CHAPTER 1

STATISTICAL PHYSICS AND STATISTICAL COMPUTING: A CRITICAL LINK

James D. Servidea
U.S. Department of Defense, Washington, DC

Xiao-Li Meng
Department of Statistics, Harvard University, Cambridge, MA 02138
meng@stat.harvard.edu

The main purpose of this chapter is to demonstrate the fruitfulness of cross-fertilization between statistical physics and statistical computation, by focusing on the celebrated Swendsen-Wang algorithm for the Ising model and its recent perfect sampling implementation by Huber. In particular, by introducing Hellinger derivative as a measure of instantaneous changes of distributions, we provide probabilistic insight into the algorithm's critical slowing down at the phase transition point. We show that at or near the phase transition, an infinitesimal change in the temperature parameter of the Ising model causes an astronomical shift in the underlying state distribution. This finding suggests an interesting conjecture linking the critical slowing down in coupling time with the grave instability of the system as characterized by the Hellinger derivative (or equivalently, by Fisher information). It also suggests that we can approximate the critical point of the Ising model, a physics quantity, by monitoring the coupling time of Huber's bounding chain algorithm, an algorithmic quantity. This finding might provide an alternative way of approximating criticality of thermodynamic systems, which is typically intractable analytically. We also speculate that whether we can turn perfect sampling from a pet pony into a workhorse for general scientific computation may depend critically on how successful we can engage, in its development, researchers from statistical physics and related scientific fields.

Keywords: Bounding chains, critical slowing, exact simulation, Fisher information, heat capacity, Hellinger distance, Ising model, Markov chain Monte Carlo, perfect sampling, Swendsen-Wang algorithm.

1. MCMC Revolution and Cross-fertilization

One of the spectacular successes in statistics, the Markov chain Monte Carlo (MCMC) revolution, began around 1990. The method, however, was in existence for more than four decades, beginning with the publication of Metropolis et. al. (1953) in the statistical physics literature. This well known fact is documented in Liu (2002), which also provides a detailed account of many recent advances made by physicists, statisticians and other scientists. Today, MCMC is used to study everything from single molecules to supernovas and from financial engineering to anti-spam software. This enormous success demonstrates the fruitfulness of cross-fertilization between statistics and an area of application, with the former serving as a "central hub" to enhance and transfer the method to essentially every field that requires quantitative investigations under uncertainty. Furthermore, the cross-fertilization also provides refinements and new methods for the application itself. This chapter documents one more example of this kind, along the way providing a brief review of Ising model, one of the most basic models in statistical physics, and of perfect sampling, a most recent development in the MCMC evolution.

2. The Ising Model

Introduced in the 1920s, the Ising model is widely used today in the areas of statistical mechanics and image analysis. It specifies a probability distribution (a Gibbs distribution, also called a Boltzmann distribution) on the magnetic spins of a lattice. In image analysis, the nodes of the lattice are viewed as image pixels and the spins on those nodes as colors. Let L denote a set of nodes (or more generally a graph). For two distinct nodes $i, j \in L$, we shall denote $i \sim j$ to mean that i and j are nearest neighbors in the lattice.

Associated with each node i of the lattice is a spin X_i that can be either down ($X_i = -1$) or up ($X_i = 1$). In image analysis, these correspond to white and black pixels, respectively. The probability mass function specified by the model is given by

$$p_\beta(x) = \frac{1}{Z(\beta)} \exp\left(J\beta \sum_{i \sim j} x_i x_j \right), \quad x \in \{-1, 1\}^{|L|}, \tag{1}$$

where $|\cdot|$ denotes cardinality and the index set "$i \sim j$" is taken to mean $\{(i, j) \in L^2 : i \sim j\}$. Since $i \sim j$ if and only if $j \sim i$, we take $(i, j) \equiv$

(j, i), and therefore each pair of adjacent nodes is counted only once in the summation. In (1), J is known as the *coupling constant*, with $J = 1$ for a ferromagnetic model, where spins tend to align themselves with their neighbors, and with $J = -1$ for an anti-ferromagnetic model, where spins tend to be opposite to their neighbors (see Thompson, 1972). The case of $J = 0$ corresponds to independence among spins. The parameter $\beta = 1/(kT)$, where T is the absolute temperature of the system and k is Boltzmann's constant, controls the amount of attraction or repulsion. Here we focus on the more common case of $J = 1$. This is the usual case in image analysis, where pixels tend to be like their neighbors.

The normalizing constant $Z(\beta)$ is known to physicists as a partition function. For some simple lattices, the partition function can be calculated analytically, but, generally speaking, solutions that are closed-form or nearly so are elusive. Consequently, the evaluation of $Z(\beta)$ often requires Monte Carlo simulation.

3. The Swendsen-Wang Algorithm and Criticality

In order to generate Monte Carlo simulations, we need to sample from (1). The earliest attempt was christened the *heat bath algorithm*, which statisticians would characterize today as a Gibbs sampler. The idea was to update the spin of each node conditionally, given the spins of all other nodes in the lattice. Due to the spatial Markovian nature of (1), this is equivalent to updating each node's spin given the spins of its nearest neighbors. This rather simple scheme proves to be painfully slow when β is large, as its approach is "local" in nature.

