A Theory for Dynamic Weighting in Monte Carlo Computation

Jun S. LIU, Faming LIANG, and Wing Hung WONG

This article provides a first theoretical analysis of a new Monte Carlo approach, the *dynamic weighting algorithm*, proposed recently by Wong and Liang. In dynamic weighting Monte Carlo, one augments the original state space of interest by a weighting factor, which allows the resulting Markov chain to move more freely and to escape from local modes. It uses a new invariance principle to guide the construction of transition rules. We analyze the behavior of the weights resulting from such a process and provide detailed recommendations on how to use these weights properly. Our recommendations are supported by a renewal theory-type analysis. Our theoretical investigations are further demonstrated by a simulation study and applications in neural network training and Ising model simulations.

KEY WORDS: Gibbs sampling; Importance sampling; Ising model; Metropolis algorithm; Neural network; Renewal theory; Simulated annealing; Simulated tempering.

1. INTRODUCTION

Optimization, integration, and system simulation are at the heart of many scientific problems, almost all but the simplest of which must be solved by numerical methods, either heuristic or semiheuristic and exact or approximate. Algorithms of a stochastic nature play a central role in these endeavors. In recent decades, Monte Carlo algorithms have received much attention from researchers in engineering and computer science (e.g., Geman and Geman 1984, Kirkpatrick, Gelatt, and Vecchi 1983), statistical physics (e.g., Goodman and Sokal 1989, Marinari and Parisi 1987, Swendsen and Wang 1987), computational biology (e.g., Lawrence et al. 1993; Leach 1996, Liu, Neuwald, and Lawrence 1999), material science (Frenkel and Smit 1996), statistics (Gelfand and Smith 1990; Tanner and Wong 1987), and many other fields.

Let $\pi(x)$ be the target density under investigation, which is often given up to a normalizing constant. Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) introduced the fundamental idea of evolving a Markov process to achieve the simulation of random samples from π. Start with any configuration, the Metropolis algorithm iterates many times of the following two steps.

Step 1: Propose a random "perturbation" of the system, (i.e., from $X \to X'$), which can be regarded as generated from a transition probability distribution $T(X, X')$; calculate the change $\Delta h = \log \pi(X') - \log \pi(X)$.

Step 2: Generate a random number U from uniform (0,1). Accept the proposal and change the configuration to X' if $\log U \le \Delta h$, and reject the proposal otherwise.

The Metropolis scheme has been used extensively in statistical physics over the last 40 years and is the cornerstone of all Markov chain Monte Carlo (MCMC) techniques recently developed in the statistics community. The Gibbs sampler (Geman and Geman 1984) can be viewed as a nontrivial variation of the Metropolis technique.

As known by many researchers, a major drawback of various MCMC algorithms is that the constructed Markov chain can mix very slowly and may be trapped indefinitely in a local mode, rendering the method ineffective. To improve mixing, techniques such as multigrid Monte Carlo (Goodman and Sokal 1989), auxiliary variables (Swendsen and Wang 1987), simulated tempering (Geyer and Thompson 1995, Marinari and Parisi 1992), and blocking and collapsing (Liu, Wong, and Kong 1994) have been proposed. These techniques can all be regarded as special variations of the basic Markov chain idea of Metropolis et al. (1953) and Hastings (1970).

In this article we study a different approach, the *dynamic weighting method* recently introduced by Wong and Liang (1997). The method extends the basic Markov chain equilibrium concept of Metropolis et al. (1953) to a more general weighted equilibrium of a Markov chain.

The purpose of introducing importance weights into the dynamic Monte Carlo process is to provide a means for the system to make large transitions not allowable by the standard Metropolis transition rules. When the distribution has regions of high density separated by barriers of very low density, the waiting time for the Metropolis process to cross over the barriers will be essentially infinite. In our dynamically weighted Monte Carlo, the process can often move against very steep probability barriers, which apparently violates the Metropolis rule. The weight variable is updated in a way that allows for an adjustment of the bias induced by such non-Metropolis moves. This device can essentially eliminate the "waiting time infinity" (i.e., the waiting time for a slow-mixing chain to converge is practically infinite) in most applications.

This advantage of using the dynamic weights comes at a price, however. There can be large variability in the resulting weighted estimates when the realized weights are very long-tailed. As we discuss in this article, many of the weighted transition rules that we propose will lead to a weight distribution with infinite mean. In short, the waiting time infinity in the standard Metropolis process now manifests itself as an "importance weight infinity" in the dynamic weighting process.

Fortunately, the standard Metropolis and Gibbs moves can be viewed as special types of weighted move—they are valid as long as the weight variable is kept constant after the move. Hence we can mix the new weighted transitions with the standard transitions so that the weighted moves are used only when we propose large changes in the system and the standard Metropolis or Gibbs moves are used for local exploration. In this way the extra variability in the weights can be greatly reduced, but the system is still capable of making large jumps.

Two key ideas involved in the approach of Wong and Liang (1997) are (1) sequential decomposition (and buildup) of the complicated target function, and (2) introduction of the importance weight, denoted by W, as a dynamic variable for the control of the Markov chain simulation in each step. Wong and Liang's tests of this method on many large-scale simulation and global optimization problems yielded promising results. Some of these problems are reviewed in Section 7. The purpose of this article is to provide a first theoretical analysis of the properties of the dynamic weighting rules and the asymptotic behaviors of the dynamically weighted Monte Carlo process. We show that asymptotically the weights will have a stationary distribution, but that this stationary distribution typically has infinite expectation. Thus the theory for the weighted estimate is nontrivial. We demonstrate that in general the weighted estimate is expected to be consistent, but its convergence rate is exceedingly slow. Fortunately, our analysis also shows that the simple device of stratified truncation of the weights before averaging (Wong and Liang 1997) is capable of generating stable and approximately unbiased estimates in reasonable sample sizes. In other words, stratified truncation seems to be an effective method for handling the "importance weight infinities" at the estimation stage. In contrast, "waiting time infinities" will preclude the possibility of any such corrections at the estimation stage.

This article is organized as follow. Section 2 defines the new transition moves, called the Q-type and the R-type. [We call the types of transitions invented by Metropolis et al. (1953) and generalized by Hastings (1970) M-type moves.] Section 2 also introduces a new invariance principle used for guiding the design of new moves. Section 3 describes the behaviors of the weights in a dynamic weighting scheme under various conditions. Section 4 studies the stochastic stability of the weight process, Section 5 provides general guidelines for the using of the method, and Section 6 gives a theoretical analysis, using some renewal theory result of Kesten (1974), for the suggestions made in Section 5. Section 7 describes a simulation study and applications of the dynamic weighting method to a few difficult problems, including the neural network training and the Ising model simulation. Section 8 concludes with a brief discussion.

2. DYNAMIC WEIGHTING SCHEMES

As with the Metropolis algorithm, the dynamic weighting scheme starts with an arbitrary Markov transition kernel $T(x, y)$, often called the "proposal chain," from which the next possible move is "suggested." In M-type moves, the system is updated by a Metropolis step, and its weight variable is not changed. We introduce two new transitions that combine the move in the original state space with the update of an extra weighting variable. The IWIW (Invariance with respect to importance weighting) principle is introduced to motivate the scheme.

2.1 Definitions

Suppose that current state is $(X_t, W_t) = (x, w)$, where X_t denotes the original system state at time t and W_t denotes the dynamic weight attached to the state. The Q-type move and the R-type move are defined as follows:

Q-Type Move

- Propose the next state $Y = y$ from the proposal $T(x, \cdot)$, and compute the *Metropolis ratio*,

$$r(x, y) = \frac{\pi(y)T(y, x)}{\pi(x)T(x, y)}. \quad (1)$$

- Choose $\theta = \theta(w, x) \geq 0$ and draw $U \sim \text{unif}(0, 1)$. Update (X_t, W_t) to (X_{t+1}, W_{t+1}) as

$$(X_{t+1}, W_{t+1})$$
$$= \begin{cases} (y, \max\{\theta, wr(x, y)\}) & \text{if } U \leq \min\{1, wr(x, y)/\theta\} \\ (x, aw) & \text{otherwise,} \end{cases}$$
$$(2)$$

where $a > 1$ can be either a constant or an independent random variable.

R-Type Move

- Draw $Y = y$ from $T(x, y)$ and compute the Metropolis ratio $r(x, y)$.
- Choose $\theta = \theta(w, x) \geq 0$, and draw $U \sim \text{unif}(0, 1)$. Update (X_t, W_t) to (X_{t+1}, W_{t+1}) as

$$(X_{t+1}, W_{t+1}) = \begin{cases} (y, wr(x, y) + \theta) & \text{if } U \leq wr(x, y)/ \\ & \quad \{wr(x, y) + \theta\} \\ (x, w(wr(x, y) + \theta)/\theta) & \text{otherwise.} \end{cases}$$
$$(3)$$

Note that θ in both types is an adjustable parameter that can depend on previous value of (X, W). Although one can play with different settings for θ, in this article we concentrate on special cases with $\theta \equiv$ constant. Because any positive constant leads to the same weight behavior, it is sufficient to consider only two cases: $\theta = 1$ versus $\theta = 0$. When $\theta = 0$, expressions (3) and (2) become identical.

The intuition of the Q-type or R-type move is that the augmented chain can escape from a local mode by automatically increasing the associated weight W. One can also try to accelerate this by adjusting θ, but we do not explore this refinement in this article.

We suggest that the two dynamic weighting moves be applied in a compact space. This can be achieved by preventing the sampler from visiting exceedingly low-probability regions. Furthermore, to guard against possible boundary effect caused by exceedingly small $r(x, y)$ (i.e., practically 0), we can modify the weight updating as follows: If $r(x, y) < \epsilon$

for a proposed y, then rejection does not induce any change of the weights.

