We shall now describe one procedure to simulate a Markov chain $(X_n)_{n\geq 0}$ with initial distribution λ and transition matrix P. Since $\sum_{i\in I}\lambda_i = 1$ we can partition $[0,1]$ into disjoint subintervals $(A_i : i \in I)$ with lengths

$$|A_i| = \lambda_i.$$

Similarly for each $i \in I$, we can partition $[0,1]$ into disjoint subintervals $(A_{ij} : j \in I)$ such that

$$|A_{ij}| = p_{ij}.$$

Now define functions

$$\begin{aligned} G_0 &: [0,1] \to I, \\ G &: I \times [0,1] \to I \end{aligned}$$

by

$$\begin{aligned} G_0(u) &= i \quad \text{if } u \in A_i, \\ G(i,u) &= j \quad \text{if } u \in A_{ij}. \end{aligned}$$

Suppose that $U_0, U_1, U_2, \dots$ is a sequence of independent random variables, uniformly distributed on $[0,1]$, and set

$$\begin{aligned} X_0 &= G_0(U_0), \\ X_{n+1} &= G(X_n, U_{n+1}) \quad \text{for } n \geq 0. \end{aligned}$$

Then

$$\begin{aligned} &\mathbb{P}(X_0 = i) = \mathbb{P}(U_0 \in A_i) = \lambda_i, \\ &\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, \dots, X_n = i_n) = \mathbb{P}(U_{n+1} \in A_{i_n i_{n+1}}) = p_{i_n i_{n+1}} \end{aligned}$$

so $(X_n)_{n\geq 0}$ is Markov(λ, P).

This simple procedure may be used to investigate empirically those aspects of the behaviour of a Markov chain where theoretical calculations become infeasible.

The remainder of this section is devoted to one application of the simulation of Markov chains. It is the application which finds greatest practical use, especially in statistics, statistical physics and computer science, known as *Markov chain Monte Carlo*. Monte Carlo is another name for computer simulation so this sounds no different from the procedure just discussed. But what is really meant is simulation *by means of* Markov chains, the object of primary interest being the invariant distribution of the Markov chain and not the chain itself. After a general discussion we shall give two examples.

The context for Markov chain Monte Carlo is a state-space in product form

$$I = \prod_{m\in\Lambda} S_m$$

where Λ is a finite set. For the purposes of this discussion we shall also assume that each component S_m is a finite set. A random variable X with values in I is then a family of component random variables $\big(X(m) : m \in \Lambda\big)$, where, for each site $m \in \Lambda$, $X(m)$ takes values in S_m.

We are given a distribution $\pi = (\pi_i : i \in I)$, perhaps up to an unknown constant multiple, and it is desired to compute the number

$$\sum_{i \in I} \pi_i f_i \tag{5.13}$$

for some given function $f = (f_i : i \in I)$. The essential point to understand is that Λ is typically a large set, making the state-space I very large indeed. Then certain operations are computationally infeasible – performing the sum (5.13) state by state for a start.

An alternative approach would be to simulate a large number of independent random variables $X_1, \dots, X_n$ in I, each with distribution π, and to approximate (5.13) by

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k).$$

The strong law of large numbers guarantees that this is a good approximation as $n \to \infty$ and, moreover, one can obtain error estimates which indicate how large to make n in practice. However, simulation from the distribution π is also difficult, unless π has product form

$$\pi(x) = \prod_{m \in \Lambda} \pi_m\big(x(m)\big).$$

For recall that a computer just simulates sequences of independent $U[0,1]$ random variables. When π does not have product form, Markov chain Monte Carlo is sometimes the only way to simulate samples from π.

The basic idea is to simulate a Markov chain $(X_n)_{n \geq 0}$, which is constructed to have invariant distribution π. Then, assuming aperiodicity and irreducibility, we know, by Theorem 1.8.3, that as $n \to \infty$ the distribution of X_n converges to π. Indeed, assuming only irreducibility, Theorem 1.10.2 shows that

$$\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \to \sum_{i \in I} \pi_i f_i$$

with probability 1. But why should simulating an entire Markov chain be easier than simulating a simple distribution π? The answer lies in the fact that the state-space is a product.

