Texts in Statistical Science

Markov Chain Monte Carlo

Stochastic Simulation for Bayesian Inference

Second Edition

Dani Gamerman
Instituto de Matemática
Universidade Federal do Rio de Janeiro, Brazil

Hedibert Freitas Lopes
Graduate School of Business
University of Chicago, U.S.A.

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Contents

Preface to the second edition

Almost a decade has elapsed since the release of the first edition. A large amount of recent work was produced on the MCMC subject but made no substantial theoretical contribution. As anticipated in the first edition, most of the ground work for the theory had been established by then. Subsequent literature has basically enabled further understanding and extensions of the previous work. In any case, the book has been updated to include the recent literature and as a result the number of references has almost doubled. We believe to have included at least a reference to most new developments in MCMC.

What has really changed in this decade is the depth of understanding and amount of applications of MCMC to the solution of inference problems. The revision we performed concentrated on this point. The reader will hopefully face a much more readable book in terms of practical aspects. The numbers of exercises, examples, numerical tables and figures have also been considerably increased. We tried to exemplify and illustrate archetypical situations to many applied areas to enable a better apprehension of the pros and cons of the variety of algorithms available in the MCMC arena.

In line with the modern resources available nowadays, the URL site `www.ufrj.br/MCMC` has been created. It contains the codes (all written in `R` language) used in many of the previously existing and new examples and exercises of the book. Readers will have free access to them and will be able to reproduce the tables and figures of the book. More importantly, the mildly self-explanatory nature of the codes will enable modification of the inputs to the codes and variation in many directions will be available for further exploration. This internet tool is planned to be constantly being updated and can also be used to compensate for any new development not included in this edition of the book.

The major changes from the previous edition are as follows. New sections on spatial models and model adequacy have been introduced in Chapter 2. Spatial models is an area that has experienced a huge development in statistics during the last decade and the writers of the book have made a few contributions there as well. A section on model adequacy should have always been there. All that was done was to minimally remedy this flaw of the first edition. Chapter 7 is the chapter that has undergone the largest change. It has moved away from its speculative flavor to a much

CHAPTER 4

Markov chains

4.1 Introduction

The presentation of Markov chains in this chapter is far from comprehensive. It tries to describe the main results by combining intuition with probabilistic reasoning. Its presence in this book is an attempt to provide in simple terms the theory governing the iterative simulation techniques used in later chapters. There are many excellent books on or including a detailed and formal treatment of the subject. Among them, the books by Feller (1968), Meyn and Tweedie (1993), Nummelin (1984) and Ross (1996) can be cited. A more statistically oriented approach is given in Guttorp (1995) and a more thorough treatment is given in Revuz (1975).

Markov dependence is a concept attributed to the Russian mathematician Andrei Andreivich Markov that at the start of the 20th century investigated the alternance of vowels and consonants in the poem *Onegin* by Poeshkin. He developed a probabilistic model where successive results depended on all their predecessors only through the immediate predecessor. The model allowed him to obtain good estimates of the relative frequency of vowels in the poem. Almost at the same time the French mathematician Henri Poincaré studied sequences of random variables that were in fact Markov chains.

A Markov chain is a special type of stochastic process, which deals with characterization of sequences of random variables. Special interest is paid to the dynamic and the limiting behaviors of the sequence. A stochastic process can be defined as a collection of random quantities $\{\theta^{(t)} : t \in T\}$ for some set T.

The set $\{\theta^{(t)} : t \in T\}$, is said to be a stochastic process with state space S and index (or parameter) set T. Throughout this book the index set T is taken as countable, defining a discrete time stochastic process. Without loss of generality, it will be assumed to be the set of natural numbers N and will mostly represent the iterations of a simulation scheme. The state space will in general be a subset of R^d representing the support of a parameter vector. For presentation purposes, it will be assumed to be discrete for the first sections of this chapter. Most important results for general Markov chains can be obtained with a discrete state space. Limiting results concerning ergodic and central limit theorems are especially relevant in this book. Later sections consider general state spaces. After the ground work has been done, connections with iterative simulation schemes are outlined and illustrated for one such specific scheme.

4.2 Definition and transition probabilities

In simple terms, a Markov chain is a stochastic process where given the present state, past and future states are independent. This property can be more formally stated through

$$\begin{aligned} & Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \ldots, \theta^{(0)} \in A_0) \\ & = Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x) \end{aligned} \tag{4.1}$$

for all sets $A_0, \ldots, A_{n-1}, A \subset S$ and $x \in S$. The Markovian property (4.1) can also be established in the equivalent forms:

1. $E[f(\theta^{(n)})|\theta^{(m)}, \theta^{(m-1)}, \ldots, \theta^{(0)}] = E[f(\theta^{(n)})|\theta^{(m)}]$ for all bounded functions f and $n > m \geq 0$;
2. $Pr(\theta^{(n+1)} = y|\theta^{(n)} = x, \theta^{(n-1)} = x_{n-1}, \ldots, \theta^{(0)} = x_0) = Pr(\theta^{(n+1)} = y|\theta^{(n)} = x)$ for all $x_0, \ldots, x_{n-1}, x, y \in S$.

The above form is clearly appropriate only for discrete state spaces. In fact, it is more appropriate than (4.1) in this case and is used as defining Markov chains for the initial sections of this chapter. In general, the probabilities in (4.1) depend on x, A and n. When they do not depend on n, the chain is said to be homogeneous. In this case, a transition function or kernel $P(x, A)$ can be defined as:

1. for all $x \in S$, $P(x, \cdot)$ is a probability distribution over S;
2. for all $A \subset S$, the function $x \mapsto P(x, A)$ can be evaluated.

It is also useful when dealing with discrete state space to identify $P(x, \{y\})$ $= P(x, y)$. This function is called a transition probability and satisfies:

- $P(x, y) \geq 0$, $\forall x, y \in S$;
- $\sum_{y \in S} P(x, y) = 1$, $\forall\, x \in S$;

as any probability distribution $P(x, \cdot)$ should.

Example 4.1 *Random walk*
Consider a particle moving independently left and right on the line with successive displacements from its current position governed by a probability function f over the integers and $\theta^{(n)}$ representing its position at instant n, $n \in N$. Initially, $\theta^{(0)}$ is distributed according to some distribution $\pi^{(0)}$. The positions can be related as

$$\theta^{(n)} = \theta^{(n-1)} + w_n = w_1 + w_2 + \ldots + w_n$$

where the w_i are independent random variables with probability function f. So, $\{\theta^{(n)} : n \in N\}$ is a Markov chain in Z.

The position of the chain at instant $t = n$ is described probabilistically by the distribution of $w_1 + \ldots + w_n$.

If $f(1) = p$, $f(-1) = q$ and $f(0) = r$ with $p+q+r = 1$ then the transition probabilities are given by

$$P(x, y) = \begin{cases} p & , \textit{if } y = x+1 \\ q & , \textit{if } y = x-1 \\ r & , \textit{if } y = x \\ 0 & , \textit{if } y \neq x-1, x, x+1 \end{cases}.$$

Example 4.2 *Branching processes*
Consider particles that can generate new particles of the same type. Each of the x particles at generation n gives birth independently to identically distributed numbers of descendants ξ_i, $i = 1, \ldots, x$ with discrete distribution F and dies. If $\theta^{(n)}$ represents the number of particles at generation n then it is a Markov chain with state space $S = \{0, 1, 2, \ldots\}$ and transition probabilities

$$P(x, y) = Pr\left(\sum_{i=1}^{x} \xi_i = y\right).$$

Note that $P(0, 0) = 1$ and once state 0 is reached, the chain does not leave it. Such states are called absorbing states.

Example 4.3 *Ehrenfest model*
Consider a total of r balls distributed in two urns with x balls in the first urn and $r - x$ in the second urn. Take one of the r balls at random and put it in the other urn. Repeat the random selection process independently and indefinitely. This procedure was used by Ehrenfest to model the exchange of molecules between two containers. If $X^{(n)}$ represents the number of balls in the first urn after n exchanges then $\{X^{(n)} : n \in N\}$ is a Markov chain with state space $S = \{0, 1, 2, \ldots, r\}$ and transition probabilities

$$P(x, y) = \begin{cases} x/d & , \textit{if } y = x+1 \\ 1 - x/d & , \textit{if } y = x-1 \\ 0 & , \textit{if } |y - x| \neq 1 \end{cases}.$$

Example 4.4 *Birth and death processes*
Consider a Markov chain that from the state x can only move in the next step to one of the neighboring states $x - 1$, representing a death, x or $x+1$, representing a birth. The transition probabilities are given by

$$P(x, y) = \begin{cases} p_x & , \textit{if } y = x+1 \\ q_x & , \textit{if } y = x-1 \\ r_x & , \textit{if } y = x \\ 0 & , \textit{if } |y - x| > 1 \end{cases}.$$

where p_x, q_x and r_x are non-negative with $p_x + q_x + r_x = 1$ and $q_0 = 0$. The Ehrenfest model is a special case of birth and death processes.