Swendsen and Wang (1987) developed an MCMC algorithm to address this problem that can be viewed as an auxiliary variables technique (e.g., Higdon, 1998). For each pair of adjacent nodes (i, j), they define a binary random variable $U_{i,j}$ that takes values in $\{-1, 1\}$. Think of $U_{i,j}$ as an edge bond between i and j that might bind the two spins in the same direction ($U_{i,j} = 1$) or leave them free ($U_{i,j} = -1$). Each update of their algorithm consists of the following two steps.

STEP I:
For each $(i, j) \in L^2$ such that $i \sim j$,
Let $U_{i,j} := X_i X_j$.
If $U_{i,j} = 1$, **Let** $U_{i,j} := -1$ with probability $\exp(-2\beta)$.
STEP II:

Let $\mathbb{C}$ be the set of all clusters $C \subset L$ such that the nodes of C are connected by edges (i,j) with $U_{i,j} = 1$.
For each $C \in \mathbb{C}$,
Let x be a random spin, -1 or 1, with probability $1/2$ each.
For all $i \in C$, **Let** $X_i := x$.

By viewing $U \equiv \{U_{i,j} : i \sim j\}$ as auxiliary variables, the Swendsen-Wang algorithm can be seen as a Gibbs sampler where STEP I samples U given $X \equiv \{X_i : i \in L\}$ and STEP II samples X given U. Here the conditional distribution of U given X is constructed to be

$$f(u|x) \propto \left(1 - e^{-2\beta}\right)^{N_u} \left(e^{-2\beta}\right)^{N_x - N_u} \prod_{i \sim j} 1_{\{u_{i,j} \le x_i x_j\}}, \tag{2}$$

where

$$N_u = \sum_{i \sim j} \frac{u_{i,j} + 1}{2} \quad \text{and} \quad N_x = \sum_{i \sim j} \frac{x_i x_j + 1}{2}. \tag{3}$$

This construction means that when $x_i \neq x_j$, we set $u_{i,j} = -1$, that is, the two nodes i and j are not bound. On the other hand, when $x_i = x_j$, it is only with probability $e^{-2\beta}$ that i and j are not bound (i.e., $u_{i,j} = -1$). The binomial part of (2) is easily seen when we note that there are exactly N_x edges with $x_i = x_j$ and N_u edges with $u_{i,j} = 1$.

Under the above joint specification $f(x,u) = f(u|x)f(x)$, noting that $f(x) \propto e^{2\beta N_x}$ from (1) and (3), we can easily see that $f(x|u) \propto \prod_{i \sim j} 1_{\{x_i x_j \ge u_{i,j}\}}$. This leads to

$$f(x|u) = \left(\frac{1}{2}\right)^{|\mathbb{C}_u|} \prod_{i \sim j} 1_{\{x_i x_j \ge u_{i,j}\}}, \tag{4}$$

where $\mathbb{C}_u$ is the set of all "clusters" defined by u and $|\mathbb{C}_u|$ is its cardinality. Therefore, conditioning on U, each cluster of lattice nodes is assigned a spin independently with equal probabilities for up and down. The product in (4) ensures $\{(i,j) : u_{i,j} = 1\} \subset \{(i,j) : x_i x_j = 1\}$, a compatibility requirement between corresponding edge bonds and nodes' spin directions.

It is not difficult to show that the Markov chain specified by the Swendsen-Wang algorithm is ergodic and has (1) as its stationary distribution. Moreover, this chain mixes substantially faster than the heat bath algorithm. But, both algorithms suffer from what physicists call *critical slowing down*, a phenomenon that we shall present in more probabilistic terms.

As mentioned earlier, the normalizing constant $Z(\beta)$ as a function of the system parameter β is an important tool in physics. By taking derivatives of simple functions of Z with respect to the parameters such as the temperature (hidden in β for our purposes), one obtains various important physical quantities. For example, the heat capacity C of a system can be computed by the formula (McCoy and Wu, 1973)

$$C = -\frac{\partial}{\partial T}\frac{\partial}{\partial \beta}\log Z(\beta) = k\beta^2\frac{\partial^2}{\partial \beta^2}\log Z(\beta), \tag{5}$$

where the second identity is due to $\beta = 1/(kT)$. That is, the heat capacity is the rate of change of the expected energy of a system as a function of its temperature.

It turns out that, on many lattices, for a certain value of β, $Z(\beta)$ nearly fails to be analytic, thus leading to astronomically large values of physical quantities, the heat capacity in particular. Such a point is called *critical*, and, in this case, corresponds to a phase transition: the proportion of up spins as opposed to down goes from 1/2 to virtually 0 or 1 with equal probability as β increases past its critical value. In the thermodynamic limit, for an infinitely large, square lattice, Z is not analytic at the critical point and the heat capacity (per node) at that point is infinite. This critical value of β is well known to be

$$\beta_c = \frac{1}{2}\operatorname{arcsinh}(1) = \frac{1}{2}\log(1+\sqrt{2}) \approx 0.44068679348. \tag{6}$$

See Baxter (1982) for a derivation.

