EUROPHYSICS LETTERS 15 July 1992

Europhys. Lett., **19** (6), pp. 451-458 (1992)

Simulated Tempering: a New Monte Carlo Scheme.

E. Marinari (*) (**) (§) and G. Parisi (*) (§§)

(*) *Dipartimento di Fisica, Università di Roma «Tor Vergata»*
Via della Ricerca Scientifica, 00173 Roma, Italy
INFN, Sezione di Roma «Tor Vergata» - Roma
(**) *Physics Department, Syracuse University - Syracuse, NY 13244, USA*

(received 17 March 1992; accepted in final form 9 June 1992)

PACS. 05.20 – Statistical mechanics.
PACS. 75.50 – Dielectric breakdown and space-charge effects.

Abstract. – We propose a new global optimization method (*Simulated Tempering*) for simulating effectively a system with a rough free-energy landscape (*i.e.*, many coexisting states) at finite nonzero temperature. This method is related to simulated annealing, but here the temperature becomes a dynamic variable, and the system is always kept at equilibrium. We analyse the method on the Random Field Ising Model, and we find a dramatic improvement over conventional Metropolis and cluster methods. We analyse and discuss the conditions under which the method has optimal performances.

Simulated annealing is an efficient heuristic method which is used to find the absolute minimum of functions with many local minima: it has been introduced independently in the framework of the Monte Carlo approach for discrete variables in ref. [1], and in the framework of stochastic differential equations (of Langevin type) for continuous variables in ref. [2].

The essence of the method consists of the following. Let us suppose that we are interested in finding the minimum of a function $H(X)$, where X denotes an element of the configuration space (which has dimension N, where N is often a very high number). In most cases we do not know any method which can guarantee to find the minimum of $H(X)$ with a computational effort that does not increase more than polynomially in N. In these cases one can try as a first guess a random search starting from a random configuration and minimizing $H(X)$ with a steepest-descent algorithm. If the number of local minima increases as $\exp[\gamma N]$, with γ different from zero, it often happens that this method also takes an exponentially large number of trials (*i.e.*, $\exp[\delta N]$, with in general $\delta < \gamma$).

In the simulated annealing method one considers a β-dependent algorithm which

(§) Marinari@Roma2.INFN.It.
(§§) Parisi@Roma2.INFN.It.

asymptotically generates the configurations X with Gibbs's probability distribution, *i.e.* $\exp[-\beta H(X)]$; for definiteness we can consider the case of Monte Carlo steps. Simulations at increasing values of β are done (eventually at $\beta = \infty$). Each time β is changed the system is driven out of equilibrium, but that does not matter since eventually we are interested in the $\beta = \infty$ result.

In general the simulated annealing method does not have any reason to converge to the exact result, *i.e.* to provide the minimum of $H(X)$. Only if we do an asymptotically large number of simulated annealing runs, or if the values of β are changed by an infinitesimal amount at each step and an infinite amount of Monte Carlo steps are done at each value of β, the simulated annealing method will converge to the exact result and will find the minimum of H. But the convergence is guaranteed only if we asymptotically invest an infinite amount of computer time. If a reasonable annealing scaling is used (β is changed by a nonzero amount and only a finite number of Monte Carlo cycles are done at a given value of β) we have no reason to believe that this procedure ends up in the global minimum; indeed in the extreme case in which β takes only two values (0 and ∞), we find the same result as the random search algorithm we have described before.

The simulated annealing algorithm can however be used as a heuristic predictor for the global minimum: one can compare the values of the energy after many simulated annealing runs and if the probability of ending with the global minimum is not too small, the simulated annealing turns out to be a rather efficient algorithm. Let us note that this efficiency depends a lot on the shape of the phase space: if the absolute minimum has a small basin of attraction, and is separated from the large local minima by very high barriers, simulated annealing does not have any reason to be a good algorithm.

Unfortunately if we want to extend the algorithm to finite temperature we are very soon in deep trouble. Indeed if we stop our simulations at a given value of $\beta < \infty$, the one we want to use to evaluate observables, different runs will give different results (if β is sufficiently large). In this case we cannot just select the runs which produce the configurations with lower energy: at $T \neq 0$ we have to minimize the free energy F and not the energy. Estimating the entropic contribution is a nontrivial task, and makes a straightforward generalization of the simulated annealing impossible. This problem is very severe in cases like spin glasses [3] or hetero-polymers folding [4] (maybe also peptides [5]) in which there are more than one equilibrium state and we are actually interested in knowing the relative weight which the different equilibrium states carry in the partition function.

