IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Statistical properties of biased sampling methods for long polymer chains

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1988 J. Phys. A: Math. Gen. 21 127

(http://iopscience.iop.org/0305-4470/21/1/020)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 129.252.86.83
The article was downloaded on 31/05/2010 at 11:18

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. **21** (1988) 127-146. Printed in the UK

Statistical properties of biased sampling methods for long polymer chains

Jannis Batoulis and Kurt Kremer

Institut für Physik, Universität Mainz, D-6500 Mainz, Federal Republic of Germany

Received 18 May 1987, in final form 20 August 1987

Abstract. We present a detailed statistical analysis of the Rosenbluth method of generating self-avoiding walks. This method became one of the standard methods for simulating long polymers. We show that this method, although very successful in yielding large samples, becomes exponentially poor with increasing chain length. This has to be taken into account for simulations and was not done yet. We describe a way to quantify the number of chains needed. However, when compared to direct simple sampling, the method still, carefully used, yields better results, especially in the vicinity of the theta point of polymers. Special care has to be taken for $d = 2$. Some extensions to improve the situation are also discussed.

1. Introduction

Monte Carlo methods have been proved to be very useful for the investigation of the statistics of polymers during the last two decades [1-3]. In particular, for the problem of the excluded volume where exact calculations are impossible ($d = 3$) [4], these methods are very valuable. There is, however, one serious problem, namely the so-called attrition. If one, for example, samples a random walk on a lattice, there are q_0^N configurations in phase space. Here q_0 is the coordination number of the lattice and N is the number of bonds of the walk; the first site is considered to be at a fixed lattice position. For self-avoiding walks it is now known that there are only $q_{\text{eff}}^N N^{\gamma-1}$ configurations with $q_{\text{eff}} < q_0$ and γ a critical exponent ($\gamma \simeq \frac{7}{6}$, $d = 3$) [4]. On the cubic lattice $q_0 = 6$ and $q_{\text{eff}} = 4.68$ [1]. Therefore, only an amount $(q_{\text{eff}}/q_0 - 1)^N N^{\gamma-1}$ of all generated random walks is self-avoiding. Thus, if one samples the chains by generating random walks and then searching for self-avoiding chains (SAW) the gain becomes exponentially small. This causes serious problems for the simulation of systems which are supposed to have somewhat realistic chain lengths, in order to compare them with polymers. It is well known that there have been many attempts to overcome these problems by the use of more intelligent sampling methods. We cannot discuss all the different ways here in detail but want to introduce briefly the various classes of methods.

First, there is the simple sampling method as described above. The first modification was introduced some 30 years ago by Rosenbluth and Rosenbluth [5]. There the chain has some knowledge of the monomer's local environment. The next step then only takes the choice out of jumps to an empty site. Within this method the same phase space as for standard SAW is sampled, but one gets a modified distribution. One has to make corrections for this. It will be one of the main results of this paper that we give an estimate of the quality of such methods. Several modifications of this, also called 'biased sampling', with a soft bias [6] or looking ahead more than one step [7]

0305-4470/88/010127+20$02.50 © 1988 IOP Publishing Ltd

were proposed and used. A significantly different approach was made by the binary assembly method of Alexandrowicz [8]†. He built chains by a *random* binary assembly of two existing chains.

The second class of methods are the so-called dynamic simulation methods ([9–12] and references therein). Here beads are selected at random and then a motion due to a random choice is attempted. These methods allow only for one chain length and, for example, one temperature in one run.

The above short description shows that the methods of generating chains directly by the use of a random algorithm are still a very important and valuable tool for many applications. It allows us to analyse all chain lengths up to the longest generated N for different quantities. By introducing a temperature in the analysis (see below), these became the classical methods to investigate the collapse transition [13–15]. Here we now want to investigate in detail the properties of the biased sampling method of Rosenbluth and Rosenbluth [5] and its various extensions. This is of special importance because these methods are used for investigations of the collapse transition [14] as well as for standard SAW, which correspond to chains in a very good solvent [16–19]. However, no detailed statistical analysis of this approach has been published up to now, in spite of the fact that recent simulations use much longer chains than the initial tests [5].

The organisation of the paper is as follows. Section 2 contains a detailed description of the biased sampling method as well as a first analysis of the data for linear polymers on the FCC lattice. In § 3 we give a detailed analysis of the statistical properties and construct a criterion for the accuracy of a sample of generated chains. Section 4 then discusses some generalisations of the method, while § 5 contains our conclusions.