Because the new moves use different rejection rules than that of the Metropolis, the detailed balance with respect to π no longer holds for either the Q-type move the R-type. Thus the equilibrium distribution of X (if it exists) is *not* π. To motivate the schemes, Wong and Liang (1997) introduced the following IWIW principle.

Definition 1. The joint distribution $f(x, w)$ of (X, W) is said to be correctly weighted with respect to π if $\sum_w \sum w \cdot f(x, w) \propto \pi(x)$. A transition rule is said to satisfy IWIW if it maintains the correctly weighted property for the joint distribution of (x, w) whenever the initial joint distribution is correctly weighted.

Clearly, the M-type move satisfies IWIW, because the state update satisfies the detailed balance and the weight is constant. In Section 3 we show that the R-type move does so as well.

2.2 Notations and Assumptions

The following notations are used in this article:

$\pi(x)$, target distribution of interest

$\mathcal{X}$, space on which $\pi(x)$ is defined

X_t, dynamic state variable (defined on $\mathcal{X}$)

W_t, dynamic weight, taking values in $(0, \infty)$

$T(x, y)$, proposal transition function, assumed to be aperiodic and irreducible

$g(x)$, an invariant measure of $T(x, y)$

$g(x, y)$, joint distribution of two consecutive steps, $g(x)T(x, y)$

$T_*(x, y)$, reversal transition function; that is, $T_*(x, y) = g(y)T(y, x)/g(x)$

$\delta(x, y)$, log-ratio between backward and forward steps; that is, $\log g(y, x) - \log g(x, y)$

$r(x, y)$, the Metropolis ratio $\pi(y)T(y, x)/\pi(x)T(x, y)$

$u(x)$, importance weight function, $\pi(x)/g(x)$

E_p or var_p, expectation or variance taken with respect to probability measure p.

The following assumptions are made throughout the article:

a. The sample space $\mathcal{X}$ is discrete and finite. (Discussions on how to relax this condition are given).

b. $T(x, y) > 0$ if and only if $T(y, x) > 0$ (so the Metropolis ratio is always defined).

c. T is irreducible and aperiodic.

d. Both $g(x)$ and the target distribution $\pi(x)$ are greater than 0 for $x \in \mathcal{X}$.

Because of our assumptions on T and $\mathcal{X}$, an invariant distribution $g(x)$ exists and is unique. Assumptions on T are not very stringent, and most practical Metropolis–Hastings schemes can achieve this with minor modification (e.g., incorporating a random component). We believe that our result can be extended to the cases when $\mathcal{X}$ is a compact space or a general metric space on which T is Harris ergodic (Asmussen 1987), and we provide some discussions on how to achieve this. If the Markov chain induced by T is reversible, which we state simply as "$T(x, y)$ is reversible," then we have $T_* = T$. But as we explain in Section 4, we are more interested in

the case where T is nonreversible. This situation arises most naturally when the proposal chain is a mixture of different types of moves (e.g., a Q-type more and a M-type move), which is also the case where dynamic weighting is most useful. Additionally, nonreversible proposal chains can arise in more advanced dynamic Monte Carlo schemes such as the Langevin diffusion, hybrid Monte Carlo, Metropolized independence sampler (Hastings 1970), and other "biased Monte Carlo" methods such as the multiple-try Metropolis algorithm (Liu, Liang, and Wong 2000).

3. INVARIANCE WITH RESPECT TO IMPORTANCE WEIGHTING PROPERTY OF DYNAMIC WEIGHTING SCHEMES

We first show that the dynamic weighting schemes satisfy an appealing property that might be a useful criterion for designing other Monte Carlo schemes.

Theorem 1. Suppose that the starting joint distribution $f_1(x, w)$ for (X, W) is correctly weighted with respect to π; that is, $\sum wf_1(x, w) = c_1\pi(x)$. After one-step transition of the R-type with $\theta = \theta(x, w) > 0$ for all (x, w), the new state (Y, W') is also correctly weighted with respect to π.

Proof. For simplicity, we work here with discrete random variables, and need to change only summations to integrations in continuous cases. Let $f_2(y, w')$ be the distribution of (Y, W'); then

$$\sum w' f_2(y, w')$$

$$= \sum_{w'} \left\{ \sum_x \sum_w w' f_1(x, w) \, I[w' = wr(x, y) + \theta] \right.$$

$$\times T(x, y) \frac{wr(x, y)}{wr(x, y) + \theta} + \sum_z \sum_{w'} w' f_1(y, w)$$

$$\times I\left[w' = \frac{w(wr(y, z) + \theta)}{\theta} \right] T(y, z) \frac{\theta}{wr(y, z) + \theta} \right\}$$

$$= \sum_x \sum_w f_1(x, w) T(x, y) \frac{wr(x, y)}{wr(x, y) + \theta}(wr(x, y) + \theta)$$

$$+ \sum_z \sum_w f_1(y, w) T(y, z) \frac{\theta}{wr(y, z) + \theta} \frac{w(wr(y, z) + \theta)}{\theta}$$

$$= \sum_x \sum_w wf_1(x, w) \frac{\pi(y)T(y, x)}{\pi(x)} + \sum_w \sum_z wf_1(y, w)T(y, z)$$

$$= \sum_x c_1 \pi(y)T(y, x) + c_1\pi(y) = 2c_1\pi(y). \tag{4}$$

In the foregoing, $I[a = b]$ is the indicator function, that is, it equals 1 if the statement $a = b$ is true and 0 otherwise. Note that θ is allowed to be a function of the previous configuration (X, W) provided that $\theta > 0$ for all (X, W).

In contrast, the Q-type move only approximately satisfies the IWIW property when $\theta > 0$. More precisely, we see that

$$\sum_{w'} w' f_2(y, w')$$

$$= \sum_{w'} \left\{ \sum_x \sum_{w} f_1(x, w) T(x, y) \min \left\{ 1, \frac{wr(x,y)}{\theta} \right\} w' \right.$$

$$\left. + \sum_{w} \sum_z a w f_1(y, w) T(y, z) q_w(y, z) \right\}$$

$$= \sum_x \left\{ \sum_{w} f_1(x, w) T(x, y) wr(x, y) \right\} + a \sum_{w} w q_w(y) f_1(y, w)$$

$$= \sum_x c_1 \pi(x) T(x, y) \frac{\pi(y) T(y, x)}{\pi(x) T(x, y)} + R_a = c_1 \pi(y) + R_a,$$

where $q_w(y, z)$ is the rejection probability when the chain proposes moving from y to z, $q_w(y)$ is the total rejection probability for moving away from y, and $R_a = a \sum_w w q_w(y) f_1(y, w)$. Clearly, if $q_w(y)$ is approximately constant in w, then $R_a \approx ac_1 \pi(y)$, and the IWIW is approximately satisfied. More generally, if we let r_0 be the smallest Metropolis ratio, that is,

$$r_0 = \min_{(x, y): T(x, y) > 0} r(x, y), \qquad (5)$$

which is greater than 0 when the state space is finite, then $q_w(y) = 0$ when $w \geq r_0^{-1}$. Hence

$$R_a = a \sum_{w < r_0^{-1}} w q_w(y) f_1(y, w).$$

In the case where $\sum_{w} w f_1(y, w)$ is very large (i.e., c_1 is large), the residue $R_a \approx 0$ in comparison with with $c_1 \pi(y)$. Hence the IWIW is also approximately satisfied.

If $\theta \equiv 0$, then all of the proposed moves are accepted in both the Q-and R-type moves. Furthermore, the two moves are identical, and the IWIW property is satisfied:

$$\sum w' f_2(y, w') = \sum_x \sum_{w'} f_1(x, w) T(x, y) wr(x, y)$$

$$= \sum_x c_1 \pi(x) \frac{\pi(y) T(y, x)}{\pi(x) T(x, y)} = c_1 \pi(y).$$

An interesting distinction between using $\theta \equiv 0$ and $\theta > 0$ is the normalizing constant ($2c_1$ vs. c_1). It is easy to see that randomly mixing any number of different types of IWIW moves also satisfies the IWIW property. However, if the change of schemes depends on the value of W or X then, IWIW can be violated.

Although the R-type move satisfies IWIW, it entails two complications. First, with $\theta > 0$, the constant c_1 is inflated to $2c_1$ after one-step transition, as shown in (4). This implies that in the long run the W sequence may diverge to infinity, rendering the scheme ineffective. Second, using $\theta = 0$ makes the expectation of W_t remain constant throughout iterations, but, as we show in the next section, W_t converges to 0 with probability 1 if the transition matrix T is nonreversible.

4. STABILITY OF THE WEIGHT PROCESS

Because weight process performance is affected by both θ and the choice of proposal transition function, we consider the following five possible scenarios. We show that for all cases with $\theta = 1$ and with suitable modification of the weight updating scheme, the weight process has a stable distribution.

4.1 Case A: $\theta = 0$ and $T(x, y)$ is Reversible

In this case the Q- and R-type moves are identical, and both can be viewed as generalizations of standard importance sampling. More precisely, every proposed move is accepted, and the weight is updated as

$$W' = W \, r(x, y).$$

Suppose that $g(x)$ is the invariant distribution for $T(x, y)$ and that $g(x, y) = g(x) T(x, y)$ is the joint distribution for the two consecutive steps in equilibrium. Let $u(x) = \pi(x)/g(x)$ be the usual 'weighting function' if importance sampling is conducted with $g(x)$ as the sampling distribution. The updating formula for the weight can be rewritten as

$$W' = W \frac{u(y)}{u(x)} \frac{g(y, x)}{g(x, y)}. \qquad (6)$$

Hence, if the transition matrix T induces a reversible chain [e.g., $g(x, y) = g(y, x)$], and we start with $X_0 = x_0$ and $W_0 = c_0 u(x_0)$, then for any $t > 0$, $W_t = c_0 u(X_t)$. These weights are identical to those from standard importance sampling using the trial distribution g.