It is well known that when the parameters are near critical, the Swendsen-Wang algorithm, like many other Markov chain samplers, tends to mix very slowly. In the next section, we offer an intuitive explanation in probabilistic terms for why this happens.

4. Instantaneous Hellinger Distance and Heat Capacity

Given a measure space $(\Omega, \mathcal{F}, \mu)$, let $\mathcal{P}$ be the set of all probability densities with respect to μ. On $\mathcal{P}$, the *Hellinger distance* is the metric defined by

$$H(p,q) = \left[\int_\Omega (\sqrt{p}-\sqrt{q})^2 \, d\mu\right]^{\frac{1}{2}}, \quad p, q \in \mathcal{P}.$$

Now suppose $\Omega \subset \mathbb{R}^k$. The subfamily of $\mathcal{P}$ whose elements are of the form

$$p_\theta(\omega) = \frac{1}{c(\theta)}\exp(\theta \cdot \omega),$$

indexed by $\theta \in \mathbb{R}^k$ such that $\int \exp(\theta \cdot \omega)\ \mu(d\omega) < \infty$, is the *exponential family* with canonical parameter θ. The Hellinger distance between two distributions whose densities belong to the same exponential family reduces to

$$H(p_{\theta_1}, p_{\theta_2}) = \left[2 - 2\frac{c(\bar{\theta})}{\sqrt{c(\theta_1)c(\theta_2)}}\right]^{\frac{1}{2}}, \tag{7}$$

where $\bar{\theta} = (\theta_1 + \theta_2)/2$.

The Hellinger distance is a bounded metric that equals $\sqrt{2}$ if the two densities have no common support on Ω (almost everywhere with respect to μ). For our purposes, it will serve as a handy way of measuring overlap in the Gibbs distributions of the Ising model for different values of β. And since (1) has an exponential form, we can express the Hellinger distance between Ising model distributions as a function of the partition function Z using (7) above.

Consider the Hellinger "derivative" given by:

$$\partial H(\beta) = \lim_{\epsilon \downarrow 0}\ \frac{1}{2\epsilon} H(p_{\beta-\epsilon}, p_{\beta+\epsilon}). \tag{8}$$

For a finite lattice, $Z(\beta)$ is a finite sum of analytic functions and therefore is itself analytic. So, we can express $Z(\beta+\epsilon)$ and $Z(\beta-\epsilon)$ as Taylor series to obtain

$$Z(\beta+\epsilon)Z(\beta-\epsilon) = Z^2(\beta) + \epsilon^2\left[Z(\beta)Z''(\beta) - [Z'(\beta)]^2\right] + O(\epsilon^3).$$

Define $K = (Z(\beta)Z''(\beta) - [Z'(\beta)]^2)/Z^2(\beta)$ and apply (7) to see that

$$\partial H(\beta) = \lim_{\epsilon \downarrow 0}\ \frac{1}{\sqrt{2}}\left[\frac{1 - [1 + \epsilon^2 K + O(\epsilon^3)]^{-1/2}}{\epsilon^2}\right]^{\frac{1}{2}}.$$

Using L'Hôpital's rule, we finally obtain

$$\partial H(\beta) = \lim_{\epsilon \downarrow 0}\ \frac{1}{\sqrt{2}}\left[\frac{[1 + \epsilon^2 K + O(\epsilon^3)]^{-3/2}[\epsilon K + O(\epsilon^2)]}{2\epsilon}\right]^{\frac{1}{2}} = \frac{\sqrt{K}}{2}.$$

Note that K is simply the second derivative of $\log Z$, thus

$$\partial H(\beta) = \frac{1}{2}\left[\frac{\partial^2}{\partial\beta^2}\log Z(\beta)\right]^{\frac{1}{2}}. \tag{9}$$

Comparing (5) and (9), we see immediately that

$$C = 4k[\beta\ \partial H(\beta)]^2. \tag{10}$$

Therefore, the heat capacity of the Ising model is large wherever the Hellinger distance changes rapidly as a function of β. Conversely, the Hellinger distance changes very quickly wherever the heat capacity is large, in particular, whenever a phase transition occurs. Consequently, it is no surprise that it is so difficult to simulate the Ising model near the critical value. A tiny perturbation in the parameter β gives a new distribution that *effectively* shares almost no support with the old one. That is, the probability masses of the new and old distributions concentrate on different parts of the state space, even though technically both distributions have the same *mathematical* support.

For example, the above result suggests that, even for parallel tempering (Geyer, 1991, Geyer and Thompson, 1995, Tesi et. al., 1996), one of today's most promising methods, we will need to use a very large number of intermediate temperatures in order to achieve a useful swapping probability at or near the critical value. Indeed, recent results by Madras and Zheng (2002) imposed the condition that the number of intermediate temperatures be proportional to the number of the nodes, in order for the parallel tempering to achieve polynomial mixing time, for a couple models simpler than (1). Our finding above makes us to believe that this condition is not only sufficient but also necessary. Moreover, because (10) is by no means limited to the Ising model—any thermodynamic model of this exponential type will have the same property—we conjecture that this necessary condition will hold quite generally. Putting it differently, it is very difficult to imagine that the parallel chains can swap with nontrivial probabilities without making the chains "infinitesimally closer" when their corresponding stationary distributions have "infinitesimally small" overlaps as measured by the Hellinger distance.

