13
Selected Theoretical Topics

Topics in this chapter include the covariance analysis of iterative conditional sampling; the comparison of Metropolis algorithms based on Peskun's theorem; the eigen-analysis of the independence sampler; perfect simulation, convergence diagnostics, and a theory for dynamic weighting. The interested reader is encouraged to read the related literature for more detailed analyses.

13.1 MCMC Convergence and Convergence Diagnostics

When the state space of the MCMC sampler is finite, theorems in Chapter 12 can be used to judge their convergence. Two key concepts are the irreducibility and aperiodicity. These two concepts can be generalized to continuous state spaces and be used to prove convergence. One of such theorem is described in (Tierney 1994), who used the techniques developed in Nummelin (1984) for the proof.

Theorem 13.1.1 (Tierney) *Suppose A is π-irreducible and $\pi A = \pi$. Then A is positive recurrent and π is the unique invariant distribution of A. If A is also aperiodic, then, for all x but a subset whose measure under π is zero (i.e., π-almost all $\mathbf{x}$),*

$$\|A^{(n)}(\mathbf{x}, \cdot) - \pi\|_{\text{var}} \to 0, \tag{13.1}$$

where $\|\cdot\|_{\text{var}}$ denote the total variation distance.

Clearly, all the MCMC algorithms covered in this book have an invariant measure π. In most cases they are also π-irreducible and aperiodic. Hence, their convergence can be "assured," except that we still have no idea how fast they converge.

Our view on the convergence diagnosis issue concurs with that of Cowles and Carlin (1996): A combination of Gelman and Rubin (1992) and Geyer (1992) can usually provide an effective, yet simple, method for monitoring convergence in MCMC sampling. Many other approaches, which typically consume a few times more computing resources, can only provide marginal improvements. The *coupling from the past* (CFTP) algorithm described in Section 13.5 is an exciting theoretical breakthrough and its value in assessing convergence of MCMC schemes has been noticed. However, the method is still not quite ready for a routine use in MCMC computation. Interested reader may find Mengersen, Robert and Cuihenneuc-Jouyaux (1999) a useful reference, which provided an extensive study on convergence diagnostics.

Based on a normal theory approximation to the target distribution $\pi(\mathbf{x})$, Gelman and Rubin (1992) propose a method that involves the following few steps:

1. Before sampling begins, obtain a simple "trial" distribution $f(\mathbf{x})$ which is overdispersed relatively to the target distribution π. Generate m (say, 10) i.i.d. samples from $f(\mathbf{x})$. A side note is that in high-dimensional problems, it is often not so easy to find a suitable over-dispersed starting distribution.

2. Start m independent samplers with their respective initial states being the ones obtained in Step 1. Run each chain for $2n$ iterations.

3. For a scalar quantity of interest (after appropriate transformation to approximate normality), say $\theta = \theta(\mathbf{x})$, we use the sample from the last n iterations to compute W, the average of m *within-chain* variances, and B, the variance between the means θ from the m parallel chains.

4. Compute the "shrink factor"

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}\right)\frac{\mathrm{df}}{\mathrm{df}-2}}$$

   Here df refers to the degree of freedom in a t-distribution approximation to the empirical distribution of θ.

Gelman and Rubin (1992) suggested using $\theta = \log \pi(\mathbf{x})$ as a general diagnosis benchmark. Other choices of θ have been reviewed in Cowles and Carlin (1996). Geyer's (1992) main criticism to Gelman and Rubin's approach is that for difficult MCMC computation, one should concentrate all the resources to a single chain iteration: The latter 9000 samples from a single run of 10,000 iterations are much more likely to come from the target distribution π than those samples from 10 parallel runs of 1000 iterations. In addition, good convergence criterion such as the integrated autocorrelation time (Section 5.8) used in physics can be produced with a single chain.

Concerning the generic use of MCMC methods, we advocate a variety of diagnostic tools rather than any single plot or statistic. In our own work, we often run a few (three to five) parallel chains with relatively scattered starting states. Then, we inspect these chains by comparing many of their aspects, such as the histogram of some parameters, autocorrelation plots, and Gelman and Rubin's $\hat{R}$.

13.2 Iterative Conditional Sampling

13.2.1 Data augmentation

It has been shown in Sections 6.4 and 6.6.1 that *data augmentation* (DA) can be viewed as a two-component Gibbs sampler with a *deterministic scan*. The transition function from $\mathbf{x}^{(0)}$ to $\mathbf{x}^{(1)}$ in DA is, then,

$$A(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) = \pi(x_1^{(1)} \mid x_2^{(0)})\pi(x_2^{(1)} \mid x_1^{(1)}), \tag{13.2}$$

where $\mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)})$. An important result proved in Section 6.6.1 is the expression (6.6) for the one-lag autocovariance of DA. Here we present more details about the theory developed in Liu, Wong, and Kong (1994, 1995).

From Figure 6.2 and the proof of Theorem 6.6.1, we make the following two observations:

(i) The *marginal chains*, $\{x_1^{(t)}, t = 1, 2, \ldots,\}$ and $\{x_2^{(t)}, t = 1, 2, \ldots,\}$, are all reversible Markov chains. In particular, the transition function for, say, the first chain is

$$A_1(x_1^{(0)}, x_1^{(1)}) = \int \pi(x_1^{(1)} \mid x_2^{(0)})\pi(x_2^{(0)} \mid x_1^{(0)})dx_2^{(0)}. \tag{13.3}$$

(ii) The two marginal chains have the *interleaving Markov property* defined below.

