integers $n \geq 0$ and all states $i_0, i_1, \dots, i_{n-1}, i, j$,

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i) \tag{1.1}$$

whenever both sides are well-defined, this stochastic process is called a *Markov chain*. It is called a *homogeneous* Markov chain (HMC) if in addition, the right-hand side of (1.1) is independent of n.

Property (1.1) is the *Markov property*. The matrix $\mathbf{P} = \{p_{ij}\}_{i,j \in E}$, where

$$p_{ij} = P(X_{n+1} = j \mid X_n = i), \tag{1.2}$$

is the *transition matrix* of the HMC. Since its entries are probabilities, and since a transition from any state i must be to *some* state, it follows that

$$p_{ij} \geq 0, \sum_{k \in E} p_{ik} = 1$$

for all states i, j. A matrix $\mathbf{P}$ indexed by E and satisfying the above properties is called a *stochastic matrix*. The state space may be infinite, and therefore such a matrix is in general not of the kind studied in linear algebra. However, the basic operations of addition and multiplication will be defined by the same formal rules. For instance, with $A = \{a_{ij}\}_{i,j \in E}$ and $B = \{b_{ij}\}_{i,j \in E}$, the product $C = AB$ is the matrix $\{c_{ij}\}_{i,j \in E}$, where $c_{ij} = \sum_{k \in E} a_{ik} b_{kj}$. The notation $x = \{x_i\}_{i \in E}$ formally represents a *column* vector, and x^T is a row vector, the transpose of x. For instance, $y = \{y_i\}_{i \in E}$ given by $y^T = x^T A$ is defined by $y_i = \sum_{k \in E} x_k a_{ki}$. Similarly, $z = \{z_i\}_{i \in E}$ given by $z = Ax$ is defined by $z_i = \sum_{k \in E} a_{ik} z_k$.

Proving the Markov property is not, in general, a difficult task, and Theorems 2.1 and 2.2 below will suffice in most situations. However, there are cases outside their scope, and the following one is quite important, both in theory and in applications.

Example 1.1. *Machine Replacement*
Let $\{U_n\}_{n \geq 1}$ be a sequence of i.i.d random variables taking their values in $\{1, 2, \dots, +\infty\}$. The random variable U_n can be interpreted as the lifetime of some machine, the nth one, which is replaced by the $(n+1)$st one upon failure. Thus at time 0, machine 1 is put in service until it breaks down at time U_1, whereupon it is immediately replaced by machine 2, which breaks down at time $U_1 + U_2$, and so on. The elapsed time in service of the current machine at time n is denoted by X_n. Thus, the process $\{X_n\}_{n \geq 0}$ takes its values in $E = \mathbb{N}$ and increases linearly from 0 at time $R_k = \sum_{i=1}^{k} U_i$ to $U_{k+1} - 1$ at time $R_{k+1} - 1$.

The sequence $\{R_k\}_{k \geq 0}$ defined in this way, with $R_0 = 0$, is called a *renewal sequence*, and X_n is called the *backward recurrence time* at time n (see Fig. 2.1.1). There is a rich and useful theory associated with renewal sequences, the so-called *renewal theory*. It will be developed in Chapter 4.

The process $\{X_n\}_{n \geq 0}$ is an HMC with state space $E = \mathbb{N}$, and the nonnull entries of its transition matrix are of the form $p_{i,i+1}$ and $p_{i,0} = 1 - p_{i,i+1}$, where

$$p_{i,i+1} = \frac{P(U_1 > i+1)}{P(U_1 > i)}. \tag{1.3}$$

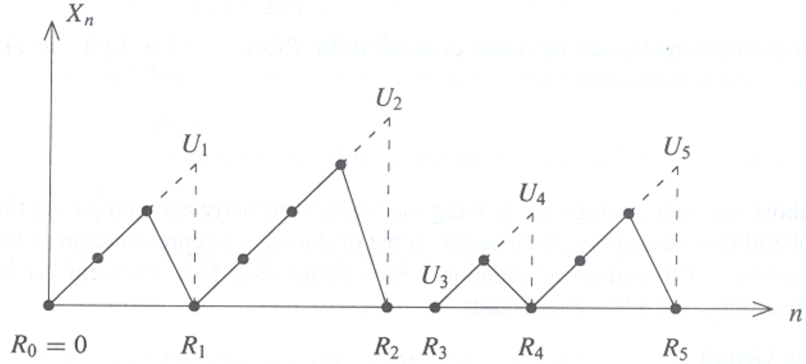


Figure 2.1.1. Backward recurrence time

To prove this, one must first verify (1.1), that is, writing $B = \{X_0 = i_0, \dots, X_{n-1} = i_{n-1}\}$,

$$\frac{P(X_{n+1} = j, X_n = i, B)}{P(X_n = i, B)} = \frac{P(X_{n+1} = j, X_n = i)}{P(X_n = i)}$$

for sequences $i_0, \dots, i_{n-1}, i, j$ such that $P(B, X_n = i, X_{n+1} = j) > 0$. In particular, $j = i + 1$ or 0, and $i_{n-1} = i - 1, \dots, i_{n-i} = 0$.

Let $\nu(n)$ be the number of renewal times R_k in the interval $[1, n]$. Taking, for instance, $j = i + 1$, and writing $D = \{X_{n-i-1} = i_{n-i-1}, \dots, X_0 = i_0\}$, we have

$$\begin{aligned} &P(X_{n+1} = j, X_n = i, B) \\ &\quad = P(X_{n+1} = i+1, X_n = i, X_{n-1} = i-1, \dots, X_{n-i} = 0, D) \\ &\quad = \sum_{k=0}^{\infty} P(X_{n+1} = i+1, X_n = i, X_{n-1} = i-1, \dots, X_{n-i} = 0, D, \nu(n) = k). \end{aligned}$$

The general term in the latter sum equals $P(U_{k+1} > i+1, R_k = n-i, D) = P(U_{k+1} > i+1) P(R_k = n-i, D) = P(U_1 > i+1) P(R_k = n-i, D)$. The independence of $\{U_n\}_{n \geq 1}$ has been used for the first equality, and the identity of the distributions of U_{k+1} and U_1 for the second one. Therefore,

$$P(X_{n+1} = i+1, X_n = i, B) = P(U_1 > i+1) \left(\sum_{k=0}^{\infty} P(R_k = n-i, D) \right).$$

Similar computations yield

$$P(X_n = i, B) = P(U_1 > i) \left(\sum_{k=0}^{\infty} P(R_k = n-i, D) \right),$$

so that

$$P(X_{n+1} = i+1 \mid X_n = i, B) = \frac{P(U_1 > i+1)}{P(U_1 > i)}.$$

The same calculations lead to the same evaluation for $P(X_{n+1} = i+1 \mid X_n = i)$. This proves the announced results. □

◇

The above example is atypical. Proving the Markov property and computing the transition probabilities are usually much easier. Most of the time, a representation of the state process in terms of a recurrence equation makes things easy (see Theorem 2.1 below). Nevertheless, there are a few tough cases.

Transition Graph

A transition matrix **P** is sometimes represented by its *transition graph* G, a graph having for nodes (or vertices) the states of E. This graph has an oriented edge from i to j if and only if $p_{ij} > 0$, in which case this edge is adorned with the label p_{ij}.

The transition graph of the Markov chain of Example 1.1 is shown in Figure 2.1.2, where

$$p_i = \frac{P(U_1 = i+1)}{P(U_1 > i)}.$$

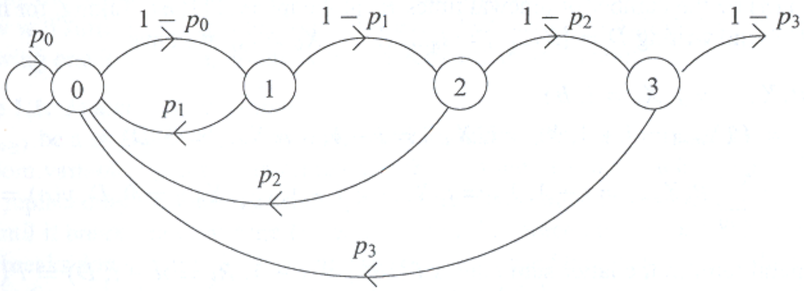


Figure 2.1.2. Transition graph of the backward recurrence chain

1.2 Distribution of an HMC

The random variable X_0 is called the *initial state*, and its probability distribution ν,

$$\nu(i) = P(X_0 = i), \tag{1.4}$$

is the *initial distribution*. From Bayes's sequential rule, $P(X_0 = i_0, X_1 = i_1, \ldots, X_k = i_k) = P(X_0 = i_0)P(X_1 = i_1 \mid X_0 = i_0) \cdots P(X_k = i_k \mid X_{k-1} = i_{k-1}, \ldots, X_0 = i_0)$, and

therefore, in view of the homogeneous Markov property and the definition of the transition matrix,

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_k = i_k) = \nu(i_0)p_{i_0 i_1} \cdots p_{i_{k-1} i_k}. \tag{1.5}$$

The data (1.5) for all $k \geq 0$, all states $i_0, i_1, \ldots, i_k$, constitute the *probability law*, or *distribution* of the HMC. Therefore we have the following result.