The method we propose in this note is meant to bypass these difficulties, and to constitute a viable scheme to minimize free energy in an effective way. It can be regarded as a very efficient global optimization scheme. The basic idea of the *Simulated Tempering* method consists of changing the temperature while remaining at equilibrium: this is in contrast with the simulated annealing method, where every change of the temperature drives the system out of equilibrium. This can be achieved by enlarging the configuration space of the system in the following way.

We define a large configuration space, which is characterized by the variables X (the original configuration space) and by a new variable m which can take M values ($m = 1 \dots M$). The probability distribution $P(X, m)$ will be chosen to be $P(X, m) \propto \exp[-H(X, m)]$, where we have absorbed the factor β in the definition of the Hamiltonian. We set $H(X, m) \equiv \beta_m H(X) - g_m$.

Here the β_m and the g_m can take arbitrary values we assign *a priori*. The g_m will be *a priori* assigned constants, and the β_m will be dynamical variables which will be allowed to span a set of values given *a priori*. For simplicity we can assume that the β_m are ordered. It is evident that the probability distribution induced by this Hamiltonian, restricted to the subspace at fixed m, is the usual Gibbs distribution for $\beta = \beta_m$. On the other hand, the probability of having a given value of m is simply given by

$$P_m \propto Z_m \exp[g_m] \equiv \exp[-(\beta_m f_m + g_m)], \tag{1}$$

where the Z_m are the partition functions at given β_m and the f_m are the corresponding free energies. If we make the choice $g_m = \beta_m f_m$, then all the P_m become equal.

If our target is to do a simulation at a given value of β, we can take $\beta_{\tilde{m}} = \beta$ and with this choice for the g_m we can perform a Monte Carlo simulation in which we also allow the change of m by 1 unit. In this case the system will be with a probability $1/\tilde{m}$ at $m = \tilde{m}$. Only a fraction $1/\tilde{m}$ of the events will be interesting for measuring directly expectation values at β (if the use of a histogram reconstruction makes also the other β values very useful). The frequent visits of the system to lower values of β_m will make it decorrelate much faster. Indeed at lower β values free-energy barriers are lower, and the system will find it much easier to jump. Then, when it decides to cool off again, it will be visiting, with the correct equilibrium probability, a different minimum. This method may be useful only if the transition from one value of β_m to another happens with non-negligible probability. It is evident that if the two contiguous values of β are too different, the probability of accepting a change will be rather small, and that, on the contrary, if they are too similar, they will not help in decorrelating.

Let us try to compute the probability of going from β_m to $\beta_{m+1} \equiv \beta_m + \delta$. If we try to modify β, the variation of the Hamiltonian is given by $\Delta H = E\delta - (g_{m+1} - g_m)$, where E is the instantaneous value of the energy $H(X)$. On the other hand, we have that $g_{m+1} - g_m$ is given by the value of the energy for some β in between β_m and β_{m+1}. More precisely

$$g_{m+1} - g_m = E_m \delta + \frac{1}{2} C_m \delta^2 + O(\delta^3), \tag{2}$$

where E_m is $E(\beta_m)$ ($E(\beta)$ is the expectation value of $H(X)$ as a function of β) and $C_m = \mathrm{d}E/\beta_m$. If we assume that E is very close to E_m, the variation ΔH will be not too large under the condition that $C_m \delta^2 = O(1)$. One should also consider that there are thermal fluctuations in the value of the energy which are of order of C_m. The condition on δ is equivalent to requiring that there is a non-negligible overlap in the values of the energy computed at contiguous values of β_m.

In the usual thermodynamic limit the energy is a quantity of order N and the condition on δ requires that δ is of order $N^{-1/2}$, which is not a very demanding condition. The main difficulty in the method is the required tuning in the choice of g_m. Indeed, if one takes for β_m an unreasonable value, the simulation could get trapped at a given value of β_m. In this respect it is interesting to note that we are not introducing any systematic bias. One can also think about the possibility of performing an iterative procedure in which the values of the g_m's are adjusted during the simulation, but we will see that already with the naive choice we are using one gets very impressive results.

