SLAC–PUB–4388
August, 1987
T

ZEROING IN ON $SU(3)$*

MAREK KARLINER AND STEPHEN R. SHARPE

Stanford Linear Accelerator Center
Stanford University, Stanford, California, 94305

Y. F. CHANG

Center for Studies of Nonlinear Dynamics
La Jolla Institute
10280 N. Torrey Pines Road
La Jolla, CA, 92037

ABSTRACT

We present an improved numerical method for calculating the density of states for lattice field theories. We use it to study the $SU(3)$ pure gauge theory at both zero and finite temperature. We also compute strong and weak coupling expansions for the density of states and find excellent agreement with our data. Using a specially developed algorithm for solving high order polynomials, we find the zeroes of the partition function. For lattices with $L_t = 2$, we test the finite size scaling prediction for the rounding of the transition by following the motion of these zeroes for $L_s = 6$, 8, 10, and 12. We find that the correlation length exponent is $1/\nu = 3.02 \pm 0.05$, in excellent agreement with the value $d = 3$ expected for a first order deconfinement transition.

Submitted to *Nuclear Physics B*

⋆ Work supported by the Department of Energy, contract DE–AC03–76SF00515.

1. INTRODUCTION

It is an old idea that one should study the thermodynamic properties of a system by following the movement of the zeroes of its partition function.[1-5] Recently, this approach has been given new impetus by the suggestion of a practical method to numerically calculate the density of states.[6,7] Given the density of states, one can reconstruct the partition function and find its zeroes. We find this approach, which we refer to as the spectral density method, to be very promising. It provides a complementary way of studying phase transitions, allowing a determination of their order and other global properties on a competitive basis with more established methods. A particularly interesting potential application is QCD with two light flavors of quark, for which there are conflicting results obtained using traditional Monte-Carlo approaches.[8]

However, the work in Refs. 6,7 considers either discrete systems, or $SU(2)$ on very small lattices. It is important to test out the method on a more complicated theory and on larger lattices. We have chosen to study pure $SU(3)$ gauge theory, on both symmetric L^4 and asymmetric $L_s^3 \times L_t$ lattices. For the symmetric, zero temperature lattices, we expect no phase transitions. For the asymmetric, finite temperature lattices, we expect to see a first order deconfining transition.

The method suggested in Refs. 6,7 involves performing a random walk through configuration space, collecting the distribution of values of the average plaquette as a histogram. Such an approach is severely hampered by systematic and statistical errors, which need to be minimized in order to study $SU(3)$ on larger lattices. To that effect, we introduce various improvements. The most important is the use of a guided random walk. By a sequential tuning of the guiding, or "weighting", function we are able to reduce the statistical errors by many orders of magnitude. This change also greatly reduces the systematic errors associated with hysteresis. The kind of simulation we are dealing with requires the resources of a supercomputer, which are best utilized if the code is vectorized. In Ref. 6,7 vectorization

was obtained through replication of the lattice. We are able to vectorize a substantial portion of the algorithm without such replication. This is essential in order to study larger lattices, for which replication is not feasible due to memory limitations. These techniques allow us to pursue our study of pure gauge $SU(3)$ with relatively modest amounts of computer time. In total we have consumed roughly 50 hours of CRAY-XMP time.

As a check on our results we have calculated the strong and the weak coupling expansions for the density of states and compared them with our data. The agreement is excellent. A further check is provided by a comparison of our results with conventional Monte-Carlo results on a 6^4 lattice.

Having calculated the density of states as a histogram, we know the partition function as a polynomial in a variable related to the coupling constant $\beta = 6/g^2$. By solving the polynomial we can find the zeroes of the partition function in the complex β plane. This is the most challenging part of the numerical calculation, for the polynomials are up to 3000-th order, with a ratio of largest to smallest coefficients ranging up to 10^{4400}, i.e. 4400 orders of magnitude. This ratio grows exponentially with the volume. We have had to develop a special purpose code to solve this problem.

As the volume of the system tends to infinity, the zeroes pinch the real β axis, if there is a phase transition.[1] Exactly how they do this depends on the order of the transition. In particular, one can extract the correlation length exponent from the motion of the zeroes closest to the real axis.[4] Thus one has a simple quantitative method of calculating critical exponents. This is particularly advantageous for first order transitions, where the zeroes give a quantitative measure of the rounding of the transition.

We have tested this by considering lattices of size $L_s^3 \times L_t$, with $L_s = 6$, 8, 10 and 12, and $L_t = 2$. We expect to see a first order deconfinement transition, with the critical exponent ν given by $1/\nu = d = 3$. We find excellent agreement with the finite size scaling formula, with the result $1/\nu = 3.02 \pm 0.05$. Although

previous authors have checked the finite size scaling in the position of the critical β,[9] which is not universal, our result is, to our knowledge, the first test of finite size scaling for the rounding of the transition.

For lattices of dimension L^4 we do not expect to see a transition as L increases. Instead there should be a crossover region. This means that there will be zeroes close to the real β axis, but at a location which does not vary with the volume. Such zeroes were investigated long ago in Ref. 2. With relatively low statistics we have examined these zeroes for L=2, 4 and 6. We find results inconsistent with a first order transition, but our data are too scant to rule out a higher order transition.

In the particular implementation we use, the method lies somewhere between traditional Monte-Carlo using the canonical ensemble at fixed β, and microcanonical methods. We would like to assess the relative strengths and weaknesses of the various methods. It is clear from our investigations that the spectral density method suffers from the usual problems of numerical simulations – i.e. hysteresis, a possible lack of ergodicity and critical slowing down.* Thus the comparison between methods is best done case by case. We argue below that the spectral density method is particularly favored for a quantitative study of finite size scaling, especially in weak first order, or higher order transitions, at finite β. For identifying a clear first order transition, or for studying the beta-function for large β,* traditional methods are to be preferred.

The rest of this paper is organized as follows. In section 2 we review the method of Refs. 6,7 and explain our improvements. We give an extensive discussion of systematic and statistical errors. In section 3 we discuss the finite size scaling behavior of zeroes, with particular emphasis on first order transitions. Section 4 contains the strong and weak coupling expansions, and the comparisons of these with our data. We present our results in section 5, and give our conclusions, as well as our view of the outlook for the method, in the final section.

---

* Here we disagree with Ref. 7.

We include two appendices. Appendix A provides a justification of the method, explains how we optimize it, and gives the details of our estimates of statistical errors. Appendix B contains an outline of the algorithm we use to find the zeroes of the polynomials.

2. METHOD

The fundamental quantity of interest is the partition function:

$$\begin{aligned} Z &= \int [dU] e^{V\beta\epsilon} \\ V\epsilon \equiv E &= \frac{1}{3}\sum_{pl} \mathrm{Re}\,\mathrm{Tr}\, U_{pl} \end{aligned} \tag{2.1}$$

Here "pl" stands for plaquette, and U_{pl} is the usual product of link matrices around the plaquette. Throughout this article we use V to denote the number of plaquettes on a four-dimensional lattice:

$$V = 6L_s^3 L_t. \tag{2.2}$$

where L_s and L_t are the linear dimensions of the lattice, in the space and time direction, respectively. We refer to V as the volume. The variable ϵ is the average plaquette (normalized to 1); it varies in the range $-1/3 \leq \epsilon \leq 1$. We refer to E as the energy, and to ϵ as the energy density.*

Pursuing further the obvious analogy with statistical mechanics, we define the entropy density $s(\epsilon)$ in terms of the density of states $N(\epsilon)$:

$$e^{Vs(\epsilon)} \equiv N(\epsilon) = \int [dU] \delta(\epsilon - \frac{\sum_{pl} \mathrm{Re}\,\mathrm{Tr}\, U_{pl}}{3V}) \tag{2.3}$$

⋆ A somewhat more frequently used notation is $Z = \int [dU] \exp(-\beta S)$ where $S = 1 - E$. The two are of course completely equivalent. The reader should therefore keep in mind that despite the plus sign in the exponential we have the usual Boltzmann weight.

We can rewrite the partition function in terms of the entropy density as

$$Z(\beta) = \int_{-1/3}^{1} d\epsilon N(\epsilon)\, e^{V\beta\epsilon} = \int_{-1/3}^{1} d\epsilon\, e^{V(s+\beta\epsilon)} \tag{2.4}$$

Thus if we know $s(\epsilon)$ for all ϵ, we can reconstruct Z and all its derivatives, for example

$$\begin{aligned} \langle\epsilon\rangle &= Z^{-1}\int [dU] e^{V\beta\epsilon}\epsilon = \frac{1}{V}\frac{d\ln Z}{d\beta} \\ C_v &= \beta^2 \frac{d\langle\epsilon\rangle}{d\beta} = \frac{\beta^2}{V}\frac{d^2\ln Z}{d\beta^2}. \end{aligned} \tag{2.5}$$

Eq. (2.4) shows that $Z(\beta)$ is just an integral transform of $s(\epsilon)$. In finite volume, the range of integration is finite, so $Z(\beta)$ is well defined for arbitrary β. Therefore, knowledge of $s(\epsilon)$ allows us to look for zeroes of Z at complex β.

In general s is a function of both ϵ and the volume, but as V increases s tends to a universal function of ϵ. We discuss the theoretical expectations for s in the following section. Here we point out only that for a given β, as $V \to \infty$, the integral over ϵ becomes dominated by a single saddle point (except for a first order transition at critical β). Thus our usages of "energy density" for ϵ rather than for $\langle\epsilon\rangle$, and of "entropy density" for $s(\epsilon)$ rather than for $s(\langle\epsilon\rangle)$, become standard as $V \to \infty$. Following the same convention, we refer to $f = s + \beta\epsilon$ as the free energy density, and $F = Vf$ as the free energy.

Refs. 6,7 suggest a numerical method for determining $s(\epsilon)$. The procedure starts by dividing the energy range up into many bins, which are grouped into overlapping sets of bins. The idea is to determine the relative number of states in adjacent bins. This is done by first bringing the configuration to an energy within the set in question. Next, random changes are made to the links; these are accepted if the resulting energy lies within the set, and rejected otherwise. Whether the change is accepted or rejected, the final energy is added to a histogram. After some number of such "events" within a set, the ratio of numbers in adjacent bins

gives an estimate of the relative number of states in the two bins. If adjacent sets of bins are chosen to overlap, these ratios can be extended over the entire energy range. Clearly, in the end we know the density of states only up to an overall factor, and thus the entropy is known up to an additive constant. Such a constant is irrelevant to physical observables, and also does not effect the zeroes of the partition function.

In Refs. 6,7 it is stressed that it is essential to record an event even if the change has been rejected. This is clearly a very important point, but the reasoning behind it is not provided in Refs. 6,7. We therefore feel is worthwhile to present here a simple heuristic argument showing that it is the correct prescription. A more rigorous argument is given in Appendix A. First consider a simple example: a set of three bins each with the same density of states. Let the random walk move either to the left or to the right by one bin. We want to show that a uniform probability distribution, $P(i) = 1/3 \quad i = 1, 2, 3$, is maintained if all events are recorded. This is true for the middle bin, since the new probability is given by $P'(2) = 1/2\, P(1) + 1/2\, P(3) = 1/3$. For the edge bins, one of these two terms is missing, but it is replaced by jumps outside the set which are rejected, yet recorded. For example, $P'(1) = 1/2\, P(1) + 1/2\, P(2) = 1/3$.

More generally, consider a "gedanken simulation" in which the system is allowed to roam freely through configuration space, with no constraint placed on the energies. Every event would be recorded, and if we waited long enough, the resulting histogram would give us the relative density of states for all energies. To obtain the relative distribution of states in a given set only, we would simply delete all the events in which the system was outside the set. A typical sequence would consist of a string of events inside the set, then a string of events outside, then inside, etc. Consider an "inside" string of events. It would end with the system going outside through an edge bin. The crucial observation is that, since E changes in small steps, after a string of outside events the system would return to the original set via the same edge bin through which it had left. Thus, when we discard the outside strings, we are left with a sequence in which every change

which would lead outside is, instead, followed by an event in the same edge bin. This sequence is almost the same as that we obtain in the actual simulation by rejecting changes which jump outside, though including them in the histogram. It is not exactly the same because in the gedanken simulation one does not enter the set at the same configuration from which one left. But, on average this does not matter, as demonstrated by the argument in Appendix A.

In the practical implementation of the method, one adjusts the number of bins so that as V changes the bin size remains constant in terms of the total energy E. In terms of ϵ the bin size is $\mathcal{O}(1/V)$.

This method, as advocated in Refs. 6,7, though straightforward, is unsatisfactory. In nearly all sets, the density of states is a rapidly varying exponential function of energy. Thus most of the events fall into one of the edge bins, with the other edge bin containing substantially fewer events. This is bad for several reasons. First, the errors in the entropy are not uniform, which could lead to systematic errors in quantities derived from the entropy. Second, and more important, there is an enormous waste of events. All physically interesting quantities depend on the entropy density in a number of adjacent sets. The error propagated across a number of sets is dominated by the errors on the results in the sparsely populated bins. The majority of events, which provide a very accurate determination of the results in certain bins, do not help reduce this error. The third problem is that a rapidly varying density of states within a set severely limits the maximum energy range (or the number of bins) that one can allow a given set to cover. The configuration space may well have energy barriers which separate configurations with the same energy. When the sets are narrow, one can never bypass such barriers.

It is easy to see why one expects a rapid variation of the density of states in a typical set. For $V \to \infty$, each energy ϵ is associated with a value of β by the saddle point equation:

$$s'(\epsilon) \equiv \left.\frac{ds(\tilde{\epsilon})}{d\tilde{\epsilon}}\right|_{\tilde{\epsilon}=\epsilon} = -\beta. \tag{2.6}$$

Thus the derivative of s is $-\beta$, and the local density of states is

$$N(\epsilon) \propto e^{-V\beta\epsilon} = e^{-\beta E}$$

For finite volume there are small corrections, of $O(1/V)$, to this behavior. If one uses, for example, 7 bins, each 1 unit of E wide, then the ratio of events in the most populated bin to that in the least populated bin is $\approx e^{-6\beta}$. The region of interest for $SU(3)$ finite temperature phase transitions turns out to be $\beta > 5.0$, so this ratio is very small. In fact, it becomes ever smaller as one approaches the continuum limit at $\beta = \infty$.

The problem, then, is that the events are distributed highly non-uniformly. Our solution is to measure not $N(\epsilon)$, but rather

$$N(\epsilon)\times W(\epsilon) = \exp\{V[s(\epsilon) - s_W(\epsilon)]\}, \tag{2.7}$$

where s_W is a weighting function. Clearly if $s_W(\epsilon)=s(\epsilon)$ then the weighted density of states will be independent of ϵ, all bins will be equally populated, the statistical error will be uniform, and the error propagated over a number of sets will be as small as possible. We introduce this weight in the standard way: a Metropolis accept/reject step. Each change in a link is accepted with probability $\min(1, e^{-V(s_W(\epsilon')-s_W(\epsilon))})$, where ϵ' is the new energy, ϵ the original energy. This accept/reject step is done in addition to the original accept/reject which confines the energies to the set. As before, one needs to record the accepted changes as well as those events where the trial change would have taken the energy outside the set. The heuristic justification of this procedure goes through unchanged with the addition of weighting, and the argument in appendix A covers weighting also.

The price of adding the Metropolis step is that some fraction of otherwise valid changes are rejected. We find that in the range of interest, the acceptance ratio from the weighting criterion alone is 20-30%. This increases our statistical errors by ~ 2. This should be compared to the increase in errors when using no

weighting. In the example mentioned above, a rough estimate of this increase is $\sim \sqrt{\exp(6\beta)/7} \approx 10^6$. A more careful estimate of the merits of weighting is given below, but it is clear that weighting gives an enormous improvement. Indeed, without it this study would not have been feasible. For $SU(3)$ the crossover occurs at much higher β than for $SU(2)$ and thus the improvement for $SU(3)$ is even more crucial than for $SU(2)$.

In order to determine the weighting function $s_W(\epsilon)$, we bootstrap ourselves up from a state of ignorance. First we determine $s(\epsilon)$ on a 2^4 lattice with $s_w = 0$, with small bins and only 4 bins per set. We then smooth the result, with cubic spline fits, and use this as a weight for runs with wider bins, more bins per set, and higher statistics. We get better and better approximations to $s(\epsilon)$, which we plug back in as the new s_W. When we move to a larger lattice, we use the results from the next smallest lattice as a starting weight. Typically we use bins of size $0.5 - 1.5$ in E. There are relatively few bins in a set; we use from $4 - 13$. We always use an overlap of 1 bin.

We find it convenient to replace the weighting function with a piecewise linear function: $s_W(\epsilon) = \beta_{set}\epsilon$ within each set, with β_{set} varying from one set to the next. For our range of sizes of the bins and sets this gives results indistinguishable from those obtained with a smooth weighting curve. Using the linear form for s_W allows a considerable saving of computer time. In practice, it means that, within each set, we are running a standard Metropolis gauge update Monte Carlo with $\beta = \beta_{set}$, except that we restrict the energy to lie within a small set, and that β varies from set to set.

A further ingredient of our method is how we construct the trial changes in the configuration space. The following procedure represents a compromise between speeding up the computer program through vectorization and making sure that the algorithm moves through all of phase space. Ref. 7 suggests making a random change to a random link. To save computer time we have done something slightly less random. We break the lattice up into 2^4 hypercubes. We then choose a random

point within the hypercube, and update a link emanating from this same point in all hypercubes. The direction of this link is chosen randomly, and separately for each hypercube. This yields a set of links whose contribution to E is non-overlapping. We order these links randomly, and then repeatedly move through the links in this order suggesting changes. We use ten such repetitions ("hits") for all the results in this paper. The changes to the link matrices are made by multiplying by an $SU(3)$ matrix drawn from a trace-biased distribution. Using a criterion described in appendix A, we optimize a parameter determining the size of the change.

It is because E must remain within bounds that the algorithm is intrinsically not vectorizable. In our implementation the loops which are not vectorized, essentially a series of logical manipulations, take 40-50% of the CPU time. The remaining time is taken up with what amounts to doing the usual manipulations in an $SU(3)$ Metropolis update code. We stress, though, that most of this time is needed whether or not we use a weighting function.

Finally, we discuss how we move between the sets. All our runs are done in a contiguous subset of the total number of sets. We begin with all links set to the unit matrix, and then move to an energy in the desired starting set at one or other end of the range. From this set, we move monotonically in energy through the sets. As discussed below, we make runs in both directions. To move from one set to the next, or from the starting configuration to the first set, we suggest changes to links and accept only those that move the energy in the desired direction. We move until we are in the central bin(s) of the set, and then begin collecting the events. We find this method to be adequate even when moving "uphill" i.e. moving in a direction of decreasing density of states. We experimented with "finite temperature" algorithms in which we accepted a certain fraction of the moves in the wrong direction, in the hope of overcoming possible metastabilities, but we found no such method to improve over the naive one.

In summary, within each set we obtain an estimate of the weighted number of states per bin. We then multiply by the inverse of the weight to obtain the

actual, binned density of states. We extend these results over the entire energy range by matching the results from adjacent sets in the overlap bin. This gives us $s(\epsilon)$ over some range of ϵ, up to an overall factor. There are three types of error in the resulting entropy density. First, there is the error due to discretization. To the extent that $s(\epsilon)$ is described by $s_W(\epsilon)$, this is a small effect. It is discussed further in the next section. Second, there are statistical errors, and, finally, there are systematic errors. It turns out that the systematic errors are most important, and we discuss these first.