Theorem 1.1. *Distribution of an HMC*
The distribution of a discrete-time HMC is determined by its initial distribution and its transition matrix.

The distribution at time n of the chain is the vector ν_n, where

$$\nu_n(i) = P(X_n = i). \tag{1.6}$$

From Bayes's rule of exclusive and exhaustive causes, $\nu_{n+1}(j) = \sum_{i \in E} \nu_n(i) p_{ij}$, that is, in matrix form, $\nu_{n+1}^T = \nu_n^T \mathbf{P}$. Iteration of this equality yields

$$\nu_n^T = \nu_0^T \mathbf{P}^n. \tag{1.7}$$

The matrix $\mathbf{P}^n$ is called the *n-step transition matrix* because its general term is

$$p_{ij}(m) = P(X_{n+m} = j \mid X_n = i). \tag{1.8}$$

Indeed, using Bayes's sequential rule and the Markov property, one finds for the right-hand side of the latter equality

$$\sum_{i_1, \ldots, i_{m-1} \in E} p_{ii_1} p_{i_1 i_2} \cdots p_{i_{m-1} j},$$

and this is the general term of the mth power of **P**.

The Markov property (1.1) extends to

$$P(X_{n+1} = j_1, \ldots, X_{n+k} = j_k \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)$$
$$= P(X_{n+1} = j_1, \ldots, X_{n+k} = j_k \mid X_n = i)$$

for all $i_0, \ldots, i_{n-1}, i, j_1, \ldots, j_k$ such that both sides of the equality are defined (Problem 2.1.2). Writing

$$A = \{X_{n+1} = j_1, \ldots, X_{n+k} = j_k\},\ B = \{X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\},$$

the last equality reads $P(A \mid X_n = i, B) = P(A \mid X_n = i)$, which is in turn equivalent to

$$P(A \cap B \mid X_n = i) = P(A \mid X_n = i)P(B \mid X_n = i). \tag{1.9}$$

In words: The future at time n and the past at time n are conditionally independent given the present state $X_n = i$. This shows in particular that the Markov property is independent of the direction of time.

Notation We shall abbreviate $P(A \mid X_0 = i)$ as $P_i(A)$. If μ is a probability distribution on E, then $P_\mu(A) = \sum_{i \in E} \mu(i) P_i(A)$ is the probability of A given that the initial state is distributed according to μ.

2 Markov Recurrences

2.1 A Canonical Representation

Many HMCs receive a natural description in terms of a recurrence equation driven by white noise.

Theorem 2.1. *HMCs Driven by White Noise.*
Let $\{Z_n\}_{n\geq 1}$ be an i.i.d sequence of random variables with values in an arbitrary space F. Let E be a countable space, and $f : E \times F \to E$ be some function. Let X_0 be a random variable with values in E, independent of $\{Z_n\}_{n\geq 1}$. The recurrence equation

$$X_{n+1} = f(X_n, Z_{n+1}) \tag{2.1}$$

then defines an HMC.

(The phrase *white noise* comes from signal theory and refers to the driving sequence $\{Z_n\}_{n\geq 1}$.)

Proof. Iteration of recurrence (2.1) shows that for all $n \geq 1$, there is a function g_n such that $X_n = g_n(X_0, Z_1, \dots, Z_n)$, and therefore $P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(f(i, Z_{n+1}) = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(f(i, Z_{n+1}) = j)$, since the event $\{X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i\}$ is expressible in terms of $X_0, Z_1, \dots, Z_n$ and is therefore independent of Z_{n+1}. Similarly, $P(X_{n+1} = j \mid X_n = i) = P(f(i, Z_{n+1}) = j)$. We therefore have a Markov chain, and it is homogeneous, since the right-hand side of the last equality does not depend on n. Explicitly

$$p_{ij} = P(f(i, Z_1) = j). \tag{2.2}$$

□

Not all homogeneous Markov chains are naturally described by the model of Theorem 2.1. A slight modification of Theorem 2.1, however, considerably enlarges its scope.

Theorem 2.2.
Let things be as in Theorem 2.1 except for the statistics of $X_0, Z_1, Z_2, \dots$. Suppose instead that for all $n \geq 0$, Z_{n+1} is conditionally independent of $Z_n, \dots, Z_1, X_{n-1}, \dots, X_0$ given X_n, that is, for all $k, k_1, \dots, k_n \in F$, $i_0, i_1, \dots, i_{n-1}, i \in E$,

$$\begin{aligned} P(Z_{n+1} &= k \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0, Z_n = k_n, \dots, Z_1 = k_1) \\ &= P(Z_{n+1} = k \mid X_n = i), \end{aligned}$$

where the latter quantity is independent of n. Then $\{X_n\}_{n\geq 0}$ is an HMC, with transition matrix $\mathbf{P}$ given by

$$p_{ij} = P(f(i, Z_1) = j \mid X_0 = i).$$

Proof. The proof is analogous to that of Theorem 2.1 (Problem 2.2.6). □

Remark 2.1. Not all homogeneous Markov chains receive a "natural" description of the type featured in Theorems 2.1 and 2.2, as Example 1.1 (machine replacement) shows. However, for any transition matrix $\mathbf{P}$ on E, there exists a homogeneous Markov chain $\{X_n\}_{n\geq 0}$ with this transition matrix and with a representation such as in Theorem 2.1, namely,

$$X_{n+1} = j \text{ if } Z_{n+1} \in [\sum_{k=0}^{j-1} p_{X_n k}, \sum_{k=0}^{j} p_{X_n k}],$$

where $\{Z_n\}_{n\geq 1}$ is i.i.d, uniform on $[0, 1]$. We can apply Theorem 2.1, and check that this HMC has the announced transition matrix. This artificial representation is useful for simulating small Markov chains and can also be helpful for the theory. ◇

2.2 A Few Famous Examples

The examples below will often be used to illustrate the theory.

Example 2.1. *1-D Random Walk*
Let X_0 be a random variable with values in $\mathbb{Z}$. Let $\{Z_n\}_{n\geq 1}$ be a sequence of i.i.d random variables, independent of X_0, taking the values $+1$ or -1, and with the probability distribution

$$P(Z_n = +1) = p,$$

where $p \in (0, 1)$. The process $\{X_n\}_{n\geq 1}$ defined by

$$X_{n+1} = X_n + Z_{n+1} \tag{2.3}$$

is, in view of Theorem 2.1, an HMC, called the *random walk* on $\mathbb{Z}$. ◇

Example 2.2. *Repair Shop*
During day n, Z_{n+1} machines break down, and they enter the repair shop on day $n + 1$. Every day one machine among those waiting for service is repaired. Therefore, denoting by X_n the number of machines in the shop on day n,

$$X_{n+1} = (X_n - 1)^+ + Z_{n+1}, \tag{2.4}$$

where $a^+ = \max(a, 0)$. In particular, if $\{Z_n\}_{n\geq 1}$ is an i.i.d sequence independent of the initial state X_0, then $\{X_n\}_{n\geq 0}$ is a homogeneous Markov chain. In terms of the probability distribution

$$P(Z_1 = k) = a_k,\ k \geq 0, \tag{2.5}$$

its transition matrix is

$$\mathbf{P} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}. \tag{2.6}$$

Indeed, from (2.2) and (2.5),

$$p_{ij} = P((i-1)^+ + Z_1 = j) = P(Z_1 = j - (i-1)^+) = a_{j-(i-1)^+}. \qquad \diamond$$

Example 2.3. *Inventory*
A given commodity is stocked in order to statisfy a continuing demand. The aggregated demand between time n and time $n+1$ is Z_{n+1} units, and it is assumed that $\{Z_n\}_{n\geq 1}$ is i.i.d, and independent of the initial value X_0 of the stock. Replenishment of the stock takes place at times $n+0$ (that is to say, immediately after time n) for all $n \geq 1$.