Definition 13.2.1 *A stationary Markov chain $\{x^{(t)}, t = 1, 2, \ldots,\}$ is said to have the interleaving Markov property if there exists a conjugate Markov chain $\{y^{(t)}, t = 1, 2, \ldots,\}$ such that*

(a) $x^{(t)}$ and $x^{(t+1)}$ are conditionally independent given $y^{(t)}$, $\forall t$,

(b) $y^{(t)}$ and $y^{(t+1)}$ are conditionally independent given $x^{(t+1)}$, $\forall k$,

(c) $(y^{(t-1)}, x^{(t)})$, $(x^{(t)}, y^{(t)})$ and $(y^{(t)}, x^{(t+1)})$ are identically distributed.

The two chains are said to be mutually interleaving. The interleaving property implies the reversibility of both chains.

Lemma 13.2.1 *The marginal chains $\{x_1^{(t)}\}$ and $\{x_2^{(t)}\}$ constructed in data augmentation are mutually interleaving.*

Consider the forward operator for the marginal chain $\{x_1^{(t)}\}$. From (13.3), we have

$$\begin{aligned} F_1 h(x_1) &= \int h(x_1^{(1)}) A_1(x_1, x_1^{(1)}) dx_1^{(1)} \\ &= E_\pi[E_\pi\{h(x_1)|x_2\}|x_1] \end{aligned}$$

If we let γ_0 be the *maximal correlation* between x_1 and x_2 under π, then the norm $\|F_1\|$ is γ_0^2. Using expressing (12.21), we can write

$$\|F_1\|^{1/2} = \gamma_0 = \sup_{h \in L_0^2(\pi)} \text{var}[E\{h(x_1) \mid x_2\}].$$

Similarly, the norm of F_2 for the companion chain is also γ_0^2. Since both marginal chains are reversible, the two forward operators are self-adjoint. Thus, their norms are equal to their spectral radii and are equal to each other. This means that the two chains *converge* at the same speed. On the other hand, the joint forward operator F has the property

$$Fh(\mathbf{x}^{(0)}) = \int\int h(x_1^{(1)}, x_2^{(1)}) \pi(x_2^{(1)} \mid x_1^{(1)}) \pi(x_1^{(1)} \mid x_2^{(0)}) dx_1^{(1)} dx_2^{(1)},$$

whose norm is shown to be $\|F\| = \gamma_0$. It is not difficult to show that the spectral radius of F is also equal to γ_0^2, as expected.

The above discussion has two implications: (a) One need only to establish convergence properties of one of the marginal chains in order for the joint chain to converge; (b) the convergence rate of data augmentation is completely determined by the maximal correlation between the two components. In statistical missing problems, one component often corresponds to missing data, $\mathbf{y}_{\text{mis}}$, and the other to the parameter θ. The maximal correlation between these two components reflects the "maximal fraction of missing information" (Little and Rubin 1987, Liu 1994b), defined as

$$r = \max_{0 < \text{var}_\pi(h) < \infty} \frac{\text{var}[E\{h(\theta) \mid \mathbf{y}_{\text{mis}}\}]}{\text{var}_\pi\{h(\theta)\}}.$$

Hence, the more the fraction of missing information, the larger the maximal correlation between θ and $\mathbf{y}_{\text{mis}}$ and the slower the corresponding data augmentation scheme. On the other hand, because of this duality, we can also estimate the *maximal faction* of missing information in a missing data problem from the output of its data augmentation scheme (Liu 1994b).

Furthermore, because

$$1 = \frac{E[\text{var}\{h(\theta) \mid \mathbf{y}_{\text{mis}}\}]}{\text{var}_\pi\{h(\theta)\}} + \frac{\text{var}[E\{h(\theta) \mid \mathbf{y}_{\text{mis}}\}]}{\text{var}_\pi\{h(\theta)\}},$$

a larger fraction of $E[\text{var}\{h(\theta) \mid \mathbf{y}_{\text{mis}}\}]$ in the total variance means that x_1 can move more freely conditional on x_2 and, hence, a faster scheme. Thus, when giving two data augmentation schemes with the same target distribution for θ, we prefer the scheme that gives a larger conditional variance for $h(\theta)$, for all functions $h(\ )$ (Liu and Wu 1999, Meng and van Dyk 1999).

13.2.2 Random-scan Gibbs sampler

As shown in Lemma 6.6.1 of Section 6.6.2, the random-scan Gibbs sampler (RSGS) has a similar expression for its first-order autocorrelation to that of data augmentation. Thus, the Markov chain produced by the RSGS also has the interleaving property, with its conjugate process $(\mathbf{i}^{(t)}, \mathbf{x}^{(t)}_{[-\mathbf{i}^{(t)}]})$, for $t = 1, 2, \ldots$. Here, $\mathbf{i}^{(t)}$ is the random variable that indicates which of the d components is updated at the t-th iteration.

The geometric convergence property of the RSGS process is not very difficult to prove and the interested reader is referred to Liu, Wong and Kong (1995) and Schervish and Carlin (1992). What is a little new here is an expression to bound the convergence rate of the RSGS. Based on the theory discussed in the previous subsection, we have, for any $\|h\| = 1$,

$$\begin{aligned} \|Fh\|^2 &= 1 - E[\text{var}\{h(\mathbf{x}) \mid \mathbf{i}, \mathbf{x}_{[-\mathbf{i}]}\}] \\ &= 1 - \sum_{i=1}^{d} \alpha_i E[\text{var}\{h(\mathbf{x}) \mid \mathbf{x}_{[-i]}\}] \\ &= \sum_{i=1}^{d} \alpha_i \text{var}[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-i]}\}]. \end{aligned} \tag{13.4}$$

Suppose we can find a pair of functions $h_i(x_i)$, and $g_i(\mathbf{x}_{[-i]})$ with unit variance such that

$$\begin{aligned} E\{h_i(x_i) \mid \mathbf{x}_{[-i]}\} &= \gamma_i g_i(\mathbf{x}_{[-i]}), \\ E\{g_i(\mathbf{x}_{[-i]} \mid x_i\} &= \gamma_i h_i(x_i). \end{aligned}$$