4.2 Case B: $\theta = 1$ and $T(x, y)$ is Reversible

If the proposal chain is reversible, then the Q-type sampler converges to a regular importance sampler with $g(x)$ as the trial density. That is, the weight W conditional on $X = x$ will converge to a degenerate distribution concentrating on $c_0 u(x)$ for some c_0. To see this, let $u_0 = \min\{\pi(x)/g(x)\}$. Then once the pair (x, w) satisfies $w = c_0 u(x)$ with $c_0 u_0 \geq 1$, the proposed transition y will *always* be accepted according to (2), and the new weight will be $c_0 u(y)$ because of (6). Thus the weight will be stabilized at $w(x) = c_0 u(x)$ once $c_0 u_0 \geq 1$, and the equilibrium distribution of x will be $g(x)$. Therefore, for any starting value of w, the weight process will have the tendency to climb until

$$W_t \geq \max_{y: \, T(X_t, y) > 0} \frac{u(X_t)}{u(y)}.$$

After that, the weight stabilizes to the degenerate distribution as described.

The behavior of the R-type move is more complicated. Here we give a simple example where W_t diverges and show how this defect can be fixed. For simplicity, we assume that T is symmetric and π is uniform on $\mathcal{X} = \{1, 2, 3\}$. Then is is easy to see that

$$W' = \begin{cases} w + 1 & \text{if } U \leq \frac{w}{w+1} \\ w(w+1) & \text{otherwise.} \end{cases}$$

Therefore, the sequence of W is monotone increasing, and it is easy to show that the W process diverges to infinity with

probability 1. A similar construction can be made for an arbitrary reversible T to show the nonexistence of the weight distribution.

A simple way to fix this problem is to modify the weight (3) by a random multiplier; that is,

$$W_{t+1} = \begin{cases} V(wr(x, y) + 1) & \text{if accepted} \\ Vw(wr(x, y) + 1) & \text{if rejected,} \end{cases} \quad (7)$$

where $V \sim \text{unif}(1 - \delta, 1 + \delta)$ is drawn independent of the X_t. It is easy to see that this modified R-type move still satisfies IWIW. The parameter δ needs to be chosen properly so that $E(\log V)$ is not too small. Using the same argument presented in Section 4.4, we can show that a stable distribution of W exists for the modified scheme.

4.3 Case C: $\theta = 0$ and $T(x, y)$ is Nonreversible

The dynamic weighting schemes are most useful when combined with the regular Metropolis–Hastings moves, which typically result in a nonreversible proposal transition. Here we simplify the situation by directly considering a nonreversible proposal chain. When $\theta = 0$, the Q-type and R-type moves are identical, and the weight process is a deterministic function of the Markov chain $\{X_t\}$ controlled by the transition function $T(x, y)$.

Lemma 1. Let $g(x_0)$ denote the marginal equilibrium distribution under transition T, let $g(x_0, x_1) = g(x_0)T(x_0, x_1)$, and let $\delta(x, y) = \log[T(y, x)/T(x, y)]$. Then

$$e_0 = E_g \delta(x_0, x_1) \leq 0.$$

The equality holds only when T induces a reversible Markov chain.

Proof. By definition, we have

$$e_0 = E_g \left\{ \log \frac{T(x_1, x_0)}{T(x_0, x_1)} \right\}$$

$$= \int \log \frac{T(x_1, x_0)}{T(x_0, x_1)} g(x_0)T(x_0, x_1) dx_0 dx_1$$

$$= \int \left\{ \log \frac{g(x_1, x_0)}{g(x_0, x_1)} + \log \frac{g(x_0)}{g(x_1)} \right\} g(x_0, x_1) dx_0 dx_1$$

$$= E_g \left\{ \log \frac{g(x_1, x_0)}{g(x_0, x_1)} \right\} + E_g \log(g(x_0)) - E_g \log(g(x_1))$$

$$= E_g \left\{ \log \frac{g(x_1, x_0)}{g(x_0, x_1)} \right\} \leq \log E_g \left\{ \frac{g(x_1, x_0)}{g(x_0, x_1)} \right\} = 0.$$

The last line follows from Jensen's inequality, in which the equality holds only when $g(x_0, x_1) = g(x_1, x_0)$. Hence the lemma is proved.

Because $\log r(x, y) = \log \pi(y) - \log \pi(x) + \delta(x, y)$, and

$$\log W_t = \log W_{t-1} + \log \pi(X_t)$$
$$- \log \pi(X_{t-1}) + \delta(X_{t-1}, X_t),$$

we have

$$\log W_t = \log \pi(X_t) - \log \pi(X_0) + \sum_{s=1}^{t} \delta(X_{s-1}, X_s)$$

$$\leq c + \sum_{s=1}^{t} \delta(X_{s-1}, X_s).$$

Thus Lemma 1 implies that under stationarity (with T as the transition function), process $\log W_t$ is bounded above by a cumulative sum of terms with negative drift e_0. Because $1/t \sum_{s=1}^{t} \delta(X_{s-1}, X_s) \to e_0$ almost surely (because of the ergodicity theorem), we see that the weight process goes to 0 almost surely. Summarizing the foregoing arguments, we have the following result.

Proposition 1. If the proposal transition $T(x, y)$ is nonreversible and the control parameter $\theta = 0$, then no stable distribution of W_t can exist for the Q- or R-type moves.

4.4 Case D: $\theta = 1$ and $T(x, y)$ is Nonreversible

Consider the log-weight process for the Q-type move:

$$\log W_t = \begin{cases} \max\{0, \log W_{t-1} + \log r(X_{t-1}, X_t)\} & \text{if accept} \\ \log W_{t-1} + \log a & \text{if reject,} \end{cases}$$

$$(8)$$

where the acceptance probability is $\min\{1, W_{t-1}r(X_{t-1}, X_t)\}$.

We observe that when W_t is large (so that $W_t r(x, y) \geq 1$, $\forall x, y$), the log-weight process is controlled by $\delta(X_{t-1}, X_t)$, which has a negative expectation according to Lemma 1, provided that the distribution of (X_{t-1}, X_t) is sufficiently close to g. This produces a negative force to prevent the process from drifting to infinity. The rejection step in the Q-type move plays the role of a reflection boundary to prevent the log-weight process from drifting to negative infinity. To avoid measure-theoretic technicality, in the rest of the article we assume that X_t is defined on a finite state space.

Theorem 2. Suppose that the sample space $\mathcal{X}$ of X_t is finite and the proposal transition $T(x, y)$ is nonreversible. Then the process $(X_t, \log W_t)$ induced by the Q-type move is positive recurrent and has a unique equilibrium distribution.

Proof. Let $Y \sim T(X_t, \cdot)$. The Q-type move induces the following updates:

$$(X_{t+1}, \log W_{t+1})$$

$$= \begin{cases} (Y, \max\{0, \log W_t + \log r(X_t, Y)\}) & \text{if accepted} \\ (Y, \log W_t + \log a) & \text{if rejected.} \end{cases}$$

Let r_0 be the minimal Metropolis ratio as defined in (5). Then the acceptance probability, $\min\{1, W_t r\}$, is at least r_0.

For the starting value (X_0, W_0), we define $V_0 = \log W_0$ and

$$V_t = V_{t-1} + \log r(X_{t-1}, X_t)$$

$$\equiv V_0 + \log \frac{\pi(X_t)}{\pi(X_0)} + \sum_{s=1}^{t} \delta(X_{s-1}, X_s).$$

We assume that $V_0 \leq a/r_0$, which is the upper bound of $\log W_t$ immediately after a rejection, and $\log[\pi(X_t)/\pi(X_0)] \leq m_0$ for all X_t. Define and let $\tau_0 = \min\{t > 0, V_t \leq 0\}$. Note that for $t < \tau_0$, $V_t \equiv \log W_t$. When τ_0 occurs, $\log W_t$ can be either 0 or equal to $V_{t-1} + \log a$.

Because the state space $\mathcal{X}$ is finite and $T(x, y) > 0$ if and only if $T(y, x) > 0$, function $\delta(x, y)$ is bounded from above. Thus for a very large N,

$$P(\tau_0 > N) \leq P\left\{\sum_{t=1}^{N} \delta(X_{t-1}, X_t) + V_0 + \log\frac{\pi(X_t)}{\pi(X_0)} > 0\right\}$$

$$\leq P\left\{\frac{1}{N}\sum_{t=1}^{N}(\delta(X_{t-1}, X_t) - e_0)\right.$$

$$\left. > -e_0 - \frac{1}{N}\left(\frac{a}{r_0} + m_0\right)\right\}$$

$$\leq c_0 \exp(-Nd_0),$$

where d_0 is related to the spectral gap of the corresponding Markov chain. The last inequality follows from theorem 3.3 of Lezaud (1998); an similar inequality was also given by Dembo and Zeitouni (1993).

