Chapter 1

Bayesian Inference and Markov Chain Monte Carlo

1.1 Bayes

Bayesian inference is a probabilistic inferential method. In the last two decades, it has become more popular than ever due to affordable computing power and recent advances in Markov chain Monte Carlo (MCMC) methods for approximating high dimensional integrals.

Bayesian inference can be traced back to Thomas Bayes (1764), who derived the inverse probability of the success probability θ in a sequence of independent Bernoulli trials, where θ was taken from the uniform distribution on the unit interval $(0, 1)$ but treated as unobserved. For later reference, we describe his experiment using familiar modern terminology as follows.

■ **Example 1.1 The Bernoulli (or Binomial) Model With Known Prior**

Suppose that $\theta \sim \text{Unif}(0, 1)$, the uniform distribution over the unit interval $(0, 1)$, and that $x_1, \ldots, x_n$ is a sample from Bernoulli(θ), which has the sample space $\mathcal{X} = \{0, 1\}$ and probability mass function (pmf)

$$\Pr(X = 1|\theta) = \theta \quad \text{and} \quad \Pr(X = 0|\theta) = 1 - \theta, \tag{1.1}$$

where X denotes the Bernoulli random variable (r.v.) with $X = 1$ for *success* and $X = 0$ for *failure*. Write $N = \sum_{i=1}^{n} x_i$, the observed number of successes in the n Bernoulli trials. Then $N|\theta \sim \text{Binomial}(n, \theta)$, the Binomial distribution with parameters size n and probability of success θ.

Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples
Faming Liang, Chuanhai Liu and Raymond J. Carroll

The inverse probability of θ given $x_1, \ldots, x_n$, known as the posterior distribution, is obtained from Bayes' theorem, or more rigorously in modern probability theory, the definition of conditional distribution, as the Beta distribution $\text{Beta}(1+N, 1+n-N)$ with probability density function (pdf)

$$\frac{1}{B(1+N, 1+n-N)} \theta^{(1+N)-1}(1-\theta)^{(1+n-N)-1} \qquad (0 \le \theta \le 1), \tag{1.2}$$

where $B(\cdot, \cdot)$ stands for the Beta function.

1.1.1 Specification of Bayesian Models

Real world problems in statistical inference involve the unknown quantity θ and observed data X. For different views on the philosophical foundations of Bayesian approach, see Savage (1967a, b), Berger (1985), Rubin (1984), and Bernardo and Smith (1994). As far as the mathematical description of a Bayesian model is concerned, Bayesian data analysis amounts to

(i) specifying a sampling model for the observed data X, conditioned on an unknown quantity θ,

$$X \sim f(X|\theta) \qquad (X \in \mathcal{X}, \theta \in \Theta), \tag{1.3}$$

where $f(X|\theta)$ stands for either pdf or pmf as appropriate, and

(ii) specifying a marginal distribution $\pi(\theta)$ for θ, called the prior distribution or simply the *prior* for short,

$$\theta \sim \pi(\theta) \qquad (\theta \in \Theta). \tag{1.4}$$

Technically, data analysis for producing inferential results on assertions of interest is reduced to computing integrals with respect to the posterior distribution, or *posterior* for short,

$$\pi(\theta|X) = \frac{\pi(\theta)L(\theta|X)}{\int \pi(\theta)L(\theta|X)d\theta} \qquad (\theta \in \Theta), \tag{1.5}$$

where $L(\theta|X) \propto f(X|\theta)$ in θ, called the likelihood of θ given X. Our focus in this book is on efficient and accurate approximations to these integrals for scientific inference. Thus, limited discussion of Bayesian inference is necessary.

1.1.2 The Jeffreys Priors and Beyond

By its nature, Bayesian inference is necessarily subjective because specification of the full Bayesian model amounts to practically summarizing available information in terms of precise probabilities. Specification of probability models is unavoidable even for frequentist methods, which requires specification

of the sampling model, either parametric or non-parametric, for the observed data X. In addition to the sampling model of the observed data X for developing frequentist procedures concerning the unknown quantity θ, Bayesian inference demands a fully specified prior for θ. This is natural when prior information on θ is available and can be summarized precisely by a probability distribution. For situations where such information is neither available nor easily quantified with a precise probability distribution, especially for high dimensional problems, a commonly used method in practice is the Jeffreys method, which suggests the prior of the form

$$\pi_J(\theta) \propto |I(\theta)|^{1/2} \qquad (\theta \in \Theta), \tag{1.6}$$

where $I(\theta)$ denotes the Fisher information

$$I(\theta) = -\int \frac{\partial^2 \ln f(x|\theta)}{\partial\theta(\partial\theta)'} f(x|\theta)dx.$$

The Jeffreys priors have the appealing property that they are invariant under reparameterization. A theoretical adjustment in terms of frequency properties in the context of large samples can be found in Welch and Peers (1963). Note that prior distributions do not need to be proper as long as the posteriors are proper and produce sensible inferential results. The following Gaussian example shows that the Jeffreys prior is sensible for single parameters.

■ **Example 1.2 The Gaussian $N(\mu, 1)$ Model**

Suppose that a sample is considered to have taken from the Gaussian population $N(\mu, 1)$ with unit variance and unknown mean μ to be inferred. The Fisher information is obtained as

$$I(\mu) = \int_{-\infty}^{\infty} \phi(x-\mu)dx = 1,$$

where $\phi(x-\mu) = (2\pi)^{-1/2}\exp\{-\frac{1}{2}(x-\mu)^2\}$ is the pdf of $N(\mu, 1)$. It follows that the Jeffreys prior for θ is the flat prior

$$\pi_J(\mu) \propto 1 \qquad (-\infty < \mu < \infty), \tag{1.7}$$

resulting in the corresponding posterior distribution of θ given X

$$\pi_J(\mu|X) = N(X, 1). \tag{1.8}$$

Care must be taken when using the Jeffreys rule. For example, it is easy to show that applying the Jeffreys rule to the Gaussian model $N(\mu, \sigma^2)$ with both mean μ and variance σ^2 unknown leads to the prior

$$\pi_J(\mu, \sigma^2) \propto \frac{1}{\sigma^3} \qquad (-\infty < \mu < \infty; \sigma^2 > 0).$$

However, this is not the commonly used prior that has better frequency properties (for inference about μ or σ) and is given by

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \qquad (-\infty < \mu < \infty; \sigma^2 > 0),$$

that is, μ and σ^2 are independent and the distributions for both μ and $\ln \sigma^2$ are flat. For high dimensional problems with small samples, the Jeffreys rule often becomes even less appealing. There are also different perspectives, provided by the extensive work on reference priors by José Bernardo and James Berger (see, e.g., Bernardo, 1979; Berger, 1985). For more discussion of prior specifications, see Kass and Wasserman (1996).

For practical purposes, we refer to Box and Tiao (1973) and Gelman *et al.* (2004) for discussion on specification of prior distributions. The general guidance for specification of priors when no prior information is available, as is typical in Bayesian analysis, is to find priors that lead to posteriors having good frequency properties (see, e.g., Rubin, 1984; Dawid, 1985). Materials on probabilistic inference without using difficult-to-specify priors are available but beyond the scope of Bayesian inference and therefore will not be discussed in this book. Readers interested in this fascinating area are referred to Fisher (1973), Dempster (2008), and Martin *et al.* (2009). We note that MCMC methods can be applied there as well.

1.2 Bayes Output

Bayesian analysis for scientific inference does not end with posterior derivation and computation. It is thus critical for posterior distributions to have clear interpretation. For the sake of clarity, probability used in this book has a long-run frequency interpretation in repeated experiments. Thus, standard probability theory, such as conditioning and marginalization, can be applied. Interpretation also suggests how to report Bayesian output as our assessment of assertions of interest on quantities in the specified model. In the following two subsections, we discuss two types of commonly used Bayes output, credible intervals for estimation and Bayes factors for hypothesis testing.

