row and column from P and have

$$Q = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad M = (I-Q)^{-1} = \begin{pmatrix} 2 & 2 & 2 & 2 & 1 \\ 2 & 4 & 4 & 4 & 2 \\ 2 & 4 & 6 & 6 & 3 \\ 2 & 4 & 6 & 8 & 4 \\ 2 & 4 & 6 & 8 & 5 \end{pmatrix}$$

and so the expected return times to 1 are $10, 9, 16, 21, 24, 25$ respectively.

3.7 Absorption probabilities

- If $X_0 = i \in T$ belongs to some transient class and if there are more than one closed classes, say $R_1, R_2, \cdots, R_L$ then the Markov may be absorbed in distinct closed class and so we wisht to compute the probabilities

$$P\{X_n \text{ reaches class } R_l \,|\, X_0 = i\}$$

- Without loss of generality we can assume that each closed class is an absorbing state $r_1, \cdots r_L$ (we can always collapse a closed class into a absorbing state since it does not matter which state in the closed class we first visit). We denote the transient states by $t_1, \cdots, t_M$ and upon reordering the state the transition matrix has the form

$$P = \begin{matrix} r_1 \\ \vdots \\ r_L \\ t_1 \\ \vdots \\ t_M \end{matrix} \begin{pmatrix} I & 0 \\ & \\ S & Q \end{pmatrix} \tag{3.3}$$

- We wish to compute

$$A(t_i, r_j) = P\{X_n \text{ eventually reaches } r_j \,|\, X_0 = t_i\}\,.$$

and it will be convenient to set $A(r_l, r_l) = 1$ and $A(r_k, r_l) = 0$ if $k \neq l$.

Theorem 3.3 (Absorption probabilities in closed classes) For a Markov chain with transition matrix Equation 3.3 where the states $t_1, \cdots t_M$ are transient we have

$$P\{X_n \text{ reaches } r_j \mid X_0 = t_i\} = A(t_i, r_j) \quad \text{with} \quad A = (I - Q)^{-1} S$$

Proof. We condition on the first step of the Markov chain to obtain

$$\begin{aligned}
A(t_i, r_j) &= P\left\{X_n = r_j \text{ eventually } | X_0 = t_i\right\} \\
&= \sum_{k \in S} P\left\{X_n = r_j \text{ eventually}, X_1 = k | X_0 = t_i\right\} \\
&= \sum_{k \in S} P\left\{X_1 = k | X_0 = t_i\right\} P\left\{X_n = r_j \text{ eventually } | X_1 = k\right\} \\
&= \sum_{k \in S} P(t_i, k) A(k, r_j) \; = \; P(t_i, r_l) + \sum_{t_k} P(t_i, t_k) A(t_k, r_j) \,.
\end{aligned}$$

If A be the $L \times M$ matrix with entries $A(t_i, r_l)$, then this can be written in matrix form as

$$A \;=\; S + QA$$

or

$$A \;=\; (I - Q)^{-1} S \;=\; MS \,.$$

Continuing with the random walk example with absorbing boundary conditions, we get

$$A = MS = \begin{pmatrix} 1.6 & 1.2 & 0.8 & 0.4 \\ 1.2 & 2.4 & 1.6 & 0.8 \\ 0.8 & 1.6 & 2.4 & 1.2 \\ 0.4 & .8 & 1.2 & 1.6 \end{pmatrix} \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix} = \begin{pmatrix} .8 & .2 \\ .6 & .4 \\ .4 & .6 \\ .2 & .8 \end{pmatrix}$$

For example from state 2 the probability to be absorbed in 0 is $.6$, and so on....

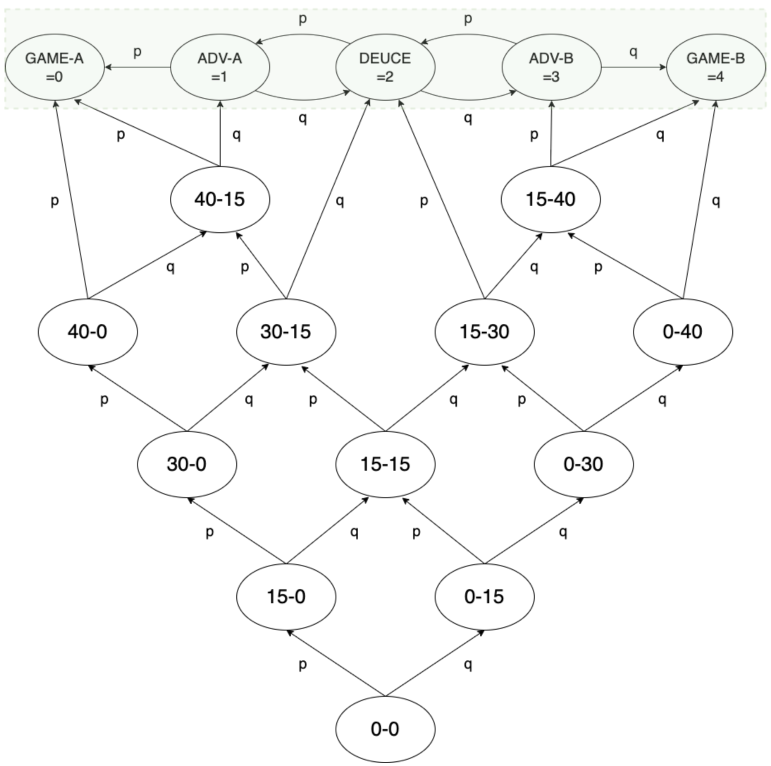
Remark You can apply similar ideas to irreducible Markov chain to compute probabilities like

$$P(\tau(i) \leq \tau(j) | X_0 = k).$$

that is the probability to return to i before returning to j by transforming i and j into absorbing states. See some examples in that spirit in the homework.

3.8 Tennis game

We model tennis as a Markov chain: For each point player A will win with probability p and player B will be win with probability $q = 1 - p$ and we assume all points are independent. Winning a game leads to the Markov chain with following communication diagram. This problem illustrates the power of compunding. For example my fellow countryman Roger Federer won 54% of all points played during his career. But he won 80% all of his game.



Tennis communication diagram

- The Markov chain "moves up" until it reaches one of the 5 states on the top (which we relabel $0, 1, 2, 3, 4$) and the chains stays on those 5 states forever and performs a random walks with absorbing boundary conditions.
- Want the probability that A wins, of course starting from the initial conditions $0 - 0$. Compute this probability by successive conditioning.
- Consider the events

$$D_i = \{X_n \text{ reaches the top row in state } i\}$$

- We compute $P(D_i)$ simply by enumerating all the paths leading up to state i starting form the initial state.

- We find then

$$P(D_0) = p^4 + 4qp^4\,, \quad P(D_1) = 4p^3q^2\,, \quad P(D_2) = 6p^2q^2\,, \quad P(D_3) = 4p^2q^3\,, \quad P(D_4) = q^4 + 4pq^4$$

- Conditioning on the B_i's: $P(A) = \sum_{i=0}^{4} P(A|D_i)P(D_i)$

- Obviously we have $P(A|D_0) = 1$ and $P(A|D_4) = 0$. By conditioning on the first step we find the system of equations

$$\begin{aligned} P(A|D_1) &= p + qP(A|B_2) \\ P(A|D_2) &= pP(A|D_1) + qP(A|D_3) \\ P(A|D_3) &= pP(A|D_2) \end{aligned} \quad \Longrightarrow \quad \begin{aligned} P(A|D_1) &= \frac{p(1-pq)}{1-2pq} \\ P(A|D_2) &= \frac{p^2}{1-2pq} \\ P(A|D_3) &= \frac{p^3}{1-2pq} \end{aligned}$$

- Putting everyting together

3.9 Gambler's ruin

- Solving gives the **Gambler's ruin probabilities**

$$\alpha_j = \frac{1-\left(\frac{1-p}{p}\right)^j}{1-\left(\frac{1-p}{p}\right)^N} \qquad \text{Gambler's ruin for } q \neq p$$

- If $p=\frac{1}{2}$ the quadratic equation Equation 3.4 has a double root 1 but we note that $\alpha_j = j$ is a second linearly independent solution of $\frac{1}{2}\alpha_{j+1} - \alpha_j + \frac{1}{2}\alpha_{j-1} = 0$. Solving the equation $\alpha_j = C_1 + C_2 j$ with the boundary conditions gives

$$\alpha(j) = \frac{j}{N} \qquad \text{Gambler's ruin for } p = \frac{1}{2}$$

- **How bad does it get?**:For example if $p=\frac{244}{495}$ and you start with a fortune a 50 and want to double it, the probability to succeed is $\frac{1-\left(\frac{251}{244}\right)^{50}}{1-\left(\frac{251}{244}\right)^{100}} = .1955$, not so great. You could also be more risky and bet an amount of 10 in which case you probability to succeed i a much better $\frac{1-\left(\frac{251}{244}\right)^{5}}{1-\left(\frac{251}{244}\right)^{10}} = .4647$. Even better bet everything and your probability to win is $p = .4929$. (In casino boldness pays, or loses less).

- **Limiting cases** To get a better handle on the gambler's ruin formula we slightly rephrase the problem:
  - We start at 0.
  - We stop whenever we reach W (W is the desired gain) 0r when we reach $-L$ (L is the acceptable losss).
  - The state are now $j \in \{-L, -L+1, \cdots, \cdots, W-1, W\}$

  We get

$$P(-L, W) \equiv P\left(\text{Reach} W \text{ before reaching } -L \text{ starting from } 0\right) = \frac{1 - \left(\frac{1-p}{p}\right)^{L}}{1 - \left(\frac{1-p}{p}\right)^{L+W}} .$$

- The limit $L \to \infty$ describes a player with infinite wealth:

$$P(-\infty, L) = \begin{cases} 1 & \text{if } p > \frac{1}{2} \\ \left(\frac{p}{1-p}\right)^{W} & \text{if } p < \frac{1}{2} \end{cases}$$

  Even with infinite wealth it is exponentially hard to win W!

- The limit $W \to \infty$ describes a player with fortune L who does not stops unless they lose.

$$P(-L, \infty) = \begin{cases} 1 - \left(\frac{1-p}{p}\right)^{L} & \text{if } p > \frac{1}{2} \\ 0 & \text{if } p < \frac{1}{2} \end{cases}$$

  The probability to play forever is, unsurprisingly, 0.

3.10 Exercises

Exercise 3.1 Consider a Markov chain with state space $\{0, \cdots, 5\}$ and transition matrix

$$P = \begin{pmatrix} .5 & .5 & 0 & 0 & 0 & 0 \\ .3 & .7 & 0 & 0 & 0 & 0 \\ 0 & 0 & .1 & 0 & .9 & 0 \\ .25 & .25 & 0 & 0 & .25 & .25 \\ 0 & 0 & .7 & 0 & .3 & 0 \\ 0 & .2 & 0 & .2 & .2 & .4 \end{pmatrix}.$$

1. What are the communication classes. Which ones are closed and which ones are transient?
2. Suppose $X_0 = 5$. What is the probability that X_n visits the state 1 before the state 4?
3. Suppose $X_0 = 5$. Compute $\lim_{n\to\infty} P^n(5, j)$ for all j?

▶ Solution

Exercise 3.2 Suppose we flip a fair coin repeatedly until we have flipped four consecutive heads. What is the expected number of flips that are needed?
Hint: Build up a suitable Markov chain with state space $\{0, 1, 2, 3, 4\}$

▶ Solution

Exercise 3.3 Consider the Markov chain with state space $\{0, \cdots, 6\}$ and transition matrix

$$P = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

1. Is this chain irreducible? Aperiodic?
2. Suppose the chain starts in state 1. What is the probability it reaches state 6 before reaching state 0?
3. Suppose the chain starts in state 3. What is the expected number of steps until it reaches 3 again?
4. Suppose the chain starts in state 0. What is the expected number of steps until it reaches state 6?

Exercise 3.4 (Random walk on the complete graph) The complete graph with vertex sets $\{1, \cdots, N\}$ is the graph such that any two vertex are joined by an edge. Let X_n be the simple random walk on this graph.

1. Let $\tau(1)$ be the first time the chain returns to state 1. Compute the p.d.f of $\tau(1)$ conditioned on $X_0 = 1$. Use this to compute $E[\tau(1)|X_0 = 1]$.
2. Compute $E[\tau(1)|X_0 = 2]$
3. Find the expected number of steps T_N until every one of the N state has been visited at least once.
   Hint: Let T_k be the time until k distinct states have been visited at least once. Compute first $E[T_k - T_{k-1}]$.

▶ Solution

Exercise 3.5 Every night Bob and Catherine go out to one of three bar A, B, or C. For each of Bob and Catherine their visits to bars are independent of each other and are governed by Markov chains with the same transition probabilities.

$$P = \begin{matrix} A \\ B \\ C \end{matrix} \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

Let us denote X_n and Y_n the bar visited by Bob and Catherine respectively on day n and let us suppose that $X_0 = A$ and $Y_0 = C$ and let T denote the first day when Bob and Catherine are in the same bar:

$$T = \min\{n\,;\, X_n = Y_n\}$$

1. Find $E[T]$.
2. What is the probability they first meet in bar C, i.e. compute $P(X_T = C)$.
3. In the long run what is the proportion of time they spend in the same bar.

Hint: You should consider the 9 state Markov chain $Z_n = (X_n, Y_n)$.

▶ Solution

Exercise 3.6 You can use absorbing state to compute *tabu probabilities*, i.e. probabilities of reaching some given state while avoiding some other states. For example the weather in the island of H is either 1=sunny, 2=cloudy, or 3=rainy and evolve according to a Markov chain with transition probabilties

$$P = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \begin{pmatrix} 0.70 & 0.10 & 0.20 \\ 0.50 & 0.25 & 0.25 \\ 0.40 & 0.30 & 0.30 \end{pmatrix}$$

1. Compute the probability that there is no rain in the next five days given that it is sunny today.
   Hint Consider the transition matrix $\tilde{P}$ where 3 is transformed into an absorbing state.
2. Compute the probability that there is no rain in the next five days given that it is rainy today.
   Hint: Condition first on tomorrow's weather and use $\tilde{P}$

Exercise 3.7 (Free casino money) You highschool friend W. B. now owns a casino in Macao. He is so happy to see you again that he gives you infinite credit at a table of craps to play the "pass line bet" which is an even money bet with probability of winning equal to 244/495. The maximal allowed bet is \$ 1,000. What is the expected size of the present your friend is giving you?

Exercise 3.8 (Roger Federer) In his commencement speech in Dartmouth college the swiss tennisman Roger Federer stated "I won almost 80% of singles matches… But I only won 54% of points. Even top ranked tennis players win barely more than half the points they play." (see the speech at here). Continue the analysis of tennis we started in Section 3.8 where we computed the probability to win a "game". A tennis match is played in five sets (best of five, the player who first wins three sets wins the match). A set is won when a player wins 6 games and two games ahead. If the scores reaches 6-5 then there are 2 possibilites. In the tie-break method when a score of 6-5 occurs then a plyer can win 7-5 or of if it reaches 6-6 a tie-break ensue. In the advantage set method games are played until until one player is to games ahead (Famously in in 2010 a set see here ended with the score 70–68 and lasted more than 6 hours. Details for the scoring rules are in https://en.wikipedia.org/wiki/Tennis_scoring_system.

Hint: Using the negative binomial distribution will be helpful.

Is the statement of Roger Federer consistent with our model?

▶ Solution

Exercise 3.9 (Computer exercise)

- Write a code which take a Markov chain and returns in canonical form.
- Add to the previous code the computation of the time until absorption and the absorption probabilities.