2. The Rosenbluth–Rosenbluth method (RR) for long chains

In the following we consider the generation of single SAW confined to a lattice. For a description of the procedure in the continuum, see [20]. The chains start at a given site. Let the coordination number of the lattice be q_0 and let us generate chains up to a length N. For standard simple sampling at each step the new bond is chosen out of q_0-1 directions at random. We take only q_0-1 lattice bonds into account in order to avoid chain termination by direct backfolding. If the selected new bond hits a lattice site that is already occupied, one has to stop the chain and start a completely new one. Of course, one can use the part of the walk generated up to that point for the statistics of shorter chains. However, the success rate of chains decays exponentially. The number of all SAW of length N is given by [4] $Z(N) = c_o q_{\text{eff}}^N N^{\gamma-1}$ with $\gamma = \frac{7}{6}$ $(d=3)$ and $\gamma = \frac{43}{32}$ $(d=2)$ [21] and effective coordination number $q_{\text{eff}} < q_0 - 1$. On the FCC lattice, which we are going to use subsequently for the actual simulations, we have $q_0 - 1 = 11$, while $q_{\text{eff}} = 10.035$. Thus, for $N = 100$, only 0.022% of the attempts are successful. A way out of this was suggested by Rosenbluth and Rosenbluth [5] as follows. Before we try to add a new bond, one looks for the empty neighbour sites. If there is no empty site available, the chain is stopped. If there are k sites available, one out of these is chosen at random with probability $p = 1/k$ (see figure 1). Here chains can only be terminated if the walk runs into a cage. For $d = 2$ we find, for example, on the square lattice that about 70% of the walks still exceed 200 bonds

† This is probably the best static method. It allows a very precise determination of the critical exponent γ.

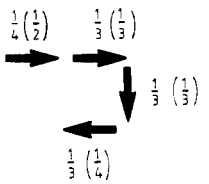

Figure 1. Biased sampling. The probabilities of the steps are indicated. The situation displayed shows that the inversely growing walk has a different probability, as indicated by the numbers in brackets. Note that, for walks on the honeycomb lattice in 2D only, this irreversibility vanishes because $q_0 - 1 = 2$.

[22]†. For $d = 3$ this termination turns out to be extremely rare. The larger the coordination number q_0, the higher is the probability of survival. For walks on the diamond lattice ($q_0 = 4$) at $N = 700$ still more than 90% of the attempts are successful [22]. For the FCC lattice the acceptance rate, due to our calculations, is 99.3% for the same length $N = 700$.

However, although the paths one samples are exactly the SAW trajectories, the distribution functions are very different. For SAW each configuration has exactly the same probability P_N, namely

$$P_N = q_0^{-1}(q_0 - 1)^{-(N-1)}. \tag{2.1}$$

This is different for the biased walk. Here the probability $P_N(\{r\})$ of a given N-step walk with configuration $\{r\}$ is

$$P_N^{RR}(\{r\}) = \prod_{i=1}^{N} (k_i)^{-1}. \tag{2.2}$$

As one can see from figure 1, 'dense' configurations have a higher probability. The RR sampling introduces a bias towards such configurations. An additional complication might occur because the sampling procedure is not reversible. The two directions in which a special walk could be generated may lead to different probabilities (but only if $q_0 - 1 > 2$). To correct for the introduced bias, each chain does not count as 1 during the sampling, but has a weight

$$W_N(\{r\}) = \prod_{i=1}^{N} k_i/(q_0 - 1) \tag{2.3}$$

because each given configuration $\{r\}$ is sampled $W_N(\{r\})^{-1}$ too often. One can interpret this weighting in the following way: the probability P_{RW}^i of a simple sampling walk to choose one of the available sites at step i is

$$P_{RW}^i = 1/(q_0 - 1). \tag{2.3a}$$

The corresponding probability of a RR walk P_{RR}^i is

$$P_{RR}^i = 1/k_i. \tag{2.3b}$$

Therefore, the weighting factor $w^i = k_i/(q_0 - 1)$ has to be introduced satisfying the relation

$$P_{RR}^i w^i = P_{RW}^i. \tag{2.3c}$$

† For walks in the SAW universality class one has $\gamma = \frac{7}{6}$ ($d = 3$). However, for the RR chains the crossover to asymptotic behaviour occurs so late that, up to $N \sim 1000$, the apparent exponent γ is not distinguishable from 1.

The product $\prod_{i=1}^{N} w^i$ which is the weight of the whole chain is then given by (2.3). Because of this weighting, the mean values for the two procedures have the following form.