Then, using (13.4) with h replaced by h_i, we have

$$\|F\|^2 \geq \langle Fh_i, h_i \rangle = 1 - \alpha_i(1 - \gamma_i^2).$$

Letting $i = 1, \ldots, d$, we have

$$\|F\|^2 \geq \max_i \{1 - \alpha_i(1 - \gamma_i^2)\}.$$

Heuristically, we may want to find the vector $(\alpha_1, \ldots, \alpha_d)$ so that this lower bound for the convergence rate is minimized. This is equivalent to finding

$$\max_{\boldsymbol{\alpha}} \left[\min_i \{\alpha_i (1 - \gamma_i^2)\} \right].$$

It is easy to see that the solution is $\alpha_i \propto (1 - \gamma_i^2)^{-1}$. This result is rather intuitive: In order to achieve a better convergence rate, one should spend more resources (number of updates roughly proportional to the inverse of the spectral gap) on those "stickier" components. On the other hand, if one is interested in estimating a particular expectation, $Ef(\mathbf{x})$, after the chain becomes stationary, one should allocate resources differently.

We can also use (13.4) to show that *grouping* and *collapsing* (Section 6.7) in a random scan are always preferable (a stronger result than that for the deterministic scans).

Theorem 13.2.1 *Suppose we can either group the first two components together or integrate out the first component so as to result in a RSGS with $d-1$ components. We also assume that the scheduling probability α_i remains unchanged for $i = 3, \ldots, d$. Then, the collapsing sampler converges faster than the grouping sampler, and the grouping sampler faster than the original RSGS.*

Proof: We directly compare (13.4) for the three samplers. For grouping,

$$\begin{aligned} \|F_g h\|^2 &= (\alpha_1 + \alpha_2)\mathrm{var}[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-1,-2]}\}] \\ &\quad + \sum_{i=3}^{d} \alpha_i \mathrm{var}[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-i]}\}]; \end{aligned}$$

for collapsing,

$$\begin{aligned} \|F_c h\|^2 &= (\alpha_1 + \alpha_2)\mathrm{var}[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-1,-2]}\}] \\ &\quad + \sum_{i=3}^{d} \alpha_i \mathrm{var}[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-1,-i]}\}]. \end{aligned}$$

For notational simplicity, here we take the test function for the collapsed sampler, which only has $d-1$ components, as $E[g(\mathbf{x}) \mid \mathbf{x}_{[-1]}]$. Because

$$E\{h(\mathbf{x}) \mid \mathbf{x}_{[-1,-i]}\} = E[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-i]}\} \mid \mathbf{x}_{[-1]}],$$

it follows that

$$\mathrm{var}[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-1,-i]}\}] \leq \mathrm{var}[E\{h(\mathbf{x}) \mid \mathbf{x}_{[-i]}\}],$$

for $i = 2, \ldots, d$. Hence, the theorem is proved. $\diamond$

13.3 Comparison of Metropolis-Type Algorithms

13.3.1 Peskun's ordering

As an alternative to the acceptance-rejection criterion of Metropolis et al. (1953), Barker (1965) proposes a more "continuous" acceptance function

$$r_B(\mathbf{x}, \mathbf{y}) = \frac{\pi(\mathbf{y})T(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{x})T(\mathbf{x}, \mathbf{y}) + \pi(\mathbf{y})T(\mathbf{y}, \mathbf{x})};$$

that is, one accepts the proposed $\mathbf{y}$ with probability r_B and rejects the proposal with probability $1 - r_B$ (Section 5.2). To understand whether Barker's proposal has any advantage, Peskun (1973) introduced a *partial ordering* among all finite-state reversible Markov transition matrices that have the same stationary distribution.

Definition 13.3.1 *Suppose two reversible transition kernels A_1 and A_2 have the same stationary distribution π. Matrix A_1 is said to dominate A_2 (i.e., $A_1 \succeq A_2$) if all the off-diagonal elements of A_1 is greater than or equal to the corresponding elements in A_2. This definition is generalized by Tierney (1998) as follows:*

$$P_1\left[\mathbf{x}^{(1)} \in S \setminus \{\mathbf{x}\} \mid \mathbf{x}^{(0)} = \mathbf{x}\right] \geq P_2\left[\mathbf{x}^{(1)} \in S \setminus \{\mathbf{x}\} \mid \mathbf{x}^{(0)} = \mathbf{x}\right]$$

for all measurable subset $S \subseteq \mathcal{X}$.

This dominance condition easily leads to a comparison between the lag-1 autocorrelations of two Markov chains.

Lemma 13.3.1 (Tierney) *Suppose A_1 and A_2 have the same invariant distribution π and satisfy $A_1 \succeq A_2$. Then, the corresponding forward operators, F_1 and F_2, satisfy*

$$\langle (F_2 - F_1)f, f \rangle \geq 0$$

for all $f \in L_0^2(\pi)$.