In the case of discrete state spaces $S = \{x_1, x_2, \ldots\}$, a transition matrix P with (i,j)th element given by $P(x_i, x_j)$ can be defined. If S is finite with r elements, the transition matrix P is given by

$$P = \begin{pmatrix} P(x_1, x_1) & \ldots & P(x_1, x_r) \\ \vdots & & \vdots \\ P(x_r, x_1) & \ldots & P(x_r, x_r) \end{pmatrix} .$$

Transition matrices have all lines summing to one. Such matrices are called stochastic and have a few interesting properties. For instance, at least one eigenvalue of a stochastic matrix equals one and the product of stochastic matrices always produces a stochastic matrix. Of course, countable state spaces will lead to an infinite number of eigenvalues.

Transition probabilities from state x to state y over m steps, denoted by $P^m(x,y)$, is given by the probability of a chain moving from state x to state y in exactly m steps. It can be obtained for $m \geq 2$ as

$$\begin{aligned} P^m(x,y) &= Pr(\theta^{(m)} = y | \theta^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(\theta^{(m)} = y, \theta^{(m-1)} = x_{m-1}, \ldots, \theta^{(1)} = x_1 | \theta^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(\theta^{(m)} = y | \theta^{(m-1)} = x_{m-1}) \ldots Pr(\theta^{(1)} = x_1 | \theta^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} P(x, x_1) P(x_1, x_2) \cdots P(x_{m-1}, y) \end{aligned}$$

where the second equality is due to the Markovian property of the process. The last equality means that the matrix containing elements $P^m(x,y)$ is also a stochastic matrix and is given by P^m obtained by the matrix product of the transition matrix P m times. Also, for completeness, $P^1(x,y) = P(x,y)$ and $P^0(x,y) = I(x = y)$. The above derivation can be used to established that

$$\begin{aligned} P^{n+m}(x,y) &= \sum_z Pr(\theta^{(n+m)} = y | \theta^{(n)} = z, \theta^{(0)} = x) Pr(\theta^{(n)} = z | \theta^{(0)} = x) \\ &= \sum_z P^n(x,z) P^m(z,y). \end{aligned} \tag{4.2}$$

Equations (4.2) are usually called Chapman-Kolmogorov equations. All summations are with respect to the elements of the state space S and results are valid for any stage of the chain due to the assumed homogeneity. Higher transition matrices can be formed with these higher transition probabilities and it can be shown that they satisfy the relation $P^{n+m} = P^n P^m$ and, in particular, $P^{n+1} = P^n P$.

The marginal distribution of the nth stage can be defined by the row vector $\pi^{(n)}$ with components $\pi^{(n)}(x_i)$, for all $x_i \in S$. For finite state spaces, this is a r-dimensional vector

$$\pi^{(n)} = (\pi^{(n)}(x_1), \cdots, \pi^{(n)}(x_r)).$$

When $n = 0$, this is the initial distribution of the chain. Then,

$$\begin{aligned} \pi^{(n)}(y) &= Pr(\theta^{(n)} = y) \\ &= \sum_{x \in S} Pr(\theta^{(n)} = y | \theta^{(0)} = x) Pr(\theta^{(0)} = x) \\ &= \sum_{x \in S} P^n(x,y) \pi^{(0)}(x) . \end{aligned}$$

The above equation can be written in matrix notation as $\pi^{(n)} = \pi^{(0)} P^n$. Also, since the same is valid for $n-1$, $\pi^{(n)} = \pi^{(0)} P^{n-1} P = \pi^{(n-1)} P$.

The probability of any event $A \subset S$ for a Markov chain starting at x is denoted by $Pr_x(A)$. The hitting time of A is defined as $T_A = \min\{n \geq 1 : \theta^{(n)} \in A\}$ if $\theta^{(n)} \in A$ for some $n > 0$. Otherwise, $T_A = \infty$. If $A = \{a\}$, the notation $T_{\{a\}} = T_a$ is used.

Example 4.5 *Consider $\{\theta^{(n)} : n \geq 0\}$, a Markov chain in $S = \{0,1\}$ with initial distribution $\pi^{(0)}$ given by $\pi^{(0)} = (\pi^{(0)}(0), \pi^{(0)}(1))$ and transition matrix P given by $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$.*

Using the relation $Pr(\theta^{(n)} = 0) = \sum_{j \in S} Pr(\theta^{(n)} = 0, \theta^{(n-1)} = j)$ for $n \geq 1$ and the Markovian property of the chain,

$$\begin{aligned} Pr(\theta^{(n)} = 0) &= (1-p) Pr(\theta^{(n-1)} = 0) + q Pr(\theta^{(n-1)} = 1) \\ &= (1-p-q)^n \pi^{(0)}(0) + q \sum_{k=0}^{n-1} (1-p-q)^k \end{aligned}$$

and $Pr(\theta^{(n)} = 1) = 1 - Pr(\theta^{(n)} = 0)$. If $p = q = 0$, the chain never moves with probability one and $Pr(\theta^{(n)} = 0) = \pi_0(0)$ and $Pr(\theta^{(n)} = 1) = \pi_0(1)$. If $p + q > 0$,

$$Pr(\theta^{(n)} = 0) = \frac{q}{p+q} + (1-p-q)^n \left(\pi^{(0)}(0) - \frac{q}{p+q} \right)$$

and $Pr(\theta^{(n)} = 1) = 1 - Pr(\theta^{(n)} = 0)$. Note that if $\pi^{(0)} = (q,p)/(p+q)$, then $Pr(\theta^{(n)} = 0) = q/(p+q)$, for all n, and the distributions at all steps are the same as the initial distribution.

If, in addition, $p + q < 2$ then

$$\lim_{n \to \infty} Pr(\theta^{(n)} = 0) = \frac{q}{p+q} \text{ and } \lim_{n \to \infty} Pr(\theta^{(n)} = 1) = \frac{p}{p+q}$$

and the initial distribution is obtained now as a limiting distribution of the chain. The value of the higher transition matrix P^n can be found via $Pr(\theta^{(n)} = 0) = \pi^{(n)}(0) = \sum_x P^n(x,0) \pi^{(0)}(x)$, for all x. Taking $\pi^{(0)}(0) = 1$

gives

$$P^n(0,0) = \frac{q}{p+q} + (1-p-q)^n \frac{p}{p+q} .$$

Analogously for $P^n(0,1)$, $P^n(1,0)$ and $P^n(1,1)$ gives

$$P^n = \frac{1}{p+q}\begin{pmatrix} q & p \\ q & p \end{pmatrix} + \frac{(1-p-q)^n}{p+q}\begin{pmatrix} p & -p \\ -q & q \end{pmatrix} . \qquad (4.3)$$

Finally $Pr_0(T_x = n)$, $x = 0,1$ can be found by $Pr_0(T_0 = n) = Pr_0(\theta^{(n)} = 0, \theta^{(j)} \neq 0,\ 1 \leq j \leq n-1) = P(1,0)P(1,1)^{n-2}P(1,0) = q(1-q)^{n-2}p$. Analogously, $Pr_0(T_1 = n) = p(1-p)^{n-1}$.

It is important to distinguish between $P^n(0,0) = Pr_0(\theta^{(n)} = 0)$, the probability of the chain, starting from state 0, hitting state 0 in n steps, and $Pr_0(T_0 = n) = Pr_0(\theta^{(n)} = 0, \theta^{(j)} \neq 0, j = 1, 2, \ldots, n-1)$, the probability of the chain, starting from state 0, hitting state 0 in n steps for the first time.

Markov chains can be extended in many directions. One extension used in other chapters of this book is provided by Markov random fields (MRF). They arise when the elements of the index set T are not ordered. More formally, a collection $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ is a MRF if the full conditional distributions of $\theta^{(i)}$ depend only on $\theta^{(j)}$ for $j \in N_i$, the set of neighbors of i, $i = 1, 2, \ldots$ In the special case $T = N$, $N_i = \{i-1, i+1\}$, for $i = 1, 2, \ldots$, and a Markov chain is obtained (see Exercise 4.6).

4.3 Decomposition of the state space

A few quantities of interest are important in the classification of states of a Markov chain with state space S and transition matrix P:
(i) The probability of the chain starting from state x hitting state y at any posterior step is $\rho_{xy} = Pr_x(T_y < \infty)$;
(ii) The number of visits of a chain to a state y is

$$N(y) = \#\{n > 0 : \theta^{(n)} = y\} = \sum_{n=1}^{\infty} I(\theta^{(n)} = y) .$$

It can be shown that $E(T_y \mid \theta^{(0)} = x) = \sum_{n=0}^{\infty} Pr_x(T_y > n)$ and $E(N(y) \mid \theta^{(0)} = x) = \sum_{n=1}^{\infty} P^n(x,y)$.

A state $y \in S$ is said to be recurrent if the Markov chain, starting in y, returns to y with probability 1 ($\rho_{yy} = 1$) and is said to be transient if it has positive probability of not returning to y ($\rho_{yy} < 1$). An absorbing state $y \in S$ is recurrent because $Pr_y(T_y = 1) = Pr_y(\theta^{(1)} = y) = P(y,y) = 1$ and therefore $\rho_{yy} = 1$. If a Markov chain starts at a recurrent state y, the hitting (or return, in this case) time of y, T_y, is a finite random quantity whose mean μ_y can be evaluated. If this mean is finite, the state y is said

to be positive recurrent and otherwise the state is said to be null recurrent. Positive recurrence is a very important property for establishing limiting results, as will be seen in the next section.

