3

Transition Probabilities

As with all stochastic processes, there are two directions from which to approach the formal definition of a Markov chain.

The first is via the process itself, by constructing (perhaps by heuristic arguments at first, as in the descriptions in Chapter 2) the sample path behavior and the dynamics of movement in time through the state space on which the chain lives. In some of our examples, such as models for queueing processes or models for controlled stochastic systems, this is the approach taken. From this structural definition of a Markov chain, we can then proceed to define the probability laws governing the evolution of the chain.

The second approach is via those very probability laws. We define them to have the structure appropriate to a Markov chain, and then we must show that there is indeed a process, properly defined, which is described by the probability laws initially constructed. In effect, this is what we have done with the forward recurrence time chain in Section 2.4.1.

From a practitioner's viewpoint there may be little difference between the approaches. In many books on stochastic processes, such as Çinlar [40] or Karlin and Taylor [122], the two approaches are used, as they usually can be, almost interchangeably; and advanced monographs such as Nummelin [202] also often assume some of the foundational aspects touched on here to be well-understood.

Since one of our goals in this book is to provide a guide to modern general space Markov chain theory and methods for practitioners, we give in this chapter only a sketch of the full mathematical construction which provides the underpinning of Markov chain theory.

However, we also have as another, and perhaps somewhat contradictory, goal the provision of a thorough and rigorous exposition of results on general spaces, and for these it is necessary to develop both notation and concepts with some care, even if some of the more technical results are omitted.

Our approach has therefore been to develop the technical detail in so far as it is relevant to specific Markov models, and where necessary, especially in techniques which are rather more measure theoretic or general stochastic process theoretic in nature, to refer the reader to the classic texts of Doob [68], and Chung [49], or the more recent exposition of Markov chain theory by Revuz [223] for the foundations we need. Whilst such an approach renders this chapter slightly less than self-contained, it is our hope that the gaps in these foundations will be either accepted or easily filled by such external sources.

Our main goals in this chapter are thus

(i) to demonstrate that the dynamics of a Markov chain $\{\Phi_n\}$ can be completely defined by its one step "transition probabilities"

$$P(x, A) = \mathsf{P}(\Phi_n \in A \mid \Phi_{n-1} = x),$$

which are well-defined for appropriate initial points x and sets A;

(ii) to develop the functional forms of these transition probabilities for many of the specific models in Chapter 2, based in some cases on heuristic analysis of the chain and in other cases on development of the probability laws; and

(iii) to develop some formal concepts of hitting times on sets, and the "Strong Markov Property" for these and related stopping times, which will enable us to address issues of stability and structure in subsequent chapters.

We shall start first with the formal concept of a Markov chain as a stochastic process, and move then to the development of the transition laws governing the motion of the chain; and complete the cycle by showing that if one starts from a set of possible transition laws then it is possible to move from these to a chain which is well defined and governed by these laws.

3.1 Defining a Markovian Process

A *Markov chain* $\mathbf{\Phi} = \{\Phi_0, \Phi_1, \ldots\}$ is a particular type of *stochastic process* taking, at times $n \in \mathbb{Z}_+$, values Φ_n in a *state space* X.

We need to know and use a little of the language of stochastic processes. A discrete time stochastic process $\mathbf{\Phi}$ on a state space is, for our purposes, a collection $\mathbf{\Phi} = (\Phi_0, \Phi_1, \ldots)$ of random variables, with each Φ_i taking values in X; these random variables are assumed measurable individually with respect to some given σ-field $\mathcal{B}(\mathsf{X})$, and we shall in general denote elements of X by letters $x, y, z, \ldots$ and elements of $\mathcal{B}(\mathsf{X})$ by A, B, C.

When thinking of the process as an entity, we regard values of the whole chain $\mathbf{\Phi}$ itself (called *sample paths* or *realizations*) as lying in the *sequence* or *path* space formed by a countable product $\Omega = \mathsf{X}^\infty = \prod_{i=0}^\infty \mathsf{X}_i$, where each X_i is a copy of X equipped with a copy of $\mathcal{B}(\mathsf{X})$. For $\mathbf{\Phi}$ to be defined as a random variable in its own right, Ω will be equipped with a σ-field $\mathcal{F}$, and for each state $x \in \mathsf{X}$, thought of as an initial condition in the sample path, there will be a probability measure P_x such that the probability of the event $\{\mathbf{\Phi} \in A\}$ is well-defined for any set $A \in \mathcal{F}$; the initial condition requires, of course, that $\mathsf{P}_x(\Phi_0 = x) = 1$.

The triple $\{\Omega, \mathcal{F}, \mathsf{P}_x\}$ thus defines a stochastic process since $\Omega = \{\omega_0, \omega_1, \ldots : \omega_i \in \mathsf{X}\}$ has the product structure to enable the projections ω_n at time n to be well defined realizations of the random variables Φ_n.

Many of the models we consider (such as random walk or state space models) have stochastic motion based on a separately defined sequence of underlying variables, namely a noise or disturbance or innovation sequence $\mathbf{W}$. We will slightly abuse notation by using $\mathsf{P}(\mathbf{W} \in A)$ to denote the probability of the event $\{\mathbf{W} \in A\}$ without specifically defining the space on which $\mathbf{W}$ exists, or the initial condition of the chain:

this could be part of the space on which the chain $\boldsymbol{\Phi}$ is defined or it could be separate. No confusion should result from this usage.

Prior to discussing specific details of the probability laws governing the motion of a chain $\boldsymbol{\Phi}$, we need first to be a little more explicit about the structure of the state space X on which it takes its values. We consider, systematically, three types of state spaces in this book:

State Space Definitions

(i) The state space X is called *countable* if X is discrete, with a finite or countable number of elements, and with $\mathcal{B}(\mathsf{X})$ the σ-field of all subsets of X.

(ii) The state space X is called *general* if it is equipped with a countably generated σ-field $\mathcal{B}(\mathsf{X})$.

(iii) The state space X is called *topological* if it is equipped with a locally compact, separable, metrizable topology with $\mathcal{B}(\mathsf{X})$ as the Borel σ-field.

It may on the face of it seem odd to introduce quite general spaces before rather than after topological (or more structured) spaces.

This is however quite deliberate, since (perhaps surprisingly) we rarely find the extra structure actually increasing the ease of approach. From our point of view, we introduce topological spaces largely because specific applied models evolve on such spaces, and for such spaces we will give specific interpretations of our general results, rather than extending specific topological results to more general contexts.

For example, after framing general properties of sets, we identify these general properties as holding for compact or open sets if the chain is on a topological space; or after framing general properties of $\boldsymbol{\Phi}$, we develop the consequences of these when $\boldsymbol{\Phi}$ is suitably continuous with respect to the topology considered.

The first formal introduction of such topological concepts is given in Chapter 6, and is exemplified by an analysis of linear and nonlinear state space models in Chapter 7. Prior to this we concentrate on countable and general spaces: for purposes of exposition, our approach will often involve the description of behavior on a countable space, followed by the development of analogous behavior on a general space, and completed by specialization of results, where suitable, to more structured topological spaces in due course.

For some readers, countable space models will be familiar: nonetheless, by developing the results first in this context, and then the analogues for the less familiar general space processes on a systematic basis we intend to make the general context more accessible. By then specializing where appropriate to topological spaces, we trust the results will be found more applicable for, say, those models which evolve on multi-dimensional Euclidean space $\mathbb{R}^k$, or one of its subsets.

There is one caveat to be made in giving this description. One of the major observations for Markov chains is that in many cases, the full force of a countable space is not needed: we merely require one "accessible atom" in the space, such as we might have with the state $\{0\}$ in the storage models in Section 2.4.1. To avoid repetition we will often assume, especially later in the book, not the full countable space structure but just the existence of one such point: the results then carry over with only notational changes to the countable case.

In formalizing the concept of a Markov chain we pursue this pattern now, first developing the countable space foundations and then moving on to the slightly more complex basis for general space chains.

3.2 Foundations on a Countable Space

3.2.1 The initial distribution and the transition matrix

A discrete time Markov chain $\boldsymbol{\Phi}$ on a countable state space is a collection $\boldsymbol{\Phi} = \{\Phi_0, \Phi_1, \ldots\}$ of random variables, with each Φ_i taking values in the countable set X. In this countable state space setting, $\mathcal{B}(\mathsf{X})$ will denote the set of all subsets of X.

We assume that for any *initial distribution* μ for the chain, there exists a probability measure which denotes the law of $\boldsymbol{\Phi}$ on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is the product σ-field on the sample space $\Omega := \mathsf{X}^\infty$. However, since we have to work with several initial conditions simultaneously, we need to build up a probability space for each initial distribution.

For a given initial probability distribution μ on $\mathcal{B}(\mathsf{X})$, we construct the probability distribution P_μ on $\mathcal{F}$ so that $\mathsf{P}_\mu(\Phi_0 = x_0) = \mu(x_0)$ and for any $A \in \mathcal{F}$,

$$\mathsf{P}_\mu(\boldsymbol{\Phi} \in A \mid \Phi_0 = x_0) = \mathsf{P}_{x_0}(\boldsymbol{\Phi} \in A) \tag{3.1}$$

where P_{x_0} is the probability distribution on $\mathcal{F}$ which is obtained when the initial distribution is the point mass δ_{x_0} at x_0.