The partition function is, for simple sampling,

$$Z_N = \frac{\text{number of chains}}{\text{number of attempts}} = c_0 \left(\frac{q_{\text{eff}}}{q_0 - 1}\right)^N N^{\gamma - 1} \tag{2.4a}$$

and, for RR sampling,

$$Z_N = \frac{\Sigma_{\text{all chains}} W_N(\{r\})}{\text{number of attempts}} = c_0 \left(\frac{q_{\text{eff}}}{q_0 - 1}\right)^N N^{\gamma - 1}. \tag{2.4b}$$

Similarly, we find for any other quantity $X(\{r\})$ of a chain of length N for simple sampling

$$\langle X \rangle = \frac{\Sigma_{\text{all chains}} X(\{r\})}{\text{number of chains of length } N} \tag{2.5a}$$

and for RR sampling

$$\langle X \rangle = \left(\sum_{\text{all chains}} X(\{r\}) W_N(\{r\}) \right) \left(\sum_{\text{all chains}} W_{N(\{r\})} \right)^{-1} \tag{2.5b}$$

where the sum runs over all generated chains. Equations (2.3)-(2.5) define the scheme for calculating all quantities of interest. Here we will first focus on the use of this approach for standard SAW. (If one wants to introduce a temperature, e.g. via nearest-neighbour energies, equations (2.4) and (2.5) hold with the obvious modifications of adding the appropriate Boltzmann factors).

The quantities we are looking for are the partition function Z_N, which is the average weight $\langle W_N \rangle$ defined as

$$\langle W_N \rangle = \left(\sum_{\{r\}} W_N(\{r\}) \right) (\text{number of successfully generated chains})^{-1} \tag{2.6}$$

as well as the whole distribution $W_N(\{r\})$ of the various samples. For the typical physical quantity X of equation (2.5) we usually take the mean squared radius of gyration $\langle R_G^2(N) \rangle$:

$$\langle R_G^2(N) \rangle = \frac{1}{N+1} \sum_{i=0}^{N} \langle (\boldsymbol{r}_i - \boldsymbol{r}_{\text{cm}})^2 \rangle \tag{2.7}$$

where $\boldsymbol{r}_i$ denotes the position of the ith monomer and $\boldsymbol{r}_{\text{cm}}$ is the centre of gravity of the whole walk. We focus our analysis on R_G instead of the mean squared end-to-end distance $\langle R^2(N) \rangle = \langle (\boldsymbol{r}_0 - \boldsymbol{r}_N)^2 \rangle$. R_G has less fluctuations [22] than R. For polymers R_G has a more general physical meaning, if one considers branched structures, for example. Using equations (2.3)-(2.7) we can analyse chains of arbitrary length N. Because our aim is not only to simulate long linear objects, but also star polymers, etc, we generate walks on a high coordination number lattice (FCC with $q_0 = 12$). To test the validity of the RR approach and to construct a criterion for its quality, standard SAW are generated with $0 < N \leqslant 480$. For comparison, for $N = 480$ about 99.8% of the attempts to generate a RR chain are successful. The samples contained up to 130 000 chains. Note that for walks, which can be understood as a critical phenomenon, the relative variance of the radius of gyration $(\Delta R_G^2) = [\langle (R_G^2)^2 \rangle - \langle R_G^2 \rangle^2]^{1/2} / \langle R_G^2 \rangle$ approaches a

constant (≈ 0.405). The distribution functions do *not* become sharper with increasing chain length. One needs the same number of simple sampling chains for all N to achieve the same relative accuracy of the results.

The first approach to check the statistical quality of a set of data is to subdivide a given sample into smaller ones. The fluctuations in the mean values of the subsamples give an estimate of the uncertainties of the whole sample. By doing this we found, for $N > 120$, strong fluctuations in $\langle R_G^2 \rangle$ for the RR method. Even for more than 10^5 generated chains for, e.g., $N = 200$ not only do the mean values of the subsamples fluctuate, but for the whole sample also the three cartesian components of $\langle R_G^2 \rangle$ showed strong fluctuations (about ±30%). However, this was only the case after correcting for the bias. Figure 2 gives an example of a run of 2^{15} attempted chains for $\langle R^2 \rangle$ and $\langle R_G^2 \rangle$.