Possible systematic errors come from the usual sources for Monte Carlo calculations: insufficient thermalization, hysteresis, and, for second order transitions, critical slowing down. Thermalization here means that when we enter a new set, it takes a certain number of events to get to that region of configuration space where there are most states for the given energy range. However, with our method of moving from one set to an adjacent set, we find this effect to be small. We are already close to the appropriate region of configuration space. Thus we have chosen not to discard any events when we begin in a new set. However, the first one or two sets are effected by thermalization, as is clear by a comparison with other runs, and we discard these edge sets.

A much more important source of systematic error is hysteresis. Since we move from one set to an adjacent set, the configurations we sample remember their history. The result in a set depends upon the direction from which we enter. This is only a problem close to the transition, which in the present instance is first order. We have tried to use this effect to our advantage. We only consider our result to be final if runs in both directions agree. Thus, for energies in the transition region, we must run long enough for both phases to have appeared, and for possible phase separated states to be included. To make sure that this happens, we find it important to use large sets. This allows our simulation to move around more freely in the configuration space, and so to avoid more easily any energy barriers. The largest sets we have used consist of 13 bins each of size ~ 1.7 in units of E. Clearly, only with a weighting function can we use such large sets. Indeed,

we must have a good first estimate of $s(\epsilon)$ in order to do this.

The third source of systematic error is critical slowing down. It is our claim, contrary to that of Refs. 6,7, that this method does suffer from critical slowing down in the vicinity of a critical point. This shows up, in energy ranges close to the critical energy, as the existence of states involving fluctuations on very different length scales. Thus, it takes more and more time to move through these states using a local algorithm.

We close this section with a discussion of the statistical error in our estimate of the density of states. Consider a range of ϵ, which we divide into B bins, partitioned into S sets each of b bins. We use a single overlap bin, so $B=(b-1)S+1$. It makes no difference how large the bins are – the same events are simply being repackaged. The only restriction on the size of the bins is that they be small enough to resolve interesting structure in $s(\epsilon)$. We place N events in total into the energy range, and record every n_{in}-th one. We are interested in the relative error in the ratio R between the number of events in the two bins at the edges of the range. This we parametrize as:

$$\frac{\delta R}{R} = \sqrt{\frac{B^2}{N}\mathcal{F}(b, n_{in})} \tag{2.8}$$

which defines a figure of merit $\mathcal{F}$, which we want as small as possible. For fixed b and n_{in}, the error is proportional to $\sqrt{B^2/N}$ because the error in each bin is $\propto \sqrt{B/N}$, and this error must be propagated over $S \approx B/b$ sets. For a fixed range of ϵ, the number of bins grows like V. Thus to maintain a constant relative error in R requires $N \propto V^2$, so that the computer time required grows like V^2, as stressed in Refs. 6,7. For a first order transition one does indeed have to study a fixed range in ϵ, as the discontinuity in ϵ tends to a constant as $V \to \infty$. For a second order transition, on the other hand, the interesting region in ϵ shrinks as V increases, and thus the scaling factor need not be so bad. This improvement will, however, be countered by the effects of critical slowing down.

What we need to know is how $\mathcal{F}$ depends on b and n_{in}, and also upon the

extent of the weighting.* It is shown in Appendix A that for perfect weighting, and at fixed n_{in}, $\mathcal{F}$ is almost independent of b (or equivalently of S). In fact $\mathcal{F}$ slightly decreases towards a limiting value as b increases, mainly because the number of overlap bins decreases. For bins of size ~ 1.5 units of E, in the region of the phase transition, and with our optimal hit matrices, this limiting value is 350-400. The large size of $\mathcal{F}$ is due to the considerable correlations between events.

Since $\mathcal{F}$ is nearly independent of b, we get almost the same $\delta R/R$ however we partition the bins into sets. A heuristic argument for this is that almost the same events are being differently packaged into sets. However, the events are not exactly the same, because the rejections which keep the events within sets are different.† Thus, with perfect weighting, we lose no statistical power if we use large sets, as is required to overcome systematic errors in $\delta R/R$ due to hysteresis.

As the weighting is removed, $\mathcal{F}$ grows rapidly. Table 3 in Appendix A shows that, for $b = 4$ the figure of merit grows from 450 for perfect weighting to 6×10^{10} for no weighting. Furthermore, with less than perfect weighting, $\mathcal{F}$ grows exponentially with b. To take an extreme case, for no weighting $\mathcal{F} \simeq 6 \times 10^{10}$ for $b = 4$ while $\mathcal{F} \simeq 10^{20}$ for $b = 7$. We stress that $\mathcal{F}$ includes the effects of Metropolis rejection which accompany weighting. These numbers show clearly how important it is to have a good weighting function. They also illustrate why there is a practical limit to the width of sets: as the set size increases the required accuracy in the weighting function eventually becomes unattainable.

The final question addressed in Appendix A is how $\mathcal{F}$ varies with n_{in}. For $n_{in} = 1$, which we use throughout, events are highly correlated because the typical

---

⋆ For simplicity we assume that the suggested changes to the link matrices move us through the energies in the same way in all sets, and that the weighted density of states varies in the same way in all sets. This is reasonable if we restrict ourselves to a limited range of ϵ, or if we are in the vicinity of the phase transition.

† A somewhat more rigorous argument goes as follows. Our algorithm moves around locally, performing a random walk, so it takes a number of events $\propto b^2$ to get an independent event in *each* bin. Now in each set there are N/S events, so the relative error in each bin is $\sim \sqrt{b^2/(N/S)} = \sqrt{B^2/(NS)}$. The matching of this ratio at the set boundaries increases this error by $\sqrt{S}$, so the total error is $\sim \sqrt{B^2/N}$, independent of S.

step size is much smaller than the size of the bins. If one increases n_{in}, correlations are reduced, so the errors per recorded event are reduced, but the penalty is that not all events are recorded. We find, as illustrated in Table 3, that, at fixed b, $\mathcal{F}$ always increases as n_{in} increases from 1. This is true with or without weighting. This increase is very slow at first, so that it may be optimal to have n_{in} somewhat larger than 1, since it takes some computer time to record an event. However, it is clearly not optimal to record only uncorrelated events, as done in Ref. 7. For $b = 4$, such a strategy roughly doubles $\mathcal{F}$ (see Table 3), and in general it increases $\mathcal{F}$ by an extra factor proportional to $\sqrt{b}$.

3. FINITE SIZE SCALING

Finite size scaling (FSS) analysis[10] is usually associated with second-order phase transitions, but it is also useful for first order transitions[11,12]. A first-order phase transition is characterized by coexistence of two phases at the critical point. Finite-volume corrections have two effects: discontinuities are rounded, with a width inversely proportional to the volume V:

$$\Delta\beta/\beta_c \sim V; \tag{3.1}$$

and the critical coupling β_c is shifted:

$$\beta_c(L) - \beta_c(\infty) \sim L^{-d_s}. \tag{3.2}$$

Here L is the linear extent of the system, $V \propto L^d$, and d_s depends on the boundary conditions. For periodic boundary conditions, which we use $d_s = d$.

One can consider a first order transition as a limiting case of a second order transition. FSS analysis of second order transitions encompasses both of the effects described above. However, it is only for the rounding of the transition that the limit of the second order analysis describes the first order FSS. The shift in β_c is

a non-universal effect. This distinction is important because our measurement of the rounding is more reliable than that of the shift in β_c. Previous studies of FSS for the finite temperature transition of pure gauge QCD,[9,13] on the other hand, have considered the shift in β_c.

FSS analysis has been extended to the zeroes of the partition function for second order transitions in Ref. 4. The basic result is that zeroes close to β_c behave like

$$\beta_0(L) - \beta_c(\infty) \sim L^{-1/\nu}. \tag{3.3}$$

Here $\beta_0(L)$ are the complex zeroes of a lattice of dimension L and ν is the correlation length exponent. The roots always come in complex conjugate pairs, so we can consider only those in the upper half plane. These lie on a straight line starting at $\beta_c(\infty)$, which is real. The position of this line is independent of L; as L increases the zeroes move along this line in the way described by Eq. (3.3). Thus the scaling behavior applies separately to the real and imaginary parts of $\beta_0 - \beta_c$. The density of zeroes along the lines is determined by the specific heat exponent α. The angle the lines make with the real axis is related to α and to the specific heat amplitude ratio.

One can extend this result to a first order transition by taking the limit $\nu \to 1/d$[11]. In this limit one finds that the line of zeroes is perpendicular to the real axis, and that the density of zeroes is constant.[14] Proceeding naively, one might expect the real part of $\beta_0(L) - \beta_c(\infty)$ to vanish. In fact, what happens is that now for each L there is a *separate* straight line perpendicular to the real β axis at $\beta_c(L)$ which depends on L in a non-universal way (cf. Eq. (3.2)). Thus in the singular limit of a first order transition the FSS result, Eq. (3.3), applies only to the imaginary part of $\beta_0 - \beta_c(\infty)$, and the universal prediction of FSS for a first order transition is

$$\mathrm{Im}\left[\beta_0(L) - \beta_c(\infty)\right] \sim L^{-d} \sim 1/V. \tag{3.4}$$

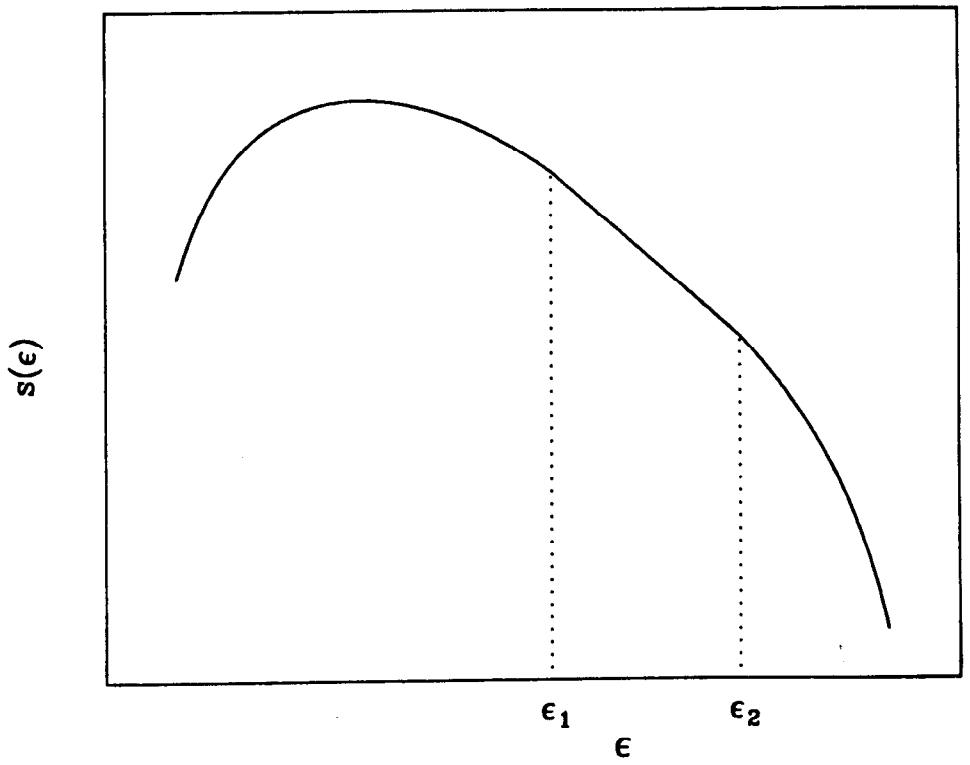

Fig. 1. A typical curve of entropy density $s(\epsilon)$ for a system with a first order phase transition.

This is transcription of Eq. (3.1) to the language of zeroes, and it is this equation which we test below.

It is instructive to understand how the scaling formula Eq. (3.4) comes about as a direct consequence of a first order transition, rather than as a limiting case of Eq. (3.3). Consider first the infinite volume limit. In a generic case the entropy density $s(\epsilon)$ has the shape depicted in Fig. 1. It is convex, $s''(\epsilon) \leq 0$, because there are at least as many states at energy density ϵ as obtained by partitioning the volume into equal parts of the states at energy $\epsilon + \delta$ and $\epsilon - \delta$. Such a partitioning gives for the entropy density a straight line joining $s(\epsilon + \delta)$ and $s(\epsilon - \delta)$, up to

corrections vanishing as $V \to \infty$.* The average energy density and β are related by the saddle point equation

$$s'(\epsilon) + \beta = 0. \tag{3.5}$$

Now $\langle \epsilon(\beta) \rangle$ has a discontinuity, jumping from ϵ_1 to ϵ_2 at $\beta = \beta_c$, jumping from ϵ_1 to ϵ_2 as shown schematically in Fig. 2. Thus, for $\beta = \beta_c$, Eq. (3.5) is satisfied by a *range* of energy values, $\epsilon_1 \leq \epsilon \leq \epsilon_2$. In other words, the free energy density $s + \beta_c \epsilon$ is *flat*, as illustrated in Fig. 3†.

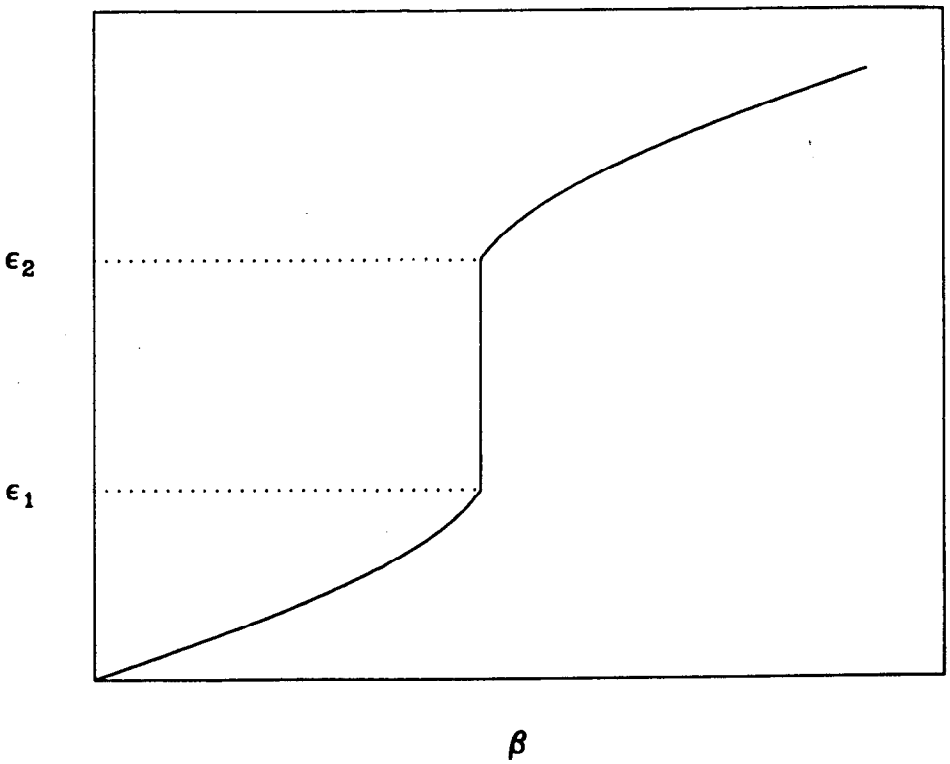

Fig. 2. The discontinuity in $\langle \epsilon(\beta) \rangle$ which occurs when $V \to \infty$ for a system whose free energy is that shown in Fig. 1.

⋆ We thank Michael Peskin for reminding us of this argument.

† Actually a form for s with two peaks at $\epsilon = \epsilon_1$ and $\epsilon = \epsilon_2$ would also yield the desired form for $\epsilon(\beta)$. We know that s is flat, however, because of its convexity.

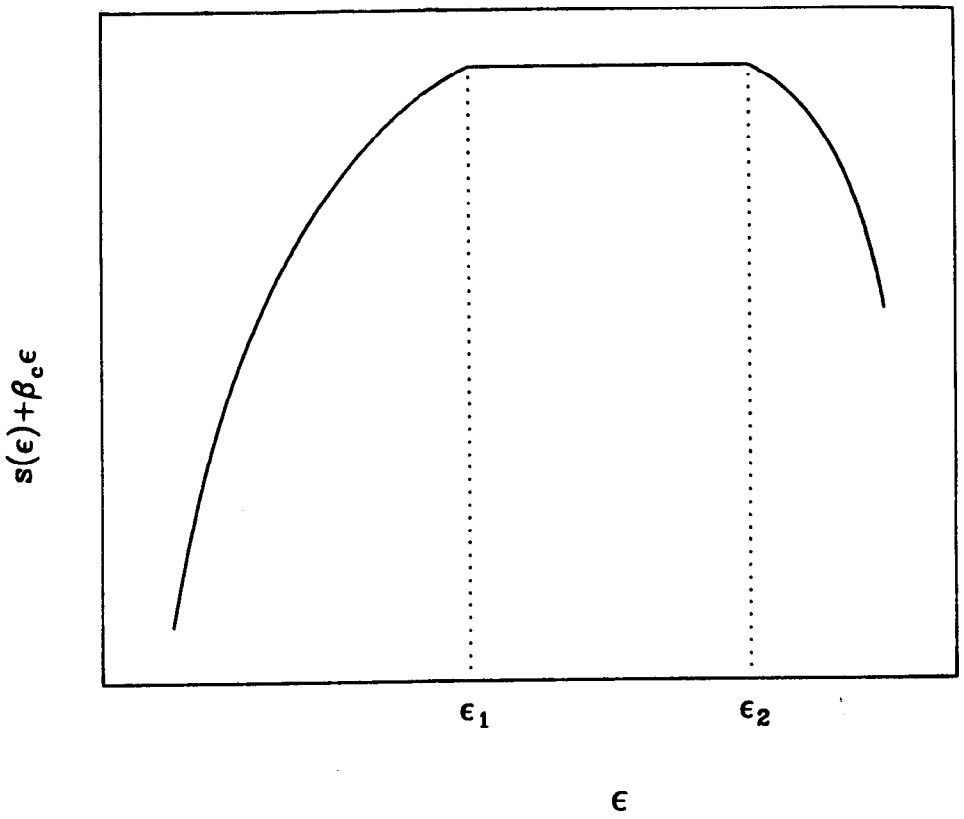

Fig. 3. The free energy density, $f = s + \beta_c \epsilon$, for a system with a first order phase transition.

For finite volume, two types of corrections arise. First, the entropy density changes. This can shift the flat region by $O(1/V)$ – the cause of the shift in β_c – and also change the shape of $s(\epsilon)$. In particular, a concavity can appear in the flat region. In this region there must be phase separation, and concavity occurs if the energy of domain walls outweighs the number of ways to produce such configurations. If it does, then the concavity in f can be as large as $O(1/L)$. In that case the free energy $F = Vf$ has a dip which grows as $O(L^{d-1})$, and only energies at the edge of the range contribute to the partition function. If, on the other hand, the dip in f is $O(1/V)$ or less, then F only has a dip of $O(1)$, and all energies in the range contribute. The second finite volume effect is that there are $O(1/V)$ corrections to the saddle point equation – a given β samples a region of ϵ.

For a typical ϵ this region has a size $\propto 1/V$. But for $|\beta - \beta_c| \lesssim 1/[(\epsilon_2 - \epsilon_1)V]$ both ends of the "flat" region, and possibly the middle too, contribute to the functional integral. This is what gives rise to the rounding of the transition.