An important result describing analytically the difference between a recurrent and a transient state is that

- if $y \in S$ is a transient state then, for all $x \in S$,

$$Pr_x(N(y) < \infty) = 1 \text{ and } E[N(y) \mid \theta^{(0)} = x] = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty.$$

- if $y \in S$ is a recurrent state then

$$Pr_y(N(y) = \infty) = 1 \text{ and } E[N(y) \mid \theta^{(0)} = y] = \infty.$$

So, recurrent states are infinitely often (i.o.) visited with probability one. The expected number of visits is finite if the state is transient.

It is interesting to investigate possible decompositions of S in subsets of recurrent and transient states. From this decomposition, probabilities of the chain hitting a given set of states can be evaluated. For states x and y in S, $x \neq y$, x is said to hit y, denoted $x \to y$, if $\rho_{xy} > 0$. A set $C \subseteq S$ is said to be closed if

$$\rho_{xy} = 0 \text{ for } x \in C \text{ and } y \notin C.$$

In obvious nomenclature, it is said to be irreducible if $x \to y$ for every pair $x, y \in C$. A chain is said to be irreducible if S is irreducible. It is not difficult to show that the condition $\rho_{xy} > 0$ is equivalent to $P^n(x,y) > 0$ for some $n > 0$. This can be used to show that if $x \in S$ is recurrent and $x \to y$ then y is also recurrent. In this case, $y \to x$ and one can write $x \leftrightarrow y$ when $x \to y$ and $y \to x$. In other words, recurrence defines an equivalence class with respect to the $\leftrightarrow$ operation. Also, $\rho_{xy} = \rho_{yx} = 1$. In fact, a stronger result is valid: null recurrence and positive recurrence also define equivalence classes (Guttorp, 1995). If $C \subseteq S$ is a closed, finite, irreducible set of states then all states of C are recurrent.

Example 4.6 *Consider the transition matrices over $S = \{0, 1, \ldots, r\}$ below. For each one of them, S can be decomposed into S_R and S_T, the sets of recurrent and transient states respectively. The symbols + and − associated with the pair (x,y) denote $x \to y$ and $x \not\to y$, respectively.*

$$a) \qquad P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/4 & 1/4 \\ 0 & 1/3 & 2/3 \end{pmatrix} \qquad \Longrightarrow \qquad \begin{pmatrix} + & + & + \\ + & + & + \\ + & + & + \end{pmatrix} .$$

Therefore, S is closed and the chain is irreducible. To show that, for instance, $0 \to 2$, it suffices to verify that $P^2(0,2) > 0$. So, $S = S_R$.

$$b)\, P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/5 & 2/5 & 1/5 & 0 & 1/5 \\ 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/4 & 0 & 3/4 \end{pmatrix} \Rightarrow \begin{pmatrix} + & - & - & - & - & - \\ + & + & + & + & + & + \\ + & + & + & + & + & + \\ - & - & - & + & + & + \\ - & - & - & + & + & + \\ - & - & - & + & + & + \end{pmatrix}.$$

In this case, $S_R = \{0\} \cup \{3,4,5\}$ and $S_T = \{1,2\}$.

If the set of recurrent states S_R is not empty then it can be written as a (finite or countable) union of disjoint, closed and irreducible sets. Whenever the chain hits a closed, irreducible set C of recurrent states, it stays in C forever and with probability one it visits all its elements i.o. It is interesting to compute ρ_{xy} for x transient and y recurrent. As y belongs to an irreducible set C, $\rho_{xy} = \rho_C(x) = Pr_x(T_C < \infty)$ is called the absortion probability. If the chain starts at $x \in S_T$, it can hit C in the first step or remain in S_T in the first step and hit C at a later step. So,

$$\rho_C(x) = \sum_{y \in C} P(x,y) + \sum_{y \in S_T} P(x,y)\rho_C(y), \quad x \in S_T. \tag{4.4}$$

The uniqueness of solutions of this system is guaranteed if $Pr_x(T_{S_R} < \infty) = 1$, for $x \in S_T$. In the case of a finite S_T it is automatically valid as transient states are only finitely visited.

Example 4.6 *(continued) Let $S_R = C_1 \cup C_2$ where $C_1 = \{0\}$ and $C_2 = \{3,4,5\}$. The absortion probabilities $\rho_{10} = \rho_{C_1}(1)$ and $\rho_{20} = \rho_{C_1}(2)$ are obtained from Equation (4.4) as*

$$\rho_{10} = \frac{1}{4} + \frac{1}{2}\rho_{10} + \frac{1}{4}\rho_{20} \text{ and } \rho_{20} = 0 + \frac{1}{5}\rho_{10} + \frac{2}{5}\rho_{20}\,.$$

Solving for ρ_{10} and ρ_{20} gives $\rho_{10} = 3/5$ and $\rho_{20} = 1/5$.

Proceeding analogously for C_2 gives $\rho_{C_2}(1) = 2/5$ and $\rho_{C_2}(2) = 4/5$. As S_T is finite, the absorption probabilities $\rho_{C_2}(x)$ can be evaluated for $x \in S_T$ through $\sum_i \rho_{C_i}(x) = 1$.

Example 4.4 *(continued) Irreducible chains are obtained when $p_x > 0$ for $x \geq 0$ and $q_x > 0$ for $x > 0$. It is possible to determine if a state y is recurrent or transient even for an infinite state space by studying the convergence of the series $\sum_{y=0}^{\infty} \gamma_y$ where*

$$\gamma_y = \begin{cases} 1 & \text{if } y = 0 \\ \frac{q_1 \cdots q_y}{p_1 \cdots p_y} & \text{if } y > 0 \end{cases}.$$

If the sum diverges, the chain is recurrent. Otherwise, the chain is transient.

If S is finite and 0 is an absorbing state, the absortion probability is

$$\rho_{\{0\}}(x) = \rho_{x0} = \frac{\sum_{y=x}^{d-1} \gamma_y}{\sum_{y=0}^{d-1} \gamma_y} \quad , \quad x = 1, \ldots, d-1.$$

Details of these calculations are given, for example, in Hoel, Port and Stone (1972).

4.4 Stationary distributions

A fundamental problem for Markov chains in the context of simulation is the study of the asymptotic behavior of the chain as the number of steps or iterations $n \to \infty$. A key concept is that of a stationary distribution π. A distribution π is said to be a stationary distribution of a chain with transition probabilities $P(x,y)$ if

$$\sum_{x \in S} \pi(x)P(x,y) = \pi(y), \quad \forall y \in S. \tag{4.5}$$

Equation (4.5) can be written in matrix notation as $\pi = \pi P$. The reason of the name is clear from the above equation. If the marginal distribution at any given step n is π then the distribution at the next step is $\pi P = \pi$. Once the chain reaches a stage where π is the distribution of the chain, the chain retains this distribution for all subsequent stages. This distribution is also known as the invariant or equilibrium distribution for similar interpretations.

It will be shown below that if the stationary distribution π exists and $\lim_{n\to\infty} P^n(x,y) = \pi(y)$ then, independently of the initial distribution of the chain, $\pi^{(n)}$ will approach π, as $n \to \infty$. In this sense, the distribution is also referred to as the limiting distribution.

Example 4.5 *(continued) The stationary distribution π is the solution of the system $\pi P = \pi$ that gives equations*

$$\pi(0)P(0,y) + \pi(1)P(1,y) = \pi(y)\,, \quad y = 0,1.$$

The solution is $\pi = (q,p)/(p+q)$, a distribution that was shown to be invariant for the stages of the chain.

Also, provided $p + q < 2$, $\lim_{n\to\infty} P^n = \frac{1}{p+q}\begin{pmatrix} q & p \\ q & p \end{pmatrix}$ from (4.3) and the distribution of $\theta^{(n)}$ converges to π at an exponential rate.

The case $p + q = 2$ still produces a stationary distribution π but this does not provide a unique limiting distribution as $Pr(\theta^{(n)} = 0) = (1/2) + (-1)^n[\pi^{(0)} - (1/2)]$, for all n. This case is somehow different because the states are always alternating through the stages. It has a periodic nature that will be addressed in the next section.

Example 4.7 *Gibbs sampler (Geman and Geman, 1984)*

This example provides a very simple special case of the Gibbs sampler and was considered by Casella and George (1992). The complete form of the Gibbs sampler will be left for the next chapter. In this special case, the state space is $S = \{0, 1\}^2$ and define a probability distribution π over S as

	θ_2	
θ_1	*0*	*1*
0	π_{00}	π_{01}
1	π_{10}	π_{11}

The probability vector π contains the above probabilities in any fixed order, say $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$.

