7

The Metropolis–Hastings Algorithm

> "What's changed, except what needed changing?" And there was something in that, Cadfael reflected. What was changed was the replacement of falsity by truth...
> —Ellis Peter, *The Confession of Brother Haluin*

This chapter is the first of a series on simulation methods based on *Markov chains*. However, it is a somewhat strange introduction because it contains a description of the most general algorithm of all. The next chapter (Chapter 8) concentrates on the more specific slice sampler, which then introduces the Gibbs sampler (Chapters 9 and 10), which, in turn, is a special case of the Metropolis–Hastings algorithm. (However, the Gibbs sampler is different in both fundamental methodology and historical motivation.)

The motivation for this reckless dive into a completely new and general simulation algorithm is that there exists no simple case of the Metropolis–Hastings algorithm that would "gently" explain the fundamental principles of the method; a global presentation does, on the other hand, expose us to the almost infinite possibilities offered by the algorithm.

Unfortunately, the drawback of this ordering is that some parts of the chapter will be completely understood only after reading later chapters. But realize that this is the pivotal chapter of the book, one that addresses the methods that radically changed our perception of simulation and opened countless new avenues of research and applications. It is thus worth reading this chapter more than once!

7.1 The MCMC Principle

It was shown in Chapter 3 that it is not necessary to directly simulate a sample from the distribution f to approximate the integral

$$\mathfrak{I} = \int h(x)f(x)dx \ ,$$

since other approaches like *importance sampling* can be used. While Chapter 14 will clarify the complex connections existing between importance sampling and Markov chain Monte Carlo methods, this chapter first develops a somewhat different strategy and shows that it is possible to obtain a sample $X_1, \ldots, X_n$ approximately distributed from f without directly simulating from f. The basic principle underlying the methods described in this chapter and the following ones is *to use an ergodic Markov chain with stationary distribution* f.

While we will discuss below some rather general schemes to produce valid transition kernels associated with arbitrary stationary distributions, the working principle of MCMC algorithms is thus as follows: For an arbitrary starting value $x^{(0)}$, a chain $(X^{(t)})$ is generated using a transition kernel with stationary distribution f, which ensures the convergence in distribution of $(X^{(t)})$ to a random variable from f. (Given that the chain is ergodic, the starting value $x^{(0)}$ is, in principle, unimportant.)

Definition 7.1. A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose stationary distribution is f.

This simple idea of using a Markov chain with limiting distribution f may sound impractical. In comparison with the techniques of Chapter 3, here we rely on more complex asymptotic convergence properties than a simple Law of Large Numbers, as we generate dependencies within the sample that slow down convergence of the approximation of $\mathfrak{I}$. Thus, the number of iterations required to obtain a good approximation seems a priori much more important than with a standard Monte Carlo method. The appeal to Markov chains is nonetheless justified from at least two points of view. First, in Chapter 5, we have already seen that some stochastic optimization algorithms (for example, the Robbins–Monro procedure in Note 5.5.3) naturally produce Markov chain structures. It is a general fact that the use of Markov chains allows for a greater scope than the methods presented in Chapters 2 and 3. Second, regular Monte Carlo and MCMC algorithms both satisfy the $O(1/\sqrt{n})$ convergence requirement for the approximation of $\mathfrak{I}$. There are thus many instances where a specific MCMC algorithm dominates, variance-wise, the corresponding Monte Carlo proposal. For instance, while importance sampling is virtually a universal method, its efficiency relies of adequate choices of the importance function and this choice gets harder and harder as the dimension increases, a practical realization of the curse of dimensionality. At a first level, some generic algorithms, like the Metropolis–Hastings algorithms, also use simulations from almost any arbitrary density g to actually generate from an equally arbitrary given density f. At a second level, however, since these algorithms allow for the dependence of g on the previous simulation, the

choice of g does not require a particularly elaborate construction a priori but can take advantage of the local characteristics of the stationary distribution. Moreover, even when an Accept–Reject algorithm is available, it is sometimes more efficient to use the pair (f, g) through a Markov chain, as detailed in Section 7.4. Even if this point is not obvious at this stage, it must be stressed that the (re)discovery of Markov chain Monte Carlo methods by statisticians in the 1990s has produced considerable progress in simulation-based inference and, in particular, in Bayesian inference, since it has allowed the analysis of a multitude of models that were too complex to be satisfactorily processed by previous schemes.

7.2 Monte Carlo Methods Based on Markov Chains

Despite its formal aspect, Definition 7.1 can be turned into a working principle: the use of a chain $(X^{(t)})$ produced by a Markov chain Monte Carlo algorithm with stationary distribution f is fundamentally identical to the use of an iid sample from f in the sense that the ergodic theorem (Theorem 6.63) guarantees the (almost sure) convergence of the empirical average

$$\Im_T = \frac{1}{T} \sum_{t=1}^{T} h(X^{(t)}) \tag{7.1}$$

to the quantity $\mathbb{E}_f[h(X)]$. A sequence $(X^{(t)})$ produced by a Markov chain Monte Carlo algorithm can thus be employed just as an iid sample. If there is no particular requirement of independence but if, rather, the purpose of the simulation study is to examine the properties of the distribution f, there is no need for the generation of n independent chains $(X_i^{(t)})$ $(i = 1, \ldots, n)$, where only some "terminal" values $X_i^{(T_0)}$ are kept: the choice of the value T_0 may induce a bias and, besides, this approach results in the considerable waste of $n(T_0 - 1)$ simulations out of nT_0. In other words, a single realization (or *path*) of a Markov chain is enough to ensure a proper approximation of $\Im$ through estimates like (7.1) for the functions h of interest (and sometimes even of the density f, as detailed in Chapter 10). Obviously, handling this sequence is somewhat more arduous than in the iid case because of the dependence structure, but some approaches to the convergence assessment of (7.1) are given in Section 7.6 and in Chapter 12. Chapter 13 will also discuss strategies to efficiently produce iid samples with MCMC algorithms.

Given the principle stated in Definition 7.1, one can propose an infinite number of practical implementations as those, for instance, used in statistical physics. The Metropolis–Hastings algorithms described in this chapter have the advantage of imposing minimal requirements on the target density f and allowing for a wide choice of possible implementations. In contrast, the Gibbs sampler described in Chapters 8–10 is more restrictive, in the sense that it

requires some knowledge of the target density to derive some conditional densities, but it can also be more effective than a generic Metropolis–Hastings algorithm.

7.3 The Metropolis–Hastings algorithm

Before illustrating the universality of Metropolis–Hastings algorithms and demonstrating their straightforward implementation, we first address the (important) issue of theoretical validity. Since the results presented below are valid for all types of Metropolis–Hastings algorithms, we do not include examples in this section, but rather wait for Sections 7.4 and 7.5, which present a collection of specific algorithms.

7.3.1 Definition

The Metropolis–Hastings algorithm starts with the objective *(target)* density f. A conditional density $q(y|x)$, defined with respect to the dominating measure for the model, is then chosen. The Metropolis–Hastings algorithm can be implemented in practice when $q(\cdot|x)$ is easy to simulate from and is either explicitly available (up to a multiplicative constant *independent of* x) or *symmetric*; that is, such that $q(x|y) = q(y|x)$. The target density f must be available to some extent: a general requirement is that the ratio

$$f(y)/q(y|x)$$

is known up to a constant *independent* of x.

The Metropolis–Hastings algorithm associated with the objective (target) density f and the conditional density q produces a Markov chain $(X^{(t)})$ through the following transition.

Algorithm A.24 –Metropolis–Hastings–

Given $x^{(t)}$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \text{with probability } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

where [A.24]

$$\rho(x, y) = \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}.$$

The distribution q is called the *instrumental* (or *proposal*) *distribution* and the probability $\rho(x, y)$ the *Metropolis–Hastings acceptance probability.*

This algorithm always accepts values y_t such that the ratio $f(y_t)/q(y_t|x^{(t)})$ is increased, compared with the previous value $f(x^{(t)})/(q(x^{(t)}|y_t)$. It is only in the symmetric case that the acceptance is driven by the objective ratio $f(y_t)/f(x^{(t)})$. An important feature of the algorithm [*A*.24] is that it may accept values y_t such that the ratio is decreased, similar to stochastic optimization methods (see Section 5.4). Like the Accept–Reject method, the Metropolis–Hastings algorithm depends only on the ratios

$$f(y_t)/f(x^{(t)}) \qquad \text{and} \qquad q(x^{(t)}|y_t)/q(y_t|x^{(t)})$$

and is, therefore, independent of normalizing constants, assuming, again, that $q(\cdot|x)$ is known up to a constant that is *independent* of x[1].

Obviously, the probability $\rho(x^{(t)}, y_t)$ is defined only when $f(x^{(t)}) > 0$. However, if the chain starts with a value $x^{(0)}$ such that $f(x^{(0)}) > 0$, it follows that $f(x^{(t)}) > 0$ for every $t \in \mathbb{N}$ since the values of y_t such that $f(y_t) = 0$ lead to $\rho(x^{(t)}, y_t) = 0$ and are, therefore, rejected by the algorithm. We will make the *convention* that the ratio $\rho(x, y)$ is equal to 0 when both $f(x)$ and $f(y)$ are null, in order to avoid theoretical difficulties.

There are similarities between [*A*.24] and the Accept–Reject methods of Section 2.3, and it is possible to use the algorithm [*A*.24] as an alternative to an Accept–Reject algorithm for a given pair (f, g). These approaches are compared in Section 7.4. However, a sample produced by [*A*.24] differs from an iid sample. For one thing, such a sample may involve repeated occurrences of the same value, since rejection of Y_t leads to repetition of $X^{(t)}$ at time $t + 1$ (an impossible occurrence in absolutely continuous iid settings). Thus, in calculating a mean such as (7.1), the Y_t's generated by the algorithm [*A*.24] can be associated with weights of the form m_t/T $(m_t = 0, 1, \ldots)$, where m_t counts the number of times the subsequent values have been rejected. (This makes the comparison with importance sampling somewhat more relevant, as discussed in Section 7.6 and Chapter 14.)

While [*A*.24] is a generic algorithm, defined for all f's and q's, it is nonetheless necessary to impose minimal regularity conditions on both f and the conditional distribution q for f to be the limiting distribution of the chain $(X^{(t)})$ produced by [*A*.24]. For instance, it is easier if $\mathcal{E}$, the support of f, is *connected*: an unconnected support $\mathcal{E}$ can invalidate the Metropolis–Hastings algorithm. For such supports, it is necessary to proceed on one connected component at a time and show that the different connected components of $\mathcal{E}$ are linked by the kernel of [*A*.24]. If the support of $\mathcal{E}$ is truncated by q, that is, if there exists $A \subset \mathcal{E}$ such that

$$\int_A f(x)dx > 0 \quad \text{and} \quad \int_A q(y|x)dy = 0\,, \qquad \forall x \in \mathcal{E}\,,$$

[1] If we insist on this independence from x, it is because forgetting a term in $q(\cdot|x)$ that depends on x does jeopardize the validity of the whole algorithm.

the algorithm [A.24] does not have f as a limiting distribution since, for $x^{(0)} \notin A$, the chain $(X^{(t)})$ never visits A. Thus, a minimal necessary condition is that

$$\bigcup_{x \in \text{supp } f} \text{supp } q(\cdot|x) \supset \text{supp } f \ .$$

To see that f is the stationary distribution of the Metropolis chain, we first examine the Metropolis kernel more closely and find that it satisfies the detailed balance property (6.22). (See Problem 7.3 for details of the proof.)

Theorem 7.2. *Let $(X^{(t)})$ be the chain produced by [A.24]. For every conditional distribution q whose support includes $\mathcal{E}$,*

(a) the kernel of the chain satisfies the detailed balance condition with f;
(b) f is a stationary distribution of the chain.

Proof. The transition kernel associated with [A.24] is

$$K(x,y) = \rho(x,y)q(y|x) + (1 - r(x))\delta_x(y) \ , \tag{7.2}$$

where $r(x) = \int \rho(x,y)q(y|x)dy$ and δ_x denotes the Dirac mass in x. It is straightforward to verify that

$$\begin{aligned} \rho(x,y)q(y|x)f(x) &= \rho(y,x)q(x|y)f(y) \\ (1 - r(x))\delta_x(y)f(x) &= (1 - r(y))\delta_y(x)f(y) \ , \end{aligned} \tag{7.3}$$

which together establish detailed balance for the Metropolis–Hastings chain. Part (*b*) now follows from Theorem 6.46. □

The stationarity of f is therefore established for almost any conditional distribution q, a fact which indicates the universality of Metropolis–Hastings algorithms.