1.2.1 Credible Intervals and Regions

Credible intervals are simply posterior probability intervals. They are used for purposes similar to those of confidence intervals in frequentist statistics and thereby are also known as Bayesian confidence intervals. For example, the 95% left-sided Bayesian credible interval for the parameter μ in the Gaussian Example 1.2 is $[-\infty, X + 1.64]$, meaning that the posterior probability that μ lies in the interval from $-\infty$ to $X + 1.64$ is 0.95. Similar to frequentist construction of two-sided intervals, for given $\alpha \in (0, 1)$, a $100(1 - \alpha)\%$ two-sided

Bayesian credible interval for a single parameter θ with equal posterior tail probabilities is defined as

$$[\theta_{\alpha/2}, \theta_{1-\alpha/2}] \tag{1.9}$$

where the two end points are the $\alpha/2$ and $1-\alpha/2$ quantiles of the (marginal) posterior distribution of θ. For the the Gaussian Example 1.2, the two-sided 95% Bayesian credible interval is $[X-1.96, X+1.96]$.

In dealing simultaneously with more than one unknown quantity, the term credible region is used in place of credible interval. For a more general term, we refer to credible intervals and regions as *credible sets*. Constructing credible sets is somewhat subjective and usually depends on the problems of interest. A common way is to choose the region with highest posterior density (h.p.d.). The $100(1-\alpha)\%$ h.p.d. region is given by

$$R_{1-\alpha}^{(\pi)} = \{\theta : \pi(\theta|X) \geq \pi(\theta_{1-\alpha}|X)\} \tag{1.10}$$

for some $\theta_{1-\alpha}$ satisfying

$$\Pr\left(\theta \in R_{1-\alpha}^{(\pi)}|X\right) = 1-\alpha.$$

For the the Gaussian Example 1.2, the 95% h.p.d. interval is $[X-1.96, X+1.96]$, the same as the two-sided 95% Bayesian credible interval because the posterior of μ is unimodal and symmetric. We note that the concept of h.p.d. can also be used for functions of θ such as components of θ in high dimensional situations.

For a given probability content $(1-\alpha)$, the h.p.d. region has the smallest volume in the space of θ. This is attractive but depends on the functional form of unknown quantities, such as θ and θ^2. An alternative credible set is obtained by replacing the posterior density $\pi(\theta|X)$ in (1.10) with the likelihood $L(\theta|X)$:

$$R_{1-\alpha}^{(L)} = \{\theta : L(\theta|X) \geq L(\theta_{1-\alpha}|X)\} \tag{1.11}$$

for some $\theta_{1-\alpha}$ satisfying

$$\Pr\left(\theta \in R_{1-\alpha}^{(L)}|X\right) = 1-\alpha.$$

The likelihood based credible region does not depend on transformation of θ. This is appealing, in particular when no prior information is available on θ, that is, when the specified prior works merely as a working prior leading to inference having good frequency properties.

1.2.2 Hypothesis Testing: Bayes Factors

While the use of credible intervals is a Bayesian alternative to frequentist confidence intervals, the use of Bayes factors has been a Bayesian alternative to classical hypothesis testing. Bayes factors have also been used to develop Bayesian methods for model comparison and selection. Here we review the basics of Bayes factors. For more discussion on Bayes factors, including its

history, applications, and difficulties, see Kass and Raftery (1995), Gelman *et al*. (2004), and references therein.

The concept of Bayes factors is introduced in the situation with a common observed data X and two competing hypotheses denoted by H_1 and H_2. A full Bayesian analysis requires

(i) specifying a prior distribution on H_1 and H_2, denoted by, $\Pr(H_1)$ and $\Pr(H_2)$, and

(ii) for each $k = 1$ and 2, specifying the likelihood $L_k(\theta_k|X) = f_k(X|\theta_k)$ and prior $\pi(\theta_k|H_k)$ for θ_k, conditioned on the truth of H_k, where θ_k is the parameter under H_k.

Integrating out θ_k yields

$$\Pr(X|H_k) = \int f_k(X|H_k)\pi(\theta_k|H_k)d\theta_k \tag{1.12}$$

for $k = 1$ and 2. The Bayes factor is the posterior odds of one hypothesis when the prior probabilities of the two hypotheses are equal. More precisely, the Bayes factor in favor of H_1 over H_2 is defined as

$$B_{12} = \frac{\Pr(X|H_1)}{\Pr(X|H_2)}. \tag{1.13}$$

The use of Bayes factors for hypothesis testing is similar to the likelihood ratio test, but instead of maximizing the likelihood, Bayesians in favor of Bayes factors average it over the parameters. According to the definition of Bayes factors, proper priors are often required. Thus, care must be taken in specification of priors so that inferential results are meaningful. In addition, the use of Bayes factors renders lack of probabilistic feature of Bayesian inference. In other words, it is consistent with the likelihood principle, but lacks of a metric or a probability scale to measure the strength of evidence. For a summary of evidence provided by data in favor of H_1 over H_2, Jeffreys (1961) (see also Kass and Raftery (1995)) proposed to interpret the Bayes factor as shown in Table 1.1.

The use of Bayes factor is illustrated by the following binomial example.

Table 1.1 Interpretation of Bayes factors.

$\log_{10}(B_{12})$	B_{12}	evidence against H_2
0 to 1/2	1 to 3.2	Barely worth mentioning
1.2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

■ **Example 1.3 The Binomial Model (continued with a numerical example)**

Suppose we take a sample of $n = 100$ from Bernoulli(θ) with unknown θ, and observe $N = 63$ successes and $n - N = 37$ failures. Suppose that two competing hypotheses are

$$H_1 : \theta = 1/2 \quad \text{and} \quad H_2 : \theta \neq 1/2. \tag{1.14}$$

Under H_1, the likelihood is calculated according to the binomial distribution:

$$\Pr(N|H_1) = \binom{n}{N} \left(\frac{1}{2}\right)^N \left(\frac{1}{2}\right)^{n-N}$$

Under H_2, instead of the uniform over the unit interval we consider the Jeffreys prior

$$\pi(\theta) = \frac{\Gamma(1/2 + 1/2)}{\Gamma(1/2)\Gamma(1/2)} \theta^{1/2-1}(1-\theta)^{1/2-1} = \frac{1}{\pi}\theta^{1/2-1}(1-\theta)^{1/2-1}$$

the proper Beta distribution with shape parameters 1/2 and 1/2. Hence, we have

$$\Pr(N|H_2) = \frac{1}{\pi}\binom{n}{N} \text{Beta}(N + 1/2, n - N + 1/2).$$

The Bayes factor $\log_{10}(B_{12})$ is then -0.4, which is 'barely worth mentioning' even if it points very slightly towards H_2.

It has been recognized that Bayes factor can be sensitive to the prior, which is related to what is known as Lindley's paradox (see Shafer (1982)).

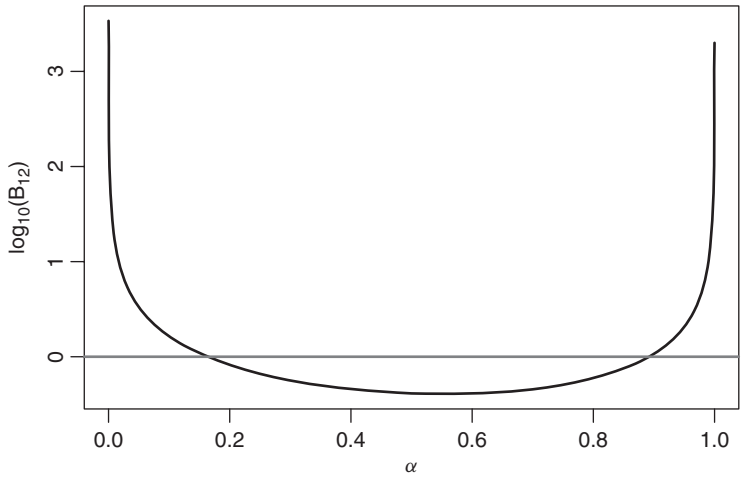

Figure 1.1 Bayes factors in the binomial example with $n = 100$, $N = 63$, and priors Beta($\alpha, 1 - \alpha$) for $0 \leq \alpha \leq 1$.

