The Annals of Applied Probability
1999, Vol. 9, No. 4, 1202–1225

IMPORTANCE SAMPLING FOR FAMILIES OF DISTRIBUTIONS

BY NEAL MADRAS[1] AND MAURO PICCIONI[2]

York University and Universitá di L'Aquila

This paper analyzes the performance of importance sampling distributions for computing expectations with respect to a whole family of probability laws in the context of Markov chain Monte Carlo simulation methods. Motivations for such a study arise in statistics as well as in statistical physics. Two choices of importance sampling distributions are considered in detail: mixtures of the distributions of interest and distributions that are "uniform over energy levels" (motivated by physical applications). We analyze two examples, a "witch's hat" distribution and the mean field Ising model, to illustrate the advantages that such simulation procedures are expected to offer in a greater generality. The connection with the recently proposed simulated tempering method is also examined.

1. Introduction. Monte Carlo methods have long been an indispensable tool in the field of statistical physics [see Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953), Sokal (1989), Binder and Heerman (1992)]. More recently, a similar view has been developing among statisticians [see, e.g., Smith and Roberts (1993), Besag and Green (1993), Tanner (1993), Gilks, Richardson and Spigelhalter (1996), Robert (1996), Gamerman (1997)]. This paper will discuss some procedures that attempt to alleviate two common problems that arise in many Monte Carlo studies:

(i) The need to perform many Monte Carlo runs that differ only in the value of some input parameter(s); and
(ii) A very slow approach to equilibrium of dynamic sampling schemes, which are usually known in statistics as Markov chain Monte Carlo methods.

Many practitioners, in statistics as well as in physics, have observed that suitable variations on the classical technique of importance sampling can often help to overcome both of these problems. For the most part, however, these observations have been largely empirical, based upon experience with a particular set of models. Our main contribution in this paper is to perform a rigorous asymptotic analysis of the behavior of such procedures in two model examples: the mean field Ising model from statistical physics, and the "witch's hat" distribution of Geyer and Thompson (1995). To this end, we establish some basic general results about the efficiency of such implementations of importance sampling, which are valid in general state spaces.

Received July 1997; revised February 1999.
[1]Supported in part by the Natural Science and Engineering Research Council of Canada.
[2]Supported in part by the Ministry of University and Research of Italy.
AMS 1991 *subject classifications*. Primary 60J05; secondary 65C05, 82B80.
Key words and phrases. Markov chain Monte Carlo, importance sampling, simulated tempering, Metropolis algorithm, spectral gap, Ising model.

The essential idea of importance sampling is the following. Suppose that μ is a known probability measure on some measurable state space $(\Omega, \mathscr{B})$ (usually a subset of R^d), and let f be a real-valued integrable function on Ω. We want to compute the expected value

$$E^\mu f = \int_\Omega f(x)\mu(dx) \tag{1.1}$$

but we are unable to evaluate it either exactly or by standard numerical approximations, typically because either d is large or μ is very complicated. The crude Monte Carlo solution is to generate a vector $\mathbf{X}_n^\mu$ of i.i.d. μ-distributed random variables $X_1^\mu, \ldots, X_n^\mu$ and to estimate $E^\mu f$ by the empirical average

$$\hat{f}(\mathbf{X}_n^\mu) = n^{-1} \sum_{i=1}^n f(X_i^\mu), \tag{1.2}$$

which is unbiased (its expectation is $E^\mu f$) and strongly consistent (it converges to $E^\mu f$ almost surely as $n \to \infty$). Moreover if the variance of $f(X_1^\mu)$, denoted by $\sigma_\mu^2(f)$, is finite, then the central limit theorem holds,

$$\sqrt{n}(\hat{f}_n - E^\mu f) \to N(0, \sigma_\mu^2(f)), \tag{1.3}$$

making it possible to evaluate the error of the estimate (the variance being likewise consistently estimated).

One can try to find an estimator with a smaller variance by sampling from a different probability distribution ν on $(\Omega, \mathscr{B})$, such that μ is absolutely continuous with respect to ν (otherwise the sampling process will always miss some nonnegligible part of Ω). If we can generate $\mathbf{X}_n^\nu = (X_1^\nu, \ldots, X_n^\nu)$ i.i.d. with distribution ν, then the empirical average $\hat{g}(\mathbf{X}_n^\nu)$, where $g = f\, d\mu/d\nu$ is the product of f with the importance sampling weights $d\mu/d\nu$, is again an unbiased and strongly consistent estimator of $E^\mu f$. Moreover, the central limit theorem still holds, provided the variance $\sigma_\nu^2(g)$ exists. It is clear that such a variance depends on ν; a good choice of ν can make it dramatically smaller than $\sigma_\mu^2(f)$. The classical guideline for a good choice is that ν should put weight where μ is concentrated and simultaneously f is large, hence the name importance sampling. However, in this paper our choice of ν is determined only by the measure(s) μ, and we adopt a "worst case" approach with respect to the variation of f. A quite different approach to importance sampling is used in rare event simulation, where the choice of ν is heavily determined by the event or function being estimated; see Bucklew (1990) for more on the subject. We emphasize that the measure ν is completely artificial; it is chosen entirely for the convenience of the Monte Carlo experimenter.

It is apparent how importance sampling can help with problem (i) from above. Consider the above procedure if μ, and perhaps f, depend on a parameter $\theta \in \Theta$. Defining $g_\theta = f_\theta\, d\mu_\theta/d\nu$, observe that a single simulation from ν enables us to compute the whole family $\hat{g}_\theta(\mathbf{X}_n^\nu)$ of estimators of $E_\nu g_\theta = E_{\mu_\theta} f_\theta$ (for all $\theta \in \Theta$). However, these estimators cannot be expected to work uniformly well for all θ with a large but fixed value of n, unless ν "covers" all the

parts of Ω where each of the μ_θ's is concentrated. This is ensured, for example, by a boundedness condition on the importance sampling weights

$$\frac{d\mu_\theta}{d\nu}(x) \leq A \quad \text{for all } x \in \Omega \text{ and } \theta \in \Theta. \tag{1.4}$$

The term "umbrella distribution" was coined by the physicists Torrie and Valleau (1977) to describe an artificial sampling distribution that simultaneously "covered" a large range of physical distributions. The general idea of reweighting a single Monte Carlo run to estimate quantities from a family of distributions goes back to Trotter and Tukey (1956). More recently, Ferrenberg and Swendsen (1988) [see also Swendsen and Ferrenberg (1990)] have popularized this method in the statistical physics community.

The principles of importance sampling continue to hold if the X_i^ν's are simulated from an ergodic Markov chain with stationary distribution ν. This basic fact is sometimes ignored in the statistical literature, where "importance sampling" usually refers only to i.i.d. simulations [Evans and Swartz (1995)]. But limiting consideration to distributions which are accessible to i.i.d. simulations is generally too restrictive to take full advantage of the method. Frequently there is no reasonable way to generate i.i.d. variables from a very complicated high-dimensional distribution, but it could be easy to implement a Markov chain whose equilibrium distribution is the desired one. A quite general recipe for accomplishing this is the Metropolis–Hastings method, introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and later generalized by Hastings (1970). Indeed, this method has a certain optimality property which justifies its choice [Peskun (1973)]. Unfortunately, this method can be inefficient if the Markov chain converges to its equilibrium very slowly, as measured for example by a small "spectral gap" (see next section for a precise definition). This is often described as a "slowly mixing" chain [e.g., Sokal (1989), Sinclair (1993)].

A typical situation in which convergence is slow is one in which the target distributions have densities with two (or more) peaks and their corresponding Markov chains tend to stay in the neighborhood of a peak for a long time. In such a case, we hope that we can improve the spectral gap by applying the Metropolis method to a suitable artificial "flattened" simulation distribution satisfying a condition such as (1.4), which guarantees that all the peaks are covered. This is done and rigorously justified in the two examples that will be presented. With the help of the basic result established in Section 2 (Proposition 2.1), it is then possible to bound the performance of the importance sampling estimator.