7.3.2 Convergence Properties

To show that the Markov chain of [A.24] indeed converges to the stationary distribution and that (7.1) is a convergent approximation to $\mathfrak{I}$, we need to apply further the theory developed in Chapter 6.

Since the Metropolis–Hastings Markov chain has, by construction, an invariant probability distribution f, if it is also an aperiodic Harris chain (see Definition 6.32), then the ergodic theorem (Theorem 6.63) does apply to establish a result like the convergence of (7.1) to $\mathfrak{I}$.

A sufficient condition for the Metropolis–Hastings Markov chain to be *aperiodic* is that the algorithm [A.24] allows events such as $\{X^{(t+1)} = X^{(t)}\}$; that is, that the probability of such events is not zero, and thus

$$P\left[f(X^{(t)})\, q(Y_t|X^{(t)}) \leq f(Y_t)\, q(X^{(t)}|Y_t)\right] < 1. \tag{7.4}$$

Interestingly, this condition implies that q is not the transition kernel of a reversible Markov chain with stationary distribution f.[2] (Note that q is not the transition kernel of the Metropolis–Hastings chain, given by (7.2), which *is* reversible.)

The fact that [*A*.24] works only when (7.4) is satisfied is not overly troublesome, since it merely states that it is useless to further perturb a Markov chain with transition kernel q if the latter already converges to the distribution f. It is then sufficient to directly study the chain associated with q.

The property of *irreducibility* of the Metropolis–Hastings chain $(X^{(t)})$ follows from sufficient conditions such as positivity of the conditional density q; that is,

$$q(y|x) > 0 \text{ for every } (x, y) \in \mathcal{E} \times \mathcal{E}, \tag{7.5}$$

since it then follows that every set of $\mathcal{E}$ with positive Lebesgue measure can be reached in a single step. As the density f is the invariant measure for the chain, the chain is *positive* (see Definition 6.35) and Proposition 6.36 implies that the chain is recurrent. We can also establish the following stronger result for the Metropolis–Hastings chain.

Lemma 7.3. *If the Metropolis–Hastings chain $(X^{(t)})$ is f-irreducible, it is Harris recurrent.*

Proof. This result can be established by using the fact that a characteristic of Harris recurrence is that the only bounded harmonic functions are constant (see Proposition 6.61).

If h is a harmonic function, it satisfies

$$h(x_0) = \mathbb{E}[h(X^{(1)})|x_0] = \mathbb{E}[h(X^{(t)})|x_0]\,.$$

Because the Metropolis–Hastings chain is positive recurrent and aperiodic, we can use Theorem 6.80, as in the discussion surrounding (6.27), and conclude that h is f-almost everywhere constant and equal to $\mathbb{E}_f[h(X)]$. To show that h is everywhere constant, write

$$\mathbb{E}[h(X^{(1)})|x_0] = \int \rho(x_0, x_1)\, q(x_1|x_0)\, h(x_1) dx_1 + (1 - r(x_0))\, h(x_0)\,,$$

and substitute $\mathbb{E}h(X)$ for $h(x_1)$ in the integral above. It follows that

$$\mathbb{E}_f[h(X)]\, r(x_0) + (1 - r(x_0))\, h(x_0) = h(x_0)\,;$$

that is, $(h(x_0) - \mathbb{E}[h(X)])\, r(x_0) = 0$ for every $x_0 \in \mathcal{E}$. Since $r(x_0) > 0$ for every $x_0 \in \mathcal{E}$, by virtue of the f-irreducibility, h is necessarily constant and the chain is Harris recurrent. □

[2] For instance, (7.4) is not satisfied by the successive steps of the Gibbs sampler (see Theorem 10.13).

We therefore have the following convergence result for Metropolis–Hastings Markov chains.

Theorem 7.4. *Suppose that the Metropolis–Hastings Markov chain* $(X^{(t)})$ *is* f*-irreducible.*

(i) If $h \in L^1(f)$*, then*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) = \int h(x)f(x)dx \qquad a.e.\ f.$$

(ii) If, in addition, $(X^{(t)})$ *is aperiodic, then*

$$\lim_{n\to\infty} \left\| \int K^n(x,\cdot)\mu(dx) - f \right\|_{TV} = 0$$

for every initial distribution μ*, where* $K^n(x,\cdot)$ *denotes the kernel for* n *transitions, as in (6.5).*

Proof. If $(X^{(t)})$ is f-irreducible, it is Harris recurrent by Lemma 7.3 , and part (i) then follows from Theorem 6.63 (the Ergodic Theorem). Part (ii) is an immediate consequence of Theorem 6.51. □

As the f-irreducibility of the Metropolis–Hastings chain follows from the above-mentioned positivity property of the conditional density q, we have the following immediate corollary, whose proof is left as an exercise.

Corollary 7.5. *The conclusions of Theorem 7.4 hold if the Metropolis–Hastings Markov chain* $(X^{(t)})$ *has conditional density* $q(x|y)$ *that satisfies (7.4) and (7.5).*

Although condition (7.5) may seem restrictive, it is often satisfied in practice. (Note that, typically, conditions for irreducibility involve the transition kernel of the chain, as in Theorem 6.15 or Note 6.9.3.)

We close this section with a result due to Roberts and Tweedie (1996) (see Problem 7.35) which gives a somewhat less restrictive condition for irreducibility and aperiodicity.

Lemma 7.6. *Assume* f *is bounded and positive on every compact set of its support* $\mathcal{E}$*. If there exist positive numbers* ε *and* δ *such that*

$$q(y|x) > \varepsilon \quad \text{if} \quad |x-y| < \delta\,, \tag{7.6}$$

then the Metropolis–Hastings Markov chain $(X^{(t)})$ *is* f*-irreducible and aperiodic. Moreover, every nonempty compact set is a small set.*

The rationale behind this result is the following. If the conditional distribution $q(y|x)$ allows for moves in a neighborhood of $x^{(t)}$ with diameter bounded from below and if f is such that $\rho(x^{(t)}, y)$ is positive in this neighborhood, then any subset of $\mathcal{E}$ can be visited in k steps for k large enough. (This property obviously relies on the assumption that $\mathcal{E}$ is connected.)

Proof. Consider $x^{(0)}$ an arbitrary starting point and $A \subset \mathcal{E}$ an arbitrary measurable set. The connectedness of $\mathcal{E}$ implies that there exist $m \in \mathbb{N}$ and a sequence $x^{(i)} \in \mathcal{E}$ $(1 \leq i \leq m)$ such that $x^{(m)} \in A$ and $|x^{(i+1)} - x^{(i)}| < \delta$. It is therefore possible to link $x^{(0)}$ and A through a sequence of balls with radius δ. The assumptions on f imply that the acceptance probability of a point $x^{(i)}$ of the ith ball starting from the $(i-1)$st ball is positive and, therefore, $P^m_{x^{(0)}}(A) = P(X^{(m)} \in A | X^{(0)} = x^{(0)}) > 0$. By Theorem 6.15, the f-irreducibility of $(X^{(t)})$ is established.

For an arbitrary value $x^{(0)} \in \mathcal{E}$ and for every $y \in B(x^{(0)}, \delta/2)$ (the ball with center $x^{(0)}$ and radius $\delta/2$) we have

$$\begin{aligned} P_y(A) &\geq \int_A \rho(y,z)\, q(z|y)\, dz \\ &= \int_{A \cap D_y} \frac{f(z)}{f(y)}\, q(y|z)\, dz + \int_{A \cap D_y^c} q(z|y)\, dz\,, \end{aligned}$$

where $D_y = \{z; f(z)q(y|z) \leq f(y)q(z|y)\}$. It therefore follows that

$$\begin{aligned} P_y(A) &\geq \int_{A \cap D_y \cap B} \frac{f(z)}{f(y)}\, q(y|z) dz + \int_{A \cap D_y^c \cap B} q(z|y) dz \\ &\geq \frac{\inf_B f(x)}{\sup_B f(x)} \int_{A \cap D_y \cap B} q(y|z) dz + \int_{A \cap D_y^c \cap B} q(z|y) dz \\ &\geq \varepsilon \frac{\inf_B f(x)}{\sup_B f(x)}\, \lambda(A \cap B)\,, \end{aligned}$$

where λ denotes the Lebesgue measure on $\mathcal{E}$. The balls $B(x^{(0)}, \delta/2)$ are small sets associated with uniform distributions on $B(x^{(0)}, \delta/2)$. This simultaneously implies the aperiodicity of $(X^{(t)})$ and the fact that every compact set is small. □

Corollary 7.7. *The conclusions of Theorem 7.4 hold if the Metropolis–Hastings Markov chain $(X^{(t)})$ has invariant probability density f and conditional density $q(x|y)$ that satisfy the assumptions of Lemma 7.6.*

One of the most fascinating aspects of the algorithm [*A*.24] is its universality; that is, the fact that an arbitrary conditional distribution q with support $\mathcal{E}$ can lead to the simulation of an arbitrary distribution f on $\mathcal{E}$. On the other hand, this universality may only hold formally if the instrumental distribution q rarely simulates points in the main portion of $\mathcal{E}$; that is to say, in the region

where most of the mass of the density f is located. This issue of selecting a good proposal distribution q for a given f is detailed in Section 7.6.

Since we have provided no examples so far, we now proceed to describe two particular approaches used in the literature, with some probabilistic properties and corresponding examples. Note that a complete classification of the Metropolis–Hastings algorithms is impossible, given the versatility of the method and the possibility of creating even more hybrid methods (see, for instance, Roberts and Tweedie 1995, 2004 and Stramer and Tweedie 1999b).

7.4 The Independent Metropolis–Hastings Algorithm

7.4.1 Fixed Proposals

This method appears as a straightforward generalization of the Accept–Reject method in the sense that the instrumental distribution q is independent of $X^{(t)}$ and is denoted g by analogy. The algorithm [A.24] will then produce the following transition from $x^{(t)}$ to $X^{(t+1)}$.

Algorithm A.25 –Independent Metropolis–Hastings–

Given $x^{(t)}$

1 Generate $Y_t \sim g(y)$.

2 Take [A.25]

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \min\left\{\dfrac{f(Y_t)\,g(x^{(t)})}{f(x^{(t)})\,g(Y_t)}, 1\right\} \\ x^{(t)} & \text{otherwise.} \end{cases}$$

Although the Y_t's are generated independently, the resulting sample is not iid: for instance, the probability of acceptance of Y_t depends on $X^{(t)}$ (except in the trivial case when $f = g$).

The convergence properties of the chain $(X^{(t)})$ follow from properties of the density g in the sense that $(X^{(t)})$ is irreducible and aperiodic (thus, ergodic according to Corollary 7.5) if and only if g is almost everywhere positive on the support of f. Stronger properties of convergence like geometric and uniform ergodicity are also clearly described by the following result of Mengersen and Tweedie (1996).

Theorem 7.8. *The algorithm* [A.25] *produces a uniformly ergodic chain if there exists a constant* M *such that*

$$f(x) \le Mg(x)\,, \quad \forall x \in \text{supp}\, f. \tag{7.7}$$

In this case,

$$\|K^n(x,\cdot) - f\|_{TV} \leq 2\left(1 - \frac{1}{M}\right)^n , \tag{7.8}$$

where $||\cdot||_{TV}$ *denotes the total variation norm introduced in Definition 6.47. On the other hand, if for every* M*, there exists a set of positive measure where (7.7) does not hold,* $(X^{(t)})$ *is not even geometrically ergodic.*

Proof. If (7.7) is satisfied, the transition kernel satisfies

$$\begin{aligned} K(x,x') &\geq g(x') \min\left\{\frac{f(x')g(x)}{f(x)g(x')}, 1\right\} \\ &= \min\left\{f(x')\frac{g(x)}{f(x)}, g(x')\right\} \geq \frac{1}{M} f(x') . \end{aligned}$$

The set $\mathcal{E}$ is therefore small and the chain is uniformly ergodic (Theorem 6.59).