The chain now consists of a bidimensional vector $\theta^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})$. Although this introduces some novelties in the presentation they can easily be removed by considering a scalar chain $\psi^{(n)}$ that assumes values that are in correspondence with the $\theta^{(n)}$ chain, e.g. $\psi^{(n)} = 10\theta_1^{(n)} + \theta_2^{(n)}$. This is always possible for discrete state spaces and from now on no distinction will be made between scalar and vector chains.

Consider the following transition probabilities:

- *For the first component θ_1, the transition probabilities are given by the conditional distribution π_1 of $\theta_1 | \theta_2 = j$,*

$$\pi_1(0|j) = \frac{\pi_{0j}}{\pi_{+j}} \text{ and } \pi_1(1|j) = \frac{\pi_{1j}}{\pi_{+j}}$$

  where $\pi_{+j} = \pi_{0j} + \pi_{1j}$, $j = 0, 1$.

- *For the second component θ_2, the transition probabilities are given by the conditional distribution π_2 of $\theta_2 | \theta_1 = i$,*

$$\pi_2(0|i) = \frac{\pi_{i0}}{\pi_{i+}} \text{ and } \pi_2(1|i) = \frac{\pi_{i1}}{\pi_{i+}}$$

  where $\pi_{i+} = \pi_{i0} + \pi_{i1}$, $i = 0, 1$.

The overall transition probability of the chain is

$$\begin{aligned} P((i,j),(k,l)) &= Pr(\theta^{(n)} = (k,l) | \theta^{(n-1)} = (i,j)) \\ &= Pr(\theta_2^{(n)} = l | \theta_1^{(n)} = k)\, Pr(\theta_1^{(n)} = k | \theta_2^{(n-1)} = j) \\ &= \frac{\pi_{kl}}{\pi_{k+}} \frac{\pi_{kj}}{\pi_{+j}} \end{aligned}$$

for $(i,j), (k,l) \in S$. Thus, a 4×4 transition matrix P can be formed.

It is evident from the above transitions that $(\theta^{(n)})_{n \geq 0}$ forms a Markov chain as transition probabilities only depend on the present value (k, l) to predict the future value (i, j). It is straightforward to ascertain that π is the stationary distribution of the chain. If all elements of π are positive, it is also a limiting distribution. The interesting message is that chains formed by the superposition of conditional distributions have a stationary distribution given by the joint distribution.

The results can be extended in the same way for cases when θ_1 can take m_1 values and θ_2 can take m_2 values. They can similarly be extended to cases where the θ consist of d components and each of them can take m_i values, $i = 1, \ldots, d$. The Gibbs sampler is a simulation scheme that will be applied in the next chapter for general (continuous, discrete and mixed) d-dimensional state spaces.

Example 4.4 *(continued) For irreducible birth and death chains, the system of equations $\sum_{x \in S} \pi(x) P(x, y) = \pi(y)$, $y \in S$, is $\pi(x) = p_{x-1}\pi(x-1) + r_x \pi(x) + q_{x+1}\pi(x+1)$ for $x > 0$ and $\pi(0) = r_0 \pi(0) + q_1 \pi(1)$ for $x = 0$. Using the fact that $r_x = 1 - p_x - q_x$, these equations reduce to*

$$\pi(x+1) = \frac{p_x}{q_{x+1}} \pi(x), \quad x \geq 0.$$

Defining

$$\pi_x = \begin{cases} 1 & , \text{if } x = 0 \\ \frac{p_0 p_1 \cdots p_{x-1}}{q_1 q_2 \cdots q_x} & , \text{if } x \geq 1 \end{cases}$$

gives that, if $\sum_x \pi_x$ converges, the birth and death chain has stationary distribution

$$\pi(x) = \frac{\pi_x}{\sum_{y=0}^{\infty} \pi_y}, \qquad x \geq 0.$$

In the case of an Ehrenfest model where $P(x, x) = r_x = 0$ for all x, $P^n(x, x) = 0$ for odd n. Hence, the chain can only return to the same state after an even number of transitions. In this case, there is no limiting distribution, as will be shown below.

The existence and uniqueness of stationary distributions can be studied through weaker results. Let $N_n(y)$ be the number of visits to state y in n steps and define $G_n(x, y) = E_x[N_n(y)]$, the average number of visits of the chain to state y and $m_y = E_y(T_y)$, the average return time to state y. Then, $G_n(x, y) = \sum_{k=1}^{n} P^k(x, y)$ and $\lim_{n \to \infty} G_n(x, y)/n$ provides the limiting occupation of state y in a chain observed for an infinitely long number of steps. It can be shown that

- If $y \in S$ is transient then $\lim_{n\to\infty} \frac{N_n(y)}{n} = 0$, with probability 1, and $\lim_{n\to\infty} \frac{G_n(x,y)}{n} = 0$ for all $x \in S$.
- If $y \in S$ is recurrent then $\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{I(T_y < \infty)}{m_y}$, with probability 1, and $\lim_{n\to\infty} \frac{G_n(x,y)}{n} = \frac{\rho_{xy}}{m_y}$, for all $x \in S$.

It follows that if π is a stationary distribution then $\pi(x) = 0$, if x is transient or null recurrent ($m_y = \infty$), and $\pi(x) = \frac{1}{m_x}$, if x is positive recurrent. Intuitively, the stationary probability of any state is given by the frequency of visits to the state. As the sets S_{Rp} and S_{Rn} of positive recurrent and null recurrent states are closed if S is finite, then $S_{Rn} = \phi$, the empty set. Therefore, an irreducible Markov chain is positive recurrent if and only if it possesses a stationary distribution π such that

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} P^k(x,y)}{n} = \lim_{n\to\infty} \frac{G_n(x,y)}{n} = \pi(y).$$

If the chain is null recurrent then Equation (4.5) is still valid but the $\pi(y)$s do not possess a finite sum and therefore do not constitute a proper distribution.

The analysis of the limiting behavior of P^n is more delicate for technical reasons and is considered in the next section. It is clear that if $y \in S_T$ or $y \in S_{Rn}$ then $\lim_{n\to\infty} P^n(x,y) = 0$, $\forall x \in S$. If $y \in S_{Rp}$, it has only been obtained that $\lim_{n\to\infty} P^n(x,y)$ must equal ρ_{xy}/m_y.

4.5 Limiting theorems

There are situations where stationary distributions are available but limiting distributions are not (Example 4.5). In order to establish limiting results, one characterization of states still absent must be introduced. This is the notion of periodicity.

The period of a state x, denoted by d_x, is the largest common divisor of the set

$$\{n \geq 1 : P^n(x,x) > 0\}.$$

It is obvious that $P(x,x) > 0$ implies that $d_x = 1$ and that if $x \leftrightarrow y$ then $d_x = d_y$. Therefore, the states of an irreducible chain have the same period. A state x is aperiodic if $d_x = 1$ and if in addition the state is positive recurrent, the state is said to be ergodic. A chain is periodic with period d if all its states are periodic with period $d > 1$ and aperiodic if all its states are aperiodic. Finally, a chain is ergodic if all its states are ergodic.

Although aperiodicity is superfluous for the existence of an equilibrium distribution, it is a condition needed to establish convergence of the transition probabilities. Let $(\theta^{(n)})_{n\geq 0}$ be an irreducible, positive recurrent chain with stationary distribution π and $\|\xi_1 - \xi_2\| = \sup_{A\subset S}|\xi_1(A) - \xi_2(A)|$ be the total variation distance between two distributions ξ_1 and ξ_2.

- If the chain is aperiodic then $\lim_{n\to\infty} P^n(x,y) = \pi(y)$ for all $x, y \in S$. In fact, irreducibility and ergodicity of the chain are equivalent to $\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0$, for all $x \in S$ (Nummelin, 1984).
- If the chain is periodic with period d then, for every $x, y \in S$, there is an integer r, $0 \leq r < d$ such that $P^n(x,y) = 0$ unless $n = md + r$ for some $m \in N$ and $\lim_{m\to\infty} P^{md+r}(x,y) = d\pi(y)$.

Example 4.3 *(continued) If $y - x$ is even then $P^{2m+1}(x,y) = 0$, $m \geq 0$ and $\lim_{m\to\infty} P^{2m}(x,y) = 2\pi(y)$, and if $y - x$ is odd then $P^{2m}(x,y) = 0$, $m \geq 0$ and $\lim_{m\to\infty} P^{2m+1}(x,y) = 2\pi(y)$. The transition matrix P is periodic with period $d = 2$.*

Let $S = \{0,1,2,3\}$ ($r = 3$). From Equation (4.5),

$$\pi = \left(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\right)$$

and the limiting distributions are obtained from

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} \end{pmatrix} \quad \textit{for even } n \textit{ and}$$

$$\lim_{n\to\infty} P^n = \begin{pmatrix} 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 \end{pmatrix} \quad \textit{for odd } n.$$

Once ergodicity of the chain is established, important limiting theorems can be stated. The first and most important one is the ergodic theorem. The ergodic average of a real-valued function $t(\theta)$ is the average $\bar{t}_n = (1/n)\sum_{i=1}^{n} t(\theta^{(i)})$.* If the chain is ergodic and $E_\pi[t(\theta)] < \infty$ for the unique limiting distribution π then

$$\bar{t}_n \xrightarrow{a.s.} E_\pi[t(\theta)] \text{ as } n \to \infty. \tag{4.6}$$

This result is a Markov chain equivalent of the law of large numbers (3.9). It states that averages of chain values also provide strongly consistent estimates of parameters of the limiting distribution π despite their dependence. If $t(\theta) = I(\theta = x)$ then the ergodic averages are simply counting the relative frequency of values of xs in realizations of the chain. By the ergodic theorem, this relative frequency converges almost surely to $\pi(x) = 1/m_x$, the average frequency of visits to state x.