For $SU(3)$, as for all theories with continuous symmetry, the density of states $N(\epsilon) = \exp[Vs]$ is continuous. In a numerical study of the roots of Z we must approximate $N(\epsilon)$ by a discrete distribution. In order to introduce the notation we need later, we will make the argument concerning the distribution of zeroes using this discrete approximation. The energy range $\epsilon_{min} \leq \epsilon \leq \epsilon_{max}$ is divided into B bins of width $\Delta\epsilon = (\epsilon_{max} - \epsilon_{min})/B$. The partition function is then approximately proportional to a polynomial of degree $B-1$:

$$\begin{aligned} Z(\beta) &= \int d\epsilon\, N(\epsilon) \exp[V\beta\epsilon] \approx \sum_{k=0}^{B-1} C_k \exp[V\beta\epsilon_k] \\ &= \exp[\beta V(\epsilon_{max} - \Delta\epsilon/2)] \sum_{k=0}^{B-1} C_k w^k \equiv \exp[\beta V(\epsilon_{max} - \Delta\epsilon/2)]\, \mathcal{P}(\beta) \end{aligned} \tag{3.6}$$

where

$$c_k = \epsilon_{max} - (k + \frac{1}{2})\Delta\epsilon\,; \quad C_k = \int_{\epsilon_{max}-(k+1)\Delta\epsilon}^{\epsilon_{max}-k\Delta\epsilon} d\epsilon N(\epsilon) \sim \Delta\epsilon\, N(\epsilon_k); \quad w = e^{-\beta V \Delta\epsilon}.$$

We first calculate the roots of $\mathcal{P}(\beta)$ for a system with a flat free energy. To do this it is useful to change variables:

$$\mathcal{P}(\beta) = \sum_k C_k w^k \propto \sum_k \exp[V(s_k + \beta_c\epsilon_k)]\, [\exp(-V\beta_c\epsilon_k)u^k] = \sum_k D_k u^k \tag{3.7}$$

where the proportionality constant is independent of β and

$$N(\epsilon_k) \equiv \exp(Vs_k); \quad D_k = \exp[V(s_k + \beta_c\epsilon_k)]; \quad u = \exp[-V(\beta - \beta_c)\Delta\epsilon].$$

The flat region $\epsilon_1 \leq \epsilon \leq \epsilon_2$ corresponds to $k_1 \leq k \leq k_2$:

$$D_k \approx \widetilde{D}: \qquad k_1 \leq k \leq k_2$$

$$D_k \ll \widetilde{D}: \qquad k < k_1 \quad \text{or} \quad k > k_2$$

The coefficients outside the flat region are suppressed by exponentials of the volume and can be neglected in the first approximation:

$$\sum_{k=0}^{B-1} D_k\, u^k \approx \sum_{k=k_1}^{k_2} D_k u^k \approx \widetilde{D}\, u^{k_1} \sum_{k=k_1}^{k_2} u^k = \widetilde{D}\, u^{k_1} \left(\frac{u^{k_2-k_1+1}-1}{u-1} \right) \tag{3.8}$$

The corresponding roots of $\mathcal{P}$ accumulate along a unit circle in the u plane, or along the line $\beta = \beta_c + i\gamma$ in the complex β plane:

$$u = \exp[-V(\beta-\beta_c)\Delta\epsilon] = \exp\left[\frac{2\pi n i}{k_2-k_1+1}\right];$$

or

$$(\beta-\beta_c) = \pm\frac{2\pi n i}{V\,\Delta\epsilon(k_2-k_1+1)} = \pm\frac{2\pi n i}{V\,(\epsilon_2-\epsilon_1)}; \qquad n = 1,2,3,\ldots,\frac{k_2-k_1}{2} \tag{3.9}$$

Thus we find, as advertised, that the zeroes lie on a line perpendicular to the real axis, with uniform density, and approach the real axis like $1/V$.

This result was derived assuming no dip in the free energy. In the opposite extreme of a large dip which grows with L, only the two points at the end of the range contribute. Thus, as $V \to \infty$ we need only include $k = k_1$ and $k = k_2$ in the sum. The resulting zeroes are given by:

$$(\beta-\beta_c) = \pm\frac{\pi(2n-1)i}{V\Delta\epsilon(k_2-k_1)}; \qquad n = 1,2,3,\ldots,\frac{k_2-k_1}{2} \tag{3.10}$$

These zeroes lie on the same line as those in Eq. (3.9), have uniform density, and approach the real axis as $1/V$. For small n, they lie half way between those obtained for the flat free energy.

We can turn the argument around, and ask what are the particular properties of the zeroes that lead to the discontinuity in $\langle\epsilon\rangle$. Here a variant of the electrostatic analogy introduced in Ref. 1 is useful. Let us give each zero in the u plane a charge $-\Delta\epsilon$. Then, the electric field along the real axis is $\langle\epsilon\rangle/u$ up to an overall constant. It is clear that the ring of zeroes described by Eqs. (3.9) will give a discontinuity in $\langle\epsilon\rangle$ in the infinite volume limit. But it is also true that similar distributions of zeroes, such as that of Eq. (3.10), will do so as well. In order to get the discontinuity, and for $\langle\epsilon\rangle$ to grow monotonically with β, what is required is that the line of zeroes must be roughly perpendicular to the real axis and that the zeroes must have a nearly uniform distribution near to the real axis. We do not, however, know of any rigorous arguments for this. The electrostatic analogy makes clear that the zeroes close to the real axis will dominate the rounding. Thus, in order to obtain Eq. (3.1) the closest zeroes must approach the real axis as $1/V$, in agreement with Eq. (3.4).

What are the errors introduced by the discretization? For the simple example discussed above, the result (3.9) holds even with no discretization, except that there is no limit to the value of n. This illustrates the general behavior. Clearly, if we discretize, there is a limit to imaginary part of β, $|\mathrm{Im}\,(\beta)| < \beta_{lim} \equiv \pi/(V\Delta\epsilon)$. As one decreases $\Delta\epsilon$ what happens is that (1) zeroes with small imaginary parts are little affected, (2) zeroes with imaginary parts close to β_{lim} do move, and (3) a whole new set of zeroes appear with $\mathrm{Im}(\beta) > \beta_{lim}$ We have explicitly verified these features. This behavior can be understood as follows: the partition function evaluated for imaginary β corresponds to a Fourier transform of $\exp\Big[V(s+\mathrm{Re}(\beta)\epsilon)\Big]$ with a frequency $V\,\mathrm{Im}(\beta)$. Thus only the distant zeroes, which correspond to the high frequencies, feel the effect of changing the resolution of the measurement of $N(\epsilon)$.

There remains one technical point concerning the discretization: the approximation $C_k \approx \Delta\epsilon\exp(Vs_k)$. We know how the density of states varies in a bin

$$N(\epsilon) = \exp\Big\{V\Big[s_k - \beta_k(\epsilon-\epsilon_k) + O\big((\epsilon-\epsilon_k)^2\big)\Big]\Big\}.$$

where our method of measurement gives us both s_k and β_k. Thus, for each β, we can calculate C_k more accurately:

$$C_k \approx \Delta\epsilon \ \exp(Vs_k) \ \frac{2 \sinh\Big[V(\beta - \beta_k)\Delta\epsilon/2\Big]}{V(\beta - \beta_k)\Delta\epsilon} \quad . \tag{3.11}$$

Consider some complex β. For bins for which $|\beta - \beta_k| \lesssim 1$, Eq. (3.11) reduces to $C_k \approx \Delta\epsilon \exp(Vs_k)$, since $V\Delta\epsilon \sim 1$. But, for a given β, it is precisely those bins with smallest $\beta - \beta_k$ that are important, so the approximation is good for nearly all β. Only for $|\mathrm{Im}\beta| \gtrsim 1$, i.e. for distant zeroes, does the approximation brake down. This is just a restatement of the fact that only the distant zeroes are sensitive to the discretization.

4. ANALYTIC EXPANSIONS

In this section we present the strong and weak coupling expansions for the entropy density and compare them with our results.

Ref. 15 provides the strong coupling expansion for the free energy density of SU(3), in infinite volume. From their result one can immediately obtain the strong coupling expansion for the energy density.* It is most convenient to use the variable $\tilde{\epsilon} = 3\ \epsilon$:

$\star$ One must take care with conventions: our β is 6 times that of Ref. 15.

$$
\begin{aligned}
\langle\tilde{\epsilon}\rangle = {} & \tfrac{1}{6}\,\beta + \tfrac{1}{72}\,\beta^2 - \tfrac{5}{31104}\,\beta^4 + \tfrac{16d-113}{5038848}\,\beta^5 + \tfrac{7(80d-133)}{302330880}\,\beta^6 \\
& + \tfrac{800d-1069}{1813985280}\,\beta^7 + \tfrac{400d-509}{14511882240}\,\beta^8 + \tfrac{6400d^2-282600d+490757}{21158324305920}\,\beta^9 \\
& + \tfrac{11(17920d^2-267624d+435299)}{592433080565760}\,\beta^{10} + \tfrac{1272320000d^2-7658871745d+10092064674}{7677932724132249600}\,\beta^{11} \\
& + \tfrac{13(157696000d^2-487188695d+364951632)}{46067596344793497600}\,\beta^{12} \\
& + \tfrac{50176000d^3+5254054400d^2-3055083115d-14239256399}{1243825101309424435200}\,\beta^{13} \\
& + \tfrac{1949696000d^3-49696460800d^2+216598878225d-250150888296}{29851802431426186444800}\,\beta^{14} \\
& + \tfrac{359667302400d^3-6538865646880d^2+21195565316042d-19208912255241}{72539879908363563306086400}\,\beta^{15} \\
& + \mathcal{O}(\beta^{16})
\end{aligned}
\tag{4.1}
$$

Here we have kept the dependence on the dimension of spacetime d.

Given the strong coupling expansion for $\langle\epsilon\rangle$, one can obtain the power series expansion for the entropy $S(\epsilon)$ around $\epsilon = 0$. The starting point is the expression for $\langle\epsilon\rangle$ in terms of the density of states:

$$
\langle\epsilon\rangle = \frac{\int \exp\{V\left[s(\epsilon)+\beta\epsilon\right]\}\,\epsilon\,d\epsilon}{\int \exp\{V\left[s(\epsilon)+\beta\epsilon\right]\}\,d\epsilon}. \tag{4.2}
$$

Since $N(\epsilon)$ is an extremely rapidly varying function, with a very sharp maximum at ϵ_0, the integrals can be evaluated using the steepest descent method. The saddle point ϵ_0 is given by

$$
s'(\epsilon_0) + \beta = 0; \qquad s(\Delta\epsilon) = -d_2\Delta\epsilon^2 - d_3\Delta\epsilon^3 + \mathcal{O}(\Delta\epsilon^4); \qquad \Delta\epsilon \equiv \epsilon - \epsilon_0 \tag{4.3}
$$

where d_0 can be taken to be zero since it corresponds to an overall multiplicative factor in $N(\epsilon)$, and $d_1 = 0$ because ϵ_0 is a an extremum of the integrand. The

result, including the leading $1/V$ corrections reads[†]

$$\epsilon = \epsilon_0 - \frac{3d_3}{4Vd_2^2} \tag{4.4}$$

We now write $s(\epsilon)$ as

$$s(\epsilon) = -\sum_{i=1}^{16} c_n \epsilon^n, \tag{4.5}$$

combine with Eqs. (4.1) and (4.4) and plug into Eq. (4.3), equating the coefficients of β to zero, order by order.

The resulting expression, including $1/V$ corrections to the first two terms, is:

$$\begin{aligned} s(\epsilon) = {} & \tfrac{1}{2V}\tilde{\epsilon} - (1 - \tfrac{1}{2V})\,\tilde{\epsilon}^2 + \tfrac{1}{3}\,\tilde{\epsilon}^3 - \tfrac{1}{4}\,\tilde{\epsilon}^4 + \tfrac{1}{6}\,\tilde{\epsilon}^5 + \tfrac{16d-275}{1944}\,\tilde{\epsilon}^6 + \tfrac{11}{120}\,\tilde{\epsilon}^7 - \tfrac{13}{192}\,\tilde{\epsilon}^8 \\ & + \tfrac{253}{5184}\,\tilde{\epsilon}^9 - \tfrac{800d+88067}{2624400}\,\tilde{\epsilon}^{10} - \tfrac{8960d-686899}{29393280}\,\tilde{\epsilon}^{11} + \tfrac{81920d^2-307855d-9894642}{725594112}\,\tilde{\epsilon}^{12} \\ & - \tfrac{61625d-7678606}{1007769600}\,\tilde{\epsilon}^{13} + \tfrac{14450688d^2-27669565d-590906590}{213324668928}\,\tilde{\epsilon}^{14} \\ & - \tfrac{65536000d^2+118450875d+364241882}{1632586752000}\,\tilde{\epsilon}^{15} \\ & + \tfrac{165150720d^3-2888990720d^2+51873015775d+169719955926}{98738846760960}\,\tilde{\epsilon}^{16} \\ & + \mathcal{O}(\tilde{\epsilon}^{17}). \end{aligned} \tag{4.6}$$

We have included the $O(1/V)$ corrections to the first two terms. These corrections are small even on a 2^4 lattice ($V = 96$), and so we decided that it was not worth calculating the $O(1/V)$ corrections to higher order terms. Also, on a finite lattice, there are $O(1/V)$ corrections to the original expansion for ϵ, equation (4.1), which come in at higher order, and which we do not know.

In Figure 4 we plot the entropy density $s(\epsilon)$ on a 4^4 lattice and compare it with the "strong coupling" series, Eq. (4.6). For $|\epsilon| \lesssim 0.3$ the agreement is excellent, providing a useful check on the correctness of our measurements in that region. We find similar agreement on other lattices.

† In the following we drop the $\langle\rangle$ sign around ϵ

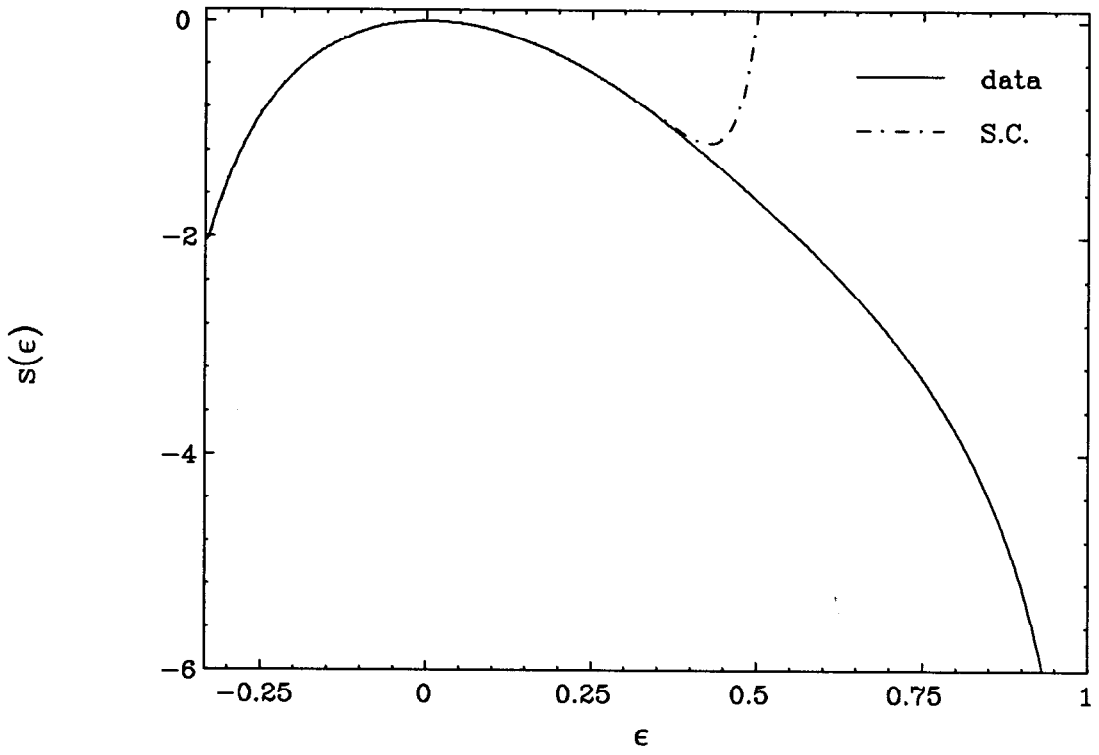

Fig. 4(a). Comparison of the measured entropy per plaquette $S(\epsilon)$ on a 4^4 lattice with the expansion (4.6) derived from the strong coupling series (4.1).

At the other extreme, the first few terms are known in the weak coupling expansion of ϵ[16]:

$$\xi = \sum_{k=1} w_k/\beta^k; \qquad \xi \equiv 1 - \epsilon. \tag{4.7}$$

ξ is the natural variable for the weak coupling expansion of $N(\epsilon)$. In Ref. 16 w_1 and w_2 have been computed analytically for all $SU(N)$ groups and the higher order coefficients have been fitted to Monte-Carlo data for $SU(2)$. Since here we are interested in $SU(3)$, we shall limit ourselves to w_1 and w_2:

$$\xi = 2/\beta + (1.2248 - 32.7V)/\beta^2. \tag{4.8}$$

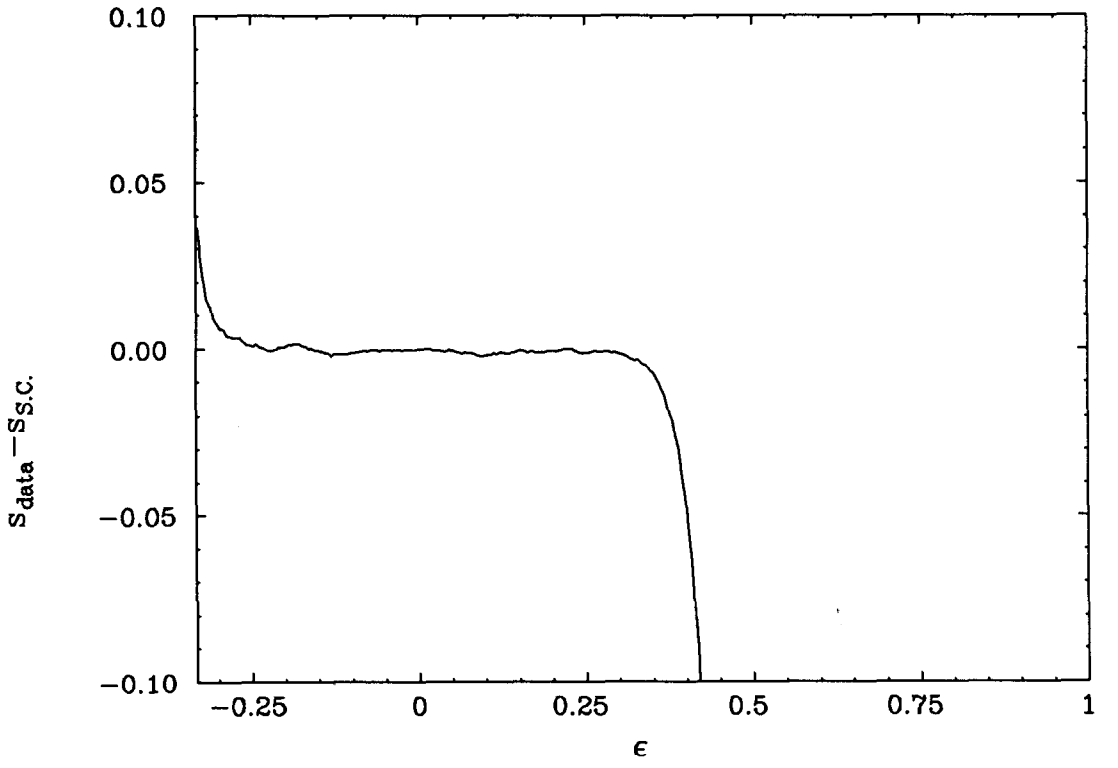

Fig. 4(b). The difference between the measured entropy and the series (4.6).