Let $\zeta_0 = \min\{t > 0, \log W_t = 0\}$ and let R_n be the total number of rejections occurring in the first n iterations. When $n < \zeta_0$ (i.e., before the occurrence of renewal event $\log W_t = 0$), R_n is also the total number of times that the "testing event" $W_t r(X_t, Y) < 1$ has occurred. (Otherwise acceptance is with 1 probability 1.) Because $W_t \geq 1$ for all t as in our design and r is bounded below by r_0, the probability for rejection in each testing event is at most $(1 - r_0)$. On the other hand, because of the effect of a negative drift in the weight update, the "testing event" must occur frequently. More precisely, we have the following computation for a sufficiently large N:

$$P(\zeta_0 > N) = P\left(\bigcup_{k=0}^{N}\{R_N = k; \zeta_0 > N\}\right)$$

$$= P\left(\bigcup_{k=0}^{L-1}\{R_N = k; \zeta_0 > N\}\right)$$

$$+ P\left(\bigcup_{k=L}^{N}\{R_N = k; \zeta_0 > N\}\right)$$

$$\leq P\left(\bigcup_{k=0}^{L-1}\{R_N = k; \zeta_0 > N\}\right) + (1 - r_0)^L/r_0$$

$$\leq \sum_{k=0}^{L}(k+1)c_0 e^{-d_0 N/k} + (1 - r_0)^L/r_0$$

$$\leq c_1 L^2 e^{-d_0 N/L} + c_2 e^{-d_1 L}$$

Letting $L = \sqrt{N}$, we have $P(\zeta_0 > N) \leq c_2 N e^{-d_2\sqrt{N}}$. Hence $E\zeta_0 < \infty$, which shows that the set $\{(x, \log w) : x \in \mathcal{X}, \log w = 0\}$ is positive recurrent. Because $\mathcal{X} \times \{0\}$ is a finite set, there must be a point $x_0 \in \mathcal{X}$ so that $(x_0, 0)$ is a regeneration set (see Asmussen 1987, pp. 150–151); thus the chain is Harris ergodic. By theorem 3.6 of Asmussen (1987, pp. 154–155), we conclude that the distribution of (X_t, W_t) converges to a unique stationary distribution in total variation.

Remark 1. The foregoing proof relies on the verification of the fact that at least one point in the joint space of X and $\log W$ is positive recurrent. When $\mathcal{X}$ is a continuous but compact state space, however, the same argument cannot be carried through. We can consider modifying the weight update in the Q-type move to

$$W_t = \delta_t W_{t-1} r(X_{t-1}, X_t), \quad \delta_t \sim \text{unif}(1 - \delta, 1 + \delta).$$

Note that this does not affect the IWIW property of the move. The purpose of this modification is to create a "continuous" component for W_t that is useful in finding a "small set" (Nummelin 1984). Briefly, we imagine cutting $\mathcal{X}$ into a large finite number of tiny pieces. Then under the same condition (bounded Metropolis ratio, etc.) and by the same argument as in Theorem 2, there exists at least one piece, B_0 say, so that the set $\{(x, 0) : x \in B_0\}$ is positive recurrent. Consider a uniform measure on $A_0 \times (-\epsilon, \epsilon)$ for a sufficiently small ϵ and subset $A_0 \subset B_0$. We can show that (3.1) of in Asmussen (1987, p. 150) holds. Thus the chain is Harris ergodic and has a unique stationary distribution.

Remark 2. For the R-type move, we need also to modify its weight transition by multiplying an independent random variable μ_t with mean 1, as suggested in Section 4.2. Because $E[\log \mu_t] < 0$, the modification produces an extra negative drift for the weight process. The new update is

$$\log W_t = \begin{cases} \log(W_{t-1} r(X_{t-1}, X_t) + 1) + \log \mu_t & \text{if accept} \\ \log W_{t-1} + \log(W_{t-1} r(X_{t-1}, X_t) + 1) \\ \quad + \log \mu_t & \text{otherwise.} \end{cases}$$

$$(9)$$

The rejection probability is $1/(W_{t-1} r(X_{t-1}, X_t) + 1)$. When W is stochastically large, the rejection probability is negligible. Thus the X process is controlled by T in this case. A similar argument to that for the Q-type move can be applied to show that $\log W_t$ comes back to, say $\{w < A\}$ for a suitable A infinitely often. Then, because μ_t has a smooth density, we can argue similarly to Asmussen (1987) that the process has a unique stationary distribution.

4.5 Case E: Mixing Different Types of Moves

Suppose that in each iteration we conduct a Q-type move with probability α and conduct a M-type transition with probability $1 - \alpha$. As in Section 4, we let $\theta = 1$. When w is sufficiently large, there will always be acceptance. Thus the actual transition when w is large is of the form

$$A(x, y) = \alpha A_1(x, y) + (1 - \alpha)A_2(x, y),$$

where $A_1(x, y)$ is just the proposal transition for the Q-type move (because there is no rejection) and $A_2(x, y)$ is a Metropolis-type transition that has π as its invariant distribution. Let $g(x)$ the the invariant distribution of $A(x, y)$, and let δ be the indicator variable that tells which type is conducted.

Then when we make an accepted move from x to y, the weight is updated as

$$w(y) = w(x)\frac{\pi(y)A_\delta(y, x)}{\pi(x)A_\delta(x, y)}.$$

Hence

$$\log\frac{w(y)}{\pi(y)} = \log\frac{w(x)}{\pi(x)} + \log\frac{A_\delta(y, x)}{A_\delta(x, y)}.$$

Similarly,

$$E\log\frac{A_\delta(y, x)}{A_\delta(x, y)} = E\log\frac{g(y)A_\delta(y, x)}{g(x)A_\delta(x, y)}$$

$$\leq \log E\frac{g(y)A_\delta(y, x)}{g(x)A_\delta(x, y)} = 0.$$

This inequality implies that the same argument as inTheorem 2 can be applied.

5. THE LAW OF LARGE NUMBERS FOR THE WEIGHTED MONTE CARLO ESTIMATES

Suppose that a set of weighted samples, $(x_1, w_1), \ldots, (x_m, w_m)$, is obtained by running either a Q-type or an R-type scheme. The quantity $\mu = E_\pi\rho(X)$ is of interest. Then the standard importance sampling estimate of μ is

$$\hat{\mu} = \frac{w_1\rho(x_1) + \cdots + w_m\rho(x_m)}{w_1 + \cdots + w_m}. \tag{10}$$

But because the weights derived from the Q-type or R-type moves may have infinite expectations, it is not clear whether estimate (10) is still valid. In Section 5.1 we show by a general weak law of large numbers that this estimate still converges, albeit very slowly. We then suggest a stratified truncation method to improve estimation. We give justifications for why stratified truncation works in Section 6.

5.1 Convergence In Probability

The most general weak law of large numbers (WLLN), due to Kolmogorov and Feller, has been given by Chung (1974, thm, 5.2.3). To suit our purposes, we state a variation of the original theorem.

Lemma 2. Let $\{Y_n\}$ be a sequence of iid random variables with distribution function F, and $S_n = \sum_{j=1}^n Y_j$. Let $\{b_n\}$ be a given sequence of real numbers increasing to $+\infty$. Suppose that we have

(a) $n\int_{|y| > b_n} dF(y) = o(1)$ and
(b) $n\int_{|y| \leq b_n} y^2 dF(y)/b_n^2 = o(1)$.

Then, if we put $a_n = n\int(b_n \wedge |y|)dF(y)$, where $a \wedge b \equiv \min(a, b)$, we have

$$\frac{1}{b_n}(S_n - a_n) \to 0 \quad \text{in probability.}$$

Proof. Let $a'_n = n\int_{|y| < b_n} ydF(y)$. Theorem 5.2.3 of Chung (1974) states that $(S_n - a'_n)/b_n \to 0$ in probability. Because $(a_n - a'_n)/b_n = o(1)$, the lemma is proved.

Suppose that one wishes to estimate the probability $P(X = x) = E_\pi I(X = x)$. We define b_n so that $b_n = a_n$; that is, b_n is the solution of

$$b_n = nE\{(b_n \wedge W)I[X = x]\}.$$

In the next section we show that for the Q-type move, the tail distribution of $\log W$ is exponential with rate 1. Hence b_n must be chosen so that $b_n = O(n + n\log b_n)$, which leads to $b_n = O(n(\log n + \log(n + \cdots)))$. For the R-type move, the tail of $\log W$ is exponential with rate $\alpha \leq 1$; then we can solve $b_n = a_n$ to derive that $b_n = O(n^{1/\alpha})$.

With b_n so chosen, we have the WLLN for the weighted estimate,

$$\lim_{n\to\infty}\frac{1}{b_n(x)}\sum_{i=1}^n(w_i I(x_i = x) - a_n) = 0.$$

The truncation point b_n is directly related to the point of interest x. If another point, say x', is of interest, then it is easily seen that

$$\lim_{n\to\infty}\frac{\sum_{i=1}^n w_i I(x_i = x)}{\sum_{i=1}^n w_i I(x_i = x')} = \lim_{n\to\infty}\frac{b_n(x)}{b_n(x')}$$

$$= \lim_{n\to\infty}\frac{E\{(W \wedge b_n(x))I(X = x)\}}{E\{(W \wedge b_n(x))I(X = x')\}}.$$

Thus we have shown that the standard estimate (10) of the target (discrete) distribution is getting closer to a truncated-distribution expectation,

$$\hat{\pi}_n(x) = \frac{E[\{W \wedge b_n(x)\}I(X = x)]}{\sum_{x'} E[\{W \wedge b_n(x')\}I(X = x')]}. \tag{11}$$

In the next section we explain why $\hat{\pi}_n$ approaches π as $n \to \infty$. The expression (11) suggests that the truncation should be stratified according to x. Finally, we use a simple example to demonstrate that $\hat{\rho}$ in (10) converges very slowly, seemingly at a rate of $\log(n)$.