Just as there is an equivalent of the law of large numbers for Markov chains, there are also versions of the central limit theorem (3.8) for Markov chains. Many forms are available depending on further conditions on the

* The average could have included the term corresponding to the initial step and all limiting results would still follow. In the sequel, it will be assumed that chains start at step 1.

chain. A chain is said to be geometrically ergodic if there is a constant $0 \leq \lambda < 1$ and a real, integrable function $M(x)$ such that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\lambda^n \tag{4.7}$$

for all $x \in S$. If the function M does not depend on x, the ergodicity is uniform. Uniform ergodicity implies geometric ergodicity which implies ergodicity (Tierney, 1994). The smallest λ satisfying (4.7) is called the rate of convergence. Roberts (1996) provides a brief discussion showing that this rate is bounded by the second largest eigenvalue of P. Of course, a geometric rate is desirable for a fast convergence to the limiting distribution. However, this speed may be offset by a very large value of $M(x)$ which may slow down convergence considerably (Polson, 1996). Also, the rate of convergence can be arbitrarily close to 1 (Example 5.5).

Before stating central limit theorems, define the autocovariance of lag k ($k \geq 0$) of the chain $t^{(n)} = t(\theta^{(n)})$ as $\gamma_k = Cov_\pi(t^{(n)}, t^{(n+k)})$, the variance of $t^{(n)}$ as $\sigma^2 = \gamma_0$, the autocorrelation of lag k as $\rho_k = \gamma_k/\sigma^2$ and $\tau_n^2/n = Var_\pi(\bar{t}_n)$. It can be shown that

$$\tau_n^2 = \sigma^2\left(1 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\rho_k\right) \tag{4.8}$$

and that $\tau_n^2 \to \tau^2$ as $n \to \infty$ where

$$\tau^2 = \sigma^2\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right) \tag{4.9}$$

if the series of autocorrelation is summable. The term between parentheses in Equation (4.9) can be called *inefficiency factor* or *integrated autocorrelation time* (Green and Han, 1992) because it measures how far $t^{(n)}$'s are from being a random sample and how much $Var_\pi(\bar{t}_n)$ increases because of that. The inefficiency factor can be used to derive the effective sample size

$$n_{\text{eff}} = \frac{n}{1 + 2\sum_{k=1}^{\infty}\rho_k} \tag{4.10}$$

which can be thought of as the size of a random sample with the same variance (see Liu and Chen, 1995) since $Var_\pi(\bar{t}_n) = \sigma^2/n_{\text{eff}}$.

It is important to distinguish between $\sigma^2 = Var_\pi[t(\theta)]$, the variance of $t(\theta)$ under the limiting distribution π and τ^2, the limiting sampling variance of $\sqrt{n}\,\bar{t}$. Note that under independent sampling they are both given by σ^2. They are both variability measures but the first one is a characteristic of the limiting distribution π whereas the second is the uncertainty of the averaging procedure.

If a chain is uniformly (geometrically) ergodic and $t^2(\theta)$ ($t^{2+\epsilon}(\theta)$) is integrable with respect to π (for some $\epsilon > 0$) then

$$\sqrt{n}\,\frac{\bar{t}_n - E_\pi[t(\theta)]}{\tau} = \sqrt{n_{\text{eff}}}\,\frac{\bar{t}_n - E_\pi[t(\theta)]}{\sigma} \xrightarrow{d} N(0,1), \tag{4.11}$$

as $n \to \infty$ (Tierney, 1996). Other versions of the central limit theorem may be found in Tierney (1994) and the subsequent discussion by Chan and Geyer (1994). Just as (4.6) provides theoretical support for the use of ergodic averages as estimates, Equation (4.11) provides support for evaluation of approximate confidence intervals. These will require estimation of the unknown quantity τ^2, a subject that will be treated in Section 4.8.

4.6 Reversible chains

Let $(\theta^{(n)})_{n\geq 0}$ be an homogeneous Markov chain with transition probabilities $P(x,y)$ and stationary distribution π. Assume that one wishes to study the sequence of states $\theta^{(n)}, \theta^{(n-1)}, \ldots$ in reversed order. It can be shown that this sequence satisfies

$$Pr(\theta^{(n)} = y \mid \theta^{(n+1)} = x, \theta^{(n+2)} = x_2, \ldots) = Pr(\theta^{(n)} = y \mid \theta^{(n+1)} = x)$$

and therefore defines a Markov chain. The transition probabilities are

$$\begin{aligned} P_n^*(x,y) &= Pr(\theta^{(n)} = y \mid \theta^{(n+1)} = x) \\ &= \frac{Pr(\theta^{(n+1)} = x \mid \theta^{(n)} = y)Pr(\theta^{(n)} = y)}{Pr(\theta^{(n+1)} = x)} \\ &= \frac{\pi^{(n)}(y)P(y,x)}{\pi^{(n+1)}(x)} \end{aligned}$$

and in general the chain is not homogeneous. If $n \to \infty$ or alternatively, $\pi^{(0)} = \pi$, then $P_n^*(x,y) = P^*(x,y) = \pi(y)P(y,x)/\pi(x)$ and the chain becomes homogeneous. If $P^*(x,y) = P(x,y)$ for all x and $y \in S$, the time reversed Markov chain has the same transition probabilities as the original Markov chain. Markov chains with such a property are said to be reversible and the reversibility condition is usually written as

$$\pi(x)P(x,y) = \pi(y)P(y,x), \text{ for all } x, y \in S. \tag{4.12}$$

It can be interpreted as saying that the rate at which the system moves from x to y when in equilibrium, $\pi(x)P(x,y)$, is the same as the rate at which it moves from y to x, $\pi(y)P(y,x)$. For that reason, (4.12) is sometimes referred to as the detailed balance equation (Guttorp, 1995); balance because it equates the rates of moves through states and detailed because it does it for every possible pair of states.

Example 4.3 *(continued) The stationary distribution has already been obtained. There are only 4 possibilities for a pair $(x,y) \in S$:*

- *if $|x-y| > 2$, then $P(x,y) = P(y,x) = 0$ and (4.12) is satisfied;*

- *if $x = y$, then (4.12) is trivially satisfied;*
- *if $x = y - 1$ then $P(x, y) = q_x$ and $P(y, x) = p_{x-1}$. It has also been shown that $\pi(x) = \pi(x-1)p_{x-1}/q_x$ for $x > 0$, which again implies that (4.12) is satisfied;*
- *if $x = y + 1$ then $P(x, y) = p_x$ and $P(y, x) = q_{x-1}$. The same relations of the previous case hold and (4.12) is satisfied.*

In conclusion, irreducible birth and death chains are reversible.

Reversible chains are useful because if there is a distribution π satisfying (4.12) for an irreducible chain, then the chain is positive recurrent, reversible with stationary distribution π. This is easily obtained by summing over y both sides of (4.12) to give (4.5). Construction of Markov chains with a given stationary distribution π reduces to finding transition probabilities $P(x, y)$ satisfying (4.12). This is always possible, as the next example shows.

Example 4.8 *Metropolis algorithm (Metropolis et al., 1953)*

Consider a given distribution p_x, $x \in S$ with $\sum_x p_x = 1$ where the state space S can be a subset of the line or even a d-dimensional subset of R^d. The problem posed and solved by Metropolis et al. (1953) was how to construct a Markov chain with stationary distribution π such that $\pi(x) = p_x$, $x \in S$. Let Q be any irreducible transition matrix on S satisfying the symmetry condition $Q(x, y) = Q(y, x)$, for $x, y \in S$.

Define a Markov chain $(\theta^{(n)})_{n \geq 0}$ as having transition from x to y proposed according to the probabilities $Q(x, y)$. This proposed value for $\theta^{(n+1)}$ is accepted with probability $\min\{1, p_y/p_x\}$ and rejected otherwise, leaving the chain in state x.