We have applied the *Simulated Tempering* method to the *Random Field Ising Model* (RFIM), which has many features that are very relevant to our case. It has a rough landscape, and the symmetry of the + and the − state of the pure Ising model is broken by the random magnetic field. This is not a trivial symmetry any longer, and the flips from the + to the − sector (and back) are an essential part of the dynamics. The state oriented in the + direction and the one oriented in the − direction, which macroscopically are very similar, are completely different from a microscopic point of view. The transition from the favoured state

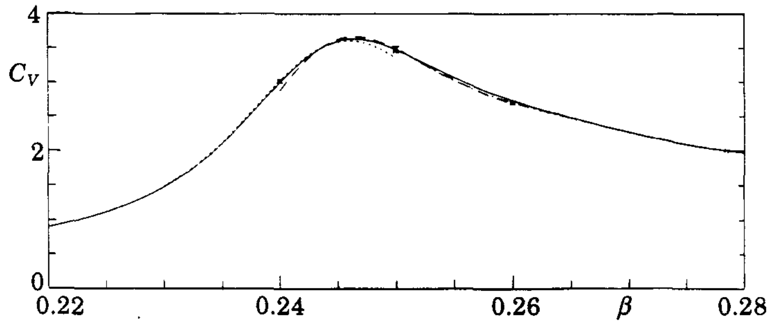

Fig. 1. – Specific heat C_V. Points from cluster algorithm data, lines from histogram reconstruction.

(which is selected by the specific realization of the magnetic field) to the suppressed one is a rare event.

For the RFIM an extension of the cluster update method [6,7] does not give any improvement over the local classical Metropolis method [8]. The system undergoes the usual pathology of freezing already at $T > T_c$, and the spins form a large cluster. In no way does the cluster method help in this case, for example, to tunnel from a + to a − state.

We have implemented the *Simulated Tempering* by proposing one β update at the end of each sweep of the lattice spins. The computational time required to compute the β update is negligible.

Let us anticipate our results: as we will show in some detail the *Simulated Tempering* method helps a lot. In our test, correlation times for observable quantities which are not sensitive to the magnetization decrease by a factor of 6 as compared to the Metropolis and the cluster method. As fas as the estimate of the magnetization is concerned, the method changes the picture dramatically, allowing tunneling where the Metropolis method is trapped in a single state, and correcting, in some cases, wrong estimates given by the Metropolis method.

The lattice Hamiltonian is the usual Ising model Hamiltonian, where the site random fields h_i take values $h_i = |h|\theta_i$ ($\theta_i = \pm 1$ with probability 1/2).

We have taken in our simulations $V = 10^3$ and $|h| = 1$. We have worked with a given realization of the random magnetic field. In order to characterize the system in fig. 1 we show its specific heat. The 3 points with errors are from 3 runs done by using the cluster algorithm, while the dotted, dashed and dot-dashed lines are done by using the reconstruction method (see ref. [9] for Ising model and $SU(2)$ gauge theory applications, ref. [10] for an earlier, independent introduction of the method, and ref. [11] for successive applications and detailed reviews). The continuous line uses the method by patching the 3 data points: The reconstruction is very reliable.

We have analysed the measured observables by means of a binning procedure, obtaining an asymptotic estimate for the errors. We have also focused our analysis on the study of τ^{int}, which is the relevant quantity related to the true error over measured observables. Following ref. [12] we use an improved estimator for τ^{int}, taking up to 20 time steps for the estimation window. The errors on τ^{int} are, when we quote an asymptotic estimate for them, always of the order of 1 on the last digit. We have also monitored that τ_{exp} gives consistent results.

In table I we give two of the measured observables: the thermal part of the energy, E_{T}, and the magnetization m. E_{T} has a behavior typical of the quantities that are Z_2 symmetric. The rows called (MC) and (CL) give information about the runs we have done with the

TABLE I. – *Thermal energy, magnetization and related integrated autocorrelation times. Errors are in round brackets* ( ). *When in square brackets,* [ ], *error and* τ^{int} *estimates are not asymptotic. The value for m given by the Metropolis method* (MC) *at* $\beta = 0.26$ *is wrong.*