This is shown in Figure 1.1 for a class of Beta priors Beta$(\alpha, 1-\alpha)$ for $0 \leq \alpha \leq 1$. The Bayes factor is infinity at the two extreme priors corresponding to $\alpha = 0$ and $\alpha = 1$. It can be shown that this class of priors is necessary in the context of imprecise Bayes for producing inferential results that have desired frequency properties. This supports the idea that care must be taken in interpreting Bayesian factors in scientific inference.

Bayesian factors are not the same as a classical likelihood ratio test. A frequentist hypothesis test of H_1 considered as a *null* hypothesis would have produced a more dramatic result, saying that H_1 could be rejected at the 1% significance level, since the probability of getting 63 or more successes from a sample of 100 if $\theta = 1/2$ is 0.0060, and as a normal approximation based two-tailed test of getting a figure as extreme as or more extreme than 63 is 0.0093. Note that 63 is more than two standard deviations away from 50, the expected count under H_1.

1.3 Monte Carlo Integration

1.3.1 The Problem

Let ν be a probability measure over the Borel σ-field $\mathcal{X}$ on the sample space $\mathcal{X} \subseteq \mathbb{R}^d$, where $\mathbb{R}^d$ denotes the d-dimensional Euclidian space. A commonly encountered challenging problem is to evaluate integrals of the form

$$E_\nu[h(X)] = \int_{\mathcal{X}} h(x)\nu(dx) \tag{1.15}$$

where $h(x)$ is a measurable function. Suppose that ν has a pdf $f(x)$. Then (1.15) can be written as

$$E_f[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx \tag{1.16}$$

For example, for evaluating the probability $\Pr(X \in \mathcal{S})$ for $S \subset \mathcal{X}$, $h(x)$ is the indicator function $h(x) = I_{x \in \mathcal{S}}$ with $h(x) = 1$ if $x \in \mathcal{S}$ and $h(x) = 0$ otherwise, and for computing the marginal distribution of $f_Y(y)$ from the joint distribution $f_{X,Y}(x, y)$, the representation in the form of (1.16) is $E_{f_X}\left[f_{Y|X}(y|x)\right]$, where $f_X(x)$ is the marginal pdf of X and $f_{Y|X}(y|x)$ is the conditional pdf of Y given X.

When the problem appears to be intractable analytically, the tool box of numerical integration methods is the next possible alternative, see, for example, Press *et al.* (1992) and references therein. For high dimensional problems, Monte Carlo methods have proved to be popular due to their simplicity and accuracy given limited computing power.

1.3.2 Monte Carlo Approximation

Suppose that it is easy to simulate a sample of size n, denoted by $X_1, \ldots, X_n$, from $f(x)$, the pdf involved in (1.16). Then the sample mean of $h(X)$,

$$\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i), \tag{1.17}$$

can be used to approximate (1.16) because $\bar{h}_n$ converges to (1.16) almost surely by the Strong Law of Large Numbers. When $h(X)$ has a finite variance, the error of this approximation can be characterized by the central limit theorem, that is,

$$\frac{\overline{h}_n - E_f[h(X)]}{\sqrt{n\mathrm{Var}(h(X))}} \sim N(0,1).$$

The variance term $\mathrm{Var}(h(X))$ can be approximated in the same fashion, namely, by the sample variance

$$\frac{1}{n-1}\sum_{i=1}^{n}(h(X_i) - \bar{h}_n)^2.$$

This method of approximating integrals by simulated samples is known as the Monte Carlo method (Metropolis and Ulam, 1949).

1.3.3 Monte Carlo via Importance Sampling

When it is hard to draw samples from $f(x)$ directly, one can resort to importance sampling, which is developed based on the following identity:

$$E_f[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx = \int_{\mathcal{X}} h(x)\frac{f(x)}{g(x)}g(x)dx = E_g[h(X)f(X)/g(X)],$$

where $g(x)$ is a pdf over $\mathcal{X}$ and is positive for every x at which $f(x)$ is positive. This identity suggests that samples from density functions different from $f(x)$ can also be used to approximate (1.16). The standard Monte Carlo theory in Section 1.3.2 applies here because of

$$E_f[h(X)] = E_g[h(X)f(X)/g(X)] = E_g[\tilde{h}(X)]$$

where $\tilde{h}(x) = h(x)\frac{f(x)}{g(x)}g(x)$. The estimator of $E_f[h(X)]$ now becomes

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m}\frac{f(x_i)}{g(x_i)}h(x_i), \tag{1.18}$$

where $x_1, \ldots, x_n$ are iid samples drawing from $g(x)$. Compared to (1.17), for each $i = 1, \ldots, m$, x_i enters with a weight $w_i = f(x_i)/g(x_i)$. For this reason,

this method is called the *importance sampling* method. Most important for this method is to choose $g(x)$ for both simplicity in generating Monte Carlo samples and accuracy in estimating $E_f[h(X)]$ by controlling the associated Monte Carlo errors.

For Monte Carlo accuracy, a natural way is to choose $g(x)$ to minimize the variance of $\tilde{h}(X)$ with $X \sim g(x)$. Theoretical results on optimal $g(x)$ are also available. The following result is due to Rubinstein (1981); see also Robert and Casella (2004).

Theorem 1.3.1 *The choice of g that minimizes the variance of the estimator of* $E_f[h(X)]$ *in (1.18) is*

$$g^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(y)|f(y)dy}.$$

The proof of Theorem 1.3.1 is left as an exercise. As always, theoretical results provide helpful guidance. In practice, balancing simplicity and optimality is more complex because human efforts and computer CPU time for creating samples from $g(x)$ are perhaps the major factors to be considered. Also, it is not atypical to evaluate integrals of multiple functions of $h(x)$, for example, in Bayesian statistics, with a common Monte Carlo sample.

1.4 Random Variable Generation

Monte Carlo methods rely on sampling from probability distributions. Generating a sample of iid draws on computer from the simplest continuous uniform Unif $(0, 1)$ is fundamentally important because all sampling methods depend on uniform random number generators. For example, for every continuous univariate distribution $f(x)$ with cdf $F(x)$ the so-called inverse-cdf method is given as follows.

Algorithm 1.1 *(Continuous Inverse-cdf)*

1. *Generate a uniform random variable U.*
2. *Compute and return* $X = F^{-1}(U)$.

where $F^{-1}(.)$ represents the inverse function of the cdf $F(.)$, provides an algorithm to create samples from $F(x)$. Similarly, for every discrete univariate distribution $p(x)$ with cdf $F(x)$ the inverse-cdf method becomes

Algorithm 1.2 *(Discrete Inverse-cdf)*

1. *Generate a uniform random variable U.*
2. *Find X such that* $F(X-1) < U \leq F(X)$.
3. *Return X.*

and provides an algorithm to create samples from $F(x)$. However, this algorithm is in general computationally expensive. More efficient methods are discussed in Sections 1.4.1 and 1.4.2, where a good and efficient uniform random generator is assumed to be available.

Unfortunately, computers are deterministic in nature and cannot be programmed to produce pure random numbers. Pseudo-random number generators are commonly used in Monte Carlo simulation. Pseudo-random number generators are algorithms that can automatically create long runs with good random properties but eventually the sequence repeats. We refer to Devroye (1986), Robert and Casella (2004), and Matsumoto and Nishimura (1998) for discussion on pseudo-random number generators, among which the Mersenne Twister of Matsumoto and Nishimura (1998) has been used as the default pseudo-random number generator in many softwares. In what follows, we give a brief review of the methods that are often used for sampling from distributions for which the inverse-cdf method does not work, including the transformation methods, acceptance rejection methods, ratio-of-uniform methods, adaptive direction sampling (Gilks, 1992), and perfect sampling (Propp and Wilson, 1996).