Proof: Here, we only prove the case when the state space is finite. Please refer to Tierney (1998) for a more general proof. Suppose the total number of states is N. In this case, we can express the target distribution as the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ and the transition function as matrices. Then, any function $f \in L_0^2(\pi)$ can be expressed as a column vector of length N, $f = (f_1, \ldots, f_N)^T$, and $F_1 f$ is simply equal to the matrix product $A_1 f$. We define an $N \times N$ matrix as

$$H = \Delta(I + A_1 - A_2),$$

where I is the identity matrix and $\Delta = \mathrm{diag}(\pi_1, \ldots, \pi_N)$. Because A_1 dominates A_2, it is easy to check that H is a probability measure on the

product space (all entries h_{ij} are non-negative and they sum to 1) and the both marginals of H are equal to π. Hence,

$$\begin{aligned}\langle (A_2 - A_1)f, f\rangle &= |!f^T \Delta(A_2 - A_1)f \\ &= f^T\{\Delta - H\}f \\ &= \sum_i f_i^2\pi_i - \sum_i\sum_j f_i f_j h_{ij}, \\ &= \frac{1}{2}\left[2\sum_i\sum_j f_i^2 h_{ij} - 2\sum_i\sum_j f_i f_j h_{ij}\right] \\ &= \frac{1}{2}\sum_i\sum_j (f_i - f_j)^2 h_{ij} \geq 0.\end{aligned}$$

$\diamond$

Suppose that we are interested in estimating $E_\pi f$ by using a MCMC sampler whose transition kernel is $A(\mathbf{x}, \mathbf{y})$. We can define the sampler's *asymptotic efficiency* as

$$v(f, A) = \lim_{n\to\infty} \frac{1}{n}\text{var}\left\{\sum_{t=1}^{n} f(\mathbf{x}^{(t)})\right\}, \tag{13.5}$$

where $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ are stationary samples obtained from this sampler. Since we are ultimately interested in using a MCMC sampler to compute quantities of interest, this asymptotic efficiency measure seems to be a sensible criterion in comparing different schemes. Peskun (1973) proved that if $A_1 \succeq A_2$, then the first chain will be asymptotically more efficient.

Theorem 13.3.1 (Peskun) *Suppose A_1 and A_2 are reversible transition kernels with the same invariant distribution and $A_1 \succeq A_2$. Then, for all $f \in L_0^2(\pi)$ (i.e., mean zero functions), we have $v(f, A_1) \leq v(f, A_2)$.*

Proof: It is easy to see that for any transition matrix A,

$$v(f, A) = \langle f, \{I + (I - A)^{-1}\}f\rangle.$$

Note that operator $(I - A)$ is invertible only in the restricted space $L_0^2(\pi)$, but not in the unrestricted space $L^2(\pi)$. Define $A(\beta) = (1 - \beta)A_1 + \beta A_2$; then,

$$\begin{aligned}\frac{\partial v(f, A(\beta))}{\partial\lambda} &= \langle f, (I - A(\beta))^{-1}(A_2 - A_1)(I - A(\beta))^{-1} f\rangle \\ &= \langle (I - A(\beta))^{-1}\}f, (A_2 - A_1)(I - A(\beta))^{-1} f\rangle \geq 0\end{aligned}$$

The second equality follows from the fact that if A is a self-adjoint operation, then all powers of A, and thus, $(I - A)^{-1}$, are also self-adjoint operators. The last inequality follows from Lemma 13.3.1. Hence, $v(f, A(\beta))$ is

a monotone nondecreasing function in β and attains its minimum at $\beta = 0$ and maximum at $\beta = 1$. $\diamond$

13.3.2 Comparing schemes using Peskun's ordering

The ordering among transition functions introduced by Peskun is very useful for comparing different schemes. Based on Theorem 13.3.1, for example, Peskun (1973) proved the following theorem.

Theorem 13.3.2 *For the same proposal transition $T(\mathbf{x}, \mathbf{y})$, the acceptance function suggested by Metropolis et al. dominates that proposed by Barker in terms of asymptotic efficiency (13.5).*

Proof: The transition function of the Metropolis algorithm is

$$A_M(\mathbf{x}, \mathbf{y}) = T(\mathbf{x}, \mathbf{y})\min\{1, r(\mathbf{x}, \mathbf{y})\} \quad \text{for} \quad \mathbf{x} \neq \mathbf{y},$$

where

$$r(\mathbf{x}, \mathbf{y}) = \frac{\pi(\mathbf{y})T(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})T(\mathbf{x}, \mathbf{y})}.$$

On the other hand, the transition function for Barker's scheme is

$$A_B(\mathbf{x}, \mathbf{y}) = T(\mathbf{x}, \mathbf{y})\frac{r(\mathbf{x}, \mathbf{y})}{1 + r(\mathbf{x}, \mathbf{y})} \quad \text{for} \quad \mathbf{x} \neq \mathbf{y}.$$

It is easy to show that

$$\min\{1, r(\mathbf{x}, \mathbf{y})\} \geq \frac{r(\mathbf{x}, \mathbf{y})}{1 + r(\mathbf{x}, \mathbf{y})}.$$

Hence, $A_M \succeq A_B$. $\diamond$

As a generalization of Peskun's result, Liu (1996a) showed that the "Metropolization" of the Gibbs sampler for a finite state space as described in Section 6.3.2 dominates the usual random-scan Gibbs sampler.

Theorem 13.3.3 *Suppose that $\mathbf{x} = (x_1, \ldots, x_d)$, where x_i takes $m_i < \infty$ possible values, and that $\pi(\mathbf{x})$ is the distribution of interest. Then, the Metropolized Gibbs sampler defined in Section 6.3.2 for discrete random variables is statistically more efficient than the random-scan Gibbs sampler.*

Proof: Suppose that in the random-scan Gibbs sampler, we choose each component with probability α_i. Then, all the nonzero elements of the transition matrix P_1 of the random scan Gibbs sampler are of the form

$$P_1(\mathbf{x}, \mathbf{y}) = \alpha_i \pi(y_i \mid \mathbf{x}_{[-i]}),$$

where $\mathbf{y} = \mathbf{x}$ except that y_i replaces x_i. In contrast, those nonzero off-diagonal elements in the transition matrix P_2 of the modified sampler are

$$P_2(\mathbf{x}, \mathbf{y}) = \alpha_i \min\left\{ \frac{\pi(y_i \mid \mathbf{x}_{[-i]})}{1 - \pi(x_i \mid \mathbf{x}_{[-i]})}, \frac{\pi(y_i \mid \mathbf{x}_{[-i]})}{1 - \pi(y_i \mid \mathbf{x}_{[-i]})} \right\}.$$