We remark that, in general, the Hellinger derivative, under mild regularity conditions, is half of the square root of the Fisher information. This can be seen by invoking a Taylor expansion of the log-likelihood function (when it is justified)

$$\ell(\theta+\delta|X) = \ell(\theta|X) + S(\theta|X)\delta - \frac{I(\theta|X)}{2}\delta^2 + O_p(\delta^3), \tag{11}$$

where $S(\theta|X)$ is the score function and $I(\theta|X)$ is the *observed* Fisher information (where X is the observed data). Applying (11) to $p(x|\theta \pm \epsilon)$ in (8) with $\delta = \pm\epsilon$ (and with θ in place of β) yields

$$\partial H(\theta) = \lim_{\epsilon \downarrow 0} \left[\int \frac{1 - \exp\{-I(\theta|X)\epsilon^2/2 + O_p(\epsilon^3)\}}{2\epsilon^2} p(X|\theta)\mu(dX) \right]^{\frac{1}{2}}. \tag{12}$$

When the limit operator in (12) can be interchanged with the integration operator, we can conclude that

$$\partial H(\theta) = \frac{1}{2}\left[\int I(\theta|X)p(X|\theta)\mu(dX)\right]^{\frac{1}{2}} = \frac{1}{2}\sqrt{I(\theta)}, \tag{13}$$

where $I(\theta)$ is the *expected* Fisher information. This general result provides additional insight into the phase transition phenomenon, because Fisher information measures the ability to distinguish different distributions and a near infinite Fisher information indicates that the systems on "opposite" sides of a critical point are fundamentally different (and therefore trivial to distinguish). This result also implies that, for the Ising model (1), the Hellinger derivative is also the half the standard deviation of the "sufficient statistic" $\sum_{i\sim j} x_i x_j$ (this can also be obtained by observing that in an exponential family, the log of the normalizing constant is also the cumulant generating function). So, near the critical value, the sufficient statistic has essentially infinite variance. For a detailed treatment of the mixing properties of the Swendsen-Wang algorithm near critical points, see Borgs *et. al.* (1999).

5. A Brief Overview of Perfect Sampling

A major contribution to Monte Carlo simulation was made by Propp and Wilson (1996), who established and popularized a general framework for converting Markov chain samplers into exact samplers. Whereas an MCMC algorithm generally produces authentic draws from the targeted stationary distribution only at the limit, Propp and Wilson showed how this can actually be achieved, for some problems, in a finite (but random) number of iterations with probability one. This seemingly impossible task was accomplished by coupling, *from the past* (i.e., with negative time), the same Markov chain with different starting points in a set $\mathcal{S}$ to detect a coalescence at time zero. As far as detecting limits goes, going from time $-\infty$ to time zero is the same as going from time zero to $+\infty$. The coupling is done by using the same random number sequence for all the chains, but by going backward in time from time zero, we do not need to go back infinitely many steps before detecting the coalescence.

One way to illustrate this is to consider a Markov chain $Y_t = \psi(Y_{t-1}, U_t)$, where $\{U_t, t = 1, 2, \ldots\}$ are i.i.d. random variables. This produces the so-called *time-forward* sequence

$$Y_t = \psi(\psi(\ldots\psi(Y_0, U_1)\ldots, U_{t-1}), U_t), \tag{14}$$

where $Y_0 \in \mathcal{S}$ is a starting point. In contrast, we can also build the *time-backward* sequence

$$\tilde{Y}_t = \psi(\psi(\ldots\psi(Y_0, U_t)\ldots, U_2), U_1). \tag{15}$$

Clearly, Y_t and $\tilde{Y}_t$ have identical distribution for any t, yet $\{\tilde{Y}_t, t \geq 1\}$ itself is not a Markov chain. As discussed in Craiu and Meng (2005), by giving up the Markovian property, we gain a better convergence property, because while $\{Y_t, t \geq 1\}$ converges in distribution, $\{\tilde{Y}_t, t \geq 1\}$ converges with probability 1 (under regularity conditions). This is because as t increases, $\tilde{Y}_t$ varies less and less as a function of Y_0, since U_t enters into deeper and deeper levels of the nested function compositions in (15). In other words, as t increases, $\tilde{Y}_t$ becomes increasingly restricted by the entire history of U's. Contrast this with (14), where U_t is always in the outside level of the composition. Indeed, it can be shown that (Foss and Tweedie, 1998) if ψ is such that Y_t is uniformly ergodic, then with probability one, $\tilde{Y}_t$, as a function of Y_0, will become a constant function with a finite (but random) t. That is, $\tilde{Y}_t$ can "hit and stay at" the limit in a *finite number* of iterations (Thorisson, 2000). This is really the essence of perfect or exact sampling—Propp and Wilson's (1996) coupling from the past (CFTP) method is just a device for finding this finite stopping time by mapping t to $-t$ in (15) and thereby creating a forward-looking sequence going from time $-t$ to time zero.