Each component $X_0(m)$ of the initial state X_0 is a random variable in S_m. It does not matter crucially what distribution X_0 is given, but we might, for example, make all components independent. The process $(X_n)_{n \geq 0}$ is made to evolve by changing components one site at a time. When the chosen site is m, we simulate a new random variable $X_{n+1}(m)$ with values in S_m according to a distribution determined by X_n, and for $k \neq m$ we set $X_{n+1}(k) = X_n(k)$. Thus at each step we have only to simulate a random variable in S_m, not one in the much larger space I.

Let us write $i \overset{m}{\sim} j$ if i and j agree, except possibly at site m. The law for simulating a new value at site m is described by a transition matrix $P(m)$, where

$$p_{ij}(m) = 0 \quad \text{unless } i \overset{m}{\sim} j.$$

We would like π to be invariant for $P(m)$. A sufficient condition is that the detailed balance equations hold: thus for all i, j we want

$$\pi_i p_{ij}(m) = \pi_j p_{ji}(m).$$

There are many possible choices for $P(m)$ satisfying these equations. Indeed, given any stochastic matrix $R(m)$ with

$$r_{ij}(m) = 0 \quad \text{unless } i \overset{m}{\sim} j$$

we can determine such a $P(m)$ by

$$\pi_i p_{ij}(m) = \big(\pi_i r_{ij}(m)\big) \wedge \big(\pi_j r_{ji}(m)\big)$$

for $i \neq j$, and then

$$p_{ii}(m) = 1 - \sum_{j \neq i} p_{ij}(m) \geq 0.$$

This has the following interpretation: if $X_n = i$ we simulate a new random variable Y_n so that $Y_n = j$ with probability $r_{ij}(m)$, then if $Y_n = j$ we set

$$X_{n+1} = \begin{cases} Y_n & \text{with probability } \big(\pi_i r_{ij}(m)/\pi_j r_{ji}(m)\big) \wedge 1 \\ X_n & \text{otherwise.} \end{cases}$$

This is called a *Hastings algorithm*.

There are two commonly used special cases. On taking

$$r_{ij}(m) = \Big(\sum_{k \overset{m}{\sim} i} \pi_k\Big)^{-1} \pi_j \qquad \text{for } i \overset{m}{\sim} j$$

we also find

$$p_{ij}(m) = \Big(\sum_{k \overset{m}{\sim} i} \pi_i\Big)^{-1} \pi_j \qquad \text{for } i \overset{m}{\sim} j.$$

So we simply resample $X_n(m)$ according to the conditional distribution under π, given the other components. This is called the *Gibbs sampler*. It is particularly useful in Bayesian statistics.

On taking $r_{ij}(m) = r_{ji}(m)$ for all i and j we find

$$p_{ij}(m) = \big((\pi_j/\pi_i) \wedge 1\big) r_{ij}(m) \qquad \text{for } i \overset{m}{\sim} j, i \neq j.$$

This is called a *Metropolis algorithm*. A particularly simple case would be to take

$$r_{ij}(m) = 1/(N_m - 1) \qquad \text{for } i \stackrel{m}{\sim} j, i \neq j$$

where $N_m = |S_m|$. This amounts to choosing another value j_m at site m uniformly at random; if $\pi_j > \pi_i$, then we adopt the new value, whereas if $\pi_j \leq \pi_i$ we adopt the new value with probability π_j/π_i.

We have not yet specified a rule for deciding which site to visit when. In practice this may not matter much, provided we keep returning to every site. For definiteness we mention two possibilities. We might choose to visit every site once and then repeat, generating a sequence of sites $(m_n)_{n\geq 0}$. Then $(m_n, X_n)_{n\geq 0}$ is a Markov chain in $\Lambda \times I$. Alternatively, we might choose a site randomly at each step. Then $(X_n)_{n\geq 0}$ is itself a Markov chain with transition matrix

$$P = |\Lambda|^{-1} \sum_{m\in\Lambda} P(m).$$

We shall stick with this second choice, where the analysis is simpler to present. Let us assume that P is irreducible, which is easy to ensure in the examples. We know that

$$\pi_i p_{ij}(m) = \pi_j p_{ji}(m)$$

for all m and all i, j, so also

$$\pi_i p_{ij} = \pi_j p_{ji}$$

and so π is the unique invariant measure for P. Hence, by Theorem 1.10.2, we have