The defining characteristic of a *Markov chain* is that its future trajectories depend on its present and its past only through the current value.

To commence to formalize this, we first consider only the laws governing a trajectory of fixed length $n \geq 1$. The random variables $\{\Phi_0 \ldots \Phi_n\}$, thought of as a sequence, take values in the space $\mathsf{X}^{n+1} = \mathsf{X}_0 \times \cdots \times \mathsf{X}_n$, the $(n+1)$-fold product of copies X_i of the countable space X, equipped with the product σ-field $\mathcal{B}(\mathsf{X}^{n+1})$ which consists again of all subsets of X^{n+1}.

The conditional probability

$$\mathsf{P}^n_{x_0}(\Phi_1 = x_1, \ldots, \Phi_n = x_n) := \mathsf{P}_{x_0}(\Phi_1 = x_1, \ldots, \Phi_n = x_n), \tag{3.2}$$

defined for any sequence $\{x_0, \ldots, x_n\} \in \mathsf{X}^{n+1}$ and $x_0 \in \mathsf{X}$, and the initial probability distribution μ on $\mathcal{B}(\mathsf{X})$ completely determine the distributions of $\{\Phi_0, \ldots, \Phi_n\}$.

Definition of a Countable Space Markov Chain

The process $\boldsymbol{\Phi} = (\Phi_0, \Phi_1, \ldots)$, taking values in the path space $(\Omega, \mathcal{F}, \mathsf{P})$, is a *Markov chain* if for every n, and any sequence of states $\{x_0, x_1 \ldots x_n\}$,

$$\begin{aligned}&\mathsf{P}_\mu(\Phi_0 = x_0, \Phi_1 = x_1, \Phi_2 = x_2, \ldots, \Phi_n = x_n)\\ &\qquad = \mu(x_0)\mathsf{P}_{x_0}(\Phi_1 = x_1)\mathsf{P}_{x_1}(\Phi_1 = x_2)\ldots\mathsf{P}_{x_{n-1}}(\Phi_1 = x_n).\end{aligned} \tag{3.3}$$

The probability μ is called the *initial distribution* of the chain.

The process $\boldsymbol{\Phi}$ is a *time-homogeneous* Markov chain if the probabilities $\mathsf{P}_{x_j}(\Phi_1 = x_{j+1})$ depend only on the values of x_j, x_{j+1} and are independent of the timepoints j.

By extending this in the obvious way from events in X^n to events in X^∞ we have that the initial distribution, followed by the probabilities of transitions from one step to the next, completely define the probabilistic motion of the chain.

If $\boldsymbol{\Phi}$ is a time-homogeneous Markov chain, we write

$$P(x, y) := \mathsf{P}_x(\Phi_1 = y);$$

then the definition (3.3) can be written

$$\begin{aligned}&\mathsf{P}_\mu(\Phi_0 = x_0, \Phi_1 = x_1, \ldots, \Phi_n = x_n)\\ &\qquad = \mu(x_0)P(x_0, x_1)P(x_1, x_2)\cdots P(x_{n-1}, x_n),\end{aligned} \tag{3.4}$$

or equivalently, in terms of the conditional probabilities of the process $\boldsymbol{\Phi}$,

$$\mathsf{P}_\mu(\Phi_{n+1} = x_{n+1} \mid \Phi_n = x_n, \ldots, \Phi_0 = x_0) = P(x_n, x_{n+1}). \tag{3.5}$$

Equation (3.5) incorporates both the "loss of memory" of Markov chains and the "time-homogeneity" embodied in our definitions. It is possible to mimic this definition, asking that the $\mathsf{P}_{x_j}(\Phi_1 = x_{j+1})$ depend on the time j at which the transition takes place; but the theory for such inhomogeneous chains is neither so ripe nor so clean as for the chains we study, and we restrict ourselves solely to the time-homogeneous case in this book.

For a given model we will almost always define the probability P_{x_0} for a fixed x_0 by defining the one-step transition probabilities for the process, and building the overall distribution using (3.4).

This is done using a *Markov transition matrix.*

Transition Probability Matrix

The matrix $P = \{P(x,y), x, y \in \mathsf{X}\}$ is called a *Markov transition matrix* if

$$P(x,y) \geq 0, \quad \sum_{z \in \mathsf{X}} P(x,z) = 1, \qquad x, y \in \mathsf{X} \tag{3.6}$$

We define the usual matrix iterates $P^n = \{P^n(x,y), x, y \in \mathsf{X}\}$ by setting $P^0 = I$, the identity matrix, and then taking inductively

$$P^n(x,z) = \sum_{y \in \mathsf{X}} P(x,y) P^{n-1}(y,z). \tag{3.7}$$

In the next section we show how to take an initial distribution μ and a transition matrix P and construct a distribution P_μ so that the conditional distributions of the process may be computed as in (3.1), and so that for any x, y,

$$\mathsf{P}_\mu(\Phi_n = y \mid \Phi_0 = x) = P^n(x,y) \tag{3.8}$$

For this reason, P^n is called the *n-step transition matrix.* For $A \subseteq \mathsf{X}$, we also put

$$P^n(x,A) := \sum_{y \in A} P^n(x,y).$$

3.2.2 Developing $\boldsymbol{\Phi}$ from the transition matrix

To define a Markov chain from a transition function we first consider only the laws governing a trajectory of fixed length $n \geq 1$. The random variables $\{\Phi_0, \dots, \Phi_n\}$, thought of as a sequence, take values in the space $\mathsf{X}^{n+1} = \mathsf{X}_0 \times \cdots \times \mathsf{X}_n$, equipped with the σ-field $\mathcal{B}(\mathsf{X}^{n+1})$ which consists of all subsets of X^{n+1}.

Define the distributions P_x of $\boldsymbol{\Phi}$ inductively by setting, for each fixed $x \in \mathsf{X}$

$$\begin{aligned} \mathsf{P}_x(\Phi_0 = x) &= 1 \\ \mathsf{P}_x(\Phi_1 = y) &= P(x,y) \\ \mathsf{P}_x(\Phi_2 = z, \Phi_1 = y) &= P(x,y)P(y,z) \end{aligned}$$

and so on. It is then straightforward, but a little lengthy, to check that for each fixed x, this gives a consistent set of definitions of probabilities P_x^n on $(\mathsf{X}^n, \mathcal{B}(\mathsf{X}^n))$, and these distributions can be built up to an overall probability measure P_x for each x on the set $\Omega = \prod_{i=0}^\infty \mathsf{X}_i$ with σ-field $\mathcal{F} = \bigvee_{i=0}^\infty \mathcal{B}(\mathsf{X}_i)$, defined in the usual way. Once we prescribe an initial measure μ governing the random variable Φ_0, we can define the overall measure by

$$\mathsf{P}_\mu(\boldsymbol{\Phi} \in A) := \sum_{x \in \mathsf{X}} \mu(x)\mathsf{P}_x(\boldsymbol{\Phi} \in A)$$

to govern the overall evolution of $\boldsymbol{\Phi}$. The formula (3.1) and the interpretation of the transition function given in (3.8) follow immediately from this construction.

A careful construction is in Chung [49], Chapter I.2. This leads to

Theorem 3.2.1 *If* X *is countable, and*

$$\mu = \{\mu(x), x \in \mathsf{X}\}, \qquad\qquad P = \{P(x,y), x, y \in \mathsf{X}\}$$

are an initial measure on X *and a Markov transition matrix satisfying (3.6) then there exists a Markov chain* $\boldsymbol{\Phi}$ *on* $(\Omega, \mathcal{F})$ *with probability law* P_μ *satisfying*

$$\mathsf{P}_\mu(\Phi_{n+1} = y \mid \Phi_n = x, \dots, \Phi_0 = x_0) = P(x,y).$$

□

3.3 Specific Transition Matrices

In practice models are often built up by constructing sample paths heuristically, often for quite complicated processes, such as the queues in Section 2.4.2 and their many ramifications in the literature, and then calculating a consistent set of transition probabilities. Theorem 3.2.1 then guarantees that one indeed has an underlying stochastic process for which these probabilities make sense.

To make this more concrete, let us consider a number of the models with Markovian structure introduced in Chapter 2, and illustrate how their transition probabilities may be constructed on a countable space from physical or other assumptions.

3.3.1 The forward and backward recurrence time chains

Recall that the forward recurrence time chain $\mathbf{V}^+$ is given by

$$V^+(n) := \inf(Z_m - n : Z_m > n), \qquad n \geq 0$$

where Z_n is a renewal sequence as introduced in Section 2.4.1.