▶ Solution

4 Markov chains with countable state space

4.1 Examples

We introduce some basic examples of Markov chains on a countable state space

Example 4.1 (Random walk on the non-negative integers) Let us consider a random walk on the set of nonnegative integers with partially reflecting boundary conditions at 0. The transition probabilities are given by

$$\begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{pmatrix} q & p & 0 & 0 & 0 & \dots \\ q & 0 & p & 0 & 0 & \dots \\ 0 & q & 0 & p & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \end{pmatrix} \qquad \text{with } q = 1-p$$

Example 4.2 (Discrete-time queueing model) At a service station (think of a cash register), during each time period there is a probability p that an additional customer comes in the queue. The first person in the queue is being served and during each time period there is a probability q that this person exits the queue.

We denote by X_n the number of people in the queue (either in being served or waiting in line). The state space is $S = \{0, 1, 2, 3, \cdots\}$ and the transition probabilities are given by

$$P = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{pmatrix} 1-p & p & 0 & 0 & \dots \\ q(1-p) & qp + (1-p)(1-q) & p(1-q) & 0 & \dots \\ 0 & q(1-p) & qp + (1-p)(1-q) & p(1-q) & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

Example 4.3 (Repair shop) A repair shop is able to repair one item on any single day. On day n Z_n items break down and are brought for repair to the repair shop and we assume that Z_n are IID random variables with pdf $P\{Z_n = k\} = a_k$ for $k = 0, 1, 2, \cdots$. If X_n denotes the number of item in the shop waiting to be repaired we have

$$X_{n+1} = \max\{(X_n - 1), 0\} + Z_n$$

The state space is $S = \{0, 1, 2, 3, \cdots\}$ and the transition probabilities are

$$P = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \qquad \text{with } \sum_{k=0}^{\infty} a_k = 1$$

Example 4.4 (Success run chain) Imagine a player taking a series of bets labelled $0, 1, 2, \cdots$. The probability to win bet j is p_j. If the player wins bet j they move up to bet $j+1$ but if they lose they move back to bet 0. If X_n denotes the number of successive winning bets then X_n has state space $S = \{0, 1, 2, 3, \cdots\}$ and transition probabilities

Example 4.5 (Simple d-dimensional random walk) The state space S of the Markov chain is the d-dimensional lattice $\mathbb{Z}^d$. We denote by $\mathbf{e}_i, i = 1, \cdots, d$ the standard orthonormal basis in $\mathbb{R}^d$. We view $\mathbb{Z}^d$ as the vertex set of a graph and any point $\mathbf{x} = (\mathbf{x_1}, \cdots, \mathbf{x_d})$ is connected by edges to $2d$ neighbors $\mathbf{x} \pm \mathbf{e}_i$. For the simple random walk we have

$$p(\mathbf{x}, \mathbf{x} \pm \mathbf{e}_i) = \frac{1}{2d}$$

and all the others $p(\mathbf{x}, \mathbf{y}) = 0$.

Example 4.6 (Branching process) The branching process, also known as the *Galton-Watson process* model the evolution over time of populations.
In a unit of time every individual in a population dies and leave behind a random number of descendents.
To describe the Markov chain we will use IID random variables $Z_n^{(k)}$ indexed by $n = 0, 1, 2 \cdots$ and $k = 0, 1, 2, \cdots$.
The branching process is given by

$$X_{n+1} = \sum_{k=1}^{X_n} Z_n^{(k)}$$

which simply says that the each of X_n individuals in the population at time n has a random number $Z_n^{(k)}$ of descendents. It is not convenient to write down the transition probabilities but we will study this process later using its moment generating function.

4.2 Transience/recurrence dichotomy

In Markov chains on a countable state space a new phenomenon occurs compared to finite state space. Suppose the the Markov chain is irreducible (or starts in a closed class), from a state i the Markov chain will return to i with positive probability but it is also possible that the Markov chain does not return to i and "wander away to infinity".

We introduce the corresponding definitions of transience and recurrence of a state. Recall that the return time to state i is given by

$$\tau(i) = \min\{n \geq 1\,;\, X_n = i\}\,. \quad \text{(return time)}$$

Definition 4.1 (recurrent and transient state)

- A **state i is recurrent** if if the Markov chain starting in i will eventually return to i with probability 1, i.e. if

$$P\left\{\tau(i) < \infty | X_0 = i\right\} = 1.$$

- A **state i is transient** if it is not recurrent, that is starting in i the Markov chain return to i with probability $q < 1$, i.e., if

$$P\left\{\tau(i) < \infty | X_0 = i\right\} = q < 1.$$

Recall also the random variable $Y(i)$ which counts the number of visits to state j:

$$Y(i) = \sum_{k=0}^{\infty} I_{\{X_k = i\}} \quad \text{with expectation} \quad E[Y(i)|X_0 = j] = \sum_{n=0}^{\infty} P^n(j,i)$$

Theorem 4.1 (Transience/recurrence dichotomy)

1. A state i is recurrent $\iff P\{Y(i) = \infty | X_0 = i\} = 1 \iff \sum_{n=0}^{\infty} P^n(i,i) = \infty$.

   Moreover if i is recurrent and $i \leftrightsquigarrow j$ then j is recurrent and we have

$$\sum_{n=0}^{\infty} P^n(i,j) = \infty \quad \text{and} \quad P\{\tau(j) < \infty | X_0 = i\} = 1$$

2. The state i is transient $\iff P\{Y(i) < \infty | X_0 = i\} = 1 \iff \sum_{n=0}^{\infty} P^n(i,i) < \infty$.

   Moreover if i is transient and $i \leftrightsquigarrow j$ then j is transient and we have

$$\sum_{n=0}^{\infty} P^n(i,j) < \infty \quad \text{and} \quad P\{\tau(j) < \infty | X_0 = i\} < 1\,.$$

Proof. If i is recurrent the Markov chain starting from i will return to i with probability 1, and then, by the (strong) Markov property, it will return a second time with probability 1 and therefore infinitely many time with probability 1. This means that $Y(i) = \infty$ almost surely and so $\sum_k P^k(i,i) = +\infty$.

If $i \leftrightsquigarrow j$ then then we can find time l and m such that $P^l(i,j) > 0$ and $P^m(j,i) > 0$ and so

$$\sum_{n=0}^{\infty} P^n(j,j) \ge \sum_{n=0}^{\infty} P^{n+l+m}(j,j) \ge P^m(j,i) \sum_{n=0}^{\infty} P^n(i,i) P^l(i,j) = \infty\,,$$

and j is recurrent. A similar argument shows that $\sum_{n=0}^{\infty} P^n(i,j) = \infty$.

It is a consequence of irreducibility that $P\{\tau(i) < \tau(j) | X_0 = j\} > 0$. (Argue by contraduction, if this probability were 0 by the Markov property, the chain would never visits i starting from j). As a consequence

$$0 \;=\; P\{\tau(j) = \infty | X_0 = j\} \;\ge\; P\{\tau(i) < \tau(j) | X_0 = j\} P\{\tau(j) = \infty | X_0 = i\}$$

and therefore $P\{\tau(j) < \infty | X_0 = i\} = 1$.

On the other hand, if i is transient, by the Markov property again, the random variable $Y(i)$ is a geometric random variable with success probability $q < 1$ which implies that $E[Y(i)] = \sum_k P^k(i,i) = \frac{1}{q} < \infty$. This implies all the other equivalence stated. ■

4.3 Transience/recurrence for the simple random walk.

We analyze the recurence and transience propeties for the simple random walk on $\mathbb{Z}^d$ (see Example 4.5).
As we will see this depends on the dimension. To prove recurrence transience here we compute/estimate directly

$$\sum_{n=0}^{\infty} P^n(0,0) = \sum_{n=0}^{\infty} P^{2n}(0,0)$$

since X_n is periodic with period 2.

$d=1$: To return to 0 in $2n$ steps the Markov chain must take exactly n steps to the left and n steps to the right and thus we have

$$P^{2n}(0,0) = \binom{2n}{n}\frac{1}{2^{2n}}$$

By Stirling's formula we have $n! \sim \sqrt{2\pi n}e^{-n}n^n$ where $a_n \sim b_n$ means that $\lim a_n/b_n = 1$. Thus we have

$$\binom{2n}{n}\frac{1}{2^{2n}} \sim \frac{1}{2^{2n}}\frac{\sqrt{2\pi 2n}e^{-2n}(2n)^{2n}}{2\pi n e^{-2n}n^{2n}} = \frac{1}{\sqrt{\pi n}}\,.$$

Recalling that if $a_n \sim b_n$ then $\sum a_n$ converges if and only if $\sum b_n$ converges we see that **the random walk in $d=1$ is recurrent**.

- $d = 2$: In dimension 2 to return to 0 in $2n$ steps the Markov chain must take exactly k steps to the left and k steps to the right and $n-k$ steps up and $n-k$ steps down. Therefore

$$P^{2n}(0,0) = \sum_{k=0}^{n} \frac{2n!}{k!k!(n-k)!(n-k)!}\frac{1}{4^{2n}} = \frac{1}{4^{2n}}\binom{2n}{n}\sum_{k=0}^{n}\binom{n}{k}\binom{n}{n-k}$$

Now we claim that $\sum_{k=0}^{n}\binom{n}{k}\binom{n}{n-k} = \binom{2n}{n}$ as can be seen by the counting the number of ways that a team of n can be formed out of n boys and n girls. Therefore

$$P^{2n}(0,0) = \binom{2n}{n}^2 \frac{1}{4^{2n}} \sim \frac{1}{\pi n}$$

and thus the simple random walk is recurrent for $d=2$ since $\sum \frac{1}{n}$ diverges.

- $d = 3$: Similarly as in 2 dimension 2 we find

$$P^{2n}(0,0) = \sum_{\substack{k,j \\ k+j\le n}} \frac{2n!}{j!j!k!k!(n-k-j)!(n-k-j)!}\frac{1}{6^{2n}} = \frac{1}{2^{2n}}\binom{2n}{n}\sum_{\substack{k,j \\ k+j\le n}}\left(\frac{1}{3^n}\frac{n!}{j!k!(n-j-k)!}\right)^2$$

To analyze this quantity we note that, by the multinomial theorem,

$$\sum_{k,j:k+j\le n}\frac{1}{3^n}\frac{n!}{j!k!(n-j-k)!} = 1$$

Moreover we have

$$\sum_i q_i = 1 \implies \sum_i q_i^2 \leq max_i q_i$$

and thus we only need to the maximum of $\frac{n!}{j!k!(n-j-k)!}$. If k_0, j_0 is the maximum then we must have for example

$$\frac{n!}{(j_0-1)!k_0!(n-j_0-k_0+1)!} \leq \frac{n!}{(j_0)!k_0!(n-j_0-k_0)!} \implies 2j_0 \leq n-k_0+1\,.$$

Repeating the same computation with $j_0 \to j_0+1, k_0 \to k_0-1, k_0 \to k_0+1$ gives the set of inequalities

$$n-j_0-1 \leq 2k_0 \leq n-j_0+1 \quad \text{and} \quad n-k_0-1 \leq 2j_0 \leq n-k_0+1$$

which implies that

$$\frac{n}{3}-1 \leq j_0, k_0 \leq \frac{n}{3}+1$$

i.e. j_0 and k_0 are of order $n/3$. Using Stirling's formula

$$P^{2n}(0,0) \leq \frac{1}{2^{2n}}\binom{2n}{n}\frac{1}{3^n}\frac{n!}{(n/3)!(n/3)!(n/3)!} \sim \frac{3\sqrt{3}}{2}\frac{1}{(\pi n)^{3/2}}$$

which shows that the random walk is transient in dimension 3.

4.4 Transience/recurrence for the succcess run chain

Continuing with Example 4.4 we consider the return time to state $0, \tau(0)$ whose pdf we can compute explicitly since to return to 0 the only possible paths are $0 \to 0, 0 \to 1 \to 0, 0 \to 1 \to 2 \to 0$, and so on. We set $u_n \equiv p_0 p_1 \cdots p_{n-1}$ and find

$$P(\tau(0) = k | X_0 = 0) = p_0 p_1 p_{k-2} q_{k-1} = p_0 p_1 p_{k-2}(1 - p_{k-1}) = u_{k-1} - u_k \,.$$

Therefore

$$P(\tau(0) \le n | X_0 = 0) = \sum_{k=1}^{n} P(\tau(0) = k | X_0 = 0) = (1 - u_0) + (u_0 - u_1) + \cdots + (u_{n-1} - u_n) = 1 - u_n$$

and

$$P(\tau(0) < \infty | X_0 = 0) = 1 - \lim_{n \to \infty} u_n$$

and we obtain that

$$\text{The success run chain is recurrent if and only if } \lim_{n \to \infty} u_n = \lim_{n \to \infty} p_0 \cdots p_{n-1} = 0$$

To have a better handle on this criterion we need a little result from analysis about infinite products.

Lemma 4.1 With $q_k = 1 - p_k$ and $u_n = \prod_{k=0}^{n-1} u_k$

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} \prod_{k=0}^{n-1} p_k = 0 \iff \sum_{k=0}^{\infty} q_k = \infty$$

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} \prod_{k=0}^{n-1} p_k > 0 \iff \sum_{k=0}^{\infty} q_k < \infty$$

Proof. We have

$\prod_{k=0}^{n-1} p_k > 0 \iff \infty > -\log(\prod_{k=0}^{n-1} p_k) = -\sum_{k=1}^{n} \log p_k = -\sum_{k=1}^{n} \log(1-q_k)$ Taking $n \to \infty$ shows that

$$\lim_{n\to\infty} \prod_{k=0}^{n-1} p_k > 0 \iff \sum_{k=1}^{\infty} \log(1-q_k) \text{ converges.}$$

For this to happen we must have $\lim_{k\to\infty} q_k = 0$. But since $\lim_{x\to 0} \log(1-x)/x = 1$ by L'Hospital rule, we have that $\sum_{k=1}^{n} \log(1-q_k)$ converges if and only if $\sum_{k=1}^{n} q_k$ converges. ■

For the chain to be transient q_k must go to 0 fast enough. If say $q_i = \frac{1}{i}$ then the chain is recurrent while if $q_i = \frac{1}{i^2}$ then it is transient.

4.5 Another criterion for transience

We add one more method to establish transience. For this pick reference state j and consider the hitting time to state j, $\sigma(j)$ (not the return time but we will play with both.)

$$\alpha(i) \;=\; P\left\{\sigma(j) < \infty | X_0 = i\right\}$$

By definition we have $\alpha(j) = 1$ since $\sigma(j) = 0$ if $X_0 = j$. If the chain is transient we must have

$$\alpha(i) < 1 \qquad \text{for } i \neq j\,.$$

since by Theorem 4.1 for $i \neq j$ we have $\alpha(i) = P\left\{\tau(j) < \infty | X_0 = i\right\} < 1$.