The two examples presented here concern two different choices of an importance sampling distribution. The first is taken from statistical physics, where we often consider a family of probability measures μ_θ having a density p_θ (with respect to some reference measure m) belonging to a one-parameter exponential family

$$p_\theta(x) = \frac{\exp(-\theta H(x))}{Z(\theta)}. \tag{1.5}$$

Physically, $\theta \geq 0$ is the reciprocal of the temperature and H represents the energy function of the system: then (1.5) is called a Gibbs distribution. Of course the density decreases as the energy H increases (unless $\theta = 0$), and this is emphasized when θ is large (i.e., when the temperature is low). In numerical experiments one is typically interested in plotting the derivatives of the free energy $\log Z(\theta)$ as a function of θ, but since these are the cumulants of H with respect to μ_θ (up to a change of sign for those of odd order), this problem is precisely of the kind mentioned at the beginning of this section. (Here we use H to refer the random variable $H(X)$ where X has the indicated distribution.) For example, physicists are often interested in values of θ where the average energy per unit volume decreases suddenly, exhibiting either a discontinuity or a slope of $-\infty$ as the size of the system grows. Alternatively one can look for values of θ where the variance of H under μ_θ, divided by the size of the system, blows up. Values of θ where such things happen are associated with phase transitions in the system [see, e.g., Thompson (1972) for an overview].

In these problems it is convenient to choose ν to be absolutely continuous with respect to m with a density which is only a function of H; in this way $d\mu_\theta/d\nu$ will also be a function of H only, and the importance sampling calculations will require only the evaluation of the energy of each sample. In many applications it is easy to compute the change in energy at each step, and so the calculation of the importance sampling weights is fast. The physicists' suggestion [Torrie and Valleau (1977), Berg and Neuhaus (1991)] is to make the distribution of H under ν uniform over a range of energy levels (which is large enough to include all the energies which are typical for the values of θ we are interested in). The physical intuition is to remove the energy barriers between states, allowing the Markov chain to explore the state space unencumbered by physically natural constraints.

In Section 3 we will show how this recipe works for the mean field Ising model on N sites. This is not a physically realistic model, but it has the qualitative features of more realistic and more complex models. We will prove that by using such a sampling distribution which is uniform over energies, in conjunction with single-site Metropolis updates, one can get good estimates of $E_\theta f$, for any inverse temperature $\theta > 0$, in a time which is polynomial in N. This is in contrast to the time needed for the crude "physical" Monte Carlo sampler which is exponential in N for $\theta > 1$ (at least for generic choices of f). Computational experience strongly suggests that similar results hold for a wide variety of physical systems that undergo phase transitions, even if it remains an open (and difficult) problem to prove this rigorously for more interesting physical models.

Before presenting the second example, in Section 4 we make a digression concerning the recently proposed remedy to the problem of slowly mixing Markov chains known as "simulated tempering" [Marinari and Parisi (1992), Geyer and Thompson (1995)]. The premise is that there are some values of θ for which a Metropolis Markov chain with stationary distribution μ_θ is slowly mixing and others at which it is rapidly mixing. This often happens in models

from statistical physics: at "high temperatures," the model has weak correlations and relaxes rapidly to equilibrium; but below some "critical temperature" there is a qualitative change in behavior, and the system becomes much harder to change, resulting in slowly mixing chains. In this case one can create a new Markov chain on the augmented state space $\Omega \times \Theta$. The chain will alternate between changing the configuration as if θ were fixed and changing the value of θ. Thus some of the time the chain will be free to mix rapidly (when θ is in the "high temperature" range), but it will also spend some time sampling parts of Ω that are typical of "low temperatures." The hope is that the augmented chain will itself be rapidly mixing. Alternatively simulated tempering can be seen as a way of accelerating convergence when there is a single distribution μ^* of interest, but the Markov chain at our disposal to sample from it is slowly mixing. In such a case, one can try to build a family of distributions μ_θ, say with $\theta \in [0, 1]$, such that $\mu_0 = \mu^*$ and the Markov chain for μ_1 is rapidly mixing. Geyer and Thompson (1995) interpret simulated tempering as a "pseudo-Bayes" approach and give several useful guidelines for its implementation in statistical problems.

Little is known rigorously about the properties of simulated tempering. The main purpose of Section 4 is to show that a natural implementation of simulated tempering is essentially equivalent to importance sampling with respect to a mixture of the μ_θ's (Proposition 4.1). It is not hard to show that a mixture with weights that are not too small will yield a bound of the form (1.4) with a constant A that is not too large (see Section 2).

In Section 5 the performance of this importance sampling technique is considered with reference to an example from Geyer and Thompson (1995) of a distribution over a high-dimensional state space for which the conventional impementation of Markov chain algorithms is slowly mixing. Our rigorous analysis explains the numerical results presented in their paper, which support the performance of simulated tempering. In the Appendix the same kind of results are proved for the mean field Ising model example as well.

2. Importance sampling. In this section we review the basic issues concerning importance sampling techniques in the dynamic sampling context and prove a simple but general result which will turn out to be extremely useful in the rest of the paper.

Suppose that for a candidate sampling distribution ν we can construct (for example, by using the Metropolis method) a time-homogeneous Markov chain $\{X_i^\nu, i = 1, 2 \ldots\}$ whose invariant distribution is ν. For the sake of simplicity, denote by $\hat{f}_n^\nu$ the estimator $\widehat{(f\, d\mu/d\nu)}(\mathbf{X}_n^\nu)$ of $E^\mu f$ defined in (1.2). Suppose that the chain is reversible, so that its Markov operator is self-adjoint on $L^2(\nu)$, and that there is a gap $\Delta(X^\nu)$ between its largest eigenvalue 1 and the rest of its spectrum. Then, provided $f\, d\mu/d\nu \in L^2(\nu)$, the central limit theorem for reversible Markov chains [Kipnis and Varadhan (1986)] ensures that, for any initial state,

$$\sqrt{n}\big(\hat{f}_n^\nu - E^\mu f\big) \to N\big(0, v^\nu(f\, d\mu/d\nu)\big), \tag{2.1}$$

where the asymptotic variance v^ν satisfies

$$v^\nu(h) \le \left(\frac{2}{\Delta(X^\nu)}\right)\sigma_\nu^2(h) \quad \text{for all } h \in L^2(\nu), \tag{2.2}$$

$\sigma_\nu^2(h)$ being the variance of h according to ν. For a discussion of more general assumptions which ensure that a central limit theorem holds for general Markov chains, see Meyn and Tweedie (1993) and Chan and Geyer (1994).

In practice, the estimator $\hat{f}_n^\nu$ cannot be computed whenever μ or ν, or both, is known only up to a normalization factor. In this case the estimator $\hat{f}_n^\nu$ is divided by the sample average of the likelihood weights $\hat{1}_n^\nu$, so that the normalization constants cancel and do not appear in the calculation. This quotient will not be unbiased but will stay strongly consistent. Moreover by using the above central limit theorem and the delta method [Ferguson (1996)], it is obtained that if $f(d\mu/d\nu)$ and $d\mu/d\nu$ are both in $L^2(\nu)$, then

$$\sqrt{n}\left(\frac{\hat{f}_n^\nu}{\hat{1}_n^\nu} - E^\mu f\right) \to N(0, s_\mu^\nu(f)), \tag{2.3}$$

where

$$s_\mu^\nu(f) := v^\nu\left((f - E^\mu f)d\mu/d\nu\right). \tag{2.4}$$

The next result clarifies how a simulation distribution ν which appropriately "covers" a family $\{\mu_\theta, \theta \in \Theta\}$ could provide a reduction of the variance with respect to more "physical" sampling schemes from each member of the family.

PROPOSITION 2.1. *Suppose that A is a constant such that*

$$\frac{d\mu_\theta}{d\nu}(x) \le A \quad \textit{for all } x \in \Omega \textit{ and } \theta \in \Theta \tag{2.5}$$

and that the samples are obtained from a reversible Markov chain $\{X_n^\nu\}$ having the stationary distribution ν with a spectral gap $\Delta(X^\nu)$. Then

$$s_{\mu_\theta}^\nu(f) \le \left(\frac{2A}{\Delta(X^\nu)}\right)\sigma_{\mu_\theta}^2(f) \tag{2.6}$$

for every $f \in L^2(\mu_\theta)$ and every $\theta \in \Theta$.

REMARK. If $f \in L^2(\nu)$, then (2.5) implies that $f \in L^2(\mu_\theta)$ for every $\theta \in \Theta$.