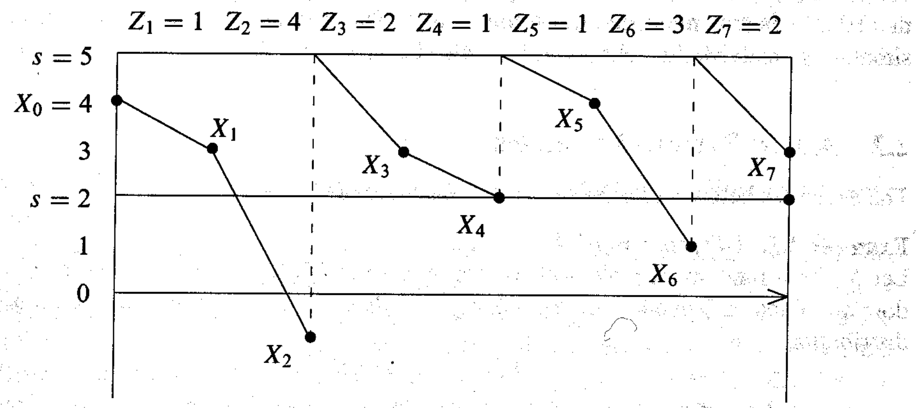


Figure 2.2.1. A sample path of the inventory Markov chain

A popular management strategy is the so-called (s, S)-strategy, where s and S are integers such that $0 < s < S$. Under this inventory policy, if the level of the stock at time n is found not larger than s, then it is brought to level S at time $n+0$. Otherwise, nothing is done. The initial stock X_0 is assumed not greater than S, and therefore $\{X_n\}_{n\geq 1}$ takes its value in $E = \{S, S-1, S-2, \ldots\}$. (See Fig. 2.2.1.) Negative values of the stock are allowed, with the interpretation that an unfilled demand is immediately satisfied upon restocking. With the above rules of operation, the evolution of the stock is governed by the dynamic equation

$$X_{n+1} = \begin{cases} X_n - Z_{n+1} & \text{if } s < X_n \leq S, \\ S - Z_{n+1} & \text{if } X_n \leq s. \end{cases} \tag{2.7}$$

In view of (2.7) and Theorem 2.1, $\{X_n\}_{n\geq 1}$ is a homogeneous Markov chain. $\diamond$

Example 2.4. *Branching Process*
In this model $Z_n = (Z_n^{(1)}, Z_n^{(2)}, \ldots)$, where the random variables $\{Z_n^{(j)}\}_{n\geq 1, j\geq 1}$ are i.i.d and integer valued. The recurrence equation

$$X_{n+1} = \sum_{k=1}^{X_n} Z_{n+1}^{(k)}, \tag{2.8}$$

with the convention $X_{n+1} = 0$ if $X_n = 0$, receives the following interpretation: X_n is the number of individuals in the nth generation of a given population (humans, particles, etc.). Individual number k of the nth generation gives birth to $Z_{n+1}^{(k)}$ descendents, and this accounts for (2.8).

If the number X_0 of ancestors is independent of $\{Z_n\}_{n\geq 1}$, then according to Theorem 2.1, $\{X_n\}_{n\geq 0}$ is a homogeneous Markov chain, called a branching process because of the genealogical tree that it generates (see Figure 2.2.2). The branching process is also known as the *Galton–Watson process*.

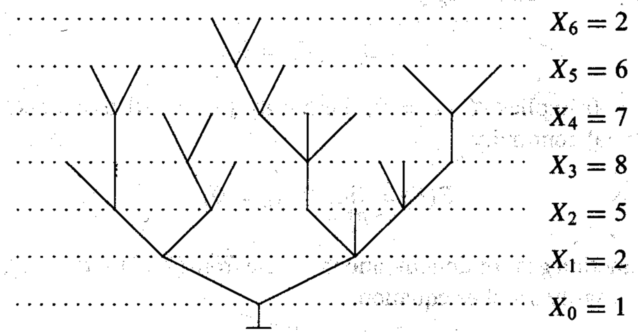


Figure 2.2.2. Sample tree of a branching process

We seek to obtain the probability of extinction of the population. For this we shall introduce g, the common generating function of the variables $Z_n^{(k)}$. The recurrence equation (2.8) provides a recurrent equation for the generating function of the number of individuals in the nth generation,

$$\psi_n(z) = E[z^{X_n}]. \tag{2.9}$$

Indeed,

$$\psi_{n+1}(z) = E[z^{X_{n+1}}] = E[z^{\sum_{k=1}^{X_n} Z^{(k)}}],$$

where $Z_{n+1}^{(k)}$ was denoted by $Z^{(k)}$ for simplicity. Since X_n is a functional of $X_0, Z_1, \ldots, Z_n$, it is independent of Z_{n+1}, and therefore, in the latter equality, X_n is independent of $Z^{(k)}$. From a computation made in Chapter 1, Exercise 5.1,

$$E\left[z^{\sum_{k=1}^{X_n} Z^{(k)}}\right] = \psi_n(g(z)).$$

Therefore,

$$\psi_{n+1}(z) = \psi_n(g(z)).$$

Iterating this equality, we obtain

$$\psi_{n+1}(z) = \psi_0(g^{(n+1)}(z)),$$

here $g^{(n)}$ is the nth iterate of g. If there is only *one ancestor*, then $\psi_0(z) = z$, and therefore $_{n+1}(z) = g^{(n+1)}(z) = g(g^{(n)}(z))$, that is,

$$\psi_{n+1}(z) = g(\psi_n(z)). \tag{2.10}$$

n particular, since $\psi_n(0) = P(X_n = 0)$, we have

$$P(X_{n+1} = 0) = g(P(X_n = 0)), \tag{2.11}$$

n equality that we shall use to discuss extinction. The event $\mathcal{E}$ = "an extinction occurs" is st "at least one generation is empty," that is,

$$\mathcal{E} = \cup_{n=1}^{\infty}\{X_n = 0\}.$$

lso, since $X_n = 0$ implies $X_{n+1} = 0$, the family $\{X_n = 0\}$ is nondecreasing, and by onotone sequential continuity,

$$P(\mathcal{E}) = \lim_{n\uparrow\infty} P(X_n = 0). \tag{2.12}$$

he generating function g is continuous, and therefore from (2.11) and (2.12), the probability f extinction necessarily satisfies equation

$$P(\mathcal{E}) = g(P(\mathcal{E})). \tag{2.13}$$

et Z be any of the random variables $Z_n^{(k)}$. It was shown in Chapter 1, Theorem 5.1, that xcluding the trivial cases where $P(Z = 0) = 1$ or $P(Z \geq 2) = 0$,

(α) if $E[Z] \leq 1$, the only solution of $x = g(x)$ in $[0, 1]$ is 1, and therefore $P(\mathcal{E}) = 1$. he branching process eventually becomes extinct.

(β) if $E[Z] > 1$, there are two solutions of $x = g(x)$ in $[0, 1]$, 1 and x_0 such that $< x_0 < 1$. From the strict convexity of $f : [0, 1] \to [0, 1]$, it follows that the sequence $= P(X_n = 0)$ that satisfies $y_0 = 0$ and $y_{n+1} = g(y_n)$ converges to x_0. Therefore, when e mean number of descendants $E[Z]$ is strictly larger than 1, then $P(\mathcal{E}) \in (0, 1)$, and in rticular, there is a nonnull probability of extinction.

Here is a small example. Suppose that

$$g(z) = \frac{1}{4} + \frac{1}{4}z + \frac{1}{2}z^2,$$

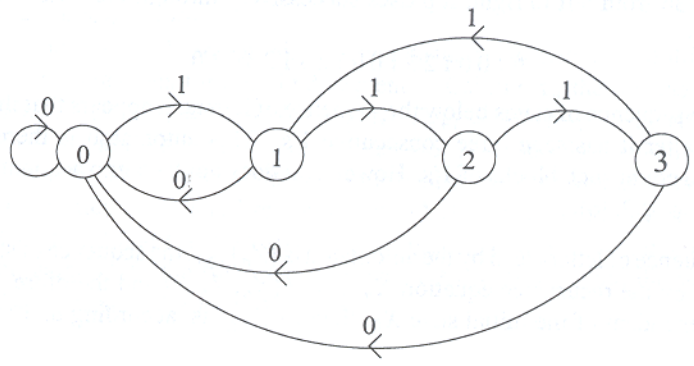
at is, the probabilities of having 0, 1, or 2 sons are respectively $\frac{1}{4}$, $\frac{1}{4}$, and $\frac{1}{2}$. In particular, $[Z] = 1.25 > 1$, and $P(\mathcal{E})$ is the solution strictly between 0 and 1 of

$$x = \frac{1}{4} + \frac{1}{4}x + \frac{1}{2}x^2,$$

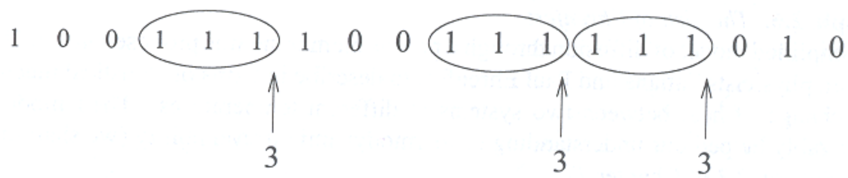
at is, $P(\mathcal{E}) = \frac{1}{2}$. ◇