Let us derive an equation for $\alpha(i)$ by conditioning of the first step. For $i \neq j$

$$\begin{aligned}
\alpha(i) &= P\left(\sigma(j) < \infty | X_0 = i\right) = P\left(\tau(j) < \infty | X_0 = i\right) \\
&= \sum_{k \in S} P\left(\tau(j) < \infty | X_1 = k\right) P(i,k) = \sum_{k \in S} P\left(\sigma(j) < \infty | X_0 = k\right) P(i,k) \\
&= \sum_{k \in S} P(i,k)\alpha(k)
\end{aligned}$$

and thus $\alpha(i)$ satisfies the equation

$$P\alpha(i) = \alpha(i)\,, \qquad i \neq j\,.$$

Theorem 4.2 (A criterion for transience) An irreducible Markov chain X_n is transient if and only if for some state j_0 there exists a solution for the equation

$$P\alpha(i) = \alpha(i) \text{ for } i \neq j_0 \tag{4.1}$$

such that

$$\alpha(j_0) = 1 \quad \text{and} \quad 0 < \alpha(i) < 1 \quad \text{for } i \neq j_0 \tag{4.2}$$

Proof. We have already established the necessity. In order to show the sufficiency assume that we have found a solution for Equation 4.1 and Equation 4.2. Then for $i \neq j_0$ we have, using repeatedly the equation $P\alpha(i) = \alpha(i)$

$$\begin{aligned}
1 > \alpha(i) = P\alpha(i) &= P(i,j_0)\alpha(j_0) + \sum_{j \neq j_0} P(i,j)\alpha(j) = P(i,j_0) + \sum_{j \neq j_0} P(i,j)P\alpha(j) \\
&= P(i,j_0) + \sum_{j \neq j_0} P(i,j)P(j,j_0) + \sum_{j \neq j_0, k \neq j_0} P(i,j)P(j,k)\alpha(k) \\
&= P(i,j_0) + \sum_{j \neq j_0} P(i,j)P(j,j_0) + \sum_{j,k \neq j_0} P(i,j)P(j,k)P(k,j_0) + \cdots \\
&= P(\tau(j_0) = 1 | X_0 = i) + P(\tau(j_0) = 2 | X_0 = i) + P(\tau(j_0) = 3 | X_0 = i) + \cdots \\
&= P(\tau(j_0) < \infty | X_0 = i) \,.
\end{aligned}$$

which establishes transience. ■

4.6 Transience/recurrence for the RW on $\{0, 1, 2, \cdots\}$

Continuing with Example 4.1 we use Theorem 4.2. We pick 0 as the reference state and for $j \neq 0$ solve the equation

$$P\alpha(j) = P(j, j-1)\alpha(j-1) + P(j, j+1)\alpha(j+1) = (1-p)\alpha(j-1) + p\alpha(j+1) = \alpha(j)$$

whose solution is (like for the Gambler's ruin problem)

$$\alpha(j) = \begin{cases} C_1 + C_2\left(\frac{1-p}{p}\right)^j & \text{if } p \neq \frac{1}{2} \\ C_1 + C_2 j & \text{if } p = \frac{1}{2} \end{cases}.$$

Using that $\alpha(0) = 0$ we find

$$\alpha(j) = \begin{cases} (1-C_2) + C_2\left(\frac{1-p}{p}\right)^j & \text{if } p \neq \frac{1}{2} \\ (1-C_2) + C_2 j & \text{if } p = \frac{1}{2} \end{cases}.$$

and we see the condition $0 < \alpha(i) < 1$ is possible only if $(1-p)/p < 1$ (that is $p > 1/2$) and by choosing $C_2 = 1$. Thus we conclude

$$\text{The random walk on } \{0, 1, 2, \cdots\} \text{ is } \begin{cases} \text{transient for } p > \frac{1}{2} \\ \text{recurrent for } p \leq \frac{1}{2} \end{cases}.$$

5 Positive recurrent Markov chains

5.1 Positive recurrence versus null recurrence

A finite state irreducible Markov chain is always recurrent, $P(\tau(i) < \infty | X_0 = j) = 1$ and we have proved Kac's formula for the invariant measure $\pi(i) = E[\tau(i)|X_0 = i]^{-1}$, that is the random variable $\tau(i)$ has finite expectation. For a countable state space it is possible for a Markov chain to be recurrent but also that $\tau(i)$ does not have finite expectation. This motivates the following definitions.

Definition 5.1 (Postive and Null recurrence)

- A state i is positive recurrent if $E[\tau(i)|X_0 = i] < \infty$
- A state i is null recurrent if it is recurrent but not positive recurrent

We first investigate the relation between recurrence and existence of invariant measures. We first show that if one state j is positive recurrent then there exists a stationary distribution. The basic idea is to decompose any path of the Markov chain into successive visits to the state j. To build up our intuition if a stationary distribution were to exists it should measure the amount of time spent in state i and to measure this we introduce

$$\mu(i) = E\left[\sum_{n=0}^{\tau(j)-1} \mathbf{1}_{\{X_n = i\}} | X_0 = j\right] \quad = \text{number of visits to } i \text{ between two successive visits to } j.$$

Note that $\mu(j) = 1$ since $X_0 = j$, and if j is positive recurrent

$$\sum_i \mu(i) = \sum_{i \in S} E\left[\sum_{n=0}^{\tau(j)-1} \mathbf{1}_{\{X_n = i\}} | X_0 = j\right] = E\left[\tau(j) \mid X_0 = j\right] < \infty$$

and thus

$$\pi(i) = \frac{\mu(i)}{E[\tau(i)|X_0 = i]}$$

is a probability distribution.

Theorem 5.1 For a recurrent irreducible Markov chain X_n and a fixed state j, $\mu(i) = E[\sum_{n=0}^{\tau(j)-1} \mathbf{1}_{\{X_n=i\}}|X_0 = j]$ is stationary in the sense that

$$\mu P = \mu$$

If the state j is positive recurrent then μ can be normalized to a stationary distribution π.

Proof. The chain visits j at time 0 and then only again at time $\tau(j)$ and thus we have the two formulas

$$\begin{aligned} \mu(i) = E\left[\sum_{n=0}^{\tau(j)-1} \mathbf{1}_{\{X_n=i\}}|X_0 = j\right] &= \sum_{n=0}^{\infty} P(X_n = i, \tau(j) > n|X_0 = j) \\ = E\left[\sum_{n=1}^{\tau(j)} \mathbf{1}_{\{X_n=i\}}|X_0 = j\right] &= \sum_{n=1}^{\infty} P(X_n = i, \tau(j) \geq n|X_0 = j) \end{aligned}$$

We have then, by *conditioning on the last step*, and using $\mu(j) = 1$

$$
\begin{aligned}
\mu(i) &= \sum_{n=1}^{\infty} P(X_n = i, \tau(j) \geq n | X_0 = j) = P(j,i) + \sum_{n=2}^{\infty} P(X_n = i, \tau(j) \geq n | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{n=2}^{\infty} P(X_n = i, X_{n-1} = k, \tau(j) \geq n | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{n=2}^{\infty} P(k,i) P(X_{n-1} = k, \tau(j) \geq n | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{n=2}^{\infty} P(k,i) P(X_{n-1} = k, \tau(j) > n-1 | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{m=1}^{\infty} P(k,i) P(X_m = k, \tau(j) > m | X_0 = j) \\
&= \mu(j) P(j,i) + \sum_{k \neq j} E\left[\sum_{m=1}^{\tau(j)-1} \mathbf{1}_{\{X_m = k\}} | X_0 = j \right] P(k,j) \\
&= \mu(j) P(j,i) + \sum_{k \neq j} \mu(k) P(k,i) \; = \; \sum_{k} \mu(k) P(k,i)
\end{aligned}
$$

which proves the invariance of μ. If the chain is positive recurrent we have already seen that μ is normalizable. $\blacksquare$.

5.2 Stationarity and irreducibility implies positive recurrence

Theorem 5.2 Assume the irreducible Markov chain has a stationary distribution $\pi(i)$ then $\pi(i) > 0$ for any i and we we have Kac's formula $\pi(i) = E\left[\tau(i)|X_0 = i\right]^{-1}$. In particular all states are positive recurrent and the stationary distribution is unique.

Proof. Let us assume that π is invariant. We first show that the chain must be recurrent. If the chain were transient then we would have $P^n(i,j) \to 0$ as $n \to \infty$ and so by dominated convergence

$$\pi(i) = \sum_j \pi(j) P^n(j,i) \to 0 \text{ as } n \to \infty \,.$$

which is impossible.

The fact that $\pi(i) > 0$ for all $i \in S$ is proved as for finite state space see the argument in Theorem 2.2.

To prove positive recurrence we use a clever argument involving the time reversed chain (more on time reversal in Chapter 8 and in the exercises). Consider the Markov chain with transition matrix $Q(i,j) = \frac{\pi(j)P(j,i)}{\pi(i)}$. It is easy to verify that $Q(i,j)$ is a transition matrix and that π is stationary for Q, $\pi Q = \pi$. We denote by Y_n the Markov chain with transition matrix Q, since $\pi(i) > 0$ this Markov chain has the same communication structure as X_n and is irreducible since X_n is. By the previous argument Y_n must be recurrent.

Next we write

$$\pi(i)E[\tau(i)|X_0 = i] \;=\; \pi(i)\sum_{n=1}^{\infty} P(\tau(i) \geq n|X_0 = i)$$

The event $\{\tau(i) \geq n\}$ conditioned on $\{X_0 = i\}$ correspond to a sequence of states

$$i_0 = i, i_1, \cdots\cdots, i_{n-1}, i_n = j$$

where $i_1, \cdots, i_{n-1}$ cannot be equal to i and $i_n = j$ can be any state. Using repeatedly the relation $\pi(i)P(i,j) = \pi(j)Q(j,i)$ the probability of such event can be written as

$$\pi(i)P(i,i_1)\cdots P(i_{n-1},i_n) = \pi(j)Q(j,i_{n-1})\cdots Q(i_1,i)$$

For the Markov chain Y_n this correspond to a path starting in j and returning to i after exactly n steps. Therefore we find

$$\begin{aligned}\pi(i)E[\tau(i)|X_0 = i] = \pi(i)\sum_{n=1}^{\infty} P(\tau(i) \geq n|X_0 = i) &= \sum_{j \in S}\pi(j)\sum_{n=1}^{\infty} P(\tau(i) = n|Y_0 = j) \\ &= \sum_{j \in S}\pi(j)P(\tau(i) < \infty|Y_0 = j) = 1\,.\end{aligned}$$

where for the last equality we have used the recurrence of the time reversed chain. This shows Kac's formula which implies the uniqueness of the stationary distribution and that all states are positive recurrent. ∎

5.3 Ergodic theorem for countable Markov chains

Theorem 5.3 (Ergodic theorem for countable state space Markov chains) If X_n is irreducible and positive recurrent then there exists a unique stationary distribution π and for any initial distribution μ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k=j\}} = \pi(j) \,,$$

with probability 1. In particular $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu P^k(j) = \pi(j)$,. Moreover π we have the Kac's formula $\pi(j) = \frac{1}{E[\tau(j)|X_0=j]}$. Conversely if an irreducible Markov has a stationary distribution then it is positive recurrent.

Proof. We have actually already proved all of it. Positive recurrence implies the existence of the stationary distribution (Theorem 5.1) and Kac's formula is from Theorem 5.2 which implies uniqueness of the stationary distribution. We can now repeat the proof of Theorem 2.6 to show that if $X_0 = i$ with i arbitrary we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k=j\}} = \pi(j) \,. \tag{5.1}$$

The reader should verify that the proof of Theorem 2.6 only use positive recurrence and not the finiteness of the state space. Taking now expectation of Equation 5.1 and summing over initial condition we have $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mu P^k(j) = \pi(j)$. ■

5.4 Convergence to equilibrium via coupling

We continue our theoretical consideration by proving that if the chain is aperiodic then the distribution of X_n converges to $\pi(j)$.

Theorem 5.4 Suppose X_n is an irreducible positive recurrent aperiodic Markov chain. Then for any initial distribution μ we have

$$\lim_{n\to\infty} \mu P^n(j) \;=\; \pi(j)\,.$$

Proof. We will use a *coupling argument*: we take two independent copies X_n and Y_n of the Markov chain where X_n is starting in the initial distribution μ while Y_n is starting in the stationary distribution π.

The idea is to consider the coupling time

$$\sigma = \inf\{n \ge 1; X_n = Y_n\}\,.$$

At the (random) time σ, X_n and Y_n are in the same state and after that time X_n and Y_n must have the same distribution by the (strong) Markov property. But since Y_n is distributed according to π so must X_n be as well and thus, at the coupling time σ, X_n has reached stationarity.

Let us now consider the chain $Z_n = (X_n, Y_n)$ with transition probabilities

$$P(Z_{n+1} = (k,l) \mid Z_n = (i,j)) = P(i,k)P(j,l)$$

and stationary distribution $\pi(i,j) = \pi(i)\pi(j)$. Since X_n and Y_n are aperiodic, given states i, j, k, l we can find n_0 such that for every $n \ge n_0$ we have $P^n(i,k) > 0$ and $P^n(j,l) > 0$. This implies that Z_n is irreducible and thus, since a stationary measure exists, by Theorem 5.3 the chain Z_n is positive recurrent. Since the coupling time is the first time the Markov chain Z_n hits a state of the form (j,j), recurrence of Z_n implies that $P(\sigma < \infty) = 1$ and thus $P(\sigma > n) \to 0$.

To conclude

$$\begin{aligned}
|\mu P^n(j) - \pi(j)| &= |P(X_n = j) - P(Y_n = j)| \\
&\le |P(X_n = j, \sigma \le n) - P(Y_n = j, \sigma \le n)| + |P(X_n = j, \sigma > n) - P(Y_n = j, \sigma > n)| \\
&= |P(X_n = j, \sigma > n) - P(Y_n = j, \sigma > n)| \\
&= \left|E[(\mathbf{1}_{\{X_n = j\}} - \mathbf{1}_{\{Y_n = j\}})\mathbf{1}_{\{\sigma > n\}}]\right| \\
&\le E[\mathbf{1}_{\{\sigma > n\}}] = P(\sigma > n) \to 0 \text{ as } n \to \infty
\end{aligned}$$

and this prove the convergence. ■.

Remark The idea of coupling used in Theorem 5.4 is an instance of powerful idea. First we can use other coupling that the one used here and if we can bound the tail behavior of the coupling time then we can control the speed of convergence to the stationary distribution! We will exploit this idea later on.

