Journal of Statistical Physics, Vol. 97, Nos. 5/6, 1999

Error Estimation in the Histogram Monte Carlo Method

M. E. J. Newman[1] and R. G. Palmer[2]

Received January 20, 1999; final July 21, 1999

We examine the sources of error in the histogram reweighting method for Monte Carlo data analysis. We demonstrate that, in addition to the standard statistical error which has been studied elsewhere, there are two other sources of error, one arising through correlations in the reweighted samples, and one arising from the finite range of energies sampled by a simulation of finite length. We demonstrate that while the former correction is usually negligible by comparison with statistical fluctuations, the latter may not be, and give criteria for judging the range of validity of histogram extrapolations based on the size of this latter correction.

KEY WORDS: Monte Carlo simulation; histogram method; error estimation.

I. INTRODUCTION

Monte Carlo simulations have a long and interesting history. As a tool for studying physical systems (rather than for performing integrals), they date back at least as far as the pioneering work on neutron diffusion by Enrico Fermi in the 1930s,[(1)] but Monte Carlo methods really came to prominence in the fifties following the calculations on hard-sphere gases and other simple systems performed by Ulam, Metropolis, von Neumann and others using the early digital computers at Aberdeen and Los Alamos.[(2)] In the last three decades, with the availability of ever-increasing amounts of computer power, the Monte Carlo method has become one of the most important tools in the statistical physicist's tool-box.[(3)]

Although the name Monte Carlo covers a multitude of different ideas and techniques, we concentrate in this paper on the simulation of classical

[1] Santa Fe Institute, Santa Fe, New Mexico 87501.
[2] Department of Physics, Duke University, Durham, North Carolina 27708.

0022-4715/99/1200-1011$16.00/0

models in thermal equilibrium. All equilibrium Monte Carlo calculations revolve around the same fundamental idea. One generates a number of states $i = 1 \cdots n$ of the system of interest and measures for each one the total energy E_i and any other quantities of interest X_i, Y_i, etc. Normally all states i are not generated with equal likelihood, but with varying probabilities p_i, a technique known as importance sampling. The best estimate of the thermal average $\langle X \rangle$ of a quantity X is then given by

$$\langle X \rangle = \frac{\sum_i X_i p_i^{-1} \mathrm{e}^{-\beta E_i}}{\sum_i p_i^{-1} \mathrm{e}^{-\beta E_i}} \tag{1}$$

where $\beta = (kT)^{-1}$ is the inverse temperature and k is the Boltzmann constant.[4]

The most common choice by far for the probabilities p_i is to make them proportional to the Boltzmann weight of the corresponding state at the temperature of interest

$$p_i \propto \mathrm{e}^{-\beta E_i} \tag{2}$$

in which case Eq. (1) reduces to a simple average over the measurements X_i. Many other choices have been investigated however, including simple or uniform sampling[5] in which p_i is a constant independent of i, entropic sampling[6] in which p_i is proportional to the reciprocal of the density of states at energy E_i, and $1/k$ sampling[7] in which p_i is proportional to the reciprocal of the integrated density of states. In the present paper we investigate the case in which the states are sampled with probabilities proportion to their Boltzmann weights, but at a temperature T_0 different from the temperature at which we wish to calculate $\langle X \rangle$. In other words, we imagine performing a normal thermal Monte Carlo simulation at a temperature T_0, and then ask for the best estimate of the expectation of X at a different temperature T. Making the replacement $\beta \to \beta_0$ in Eq. (2) and substituting into (1), we obtain

$$\langle X \rangle = \frac{\sum_i X_i \mathrm{e}^{-(\beta - \beta_0) E_i}}{\sum_i \mathrm{e}^{-(\beta - \beta_0) E_i}} \tag{3}$$

This is not a new result. Already in 1972, Valleau and Card[8] pointed out that it is possible in theory to extract a value for $\langle X \rangle$ at any temperature from the results of a single thermal Monte Carlo simulation using an equation of this type. Their results were rediscovered and extended in 1988 by Ferrenberg and Swendsen,[9] who dubbed this technique the "single histogram method." The name is something of a misnomer, since the

method's application does not necessarily involve the construction of any histograms. Ferrenberg and Swendsen's formulation however was in terms of histograms and, as we will see, it is often convenient to represent the method in this way.

Defining the double histogram $H(E, X)$ to be the number of states i sampled for which $E_i = E$ and $X_i = X$, we can rewrite Eq. (3) in terms of sums over the possible values of E and X thus:

$$\langle X \rangle = \frac{\sum_{E,X} XH(E, X)\, e^{-(\beta-\beta_0)E}}{\sum_{E,X} H(E, X)\, e^{-(\beta-\beta_0)E}} \tag{4}$$

If we define a set of weights

$$W(E, X) = H(E, X)\, e^{-(\beta-\beta_0)E} \tag{5}$$

then Eq. (4) can be rewritten as a weighted average over X:

$$\langle X \rangle = \frac{\sum_{E,X} XW(E, X)}{\sum_{E,X} W(E, X)} \tag{6}$$

Note that $W(E, X)$ and $H(E, X)$ become equal when $\beta = \beta_0$. In effect, $W(E, X)$ is an estimate of the value of the histogram $H(E, X)$ at the temperature of interest.