The saddle point equations (4.3) and (4.5) are replaced by:

$$s'(\xi_0)-\beta = 0; \qquad s(\Delta\epsilon) = -d_2\Delta\xi^2 - d_3\Delta\xi^3 + \mathcal{O}(\Delta\xi^4); \qquad \Delta\xi \equiv \xi - \xi_0; \quad (4.9)$$

$$s(\xi) = a_0 \log \xi + \sum_{k=1} a_k \xi^k \quad . \qquad (4.10)$$

The $\log(\xi)$ term in (4.10) is necessary for consistency of Eq. (4.9), since $\xi \underset{\beta\to\infty}{\longrightarrow} 0$.

We now combine Eqs. (4.8), (4.10) and (4.9), just as we did for the strong coupling, and obtain

$$s(\xi) = (2 - 1/V)\log(\xi) + (0.61259 - 16.4/V)\xi + a_2\xi^2 + \mathcal{O}(\xi^3)\,. \qquad (4.11)$$

a_2 can be fitted to the data. Figure 5 clearly shows that with $a_2 \approx 0.36$ the data

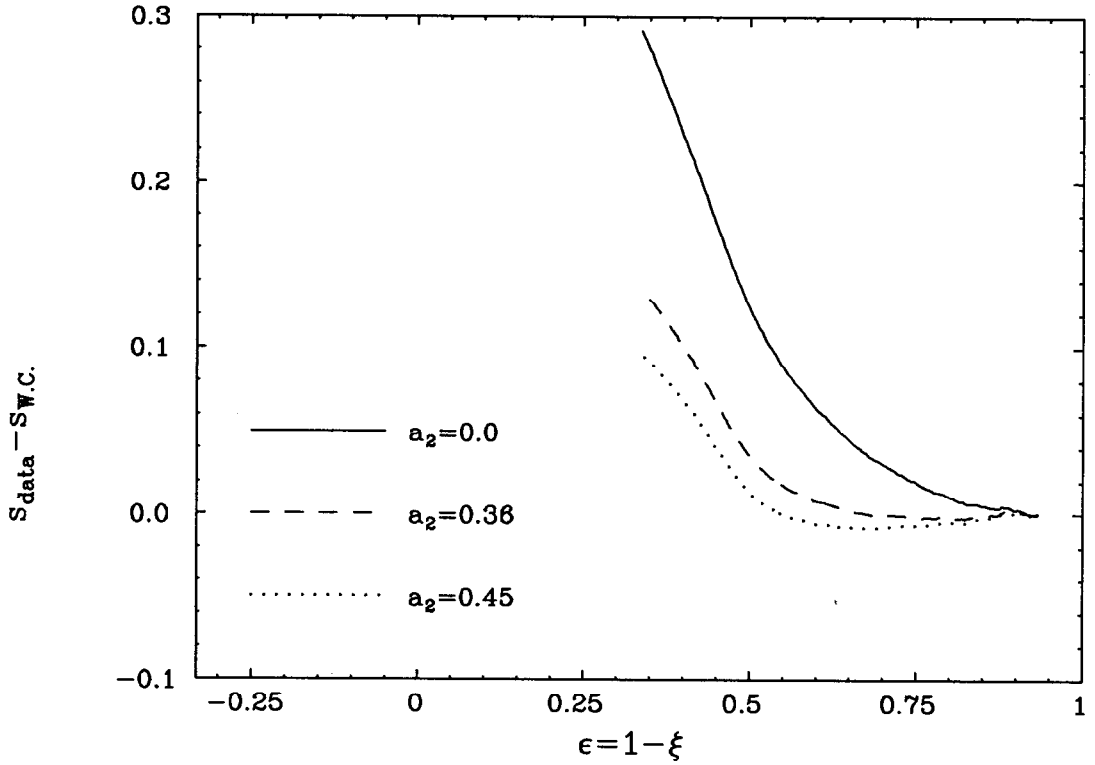

Fig. 5. The difference between data on a 4^4 lattice, and the weak coupling expansion, Eq. (4.11), for several values of a_2.

is well matched by (4.11) up to $\xi \approx .47$.

In summary, existing strong and weak coupling expansions predict $s(\epsilon)$ for $-1/3 \lesssim \epsilon \lesssim 0.3$ and for $.55 \lesssim \epsilon \lesssim 1.0$, respectively. The remaining range, $.3 \lesssim \epsilon \lesssim .55$, is where we concentrate our numerical calculations. This range is the "crossover" region, and it also contains, for the lattices we consider, the finite temperature phase transition. Of course, as one considers lattices with larger extent in the time direction, the finite temperature transition moves towards $\beta = \infty$. We mention this just to emphasize that there is nonperturbative information hidden in $s(\epsilon)$ for $\epsilon \sim 1$, though it is swamped by the perturbative "background".

5. RESULTS

We begin by illustrating the quantities that can be calculated given the numerically determined entropy density. For the 2^4 lattice, we have determined s for all ϵ except in the region corresponding to extremely weak coupling, i.e. $\epsilon \sim 1$. The weak coupling expansion, Eq. (4.10) shows that as $\epsilon \longrightarrow 1$ the slope of s diverges $\propto 1/(1-\epsilon)$. Thus, even with weighting, there is some value of ϵ beyond which numerical methods fail. Our best calculation for a 2^4 lattice uses 400 sets of 4 bins each, each set overlapping the next by 1 bin. Thus each of the 1201 bins is ~ 0.1 units of E wide, which is very narrow compared to the bins we use for the larger lattices. There are 10^5 events in each set, and for this run no weighting is used. To describe the parameters of this run we use the notation: [400 S (29-400), $4B$, 10^5 ev]. The numbers in parentheses give the subset of sets in which data is taken – the numbering begins at $\epsilon = 1$ and extends to $\epsilon = -1/3$. We stress that we are quoting the number of events per set, and that each event consists of a single hit on a single link.

Fig. 6 shows the resulting curve for the entropy density. The maximum is almost at $\epsilon = 0$ – this is true up to $O(1/V)$ corrections in the strong coupling expansion, cf. Eq.(4.6). The expected logarithmic divergence at $\epsilon = 1$ is clearly seen. What is most interesting is the region $.3 \lesssim \epsilon \lesssim .55$, where neither of the analytic expansions apply, and where the curve shows the smallest second derivative. To examine this "crossover" region we use the data for $s(\epsilon)$ to construct $\langle\epsilon\rangle$:

$$\langle\epsilon(\beta)\rangle \approx \frac{\sum_k \exp\left[V(s_k+\beta\epsilon_k)\right]\epsilon_k}{\sum_k \exp\left[V(s_k+\beta\epsilon_k)\right]} \quad . \tag{5.1}$$

Here the sum runs only over those bins in which we have data. Fig. 7 shows $\langle\epsilon(\beta)\rangle$. The crossover region becomes, in this plot, the range of β with the largest derivative, $\beta \sim 5.0$.

As explained in section 2, only a limited number of terms contribute to the sums in Eq. (5.1) for a given β. If we normalize the largest term in the sum in

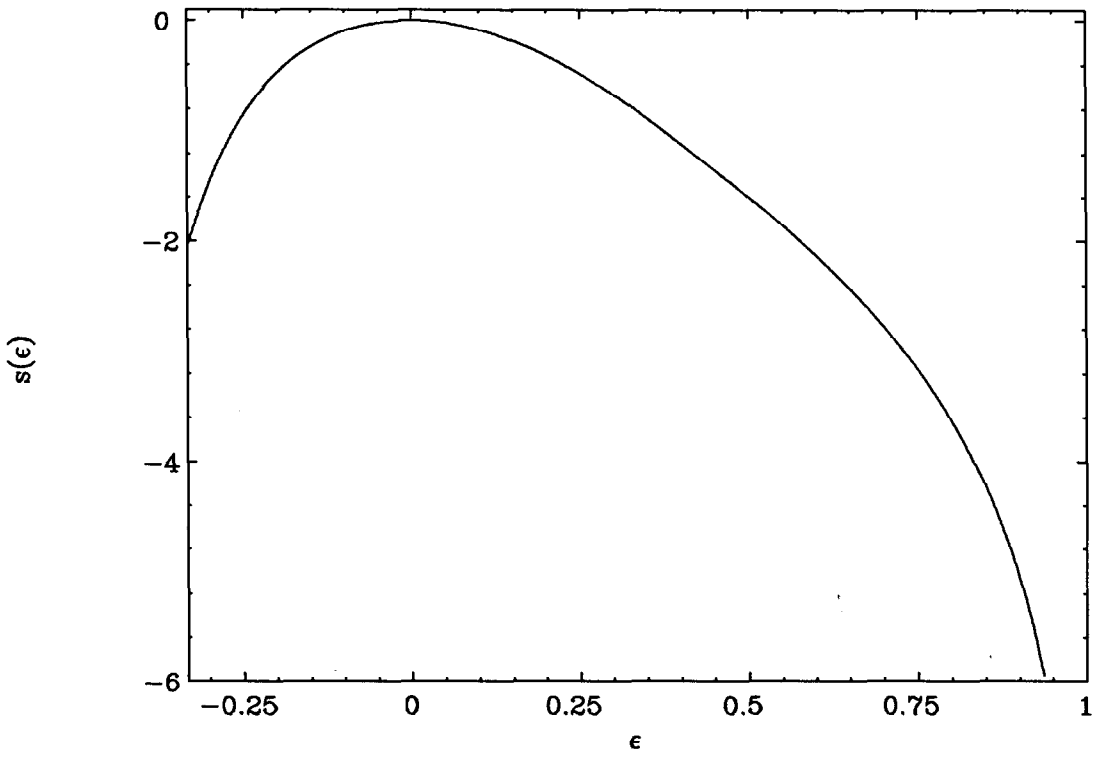

Fig. 6. The entropy density $s(\epsilon)$ for a 2^4 lattice. The absolute position of the vertical scale is arbitrary

the denominator to be 1, then we say that a term contributes to this sum if it is larger than some small number. Because the free energy varies rapidly with ϵ away from its maximum, it matters little what we choose for this number. Using 10^{-3}, we find for the 2^4 lattice that the maximum range of ϵ occurs for $\beta \sim 5.0$, and is $\simeq 1.0$. For β away from the crossover, the range which contributes is smaller. For example, at $\beta = 0$, the range is $\sim .4$, while at $\beta = 12$ it is $\sim .1$. Thus, although we do not have data for $\epsilon \simeq 1$, it is only for large β that this effects the calculation of $\langle\epsilon\rangle$. Here the largest β allowed is ~ 20. For larger lattices, with correspondingly larger V, these ranges of ϵ decrease.

What we are most interested in extracting from the entropy density are the zeroes of the partition function. We have Z as a polynomial in the variable

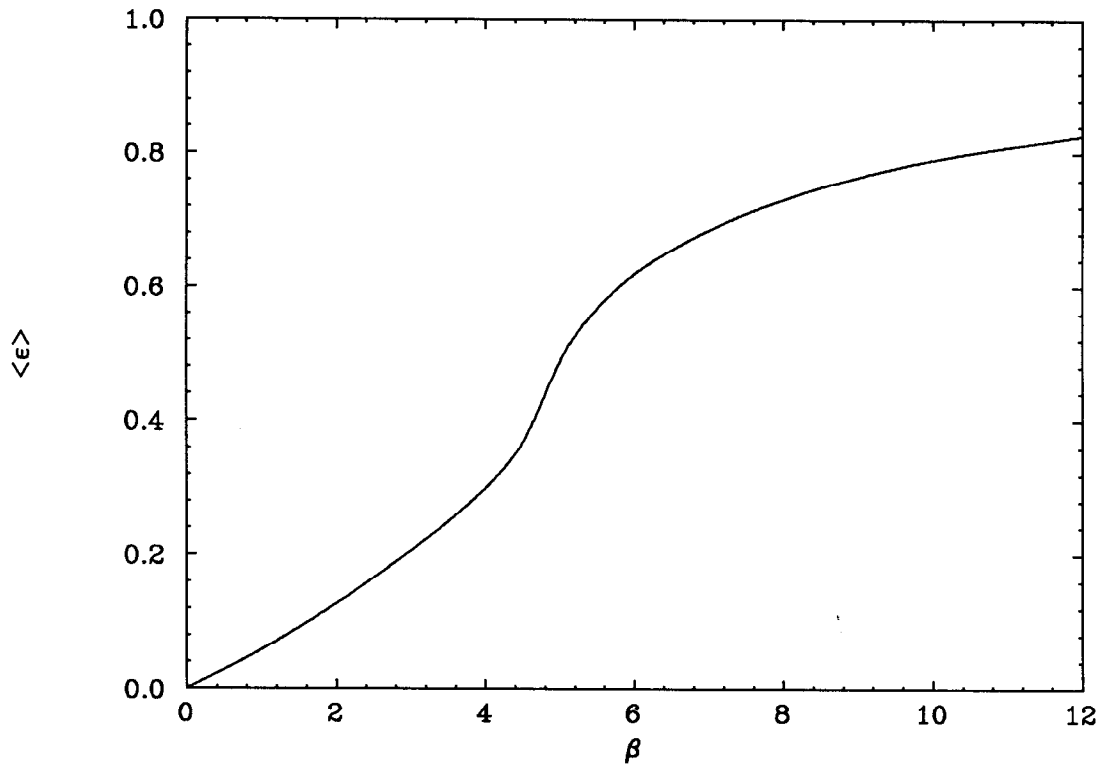

Fig. 7. $\langle\epsilon\rangle$ as a function of β on a 2^4 lattice.

$$u = \exp\left[-(\beta - \beta_{eff})\frac{4V}{3B}\right]. \tag{5.2}$$

Here $B = S(b-1)+1$ is the total number of bins, which have been partitioned into S sets each of b bins. β_{eff} is in principle arbitrary, but we usually choose it to lie as close as possible to the critical coupling β_c. For the 2^4 lattice data, however, we use $\beta_{eff} = 0$.

Z is in principle a polynomial of $(B\text{–}1)$-th order in u, and thus has B–1 zeroes. However, because we do not know s_k for all k, we find the zeroes only of a truncated polynomial. For the 2^4 lattice this truncated polynomial is of order 1188. The resulting zeroes are shown in Fig. 8. Just as for $\langle\epsilon\rangle$, the truncation only effects the zeroes with large $\mathrm{Re}(\beta)$, i.e. with small $|u|$. Had we used the full polynomial, the

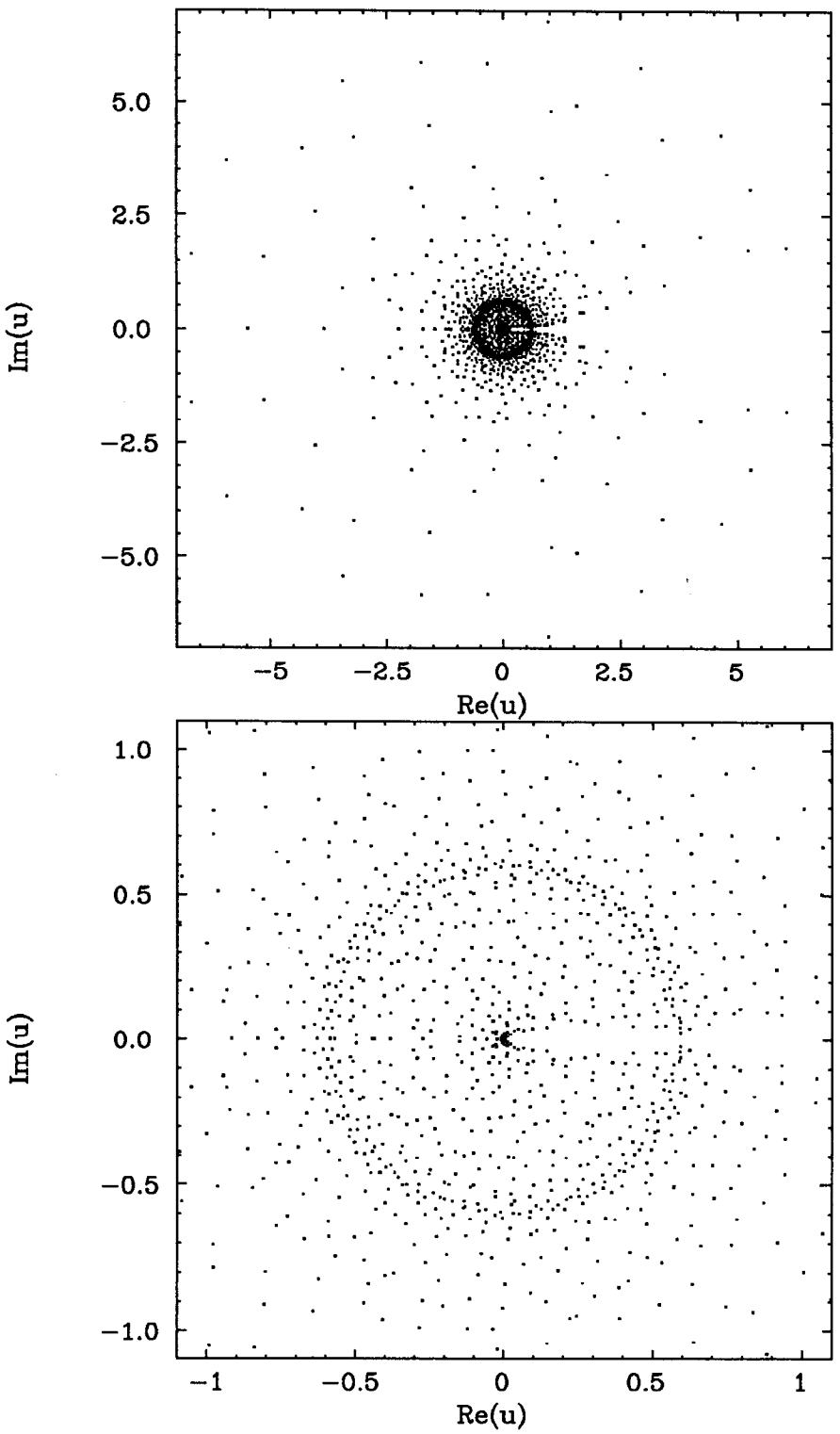

Fig. 8. Zeroes of the partition function on a 2^4 lattice in the u plane. (a) All 1188 zeroes. (b) An expanded view of the region near the origin.

zeroes closest to the origin would have moved, and additional zeroes with smaller $|u|$ would have appeared.