5.5 Examples

Positive recurrence for the random walk on $\{0,1,2,\cdots\}$. Continuing with Example 4.1, we use Theorem 5.3 to establish positive recurrence by computing the stationary distribution. We find the equations

$$\pi(0)(1-p)+\pi(1)(1-p)=\pi(0) \quad \text{and} \quad \pi(j+1)(1-p)+\pi(j-1)p=\pi(j) \quad \text{for } j\ge 1$$

The second equation has the general solution

$$\pi(n)=C_1+C_2\left(\frac{p}{1-p}\right)^n \ (\text{if } p\neq \frac{1}{2}) \quad \text{and} \quad \pi(n)=C_1+C_2 n \ (\text{ if } p=\frac{1}{2})$$

and the first equation gives $\pi(1)=\frac{p}{1-p}\pi(0)$. For $p=\frac{1}{2}$ we cannot find a normalized solution. For $p<\frac{1}{2}$ we can choose $C_1=0$ and $\pi(n)=\left(\frac{p}{1-p}\right)^n\pi(0)$ can be normalized toi find $\pi(n)=\frac{1-2p}{1-p}\left(\frac{p}{1-p}\right)^n$. If $p>\frac{1}{2}$ we already know the Markov chain is transient and thus

$$\text{The random walk on } \{0,1,2,\cdots\} \text{ is } \begin{cases} \text{transient for } p>\frac{1}{2} \\ \text{null recurrent for } p=\frac{1}{2} \\ \text{positive recurrent for } p<\frac{1}{2} \end{cases} .$$

Positive recurrence for the success run chain. Continuing with Examples Example 4.4 we determine recurrence by solving $\pi P = \pi$. This gives the equations

$$\pi(0) = \pi(0)q_0 + \pi(1)q_1 + \cdots = \sum_{n=0}^{\infty} \pi(n) q_n$$
$$\pi(1) = \pi(0)p_0\,, \quad \pi(2) = \pi(1)p_1, \quad \cdots$$

From the second line we find $\pi(n) = \pi(0)p_0p_1\cdots p_{n-1}$ and inserting into the first equation

$$\pi(0) = \pi(0)\left[(1-p_0) + p_0(1-p_1) + p_0p_1(1-p_2) + \cdots\right] = \pi(0)\left[1 - \lim_{n\to\infty} p_0p_1\cdots p_{n-1}\right]$$

Recall that recurrence occurs provided $\lim_{n\to\infty}\prod_{j=0}^{n-1} p_j = 0$ and so if X_n is recurrent there exists a solution of $\pi P = \pi$ and we can normalize π provided

$$\sum_{n=1}^{\infty}\prod_{j=0}^{n-1} p_j < \infty$$

Therefore we obtain

$$\text{The success run chain is} \left\{ \begin{array}{l} \text{transient if } \lim_n \prod_{j=0}^{n-1} p_j > 0 \\ \text{recurrent if } \lim_n \prod_{j=0}^{n-1} p_j = 0 \\ \text{positive recurrent if } \sum_n \prod_{j=0}^{n-1} p_j < \infty \end{array} \right. .$$

5.6 Exercises

These exercises will use the material of both Section 4 and Section 5

Exercise 5.1 Consider the following queueing model. In any given time unit, there is probability q that an item being serviced is repaired and there is a probability p that new item is added to the queue. Denote X_n to be the numbers of items in the queue (being repaired or waiting to be repaired). The transition probabilities are given by

$$P = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{pmatrix} 1-p & p & 0 & 0 & \dots \\ q(1-p) & qp+(1-p)(1-q) & p(1-q) & 0 & \dots \\ 0 & q(1-p) & qp+(1-p)(1-q) & p(1-q) & \\ \vdots & & \ddots & \ddots & \ddots \end{pmatrix}$$

a. Determine when the Markov chain is transient, recurrent, or positive recurrent.

b. In the positive recurrent case compute the stationary distribution and the length of the queue in equilibrium.

c. In the transient case compute the probability $\alpha(j)$ of ever reaching 0 starting from j.

▶ Solution

Exercise 5.2 Let $p(k), k = 0, 1, 2, 3, \cdots$ be such that $\sum_{k=0}^{\infty} p(k) = 1$. Consider the Markov chain on the state space $S = \{0, 1, 2, 3, \cdots\}$ with transition probabilities

$$P = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{pmatrix} p(0) & p(1) & p(2) & p(3) & \dots & \\ 1 & 0 & 0 & 0 & \dots & \\ 0 & 1 & 0 & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

Starting from zero the Markov chain X_n visits state k with probability $p(k)$ and then falls back to 0 one step at a time.

Determine under which conditions on p_k is the Markov chain positive recurrent, null recurrent, transient? Compute the stationary distribution π in the positive recurrent case.
Hint: Study the distribution of the return time $\tau(0)$.

▶ Solution

Exercise 5.3 Consider the Markov chain with state space $\{0, 1, 2, \cdots\}$ with transition probabilities

$$P = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{pmatrix} 1-p & 0 & p & 0 & 0 & \dots \\ 1-p & 0 & 0 & p & 0 & \dots \\ 0 & (1-p) & 0 & 0 & p & \dots \\ \vdots & & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

For which values of p is the chain transient, recurrent, positive recurrent?

Exercise 5.4 For each of the following success run Markov chain with state space $\{0, 1, 2, \cdots\}$ determine if the chain is transient, recurrent, positive recurrent. (If it is positive recurrent compute the stationary distribution π).

a. $p(j, 0) = \frac{1}{j+2}, \quad p(j, j+1) = \frac{j+1}{j+2}$

b. $p(j, 0) = \frac{j+1}{j+2}, \quad p(j, j+1) = \frac{1}{j+2}$

c. $p(j, 0) = \frac{1}{j^2+2}, \quad p(j, j+1) = \frac{j^2+1}{j^2+2}$

Exercise 5.5 (Hitting times)

1. Consider the gambler's ruin problem again. The goal is to compute the duration of play, that is let τ be the stopping time given by

$$\tau = \inf\{n \geq 0, X_n \in \{0, N\}\}$$

Compute $\alpha(j) = E[\tau | X_0 = j]$.
To do this condition on the frist step to find a system of (inhomogeneous) linear equations for $\alpha(j)$ and solve them (you will need some help from linear ODEs)

2. Consider a positive recurrent random walk on $\{0, 1, 2, \cdots\}$. Using the same ideas as in 1. compute the expected hitting time to 0, $E[\tau(0) | X_0 = j]$.

▶ Solution

Exercise 5.6 In this problem we are interested in how long should we wait until we a series of n consecutive occurences in a sequence of independent trials. Suppose $B_1, B_2, \cdots$ are independent Bernoulli random variables random variables with $P(B_j = 1) = 1 - P(B_j = 0) = p$. Let N be a positive integer and let T_N be the first time that N consecutive 1 have appeared.

Consider the Markov chain with $X_0 = 0$ and X_n to be the number of consecutive 1 in the last run.

1. Explain why X_n is a Markov chain with state space $\{0, 1, 2, \cdots\}$ and give the transition probabilities.
2. Show that the chain is irreducible and positive recurrent and give the invariant probability.
3. Find $E(T_N)$ by writing an equation for $E(T_N)$ in terms of $E[T_{N-1}]$ and then solving the recursive equation.
4. Find $E(T_N)$ is a different way. Suppose the chain starts in state N, and let S_N be the the time until returning to state N and S_0 the time until the chain reaches state 0. Explain why

$$E[S_N] = E[S_0] + E[T_N]$$

   Find $E[S_0]$, and use part (b) to determine $E[T_N]$.

▶ Solution

6 Branching processes

6.1 Moment generating function

- For a random variable X taking values in $\{0, 1, 2, 3, \cdots\}$ the generating function of X is the function $\phi_X(s) = E[s^X]$, defined for $s \geq 0$. We have

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

  It is closely related to the moment generating function $E[e^{tX}]$ but the parametrization $s = e^t$ is more useful her.

- Elementary properties of $\phi_X(s)$:
  - $\phi_X(s)$ is an increasing function of s with $\phi_X(0) = P\{X = 0\}$ and $\phi_X(1) = 1$.
  - Differentiating gives

$$\phi'_X(s) = \sum_{k=1}^{\infty} k s^{k-1} P\{X = k\}, \qquad \phi''_X(s) = \sum_{k=2}^{\infty} k(k-1) s^{k-2} P\{X = k\}$$

    and thus $\phi'_X(1) = E[X]$.
    If $P\{X \geq 2\}$ then $\phi''_X(s) > 0$ and ϕ_X is strictly convex.
  - If $X_1, X_2, \cdots X_m$ are independent random variables then

$$\phi_{X_1+\cdots+X_m}(s) = \phi_{X_1}(s) \cdots \phi_{X_1}(s)$$

6.2 Branching process

- In a branching process individuals reproduce independently of each other.
- During one time period each indvidual dies and leaves k offsprings with probability $p(k)$ for $k = 0, 1, 2, \cdots$. We denote by X_n the total population after n time period.
- 0 is an absorbing state and correspond to the extinction of the population. We assume $p(0) > 0$ so the absorbing state can be reached from other states.
- The transition probabilities are not easy to write down explicitly. If $X_n = k$ then X_{n+1} will be the sum of the offsprings of the k indviduals. That is

$$P\{X_{n+1} = j | X_n = k\} = P\{Y_1 + \cdots + Y_k = j\}$$

  where $Y_1, \cdots, Y_k$ are IID random variables with pdf $p(k)$.

- Mean population $E[X_n]$ is easy to compute by conditioning. Denoting by $\mu = E[Y_1] = \sum_{n=0}^{\infty} np(n)$ the mean number of offsprings of the individuals we have

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

  and thus

$$E[X_n] = E[E[X_n | X_{n-1}]] = \mu E[X_{n-1}] = \cdots = \mu^n E[X_0]$$

- If $\mu < 1$ then $E[X_n] \to 0$ and so the population goes extinct since

$$P\{X_n \geq 1\} = \sum_{k=1}^{\infty} P\{X_n = k\} \leq \sum_{k=0}^{\infty} kP\{X_n = k\} = E[X_n] \to 0$$

and so $\lim_{n \to \infty} P\{X_n = 0\} = 1$.

- If $\mu = 1$ then the population stays constant but it could go extinct with probability 1 nonetheless (that is $P(X_n = 0$ goes to 1 but $E[X_n]$ is not small). If $\mu > 1$ then the population grows on average but it still could go extinct with some non-zero probability.

- To avoid trival cases we assume that $p(0) > 0$ (the population can go extinct) and $p(0) + p(1) < 1$ (the population can grow). We define the extinction probability

$$\begin{aligned} a_n(k) &= P\{X_n = 0 | X_0 = k\} \quad \text{this is increasing in } n \\ a(k) &= \lim_{n \to \infty} P\{X_n = 0 | X_0 = k\} = P\{\text{population goes extinct} | X_0 = k\} \end{aligned}$$

- Since for the population to go extinct all branches must go extinct, by independence, we have

$$a(k) = a(1)^k \quad \text{so we set } a \equiv a(1)$$

and we assume from now on that $X_0 = 1$. Note this formula is also correct for $k = 0$ since $a(0) = 1$.

- Equation for the extinction probability a. By conditioning on the first step

$$
\begin{aligned}
a &= P\left\{\text{population goes extinct} | X_0 = 1\right\} \\
&= \sum_{k=0}^{\infty} P\left\{\text{population goes extinct} | X_1 = k\right\} P\{X_1 = k | X_0 = 1\} \\
&= \sum_{k=0}^{\infty} a(k)p(k) \\
&= \sum_{k=0}^{\infty} a^k p(k) = \phi_Z(a)
\end{aligned}
$$

where $\phi_Z(s)$ is the generating function for the offspring distribution Z. So we have

$$\text{The extinction probability solves the fixed point equation } \phi_Z(a) = a\,.$$

- this is a nice illustration of why moment generating functions are useful!

- Compute now the generating function for the population X_n after n generations.
- If $X_0 = 1, \phi_{X_0}(s) = s$ and $\phi_{X_1}(s) = \phi_Z(s)$ since X_1 are the descendents of a single individiual. For $n \geq 2$

$$
\begin{aligned}
\phi_{X_n}(s) &= \sum_{k=0}^{\infty} P\left\{X_n = k\right\} s^k \\
&= \sum_{k=0}^{\infty} \left[\sum_{j=0}^{\infty} P\left\{X_n = k | X_1 = j\right\} P\left\{X_1 = j\right\}\right] s^k \\
&= \sum_{j=0}^{\infty} p(j) \sum_{k=0}^{\infty} P\left\{X_n = k | X_1 = j\right\} s^k
\end{aligned}
$$

Now note that X_n conditioned on $X_1 = j$ is the sum of j independent copies of X_n conditioned on $X_1 = 1$. Therefore the generating of function X_n conditioned on $X_1 = j$ is the generating function of X_n conditioned on $X_1 = 1$ to the j^{th} power. So we find

$$
\phi_{X_n}(s) = \sum_{j=0}^{\infty} p(j) [\phi_{X_{n-1}}(s)]^j = \phi_Z(\phi_{X_{n-1}}(s))
$$

and thus we get

$$
\phi_{X_n}(s) = \phi_Z^n(s) = \phi_Z(\phi_Z(\cdots(\phi_Z(s))))
$$

Using these computations we are ready to derive the main result

Theorem 6.1 Let Z be random variable describing the distribution of descendants of a single individual in a branching process and let us assume that $p(0) > 0$ and $p(0) + p(1) < 1$. Then if $X_0 = 1$ the *extinction probability is the smallest root of the equation* $\phi_Z(a) = a$.

- If $\mu \leq 1$ then $a = 1$ and the population eventually dies out.
- If $\mu > 1$ then the extinction probability $a < 1$ is the unique root of the equation $\phi_Z(s) = s$ with $0 < s < 1$.

Proof. We have aready established the extinction probability a is a root of $\phi_Z(s) = s$. But we also know that 1 is a root since $\phi_Z(1) = 1$ and the slope of ϕ_Z is equal to μ at $s = 1$. Since $p(0) + p(1) < 1$ then $\phi_Z(s)$ is strictly convex and thus $\phi_Z(s)$ has at most two roots. We have the following cases (illustrated in Figure 6.1 on the next slide).

1. If $\mu < 1$ the equation $\phi_Z(s) = s$ has two roots 1 and $s > 1$ and the extinction probability is 1.
2. If $\mu = 1$ the line s is a tangent to the curve $\phi_Z(s)$ at $s = 1$ and so $\phi_Z(s) = s$ has one roots 1 and the extinction probability is 1.
3. Finally if $\mu > 1$ the equation $\phi_Z(s) = s$ has two roots 1 and $a < 1$. Since $\phi_Z(0) = p(0) > 0$ the second root satisfies $a > 0$. To show that the smallest root is the extinction probability note that we have

$$a_n(1) = P\{X_n = 0 | X_0 = 1\} = \phi_Z^n(0)$$

By induction we show that $a_n(1) \leq a$. True for $n = 0$ since $a_0(1) = 0$. Assuming that $a_{n-1}(1) \leq a$ we have

$$a_n(1) = P\left\{X_n = 0 | X_0 = 1\right\} = \phi_Z^n(0) = \phi_Z(\phi_Z^{n-1}(0)) = \phi_Z(a_{n-1}(1)) \leq \phi_Z(a) = a$$

where we used that ϕ_Z is increasing.This shows that the smallest root is the extinction probability ■.

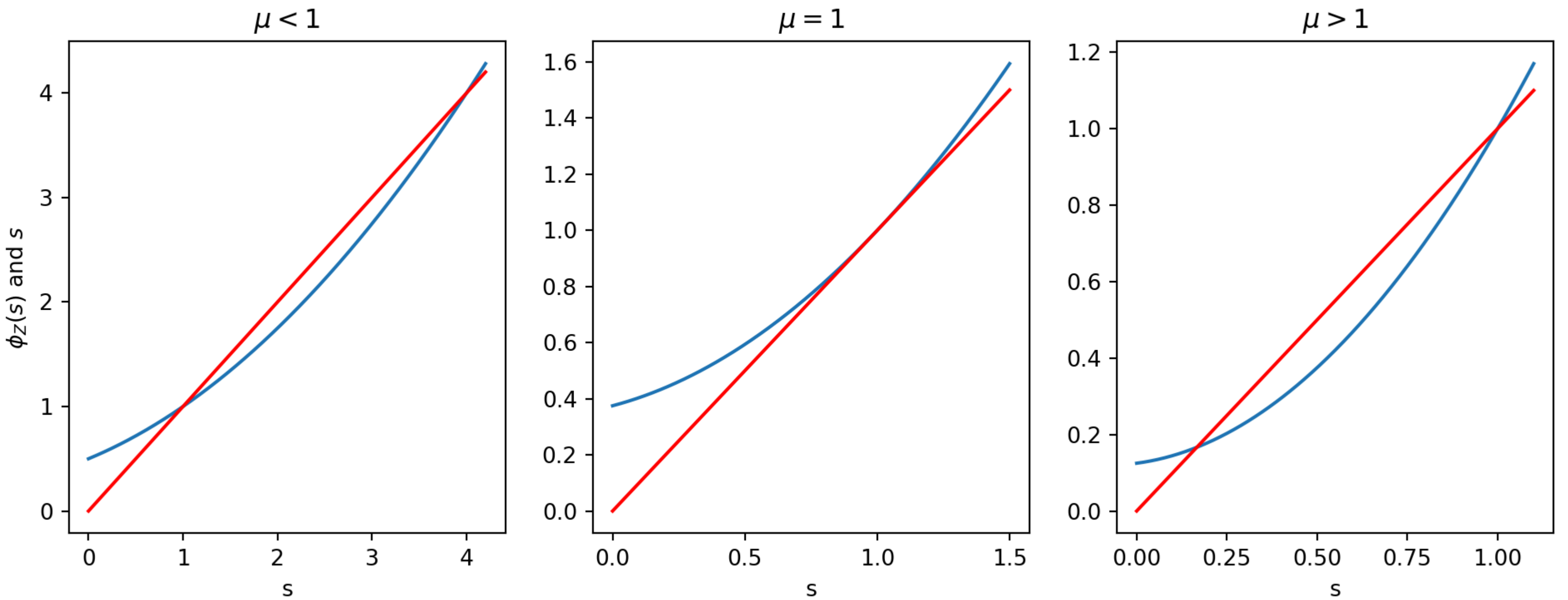


Figure 6.1: Extinction probabilities for branching processes

6.3 Examples

- If

$$p(0) = \frac{1}{4}, p(1) = \frac{1}{4}, p(2) = \frac{1}{2} \implies \phi(s) = \frac{1}{4} + \frac{1}{4}s + \frac{1}{2}s^2$$

then $\mu = \frac{5}{4} > 1$ and solving $\phi(a) = a$ gives $a = 1, \frac{1}{2}$ so the extinction probability is $1/2$.