The transition probabilities $P(x, y)$ of the above chain $(\theta^{(n)})_{n \geq 0}$ are

$$\begin{aligned} P(x,y) &= Pr(\theta^{(n+1)} = y, TA|\theta^{(n)} = x) \\ &= Pr(\theta^{(n+1)} = y|\theta^{(n)} = x)Pr(TA) \\ &= Q(x,y)\min\{1, p_y/p_x\} \end{aligned}$$

for $y \neq x$ and TA denotes the event [transition is accepted]. If $y = x$, then

$$\begin{aligned} P(x,x) \;=\; & Pr(\theta^{(n+1)} = x, TA|\theta^{(n)} = x) + Pr(\theta^{(n+1)} \neq x, \bar{T}A|\theta^{(n)} = x) \\ = \;& Pr(\theta^{(n+1)} = x|\theta^{(n)} = x)Pr(TA) + \sum_{y \neq x} Pr(\theta^{(n+1)} = y, \bar{T}A|\theta^{(n)} = x) \\ = \;& Q(x,x) + \sum_{y \neq x} Q(x,y)[1 - \min\{1, p_y/p_x\}] \,. \end{aligned}$$

The first step to obtaining the stationary distribution of this chain is to prove that the probabilities p_x satisfy the reversibility condition. For $x = y$, Equation (4.12) is trivially satisfied. For $x \neq y$, suppose first that $p_y > p_x$.

Then

$$p_x P(x,y) = p_x Q(x,y) = Q(y,x)\min\left\{1, \frac{p_x}{p_y}\right\} p_y = p_y P(y,x).$$

Analogous calculations follow for the case $p_y < p_x$. Therefore, the chain is reversible and the probabilities p_x, $x \in S$ provide the stationary distribution of the chain.

If Q is aperiodic, so will be P and the stationary distribution is also the limiting distribution. It is not difficult to find examples of symmetric transition matrices. The random walk chain (Example 4.1) with f symmetric around 0 is an example. The birth and death model with $p_x = q_{x+1}$ is another example.

4.7 Continuous state spaces

This section considers sequences of random quantities that form a Markov chain in R but still retain a discrete parameter space T. Although not explored in as many textbooks as the case of discrete state spaces, it may be found in a few recent books (Gillespie, 1992; Medhi, 1994; Meyn and Tweedie, 1993). There are a few changes required with respect to the discrete case but the main results of the previous sections are still valid. In particular, convergence to the limiting distribution, the ergodic theorem and the central limit theorem need basically technical changes in the conditions of the chain to hold.

4.7.1 Transition kernels

Markov chains are still defined in terms of Equation (4.1). If the conditional probabilities do not depend on the step n, the chain is homogeneous. Then the transition kernel $P(x, A)$ (Section 4.2) is again used to define the chain. The analogy with the discrete case breaks when trying to consider $P(x, \{y\})$, which is always null in the continuous case and not useful in this context. Therefore, transition matrices cannot be constructed and transition kernels must be used instead. However, given that $P(x, \cdot)$ defines a probability distribution, the notation $P(x, y)$ can be used as

$$P(x,y) = Pr(\theta^{(n+1)} \leq y \mid \theta^{(n)} = x) = Pr(\theta^{(1)} \leq y \mid \theta^{(0)} = x), \text{ for } x, y \in S,$$

when P is absolutely continuous with respect to y. Also associated with this conditional distribution, one can obtain the conditional density

$$p(x,y) = \frac{\partial P(x,y)}{\partial y}, \text{ for } x, y \in S.$$

This density can be used to define the transition kernel of the chain instead of $P(x, A)$. The state space S does not need to be the entire line. It can be any interval or collection of intervals for results below to hold.

The conditional transition probability over m steps is given by

$$P^m(x,y) = Pr(\theta^{(m+n)} \le y \mid \theta^{(n)} = x), \text{ for } x, y \in S,$$

the transition kernel over m steps is given by

$$p^m(x,y) = \frac{\partial P^m(x,y)}{\partial y}, \text{ for } x, y \in S$$

and the equivalent equation to (4.2) has the form

$$P^{n+m}(x,y) = \int_{-\infty}^{\infty} P^m(z,y) p^n(x,z) dz, \quad m, n \ge 0.$$

This is the continuous version of the Chapman-Kolmogorov equations. For $m = 1$, it reduces to

$$P^{n+1}(x,y) = \int_{-\infty}^{\infty} P(z,y) p^n(x,z) dz, \quad n \ge 0.$$

The marginal distribution at any step n has density $\pi^{(n)}$ $(n \ge 0)$ that can be obtained from the marginal distribution at the previous step as

$$\pi^{(n)}(y) = \int_{-\infty}^{\infty} p(x,y) \pi^{(n-1)}(x) dx. \tag{4.13}$$

Example 4.1 *(continued) Assume now that the random displacements w_n are continuous quantities with common density $f(w)$ and $\pi^{(0)}$ is some continuous distribution. As in the discrete case,*

$$\theta^{(n+1)} = \theta^{(n)} + w_{n+1} = w_1 + \ldots + w_{n+1}.$$

Note that this is the model used for the system equation in Example 2.9 with $f(\cdot) = f_N(\cdot; 0, W)$, assuming a constant system variance W. The transition probabilities are

$$\begin{aligned} P(x,y) &= Pr(\theta^{(n)} + w_{n+1} \le y \mid \theta^{(n)} = x) \\ &= Pr(w_{n+1} \le y - x) \\ &= \int_{-\infty}^{y-x} f(w) dw \end{aligned}$$

and $p(x,y) = f(y-x)$. The marginal distribution at each step is recursively obtained as

$$\pi^{(n)}(y) = \int_{-\infty}^{\infty} f^n(y-x) \pi^{(0)}(x)\, dx$$

where f^n is the nth convolution of f. In the case of Example 2.9, these calculations simplify due to the normality assumptions to give $\pi^{(n)} = N(a, R+ nW)$ if $\pi^{(0)} = N(a, R)$.

Example 4.9 *A numerical sequence $y_1, y_2, \cdots$ is split by bars whenever*

$y_j > y_{j+1}$ and starts with a bar. A run is a collection of numbers limited by bars. For example, the portion 3,6,9,2,3,1,5,2 of a sequence is split as

$$\mid 3, 6, 9 \mid 2, 3 \mid 1, 5 \mid 2.$$

and (3,6,9), (2,3) and (1,5) are runs.

Consider a sequence $\phi_1, \phi_2, \cdots$ of independent random variables with identical distribution $U[0,1]$ and let $(\theta^{(n)})_{n \ge 1}$ be a Markov chain on $S = (0,1)$ formed by the initial values of the runs. The transition kernel of the chain is obtained from

$$P(x,y) = \sum_{m=1}^{\infty} Pr(\theta^{(n+1)} \le y, \psi_n = m | \theta^{(n)} = x)$$

where ψ_n is the length of the nth run. It can be shown that this leads to

$$p(x,y) = \begin{cases} e^{1-x} & \text{, if } y < x \\ e^{1-x} - e^{y-x} & \text{, if } y > x \end{cases}.$$

Assume now that $\theta^{(n)} = (\theta_1^{(n)}, \ldots, \theta_d^{(n)})'$ is a random vector in R^d. A sequence $(\theta^{(n)})_{n \ge 0}$ in $S \subset R^d$ is a Markov chain with continuous state space if

$$\begin{aligned} & Pr(\theta^{(n+1)} \le y \mid \theta^{(n)} = x\ , \theta^{(n-1)} = x^{(n-1)}, \ldots, \theta^{(0)} = x^{(0)}) \\ = \ & Pr(\theta^{(n+1)} \le y \mid \theta^{(n)} = x), \end{aligned}$$

where $x^{(0)}, \ldots, x^{(n-1)}, x$ and $y \in R^d$ and $z \le w$ for d-dimensional vectors stands for $z_i \le w_i$, $i = 1, \ldots, d$. Homogeneous chains are defined in the same way and transition probabilities are given by

$$P(x,y) = Pr(\theta^{(n+1)} \le y \mid \theta^{(n)} = x) = Pr(\theta^{(1)} \le y \mid \theta^{(0)} = x).$$

As this transition defines a d-dimensional conditional distribution, the transition kernel given by the conditional density associated with this distribution is

$$p(x,y) = \frac{\partial P(x,y)}{\partial y}.$$

4.7.2 Stationarity and limiting results

The stationary or invariant distribution π of a chain with transition kernel $p(x,y)$ must satisfy

$$\pi(y) = \int_{-\infty}^{\infty} \pi(x)\, p(x,y)\, dx \tag{4.14}$$

which is the continuous version of Equation (4.5). The interpretation remains the same, as is clear from Equation (4.13).

To study convergence and limiting results, the classification of states must be revisited. Instead of considering hitting time to a given state x, one must consider hitting time T_A to a given set $A \subset S$ and a distribution ν. A chain is said to be ν-irreducible if for a set A with positive probability under ν, $\rho_{xA} = Pr_x(T_A < \infty) > 0$, for all $x \in S$. This is equivalent to imposing the existence of an integer n such that $P^n(x, A) > 0$. A chain is irreducible if there is at least one distribution ν ensuring that it is ν-irreducible. Usually, irreducibility is simpler to verify through this last condition, with $\nu = \pi$. Also, for most of the chains of interest for simulation, $P(x, A) > 0$.

The other vital properties for establishment of limiting results are aperiodicity and positive recurrence. These are defined as in the discrete case but considering sets A with positive probability under ν to replace atoms $\{y\}$ for which transition probabilities are always null in the continuous case. For the specific case of recurrence, a slightly stronger notion of Harris recurrence is used to replace positive recurrence (Tierney, 1994). Ergodic chains are defined as aperiodic, Harris recurrent chains.