For the RR walk without correcting for the bias, taken as a different physical system (using equation (2.5*a*)), the error for $\langle R_G^2 \rangle$ was only ±1%. We checked different random-number generators to make sure that dangerous correlations are not hidden in our Markov process. This result, on a first glance, was somewhat surprising to us. Various authors generated chains of similar length N by the RR method and used the results for SAW properties [16-19]. Since the fluctuations disappear if one does not correct for the bias, it must be the weighting (2.5*b*) itself which causes problems. Subsequently we shall try to quantify this rather qualitative statement. As has been pointed out earlier [14], the RR method gives a systematic error which vanishes with increasing sample size. Our intention, however, is to quantify the statistical error of this procedure for sample sizes where no systematic errors remain.

Let us first look at the distribution of the weights $W_N(\{r\})$. The situation becomes worse the more the most relevant weights are pushed to the tails of the distributions. Figure 3 illustrates the influence of the tails of the distribution of weights for different chain lengths $30 \leq N \leq 480$. The sample used here contains 2^{15} (= 32 768) chains. The

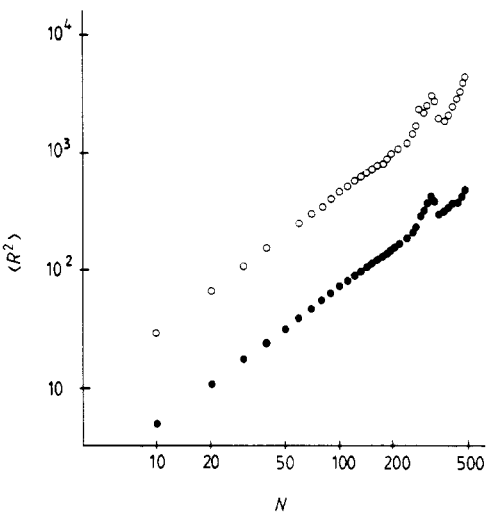

Figure 2. End-to-end distance $\langle R^2 \rangle$(○) and radius of gyration $\langle R_G^2 \rangle$ (●) plotted against N for a sample of 2^{15} chains. Since only 72 chains ran into cages for $N = 240$ (acceptance rate of 99.8%), we quote the number of attempts as the number of chains when describing the plots. For the actual calculations, of course, the real number of generated chains for each length N was used.

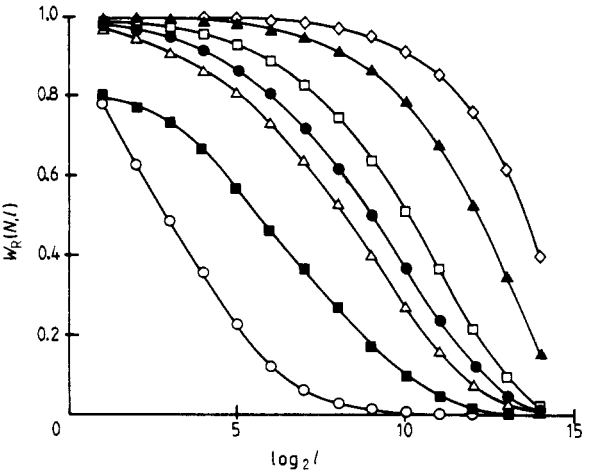

Figure 3. $W_R(N, l)$ against $\log_2 l$ for different lengths N (◇, 30; ▲, 60; □, 120; ●, 150; △, 180; ■, 240; ○, 480).

reduced weight $W_R(N, l)$ is defined as

$$W_R(N,l) = \left(\sum_{\{r\}} W_N(\{r\}) - \sum_{i=1}^{l} W_N(\{r\}, i)\right)\left(\sum_{\{r\}} W_N(\{r\})\right)^{-1} \tag{2.8}$$

with $W_N(\{r\}, 1)$ equal to the weight of the chain with the highest weight, while $W_N(\{r\}, 2)$ denotes the second highest and so forth. $W_R(N, l)$ measures the relative contribution of the l chains with the highest weight to the overall sample. The figure clearly shows that the whole distribution for $N = 480$, 240 is dominated by the few configurations with the biggest weight. Only for chain lengths $N \leq 120$ are we allowed to expect reasonable results for this sample size. One can, of course, argue that the sample size is quite small and one should not expect to get good results. However, the above discussion of the variance of, e.g., $\langle R_G^2 \rangle$ shows that the relative error should not depend on the lengths of the walks at a given sample size, which is the case here. With figure 3 the reason for this loss of reliability now becomes more clear: the relative size of the dominant part of the sample decreases markedly with increasing chain length. The most relevant chains are in the tails of the sample distribution. To get this tail, the sample has to be extremely large. For figure 3 this requires a zero slope for small l. How such a distribution develops to the desired behaviour can be seen from figure 4. As expected, a given value of $W_R(N, l)$ corresponds to varying l proportionally to the sample size. Therefore, the problems of RR sampling arise from the difficulties in sampling tails of distribution functions. This will now be nicely illustrated by a similar analysis of the radius of gyration $\langle R_G^2 \rangle$. For the uncorrected part we calculated $\langle R_G^2 \rangle$ analogously to equation (2.5a), which is simply counting the number of walks for a given interval of R_G^2. The corrected distribution counts the weights of all walks having R_G^2 in a given interval. Figure 5 clearly indicates that the overlap between these two distributions decreases and that the error bars of the mean values of the corrected sample increase greatly with increasing system size. In order to understand the reason for this effect and to develop a strategy to overcome this problem, we have to look more into the theoretical background of the RR method.