5.2 Stratified Truncation

To deal with the large variance of importance weights, we recommend using the method of stratified truncation for estimation. Suppose that we wish to estimate $\mu = E_\pi\rho(X)$. First, the sample points are stratified according to the value of the function $\rho(x)$. Within each stratum function, ρ should be as close to constant as possible and the sizes of the strata are comparable. The highest $k\%$ (usually $k = 1$ or 2) of the weights within each stratum are then trimmed down to the value of the $(100 - k)$th percentile of the weights within stratum.

In Section 7.3 we are interested in estimating the expected value of the spontaneous magnetization defined as $\mu = E\rho(X) = E \mid \sum_i \sigma_i \mid /d^2$, where $X = (\sigma_i, i \in d \times d$ lattice points) follows an Ising model. First, the range of $\rho(x)$ is divided into small intervals $b_0 < b_1 < \cdots < b_k$, and stratify the weighted samples (x_t, w_t), $t = 1, \ldots, m$, according to the values of $\rho(x)$; that is, construct $S_j = \{(x_t, w_t) : \rho(x_t) \in (b_{j-1}, b_j)\}$. Then the w_t in each stratum S_j is truncated to $\tilde{w}_t = w_t \wedge w_j^*$, where w_j^* is the $(100-k)$th percentile of the weights in S_j. Finally, an estimation of μ is

given by (10) with the w_t replaced by the $\tilde{w}_t$. In light of the WLLN, this estimate after truncation makes sense and converges to something of a similar form as the limit of the raw weighted estimate. However, it is less clear why either the estimate with stratified truncation or the raw estimate (10) gives desirable results. The next section provides a theoretical basis for the appropriateness of using the former.

If the function of interest $\rho(X)$ is too degenerate (e.g., a step function with only a few different values), then the stratification should be further refined. A general strategy is to stratify according to both values of $\rho(X)$ and the sampled probability values, $\pi(X)$. Knowing π only up to a normalizing constant does not affect the stratification.

6. TAIL BEHAVIOR OF THE WEIGHTS AND STRATIFIED TRUNCATION

To investigate the theory behind stratified truncation, we consider stratifications based on the value of x. For simplicity, we analyze the Q-type move only. Suppose that the nonreversible Markov chain $\{X_t\}$ is controlled by the transition function $T(x, y)$, and define a cumulative sum process of the form

$$V_t = V_{t-1} + \varphi(X_{t-1}, X_t),$$

where φ is some smooth function such as a log-likelihood ratio. Kesten (1974) provided a renewal theory for the behavior of such a process. His result can be adapted to derive the extremal behavior of V_t.

Lemma 3. Suppose that $X_0, X_1, \ldots$ is an irreducible and aperiodic Markov chain induced by the transition function $T(x, y)$. Let $V_t = \sum_{s=1}^{n} \varphi(X_{s-1}, X_s)$ for some function φ. Suppose that we can find a function $r: \mathcal{X} \to (0, \infty)$ and $\kappa > 0$ such that

$$r(x) = E_T[e^{\kappa \varphi(x, X_1)} r(X_1) | X_0 = x]. \tag{12}$$

Then

$$\lim_{c \to \infty} e^{\kappa c} P[\max_t V_t > c \mid X_0 = x] = K(\varphi) r(x),$$

where K is some constant independent of x (but may depend on function φ).

Proof. The proof has been given by Kesten (1974, sec. 4).

Lemma 4. Suppose that $V_t = \sum_{s=1}^{t} \log r(X_{s-1}, X_s)$, where $r(x, y)$ is the Metropolis ratio as defined in (1). Then

$$\lim_{c \to \infty} e^c P\left[\max_t V_t > c \mid X_0 = x\right] = Kg(x)/\pi(x),$$

where $g(x)$ is the stationary distribution of the X_t.

Proof. If we let $\varphi(x, y) = \log r(x, y)$, $\kappa = 1$, and $r(x) = g(x)/\pi(x)$ in Lemma 3, then it is easy to verify that (12) is satisfied. Hence the result holds.

Now we define the process Z_t, which differs from V_t by having a reflecting boundary at 0:

$$Z_t = \begin{cases} Z_{t-1} + \log r(X_{t-1}, X_t) & \text{if } Z_{t-1} + \log r(X_{t-1}, X_t) > 0 \\ 0 & \text{if } Z_{t-1} + \log r(X_{t-1}, X_t) < 0. \end{cases}$$

Because the behavior of the process Z_t is similar to V_t, we expect its tail probability to also have an exponential decay.

Lemma 5. Suppose that $Z_0 = 0$ and that X_0 starts from its equilibrium distribution, $g(x)$. Then the tail probability $P(Z_t > c \mid X_t = x)$ decays exponentially with rate 1, and $\lim_{c \to \infty} e^c P(Z_t > c \mid X_t = x) = Ku(x)$, where $u(x) = \pi(x)/g(x)$.

Proof. A way to look at Z_t is to relate it to the process V_t. Specifically, we have

$$Z_t = \max_{1 \le k \le t}(V_t - V_k) = \max_{1 \le k \le t} \sum_{s=k+1}^{t} \log r(X_{s-1}, X_s).$$

Now if we imagine a "reversal process" $Y_0^*, Y_1^*, \ldots$, governed by the conjugate transition $T_*(x, y) = g(y)T(y, x)/g(x)$, then the joint distribution of $(X_0, \ldots, X_t)$ is the same as the joint distribution of $(Y_t^*, \ldots, Y_0^*)$. Furthermore, we can identify X_s with Y_{t-s}^* so that

$$\log r(X_{s-1}, X_s) = \log \frac{\pi(X_s)T(X_s, X_{s-1})}{\pi(X_{s-1})T(X_{s-1}, X_s)}$$

$$= \log \frac{u(X_s)T(X_s, X_{s-1})}{u(X_{s-1})T_*(X_s, X_{s-1})}$$

$$= \log \frac{u(Y_{t-s}^*)T(Y_{t-s}^*, Y_{t-s+1}^*)}{u(Y_{t-s+1}^*)T_*(Y_{t-s}^*, Y_{t-s+1}^*)}$$

$$\equiv \delta_*(Y_{t-s}^*, Y_{t-s+1}^*).$$

Thus

$$Z_t \stackrel{\mathcal{L}}{=} Z_t^* \equiv \max_{1 \le k \le t} \sum_{s=1}^{k} \delta_*(Y_{s-1}^*, Y_s^*).$$

By Lemma 3 applied to the process Z_t^* with $\varphi(x, y) = \delta_*(x, y)$, $\kappa = 1$ and $r(x) = u(x)$, we obtain that $\lim_{c \to \infty} e^c P(Z_t > c \mid X_t = x) = \lim_{c \to \infty} e^c P(\max_t Z_t^* > c \mid Y_0^* = x) = Ku(x)$, which proves the result.

Now we go back to our Q-type moves. If no rejection occurs, then $\log W_{t+1} = \log W_t + \log r(X_t, X_{t+1})$. Because rejection occurs only when the weight is relatively small, for large c we expect the $\log W_t$ process to behave similarly to the Z_t process. Thus we expect the process $\log W_t$ to satisfy

$$\lim_{c \to \infty} P(\log W_t > c \mid X_t = x) = K' u(x).$$

To show that the $\log W_t$ process behaves similarly to the Z_t process, we study sojourns of both the Z_t and the U_t^* above a large positive value A. Consider event $\{Z_t > A\}$, and let τ_j be the jth crossing time of the process (i.e., the jth time that event $\{Z_{\tau_j-1} < A \le Z_{\tau_j}\}$ occurs). Then the random variable $S_j = Z_{\tau_j} - A$ has a stationary distribution. Because $\log r(x, y)$ is bounded, S_j is also bounded. Then if the limits exist, they satisfy

$$\lim_{B \to \infty} e^B P(\log W_t > A + B \mid U_t > A, X_t = x)$$

$$= \lim_{B \to \infty} e^B P(Z_t > B - S_j \mid X_t = x) = KE(e^{S_j}).$$

Similarly, we let ζ_j be the jth crossing time of the $\log W_t$ process and let $T_j = \log W_{\zeta_j} - A$. Then if A is sufficiently large,

the behavior of $\log W_t$ conditional on that of $\log W_t > A$ is the same as that of Z_t conditional on $Z_t > A$. Thus

$$\lim_{B \to \infty} e^B P(\log W_t > A + B, X_t = x \mid U_t > A)$$

$$= \lim_{B \to \infty} e^B P(Z_t > B - T_A \mid X_t = x)$$

$$= K u(x) E(e^{T_j}).$$

Hence, up to a constant, the limiting behavior of the tail probability of $\log W_t$ is identical to that of Z_t.

The forgoing argument demonstrates that when k is small, the conditional $(100\text{-}k)$th percentile $q_k(x)$ of the weights satisfies approximately the relationship

$$q_k(x) \propto u(x) = \pi(x)/g(x).$$

If we draw lines to connect the $(100\text{-}k)$th percentiles of the weights for, say, x and x', the lines should parallel to each other for different k's. Because no rejections for those X_t's are associated with the large W_t, the distribution of these X_t is close to g.