It is possible to write an equation similar to (3) for parameters other than the temperature, allowing us to extrapolate the results of a single simulation to other values of any external field appearing in the Hamiltonian. It is also straightforward to generalize the histogram method to non-Boltzmann sampling schemes. Here however we concentrate on the simple case described above.

In this paper we explore the sources of error in histogram extrapolations. The statistical errors inherent in the method have beep discussed at some length elsewhere,[10] and it is not our intention to reproduce previous results here. We focus instead on two important sources of error which have been neglected in previous studies. In Section II we discuss errors introduced as a result of the finite range of energies sampled in a simulation of finite length, and show that in certain temperature regimes this, and not statistical fluctuation, is the dominant source of error. In Section III we discuss errors introduced by the correlation between fluctuations in the numerator and denominator of Eq. (3). In Section IV we discuss corrections to the normal expression for the statistical errors arising from the previous analysis and show that to leading order these corrections are negligible. In Section V we give our conclusions.

II. FINITE-SAMPLE-SIZE ERRORS

Suppose that we perform a single Monte Carlo simulation at temperature T_0 on some system-of interest, and that this simulation samples n states of the system at intervals of τ_s Monte Carlo steps. We assume in this paper that τ_s is much greater than the correlation time τ of the simulation algorithm used (also measured in Monte Carlo steps) so that the states may be considered to be statistically independent. More generally, if τ_s and τ are comparable, then the variance in a measured quantity is increased by a factor of $1+2\tau/\tau_s$ over its value for uncorrelated samples.[11] All the results given in this paper can be generalized to this case in a straightforward manner; see ref. 10 for a thorough exploration of this issue.

In the limit of an infinite number of independent samples, $n \to \infty$, Eq. (3) is exact and correctly gives the value of $\langle X \rangle$ at all temperatures. In practice, however, n is always finite, and this limits the range over which the extrapolation is valid. In Fig. 1 we show an example of the use of the single histogram method to calculate the internal energy of a two-dimensional

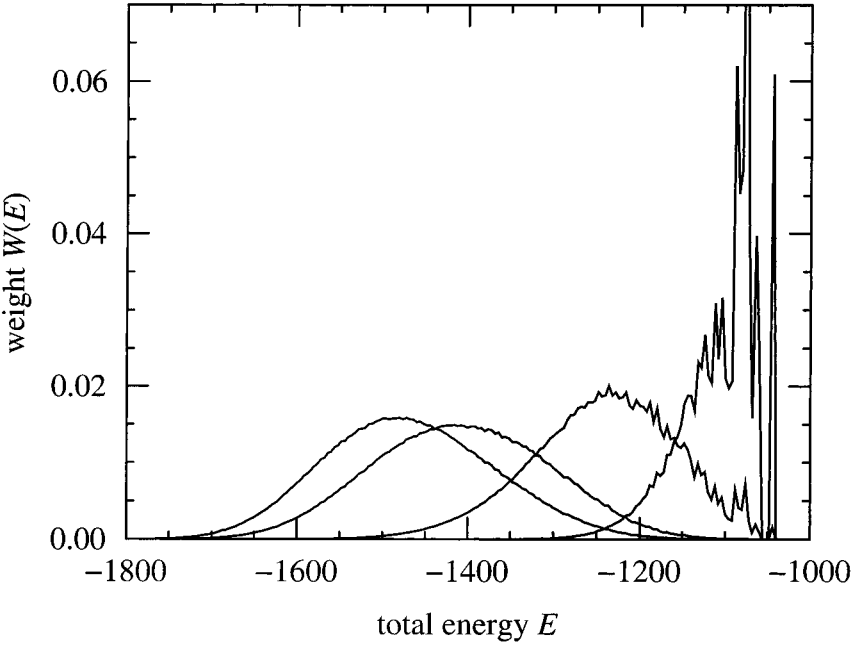

Fig. 1. The weight function $W(E)$ for a 32×32 Ising ferromagnet on a square lattice in two dimensions, calculated at four different temperatures from a single simulation at the critical temperature $T_c = 2.269$ of the infinite system. The curves shown are (left to right) for $T = T_c$, 2.3, 2.4, and 2.6.