PROOF OF PROPOSITION 2.1. Fix $\theta \in \Theta$ and $f \in L^2(\mu_\theta)$. The bound (2.5) implies that $d\mu_\theta/d\nu$ and $f\, d\mu_\theta/d\nu$ are both in $L^2(\nu)$, and that

$$\sigma_\nu^2\left((f - E^{\mu_\theta} f)\frac{d\mu_\theta}{d\nu}\right) \le A\sigma_{\mu_\theta}^2(f). \tag{2.7}$$

From the definition (2.4) and the inequality (2.2), $s_\theta^\nu(f)$ is bounded above by the left-hand side of (2.7) times $2/\Delta(X^\nu)$, from which (2.6) is immediately obtained. □

A simple application of this bound is the case of i.i.d. sampling from a uniformly weighted mixture

$$\nu = \frac{1}{D}(\mu_{\theta_1} + \cdot + \mu_{\theta_D}).$$

Clearly $d\mu_{\theta_i}/d\nu \le D$. Then (2.6) holds with $A = D$, and thus the D estimates for $E^{\mu_{\theta_i}} f$ obtained using nD i.i.d. samples from ν are at least as good as what one gets by estimating $E^{\mu_{\theta_i}} f$ using n i.i.d. samples from μ_{θ_i} independently for each $i = 1, \ldots, D$. In general, if the dynamic sampling scheme available for ν is not too bad and if the importance sampling weights are not too big, then importance sampling from ν via the ratio estimator $\hat{f}_n^\nu / \hat{1}_n^\nu$ is not much worse than independent sampling from the "physical distribution" μ_θ (and could in fact be much better than a slowly mixing Markov chain from it).

For another application, consider an exponential family of the type (1.4) when the reference measure m is the counting measure on a discrete set. Let $M(h)$ be the number of sample points x whose "energy" $H(x)$ is equal to h, and let L be the number of values in the range of H. Define the probability measure

$$\nu(\{x\}) = \frac{1}{L\, M(H(x))}.$$

That is, the probability is divided up equally among the L possible values of the energy, and at each such value this probability is further divided up equally among all x's that share this particular energy value. This ν is the distribution that is "uniform over energies" that was referred to in the Introduction. Since

$$M(H(x))\mu_\theta(\{x\}) = \mu_\theta\{H = H(x)\} \le 1$$

for every x, we see that (2.5) holds with $A = L$. The usual situation with discrete spin systems is that $|\Omega|$ grows exponentially while L grows only polynomially in the number N of sites.

In the following we will consider only reversible Markov chains. Our main interest will be chains constructed by the Metropolis method, which is briefly introduced below.

Suppose ν has a density p with respect to a reference measure m on $(\Omega, \mathscr{B})$ and let $Q(x, dy)$ be a transition kernel such that the measure $m(dx)Q(x, dy)$ is symmetric. Also let α be a measurable function from $\Omega \times \Omega$ to $[0, 1]$. We use these to construct the following randomized algorithm: the transition kernel Q "proposes" a move from x to y which is then "accepted" with probability $\alpha(x, y)$. This algorithm produces a Markov chain whose transition probability kernel is

$$P(x, dy) = \alpha(x, y)Q(x, dy) + w(x)\delta_x(dy), \tag{2.8}$$

where $w(x) = 1 - \int \alpha(x, z)Q(x, dz)$ is the probability of not accepting a proposal from x. By construction, P is reversible with respect to ν if and only if

$$p(x)\alpha(x, y) = p(y)\alpha(y, x) \quad \text{a.e. } [\nu(dx)Q(x, dy)] \tag{2.9}$$

[Tierney (1998)]. When this holds we define the Dirichlet form

$$E(g) = \tfrac{1}{2}\int_{\Omega\times\Omega} |g(x) - g(y)|^2 p(x)m(dx)P(x, dy) \tag{2.10}$$

for $g \in L^2(\nu)$. For kernels of the type (2.8), a popular choice for α is the Metropolis kernel

$$\alpha_{\text{Met}}(x, y) = \min\{p(y)/p(x), 1\}. \tag{2.11}$$

This is clearly maximal among all α's satisfying (2.9), and it is known to be optimal in the sense that it minimizes asymptotic variances and maximizes the spectral gap among all such α's. This optimality follows from the maximality of α_{Met} and the following result.

PROPOSITION 2.2. *Let P_A and P_B be the transition kernels of two Markov chains on Ω that are both reversible with respect to the same probability measure ν. Suppose also that $P_A(x, D \setminus \{x\}) \le P_B(x, D \setminus \{x\})$ for every $x \in \Omega$ and every measurable set D. Then*:

(i) *For every $f \in L^2(\nu)$, the asymptotic variance $v^\nu(f)$ is smaller for the P_B chain than for the P_A chain.*

(ii) *The spectral gap of the P_B chain is greater than that of the P_A chain.*

Proposition 2.2(i) was originally proven by Peskun (1973) for finite state spaces, and was extended to general state spaces by Caracciolo, Pelissetto and Sokal (1990), Appendix, as well as by Tierney (1998). Part (ii) is also in Caracciolo, Pelissetto and Sokal (1990), among other places, and is a consequence of the following property that we shall need later. For chains X^ν that are reversible with respect to ν, it is well known that the spectral gap $\Delta(X^\nu)$ is related to the Dirichlet form as follows:

$$\Delta(X^\nu) = \inf \frac{E(g)}{\sigma_\nu^2(g)}, \tag{2.12}$$

where the inf is over all nonconstant functions g in $L^2(\nu)$. This is discussed for example in Diaconis and Stroock (1991) for the case that Ω is finite; the general case is very similar and is treated in the Appendix of Caracciolo, Pelissetto and Sokal (1990).

For the remainder of this section we shall consider only Metropolis Markov chains, that is, chains of the form (2.8) with the choice (2.11) for α. For such chains, we can express the Dirichlet form (2.10) as

$$E(g) = \tfrac{1}{2}\int_{\Omega\times\Omega} |g(x) - g(y)|^2 m(dx)\min\{p(y), p(x)\}Q(x, dy). \tag{2.13}$$

It is useful to notice that the spectral gap for Metropolis chains does not depend too sensitively on ν, as the following proposition shows.

PROPOSITION 2.3. *Let $\nu_i(dy) = p_i(y)m(dy)$, $i = 1, 2$ be two probability distributions and let $\{X_n^{\nu_1}\}$ and $\{X_n^{\nu_2}\}$ be the corresponding stationary Metropolis*

Markov chains with kernels $P_i(x, dy)$ [*as defined in* (2.8) *and* (2.11)] *with the same proposal kernel* Q *and* $p = p_i$, *for* $i = 1, 2$. *Assume*

$$a \leq \frac{p_1(x)}{p_2(x)} \leq b \tag{2.14}$$

for all $x \in \Omega$ *such that* $p_1(x)$ *and* $p_2(x)$ *do not vanish simultaneously. Then*

$$ab^{-1}\Delta(X^{\nu_2}) \leq \Delta(X^{\nu_1}) \leq ba^{-1}\Delta(X^{\nu_2}). \tag{2.15}$$

PROOF. Notice that $L^2(\nu_1) = L^2(\nu_2)$ because of (2.14). Let E_i be the Dirichlet form for P_i. Then it is immediately seen from (2.13) that (2.14) implies

$$aE_2(g) \leq E_1(g) \leq bE_2(g) \quad \text{for } g \text{ in } L^2(\nu_1) = L^2(\nu_2).$$

The proposition is then a straightforward consequence of Lemma 3.3 in Diaconis and Saloff-Coste (1996). □

3. Example 1. The mean field Ising model. In this section we apply the results of Section 2 to the simplest example of a spin system which exhibits a phase transition and the corresponding slow mixing rate for local dynamics. Ising models are introduced below; for more details, the reader can consult Thompson (1972) or Ellis (1985). We will consider here a simulation distribution which is uniform over energy levels.

The general Ising model represents N particles (magnets) which can have either positive or negative spin $x_i \in \{-1, +1\}$, $i = 1, \ldots, N$. These particles sit on the nodes of a graph, and each particle interacts with all of its neighbors. In the *mean field Ising model*, the graph is the complete graph on N nodes (i.e., every pair of nodes are neighbors). Thus every particle interacts equally with every other particle. We will take N to be even, since this will simplify the notation in the sequel. The energy of the configuration $x = (x_1, \ldots, x_N) \in \Omega_N = \{-1, +1\}^N$ is given by

$$H_N(x) = -\sum_{(i,\, j)} x_i x_j - N/2 = -S_N^2(x)/2, \tag{3.1}$$

where the sum is over all pairs (i, j), and $S_N(x) = \sum_{i=1}^N x_i$ is the *total spin*.

