4.1 Markov Chains

Informally, Markov chains (MCs) serve as theoretical models for describing a "system" which can be in various "states", the fixed set of possible states being countable (i.e. finite, or denumerably infinite). The system "jumps" at unit time intervals from one state to another, and the probabilistic law according to which jumps occur is

> "If the system is in the ith state at time $k - 1$, the next jump will take it to the jth state with probability $p_{ij}(k)$."

The set of transition probabilities $p_{ij}(k)$ is prescribed for all i, j, k and determines the probabilistic behavior of the system, once it is known how it starts off "at time 0".

A more formal description is as follows. We are given a countable set $\mathscr{S} = \{s_1, s_2, \cdots\}$ or, sometimes, more conveniently $\{s_0, s_1, s_2, \ldots\}$ which is known as the state space, and a sequence of random variables $\{X_k\}$, $k = 0, 1, 2, \ldots$ taking values in $\mathscr{S}$, and having the following *probability property*: if $x_0, x_1, \ldots, x_{k+1}$ are elements of $\mathscr{S}$, then

$$P(X_{k+1} = x_{k+1} \mid X_k = x_k, X_{k-1} = x_{k-1}, \ldots, X_0 = x_0) = P(X_{k+1} = x_{k+1} \mid X_k = x_k)$$

if

$$P(X_k = x_k, \ldots, X_0 = x_0) > 0$$

(if $P(B) = 0$, $P(A \mid B)$ is undefined).

This property which expresses, roughly, that future probabilistic evolution of the process is determined once the *immediate past* is known, is the Markov property, and the stochastic process $\{X_k\}$ possessing it is called *a Markov chain*.

Moreover, we call the probability

$$P(X_{k+1} = s_j \mid X_k = s_i)$$

the transition probability from state s_i to state s_j, and write it succinctly as

$$p_{ij}(k+1), \; s_i, s_j \in \mathscr{S}, \qquad k = 0, 1, 2, \ldots$$

Now consider

$$P[X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_k = s_{i_k}].$$

Either *this is positive*, in which case, by repeated use of the Markov property and conditional probabilities it is in fact

$$P[X_k = s_{i_k} \mid X_{k-1} = s_{i_{k-1}}] \cdots P[X_1 = s_{i_1} \mid X_0 = s_{i_0}] P[X_0 = s_{i_0}] = p_{i_{k-1}, i_k}(k) p_{i_{k-2}, i_{k-1}}(k-1) \cdots p_{i_0, i_1}(1) \Pi_{i_0}$$

where $\Pi_{i0} = P[X_0 = s_{i_0}]$

or it is zero, in which case for some $0 \le r \le k$ (and we take such minimal r)

$$P[X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_r = s_{i_r}] = 0.$$

Considering the cases $r = 0$ and $r > 0$ separately, we see (repeating the above argument), that it is *nevertheless* true that

$$P[X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_k = s_{i_k}] = \Pi_{i_0} p_{i_0, i_1}(1) \cdots p_{i_{k-1}, i_k}(k)$$

since the product of the first $r + 1$ elements on the right is zero. Thus we see that the probability structure of any finite sequence of outcomes is *completely defined* by a knowledge of the *non-negative quantities*

$$p_{ij}(k);\ s_i, s_j \in \mathscr{S}, \qquad \Pi_i;\ s_i \in \mathscr{S}.$$

The set $\{\Pi_i\}$ of probabilities is called the *initial probability distribution* of the chain. We consider these quantities as specified, and denote the row vector of the initial distribution by $\mathbf{\Pi}_0'$.

Now, for fixed $k = 1, 2, \ldots$ the matrix

$$P_k = \{p_{ij}(k)\},\ s_i, s_j \in \mathscr{S}$$

is called the *transition matrix* of the MC at time k. It is clearly a square matrix with non-negative elements, and will be doubly infinite if $\mathscr{S}$ is denumerably infinite.

Moreover, its row sums (understood in the limiting sense in the denumerably infinite case) are unity, for

$$\sum_{j \in \mathscr{S}} p_{ij}(k) = \sum_{j \in \mathscr{S}} P[X_k = s_j \mid X_{k-1} = s_i]$$
$$= P[X_k \in \mathscr{S} \mid X_{k-1} = s_i]$$

by the addition of probabilities of disjoint sets;

$$= 1.$$

Thus the matrix P_k is *stochastic*.

Definition 4.1. If $P_1 = P_2 = \cdots = P_k = \cdots$ the Markov chain is said to have stationary transition probabilities or is said to be *homogeneous*. Otherwise it is *non-homogeneous* (or: *inhomogeneous*).

In the homogeneous case we shall refer to the common transition matrix as *the* transition matrix, and denote it by P.