Now for any pair x and x', we let $\log b$ and $\log b'$ be the $(100\text{-}k)$th percentiles of $[W_t \mid X_t = x]$ and $[W_t \mid X_t = x']$, with $k \to 0$. Let a and a' be the $(100\text{-}k')$th percentiles, where k' goes to 0 at a slower rate than k. Then we have

$$\frac{E[W_t \wedge b \mid I(X_t = x)]}{E[W_t \wedge b' \mid I(X_t = x')]}$$

$$= \frac{E[e^{\log W_t \wedge \log b} \mid I(X_t = x)]}{E[e^{\log W_t \wedge \log b'} \mid I(X_t = x')]}$$

$$\approx \frac{\sum_{j=[a]}^{[\log b]} E\{e^{\log W_t \wedge \log b} I[\log W_t \in (j, j+1)] \mid I(X_t = x)\}}{\sum_{j=[a']}^{[\log b']} E\{e^{\log W_t \wedge \log b'} I[\log W_t \in (j, j+1)] \mid I(X_t = x')\}}$$

$$\approx \frac{\sum_{j=[a]}^{[\log b]} e^1 \times e^{-1} u(x)}{\sum_{j=[a']}^{[\log b']} e^1 \times e^{-1} u(x')} \approx \frac{(\log b - a) u(x)}{(\log b' - a') u(x')} \approx \frac{u(x)}{u(x')}.$$

The final approximation holds because the $\log W_t$ is approximately exponentially distributed, implying that $(\log b - a)/(\log b' - a') \to 1$. The foregoing argument explains why $\hat{\pi}(x)$ in (11) approaches $\pi(x)$ as $n \to \infty$. Thus the stratified truncation method outlined in Section 4.2 gives the desirable estimate. This conclusion is further supported by a simulation study and some real examples in Section 7.

Because the exponential decay rate of the log weight is 1, the expectation of W_t for the Q-type process is infinite. It should be noted that the infinite weight expectation is not necessarily a bad thing; it helps the chain escape from a local mode effectively. The phenomenon is also a logical consequence of the dynamic weighting philosophy. The method transforms a waiting time (i.e., the time for the chain to reach equilibrium) infinity to an importance weight infinity.

7. SIMULATION STUDIES AND REAL EXAMPLES

7.1 Simulation Studies

To understand detailed performances of both the Q-type and R-type moves, we designed a simulation to check several predictions of our theory:

a. The tail distribution of the log-weight in a Q-type move is exponential with decay rate $=1$, and that of a R-type is exponential with decay rate $\beta \le 1$.
b. Upper percentiles of the stratified weights are approximately proportional to $u(X_t)$.
c. The plain importance sampling estimate (10) converges slowly, but to the correct mean.
d. Estimation with stratified truncation gives us an approximately correct answer.

To achieve the stated purposes, we let the state space of X be $\{1, 2, 3, 4, 5\}$, and generated a random 5×5 transition matrix [with each row drawn independently from Dirichlet $(1, 1, 1, 1, 1)$]:

$$T = \begin{pmatrix} .00370 & .15436 & .55588 & .15998 & .12608 \\ .18506 & .34190 & .17511 & .14471 & .15322 \\ .27798 & .26276 & .16575 & .21687 & .07664 \\ .29265 & .28028 & .22982 & .15994 & .03731 \\ .25206 & .23105 & .02426 & .22976 & .26287 \end{pmatrix}.$$

It is easy to verify that the invariant distribution of T is $g = (.1987, .2611, .2398, .1782, .1222)$. We took the target distribution $\pi = (.25, .1, .2, .4, .05)$. With $a = 2$, a Q-type process was initiated with $W_0 = 1$, and $X_0 \sim g$. A total of 200,000 iterations were carried out. Figure 1(a) shows the percentiles of weights stratified according to the state space. The percentages range from 70% to 99%. The q-q plot [Fig. 1(c)] shows that the tail of the weights is like that of an exp(1) distribution. Estimating π by using stratified truncation at $k\% = 1\%$ and 5% gives $\hat{\pi} = (.2453, .0984, .2001, .4071, .0491)$ and $(.2449, .1023, .1994, .4049, .0485)$. These results confirmed our predictions a–c. To show the slow convergence of the raw estimate, we ran 2^{30} iterations, estimating π by (10) at every 2^k epoch. Figure 1(d) shows the plot of the standardized error of these estimates [i.e., $(\sum_{i=1}^{5} (\pi_i - \hat{\pi})^2/\pi_i)^{1/2}$] versus the logarithm of the number of iterations.

We also applied the R-type moves to the same problem. The corresponding results are similar to those of the Q-type moves (figures omitted). The weights resulting from R-type moves are appreciably greater than those from Q-type moves, and the tail distribution of the weights seems to still be exponential but with a changing rate α that approaches to one as the quantiles become extreme.

To verify that our analysis can be extrapolated to more complicated cases, we considered a Bayesian testing problem. Suppose that we wish to test whether a sequence of binary observations, $y = (y_1, \ldots, y_n)$, is iid (null model) or whether they form a stationary Markov chain (alternative model). The Markov model can be parameterized by $\theta = (\theta_0, \theta_1)$, where $\theta_i = P(y_{s+1} = 1 \mid y_s = i)$. Then the parameters in the iid model lie in the subspace corresponding to $\theta_0 = \theta_1$. Let M be the model indicator, 0 for the null and 1 for the alternative. Suppose that the two models are equally likely a priori. Then jointly we have

$$P(y, \theta, M) = \theta_0^{n_{01}} (1 - \theta_0)^{n_{00}} \theta_1^{n_{11}} (1 - \theta_1)^{n_{10}}$$

$$\times \left[\frac{1}{2} (1 - M) \delta_{\theta_0 = \theta_1} + \frac{1}{2} M \right],$$
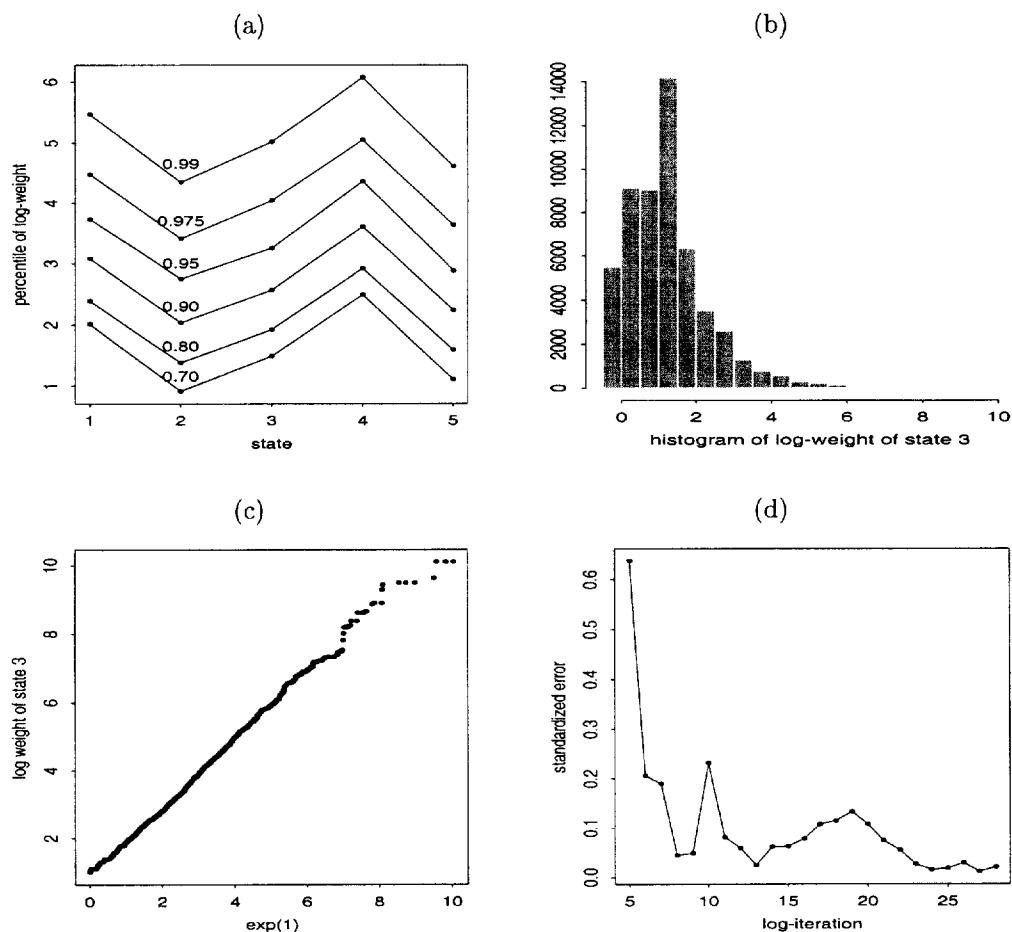
Figure 1. Results for the Simulation Study With Q-Type Moves. (a) The conditional percentiles of the weights. The parallelness of these quantile locations is predicted by theory in Section 5 and is the basis for stratified truncation estimate. (b) The histogram of the log-weights corresponding to X = 3. (c) The q-q plot of the upper tail of the log-weights versus exp(1). (d) Convergence of the raw weighted estimates, even though the weight has an infinite expectation.

where $n_{ij} = \sum_{s=2}^{n} I_{(y_{s-1}=i)} I_{(y_s=j)}$. The Bayes factor $B = P(M = 0 \mid y)/P(M = 1 \mid y)$ can be computed from this joint distribution. In many applications, however, an analytical expression for $P(M \mid y)$ is impossible. An MCMC procedure is also difficult to implement because the parameter space involves a degenerate part. Green (1995) described a variant of the Metropolis–Hastings algorithm, termed *reversible jumps*, that can be used overcome this difficulty. Besides the usual Metropolis-type moves, the reversible jump algorithm also specifies a pair of proposals for "jumping" between two spaces. For example, we can propose a jump from $M = 1$ to $M = 0$ as $(\theta_0, \theta_1) \rightarrow \theta_0$ and a reverse jump as

$$\theta_0 \rightarrow (\theta_0, \theta_1^*), \quad \text{with } \theta_1^* \sim \text{unif}[0, 1].$$

Whether to accept or to reject this proposal is determined by the usual Metropolis–Hastings rule. As noted by Liu and Sabatti (1998), the jump proposals are rarely accepted in complicated applications, and dynamic weighting schemes can be applied to help.