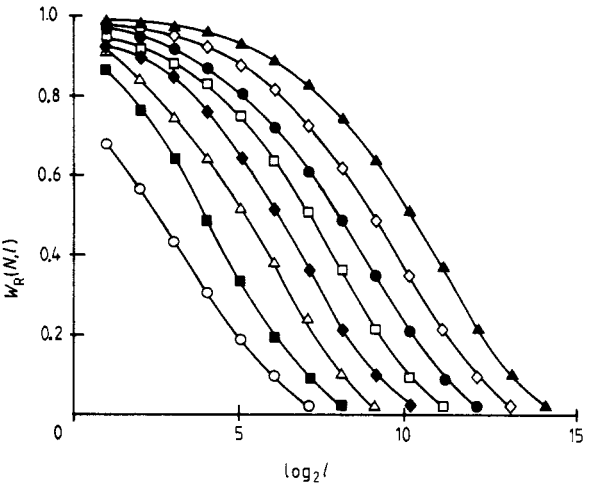

Figure 4. $W_R(N, l)$ against $\log_2 l$ for different sample sizes and fixed chain length $N = 120$. The number of chains considered starts from $2^8 = 256$ up to 2^{15} in powers of 2, where the lower left curve gives the data for the smallest sample.

3. Theoretical analysis

In the previous section, we saw that the biased sampling introduces an effective attraction among the monomers. This means that we have more nearest-neighbour contacts in the RR walk than in the SAW. However, these contacts can be interpreted as the potential energy of the chains. Since the energy itself is an extensive quantity, this gives (for $N \gg 1$)

$$\langle E_{RR} \rangle = a_{RR} N$$

and

$$\langle E_{SAW} \rangle = a_{SAW} N.$$

Here a_{RR}, a_{SAW} are constants with $a_{RR} > a_{SAW}$. $a_{RR} N$ denotes the energy of a RR walk without correcting for the bias (using equation (2.5a)). Since $a_{RR} > a_{SAW}$, the distance in the peaks of the energy distributions diverges linearly with N. Simultaneously, the typical width grows only with $\sqrt{N}$. This is because the heat capacity itself is extensive and just the square fluctuations of E:

$$cT^2 = (\langle E^2 \rangle - \langle E \rangle^2) \propto N. \tag{3.1b}$$

Thus the relative fluctuations in E vanish with $N^{-1/2}$. Figure 6(a, b) displays equation (3.1). It shows that the chains considered are long enough for this argument. Since the distance in the distribution peaks run apart with N, while the width of these distributions goes like $N^{+1/2}$, the overlap vanishes with increasing N. This must cause serious problems for biased sampling. To analyse this we first look at the partition function.

3.1. Partition function

In the previous section we found that a walk with configuration $\{r_i\}$, generated by the RR algorithm, has a probability which is too high by a factor $\{W_N(\{r_i\})\}^{-1}$ compared