While the idea of using negative time Markov chains was known previously in the theoretical context, Propp and Wilson (1996) appear to be the first to have brought this theoretical tool to the attention of the general Monte Carlo simulation community. One key observation made by Propp and Wilson (1996) is that, for many monotone, discrete-state space Markov chains, we do not need to choose the "starting set" $\mathcal{S}$ to be the whole state space. Here by "monotone" we mean $Y_t = \psi(Y_{t-1}, U_t)$ is monotone in Y_{t-1} with respect to a partial ordering. In such cases, if there is also a maximum state and minimum state with respect to the same partial ordering, we can choose $\mathcal{S}$ to contain only these two extreme states, that is, to trace only the maximum chain and minimum chain. The monotonicity of the updating function then ensures that if the two coupled chains from these two extreme states coalesce at time zero, all coupled chains starting from other states will automatically reach the same state at time zero because they will all be "squeezed" between the two extreme chains. For other clever tricks and strategies for implementing perfect sampling, consult the web site maintained by David Wilson, `http://dimacs.rutgers.edu/~dbwilson/exact`,

which contains a comprehensive and up-to-date list of research papers on perfect sampling and related topics.

6. Huber's Bounding Chain Algorithm

When there is no such monotonicity that can be used, it is still possible to use the "squeezing" idea, as discussed in Häggström and Nelander (1998) and Huber (1998). The idea is to introduce an appropriate bounding chain such that the coalescence of the bounding chain implies the coalescence of the original chain, and the coalesced value from the original chain is a genuine draw from its stationary distribution. As an example, Huber (2002) provides such a bounding chain for the Ising model, using the Swendsen-Wang algorithm as its core. The key here is to design a Markov chain on the set of all nonempty subsets of the state space, such that when this chain arrives at a set of cardinality one, the one element of that set is a draw from the Ising model. See Huber (2002) for the derivation and proof. Here we give only the pseudocode. Note that in order to obtain exact samples from the Ising model, it should be implemented in the "backward" fashion as in Propp and Wilson (1996). However, to investigate empirically the coalescence time, we can run it forward as the forward and backward coalescence times have the same distribution (Propp and Wilson, 1996).

Specifically, for each $i \in L$, let Y_i be a set-valued random variable taking three possible values: $\{-1\}$, $\{1\}$, and $\{-1, 1\}$. Each update of Huber's bounding chain algorithm for the Ising model takes the following form:

For each $(i, j) \in L^2$ such that $i \sim j$,
  If $Y_i = Y_j$ and $|Y_i| = |Y_j| = 1$, **Let** $U_{i,j} := 1$.
    Else If $|Y_i \cap Y_j| \geq 1$, **Let** $U_{i,j} := 0$.
      Else Let $U_{i,j} := -1$.
  If $U_{i,j} \geq 0$, **Let** $U_{i,j} := -1$ with probability $e^{-2\beta}$.
Let $\mathbb{C}$ be the set of all clusters $C \subset L$ such that the nodes of C are connected by edges (i, j) with $U_{i,j} = 1$.
Let $\mathbb{C}'$ be the set of all clusters $C \subset L$ such that the nodes of C are connected by edges (i, j) with $U_{i,j} = 0$ or $U_{i,j} = 1$.
For each $C \in \mathbb{C}$,
  Let x be a random spin, -1 or 1, with probability $1/2$ each.
  For all $i \in C$, **Let** $X_i := x$.
Let π be a uniformly random ordering on $\mathbb{C}$.
For each $C \in \mathbb{C}$,

Let C' be the unique set in $\mathbb{C}'$ such that $C \subset C'$.
For each $i \in C$, **Let** X_D be the set (either $\{-1\}$ or $\{1\}$) that contains only the single common value of all x_i in the cluster D, and **Let**

$$Y_i := \bigcup_{\substack{D \in \mathbb{C} \cap C': \\ \pi(D) > \pi(C)}} X_D.$$

This algorithm's roots in Swendsen-Wang shine through. The only complication is the classification of edges. In Swendsen-Wang, each edge either binds its end nodes ($U_{i,j} = 1$) or does not ($U_{i,j} = -1$). Here, there exists the third possibility ($U_{i,j} = 0$) that the current values of Y_i and Y_j do not conclusively determine whether nodes i and j should have the same spin.

In a private communication, Huber pointed out that the ordering π on $\mathbb{C}$ can be omitted. In such a case, the assignment of the new value of Y_i is the union over all $D \in \mathbb{C} \cap C'$. This modification also produces a valid bounding chain, but this chain will take longer to couple.

As mentioned, to actually obtain draws from (1), one must implement the updating algorithm above using the coupling from the past protocol of Propp and Wilson (1996). The easiest way to do this is to carefully keep track of the seeds used by the pseudorandom number generators for the previous pass through the algorithm (note each pass uses no more than $4|L|$ uniforms). Our implementation was essentially as follows.

Let $k := 1$.
Repeat
  Generate new value for `seed` and
  Store it as the kth element in the list.
  Let $m := k$.
  Repeat
    Let `seed` take the value of the mth element in the list.
    Repeat 2^{m-1} times:
      Do one update according to the code above.
    Let $m := m - 1$.
  Until $m = 0$.
  Let $k := k + 1$.
Until $|Y_i| = 1$ for all $i \in L$.
Return X such that $X_i \in Y_i$ for all $i \in L$.
End.