Example 2.5. *Stochastic Automata*

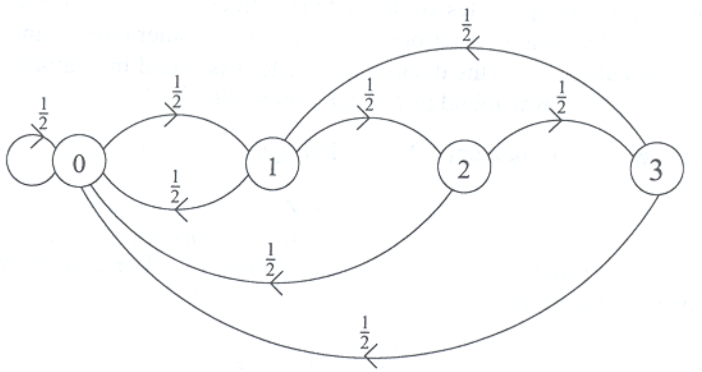
A finite automaton $(E, \mathcal{A}, f)$ can read sequences of letters from a finite alphabet $\mathcal{A}$ written on some infinite tape. It can be in any state of a finite set E, and its evolution is governed by a function $f : E \times \mathcal{A} \to E$, as follows. When the automaton is in state $i \in E$ and reads letter $a \in \mathcal{A}$, it switches from state i to state $j = f(i, a)$ and then reads on the tape the next letter to the right.



a



b



c

Figure 2.2.3. The automaton: the recognition process and the Markov chain

An automaton can be represented by its transition graph G having for nodes the states of E. There is an oriented edge from the node (state) i to the node j if and only if there exists $a \in \mathcal{A}$ such that $j = f(i, a)$, and this edge then receives label a. If $j = f(i, a_1) = f(i, a_2)$ for $a_1 \neq a_2$, then there are two edges from i to j with labels a_1 and a_2, or, more economically, one such edge with label (a_1, a_2). More generally, a given oriented edge can have multiple labels of any order.

Consider, for instance, the automaton with alphabet $\mathcal{A} = \{0, 1\}$ corresponding to the transition graph of Figure 2.2.3a. As the automaton, initialized in state 0, reads the sequence of Figure 2.2.3b from left to right, it passes successively through the states (including the initial state 0)

$$0\,1\,0\,0\,1\,2\,3\,1\,0\,0\,1\,2\,3\,1\,2\,3\,0\,1\,0\,.$$

Rewriting the sequence of states below the sequence of letters, it appears that the automaton is in state 3 after it has seen three consecutive 1's. This automaton is therefore able to recognize and count such blocks of 1's. However, it does not take into account overlapping blocks (see Fig. 2.2.3b).

If the sequence of letters read by the automaton is $\{Z_n\}_{n\geq 1}$, the sequence of states $\{X_n\}_{n\geq 0}$ is then given by the recurrence equation $X_{n+1} = f(X_n, Z_{n+1})$ and therefore, if $\{Z_n\}_{n\geq 1}$ is i.i.d and independent of the initial state X_0, then $\{X_n\}_{n\geq 1}$ is, according to Theorem 2.1 an HMC. ◇

Example 2.6. *The Urn of Ehrenfest*
This simplified model of diffusion through a porous membrane was proposed in 1907 by the Austrian physicists Tatiana and Paul Ehrenfest to describe in terms of statistical mechanics the exchange of heat between two systems at different temperatures. Their model also considerably helped our understanding of thermodynamic irreversibility (we shall discuss this in Section 2.3 of Chapter 4).

There are N particles that can be either in compartment A or in compartment B. Suppose that at time $n \geq 0$, $X_n = i$ particles are in A. One then chooses a particle at random, and this particle is moved at time $n+1$ from where it is to the other compartment. Thus, the next state X_{n+1} is either $i-1$ (the displaced particle was found in compartment A) with probability $\frac{i}{N}$, or $i+1$ (it was found in B) with probability $\frac{N-i}{N}$.

This model pertains to Theorem 2.2. For all $n \geq 0$,

$$X_{n+1} = X_n + Z_{n+1}, \tag{2.14}$$

where $Z_n \in \{-1, +1\}$ and $P(Z_{n+1} = -1 \mid X_n = i) = \frac{i}{N}$. The nonzero entries of the transition matrix are therefore

$$p_{i,i+1} = \frac{N-i}{N}, \quad p_{i,i-1} = \frac{i}{N}. \tag{2.15}$$

◇

3 First-Step Analysis

3.1 Absorption Probability

Many functionals of homogeneous Markov chains, in particular probabilities of absorption by a closed set (A is called *closed* if $\sum_{j\in A} p_{ij} = 1$ for all $i \in A$) and average times before absorption, can be evaluated by a technique called *first-step analysis*. This technique, which is the motor of most computations in Markov chain theory, is best illustrated by the following example.

Example 3.1. *Gambler's Ruin*
Two players A and B play "heads or tails", where heads occur with probability $p \in (0, 1)$ and the successive outcomes form an i.i.d sequence. Calling X_n the fortune in dollars of player A at time n, then $X_{n+1} = X_n + Z_{n+1}$, where $Z_{n+1} = +1$ (resp., -1) with probability p (resp., $q = 1 - p$), and $\{Z_n\}_{n\geq 1}$ is i.i.d. In other words, A bets \$1 on heads at each toss, and B bets \$1 on tails. The respective initial fortunes of A and B are a and b. The game ends when a player is ruined, and therefore the process $\{X_n\}_{n\geq 1}$ is a random walk as described in Example 2.1, except that it is restricted to $E = \{0, \dots, a, a+1, \dots, a+b = c\}$. The duration of the game is T, the first time n at which $X_n = 0$ or c, and the probability of winning for A is $u(a) = P(X_T = c \mid X_0 = a)$.

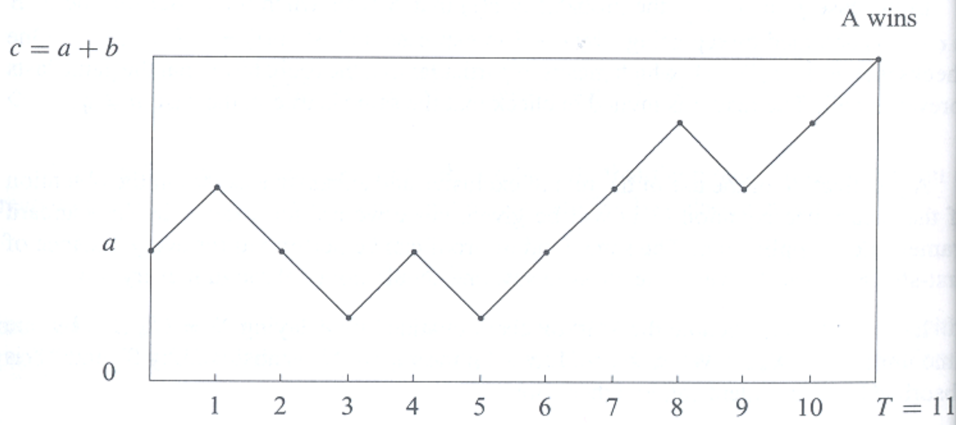


Figure 2.3.1. The basic random walk and the gambler's ruin

Instead of computing $u(a)$ alone, first-step analysis computes

$$u(i) = P(X_T = c \mid X_0 = i)$$

for all states $i \in [0, c]$, and for this, it first generates a recurrence equation for the $u(i)$'s by breaking down event "A wins" according to what can happen after the first step (the first toss) and using the rule of exclusive and exhaustive causes. If $X_0 = i \in [1, c-1]$, then $X_1 = i + 1$ (resp., $X_1 = i - 1$) with probability p (resp., q), and the probability of ruin of