- If

$$p(0) = \frac{1}{4}, p(1) = \frac{1}{2}, p(2) = \frac{1}{1} \implies \phi(s) = \frac{1}{4} + \frac{1}{2}s + \frac{1}{4}s^2$$

then $\mu = 1$ and solving $\phi(a) = a$ gives $a = 1$ so the extinction probability is 1.

- If

$$p(0) = \frac{1}{2}, p(1) = \frac{1}{4}, p(2) = \frac{1}{4} \implies \phi(s) = \frac{1}{2} + \frac{1}{4}s + \frac{1}{4}s^2$$

then $\mu = \frac{3}{4} < 1$ and solving $\phi(a) = a$ gives $a = 1, 2$ so the extinction probability is 1.

Remark: The theorem provides a numerical algorithm to find the extinction probability (if $\mu < 1$). Indeed we have shown that the sequence $0, \phi_Z(0), \phi^2(0), \phi^3(0), \cdots$ converges to a.

▶ Code

```
0.125
0.17413330078125
0.19498016827133297
0.20415762609873456
0.20826713424631457
0.2101216307743601
0.21096146654108697
0.21134240863922932
0.21151532651001245
0.2115938436492296
0.21162950143053777
0.21164569616414142
0.2116530515720635
0.21165639233633957
0.21165790969301684
0.21165859887003846
0.21165891189175637
0.21165905406517702
0.211659118639882
0.21165914796951835
```

6.4 Repair shop example

Recall the Markov chain in Example 4.3. To establish if the system is postive recurrent we try to solve $\pi P = \pi$ and find the system of equations

$$\begin{aligned}
\pi(0) &= \pi(0)a_0 + \pi(1)a_0 \\
\pi(1) &= \pi(0)a_1 + \pi(1)a_1 + \pi(2)a_0 \\
\pi(2) &= \pi(0)a_2 + \pi(1)a_2 + \pi(2)a_1 + \pi(3)a_0 \\
&\vdots \\
\pi(n) &= \pi(0)a_n + \sum_{j=1}^{n+1} \pi(j)a_{n+1-j}
\end{aligned}$$

We solve this using the generating functions $\psi(s) = \sum_{n=0}^{\infty} s^n \pi(n)$ and $\phi(s) = \sum_{k=0} s^k a_k$:

$$\begin{aligned}
\psi(s) = \sum_{n=0}^{\infty} s^n \pi(n) &= \pi(0)\sum_{n=0}^{\infty} s^n a_n + \sum_{n=0}^{\infty} s^n \sum_{j=1}^{n+1} \pi(j)a_{n+1-j} \\
&= \pi(0)\sum_{n=0}^{\infty} s^n a_n + s^{-1}\sum_{j=1}^{\infty} \pi(j)s^j \sum_{n=j-1}^{\infty} s^{n+1-j}a_{n+1-j} \\
&= \pi(0)\phi(s) + s^{-1}(\psi(s) - \pi(0))\phi(s)
\end{aligned}$$

Solving for $\psi(s)$ we find, after some algebra, the equation

$$\psi(s) = \frac{\pi(0)\phi(s)}{1 - \frac{1-\phi(s)}{1-s}}$$

To see if we can find $\pi(0)$ such that this equation canbe solved we take $s \to 1$. We have $\psi(1) = \phi(1) = 1$ and we have $\lim_{s \to 1} \frac{1-\phi(s)}{1-s} = \phi'(1) = \sum_{k=0}^{\infty} k a_k = \mu$ which is the mean number of object arriving in the repair shop in a single day.

We find the equation

$$1 = \frac{\pi(0)}{1-\mu}$$

and so we can find a solution $0 < \pi(0) \leq 1$ if and only if $\mu < 1$ and so $\pi(0) = 1 - \mu$.

$$\text{The repair shop Markov chain is positive recurrent iff } \mu < 1$$

To study transience we use Theorem 4.2 and try to find a solutinon $P\alpha(j) = \alpha(j)$ for $j \geq 1$ with $\alpha(0) = 1$ and $0 < \alpha(j) < 1$ for $j > 1$. We find the system of equation

$$\begin{aligned} \alpha(1) &= a_0\alpha(0) + a_1\alpha(1) + a_2\alpha(2) + \cdots \\ \alpha(2) &= a_0\alpha(1) + a_1\alpha(2) + a_2\alpha(3) + \cdots \\ \alpha(3) &= a_0\alpha(2) + a_1\alpha(3) + a_2\alpha(4) + \cdots \\ &\vdots \\ \alpha(n) &= \sum_{j=0}^{\infty} a_j\alpha(j+n-1) \end{aligned}$$

We try for a solution of the form $\alpha(j) = s^j$ which gives

$$s^n = \sum_{k=0}^{\infty} a_j s^{j+n-1} = s^{n-1}\phi(s) \implies \phi(s) = s$$

and from our analys of branching processes we know that a solution with $s < 1$ exists iff $\mu = \sum_{k=0}^{\infty} ka_k > 1$. So we get

$$\text{The repair shop Markov chain is transient iff } \mu > 1$$

and it is null-recurrent if $\mu = 1$.

6.5 Exercises

Exercise 6.1 Consider a branching process with offspring distribution given by p_n. Make this process irreducible by asserting that if the the population ever dies out, then in the next generation one new individual appears (i.e. $P_{01} = 1$). Determine for which values of p_n the chain is positive recurrent and transient.

Exercise 6.2 Given a branching process with the following offspring distributions determine the extinction probability a.

a. $p(0) = .25, p(1) = .4, p(2) = .35$

b. $p(0) = 5, p(1) = .1, p(3) = .4$

c. $p(0) = .62, p(1) = .30, p(2) = .02, p(6) = .02, p(13) = .04$

d. $p(i) = (1 - q)q^i$

e. $p(0) = 1/10, p(1) = 3/10, p(2) = 2/10, p(4) = 1/20, p(5) = 1/20, p(8) = 1/10, p(12) = 2/10$

f. $p(k)$ follows a Poisson distribution with parameter $\lambda = 3/2$.

Hint: For the last two you need to do it numerically.

Exercise 6.3 Consider the following variant of the branching process. During each time period each individual produces k offsprings with probability $p(k)$ and has probability $0 < q < 1$ of dying. Hence an individual will reproduce j times where j is the lifetime of the individual. For which choice of values of q and $p(n)$ do we have eventual exctinction probability equal to 1.

Exercise 6.4 We consider the Moran strorage model. Imagine a reservoir of water capacity $c \le \infty$. The levels X_n of the resevoir are observed at time $n, n+1, \cdots$. During each period an amount m units of water (if available, and, with $m \le c$) are removed from the reservoir. In addition a random amount A_n is added to the reservoir. The A_n are supposed to be independent IID random variables and are also independent of X_0. The corresponding Markov chain is given by

$$X_{n+1} = \max\{(X_n + A_{n+1} - m)_+, c\}$$

where Y_+ is the positive part of Y.

Consider the special case $m = 1$ and $c = \infty$ (infinite reservoir). Using a moment generating function show that this Markov chain is positive recurrent provided $E[A_1] < 1$. *Hint:* If X_n is positive recurrent then X_n must converges to a random variable X_∞ whose distribution is π. Then X_∞ must satisfy the equation

$$X_\infty = (X_\infty + A_\infty - m)_+$$

7 Reversible Markov chains

7.1 Balance equation

Consider a Markov chain with transition probabilities $P(i,j)$ and stationary distribution $\pi(i)$. We can rewrite the equation for stationarity,

$$\pi(i) = \sum_j \pi(j)P(j,i)$$

as

$$\sum_j \pi(i)P(i,j) = \sum_j \pi(j)P(j,i)\,. \tag{7.1}$$

which we are going to interpret as a *balance equation*.

We introduce the (stationary) probability current from i to j as

$$J(i,j) \equiv \pi(i)P(i,j) \tag{7.2}$$

and Equation 7.1 can be rewritten as

$$\underbrace{\sum_j J(i,j)}_{\text{total current out of state } i} = \underbrace{\sum_j J(j,i)}_{\text{total current into of state i}} \qquad \text{balance equation} \tag{7.3}$$

i.e., to be stationary the total probability current out of i must be equal to the total probability current into i.

7.2 Detailed balance

A stronger condition for stationarity can be expressed in terms of the balance between the currents $J(i,j)$. A Markov chain X_n satisfies detailed balance if there exists $\pi(i) \ge 0$ with $\sum_i \pi(i) = 1$ such that for all i, j we have

$$\pi(i)P(i,j) = \pi(j)P(j,i) \quad \text{or equivalently} \quad J(i,j) = J(j,i) \tag{7.4}$$

This means that for every pair i, j the probability currents $J(i,j)$ and $J(j,i)$ balance each other.
Clearly Equation 7.4 is a stronger condition than Equation 7.3 and thus we have

Lemma 7.1 If the Markov chain satisfies detailed balance for a probability distribution π then π is a stationary distribution.

But it is easy to see that detailed balance is a stronger condition than stationarity. The property of detailed balance is often also called (time)-reversibility since we have the following results which states that the probability of any sequence is the same as the probability of the time reversed sequence.

Theorem 7.1 (Time reversibility) Suppose the Markov chain X_n satisfies detailed balance and assume that the initial distribution is the stationary distribution π. Then for any sequnce of states $i_0, \cdots i_n$ we have

$$P\{X_0 = i_0\,, X_1 = i_1\,, \cdots\,, X_n = i_n\} = P\{X_0 = i_n\,, X_1 = i_{n-1}\,, \cdots\,, X_n = i_0\}$$

Proof. Using Equation 7.4 repeatedly we find

$$
\begin{aligned}
P\{X_0 = i_0\,, X_1 = i_1\,, \cdots\,, X_n = i_n\} &= \pi(i_0)P(i_0,i_1)P(i_1,i_2)\cdots P(i_{n-1},i_n) \\
&= P(i_1,i_0)\pi(i_1)P(i_1,i_2)\cdots P(i_{n-1},i_n) \\
&= P(i_1,i_0)P(i_2,i_1)\pi(i_2)\cdots P(i_{n-1},i_n) \\
&= \cdots \\
&= P(i_1,i_0)P(i_2,i_1)\cdots \pi(i_{n-1})P(i_{n-1},i_n) \\
&= P(i_1,i_0)P(i_2,i_1)\cdots P(i_n,i_{n-1})\pi(i_n) \\
&= P\{X_0 = i_n\,, X_1 = i_{n-1}\,, \cdots\,, X_n = i_0\}
\end{aligned}
$$

■

The next result is very easy and very useful.

Theorem 7.2 Suppose X_n is a Markov chain with finite state space S and with a *symmetric* transition matrix, i.e, $P(i,j) = P(j,i)$. Then X_n satisfies detailed balance with $\pi(j) = 1/|S|$, i.e., the stationary distribution is uniform on S.

7.3 Examples

- *Random walk on the hypercube* $\{0,1\}^m$. The state space S is

$$S = \{0,1\}^m \;=\; \{\sigma = (\sigma_1, \cdots, \sigma_m)\,;\, \sigma_i \in \{0,1\}\}$$

To define the move of the random walk, just pick one coordinate $j \in \{1, \cdots, m\}$ and flip the j^{th} coordinate, i.e., $\sigma_j \to 1 - \sigma_j$. We have thus

$$P(\sigma, \sigma') = \begin{cases} \frac{1}{m} & \text{if } \sigma \text{ and } \sigma' \text{ differ by one coordinate} \\ 0 & \text{otherwise} \end{cases}$$

Clearly P is symmetric and thus $\pi(\sigma) \;=\; 1/2^m$.

- *Random walk on the graph* $G = (E, V)$ has transition probabilities $p(v,w) = \frac{1}{\deg(v)}$. This Markov chain satifies detailed balance with the (unnormalized) $\mu(v) = deg(v)$. Indeed we have $P(v,w) > 0 \iff P(w,v) > 0$ and thus if $P(v,w) > 0$ we have

$$\mu(v)P(v,w) \;=\; \deg(v)\frac{1}{\deg(v)} = 1 \;=\; \deg(w)\frac{1}{\deg(w)} = \mu(w)P(w,v)\,.$$

This is slightly easier to verify that the stationary equation $\pi P = \pi$. After normalization we find $\pi(v) = \deg(v)/2|E|$.
For example for the simple random walk on $\{0, 1, \cdots, N\}$ with reflecting boundary conditions we find $\pi \;=\; \left(\frac{1}{2N}, \frac{2}{2N}, \cdots, \frac{1}{2N}\right)$.

- *Network Markov chain*: The previous example can be generalized as follows. For a given graph $G = (E, V)$ let us assign a positive weight $c(e) > 0$ to each (undirected) edge $e = \{v, w\}$, that is we choose numbers $c(v,w) = c(w,v)$ with $c(v,w) = 0$ iff v and w are not connected by an edge. If the transition probabilities are given by

$$P(v,w) = \frac{c(v,w)}{c(v)} \quad \text{with } c(v) = \sum_w c(v,w)\,,$$

then it is easy to verify that the Markov chain satisfies detailed balance with

$$\pi(v) = \frac{c(v)}{c_G} \quad \text{with } c_G = \sum_v c(v)\,.$$

- *Birth-Death Processes Markov chain:* Let us consider a Markov chain on the state space $S = \{0, \cdots, N\}$ (N could be infinite) with transition probabilities have the following tridiagonal structure

$$\begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ \dots \end{matrix} \begin{pmatrix} r_0 & p_0 & 0 & 0 & \cdots \\ q_1 & r_1 & p_1 & 0 & \cdots \\ 0 & q_2 & r_2 & p_2 & \cdots \\ \vdots & & \ddots & \ddots & \ddots \end{pmatrix}$$

This is called a *birth and death process* since the only possible transition are to move up or down by unit or stay unchanged. Random walks (see Example 1.5 and Example 4.1), discrete queueing models (Example 4.2), the Ehrenfest urn model Example 1.6 are special case of birth and death processes.