The corresponding Gibbs distribution $\mu_{\beta,\, N}$ with inverse temperature $\beta > 0$ is defined by the probability masses

$$p_{\beta,\, N}(x) = \frac{\exp\bigl(-\beta N^{-1} H_N(x)\bigr)}{Z_N(\beta)}, \qquad x \in \Omega_N, \tag{3.2}$$

where $Z_N(\beta)$ is the appropriate normalizing constant ("partition function").

The total spin S_N assumes even values not larger than N in absolute value, with probabilities

$$\rho_{\beta,\, N}(j) := \mu_{\beta,\, N}\{S_N = j\} = \binom{N}{(N+j)/2} \frac{\exp\bigl(\beta j^2/(2N)\bigr)}{Z_N(\beta)}. \tag{3.3}$$

We are interested in the asymptotic behavior of $\rho_{\beta,N}(\lfloor sN \rfloor)$ as $N \to \infty$ for a fixed $s \in (-1, 1)$. By using Stirling's formula it is easily obtained that

$$\rho_{\beta,N}(\lfloor sN \rfloor) \sim \sqrt{2}\big(\pi(1-s^2)N\big)^{-1/2} \frac{\exp\big(Ni_\beta(s)\big)}{Z_N(\beta)}, \tag{3.4}$$

where

$$i_\beta(s) = \log 2 + 2^{-1}\big\{\beta s^2 - (1-s)\log(1-s) - (1+s)\log(1+s)\big\}. \tag{3.5}$$

Moreover, since the number of values for the total spin is $O(N)$, it is easily seen that

$$\lim_{N\to\infty} N^{-1} \log Z_N(\beta) = \max_{s\in[-1,1]} i_\beta(s). \tag{3.6}$$

We are thus led to the study of the function i_β. This is clearly symmetric around zero with $i_\beta(0) = \log 2$ and $i_\beta(-1) = i_\beta(1) = \beta/2$. Moreover the first and second derivatives are

$$i'_\beta(s) = \beta s - 2^{-1}\log(1+s)/(1-s),$$
$$i''_\beta(s) = \beta - [1-s^2]^{-1}.$$

Since $i''_\beta(s)$ is always negative for $\beta \le 1$ (except for $s = 0$ when $\beta = 1$), the function i_β is strictly concave and uniquely maximized at 0. This is not true anymore for $\beta > 1$. In fact by equating $i'_\beta(a)$ to zero, it is obtained that

$$2\beta s = \log \frac{1+s}{1-s},$$

that is, $\tanh \beta s = s$. This equation always has 0 as a solution, but for $\beta > 1$ two other solutions appear, which we call $s^{[\beta]}$ and $-s^{[\beta]}$. It is easily seen that these are maximum points for i_β when $\beta > 1$, whereas 0 becomes a local minimum point. This is an example of a phase transition with critical value $\beta = 1$.

In particular, for $\beta > 1$ and $s = 0$ it is obtained from (3.4) and (3.6) that balanced configurations became quite rare, that is,

$$\lim_{N\to\infty} N^{-1} \log \rho_{\beta,N}(0) = i_\beta(0) - i_\beta\big(s^{[\beta]}\big), \tag{3.7}$$

where the right-hand side is strictly negative. For an explicit computation see Thompson (1972).

Next consider the Metropolis dynamics associated to $\mu_{\beta,N}$ where the underlying symmetric transition matrix is the single spin flip: $Q(x, y) = N^{-1}$ if x and y differ for a single spin and $= 0$ otherwise. First, we show that for $\beta > 1$ this chain is slowly mixing, since its spectral gap can be bounded from above by a function which decreases exponentially in N. To do this, it is enough to show the same property for the conductance [Sinclair (1993)]

$$C_{\beta,N} = \min_{A:\mu_{\beta,N}(A)\le 1/2} \frac{\sum_{x\in A}\sum_{y\in A^c} p_{\beta,N}(x)P_{\beta,N}(x, y)}{\mu_{\beta,N}(A)}. \tag{3.8}$$

In fact, consider the set $A = A_N$ of all configurations that have positive total spin. It is clear that A has probability tending to $1/2$ from below, because of (3.7). Moreover,

$$\sum_{x \in A} \sum_{\sigma \in A^c} p_{\beta, N}(x) P_{\beta, N}(x, \sigma) = 2^{-1} \mu_{\beta, N}\{S_N = 0\} = 2^{-1} \rho_{\beta, N}(0). \tag{3.9}$$

Since $\rho_{\beta, N}(0)$ decays exponentially by (3.7), this establishes the promised result. In fact, this kind of argument applies to more general spin systems. On the other hand, by using the Dobrushin criterion described below, it is easy to prove that for $\beta < 1$ the spectral gap decays like $O(N^{-1})$.

From the above result and the discussion of Section 2, we conclude that for any specified $\beta > 1$ there exists a function f_N with $\sigma^2_{\beta, N}(f_N) = 1$ whose asymptotic variance $v^{\mu_{\beta, N}}(f_N)$ (corresponding to the Metropolis' sampling scheme described above) grows (at least) exponentially in N. We now proceed to exhibit a simulation distribution such that the asymptotic variance of any such function cannot grow more than polynomially in N.

First we shall set some notation. Suppose that ν is a probability measure on Ω_N. For any $x = (x_1, \ldots, x_N) \in \Omega_N$ and any set $A \subset \{1, \ldots, N\}$, let x^A be the configuration obtained by flipping the spins of x that lie in A (i.e., $x_i \neq x_i^A$ if and only if $i \in A$). For $i = 1, \ldots, N$, let $P^{(i)}(x, \cdot)$ be the heat bath (Gibbs sampler) kernel on $\{-1, +1\}$ for the update of the ith spin

$$P^{(i)}(x, y) = \frac{\nu(y)}{\nu(x) + \nu(x^i)} \prod_{j \neq i} \delta_{x_j, y_j}.$$

Now let $X^{\nu, HB}$ be the Markov chain obtained by applying the heat bath to a randomly chosen site; in other words, it is the chain whose transition probability kernel is

$$P_N(x, dy) = \frac{1}{N} \sum_{i=1}^{N} P^{(i)}(x, dy). \tag{3.10}$$

Let $X^{\nu, \mathrm{Met}}$ be the Markov chain corresponding to the Metropolis scheme whose proposals are randomly chosen single-site flips and equilibrium measure ν.

Now we let ν_N be the distribution "uniform over energy levels" described in Section 2. By (3.1), values of H_N correspond to values of S_N, except for ambiguity of sign; so for simplicity we shall modify the construction of Section 2 slightly. Let $M(h)$ be the number of configurations whose total spin S_N is equal to h and define the masses

$$\nu_N(x) = \{L_N M(S_N(x))\}^{-1},$$

where $L_N = N + 1$ is the number of values assumed by S_N. As in Section 2, we get

$$\frac{d\mu_{\beta, N}}{d\nu}(x) \leq N + 1 \quad \text{for every } \beta \geq 0 \text{ and } x \in \Omega_N. \tag{3.11}$$

Moreover, $M(s) = \binom{N}{(N+s)/2}$. Then the transition probability of the heat bath chain $X^{\nu_N, \mathrm{HB}}$ is

$$P_N(x, x^i) = N^{-1}\left(\frac{S_N(x) + N + 2}{2(N+1)}\delta_{-1, x_i} + \frac{N - S_N(x) + 2}{2(N+1)}\delta_{1, x_i}\right).$$

We are now ready to prove the promised bound.

THEOREM 3.1. *Let ν_N be the probability measure which is uniform over energy levels for the N-site mean field Ising model. Then*

$$\Delta\big(X^{\nu_N, \mathrm{Met}}\big) \geq \Delta\big(X^{\nu_N, \mathrm{HB}}\big) \geq \frac{2}{N(N+1)}. \tag{3.12}$$

Therefore, for either the heat bath or Metropolis chain, the asymptotic variance of the importance sampling estimator for $E^{\mu_{\beta, N}} f$ satisfies

$$s^{\nu_N}_{\mu_{\beta, N}}(f) \leq (N+1)^3 \sigma^2_{\mu_{\beta, N}}(f) \quad \textit{for every } \beta \geq 0 \textit{ and every } f. \tag{3.13}$$

PROOF. The key to the proof is the classical Dobrushin's criterion, which gives a bound for a heat bath kernel P_N with respect to some measure ν on Ω_N. Namely, let $c_N = \sup_i \sum_{j \neq i} c_{ij}$, where

$$c_{ij} = \sup_x \big| P_N(x, x^i) - P_N(x^j, x^{i,j}) \big|;$$

then

$$\Delta(X^{\nu_N, HB}) \geq \frac{1 - c_N}{N} \tag{3.14}$$

(see Lemma A.1 in the Appendix).