Clearly, $P_2 \succeq P_1$. $\diamond$

Although Theorem 13.3.3 does not even require m_i to be finite, the modification is likely to be most useful for components with m_i rather small. It is easily shown from inequality $v(f, P_1) \geq v(f, P_2)$, $\forall f \in L^2(\pi)$, that the second largest eigenvalue of P_1 is greater than or equal to that of P_2. Frigessi et al. (1993) proved that for the binary Ising model, Metropolis converges faster than Gibbs for strong interaction and more slowly for weak interaction. This does not conflict with our result, which concerns statistical efficiency in equilibrium, rather than rate of convergence. Whereas the eigenvalues of the Gibbs sampler are necessarily non-negative (Liu, Wong and Kong 1995), slow Metropolis convergence under weak interaction is the product of a large negative eigenvalue.

Using the same technique, Besag et al. (1995) and Tierney (1998) proved another interesting result regarding the use of a mixture proposal in a Metropolis sampler. Let T_i be a sequence of proposal kernels and let $\alpha_i > 0$ with $\sum_i \alpha_i = 1$. Let A_i be the corresponding Metropolis transition kernel:

$$A_i(\mathbf{x}, \mathbf{y}) = T_i(\mathbf{x}, \mathbf{y}) \min\left\{ 1, \frac{\pi(\mathbf{y}) T_i(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x}) T_i(\mathbf{x}, \mathbf{y})} \right\}.$$

Furthermore, we define a mixture proposal

$$T^*(\mathbf{x}, \mathbf{y}) = \sum_i \alpha_i T_i(\mathbf{x}, \mathbf{y})$$

and its corresponding Metropolis transition function $A^*(\mathbf{x}, \mathbf{y})$.

Theorem 13.3.4 *The Metropolis transition with a mixture proposal dominates the corresponding mixture of Metropolis transitions; that is,*

$$A^* \succeq \sum_i \alpha_i A_i$$

Proof: Because of the simple inequality

$$\min(A_1, B_1) + \min(A_2, B_2) \leq \min(A_1 + A_2, B_1 + B_2),$$

we have that

$$\begin{aligned} \sum_i \alpha_i A_i(\mathbf{x}, \mathbf{y}) &= \sum_i \min\left\{ \alpha_i T_i(\mathbf{x}, \mathbf{y}), \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})} \alpha_i T_i(\mathbf{y}, \mathbf{x}) \right\} \\ &\leq \min\left\{ \sum_i \alpha_i T_i(\mathbf{x}, \mathbf{y}), \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})} \sum_i \alpha_i T_i(\mathbf{y}, \mathbf{x}) \right\} \\ &= \min\left\{ T^*(\mathbf{x}, \mathbf{y}), \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})} T^*(\mathbf{y}, \mathbf{x}) \right\} = A^*(\mathbf{x}, \mathbf{y}) \end{aligned}$$

Hence, the theorem is proved. $\diamond$

This theorem may also shed some light on the issue of whether the multi-point Metropolis method is superior to the ordinary Metropolis algorithm (with a comparable number of proposals).

13.4 Eigenvalue Analysis for the Independence Sampler

In the special case that the proposal is an *independent* transition function (Section 5.4.2), we have a rather clean result on the analysis of all the eigenvalues of the Metropolis-Hastings transition matrix (Liu 1996a). In this section, we assume that the state space $\mathcal{X}$ is finite. Without loss of generality, we let $\mathcal{X} = \{1, \ldots, m\}$. Two probability measures $\pi(\cdot)$ and $p(\cdot)$ are then abbreviated as $\pi_i = \pi(i)$, and $p_i = p(i)$, $i = 1, \ldots, m$. We introduce the following four notations: $F_\pi(k) = \pi_1 + \cdots + \pi_k$, $S_\pi(k) = 1 - F_\pi(k-1) = \pi_k + \cdots + \pi_m$, $F_p(k) = p_1 + \cdots + p_k$, and $S_p(k) = 1 - F_p(k-1)$.

For any $i, j \in \mathcal{X}$, we can write down the transition probability from i to j for the Metropolized independence sampler

$$A(i,j) = \begin{cases} p_j \min\{1, w_j/w_i\} & \text{if } j \neq i \\ p_i + \sum_k p_k \max\{0, 1 - w_k/w_i\} & \text{if } j = i, \end{cases}$$

where $w_i = \pi_i / p_i$ is the *importance ratio*. Without loss of generality, we further assume that the states are sorted according to the magnitudes of their importance ratios; that is, the elements in $\mathcal{X}$ are labeled so that

$$w_1 \geq w_2 \geq \cdots \geq w_m.$$

The transition matrix can then be written as

$$A = \begin{pmatrix} p_1 + \lambda_1 & \pi_2/w_1 & \pi_3/w_1 & \cdots & \pi_{m-1}/w_1 & \pi_m/w_1 \\ p_1 & p_2 + \lambda_2 & \pi_3/w_2 & \cdots & \pi_{m-1}/w_2 & \pi_m/w_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1 & p_2 & p_3 & \cdots & p_{m-1} + \lambda_{m-1} & \pi_m/w_{m-1} \\ p_1 & p_2 & p_3 & \cdots & p_{m-1} & p_m \end{pmatrix},$$

where

$$\lambda_k = \sum_{i=k}^{m}(p_i - \pi_i/w_k) = S_p(k) - S_\pi(k)/w_k, \tag{13.6}$$

which is just the probability of being rejected in the next step if the chain is currently at state k.