The transition matrix for $\mathbf{V}^+$ is particularly simple. If $V^+(n) = k$ for some $k > 0$, then after one time unit $V^+(n+1) = k-1$. If $V^+(n) = 1$ then a renewal occurs at $n+1$ and $V^+(n+1)$ has the distribution p of an arbitrary term in the renewal sequence. This gives the sub-diagonal structure

$$P = \begin{bmatrix} p(1) & p(2) & p(3) & p(4) & \dots \\ 1 & 0 & 0 & & \dots \\ 0 & \ddots & \ddots & & \\ & 0 & 1 & 0 & \\ \vdots & \vdots & & \ddots & \ddots \end{bmatrix}$$

The backward recurrence time chain $\mathbf{V}^-$ has a similarly simple structure. For any $n \in \mathbb{Z}_+$, let us write

$$\overline{p}(n) = \sum_{j \geq n+1} p(j). \tag{3.9}$$

Write $M = \sup(m \geq 1 : p(m) > 0)$; if $M < \infty$ then for this chain the state space $\mathsf{X} = \{0, 1, \dots, M-1\}$; otherwise $\mathsf{X} = \mathbb{Z}_+$. In either case, for $x \in \mathsf{X}$ we have (with Y as a generic increment variable in the renewal process)

$$\begin{aligned} P(x, x+1) &= \mathsf{P}(Y > x+1 \mid Y > x) = \overline{p}(x+1)/\overline{p}(x) \\ P(x, 0) &= \mathsf{P}(Y = x+1 \mid Y > x) = p(x+1)/\overline{p}(x) \end{aligned} \tag{3.10}$$

and zero otherwise. This gives a superdiagonal matrix of the form

$$P = \begin{bmatrix} b(1) & 1-b(1) & 0 & 0 & \dots \\ b(2) & 0 & 1-b(2) & 0 & \dots \\ b(3) & 0 & \ddots & 1-b(3) & \\ \vdots & \vdots & & \ddots & \ddots \end{bmatrix}$$

where we have written $b(j) = p(j+1)/\overline{p}(j)$.

These particular chains are a rich source of simple examples of stable and unstable behaviors, depending on the behavior of p; and they are also chains which will be found to be fundamental in analyzing the asymptotic behavior of an arbitrary chain.

3.3.2 Random walk models

Random walk on the integers Let us define the random walk $\boldsymbol{\Phi} = \{\Phi_n; n \in \mathbb{Z}_+\}$ by setting, as in (RW1), $\Phi_n = \Phi_{n-1} + W_n$ where now the increment variables W_n are i.i.d. random variables taking only integer values in $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$. As usual, write $\Gamma(y) = \mathsf{P}(W = y)$.

Then for $x, y \in \mathbb{Z}$, the state space of the random walk,

$$\begin{aligned} P(x, y) &= \mathsf{P}(\Phi_1 = y \mid \Phi_0 = x) \\ &= \mathsf{P}(\Phi_0 + W_1 = y \mid \Phi_0 = x) \\ &= \mathsf{P}(W_1 = y - x) \\ &= \Gamma(y - x). \end{aligned} \tag{3.11}$$

The random walk is distinguished by this translation invariant nature of the transition probabilities: the probability that the chain moves from x to y in one step depends only on the difference $x - y$ between the values.

Random walks on a half line It is equally easy to construct the transition probability matrix for the random walk on the half-line $\mathbb{Z}_+$, defined in (RWHL1).

Suppose again that $\{W_i\}$ takes values in $\mathbb{Z}$, and recall from (RWHL1) that the random walk on a half line obeys

$$\Phi_n = [\Phi_{n-1} + W_n]^+. \tag{3.12}$$

Then for $y \in \mathbb{Z}_+$, the state space of the random walk on a half line, we have as in (3.11) that for $y > 0$

$$P(x, y) = \Gamma(y - x); \tag{3.13}$$

whilst for $y = 0$,

$$\begin{aligned} P(x, 0) &= \mathsf{P}(\Phi_0 + W_1 \leq 0 \mid \Phi_0 = x) \\ &= \mathsf{P}(W_1 \leq -x) \\ &= \Gamma(-\infty, -x]. \end{aligned} \tag{3.14}$$

The simple storage model The storage model given by (SSM1)-(SSM2) is a concrete example of the structure in (3.13) and (3.14), provided the release rate is $r = 1$, the inter-input times take values $n \in \mathbb{Z}_+$ with distribution G, and the input values are also integer valued with distribution H.

The random walk on a half line describes the behavior of this storage model, and its transition matrix P therefore defines its one-step behavior. We can calculate the values of the increment distribution function Γ in a different way, in terms of the basic parameters G and H of the models, by breaking up the possibilities of the input time and the input size: we have

$$\begin{aligned}\Gamma(x) &= \mathsf{P}(S_n - J_n = x) \\ &= \textstyle\sum_{i=0}^{\infty} H(i)G(x+i).\end{aligned}$$

We have rather forced the storage model into our countable space context by assuming that the variables concerned are integer valued. We will rectify this in later sections.

3.3.3 Embedded queueing models

The GI/M/1 Queue The next context in which we illustrate the construction of the transition matrix is in the modeling of queues through their embedded chains.

Consider the random variable $N_n = N(T'_n-)$, which counts customers immediately before each arrival in a queueing system satisfying (Q1)-(Q3).

We will first construct the matrix $P = (P(x,y))$ corresponding to the number of customers $\mathbf{N} = \{N_n\}$ for the GI/M/1 queue; that is, the queue satisfying (Q4).

Proposition 3.3.1 *For the GI/M/1 queue, the sequence* $\mathbf{N} = \{N_n, n \geq 0\}$ *can be constructed as a Markov chain with state space* $\mathbb{Z}_+$ *and transition matrix*

$$P = \begin{bmatrix} q_0 & p_0 & & & \\ q_1 & p_1 & p_0 & & 0 \\ q_2 & p_2 & p_1 & p_0 & \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

where $q_j = \sum_{i=j+1}^{\infty} p_i$, *and*

$$p_0 = \mathsf{P}(S > T) = \int_0^{\infty} e^{-\mu t}\, G(dt) \tag{3.15}$$

$$\begin{aligned} p_j &= \mathsf{P}\{S'_j > T > S'_{j-1}) \\ &= \int_0^{\infty} \{e^{-\mu t}(\mu t)^j / j!\}\, G(dt), \qquad j \geq 1. \end{aligned} \tag{3.16}$$

Hence $\mathbf{N}$ *is a random walk on a half line.*

PROOF In Section 2.4.2 we established the Markovian nature of the increases at T'_n-, in (2.27), under the assumption of exponential service times.

Since we consider $N(t)$ immediately before every arrival time, N_{n+1} can only increase from N_n by one unit at most; hence for $k > 1$ it is trivial that

$$\mathsf{P}(N_{n+1} = j + k \mid N_n = j, N_{n-1}, N_{n-2}, \ldots, N_0) = 0. \tag{3.17}$$

The independence and identical distribution structure of the service times show as in Section 2.4.2 that, no matter which previous customer was being served, and when their service started,

$$\mathsf{P}(N_{n+1} = j+1 \mid N_n = j, N_{n-1}, N_{n-2}, \dots, N_0) = \int_0^\infty e^{-\mu t}\, G(dt) = p_0 \qquad (3.18)$$

as shown in equation (2.31). This establishes the upper triangular structure of P.

If $N_n = j$, then for $0 < i \le j$, we have $N_{n+1} = i$ provided exactly $(j - i + 1)$ jobs are completed in an inter-arrival period. It is an elementary property of sums of exponential random variables (see, for example, Çinlar [40], Chapter 4) that for any t, the number of services completed in a time $[0,t]$ is Poisson with parameter μt, so that

$$\mathsf{P}(S_0 + \cdots + S_{j+1} > t > S_0 + \cdots + S_j) = e^{-\mu t}(\mu t)^j / j! \qquad (3.19)$$

from which we derive (3.16).

It remains to show that $P(j,0) = q_j = \sum_{i=j+1}^{\infty} p_i$; but this follows analogously with equation (3.16), since the queue empties if more than $(j+1)$ customers complete service between arrivals.

Finally, to assert that $\mathbf{N} = \{N_n\}$ can actually be constructed in its entirety as a Markov chain on $\mathbb{Z}_+$, we appeal to the general results of Theorem 3.2.1 above to build $\mathbf{N}$ from the probabilistic building blocks $P = (P(i,j))$, and any initial distribution μ. □

The M/G/1 queue Next consider the random variables N_n^*, which count customers immediately after each service time ends in a queueing system satisfying (Q1)-(Q3).

We showed in Section 2.4.2 that this is Markovian when the inter-arrival times are exponential: that is, for an M/G/1 model satisfying (Q5).