1.4.1 Direct or Transformation Methods

Transformation methods are those based on transformations of random variables. Algorithms 1.1 and 1.2 provide such examples. Except for a few cases, including the exponential and Bernoulli distributions, Algorithms 1.1 and 1.2 are often inefficient. Better methods based on transformations can be obtained, depending on the target distribution $f(x)$. Table 1.2 provides a few examples that are commonly used in practice.

1.4.2 Acceptance-Rejection Methods

Acceptance-Rejection (AR), or Accept-Reject, methods are very useful for random number generation, in particular when direct methods do not exist

Table 1.2 Examples of transformation methods for random number generation.

method	transformation	distribution
Exponential	$X = -\ln(U)$	$X \sim \text{Expo}(1)$
Cauchy	$X = \tan(\pi U - \pi/2)$	$X \sim \text{Cauchy}(0,1)$
Box-Muller (1958)	$X_1 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$ $X_2 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$	$X_i \overset{iid}{\sim} N(0,1)$
Beta	$X_i \overset{ind}{\sim} \text{Gamma}(\alpha_i)$, $i = 1, 2$	$\frac{X_1}{X_1+X_2} \sim \text{Beta}(\alpha_1, \alpha_2)$

where $U \sim \text{Unif}(0,1)$ and $U_i \overset{ind}{\sim} \text{Unif}(0,1)$ for $i = 1, 2$.

or are computationally inefficient. We discuss the AR methods via a geometric argument.

Suppose that the distribution to sample from is of d-dimension with the sample space $\mathcal{X} \subseteq \mathbb{R}^d$. According to the definition of pdf (or pmf), the region under the pdf curve/surface

$$\mathbb{C}_f = \{(x, u) : 0 \leq u \leq f(x)\} \subset \mathbb{R}^{d+1} \tag{1.19}$$

has unit volume. Thus, if (X, U) is uniform in the region $\mathbb{C}_f$ then $X \sim f(x)$. Note that $X \sim f(x)$ still holds when $f(x)$ in (1.19) is multiplied by an arbitrary positive constant, that is,

$$\mathbb{C}_h = \{(x, y) : 0 \leq u \leq h(x)\} \subset \mathbb{R}^{d+1}, \tag{1.20}$$

where $h(x) \propto f(x)$, because rescaling on U will not affect the marginal distribution of X.

This fact suggests a possible way of generating X by simulating points distributed uniformly over $\mathbb{C}_f$ or $\mathbb{C}_h$. When it is difficult to sample from $\mathbb{C}_h$ directly, samples from $\mathbb{C}_h$ can be obtained indirectly by (i) generating points uniformly over an enlarged and easy-to-sample region $\mathbb{D} \supseteq \mathbb{C}_h$ and (ii) collecting those falling inside of $\mathbb{C}_h$. Such an enlarged region $\mathbb{D}$ can be constructed by an easy-to-sample distribution with pdf $g(x)$ with the restriction that $f(x)/g(x)$ is bounded from above by some finite constant M so that $\mathbb{C}_h$ can be enclosed in the region

$$\mathbb{C}_g = \{(x, u) : 0 \leq u \leq g(x)\} \subset \mathbb{R}^{d+1}, \tag{1.21}$$

for some $h(x) \propto f(x)$. The distribution $g(x)$ is called the *envelope* or instrumental distribution, while $f(x)$ the *target*.

To summarize, we have the following AR algorithm to generate random numbers from $f(x)$ using an envelope distribution $g(x)$, where $\sup_x h(x)/g(x) \leq M < \infty$.

Algorithm 1.3 *(Acceptance-Rejection)*

Repeat the following two steps until a value is returned in Step 2:

1. *Generate X from $g(x)$ and U from $Unif(0, 1)$.*
2. *If $U \leq \frac{h(X)}{Mg(X)}$, return X (as a random deviate from $f(x)$).*

The acceptance rate is the ratio of the volume of the target region to the volume of the proposal region, that is,

$$r = \frac{1}{M} \frac{\int h(x)dx}{\int g(x)dx}.$$

In the case when both $h(x)$ and $g(x)$ are normalized, the acceptance ratio is $1/M$, suggesting the use of $M = \sup_x h(x)/g(x)$ when it is simple to compute.

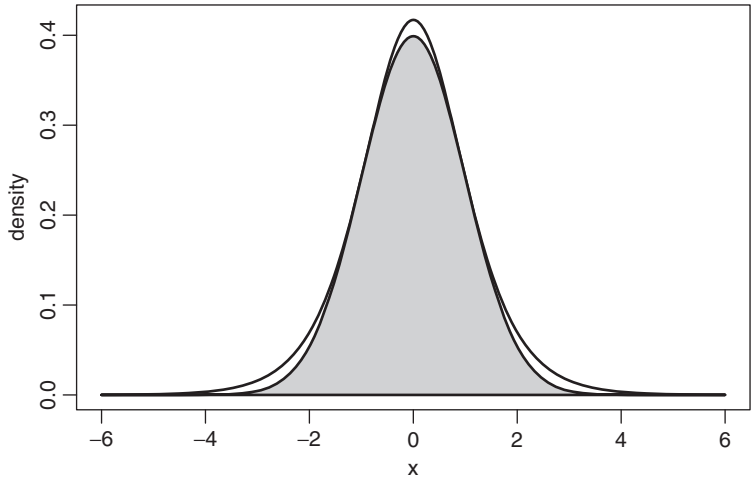

Figure 1.2 The target and proposal distributions $\phi(x)$ and $Mh_s(x)$ with $s = .648$ and $M = 1.081$ in Example 1.4.

■ **Example 1.4 The Standard Normal $N(0, 1)$**

This example illustrates the AR algorithm of using the logistic distribution with density

$$h_s(x) = \frac{e^{-x/s}}{s(1 + e^{-x/s})^2} = \frac{1}{4s\cosh^2(-x/s)} \qquad (-\infty < x < \infty)$$

where $s = .648$, as the proposal distribution to generate samples from the standard normal $N(0, 1)$. Note that $N(0, 1)$ has the pdf $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ $(-\infty < x < \infty)$. The maximum value

$$M = \max_{-\infty<x<\infty} \frac{\phi(x)}{h_s(x)},$$

obtained using the Newtow-Raphson method, is about 1.0808. Take $M = 1.081$. The target density and $Mh_s(x)$ are shown in Figure 1.2. The acceptance rate of the corresponding AR algorithm is given by $1/M = .925$.

When it is expensive to evaluate $h(x)$, an easy-to-compute function $s(x)$ $(0 \le s(x) \le h(x))$, called the squeeze function, can be used to reduce the frequency of computing $h(x)$. More specifically, the modified AR method is as follows.

Algorithm 1.4 *(Acceptance-Rejection With a Squeezer)*
Repeat the following two steps until a value is returned in Step 2:

1. *Generate X from $g(x)$ and U from Unif$(0, 1)$.*

2. *If* $U \leq \frac{s(X)}{Mg(X)}$ *or* $\frac{s(X)}{Mg(X)} < U \leq \frac{h(X)}{Mg(X)}$, *return* X *(as a random deviate from* $f(x)$*).*

Thus, in the case $U \leq \frac{s(X)}{Mg(X)}$, the algorithm does not evaluate $h(x)$.

1.4.3 The Ratio-of-Uniforms Method and Beyond

The ratio-of-uniforms method of Kinderman and Monahan (1977) is a popular method for random number generations of many standard distributions, including the Gamma, normal, and student-t. It can be considered as obtained from the rejection method via transformation subject to some kind of simplicity. Here we discuss the general idea behind the ratio-of-uniforms method and derive the method of Kinderman and Monahan (1977) and its extension proposed by Wakefield, Gelfand, and Smith (1991) as special cases.