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \to \sum_{i\in I} \pi_i f_i$$

as $n \to \infty$ with probability 1. Thus the algorithm works eventually. In practice one is concerned with how fast it works, but useful information of this type cannot be gained in the present general context. Given more information on the structure of S_m and the distribution π to be simulated, much more can be said. We shall not pursue the matter here. It should also be emphasised that there is an empirical side to simulation: with due caution informed by the theory, the computer output gives a good idea of how well we are doing. For further reading we recommend *Stochastic Simulation* by B. D. Ripley (Wiley, Chichester, 1987), and *Markov Chain Monte Carlo in practice* by W. R. Gilks, S. Richardson and D. J. Spiegelhalter (Chapman and Hall, London, 1996). The recent survey article *Bayesian computation and stochastic systems* by J. Besag, P. Green, D. Higdon and K. Mengersen (*Statistical Science*, 10 (1), pp.3–40, 1995) contains many interesting references. We finish with two examples.

Example 5.5.1 (Bayesian statistics)

In a statistical problem one may be presented with a set of independent observations $Y_1, \dots, Y_n$, which it is reasonable to assume are normally distributed, but with unknown mean μ and variance τ^{-1}. One then seeks to draw conclusions about μ and τ on the basis of the observations. The Bayesian approach to this problem is to assume that μ and τ are themselves random variables, with a given prior distribution. For example, we might assume that

$$\mu \sim N(\theta_0, \phi_0^{-1}), \quad \tau \sim \Gamma(\alpha_0, \beta_0),$$

that is to say, μ is normal of mean θ_0 and variance ϕ_0^{-1}, and τ has gamma distribution of parameters α_0 and β_0. The parameters θ_0, ϕ_0, α_0 and β_0 are known. Then the prior density for (μ, τ) is given by

$$\pi(\mu, \tau) \propto \exp\{-\phi_0(\mu - \theta_0)^2/2\}\tau^{\alpha_0 - 1}\exp\{-\beta_0\tau\}.$$

The posterior density for (μ, τ), which is the conditional density given the observations, is then given by Bayes' formula

$$\begin{aligned}\pi(\mu, \tau \mid y) &\propto \pi(\mu, \tau) f(y \mid \mu, \tau)\\ &\propto \exp\{-\phi_0(\mu - \theta_0)^2/2\}\exp\left\{-\tau\sum_{i=1}^n (y_i - \mu)^2/2\right\}\tau^{\alpha_0 - 1 + n/2}\exp\{-\beta_0\tau\}.\end{aligned}$$

Note that the posterior density is no longer in product form: the conditioning has introduced a dependence between μ and τ. Nevertheless, the *full conditional distributions* still have a simple form

$$\begin{aligned}\pi(\mu \mid y, \tau) &\propto \exp\{-\phi_0(\mu - \theta_0)^2/2\}\exp\left\{-\tau\sum_{i=1}^n (y_i - \mu)^2/2\right\} \sim N(\theta_n, \phi_n^{-1}),\\ \pi(\tau \mid y, \mu) &\propto \tau^{\alpha_0 - 1 + n/2}\exp\left\{-\tau\left(\beta_0 + \sum_{i=1}^n (y_i - \mu)^2/2\right)\right\} \sim \Gamma(\alpha_n, \beta_n)\end{aligned}$$

where

$$\begin{aligned}\theta_n &= \left(\phi_0\theta_0 + \tau\sum_{i=1}^n y_i\right)/(\phi_0 + n\tau), \quad \phi_n = \phi_0 + n\tau,\\ \alpha_n &= \alpha_0 + n/2, \quad \beta_n = \beta_0 + \sum_{i=1}^n (y_i - \mu)^2/2.\end{aligned}$$