With a data sequence 0001010101000011100010101100, we used both the Q-type and R-type rules for between-space moves and reserved the M-type rules for within-space moves. When we stratified the dynamic weights only according to the value of M for weight truncations, the results had a very large

bias. We made a refinement by stratifying the weights on both the M value and the log-likelihood value (i.e., divided the lower-dimensional space into 10 parts and the higher dimensional space into 15 pieces according to the $\log P(\theta, y, M)$ value). The estimates based on 99.9%, 99%, and 95% truncations were 1.19, 1.22, and 1.23. For this simple example, the exact answer for the Bayes factor is $B = 1.187$. Figure 2 shows the parallel plot of the percentiles of the weights in each stratum. This example clearly demonstrates that finer stratification reduces the bias in estimation and but it will typically increase the variance. A balance between bias and variance is often important in practice.

7.2 Neural Network Training

The artificial neural network is a simple mathematical model motivated by neuron functions and has been widely used in learning and classification problems (Hopfield 1982; Rumelhart and McClelland 1986). The most popular of these networks is *multilayer perceptrons* (MLP), a type of feedforward network. Our stochastic learning algorithm focuses on the MLP.

In an MLP, all the units (nodes) are grouped into layers (typically three layers). The layers are ordered (i.e., input-hidden-output) so that the units in lower layers (input) connects only
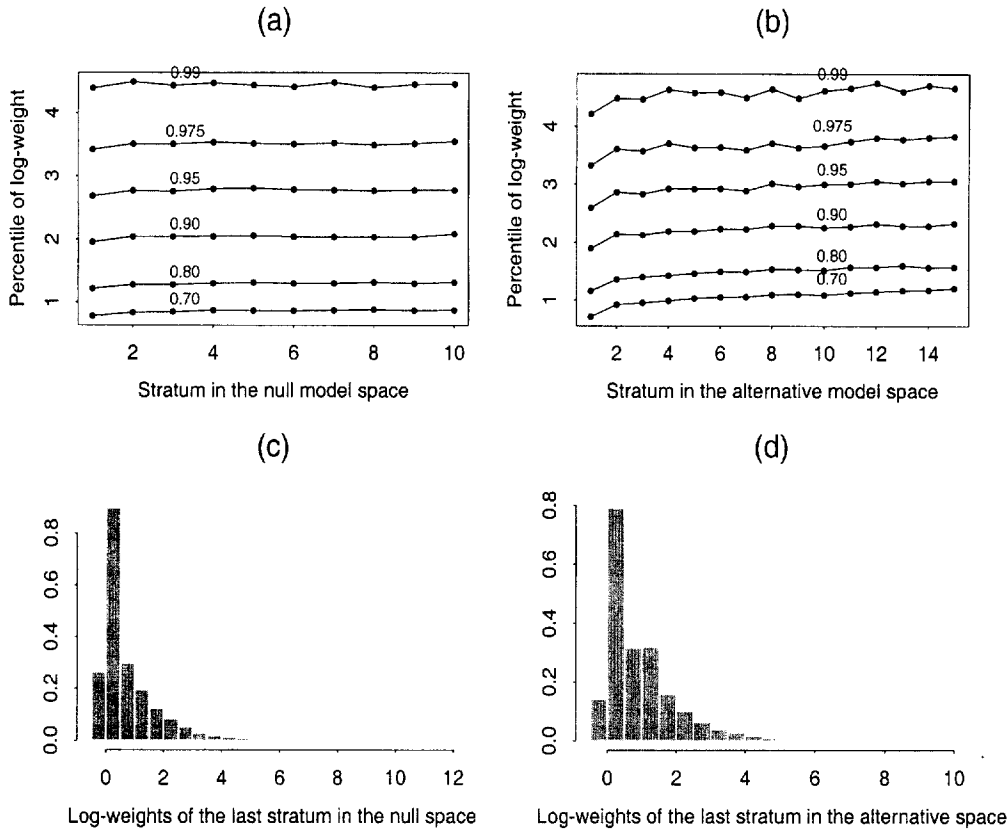
Figure 2. *Parallel Plots of the Stratified Dynamic Weights for the Model Selection Example (a) and (b), and Histograms of the Weights in Two of the Strata (c) and (d).*

with the units in the higher layer (Ripley 1996). Each node in a higher layer independently processes the values fed to it by nodes in the lower layer in the form

$$y_k = f_k\left(\alpha_k + \sum_{j \sim k} w_{jk} x_j\right),$$

where the x_j are inputs, and then present the output y_k as an input for the next layer. Here we take f_k as the same sigmoidal function [i.e., $f(s) = 1/(1 + \exp(-s))$] throughout the network. Neural network "learning" is accomplished by choosing the connection strengths w_{jk} so that the network outputs match the desired output in the training data as closely as possible. Currently, the most popular learning algorithm is the back-propagation and its variants (Rumelhart, Hinton, and Williams 1986). But the back-propagation method can fail badly in some cases, one of which is the two-spiral problem (Lang and Witbrock 1989). By using the dynamic weighting method together with the tempering idea (Geyer and Thompson 1995; Marinari and Parisi 1992), Wong and Liang (1997) treated the two-spiral problem with considerable success. (Both the 2-25-1 and 2-14-4-1 networks have been fitted, and the results were close to perfect, whereas the error rate for back propagation is generally greater than 40%.)

In training programs such as back propagation and Learning Vector Quantization algorithms (Kohonen 1989), the total mean squared error,

$$E_p = \sum_p \|O_p - T_p\|^2,$$

where T_p is the pth training case's ideal output and O_p is the output of the network, is used as the cost function. We use the same cost function and define a probability distribution jointly for the connection strengths w_{jk} and a temperature parameter T so that

$$\pi(w_{jk}, \text{ all } j, k; \ T) \propto \alpha_T \exp(E_p/T),$$

and T represents a finite number of temperature levels, $t_1 > t_2 > \cdots > t_L$. Wong and Liang chose $L = 4$ for the two-spiral problem. Conditional on $T = t_l$, we use a standard Metropolis move to do local changes on the connection strengths (Neal 1996), whereas conditional on the w_{jk}, we use a Q-type move to jump across two temperature levels. After obtaining reasonable configurations of the connection strengths from the lowest-temperature level, we conduct a *postoptimization* to zoom in for the local optimum. Commonly used postoptimization methods include steepest-gradient descent and conjugate gradient. More details of the method were given by Liang (1997).

We now illustrate this method in the encoder problem (Ackley, Hinton, and Sejnowski 1985) and the parity problem (Rumelhart et al. 1986). These two problems have been regarded as classic benchmarks for testing new methods in the neural network community. Their difficulties stem from the stringent noiseless output requirement. The input in the encoder problem is a length-d binary sequence, and the output is desired to be identical to the input. A requirement for the network designed for the task is that the hidden layers cannot have more than $\log_2(d)$ nodes. Apparently, a network with a

hidden layer of d nodes is trivial to design. We trained a three-layer network with five hidden units for $d = 32$ (constituting a 32-5-32 network) without the constant term. Sigmoid was used as the activation function. In this example, we are dealing with a $5 \times 32 \times 2 = 320$–dimensional optimization problem. Our algorithm achieved perfect learning in about 5 minutes on a Sun SPARC-20 workstation. With a longer running time, perfect learning was also achieved on the much more difficult 32-4-32 (with 4 hidden units and 256 scalar parameters involved) encoder problem.

The input of a d-parity problem is also a binary sequence of length d. The output is required to be 1 if the input sequence contains an odd number of 1's, and is 0 otherwise. So this exercise is meant to show how a "black box" network can mimic a very nonlinear and noncontinuous function. Rumelhart et al. (1986) showed that *at least d* hidden units are required for a three-layer MLP to solve this problem. Our method had no difficulty solving this problem with a d-d-1 ($2 \le d \le 8$) network. A perfect solution for $d = 8$ (a 72-dimensional optimization problem) was obtained by Liang (1997).

7.3 Ising Model Simulation at Subcritical Temperatures

Simulations of two-dimensional Ising models and investigating phase transition phenomena present yet another challenge and also a good test to our method. A two-dimensional Ising model on a $L \times L$ lattice is a probability distribution on $x = \{\sigma_i, \text{ with } i = (a, b) \text{ and } 1 \le a, b \le L\}$,

$$\pi(x) = \frac{1}{Z(K)} \exp\left\{ K \sum_{\langle i,j \rangle} \sigma_i \sigma_j \right\},$$

where the spins $\sigma_i = \pm 1$, $\langle i, j \rangle$ denotes the nearest neighbors on the lattice, K is the coupling constant (inverse temperature), and $Z(K)$ is the partition function. This problem and other spin glass models have been extensively studied in the statistical physics literature. Among the proposed Monte Carlo methods for this problem, the clustering approach of Swendsen and Wang (1987) greatly increased the mixing rate but is difficult to generalize to other systems, such as random field Ising models (Marinari and Parisi 1992). Other successful methods include the simulated tempering (Marinari and Parisi 1992) and multicanonical method (Berg and Neuhaus 1991). But the methods may encounter difficulties when simulating an Ising system at a temperature below the critical point (where the energy variation is huge). The multigrid Monte Carlo method of Goodman and Sokal (1989) can be successful for some other models, but is not suitable for the Ising model.