Example 3.4. *Gambler's Ruin*
This example continues Example 3.1. The average duration $m(i) = E[T \mid X_0 = i]$ of the game when the initial fortune of player A is i satisfies the recurrence equation

$$m(i) = 1 + pm(i+1) + qm(i-1) \tag{3.4}$$

for $i \in [1, c-1]$. Indeed, the coin will be tossed at least once, and then with probability p (resp., q) the fortune of player A will be $i+1$ (resp., $i-1$), and therefore $m(i+1)$ (resp., $m(i-1)$) more tosses will be needed on average before one of the players goes broke. The boundary conditions are

$$m(0) = 0, m(c) = 0. \tag{3.5}$$

In order to solve (3.4) with the boundary conditions (3.5), write (3.4) in the form $-1 = p(m(i+1) - m(i)) - q(m(i) - m(i-1))$. Defining

$$y_i = m(i) - m(i-1),$$

we have, for $i \in [1, c-1]$,

$$-1 = py_{i+1} - qy_i \tag{3.6}$$

and

$$m(i) = y_1 + y_2 + \cdots + y_i. \tag{3.7}$$

We now solve (3.6) with $p = q = \frac{1}{2}$. From (3.6),

$$\begin{aligned} -1 &= \frac{1}{2}y_2 - \frac{1}{2}y_1, \\ -1 &= \frac{1}{2}y_3 - \frac{1}{2}y_2, \\ &\vdots \\ -1 &= \frac{1}{2}y_i - \frac{1}{2}y_{i-1}, \end{aligned}$$

and therefore, summing up,

$$-(i-1) = \frac{1}{2}y_i - \frac{1}{2}y_1,$$

that is, for $i \in [1, c]$,

$$y_i = y_1 - 2(i-1).$$

Reporting this expression in (3.7), and observing that $y_1 = m(1)$, we obtain

$$m(i) = im(1) - 2[1 + 2 + \cdots + (i-1)] = im(1) - i(i-1).$$

The boundary condition $m(c) = 0$ gives $cm(1) = c(c-1)$ and therefore, finally,

$$m(i) = i(c-i). \tag{3.8}$$

◇

First-step analysis leads to necessary conditions in the form of a system of linear equations. In the above examples, it turns out that the system in question has a unique solution, a situation that prevails when the state space is finite but that is not the general case with an infinite state space. The issue of uniqueness, and of which solution to choose in case of nonuniqueness, is addressed in Chapter 4, where absorption is studied in more detail.

4 Topology of the Transition Matrix

4.1 Communication

All the properties defined in the present section are *topological* in the sense that they concern only the *naked* transition graph (without the labels).

Definition 4.1. *Communication*
State j is said to be *accessible* from state i if there exists $M \geq 0$ such that $p_{ij}(M) > 0$. In particular, a state i is always accessible from itself, since $p_{ii}(0) = 1$. States i and j are said to *communicate* if i is accessible from j *and* j is accessible from i, and this is denoted by $i \leftrightarrow j$.

For $M \geq 1$, $p_{ij}(M) = \sum_{i_1,\dots,i_{M-1}} p_{ii_1} \cdots p_{i_{M-1}j}$, and therefore $p_{ij}(M) > 0$ if and only if there exists at least one path $i, i_1, \dots, i_{M-1}, j$ from i to j such that

$$p_{ii_1} p_{i_1i_2} \cdots p_{i_{M-1}j} > 0,$$

or, equivalently, if there is an oriented path from i to j in the transition graph G. Clearly,

$$\begin{aligned} i &\leftrightarrow i \quad &&\text{(reflexivity)}, \\ i \leftrightarrow j &\Rightarrow j \leftrightarrow i \quad &&\text{(symmetry)}, \\ i \leftrightarrow j, j \leftrightarrow k &\Rightarrow i \leftrightarrow k \quad &&\text{(transivity)}. \end{aligned}$$

Therefore, the communication relation ($\leftrightarrow$) is an equivalence relation, and it generates a partition of the state space E into disjoint equivalence classes called *communication classes*.

Definition 4.2. *Closed Sets*
A state i such that $p_{ii} = 1$ is called *closed*. More generally, a set C of states such that for all $i \in C$, $\sum_{j \in C} p_{ij} = 1$ is called *closed*.

Example 4.1.
The transition graph of Figure 2.4.1 has 3 communication classes: $\{1, 2, 3, 4\}$, $\{5, 7, 8\}$, and $\{6\}$. State 6 is closed. The communication class $\{5, 7, 8\}$ is not closed, but the set $\{5, 6, 7, 8\}$ is.
◇

Observe in this example that there may exist oriented edges linking two different communication classes E_k and E_ℓ. However, all the oriented edges between two communication classes have the same orientation (all from E_k to E_ℓ or all from E_ℓ to E_k). Why?

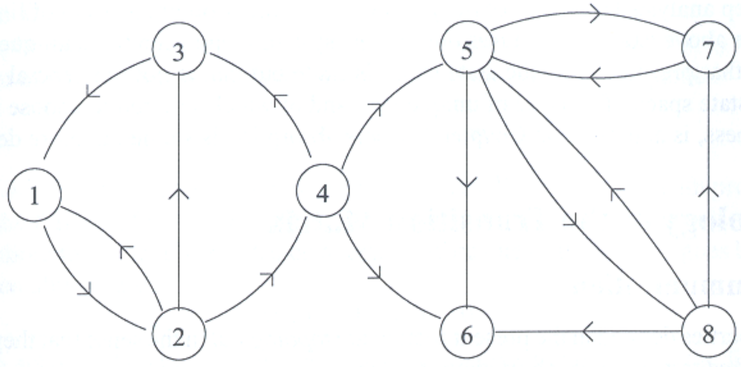


Figure 2.4.1. A transition graph with 3 communication classes

Definition 4.3. *Irreducibility*
If there exists only one communication class, then the chain, its transition matrix, and its transition graph, are said to be *irreducible*.

4.2 Period

Consider the random walk on $\mathbb{Z}$ (Example 2.1). Since $p \in (0, 1)$, it is irreducible. Observe that $E = C_0 + C_1$, where C_0 and C_1, the set of even and odd relative integers respectively, have the following property. If you start from $i \in C_0$ (resp., C_1), then in one step you can go only to a state $j \in C_1$ (resp., C_0). The chain $\{X_n\}$ passes alternately from one class to the other. In this sense, the chain has a periodic behavior, corresponding to the period 2. More generally, we have the following.

Theorem 4.1. *Cyclic Structure*
For any *irreducible* Markov chain, one can find a *unique partition* of E into d classes C_0, $C_1, \ldots, C_{d-1}$ such that for all k, $i \in C_k$,

$$\sum_{j \in C_{k+1}} p_{ij} = 1,$$

where by convention $C_d = C_0$, and where d is maximal (that is, there is no other such partition $C'_0, C'_1, \ldots, C'_{d'-1}$ with $d' > d$).

Proof. A direct consequence of Theorem 4.3 below. □

The number $d \geq 1$ is called the *period of the chain* (resp., of the transition matrix, of the transition graph). The classes $C_0, C_1, \ldots, C_{d-1}$ are called the *cyclic classes*.

The chain therefore moves from one class to the other at each transition, and this cyclically, as shown in Figure 2.4.2.

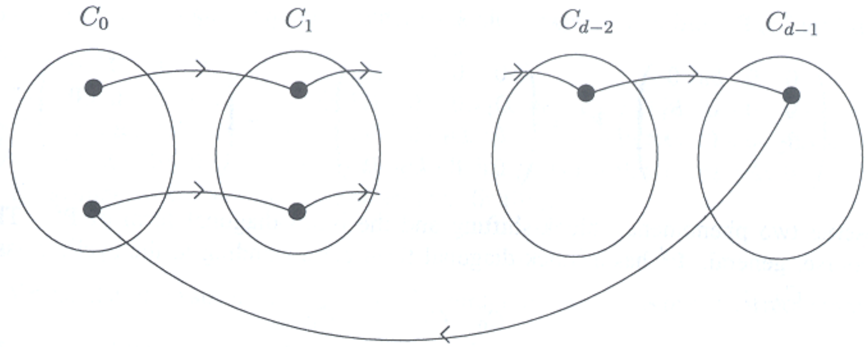


Figure 2.4.2. Cycles

Example 4.2.
The chain with the transition graph depicted in Figure 2.4.3 is irreducible and has period $d = 3$, with the cyclic classes $C_0 = \{1, 2\}$, $C_1 = \{4, 7\}$, $C_3 = \{3, 5, 6\}$. ◇

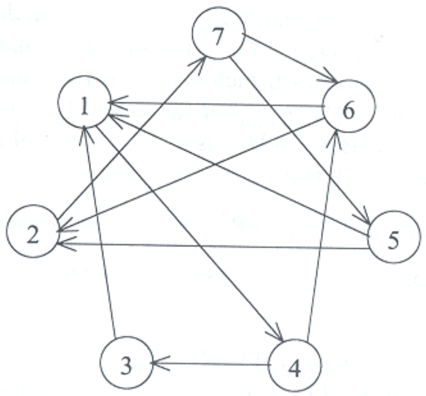


Figure 2.4.3. An irreducible transition graph with period 3

Consider now an irreducible chain of period d with the cyclic classes $C_0, C_1, \ldots, C_d$. Renumbering the states of E if necessary, the transition matrix has the block structure below (where $d = 4$, to be explicit),

$$\mathbf{P} = \begin{array}{c} \\ C_0 \\ C_1 \\ C_2 \\ C_3 \end{array} \begin{array}{c} \begin{array}{cccc} C_0 & C_1 & C_2 & C_3 \end{array} \\ \begin{pmatrix} 0 & A_0 & 0 & \\ 0 & 0 & A_1 & 0 \\ 0 & 0 & 0 & A_2 \\ A_3 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

and therefore $\mathbf{P}^2$, $\mathbf{P}^3$, and $\mathbf{P}^4$ also have a block structure corresponding to C_0, C_1, C_2, C_3:

$$\mathbf{P}^2 = \begin{pmatrix} 0 & 0 & B_0 & 0 \\ 0 & 0 & 0 & B_1 \\ B_2 & 0 & 0 & 0 \\ 0 & B_3 & 0 & 0 \end{pmatrix}, \mathbf{P}^3 = \begin{pmatrix} 0 & 0 & 0 & D_0 \\ D_1 & 0 & 0 & 0 \\ 0 & D_2 & 0 & 0 \\ 0 & 0 & D_3 & 0 \end{pmatrix}, \mathbf{P}^4 = \begin{pmatrix} E_0 & 0 & 0 & 0 \\ 0 & E_1 & 0 & 0 \\ 0 & 0 & E_2 & 0 \\ 0 & 0 & 0 & E_3 \end{pmatrix}.$$