Our final belief about μ and τ is regarded as measured by the posterior density. We may wish to compute probabilities and expectations. Here the *Gibbs sampler* provides a particularly simple approach. Of course, numerical integration would also be feasible as the dimension is only two. To make the connection with our general discussion we set

$$I = S_1 \times S_2 = \mathbb{R} \times [0, \infty).$$

We wish to simulate $X = (\mu, \tau)$ with density $\pi(\mu, \tau \mid y)$. The fact that $\mathbb{R}$ and $[0, \infty)$ are not finite sets does not affect the basic idea. In any case the computer will work with finite approximations to $\mathbb{R}$ and $[0, \infty)$. First we simulate X_0, say from the product form density $\pi(\mu, \tau)$. At the kth stage, given $X_k = (\mu_k, \tau_k)$, we first simulate μ_{k+1} from $\pi(\mu \mid y, \tau_k)$ and then τ_{k+1} from $\pi(\tau \mid y, \mu_{k+1})$, then set $X_{k+1} = (\mu_{k+1}, \tau_{k+1})$. Then $(X_k)_{k \geq 0}$ is a Markov chain in I with invariant measure $\pi(\mu, \tau \mid y)$, and one can show that

$$\frac{1}{k} \sum_{j=0}^{k-1} f(X_j) \to \int_I f(x) \pi(x \mid y) dx \quad \text{as } k \to \infty$$

with probability 1, for all bounded continuous functions $f : I \to \mathbb{R}$. This is not an immediate consequence of the ergodic theorem for discrete state-space, but you may find it reasonable at an intuitive level, with a rate of convergence depending on the smoothness of π and f.

We now turn to an elaboration of this example where the Gibbs sampler is indispensible. The model consists of m copies of the preceding one, with different means but a common variance. Thus there are mn independent observations Y_{ij}, where $i = 1, \ldots n$, and $j = 1, \ldots, m$, normally distributed, with means μ_j and common variance τ^{-1}. We take these parameters to be independent random variables as before, with

$$\mu_j \sim N(\theta_0, \phi_0^{-1}), \quad \tau \sim \Gamma(\alpha_0, \beta_0).$$

Let us write $\mu = (\mu_1, \ldots, \mu_n)$. The prior density is given by

$$\pi(\mu, \tau) \propto \exp\left\{ -\phi_0 \sum_{j=1}^{m} (\mu_j - \theta_0)^2 / 2 \right\} \tau^{\alpha_0 - 1} \exp\{-\beta_0 \tau\}$$

and the posterior density is given by

$$\begin{aligned} \pi(\mu, \tau \mid y) \propto \exp & \left\{ -\phi_0 \sum_{j=1}^{m} (\mu_j - \theta_0)^2 / 2 \right\} \\ & \times \exp\left\{ -\tau \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \mu_j)^2 / 2 \right\} \tau^{\alpha_0 - 1 + mn/2} \exp\{-\beta_0 \tau\}. \end{aligned}$$

Hence the full conditional distributions are

$$\pi(\mu_j \mid y, \tau) \sim N(\theta_{jn}, \phi_n^{-1}), \quad \pi(\tau \mid y, \mu) \sim \Gamma(\alpha_n, \beta_n)$$

where

$$\begin{aligned} \theta_{jn} &= \left(\phi_0 \theta_0 + \tau \sum_{i=1}^{n} y_{ij} \right) / (\phi_0 + n\tau), \quad \phi_n = \phi_0 + n\tau, \\ \alpha_n &= \alpha_0 + mn/2, \quad \beta_n = \beta_0 + \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \mu_j)^2 / 2. \end{aligned}$$

We can construct approximate samples from $\pi(\mu, \tau \mid y)$, just as in the case $m = 1$ discussed above, by a Gibbs sampler method. Note that, conditional on τ, the means μ_j, for $j = 1, \dots, m$, remain independent. Thus one can update all the means simultaneously in the Gibbs sampler. This has the effect of speeding convergence to the equilibrium distribution. In cases where m is large, numerical integration of $\pi(\mu, \tau \mid y)$ is infeasible, as is direct simulation from the distribution, so the Markov chain approach is the only one available.