To establish the bound on $\|K^n(x,\cdot) - f\|_{TV}$, first write

$$\begin{aligned} \|K(x,\cdot) - f\|_{TV} &= 2\sup_A \left|\int_A (K(x,y) - f(y))dy\right| \\ &= 2\int_{\{y;f(y)\geq K(x,y)\}} (f(y) - K(x,y))dy \\ &\leq 2\left(1 - \frac{1}{M}\right)\int_{\{y;f(y)\geq K(x,y)\}} f(y)dy \\ &\leq 2\left(1 - \frac{1}{M}\right). \end{aligned} \tag{7.9}$$

We now continue with a recursion argument to establish (7.8). We can write

$$\begin{aligned} \int_A (K^2(x,y) - f(y))dy = \int_{\mathcal{E}} \left[\int_A (K(u,y) - f(y))dy\right] \\ \times (K(x,u) - f(u))du, \end{aligned} \tag{7.10}$$

and an argument like that in (7.9) leads to

$$\|K^2(x,\cdot) - f\|_{TV} \leq 2\left(1 - \frac{1}{M}\right)^2 . \tag{7.11}$$

We next write a general recursion relation

$$\begin{aligned} &\int_A (K^{n+1}(x,y) - f(y))dy \\ &\quad = \int_{\mathcal{E}} \left[\int_A (K^n(u,y) - f(y))dy\right] (K(x,u) - f(u))du, \end{aligned} \tag{7.12}$$

and proof of (7.8) is established by induction (Problem 7.11).

If (7.7) does not hold, then the sets

$$D_n = \left\{x; \frac{f(x)}{g(x)} \geq n\right\}$$

satisfy $P_f(D_n) > 0$ for every n. If $x \in D_n$, then

$$\begin{aligned} P(x,\{x\}) &= 1 - \mathbb{E}_g\left[\min\left\{\frac{f(Y)g(x)}{g(Y)f(x)}, 1\right\}\right] \\ &= 1 - P_g\left(\frac{f(Y)}{g(Y)} \geq \frac{f(x)}{g(x)}\right) - \mathbb{E}_g\left[\frac{f(Y)g(x)}{f(x)g(Y)}\, \mathbb{I}_{\frac{f(Y)}{g(Y)} < \frac{f(x)}{g(x)}}\right] \\ &\geq 1 - P_g\left(\frac{f(Y)}{g(Y)} \geq n\right) - \frac{g(x)}{f(x)} \geq 1 - \frac{2}{n}, \end{aligned}$$

since Markov inequality implies that

$$P_g\left(\frac{f(Y)}{g(Y)} \geq n\right) \leq \frac{1}{n}\mathbb{E}_g\left[\frac{f(Y)}{g(Y)}\right] = \frac{1}{n}.$$

Consider a small set C such that $D_n \cap C^c$ is not empty for n large enough and $x_0 \in D_n \cap C^c$. The return time to C, τ_C, satisfies

$$P_{x_0}(\tau_C > k) \geq \left(1 - \frac{2}{n}\right)^k;$$

therefore, the radius of convergence of the series (in κ) $\mathbb{E}_{x_0}[\kappa^{\tau_C}]$ is smaller than $n/(n-2)$ for every n, and this implies that $(X^{(t)})$ cannot be geometrically ergodic, according to Theorem 6.75. □

This particular class of Metropolis–Hastings algorithms naturally suggests a comparison with Accept–Reject methods since every pair (f, g) satisfying (7.7) can also induce an Accept–Reject algorithm. Note first that the expected acceptance probability for the variable simulated according to g is larger in the case of the algorithm [*A*.25].

Lemma 7.9. *If (7.7) holds, the expected acceptance probability associated with the algorithm* [*A*.25] *is at least* $\frac{1}{M}$ *when the chain is stationary.*

Proof. If the distribution of $f(X)/g(X)$ is absolutely continuous,[3] the expected acceptance probability is

[3] This constraint implies that f/g is not constant over some interval.

$$\begin{aligned}
\mathbb{E}\left[\min\left\{\frac{f(Y_t)g(X^{(t)})}{f(X^{(t)})g(Y_t)}\,,1\right\}\right] &= \int \mathbb{I}_{\frac{f(y)g(x)}{g(y)f(x)}>1}\, f(x)g(y)\,dxdy \\
&\quad + \int \frac{f(y)g(x)}{g(y)f(x)}\;\mathbb{I}_{\frac{f(y)g(x)}{g(y)f(x)}\leq 1}\, f(x)g(y)\,dxdy \\
&= 2\int \mathbb{I}_{\frac{f(y)g(x)}{g(y)f(x)}\geq 1}\, f(x)g(y)\,dxdy \\
&\geq 2\int \mathbb{I}_{\frac{f(y)}{g(y)}\geq\frac{f(x)}{g(x)}}\, f(x)\,\frac{f(y)}{M}\,dxdy \\
&= \frac{2}{M}\,P\left(\frac{f(X_1)}{g(X_1)}\geq\frac{f(X_2)}{g(X_2)}\right) = \frac{1}{M}\,.
\end{aligned}$$

Since X_1 and X_2 are independent and distributed according to f, this last probability is equal to $1/2$, and the result follows. □

Thus, the independent Metropolis–Hastings algorithm [*A*.25] is more efficient than the Accept–Reject algorithm [*A*.4] in its handling of the sample produced by g, since, on the average, it accepts more proposed values. A more advanced comparison between these approaches is about as difficult as the comparison between Accept–Reject and importance sampling proposed in Section 3.3.3, namely that the size of one of the two samples is random and this complicates the computation of the variance of the resulting estimator. In addition, the correlation between the X_i's resulting from [*A*.25] prohibits a closed-form expression of the joint distribution. We therefore study the consequence of the correlation on the variance of both estimators through an example. (See also Liu 1996b and Problem 7.33 for a comparison based on the eigenvalues of the transition operators in the discrete case, which also shows the advantage of the Metropolis–Hastings algorithm.)

Example 7.10. Generating gamma variables. Using the algorithm of Example 2.19 (see also Example 3.15), an Accept–Reject method can be derived to generate random variables from the $\mathcal{G}a(\alpha,\beta)$ distribution using a Gamma $\mathcal{G}a(\lfloor\alpha\rfloor,b)$ candidate (where $\lfloor a\rfloor$ denotes the integer part of a). When $\beta=1$, the optimal choice of b is

$$b = \lfloor\alpha\rfloor/\alpha\,.$$

The algorithms to compare are then

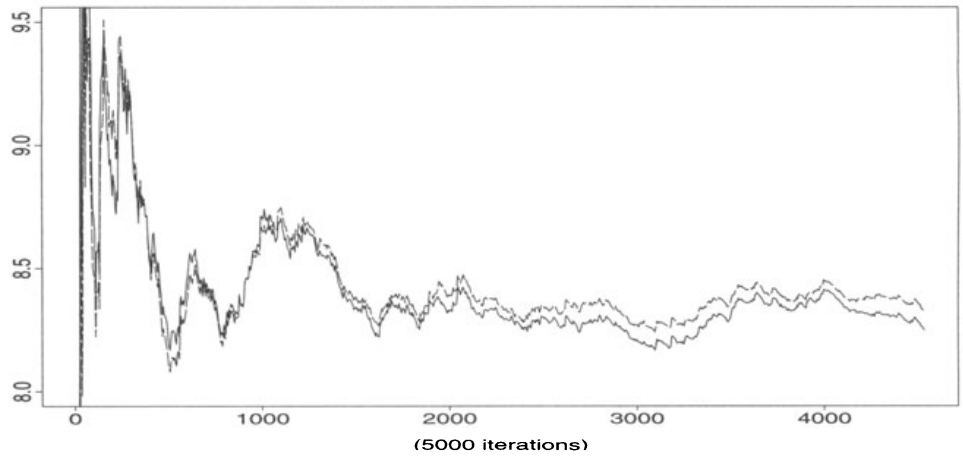

Fig. 7.1. Convergence of Accept–Reject (solid line) and Metropolis–Hastings (dashed line) estimators to $\mathbb{E}_f[X^2] = 8.33$, for $\alpha = 2.43$ based on the same sequence $y_1, \ldots, y_{5000}$ simulated from $\mathcal{G}a(2, 2/2.43)$. The number of acceptances in [*A*.27] is then random. The final values of the estimators are 8.25 for [*A*.27] and 8.32 for [*A*.26].

Algorithm A.26 –Gamma Metropolis–Hastings–

1. Generate $Y_t \sim \mathcal{G}a(\lfloor\alpha\rfloor, \lfloor\alpha\rfloor/\alpha)$.
2. Take [*A*.26]

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \varrho_t \\ x^{(t)} & \text{otherwise,} \end{cases}$$

where

$$\varrho_t = \min\left[\left(\frac{Y_t}{x^{(t)}}\,\exp\left\{\frac{x^{(t)} - Y_t}{\alpha}\right\}\right)^{\alpha-\lfloor\alpha\rfloor}, 1\right].$$

and

Algorithm A.27 –Gamma Accept–Reject–

1. Generate $Y \sim \mathcal{G}a(\lfloor\alpha\rfloor, \lfloor\alpha\rfloor/\alpha)$.
2. Accept $X = Y$ with probability [*A*.27]

$$\left(\frac{ey\exp(-y/\alpha)}{\alpha}\right)^{\alpha-\lfloor\alpha\rfloor}$$

Note that (7.7) does apply in this particular case with $\exp(x/\alpha)/x > e/\alpha$.

A first comparison is based on a sample $(y_1, \ldots, y_n)$, of fixed size n, generated from $\mathcal{G}a(\lfloor\alpha\rfloor, \lfloor\alpha\rfloor/\alpha)$ with $x^{(0)}$ generated from $\mathcal{G}a(\alpha, 1)$. The number

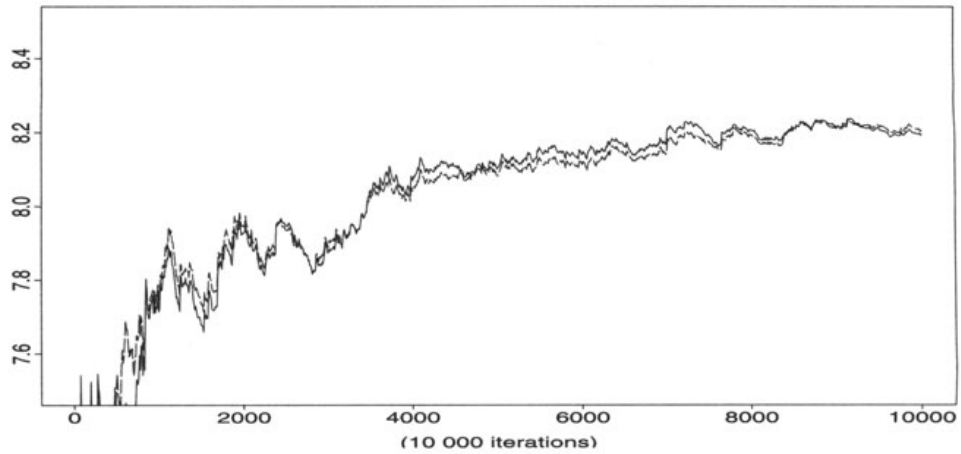

Fig. 7.2. Convergence to $\mathbb{E}_f[X^2] = 8.33$ of Accept–Reject (full line) and Metropolis–Hastings (dots) estimators for 10,000 acceptances in [*A*.27], the same sequence of y_i's simulated from $\mathcal{G}a(2, 2/2.43)$ being used in [*A*.27] and [*A*.26]. The final values of the estimators are 8.20 for [*A*.27] and 8.21 for [*A*.26].

t of values accepted by [*A*.27] is then random. Figure 7.1 describes the convergence of the estimators of $\mathbb{E}_f[X^2]$ associated with both algorithms for the same sequence of y_i's and exhibits strong agreement between the approaches, with the estimator based on [*A*.26] being closer to the exact value 8.33 in this case.

On the other hand, the number t of values accepted by [*A*.27] can be fixed and [*A*.26] can then use the resulting sample of random size n, $y_1, \ldots, y_n$. Figure 7.2 reproduces the comparison in this second case and exhibits a behavior rather similar to Figure 7.1, with another close agreement between estimators and, the scale being different, a smaller variance (which is due to the larger size of the effective sample).