The general idea of the ratio-of-uniforms method is to find a pair of differentiable transformations

$$U = u(Y) \quad \text{and} \quad X = x(Z, Y)$$

with $U = u(Y)$ strictly increasing to propagate the inequality in (1.20) and with constant Jacobian so that (Y, Z) is also uniform over the corresponding image of $\mathbb{C}_h$:

$$\mathbb{C}_h^{(Y,Z)} = \{(y, z) : u^{-1}(0) \leq y = u^{-1}(u) \leq u^{-1}(h(x(z, y)))\} \subset \mathbb{R}^{d+1}, \quad (1.22)$$

where $u^{-1}(.)$ denotes the inverse of $u(.)$. This leads to the following generic rejection algorithm to sample from $f(x)$ with a chosen easy-to-sample region $\mathbb{D}$ enclosing $\mathbb{C}_h^{(Y,Z)}$.

Algorithm 1.5 *(The Generic Acceptance-Rejection of Uniforms Algorithm)*

Repeat the following two steps until a value is returned in Step 2:

Step 1. Generate (Y, Z)*, uniform deviates over* $\mathbb{D} \supseteq \mathbb{C}_h^{(Y,Z)}$.

Step 2. If $(Y, Z) \in \mathbb{C}_h^{(Y,Z)}$*, return* $X = x(Y, Z)$ *as the desired deviate.*

This algorithm has the acceptance rate

$$r = \frac{\int_{\mathbb{C}_h^{(Y,Z)}} dydz}{\int_{\mathbb{D}} dydz} = \frac{\int_{\mathcal{X}} h(x)dx}{J_{x,u}(z, y)| \int_{\mathbb{D}} dydz},$$

where

$$J_{x,u}(z, y) = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial y} \\ 0 & \frac{\partial u}{\partial y} \end{vmatrix} = u'(y) \left| \frac{\partial x}{\partial z} \right|$$

denotes the Jacobian of the transformations.

It is state-of-the-art to choose the pair of transformation and construct $\mathbb{D}$. Thus, simplicity plays an important role. Let x_i be a function of z_i and y, for

example, the Jacobian of the transformation has a simple form

$$J_{x,u}(z,y) = u'(y)\left|\frac{\partial x}{\partial z}\right| = u'(y)\prod_{i=1}^{d}\left|\frac{\partial x_i(z_i,y)}{\partial z_i}\right| = \text{const.}$$

Hence, $x_i(z,y)$ is linear in z_i

$$x_i(z,y) = a_i(y)z_i + b_i(y) \qquad (i = 1,\ldots,d),$$

with the restriction $\prod_{i=1}^{d} a_i(y) = 1/u'(y)$. For example, the result of Wakefield *et al.* (1991) is obtained by letting $a_i(y) = [u'(y)]^{-1/d}$ and $b_i(y) = 0$ for $i = 1,\ldots,d$. The uniform region is

$$\{(y,z) : u^{-1}(0) \le y \le u^{-1}(h(z/[u'(y)]^{1/d}))\}.$$

The method of Wakefield *et al.* (1991) for generating multivariate distributions is obtained by taking the power transformation on Y: $u(y) = y^{r+1}/(r+1)$, $r \ge 0$. This results in the target region

$$\mathbb{C}_h^{(r)} = \left\{(y,z) : 0 \le y \le \left[(r+1)h\left(\frac{z}{y^{r/d}}\right)\right]^{1/(r+1)}\right\},$$

or equivalently

$$\mathbb{C}_h^{(r)} = \left\{(y,z) : 0 \le y \le \left[h\left(\frac{z}{y^{r/d}}\right)\right]^{1/(r+1)}\right\}, \tag{1.23}$$

because $h(x)$ is required to be known up to a proportionality constant. Wakefield *et al.* (1991) consider $(d+1)$-boxes $\mathbb{D}$ to bound $\mathbb{C}_h^{(r)}$, provided that $\sup_x h(x)$ and $\sup_x |x_i|[h(x)]^{r/(dr+d)}$, $i = 1,\ldots,d$, are all finite. Pérez *et al.* (2008) proposed to use ellipses as $\mathbb{D}$ in place of $(d+1)$-boxes.

In the univariate case, the algorithm based on (1.23) with the choice of $r = 1$ reduces to the famous ratio-of-uniforms method of Kinderman and Monahan (1977):

Algorithm 1.6 *(Ratio-of-Uniforms Algorithm of Kinderman and Monahan, 1977)*

Repeat the following two steps until a value is returned in Step 2:

1. *Generate (y,z) uniformly over $\mathbb{D} \supseteq \mathbb{C}_h^{(1)}$.*
2. *If $(Y,Z) \in \mathbb{C}_h^{(1)}$ return $X = Z/Y$ as the desired deviate.*

The uniform region is

$$\mathbb{C}_h^{(1)} = \left\{(y,z) : 0 \le y \le \left[h\left(\frac{z}{y}\right)\right]^{1/2}\right\}. \tag{1.24}$$

When $\sup_x h(x)$ and $\sup_x |x|[h(x)]^{1/2}$ are finite, the easy-to-sample bounding region $\mathbb{D}$ can be set to the tightest rectangle enclosing $\mathbb{C}_h^{(1)}$. For more efficient

algorithms, more refined enclosing regions such as polygons could be used. This is potentially useful in simulating truncated distributions. Here we provide an illustrative example, which shows that efficient ratio-of-uniforms methods can be derived by considering transformations of the random variable X, including the relocation technique proposed by Wakefield *et al.* (1991), before applying the ratio-of-uniforms.

■ **Example 1.5 Gamma(α) With $\alpha < 1$**

Consider the ratio-of-uniforms method for generating a variate X from the Gamma density $f_\alpha(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$, $x > 0$. To generate Gamma variates with density $f_\alpha(x)$, Kinderman and Monahan (1977) and Cheng and Feast (1979) used the ratio-of-uniforms method by setting $h(x) = x^{\alpha-1} e^{-x}$. The method is valid only for $\alpha > 1$ and gives a uniform region whose shape changes awkwardly when α is near 1. Cheng and Feast (1980) got around this problem by using the transformation $x = y^n$, $y > 0$, and setting $h(y) = y^{n\alpha-1} e^{-y^n}$. This effectively extends the range of α from 1 down to $1/n$. For more discussion on Gamma random number generators, see Tanizaki (2008) and the references therein.

We now use the transformation $X = e^{T/\alpha}$ (or $T = \alpha \ln X$), $-\infty < t < \infty$, for $\alpha < 1$. The random variable T has density $f(t) = e^{t - e^{t/\alpha}}/(\alpha\Gamma(\alpha))$. Let $h(t) = e^{t - e^{t/\alpha}}$. Then the uniform region is

$$\mathbb{C}_h^{(1)} = \left\{ (y, z) : 0 \le y \le e^{\left(t - e^{t/\alpha}\right)/2}, t = \frac{z}{y} \right\}.$$

Hence, y has the upper bound $\max_t [h(t)]^{1/2} = (\alpha e^{-1})^{\alpha/2}$. The upper bound $\max_{t>0} t[h(t)]^{1/2}$ of z requires us to solve the equation

$$\frac{1}{t} - \frac{1}{2} - \frac{1}{2} e^{t/\alpha} = 0 \tag{1.25}$$

for $t > 0$. As a simple alternative, a slightly loose upper bound can be obtained by making use the inequality $\ln(t/\alpha) \le t/\alpha - 1$. That is, $e^{t/\alpha} \ge (e^1 t)/\alpha$ and, thereby,

$$t[h(t)]^{1/2} = t e^{\frac{t}{2} - \frac{1}{2} e^{t/\alpha}} \le t e^{\frac{t}{2} - \frac{et}{2\alpha}} \le \frac{2\alpha}{e(e - \alpha)} \qquad (t > 0)$$

The lower bound of z also exists and requires one to solve (1.25) for $t < 0$. A simple lower bound is obtained as follows:

$$t[h(t)]^{1/2} = t e^{\frac{t}{2} - \frac{1}{2} e^{t/\alpha}} \ge t e^{\frac{t}{2}} \ge -\frac{2}{e} \qquad (t < 0).$$

Although it is not very tight for α near 1 and better ones can be found, this lower bound works pretty well, as indicated by the boundary plots in Figure 1.3 for a selected sequence of α values. The following experimental computer code written in R (`http://www.r-project.com`) demonstrates the simplicity of this method.