Example 5.5.2 (Ising model and image analysis)

Consider a large box $\Lambda = \Lambda_N$ in $\mathbb{Z}^2$

$$\Lambda = \{-N, \dots, -1, 0, 1, \dots, N\}^2$$

with boundary $\partial\Lambda = \Lambda_N \backslash \Lambda_{N-1}$, and the *configuration space*

$$I = \{-1, 1\}^\Lambda.$$

For $x \in \Lambda$ define

$$H(x) = \tfrac{1}{2} \sum \big(x(m) - x(m')\big)^2$$

where the sum is taken over all pairs $\{m, m'\} \subseteq \Lambda$ with $|m - m'| = 1$. Note that $H(x)$ is small when the values taken by x at neighbouring sites are predominantly the same. We write

$$I^+ = \{x \in I : x(m) = 1 \quad \text{for all } m \in \partial\Lambda\}$$

and for each $\beta > 0$ define a probability distribution $\big(\pi(x) : x \in I^+\big)$ by

$$\pi(x) \propto e^{-\beta H(x)}.$$

As $\beta \downarrow 0$ the weighting becomes uniform, whereas, as $\beta \uparrow \infty$ the mass concentrates on configurations x where $H(x)$ is small. This is one of the fundamental models of statistical physics, called the *Ising model*. A famous and deep result of Onsager says that if X has distribution π, then

$$\lim_{N \to \infty} \mathbb{E}\big(X(0)\big) = \big[\big(1 - (\sinh 2\beta)^{-4}\big)^+\big]^{1/8}.$$

In particular, if $\sinh 2\beta \le 1$, the fact that X is forced to take boundary values 1 does not significantly affect the distribution of $X(0)$ when N is large, whereas if $\sinh 2\beta > 1$ there is a residual effect of the boundary values on $X(0)$, uniformly in N.

Here we consider the problem of simulating the Ising model. Simulations may sometimes be used to guide further developments in the theory, or even to detect phenomena quite out of reach of the current theory. In fact, the Ising model is rather well understood theoretically; but there are many related models which are not, where simulation is still possible by simple modifications of the methods presented here.

First we describe a Gibbs sampler. Consider the sets of even and odd sites

$$\begin{aligned}\Lambda^+ &= \{(m_1, m_2) \in \Lambda : m_1 + m_2 \text{ is even}\},\\ \Lambda^- &= \{(m_1, m_2) \in \Lambda : m_1 + m_2 \text{ is odd}\}\end{aligned}$$

and for $x \in I$ set

$$x^\pm = \big(x(m) : m \in \Lambda^\pm\big).$$

We can exploit the fact that the conditional distribution $\pi(x^+ \mid x^-)$ has product form

$$\pi(x^+ \mid x^-) \propto \prod_{m \in \Lambda^+ \backslash \partial\Lambda} e^{\beta x(m) s(m)}$$

where, for $m \in \Lambda^+ \backslash \partial\Lambda$

$$s(m) = \sum_{|m'-m|=1} x^-(m').$$

Therefore, it is easy to simulate from $\pi(x^+ \mid x^-)$ and likewise from $\pi(x^- \mid x^+)$. Choose now some simple initial configuration X_0 in I^+. Then inductively, given $X_n^- = x^-$, simulate firstly X_{n+1}^+ with distribution $\pi(\cdot \mid x^-)$ and then given $X_{n+1}^+ = x^+$, simulate X_{n+1}^- with distribution $\pi(\cdot \mid x^+)$. Then according to our general discussion, for large n, the distribution of X_n is approximately π. Note that we did not use the value of the normalizing constant

$$Z = \sum_{x \in I^+} e^{-\beta H(x)}$$

which is hard to compute by elementary means when N is large.