Let us denote by $\mathbf{\Pi}_k'$ the row vector of the probability distribution of X_k; then it is easily seen from the expression for a single finite sequence of outcomes in terms of transition and initial probabilities that

$$\mathbf{\Pi}_k' = \mathbf{\Pi}_0' P_1 \cdots P_k$$

by summing (possibly in the limiting sense) over all sample paths for any fixed state at time k. In keeping with the notation of Chapter 3, we might now adopt the notation

$$T_{p,r} = P_{p+1} P_{p+2} \cdots P_{p+r}$$

and write

$$\Pi'_k = \Pi'_0 T_{0,k}.$$

[We digress for a moment to stress that, even in the case of infinite transition matrices, the above products are well defined by the natural extension of the rule of matrix multiplication, and are themselves stochastic. For: let

$$P_\alpha = \{p_{ij}(\alpha)\} \qquad \text{and} \qquad P_\beta = \{p_{ij}(\beta)\}$$

be two infinite stochastic matrices defined on the index set $\{1, 2, \ldots\}$. Define their product $P_\alpha P_\beta$ as the matrix with i, j entry given by the (non-negative) number:

$$\sum_{k=1}^{\infty} p_{ik}(\alpha) p_{kj}(\beta).$$

This sum converges, since the summands are non-negative, and

$$\sum_{k=1}^{\infty} p_{ik}(\alpha) p_{kj}(\beta) \leq \sum_{k=1}^{\infty} p_{ik}(\alpha) \leq 1$$

since probabilities always take on values between 0 and 1. Further the ith row sum of the new matrix is

$$\begin{aligned} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} p_{ik}(\alpha) p_{kj}(\beta) &= \sum_{k=1}^{\infty} p_{ik}(\alpha) \left(\sum_{j=1}^{\infty} p_{kj}(\beta) \right) \\ &= \sum_{k=1}^{\infty} p_{ik}(\alpha) = 1 \end{aligned}$$

by stochasticity of both P_α and P_β. (The interchange of summations is justified by the non-negativity of the summands.)]

It is also easily seen that for $k > p$

$$\Pi'_k = \Pi'_p T_{p, k-p}.$$

We are now in a position to see why the theory of homogeneous chains is substantially simpler than that of non-homogeneous ones: for then

$$T_{p,k} = P^k$$

so we have only to deal with powers of the common transition matrix P, and further, the probabilistic evolution is *homogeneous in reference to any initial time point* p.

In the remaining section of this chapter we assume that we are dealing with finite ($n \times n$) *matrices as before*, so that the index set is $\{1, 2, \ldots, n\}$ as before (or perhaps, more conveniently, $\{0, 1, \ldots, n-1\}$).

Examples

(1) *Bernoulli scheme*. Consider a sequence of independent trials in each of which a certain event has fixed probability, p, of occurring (this outcome being called a "success") and therefore a probability $q = 1 - p$ of not occurring (this outcome being called a "failure"). We can in the usual way equate success with the number 1 and failure with the number 0; then $\mathcal{S} = \{0, 1\}$, and the transition matrix at any time k is

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

so that we have here a homogeneous 2-state Markov chain. Notice that here the rows of the transition matrix are identical, which must in fact be so for any "Markov chain" where the random variables $\{X_k\}$ are independent.

(2) *Random walk between two barriers*. A particle may be at any of the points $0, 1, 2, 3, \ldots, s$ $(s \geq 1)$ on the x-axis. If it reaches point 0 it remains there with probability a and is reflected with probability $1 - a$ to state 1; if it reaches point s it remains there with probability b and is reflected to point $s - 1$ with probability $1 - b$. If at any instant the particle is at position i, $1 \leq i \leq s - 1$, then at the next time instant it will be at position $i + 1$ with probability p, or at $i - 1$ with probability $q = 1 - p$.

It is again easy to see that we have here a homogeneous Markov chain on the finite state set $\mathcal{S} = \{0, 1, 2, \ldots, s\}$ with transition matrix

$$P = \begin{bmatrix} a & 1-a & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 & 0 & 0 \\ & & & & \vdots & & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1-b & b \end{bmatrix}; \qquad p + q = 1, \quad 0 < p < 1.$$

If $a = 0$, 0 is a *reflecting barrier*, if $a = 1$ it is an *absorbing barrier*, otherwise i.e. if $0 < a < 1$ it is an *elastic barrier*; and similarly for state s.

(3) *Random walk unrestricted to the right*. The situation is as above, except that there is no "barrier" on the right, i.e. $\mathcal{S} = \{0, 1, 2, 3, \ldots\}$ is denumerably infinite, and so is the transition matrix P.

(4) *Recurrent event*. Consider a "recurrent event", described as follows. A system has a variable lifetime, whose length (measured in discrete units) has probability distribution $\{f_i\}$, $i = 1, 2, \ldots$. When the system reaches age $i \geq 1$, it either continues to age, or "dies" and starts afresh from age 0. The movement of the system if its age is $i - 1$ units, $i \geq 2$ is thus to i, with (conditional) probability $(1 - f_1 - \cdots - f_i)/(1 - f_1 - \cdots - f_{i-1})$ or to age 0, with probability $f_i/(1 - f_1 - \cdots - f_{i-1})$. At age $i = 0$, it either reaches age 1 with probability $1 - f_1$, or dies with probability f_1.