β	E_{T}	$\tau_{E_{\text{T}}}^{\text{int}}$	m	τ_m^{int}	$N_{\text{iter}} \cdot 10^{-3}$
0.24 (MC)	1.1980 (18)	10	– 0.161 [12]	[70]	200
0.24 (CL)	1.2059 (22)	14	– 0.180 [10]	[90]	200
0.24 (B)	1.2045 (19)	6	– 0.187 (10)	60	145
0.24 (E)	1.2025 (13)	3.7	– 0.159 (10)	40	160
0.24 (F)	1.2015 (11)	5.5	– 0.175 (5)	32	290
0.25 (MC)	1.5286 (15)	7	– 0.37 [6]	[700]	200
0.25 (CL)	1.5252 (25)	11	– 0.32 [4]	[660]	200
0.25 (B)	1.5311 (10)	3.9	– 0.363 (15)	150	297
0.25 (C)	1.5303 (12)	4.8	– 0.351 (11)	70	226
0.25 (D)	1.5299 (9)	3.5	– 0.350 (20)	[370]	300
0.25 (E)	1.5279 (8)	2.4	– 0.320 (12)	105	301
0.25 (F)	1.5281 (8)	3.3	– 0.352 (9)	52	290
0.255 (MC)	1.6723 (12)	9	– 0.35 (13)	[6000]	200
0.255 (D)	1.6723 (8)	1.5	– 0.414 (22)	180	151
0.255 (E)	1.6718 (6)	1.8	– 0.382 (13)	108	301
0.26 (MC)	1.7954 (8)	2.8	– 0.7016 (3)	3.8	200
0.26 (CL)	1.7942 (11)	7.6	– 0.53 [5]	[1000]	200
0.26 (B)	1.7925 (7)	1.6	– 0.476 [18]	[81]	158
0.26 (E)	1.7924 (6)	1.15	– 0.433 (13)	52	150
0.26 (F)	1.7928 (5)	1.75	– 0.473 (10)	64	307

Metropolis method and with the cluster algorithm. These runs have been used to get a preliminary estimate of the system energy and to determine the values of the g_m. It is in no way necessary to get, for estimating the g_m, more than a rough estimate of the E_m, and in a practical application of the method the preliminary MC runs can be very short. It is possible to determine directly the values of the $\exp[-f_n]$, by using the energy histograms taken in the preliminary runs. Although we stress that this possibility exists, we do not think that it could dramatically increase the efficiency of the method. When, in table I, we put errors and τ_{int} in square brackets we mean that we did not get an asymptotic estimate. Let us also note now

TABLE II. – β *values allowed in each of our «Simulated Tempering» runs, and number of iterations* (*in units of* 10^3) *the system spent at each* β *value. For historical reasons we label the runs with the capital letters B, C, D, E, F.*

Run	β_1, n_1	β_2, n_2	β_3, n_3	β_4, n_4	β_5, n_5
B	0.24, 145	0.25, 297	0.26, 158		
C	0.23, 206	0.25, 226	0.27, 167		
D	0.245, 148	0.25, 301	0.255, 151		
E	0.24, 149	0.245, 300	0.25, 301	0.255, 301	0.26, 150
F	0.23, 159	0.24, 290	0.25, 290	0.26, 306	0.27, 155

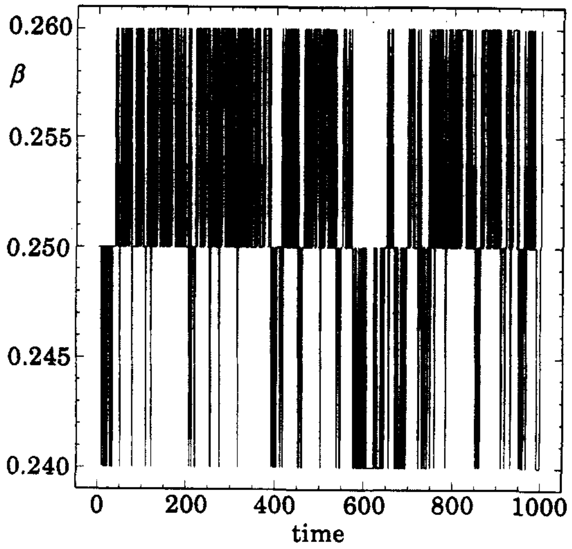

Fig. 2. – β as a function of the computer time for the runs of series B (3β values allowed, 50% acceptance ratio).

that the MC run at $\beta = 0.26$ gets a *wrong expectation value* for m. In this case the standard Metropolis does not produce any tunneling event, and always stays in the – phase.

In table II we give details about our *Simulated Tempering* runs. We have tried different combinations, allowing the system to take 3 or 5 β values, always centred around $\beta = 0.25$. In table II we check the performance of our method at the different β values we have allowed in the different simulations. The choice of the β values has been dictated, as we have discussed before, by the requirement of having a non-negligible overlap in the energy histograms of the preliminary MC runs. Runs D and E have a very small δ value, and a high acceptance factor for a β update, of $\simeq 70\%$. Runs B and F have a medium δ, and a β acceptance factor of $(40 \div 50)\%$. Run C has a higher δ value and a very low acceptance factor for the β update, $(10 \div 15)\%$.