Algorithm 1.7

```
EXP1 = exp(1) # define the constant
rou.gamma(n, shape, log=FALSE){ #arguments: n = sample size, shape = α
        # log = flag; if log=TRUE, it returns the deviates in log scale
    if(shape<=0 || shape>=1) stop("shape is not in (0, 1)")
    if(shape <0.01 && !log)
        warning("It is recommended to set log=TRUE for shape < 0.01")
    y.max = (shape/EXP1)^(shape/2)
    z.min = -2/EXP1
    z.max = 2*shape/EXP1/(EXP1-shape)
    s = numeric(n) #allocate space for the generated desired deviates
    for(i in 1:n) {
        repeat {
           y = runif(1, 0, y.max)          # y ~ Unif(0,y.max)
           t = runif(1, z.min, z.max)/y    # t = z/y
           x = exp(t/shape)
           if(2*log(y) <= t-x){
              s[i] = if(log) t/shape else x
              break
           }
        }
    }
    return(s)
}
```

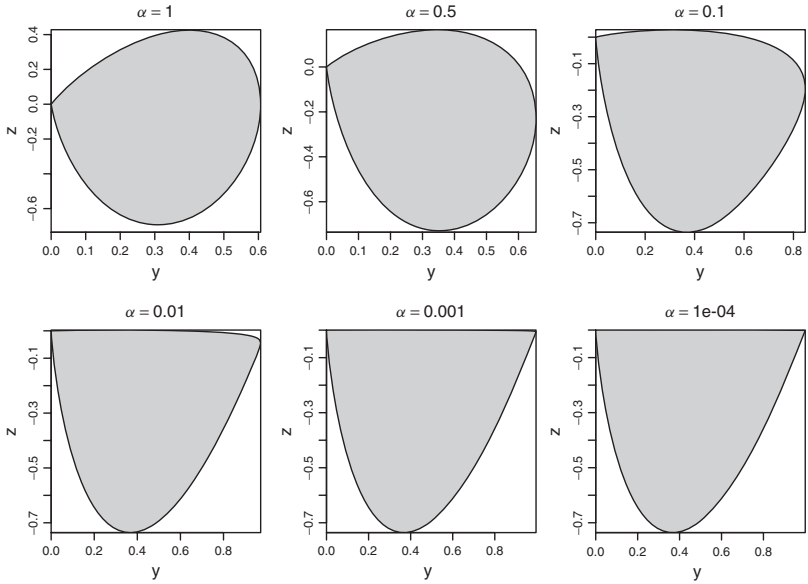

Figure 1.3 The uniform regions and their boundaries of the ratio-of-uniforms method for generating deviate X from Gamma(α), $\alpha < 1$, by letting $x = e^{t/\alpha}$, $-\infty < t < \infty$, and setting $h(t) = e^{t-e^{t/\alpha}}$.

We note that care must be taken in implementing Gamma deviate generators for small values of α. For example, we recommend in the above experimental code to return the output in a logarithmic scale when $\alpha < 0.01$.

1.4.4 Adaptive Rejection Sampling

Adaptive rejection sampling (ARS), introduced by Gilks (1992) (see also Gilks and Wild, 1992, and Wild, 1993), is a useful sampling method for log-concave densities. ARS works by constructing an envelope function of the log of the target density, which is then used in rejection sampling. For log-concave density functions, ARS is simple and efficient, especially when sampling from the same distributions occurs frequently. In this case, an adaptive squeezer can also be easily constructed to further improve its performance. For densities that are not log-concave, Gilks, Best, and Tan (1995) propose an Adaptive Metropolis sampling method.

1.4.5 Perfect Sampling

Propp and Wilson (1996) propose a perfect (exact) sampling MCMC algorithm, called *coupling from the past* (CFTP), for sampling from certain discrete distributions with finite number of states, for example, the Ising model. The algorithm uses a clever scheme to determine the time at which the Markov chain has reached its equilibrium. Later, it was extended by Murdoch and Green (1998) for sampling from a continuous state space. Fill (1998) proposes an alternative to CFTP, known as Fill's perfect rejection sampling algorithm. Interested readers are referred to the review paper by Djurić, Huang, and Ghirmai (2002), the website maintained by Wilson on perfect sampling, and the references therein (`http://www.dimacs.rutgers.edu/~/dbwilson/exact/`).

1.5 Markov Chain Monte Carlo

1.5.1 Markov Chains

When generating iid samples from the target distribution π is infeasible, dependent samples $\{X_i\}$ can be used instead, provided that the sample mean (1.17) converges to (1.16) at a satisfactory rate. A particular class of such dependent sequences that can be simulated is the class of Markov chains. A Markov chain, named after Andrey Markov, is a sequence of random variables $\{X_i : i = 0, 1, 2, \ldots\}$ with the Markov property that given the present state, the future and past states are independent, that is, for all measurable sets A in $\mathcal{X}$,

$$\Pr\left(X_{t+1} \in A | X_0 = x_0, \ldots, X_t = x_t\right) = \Pr\left(X_{t+1} \in A | X_t = x_t\right) \qquad (1.26)$$

holds for time $t = 0, 1, \ldots$

A convenient way of handling both discrete and continuous variables is to use the notation $\pi(dy)$ to denote the probability measure π on $(\mathbb{X}, \mathcal{X})$. For a continuous r.v. X, its pdf $f(x)$ is the Radon-Nikodym derivative of the probability measure $\pi(dx)$ with respect to the Lebesgue measure, while for discrete r.v. X, its pdf $f(x)$ is the derivative of $\pi(dx)$ with respect to the counting measure. Thus, we write $P_t(dx)$ for the marginal distribution of X_t over states $\mathcal{X}$ at time t. Starting with the distribution $P_0(dx)$, called the initial distribution, the Markov chain $\{X_t\}$ evolves according to

$$P_{t+1}(dy) = \int_{\mathbb{X}} P_t(dx)P_t(x, dy). \tag{1.27}$$

The distribution $P_t(x, dy)$ is the probability measure for X_{t+1} given $X_t = x$, called the transition kernel distribution at time t. In real life, this is the conditional density/mass function of X_{t+1} given $X_t = x$.

A primary class of Markov chains commonly used in MCMC is the class of time-homogeneous Markov chains or stationary Markov chains where

$$P_t(x, dy) = P(x, dy) \tag{1.28}$$

holds for $t = 1, 2, \ldots$. In this case, (1.27) becomes

$$P_{t+1}(dy) = \int_{\mathbb{X}} P_t(dx)P(x, dy) \tag{1.29}$$

and $P_t(dx)$ is uniquely determined by the initial distribution $P_0(dx)$ and the transition kernel $P(x, dy)$. For this reason, we write $P^n(x, .)$ for the conditional distribution of X_{t_0+n} given $X_{t_0} = x$. The basic idea of creating Markov chains for approximating $E_\pi(h(X))$ is to construct a transition kernel $P(x, dy)$ with $\pi(dx)$ as its invariant distribution, that is, $P(x, dy)$ and $\pi(dx)$ satisfy the *balance condition*

$$\pi(dy) = \int_{\mathbb{X}} \pi(dx)P(x, dy) \tag{1.30}$$

When the target distribution π has the density $f(x)$ and the transition kernel $P(x, dy)$ has the conditional density $p(y|x)$, this balance condition can be written as

$$f(y) = \int_{\mathbb{X}} p(y|x)f(x)dx.$$

The balance condition (1.30) can be viewed as obtained by requiring $P_{t+1}(dx) = P_t(dx) = \pi(dx)$ in (1.29). It says that if X_t is a draw from the target $\pi(x)$ then X_{t+1} is also a draw, possibly dependent on X_t, from $\pi(x)$. Moreover, for almost any $P_0(dx)$ under mild conditions $P_t(dx)$ converges to $\pi(dx)$. If for π-almost all x, $\lim_{t\to\infty} \Pr(X_t \in A|X_0 = x) = \pi(A)$ holds for all measurable sets A, $\pi(dx)$ is called the *equilibrium distribution* of the Markov chain. The relevant theoretical results, including those on the convergence behavior of the MCMC approximation $\bar{h}_n$ to $E(h(X))$, are summarized in Section 1.5.2.