We observe two phenomena: block-shifting and the block-diagonal form of $\mathbf{P}^4$. This is, of course, general: $\mathbf{P}^d$ has a block-diagonal form corresponding to the cyclic classes $C_0, C_1, \dots, C_{d-1}$:

$$\mathbf{P}^d = \begin{array}{c} \\ C_0 \\ C_1 \\ \vdots \\ C_{d-1} \end{array} \begin{array}{c} \begin{array}{cccc} C_0 & C_1 & \cdots & C_{d-1} \end{array} \\ \begin{pmatrix} E_0 & & 0 & \\ & E_1 & & \\ & & \ddots & \\ 0 & & & E_{d-1} \end{pmatrix} \end{array}. \tag{4.1}$$

The d-step transition matrix $\mathbf{P}^d$ is also a stochastic matrix, and obviously the $\mathbf{P}$-cyclic classes $C_0, C_1, \dots, C_{d-1}$ are all in different $\mathbf{P}^d$-communication classes, as the diagonal block structure shows.

The question is this: Is there, in C_0 for instance, more than one $\mathbf{P}^d$-communication class? The answer is no, and therefore matrix E_0 in (4.1) is an irreducible stochastic matrix. To see this, take two different states $i, j \in C_0$. Since they $\mathbf{P}$-communicate ($\mathbf{P}$ was assumed irreducible), there exist $m > 0$ and $n > 0$ such that $p_{ij}(m) > 0$ and $p_{ji}(n) > 0$. But since $\mathbf{P}$ has period d, necessarily $m = Md, n = Nd$ for some $M > 0$ and $N > 0$. Therefore, $p_{ij}(Md) > 0$ and $p_{ji}(Nd) > 0$. But $p_{ij}(Md)$ is the (i, j) term of $(\mathbf{P}^d)^M$, and similarly for $p_{ji}(Nd)$. We therefore have proven that i and j $\mathbf{P}^d$-communicate.

Also (Problem 2.4.7), the restriction of $\mathbf{P}^d$ to any cyclic class is aperiodic.

For an arbitrary transition matrix, not necessarily irreducible, the formal notion of *period* is the following.

Definition 4.4. *Arithmetic Definition of Period*
The period d_i of state $i \in E$ is, by definition,

$$d_i = \gcd\{n \geq 1\,;\ p_{ii}(n) > 0\}, \tag{4.2}$$

with the convention $d_i = +\infty$ if there is no $n \geq 1$ with $p_{ii}(n) > 0$. If $d_i = 1$, the state i is called *aperiodic* .

Theorem 4.2. *Period Is a Class Property*
If states i and j communicate they have the same period.

Proof. As i and j communicate, there exist integers N and M such that $p_{ij}(M) > 0$ and $p_{ji}(N) > 0$. For any $k \geq 1$,

$$p_{ii}(M + nk + N) \geq p_{ij}(M)(p_{jj}(k))^n p_{ji}(N)$$

(indeed, the path $X_0 = i, X_M = j, X_{M+k} = j, \dots, X_{M+nk} = j, X_{M+nk+N} = i$ is just one way of going from i to i in $M + nk + N$ steps).

Therefore, for any $k \geq 1$ such that $p_{jj}(k) > 0$, we have $p_{ii}(M + nk + N) > 0$ for all $n \geq 1$. Therefore, d_i divides $M + nk + N$ for all $n \geq 1$, and in particular, d_i divides k. We have therefore shown that d_i divides all k such that $p_{jj}(k) > 0$, and in particular, d_i divides d_j. By symmetry, d_j divides d_i, and therefore, finally, $d_i = d_j$. □

We can therefore speak of the period of a communication class or of an irreducible chain.

The important result concerning periodicity is the following.

Theorem 4.3. *Lattice Theorem*
Let $\mathbf{P}$ be an irreducible stochastic matrix with period d. Then for all states i, j there exist $m \geq 0$ and $n_0 \geq 0$ (m and n_0 possibly depending on i, j) such that

$$p_{ij}(m + nd) > 0,\ \forall n \geq n_0. \tag{4.3}$$

Proof. First observe that it suffices to prove this for $i = j$. Indeed, there exists m such that $p_{ij}(m) > 0$, because j is accessible from i, the chain being irreducible, and therefore, if for some $n_0 \geq 0$ we have $p_{jj}(nd) > 0$ for all $n \geq n_0$, then $p_{ij}(m + nd) \geq p_{ij}(m)p_{jj}(nd) > 0$ for all $n \geq n_0$. The gcd of the set $A = \{k \geq 1; p_{jj}(k) > 0\}$ is d, and A is closed under addition. The set A therefore contains all but a finite number of the positive multiples of d (see Theorem 1.1 of the Appendix). In other words, there exists n_0 such that $n > n_0$ implies $p_{jj}(nd) > 0$. □

5 Steady State

5.1 Stationarity

We now introduce the central notion of the stability theory of discrete-time HMCs.

Definition 5.1. *Stationary Distribution*
A probability distribution π satisfying

$$\pi^T = \pi^T \mathbf{P} \tag{5.1}$$

is called a *stationary distribution* of the transition matrix $\mathbf{P}$, or of the corresponding HMC.

The *global balance equation* (5.1) says that for all states i,

$$\pi(i) = \sum_{j \in E} \pi(j) p_{ji}. \tag{5.2}$$

Iteration of (5.1) gives $\pi^T = \pi^T \mathbf{P}^n$ for all $n \geq 0$, and therefore, in view of (1.7), if the initial distribution $\nu = \pi$, then $\nu_n = \pi$ for all $n \geq 0$. Thus, if a chain is started with a stationary distribution, it keeps the same distribution forever. But there is more, because then,

$$P(X_n = i_0, X_{n+1} = i_1, \dots, X_{n+k} = i_k) = \pi(i_0) p_{i_0 i_1} \cdots p_{i_{k-1} i_k} \tag{5.3}$$

does not depend on n. In this sense the chain is *stationary*. One also says that the chain is in a *stationary regime*, or in *equilibrium*, or in *steady state*. In summary:

Theorem 5.1. *Steady State*
A chain started with a stationary distribution is stationary.

Remark 5.1. The balance equation $\pi^T \mathbf{P} = \pi^T$, together with the requirement that π be a probability vector, i.e., $\pi^T \mathbf{1} = 1$ (where $\mathbf{1}$ is a column vector with all its entries equal to 1), constitute when E is finite, $|E| + 1$ equations for $|E|$ unknown variables. One of the $|E|$ equations in $\pi^T \mathbf{P} = \pi^T$ is superfluous given the constraint $\pi^T \mathbf{1} = 1$. Indeed, summing up all equalities of $\pi^T \mathbf{P} = \pi^T$ yields the equality $\pi^T \mathbf{P} \mathbf{1} = \pi^T \mathbf{1}$, that is, $\pi^T \mathbf{1} = 1$. ◇

5.2 Examples

Example 5.1. *Two-State Markov Chain*
Take $E = \{1, 2\}$ and define the transition matrix

$$\mathbf{P} = \begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \end{array},$$

where $\alpha, \beta \in (0, 1)$. The global balance equations are

$$\begin{aligned} \pi(1) &= \pi(1)(1-\alpha) + \pi(2)\beta, \\ \pi(2) &= \pi(1)\alpha + \pi(2)(1-\beta). \end{aligned}$$

This is a dependent system which reduces to the single equation $\pi(1)\alpha = \pi(2)\beta$, to which must be added $\pi(1) + \pi(2) = 1$ expressing that π is a probability vector. We obtain

$$\pi(1) = \frac{\beta}{\alpha + \beta}, \quad \pi(2) = \frac{\alpha}{\alpha + \beta}.$$