Proposition 3.3.2 *For the M/G/1 queue, the sequence $\mathbf{N}^* = \{N_n^*, n \ge 0\}$ can be constructed as a Markov chain with state space $\mathbb{Z}_+$ and transition matrix*

$$P = \begin{bmatrix} q_0 & q_1 & q_2 & q_3 & q_4 & \cdots \\ q_0 & q_1 & q_2 & q_3 & q_4 & \cdots \\ & q_0 & q_1 & q_2 & q_3 & \cdots \\ & & q_0 & q_1 & q_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \end{bmatrix}$$

where for each $j \ge 0$

$$q_j = \int_0^\infty \{e^{-\lambda t}(\lambda t)^j / j!\}\, H(dt) \qquad j \ge 1. \qquad (3.20)$$

Hence $\mathbf{N}^*$ is similar to a random walk on a half line, but with a different modification of the transitions away from zero.

Proof Exactly as in (3.19), the expressions q_k represent the probabilities of k arrivals occurring in one service time with distribution H, when the interarrival times are independent exponential variables of rate λ. □

3.3.4 Linear models on the rationals

The discussion of the queueing models above not only gives more explicit examples of the abstract random walk models, but also indicates how the Markov assumption may or may not be satisfied, depending on how the process is constructed: we need the exponential distributions for the basic building blocks, or we do not have probabilities of transition independent of the past.

In contrast, for the simple scalar linear AR(1) models satisfying (SLM1) and (SLM2), the Markovian nature of the process is immediate. The use of a countable space here is in the main inappropriate, but some versions of this model do provide a good source of examples and counterexamples which motivate the various topological conditions we introduce in Chapter 6. Recall then that for an AR(1) model X_n and W_n are random variables on $\mathbb{R}$, satisfying

$$\mathsf{X}_n = \alpha \mathsf{X}_{n-1} + W_n,$$

for some $\alpha \in \mathbb{R}$, with the "noise" variables $\{W_n\}$ independent and identically distributed. To use the countable structure of Section 3.2 we might assume, as with the storage model in Section 3.3.2 above, that α is integer valued, and the noise variables are also integer valued.

Or, if we need to assume a countable structure on X we might, for example, find a better fit to reality by supposing that the constant α takes a rational value; and that the generic noise variable W also has a distribution on the rationals $\mathbb{Q}$, with $\mathsf{P}(W = q) = \Gamma(q)$, $q \in \mathbb{Q}$. We then have, in a very straightforward manner

Proposition 3.3.3 *Provided $x_0 \in \mathbb{Q}$, the sequence $\mathbf{X} = \{X_n, n \geq 0\}$ can be constructed as a time homogeneous Markov chain on the countable space $\mathbb{Q}$, with transition probability matrix*

$$\begin{aligned} P(r,q) &= \mathsf{P}(X_{n+1} = q \mid X_n = r) \\ &= \Gamma(q - \alpha r), \qquad r, q \in \mathbb{Q}. \end{aligned}$$

Proof We have established that $\mathbf{X}$ is Markov. Clearly, from (SLM1), when $X_0 \in \mathbb{Q}$, the value of X_1 is in $\mathbb{Q}$ also; and $P(r,q)$ merely describes the fact that the chain moves from r to αr in a deterministic way before adding the noise with distribution W.

Again, once we have $P = \{P(r,q), r, q \in \mathbb{Q}\}$, we are guaranteed the existence of the Markov chain $\mathbf{X}$, using the results of Theorem 3.2.1 with P as transition probability matrix. □

This autoregression highlights immediately the shortcomings of the countable state space structure. Although $\mathbb{Q}$ is countable, so that in a formal sense we can construct a linear model satisfying (SLM1) and (SLM2) on $\mathbb{Q}$ in such a way that we can use countable space Markov chain theory, it is clearly more natural to take, say, α as real and the variable W as real-valued also, so that X_n is real-valued for any initial $x_0 \in \mathbb{R}$.

To model such processes, and the more complex autoregressions and nonlinear models which generalize them in Chapter 2, and which are clearly Markovian but continuous-valued in conception, we need a theory for continuous-valued Markov chains. We turn to this now.

3.4 Foundations for General State Space Chains

3.4.1 Developing Φ from transition probabilities

The countable space approach guides the development of the theory we shall present in this book for a much broader class of Markov chains, on quite general state spaces: it is one of the more remarkable features of this seemingly sweeping generalization that the great majority of the countable state space results carry over virtually unchanged, without assuming any detailed structure on the space.

We let X be a general set, and $\mathcal{B}(\mathsf{X})$ denote a countably generated σ-field on X: when X is topological, then $\mathcal{B}(\mathsf{X})$ will be taken as the Borel σ-field, but otherwise it may be arbitrary.

In this case we again start from the one-step transition probabilities and construct Φ much as in Theorem 3.2.1.

Transition Probability Kernels

If $P = \{P(x,A), x \in \mathsf{X}, A \in \mathcal{B}(\mathsf{X})\}$ is such that

(i) for each $A \in \mathcal{B}(\mathsf{X})$, $P(\,\cdot\,,A)$ is a non-negative measurable function on X

(ii) for each $x \in \mathsf{X}$, $P(x,\,\cdot\,)$ is a probability measure on $\mathcal{B}(\mathsf{X})$

then we call P a *transition probability kernel* or *Markov transition function.*

On occasion, as in Chapter 6, we may require that a collection $T = \{T(x,A), x \in \mathsf{X}, A \in \mathcal{B}(\mathsf{X})\}$ satisfies (i) and (ii), with the exception that $T(x,\mathsf{X}) \le 1$ for each x: such a collection is called a *substochastic* transition kernel. In the other direction, there will be times when we need to consider completely non-probabilistic mappings $K: \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}_+$ with $K(x,\,\cdot\,)$ a measure on $\mathcal{B}(\mathsf{X})$ for each x, and $K(\,\cdot\,,B)$ a measurable function on X for each $B \in \mathcal{B}(\mathsf{X})$. Such a map is called a *kernel* on $(\mathsf{X},\mathcal{B}(\mathsf{X}))$.

We now imitate the development on a countable space to see that from the transition probability kernel P we can define a stochastic process with the appropriate Markovian properties, for which P will serve as a description of the one-step transition laws.

We first define a finite sequence $\Phi = \{\Phi_0, \Phi_1, \dots, \Phi_n\}$ of random variables on the product space $\mathsf{X}^{n+1} = \prod_{i=0}^{n} \mathsf{X}_i$, equipped with the product σ-field $\bigvee_{i=0}^{n} \mathcal{B}(\mathsf{X}_i)$, by an inductive procedure.

For any measurable sets $A_i \subseteq \mathsf{X}_i$, we develop the set functions $\mathsf{P}^n_x(\cdot)$ on X^{n+1} by setting, for a fixed starting point $x \in \mathsf{X}$ and for the "cylinder sets" $A_1 \times \cdots \times A_n$

$$\begin{aligned}
\mathsf{P}^1_x(A_1) &= P(x, A_1), \\
\mathsf{P}^2_x(A_1 \times A_2) &= \int_{A_1} P(x, dy_1) P(y_1, A_2), \\
&\vdots \\
\mathsf{P}^n_x(A_1 \times \cdots \times A_n) &= \int_{A_1} P(x, dy_1) \int_{A_2} P(y_1, dy_2) \cdots P(y_{n-1}, A_n).
\end{aligned}$$

These are all well-defined by the measurability of the integrands $P(\,\cdot\,,\,\cdot\,)$ in the first variable, and the fact that the kernels are measures in the second variable.

If we now extend P^n_x to all of $\bigvee_0^n \mathcal{B}(\mathsf{X}_i)$ in the usual way [25] and repeat this procedure for increasing n, we find

Theorem 3.4.1 *For any initial measure μ on $\mathcal{B}(\mathsf{X})$, and any transition probability kernel $P = \{P(x, A), x \in \mathsf{X}, A \in \mathcal{B}(\mathsf{X})\}$, there exists a stochastic process $\boldsymbol{\Phi} = \{\Phi_0, \Phi_1, \ldots\}$ on $\Omega = \prod_{i=0}^{\infty} \mathsf{X}_i$, measurable with respect to $\mathcal{F} = \bigvee_{i=0}^{\infty} \mathcal{B}(\mathsf{X}_i)$, and a probability measure P_μ on $\mathcal{F}$ such that $\mathsf{P}_\mu(B)$ is the probability of the event $\{\boldsymbol{\Phi} \in B\}$ for $B \in \mathcal{F}$; and for measurable $A_i \subseteq \mathsf{X}_i, i = 0, \ldots, n$, and any n*

$$\begin{aligned}
&\mathsf{P}_\mu(\Phi_0 \in A_0, \Phi_1 \in A_1, \ldots, \Phi_n \in A_n) \qquad (3.21)\\
&\quad = \int_{y_0 \in A_0} \cdots \int_{y_{n-1} \in A_{n-1}} \mu(dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, A_n).
\end{aligned}$$

Proof Because of the consistency of definition of the set functions P^n_x, there is an overall measure P_x for which the P^n_x are finite dimensional distributions, which leads to the result: the details are relatively standard measure theoretic constructions, and are given in the general case by Revuz [223], Theorem 2.8 and Proposition 2.11; whilst if the space has a suitable topology, as in (MC1), then the existence of $\boldsymbol{\Phi}$ is a straightforward consequence of Kolmogorov's Consistency Theorem for construction of probabilities on topological spaces. □

The details of this construction are omitted here, since it suffices for our purposes to have indicated why transition probabilities generate processes, and to have spelled out that the key equation (3.21) is a reasonable representation of the behavior of the process in terms of the kernel P.