In the case of the heat bath chain, it is easy to compute that the coefficient c_N satisfies

$$c_N \leq (N-1) \sup_{x, i, j} \big| P_N(x, x^i) - P_N(x^j, x^{i,j}) \big| = \frac{N-1}{N+1}. \tag{3.15}$$

Inserting the right-hand side of (3.15) back into (3.14), we obtain the right-hand inequality of (3.12). The left-hand inequality of (3.12) is a consequence of Proposition 2.2(ii), as was first observed by Peskun (1973). Now (3.13) follows from (3.11), (3.12) and Proposition 2.1. □

We finally observe that an analogous theorem holds for the case where ν is a finite mixture of fixed-temperature distributions $\mu_{\beta, N}$. The proof of this result is more technical, so it is left to the Appendix.

THEOREM 3.2. *Let N be even. There exist positive numbers $\tilde{\beta}^N_k$ $(k = 0, \ldots, N/2)$ such that the mixture*

$$\mu_{F, N} = \left(\frac{N}{2} + 1\right)^{-1} \sum_{k=0}^{N/2} \mu_{\tilde{\beta}^N_k, N} \tag{3.16}$$

satisfies

$$\frac{d\mu_{\beta,N}}{d\mu_{F,N}} \leq \tfrac{1}{2}(N+2)^2 \quad \textit{for every } \beta \geq 0 \tag{3.17}$$

and

$$\Delta(X^{\mu_{F,N},\,\mathrm{Met}}) \geq 4(N+2)^{-4}. \tag{3.18}$$

Therefore, for the Metropolis chain $X^{\mu_{F,N},\,\mathrm{Met}}$, *the asymptotic variance of the importance sampling estimator for* $E^{\mu_{\beta,N}} f$ *satisfies*

$$s^{\mu_{F,N}}_{\mu_{\beta,N}}(f) \leq \tfrac{1}{4}(N+2)^6 \sigma^2_{\mu_{\beta,N}}(f) \quad \textit{for every } \beta \geq 0 \textit{ and every } f.$$

4. Simulated tempering and sampling from mixtures. Consider again the problem of the computation of the expectation of some function f with respect to D probability distributions μ_{θ_i} where $\mu_{\theta_i}(dx) = p_i(x)m(dx)$, for $i = 1, \ldots, D$. From some symmetric transition kernel $Q(x, dy)$ with respect to m we can construct Metropolis' algorithms for each of the μ_{θ_i}'s, some of which could be slowly mixing. We have already noticed that in order to cover all the important parts of the sample space it is quite natural to take simulations from a mixture $\nu = \sum_{i=1}^{D} a_i \mu_{\theta_i}$, for some roughly uniform choice of the weights a_i. A sampling process from ν can be obtained by the Metropolis kernel P^{mix} defined by (2.8) and (2.11) with density

$$p(x) \equiv p^{\mathrm{mix}}(x) = \sum_{i=1}^{d} a_i p_i(x) \tag{4.1}$$

with respect to m. An alternative to importance sampling is offered by simulated tempering, a technique recently introduced by Marinari and Parisi (1992). The idea behind simulated tempering is to augment the sample space Ω by including a label variable taking values in the set of labels $\{1, \ldots, D\}$ of the physical distributions μ_{θ_i} of interest. The overall probability is given by its density ϕ with respect to the product of m and the counting measure, which is taken to be

$$\phi(i, x) = a_i p_i(x). \tag{4.2}$$

It is then obvious that the a_i's represent exactly the probabilities of the various labels and that the marginal distribution of the configuration variable x is exactly the mixture p^{mix} defined in (4.1). Next a Markov process on $\Omega \times \{1, \ldots, D\}$ for sampling from ϕ is obtained in the following way. Given the configuration $x \in \Omega$ at the kth stage of the process, a label is selected by giving to $i = 1, \ldots, D$ the probabilities

$$\frac{a_i p_i(x)}{\sum_{k=1}^{D} a_k p_k(x)} = \phi(i \mid x), \tag{4.3}$$

which are the conditional probabilities of the various labels given the configuration x, according to the joint distribution ϕ. Then a step of the Metropolis

method is performed from the transition kernel P_i corresponding to the selected value of i, producing the $(k+1)$th configuration.

It is quite clear that the choice (4.3) preserves the marginal distribution of the labels and since the Metropolis step preserves the distribution over the configurations conditional to label i (which is clearly μ_{θ_i} itself), the joint distribution ϕ is preserved. Of course, (4.3) is not the only possible choice; but it simplifies our analysis since the label selected at the kth step is not taken into account for producing either the label or the configuration at step $(k+1)$. By consequence *the sequence of configurations produced by the sampling process is by itself a Markov process*, with transition probabilities given by

$$P^{\mathrm{st}}(x, dy) = p^{\mathrm{st}}(x, y)Q(x, dy) + \delta_x(dy)w^{\mathrm{st}}(x), \tag{4.4a}$$

where

$$p^{\mathrm{st}}(x, y) = \frac{\sum_{k=1}^{D} a_k p_k(x) \min\{p_k(y)/p_k(x), 1\}}{p(x)} \tag{4.4b}$$

and where w^{st} is obviously defined in order to make $P^{\mathrm{st}}(x, \cdot)$ a probability. Here and in the following we will use $p^*(x, y)$ to denote the density of a transition probability kernel $P^*(x, \cdot)$ with respect to $Q(x, \cdot)$.

PROPOSITION 4.1. (i) *For any $x \neq y$, the transition densities with respect to $Q(x, \cdot)$ are related by*

$$p^{\mathrm{mix}}(x, y) \geq p^{\mathrm{st}}(x, y), \tag{4.5}$$

from which the corresponding asymptotic variance forms satisfy

$$v^{\mathrm{st}}(h) \geq v^{\mathrm{mix}}(h) \qquad \forall\, h \in L^2(\nu). \tag{4.6}$$

(ii) *Suppose that for all pairs of configurations x and y and labels i and j,*

$$(p_i(x) - p_i(y))(p_j(x) - p_j(y)) \geq 0. \tag{4.7}$$

Then $P^{\mathrm{mix}} = P^{\mathrm{st}}$.

PROOF. It suffices to notice that, from its definition (4.4a), for $x \neq y$ the transition density with respect to $Q(x, dy)$ satisfies

$$p^{\mathrm{st}}(x, y) \leq \frac{\min\{\sum_{k=1}^{D} a_k p_k(y), \sum_{k=1}^{D} a_k p_k(x)\}}{p(x)} = \min\left\{1, \frac{p(y)}{p(x)}\right\},$$

where the inequality is obtained by shifting the minimum out of the sum. From the above, (4.5) is obtained. Under the additional assumption of (ii), it is immediately obtained that

$$\frac{p(y)}{p(x)} \geq 1 \quad \text{if and only if} \quad \frac{p_i(y)}{p_i(x)} \geq 1 \quad \text{for every } i = 1, \ldots, D,$$

meaning that the two algorithms will accept a move proposed by the transition Q only simultaneously. Thus the equality of the two kernels holds. The assertion (4.6) follows from Proposition 2.2(i). □

The condition of Proposition 4.1(ii), under which simulated tempering becomes equivalent to a direct Metropolis algorithm for the desired mixture, is of course fulfilled whenever each p_i is of the form p_{θ_i} as given by (1.5) and the θ_i's all have the same sign.

Proposition 4.1 is formally restricted in scope, yet it is highly suggestive. On the one hand, it is somewhat restricted because it only addresses a specific implementation of simulated tempering. For example, Geyer and Thompson (1995) updated the labels by Metropolis instead of heat bath; and Marinari and Parisi (1992) did not use Metropolis to update the configurations for fixed θ.

On the other hand, the proposition is suggestive because it says that if we are considering using such an implementation of simulated tempering, then we cannot do worse by using importance sampling (via Metropolis) with a distribution of the form $\sum_i a_i p_{\theta_i}$, and perhaps we can do much better with an importance sampling distribution of some other form (e.g., the distribution that is "uniform over energies" introduced before). In practice, even when a theoretical analysis is difficult, distributions of this kind are something to aim for.