◇

Example 5.2. *The Urn of Ehrenfest*
The corresponding HMC was described in Example 2.6. The global balance equations are, for $i \in [1, N-1]$,

$$\pi(i) = \pi(i-1)\left(1 - \frac{i-1}{N}\right) + \pi(i+1)\frac{i+1}{N}$$

and, for the boundary states,

$$\pi(0) = \pi(1)\frac{1}{N}, \quad \pi(N) = \pi(N-1)\frac{1}{N}.$$

Leaving $\pi(0)$ undetermined, one can solve the balance equations for $i = 0, 1, \dots, N$ successively, to obtain

$$\pi(i) = \pi(0)\binom{N}{i}.$$

The value of $\pi(0)$ is then determined by writing that π is a probability vector:

$$1 = \sum_{i=0}^{N} \pi(i) = \pi(0) \sum_{i=0}^{N} \binom{N}{i} = \pi(0) 2^N.$$

This gives for π the binomial distribution of size N and parameter $\frac{1}{2}$:

$$\pi(i) = \frac{1}{2^N}\binom{N}{i}. \tag{5.4}$$

This is the distribution one would obtain by placing independently each particle in the compartments, with probability $\frac{1}{2}$ for each compartment. ◇

Example 5.3. *Symmetric Random Walk*
A symmetric random walk on $\mathbb{Z}$ cannot have a stationary distribution. Indeed, the solution of the balance equation

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

for $i \geq 0$, with initial data $\pi(0)$ and $\pi(1)$, is

$$\pi(i) = \pi(0) + (\pi(1) - \pi(0))\, i.$$

Since $\pi(i) \in [0, 1]$, necessarily $\pi(1) - \pi(0) = 0$. Therefore, $\pi(i)$ is a constant, necessarily 0 because the total mass of π is finite. Thus for all $i \geq 0$, and therefore, in view of the global balance equation, for all i, $\pi(i) = 0$, a contradiction if we want π to be a probability distribution. ◇

Example 5.4. *Stationary Distributions May Be Many*
Take the identity as transition matrix. Then any probability distribution on the state space is a stationary distribution. ◇

Recurrence equations can be used to obtain the stationary distribution when the latter exists and is unique. Generating functions sometimes usefully exploit the dynamics.

xample 5.5. *Repair Shop*
his example continues Example 2.2. For any complex number z with modulus not larger han 1, it follows from the recurrence equation (2.4) that

$$\begin{aligned} z^{X_{n+1}+1} &= \left(z^{(X_n-1)^+ +1}\right) z^{Z_{n+1}} \\ &= \left(z^{X_n} 1_{\{X_n>0\}} + z 1_{\{X_n=0\}}\right) z^{Z_{n+1}} \\ &= \left(z^{X_n} - 1_{\{X_n=0\}} + z 1_{\{X_n=0\}}\right) z^{Z_{n+1}}, \end{aligned}$$

nd therefore

$$z z^{X_{n+1}} - z^{X_n} z^{Z_{n+1}} = (z-1) 1_{\{X_n=0\}} z^{Z_{n+1}}.$$

rom the independence of X_n and Z_{n+1}, $E[z^{X_n} z^{Z_{n+1}}] = E[z^{X_n}] g_Z(z)$, where $g_Z(z)$ is the enerating function of Z_{n+1}, and $E[1_{\{X_n=0\}} z^{Z_{n+1}}] = \pi(0) g_Z(z)$, where $\pi(0) = P(X_n = 0)$. herefore,

$$zE[z^{X_{n+1}}] - g_Z(z)E[z^{X_n}] = (z-1)\pi(0)g_Z(z).$$

ut in steady state, $E[z^{X_{n+1}}] = E[z^{X_n}] = g_X(z)$, and therefore

$$g_X(z)(z - g_Z(z)) = \pi(0)(z-1)g_Z(z). \tag{5.5}$$

his gives the generating function $g_X(z) = \sum_{i=0}^{\infty} \pi(i) z^i$, as long as $\pi(0)$ is available. To btain $\pi(0)$, differentiate (5.5):

$$g'_X(z)(z - g_Z(z)) + g_X(z)\left(1 - g'_Z(z)\right) = \pi(0)\left(g_Z(z) + (z-1)g'_Z(z)\right),$$

nd let $z = 1$, to obtain, taking into account the equalities $g_X(1) = g_Z(1) = 1$ and $'_Z(1) = E[Z]$,

$$\pi(0) = 1 - E[Z]. \tag{5.6}$$

ince $\pi(0)$ must be nonnegative, this immediately gives the necessary condition $E[Z] \le 1$. ctually, one must have, if the trivial case $Z_{n+1} \equiv 1$ is excluded,

$$E[Z] < 1. \tag{5.7}$$

ndeed, if $E[Z] = 1$, implying $\pi(0) = 0$, it follows from (5.5) that

$$g_X(x)(x - g_Z(x)) = 0$$

or all $x \in [0, 1]$. But excluding the case $Z_{n+1} \equiv 1$ (that is, $g_Z(x) \equiv x$), the equation $- g_Z(x) = 0$ has only $x = 1$ for a solution when $g'_Z(1) = E[Z] \le 1$ (see Chapter 1, heorem 5.1). Therefore, $g_X(x) \equiv 0$ for all $x \in [0, 1)$, and consequently $g_X(z) \equiv 0$ on $|z| < 1\}$. This leads to a contradiction, since the generating function of an integer-valued andom variable cannot be identically null.

We shall prove later that $E[Z] < 1$ is also a sufficient condition for the existence of steady state. For the time being, we learn from (5.5) and (5.6) that, if the stationary istribution exists, then its generating function is given by the formula

$$\sum_{i=0}^{\infty} \pi(i) z^i = (1 - E[Z]) \frac{(z-1)g_Z(z)}{z - g_Z(z)}. \tag{5.8}$$

◇

Example 5.6. *Birth and Death with Two Reflecting Barriers*
The Ehrenfest model is a special case of a birth-and-death HMC with reflecting barriers at 0 and N. The state space of such a chain is $E = \{0, 1, \ldots, N\}$, and its transition matrix is

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & & & & & \\ q_1 & r_1 & p_1 & & & & \\ & q_2 & r_2 & p_2 & & & \\ & & \ddots & & & & \\ & & & q_i & r_i & p_i & \\ & & & & \ddots & \ddots & \ddots \\ & & & & q_{N-1} & r_{N-1} & p_{N-1} \\ & & & & & 1 & 0 \end{pmatrix},$$

where $p_i > 0$, $q_i > 0$, and $p_i + q_i + r_i = 1$ for all states $i \in [1, N-1]$. The global balance equations for the states $i \in [1, N-1]$ are

$$\pi(i) = p_{i-1}\pi(i-1) + r_i\pi(i) + q_{i+1}\pi(i+1),$$

and for the boundary states,

$$\pi(0) = \pi(1)q_1, \ \pi(N) = \pi(N-1)p_{N-1}.$$

Of course, π must be a probability, whence

$$\sum_{n=0}^{N} \pi(i) = 1.$$

Writing $r_i = 1 - p_i + q_i$ and regrouping terms gives for $i \in [2, N-1]$,

$$\pi(i+1)q_{i+1} - \pi(i)p_i = \pi(i)q_i - \pi(i-1)p_{i-1}$$

and

$$\begin{aligned} \pi(1)q_1 - \pi(0) &= 0, \\ \pi(2)q_2 - \pi(1)p_1 &= \pi(1)q_1 - \pi(0). \end{aligned}$$

Therefore, $\pi(1)q_1 = \pi(0)$, and for $i \in [2, N-1]$,

$$\pi(i)q_i = \pi(i-1)p_{i-1}.$$

This gives

$$\pi(1) = \pi(0)\frac{1}{q_1},$$

and for $i \in [2, N]$,

$$\pi(i) = \pi(0)\frac{p_1 p_2 \cdots p_{i-1}}{q_1 q_2 \cdots q_i}. \tag{5.9}$$

The unknown $\pi(0)$ is obtained by $\sum_{i=0}^{N} \pi(i) = 1$, that is,

$$\pi(0)\left\{1+\frac{1}{q_1}+\frac{p_1}{q_1 q_2}+\cdots+\frac{p_1 p_2 \cdots p_{N-1}}{q_1 q_2 \cdots q_{N-1} q_N}\right\}=1. \tag{5.10}$$

◇

Example 5.7. *Birth and Death with One Reflecting Barrier*
The model is the same as above, except that the state space is $E = \mathbb{N}$, and therefore the upper barrier is at infinity. The same computations as above lead to the expression (5.9) for the general solution of $\pi^T \mathbf{P} = \pi^T$, which depends on the initial condition $\pi(0)$. For this solution to be a probability, we must have $\pi(0) > 0$. Also, writing $\sum_{i=1}^{\infty} \pi(i) = 1$,

$$\pi(0)\left\{1+\frac{1}{q_1}+\sum_{j=1}^{\infty}\frac{p_1 p_2 \cdots p_j}{q_1 q_2 \cdots q_{j+1}}\right\}=1. \tag{5.11}$$

Thus a stationary distribution exists if and only if

$$\sum_{j=1}^{\infty}\frac{p_1 p_2 \cdots p_j}{q_1 q_2 \cdots q_{j+1}} < \infty. \tag{5.12}$$

In this case $\pi(i)$ is given by the expressions in (5.9), where $\pi(0)$ is determined by (5.11). ◇

6 Time Reversal

6.1 Reversed Chain

The notions of time-reversal and time-reversibility are very productive, in particular in the theory of Markov chains, and especially in Monte Carlo simulation (Chapter 7) and queuing theory (Chapter 9).