1.5.2 Convergence Results

Except for rare cases where it is satisfactory to have one or few draws from the target distribution $f(x)$, most MH applications provide approximations to characteristics of $f(x)$, which can be represented by integrals of the form

$$E_\pi(h) = \int h(x)\pi(dx).$$

Assuming $E_\pi(|h|) < \infty$, this is to be approximated by

$$\bar{h}_{m,n} = \frac{1}{n}\sum_{i=1}^{n} h(X_{i+m}) \tag{1.31}$$

where $\{X_i\}$ denotes a simulated Markov chain and m is non-negative integer denoting the length of what is called the *burn-in* period; see Section 1.5.3.

This subsection includes theoretical results on (1.31) and is mainly based on Tierney (1994, Section 3). With some necessary preliminary theoretical results summarized in Section 1.5.2.2, Section 1.5.2.3 provides needed theoretical results concerning limiting behavior of (1.31) as $n \to \infty$ and m fixed. To proceed, Section 1.5.2.1 gives some key concepts and notations, in addition to those introduced in Section 1.5.1.

1.5.2.1 Notation and Definitions

The most important concepts are irreducibility and aperiodicity. A Markov chain with invariant distribution $\pi(dx)$ is *irreducible* if for any initial state, it has positive probability of entering any set to which $\pi(dx)$ assigns positive probability. A chain is *periodic* if there are portions of the state space $\mathcal{X}$ it can only visit at certain regularly spaced times; otherwise, the chain is *aperiodic*. A fundamental result is established in Theorem 1.5.1: If a chain has a proper invariant distribution $\pi(dx)$ and is irreducible and aperiodic, then $\pi(dx)$ is the unique invariant distribution and is also the equilibrium distribution of the chain.

Two additional crucial concepts in the theory of Markov chains are *recurrence* and *ergodicity*.

Definition 1.5.1 *Let X_n be a π-irreducible chain with invariant distribution $\pi(.)$ and let the notation $\{A_n \ i.o.\}$ mean that the sequence occurs infinitely often, that is, $\sum_i I_{A_i} = \infty$ (with probability one).*

(a) The chain is recurrent if, for every B with $\pi(B) > 0$,

$$Pr(X_n \in B \ i.o. | X_0 = x) > 0 \text{ for all } x$$

and

$$Pr(X_n \in B \ i.o. | X_0 = x) = 1 \text{ for } \pi\text{-almost all } x.$$

(b) The chain is Harris recurrent if $Pr(X_n \in B \ i.o. | X_0 = x) = 1$ for all x.

To define different forms of ergodicity, the total variation distance between two measures on $\mathcal{X}$ is used. The total variation distance between two measures on $(\mathbb{X}, \mathcal{X})$ is defined by the total variation norm of a signed measure λ on $(\mathbb{X}, \mathcal{X})$

$$\| \lambda \| = \sup_{A \in \mathcal{X}} \lambda(A) - \inf_{A \in \mathcal{X}} \lambda(A). \tag{1.32}$$

The concept of *hitting time* is also used. The *hitting time* of the subset $B \in \mathcal{X}$ is the random variable

$$H_B = \inf\{t \geq 0 : X_t \in B\}$$

where the infimum of the empty set is taken to be ∞.

Definition 1.5.2 *Different forms of ergodicity are as follows:*

(a) A Markov chain is said to be ergodic if it is positive Harris recurrent and aperiodic.

(b) Let H_B denote the hitting time for the set B. An ergodic chain with invariant distribution $\pi(x)$ is said to be ergodic of degree 2 if

$$\int_B E_x(H_B^2)\pi(dx) < \infty$$

holds for all $H \in \mathcal{X}$ with $\pi(H) > 0$.

(c) An ergodic chain with invariant distribution $\pi(x)$ is said to be geometrically ergodic if there exists a nonnegative extended real-valued function M with $E(|M(X)|) < \infty$ and a positive constant $r < 1$ such that

$$\| P^n(x, .) - \pi \| \leq M(x) r^n$$

for all x.

(d) The chain in (c) is said to be uniformly ergodic if there exist a constant M and a positive constant $r < 1$ such that

$$\| P^n(x, .) - \pi \| \leq M r^n.$$

1.5.2.2 Convergence of Distributions

The total variation distance between two measures on $(\mathbb{X}, \mathcal{X})$ is used to describe the convergence of a Markov chain in the following theorem (Theorem 1 of Tierney, 1994).

Theorem 1.5.1 *Suppose that $P(x, dy)$ is π-irreducible and π-invariant. Then $P(x, dy)$ is positive recurrent and $\pi(dx)$ is the unique invariant distribution of $P(x, dy)$. If $P(x, dy)$ is also aperiodic, then for π-almost all x,*

$$\| P^n(x, .) - \pi \| \to 0,$$

with $\| . \|$ denoting the total variation distance. If $P(x, dy)$ is Harris recurrent, then the convergence occurs for all x.

Tierney (1994) noted that the assumptions in Theorem 1.5.1 are essentially necessary and sufficient: if

$$||P_t(x,.) - \pi|| \to 0$$

for all x, then the chain is π-irreducible, aperiodic, positive Harris recurrent, and has the invariant distribution $\pi(dx)$. We refer to Tierney (1994) and Hernandez-Lerma and Lasserre (2001) for more discussion on sufficient conditions for Harris recurrence. Relevant theoretical results on the rate of convergence can also be found in Nummelin (1984), Chan (1989), and Tierney (1994).

1.5.2.3 Limiting Behavior of Averages

Tierney (1994) noted that a law of large numbers can be obtained from the ergodic theorem or the Chacon-Ornstein theorem. The following theorem is a corollary to Theorem 3.6 in Chapter 4 of Revuz (1975).

Theorem 1.5.2 *Suppose that X_n is ergodic with equilibrium distribution $f(x)$ and suppose $h(x)$ is real-valued and $E_f(|h(X)|) < \infty$. Then for any initial distribution, $\bar{h}_n \to E_f(h(X))$ almost surely.*

The central limit theorems that are available require more assumptions. Tierney (1994) gives the following central limit theorem.

Theorem 1.5.3 *Suppose that X_n is ergodic of degree 2 with equilibrium distribution $f(x)$ and suppose $h(x)$ is real-valued and bounded. Then there exists a real number σ_h such that the distribution of*

$$\sqrt{n}\left(\bar{h}_n - E_f(h(X))\right)$$

converges weakly to a normal distribution with mean 0 and variance σ_h^2 for any initial distribution.

The boundedness assumption on $h(x)$ can be removed if the chain is uniformly ergodic, provided $\mathrm{E}_f(h^2(X)) < \infty$.

Theorem 1.5.4 *Suppose that X_n is uniformly ergodic with equilibrium distribution $f(x)$ and suppose $h(x)$ is real-valued and $E_f(h^2(X)) < \infty$. Then there exists a real number σ_h such that the distribution of*

$$\sqrt{n}\left(\bar{h}_n - E_f(h(X))\right)$$

converges weakly to a normal distribution with mean 0 and variance σ_h^2 for any initial distribution.

1.5.3 Convergence Diagnostics

The theoretical results provide a useful guidance for designing practically valid and efficient MCMC (sampling) algorithms. It is difficult to make use of them to decide when it is safe to terminate MCMC algorithms and report MCMC approximations with their associated errors. More specifically, two critical issues arising in the context of computing $\bar{h}_{m,n}$ in (1.31) are (i) how to choose m, and (ii) how large should n be taken so that $\bar{h}_{m,n}$ has about the same precision as the corresponding Monte Carlo approximation based on an iid sample of a prespecified sample size, say, n_0. In theory, there are no definite answers to these two questions based on the simulated finite sequence $\{X_t : t = 0, \ldots, T\}$ because Markov chains can be trapped into local modes for arbitrarily long periods of time, if not indefinitely. While designing problem-specific efficient MCMC sampling algorithms is desirable, and is a major focus of this book, there have also been many proposed convergence diagnostic methods.