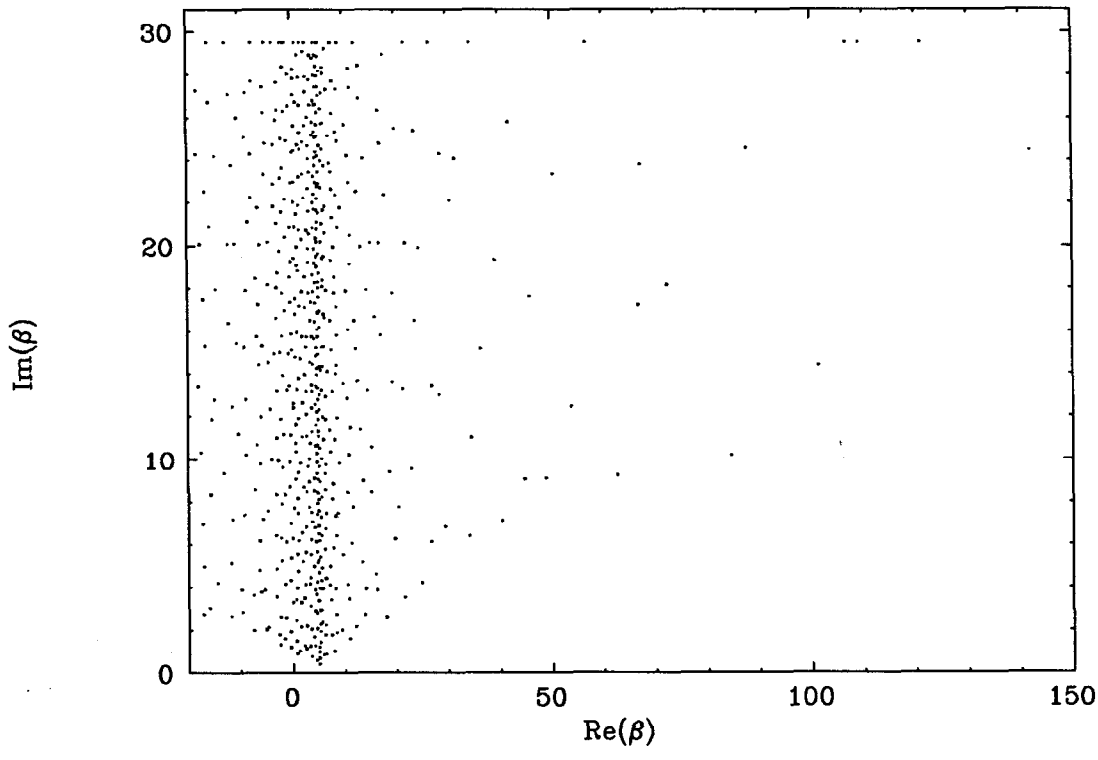

Fig. 9. Zeroes of the 2^4 partition function in the β plane. Only zeroes in the upper half plane are shown.

The most striking features of Fig. 8 are the almost total absence of zeroes close to the positive real axis, the accumulation of zeroes close to the origin, and the ring of roots at $|u| \simeq 0.6$. The ring is the manifestation of the crossover in the distribution of zeroes. If we show the zeroes in the β plane, as in Fig 9, then the ring becomes a vertical band. Notice that the band contains those zeroes which most closely "pinch" the real axis. These figures should be compared to those for $SU(2)$[7]on a 2^4 lattice, for which the pinch is less pronounced.

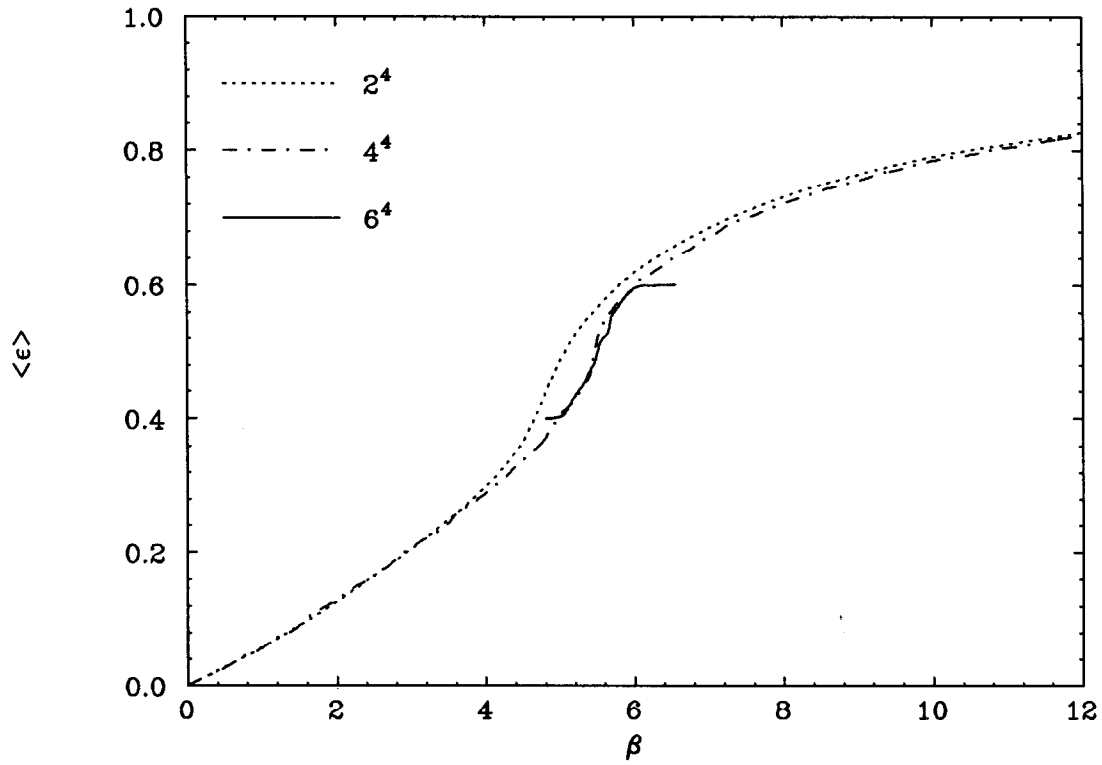

Fig. 10. $\langle \epsilon(\beta) \rangle$ for 2^4, 4^4 and 6^4 lattices.

To investigate zero temperature QCD we have calculated s for 4^4 and 6^4 lattices. The parameters we use for the 4^4 lattice are: $[1000\ S\ (50\text{-}1000),\ 4B,\ 5 \times 10^4\ ev]$. Thus there are 3001 bins of size $\simeq 0.7$, and we evaluate s_k for all but the 49 sets close to $\epsilon = 1$. This calculation uses a weighting function, as do all subsequent ones. For the 6^4 lattice we use $[1000\ S\ (300\text{-}450),\ 7B,\ 8.1 \times 10^5\ ev]$. Notice that the bins are larger, about 1.73 units of E, and that we use 7 bins per set. Furthermore, we only calculate s for values of ϵ in the crossover region. We show in Fig 10 the resulting $\langle \epsilon \rangle$ for these two lattices, as well as that for the 2^4 lattice. Notice the considerable movement of the crossover region between 2^4 and 4^4 lattices. Fig 11 compares our 6^4 data with that obtained by traditional Monte Carlo.[17,18] The agreement is good, within statistical errors, which, in our data, show up as

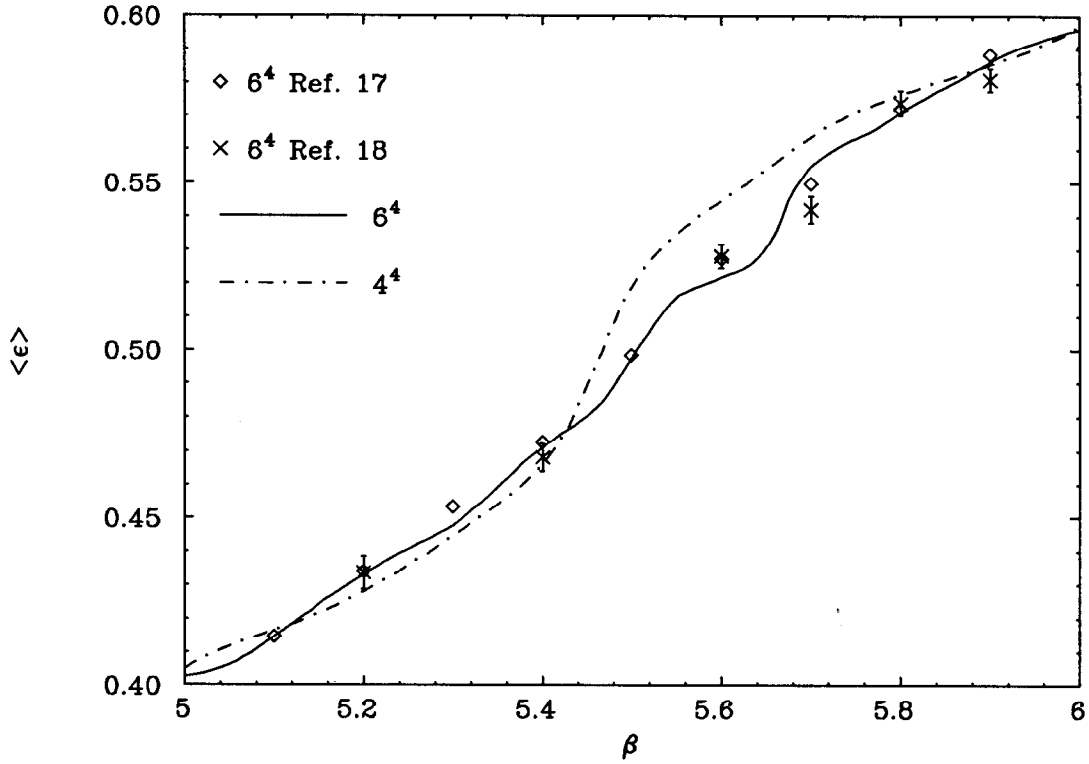

Fig. 11. Comparison of results for $\langle\epsilon\rangle$ between our results and data from traditional Monte-Carlo calculations, both on 6^4 lattices. Also shown is our data on a 4^4 lattice.

the wiggles. Together with the comparisons with analytic expansions, this means that we have checked our method over the entire range of ϵ. The 4^4 data are also shown on the graph to indicate how they differ from the 6^4 data.

As one moves along the sequence of L^4 lattices to larger L, we do not expect a phase transition to appear for non-zero β. Thus, away from the origin, the zeroes should not approach the real axis as $L \to \infty$. Were there a phase transition, on the other hand, we would expect the zeroes closest to the real axis, β_0, to behave as $\mathrm{Im}(\beta_0) \simeq L^{-1/\nu}$. We have checked this for $L = 2$, 4 and 6. We show in Fig. 12 that the data from these three lattices are actually compatible with $\nu \approx 2.5$. This rules out a first order bulk transition ($\nu = 4$), but cannot exclude a lower order transition. To do that we would need to calculate for larger L, which is beyond

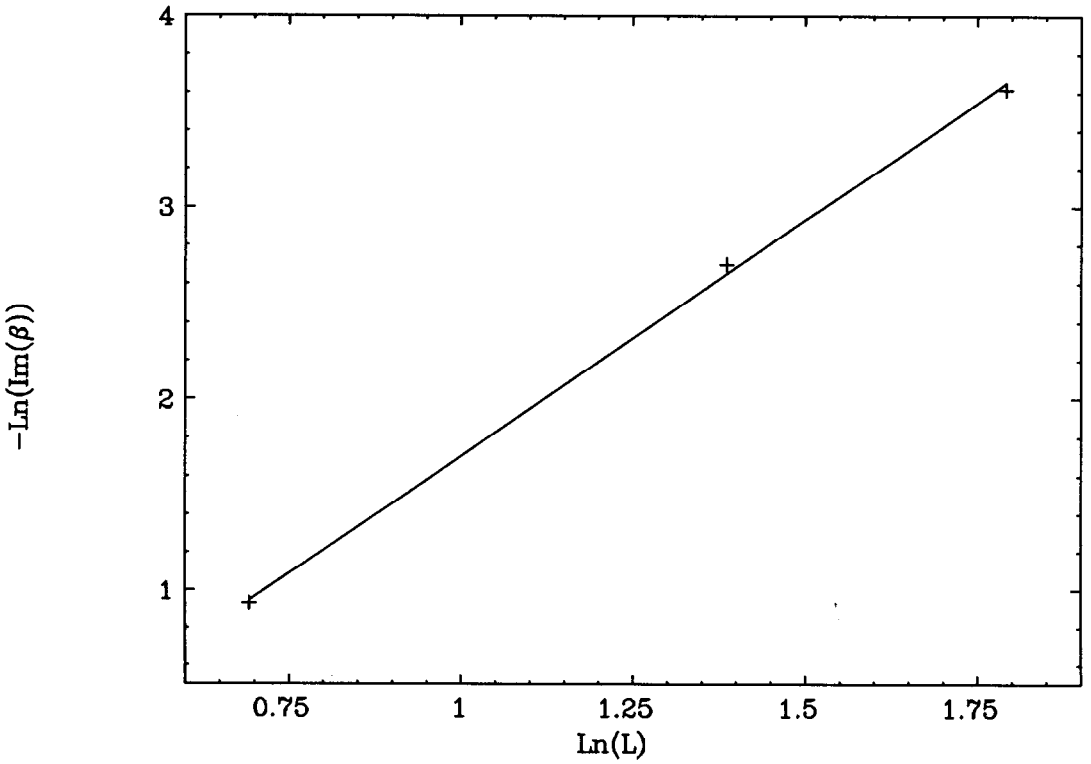

Fig. 12. The scaling of the imaginary part of the zero closest to the real axis in the β plane is shown by plotting $\ln[\mathrm{Im}(\beta_0)]$ versus $\ln L$ for L^4 lattices with $L = 2, 4, 6$. The curve shows a best fit to a straight line, and has slope –2.46.

our present resources.

To investigate the continuum limit of zero temperature SU(3) one would have to study the behavior of the zeroes with large $\mathrm{Re}(\beta)$ as $L \to \infty$. In the u plane, these are the zeroes closest to the origin. As pointed out in Ref. 7, one can deduce the beta-function from the way in which the zeroes closest to the origin move with L. However, as pointed out above, it is very hard to study these zeroes numerically because of the steepness of $s(\epsilon)$, and it becomes progressively more difficult as L increases. It seems to us that traditional numerical methods of calculating the non-perturbative $\beta-$function are to be preferred.

Because of this, we have concentrated our attention on the finite temperature

phase transition of SU(3). We can study this by considering asymmetric lattices: $L_s^3 \times L_t$, where $L_s > L_t$. It is well established that such systems undergo a first-order deconfining phase transition for $L_s \to \infty$. As noted earlier, such transitions are expected to a have a particularly clean signature in the form of finite size scaling, Eq. (3.4). We have made our most detailed study for $L_t = 2$. This means that we are looking at the infinite volume limit of a strong coupling lattice system, with no pretence that we are taking the continuum limit. Nevertheless, this system is known to have a first order phase transition, and thus it provides a good testing ground for our method.

We have taken high statistics data for $L_s^3 \times 2$ lattices for $L_s = 6$, 8, 10 and 12. Given that the system is expected to undergo a first-order phase transition, there will be hysteresis effects. To control these, we employ a three-pronged attack. First, we make the sets as large as possible. Second, we concentrate most of our events in the region of ϵ where both phases are coexisting. Third, we do both cooling and heating runs. Here heating means that we step through the sets in the direction $\epsilon : [1 \to -1/3]$, while for cooling the direction is $\epsilon : [-1/3 \to 1]$. In practice it turns out that the cooling and heating data for $s(\epsilon)$ differ only in the coexistence region. We increase the statistics in this region until the difference between cooling and heating runs is reduced to a tolerable level. We then splice this data onto that from lower statistics runs on either side of the coexistence region. When $s(\epsilon)$ obtained by this procedure is plotted on a scale showing its full span (such as in Fig. 6), the cooling and heating runs are virtually indistinguishable. However, what appears in the partition function for a given β is the exponential of the free energy $F = Vf$. Thus it is better to plot $f = s + \beta\epsilon$ with β chosen to be in the critical region. Furthermore, as discussed above, only a finite range of ϵ is important. Thus, in Fig. 13, we plot the free energy densities for $\beta = 5.088$ on an expanded scale. This value of β is chosen to make $f(\epsilon)$ as flat as possible near the maximum. The differences between the cooling and heating runs are now clear.

There are various points to notice in Fig. 13. As expected, the curves have nearly the same shape for all L_s. The width of the flat regions, which corresponds

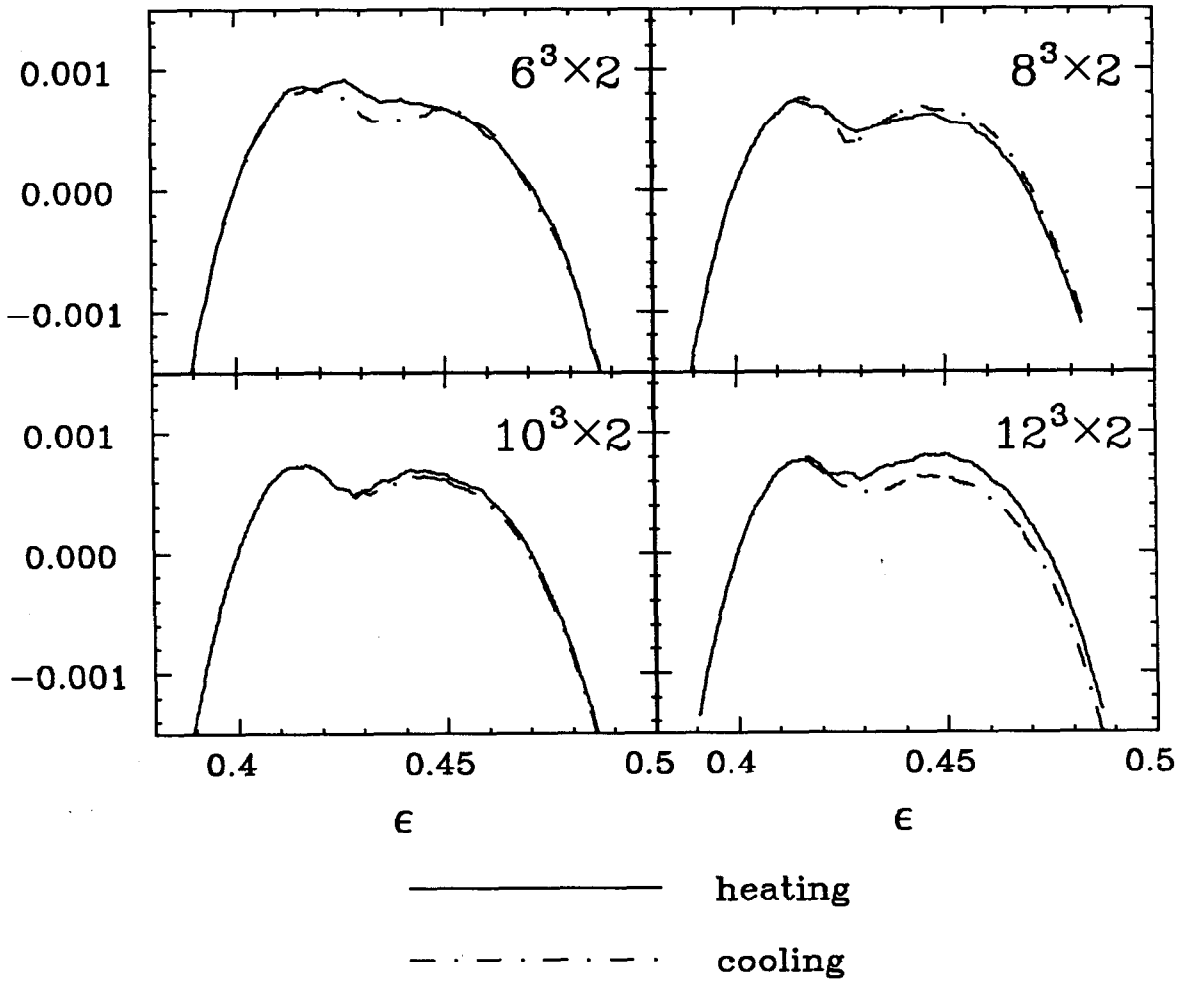

Fig. 13. Free energy density, $f(\epsilon) \equiv s(\epsilon) + \beta$, $\beta = 5.088$ on $L_s^3 \times 2$ lattices, $L = 6, 8, 10, 12$. Continuous and dash-dot lines denote heating and cooling runs, respectively. All the curves are normalized to pass through zero at $\epsilon = 0.4$.

to the jump in $\langle\epsilon\rangle$ across the transition, is almost the same for all curves. All the curves show a dip at $\epsilon \sim .43$, the expected finite volume concavity in the entropy density. For $L_s > 6$, the depth of the dips decreases with increasing L_s. This decrease appears to be slower than $1/V$, so that the free energy F has a dip which

deepens as L_s increases. There are also differences in the relative heights of the two peaks. These correspond to differences in the critical β, and we discuss these differences below.