We can now formally define

Markov Chains on General Spaces

The stochastic process $\boldsymbol{\Phi}$ defined on $(\Omega, \mathcal{F})$ is called a time-homogeneous Markov chain with transition probability kernel $P(x, A)$ and initial distribution μ if the finite dimensional distributions of $\boldsymbol{\Phi}$ satisfy (3.21) for every n.

3.4.2 The n-step transition probability kernel

As on countable spaces the *n-step transition probability kernel* is defined iteratively. We set $P^0(x, A) = \delta_x(A)$, the Dirac measure defined by

$$\delta_x(A) = \begin{cases} 1 & x \in A \\ 0 & x \notin A, \end{cases} \tag{3.22}$$

and, for $n \geq 1$, we define inductively

$$P^n(x, A) = \int_{\mathsf{X}} P(x, dy) P^{n-1}(y, A), \quad x \in \mathsf{X}, A \in \mathcal{B}(\mathsf{X}). \tag{3.23}$$

We write P^n for the n-step transition probability kernel $\{P^n(x, A), x \in \mathsf{X}, A \in \mathcal{B}(\mathsf{X})\}$: note that P^n is defined analogously to the n-step transition probability matrix for the countable space case.

As a first application of the construction equations (3.21) and (3.23), we have the celebrated Chapman-Kolmogorov equations. These underlie, in one form or another, virtually all of the solidarity structures we develop.

Theorem 3.4.2 *For any m with $0 \leq m \leq n$,*

$$P^n(x, A) = \int_{\mathsf{X}} P^m(x, dy) P^{n-m}(y, A), \quad x \in \mathsf{X}, \ A \in \mathcal{B}(\mathsf{X}). \tag{3.24}$$

Proof In (3.21), choose $\mu = \delta_x$ and integrate over sets $A_i = \mathsf{X}$ for $i = 1, \ldots, n-1$; and use the definition of P^m and P^{n-m} for the first m and the last $n-m$ integrands. □

We interpret (3.24) as saying that, as $\boldsymbol{\Phi}$ moves from x into A in n steps, at any intermediate time m it must take (obviously) some value $y \in \mathsf{X}$; and that, being a Markov chain, it forgets the past at that time m and moves the succeeding $(n-m)$ steps with the law appropriate to starting afresh at y. We can write equation (3.24) alternatively as

$$\mathsf{P}_x(\Phi_n \in A) = \int_{\mathsf{X}} \mathsf{P}_x(\Phi_m \in dy) \mathsf{P}_y(\Phi_{n-m} \in A). \tag{3.25}$$

Exactly as the one-step transition probability kernel describes a chain $\boldsymbol{\Phi}$, the m-step kernel (viewed in isolation) satisfies the definition of a transition kernel, and thus defines a Markov chain $\boldsymbol{\Phi}^m = \{\Phi_n^m\}$ with transition probabilities

$$\mathsf{P}_x(\Phi_n^m \in A) = P^{mn}(x, A). \tag{3.26}$$

This, and several other transition functions obtained from P, will be used widely in the sequel.

Skeletons and Resolvents

The chain $\boldsymbol{\Phi}^m$ with transition law (3.26) is called the *m-skeleton* of the chain $\boldsymbol{\Phi}$.

The *resolvent* K_{a_ε} is defined for $0 < \varepsilon < 1$ by

$$K_{a_\varepsilon}(x, A) := (1 - \varepsilon) \sum_{i=0}^{\infty} \varepsilon^i P^i(x, A), \qquad x \in \mathsf{X},\ A \in \mathcal{B}(\mathsf{X}).$$

The Markov chain with transition function K_{a_ε} is called the *K_{a_ε}-chain.*

This nomenclature is taken from the continuous-time literature, but we will see that in discrete time the m-skeletons and resolvents of the chain also provide a useful tool for analysis.

There is one substantial difference in moving to the general case from the countable case, which flows from the fact that the kernel P^n can no longer be viewed as symmetric in its two arguments.

In the general case the kernel P^n operates on quite different entities from the left and the right. As an operator P^n acts on both bounded measurable functions f on X and on σ-finite measures μ on $\mathcal{B}(\mathsf{X})$ via

$$P^n f\,(x) = \int_{\mathsf{X}} P^n(x, dy) f(y), \qquad \mu P^n\,(A) = \int_{\mathsf{X}} \mu(dx) P^n(x, A),$$

and we shall use the notation $P^n f, \mu P^n$ to denote these operations. We shall also write

$$P^n(x, f) := \int P^n(x, dy) f(y) := \delta_x P^n f$$

if it is notationally convenient. In general, the functional notation is more compact: for example, we can rewrite the Chapman-Kolmogorov equations as

$$P^{m+n} = P^m P^n, \qquad m, n \in \mathbb{Z}_+.$$

On many occasions, though, where we feel that the argument is more transparent when written in full form we shall revert to the more detailed presentation.

The form of the Markov chain definitions we have given to date concern only the probabilities of events involving $\boldsymbol{\Phi}$. We now define the expectation operation E_μ corresponding to P_μ.

For cylinder sets we define E_μ by

$$\mathsf{E}_\mu[\mathbb{1}_{A_0 \times \cdots \times A_n}(\Phi_0, \ldots, \Phi_n)] := \mathsf{P}_\mu(\{\Phi_0, \ldots, \Phi_n\} \in A_0 \times \cdots \times A_n),$$

where $\mathbb{1}_B$ denotes the indicator function of a set B. We may extend the definition to that of $\mathsf{E}_\mu[h(\Phi_0, \Phi_1, \ldots)]$ for any measurable bounded real-valued function h on Ω by requiring that the expectation be linear.

By linearity of the expectation, we can also extend the Markovian relationship (3.21) to express the Markov property in the following equivalent form. We omit the details, which are routine.

Proposition 3.4.3 *If $\boldsymbol{\Phi}$ is a Markov chain on $(\Omega, \mathcal{F})$, with initial measure μ, and $h\colon \Omega \to \mathbb{R}$ is bounded and measurable, then*

$$\mathsf{E}_\mu[h(\Phi_{n+1}, \Phi_{n+2}, \dots) \mid \Phi_0, \dots, \Phi_n;\ \Phi_n = x] = \mathsf{E}_x[h(\Phi_1, \Phi_2, \dots)]. \tag{3.27}$$

□

The formulation of the Markov concept itself is made much simpler if we develop more systematic notation for the information encompassed in the past of the process, and if we introduce the "shift operator" on the space Ω.

For a given initial distribution, define the σ-field

$$\mathcal{F}_n^\Phi := \sigma(\Phi_0, \dots, \Phi_n) \subseteq \mathcal{B}(\mathsf{X}^{n+1})$$

which is the smallest σ-field for which the random variable $\{\Phi_0, \dots, \Phi_n\}$ is measurable. In many cases, $\mathcal{F}_n^\Phi$ will coincide with $\mathcal{B}(\mathsf{X}^n)$, although this depends in particular on the initial measure μ chosen for a particular chain.

The *shift operator* θ is defined to be the mapping on Ω defined by

$$\theta(\{x_0, x_1, \dots, x_n, \dots\}) = \{x_1, x_2, \dots, x_{n+1}, \dots\}.$$

We write θ^k for the k^{th} iterate of the mapping θ, defined inductively by

$$\theta^1 = \theta, \qquad \theta^{k+1} = \theta \circ \theta^k, \quad k \geq 1.$$

The shifts θ^k define operators on random variables H on $(\Omega, \mathcal{F}, P_\mu)$ by

$$(\theta^k H)(w) = H \circ \theta^k(\omega).$$

It is obvious that $\Phi_n \circ \theta^k(\omega) = \Phi_{n+k}$. Hence if the random variable H is of the form $H = h(\Phi_0, \Phi_1, \dots)$ for a measurable function h on the sequence space Ω then

$$\theta^k H = h(\Phi_k, \Phi_{k+1}, \dots)$$

Since the expectation $\mathsf{E}_x[H]$ is a measurable function on X, it follows that $\mathsf{E}_{\Phi_n}[H]$ is a random variable on $(\Omega, \mathcal{F}, \mathsf{P}_\mu)$ for any initial distribution. With this notation the equation

$$\mathsf{E}_\mu[\theta^n H \mid \mathcal{F}_n^\Phi] = \mathsf{E}_{\Phi_n}[H] \qquad \text{a.s. } [\mathsf{P}_\mu] \tag{3.28}$$

valid for any bounded measurable h and fixed $n \in \mathbb{Z}_+$, describes the *time homogeneous Markov property* in a succinct way.

It is not always the case that $\mathcal{F}_n^\Phi$ is complete: that is, contains every set of P_μ-measure zero. We adopt the following convention as in [223]. For any initial measure μ we say that an event A occurs P_μ-a.s. to indicate that A^c is a set contained in an element of $\mathcal{F}_n^\Phi$ which is of P_μ-measure zero.