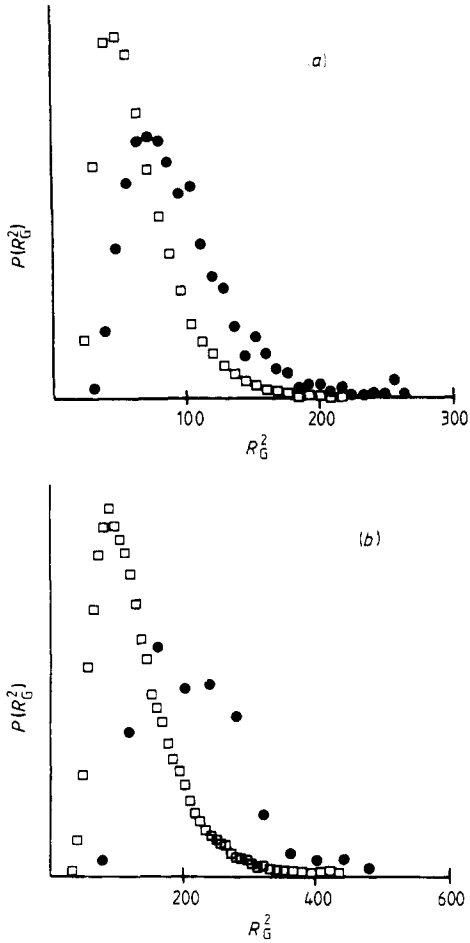

Figure 5. Probability distribution $P(R_G^2)$ against R_G^2 for the corrected and uncorrected distributions. (*a*) gives the results for $N = 120$ while (*b*) corresponds to $N = 240$. Both distributions are taken from a sample of $2^{17} \approx 1.3 \times 10^5$ chains. The circles give the corrected and the squares the uncorrected data.

to the corresponding non-reversal random walk (NRRW). The NRRW is a random walk without direct backfolding and a coordination number $q_1 = q_0 - 1$. However, this compares probabilities of two kinds of systems which have a different set of configurations. What is needed is a comparison of the probabilities of the SAW of the same length. To be specific, we are looking for the probability of a given SAW of length N relative to all configurations of SAW of length N. Here we must not look for the stochastic probability of generating a given configuration by a given algorithm, because this includes all the unsuccessful attempts to build a chain. Taking this into account, a chain generated by biased sampling is too probable by a factor

$$1/W_N^*(\{r_i\}) = [W_N(\{r_i\})(q_0-1)^N/(q_{\text{eff}}^N N^{\gamma-1})]^{-1} \tag{3.2}$$

where $q_{\text{eff}}^N N^{\gamma-1}$ is the number of all the possible SAW configurations.

Following equation (3.2) we can state that a given RR chain counts for $W_N^*(\{r_i\})$ chains which is a number much smaller than 1. To quantify this we rewrite equation

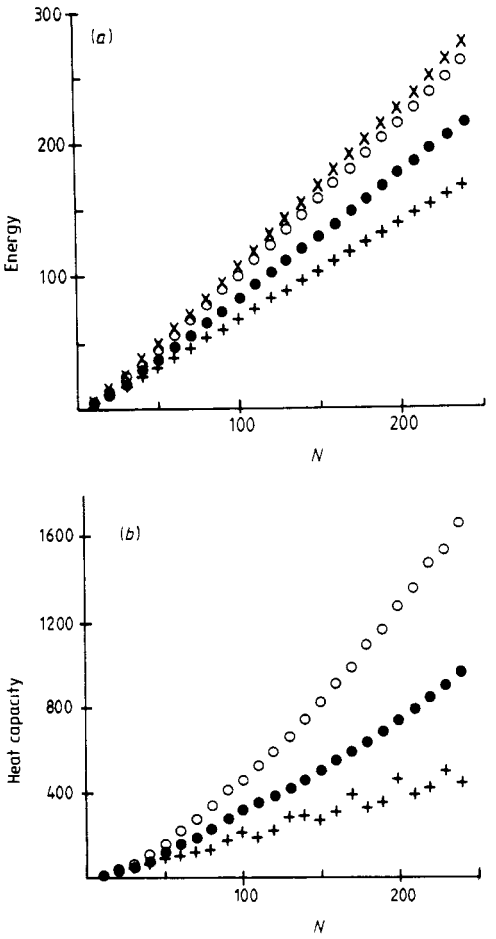

Figure 6. (*a*) Plot of $\langle E(N)\rangle$ against N for $N \leqslant 240$. (*b*) Plot of $\langle E(N)^2\rangle - \langle E(N)\rangle^2$ for $N < 240$. ○, biased sampling; ●, extended biased sampling; +, simple sampling; ×, $T = 7.8$ (SAW in the neighbourhood of $T = \theta$).