Note, however, that both comparisons are biased. In the first case, the sample of $X^{(i)}$ produced by [*A*.27] does not have the distribution f and, in the second case, the sample of Y_i's in [*A*.26] is not iid. In both cases, this is due to the use of a stopping rule which modifies the distribution of the samples.||

Example 7.11. Logistic Regression. We return to the data of Example 1.13, which described a logistic regression relating the failure of O-rings in shuttle flights to air temperature. We observe (x_i, y_i), $i = 1, \ldots, n$ according to the model

$$Y_i \sim \text{Bernoulli}(p(x_i)), \quad p(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

where $p(x)$ is the probability of an O-ring failure at temperature x. The likelihood is

$$L(\alpha, \beta|\mathbf{y}) \propto \prod_{i=1}^{n} \left(\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(\alpha + \beta x_i)} \right)^{1-y_i}$$

and we take the prior to be

$$\pi_\alpha(\alpha|b)\pi_\beta(\beta) = \frac{1}{b} e^{\alpha} e^{-e^{\alpha}/b} d\alpha d\beta,$$

which puts an exponential prior on $\log \alpha$ and a flat prior on β, and insures propriety of the posterior distribution (Problem 7.25). To complete the prior specification we must give a value for b, and we choose the data-dependent value that makes $\mathbb{E}\alpha = \hat{\alpha}$, where $\hat{\alpha}$ is the MLE of α. (This also insures that the prior will not have undue influence, as it is now centered near the likelihood.) It can be shown that

$$\mathbb{E}[\alpha] = \int_0^\infty \frac{1}{b} e^{\alpha} e^{-e^{\alpha}/b} d\alpha = \int_0^\infty \log(w) \frac{1}{b} e^{-w/b} dw = \log(b) - \gamma,$$

where γ is *Euler's Constant*, equal to .577216. Thus we take $\hat{b} = e^{\hat{\alpha}+\gamma}$.

The posterior distribution is proportional to $L(\alpha, \beta|\mathbf{y})\pi(\alpha, \beta)$, and to simulate from this distribution we take an independent candidate

$$g(\alpha, \beta) = \pi_\alpha(\alpha|\hat{b})\phi(\beta),$$

where $\phi(\beta)$ is a normal distribution with mean $\hat{\beta}$ and variance $\hat{\sigma}_\beta^2$, the MLEs. Note that although basing the prior distribution on the data is somewhat in violation of the formal Bayesian paradigm, nothing is violated if the candidate depends on the data. In fact, this will usually result in a more effective simulation, as the candidate is placed close to the target.

Generating a random variable from $g(\alpha, \beta)$ is straightforward, as it only involves the generation of a normal and an exponential random variable. If we are at the point (α_0, β_0) in the Markov chain, and we generate (α', β') from $g(\alpha, \beta)$, we accept the candidate with probability

$$\min \left\{ \frac{L(\alpha', \beta'|\mathbf{y})}{L(\alpha_0, \beta_0|\mathbf{y})} \frac{\phi(\beta_0)}{\phi(\beta')}, 1 \right\}.$$

Figure 7.3 shows the distribution of the generated parameters and their convergence. ||

Example 7.12. Saddlepoint tail area approximation. In Example 3.18, we saw an approximation to noncentral chi squared tail areas based on the regular and renormalized saddlepoint approximations. Such an approximation requires numerical integration, both to calculate the constant and to evaluate the tail area.

An alternative is to produce a sample $Z_1, \ldots, Z_m$, from the saddlepoint distribution, and then approximate the tail area using

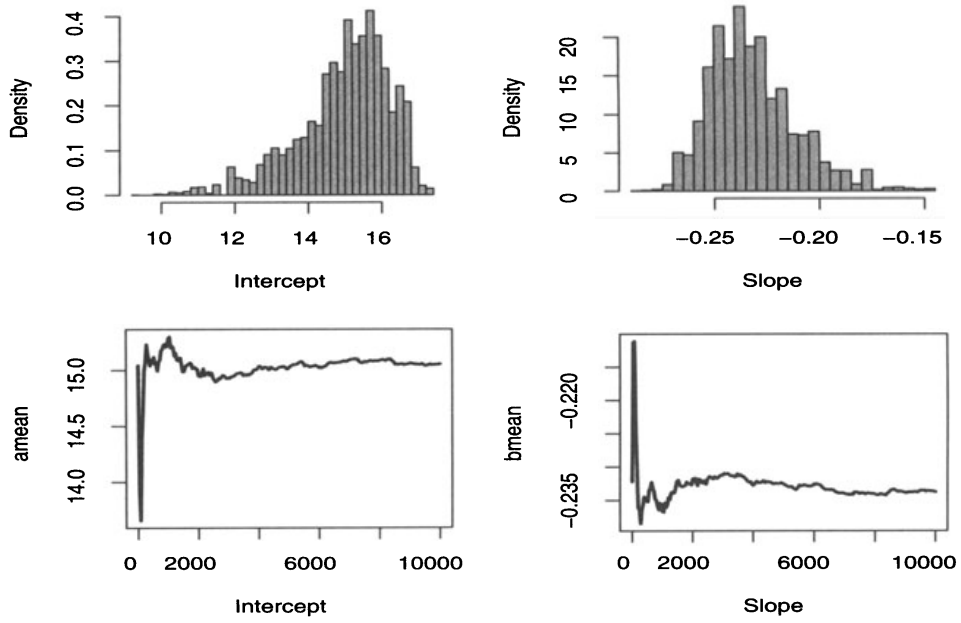

Fig. 7.3. Estimation of the slope and intercept from the Challenger logistic regression. The top panels show histograms of the distribution of the coefficients, while the bottom panels show the convergence of the means.

$$P(\bar{X} > a) = \int_{\hat{\tau}(a)}^{1/2} \left(\frac{n}{2\pi}\right)^{1/2} [K_X''(t)]^{1/2} \exp\{n[K_X(t) - tK_X'(t)]\}\, dt$$

$$\approx \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}[Z_i > \hat{\tau}(a)]\,, \tag{7.13}$$

where $K_X(\tau)$ is the cumulant generating function of X and $\hat{\tau}(x)$ is the solution of the saddlepoint equation $K'(\hat{\tau}(x)) = x$ (see Section 3.6.2).

Note that we are simulating from the transformed density. It is interesting (and useful) to note that we can easily derive an instrumental density to use in a Metropolis–Hastings algorithm. Using a Taylor series approximation, we find that

$$\exp\{n[K_X(t) - tK_X'(t)]\} \approx \exp\left\{-nK_X''(0)\frac{t^2}{2}\right\}\,, \tag{7.14}$$

so a first choice for an instrumental density is the $\mathcal{N}(0, 1/nK_X''(0))$ distribution (see Problem 7.26 for details). Booth et al. (1999) use a Student's t approximation instead.

We can now simulate the noncentral chi squared tail areas using a normal instrumental density with $K_X''(t) = 2[p(1-2t) + 4\lambda]/(1-2t)^3$. The results are presented in Table 7.1, where we see that the approximations are quite good. Note that the same set of simulated random variables can be used for all the tail area probability calculations. Moreover, by using the Metropolis–

Hastings algorithm, we have avoided calculating the normalizing constant for the saddlepoint approximation. ||

Interval	Renormalized	Exact	Monte Carlo
$(36.225, \infty)$	0.0996	0.1	0.0992
$(40.542, \infty)$	0.0497	0.05	0.0497
$(49.333, \infty)$	0.0099	0.01	0.0098

Table 7.1. Monte Carlo saddlepoint approximation of a noncentral chi squared integral for $p = 6$ and $\lambda = 9$, based on $10,000$ simulated random variables.

As an aside, note that the usual classification of "Hastings" for the algorithm [*A*.25] is somewhat inappropriate, since Hastings (1970) considers the algorithm [*A*.24] in general, using random walks (Section 7.5) rather than independent distributions in his examples. It is also interesting to recall that Hastings (1970) proposes a theoretical justification of these methods for finite state-space Markov chains based on the finite representation of real numbers in a computer. However, a complete justification of this physical discretization needs to take into account the effect of the approximation in the entire analysis. In particular, it needs to be verified that the computer choice of discrete approximation to the continuous distribution has no effect on the resulting stationary distribution or irreducibility of the chain. Since Hastings (1970) does not go into such detail, but keeps to the simulation level, we prefer to study the theoretical properties of these algorithms by bypassing the finite representation of numbers in a computer and by assuming flawless pseudo-random generators, namely algorithms producing variables which are uniformly distributed on $[0, 1]$. See Roberts et al. (1995) for a theoretical study of some effects of the computer discretization.

A final note about independent Metropolis–Hastings algorithms is that they cannot be omniscient: there are settings where an independent proposal does not work well because of the complexity of the target distribution. Since the main purpose of MCMC algorithms is to provide a crude but easy simulation technique, it is difficult to imagine spending a long time on the design of the proposal distribution. This is specially pertinent in high-dimensional models where the capture of the main features of the target distribution is most often impossible. There is therefore a limitation of the independent proposal, which can be perceived as a *global* proposal, and a need to use more *local* proposals that are not so sensitive to the target distribution, as presented in Section 7.5. Another possibility, developed in Section 7.6.3, is to validate *adaptive* algorithms that learn from the ongoing performances of the current proposals to refine their construction. But this solution is delicate, both from a theoretical (*"Does ergodicity apply?"*) and an algorithmic (*"How does one*

tune the adaptation?") point of view. The following section first develops a specific kind of adaptive algorithm.

7.4.2 A Metropolis–Hastings Version of ARS

The ARS algorithm, which provides a general Accept–Reject method for log-concave densities in dimension one (see Section 2.4.2), can be generalized to the ARMS method (which stands for *Adaptive Rejection Metropolis Sampling*) following the approach developed by Gilks et al. (1995). This generalization applies to the simulation of arbitrary densities, instead of being restricted to log-concave densities as the ARS algorithm, by simply adapting the ARS algorithm for densities f that are not log-concave. The algorithm progressively fits a function g, which plays the role of a pseudo-envelope of the density f. In general, this function g does not provide an upper bound on f, but the introduction of a Metropolis–Hastings step in the algorithm justifies the procedure.

Using the notation from Section 2.4.2, take $h(x) = \log f_1(x)$ with f_1 proportional to the density f. For a sample $S_n = \{x_i, 0 \leq i \leq n+1\}$, the equations of the lines between $(x_i, h(x_i))$ and $(x_{i+1}, h(x_{i+1}))$ are denoted by $y = L_{i,i+1}(x)$. Consider

$$\tilde{h}_n(x) = \max\{L_{i,i+1}(x), \min[L_{i-1,i}(x), L_{i+1,i+2}(x)]\} ,$$

for $x_i \leq x < x_{i+1}$, with

$$\begin{aligned}
& \tilde{h}_n(x) = L_{0,1}(x) && \text{if} \quad x < x_0, \\
& \tilde{h}_n(x) = \max[L_{0,1}(x), L_{1,2}(x)] && \text{if} \quad x_0 \leq x < x_1, \\
& \tilde{h}_n(x) = \max[L_{n,n+1}(x), L_{n-1,n}(x)] && \text{if} \quad x_n \leq x < x_{n+1}, \\
\text{and} \quad & \tilde{h}_n(x) = L_{n,n+1}(x) && \text{if} \quad x \geq x_{n+1}.
\end{aligned}$$

The resulting proposal distribution is $g_n(x) \propto \exp\{\tilde{h}_n(x)\}$. The ARMS algorithm is based on g_n and it can be decomposed into two parts, a first step which is a standard Accept–Reject step for the simulation from the instrumental distribution

$$\psi_n(x) \propto \min\left[f_1(x), \exp\{\tilde{h}_n(x)\}\right] ,$$

based on g_n, and a second part, which is the acceptance of the simulated value by a Metropolis–Hastings procedure:

Algorithm A.28 –ARMS Metropolis–Hastings–

1. Simulate Y from $g_n(y)$ and $U \sim \mathcal{U}_{[0,1]}$ until
$$U \leq f_1(Y)/\exp\{\tilde{h}_n(Y)\}.$$
2. Generate $V \sim \mathcal{U}_{[0,1]}$ and take [A.28]

$$X^{(t+1)} = \begin{cases} Y & \text{if } V < \dfrac{f_1(Y)\,\psi_n(x^{(t)})}{f_1(x^{(t)})\,\psi_n(Y)} \wedge 1 \\ x^{(t)} & \text{otherwise.} \end{cases} \tag{7.15}$$

The Accept–Reject step indeed produces a variable distributed from $\psi_n(x)$ and this justifies the expression of the acceptance probability in the Metropolis–Hastings step. Note that [A.28] is a particular case of the approximate Accept–Reject algorithms considered by Tierney (1994) (see Problem 7.9). The probability (7.15) can also be written

$$\begin{cases} \min\left[1, \dfrac{f_1(Y)\exp\{\tilde{h}_n(x^{(t)})\}}{f_1(x^{(t)})\exp\{\tilde{h}_n(Y)\}}\right] & \text{if } f_1(Y) > \exp\{\tilde{h}_n(Y)\}, \\ \min\left[1, \dfrac{\exp\{\tilde{h}_n(x^{(t)})\}}{f_1(x^{(t)})}\right] & \text{otherwise,} \end{cases}$$

which implies a sure acceptance of Y when $f_1(x^{(t)}) < \exp\{\tilde{h}_n(x^{(t)})\}$; that is, when the bound is correct.