If A occurs P_x-a.s. for all $x \in \mathsf{X}$ then we write that A occurs P_*-a.s.

3.4.3 Occupation, hitting and stopping times

The distributions of the chain $\boldsymbol{\Phi}$ at time n are the basic building blocks of its existence, but the analysis of its behavior concerns also the distributions at certain random times in its evolution, and we need to introduce these now.

Occupation Times, Return Times and Hitting Times

(i) For any set $A \in \mathcal{B}(\mathsf{X})$, the *occupation time* η_A is the number of visits by $\boldsymbol{\Phi}$ to A after time zero, and is given by

$$\eta_A := \sum_{n=1}^{\infty} \mathbb{1}\{\Phi_n \in A\}.$$

(ii) For any set $A \in \mathcal{B}(\mathsf{X})$, the variables

$$\begin{array}{rcl} \tau_A & := & \min\{n \ge 1 : \Phi_n \in A\} \\ \sigma_A & := & \min\{n \ge 0 : \Phi_n \in A\} \end{array}$$

are called the *first return* and *first hitting* times on A, respectively.

For every $A \in \mathcal{B}(\mathsf{X})$, η_A, τ_A and σ_A are obviously measurable functions from Ω to $\mathbb{Z}_+ \cup \{\infty\}$.

Unless we need to distinguish between different returns to a set, then we call τ_A and σ_A the return and hitting times on A respectively. If we do wish to distinguish different return times, we write $\tau_A(k)$ for the random time of the k^{th} visit to A: these are defined inductively for any A by

$$\begin{array}{rcl} \tau_A(1) & := & \tau_A \\ \tau_A(k) & := & \min\{n > \tau_A(k-1) : \Phi_n \in A\}. \end{array} \tag{3.29}$$

Analysis of $\boldsymbol{\Phi}$ involves the kernel U defined as

$$\begin{array}{rcl} U(x,A) & := & \displaystyle\sum_{n=1}^{\infty} P^n(x,A) \\ & = & \mathsf{E}_x[\eta_A] \end{array} \tag{3.30}$$

which maps $\mathsf{X} \times \mathcal{B}(\mathsf{X})$ to $\mathbb{R} \cup \{\infty\}$, and the return time probabilities

$$\begin{array}{rcl} L(x,A) & := & \mathsf{P}_x(\tau_A < \infty) \\ & = & \mathsf{P}_x(\boldsymbol{\Phi}\text{ ever enters } A). \end{array} \tag{3.31}$$

In order to analyze numbers of visits to sets, we often need to consider the behavior after the first visit τ_A to a set A (which is a random time), rather than behavior after fixed times. One of the most crucial aspects of Markov chain theory is that the "forgetfulness" properties in equation (3.21) or equation (3.27) hold, not just for fixed times n, but for the chain interrupted at certain random times, called *stopping times*, and we now introduce these ideas.

Stopping Times

A function $\zeta\colon \Omega \to \mathbb{Z}_+ \cup \{\infty\}$ is a *stopping time* for $\boldsymbol{\Phi}$ if for any initial distribution μ the event $\{\zeta = n\} \in \mathcal{F}_n^{\Phi}$ for all $n \in \mathbb{Z}_+$.

The first return and the hitting times on sets provide simple examples of stopping times.

Proposition 3.4.4 *For any set $A \in \mathcal{B}(\mathsf{X})$, the variables τ_A and σ_A are stopping times for $\boldsymbol{\Phi}$.*

Proof Since we have

$$\begin{aligned} \{\tau_A = n\} &= \cap_{m=1}^{n-1}\{\Phi_m \in A^c\} \cap \{\Phi_n \in A\} \in \mathcal{F}_n^{\Phi}, \quad n \geq 1 \\ \{\sigma_A = n\} &= \cap_{m=0}^{n-1}\{\Phi_m \in A^c\} \cap \{\Phi_n \in A\} \in \mathcal{F}_n^{\Phi}, \quad n \geq 0 \end{aligned}$$

it follows from the definitions that τ_A and σ_A are stopping times. □

We can construct the full distributions of these stopping times from the basic building blocks governing the motion of $\boldsymbol{\Phi}$, namely the elements of the transition probability kernel, using the Markov property for each fixed $n \in \mathbb{Z}_+$. This gives

Proposition 3.4.5 (i) *For all $x \in \mathsf{X}$, $A \in \mathcal{B}(\mathsf{X})$*

$$\mathsf{P}_x(\tau_A = 1) = P(x, A),$$

and inductively for $n > 1$

$$\begin{aligned} \mathsf{P}_x(\tau_A = n) &= \int_{A^c} P(x, dy)\mathsf{P}_y(\tau_A = n-1) \\ &= \int_{A^c} P(x, dy_1) \int_{A^c} P(y_1, dy_2) \cdots \\ &\qquad \int_{A^c} P(y_{n-2}, dy_{n-1}) P(y_{n-1}, A). \end{aligned}$$

(ii) *For all $x \in \mathsf{X}$, $A \in \mathcal{B}(\mathsf{X})$*

$$\mathsf{P}_x(\sigma_A = 0) = \mathbb{1}_A(x)$$

and for $n \geq 1$, $x \in A^c$

$$\mathsf{P}_x(\sigma_A = n) = \mathsf{P}_x(\tau_A = n).$$

□

If we use the kernel I_B defined as $I_B(x, A) := \mathbb{1}_{A \cap B}(x)$, we have, in more compact functional notation,

$$\mathsf{P}_x(\tau_A = k) = \left[(P I_{A^c})^{k-1} P\right](x, A).$$

From this we obtain the formula

$$L(x, A) := \sum_{k=1}^{\infty} \left[(P I_{A^c})^{k-1} P\right](x, A)$$

for the return time probability to a set A starting from the state x.

The simple Markov property (3.28) holds for any bounded measurable h and fixed $n \in \mathbb{Z}_+$. We now extend (3.28) to stopping times.

If ζ is an arbitrary stopping time, then the fact that our time-set is $\mathbb{Z}_+$ enables us to define the random variable Φ_ζ by setting $\Phi_\zeta = \Phi_n$ on the event $\{\zeta = n\}$. For a stopping time ζ the property which tells us that the future evolution of $\boldsymbol{\Phi}$ after the stopping time depends only on the value of Φ_ζ, rather than on any other past values, is called the Strong Markov Property.

To describe this formally, we need to define the σ-field $\mathcal{F}_\zeta^\Phi := \{A \in \mathcal{F} : \{\zeta = n\} \cap A \in \mathcal{F}_n^\Phi, n \in \mathbb{Z}_+\}$, which describes events which happen "up to time ζ".

For a stopping time ζ and a random variable $H = h(\Phi_0, \Phi_1, \ldots)$ the shift θ^ζ is defined as

$$\theta^\zeta H = h(\Phi_\zeta, \Phi_{\zeta+1}, \ldots),$$

on the set $\{\zeta < \infty\}$. The required extension of (3.28) is then

The Strong Markov Property

We say $\boldsymbol{\Phi}$ has the *Strong Markov Property* if for any initial distribution μ, any real-valued bounded measurable function h on Ω, and any stopping time ζ,

$$\mathsf{E}_\mu[\theta^\zeta H \mid \mathcal{F}_\zeta^\Phi] = \mathsf{E}_{\Phi_\zeta}[H] \quad \text{a.s. } [\mathsf{P}_\mu], \tag{3.32}$$

on the set $\{\zeta < \infty\}$.

Proposition 3.4.6 *For a Markov chain $\boldsymbol{\Phi}$ with discrete time parameter, the Strong Markov Property always holds.*

Proof This result is a simple consequence of decomposing the expectations on both sides of (3.32) over the set where $\{\zeta = n\}$, and using the ordinary Markov property, in the form of equation (3.28), at each of these fixed times n. □

We are not always interested only in the times of visits to particular sets. Often the quantities of interest involve conditioning on such visits being in the future.

Taboo Probabilities

We define the n-step *taboo probabilities* as

$$ {}_AP^n(x,B) := \mathsf{P}_x(\Phi_n \in B, \tau_A \geq n), \qquad x \in \mathsf{X},\ A, B \in \mathcal{B}(\mathsf{X}). $$

The quantity ${}_AP^n(x,B)$ denotes the probability of a transition to B in n steps of the chain, "avoiding" the set A. As in Proposition 3.4.5 these satisfy the iterative relation

$$ {}_AP^1(x,B) = P(x,B) $$

and for $n > 1$

$$ {}_AP^n(x,B) = \int_{A^c} P(x,dy)\,{}_AP^{n-1}(y,B), \qquad x \in \mathsf{X},\ A, B \in \mathcal{B}(\mathsf{X}), \tag{3.33} $$

or, in operator notation, ${}_AP^n(x,B) = [(PI_{A^c})^{n-1}P](x,B)$.