(3.2) as a ratio of probabilities,

$$W_N^*(\{r_i\}) = P_N / P_N^{\mathrm{RR}}(\{r_i\}) \tag{3.3}$$

where $P_N = (q_{\mathrm{eff}}^N N^{\gamma-1})^{-1}$ and $P_N^{\mathrm{RR}}(\{r_i\})$ is the probability due to equation (2.2). With $\langle P_N^{\mathrm{RR}}(\{r_i\})\rangle$ being the average probability of an N-step RR chain, equation (3.3) becomes

$$W_N^* = P_N / \langle P_N^{\mathrm{RR}}\rangle. \tag{3.4}$$

The behaviour of $\langle P_N^{\mathrm{RR}}\rangle$ is known quite accurately for certain cases. Although the uncorrected walk belongs to the SAW universality class, $N \leqslant 1000$ [22] to very good accuracy, we can write

$$\langle P_N^{\mathrm{RR}}\rangle = q_{\mathrm{eff,RR}}^{-N} \tag{3.5}$$

with $q_{\mathrm{eff,RR}} < q_{\mathrm{eff}}$. $q_{\mathrm{eff,RR}}$ plays the same role as q_{eff} in (2.4*a*). Thus we write for W_N^*

$$W_N^* = \left(\frac{q_{\mathrm{eff,RR}}}{q_{\mathrm{eff}}}\right)^N N^{1-\gamma} \tag{3.6}$$

with $q_{\text{eff,RR}} = 9.76$ and $q_{\text{eff}} = 10.035$ for the FCC lattice ($N \to \infty$). This results in $W_N^* \simeq N^{-1/6}(1.03)^{-N}$. This suggests that we need $(W_N^*)^{-1}$ times as many chains for the RR method than for simple sampling. However, as figures 3-5 suggest, this estimate is too optimistic.†

Equation (3.5) gives a formal description for the problem occurring but no real physical picture. What does this change in q_{eff} mean? As mentioned earlier, the biased sampling of RR introduces an effective attraction between the monomers of the chain. We use this to rewrite the weight of a chain as

$$W_N(\{r_i\}) = \prod_{i=1}^{N} \left(1 - \frac{m_i}{q_0 - 1}\right) \tag{3.7}$$

where m_i is the number of neighbour sites occupied by other monomers of the chain. This leaves $k_i = q_1 - m_i$ directions for the ith bond with $q_1 = q_0 - 1$. Note that the previous monomer is not counted as a contact in order to be consistent with the usual notation. These contacts are the energy of such a chain if nearest-neighbour interactions are considered (the energy of the last site as starting point for the $(i+1)$th bond is not counted here). Since E is an extensive quantity, the average number of contacts E in such a chain is proportional to the number of monomers. Therefore we can write

$$\langle E \rangle = \left\langle \sum_{i=0}^{N-1} m_i \right\rangle = \langle m \rangle N. \tag{3.8}$$

The energy is governed by the typical density around a given monomer measured with respect to a volume of the order l^3 where l is the bond length. This density along the chain is, to leading order, independent of the length N of the chain. We now make the assumption that to good accuracy the energy $E\{r_i\}$ of a given configuration is given by contacts homogeneously distributed along the chain. With $E(\{r_i\}) = \Sigma_{i=1}^{N-1} m_i$ we then write for equation (3.7)

$$W_N(\{r_i\}) \simeq g(\{r_i\}) = \prod_{i=1}^{N} \left(1 - \frac{\langle m \rangle}{q_1}\right). \tag{3.9}$$

For the FCC lattice this ratio $\langle m \rangle / q_1$ is typically of the order $\frac{1}{10}$. Note that this is a kind of mean-field argument on the single-chain level. On the level of different chains we do take into account fluctuations. They will lead us to our quality criterion. Taking the logarithm of g and expanding to first order yields

$$\ln g(\{r_i\}) = -\langle m \rangle N / q_1$$

and (3.10)

$$g(\{r_i\}) = \exp(-\langle m \rangle N / q_1).$$

This is a quite important result, because this means that the distribution function for the statistical weights W_N plotted on a logarithmic scale is essentially the same as the distribution of the uncorrected energy E. Figure 7 gives a check of this statement and shows that this is a good approximation. The deviation is a consequence of the mean-field-like argument above. By assuming a completely homogeneous contact distribution along the chain, the calculated mean weight is slightly higher than the

† For $d = 2$, equation (3.5) has to be modified to $\langle P_N^{\text{RR}} \rangle = q_{\text{eff}}^{-N} N^{-(\gamma-1)}$ where γ, due to the slow crossover into the asymptotic regime, is an N-dependent quantity varying typically between 1.15 and $\frac{43}{34}$ ($N \to \infty$). For a first criterion, however, it is probably sufficient to omit this N dependence and take a value around 1.2.