The statistical errors in the data are evident from the wiggles in the curves. We estimate the expected statistical errors using Eq. (2.8), with $\mathcal{F} = 400$. The results are given in Table 1, together with the parameters of the runs. For each L_s we have runs in both inner and outer ranges of ϵ. It is only in the inner range that hysteresis is important, as we have checked explicitly. Thus we need only consider a single run in the outer range, while we need both a heating and a cooling run in the inner range. Splicing the data together yields the polynomials for which we find the zeroes, the order of which is also given in Table 1.

For each L_s we quote two errors. $(\delta R/R)_{inner}$ is the estimated statistical error propagated across the inner range. The corresponding error in f is given by

$$(\delta f)_{inner} = \ln\left(1 + (\delta R/R)_{inner}\right)/V. \tag{5.3}$$

If f is held fixed at one end of the range, the fluctuations at the other end of the range are given by δf. Half way along the range, the fluctuations are roughly $\sqrt{2}$ smaller. For all L_s, the differences between heating and cooling curves in the vicinity of the dip are larger than $(\delta f)_{inner}/\sqrt{2}$. This confirms what is apparent to the eye, namely that the wiggles in the curves are not large enough to explain the differences between heating and cooling runs. These differences are systematic errors due to hysteresis.

Table 1				
L_s	6	8	10	12
V	2592	6144	12000	20736
S	200	500	800	1400
b	13	13	13	13
bin size	1.439	1.365	1.666	1.646
inner sets	78 - 90	207 - 220	340 - 350	582 - 615
inner ϵ	.400 - .487	.413 - .451	.417 - .435	.414 - .448
inner ev	2.7×10^7	1.9×10^7	5.0×10^7	1.0×10^7
$(\delta R/R)_{inner}$	0.17	0.21	0.11	0.44
$(\delta f)_{inner}$	6.1×10^{-5}	3.1×10^{-5}	$.87 \times 10^{-5}$	1.8×10^{-5}
outer sets	50 - 100	195 - 230	300 - 370	540 - 640
outer ϵ	.333 - .673	.387 - .483	.383 - .502	.390 - .487
outer ev	$.16 \times 10^7$	$.96 \times 10^7$	1.25×10^7	$.51 \times 10^7$
order of $\mathcal{P}$	612	432	852	1212
$\delta R/R$ range	.393 - .493	.397 - .477	.402 - .470	.405 - .465
$(\delta R/R)_{total}$	0.30	0.37	0.39	0.71
$(\delta f)_{total}$	1.0×10^{-4}	$.50 \times 10^{-4}$	$.28 \times 10^{-4}$	$.26 \times 10^{-4}$
CPU hours	3.2	4.2	6.1	3.8

Table 1: Parameters of the runs used in the FSS analysis. The notation is defined in the text. The inner and outer ranges are given both in terms of the sets which they contain, and as ranges of ϵ. ev refers to number of events per set. The "$\delta R/R$ range" is that for which $(\delta R/R)_{total}$ applies. The CPU time is the total for a single run in the outer range and both a heating and a cooling run in the inner range.

The other error quoted in Table 1, $(\delta R/R)_{total}$, is that propagated across the range of ϵ which contributes to $\langle\epsilon\rangle$ for $\beta = 5.088$, in the sense defined above. This range, which is called the "$\delta R/R$ range " in the Table, lies between the inner and outer ranges. The errors are calculated taking into account the differing statistics in inner and outer ranges. The corresponding δf is also quoted. Because the curves in Fig. 13 are all forced to agree at $\epsilon = .4$, which is very close to the beginning of the contributing range for all L_s, δf is the expected fluctuation in the value of f at the upper end of the respective ranges. Even though δf is large enough to be visible in Fig. 13, it is still a very small effect for $\beta \approx \beta_c \approx 5.088$ because it is a shift in a region of the free energy which makes a very small contribution to $\langle\epsilon\rangle$. We stress that for β away from β_c the range of ϵ that contributes is much smaller,

and so the relative error in R propagated across this range is much smaller than that for $\beta = \beta_c$.

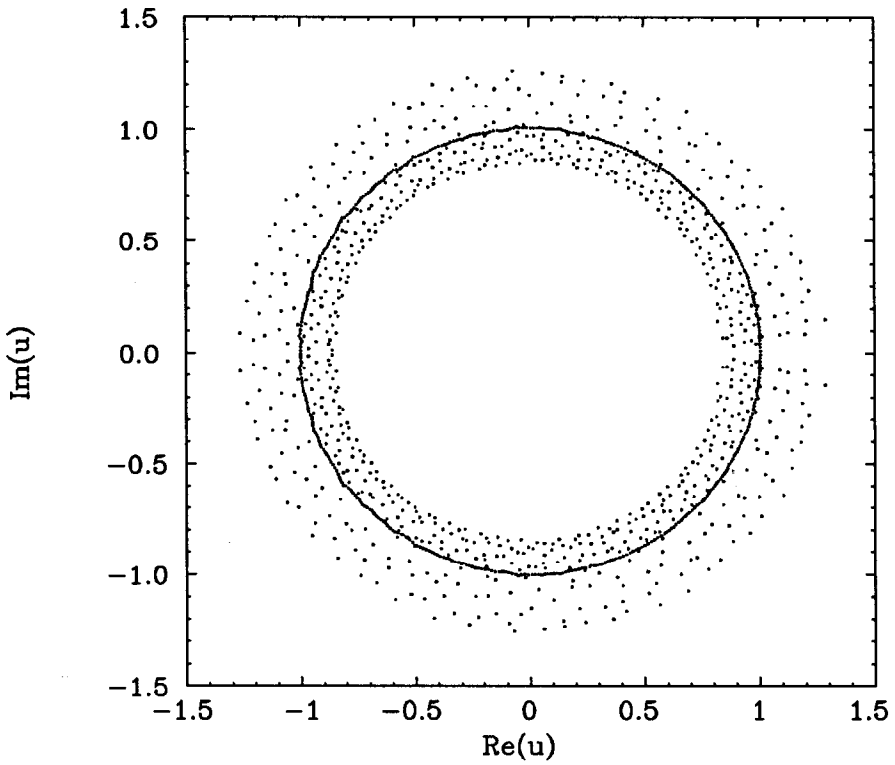

Fig. 14. Zeroes of the $SU(3)$ partition function on a $12^3 \times 2$ lattice plotted on the u plane with $\beta_{eff} = 5.088$. The results are from the heating run, the parameters of which are given in Table 1. The polynomial is of order 1212.

Fig. 13 shows that the shapes of the free energy density curves are nearly independent of L_s, and that f is nearly flat. This provides a qualitative confirmation that the system undergoes a first order transition when $L_s \to \infty$. To obtain a quantitative confirmation we use the movement of the zeroes of the partition function. In Fig. 14, we show the zeroes for the heating run on the $12^3 \times 2$ lattice. Notice how on this large lattice the flat region in f manifests itself as a crisp ring of zeroes in the u plane (cf. Eq. (3.8)). The fact that the zeroes form a band,

rather than covering the whole plane, as in Fig. 8, is due to our truncation of the complete polynomial. This truncation does not effect the zeroes at the center of the band, including those in the ring, though zeroes at the edge of the band are not reliable. For discrete systems like the Ising model, zeroes often lie on lines. In particular, there is a line of zeroes which pinches the real axis. For our case, however, there is a general background of zeroes in which the ring sits, and from which only a few zeroes extend towards the real axis. The figure shows that the ring is not a perfect circle, and that the distribution of zeroes along the circle is not precisely uniform. In other words, Eq. (3.8), which was derived for an exactly flat free energy, is only a rough description of our data.

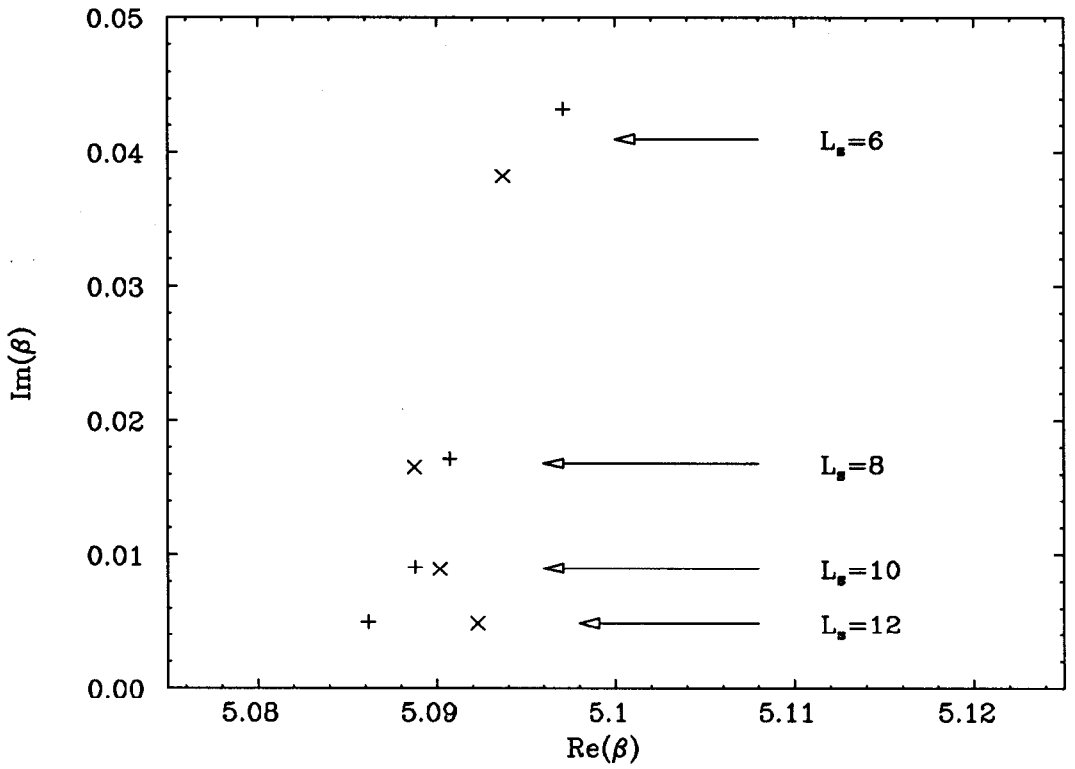

Fig. 15. Zeroes closest to the real β axis on $L_s^3 \times 2$ lattices, $L_s = 6, 8, 10, 12$. For each L_s we show two zeroes corresponding to the cooling ($\times$) and heating ($+$) runs. Note that the scale is the same for both real and imaginary parts.

In view of this, we concentrate on the zero closest to the real axis for our quantitative check of finite size scaling. We give in Table 2 the positions of these zeroes for the eight runs (cooling and heating for $L_s = 6$, 8, 10 and 12). Fig. 15 plots these positions in the complex β plane. The major difference between cooling and heating runs is in the real part of β. What we expect is that a heating run, which starts in the high temperature phase, remains in that phase slightly beyond the transition point, so that the zero has a smaller real part. A cooling run should display the opposite metastability. This is indeed the pattern we observe for $L_s = 10$ and 12. However, for $L_s = 6$ and 8, the opposite behavior is seen. It is still true for these lattices, however, that the cooling run dips down more quickly, and for longer, to the right of the left peak. This is what one would expect from hysteresis. What one would not expect is the sharper rise, in the cooling runs, up to the right peak. We do not understand this behavior, though we are confident that it is a systematic effect, not due to statistical fluctuations.

In view of these uncertainties we can say little about the variation of Reβ with L.[⋆] This is, in any case, not a universal phenomenon. In contrast, the imaginary parts of the zeroes are much less sensitive to hysteresis, since they are related to the latent heat of the transition. Indeed, Imβ is expected to satisfy the universal scaling formula, Eq. (3.4). We test this by fitting $\mathrm{Im}\,\beta_c(L_s)$ to the form:

$$\mathrm{Im}\,\beta_c(L_s) \propto L_s^{-d_{eff}}. \tag{5.4}$$

This corresponds to a straight line on a log-log plot, and we show our data on such a plot in Fig. 16. We assume that the correct answer is the average of the heating

⋆ Ref. 9 gives the data for Reβ_c for $L_t = 2$ from which a FSS analysis can be done. Their results are 5.071, 5.086, 5.092 and 5.0945 for $L_s = 5$, 7, 9 and 11, respectively. These numbers show a different trend from ours, and appear to extrapolate to a higher infinite volume limit. However, we cannot make a direct comparison, because Ref. 9 used helical boundary conditions, whereas we use periodic boundary conditions. Furthermore, the criterion used in Ref. 9 to determine β_c involves the Polyakov line, and differs from ours. Nevertheless, the infinite volume limits should agree. It may be that either we or the authors of Ref. 9 have underestimated the systematic errors.

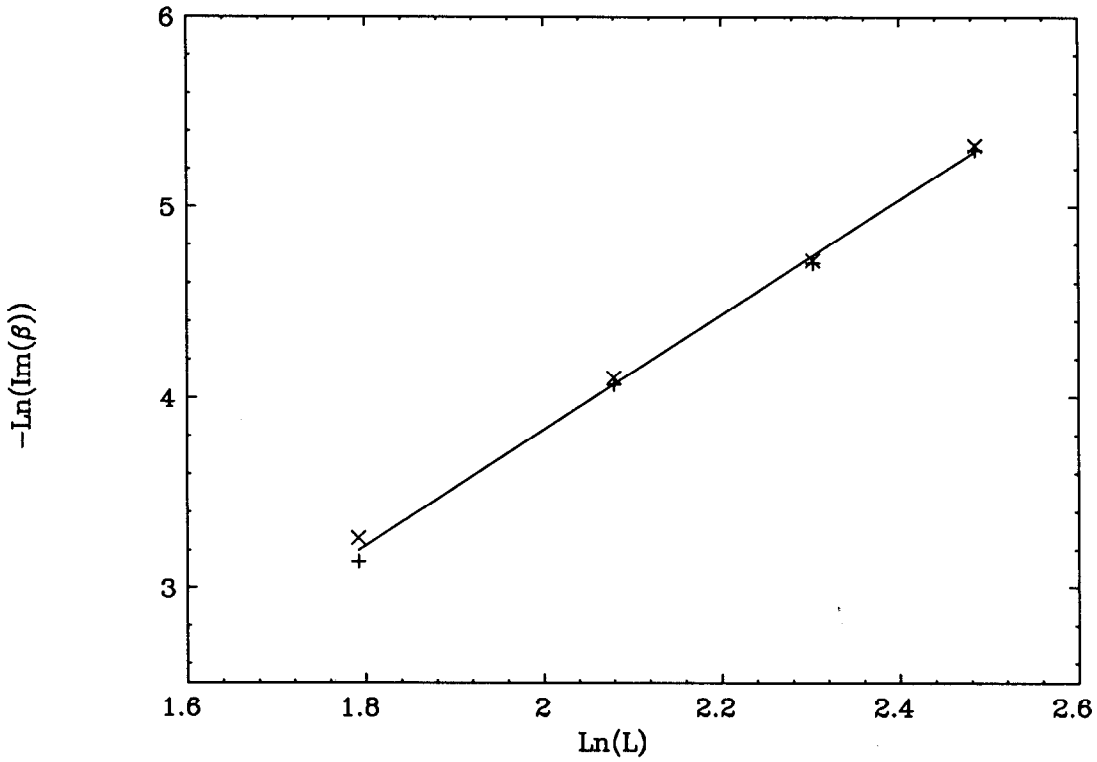

Fig. 16. The scaling of the imaginary part of the zeroes closest to the real axis in the β plane for $L_s^3 \times 2$ lattices with $L_s = 6, 8, 10, 12$. We plot $\ln\left[\mathrm{Im}(\beta_c(L_s))\right]$ *vs.* $\ln(L_s)$. The straight line is the fit corresponding to Eq. (5.4), with $d_{eff} = 3.016$. $(\times)$ and $(+)$ denote cooling and heating runs, respectively.

and cooling results, and therefore we fit to the eight points giving equal weight to each. The error estimate is based on assuming that for each L both the heating and the cooling results are separated by one standard deviation from the central value. The fit yields the result $d_{eff} = 3.02 \pm 0.05$. The expected result for a first order transition is $d_{eff} = 1/\nu = 3.0$. Thus our data provides a good quantitative confirmation that the $SU(3)$ deconfining transition is first order.

Table 2		
L_s	Sweep direction	β_c
6	H	$5.0971 + 0.04325i$
	C	$5.0937 + 0.03823i$
8	H	$5.0907 + 0.01714i$
	C	$5.0888 + 0.01652i$
10	H	$5.0888 + 0.009085i$
	C	$5.0902 + 0.008926i$
12	H	$5.0862 + 0.004993i$
	C	$5.0923 + 0.004879i$

Table 2: Partition function of $SU(3)$ on a $L_s \times 2$ lattice: zeroes closest to the Re β axis as function of L_s. The two measurements for each L_s are from a heating and a cooling run, respectively.

6. CONCLUSIONS

The calculation of the zeroes of the partition function by measuring the spectral density is a novel and powerful tool with which to study statistical systems and their phase transitions.

The partition function as function of temperature is defined in the whole complex plane, but with traditional Monte Carlo methods one can only simulate statistical systems at real temperature or coupling. Even though in the laboratory there are no complex temperatures, it is important to realize that what is observed at real temperature results from the analytic structure in the complex plane. A useful analogy is a three-dimensional object which is first observed from one particular angle. What one sees is a two dimensional projection, with many features completely obscured, until one is given a chance to observe the full three dimensional structure.

Thus the singularities which, in the thermodynamic limit, manifest themselves as phase transitions at real temperature, exist even in relatively small systems, albeit in the complex plane. In addition, many phenomena which are observed on the real axis, such as the rapid crossover from strong to weak coupling, can be traced to the complex zeroes of the partition function.

It is only with the advent of the spectral density method (SDM)[6,7] that the study of the complex analytic structure for complicated systems has become feasible.

By improving upon the method of Refs. 6,7 we have been able to carry out a high-precision FSS analysis of the rounding of the finite temperature deconfinement transition for pure gauge $SU(3)$.

What is the future of the spectral density method? For the remainder of this section we explain our tentative answers to this question. Traditional methods are probably favored for mapping out phase boundaries, for identifying strong first order transitions, and for calculating correlation functions. On the other hand, the SDM is the method of choice for detailed quantitative studies of phase transitions. Thus we imagine the SDM as a tool to be used once traditional methods have mapped out the interesting range of coupling constants.

An important exception are systems with complex actions. These have been very hard to study using existing methods. From the standpoint of the spectral density method, however, there is no fundamental difficulty. After all, in finding the zeroes we are using a complex action already. It will be very interesting to see how the spectral density method fares on some simple examples.

The main reason why we think that the SDM should be used only for detailed studies is the rapid increase in the computer time needed as V is increased. As we discussed in section 2, the time required grows as V^2 for a first order transition. It is simple to generalize this to higher order transitions: the time scales as $V^{2/\nu d}$, where d is the dimension in which the transition is occurring.* For a first order transition $\nu d = 1$, while for second order transitions $1 \leq \nu d \leq 2$. Thus at the boundary between second and third order transitions the time required scales as V.