Implementation of this algorithm is not especially difficult when using a lattice with a regular structure. The only tricky parts are those operations performed on the clusters, that is, the subsets of L in $\mathbb{C}$ and $\mathbb{C}'$. We warn anyone interested in implementing this algorithm to be especially careful when programming. Telling the difference between Ising states with the right distribution and those with a wrong one (but not "too wrong") is nearly impossible when merely looking at the output.

Moreover, note that the implementation above is different from merely running the updating sequence until the bounding chains coalesce. Such an algorithm would be "forward" in nature and is not guaranteed to produce draws from (1). However, the draws from the forward implementation may not appear to be different from those of a backward implementation, at least on casual inspection. The following example illustrates how we can find lower dimensional summaries that can reveal the biases in a high-dimensional sampler, even when the biases arise in subtle ways.

Specifically, we made 1000 draws from each implementation with $\beta = 0.25$ for a 128×128 toroidal lattice, and recorded the sufficient statistic $\sum_{i \sim j} x_i x_j$ for each draw. Unsurprisingly, the sample statistics are approximately normally distributed, and therefore their first two moments are good summaries of their distributions. The forward draws had mean 2281 and sample variance 11,915. The backward draws had mean 2286 and sample variance 11,780. But, examination of a quantile-quantile plot (see Figure 1) reveals some differences in the tails. To see this more clearly, we divided our sample into 100 subsamples of 10 draws of each kind. For each subsample we computed the minimum and the maximum. The minima had mean 2113 and sample variance 3810 for the forward draws and mean 2122 and sample variance 5404 for the backward draws. Similarly for the maxima, we observed a mean of 2449 and sample variance 3666 for the forward draws and a mean of 2451 and a sample variance of 4595 for the backward draws. A common obstacle with most forward implementations is that the distribution of the resulting draw is condensed onto a subset of the draw's state space (i.e., it "gets stuck" in a subspace). We suspect that the diminished variances of the order statistics in the forward implementation reflects this problem.

7. Approximating Criticality via Coupling Time

Figure 2 shows six Ising draws on a 128×128 square lattice with toroidal ("wrap around") boundary conditions to reduce boundary effects. A white

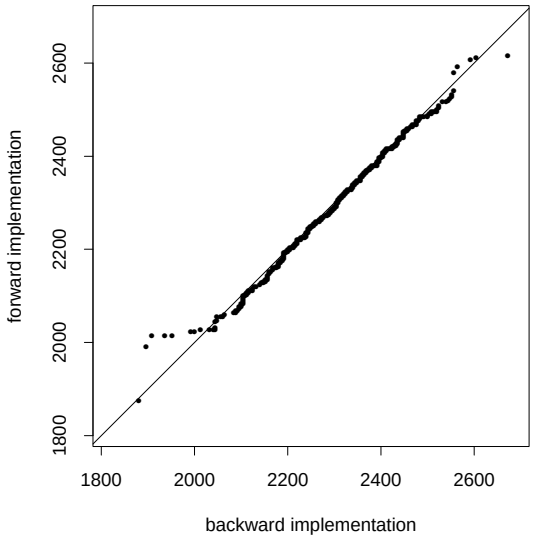

Fig. 1. Comparisons of forward and backward implementations.

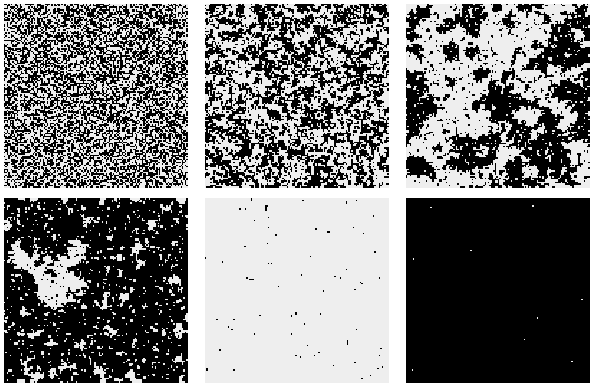

Fig. 2. Ising draws from Huber's algorithm, with $\beta = 0.05, 0.3, 0.42, 0.44, 0.7$ and 0.9, respectively.

pixel indicates a down spin and a black pixel an up spin, though the distinction is unimportant given the symmetry of the distribution. The values of β corresponding to the draws are 0.05, 0.3, 0.42, 0.44, 0.7, and 0.9, respectively. Note, that for small β, there are approximately as many down spins as up spins. Also note that they are clumped together more than one would expect if each spin were assigned independently. In the $\beta = 0.42$ picture the clusters have grown larger, but we still have approximate parity of spin directions. This indicates that the phase transition has not occurred. When $\beta = 0.44$, we see that the parity of spin directions is long gone, indicating that the phase transition has occurred. As β increases, one spin direction dominates the lattice, with only a few brave spins in the opposite direction.