Once these definitions are given, all important convergence results established for discrete chains are valid here. For convenience, they are reviewed below for a continuous Markov chain $\theta^{(n)}$ with state space $S \subset R^d$ and stationary distribution π:

- Irreducibility and aperiodicity of the chain is equivalent to ergodicity and the uniqueness of π as the limiting distribution in total variation norm.
- The ergodic averages of real-valued functions $t(\theta)$ converge almost surely to their limiting expectations (when they exist) as stated in (4.6).
- The central limit theorem stated in (4.11) with τ^2 given by (4.9) applies to ergodic averages.

Verification of ergodicity may be difficult for some chains. Tierney (1994) proved that most Markov chains used nowadays for simulation are ergodic and the above results can be applied. Finally, the important condition of reversibility of a chain is given by

$$\pi(x)p(x,y) = \pi(y)p(y,x), \text{ for all } x, y \in S \tag{4.15}$$

in direct analogy with the discrete case. Note that (4.14) follows directly from (4.15) by integrating both sides with respect to x. Reversible chains have proved to be very useful in helping specification of a Markov chain with limiting distribution π.

4.8 Simulation of a Markov chain

Consider an ergodic Markov chain $(\theta^{(n)})_{n \geq 0}$ with state space $S \subset R^d$, transition kernel $p(x, y)$ and initial distribution $\pi^{(0)}$. Generation of a value of this chain starts with a value for $\theta^{(0)}$ sampled from $\pi^{(0)}$. The value of $\theta^{(1)}$ is then distributed with density $p(\theta^{(0)}, \cdot)$ and can be generated from it (Section 1.3). For $\theta^{(2)}$, this procedure is repeated by drawing from a distribution with density $p(\theta^{(1)}, \cdot)$. Iterating this scheme through the steps of the chain leads to drawing $\theta^{(n)}$ from a distribution with density $p(\theta^{(n-1)}, \cdot)$, for all n.

As the value of n gets large, the draws become increasingly closer to draws from the limiting distribution π and can be considered as approximate draws from π. Note that all chain values sampled after convergence is reached are also draws from π due to stationarity. Here and throughout this book, convergence is assumed to hold approximately for an iteration whose marginal distribution is arbitrarily close to the equilibrium distribution π and not in the formal sense.

These simple results have a far-reaching impact on simulation well beyond the study of Markov chains. They provide sophisticated machinery with which to approach sampling from any (possibly highly dimensional) distribution π. This area of study is collectively known as Markov chain Monte Carlo (MCMC) methods. Chapter 2 evidenced the need for summarization of posterior distributions π. Many examples showed that this task is far from trivial in complex models and Chapter 3 presented some approximating alternatives, including simulation. It was shown there that non-iterative techniques have a limited scope. As the dimension of the model gets large, they become less reliable.

This chapter provides the means by which sampling from virtually any posterior distribution π can be approached. One simply has to embed π as the limiting distribution of an ergodic Markov chain with transition kernel p. The main requirement from $p(x, \cdot)$ is to provide distributions that can be sampled from.

The remainder of this book is devoted to presenting and studying Markov chains whose simulation lead to draws from a limiting distribution of interest π. Many important questions arise and will be tackled during the next chapters. The most basic one is whether such chains can always be constructed and sampled from. It is remarkable that there are many such chains for any posterior distribution π, no matter how complex the model is. Examples 4.7 and 4.8 have provided simple cases that hint at a positive answer. Although these examples considered only the discrete context, they will be extended in the next chapters to accommodate highly dimensional continuous distributions.

Other relevant questions regard the criteria for selecting the iteration to stop sampling and choice of kernel among the possible alternatives. The first point deals with determination of convergence of the chain. In more precise terms, one would wish to ascertain how close the marginal distribution $\pi^{(n)}$ of the current iteration is to the target distribution π. Many aspects are involved including theoretical bounds on probability distances and the computational complexity of calculations required. In broad terms, the initial chain values are far from the stationary distribution and should be discarded. This period is referred to as the warm-up or burn-in period for obvious reasons. Deletion of these values hopefully improves the accuracy of ergodic averages but enlarges their variance by the reduction of the sample size. This point will be returned to when convergence diagnostics are discussed in the next chapter. In any case, Markov chain simulation opens up a host of possibilities for sampling in situations where the direct sampling methods of Chapter 1 and the indirect sampling methods of Chapter 3 do not apply.

Before concluding this section, it is important to comment on methods for assessing the accuracy of the estimates provided by ergodic averages as measured by their variance τ_n^2/n or its limiting approximation τ^2/n. There are many methods reviewed by Ripley (1987) and Geyer (1992). They can be broadly divided into direct, time series and batching methods. Assume that interest lies in the estimation of $E_\pi[t(\theta)]$ and a stream of simulated values $t^{(n)} = t(\theta^{(n)})$ is available from a Markov chain with stationary distribution π.

Direct methods are based on estimates $\hat\tau_1^2$ of τ^2 obtained by respective replacements of σ^2 and ρ_k in (4.9) by moment estimates $\hat\sigma^2 = \hat\gamma_0$ and $\hat\rho_k = \hat\gamma_k/\hat\gamma_0$, $k \le k^*$ where

$$\hat\gamma_k = \frac{1}{n}\sum_{j=1}^{n-k} t^{(j)}t^{(j+k)} - \bar t^2 \text{ , for } k \ge 0 \tag{4.16}$$

and for $k > k^*$, $\hat\rho_k = 0$ where k^* is chosen to limit the sum to include relevant terms. Unfortunately, the resulting estimate $\hat\tau^2$ will not be consistent. Many variants of the above estimates considering other multiplying constants instead of $1/n$ have been proposed and are reviewed by Priestley (1981). Consistency is ensured by appropriately downweighting higher order autocorrelation (Geyer, 1992).

There are many methods of estimation based on time-series ideas. One approach is to fit an autoregressive structure to the time series $t^{(n)}$ and estimate τ^2 from the estimated residual variance. More generally, ARMA models can be used. Geweke (1992) used estimates based on the spectral density $S(w)$ of the series evaluated at frequency $w = 0$. These are commonly used in the study of time series in the frequency domain (Priestley, 1981). Geyer (1992) considers other estimators based on the autocovariance structure of Markov chains.

Batching estimators are based on the simple idea of dividing the stream of $n = mk$ chain values into k batches of m successive values. The rationale behind it is to seek approximate independence between batches and therefore take the batch averages $\bar t_1, \ldots, \bar t_k$ as approximately independent quantities. The value of k is chosen to enforce approximate independence or at least very low autocorrelation of the sequence. Generally, values of k should be between 10 and 30 (Schmeiser, 1982). Then, for large m, each $\bar t_i$ has common approximate mean $E_\pi[t(\theta)]$ and variance $\tau^2/m = kVar(\bar t)$. Hence, the sample variance of the $\bar t_i$ estimates $kVar(\bar t)$ and τ^2 is estimated as

$$\frac{\hat\tau_2^2}{n} = \frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar t_i - \bar{\bar t})^2$$

where $\bar{\bar t} = \bar t$. Inference about $E_\pi[t(\theta)]$ is based on an approximate sampling distribution $\sqrt{n}\{\bar t - E_\pi[t(\theta)]\}/\hat\tau_2 \sim t_{k-1}(0,1)$, to account for the extra variability due to the estimation of τ^2. The simplicity of the method has made it a popular choice for estimation of sampling variance in Markov chains.

4.9 Data augmentation or substitution sampling

This chapter concludes with an example of a Markov chain constructed by Tanner and Wong (1987) to have π as a limiting distribution. Assume that $\theta = (\phi, \psi)$ has posterior distribution π and components ϕ and ψ can have any dimension. Assume also that interest lies mostly in inference about ϕ, ψ being a set of constructed parameters, latent variables or additional data. Under the last interpretation data already available is augmented by ψ, hence the first name of the method.

The marginal posterior densities of ϕ and ψ are obtained as

$$\begin{aligned} \pi(\phi) &= \int \pi(\phi|\psi)\pi(\psi)d\psi \quad \text{and} \\ \pi(\psi) &= \int \pi(\psi|x)\pi(x)dx \end{aligned}$$

where x is a dummy argument playing the role of the parameter ϕ. Substituting the second into the first equation and interchanging integration signs leads to

$$\begin{aligned} \pi(\phi) &= \int \pi(\phi|\psi)\left[\int \pi(\psi|x)\pi(x)dx\right]d\psi \\ &= \int p(x,\phi)\pi(x)dx \end{aligned} \tag{4.17}$$

where $p(x,\phi) = \int \pi(\phi|\psi)\pi(\psi|x)d\psi$. Equation (4.17) suggests that successive substitutions of $\pi(\phi)$ form an iterative algorithm. Since the integrations involved will generally not be feasible analytically, they can be replaced by sampling approximations. Hence the second name of the method. More importantly, Equation (4.17) is Equation (4.14), satisfied by the stationary distribution $\pi(\phi)$ of a Markov chain $\phi^{(n)}$ with transition kernel $p(x,\phi)$.