An alternative approach is to use a Metropolis algorithm. We can again exploit the even/odd partition. Given that $X_n = x$, independently for each $m \in \Lambda^+ \backslash \partial\Lambda$, we change the sign of $X_n^+(m)$ with probability

$$p(m, x) = \big(\pi(\hat{x})/\pi(x)\big) \wedge 1 = e^{2\beta x(m) s(m)} \wedge 1$$

where $\hat{x} \overset{m}{\sim} x$ with $\hat{x}(m) = -x(m)$. Let us call the resulting configuration Y_n. Next we apply the corresponding transformation to $Y_n^-(m)$ for the odd sites $m \in \Lambda^- \backslash \partial\Lambda$, to obtain X_{n+1}. The process $(X_n)_{n \ge 0}$ is then a Markov chain in I^+ with invariant distribution π.

Both methods we have described serve to simulate samples from π; there is little to choose between them. Convergence is fast in the subcritical case $\sinh 2\beta < 1$, where π has an approximate product structure on large scales.

In a Bayesian analysis of two-dimensional images, the Ising model is sometimes used as a prior. We may encode a digitized image on a two-dimensional grid as a particular configuration $\big(x(m) : m \in \Lambda\big) \in I$, where $x(m) = 1$ for a white pixel and $x(m) = -1$ for a black pixel. By varying the parameter β in the Ising model, we vary the tendency of black pixels to clump together; the same for white pixels.

Thus β is a sort of texture parameter, which we choose according to the sort of image we expect, thus obtaining a prior $\pi(x)$. Observations are now made at each site which record the true pixel, black or white, with probability $p \in (0,1)$. The posterior distribution for X given observations Y is then given by

$$\pi(x \mid y) \propto \pi(x) f(y \mid x) \propto e^{-\beta H(x)} p^{a(x,y)} (1-p)^{d(x,y)}$$

where $a(x,y)$ and $d(x,y)$ are the numbers of sites at which x and y agree and disagree respectively. 'Cleaned-up' versions of the observed image Y may now be obtained by simulating from the posterior distribution. Although this is not exactly the Ising model, the same methods work. We describe the appropriate Metropolis algorithm: given that $X_n = x$, independently for each $m \in \Lambda^+ \backslash \partial\Lambda$, change the sign of $X_n^+(m)$ with probability

$$\begin{aligned} p(m,x,y) &= \big(\pi(\hat{x} \mid y)/\pi(x \mid y)\big) \wedge 1 \\ &= e^{-2\beta x(m) s(m)} \big((1-p)/p\big)^{x(m)y(m)} \end{aligned}$$

where $\hat{x} \overset{m}{\sim} x$ with $\hat{x}(m) = -x(m)$. Call the resulting configuration $X_{n+1/2}$. Next apply the corresponding transformation to $X_{n+1/2}^-$ for the odd sites to obtain X_{n+1}. Then $(X_n)_{n\geq 0}$ is a Markov chain in I^+ with invariant distribution $\pi(\cdot \mid y)$.

6

Appendix: probability and measure

Section 6.1 contains some reminders about countable sets and the discrete version of measure theory. For much of the book we can do without explicit mention of more general aspects of measure theory, except an elementary understanding of Riemann integration or Lebesgue measure. This is because the state-space is at worst countable. The proofs we have given may be read on two levels, with or without a measure-theoretic background. When interpreted in terms of measure theory, the proofs are intended to be rigorous. The basic framework of measure and probability is reviewed in Sections 6.2 and 6.3. Two important results of measure theory, the monotone convergence theorem and Fubini's theorem, are needed a number of times: these are discussed in Section 6.4. One crucial result which we found impossible to discuss convincingly without measure theory is the strong Markov property for continuous-time chains. This is proved in Section 6.5. Finally, in Section 6.6, we discuss a general technique for determining probability measures and independence in terms of π-systems, which are often more convenient than σ-algebras.

6.1 Countable sets and countable sums

A set I is *countable* if there is a bijection $f : \{1, \dots, n\} \to I$ for some $n \in \mathbb{N}$, or a bijection $f : \mathbb{N} \to I$. In either case we can *enumerate* all the elements of I

$$f_1, f_2, f_3, \dots$$

where in one case the sequence terminates and in the other it does not. There would have been no loss in generality had we insisted that all our Markov chains had state-space $\mathbb{N}$ or $\{1, \dots, n\}$ for some $n \in \mathbb{N}$: this just corresponds to a particular choice of the bijection f.