Ising model in zero field. The case of the internal energy is particularly simple, since the weight function $W(E, X)$ reduces in this case to a function $W(E)$ of a single variable E, the energy of the states sampled in the simulation. The figure shows the calculated value of this function for a variety of different temperatures at distances increasingly far from the temperature T_0 of the original simulation. For small deviations from T_0 the calculated value of $W(E)$ is a good approximation to the histogram $H(E)$ which would be generated by a simulation performed at temperature T. However as T strays farther from T_0, the value of $W(E)$ becomes an increasingly poor representation of the correct histogram, as can be seen in the figure. The source of this problem is clear: a finite-n Monte Carlo simulation samples energies in only a rather narrow range around the value $U(T_0)$ of the equilibrium internal energy of the system at T_0. Extrapolation of the results to temperatures T for which the true histogram $H(E)$ would possess significant contributions at energies outside this range is therefore guaranteed to give poor results. In the particular case of the internal energy, it is clear that if the highest energy sampled by our simulation is E_+, then no reweighting of our histogram can ever produce an estimate of $U(T) \equiv \langle E \rangle$ greater than E_+, regardless of the true value.

The usual rule of thumb for estimating the range of validity of the extrapolation is to require that the mean of the reweighted distribution $W(E)$, which is just the internal energy $U(T)$, should be less than σ_E away from the mean $U(T_0)$ of the histogram $H(E)$, where σ_E is the standard deviation of $H(E)$. Since σ_E is related to the specific heat C at T_0 according to $C(T_0) = k\beta_0^2\sigma_E^2$, we can also express this condition in terms of $C(T_0)$ as

$$[U(T) - U(T_0)]^2 < kT_0^2 C(T_0) \tag{7}$$

This inequality can be simplified further if we make the derivative approximation

$$U(T) - U(T_0) \simeq (T - T_0)\left.\frac{dU}{dT}\right|_{T_0} = \Delta T\, C(T_0) \tag{8}$$

where $\Delta T \equiv T - T_0$ is the temperature range over which we are extrapolating. Employing this approximation, our condition becomes

$$\left[\frac{\Delta T}{T_0}\right]^2 < \frac{k}{C(T_0)} \tag{9}$$

This condition is intuitively easy to understand and in most cases is a reasonable guide for applying the histogram method. However, as we will

demonstrate, the actual range of validity of the method can deviate arbitrarily far from the value of ΔT given by Eq. (9), depending on the number n of samples generated by the Monte Carlo simulation.

We now construct a more accurate criterion for the extrapolation range. The basic idea is to make an estimate of the energy E_+ above which there are no samples, and then to approximate the error introduced into our extrapolation by assuming that the histogram is accurate up to E_+, and contains no samples thereafter. We do the same for the lower limit E_- of the histogram. A variation on this idea would be to restrict the extrapolation to a range of energies such that some prescribed fraction of the samples in the histogram fall within that range. However, since the tails of the histogram typically decay exponentially or faster, these two approaches give approximately the same results.

Consider the ideal histogram $\overline{H(E)}$, which we define to be the value of the histogram $H(E)$ averaged, bin by bin, over an infinite number of simulations which generate n samples each. We then approximate the histogram resulting from a single simulation by

$$H(E) = \begin{cases} (n/n')\,\overline{H(E)} & \text{if} \quad E_- < E < E_+ \\ 0 & \text{otherwise} \end{cases}$$

The factor n/n', where $n' = \int_{E_-}^{E_+} \overline{H(E)}\,\mathrm{d}E$, is a normalizing factor which ensures that the integral of $H(E)$ over E is correctly equal to n. The values of E_+ and E_- are given by

$$\overline{H(E_\pm)} = a \tag{11}$$

where a is a constant of order unity.

Making this approximation, the error in the extrapolated internal energy can be written as

$$\Delta U = \overline{U(T)} - U(T) = \frac{\int E \mathrm{e}^{(\beta-\beta_0)E}\,\overline{H(E)}\,\mathrm{d}E}{\int \mathrm{e}^{(\beta-\beta_0)E}\,\overline{H(E)}\,\mathrm{d}E} - \frac{\int E \mathrm{e}^{(\beta-\beta_0)E} H(E)\,\mathrm{d}E}{\int \mathrm{e}^{(\beta-\beta_0)E} H(E)\,\mathrm{d}E}$$

$$= \frac{\partial}{\partial\beta} \log \frac{\int_{E_-}^{E_+} \mathrm{e}^{-(\beta-\beta_0)E}\,\overline{H(E)}\,\mathrm{d}E}{\int_{-\infty}^{\infty} \mathrm{e}^{-(\beta-\beta_0)E}\,\overline{H(E)}\,\mathrm{d}E} \tag{12}$$

In order to proceed we make a Gaussian approximation for $\overline{H(E)}$:

$$\overline{H(E)} = \frac{n}{\sqrt{2\pi\sigma_E^2}} \exp\left(-\frac{[E - U(T_0)]^2}{2\sigma_E^2} \right) \tag{13}$$

This assumption is an excellent guide for the behavior of most systems at temperatures well above $T = 0$. For instance, in the Ising system of Fig. 1 it gives $\log \overline{H(E)}$ within a few percent over more than a hundred orders of magnitude of $\overline{H(E)}$.