It may also seem that the above proposition proves that the specified implementation of simulated tempering can have no advantage as a sampling process. This is not necessarily true, because it must be taken into account that in order to fit our analysis into the scheme discussed before, we are not allowed to use the samples from the label process in the estimator. But in principle these can carry additional information. In order to appreciate this point, let us imagine that we need to compute $E^{\mu_{\theta_1}} f$. If (X_i^ν, I_i) is the output of the simulated tempering process, then our recipe to estimate $E^{\mu_{\theta_1}} f$ is to use $\hat{f}_n = \hat{f}_n^\nu / \hat{1}_n^\nu$, where

$$\hat{f}_n^\nu = n^{-1} \sum_{i=1}^{n} f(X_i^\nu) \frac{p_1(X_i^\nu)}{\sum_{k=1}^{D} a_k p_k(X_i^\nu)}. \tag{4.8}$$

However, there is another natural asymptotically unbiased estimator which makes use of the label process, which is $\tilde{f}_n = \tilde{f}_n^1 / \tilde{1}_n^1$, where

$$\tilde{f}_n^1 = n^{-1} \sum_{i=1}^{n} f(X_i^\nu) a_1^{-1} \delta_{I_i, 1}, \tag{4.9}$$

which is the average of f over the samples labeled 1. Notice by a direct inspection that (4.8) is obtained by taking the conditional expectation of each summand in (4.9) with respect to the current configuration X_i^ν, $i = 1, \ldots, n$. This Rao–Blackwellization technique [Arnold (1993)] certainly reduces the variance in the i.i.d. case, since then

$$\hat{f}_n^\nu = E\big(\tilde{f}_n^1 \mid X_i^\nu, i = 1, \ldots, n\big). \tag{4.10}$$

But this is not true anymore when the sampling scheme for the configurations is Markovian [for a particular situation in which this continues to hold, see Liu, Wong and Kong (1994)].

Another open question concerns the choice of the weights a_i, $i = 1, \ldots, D$. The choice of equal weights is not necessarily optimal: in fact, some of the θ_i's could have been added only in order to speed up the simulated tempering process. On the other hand a direct minimization of the asymptotic variance of (4.8) with respect to the weights seems completely unfeasible. Even if the optimal weights were known, the fact that the densities p_{θ_i} are known only up to normalization constants makes the implementation of both simulated tempering and direct Metropolis from the mixture technically impossible. It is easy to see that at least the ratios between these normalization constants need to be known. But in many applications, the problem of interest is precisely the evaluation of these ratios. This is true in statistical physics as well as in statistics [e.g., Bennett (1976), Meng and Wong (1996)]. It is however easy to get first estimates for these quantities through a preliminary sampling process. More precisely, observe that if $p_k(x) = h_k(x)/Z_k$, for $k = 1, \ldots, D$, the Z_k's being unknown, we are forced to choose $a_k = P(I_1 = k)$ to be proportional to $c_k Z_k$ for some choice of the positive constants c_k, $k = 1, \ldots, D$. This means that strongly consistent estimators of Z_k/Z_j are given by

$$\frac{\sum_{i=1}^{n}\left(h_k(X_i)/\sum_{l=1}^{D} c_l h_l(X_i)\right)}{\sum_{i=1}^{n}\left(h_j(X_i)/\sum_{l=1}^{D} c_l h_l(X_i)\right)} \tag{4.11}$$

if Metropolis sampling from the marginal mixture distribution P^{mix} is used, or

$$\frac{c_j \sum_{i=1}^{n} \delta_{I_i, k}}{c_k \sum_{i=1}^{n} \delta_{I_i, j}} \tag{4.12}$$

if simulated tempering P^{st} is used.

By Proposition 2.3 it is not so important for these estimates to be very precise. A more elegant approach would be to estimate such constants on the same sample we are using for the estimation of the desired expected values. The theory of recursive stochastic algorithms [Duflo (1996)] seems an indispensable tool for this purpose.

Likewise, if one is working with importance sampling from a general distribution, one can start from an initial guess and then adjust the distribution itself during preliminary runs so as to get some desired property, such as a uniform histogram of energies [see Valleau (1991) and Janse van Rensburg and Madras (1997)].

Finally, we remark that there are closely related methods which do not require knowledge of the normalizing constants of the p_{θ_i}'s. Geyer (1991) proposed running a chain at each θ_i and trying to swap states of different chains. Another approach is Neal (1996).

5. Example 2. The witch's hat. In this section we examine the performance of mixture sampling for the "witch's hat" distribution, following the implementation of Geyer and Thompson (1995).

The target density is a mixture of the uniform distribution on the unit d-dimensional hypercube $[0, 1]^d$ and the uniform distribution on a smaller

hypercube $[0, \alpha_0]^d$, with weights such that this latter hypercube has probability α_0. The probability α_0 of this subcube is much larger than its volume α_0^d, and this slows down the convergence of the Gibbs' sampler and similar single-coordinate Markov chain updating schemes. For example, for the case considered by Geyer and Thompson ($\alpha_0 = 1/3$ and $d = 30$), the mixing time of the Metropolis algorithm (with uniform choice of the variable to be updated and its value) is at least $3^{29} \simeq 7 \times 10^{12}$ (see next paragraph).

For α in $(0, 1]$, let $u(\cdot; \alpha)$ be the uniform density (with respect to Lebesgue measure) on $[0, \alpha]^d$ and let

$$r(\cdot; \alpha) = \frac{\alpha - \alpha^d}{1 - \alpha^d} u(\cdot; \alpha) + \frac{1 - \alpha}{1 - \alpha^d} u(\cdot; 1) \tag{5.1}$$

be the family of possible target or interpolating densities. Notice that $[0, \alpha]^d$ has probability α under $r(\cdot; \alpha)$. Our choice for the proposal kernel in the Metropolis algorithm is to let $Q(x, dy)$ be the uniform distribution on the points of the hypercube $[0, 1]^d$ that differ from x in exactly one coordinate. We now show that if $\alpha < 1$ and d is large, then the Metropolis algorithm for $r(\cdot; \alpha)$ is slowly mixing. By (2.12), the spectral gap is the infimum of

$$\frac{\int\int \big(f(x) - f(y)\big)^2 \min\big\{r(x; \alpha), r(y; \alpha)\big\}\, Q(x, dy) m(dx)}{2\sigma^2_{r(\cdot;\alpha)}(f)} \tag{5.2}$$

over all nonconstant square integrable functions on $[0, 1]^d$ [recall that $\sigma_q^2(f)$ is the variance of $f(X)$ when X is distributed according to the density q]. Taking f to be the indicator function of $[0, \alpha]^d$, we see that $\sigma^2_{r(\cdot;\alpha)}(f) = \alpha(1 - \alpha)$, and the numerator of (5.2) equals

$$\int_{x \in [0, \alpha]^d} \int_{y \notin [0, \alpha]^d} \frac{1 - \alpha}{1 - \alpha^d} Q(x, dy) m(dx) = \frac{\alpha^d (1 - \alpha)^2}{1 - \alpha^d} \le \alpha^d (1 - \alpha).$$

Therefore the spectral gap is bounded above by $2\alpha^{d-1}$. So, for fixed α, the Metropolis algorithm applied directly to $r(\cdot; \alpha)$ is exponentially slow in d.

Next choose an integer $D > 1$, and define the constants $\alpha_i = \alpha_0^{1 - i/D}$, for $i = 0, \ldots, D$, and the mixture density

$$p_{\text{mix}}(x) = \frac{1}{D + 1} \sum_{i=0}^{D} r(x; \alpha_i). \tag{5.3}$$

As viewed by Geyer and Thompson (1995), $r(\cdot; \alpha_0)$ is the target density, $r(\cdot; \alpha_D) = r(\cdot; 1)$ is the "hot" distribution (uniform on $[0, 1]^d$ in our case, which permits rapid mixing) and the $r(\cdot; \alpha_i)$'s are the interpolating densities. The crucial step for our analysis is to rewrite p_{mix} as a new mixture

$$p_{\text{mix}}(x) = \sum_{i=0}^{D} b_i p_i(x), \tag{5.4}$$

with $p_i(x) = u(x; \alpha_i)$ $(i = 0, \ldots, D)$,

$$b_i = \frac{\alpha_i - \alpha_i^d}{(D+1)(1-\alpha^d)}, \qquad i = 0, \ldots, D-1$$

and $b_D = 1 - \sum_{j=0}^{D-1} b_i$.