In fig. 2 we give β_m as a function of the computer time for system B. Let us start by commenting on the results for m, which are quite spectacular. At $\beta = 0.24$ (not so low T) τ_m is higher than $O(100)$ for the Metropolis and Cluster methods, and gets down to 32 in the F run. In general runs with a larger δ value seem to be more effective for improving the estimate of m. Things are better and better at lower temperatures. At $\beta = 0.25$ from $\tau_m > 700$ we go down to $\tau_m = 52$ in run F, with a gain of a factor larger than 12. At $\beta = 0.255$ from $\tau_m > 6000$ we go down to 108 in run E, with a gain of a factor better than 60. At $\beta = 0.26$ after 200 000 steps the Metropolis method does not succeed in getting a single tunneling event, while our run E has $\tau_m = 52$. In fig. 3*a*)-*c*) we show what happens. In fig. 3*a*) we give the magnetization as a function of computer time for the Metropolis method, for 200 000 steps. The system stays in the – state, with very large fluctuations which never succeed in getting a complete flip. In fig. 3*b*) we plot m for our F system, only 10 000 steps. Here the data points are at different β values, and it is clear that going to different β values allows an easy flipping. In order to make the situation clear in fig. 3*c*) we have selected only the first 10 000 configurations, of the F dynamics, which happen to be at $\beta = 0.26$. The picture speaks for itself.

Also for E_{T} there is a large gain at all β values. One gains a factor 3 at $\beta = 0.24$, 0.25, a factor 6 at $\beta = 0.255$, and a factor 2.5 at $\beta = 0.26$. In this case the best performances are obtained for small δ values.