Using Eq. (13) and another derivative approximation:

$$(\beta - \beta_0)\, \sigma_E^2 = -(\beta - \beta_0) \frac{\mathrm{d}U}{\mathrm{d}\beta}\bigg|_{\beta_0} \simeq U(T_0) - U(T) \tag{14}$$

we complete the square to obtain

$$\overline{H(E)}\, \mathrm{e}^{-(\beta - \beta_0) E} \simeq \frac{n}{\sqrt{2\pi\sigma_E^2}} f(\beta) \exp\left(-\frac{[E - U(T)]^2}{2\sigma_E^2} \right) \tag{15}$$

where

$$f(\beta) = \exp\left(\frac{U^2(T) - U^2(T_0)}{2\sigma_E^2} \right) \tag{16}$$

is a shorthand for all the terms in the exponential which depend on β but not on E. Substituting Eq. (15) into (12) and performing the integral leads to

$$\begin{aligned} \Delta U &= \frac{\partial}{\partial\beta} \log\left[\frac{1}{2} \operatorname{erf}\left(\frac{E - U(T)}{\sqrt{2}\, \sigma_E} \right) \right]_{E_-}^{E_+} \\ &= \sqrt{\frac{2\sigma_E^2}{\pi}} \frac{\exp(-x_+^2) - \exp(-x_-^2)}{\operatorname{erf}(x_+) - \operatorname{erf}(x_-)} \end{aligned} \tag{17}$$

where $\operatorname{erf}(x) = (2/\sqrt{\pi}) \int_0^x \mathrm{e}^{-t^2}\, \mathrm{d}t$ is the Gaussian error function, and

$$x_\pm \equiv \frac{E_\pm - U(T)}{\sqrt{2}\, \sigma_E} = \pm \sqrt{\log \frac{n}{\sqrt{2\pi}\, \sigma_E a}} - \frac{\sigma_E \Delta T}{\sqrt{2}\, k T_0^2} \tag{18}$$

using Eqs. (8), (11) and (13).

Between them, Eqs. (17) and (18) give us an estimate of the deviation of the extrapolation of U from its true value as a function of the number of samples n and the temperature range ΔT over which we extrapolate.

As a test of this calculation we have plotted in Fig. 2 the value of ΔU measured in simulations of a 100×100 Ising model on a square lattice in two dimensions. The data points with error bars show the difference between the true internal energy (obtained from further independent simulations) and those calculated via Eq. (3) from simulations with $n = 100$ samples at temperature $T_0 = 2.269$ (the critical temperature of the infinite system). These points are averaged over 1000 repetitions of the simulation at T_c. The solid line is from Eqs. (17) and (18) with the constant a chosen so as to best fit the data. As the figure shows, the agreement between the two is good.

In a typical Monte Curio calculation we want to know the range of temperature ΔT over which we can extrapolate from a single histogram to a given degree of accuracy ΔU as a function of the sample size n. In the

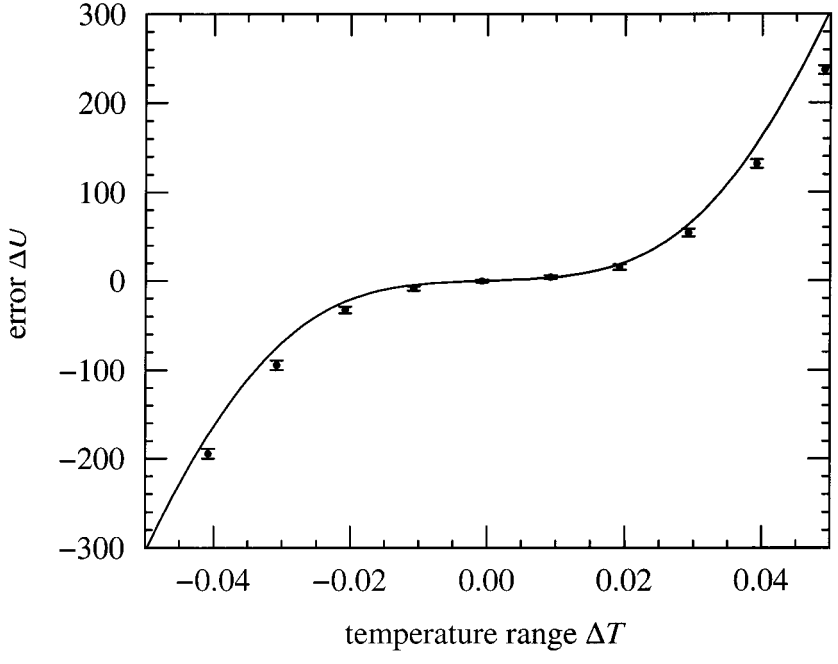

Fig. 2. The difference ΔU between the true internal energy of a 100×100 Ising ferromagnet and an extrapolation using Eq. (3) of the same quantity from simulations with $n = 100$ samples performed at a single temperature $T_0 = 2.269$. The line is a fit using Eqs. (17) and (18). Energies are in units of the coupling constant J, and may be compared to $U(T_0) = -1.4 \times 10^4$.