We shall write Δ_{mix} for the spectral gap for the Metropolis algorithm for p_{mix}, with proposal kernel Q. To bound Δ_{mix} we use the following result, due to Madras and Randall (1999).

PROPOSITION 5.1. *Let $p_0, \ldots, p_D$ be probability densities (with respect to a common reference measure m) and let $b_0, \ldots, b_D$ be positive numbers that add up to 1. Define the mixture density*

$$p_{\text{mix}} = \sum_{j=0}^{D} b_j p_j. \tag{5.5}$$

Let $Q(x, dy)$ be a proposal kernel symmetric with respect to m. Let Δ_j (respectively, Δ_{mix}) be the spectral gap of the Metropolis chain for p_j (respectively, p_{mix}) whose proposal chain is Q. Finally, assume that neighboring p_j's have some "overlap": that is, assume

$$\int \min\{p_j(x), p_{j+1}(x)\}\, dx \geq \delta, \qquad j = 0, \ldots, D-1 \tag{5.6}$$

for some $\delta > 0$. Then

$$\Delta_{\text{mix}} \geq \frac{\delta}{2D} \min_{j=0,\ldots,D} b_j \Delta_j. \tag{5.7}$$

In order to apply Proposition 5.1 to our case, observe first that

$$\frac{1}{D+1} \geq b_i \geq \frac{\alpha_i}{(D+1)(1+\alpha_i)} \geq \frac{\alpha_0}{(D+1)(1+\alpha_0)}, \qquad i = 0, \ldots, D-1 \tag{5.8}$$

and $b_D \geq 1/(D+1)$ [by the leftmost inequality in (5.8)]. Therefore

$$\min\{b_0, \ldots, b_D\} \geq \frac{\alpha_0}{(1+\alpha_0)(D+1)}. \tag{5.9}$$

We now check the "overlap" condition: for $j = 0, \ldots, D-1$,

$$\begin{aligned} \int \min\{p_j(x), p_{j+1}(x)\}\, dx &= \int_{[0,\alpha_j]^d} p_{j+1}(x)\, dx \\ &= \frac{\alpha_j^d}{\alpha_{j+1}^d} \\ &= \alpha_0^{d/D}. \end{aligned} \tag{5.10}$$

Now we must compute the spectral gap Δ_i for the Metropolis algorithm sampling from p_i with the proposal kernel Q above. We claim that

$$\Delta_i = \alpha_i \Delta_D = \frac{\alpha_i}{d}, \qquad i = 0, \ldots, D. \tag{5.11}$$

The first equality of (5.11) is a consequence of $p_i(x) = p_D(x/\alpha_i)$. For the second equality, we shall now prove that $\Delta_D = 1/d$.

Let us write the proposal operator as $Q = (\sum_{k=1}^{d} R_k)/d$, where for each $k = 1, \ldots, d$, we define the operator R_k on $L^2([0,1]^d)$ by

$$R_k f(a_1, \ldots, a_d) = \int_0^1 f(a_1, \ldots, a_{k-1}, x_k, a_{k+1}, \ldots, a_d)\, dx_k.$$

By definition, Δ_D is the spectral gap of the operator Q on $L^2([0,1]^d)$. Let V_k be the set of functions in $L^2([0,1]^d)$ with the property that $f(x_1, \ldots, x_d)$ does not depend on x_k. It is easy to see that f is in V_k if and only if $R_k f = f$; in fact, R_k is the operator of orthogonal projection onto V_k. Next, for each $S \subset \{1, \ldots, d\}$, define

$$U_S = \Big(\bigcap_{i \in S} V_i\Big) \cap \Big(\bigcap_{i \notin S} V_i^{\perp}\Big)$$

In particular $U_{\{1, \ldots, d\}}$ is the set of constant functions. Notice that if $f \in U_S$, then $R_i f = f$ for every $i \in S$, and $R_i f = 0$ for every other i; hence $Qf = (|S|/d)f$. Since the U_S's are mutually orthogonal subspaces whose direct sum is all of L^2, we have completely determined the spectrum of Q, and it follows that $\Delta_D = 1/d$.

Inserting (5.9), (5.10) and (5.11) into the bound of Proposition 5.1, we get that the spectral gap Δ_{mix} for the Metropolis sampler from p_{mix} has a lower bound

$$\Delta_{\text{mix}} \geq \frac{\alpha_0^{(d/D)+2}}{2dD(D+1)(1+\alpha_0)}.$$

The last step is to apply Proposition 2.1, plugging Δ_{mix} into the estimate (2.6) with $A = D + 1$, this time using the representation of p_{mix} as a mixture with uniform weights. The result is summarized in the following theorem.

THEOREM 5.1. *Let p_{mix} be the probability distribution defined by* (5.3), *and let f be any function in $L^2([0,1]^d)$. Then the asymptotic variance of the Metropolis importance sampling estimator for $E^{r(\cdot;\alpha_0)} f$ satisfies*

$$s^{p_{\text{mix}}}_{r(\cdot;\alpha_0)}(f) \leq \frac{4dD(D+1)^2(1+\alpha_0)}{\alpha_0^{(d/D)+2}} \sigma^2_{r(\cdot;\alpha_0)}(f).$$

Thus, if the number of distributions D is of the order of d, then we need $O(dD^2)$ steps of our chain [i.e., $O(D^2)$ sweeps through the d coordinates] to get an "independent observation" from p_{mix}; therefore we need $O(dD^3)$ steps (or $O(D^3)$ sweeps) to get an "independent observation" from the target distribution $r(\cdot;\alpha_0)$.

APPENDIX

In this Appendix, we give proofs of two technical results. The first of these, Dobrushin's criterion, is well known in various forms. Our presentation here is motivated by completeness as well as by our inability to find it in the literature expressed in the precise form that we need.

LEMMA A.1 (Dobrushin's criterion). *Consider a heat bath Markov operator* P_N *on* $\Omega_N = \{-1, +1\}^N$. *Define* $c_{ij} = \sup_{x \in \Omega_N} |P_N(x, x^i) - P_N(x^j, x^{i,j})|$ *for* $i \neq j$, *and let* $c_N = \sup_i \sum_{j \neq i} c_{ij}$. *Then*

$$\Delta(X^{\nu, HB}) \geq \frac{1 - c_N}{N}. \tag{A.1}$$

PROOF. For any index j and function f on Ω_N, define

$$\delta_j(f) := \sup_{x \in \Omega_N} |f(x) - f(x^j)|.$$

It is then not hard to show [e.g., equations (4.3)–(4.5) of Gross (1979)] that

$$\delta_j(P^{(i)} f) \leq \begin{cases} \delta_j(f) + c_{ij}\delta_i(f), & \text{if } j \neq i, \\ 0, & \text{if } j = i. \end{cases} \tag{A.2}$$

Next, on the subspace of $L^2(\nu)$ of functions having zero mean, define the norm

$$|||f||| = \sum_j \delta_j(f).$$

By summing (A.2) over j, we get

$$|||P^{(i)} f||| = \sum_j \delta_j(P^{(i)} f) \leq \sum_{j \neq i} \{\delta_j(f) + c_{ij}\delta_i(f)\}$$

and summing this over i yields

$$|||Pf||| \leq \frac{1}{N} \sum_{i,\, j:\, i \neq j} \{\delta_j(f) + c_{ij}\delta_i(f)\} \leq \left(1 - \frac{1 - c_N}{N}\right) |||f|||.$$

Since the heat bath dynamics is irreducible and reversible, the spectral gap equals $1 - \lambda_2$, where λ_2 is the second largest eigenvalue of P. By taking f to be the corresponding eigenvector, we immediately get (A.1). □

REMARK. By suitably generalizing the definition of c_{ij}, Lemma A.1 extends to Gibbs sampler dynamics for much more general multivariate state spaces. See, for example, Gross (1979).