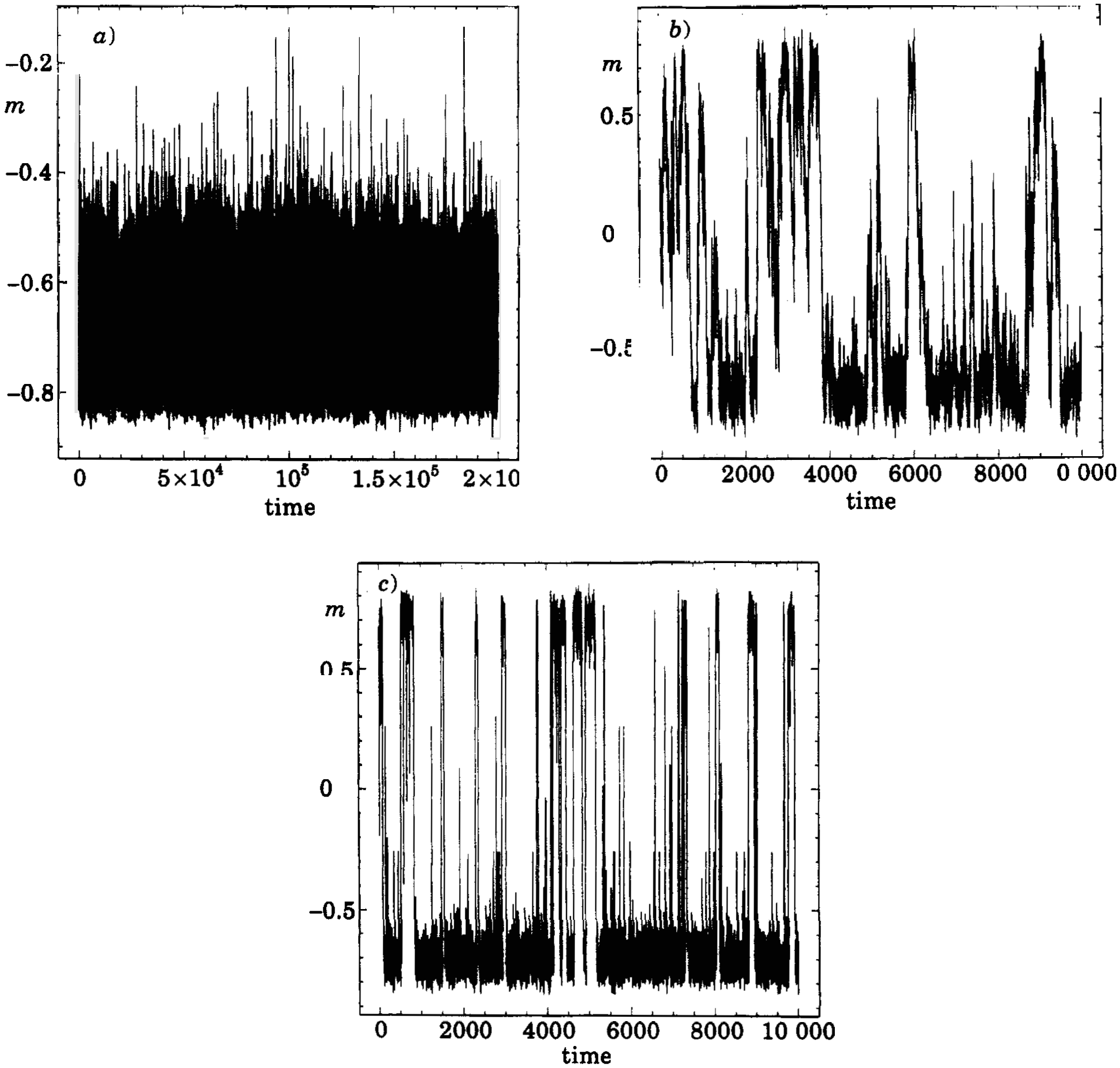

Fig. 3. – *a*)-*c*) Magnetization m as a function of computer time. In *a*) for the Metropolis method at $\beta = 0.26$, in *b*) m for the F systems (β is here a dynamical variable which is allowed to take 5 values during the course of the dynamics), in *c*) the configurations of run F which have $\beta = 0.26$.

* * *

It is a pleasure for us to thank M. FUKUGITA for intersting discussions, and P. CODDINGTON for a critical reading of the manuscript. Hardware and software computer support has been given by INFN (Roma «Tor Vergata») and NPAC (Syracuse).

Additional remark.

After submitting this note we learnt about ref. [13], which propose a different but related method. For our method we do not need any patching, and we just get the correct probability distribution at each β value.

REFERENCES

[1] KIRCKPATRICK S., GELATT C. D. and VECCHI M. P., *Science*, **220** (1983) 671.

[2] ALUFFI-PENTINI F., PARISI V. and ZIRILLI F., *J. Optim. Theory Appl.*, **17** (1985) 1.

[3] MEZARD M., PARISI G. and VIRASORO M. A., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987.

[4] IORI G., MARINARI E. and PARISI G., *J. Phys. A*, **24** (1991) 5349.

[5] FUKUGITA M., KAWAI H., NAKAZAWA T. and OKAMOTO Y., *Nucl. Phys. B* (*Proc. Suppl.*, **20** (1991) 766.

[6] SWENDSEN R. H. and WANG J. S., *Phys. Rev. Lett.*, **58** (1987) 86.

[7] EDWARDS R. G. and SOKAL A., *Phys. Rev. D*, 38 (1988) 2009.

[8] GUAGNELLI M., MARINARI E. and PARISI G., unpublished.

[9] FALCIONI M., MARINARI E., PACIELLO M. L., PARISI G. and TAGLIENTI B., *Phys. Lett. B*, **108** (1982) 331; MARINARI E., *Nucl. Phys. B*, **235** (1984) 123.

[10] SALSBURG Z. W., JACKSON J. D., FICKETT W. and WOOD W. W., *J. Chem. Phys.*, **30** (1959) 65; MCDONALD I. R. and SINGER K., *Discuss. Faraday Soc.*, **43** (1967) 40; VALLEAU J. P. and CARD D. N., *J. Chem. Phys.*, **57** (1972) 5457.

[11] For good reviews, extensions and detailed applications of the method of ref. [9] see for example BHANOT G., BLACK S., CARTER P. and SALVADOR R., *Phys. Lett. B*, **183** (1987) 331; FERRENBERG A. M. and SWENDSEN R. H., *Phys. Rev. Lett.*, **61** (1988) 2635; *Erratum*, **63** (1989) 1658; FERRENBERG A. M. and SWENDSEN R. H., *Phys. Rev. Lett.*, **63** (1989) 1196; ALVES N. A., BERG B. A. and SANIELEVICI S., *Spectral Density Study of the SU*(3) *Deconfining Phase Transition*, to be published in *Nucl. Phys. B*.

[12] WOLFF U., *Phys. Lett. B*, **228** (1989) 379.

[13] BERG B. A. and NEUHAUS T., *Phys. Lett. B*, **267** (1991) 249; *Phys. Rev. Lett.*, **68** (1992) 9; BERG B. A. and CELIK T., *A New Approach to Spin Glass Simulations*, preprint FSU-SCRI-92-58 (April 1992), and references therein.