We will also use extensively the notation

$$ U_A(x,B) := \sum_{n=1}^{\infty} {}_AP^n(x,B), \qquad x \in \mathsf{X},\ A, B \in \mathcal{B}(\mathsf{X}); \tag{3.34} $$

note that this extends the definition of L in (3.31) since

$$ U_A(x,A) = L(x,A), \qquad x \in \mathsf{X}. $$

3.5 Building Transition Kernels For Specific Models

3.5.1 Random walk on a half line

Let $\boldsymbol{\Phi}$ be a random walk on a half line, where now we do not restrict the increment distribution to be integer-valued. Thus $\{W_i\}$ is a sequence of i.i.d. random variables taking values in $\mathbb{R} = (-\infty, \infty)$, with distribution function $\Gamma(A) = \mathsf{P}(W \in A)$, $A \in \mathcal{B}(\mathbb{R})$.

For any $A \subseteq (0,\infty)$, we have by the arguments we have used before

$$\begin{aligned} P(x,A) &= \mathsf{P}(\Phi_0 + W_1 \in A \mid \Phi_0 = x) \\ &= \mathsf{P}(W_1 \in A - x) \\ &= \Gamma(A - x), \end{aligned} \tag{3.35}$$

whilst

$$\begin{aligned} P(x,\{0\}) &= \mathsf{P}(\Phi_0 + W_1 \leq 0 \mid \Phi_0 = x) \\ &= \mathsf{P}(W_1 \leq -x) \\ &= \Gamma(-\infty, -x]. \end{aligned} \tag{3.36}$$

These models are often much more appropriate in applications than random walks restricted to integer values.

3.5.2 Storage and queueing models

Consider the Moran dam model given by (SSM1)-(SSM2), in the general case where $r > 0$, the inter-input times have distribution G; and the input values have distribution H.

The model of a random walk on a half line with transition probability kernel P given by (3.36) defines the one-step behavior of the storage model. As for the integer valued case, we calculate the distribution function Γ explicitly by breaking up the possibilities of the input time and the input size, to get a similar convolution form for Γ in terms of G and H:

$$\begin{aligned} \Gamma(A) &= \mathsf{P}(S_n - J_n \in A) \\ &= \int_0^\infty G(A/r + y/r)\, H(dy), \end{aligned} \tag{3.37}$$

where as usual the set $A/r := \{y : ry \in A\}$.

The model (3.37) is of a storage system, and we have phrased the terms accordingly. The same transition law applies to the many other models of this form: inventories, insurance models, and models of the residual service in a GI/G/1 queue, which were mentioned in Section 2.5.

In Section 3.5.4 below we will develop the transition probability structure for a more complex system which can also be used to model the dynamics of the GI/G/1 queue.

3.5.3 Renewal processes and related chains

We now consider a real-valued renewal process: this extends the countable space version of Section 2.4.1 and is closely related to the residual service time mentioned above.

Let $\{Y_1, Y_2, \ldots\}$ be a sequence of independent and identical random variables, now with distribution function Γ concentrated, not on the whole real line nor on $\mathbb{Z}_+$, but rather on $\mathbb{R}_+$. Let Y_0 be a further independent random variable, with the distribution of Y_0 being Γ_0, also concentrated on $\mathbb{R}_+$. The random variables

$$Z_n := \sum_{i=0}^{n} Y_i$$

are again called a *delayed renewal process*, with Γ_0 being the distribution of the delay described by the first variable. If $\Gamma_0 = \Gamma$ then the sequence $\{Z_n\}$ is again referred to as a renewal process.

As with the integer-valued case, write $\Gamma_0 * \Gamma$ for the convolution of Γ_0 and Γ given by

$$\Gamma_0 * \Gamma\,(dt) := \int_0^t \Gamma(dt-s)\,\Gamma_0(ds) = \int_0^t \Gamma_0(dt-s)\,\Gamma(ds) \tag{3.38}$$

and Γ^{n*} for the n^{th} convolution of Γ with itself. By decomposing successively over the values of the first n variables $Z_0, \dots, Z_{n-1}$ we have that

$$\mathsf{P}(Z_n \in dt) = \Gamma_0 * \Gamma^{n*}\,(dt)$$

and so the *renewal measure* given by $U(-\infty, t] = \sum_0^\infty \Gamma^{n*}\,(-\infty, t]$ has the interpretation

$$U[0,t] = \mathsf{E}_0[\text{number of renewals in } [0,t]]$$

and

$$\Gamma_0 * U\,[0,t] = \mathsf{E}_{\Gamma_0}[\text{number of renewals in } [0,t]],$$

where E_0 refers to the expectation when the first renewal is at 0, and E_{Γ_0} refers to the expectation when the first renewal has distribution Γ_0.

It is clear that Z_n is a Markov chain: its transition probabilities are given by

$$P(x,A) = \mathsf{P}(Z_n \in A \mid Z_{n-1} = x) = \Gamma(A-x)$$

and so Z_n is a random walk. It is not a very stable one, however, as it moves inexorably to infinity with each new step.

The forward and backward recurrence time chains, in contrast to the renewal process itself, exhibit a much greater degree of stability: they grow, then they diminish, then they grow again.

Forward and backward recurrence time chains

If $\{Z_n\}$ is a renewal process with no delay, then we call the process

(RT3)

$$V^+(t) := \inf(Z_n - t : Z_n > t,\ n \geq 1), \qquad t \geq 0 \qquad (3.39)$$

the *forward recurrence time process*; and for any $\delta > 0$, the discrete time chain $\mathbf{V}^+_\delta = \{V^+_\delta(n) = V^+(n\delta),\, n \in \mathbb{Z}_+\}$ is called the *forward recurrence time δ-skeleton.*

We call the process

(RT4)

$$V^-(t) := \inf(t - Z_n : Z_n \leq t,\ n \geq 1), \qquad t \geq 0$$

the *backward recurrence time process*; and for any $\delta > 0$, the discrete time chain $\mathbf{V}^-_\delta = \{V^-_\delta(n) = V^-(n\delta),\ n \in \mathbb{Z}_+\}$ is called the *backward recurrence time δ-skeleton.*

No matter what the structure of the renewal sequence (and in particular, even if Γ is not exponential), the forward and backward recurrence time δ-skeletons $\mathbf{V}^+_\delta$ and $\mathbf{V}^-_\delta$ are Markovian.

To see this for the forward chain, note that if $x > \delta$, then the transition probabilities P^δ of $\mathbf{V}^+_\delta$ are merely

$$P^\delta(x, \{x - \delta\}) = 1$$

whilst if $x \leq \delta$ we have, by decomposing over the time and the index of the last renewal in the period after the current forward recurrence time finishes, and using the independence of the variables Y_i

$$\begin{aligned} P^\delta(x, A) &= \int_0^{\delta - x} \sum_{n=0}^{\infty} \Gamma^{n*}(dt)\Gamma(A - [\delta - x] - t) \\ &= \int_0^{\delta - x} U(dt)\Gamma(A - [\delta - x] - t). \end{aligned} \qquad (3.40)$$

For the backward chain we have similarly that for all x

$$\mathsf{P}(V^-(n\delta) = x + \delta \mid V^-((n-1)\delta) = x) = \Gamma(x + \delta, \infty)/\Gamma(x, \infty)$$

whilst for $dv \subset [0, \delta]$

$$\mathsf{P}(V^-(n\delta) \in dv \mid V^-((n-1)\delta) = x) = \int_x^{x+\delta} \Gamma(du) U(dv - (u - x) - \delta) \frac{\Gamma(v, \infty)}{[\Gamma(x, \infty)]^{-1}}.$$

3.5.4 Ladder chains and the GI/G/1 queue

The GI/G/1 queue satisfies the conditions (Q1)-(Q3). Although the residual service time process of the GI/G/1 queue can be analyzed using the model (3.37), the more detailed structure involving actual numbers in the queue in the case of general (i.e. non-exponential) service and input times requires a more complex state space for a Markovian analysis.

We saw in Section 3.3.3 that when the service time distribution H is exponential, we can define a Markov chain by

$$N_n = \{ \text{ number of customers at } T'_n-, n = 1, 2, \dots \},$$

whilst we have a similarly embedded chain after the service times if the inter-arrival time is exponential. However, the numbers in the queue, even at the arrival or departure times, are not Markovian without such exponential assumptions.

The key step in the general case is to augment $\{N_n\}$ so that we do get a Markov model. This augmentation involves combining the information on the numbers in the queue with the information in the residual service time

To do this we introduce a bivariate "ladder chain" on a "ladder" space $\mathbb{Z}_+ \times \mathbb{R}$, with a countable number of rungs indexed by the first variable and with each rung constituting a copy of the real line.

This construction is in fact more general than that for the GI/G/1 queue alone, and we shall use the ladder chain model for illustrative purposes on a number of occasions.