PROOF OF THEOREM 3.2. Our goal is to make the induced distribution of energy levels

$$P_{F,N}\{S_N = 2i - N\} = P_{F,N}\{S_N = N - 2i\} =: \rho_{F,N}(i), \qquad i = N/2, \ldots, N$$

as close as needed to a uniform distribution. Note that

$$\frac{\rho_{\beta,N}(i+1)}{\rho_{\beta,N}(i)} = \frac{\binom{N}{i+1}}{\binom{N}{i}} \exp\left(\frac{\beta}{2N}((2i+2-N)^2-(2i-N)^2)\right)$$
$$= \frac{N-i}{i+1}\exp\left(\frac{2\beta}{N}(2i-N+1)\right)$$

for any $i \in \{N/2, \ldots, N-1\}$. By using some simple algebra it is easily obtained that for $i = N/2, \ldots, N-1$,

$$\rho_{\beta,N}(i+1) \le \rho_{\beta,N}(i) \quad \text{if and only if } \beta \le \beta_{i+1}^N, \tag{A.3}$$

where we have defined

$$\beta_{i+1}^N := \frac{N}{2(2i-N+1)} \log \frac{i+1}{N-i}.$$

Next, we claim that β_i^N is increasing in $i = N/2+1, \ldots, N$. Set $A := 1 + N^{-1}$, and define the function

$$b(r) := \frac{1}{2(2r-A)} \log \frac{r}{A-r} \quad \text{for } r \in (A/2, A).$$

Note that $b(i/N) = \beta_i^N$ for $i = N/2+1, \ldots, N$. Next for $D > C > 0$,

$$\log D - \log C < 2^{-1}(D-C)\left(\frac{1}{D} + \frac{1}{C}\right),$$

which follows from convexity: the left-hand side is the area below the curve $y = 1/x$ over the interval $[C, D]$, whereas the right-hand side is the area below the line joining the endpoints $(C, 1/C)$ and $(D, 1/D)$. Now let us differentiate the function $b(r)$ and apply this last inequality with $D = r$ and $C = A - r$ to obtain

$$b'(r) = \frac{-1}{(2r-A)^2} \log \frac{r}{A-r} + \frac{1}{2(2r-A)}\left(\frac{1}{r} + \frac{1}{A-r}\right) > 0$$

for $A/2 < r < A$. This proves the claim.

Now define $\beta_{N/2}^N = 0$ and $\beta_{N+1}^N = +\infty$. Suppose that $i \in \{N/2, \ldots, N\}$ and $\beta_i^N \le \beta \le \beta_{i+1}^N$. Then it follows from the claim of the preceding paragraph and from (A.3) that $\rho_{\beta,N}(j)$ is decreasing in j for $i \le j \le N$ and increasing in j for $N/2 \le j \le i$. In particular, the maximum occurs at $j = i$, and so

$$\rho_{\beta,N}(i) \ge \frac{1}{N+1}. \tag{A.4}$$

Next, for $k = 0, \ldots, N/2$, choose $\tilde{\beta}_k^N$ to be any point belonging to the interval $(\beta_{N/2+k}^N, \beta_{N/2+k+1}^N)$. Then for $i = N/2, \ldots, N$, we have because of (A.4)

$$\rho_{F,N}(i) \geq \frac{1}{(N/2+1)} \rho_{\tilde{\beta}_{i-N/2}^N}(i) \geq \frac{1}{(N/2+1)(N+1)}. \tag{A.5}$$

Now we consider the ratio of probability masses of the mixture $\mu_{F,N}$ defined by (3.16) and the measure ν_N uniform over energies from Theorem 3.1. Because of (A.5) we see that

$$\frac{\mu_{F,N}(\{x\})}{\nu_N(\{x\})} = (N+1)\rho_{F,N}\left(\frac{N+S_N(x)}{2}\right) \in \left[\frac{1}{(N/2)+1}, N+1\right]. \tag{A.6}$$

Now (3.17) follows from (A.6) and (3.11). Also, (A.6) allows us to apply Proposition 2.3 (with $a = (N/2+1)^{-1}$ and $b = N+1$) and use Theorem 3.1 to obtain (3.18). The final assertion of the theorem follows immmediately from (3.18) and Proposition 2.1. □

Acknowledgments. We have benefitted from valuable discussions with Alan Gelfand, Charlie Geyer, Andrea Pelissetto, Buks van Rensburg, Jeff Rosenthal, Alan Sokal and John Valleau.

REFERENCES

ARNOLD, S. F. (1993). Gibbs Sampling. In *Handbook of Statistics* (C. R. Rao ed.) **10** North-Holland, Amsterdam.

BENNETT, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22** 245–268.

BERG, B. A. and NEUHAUS, T. (1991). Multicanonical algorithms for first order phase transitions. *Phys. Lett. B* **267** 249–253.

BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55** 25–37.

BINDER, K. and HEERMANN, D. W. (1992). *Monte Carlo Simulation in Statistical Physics*, 2nd ed. Springer, New York.

BUCKLEW, J. A. (1990). *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley, New York.

CARACCIOLO, S., PELISSETTO, A. and SOKAL, A. D. (1990). Nonlocal Monte Carlo algorithm for self-avoiding walks with fixed endpoints. *J. Statist. Phys.* **60** 1–53.

CHAN, K. S. and GEYER, C. J. (1994). Discussion of "Markov chains for exploring posterior distributions" by L. Tierney. *Ann. Statist.* **22** 1747–1758.

DIACONIS, P. and SALOFF-COSTE, L. (1996). Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.* **6** 695–750.

DIACONIS, P. and STROOCK, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1** 36–61.

DUFLO, M. (1996). *Algorithmes Stochastiques*. Springer, Berlin.

ELLIS, R. (1985). *Entropy, Large Deviations and Statistical Mechanics*. Springer, New York.
EVANS, M. and SWARTZ, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statist. Sci.* **10** 254–272.
FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, London.
FERRENBERG, A. M. and SWENDSEN, R. H. (1988). New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.* **61** 2635–2638.
GAMERMAN, D. (1997). *Monte Carlo Markov Chains: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, London.
GEYER, C. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E. M. Keramidas, ed.) 156–163. Interface Foundation, Fairfax Station.
GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.
GILKS, W. R., RICHARDSON, S. and SPIGELHALTER, D. J. (eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.
GROSS, L. (1979). Decay of correlations in classical lattice models at high temperature. *Comm. Math. Phys.* **68** 9–27.
HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.
JANSE VAN RENSBURG, E. J. and MADRAS, N. (1997). Monte Carlo study of the θ-point for collapsing trees. *J. Statist. Phys.* **86** 1–36.
KIPNIS, C. and VARADHAN, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.* **104** 1–19.
LIU, J., WONG, W. H. and KONG, A. (1994). A covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes. *Biometrika* **81** 27–40.
MADRAS, N. and RANDALL, D. (1999). Markov chain decomposition for convergence rate analysis. Preprint.
MARINARI, E. and PARISI, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19** 451–458.
MENG, X. L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical explanation. *Statist. Sinica* **6** 831–860.
METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER E. (1953). Equations of state calculation by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.
MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.
NEAL, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statist. Comput.* **6** 353–366.
PESKUN, P. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika* **60** 607–612.
ROBERT, C. (1996). *Methodes de Monte Carlo par chaînes de Markov*. Economica, Paris.
SINCLAIR, A. (1993). *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhäuser, Boston.
SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 3–23.
SOKAL, A. D. (1989). Monte Carlo methods in statistical mechanics: Foundations and new algorithms. Lecture notes: Cours de Troisième Cycle de la Physique en Suisse Romande (Lausanne, June 1989). Unpublished manuscript
SWENDSEN, R. H. and FERRENBERG, A. M. (1990). Histogram methods for Monte Carlo data analysis. In *Computer Studies in Condensed Matter Physics II* (D. P. Landau, K. K. Man and H. B. Schüttler, eds.) 179–183. Springer, Berlin.
TANNER, M. A. (1993). *Tools for Statistical Inference*. Springer, New York.
THOMPSON, C. J. (1972). *Mathematical Statistical Mechanics*. Macmillan, New York.

TIERNEY, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9.

TORRIE, G. M. and VALLEAU, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free energy estimation: Umbrella sampling. *J. Comput. Phys.* **23** 187–199.

TROTTER, H. F. and TUKEY, J. W. (1956). Conditional Monte Carlo for normal samples. In *Symposium on Monte Carlo Methods* (H. A. Meyer, ed.) 64–79. Wiley, New York.

VALLEAU, J. P. (1991). Density-scaling: a new Monte Carlo technique in statistical mechanics. *J. Comput. Phys.* **96** 193–216.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
YORK UNIVERSITY
4700 KEELE STREET
TORONTO, ONTARIO M3J 1P3
CANADA
E-MAIL: madras@mathstat.yorku.ca

DIPARTIMENTO DI MATEMATICA
PURA E APPLICATA
UNIVERSITÁ DI L'AQUILA
67100, L'AQUILA
ITALY
E-MAIL: piccioni@aquila.infn.it