For any function $f(x)$ defined on $\mathcal{X}$, we denote

$$f^+(x) = \begin{cases} f(x) & \text{if } f(x) > 0; \\ 0 & \text{if } f(x) \le 0. \end{cases}$$

It is noted that λ_k has another expression:

$$\lambda_k = \sum_{i \ge k}(\pi_i/w_i - \pi_i/w_k) = E_\pi\left\{\frac{1}{w(X)} - \frac{1}{w_k}\right\}^+,$$

where the expectation is taken with respect to $X \sim \pi$. Apparently, if two states i and $i+1$ have equal importance ratios, then $\lambda_i = \lambda_{i+1}$. Let $\boldsymbol{p} = (p_1, \ldots, p_m)^T$ denote the column vector of the trial distribution and let $\mathbf{e} = (1, \ldots, 1)^T$. Then, A can be expressed as

$$A = G + \mathbf{e}\boldsymbol{p}^T,$$

where G is an upper triangular matrix of the form

$$G = \begin{pmatrix} \lambda_1 & \frac{p_2(w_2 - w_1)}{w_1} & \cdots & \frac{p_{m-1}(w_{m-1} - w_1)}{w_1} & \frac{p_m(w_m - w_1)}{w_1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{m-1} & \frac{p_m(w_m - w_{m-1})}{w_{m-1}} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

Note that $\mathbf{e}$ is a common right eigenvector for both A and $A - G$, corresponding to the largest eigenvalue 1. Since $A - G$ is of rank 1, the rest of the eigenvalues of A and G have to be the same. Hence, the eigenvalues for A are $1 > \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{m-1}$.

When m is fixed and the number of iterations goes to infinity, the mixing rate of this Metropolis Markov chain is asymptotically dominated by the second largest eigenvalue λ_1, which is equal to $1 - 1/w_1$. All the eigenvectors of G can also be found explicitly. We first note that the vector $\tilde{\mathbf{v}}_1 = (1, 0, \ldots, 0)^T$ is a right eigenvector corresponding to λ_1. Checking one more step, we find that $\tilde{\mathbf{v}}_2 = (\pi_2, 1 - \pi_1, 0, \ldots, 0)^T$ is a right eigenvector of λ_2. Generalizing the result, we obtain the following result.

Lemma 13.4.1 *The eigenvectors and eigenvalues of G are λ_k, and $\tilde{\mathbf{v}}_k = (\pi_k, \ldots, \pi_k, S_\pi(k), 0, \ldots, 0)^T$, for $k = 1, \ldots, m-1$, where there are k nonzero entries in $\tilde{\mathbf{v}}_k$.*

Theorem 13.4.1 *For the Metropolized independence sampler, all the eigenvalues of its transition matrix are $1 > \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{m-1} \ge 0$, where $\lambda_k = \sum_{i=k}^{m}(p_i - \pi_i/w_k) = E_\pi\{1/w(X) - 1/w_k\}^+$. The right eigenvector $\mathbf{v}_k$ corresponding to λ_k is*

$$\mathbf{v}_k \propto (0, \ldots 0, S_\pi(k+1), -\pi_k, \ldots, -\pi_k)^T,$$

where there are $k-1$ zero entries.

Proof: Since $A = G + \mathbf{e}\boldsymbol{p}^T$, $A\tilde{\mathbf{v}}_k = G\tilde{\mathbf{v}}_k + \mathbf{e}(\boldsymbol{p}^T\tilde{\mathbf{v}}_k)$. It is further noted that

$$\boldsymbol{p}^T\tilde{\mathbf{v}}_k = S_\pi(k)\pi_k + p_k S_p(k) = \pi_k(1 - \lambda_k).$$

Hence, $A\tilde{\mathbf{v}}_k = \lambda_k\tilde{\mathbf{v}}_k + \pi_k(1 - \lambda_k)\mathbf{e}$. Since $\mathbf{e}$ is a right eigenvector of A with eigenvalue 1, we have, for any t,

$$A(\tilde{\mathbf{v}}_k - t\mathbf{e}) = \lambda_k\left\{\tilde{\mathbf{v}}_k - \frac{t - \pi_k(1 - \lambda_k)}{\lambda_k}\mathbf{e}\right\}.$$

Solving $t = \{t - \pi_k(1 - \lambda_k)\}/\lambda_k$, we find that $\mathbf{v}_k = \tilde{\mathbf{v}}_k - \pi_k\mathbf{e}$ is a right eigenvector of A corresponding to λ_k. $\diamond$

The coupling method can also be used to bound the convergence rate for this sampler and the argument does not require that the state space of the chain is discrete. Suppose two independence sampler chains $\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots\}$ and $\{\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \ldots\}$ are simulated, of which the $\mathbf{x}$ chain starts from a fixed point $\mathbf{x}^{(0)} = \mathbf{x}_0$ (or a distribution) and the $\mathbf{y}$ chain starts from the equilibrium distribution π. The two chains can be "coupled" in the following way. Suppose $\mathbf{x}^{(t)} = \mathbf{x}$ and $\mathbf{y}^{(t)} = \mathbf{y}$. At step $t+1$, a new state $\mathbf{z}$ is drawn according to distribution $p(\cdot)$, its associated importance ratio $w_z \equiv \pi(\mathbf{z})/p(\mathbf{z})$ is computed and a uniform random variable u is generated independently. There are three scenarios: (i) If $u \le \min\{w_z/w_x, w_z/w_y\}$, then both chains accept $\mathbf{x}$ as their next states (i.e., $\mathbf{x}^{(t+1)} = \mathbf{y}^{(t+1)} = \mathbf{z}$); (ii) if $u \ge \max\{w_z/w_x, w_z/w_y\}$, then both chains reject so that $\mathbf{x}^{(t+1)} = \mathbf{x}$ and $\mathbf{y}^{(t+1)} = \mathbf{y}$; and (iii) if u lies between w_z/w_x and w_z/w_y, then the chain with larger ratio accepts and the chain with smaller ratio rejects. It is clear that the first time when scenario (i) occurs is the *coupling time*, the time at and after which the realizations of the two chains become identical.