A natural question to ask is how many iterations are typically required for coalescence, particularly when β is near criticality. To answer this, for each β on a grid from 0.01 to 0.99 with increments of 0.01, and for each of four increasingly larger lattices, we made 100 draws using Huber's bounding chain *with a forward implementation*. This is justified by the fact that the distribution of the coupling time is the same for the forward or backward implementation. Moreover, a forward implementation counts the precise number of iterations, whereas a typical backward implementation counts on an exponential base 2 scale, the so-called "binary back-off" scale (Propp and Wilson, 1996). For each forward draw, we recorded the number of iterations.

The average number of iterations till coalescence are shown in Figure 3. The bottom curve was generated using a 16×16 square, toroidal lattice. The curves above it correspond to 32×32, 64×64, and 128×128 lattices, in that order. The scale of the plot is logarithmic, base 4. We chose base 4, since each lattice contains four times as many nodes as the next smaller one. The vertical line represents β_c of (6).

It is not difficult to conjecture that the peak corresponds to critical slowing down, although it does not occur precisely at β_c, which is for an infinite lattice, not for our finite lattices with toroidal boundary conditions. Given the well developed theory of "finite-size scaling" in statistical physics (e.g., Binder and Heermann, 1982; Landau and Binder, 2000), one way to empirically check this conjecture is to see if the peak as identified by the coupling method will approach the limiting value β_c as the lattice size increases, though we have neither the patience, nor the computational power, to verify this. The critical slowing is not surprising, since there is a strong relationship between the coupling time of a backward sampler and the mixing time of its underlying Markov chain, and the Swendsen-Wang

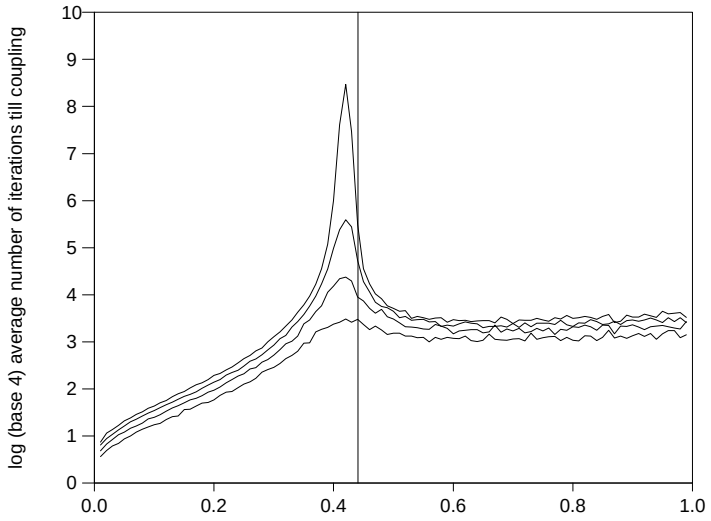

Fig. 3. Average number of iterations till coupling of Huber's algorithm as a function of β for four lattice sizes: $2^k \times 2^k, k = 4, 5, 6, 7$ (from bottom up). The vertical line corresponds to the limiting critical value β_c with infinite lattice.

algorithm is known to be subject to critical slowing, as discussed previously. Away from the peak, the graphs are more or less evenly spaced, suggesting a linear increase in the number of iterations till coupling as a function of lattice size. The bumpiness of the graphs on the high end indicates larger variation in the coupling time for big values of β. On the whole, we feel that the general shape of the graphs at that end is clear.

Using the information obtained when creating Figure 3 also enabled us to improve the quality of our backward implementation. Except when β is very small, starting with a negative t near zero for the backward coupling is usually a poor choice. In our implementation, we estimated the 99th percentile of the required number of steps till coupling, for each value of β. The initial number of steps was set to be this percentile depending on β. This allowed us to make only one pass through the outer loop for generating each picture of Figure 2. On occasion, of course, more than one pass may be necessary.

We emphasize that, although Huber's bounding chain algorithm does slow considerably near criticality, it is nevertheless a very useful algorithm.

Our earlier discussion of criticality leaves little hope for the discovery of any algorithm that does not exhibit some kind of critical slowing. Moreover, by making plots similar to Figure 3, one can empirically determine critical values on any lattice for which the algorithm is implemented. This potentially can be useful, because the analytical computation of critical values quickly gets out of hand when the lattice is made even slightly more complex.

We conclude this section by noting that if our earlier conjecture is true, it should hold for MCMC algorithms other than Huber's algorithm, because our general results in Section 4 refer to a property of the target distribution, not any particular MCMC algorithm. That is, our general conjecture is that there is a close tie between the magnitude of the Hellinger derivative, with respect to a parameter indexing a family of distributions, and an MCMC algorithm whose stationary distribution is from the same family (and with the same parameter value). Of course, we are at a very early stage of exploring this relationship, for we yet need to formulate a measure to quantify this tie.

8. A Speculation

A perfect sampling algorithm such as Huber's is very useful even if one is unwilling to wait long to use it to obtain a large sample of independent draws, because it can be used to generate the initial state for a Markov chain sampler like Swendsen-Wang. Indeed, using a perfect sampling algorithm to generate an initial state is a generally recommended strategy because it eliminates potential biases due to insufficient "burn-in." For high-dimensional problems such as Ising models, this is particularly important because it is generally not possible to just choose a "reasonable looking" starting point (which is often feasible for univariate distributions), nor is it easy, if at all possible, to perform reliable diagnostic tests to determine an adequate burn-in period (for the entire state distribution, not just for some summary statistics).