$\star$ This scaling law is derived by assuming that one need only measure s for those ϵ which contribute, in the sense defined in section 5, to the partition function at the critical coupling.

These scaling laws should be compared with that for traditional Monte-Carlo methods, for which the time required increases as V. For first and second order transitions, it seems naively as though the SDM will be slower for large V, with the discrepancy getting smaller as the order of the transition is decreased. However, this comparison is somewhat misleading. The scaling law for traditional Monte-Carlo assumes that one need simulate only for a fixed number of couplings as V increases. This may be true if one is interested only in a qualitative picture of the phase structure, or in calculating correlation functions. But, as we argue below, to make a quantitative study of a phase transition one may need to use more couplings for larger V, making the time required the same for all methods. If so, then the SDM is favored, because it allows the most straightforward calculation of the critical exponents.

That the spectral density method may be competitive for detailed studies can be seen as follows. Operationally, the procedure we use has close similarities with both canonical and microcanonical methods. The energies are restricted to a small range, as with a microcanonical method, but within each set we generate a canonical ensemble. To determine the order of the transition one needs to know, directly or indirectly, the functional form of $s(\epsilon)$ close to the critical point. One does this in the canonical method by scanning in β, while in the SDM and microcanonical methods one scans in ϵ. Whichever method one uses, however, one is collecting data from the same range of ϵ. The different methods are simply packaging this data in different ways. Thus, even though it is hard to compare the various methods directly, one should need roughly the same amount of data for all methods.

It is worthwhile illustrating these general arguments with a specific example. Consider a system with a strong first order transition. To search for this transition in the canonical ensemble, one looks for flip-flops between states and/or coexistent phases. In a microcanonical approach one searches for S-shapes in the plot of $\langle\epsilon\rangle$ versus β.[21] The corresponding phenomena in the SDM are the flat free energy and the concavity in the entropy. Taken together, they provide a clear qualitative signal for a strong first order transition. However, one needs to calculate the

density of states across the *entire* flat region in order to get this signal, and it is this which makes the time required grow like V^2. Most of the time is spent sampling configurations containing both phases in varying proportions. This is an overkill if one simply wants to establish the existence of the transition. In fact, one need only sample configurations containing wholly one phase or the other, and show that they can coexist, as is done in the canonical ensemble. This is why traditional methods are favored for identifying strong first order transitions. But to investigate the rounding of the transition, as done in this paper, one needs to know the relative proportions of phases as a function of β. Thus, directly or indirectly, one must fully investigate the flat region, and this will take roughly the same time with all methods. The SDM is then favored because of its quantitative measurement of the rounding, as illustrated by the results presented here.

As part of a detailed study of a phase transition it may be necessary to make the SDM more sensitive by using multi-dimensional bins. For example, in our $SU(3)$ calculation we could have made use of the Polyakov loop. This is the order parameter for the deconfinement transition.[19] In conventional Monte Carlo calculations it has been essential for determining the position of the transition on large lattices.[9,13,20] It can be included in the spectral density approach by using two dimensional bins in the energy—Polyakov line plane. In each such bin the weighting function would use an effective β, and a source term coupling to the Polyakov line. The main advantage of two-dimensional binning would be an improvement in the sampling of configuration space. In particular, there are three components of the high temperature phase, characterized by different values of the phase of the Polyakov line. If one bins only in ϵ, as we did above, the system has to wander between these phases on its own. Using two-dimensional bins, one can force the system to be in one or other of the phases. This will reduce the systematic errors, though inevitably at the expense of increasing the amount of computer time needed.

There are two systems for which the sensitivity of the SDM may be needed, and for which two dimensional bins may be essential. Both inhabit the murky area

at the boundary of first and second order transitions. The first is the $U(1)$ pure gauge theory in the fundamental/adjoint coupling constant plane. The nature of the line of phase transitions remains controversial despite much work. To use the SDM, one should tile, with two dimensional bins, the critical region of the coupling constant plane. We are beginning a study of this model.

The second theory is QCD with two or three light dynamical fermions. One would have to use two dimensional bins in the energy—$\langle\overline{\psi}\psi\rangle$ plane. In each such bin one would use an effective β and an effective quark mass. In at least some existing algorithms, e.g. the exact algorithm of Ref. 22, it is straightforward to calculate $\langle\overline{\psi}\psi\rangle$ after each link is changed, as would be necessary. This is a very challenging project, which should probably wait until two dimensional binning has been tested on a simpler system. Nevertheless the potential rewards are very large: one would be able to study the transition for all quark masses at once.

Note added:
After the completion of this work we received a paper by K. Bitar (FSU-SCRI-87-33) in which the method of Ref. 7 is applied to $SU(2)$ at finite temperature.

Acknowledgements:

We thank Dick Blankenbecler, Rajan Gupta, Enzo Marinari, Bob Pearson, Michael Peskin and Bob Sugar for helpful conversations, and Tom Banks and Yosef Nir for comments on the manuscript. The numerical work was done at the MFE computing center using time granted by the DOE.

APPENDIX A

This appendix serves three purposes. First, it gives a more rigorous derivation that the weighted random walk yields the correct density of states. Second, it explains how we optimize the parameters of our hit matrix. And, third, it fills in the details of our discussion of statistical errors.

To discuss the justification for our method, we first imagine that there are no restrictions on the energy. Assuming that the algorithm is ergodic, imposition of the Metropolis step means that configurations appear with probability proportional to the weight $W(\epsilon)$. The distribution with respect to ϵ will thus be $N(\epsilon)W(\epsilon)d\epsilon$. Let the hit matrix, which moves from one configuration (U) to the next (U') be $C(U,U')$. It is a probability distribution:

$$\int [dU']C(U,U') = 1 \ ,$$

and, for simplicity, we take it to be symmetric. The probability that a configuration of energy density ϵ_1 will jump to ϵ_2, is given by the kernel

$$K(\epsilon_1,\epsilon_2) = \frac{1}{N(\epsilon_1)} \int [dU]\delta(\epsilon_1 - \epsilon) \int [dU']C(U,U')\min\left[1, \frac{W(\epsilon')}{W(\epsilon)}\right] \delta(\epsilon_2 - \epsilon') + \frac{\delta(\epsilon_1 - \epsilon_2)}{N(\epsilon_1)} \int [dU]\delta(\epsilon_1 - \epsilon) \int [dU']C(U,U')\max\left[0, 1 - \frac{W(\epsilon')}{W(\epsilon)}\right]. \tag{A.1}$$

Here ϵ is the energy density of the configuration U, ϵ' that of configuration U'. The second term represents rejection in the Metropolis step. The kernel acts on a probability distribution $D(\epsilon_1)$ representing an ensemble of Monte Carlo simulations, probability conservation being guaranteed by

$$\int d\epsilon' K(\epsilon,\epsilon') = 1. \tag{A.2}$$

It is convenient to define the hit matrix in energy density space:

$$\tilde{C}(\epsilon_1,\epsilon_2) = \int [dU] \int [dU']\delta(\epsilon_1 - \epsilon)C(U,U')\delta(\epsilon_2 - \epsilon').$$

This is still symmetric, and satisfies

$$\int [d\epsilon_2] \tilde{C}(\epsilon_1, \epsilon_2) = N(\epsilon_1).$$

We can now rewrite the kernel

$$\begin{aligned} K(\epsilon_1, \epsilon_2) = & \frac{1}{N(\epsilon_1)W(\epsilon_1)} \min\Big[W(\epsilon_1), W(\epsilon_2)\Big] \tilde{C}(\epsilon_1, \epsilon_2) \\ & + \frac{\delta(\epsilon_1 - \epsilon_2)}{N(\epsilon_1)W(\epsilon_1)} \int [d\epsilon'] \max\Big[0, W(\epsilon_1) - W(\epsilon')\Big] \tilde{C}(\epsilon_1, \epsilon'). \end{aligned} \tag{A.3}$$

In this form the symmetry of K is manifest:

$$N(\epsilon_1)W(\epsilon_1)K(\epsilon_1, \epsilon_2) = N(\epsilon_2)W(\epsilon_2)K(\epsilon_2, \epsilon_1). \tag{A.4}$$

It is straightforward to verify that the ensemble distribution $D(\epsilon_1) = N(\epsilon_1)W(\epsilon_1)$ is an eigenvector of this kernel with eigenvalue 1.

We now consider energies confined within a set. The kernel taking such confinement into account, K^{con}, is defined implicitly by:

$$D_{n+1}(\epsilon) = \int_{in} D_n(\epsilon') K^{con}(\epsilon', \epsilon) = \int_{in} d\epsilon' D_n(\epsilon') K(\epsilon', \epsilon) + D_n(\epsilon) \int_{out} d\epsilon' K(\epsilon, \epsilon'). \tag{A.5}$$

$D_n(\epsilon)$ is the ensemble probability distribution after the $n-$th application of K^{con}, with ϵ is restricted to be within the set. Integrals labelled "*in*" and "*out*" run over energies within and outside the set, respectively. The last integral represents the rejection of a change when the energy goes outside the set. K^{con} satisfies the same symmetry property as K, Eq. (A.4).

Using the fact that the original kernel conserves probability, we can rewrite (A.5) as

$$D_{n+1}(\epsilon) - D_n(\epsilon) = \int_{in} d\epsilon' \Big[D_n(\epsilon') K(\epsilon', \epsilon) - D_n(\epsilon) K(\epsilon, \epsilon')\Big].$$

It is straightforward to verify that the r.h.s. vanishes if $D_n(\epsilon) \propto N(\epsilon)W(\epsilon)$, i.e. this form is an eigenvector of K^{con} with eigenvalue 1. All other eigenvalues have

modulus less than one. Since equation (A.5) represents the procedure that we actually use, it is evident that no complications are introduced by restricting oneself to a limited range of ϵ, or by using a weighting function.

We next turn to the optimization of our hit matrix C. We want to move as quickly as possible through configuration space, but to do so we must balance two competing effects. Clearly, if we make the hit matrices differ more from the unit matrix each step is larger. On the other hand, large steps are more likely to be rejected by the Metropolis criterion, and/or by the requirement that the energy stay within the set. If too many steps are rejected we move more slowly through configuration space. The criterion we use to optimize the step size is that the second largest eigenvalue of the confined kernel be minimized. The second largest eigenvalue, λ_1, represents the dominant transient effect. After a number of events n such that $\lambda_1^n \ll 1$, the probability distribution for events is essentially independent of the starting position, i.e. each such event is independent. For λ_1 such that $1 - \lambda_1 \ll 1$ a rough estimate of the number of events needed is $n \sim 1/(1 - \lambda_1)$. By minimizing this number, we minimize correlations, and thus minimize the statistical errors in our result.

To calculate λ_1, we partition the set into many energy subintervals, each much smaller than the bins, and collect the matrix K as a histogram. The second eigenvalue and eigenvector are determined by numerical iteration of equation (A.5), having projected against the leading eigenvector. Let δ denote the "hit size", i.e. the parameter(s) determining the distance of the hit matrices from unity. For a given set we minimize λ_1 with respect to δ. This procedure is repeated for a series of different sets in the range of interest, and for different lattice sizes. We find that δ_{min} varies significantly between sets, while near the minimum λ_1 is a very shallow function of δ. Thus we can choose one value of δ which is nearly optimal for all sets of interest.

We will present numerical results for a "typical" set in the region of the phase transition. This set contains b bins of size 1.5 units of E. To a very good approxima-

tion the density of states varies exponentially within this set, and by construction our weighting functions are pure exponentials:

$$N(\epsilon) \propto \exp(-V\beta_{set}\epsilon); \qquad W(\epsilon) \propto \exp(V\beta_W\epsilon).$$

If $\epsilon_k = k\Delta\epsilon + \text{const.}$ are the energies at the centers of the bins, then we can characterize the weighted density of states, by the parameter x

$$D_\infty(\epsilon_k) = N(\epsilon_k)W(\epsilon_k) \propto x^k; \qquad x = \exp\left[V(\beta_W - \beta_{set})\Delta\epsilon\right].$$

In this notation, perfect weighting corresponds to $x = 1$. We take our typical set to have $\beta_{set} = 5.0$, so that with no weighting $x = 1/1808$. Partial weighting corresponds to values of x between these extremes.

Fixing the hit matrix parameters to their optimal values for $x = 1$, we have calculated λ_1 for a variety of values of x and b. The results for $b = 4$ are given in the first 6 rows of Table 3 (only the first and third columns are relevant for the moment). For $x = 1$ the result is $\lambda_1 \simeq .9965$, so it takes about $1/(1-\lambda_1) \sim 300$ events to decorrelate, with each event moving $\sim 6/\sqrt{300} \sim 0.35$ units of energy. As we increase b we find to very good accuracy that $1/(1-\lambda_1) \propto b^2$, as expected from a random walk. Thus for 13 bins of size 1.5, roughly the largest set we use, $\lambda_1 \simeq .99965$, so it takes 3000 events to decorrelate. As x is decreased, λ_1 decreases to 0.825, its value for no weighting. For all $x \ll 1$, we find that λ_1 is nearly independent of b. Thus, when there is little or no weighing, it takes only $5-10$ events to decorrelate, however large the set. This is because the non-leading eigenvector, like the leading eigenvector, is concentrated in a small, fixed range of ϵ at one end of the set. Given this quick decorrelation, one might be concerned that the unweighted method is competitive with that using weighting. The remainder of the appendix is intended to allay such concerns.

Our measurements are not made from an ensemble, but rather by averaging along a particular Markov chain. To calculate the dispersion, we need to average

over all possible chains. In order to make the analysis more simple we break up the continuous energy interval of the set into bins. K^{con} is then represented by a "transfer" matrix T_{ij} which gives the probability of moving from bin i to bin j. Although the following analysis goes through whatever the size of the bins, we will apply the results to bins of the size that we use in our numerical work. Thus we will continue to use b for the number of bins per set, and to characterize the weighted density of states by the parameter x introduced above. In general, x is the amount by which the weighted density of states decreases from the center of one bin to that of the next.

The transfer matrix must have two properties. Conservation of probability requires

$$\sum_j T_{ij} = 1. \tag{A.6}$$

T must also satisfy the same symmetry condition as K^{con}, equation (A.4). The required symmetry property is (no sum on indices)

$$x^i T_{ij} = x^j T_{ji}. \tag{A.7}$$

These two properties insure that x^i is a left eigenvector of T with eigenvalue 1. The corresponding right eigenvector has all components equal.

To calculate the errors we introduce the generating function of chains

$$Z_i(\vec{\alpha}) = \sum_j \Big(M\big[TM\big]^{n-1} \Big)_{ij}; \qquad M \equiv diag\big[e^{\alpha_1}, \ldots, e^{\alpha_b}\big]. \tag{A.8}$$

Z_i is a sum over all possible chains of length n, which start at the $i-$th bin:

$$Z_i(\vec{\alpha}) = \sum_{\mathcal{C}} P(\mathcal{C}) \exp\Big[\sum_k \alpha_k n_k(\mathcal{C})\Big] \tag{A.9}$$

In the sum each chain $\mathcal{C}$ is weighted by its probability $P(\mathcal{C})$, while $n_k(\mathcal{C})$ denotes the number of times the $k-$th bin appears in $\mathcal{C}$ and $\vec{\alpha}$ represents the α_i collected

into a vector. In terms of Z_i we have

$$\begin{aligned} f_k &\equiv \frac{\langle n_k \rangle}{n} = \frac{1}{n} \frac{\partial \ln Z_i(\vec{\alpha})}{\partial \alpha_k}\Big|_{\vec{\alpha}=0}; \\ C_{kl} &\equiv \frac{\langle n_k n_l \rangle - \langle n_k \rangle \langle n_l \rangle}{n^2} = \frac{1}{n^2} \frac{\partial^2 \ln Z_i(\vec{\alpha})}{\partial \alpha_k \partial \alpha_l}\Big|_{\vec{\alpha}=0}. \end{aligned} \tag{A.10}$$

Introducing the diagonal matrix X, $X_{ii} = x^i$, we can construct a symmetric matrix $\widetilde{T}$

$$\widetilde{T}(\vec{\alpha}) = M^{\frac{1}{2}} X^{\frac{1}{2}} T X^{-\frac{1}{2}} M^{\frac{1}{2}},$$

in terms of which

$$Z_i(\vec{\alpha}) = M^{\frac{1}{2}} X^{-\frac{1}{2}} \left[\widetilde{T}(\vec{\alpha})\right]^n X^{\frac{1}{2}} M^{\frac{1}{2}}.$$

Let the eigenvalues of $\widetilde{T}(\vec{\alpha})$ be $\lambda_i(\vec{\alpha})$, with $\lambda_0(\vec{\alpha})$ the largest in magnitude. Then, for large n, one can show that

$$f_k = \frac{\partial \ln \lambda_0(\vec{\alpha})}{\partial \alpha_k}\Big|_{\vec{\alpha}=0}; \qquad C_{kl} = \frac{1}{n} \frac{\partial^2 \ln \lambda_0(\vec{\alpha})}{\partial \alpha_k \partial \alpha_l}\Big|_{\vec{\alpha}=0}, \tag{A.11}$$

up to corrections of $O(1/n)$ and $O(\lambda_1^n/\lambda_0^n)$.

To evaluate Eq. (A.11) we need λ_0 for small $\vec{\alpha}$. This can be obtained using perturbation theory in $\widetilde{T} - \widetilde{T}(\vec{\alpha} = 0)$. We have constructed $\widetilde{T}(\vec{\alpha} = 0)$ to have the same eigenvalues as T, so that $\lambda_0(\vec{\alpha} = 0) = 1$. First order perturbation theory yields

$$f_k = \frac{x^k}{\sum_{l=1}^{b} x^l}.$$

f_k is the leading left eigenvector of T, which shows that one gets the same answer from averaging over chains as one does from an ensemble average. Second order

perturbation theory gives

$$C_{kl} = \frac{1}{n}\left(\delta_{kl} f_k - f_k f_l + 2\sqrt{f_k f_l}\sum_{p=1}^{b-1}\frac{\lambda_p}{1-\lambda_p}\psi_k^{(p)}\psi_l^{(p)}\right). \qquad (A.12)$$

Here, $\psi^{(p)}$ is the $p-$th eigenvector of $\widetilde{T}(\vec{\alpha}=0)$, with eigenvalue λ_p, and the sum over p excludes the leading eigenvector $p=0$. As required, $\sum_k f_k = 1$, and $\sum_k C_{kl} = 0$, i.e. the fractions sum to 1 with no dispersion.

The first term in Eq. (A.12) is the dispersion expected for uncorrelated events. The second term is the correlation in the dispersion introduced by the constraint $\sum_k f_k = 1$. The final term contains the effects of the correlations between events. If there are no such correlations, then $\lambda_p = 0$ for $p \geq 1$, and the last term vanishes. As the correlations increase, and in particular as $\lambda_1 \to 1$, this term increases because of the factor $1-\lambda_p$ in the denominator. It turns out that, for our typical set, this term is completely dominant for $x = 1$, and gives the largest contribution for $x \ll 1$.