We have here a homogeneous Markov chain on the state set $\mathscr{S} = \{0, 1, 2, \ldots\}$ describing the movement of the age of the system. The transition matrix is then the denumerably infinite one:

$$\begin{bmatrix} f_1 & 1 - f_1 & 0 & 0 & 0 & \cdots \\ \dfrac{f_2}{1 - f_1} & 0 & \dfrac{1 - f_1 - f_2}{1 - f_1} & 0 & 0 & \cdots \\ \dfrac{f_3}{1 - f_1 - f_2} & 0 & 0 & \dfrac{1 - f_1 - f_2 - f_3}{1 - f_1 - f_2} & 0 & \cdots \\ & & \vdots & & & \end{bmatrix}.$$

It is customary to specify only that $\sum_{i=1}^{\infty} f_i \le 1$, thus allowing for the possibility of an infinite lifetime.

(5) *Pólya Urn scheme*. Imagine we have a white and b black balls in an urn. Let $a + b = N$. We draw a ball at random and before drawing the next ball we replace the one drawn, adding also s balls of the same colour.

Let us say that after r drawings the system is in state i, $i = 0, 1, 2, \ldots$ if i is the number of white balls obtained in the r drawings. Suppose we are in state i ($\le r$) after drawing number r. Thus $r - i$ black balls have been drawn to date, and the number of white balls in the urn is $a + is$, and the number of black is $b + (r - i)s$. Then at the next drawing we have movement to state $i + 1$ with probability

$$p_{i,\,i+1}(r + 1) = \frac{a + is}{N + rs}$$

and to state i with probability

$$p_{i,\,i}(r + 1) = \frac{b + (r - i)s}{N + rs} = 1 - p_{i,\,i+1}(r + 1).$$

Thus we have here a *non-homogeneous* Markov chain (if $s > 0$) with transition matrix P_k at "time" $k \equiv r + 1 \ge 1$ specified by

$$\begin{aligned} p_{ij}(k) &= \frac{a + is}{N + (k - 1)s}, \qquad j = i + 1 \\ &= \frac{b + (k - 1 - i)s}{N + (k - 1)s}, \qquad j = i \\ &= \text{otherwise}, \end{aligned}$$

where $\mathscr{S} = \{0, 1, 2, \ldots\}$.

N.B. This example is given here because it is a good illustration of a non-homogeneous chain; the non-homogeneity clearly occurring because of the addition of s balls of colour like the one drawn at each stage. Nevertheless, the reader should be careful to note that this example does not fit into the framework in which we have chosen to work in this chapter, since the matrix P_k is really *rectangular*, viz. $k \times (k + 1)$ in this case, a situation which can occur with non-homogeneous chains, but which we omit from further theoretical consideration. Extension in both directions to make each P_k doubly infinite corresponding to the index set $\{0, 1, 2, \ldots\}$ is not necessarily a good idea, since matrix dimensions are equalized at the cost of zero rows (beyond the $(k - 1)$th) thus destroying stochasticity.

4.2 Finite Homogeneous Markov Chains

Within this section we are in the framework of the bulk of the matrix theory developed hitherto.

It is customary in Markov chain theory to classify states and chains of various kinds. In this respect we shall remain totally consistent with the classification of Chapter 1.

Thus a chain will be said to be *irreducible*, and, further, *primitive* or *cyclic* (*imprimitive*) according to whether its transition matrix P is of this sort. Further, states of the set

$$\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$$

(or $\{s_0, s_1, \ldots, s_{n-1}\}$) will be said to be *periodic, essential and inessential*, to *lead* one to another, to *communicate*, to form *essential and inessential classes* etc. according to the properties of the corresponding indices of the index set $\{1, 2, \ldots, n\}$ of the transition matrix.

In fact, as has been mentioned earlier, this terminology was introduced in Chapter 1 in accordance with Markov chain terminology. The reader examining the terminology in the present framework should now see the logic behind it.

Irreducible MCs

Suppose we consider an irreducible MC $\{X_k\}$ with (irreducible) transition matrix P. Then putting as usual $\mathbf{1}$ for the vector with unity in each position,

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

by stochasticity of P; so that 1 is an eigenvalue and $\mathbf{1}$ a corresponding eigenvector. Now, since all row sums of P are equal and the Perron–Frobenius eigenvalue lies between the largest and the smallest, 1 is the Perron–Frobenius eigenvalue of P, and $\mathbf{1}$ may be taken as the corresponding right Perron–Frobenius eigenvector. Let $\boldsymbol{v}'$, normed so that $\boldsymbol{v}'\mathbf{1} = 1$, be the corresponding positive left eigenvector. Then we have that

$$\boldsymbol{v}'P = \boldsymbol{v}', \tag{4.1}$$

where $\boldsymbol{v}$ is the column vector of a probability distribution.