Each simulation of $Y \sim g_n$ in Step 1 of [A.28] provides, in addition, an update of S_n in $S_{n+1} = S_n \cup \{y\}$, and therefore of g_n, when Y is rejected. As in the case of the ARS algorithm, the initial S_n set must be chosen so that g_n is truly a probability density. If the support of f is not bounded from below, $L_{0,1}$ must be increasing and, similarly, if the support of f is not bounded from above, $L_{n,n+1}$ must be decreasing. Note also that the simulation of g_n detailed in Section 2.4.2 is valid in this setting.

Since the algorithm [A.28] appears to be a particular case of independent Metropolis–Hastings algorithm, the convergence and ergodicity results obtained in Section 7.4 should apply for [A.28]. This is not the case, however, because of the lack of time homogeneity of the chain (see Definition 6.4) produced by [A.28]. The transition kernel, based on g_n, can change at each step with a positive probability. Since the study of nonhomogeneous chains is quite delicate, the algorithm [A.28] can be justified only by reverting to the homogeneous case; that is, by fixing the function g_n and the set S_n after a warm-up period of length n_0. The constant n_0 need not be fixed in advance as this warm-up period can conclude when the approximation of f_1 by g_n is satisfactory, for instance when the rejection rate in Step 1 of [A.28] is sufficiently small. The algorithm [A.28] must then start with an initializing (or

calibrating) step which adapts the parameters at hand (in this case, g_n) to the function f_1. This adaptive structure is generalized in Section 7.6.3.

The ARMS algorithm is useful when a precise analytical study of the density f is impossible, as, for instance, in the setup of generalized linear models. In fact, f (or f_1) needs to be computed in only a few points to initialize the algorithm, which thus does not require the search for "good" density g which approximates f. This feature should be contrasted to the cases of the independent Metropolis–Hastings algorithm and of sufficiently fast random walks as in the case of [*A*.29].

Example 7.13. Poisson logistic model. For the generalized linear model in Example 2.26, consider a logit dependence between explanatory and dependent (observations) variables,

$$Y_i|x_i \sim \mathcal{P}\left(\frac{\exp(bx_i)}{1+\exp(bx_i)}\right), \qquad i = 1, \ldots, n,$$

which implies the restriction $\lambda_i < 1$ on the parameters of the Poisson distribution, $Y_i \sim \mathcal{P}(\lambda_i)$. When b has the prior distribution $\mathcal{N}(0, \tau^2)$, the posterior distribution is

$$\pi(b|\mathbf{x}, \mathbf{y}) \propto \frac{\exp\{\sum_i y_i(bx_i)\}}{\prod_i(1+\exp(bx_i))} \exp\left\{-\sum_i \frac{e^{bx_i}}{1+e^{bx_i}}\right\} e^{-b^2/2\tau^2}.$$

This posterior distribution $\pi(b|\mathbf{x})$ is not easy to simulate from and one can use the ARMS Metropolis–Hastings algorithm instead. ||

7.5 Random Walks

A natural approach for the practical construction of a Metropolis–Hastings algorithm is to take into account the value previously simulated to generate the following value; that is, to consider a *local* exploration of the neighborhood of the current value of the Markov chain. This idea is already used in algorithms such as the simulated annealing algorithm [*A*.19] and the stochastic gradient method given in (5.4).

Since the candidate g in algorithm [*A*.24] is allowed to depend on the current state $X^{(t)}$, a first choice to consider is to simulate Y_t according to

$$Y_t = X^{(t)} + \varepsilon_t,$$

where ε_t is a random perturbation with distribution g, independent of $X^{(t)}$. In terms of the algorithm [*A*.24], $q(y|x)$ is now of the form $g(y-x)$. The Markov chain associated with q is a *random walk* (see Example 6.39) on $\mathcal{E}$.

δ	0.1	0.5	1.0
Mean	0.399	−0.111	0.10
Variance	0.698	1.11	1.06

Table 7.2. Estimators of the mean and the variance of a normal distribution $\mathcal{N}(0,1)$ based on a sample obtained by a Metropolis–Hastings algorithm using a random walk on $[-\delta,\delta]$ (15,000 simulations).

The convergence results of Section 7.3.2 naturally apply in this particular case. Following Lemma 7.6, if g is positive in a neighborhood of 0, the chain $(X^{(t)})$ is f-irreducible and aperiodic, therefore ergodic. The most common distributions g in this setup are the uniform distributions on spheres centered at the origin or standard distributions like the normal and the Student's t distributions. All these distributions usually need to be scaled; we discuss this problem in Section 7.6. At this point, we note that the choice of a *symmetric function* g (that is, such that $g(-t) = g(t)$), leads to the following original expression of [*A*.24], as proposed by Metropolis et al. (1953).

Algorithm A.29 –Random walk Metropolis–Hastings–

Given $x^{(t)}$,

1. Generate $Y_t \sim g(|y - x^{(t)}|)$.
2. Take [*A*.29]

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \min\left\{1, \dfrac{f(Y_t)}{f(x^{(t)})}\right\} \\ x^{(t)} & \text{otherwise.} \end{cases}$$

Example 7.14. A random walk normal generator. Hastings (1970) considers the generation of the normal distribution $\mathcal{N}(0,1)$ based on the uniform distribution on $[-\delta,\delta]$. The probability of acceptance is then $\rho(x^{(t)}, y_t) = \exp\{({x^{(t)}}^2 - y_t^2)/2\} \wedge 1$. Figure 7.4 describes three samples of 15,000 points produced by this method for $\delta = 0.1, 0.5$, and 1. The corresponding estimates of the mean and variance are provided in Table 7.2. Figure 7.4 clearly shows the different speeds of convergence of the averages associated with these three values of δ, with an increasing regularity (in δ) of the corresponding histograms and a faster exploration of the support of f. ||

Despite its simplicity and its natural features, the random walk Metropolis–Hastings algorithm does not enjoy uniform ergodicity properties. Mengersen and Tweedie (1996) have shown that in the case where supp $f = \mathbb{R}$, this algorithm cannot produce a uniformly ergodic Markov chain on $\mathbb{R}$ (Problem 7.16). This is a rather unsurprising feature when considering the *local* character of

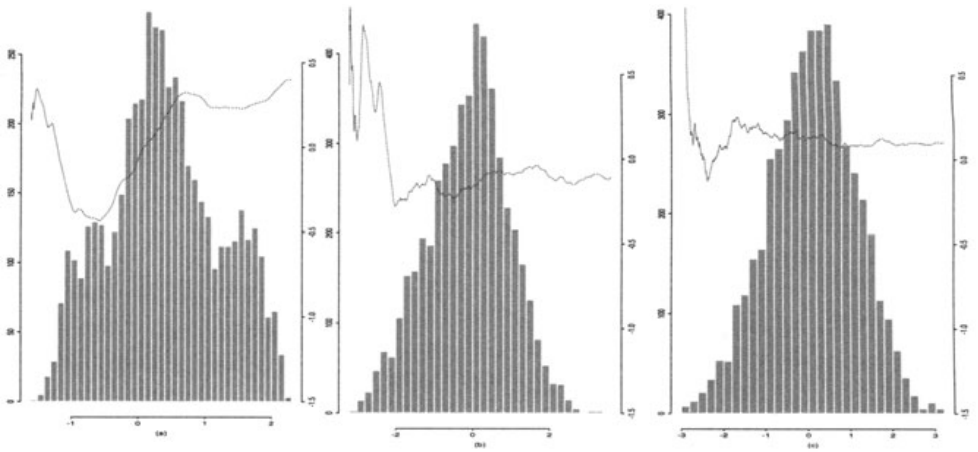

Fig. 7.4. Histograms of three samples produced by the algorithm [A.29] for a random walk on $[-\delta, \delta]$ with (a) $\delta = 0.1$, (b) $\delta = 0.5$, and (c) $\delta = 1.0$, with the convergence of the means (7.1), superimposed with scales on the right of the graphs (15,000 simulations).

the random walk proposal, centered at the current value of the Markov chain.

Although uniform ergodicity cannot be obtained with random walk Metropolis–Hastings algorithms, it is possible to derive necessary and sufficient conditions for geometric ergodicity. Mengersen and Tweedie (1996) have proposed a condition based on the *log-concavity of* f in the tails; that is, if there exist $\alpha > 0$ and x_1 such that

$$\log f(x) - \log f(y) \geq \alpha |y - x| \tag{7.16}$$

for $y < x < -x_1$ or $x_1 < x < y$.

Theorem 7.15. *Consider a symmetric density f which is log-concave with associated constant α in (7.16) for $|x|$ large enough. If the density g is positive and symmetric, the chain $(X^{(t)})$ of [A.29] is geometrically ergodic. If f is not symmetric, a sufficient condition for geometric ergodicity is that $g(t)$ be bounded by $b\exp\{-\alpha|t|\}$ for a sufficiently large constant b.*

The proof of this result is based on the use of the drift function $V(x) = \exp\{\alpha|x|/2\}$ (see Note 6.9.1) and the verification of a geometric drift condition of the form

$$\Delta V(x) \leq -\lambda V(x) + b\mathbb{I}_{[-x^*,x^*]}(x)\ , \tag{7.17}$$

for a suitable bound x^*. Mengersen and Tweedie (1996) have shown, in addition, that this condition on g is also necessary in the sense that if $(X^{(t)})$ is geometrically ergodic, there exists $s > 0$ such that

$$\int e^{s|x|}\, f(x)dx < \infty\ . \tag{7.18}$$

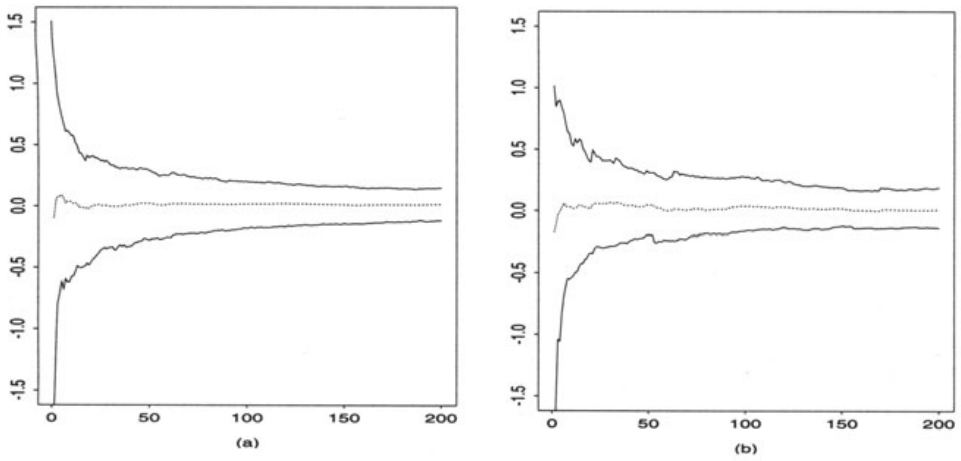

Fig. 7.5. 90% confidence envelopes of the means produced by the random walk Metropolis–Hastings algorithm [*A*.24] based on a instrumental distribution $\mathcal{N}(0,1)$ for the generation of (a) a normal distribution $\mathcal{N}(0,1)$ and (b) a distribution with density ψ. These envelopes are derived from 500 parallel independent chains and with identical uniform samples on both distributions.

Example 7.16. A comparison of tail effects. In order to assess the practical effect of this theorem, Mengersen and Tweedie (1996) considered two random walk Metropolis–Hastings algorithms based on a $\mathcal{N}(0,1)$ instrumental distribution for the generation of (a) a $\mathcal{N}(0,1)$ distribution and (b) a distribution with density $\psi(x) \propto (1+|x|)^{-3}$. Applying Theorem 7.15 (see Problem 7.18), it can be shown that the first chain associated is geometrically ergodic, whereas the second chain is not. Figures 7.5(a) and 7.5(b) represent the average behavior of the sums

$$\frac{1}{T}\sum_{t=1}^{T} X^{(t)}$$

over 500 chains initialized at $x^{(0)} = 0$. The 5% and 95% quantiles of these chains show a larger variability of the chain associated with the distribution ψ, in terms of both width of the confidence region and precision of the resulting estimators. ‖

We next look at a discrete example where Algorithm [*A*.29] generates a geometrically ergodic chain.