Define the Markov chain $\boldsymbol{\Phi} = \{\Phi_n\}$ on $\mathbb{Z}_+ \times \mathbb{R}$ with motion defined by the transition probabilities $P(i, x; j \times A)$, $i, j \in \mathbb{Z}_+$, $x \in \mathbb{R}$, $A \in \mathcal{B}(\mathbb{R})$ given by

$$\begin{aligned} P(i,x;j \times A) &= 0 \quad j > i+1 \\ P(i,x;j \times A) &= \Lambda_{i-j+1}(x,A), \quad j = 1,\dots,i+1 \\ P(i,x;0 \times A) &= \Lambda_i^*(x,A). \end{aligned} \tag{3.41}$$

where each of the Λ_i, Λ_i^* is a substochastic transition probability kernel on $\mathbb{R}$ in its own right.

The translation invariant and "skip-free to the right" nature of the movement of this chain, incorporated in (3.42), indicates that it is a generalization of those random walks which occur in the GI/M/1 queue, as delineated in Proposition 3.3.1. We have

$$P = \begin{bmatrix} \Lambda_0^* & \Lambda_0 & & & \\ \Lambda_1^* & \Lambda_1 & \Lambda_0 & 0 & \\ \Lambda_2^* & \Lambda_2 & \Lambda_1 & \Lambda_0 & \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

where now the Λ_i, Λ_i^* are substochastic transition probability kernels rather than mere scalars.

To use this construction in the GI/G/1 context we write

$$\Phi_n = (N_n, R_n), \quad n \geq 1$$

where as before N_n is the number of customers at T'_n- and

$$R_n = \{\text{total residual service time in the system at } T'_n+\} :$$

then $\boldsymbol{\Phi} = \{\Phi_n; n \in \mathbb{Z}_+\}$ can be realised as a Markov chain with the structure (3.42), as we now demonstrate by constructing the transition kernel P explicitly.

As in (Q1)-(Q3) let H denote the distribution function of service times, and G denote the distribution function of interarrival times; and let $Z_1, Z_2, Z_3, \dots$ denote an undelayed renewal process with $Z_n - Z_{n-1} = S_n$ having the service distribution function H, as in (2.26). This differs from the process of completion points of services in that the latter may have longer intervals when there is no customer present, after completion of a busy cycle.

Let R_t denote the forward recurrence time in the renewal process $\{Z_k\}$ at time t in this process, i.e., $R_t = Z_{N(t)+1} - t$, where $N(t) = \sup\{n : Z_n \le t\}$ as in (3.39). If $R_0 = x$ then $Z_1 = x$. Now write

$$P^t_n(x, y) = \mathsf{P}(Z_n \le t < Z_{n+1}, R_t \le y \mid R_0 = x) \tag{3.42}$$

for the probability that, in this renewal process n "service times" are completed in $[0, t]$ and that the residual time of current service at t is in $[0, y]$, given $R_0 = x$.

With these definitions it is easy to verify that the chain $\boldsymbol{\Phi}$ has the form (3.42) with the specific choice of the substochastic transition kernels Λ_i, Λ^*_i given by

$$\Lambda_n(x, [0, y]) = \int_0^\infty P^t_n(x, y)\, G(dt) \tag{3.43}$$

and

$$\Lambda^*_n(x, [0, y]) = \Bigl[\sum_{n+1}^{\infty} \Lambda_j(x, [0, \infty))\Bigr] H[0, y]. \tag{3.44}$$

3.5.5 State space models

The simple nonlinear state space model is a very general model and, consequently, its transition function has an unstructured form until we make more explicit assumptions in particular cases. The general functional form which we construct here for the scalar SNSS(F) model of Section 2.2.1 will be used extensively, as will the techniques which are used in constructing its form.

For any bounded and measurable function $h\colon \mathsf{X} \to \mathbb{R}$ we have from (SNSS1),

$$h(X_{n+1}) = h(F(X_n, W_{n+1}))$$

Since $\{W_n\}$ is assumed i.i.d. in (SNSS2) we see that

$$\begin{aligned} Ph\,(x) &= \mathsf{E}[h(X_{n+1}) \mid X_n = x] \\ &= \mathsf{E}[h(F(x, W))] \end{aligned}$$

where W is a generic noise variable. Since Γ denotes the distribution of W, this becomes

$$Ph\,(x) = \int_{-\infty}^{\infty} h(F(x, w))\, \Gamma(dw)$$

and by specializing to the case where $h = \mathbb{1}_A$, we see that for any measurable set A and any $x \in \mathsf{X}$,

$$P(x, A) = \int_{-\infty}^{\infty} \mathbb{1}\{F(x, w) \in A\}\, \Gamma(dw).$$

To construct the k-step transition probability, recall from (2.5) that the transition maps for the SNSS(F) model are defined by setting $F_0(x) = x$, $F_1(x_0, w_1) = F(x_0, w_1)$, and for $k \geq 1$,

$$F_{k+1}(x_0, w_1, \dots w_{k+1}) = F(F_k(x_0, w_1, \dots w_k), w_{k+1})$$

where x_0 and w_i are arbitrary real numbers. By induction we may show that for any initial condition $X_0 = x_0$ and any $k \in \mathbb{Z}_+$,

$$X_k = F_k(x_0, W_1, \dots, W_k),$$

which immediately implies that the k-step transition function may be expressed as

$$\begin{aligned} P^k(x, A) &= \mathsf{P}(F_k(x, W_1, \dots, W_k) \in A) \\ &= \int \cdots \int \mathbb{1}\{F_k(x, w_1, \dots, w_k) \in A\}\, \Gamma(dw_1) \dots \Gamma(dw_k) \qquad (3.45) \end{aligned}$$

3.6 Commentary

The development of foundations in this chapter is standard. The existence of the excellent accounts in Chung [49] and Revuz [223] renders it far less necessary for us to fill in specific details.

The one real assumption in the general case is that the σ-field $\mathcal{B}(\mathsf{X})$ is countably generated. For many purposes, even this condition can be relaxed, using the device of "admissible σ-fields" discussed in Orey [208], Chapter 1. We shall not require, for the models we develop, the greater generality of non-countably generated σ-fields, and leave this expansion of the concepts to the reader if necessary.

The Chapman-Kolmogorov equations, simple though they are, hold the key to much of the analysis of Markov chains. The general formulation of these dates to Kolmogorov [139]: David Kendall comments [132] that the physicist Chapman was not aware of his role in this terminology, which appears to be due to work on the thermal diffusion of grains in a non-uniform fluid.

The Chapman-Kolmogorov equations indicate that the set P^n is a semigroup of operators just as the corresponding matrices are, and in the general case this observation enables an approach to the theory of Markov chains through the mathematical structures of semigroups of operators. This has proved a very fruitful method, especially for continuous time models. However, we do not pursue that route directly in this book, nor do we pursue the possibilities of the matrix structure in the countable case.

This is largely because, as general non-negative operators, the P^n often do not act on useful spaces for our purposes. The one real case where the P^n operate successfully on a normed space occurs in Chapter 16, and even there the space only emerges after a probabilistic argument is completed, rather than providing a starting point for analysis.

Foguel [79, 81] has a thorough exposition of the operator-theoretic approach to chains in discrete time, based on their operation on L^1 spaces. Vere-Jones [283, 284] has a number of results based on the action of a matrix P as a non-negative operator on sequence spaces suitably structured, but even in this countable case results are limited. Nummelin [202] couches many of his results in a general non-negative operator context, as does Tweedie [272, 273], but the methods are probabilistic rather than using traditional operator theory.

The topological spaces we introduce here will not be considered in more detail until Chapter 6. Very many of the properties we derive will actually need less structure than we have imposed in our definition of "topological" spaces: often (see for example Tuominen and Tweedie [269]) all that is required is a countably generated topology with the T_1 separability property. The assumptions we make seem unrestrictive in practice, however, and avoid occasional technicalities of proof.

Hitting times and their properties are of prime importance in all that follows. On a countable space Chung [49] has a detailed account of taboo probabilities, and much of our usage follows his lead and that of Nummelin [202], although our notation differs in minor ways from the latter. In particular our τ_A is, regrettably, Nummelin's S_A and our σ_A is Nummelin's T_A; our usage of τ_A agrees, however, with that of Chung [49] and Asmussen [10], and we hope is the more standard.

The availability of the Strong Markov Property is vital for much of what follows. Kac is reported as saying [35] that he was fortunate, for in his day all processes had the Strong Markov Property: we are equally fortunate that, with a countable time set, all chains still have the Strong Markov Property.

The various transition matrices that we construct are well-known. The reader who is not familiar with such concepts should read, say, Çinlar [40], Karlin and Taylor [122] or Asmussen [10] for these and many other not dissimilar constructions in the queueing and storage area. For further information on linear stochastic systems the reader is referred to Caines [39]. The control and systems areas have concentrated more intensively on controlled Markov chains which have an auxiliary input which is chosen to control the state process $\boldsymbol{\Phi}$. Once a control is applied in this way, the "closed loop system" is frequently described by a Markov chain as defined in this chapter. Kumar and Varaiya [143] is a good introduction, and the article by Arapostathis et al [7] gives an excellent and up to date survey of the controlled Markov chain literature.