regime where $U(T)$ approaches either of the limits E_- or E_+, one or other of the terms on the top and bottom of Eq. (17) becomes a constant (either zero or one) and the variation in ΔU resides entirely in the remaining terms. In this case a line of constant ΔU is also a line of constant x_+ or x_- (for ΔT positive or negative respectively) which means that

$$\pm \sqrt{\log \frac{n}{\sqrt{2\pi}\,\sigma_E a}} - \frac{\sigma_E \Delta T}{\sqrt{2}\,kT_0^2} = b \tag{19}$$

with the value of the constant b depending on the size of error ΔU we are willing to live with. Thus, for given ΔU, the temperature range ΔT over which the extrapolation is valid increases at most logarithmically with increasing sample size n.

In Fig. 3 we demonstrate this formula for the 100×100 two-dimensional Ising model. The inset shows extrapolations from the critical temperature

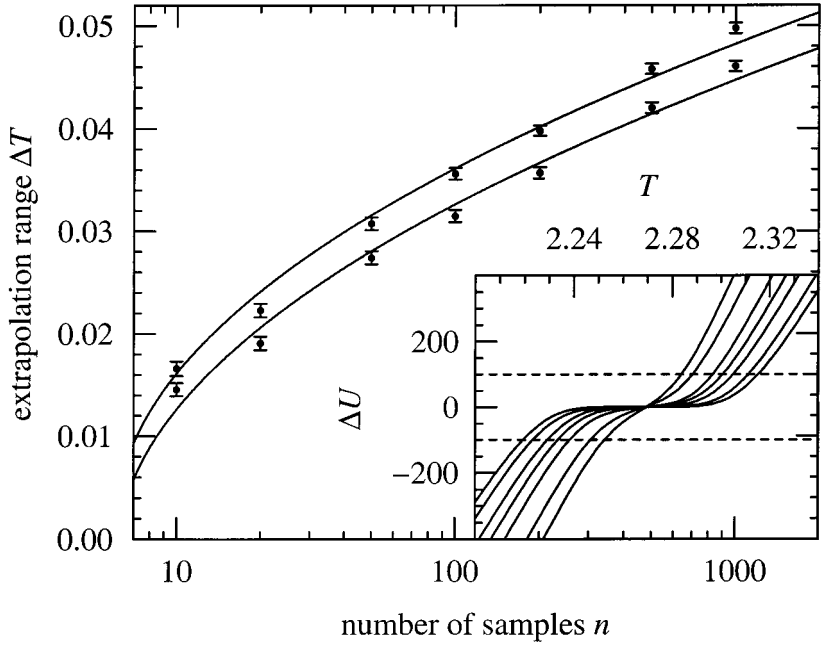

Fig. 3. Inset: the difference between the true and extrapolated internal energies of a 100×100 Ising ferromagnet for a variety of different sample sizes n. Main figure: the range ΔT over which the extrapolation is accurate to ± 100, as a function of n. The points are the values from the simulations shown in the inset and the two solid lines are Eq. (19), taking the $+$ and $-$ signs separately. The upper curve and points are for positive ΔT, the lower ones for negative ΔT.

of the infinite system for sample sizes $n = 10$, 20, 50, 100, 200, 500, and 1000, using Eq. (3). The errors in these results are comparable to the widths of the lines. The dashed lines show an arbitrarily-chosen deviation of $\Delta U = \pm 100$ from the true value as our limit of acceptable accuracy—a relative error of about 0.7%. The intersections of the solid curves and dashed lines give the ranges ΔT over which simulations with different n give acceptable results. The main figure shows these ranges as points with error bars, the upper points corresponding to values $\Delta T > 0$ (i.e., extrapolation above T_0), the lower ones to $\Delta T < 0$. The solid lines are Eq. (19) with the constants a and b chosen by a least squares fit to the data. As the figure shows, simulation and theory are in good agreement.

As an example of the use of Eq. (19), consider the results of Münger and Novotny[12] who performed an extensive numerical study of the accuracy of the single histogram extrapolation method for the case of the $q = 3$ Potts ferromagnet in two dimensions. They concluded that the specific heat predicted by the method shows systematic deviations from the true specific heat, and presented evidence indicating that the size of these deviations decrease with increasing n. As we now show, our Eqs. (9) and (19) are completely in agreement with this finding.