Example 7.17. Random walk geometric generation. Consider generating a geometric[4] distribution, $\mathcal{G}eo(\theta)$ using [*A*.29] with $(Y^{(t)})$ having transition probabilities $q(i,j) = P(Y^{(t+1)} = j | Y^{(t)} = i)$ given by

[4] The material used in the current example refers to the drift condition introduced in Note 6.9.1.

$$q(i,j) = \begin{cases} 1/2 & i = j-1, j+1 \text{ and } j = 1,2,3,\ldots \\ 1/2 & i = 0,1 \text{ and } j = 0 \\ 0 & \text{otherwise;} \end{cases}$$

that is, q is the transition kernel of a symmetric random walk on the non-negative integers *with reflecting boundary at* 0.

Now, $X \sim \mathcal{G}eo(\theta)$ implies $P(X = x) = (1-\theta)^x\theta$ for $x = 0,1,2,\ldots$. The transition matrix has a band diagonal structure and is given by

$$T = \begin{pmatrix} \frac{1+\theta}{2} & \frac{1-\theta}{2} & 0 & 0 & \cdots \\ \frac{1}{2} & \frac{\theta}{2} & \frac{1-\theta}{2} & 0 & \cdots \\ 0 & \frac{1}{2} & \frac{\theta}{2} & \frac{1-\theta}{2} & \cdots \\ & \ddots & \ddots & \ddots & \cdots \end{pmatrix} .$$

Consider the potential function $V(i) = \beta^i$ where $\beta > 1$, and recall that $\Delta V(y^{(0)}) = \mathbb{E}[V(Y^{(1)})|y^{(0)}] - V(y^{(0)})$. For $i > 0$, we have

$$\begin{aligned} \mathbb{E}[V(Y^{(1)})|Y^{(0)} = i] &= \frac{1}{2}\,\beta^{i-1} + \frac{\theta}{2}\,\beta^i + \frac{1-\theta}{2}\,\beta^{i+1} \\ &= V(i)\left(\tfrac{1}{2\beta} + \tfrac{\theta}{2} + \tfrac{1-\theta}{2}\beta\right) . \end{aligned}$$

Thus, $\Delta V(i) = V(i)(1/(2\beta) + \theta/2 - 1 + \beta(1-\theta)/2) = V(i)g(\theta,\beta)$. For a fixed value of θ, $g(\theta,\beta)$ is minimized by $\beta = 1/\sqrt{1-\theta}$. In this case, $\Delta V(i) = (\sqrt{1-\theta} + \theta/2 - 1)V(i)$ and $\lambda = \sqrt{1-\theta} + \theta/2 - 1$ is the geometric rate of convergence. The closer θ is to 1, the faster the convergence. ||

Tierney (1994) proposed a modification of the previous algorithm with a proposal density of the form $g(y - a - b(x-a))$; that is,

$$y_t = a + b(x^{(t)} - a) + z_t\,, \qquad z_t \sim g\,.$$

This autoregressive representation can be seen as intermediary between the independent version ($b = 0$) and the random walk version ($b = 1$) of the Metropolis–Hastings algorithm. Moreover, when $b < 0$, $X^{(t)}$ and $X^{(t+1)}$ are negatively correlated, and this may allow for faster excursions on the surface of f if the symmetry point a is well chosen. Hastings (1970) also considers an alternative to the uniform distribution on $[x^{(t)} - \delta, x^{(t)} + \delta]$ (see Example 7.14) with the uniform distribution on $[-x^{(t)} - \delta, -x^{(t)} + \delta]$: The convergence of the empirical average to 0 is then faster in this case, but the choice of 0 as center of symmetry is obviously crucial and requires some a priori information on the distribution f. In a general setting, a and b can be calibrated during the first iterations. (See also Problem 7.23.) (See also Chen and Schmeiser 1993, 1998 for the alternative "hit-and-run" algorithm, which proceeds by generating a random direction in the space and moves the current value by a random distance along this direction.)

7.6 Optimization and Control

The previous sections have established the theoretical validity of the Metropolis–Hastings algorithms by showing that under suitable (and not very restrictive) conditions on the transition kernel, the chain produced by [*A*.24] is ergodic and, therefore, that the mean (7.1) converges to the expectation $\mathbb{E}_f[h(X)]$. In Sections 7.4 and 7.4, however, we showed that the most common algorithms only rarely enjoy strong ergodicity properties (geometric or uniform ergodicity). In particular, there are simple examples (see Problem 7.5) that show how slow convergence can be.

This section addresses the problem of choosing the transition kernel $q(y|x)$ and illustrates a general acceleration method for Metropolis–Hastings algorithms, which extends the conditioning techniques presented in Section 4.2.

7.6.1 Optimizing the Acceptance Rate

When considering only the classes of algorithms described in Section 7.4, the most common alternatives are to use the following:

(a) a fully automated algorithm like ARMS ([*A*.28]);
(b) an instrumental density g which approximates f, such that f/g is bounded for uniform ergodicity to apply to the algorithm [*A*.25];
(c) a random walk as in [*A*.29].

In case (a), the automated feature of [*A*.28] reduces "parameterization" to the choice of initial values, which are theoretically of limited influence on the efficiency of the algorithm. In both of the other cases, the choice of g is much more critical, as it determines the performances of the resulting Metropolis–Hastings algorithm. As we will see below, the few pieces of advice available on the choice of g are, in fact, contrary! Depending on the type of Metropolis–Hastings algorithm selected, one would want high acceptance rates in case (b) and low acceptance rates in case (c).

Consider, first, the independent Metropolis–Hastings algorithm introduced in Section 7.4. Its similarity with the Accept–Reject algorithm suggests a choice of g that maximizes the average *acceptance rate*

$$\begin{aligned}\rho &= \mathbb{E}\left[\min\left\{\frac{f(Y)\,g(X)}{f(X)\,g(Y)}, 1\right\}\right] \\ &= 2P\left(\frac{f(Y)}{g(Y)} \geq \frac{f(X)}{g(X)}\right), \qquad X \sim f,\ Y \sim g,\end{aligned}$$

as seen[5] in Lemma 7.9. In fact, the optimization associated with the choice of g is related to the speed of convergence of $\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$ to $\mathbb{E}_f[h(X)]$

[5] Under the same assumption of no point mass for the ratio $f(Y)/g(Y)$.

and, therefore, to the ability of the algorithm [*A*.25] to quickly explore any complexity of f (see, for example, Theorem 7.8).

If this optimization is to be generic (that is, independent of h), g should reproduce the density f as faithfully as possible, which implies the maximization of ρ. For example, a density g that is either much less or much more concentrated, compared with f, produces a ratio

$$\frac{f(y)\, g(x)}{f(x)\, g(y)} \wedge 1$$

having huge variations and, therefore, leads to a low acceptance rate.

The acceptance rate ρ is typically impossible to compute, and one solution is to use the minorization result $\rho \geq 1/M$ of Lemma 7.9 to minimize M as in the case of the Accept–Reject algorithm.

Alternatively, we can consider a more *empirical* approach that consists of choosing a parameterized instrumental distribution $g(\cdot|\theta)$ and adjusting the corresponding parameters θ based on the evaluated acceptance rate, now $\hat{\rho}(\theta)$; that is, first choose an initial value for the parameters, θ_0, and estimate the corresponding acceptance rate, $\hat{\rho}(\theta_0)$, based on m iterations of [*A*.25], then modify θ_0 to obtain an increase in $\hat{\rho}$.

In the simplest cases, θ_0 will reduce to a scale parameter which is increased or decreased depending on the behavior of $\hat{\rho}(\theta)$. In multidimensional settings, θ_0 can also include a position parameter or a matrix acting as a scale parameter, which makes optimizing $\rho(\theta)$ a more complex task. Note that $\hat{\rho}(\theta)$ can be obtained by simply counting acceptances or through

$$\frac{2}{m} \sum_{i=1}^{m} \mathbb{I}_{\{f(y_i)g(x_i|\theta) > f(x_i)g(y_i|\theta)\}} ,$$

where $x_1, \ldots, x_m$ is a sample from f, obtained, for instance, from a first MCMC algorithm, and $y_1, \ldots, y_m$ is an iid sample from $g(\cdot|\theta)$. Therefore, if θ is composed of location and scale parameters, a sample $((x_1, y_1), \ldots, (x_m, y_m))$ corresponding to a value θ_0 can be used repeatedly to evaluate different values of θ by a deterministic modification of y_i, which facilitates the maximization of $\rho(\theta)$.

Example 7.18. Inverse Gaussian distribution. The *inverse Gaussian distribution* has the density

$$f(z|\theta_1, \theta_2) \propto z^{-3/2} \exp\left\{-\theta_1 z - \frac{\theta_2}{z} + 2\sqrt{\theta_1\theta_2} + \log \sqrt{2\theta_2}\right\} \tag{7.19}$$

on $\mathbb{R}_+$ ($\theta_1 > 0, \theta_2 > 0$). Denoting $\psi(\theta_1, \theta_2) = 2\sqrt{\theta_1\theta_2} + \log\sqrt{2\theta_2}$, it follows from a classical result on exponential families (see Brown 1986, Chapter 2, Robert 2001, Lemma 3.3.7, or Problem 1.38) that

β	0.2	0.5	0.8	0.9	1	1.1	1.2	1.5
$\hat{\rho}(\beta)$	0.22	0.41	0.54	0.56	0.60	0.63	0.64	0.71
$\mathbb{E}[Z]$	1.137	1.158	1.164	1.154	1.133	1.148	1.181	1.148
$\mathbb{E}[1/Z]$	1.116	1.108	1.116	1.115	1.120	1.126	1.095	1.115

Table 7.3. Estimation of the means of Z and of $1/Z$ for the inverse Gaussian distribution $\mathcal{IN}(\theta_1, \theta_2)$ by the Metropolis–Hastings algorithm [A.25] and evaluation of the acceptance rate for the instrumental distribution $\mathcal{G}a(\sqrt{\theta_2/\theta_1}\ \beta, \beta)$ ($\theta_1 = 1.5, \theta_2 = 2$, and $m = 5000$).

$$\begin{aligned}\mathbb{E}\left[(Z, 1/Z)\right] &= \nabla\psi(\theta_1, \theta_2) \\ &= \left(\sqrt{\frac{\theta_2}{\theta_1}}, \sqrt{\frac{\theta_1}{\theta_2}} + \frac{1}{2\theta_2}\right).\end{aligned}$$

A possible choice for the simulation of (7.19) is the Gamma distribution $\mathcal{G}a(\alpha, \beta)$ in algorithm [A.25], taking $\alpha = \beta\sqrt{\theta_2/\theta_1}$ so that the means of both distributions coincide. Since

$$\frac{f(x)}{g(x)} \propto x^{-\alpha-1/2} \exp\left\{(\beta - \theta_1)x - \frac{\theta_2}{x}\right\},$$

the ratio f/g is bounded for $\beta < \theta_1$. The value of x which maximizes the ratio is the solution of

$$(\beta - \theta_1)x^2 - \left(\alpha + \frac{1}{2}\right)x + \theta_2 = 0\,;$$

that is,

$$x^*_\beta = \frac{(\alpha + 1/2) - \sqrt{(\alpha + 1/2)^2 + 4\theta_2(\theta_1 - \beta)}}{2(\beta - \theta_1)}.$$

The analytical optimization (in β) of

$$M(\beta) = (x^*_\beta)^{-\alpha-1/2} \exp\left\{(\beta - \theta_1)x^*_\beta - \frac{\theta_2}{x^*_\beta}\right\}$$

is not possible, although, in this specific case the curve $M(\beta)$ can be plotted for given values of θ_1 and θ_2 and the optimal value $\beta^\star$ can be approximated numerically. Typically, the influence of the choice of β must be assessed empirically; that is, by approximating the acceptance rate ρ via the method described above.