Given this utility, we naturally wonder if we can construct perfect sampling algorithms for many, if not all, practical applications, such as those encountered in *routine* Bayesian computation (Gelman et. al. 2004; van Dyk and Meng, 2001). This turns out to be a very challenging problem in general, as demonstrated in Murdoch and Meng (2001). Indeed, the jury is still out on whether perfect sampling can ever be a tool as general as the general MCMC framework itself.

Historically, essentially all major Monte Carlo methods were origi-

nated, (re)discovered independently, or at least researched extensively by researchers in statistical physics or in related scientific areas. Perfect sampling seems to be the only exception, at least so far. While some of us might take pride in this observation, we wonder if this lack of involvement of these researchers, so far, is somewhat responsible for the fact that perfect sampling is still largely a pet pony, not a workhorse, for general scientific computation. Perhaps it can never be a workhorse as strong as Metropolis-Hastings or the Gibbs sampler. But nevertheless given the amazing success of statistical physicists and the like in the entire history of Monte Carlo development, it seems logical to speculate that the real breakthrough of perfect sampling as a general tool might depend on, possibly critically, the direct involvement of researchers from statistical physics and related scientific areas.

Acknowledgments We thank J. Blanchet, G. Goswami, M. Huber, J.S. Liu, and S. Lalley for helpful exchanges, and a referee for a set of very insightful comments. This research was supported in part by a U.S. Department of Education grant (Servidea) and several National Science Foundation grants (Meng).

References

1. R. J. Baxter, *Exactly Solved Models in Statistical Mechanics*, New York: Academic Press (1982), §6.1.
2. C. Borgs, J. T. Chayes, A. Frieze, J. H. Kim, P. Tetali, E. Vigoda, and V. H. Vu, "Torpid mixing of some Monte Carlo Markov chain algorithms in statistical physics," Proc. 40th IEEE Symposium on Foundations of Computer Science (1999), pp. 218–229.
3. R. Craiu and X.-L. Meng, "Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo," Annals of Statistics 33 (2005), pp. 661 - 697.
4. Foss, S.G. and Tweedie, R.L., "Perfect simulation and backward coupling," Comm. Statist. Stochastic Models 14 (1998), pp. 187-203.
5. C. J. Geyer, "Markov chain Monte Carlo maximum likelihood," Computing Science and Statistics: Proc. 23rd Symp. Interface (1991), pp. 156-163.
6. C. J. Geyer and E. A. Thompson, "Annealing Markov chain Monte Carlo with applications to ancestral inference," J Amer Statist Assoc 90 (1995), pp. 909-920.
7. A. Gelman, J. Carlin, H. Stern and D.B. Rubin, *Bayesian Data Analysis* (2nd Ed). Chapman and Hall. (2004)
8. O. Häggström and K. Nelander, "Exact sampling from anit-monotone systems," Statistica Neerlandica 52 (1998), pp. 360–380.
9. D. M. Higdon, "Auxiliary variable methods for Markov chain Monte Carlo with applications," J Amer Statist Assoc 93 (1998), pp. 585–595.

10. M. L. Huber, "Exact sampling and approximate counting techniques," Proc 30th Symp Theory of Computing (1998), pp. 31-40.
11. M. L. Huber, "A bounding chain for Swendsen-Wang," Random Struct Alg 22 (2002), pp. 43-59.
12. J. S. Liu. *Monte Carlo Methods for Scientific Computation*, Springer Verlag, Berlin (2002).
13. B. M. McCoy, and T. T. Wu, *The Two-Dimensional Ising Model*, Cambridge, Mass., Harvard University Press (1973), §2.4-§2.5.
14. Metropolis, N, Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller E. "Equations of state calculations by fast computing machines," J Chemical Physics 21(6) (1953), pp. 1087–1091.
15. N. Madras and Z. Zheng, "On the swapping algorithm," Random Struct Alg 22 (2002), pp. 66-97.
16. D. Murdoch and X.-L. Meng, "Towards perfect sampling for Bayesian mixture prior," In *Bayesian Methods, with Applications to Science, Policy and Official Statistics* (E. George, ed.) (2001), pp 318-390. Eurostat.
17. J. G. Propp and D. B. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics," Random Struct Alg 9 (1996), pp. 223–252.
18. R. H. Swendsen and J. S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," Phys Rev Letters 58 (1987), pp. 86–88.
19. C. J. Thompson. *Mathematical Statistical Mechanics.* Princeton University Press (1972).
20. Thorisson, H. *Coupling, Stationary, and Regeneration.* Springer, New York (2000).
21. M.C. Tesi, E.J. Janse van Rensburg, E. Orlandini and S.G. Whittington, "Monte Carlo study of the interacting self-avoiding walk model in three dimensions," J. Stat. Phys. 82 (1996), pp. 155-181.
22. D. van Dyk and X.-L. Meng "The art of data augmentation (with discussion)," J. Computational and Graphic Statistics 10 (2001), pp. 1-111.