Münger and Novotny performed simulations on 8×8 square systems with $n = 50$ and $n = 500$, at three temperatures $0.8T_c$, T_c, and $1.3T_c$, where $T_c = 0.9950$ is the critical temperature of the infinite system. They attempted to calculate the position of the maximum of the specific heat of the system by histogram extrapolation. The maximum in a system of this size occurs at about $1.04T_c$. Column 3 of Table I shows the values of ΔT calculated using the simple rule of thumb, Eq. (9). As the values show, whilst the required extrapolation range for the simulation at T_c is comfortably within that allowed by the criterion, the ranges for the other two simulations are not nearly enough to reach the temperature of the specific heat peak. (Column 2 of the same table shows the value of ΔT required to reach the peak.) So, it should not come as a surprise that extrapolations

Table I. Values of the Extrapolation range ΔT from Eqs. (9) and (19) for the Simulations Performed by Münger and Novotny[12]

Temperature	Required ΔT	Simple estimate	$n = 50$	$n = 500$
0.8	0.24	0.13	−0.23	0.35
1.0	0.04	0.08	−0.29	0.07
1.3	0.26	0.19	−0.57	0.33

using the results of these two simulations give very poor estimates of the position of the peak. As we have argued however, if n is large, Eq. (9) underestimates ΔT, and so it may be possible that for the particular values of n used in these simulations the extrapolation will work. We now show, as Münger and Novotny also found, that this is not the case. Even making the most generous estimate possible of ΔT, the temperature range of the extrapolation is just too large to allow us to calculate the position of the peak accurately using the histogram method.

Let us then apply Eq. (19) to the calculations of ref. 12. Before we can do this, we have to choose values for the constants a and b. The most generous choice we can make for b is to set $b = 0$, which implies that the histogram extrapolation is accurate all the way out to $E_{\pm}$. Experience suggests that this is probably over-generous by about a factor of 2 in ΔT, but let us go with this value for the moment and see where it leads. With this value of b, Eq. (19) can be rewritten in the form

$$\left[\frac{\Delta T}{T_0}\right]^2 < \frac{k}{C(T_0)} \log \frac{n^2}{2\pi\sigma_E^2 a^2} \tag{20}$$

Comparing with Eq. (9), we see that the net result is to multiply our simple rule-of-thumb estimate of ΔT by a factor depending logarithmically on n. Münger and Novotny noted in their paper that the convergence of the specific heat peak to its correct position with increasing n appeared to be very slow, and this equation indicates that the convergence is in fact logarithmic.

The value of a is harder to judge. In Eq. (11) we defined a to be the number of samples in a bin which just suffices to make an estimate of the value of the weight function $W(E)$ at that energy. Clearly this number should be greater than one, but probably not much greater. We have chosen a value of $a = 10$ for our calculations, which, if we assume Poisson fluctuations in the number of samples per bin, implies we know $W(E)$ to within about 30%. As with our choice of value for b, this is quite a generous estimate. Better accuracy is probably desirable for most real Monte Carlo calculations.

The fourth and fifth columns of Table I give the values of ΔT calculated from Eq. (20) for the two values of n used in the simulations. The values for $n = 50$ are in all cases negative. This is an indication that the Gaussian approximation we have used in the calculation has broken down because there is less than one sample per bin on average at energy $U(T)$ for the temperature T which we want to extrapolate to. One should definitely not expect these calculations to give a good estimate of the position of the specific heat peak. In the case where $n = 500$, the values are

larger and in fact just sufficient to encompass the desired temperature range. However, given that we have here made the most generous possible estimate of ΔT—the true values are probably a factor of 2 or more smaller—it would not be prudent to use these calculations to find the specific heat peak either. It should come as no surprise therefore that Münger and Novotny found that the histogram method gave a poor estimate of the position of the peak in this case. The values in Table I are a strong indication that larger values of n are needed to give reliable results.

Münger and Novotny deliberately performed simulations with small values of n in order to investigate the inaccuracies of the histogram method. However, in normal use, the method is applied to simulations with large n, and in the region close to T_0 where the deviation ΔU is small. We can characterize this regime as one in which $|x_\pm| \gg 1$, in which case the value of the denominator in Eq. (17) is close to 2 and the primary variation in ΔU comes from the Gaussians in the numerator:

$$\Delta U \simeq \sqrt{\frac{\sigma_E^2}{2\pi}} \left[\exp(-x_+^2) - \exp(-x_-^2) \right] \tag{21}$$

Since E_+ and E_- are symmetrically distributed about $U(T_0)$, we have $x_+(T_0) = -x_-(T_0)$, and the two terms cancel to give $\Delta U = 0$ at $T = T_0$, as expected. The leading term in the expansion of ΔU about this point is linear in ΔT with coefficient

$$\left. \frac{\partial \Delta U}{\partial T} \right|_{T_0} = \frac{2a\beta_0^2 \sigma_E^3}{n} \sqrt{\log \frac{n^2}{2\pi \sigma_E^2 a^2}} \tag{22}$$

Thus ΔU tends to zero roughly as $1/n$ to leading order, and the higher order terms vanish faster than this. As we will see in Section IV, the statistical errors in extrapolated quantities fall off in the normal $1/\sqrt{n}$ fashion, so that in the region close to T_0, finite sample size errors always become negligible for sufficiently large n.