Let $\{X_n\}_{n\geq 0}$ be an HMC with transition matrix $\mathbf{P}$ and admitting a stationary distribution π such that

$$\pi(i) > 0 \tag{6.1}$$

for all states i. Define the matrix $\mathbf{Q}$, indexed by E, by

$$\pi(i) q_{ij} = \pi(j) p_{ji}. \tag{6.2}$$

This matrix is stochastic, since

$$\sum_{j\in E} q_{ij} = \sum_{j\in E}\frac{\pi(j)}{\pi(i)} p_{ji} = \frac{1}{\pi(i)}\sum_{j\in E}\pi(j) p_{ji} = \frac{\pi(i)}{\pi(i)} = 1,$$

where the third equality uses the balance equations. Its interpretation is the following: Suppose that the initial distribution of $\{X_n\}$ is π, in which case for all $n \geq 0$, all $i \in E$,

$$P(X_n = i) = \pi(i). \tag{6.3}$$

Then, from Bayes's retrodiction formula,

$$P(X_n = j \mid X_{n+1} = i) = \frac{P(X_{n+1} = i \mid X_n = j) P(X_n = j)}{P(X_{n+1} = i)},$$

that is, in view of (6.2) and (6.3),

$$P(X_n = j \mid X_{n+1} = i) = q_{ji}. \tag{6.4}$$

We see that $\mathbf{Q}$ is the transition matrix of the initial chain when time is reversed.

The following is a very simple observation that will be promoted to the rank of a theorem in view of its usefulness and also for the sake of easy reference.

Theorem 6.1. *Reversal Test*
Let $\mathbf{P}$ be a stochastic matrix indexed by a countable set E, and let π be a probability distribution on E. Let $\mathbf{Q}$ be a stochastic matrix indexed by E such that for all $i, j \in E$,

$$\pi(i) q_{ij} = \pi(j) p_{ji}. \tag{6.5}$$

Then π is a stationary distribution of $\mathbf{P}$.

Proof. For fixed $i \in E$, sum equalities (6.5) with respect to $j \in E$ to obtain

$$\sum_{j\in E}\pi(i) q_{ij} = \sum_{j\in E}\pi(j) p_{ji}.$$

But the left-hand side is equal to $\pi(i)\sum_{j\in E} q_{ij} = \pi(i)$, and therefore, for all $i \in E$,

$$\pi(i) = \sum_{j\in E}\pi(j) p_{ji}.$$

□

Example 6.1. *Extension of a Stationary Chain to Negative Times*
Time reversal can also be used to extend to negative times a chain $\{X_n\}_{n\geq 0}$ in steady state corresponding to a stationary distribution π such that $\pi(i) > 0$ for all $i \in E$. See Problem 2.6.2. ◇

6.2 Time Reversibility

Definition 6.1. *Reversible Chain*
One calls *reversible* a stationary Markov chain with initial distribution π (a stationary distribution) assumed positive if for all $i, j \in E$,

$$\pi(i) p_{ij} = \pi(j) p_{ji}. \tag{6.6}$$

In this case, $q_{ij} = p_{ij}$, and therefore the chain and the time-reversed chain are statistically the same, since the distribution of a homogeneous Markov chain is entirely determined by its initial distribution and its transition matrix (Theorem 1.1). Equations (6.6) are called the *detailed* balance equations. The following is an immediate corollary of Theorem 6.1.

Corollary 6.1. *Detailed Balance Test*
Let **P** be a transition matrix on the countable state space E, and let π be some probability distribution on E. If for all $i, j \in E$, the detailed balance equations (6.6) are satisfied, then π is a stationary distribution of **P**.

Example 6.2. *The Urn of Ehrenfest*
This example continues Examples 2.6 and 5.2. Recall that we obtained the expression

$$\pi(i) = \frac{1}{2N}\binom{N}{i}$$

for the stationary distribution. Checking the detailed balance equations

$$\pi(i)p_{i,i+1} = \pi(i+1)p_{i+1,i}$$

is immediate. ◇

Example 6.3. *The Generalized Mouse*
The reason for the title of this example is that it is an abstract form of the motion of a mouse in a maze; see Example 3.2. However, the professionals call this a *random walk on a graph*. Consider a finite nonoriented graph and call E the set of vertices, or nodes, of this graph. Call d_i the number of edges "adjacent" to node i. Transform this graph into an oriented graph by splitting each edge into two oriented edges of opposite directions, and make it a transition graph by associating to the oriented edge from i to j the transition probability $\frac{1}{d_i}$ (see Fig. 2.6.1).

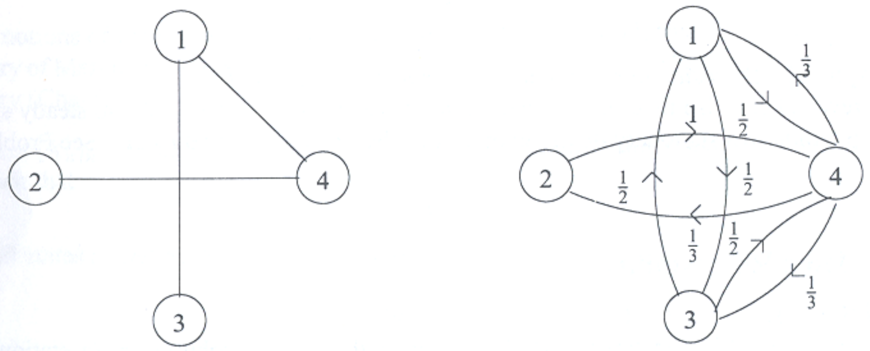


Figure 2.6.1. A random walk on a graph

It will be assumed, as is the case in Figure 2.6.1, that $d_i > 0$ for all states i. A stationary distribution (in fact, *the* stationary distribution, as we shall see later, in Chapter 3) is given by

$$\pi(i) = \frac{d_i}{\sum_{j \in E} d_j}.$$

For this, we can use Corollary 6.1, making the insider's guess that the chain is reversible. We just have to check that

$$\pi(i)\frac{1}{d_i} = \pi(j)\frac{1}{d_j}.$$

Hence we have obtained the stationary distribution and proved the reversibility of the chain. ◇

Example 6.4. *Birth and Death*
One verifies that for both Examples 5.6 and 5.7, when the stationary distribution π exists, the detailed balance equations $\pi(i)p_i = \pi(i+1)q_{i+1}$ hold for all $i \in E$. ◇

7 Regeneration

7.1 Strong Markov Property

Definition 7.1. *Stopping Times*
A *stopping time* with respect to a stochastic process $\{X_n\}_{n\geq 0}$ is, by definition, a random variable τ taking its values in $\mathbb{N} \cup \{+\infty\}$ and such that for all integers $m \geq 0$, the event $\{\tau = m\}$ can be expressed in terms of $X_0, X_1, \dots, X_m$.

The latter property is symbolized by the notation

$$\{\tau = m\} \in X_0^m. \tag{7.1}$$

When the state space is countable, (7.1) means that

$$1_{\{\tau=m\}} = \psi_m(X_0, \dots, X_m),$$

for some function ψ_m with values in $\{0, 1\}$.

Example 7.1. *Return Times*
In the theory of Markov chains, a typical and most important stopping time is the *return time* to state $i \in E$,

$$T_i = \inf\{n \geq 1;\ X_n = i\}, \tag{7.2}$$

where $T_i = \infty$ if $X_n \neq i$ for all $n \geq 1$. It is indeed a stopping time. Do a direct proof, or wait for Example 7.5. ◇

Observe that $T_i \geq 1$, and in particular, $X_0 = i$ does *not* imply $T_i = 0$. This is why T_i is called the *return* time to i, and not the *hitting* time of i. The latter is $S_i = T_i$ if $X_0 \neq i$, and $S_i = 0$ if $X_0 = i$. It is also a stopping time.

Example 7.2. *Deterministic Times*
A constant time is a stopping time (check this). ◇