Note that a new sample $(y_1, \ldots, y_m)$ must be simulated for every new value of β. Whereas $y \sim \mathcal{G}a(\alpha, \beta)$ is equivalent to $\beta y \sim \mathcal{G}a(\alpha, 1)$, the factor α depends on β and it is not possible to use the same sample for several values of β. Table 7.3 provides an evaluation of the rate ρ as a function of β and gives

estimates of the means of Z and $1/Z$ for $\theta_1 = 1.5$ and $\theta_2 = 2$. The constraint on the ratio f/g then imposes $\beta < 1.5$. The corresponding theoretical values are respectively 1.155 and 1.116, and the optimal value of β is $\beta^* = 1.5$. ||

The *random walk* version of the Metropolis–Hastings algorithm, introduced in Section 7.5, requires a different approach to acceptance rates, given the dependence of the instrumental distribution on the current state of the chain. In fact, a high acceptance rate does not necessarily indicate that the algorithm is moving correctly since it may indicate that the random walk is moving too slowly on the surface of f. If $x^{(t)}$ and y_t are close, in the sense that $f(x^{(t)})$ and $f(y_t)$ are approximately equal, the algorithm [A.29] leads to the acceptance of y with probability

$$\min\left(\frac{f(y_t)}{f(x^{(t)})}, 1\right) \simeq 1 .$$

A higher acceptance rate may therefore correspond to a slower convergence as the moves on the support of f are more limited. In the particular case of multimodal densities whose modes are separated by zones of extremely small probability, the negative effect of limited moves on the surface of f clearly shows. While the acceptance rate is quite high for a distribution g with small variance, the probability of jumping from one mode to another may be arbitrarily small. This phenomenon occurs, for instance, in the case of mixtures of distributions (see Section 9.7.1) and in overparameterized models (see, e.g., Tanner and Wong 1987 and Besag et al. 1995). In contrast, if the average acceptance rate is low, the successive values of $f(y_t)$ tend to be small compared with $f(x^{(t)})$, which means that the random walk moves quickly on the surface of f since it often reaches the "borders" of the support of f (or, at least, that the random walk explores regions with low probability under f).

The above analysis seems to require an advanced knowledge of the density of interest, since an instrumental distribution g with too narrow a range will slow down the convergence rate of the algorithm. On the other hand, a distribution g with a wide range results in a waste of simulations of points outside the range of f without improving the probability of visiting all of the modes of f. It is unfortunate that an automated parameterization of g cannot guarantee uniformly optimal performances for the algorithm [A.29], and that the rules for choosing the rate presented in Note 7.8.4 are only heuristic.

7.6.2 Conditioning and Accelerations

Similar[6] to the Accept–Reject method, the Metropolis–Hastings algorithm does not take advantage of the total set of random variables that are generated. Lemma 7.9 shows that the "rate of waste" of these variables y_t is lower than for the Accept–Reject method, but it still seems inefficient to ignore the

[6] This section presents material related to nonparametric Rao–Blackwellization, as in Section 4.2, and may be skipped on a first reading.

rejected y_t's. As the rejection mechanism relies on an independent uniform random variable, it is reasonable to expect that the rejected variables bring, although indirectly, some relevant information on the distribution f. As in the conditioning method introduced in Section 4.2, the *Rao–Blackwellization* technique applies in the case of the Metropolis–Hastings algorithm. (Other approaches to Metropolis–Hastings acceleration can be found in Green and Han 1992, Gelfand and Sahu 1994, or McKeague and Wefelmeyer 2000.)

First, note that a sample produced by the Metropolis–Hastings algorithm, $x^{(1)}, \ldots, x^{(T)}$, is based on two samples, $y_1, \ldots, y_T$ and $u_1, \ldots, u_T$, with $y_t \sim q(y|x^{(t-1)})$ and $u_t \sim \mathcal{U}_{[0,1]}$. The mean (7.1) can then be written

$$\begin{aligned}\delta^{MH} = \frac{1}{T}\sum_{t=1}^{T} h(x^{(t)}) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{t} \mathbb{I}_{x^{(t)}} \\ &= \frac{1}{T}\sum_{t=1}^{T} h(y_t)\sum_{i=t}^{T} \mathbb{I}_{x^{(i)}=y_t}\end{aligned}$$

and the conditional expectation

$$\begin{aligned}\delta^{RB} &= \frac{1}{T}\sum_{t=1}^{T} h(y_t)\,\mathbb{E}\left[\sum_{i=t}^{T} \mathbb{I}_{X^{(i)}=y_t}\middle| y_1, \ldots, y_T\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} h(y_t)\left(\sum_{i=t}^{T} P(X^{(i)} = y_t|y_1, \ldots, y_T)\right)\end{aligned}$$

dominates the empirical mean, δ^{MH}, under quadratic loss. This is a consequence of the Rao–Blackwell Theorem (see Lehmann and Casella 1998, Section 1.7), resulting from the fact that δ^{RB} integrates out the variation due to the uniform sample.

The practical interest of this alternative to δ^{MH} is that the probabilities $P(X^{(i)} = y_t|y_1, \ldots, y_T)$ can be explicitly computed. Casella and Robert (1996) have established the two following results, which provide the weights for $h(y_t)$ in δ^{RB} both for the independent Metropolis–Hastings algorithm and the general Metropolis–Hastings algorithm. In both cases, the computational complexity of these weights is of order $\mathcal{O}(T^2)$, which is a manageable order of magnitude.

Consider first the case of the independent Metropolis–Hastings algorithm associated with the instrumental distribution g. For simplicity's sake, assume that $X^{(0)}$ is simulated according to the distribution of interest, f, so that the chain is stationary, and the mean (7.1) can be written

$$\delta^{MH} = \frac{1}{T+1}\sum_{t=0}^{T} h(x^{(t)})\,,$$

with $x^{(0)} = y_0$. If we denote

n	10	25	50	100
h_1	50.11	49.39	48.27	46.68
h_2	42.20	44.75	45.44	44.57

Table 7.4. Decrease (in percentage) of squared error risk associated with δ^{RB} for the evaluation of $\mathbb{E}[h_i(X)]$, evaluated over 7500 simulations for different sample sizes n. (*Source*: Casella and Robert 1996).

$$w_i = \frac{f(y_i)}{g(y_i)}, \qquad \rho_{ij} = \frac{w_j}{w_i} \wedge 1 \qquad (0 \le i < j),$$

$$\zeta_{ii} = 1, \qquad \zeta_{ij} = \prod_{t=i+1}^{j} (1 - \rho_{it}) \qquad (i < j),$$

we have the following theorem, whose proof is left to Problem 7.31.

Theorem 7.19. *The estimator δ^{RB} can be written*

$$\delta^{RB} = \frac{1}{T+1} \sum_{i=0}^{T} \varphi_i \, h(y_i),$$

where

$$\varphi_i = \tau_i \sum_{j=i}^{T} \zeta_{ij} \ ,$$

and the conditional probability $\tau_i = P(X^{(i)} = y_i | y_0, y_1, \ldots, y_T)$, $i = 0, \ldots, T$, is given by $\tau_0 = 1$ and $\tau_i = \sum_{j=0}^{i-1} \tau_j \, \zeta_{j(i-1)} \rho_{ji}$ for $i > 0$.

The computation of ζ_{ij} for a fixed i requires $(T - i)$ multiplications since $\zeta_{i(j+1)} = \zeta_{ij}(1 - \rho_{i(j+1)})$; therefore, the computation of all the ζ_{ij}'s require $T(T+1)/2$ multiplications. The derivations of τ_i and φ_i are of the same order of complexity.

Example 7.20. Rao–Blackwellization improvement for a $\mathcal{T}_3$ simulation. Suppose the target distribution is $\mathcal{T}_3$ and the instrumental distribution is Cauchy, $\mathcal{C}(0, 1)$. The ratio f/g is bounded, which ensures a geometric rate of convergence for the associated Metropolis–Hastings algorithm. Table 7.4 illustrates the improvement brought by δ^{RB} for some functions of interest $h_1(x) = x$ and $h_2(x) = \mathbb{I}_{(1.96,+\infty)}(x)$, whose (exact) expectations $\mathbb{E}[h_i(X)]$ $(i = 1, 2)$ are 0 and 0.07, respectively. Over the different sample sizes selected for the experiment, the improvement in mean square error brought by δ^{RB} is of the order 50%. ||

We next consider the general case, with an arbitrary instrumental distribution $q(y|x)$. The dependence between Y_i and the set of previous variables Y_j $(j < i)$ (since $X^{(i-1)}$ can be equal to Y_0, $Y_1, \ldots$, or Y_{i-1}) complicates the expression of the joint distribution of Y_i and U_i, which cannot be obtained in closed form for arbitrary n. In fact, although $(X^{(t)})$ is a Markov chain, (Y_t) is not.

Let us denote

$$\begin{aligned}
&\rho_{ij} = \frac{f(y_j)/q(y_j|y_i)}{f(y_i)/q(y_i|y_j)} \wedge 1 \qquad (j > i),\\
&\overline{\rho}_{ij} = \rho_{ij} q(y_{j+1}|y_j), \quad \underline{\rho}_{ij} = (1-\rho_{ij}) q(y_{j+1}|y_i) \qquad (i < j < T),\\
&\zeta_{jj} = 1, \quad \zeta_{jt} = \prod_{l=j+1}^{t} \underline{\rho}_{jl} \qquad (i < j < T),\\
&\tau_0 = 1, \quad \tau_j = \sum_{t=0}^{j-1} \tau_t \zeta_{t(j-1)}\, \overline{\rho}_{tj}, \quad \tau_T = \sum_{t=0}^{T-1} \tau_t \zeta_{t(T-1)} \rho_{tT} \qquad (i < T),\\
&\omega_T^i = 1, \quad \omega_i^j = \overline{\rho}_{ji}\omega_{i+1}^i + \underline{\rho}_{tj}\omega_{i+1}^j \qquad (0 \le j < i < T).
\end{aligned}$$

Casella (1996) derives the following expression for the weights of $h(y_i)$ in δ^{RB}. We again leave the proof as a problem (Problem 7.32).

Theorem 7.21. *The estimator δ^{RB} satisfies*

$$\delta^{RB} = \frac{\sum_{i=0}^{T} \varphi_i\, h(y_i)}{\sum_{i=0}^{T-1} \tau_i\, \zeta_{i(T-1)}},$$

with $(i < T)$

$$\varphi_i = \tau_i \left[\sum_{j=i}^{T-1} \zeta_{ij}\omega_{j+1}^i + \zeta_{i(T-1)}(1-\rho_{iT}) \right]$$

and $\varphi_T = \tau_T$.

Although these estimators are more complex than in the independent case, the complexity of the weights is again of order $\mathcal{O}(T^2)$ since the computations of $\overline{\rho}_{ij}$, $\underline{\rho}_{ij}$, ζ_{ij}, τ_i, and ω_j^i involve $T(T+1)/2$ multiplications. Casella and Robert (1996) give algorithmic advice toward easier and faster implementation.

Example 7.22. (Continuation of Example 7.20) Consider now the simulation of a $\mathcal{T}_3$ distribution based on a random walk with perturbations distributed as $\mathcal{C}(0, \sigma^2)$. The choice of σ determines the acceptance rate for the Metropolis–Hastings algorithm: When $\sigma = 0.4$, it is about 0.33, and when $\sigma = 3.0$, it increases to 0.75.

As explained in Section 7.6.1, the choice $\sigma = 0.4$ is undoubtedly preferable in terms of efficiency of the algorithm. Table 7.5 confirms this argument,

	n	10	25	50	100
$\sigma = 0.4$	h_1	10.7	8.8	7.7	7.7
		(1.52)	(0.98)	(0.63)	(0.3)
	h_2	23.6	25.2	25.8	25.0
		(0.02)	(0.01)	(0.006)	(0.003)
$\sigma = 3.0$	h_1	0.18	0.15	0.11	0.07
		(2.28)	(1.77)	(1.31)	(0.87)
	h_2	0.99	0.94	0.71	1.19
		(0.03)	(0.02)	(0.014)	(0.008)

Table 7.5. Improvement brought by δ^{RB} (in %) and quadratic risk of the empirical average (in parentheses) for different sample sizes and 50,000 simulations of the random walk based on $\mathcal{C}(0, \sigma^2)$ (*Source*: Casella and Robert 1996).

since the quadratic risk of the estimators (7.1) is larger for $\sigma = 3.0$. The gains brought by δ^{RB} are smaller, compared with the independent case. They amount to approximately 8% and 25% for $\sigma = 0.4$ and 0.1% and 1% for $\sigma = 3$. Casella and Robert (1996) consider an additional comparison with an importance sampling estimator based on the same sample $y_1, \ldots, y_n$. ‖

7.6.3 Adaptive Schemes

Given the range of situations where MCMC applies, it is unrealistic to hope for a *generic* MCMC sampler that would function in every possible setting. The more generic proposals like random walk Metropolis–Hastings algorithms are known to fail in large dimension and disconnected supports, because they take too long to explore the space of interest (Neal 2003). The reason for this impossibility theorem is that, in realistic problems, the complexity of the distribution to simulation is the very reason why MCMC is used! So it is difficult to ask for a prior opinion about this distribution, its support or the parameters of the proposal distribution used in the MCMC algorithm: intuition is close to void in most of these problems.