Birth and death Markov chain always satisfy detailed balance. Indeed the only non trivial detailed balance conditions are

$$\pi(j)P(j,j+1) = \pi(j+1)P(j+1,j) \implies \pi(j)p_j \;=\; \pi(j+1)q_{j+1}\,, \qquad \text{for } j=0,\cdots,N-1\,.$$

and this can be solved recursively. We obtain

$$\begin{aligned}
\pi(1) &= \pi(0)\frac{p_0}{q_1} \\
\pi(2) &= \pi(1)\frac{p_1}{q_2} = \pi(0)\frac{p_0p_1}{q_1q_2} \\
&\vdots \\
\pi(N) &= \pi(0)\frac{p_0p_1\cdots p_{N-1}}{q_1q_2\cdots q_{N-1}}
\end{aligned}$$

and with normalization $\pi(j) = \frac{\prod_{k=1}^{j}\frac{p_{k-1}}{q_k}}{\sum_{l=0}^{N}\prod_{k=1}^{l}\frac{p_{k_1}}{q_k}}$. If N is infinite we need the sum in the deominator to be finite.

For example the Ehrenfest urn in Example model has

$$p_j = \frac{N-j}{N}\,, \quad q_j = \frac{j}{N}$$

and thus we obtain

$$\pi(j) = \pi(0)\frac{\frac{N}{N}\frac{N-1}{N}\cdots\frac{N-(j-1)}{N}}{\frac{1}{N}\frac{2}{N}\cdots\frac{j}{N}} = \pi(0)\binom{N}{j} \quad \text{and also } \pi(0) = \sum_{j=0}^{N}\binom{N}{j} = 2^N.$$

7.4 Exercises

Exercise 7.1 A total of m white balls and m black balls are distributed among two urns, each of which contains m balls. At each stage a ball is randomly selected from each urn and the two selected balls are interchanged. Let X_n be the number of black balls in urn 1 after n stages.

1. Give the transition probabilities of the Markov chain X_n.
2. Can you guess without any computation guess what the stationary distribution is?
3. Find the stationary distribution using the detailed balance condition.

▶ Solution

Exercise 7.2 (Cycle condition for detailed balance) Suppose X_n an irreducible Markov chain with stationary distribution π. Assume that for all pairs i, j we have $P(i,j) > 0 \iff P(j,i) > 0$. Show that X_n satisfies detailed balance if and only if for any n and any sequence of state $i_0, i_1, i_2, \cdots i_n$

$$P(i_0,i_1)P(i_1,i_2)\cdots P(i_{n-1},i_n)P(i_n,i_0) = P(i_0,i_n)P(i_n,i_{n-1})\cdots P(i_2,i_1)P(i_1,i_0)$$

Hint: For the “if” part use the convergence to stationary distribution.

▶ Solution

Exercise 7.3 (The time reversed chain) Let X_n be an irreducible Markov chain with transition probabilities $P(i,j)$ and stationary distribution $\pi(i)$. Define $\widehat{P}(i,j) = \frac{\pi(j)}{\pi(i)} P(j,i)$.

1. Show that $\widehat{P}(i,j)$ is a stochastic matrix and that the corresponding Markov chain $\widehat{X}_n$ is irreducible with stationary distribution π.

2. Show that if the initial distribution is π then we have

$$P\left\{\widehat{X}_0 = i_0, \widehat{X}_1 = i_1, \cdots, \widehat{X}_n = i_n\right\} = P\left\{X_0 = i_n, X_1 = i_{n-1}, \cdots X_n = i_O\right\}.$$

   which justifies calling the Markov chain $\widehat{X}_n$ to be the *time reversed* chain.

3. What is the time reversed chain for the Markov chain with transition matrix

$$P = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{pmatrix} 1-p & p & 0 & 0 & \cdots \\ 1-p & 0 & p & 0 & \cdots \\ 1-p & 0 & 0 & p & \cdots \\ \vdots & & & \ddots & \ddots \end{pmatrix}$$

4. Show that the Markov chains with transition matrices $\frac{P+\widehat{P}}{2}$ and $\widehat{P}P$ satisfy detailed balance with stationary distribution π.

▶ Solution

Exercise 7.4 (Eigenvalues of P) Let P be the transition matrix of an irreducible Markov chain with finite state space S and stationary distribution π.

1. Show that if P is aperiodic then P has an eigenvalue 1 which is simple (i.e. 0 is a simple root of $\det(I-P)$) and that all other eigenvalues have absolute value strictly less than 1. *Hint*: Use the convergence theorem

2. Prove that if P is periodic with period 2 then P has eigenvalues 1 and -1 each of which is simple.
   Hint: WLOG you can write P is the block form $P = \begin{pmatrix} 0 & P_{01} \\ P_{10} & 0 \end{pmatrix}$. Show that 1 is an eigenvalue of P^2 with multiplcity 2. If $f = \begin{pmatrix} f_0 \\ f_1 \end{pmatrix}$ is an eigenvector for P, then consider $g = \begin{pmatrix} f_0 \\ -f_1 \end{pmatrix}$.

3. For column vectors f and g define a scalar product on $\mathbb{R}^{|S|}$ by $\langle f, g \rangle_\pi = \sum_{i \in S} \pi(i) f(i) g(i)$. Let $P^\dagger$ to be the adjoint of P with respect to the scalar product $\langle f, g \rangle_\pi$, i.e., $\langle f, Pg \rangle_\pi = \langle P^\dagger f, g \rangle_\pi$ for all f, g.
   Show that $P^\dagger$ are the transition probabilities of the reversed Markov chain from Exercise 7.3 and that the Markov chain is reversible if and only $P = P^\dagger$.

4. Show that if the Markov chain is reversible, the eigenvalues of P are real with $\lambda_0 = 1 > \lambda_2 \geq \lambda_3 \cdots \geq \lambda_N \geq -1$ and that if the chain is aperiodic then $\lambda_N > -1$.

▶ Solution

Exercise 7.5 A Markov chain (on a finite state space) is a called a tree process if

- $P(i,j) > 0 \iff P(j,i) > 0$
- For any pair (i,j) there is a unique path $i_0 = i, i_1, i_2, \cdots, i_N = j$ such that $P(i_k, i_k + 1) > 0$ for $k = 0, \cdots, N-1$.

Show that a irreducible tree process satisfies detailed balance.

▶ Solution

8 Markov chain Monte-Carlo

8.1 MCMC

Suppose you are given a certain probability distribution π on a set S and you goal is to generate a sample from this distribution. The **Monte-Carlo Markov chain (MCMC)** method consists in constructing an irreducible Markov chain X_n whose stationary distribution is π. Then to generate π one simply runs the Markov chains X_n long enough such that it is close to its equilibrium distribution. It turns out that using the detailed balance condition is a very useful tool to construct the Markov chain in this manner.

A-priori this method might seems an unduly complicated way to sample from π. Indeed why not simply simulate from π directly?

To dispel this impression we will consider some concrete examples but for example we will see that often one want to generate a uniform distribution on a set S whose cardinality $|S|$ might be very difficult.

A large class of model are so-called "energy models" which are described by an explicit function $f : S \to \mathbb{R}$ interpreted as the energy (or weight) of the state. Then one is interested in the stationary distribution

$$\pi(i) = Z^{-1} e^{-f(i)} \quad \text{where } Z = \sum_i e^{-f(i)} \text{ is a normalization constant}$$

As we will see computing the normalization constant Z can be very difficult and so one cannot directly simulate from π!

The measure π assign biggest probability to the states i where $f(i)$ is the smallest, that is they have the smallest energy. Often (e.g. in economics) the measure is written with a different sign convention $\pi(i) = Z^{-1} e^{f(i)}$ which now favor the states with the highest values of f.

8.2 Proper q-coloring of a graph

For a graph $G = (E, V)$ a proper q-coloring consists of assigning to each vertex v of the graph one of q colors subject to the constraint that if 2 vertices are linked by an edge they should have different colors. The set S' of all proper q-coloring which is a subset of $S = \{1, \cdots, q\}^V$. We denote the elements of S by $\sigma = \{\sigma(v)\}_{v \in V}$ with $\sigma(v) \in \{1, \cdots, q\}$. The uniform distribution on all such proper coloring is $\pi(\sigma) = 1/|S'|$ for all $\sigma \in S'$. Even for moderately complicated graph it can be very difficult to compute to $|S'|$.

Consider the following transition rules:
1. Choose a vertex v of at random and choose a color a at random. 2. Set $\sigma'(v) = a$ and $\sigma'(w) = \sigma(w)$ for $w \neq v$. 3. If σ' is a proper q-coloring then set $X_{n+1} = \sigma'$. Otherwise set $X_n = \sigma$.

The transition probabilities are then given by

$$P(\sigma, \sigma') = \begin{cases} \frac{1}{q|V|} & \text{if } \sigma \text{ and } \sigma' \text{ differ at exactly one vertex} \\ 0 & \text{if } \sigma \text{ and } \sigma' \text{ differ at more than one vertex} \\ 1 - \sum_{\sigma' \neq \sigma} P(\sigma, \sigma') & \text{if } \sigma' = \sigma \end{cases}$$

Note that $P(\sigma, \sigma)$ is not known explicitly either but is also not used to run the algorithm. The cardinality $|S'|$ is also not needed.

In order to check that the uniform distribution is stationary for this Markov chain it note that $P(\sigma, \sigma')$ is symmetric matrix. If one can change σ into σ' by changing one color then one can do the reverse transformation too.

8.3 Knapsack problem

This is a classical optimization problem. You own m books and the i^{th} book has weight w_i lb and is worth \$ v_i. In your knapsack you can put at most a total of b pounds and you are looking to pack the most valuable knapsack possible.

To formulate the problem mathematically we introduce

$$\begin{aligned} w &= (w_1, \cdots w_m) \in \mathbf{R}^m && \text{weight vector} \\ v &= (v_1, \cdots v_m) \in \mathbf{R}^m && \text{value vector} \\ \sigma &= (\sigma_1, \cdots \sigma_m) \in \{0,1\}^m && \text{decision vector} \end{aligned}$$

where we think that $\sigma_i = 1$ if the i^{th} book is in the knapsack. The state space is

$$S' = \{\sigma \in \{0,1\}^m \, ; \, \sigma \cdot w \le b\}$$

and the optimization problem is

$$\text{Maximize } v \cdot \sigma \text{ subject to the constraint } \sigma \in S'$$

As a first step we generate a random element of S' using the simple algorithm. If $X_n = \sigma$ then

1. Choose $j \in \{1, \cdots m\}$ at random.
2. Set $\sigma' = (\sigma_1, \cdots, 1 - \sigma_j, \cdots, \sigma_m)$.
3. If $\sigma' \in S'$, i.e., if $\sigma' \cdot w \le b$ then let $X_{n+1} = \sigma'$. Otherwise $X_{n+1} = \sigma$.

In other words, choose a random book. If it is in the sack already remove it. If it is not in the sack add it provided you do not exceed the the maximum weight. Note that the Markov chain X_n is irreducible, since each state communicates with the state $\sigma = (0, \cdots, 0)$. It is aperiodic except in the uninteresting case where $\sum_i w_i \le b$. Finally the transition probabilities are symmetric and thus the uniform distribution the unique stationary distribution.

In the knapsack problem we want to maximize the function $f(\sigma) = \sigma \cdot v$ on the state space. One possible algorithm would be to generate an uniform distribution on the state space and then to look for the maximum value of the function. But it would be a better idea to sample from a distribution which assign higher probabilities to the states which we are interested in, the ones with a high value of f.

Let S be the state space and let $f : S \to \mathbb{R}$ be a function. It is convenient to introduce the probability distributions define for $\beta > 0$ by

$$\pi_\beta(i) = \frac{e^{\beta f(i)}}{Z_\beta} \quad \text{with} \quad Z_\beta = \sum_{j \in S} e^{\beta f(j)} .$$

Clearly π_β assign higher weights to the states i with bigger values of $f(i)$. Let us define

$$S^* = \left\{ i \in S\,;\, f(i) = f^* \equiv \max_{j \in S} f(j) \right\} .$$

If $\beta = 0$ then π_0 is simply the uniform distribution on S. For $\beta \to \infty$ we have

8.4 Metropolis algorithm

A fairly general method to generate a distribution π on the state space S is given the **Metropolis algorithm**. This algorithm assumes that you already know how to generate the uniform distribution on S by using a symmetric transition matrix Q.

Theorem 8.1 (Metropolis algorithm) Let π be a probability distribution on S with $\pi(i) > 0$ for all i and let Q be a symmetric transition matrix. Consider the Markov chain with the following transition matrix (the Metropolis algorithm). If $X_n = i$

1. Choose $Y \in S$ according to Q, i.e., $P\{Y = j \,|\, X_n = i\} = Q(i,j)$
2. Define the acceptance probability $\alpha(i,j) = \min\left\{1\,,\, \frac{\pi(j)}{\pi(i)}\right\}$
3. Accept Y with probability $\alpha = \alpha(i,j)$ by generating random number U. If $U \le \alpha$ then set $X_{n+1} = Y$ (i.e., accept the move) and if $U > \alpha$ then $X_{n+1} = X_n$ (i.e., reject the move).

If Q is an irreducible transition probability matrix on S then the Metropolis algorithm defines an irreducible Markov chain on S which satisfies detailed balance with stationary distribution π.

Proof. Let $P(i,j)$ be the transition probabilities for the Metropolis Markov chain. Then we have

$$P(i,j) = Q(i,j)\alpha(i,j) = Q(i,j)\min\left\{1\,,\, \frac{\pi(j)}{\pi(i)}\right\}.$$

Since we assume $\pi(i) > 0$ for all states i, the acceptance probability α never vanishes. Thus if $P(i,j) > 0$ whenever $Q(i,j) > 0$ and thus P is irreducible if Q is itself irreducible.

In order to check the reversibility we note that

$$\pi(i)P(i,j) \;=\; Q(i,j)\pi(i)\min\left\{1, \frac{\pi(j)}{\pi(i)}\right\} \;=\; Q(i,j)\min\left\{\pi(i)\,,\,\pi(j)\right\}$$

and the r.h.s is symmetric in i, j and thus $\pi(i)P(i,j) = \pi(j)P(j,i)$. ■

- Note that only the ratios $\frac{\pi(j)}{\pi(i)}$ are needed to run the algorithm, in particular we do not need the normalization constant. This is a very important feature of the Metropolis algorithm.
- We could have chosen another acceptance probability $\alpha(i,j)$. By inspection of the proof it is enough to pick $\alpha(i,j)$ such that $\alpha(i,j) \le 1$ and $\pi(i)\alpha(i,j) = \pi(j)\alpha(j,i)$ is symmetric. Some such examples will be considered in the homework and the choice given in Theorem 8.1 is actually optimal in the sense it gives the highest acceptance probability.
- The general case with a non-symmetric proposal matrix Q is called the **Metropolis-Hastings algorithm**. In this case we use the acceptance probability is chosen to be

$$\alpha(i,j) = \min\left\{1, \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}\right\}$$

  and it is not difficult to check (see Homework) that the Metropolis-Hasting algorithm yields a reversible Markov chain with stationary distribution π.

8.5 Uniform distribution on a graph

Suppose we are given a graph $G = (E, V)$. Often in application to computer networks or social networks the graph is not fully known. Only local information is available: given a vertex one knows the vertices one is connected to. Nonetheless you can run a random walk on the graph by choosing a edge at random and moving to the corresponding vertex, i.e.

$$Q(v,w) = \frac{1}{\deg(v)} \text{ if } v \sim w$$

and the stationary distribution is $\pi(v) \propto \mathrm{deg(v)}$.