We now review the results obtained on Ising model simulations by dynamic weighting with R-type moves (Liang and Wong 1999). The simulations were done on lattices of size 32^2, 64^2, and 128^2. As with simulated tempering, we treat the inverse temperature K as a dynamic variable taking values in a ladder of suitable chosen levels near the critical point (known to be .44). We applied the R-type moves to cross various temperature levels uniformly spaced in the range [.4, .5], and used the M-type moves within each temperature level. In each of the three lattice sizes, we started a single run with the configuration that all spins are $+1$. The run continued until we obtained 10,000 configurations at the final temperature level. Figure 3(a) plots the estimate of the expected absolute value of the spontaneous magnetization (defined as $E \mid \sum \sigma_i \mid /d^2$, where d is the lattice size) at various inverse temperatures K for the different sizes of lattices. Estimation was done by
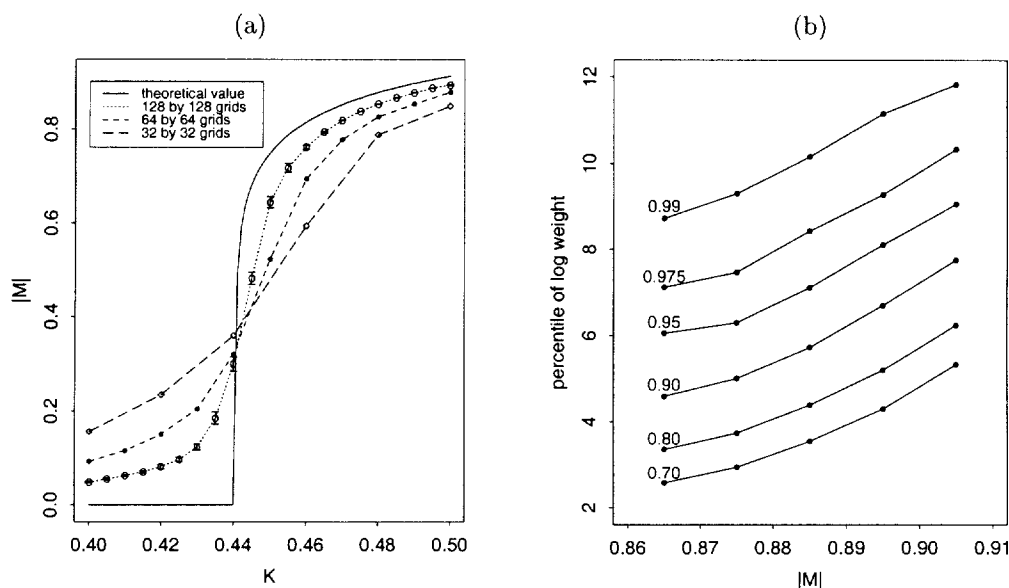

(a)   (b)

Figure 3. The Expected Absolute Value of the Spontaneous Magnetization (defined as $E \mid \sum_i \sigma_i \mid /d^2$, where d is the lattice size) Plotted Against Various Temperature K in (a) for a Two-Dimensional Ising Model With Lattices of Size 32^2 (long-dashed line), 64^2 (short-dashed line), 128^2 (dotted line), and Infinite. The solid line in (a) corresponds to the theoretical infinite lattice result. The plot of the conditional quantiles of the weights at five typical magnetization values is shown in (b). Our theory in Section 5 predicts that the lines connecting the same quantiles should be approximately parallel.

weighted averaging, with the weights stratified according to spontaneous magnetization and then truncated at 99%.

Because the model is many orders of magnitude more complex than the examples in Section 7.1, it is of interest to see if our theory on the behavior of the weights (Sec. 5) still holds in this case. Figure 3(b) shows the upper quantiles of the conditional weight distributions with σ stratified according to five typical values of the spontaneous magnetization for the 128^2 model. The weights behave very much as predicted. The phenomena suggest that for any application of the dynamic weighting method, this parallel graph of conditional quantiles of the weights can serve as a diagnostic tool for judging how well the method works.

The smooth curve in Figure 3(a) is the celebrated infinite lattice result (i.e., the "truth" when the lattice size is infinite) discovered by Onsager (1949) and proved by Yang (1952). It is seen that the critical point (.44) can be estimated quite well from our simulation by the crossing of the curves for the 64^2 and 128^2 model. A major strength of our method is that a single run of the process can yield accurate estimates over the entire temperature range extending well below the critical point. As a comparison, we also applied simulated tempering in the same setting. The scheme was not able to sample both energy wells in the same run in the 64^2 and 128^2 models (see Liang and Wong 1999 for more details).

8. DISCUSSION

This article has presented some theory underlying a new Monte Carlo strategy that combines importance weighting and Markov chain moves. The advantage of the new scheme is that it enables the sampler to search a much larger part of the state space and in the same time respects the constraints given by the target function π. In other words, it moves more freely than a standard MCMC, but is much more "disciplined" than a random walk. This Monte Carlo strategy not only is effective in optimization, but also is useful for Monte Carlo integration/estimation. As was shown by many examples on which we have tried this method, the improvements over existing methods can be substantial. The theory presented in this article can only be regarded as a preliminary understanding of the dynamic weighting method, which we hope will stimulate further research and development of this promising methodology.

REFERENCES

Ackley, D. H., Hinton, G. R., and Sejnowski, T. J. (1985), "A Learning Algorithm for Boltzmann Machines," *Cognitive Science*, 9, 147–169.

Asmussen, S. (1987), *Applied Probability and Queues*, New York: Wiley.

Berg, B. A., and Neuhaus, T. (1991), "Multicanonical Algorithms for First-Order Phase Transitions," *Physics Letters*, Ser. B, 267, 249.

Chung, K. L. (1974), *A Course in Probability Theory*, New York: Academic Press.

Dembo, A., and Zeitouni, O. (1993), *Large Deviations Techniques*, Boston: Jones and Bartlett.

Frenkel, D., and Smit, B. (1996), *Understanding Molecular Simulation: From Algorithms to Applications*, New York: Academic Press.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geyer, C. J., and Thompson, E. A. (1995), "Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference," *Journal of the American Statistical Association*, 90, 909–920.

Goodman, J., and Sokal, A. D. (1989), "Multigrid Monte Carlo Method. Conceptual Foundations," *Physical Review D*, 40, 2035–2071.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Hopfield, J. J. (1982), "Neural Networks and Physical Systems With Emergent Collective Computational Abilities," *Proceedings of the National Academy of Sciences*, 79, 2554–2558.

Kesten, H. (1974) "Renewal Theory for Markov Chains," *The Annals of Probability*, 2, 355–387.

Kirkpatrick, S., Gelatt, C. D. Jr., and Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680.

Kohonen, T. (1989), *Self-Organizing and Associative Memory*, Berlin: Springer-Verlag.

Lang, K. J., and Witbrock, M. J. (1989), "Learning to Tell Two Spirals Apart," in *Proceedings of 1988 Connectionist Models Summer School*, pp. 52–59.

Lawrence C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993), "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment," *Science*, 262, 208–214.

Leach, A. R. (1996), *Molecular Modeling: Principles and Applications*, Singapore: Addison Wesley Longman.

Lezaud, P. (1998), "Chernoff-Type Bound for Finite Markov Chains," *Annals of Applied Probability*, 8, 849–867.

Liang, F. (1997), "Weighted Markov Chain Monte Carlo and Optimization," unpublished doctorial thesis, The Chinese University of Hong Kong.

Liang, F., and Wong, W. H. (1999), "Dynamic Weighting in Simulations of Spin Systems," *Physics Letters*, Ser. A, 252, 257–262.

Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1999), "Markovian Structures in Biological Sequence Alignments," *Journal of the American Statistical Association*, 94, 1–15.

Liu, J. S., and Sabattti, C. (1998), "Simulated Sintering: Markov Chain Monte Carlo with Spaces of Varying Dimension," in *Bayesian Statistics*, 6, 389–413, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York, Oxford University Press.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

Liu, J. S., Liang, F., and Wong, W. H. (2000), "The Multiple-try Method and Local Optimization in Metropolis Sampling," *Journal of the American Statistical Association*, 95, 121–134.

Marinari, E., and Parisi, G. (1992), "Simulated Tempering: A New Monte Carlo Scheme," *Europhysics Letters*, 19, 451.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087-1091.

Neal, R. M. (1996), *Bayesian Learning for Neural Networks*, New York: Springer.

Nummelin, E. (1984), *General Irreducible Markov Chains and Non-Negative Operators*, New York: Cambridge University Press.

Onsager, L. (1949), "Statistical Hydrodynamics," *Nuovo Cimento* (Suppl.), 6, 261.

Ripley, D. (1996), *Pattern Recognition and Neural Networks*, Cambridge, U.K.: Cambridge University Press.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986), "Learning Representations by Back-Propagating Errors," *Nature*, 323, 533–536.

Rumelhart, D. E., and McClelland, J. (1986), *Parallel Distributed Processing: Exploitations in the Micro-Structure of Cognition*, Vols. 1 and 2, Cambridge, MA: MIT Press.

Swendsen, R. H., and Wang, J. S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physics Review Letters*, 58, 86.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distribution by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Wong, W. H., and Liang, F. (1997), "Dynamic Weighting in Monte Carlo and Optimization," *Proceedings of the National Academy of Science*, 94, 14220–14224.

Yang, C. N. (1952), "The Spontaneous Magnetization of a Two-Dimensional Ising Model," *Physics Review*, 85, 808.