However, the performances of off-the-shelf algorithms like the random walk Metropolis–Hastings scheme bring information about the distribution of interest and, thus, should be incorporated in the design of better and more powerful algorithms. The problem is that we usually miss the time to train the algorithm on these previous performances and are looking for the Holy Grail of automated MCMC procedures! While it is natural to think that the information brought by the first steps of an MCMC algorithm should be used in later steps, there is a severe catch: using the whole past of the "chain" implies that this is not a Markov chain any longer. Therefore, usual convergence theorems do not apply and the validity of the corresponding algorithms

is questionable. Further, it may be that, in practice, such algorithms do degenerate to point masses because of a too-rapid decrease in the variation of their proposal.

Example 7.23. t-distribution. Consider a $\mathcal{T}(\nu, \theta, 1)$ sample $(x_1, \ldots, x_n)$ with ν known. Assume in addition a flat prior $\pi(\theta) = 1$ on θ as in a noninformative environment. While the posterior distribution can be easily computed at an arbitrary value of θ, direct simulation and computation from this posterior is impossible. In a Metropolis–Hastings framework, we could fit a normal proposal from the empirical mean and variance of the previous values of the chain,

$$\mu_t = \frac{1}{t}\sum_{i=1}^{t}\theta^{(i)} \quad \text{and} \quad \sigma_t^2 = \frac{1}{t}\sum_{i=1}^{t}(\theta^{(i)} - \mu_t)^2\,.$$

Notwithstanding the dependence on the past, we could then use the Metropolis-Hastings acceptance probability

$$\prod_{j=2}^{n}\left[\frac{\nu + (x_j - \theta^{(t)})^2}{\nu + (x_j - \xi)^2}\right]^{-(\nu+1)/2} \frac{\exp -(\mu_t - \theta^{(t)})^2/2\sigma_t^2}{\exp -(\mu_t - \xi)^2/2\sigma_t^2}\,,$$

where ξ is the proposed value from $\mathcal{N}(\mu_t, \sigma_t^2)$. The invalidity of this scheme (because of the dependence on the whole sequence of $\theta^{(i)}$'s till iteration t) is illustrated in Figure 7.6: when the range of the initial values is too small, the sequence of $\theta^{(i)}$'s cannot converge to the target distribution and concentrates on too small a support. But the problem is deeper, because even when the range of the simulated values is correct, the (long-term) dependence on past values modifies the distribution of the sequence. Figure 7.7 shows that, for an initial variance of 2.5, there is a bias in the histogram, even after 25,000 iterations and stabilization of the empirical mean and variance. ||

Even though the Markov chain is converging *in distribution* to the target distribution (when using a proper, i.e., time-homogeneous, updating scheme), using past simulations to create a nonparametric approximation to the target distribution does not work either. For instance, Figure 7.8 shows the output of an adaptive scheme in the setting of Example 7.23 when the proposal distribution is the Gaussian kernel based on earlier simulations. A very large number of iterations is not sufficient to reach an acceptable approximation of the target distribution.

The overall message is thus that one should not *constantly* adapt the proposal distribution on the past performances of the simulated chain. Either the adaptation must cease after a period of *burn in* (not to be taken into account for the computations of expectations and quantities related to the target distribution), or the adaptive scheme must be theoretically assessed in its own right. This latter path is not easy and only a few examples can be found (so

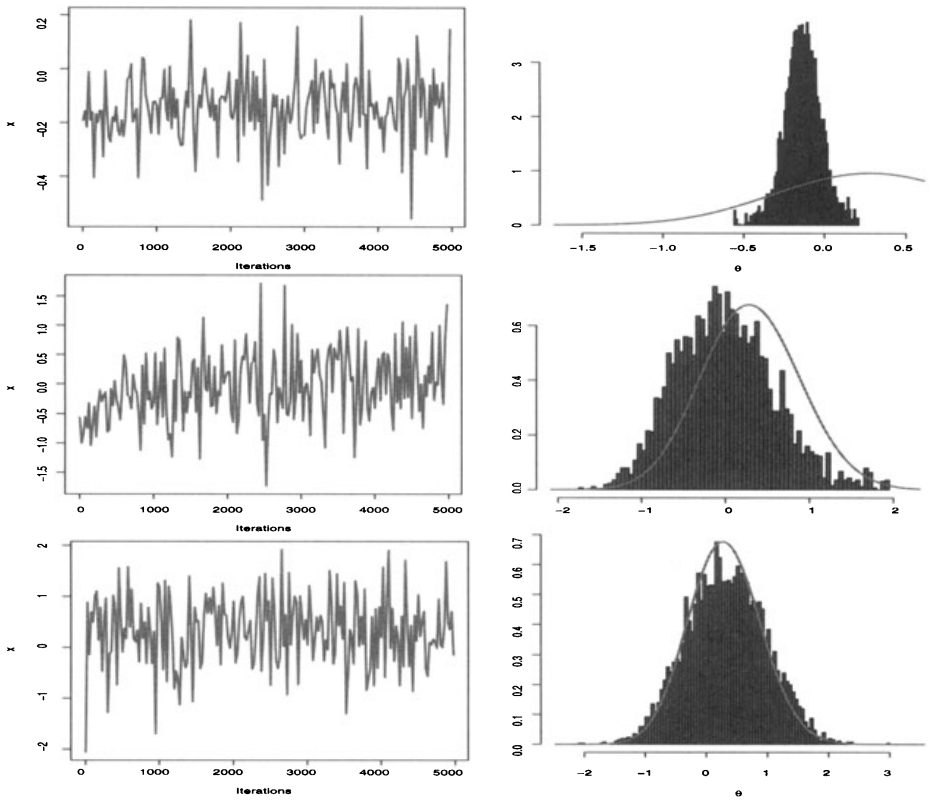

Fig. 7.6. Output of the adaptive scheme for the t-distribution posterior with a sample of 10 $x_j \sim \mathcal{T}_3$ and initial variances of *(top)* 0.1, *(middle)* 0.5, and *(bottom)* 2.5. The left column plots the sequence of $\theta^{(i)}$'s while the right column compares its histogram against the true posterior distribution *(with a different scale for the upper graph)*.

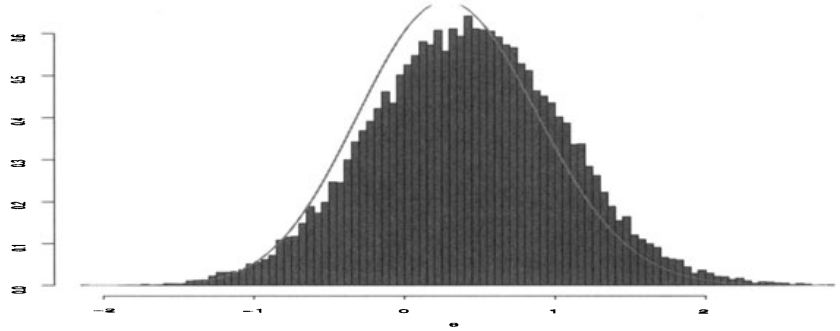

Fig. 7.7. Comparison of the distribution of an adaptive scheme sample of 25,000 points with initial variance of 2.5 and of the target distribution.

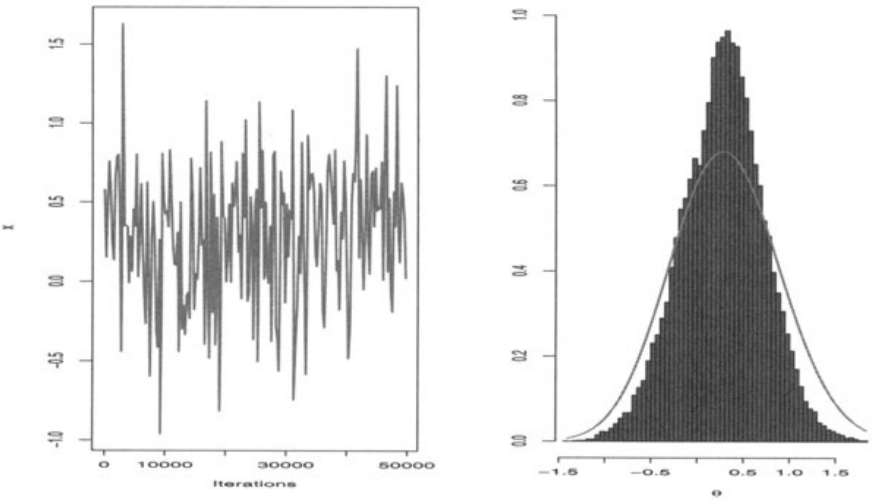

Fig. 7.8. Sample produced by 50,000 iterations of a nonparametric adaptive MCMC scheme and comparison of its distribution with the target distribution.

far) in the literature. See, e.g., Gilks et al. (1998), who use regeneration to create block independence and preserve Markovianity on the paths rather than on the values (see also Sahu and Zhigljavsky 1998 and Holden 1998), Haario et al. (1999, 2001), who derive a proper[7] adaptation scheme in the spirit of Example 7.23 by using a Gaussian proposal and a ridge-like correction to the empirical variance,

$$\Sigma_t = s\text{Cov}(\theta^{(0)}, \ldots, \theta^{(t)}) + s\epsilon I_d\,,$$

where s and ϵ are constant, and Andrieu and Robert (2001), who propose a more general framework of valid adaptivity based on stochastic optimization and the Robbin–Monro algorithm. (The latter actually embeds the chain of interest $\theta^{(t)}$ in a larger chain $(\theta^{(t)}, \xi^{(t)})$ that also includes the parameter of the proposal distribution as well as the gradient of a performance criterion.) We will again consider adaptive algorithms in Chapter 14, with more accessible theoretical justifications.

7.7 Problems

7.1 Calculate the mean of a Gamma(4.3, 6.2) random variable using
(a) Accept-Reject with a Gamma(4, 7) candidate.
(b) Metropolis–Hastings with a Gamma(4, 7) candidate.
(c) Metropolis–Hastings with a Gamma(5, 6) candidate.

[7] Since the chain is not Markov, the authors need to derive an ergodic theorem on their own.

In each case monitor the convergence.

7.2 Student's $\mathcal{T}_\nu$ density with ν degrees of freedom is given by

$$f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} .$$

Calculate the mean of a t distribution with 4 degrees of freedom using a Metropolis–Hastings algorithm with candidate density

(a) $\mathrm{N}(0, 1)$

(b) t with 2 degrees of freedom.

Monitor the convergence of each.

7.3 Complete some details of Theorem 7.2:

(a) To establish (7.2), show that

$$\begin{aligned} K(x, A) &= P(X_{t+1} \in A | X_t = x) \\ &= P(Y \in A \text{ and } X_{t+1} = Y | X_t = x) + P(x \in A \text{ and } X_{t+1} = x | X_t = x) \\ &= \int_A q(y|x)\varrho(x, y)dy + \int_{\mathcal{Y}} \mathbb{I}(x \in A)(1 - \varrho(x, y))q(y|x)dy, \end{aligned}$$

where $q(y|x)$ is the instrumental density and $\varrho(x, y) = P(X_{t+1} = y | X_t = x)$. Take the limiting case $A = \{y\}$ to establish (7.2).

(b) Establish (7.3). Notice that $\delta_y(x)f(y) = \delta_x(y)f(x)$.

7.4 For the transition kernel,

$$X^{(t+1)}|x^{(t)} \sim \mathcal{N}(\rho x^{(t)}, \tau^2)$$

gives sufficient conditions on ρ and τ for the stationary distribution π to exist. Show that, in this case, π is a normal distribution and that (7.4) occurs.

7.5 (Doukhan et al. 1994) The algorithm presented in this problem is used in Chapter 12 as a benchmark for slow convergence.

(a) Prove the following result:

Lemma 7.24. *Consider a probability density g on $[0, 1]$ and a function $0 < \rho < 1$ such that*

$$\int_0^1 \frac{g(x)}{1 - \rho(x)} dx < \infty .$$

The Markov chain with transition kernel

$$K(x, x') = \rho(x)\, \delta_x(x') + (1 - \rho(x))\, g(x') ,$$

where δ_x is the Dirac mass at x, has stationary distribution

$$f(x) \propto g(x)/(1 - \rho(x)).$$

(b) Show that an algorithm for generating the Markov chain associated with Lemma 7.24 is given by

Algorithm A.30 –Repeat or Simulate–

1. Take $X^{(t+1)} = x^{(t)}$ with probability $\rho(x^{(t)})$
2. Else, generate $X^{(t+1)} \sim g(y)$. [A.30]