The probability of the occurrence of (i) can be bounded from below:

$$\begin{aligned} P(\text{accept} \mid X_k = x, Y_k = y) &= \sum_{i=1}^{m} p_i \min\left\{1, \frac{w_i}{w_x}, \frac{w_i}{w_y}\right\} \\ &= \sum_{i=1}^{m} \pi_i \min\left\{\frac{1}{w_i}, \frac{1}{w_x}, \frac{1}{w_y}\right\} \ge \frac{1}{w_1}, \end{aligned}$$

where w_1 is the largest importance ratio. Hence, from the Markov property, the number of steps for the chains to be coupled is bounded by a geometric

distribution

$$P(N > n) \le (1 - w_1^{-1})^n.$$

Consequently, for any measurable subset $S \subset \mathcal{X}$,

$$\begin{aligned} |p^{(n)}(S) - \pi(S)| &= |P(X_n \in S) - P(Y_n \in S)| \\ &\le P(X_n \neq Y_n) = P(N > n) \le (1 - w_1^{-1})^n. \end{aligned}$$

13.5 Perfect Simulation

When running a MCMC sampler, we have always to wait for a period of "burn-in" time (or called the time for *equilibration*). Samples obtained after this period of time can be regarded as approximately following the target distribution π and be used in Monte Carlo estimation. In practice, however, one is never sure how long the "burn-in" period should be and it is always a distracting question for researchers to know when to declare "convergence" of the chain. A surprising discovery recently made by Propp and Wilson (1996) is that *perfect* random samples can be obtained, in finite (but stochastic) time, from many Markov chain samplers. Their algorithm is also called *coupling from the past* (CFTP).

Under mild conditions (irreducibility, aperiodicity, and a drift condition) which we have assumed throughout of the book, the Markov chain underlying a MCMC sampler would have been in its stationary distribution had it been iterated for infinite steps. Thus, if the chain had been started from $t = -\infty$, the infinite past, then at time $t = 0$, the chain would have been in equilibrium and a sample produced at $t = 0$ would have been an exact sample from the target distribution π. This fact has already been known to all the probabilists a long time ago. What Propp and Wilson discovered is that one can figure out what the *current sample* is without actually tracing back to the infinite past. The strategy they took was also known to probabilists a long time ago: *coupling* and *coalescence*.

Suppose the Markov chain under consideration is defined on a finite space $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$. Let the transition matrix be $A(\mathbf{x}, \mathbf{y})$ and the equilibrium distribution be π. Consider all possible ways from $\mathbf{x}^{(-1)} \to \mathbf{x}^{(0)}$. The transition function tells us that

$$\Pr(\mathbf{x}^{(0)} = j \mid \mathbf{x}^{(-1)} = i) = A(i, j).$$

If we want to simulate this step on a computer, we will first compute the *cumulative transition probabilities*:

$$G(i, j) = \sum_{k=1}^{j} A(i, k) \equiv \Pr(\mathbf{x}^{(0)} \le j \mid \mathbf{x}^{(-1)} = i).$$

Then, we generate a uniform random number [i.e., $u_0 \sim \text{Uniform}(0, 1)$]. Finally, we let $\mathbf{x}^{(0)} = j$ if $G(i, j-1) < u_0 \le G(i, j)$. In a usual computer algorithm for realizing a forward Markov transition, one generates a random number whenever needed and they may be different had $\mathbf{x}^{(-1)}$ been $i' \neq i$. However, there is no reason why we cannot use the *same* random number u_0 generated beforehand and use it for all possible states $i \in \mathcal{X}$ at time -1. More abstractly, we can think of the above sampling step as a mapping

$$\mathbf{x}^{(0)} = \phi(u_0, \mathbf{x}^{(-1)}). \tag{13.7}$$

A distinctive feature of (13.7) is that the chains starting from all possible states are *coupled* by the same random number u_0.

If it so happens that our random number u_0 makes all the chain "coupled;" that is,

$$\phi(u_0, i) \equiv j_0 \quad \text{for all } i, \tag{13.8}$$

then $\mathbf{x}^{(0)} = j_0$ must be a perfect sample from the target distribution π. To see this point, imagine that the Markov chain has been run from $t = -\infty$ and entered into time $t = -1$. Then, it must be in stationarity at time $t = -1$. Because of the construction of ϕ, the next step $\mathbf{x}^{(0)}$ must still be in stationarity. Because of (13.8), $\mathbf{x}^{(0)}$ has to take value j_0 no matter what state $\mathbf{x}^{(-1)}$ takes.

Of course, the chance that the chains are all coupled in one step is too small. If they are not all coupled, we can iterate (13.7) backward. Since

$$\mathbf{x}^{(-n)} = \phi(u_{-n}, \mathbf{x}^{(-n-1)}) \tag{13.9}$$

for all n, we have

$$\mathbf{x}^{(0)} = \phi(u_0, \phi(u_{-1}, \ldots, \phi(u_{-N+1}, \mathbf{x}^{(-N)}) \cdots)).$$

In fact, we can even imagine that the sequence of uniform random numbers, $\ldots, u_{-N}, \ldots, u_{-1}, u_0$, has been given in advance, and we realize a stationary Markov chain by composing (13.9) from the infinite past.