Suppose you wish to generate a uniform distribution on the graph. Then we can use the Metropolis-Hasting to generate a Markov chain with a uniform stationary distribution. We have

$$\alpha(v,w) = \min\left\{1, \frac{Q(w,v)}{Q(v,w)}\right\} = \min\left\{1, \frac{\deg(v)}{\deg(w)}\right\} \text{ if } v \sim w$$

and thus the transition matrix

$$P(v,w) = Q(v,w)\alpha(v,w) = \frac{1}{\deg(v)} \min\left\{1, \frac{\deg(v)}{\deg(w)}\right\} = \min\left\{\frac{1}{\deg(v)}, \frac{1}{\deg(w)}\right\}$$

generates a uniform distribution on a graph.

8.6 Metropolis algorithm for the knapsack problem

Consider the distribution $\pi_\beta(\sigma) = \frac{e^{\beta v\cdot\sigma}}{Z_\beta}$ where the normalization constant $Z_\beta = \sum_{\sigma\in S'} e^{\beta v\cdot\sigma}$ is almost always impossible to compute in practice. However the ration

$$\frac{\pi(\sigma')}{\pi(\sigma)} = e^{\beta v\cdot(\sigma'-\sigma)}$$

does not involve Z_β and is easy to compute.

For this distribution we take as the proposal Q matrix used to generate a uniform distribution on the allowed states of the knapsack (see Knapsack problem) and the Metropolis algorithm now reads as follows. If $X_n = \sigma$ then

1. Choose $j \in \{1, \cdots m\}$ at random.
2. Set $\sigma' = (\sigma_1, \cdots, 1-\sigma_j, \cdots, \sigma_m)$.
3. If $\sigma'\cdot w > b$ (i.e. if $\sigma \notin S'$) then $X_{n+1} = \sigma$. (i.e. reject)
4. If $\sigma'\cdot w \le b$ then let $\alpha = \min\left\{1, \frac{\pi(\sigma')}{\pi(\sigma)}\right\} = \min\left\{1, e^{\beta v\cdot(\sigma'-\sigma)}\right\} = \begin{cases} e^{-\beta v_j} & \text{if } \sigma_j = 1 \\ 1 & \text{if } \sigma_j = 0 \end{cases}$
5. Generate a random number U, If $U \le \alpha$ then $X_{n+1} = \sigma'$. Otherwise $X_{n+1} = \sigma$.

In short, if you can add a book to your knapsack you always do so, while you remove a book with a probability which is exponentially small in the weight.

8.7 Glauber algorithm

Another algorithm which is widely used for Monte-Carlo Markov chain is the **Glauber algorithm** which appear in the literature under a variety of other names such as Gibbs sampler in statistical applications, logit rule in economics and social sciences, heat bath in physics, and undoubtedly under various other names.

The Glauber algorithm is not quite as general as the Metropolis algorithm. Indeed we *assume* that the state space S has the following structure

$$S \subset \Omega^V$$

where both Ω and V are finite sets. For example $S \subset \{0,1\}^m$ in the case of the knapsack problem or $S \subset \{1, \cdots, q\}^V$ for the case of the proper q-coloring of a graph. We denote by

$$\sigma \; = \; \{\sigma(v)\}_{v \in V} \,, \quad \sigma(v) \in \Omega \,.$$

the elements of S.

It is useful to introduce the notation

$$\sigma_{-v} \; = \; \{\sigma(w)\}_{w \in V, w \neq v}$$

and we write

$$\sigma \; = \; (\sigma_{-v}, \sigma(v)) \,.$$

to single out the v entry of the vector σ.

Theorem 8.2 (Glauber algorithm) Let π be a probability distribution on $S \subset \Omega^V$ and extend π to Ω^V by setting $\pi(\sigma) = 0$ if $\sigma \in \Omega^V \setminus S$. If $X_n = \sigma$ then

1. Choose $v \in V$ at random.
2. Replace $\sigma(v)$ by a new value $a \in \Omega$ (provided $(\sigma_{-v}, a) \in S$) with probability

$$\frac{\pi(\sigma_{-v}, a)}{\sum_{b \in \Omega} \pi(\sigma_{-v}, b)} \,.$$

The Glauber algorithm defines a Markov chain on S which satisfies detailed balance with stationary distribution π.

The irreducibility of the algorithm is not guaranteed a-priori and needs to be checked on a case-by-case basis.

Proof. The transition probabilities are given by

$$P(\sigma, \sigma') = \begin{cases} \frac{1}{|V|} \frac{\pi(\sigma_{-v}, \sigma'(v))}{\sum_{b\in\Omega} \pi(\sigma_{-v}, b)} & \text{if } \sigma_{-v} = \sigma'_{-v} \text{ for some } v \\ 0 & \text{if } \sigma_{-v} \neq \sigma'_{-v} \text{ for all } v \\ 1 - \sum_{\sigma'} P(\sigma, \sigma') & \text{if } \sigma = \sigma' \end{cases}$$

To check detailed balance we note that if $P(\sigma, \sigma') \neq 0$

$$\pi(\sigma) P(\sigma, \sigma') = \frac{\pi(\sigma)\pi(\sigma')}{\sum_{b\in\Omega} \pi(\sigma_{-v}, b)} \,,$$

and this is symmetric in σ and σ'. ■.

8.8 Ising model on a graph

Let $G = (E, V)$ be a graph and let $S = \{-1, 1\}^V$. That is to each vertex assign the value ± 1, you can think of a magnet at each vertex pointing either upward ($+1$) or downward (-1). To each $\sigma \in S$ we assign an "energy" $H(\sigma)$ given by

$$H(\sigma) = - \sum_{e=(v,w) \in E} \sigma(v)\sigma(w) \,.$$

The energy σ is minimal if $\sigma(v)\sigma(w) = 1$ i.e., if the magnets at v and w are aligned. Let us consider the probability distribution

$$\pi_\beta(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z_\beta}\,, \quad Z_\beta = \sum_{\sigma} e^{-\beta H(\sigma)} \,.$$

The distribution π_β is concentrated around the minima of $H(\sigma)$. To describe the Glauber dynamics note that

$$H(\sigma_{-v}, 1) - H(\sigma_{-v}, -1) = -2 \sum_{w\,;\, w \sim v} \sigma(w)$$

and this can be computed simply by looking at the vertices connected to v and not at all the graph. So the transition probabilities for the Glauber algorithm are given by picking a vertex at random and then updating with probabilities

$$\frac{\pi(\sigma_{-v}, \pm 1)}{\pi(\sigma_{-v}, 1) + \pi(\sigma_{-v}, -1)} = \frac{1}{1 + e^{\pm\beta[H(\sigma_{-v},1) - H(\sigma_{-v},-1)]}} = \frac{1}{1 + e^{\mp 2\beta \sum_{w\,;\, w\sim v} \sigma(w)}} \,.$$

By comparison for the Metropolis algorithm we pick a vertex at random and switch $\sigma(v)$ to $-\sigma(v)$ and accept the move with probability

$$\min\left\{1, \frac{\pi(\sigma_{-v}, -\sigma(v))}{\pi(\sigma_{-v}, \sigma(v))}\right\} = \min\left\{1, \frac{\pi(\sigma_{-v}, -\sigma(v))}{\pi(\sigma_{-v}, \sigma(v))}\right\} = \min\left\{1, e^{2\beta \sum_{w\,;\,w\sim v} \sigma(w)\sigma(v)}\right\} .$$

8.9 Simulated annealing

- Consider again the problem of finding the minimum of a function $f(j)$, that is

$$f^* = \min\{f(j), j \in S\}$$

and we denote by $S^* \subset S$ the location of the global minima. You should think of f as a complicated (and non-convex) function with complicated level sets and various “local” minima.

- To perform this task we sampling a distribution of the form

$$\pi_T(j) = e^{-\frac{f(j)}{T}}$$

which concentrates on the minima of $f(j)$, and the more so as $T \to 0$.

- The idea of simulated annealing comes from physics. The concept of annealing in physics is to obtain a low energy state of a solid (typically a crystal) you first heat it up to reach a liquid state and then, slowly, decrease the temperature to let the particles arrange themselves.

- For a Markov chain the idea is to pick a *temperature schedule*

$$T_1 > T_2 > T_3 > \cdots \quad \text{with } \lim_{k\to\infty} T_k = 0$$

with T_1 sufficiently large.

The following algorithm is a Markov chain with nonstationary transition probabilities.

Simulated annealing algorithm:
+ Initialise the Markov chain X_0 and the temperature T_1.
+ For each k run N_k steps of the Metropolis or Gibbs sampler with invariant distribution π_{T_k}.
+ Update the temperature to T_{k+1} starting with the final configuration.

A nice result about Metropolis sampler can be found for example in Hajek (1988) (a more precsie version is given there)

Theorem 8.3 (Convergence of simulated annealing) For the simulated annealing of the Metropolis algorithm ($N_k = 1$) there exists a constant d^* such that we have

$$\lim_{n\to\infty} P\{X_n \in S^*\} = 1$$

if and only if

$$\sum_{k=1}^{\infty} e^{-\frac{d^*}{T_k}} = \infty$$

The constant d^* measure the depths of the local minima of f where locality is measure in terms of the proposal matrix Q.

The theorem tells us that if decrease the temperature very very slowly, e.g. like

$$T_k = \frac{c}{\log(k)}$$

with $c > d^*$ then the Markov chain will converge, with probabilty 1, to a minima.
This an extremely slow cooling schedule which makes it not very practical and other schedules are used like $T_{k+1} = 0.99T_k$.

8.10 Parallel tempering

- The idea of parallel tempering is not use a temperature schedule but rather to use several copies of the system running at different temperatures. The copies at high temperature will explore the state space more efficiently and there is a switching mechanism which *exchange* the different copies so that the lower temperature copies can take advantage of the exploration done at high temperature.

- The state space is the product of k copies of S

$$S^{(K)} = \underbrace{S \times \cdots \times S}_{\text{k times}}$$

and we denote a state by the vector $\mathbf{i} = (i_1, \cdots, i_K)$.

- One picks a set of temperatures $T_1 < T_2 < \cdots < T_k$ and a probability distribution

$$\pi^{(k)}(\mathbf{i}) = \prod_{l=1}^{K} \pi_{T_l}(i_l) = \prod_{l=1}^{K} Z_l e^{-\frac{f(i_l)}{T_l}}$$

which is the product of the distribution at different temperatures.

- The parallel tempering consists if two kinds of moves. The parallel move update each component of $\mathbf{X}_n$ independently with the Metropolis algorithms at different temperature and there is a swapping move which exchanges a pair of components of the state vector $\mathbf{X}_n$ in such a way as not to disturb the invariant measure. The component at the lowest temperature can be used to find the desired minimum.

Parallel tempering algorithm:

- Suppose the state $\mathbf{X} = \mathbf{i}$, pick a random number U.
- If $U < \alpha$ then we do the *parallel step* and update each compoent of $\mathbf{X}$ according to a Metropolis move at the corresponding temperatures T_l.
- If $U \geq \alpha$ we do the *swapping step*. We randomly chose a neighboring pair $l, l+1$ and propose to swap the components $X_{n,l}$ and $X_{n,l+1}$. We accept the swap with probability

$$\min\left\{1, \frac{\pi_{T_l}(i_{l+1})\pi_{T_{l+1}}(i_l)}{\pi_{T_l}(i_l)\pi_{T_{l+1}}(i_{l+1})}\right\}$$

The parallel moves clearly satisfy satisfy the detailed balance since each component does. As for a swap move which swaps the component i_l and i_{l+1} in the state vector, we also have detailed balance since

$$\begin{aligned}
&\pi_{T_l}(i_l)\pi_{T_{l+1}}(i_{l+1})(1-\alpha)\frac{1}{K-1}\min\left\{1, \frac{\pi_{T_l}(i_{l+1})\pi_{T_{l+1}}(i_l)}{\pi_{T_l}(i_l)\pi_{T_{l+1}}(i_{l+1})}\right\} \\
&= (1-\alpha)\frac{1}{K-1}\min\left\{\pi_{T_l}(i_l)\pi_{T_{l+1}}(i_{l+1}), \pi_{T_l}(i_{l+1})\pi_{T_{l+1}}(i_l)\right\}
\end{aligned}$$

which is symmetric in i_l, i_{l+1}.

8.11 Exercises

Exercise 8.1 Consider a symmetric random walk $p = \frac{1}{2}$ on $\{0, 1, 2, \cdots\}$ as a proposal matrix. What is the Metropolis-Hasting algorithm to generate a Poisson distribution with parameter λ.

Exercise 8.2 (General Metropolis-Hastings algorithm) Let $\pi(i) > 0$ be a probability distribution on the state space S. Let $Q(i,j)$ be a transition probability matrix (not necessarily symmetric). Set

$$T(i,j) \;=\; \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}$$

and suppose $A : [0,\infty] \to [0,1]$ be any function such that

$$A(z) = zA(1/z)$$

for all $z \in [0,\infty]$. Set

$$P(i,j) \;=\; Q(i,j)A(T(i,j)) \text{ for } i \neq j$$

and $P(i,i) = 1 - \sum_{j \neq i} P(i,j)$. Think of $A(T(i,j))$ has the acceptance probability for the proposed transition from i to j.

Exercise 8.3 (N-queens problem) The N-queens problems consist at arranging N queens on a $N \times N$ checkerboord so that no two queens can attack each other. You should formulate this problem as a MCMC method as follows.

- Clearly no 2 queens can be on the same row and so, without loss of generality, we can represent a state of the system by a vector $\sigma = (\sigma_1, \cdots, \sigma_N)$ where $\sigma_j \in \{1, 2, \cdots, N\}$. Here $\sigma_j = k$ if the row on row j is on the k^{th} position on that row.
- Define a function which $S(\sigma)$ which counts the number of attacking pair of queens. Solving the N-queens problem consists in finding solutions such that $S(\sigma) = 0$.

1. Write down the Metropolis and Glauber algorithms for the distribution $\pi(\sigma) \propto e^{-S(\sigma)/T}$.
2. Code it, e.g. the Metropolis and try to find all the solutions for the N-queens problems. For $N = 8$ there are 92 solutions.

Exercise 8.4 (Gibbs sampler) The Gibbs sampler is a variation on Glauber sampler to sample distribution $() on vectors $\sigma = (\sigma_1, \cdots, \sigma_d) \in S = \Omega^m$.
It is atttributed to Geman and Geman (one of which used to work here at UMass Amherst)

The algorithm goes as follows: Suppose $\pi(\sigma)$ is the target distribution.

- Given $X_n = (X_{n,1}, \cdots, X_{n,d})$ generate Y as follows.
  - Generate Y_1 from the conditional distribution $\pi(\sigma_1 | X_{n,2}, \cdots, X_{n,d})$.
  - Generate Y_2 from the conidtional distribution $\pi(\sigma_2 | Y_1, X_{n,3}, \cdots, X_{n,d})$.
  - $\ldots$
  - Generate Y_d from the conidtional distribution $\pi(\sigma_2 | Y_1, Y_2, \cdots, Y_{n-1})$.
- Set $X_{n+1} = Y$.

1. Write down the algorithm explicitly, say for the knapsack problem.
2. What is the connection between the Glauber and Gibbs samplers?
3. Prove that π is a stationary distribution for the resulting Markov chain. Is it reversible?

:::

9 Coupling methods

9.1 Total variation norm

- Given two probability measure μ and ν on S the total variation distance between μ and ν is given by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A \subset \Omega} |\mu(A) - \nu(A)|$$

  that is the largest distance between the measures of sets.