On the other hand, when we get far away from T_0, the extrapolated value of U becomes roughly equal to E_+ or E_- (depending on the direction in which we extrapolate) and hence approximately independent of n, since $E_\pm$ only varies slowly with n. Thus the error ΔU is approximately n-independent in this regime and dominates over statistical errors for sufficiently large n. The point of crossover between the two regimes is given by Eq. (19).

A similar argument can be made for the extrapolation of quantities other than the energy. The limiting extrapolated values of any quantity Y

are set by the values $Y_{\pm}$ corresponding to the highest and lowest energies sampled in the simulation, and since these energies are approximately n-independent, so normally will $Y_{\pm}$ be. Thus Eq. (19) tells us for any quantity Y the point of crossover at which errors due to the finite number of samples in the histogram become the dominant source of inaccuracy in the histogram method.

III. DISTRIBUTION ERRORS

There is another source of systematic error in the estimates given by the single histogram method which has not, to our knowledge, been remarked upon before. Even ignoring the corrections discussed in the last section, which were due to the imperfect sampling of the histogram $H(E)$, Eq. (3) is not in fact a correct expression for the best estimate of $\langle X \rangle$ for any finite n. To understand this, consider again the hypothetical situation in which we perform a large number N of simulations of the system of interest, each one generating n statistically independent samples drawn from the Boltzmann distribution at T_0. For each one we calculate an estimate

$$\langle X \rangle_i = \frac{\sum_j X_{ij} \, e^{-(\beta - \beta_0) E_{ij}}}{\sum_j e^{-(\beta - \beta_0) E_{ij}}} = \frac{P_i}{Q_i} \tag{23}$$

where $i = 1 \cdots N$ labels the different simulations and X_{ij} is the value of X in the jth state sampled by the ith simulation. The new quantities P and Q will provide a convenient shorthand for the numerator and denominator of this equation.

Now we want to compute the best estimate of $\langle X \rangle$ over all N simulations. Since the samples in each simulation were drawn from the same distribution, we can just as well regard them all as being one large set of samples of size nN drawn from a single simulation, in which case it is clear that in the limit of large N the correct answer for $\langle X \rangle$ is

$$\langle X \rangle = \frac{\sum_{ij} X_{ij} \, e^{-(\beta - \beta_0) E_{ij}}}{\sum_{ij} e^{-(\beta - \beta_0) E_{ij}}} = \frac{\bar{P}}{\bar{Q}} \tag{24}$$

where $\bar{P}$ and $\bar{Q}$ indicate the averages of P_i and Q_i over all N simulations. (We use the barred notation to avoid confusion with the notation $\langle X \rangle$ for thermal expectation values.) This equation indicates that the best estimate of $\langle X \rangle$ is calculated by separately averaging the numerator and denominator of Eq. (23) over our many simulations. In practice, one does

not perform many simulations, one performs only one simulation with finite n and then calculates the ratio P/Q for that one simulation. The mean value of this ratio however is not the same as the ratio of the means, Eq. (24), which gives the correct answer. This difference leads to a systematic error in the predictions of the single histogram method for finite sample sizes. In this section we calculate the size of this error.

Consider the double Taylor expansion of the quantity P/Q around $\bar{P}/\bar{Q}$:

$$\frac{P}{Q} = \frac{\bar{P}}{\bar{Q}} + (P-\bar{P})\frac{1}{\bar{Q}} - (Q-\bar{Q})\frac{\bar{P}}{\bar{Q}^2} + (Q-\bar{Q})^2\frac{\bar{P}}{\bar{Q}^3} - (P-\bar{P})(Q-\bar{Q})\frac{1}{\bar{Q}^2} + \cdots \tag{25}$$

Taking the average of both sides over many repetitions of the simulation, the linear terms vanish and to leading order we are left with

$$\overline{P/Q} = \frac{\bar{P}}{\bar{Q}}\left[1 + \frac{\sigma_Q^2}{\bar{Q}^2} - \frac{\mathrm{cov}(P,Q)}{\bar{P}\bar{Q}}\right] \tag{26}$$

where σ_Q^2 is the variance of Q over simulations i and $\mathrm{cov}(P,Q)$ is the covariance of P and Q. Thus the mean value of the quantity P/Q, which is the quantity measured in our Monte Carlo calculations, differs from the true value of $\langle X\rangle = \bar{P}/\bar{Q}$ by the factor enclosed in the square brackets $[\cdots]$. One should take this factor into account in order to calculate the extrapolation of a quantity correctly.