Now, consider starting $|\mathcal{X}|$ parallel Markov chains at time $t = -N$, each with a different starting state [i.e., $\mathbf{x}^{(-N,j)} = j$]. Then, after one iteration of the ϕ function, we have, for all the chains,

$$\mathbf{x}^{(-N+1,j)} = \phi(u_{-N+1}, j) \quad \text{for } j = 1, \ldots, |\mathcal{X}|. \tag{13.10}$$

Hence, the $\mathbf{x}^{(-N+1,j)}$ have fewer distinct values than that of $\mathbf{x}^{(-N,j)}$. This means that each iteration of the ϕ function will "coalesce" some chains. If N is large enough, then all the chains starting at $t = -N$ will coalescence and produce a single random sample, $\mathbf{x}_0$, at time $t = 0$. Since a Markov chain that comes from the infinite past has to get into time $-N$ and then passes through recursion (13.10), the sample obtained at time $t = 0$ has to be

$\mathbf{x}^{(0)}$ (if the same set of uniform random numbers has been used from time $-N$ to 0). Thus, this $\mathbf{x}^{(0)}$ is an exact draw from the stationary distribution π. If N is not large enough, we will need to move K steps backward to time $-N-K$ and try again, reusing all the previously generated random numbers.

The CFTP algorithm can be implemented, at least conceptually, as follows:

1. Generate $u_0 \sim \text{Uniform}(0,1)$ and compute $f_{-1}(i) = \phi(u_0, i)$, for $i = 1, \ldots, |\mathcal{X}|$.
   (a) If the $f_{-1}(i)$ are all equal, then the common value $f_{-1}(i)$ is retained as a random sample from π and the algorithm is stopped.
   (b) If not all the $f_{-1}(i)$ are the same, set $n = 2$ and go to Step 2.
2. At time $-n$, we generate $u_{-n+1} \sim \text{Uniform}(0,1)$ and update
$$f_{-n}(i) = f_{-n+1}\{\phi(u_{-n+1}, i)\} \text{ for } i = 1, \ldots, |\mathcal{X}|. \tag{13.11}$$
   (a) If all the $f_{-n}(i)$ are the same, return the common value $f_{-n}(i)$ as a sample from π and stop.
   (b) If not all the $f_{-n}(i)$ are the same, set $n \leftarrow n+1$ and return to Step 2.

It is important to notice the difference between the forward coupling expression (13.9) and the backward coupling formula (13.11). More explicitly, $f_{-n}(i)$, for all n, refers to a possible state for $\mathbf{x}^{(0)}$ at time 0 instead of that for $\mathbf{x}^{(-n)}$ at time $-n$.

It is often too slow to move one-step backward a time. A preferable approach is to modify Step 2(b) in the foregoing CFTP algorithm by setting $n \leftarrow 2n$; that is, one doubles the backward steps if not all the chains coalesce at time 0 in n steps. A main difficulty in applying the CFTP algorithm in interesting cases is that it is often impossible to monitor simultaneously all the chains starting from *all* possible states. For example, an Ising model on a 64×64 lattice has 2^{64^2} possible states, which are impossible to follow. A useful method (Propp and Wilson 1996) is to establish an *ordering* "$\preceq$" among all the states, so that this ordering is maintained after the one-step coupled Markov transition:

$$\mathbf{x} \preceq \mathbf{y} \quad \to \quad \phi(u, \mathbf{x}) \preceq \phi(u, \mathbf{y}),$$

for all $0 < u < 1$. Suppose a "maximal state" and a "minimal state" under this ordering exist. Then, one needs only to monitor two chains on the computer: one started from the maximal state and the other from the minimal state. When these two chains are coupled, then chains from all other states must be coupled to the same state.

The work of Propp and Wilson (1996) has stimulated a lot of interest from computer scientists, probabilists, and statisticians. Many new tricks have been developed to tackle various situations. One of the main concerns regarding the CFTP algorithm is the "user impatience" bias; that is, the user may stop the algorithm when it takes too long to find an appropriate past time $-N$ or the algorithm is stopped before schedule because of emergency (an electricity outage, say). Both cases will create a bias in the produced samples. In a sense, the CFTP cannot be *interrupted*. Fill (1998) recently proposed an interruptible algorithm that alleviates this concern. See Green and Murdoch (1998), Fill (1998), and Propp and Wilson (1998).

13.6 A Theory for Dynamic Weighting

13.6.1 Definitions

Suppose the configuration state $\mathcal{X}$ is augmented by a one-dimensional weight space so that the *current* state in a dynamic weighting Monte Carlo scheme is $(\mathbf{x}, w)$. Most of the analysis presented in this section are adapted from Liu et al. (2001).

Let constant $\theta \geq 0$ be given in advance. The Q-type and the R-type moves that we will study in this section are defined as follows.

Q-type Move:

- Propose the next state $\mathbf{y}$ from the proposal $T(\mathbf{x}, \cdot)$ and compute the *Metropolis ratio*
$$r(\mathbf{x}, \mathbf{y}) = \frac{\pi(\mathbf{y})T(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})T(\mathbf{x}, \mathbf{y})}. \tag{13.12}$$
- Draw $U \sim \text{Uniform}(0,1)$. Update $(\mathbf{x}, w)$ to $(\mathbf{x}', w')$ as
$$(\mathbf{x}', w') = \begin{cases} (\mathbf{y}, \max\{\theta, wr(\mathbf{x}, \mathbf{y})\}) & \text{if } U \leq \min\{1, wr(\mathbf{x}, \mathbf{y})/\theta\} \\ (\mathbf{x}, aw) & \text{otherwise.} \end{cases} \tag{13.13}$$
  where $a > 1$ can be either a constant or a random variable independent of all other variables.

R-type Move

- Propose $\mathbf{y}$ and compute $r(\mathbf{x}, by)$ as in the Q-type move.