- The supremum over all set is not convenient to compute but we have the formula

Theorem 9.1 We have the formula

$$\|\mu - \nu\|_{\mathrm{TV}} = \sum_{i:\mu(i) \geq \nu(i)} (\mu(i) - \nu(i))(x) = \frac{1}{2} \sum_i |\mu(i) - \nu(i)| \tag{9.1}$$

Proof. First note that the second equality in Equation 9.1 follows from the first. Indeed, if the first equality holds, then by interchanging μ and ν we also have

$$\|\mu - \nu\|_{\mathrm{TV}} = \sum_{i:\mu(i) \geq \nu(i)} |\mu(i) - \nu(i)| = \sum_{i:\nu(i) \geq \mu(i)} |\nu(i) - \mu(i)|$$

which proves the second equality.

To prove the first equality in Equation 9.1 we consider the set $B = \{i : \mu(i) \ge \nu(i)\}$. For any event A we have

$$\mu(A) - \nu(A) = \sum_{i\in A} \mu(i) - \nu(i) \le \sum_{i\in A\cap B} \mu(i) - \nu(i) \le \sum_{i \in B} \mu(i) - \nu(i) = \mu(B) - \nu(B)$$

By interchanging the role of μ and ν we find

$$\nu(A) - \mu(A) \le \nu(B^c) - \mu(B^c) = \mu(B) - \nu(B)$$

and thus for any set A we have $|\mu(A) - \nu(A)| \le \mu(B) - \nu(B)$. ■

- The total variation is also intimately related to the notion of coupling between probability measures. A coupling between the probability measure μ and ν is a probability measure $q(i,j)$ on the product space $S \times S$ such that

$$\sum_j q(i,j) = \mu(i) \quad \text{and} \quad \sum_i q(i,j) = \nu(j)$$

  i.e. the marginals of q are μ and ν.

- Coupling are nicely expressed in terms of random variables. We can think of $q(i,j)$ has the (joint) pdf of the rv (X, Y) where X has pdf μ and Y has pdf ν.

- There always exists a coupling since we can always hoose X and Y independent, i.e. $q(i,j) = \mu(i)\nu(j)$.

- On the oppposite extreme if $\mu = \nu$ then we can pick $X = Y$ has a coupling i.e $q(i,i) = \mu(i) = \nu(i)$ and $q(i,j) = 0$ if $i \ne j$.

9.2 Total variation and coupling

Theorem 9.2 (Coupling representation of total variation) We have

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf \{P\{X \neq Y\}\,;(X,Y) \text{ coupling of } \mu \text{ and } \nu\} \tag{9.2}$$

Proof. We first prove an inequality:

$$\begin{aligned}\mu(A) - \nu(A) = P\{X \in A\} - P\{Y \in A\} &\leq P\{X \in A\} - P\{X \in A, Y \in A\} \\ = P\{X \in A, Y \notin A\} &\leq P\{X \neq Y\}\end{aligned}$$

and thus $\|\mu - \mu\|_{\mathrm{TV}} \leq \inf P\{X \neq Y\}$.

To show the equality we construct an optimal coupling.

Recall from Theorem 9.1 that $\|\mu - \nu\|_{\mathrm{TV}} = \sum_{\mu(i)\geq\nu(i)} \mu(i) - \nu(i) = \sum_{\nu(i)\geq\mu(i)} \nu(i) - \mu(i)$ (see regions A and B in Figure 9.1) and we set

$$p = \sum_i \mu(i) \wedge \nu(i) = 1 - \|\mu - \nu\|_{\mathrm{TV}}$$

(see region C in Figure 9.1).

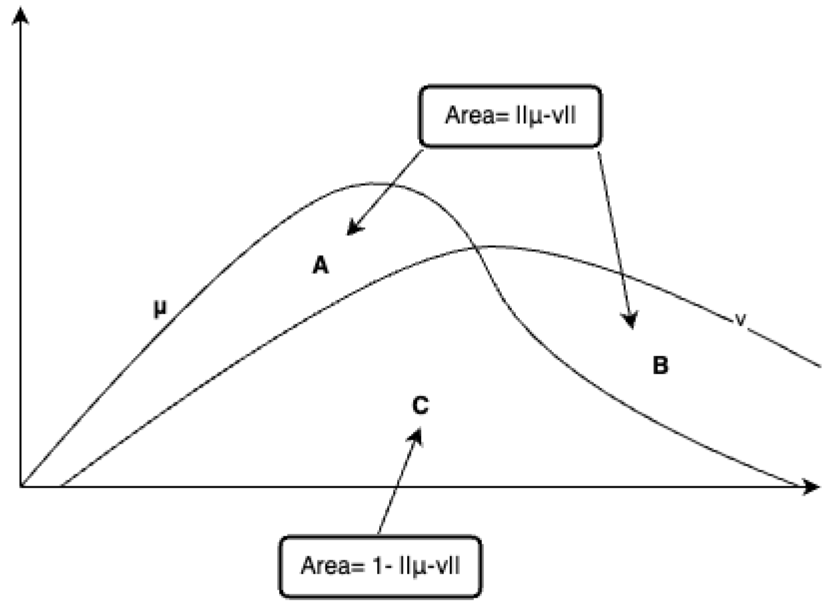


Figure 9.1: schematics of the optimal coupling

Consider now the coupling defined as follows. Pick a random number U

If $U \le p$ then let Z be the RV with pdf

$$\gamma_C(i) = \frac{\mu(i) \wedge \nu(i)}{p}$$

and set $X = Y = Z$.

If $U > p$ then let X be the random variable with pdf $\gamma_A(i)$ and Y be the random variable with pdf $\gamma_B(i)$ where

$$\gamma_A(i) = \begin{cases} \frac{\mu(i)-\nu(i)}{\|\mu-\nu\|_{\rm TV}} & \mu(i) \ge \nu(i) \\ 0 & \text{otherwise} \end{cases} \qquad \gamma_B(i) = \begin{cases} \frac{\nu(i)-\mu(i)}{\|\mu-\nu\|_{\rm TV}} & \nu(i) \ge \mu(i) \\ 0 & \text{otherwise} \end{cases}$$

Since $p\gamma_C + (1-p)\gamma_A = \mu$ and $p\gamma_C + (1-p)\gamma_B = \nu$ this defines a coupling and we have $P\{X \neq Y\} = 1 - p = \|\mu - \nu\|_{\rm TV}$. ■.

9.3 Coupling of Markov chains

- A coupling of a Markov chain is a stochastic process (X_n, Y_n) with state space $S \times S$ such that both X_n and Y_n are Markov chains with transition matrix P but with possibly different initial conditions.

- A Markovian coupling is a coupling such that the joint (X_n, Y_n) is itself a Markov chain with some transition matrix $Q((i,j),(k,l))$ which must then satisfy

$$\sum_k Q((i,j),(k,l)) = P(j,l) \quad \text{and} \quad \sum_l Q((i,j),(k,l)) = P(i,k)\,.$$

- In Theorem 2.3 where we proved the convergence of aperiodic irreducible Markov chains to their stationary distribution using an independent coupling. That is we pick $Q((i,j),(k,l)) = P(i,k)P(j,l)$.

- Another useful coupling is the common noise coupling. To understand this recall that any Markov chain has the representation $X_n = f(X_{n-1}, Z_n)$ where the Z_n are IID random variable. For the common noise coupling we use

$$X_n = f(X_{n-1}, Z_n) \qquad Y_n = f(X_{n-1}, Z_n)$$

  where we use the same noise Z_n to evolve both X_n and Y_n.

- The coupling time is defined by

$$\sigma = \inf\{n \ge 0 \,:\, X_n = Y_n\}$$

  that is, it is the first time that the Markov chain visit the same state.

- After the coupling time σ, X_n and Y_n have the same distribution (by the strong Markov property) so we can always modified a coupling so that, after time σ, X_n and Y_n are moving together, i.e. $X_n = Y_n$ for all $n \ge \sigma$ (that is we use the common noise coupling). We will *always* assume this to be true in what follows.

9.4 Speed of convergence via coupling

Theorem 9.3 Suppose (X_n, Y_n) is a coupling of a Markov chain such that $X_0 = i$ and $Y_0 = j$ and σ is the coupling time. Then we have

$$\|P^n(i,\cdot) - P^n(j,\cdot)\|_{\mathrm{TV}} \le P\left\{\sigma > n | X_0 = i, Y_0 = j\right\}$$

Proof. We have $P\{X_n = l | X_0 = i\} = P^n(i,l)$ and $P\{Y_n = l | Y_0 = j\} = P^n(j,l)$ and therefore X_n and Y_n is a coupling of the the probability distributions $P^n(i,\cdot)$ and $P^n(j,\cdot)$. So by Theorem 9.2 we have

$$\|P^n(i,\cdot) - P^n(j,\cdot)\|_{\mathrm{TV}} \le P\left\{X_n \neq Y_n | X_0 = i, Y_0 = j\right\} = P\left\{\sigma > n | X_0 = i, Y_0 = j\right\} .$$

■

We can use this result to bound the distance to the stationary measure

Theorem 9.4 We have

$$\sup_{i \in S} \|P^n(i,\cdot) - \pi\|_{\mathrm{TV}} \le \sup_{i,j \in S} \|P^n(i,\cdot) - P^n(j,\cdot)\|_{\mathrm{TV}}$$

Proof. Using the stationarity and the triangle inequality we have

$$\begin{aligned}\|P^n(i,\cdot)-\pi\|_{\mathrm{TV}} &= \sup_A |P^n(i,A)-\pi(A)| \\ &= \sup_A |\sum_j \pi(j)(P^n(i,A)-P^n(j,A))| \\ &\le \sum_j \pi(j) \sup_A |P^n(i,A)-P^n(j,A)| \\ &= \sum_j \pi(j) \|P^n(i,\cdot)-P^n(j,\cdot)\|_{TV} \\ &\le \sup_{i,j\in S} \|P^n(i,\cdot)-P^n(j,\cdot)\|_{TV} \quad \blacksquare\end{aligned}$$

We set $d(n) = \sup_i \|P^n(i,\cdot)-\pi\|_{\mathrm{TV}}$ which is the maximal distance to stationarity starting from an arbitrary initial state. (It is not hard to see that $\|\mu P^n - \pi\|_{\mathrm{TV}} \le d(n)$ for abitrary initial distribution μ as well).
We define then the mixing time $t_{\mathrm{mix}}(\varepsilon)$ of a Markov chain to be

$$t_{\mathrm{mix}}(\varepsilon) = \min\{n, d(n) \le \varepsilon\}\,.$$

That is, if $n > t_{\mathrm{mix}}(\varepsilon)$ then μP^n is less than ε close to the stationary distribution.

9.5 Mixing time for the sucess run chain

We start with a countable state space example, the success run chain which is very easy to analyze.

Parenthetically, it should be noted that the supremum over i in $d(n)$ is often not well suited for countable state space. It may often happen that the number of steps it take to be close to the stationary distribution may depend on where you start.

We consider a special case of Example 4.4 with constant succes probability.

$$P(n,0) = (1-p)\,, \quad P(n,n+1) = p\,, \quad n = 0,1,2,\cdots$$

Suppose $X_0 = i$ and $Y_0 = j$ (with $i \neq j$) we couple the two chains by moving them together.

- Pick a random number U, if $U \leq (1-p)$ then set $X_1 = Y_1 = 0$ the coupling time $\sigma = 1$.
- If $U \geq 1-p$ then move $X_1 = i+1, Y_1 = j+1$.

Clearly we have

$$P(\sigma > n | X_0 = i, Y_0 = j) = p^n$$

and thus we find for the mixing time

$$d(n) = p^n < \varepsilon \iff n \geq \frac{\ln(p)}{\ln \varepsilon}$$

9.6 Mixing time for the random walk on the hypercube

Recall that for the random walk on the hypercube (see Example 1.4) a state is a d-bit $\sigma = (\sigma_1, \cdots, \sigma_d)$ with $\sigma_j \in \{0,1\}$.

The Markov chain is periodic with period 2 so to make it aperiodic we consider its "lazy" version in which we do not move with probability $1/2$, that is we consider instead the transition matrix $\frac{P+I}{2}$ which makes it periodic. The Markov chain moves then as follows, we pick a coordinate σ_j at random and then replace by a random bit 0 or 1.

To couple the Markov chain we simply move *move them together*: If $X_n = \sigma$ and $Y_n = \sigma'$ we pick a coordinate $j \in \{1, \cdots, d\}$ at random and then replace both σ_j and σ'_j by the *same* random bit. This is a coupling and after a move the chosen coordinates coincide.

Under this coupling X_n and Y_n will get closer to each other if we select a j such that $\sigma_j \neq \sigma'_j$ and we will couple when all the coordinates have been selected. The distribution of the coupling time is thus exactly the same as the time τ need to collect d toys in the coupon collector problem.

We now get a bound on the tail as follows. We denote by A_i the events that coordinate i has not been selected after m steps. We have, using the inequality $(1-x) \le e^{-x}$

$$P\{\sigma > n\} = P\left(\bigcup_{i=1}^{d} A_i\right) \le \sum_{i=1}^{d} P(A_i) = \sum_{i=1}^{d} \left(1 - \frac{1}{d}\right)^n \le de^{-\frac{n}{d}}$$

So if we pick $n = d\ln(d) + cd$ we find

$$P\{\sigma > d\ln(d) + cd\} \le de^{-\frac{d\ln(d)+cd}{d}} = e^{-c} \implies t_{\text{mix}}(\varepsilon) \le d\ln(d) + \ln(\varepsilon^{-1})d\,.$$

9.7 Exercise

Exercise 9.1 (Properties of the total variation norm)

1. Prove the following formula for the total variation norm

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{f\,:\,\|f\|_\infty = \sup_i |f(i)| \le 1} |\mu f - \nu f|$$

2. Prove that the map $(\mu, \nu) \mapsto \|\mu - \nu\|_{\mathrm{TV}}$ is a (jointly) convex in $(\mu\,\nu)$.
3. Prove that $\|\mu P^n - \nu P^n\|_{TV} \le \sum_{i,j} \mu(i)\nu(j)\|P^n(i,\cdot) - P^n(j,\cdot)\|_{\mathrm{TV}}$

Exercise 9.2 (Examples of optimal coupling)

1. Suppose X and Y are two Bernoulii random variables with paramter p and q. Using random number U we can represent $X = 1_{U<p}$. Consider the coupling between X and Y by using the same random variable U to generate X and Y. Compute the joint distribution of (X, Y).
2. What is the optimal coupling between X and Y in the sense of Theorem 9.2

Exercise 9.3 (Coupling for card shuffling) We consider the following shuffling procedure: given a deck of 52 cards pick one number j at random between 1 and 52 and take the card in poistion j and move itto the top of the card (this is called the random to top shuffling).

Now consider the following coupling: you have two decks of cards. Do a random to top shuffling for the first deck. Look at which card you just put on top (say a seven of spades). Now in the second deck take the seven of spades and put it on top.

Show that this is a coupling and use this to show that after $n \ln(n/\epsilon)$ the random to top shuffling is at total variation distance no more than ϵ from the stationary distribution.

:::

:::

9.8 References

Hajek, Bruce. 1988. "Cooling Schedules for Optimal Annealing." *Mathematics of Operations Research* 13 (2): 311–29. http://www.jstor.org/stable/3689827.

Lawler, Gregory F. 2006. *Introduction to Stochastic Processes.*

Levin, David Asher, James Propp, David B. Wilson, Elizabeth L. Wilmer, and Y. Peres. 2017. *Markov Chains and Mixing Times.* American Mathematical Society.