The iterative solution proposed by Tanner and Wong (1987) is to update an approximation $\pi^{(n)}$ of $\pi(\phi)$ to $\pi^{(n+1)}$ as follows:

1. Draw a sample $\phi_1, \ldots, \phi_m$ from $\pi^{(n)}(\phi)$.

2. Draw a sample $\psi_1, \ldots, \psi_m$ from $\pi(\psi)$. This is approximately achieved by drawing ψ_i from $\pi(\psi|\phi_i)$, $i = 1, \ldots, m$ (Section 1.3).
3. Form a Monte Carlo approximation

$$\pi^{(n+1)}(\phi) = \frac{1}{m} \sum_{i=1}^{m} \pi(\phi|\psi_i) \; .$$

For large m, these steps form a sampling-based approximation to an iteration of the transition kernel $p(x, \phi)$, that only depends on the conditional distributions. It has been stressed in the previous chapters that despite the difficulty in direct sampling from the marginal distributions, sampling from the conditionals is generally easy for many models. So, step 2 can in these cases be performed. Step 3 informs that sampling required at step 1 will be from an approximation to the marginal given by a discrete mixture of conditionals. If it is easy to sample from the conditionals, then all sampling procedures are easily carried out.

Tanner and Wong (1987) showed that the data augmentation algorithm is uniformly ergodic and $\pi(\phi)$ is the unique distribution satisfying (4.17), provided the transition kernel $p(x, \phi)$ is positive for all pairs of points (x, ϕ) in the support of $\pi(\phi)$ and is uniformly bounded and equicontinuous. These results are valid for any choice of m. Also, note that the algorithm is symmetric in terms of ϕ and ψ. Therefore, after convergence, samples of size m are available from the marginal distributions of both ϕ and ψ.

It is also interesting to consider the case $m = 1$. An iteration of the algorithm simply alternates single draws from the conditional distributions. This structure resembles Example 4.7 and is the basis of Gibbs sampling scheme, introduced in the next chapter. Also, this iterative algorithm can be extended to more than two components (Gelfand and Smith, 1990).

4.10 Exercises

4.1 *Obtain the transition matrices for the chains described in Examples 4.1 to 4.4.*

4.2 *Consider the problem of sending binary messages of length d through a channel consisting of various stages. The transmission through each stage has error probability α. Let θ_0 be the message originally sent and θ_n the message received at the nth stage.*

(a) Obtain the transition matrix.

(b) What is the probability that a signal is correctly received at the second stage?

(c) What is the probability that a signal is incorrectly received for the first time at the second stage?

4.3 *Let P be a transition matrix. Prove that P^k is stochastic and has at least one eigenvalue equal to one, $k \geq 1$.*

4.4 *Consider the Ehrenfest model for $r = 3$.*

(a) Obtain $Pr_x(T_0 = n)$ for $x \in S$ and $1 \leq n \leq 3$.

(b) Obtain the matrices P, P^2 and P^3.

(c) If $\pi_0 = (1, 1, 1, 1)/4$, calculate π_1, π_2 and π_3.

4.5 *Consider a modified Ehrenfest model with $S = \{0, 1, \cdots, r\}$ and transition probabilities given by*

$$P(x, y) = \begin{cases} (d - x)/2d & \text{, if } y = x + 1 \\ 1/2 & \text{, if } y = x \\ x/2d & \text{, if } y = x - 1 \\ 0 & \text{, if } |y - x| \neq 1 \end{cases} .$$

If initially the first urn has on average d/2 balls, how many balls can be expected to lie in the first urn at the next step?

4.6 *Show that the full conditional distribution of $\theta^{(i)}$ depends only on $\theta^{(i-1)}$ and $\theta^{(i+1)}$, for $i = 1, 2, \ldots$ (You may find it easier to prove the result for the special cases of absolutely continuous or discrete state spaces.)*

4.7 *Show that*

$$E(T_y \,|\, \theta^{(0)} = x) = \sum_{n=0}^{\infty} Pr_x(T_y > n) \;\; \text{and} \;\; E(N(y) \,|\, \theta^{(0)} = x) = \sum_{n=1}^{\infty} P^n(x, y) \; .$$

4.8 *Show that if $y \in S$ is a transient state then, for all $x \in S$,*

$$Pr_x(N(y) < \infty) = 1 \;\; \text{and} \;\; E[N(y) \mid \theta^{(0)} = x] = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty \; .$$

Show also that if $y \in S$ is a recurrent state then

$$Pr_y(N(y) = \infty) = 1 \;\; \text{and} \;\; E[N(y) \mid X_0 = y] = \infty \; .$$

4.9 *Show that birth and death processes are irreducible Markov chains when $p_x > 0$ for $x \geq 0$ and $q_x > 0$ for $x > 0$.*

4.10 *Consider again a birth and death process in $S = \{0, 1, 2, \cdots\}$ with $p_x = \frac{x+2}{2(x+1)}$ and $q_x = \frac{x}{2(x+1)}$.*

(a) Determine whether the chain is recurrent or transient.

(b) Determine $Pr_x(T_a < T_b)$ for $a < x < b$.

4.11 *Show that*

(a) $\rho_{xy} > 0$ is equivalent to $P^n(x, y) > 0$ for some $n > 0$;

(b) if $x \in S$ is recurrent, $x \rightarrow y$ and $y \rightarrow x$ then $y \in S$ is also recurrent;

(c) if $x \in S$ is recurrent, $x \not\rightarrow y$ and $y \not\rightarrow x$ then $P(x, y) = 0$.

4.12 *Consider the 2×2 version of the Gibbs sampler presented in Example 4.7. Obtain the 4×4 transition matrix P and show that π is the stationary distribution of the chain.*

4.13 *The exports of a country can be modelled, under economic stability, as a Markov chain with states +1, 0 and -1 representing, respectively, growth of 5% or more, variation smaller than 5% and decline of 5% or more with respect to the previous year. Let the transition matrix be*

$$P = \begin{pmatrix} 0.8 & 0.2 & 0 \\ 0.35 & 0.3 & 0.35 \\ 0 & 0.4 & 0.6 \end{pmatrix} .$$

(a) Does this chain have a limiting distribution?

(b) Determine the average return times for all states.

4.14 *Points 0,1,2,3 and 4 are marked clockwise in a circle. At each step, a particle moves with probability p to the right (clockwise) and $1-p$ to the left (anti-clockwise). Let $\theta^{(n)}$ be the position of the particle in the circle at the nth step. Obtain the transition matrix and the limiting distribution, if it exists. What is the expected number of steps for a return to the initial state?*

4.15 *Analyse the limiting behavior of the transition matrices*

$$P_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} , \quad P_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad P_3 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix} .$$

4.16 *Prove the ergodic theorem for $t(\theta) = \theta_i$, $i = 1, 2$, under the conditions of Example 4.7 with $p_{00} = p_{11} = p/2$ and $p_{01} = p_{10} = (1-p)/2$. In other words, obtain that*

$$Pr\left(\lim_{j\to\infty} \frac{1}{j} \sum_{l=1}^{j} \theta_{il} = \frac{1}{2}\right) = 1, \quad i = 1, 2 .$$

4.17 *Consider a Markov chain $\theta^{(n)}$ and define the autocovariance of lag k $(k \geq 0)$ of the chain as $\gamma_k = Cov_\pi(\theta^{(n)}, \theta^{(n+k)})$, the variance of $\theta^{(n)}$ as $\sigma^2 = \gamma_0$, the autocorrelation of lag k as $\rho_k = \gamma_k/\sigma^2$ and $\tau_n^2/n = Var_\pi(\bar{\theta}_n)$. Show that*

$$\tau_n^2 = \sigma^2 \left(1 + 2 \sum_{k=1}^{n-1} \frac{n-k}{n} \rho_k\right)$$

and that $\tau_n^2 \to \tau^2$ as $n \to \infty$ where $\tau^2 = \sigma^2(1 + 2\sum_{k=1}^{\infty} \rho_k)$ if the series of autocorrelation is summable.

4.18 *Obtain for continuous state spaces that*

(a) $P^{n+m}(x,y) = \int_{-\infty}^{\infty} P^m(z,y) p^n(x,z) dz$;

(b) $\pi^{(n)}(y) = \int_{-\infty}^{\infty} p(x,y)\pi^{(n-1)}(x)dx = \int_{-\infty}^{\infty} p^n(x,y)\pi^{(0)}(x)dx$.

4.19 *Consider a chain $(\theta^{(n)})_{n\geq 1}$ formed according to the process described in Example 4.9. Show that the transition kernel and the density of the limiting distribution of the chain are respectively given by*

$$p(x,y) = \begin{cases} e^{1-x} & , \text{if } y < x \\ e^{1-x} - e^{y-x} & , \text{if } y > x \end{cases} \quad \text{and} \quad \pi(y) = \begin{cases} 2(1-y) & , \text{if } 0 < y < 1 \\ 0 & , \text{otherwise} \end{cases} .$$

4.20 *Discuss the sense in which an iteration of the data augmentation method provides a sampling-based version of the transition kernel in (4.17).*