We apply this formalism to transfer matrices which act on bins of our typical size, 1.5 units of E. This gives us directly the expected statistical errors on the numbers we measure. By working with such large bins we are ignoring the detailed statistical fluctuations of the data, but these are of little interest. Since each event moves an energy much smaller than the bin size, these "coarse-grained" transfer matrices only contain entries for movement to the same bin or to adjacent bins. T should also be independent of the position within the set, except for edge effects, since the sets cover only a tiny range of ϵ. These requirements, together with equations (A.6) and (A.7), determine T uniquely to be

$$T(x,\Lambda)_{i,i+1} = \frac{(1-\Lambda)x}{2}, \quad T(x,\Lambda)_{i+1,i} = \frac{(1-\Lambda)}{2}: \quad i = 1, b-1;$$

$$T(x,\Lambda)_{1,1} = 1 - \frac{(1-\Lambda)x}{2}; \quad T(x,\Lambda)_{b,b} = 1 - \frac{1}{2}:$$

$$T(x,\Lambda)_{i,i} = 1 - \frac{(1-\Lambda)(1+x)}{2}: \quad i = 2, b-1.$$

Only non-zero elements are shown. For each x there is a single parameter Λ. This we adjust so that the λ_1 matches onto that found by the numerical method described above. This gives us a coarse-grained kernel which has the same leading eigenvector and the same dominant transient effects as the continuous kernel we actually use.

The kernels $T(x,\Lambda)$ are sufficiently simple that we can evaluate (A.12) analytically. In particular, the eigenvalues are

$$(1-\lambda_p) = (1-\Lambda)\left(-\cos(p\pi/b)\sqrt{x} + \frac{1+x}{2}\right); \qquad p = 1, b-1. \tag{A.13}$$

Notice that for $x = 1$, and fixed Λ, $1-\lambda_1 \propto 1/b^2$ for large enough b. This is the same behavior as shown by the numerically determined kernels. Thus, Λ can be kept constant as b is changed. This is physically reasonable, since Λ determines the size of the off-diagonal elements of T, and these should be independent of b. Eq. (A.13) shows that for $x \ll 1$ all the eigenvalues collapse to a common value, $1-\lambda_p = (1-\Lambda)/2$, independent of b. This is again in accord with the numerically determined results for λ_1, as long as Λ is held fixed as b is varied.

We use these kernels to evaluate the error in R, the ratio of events in the edge bins of a range of B bins. For a given x we fix Λ once and for all by matching the numerically determined value of λ_1 for $b = 4$. We then vary b, and also n_{in}, the period with which we record events. It is straightforward to include the effect of $n_{in} \neq 1$ in the above formalism. The result for the dispersion in R, for N total events, is

$$\begin{aligned} \frac{\langle(\delta R)^2\rangle}{R^2} &\equiv \frac{B^2}{N}\mathcal{F}(b, n_{in}) \\ \mathcal{F}(b, n_{in}) &= \frac{n_{in}}{(b-1)^2}\left(\frac{1}{f_1} + \frac{1}{f_b} + 2\sum_{p=1}^{b-1}\frac{\lambda_p^{n_{in}}}{1-\lambda_p^{n_{in}}}\left(\frac{\psi_1^{(p)}}{\sqrt{f_1}} - \frac{\psi_b^{(p)}}{\sqrt{f_b}}\right)^2\right), \end{aligned} \tag{A.14}$$

up to corrections of $O(1/B)$. The figure of merit $\mathcal{F}$ should be as small as possible.

A selection of the results are given in Table 3. These are all for typical bins of width 1.5 units of E, and $\beta_{set} = 5.0$. The Table is broken up into five sections illustrating the effects of: (1) reducing and eventually removing the weighting (no weighting corresponds to $1/x = 1808$), for $b = 4$ and $n_{in} = 1$; (2) increasing the number of bins holding $x = 1$ (perfect weighting); (3) increasing n_{in} holding $x = 1$; (4) increasing the number of bins with no weighting; and (5) increasing n_{in} with no weighting. These results are discussed in section 2.

Table 3					
$1/x$	Λ	λ_1	n_{in}	b	$\mathcal{F}$
1	.988	.9965	1	4	4.5×10^2
2	.982	.9956			6.6×10^2
10	.953	.9848			1.2×10^4
100	.834	.928			2.6×10^6
1000	.627	.825			1.1×10^9
1808	.638	.825			6.6×10^9
1	.988	.9965	1	4	445.42
		.99882		7	390.14
		.99965		13	362.45
1	.988	.9965	1	4	445.42
		.9930	2	4	445.43
		.9635	10	4	445.51
		.7043	100	4	454.03
		.3493	300	4	511.19
		.1220	600	4	659.83
		.0149	1200	4	1094.2
1808	.638	.825	1	4	6.6×10^9
		.827		7	9.8×10^{19}
		.827		13	3.7×10^{37}
1808	.638	.825	1	4	6.61×10^9
		.681	2	4	6.67×10^9
		.146	10	4	8.64×10^9
		.021	20	4	1.36×10^{11}
		.0005	40	4	2.63×10^{11}

Table 3: Results for the figure of merit $\mathcal{F}$ defined in equation (A.12). For a complete explanation see the text. For a blank entry one should read the last non-blank entry above it.

APPENDIX B

In this appendix, we describe the numerical methods that allow us to solve the polynomials. Typically, the polynomials have more than 1,000 coefficients, perhaps even as many as 10,000 terms. The dynamic range of a polynomial is defined to be the maximum magnitude difference of those terms. The dynamic range of the polynomials experienced in this research has been as large as 4,400 orders of magnitude – 10^{4400}.

There are seven important factors in the development of a robust computer program for solution of such polynomials: (1) The iteration algorithm; (2) the deflation algorithm; (3) the search algorithm; (4) the need for an internal representation of very large numbers; (5) the implementation of arithmetic operations for such an internal representation; (6) a scheme to estimate the error of the results; and (7) a determination of the amount of precision needed by the solution. Of these seven, the first three are of general interest while the remaining four are specific programming details of purely technical interest. We shall discuss the first three topics in detail below. The most important is the search procedure. This is because both the iteration and the deflation algorithms suffer from severe inadequacies.

The Iteration Algorithm

The iteration algorithm of first choice is Newton's method. This method is very easy to program and it converges quite fast. Its major problem is its lack of absolute convergence. Unless the starting point is close to the zero sought, there is no assurance that Newton's method will converge to that zero. Thus, it is most important to have a good search algorithm.

Given a good search algorithm, capable of locating a zero of the polynomial to an accuracy of at least 4 decimal figures, Newton's iteration will yield 16 decimal figures of accuracy in just one or two steps.

Consider the following polynomial with real or complex coefficients:

$$c_0 + c_1 x + c_2 x^2 + c_3 x^3 + ... + c_n x^n \quad . \tag{B.1}$$

Assuming that there is a single zero, real or complex, close to the origin, Newton's method will give the first estimate of the root as

$$x_0 = -c_0/c_1 \quad . \tag{B.2}$$

The method proceeds by substituting x_0 into the polynomial and its derivative to find the next values for c_0 and c_1. This step is repeated until one knows the root to the desired accuracy.

Three things should be noted in the above discussion: (1) One must move off the real axis into the complex plane to find complex zeroes. Even if the coefficients of the polynomial are real, so that zeroes appear in complex conjugate pairs, one cannot simply "sit" on the real axis and hope to evaluate both zeroes of a pair. (2) The assumption of a single zero near the origin is vital. If there should be more than one zero near the origin, the method will not converge. In our solution of polynomials, at the point of application of Newton's method, the next nearest zero is several orders of magnitude farther away than the nearest zero. (3) A multiple zero is not dealt with in the above discussion.

The Deflation Algorithm

Once a zero of the polynomial has been found, the degree of the polynomial can be reduced by one. This is done by factoring out, by synthetic division, the calculated zero from the original polynomial. The resultant polynomial of reduced degree is called the deflated polynomial and the reduction process is called deflation.

It is well known that deflation must be done with great care. In general, accumulation of errors in the estimate of the known roots will cause the deflated polynomial to separate away from the original one: the zeroes of the former will no longer coincide with the remaining unknown zeroes of the latter. Furthermore, in general the dynamic range of the deflated polynomial will be larger than that of the original polynomial. Conventional wisdom has it that the proper way to minimize these problems is to deflate the zeroes in order of increasing magnitude.

This is much easier said than done. For example, the polynomial

$$1 - x^{500},$$

has 500 zeroes of equal magnitude, uniformly spaced over the unit circle. Any attempt to solve this example by deflating the polynomial as the zeroes are found in an orderly sequence around the unit circle will fail after about the 30-th zero. This is in spite of the fact that those 30 or so zeroes are calculated to a precision better than 12 decimal digits. The deflated polynomial will have completely separated from the original by that time. The "size" criterion in the determination of the deflation order is useless when all the zeroes have the same, or nearly the same, magnitude. Therefore there must be some additional criterion or strategy with which the search is to proceed if the deflation algorithm is to remain stable. In principle, one could attempt to develop a computer program to solve polynomials without deflation. In practice, however, such programs are not feasible, as they would suffer from a host of extremely severe difficulties.

We shall now discuss such a search procedure which is suited for deflation. As far as we know, there is no general theory which addresses this issue. On the other hand, we have discovered a heuristic algorithm which provides a robust and stable search strategy, on top of the basic magnitude criterion. We refer to this algorithm as the "balanced" approach, drawing on the analogy between the zeroes of the polynomial around a circle and some weights around the rim of a wheel. If one finds the zeroes and deflates the polynomial in a sequential order around the circle, the analogous weights around the wheel are removed primarily from one side. Such a wheel will become unbalanced and no longer spin true. In the following, we describe the intuitive reasoning which went into the development of the heuristics.

The above mentioned analogy is quite proper, because in general the deflated polynomial has a larger dynamic range than the original one. That will be true unless the zeroes are removed in a balanced manner, in which case the resulting

dynamic range will be no larger than the initial one. Since on any real computer the precision and magnitude are limited, an increasing dynamic range is disastrous, because in practice it will eventually cause the deflation process to fail.

The requirement for a balanced solution is further reinforced by the fact that one cannot deflate the complex zeroes in conjugate pairs. We have repeatedly attempted to implement a scheme whereby the conjugate pair is deflated together, while more or less maintaining the balanced approach. All such attempts failed.

We have examined various means of implementing the balanced search order. The simplest approach is to rotate forward a certain angle from the last zero found. For example, a 90-degree rotation implies that if the first zero is on the positive real axis, the place to search for the next one will be on the positive imaginary axis. The third place would be on the negative real axis, etc. Such a scheme is certainly balanced but it fails. The result of a 90-degree rotation has the appearance of a 4-bladed propeller. This propeller may be balanced, but the weights' analogy suggests that the deflation should proceed in a manner which is not only balanced but also uniformly distributed.

In order to find an optimal rotation angle an experiment was performed using a 3,000-degree polynomial. The rotation angle was varied from zero to 90-degrees, in 1-degree increments. Nearly all the runs failed. Some failed after only about 30 zeroes, some after about 100 zeroes, some after several hundred zeroes, and some after more than 2,000 zeroes. Rotation angles smaller than 25-degrees produced very poor results. The rotation angles near 30, 45, 60, 75, and 90-degrees also failed fairly early. That was taken as an indication that one must avoid rotation angles which are at or near simple integer fractions of 360 degrees, such as 360/4, 360/5, 360/6, etc. The most likely reason is that in these cases the search order is balanced but not uniform enough. We have therefore examined the higher fractions in the immediate neighborhood of 1/9 and empirically determined the optimal rotation angle to be 40.7-degrees. As a fraction of 360-degrees, this lies between 6/53 and 7/62, both relatively far removed from simple fractions.

This has proven to be the correct search procedure for a deflation-based solution of polynomials. It has been successful for all the polynomials tested to date, including an example of degree 8,000.

The Search Algorithm

As the preceding discussions indicates, the search algorithm is of utmost importance for a robust polynomial solver. Yet, not one text on numerical analysis deals with this subject. The requirements for a good search procedure are: (1) absolute convergence; (2) insensitivity to nearby zeroes; (3) capability of locating multiple-order zeroes; (4) ability to self-start ; and (5) fast convergence. In addition, a good search algorithm should be also a good iteration procedure.

We would like to remind the reader that Newton's method is an "inverse" method. All such methods suffer from a lack of absolute convergence. This is because in between the roots there are many points of zero derivative, where an inverse method will yield a next guess that is infinitely far away.

In order to avoid that problem, we have developed a "reciprocal" method based on a simple ratio test for convergence. This method fulfills all five requirements listed above. It is used in the search algorithm and not in the iteration algorithm, because Newton's method is so simple. Once a starting point close enough to the root is obtained by the reciprocal method, Newton's method is fast and reliable.

The reciprocal of a given polynomial, Eq.(B.1), can be written as the sum of partial fractions.* The partial fractions are:-

$$r(x) = \frac{k_1}{x - a_1} + \frac{k_2}{x - a_2} + \frac{k_3}{x - a_3} \tag{B.3}$$

Here, a_1, a_2, a_3 are the roots of the original polynomial (real and/or complex), and k_1, k_2, k_3 are the residues. We further assume that these partial fractions are ordered so that $a_1 \leq a_2 \leq a_3$, etc.

⋆ The partial fractions are used here for derivation purposes only, and are never actually evaluated.

Concentrating our attention for the moment on the first fraction and, without loss of generality, considering the case of $k_1 = 1$, we have the Taylor series

$$r(h) = -\frac{1}{a_1} - \frac{h}{a_1^2} - \frac{h^2}{a_1^3} - \ldots - \frac{(h/a_1)^n}{a_1} - \ldots \quad , \tag{B.4}$$

where h is the step size used to expand the series. We apply the ratio test for convergence, and eliminate the common negative signs and the a_1 in the denominator. The ratio R is given by

$$R = \frac{(h/a_1)^n}{(h/a_1)^{n-1}} = \frac{h}{a_1}. \tag{B.5}$$

This simple result has two important implications: (1) The Taylor series for an isolated simple pole (the reciprocal of a single zero) is a perfect exponential. On a semi-logarithm plot, this Taylor series will appear as a straight line. (2) When the step size for expansion is equal to the location of the zero ($h = a_1$), this straight line will be perfectly horizontal. Therefore, the reciprocal algorithm is an extremely simple one. If we can assume that there is an isolated single zero, two adjacent terms of the Taylor series of the reciprocal of the polynomial can yield the location of that zero.

$$a_1 = \frac{h}{R} \tag{B.6}$$

Since the step size h is under our control and therefore known, the location of the zero can be easily found.

Of course, this is an oversimplified picture. Widening the scope a bit, we include the next nearest zero as follows. The general Taylor term for a sum of two partial fractions is given by

$$\begin{aligned} n - \text{th term} &= -\frac{k_1(h/a_1)^n}{a_1} - \frac{k_2(h/a_2)^n}{a_2} \\ &= -\frac{k_1(h/a_1)^n}{a_1}\left[1 + \frac{k_2}{k_1}(a_1/a_2)^{n+1}\right] \quad . \end{aligned} \tag{B.7}$$

Since $a_1 < a_2$, the second term within the square brackets will be negligibly small

when n is large. This is independent of the relative magnitudes of the two k's. One merely needs to advance to a higher order term of the Taylor series to overcome the necessity of including the Taylor term for the second zero.

Thus the reciprocal method consists of (1) evaluating the Taylor series for the reciprocal of the polynomial and (2) examining the ratio R for a sufficiently high order, where it becomes a constant. All the effects of the secondary zeroes will be washed out at such a higher order. Equation (B.6) is the basis of the reciprocal algorithm.

Of course, real life is not that simple. However, in the reciprocal method there are no inherently debilitating conditions similar to the non-convergence of the inverse methods. We do have to be concerned with finite computers, finite length Taylor series, closeness of the secondary zeroes, multiple zeroes, and the like. But, they can all be accounted for in the fine tuning of the final computer code for the algorithm. Details of such fine tuning do not deserve attention here.

REFERENCES

1. C.N. Yang and T.D. Lee, *Phys. Rev.* **87**(1952), 404; *ibid* **87**(1952),410.
2. M. Falcioni, E. Marinari, M.L. Paciello, G. Parisi and B. Taglienti, *Phys. Lett.* **108B**(1982), 331.
3. R. B. Pearson, *Phys. Rev.* **D26**(1982), 6285.
4. C. Itzykson, R.B. Pearson and J.B. Zuber, *Nucl. Phys.* **B220**(1983), 415.
5. E. Marinari, *Nucl. Phys.* **B235**(1984), 123.
6. G. Bhanot, S. Black, P. Carter and R. Salvador, *Phys. Lett.* **183B**(1987), 331;
   G. Bhanot, K. Bitar and R. Salvador, *Phys. Lett.* **187B**(1987), 381.
7. G. Bhanot, K. Bitar and R. Salvador, *Phys. Lett.* **188B**(1987), 246.

8. M. Fukugita, S. Ohta, Y. Oyanagi and A. Ukawa, *Phys. Rev. Lett.* **58**(1987), 2515;
R. V. Gavai, J. Potvin and S. Sanielevici, *Phys. Rev. Lett.* **58**(1987), 2519, and preprint TIFR/TH/87-27 (July 1987);
S. Gottlieb, W. Liu, D. Toussaint, R. L. Renken and R. L. Sugar, preprint UCSD-PTH-87/10 (June 1987);
R. Gupta, lectures given at Symposium on Lattice Gauge Theory using Parallel Processors, Beijing, China, May 1987, and Los Alamos Preprint.

9. S.A. Gottlieb, J. Kuti, D. Toussaint, A.D. Kennedy, S. Meyer, B.J. Pendleton and R.L. Sugar, *Phys. Rev. Lett.* **55**(1985), 1958.

10. M. E. Fisher in *Critical Phenomena*, Proceedings of the 51^{st} International School of Physics, Enrico Fermi. Edited by M.S. Green. N.Y., Academic Press, 1971.

11. B. Nienhuis and M. Nauenberg, *Phys. Rev. Lett.* **35**(1975), 477;
Y. Imry, *Phys. Rev.* **B21**(1980), 2042;
M. E. Fisher and A. N. Berker, *Phys. Rev.* **B26**(1982), 2502.

12. E. Brezin and J. Zinn-Justin, *Nucl. Phys.* **257B**(1985), 867.

13. K. Kajantie, C. Montonen, E. Pietarinen, *Zeit. Phys.* **C9**(1981), 253;
L.G. Yaffe, B. Svetitsky, *Phys. Rev.* **D26**(1982), 963;
S. Gottlieb, A.D. Kennedy, J. Kuti, D. Toussaint, S. Meyer, B.J. Pendleton and R.L. Sugar, *Phys. Lett.* **189B**(1987), 181.

14. M.L. Glasser, V. Privman and L.S. Schulman, *Phys. Rev.* **B35**(1987), 1841.

15. R. Balian, J.M. Drouffe, C. Itzykson, *Phys. Rev.* **D19**(1979), 2514.

16. A. Di Giacomo and G. Paffuti, *Phys. Lett.* **108B**(1981), 327

17. M. Creutz and K. J. M. Moriarty, *Phys. Rev.* **D26**(1982), 2166.

18. "Wilson Loop Expectation Values in $SU(3)$", compilation by A. Hasenfratz and P. Hasenfratz, November 1984.

19. A. M. Polyakov, *Phys. Lett.* **72B**(1978), 477;
L. Susskind, *Phys. Rev.* **D20**(1979), 2610.

20. L.D. McLerran and B. Svetitsky, *Phys. Lett.* **98B**(1981), 195;
J. Kuti, J. Polonyi and K. Szlachanyi, *Phys. Lett.* **98B**(1981), 199.

21. J. Kogut, J. Polonyi, H.W. Wyld, J. Shigemitsu and D.K. Sinclair, *Nucl. Phys.* **B251**(1985), 311;
U. Heller and N. Seiberg, *Phys. Rev.* **D27**(1983), 2980.

22. D. J. Scalapino and R. L. Sugar, *Phys. Rev. Lett.* **46**(1981), 519.