Given that in a typical situation we only perform one simulation of our system, what is the best estimate we can make of this factor from our Monte Carlo results? Clearly the best estimates of $\bar{P}$ and $\bar{Q}$ are simply the values of P and Q measured in the simulation: $\bar{P} = P$, $\bar{Q} = Q$. The best estimates of the variance and covariance terms are

$$\sigma_Q^2 = \frac{1}{n-1}\left\{\sum_j \mathrm{e}^{-2(\beta-\beta_0)E_j} - \left[\sum_j \mathrm{e}^{-(\beta-\beta_0)E_j}\right]^2\right\} \tag{27}$$

and

$$\mathrm{cov}(P,Q) = \frac{1}{n-1}\left\{\sum_j X_j \mathrm{e}^{-2(\beta-\beta_0)E_j} - \sum_j X_j \mathrm{e}^{-(\beta-\beta_0)E_j}\sum_j \mathrm{e}^{-(\beta-\beta_0)E_j}\right\} \tag{28}$$

Substituting these into Eq. (26) we see that the correction term scales as $1/n$ with sample size. But, as shown below, statistical errors scale as $1/\sqrt{n}$ and therefore dominate for large n. Thus it should be safe to ignore errors of the type described by Eq. (26) for simulations of sufficient length.

IV. STATISTICAL ERRORS

The third and final source of error which we consider is statistical fluctuation in the extrapolation due to the essential random nature of a Monte Carlo simulation. We can calculate the variance $\sigma^2_{\overline{P/Q}}$ of the quantity $\overline{P/Q}$ by a technique similar to that used to derive Eq. (26); we perform a Taylor expansion of $\overline{P^2/Q^2}$ about $\bar{P}/\bar{Q}$ and take the average over many simulations. Then we calculate the variance as $\sigma^2_{\overline{P/Q}} = \overline{P^2/Q^2} - \overline{P/Q}^2$. The variance σ^2_X of the best estimate of $\langle X \rangle$ is then $\sigma^2_{\overline{P/Q}}$ times the square of the correction factor in Eq. (26). To leading order this gives

$$\frac{\sigma_X^2}{\langle X \rangle^2} = \frac{\sigma_P^2}{\bar{P}^2} + \frac{\sigma_Q^2}{\bar{Q}^2} - 2\frac{\mathrm{cov}(P, Q)}{\bar{P}\bar{Q}} \tag{29}$$

This expression is identical to that given by Ferrenberg *et al.*,[10] for the error on the uncorrected estimate $\overline{P/Q}$.

Using Eqs. (27) and (28), along with the obvious extension

$$\sigma_P^2 = \frac{1}{n-1}\left\{\sum_j X_j^2 \mathrm{e}^{-2(\beta-\beta_0)E_j} - \left[\sum_j X_j \mathrm{e}^{-(\beta-\beta_0)E_j}\right]^2\right\} \tag{30}$$

it is clear that σ_X^2 scales as $1/n$, and hence that σ_X scales as $1/\sqrt{n}$, as claimed earlier. This is a slower scaling than the $1/n$ of the previous section, but still much better than the approximately constant value of the finite sample size error of Section II for large extrapolation range ΔT. This means that we must use an equation such as (19) to decide which of these two latter sources of error is the dominant one under given circumstances.

V. CONCLUSIONS

In this paper we have examined in detail the sources of error in the Monte Carlo extrapolation method known as the single histogram method. We have discussed three sources of error: finite sample size errors, systematic errors due to the approximations made in the calculation of the extrapolation, and finally statistical errors. The first two of these have not

to our knowledge been discussed previously, and in particular we find that the finite sample size errors are, under commonly encountered conditions, significantly larger than either of the other sources of error.

ACKNOWLEDGMENTS

The authors would like to thank Gerard Barkema, Jon Machta and Catherine Macken for useful discussions. This research was funded in part by the Santa Fe Institute and DARPA under grant number ONR N00014-95-1-0975.

REFERENCES

1. E. Segrè, *From X-Rays to Quarks* (Freeman, San Francisco, 1980).
2. N. Metropolis, J. Howlett, and G.-C. Rota, *A History of Computing in the Twentieth Century* (Academic Press, New York, 1980).
3. M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, Oxford, 1999).
4. In some cases, particularly in systems which display symmetry- or ergodicity-breaking, we may not in fact wish to calculate an average over all states in this way (see, for example, R. G. Palmer, *Adv. Phys.* **31**:669 (1982)). For the purposes of this paper however, we assume that we are working with ergodic systems for which expectations of the form (1) are physically meaningful.
5. K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics* (Springer-Verlag, Berlin, 1992).
6. J. Lee, *Phys. Rev. Lett.* **71**:211 (1993).
7. B. Hesselbo and R. B. Stincdcombe, *Phys. Rev. Lett.* **74**:2151 (1995).
8. J. P. Valleau and D. N. Card, *J. Chem. Phys.* **57**:5457 (1972).
9. A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **61**:2635 (1988).
10. A. M. Ferrenberg, D. P. Landau, and R. H. Swendsen, *Phys. Rev. E* **51**:5092 (1995).
11. H. Müller-Krumbhaar and K. Binder, *J. Stat. Phys.* **8**:1 (1973).
12. E. P. Münger and M. A. Novotny, *Phys. Rev. B* **43**:5773 (1991).