Gelman and Rubin (1992) is one of the most popular convergence diagnostic tools. The Gelman and Rubin method requires running multiple sequences $\{X_t^{(j)} : t = 0, 1, \ldots; j = 1, \ldots, J\}$, $J \geq 2$, with the starting sample $X_0^{(1)}, \ldots, X_0^{(J)}$ generated from an overdispersed estimate of the target distribution $\pi(dx)$. Let n be the length of each sequence after discarding the first half of the simulations. For each scalar estimand $\psi = \psi(X)$, write

$$\psi_i^{(j)} = \psi(X_i^{(j)}) \qquad (i = 1, \ldots, n; j = 1, \ldots, J).$$

Let

$$\bar{\psi}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \psi_i^{(j)} \quad \text{and} \quad \bar{\psi} = \frac{1}{J}\sum_{j=1}^{J} \bar{\psi}^{(j)},$$

for $j = 1, \ldots, J$. Then compute B and W, the between- and within-sequence variances:

$$B = \frac{n}{J-1}\sum_{j=1}^{J} \left(\bar{\psi}^{(j)} - \bar{\psi}\right)^2 \quad \text{and} \quad W = \frac{1}{J}\sum_{j=1}^{J} s_j^2,$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\psi_i^{(j)} - \bar{\psi}^{(j)}\right)^2 \qquad (j = 1, \ldots, J).$$

Suppose that the target distribution of ψ is approximately normal and assume that the jumps of the Markov chains are local, as is often the case in practical iterative simulations. For any finite n, the within variance W underestimates the variance of ψ, σ_ψ^2; while the between variance B overestimates σ_ψ^2. In the limit as $n \to \infty$, the expectations of both B and W

approach σ_ψ^2. Thus, the Gelman and Rubin reduction coefficient,

$$\sqrt{\hat{R}} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}}, \tag{1.33}$$

should be expected to decline to one as $n \to \infty$. Gelman *et al.* (2004) recommend computing the reduction coefficient $\sqrt{\hat{R}}$ for all scalar estimands of interest; if $\sqrt{\hat{R}}$ is not near one for all of them, continue the simulation runs. Once $\sqrt{\hat{R}}$ is near one for all scalar estimands of interest, just collect the $J \times n$ samples from the second halves of the J sequences together and treat them as (dependent) samples from the target distribution $\pi(dx)$.

A number of criticisms of the GR method have been made in the literature. Readers are referred to Cowles and Carlin (1996) for these criticisms. Ideas of constructing overdispersed starting values can be found in Gelman and Rubin (1992) and Liu and Rubin (1996). For a review of other convergence diagnostic methods, see Cowles and Carlin (1996), Brooks and Gelman (1998), Mengersen *et al.* (1999), and Plummer *et al.* (2006) and the references therein.

Exercises

1.1 Suppose that a single observation $X = 2.0$ is considered to have been drawn from the Gaussian model $N(\theta, 1)$ with unknown θ. Consider the hypothesis $H_0 : \theta = 0$ versus the alternative hypothesis $H_a : \theta \neq 0$. Apply the Bayes approach using Bayes factors.

1.2 Consider inference about the binomial proportion θ in Binomial(n, θ) from an observed count X.

- **(a)** Show that the Jeffreys prior for the binomial proportion θ is the Beta distribution Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$.
- **(b)** Derive the posterior $\pi(\theta|X)$.
- **(c)** For the case of $n = 1$, evaluate the frequency properties of the 95% credible interval for each of $\theta = .0001, 0.001, 0.01, 0.1, .25, .5, .75$, .9, .99, .999, and .9999.

1.3 Suppose that the sample density function of a single observation $X \in \mathcal{R}$ has the density of the form $f(x-\theta)$, where $\theta \in \mathcal{R}$ is unknown parameter to be estimated.

- **(a)** Show that the Jeffreys prior is $\pi(\theta) \propto 1$.
- **(b)** Consider frequency properties of one-sided credible intervals.
- **(c)** Discuss the case where θ is known to be on a sub-interval of $\mathcal{R}$.

1.4 Prove Theorem 1.3.1.

1.5 Verify the results in Table 1.2.

1.6 Extend the method of generating Beta to simulate the Dirichlet random variable.

1.7 Consider the problem of generating Poisson variates.

- **(a)** Design an efficient discrete inverse-cdf algorithm using a mapping.
- **(b)** Develop an Acceptance-Rejection method with a continuous envelope distribution.
- **(c)** Investigate efficient ratio-of-uniforms methods; see Stadlober (1989).

1.8 Consider the problem of Example 1.5, that is, to generate a deviate from the Gamma distribution with shape parameter less than one.

- **(a)** Create a mixture distribution as envelope for implementing the Acceptance-Rejection algorithm.
- **(b)** Compare your methods against the ratio-of-uniforms methods in Example 1.5.

1.9 Develop the ratio-of-uniforms method for generating Gamma deviates (for all $\alpha > 0$) by replacing the transformation $T = \alpha \ln X$ in the problem of Example 1.5 with $T = \sqrt{\alpha} \ln \frac{X}{\alpha}$.

1.10 The standard student-t distribution with ν degrees of freedom has density

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} \qquad (-\infty < x < \infty)$$

where ν is the number of degrees of freedom and $\Gamma(.)$ is the Gamma function.

- **(a)** Implement the ratio-of-uniforms method for generating random deviates from $f_\nu(x)$.
- **(b)** Develop an efficient ratio-of-uniforms algorithm to generate random variables from interval-truncated Student-t distribution.

1.11 This is the same as Exercise 1.10 but for the standard normal distribution $N(0, 1)$.

1.12 Suppose that $D = \{y_i = (y_{1i}, y_{2i})' : i = 1, \ldots, n\}$ is a random sample from a bivariate normal distribution $N_2(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

with unknown correlation coefficient $\rho \in (-1, 1)$ (Pérez *et al.*, 2008)

- **(a)** Assuming the prior $\pi(\rho) \propto 1/(1 - \rho^2)$, derive the posterior distribution $\pi(\rho|D)$.

(b) Implement the ratio-of-uniforms method to generate ρ from $\pi(\rho|D)$.

(c) Implement the ratio-of-uniforms method to generate η from $\pi(\eta|D)$, which is obtained from $\pi(\rho|D)$ via the one-to-one transformation $\eta = \ln \frac{1+\rho}{1-\rho}$.

(d) Conduct a simulation study to compare the two implementations in **(b)** and **(c)**.

1.13 Consider the simple random walk Markov chain with two reflecting boundaries on the space $\mathbb{X} = \{a, a+1, \ldots, b\}$ with the transition kernel distribution (matrix) $P = (p_{ij})$, where

$$p_{ij} = \Pr(X_{t+1} = j | X_t = i) = \begin{cases} p & \text{if } j = i+1 \text{ and } a < i < b; \\ q & \text{if } j = i-1 \text{ and } a < i < b; \\ 1 & \text{if } i = a \text{ and } j = a+1; \\ 1 & \text{if } i = b \text{ and } j = b-1, \end{cases}$$

with $0 < p < 1$ and $q = 1 - p$.

(a) Find the invariant distribution π.

(b) Show that the invariant distribution π is also the equilibrium distribution of the Markov chain.

1.14 Let π_i $(i = 1, 2)$ be the probability measure for $N(\mu_i, 1)$. Find the total variation distance between π_1 and π_2.

Hint: Let $\lambda = \pi_2 - \pi_1$ and let $\phi(x - \mu_i)$ be the density of π_i for $i = 1$ and 2.

Then $\sup_A \lambda(A) = \inf_{\phi(x-\mu_2)-\phi(x-\mu_1)>0} [\phi(x - \mu_2) - \phi(x - \mu_1)]\, dx$.