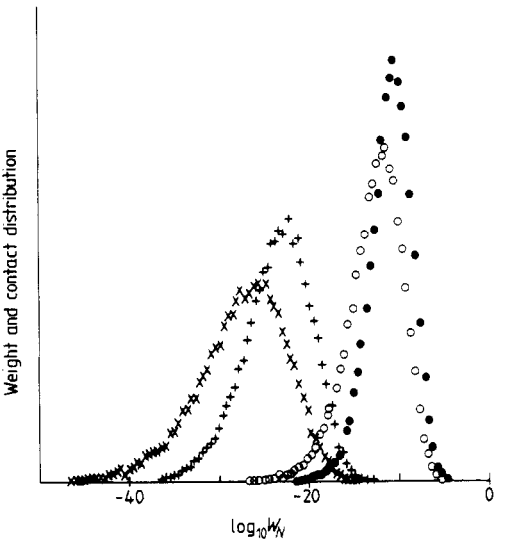

Figure 7. Distribution function of the statistical weights $P(W)$ against ln W for $N = 120$, 240 and of the energy due to (3.10). The normalisation is such that the area under the curves is set to one. $N = 120$: ○, weights; ●, contacts; $N = 240$: ×, weights; +, contacts.

actual one. However, we are able to derive a functional form of the weight distribution. The approximation becomes better the higher the coordination number is. The mapping of weights to the contact energy of a specific generated walk leads us to a direct physical interpretation and quality criterion. Since the weight distribution function is given by the energy contact distribution we can estimate the soundness of the RR sampling.

To quantify this we assume that the energy distribution is a Gaussian and similarly the distribution of the logarithms of the weights. Note that the weight distribution functions are always normalised to one with respect to the real (non-logarithmic weight) scale. With

$$x = \ln W_N(\{r_i\})$$

$$\bar{x} = \langle \ln W_N(\{r_i\}) \rangle$$

$$s = (\langle x^2 \rangle - \langle x \rangle^2)^{1/2} (= \ln b) \tag{3.11a}$$

we get for the normalised distribution function of the logarithms of the weights

$$h(W) = H_0[(2\pi)^{1/2} \ln b]^{-1} \exp[(-1/2 \ln^2 b)(\ln W - \langle \ln W \rangle)^2] \tag{3.11b}$$

which translates to

$$h(x) = H_0(\sqrt{2\pi}s)^{-1} \exp[(-1/2s^2)(x - \bar{x})^2]. \tag{3.11c}$$

H_0 is determined by the condition $\int h(W)\, \mathrm{d}W = 1$. This gives

$$H_0 = \exp(-\bar{x} - \tfrac{1}{2}s^2). \tag{3.12}$$

Using this we now give a direct way of calculating the relative accuracy of a biased sampling sample.

First let us concentrate on the partition function. Using equation (2.4*a*) for the unbiased sample, the probability that a started walk succeeds up to N steps is

$$P_N = \frac{Z_N^{\mathrm{SAW}}}{Z_N^{\mathrm{NRRW}}} = \left(\frac{q_{\mathrm{eff}}}{q_0 - 1}\right)^N \frac{q_0 - 1}{q_0} N^{\gamma - 1}. \tag{3.13}$$

Thus after n_a attempts we have on average

$$\langle n \rangle = n_a P_N \tag{3.14}$$

chains. The deviation from this mean value is then given via $\langle n^2 \rangle = \langle n \rangle^2 + n_a P_N (1 - P_N)$ for uncorrelated events with probability P_N. The relative error in calculating the partition function therefore gives, by means of simple sampling,

$$\frac{\Delta n}{\langle n \rangle} = \frac{(\langle n^2 \rangle - \langle n \rangle^2)^{1/2}}{\langle n \rangle} = \frac{1}{\sqrt{n_a}} \left(\frac{1 - P_N}{P_N}\right)^{1/2}. \tag{3.15}$$

For $P_N \to 0$ this is $\Delta n / \langle n \rangle = \langle n \rangle^{-1/2}$.

To calculate the error for biased sampling we proceed in the same way but using the weight distribution of equation (3.11*b*). From the proposed functional form of the weight distribution, we are able to give an analytic expression for the error. For biased sampling we now follow equation (2.4*b*). To calculate the partition function, we need the average weight $\langle W_N \rangle$ of an N-step walk. With (2.4*b*) this gives

$$\langle W_N \rangle = \frac{1}{n_a} \sum_{i=1}^{n} W_N(\{r_i\}). \tag{3.16}$$

Note that n_a and n are not necessarily the same. For $d = 3$ we can as an excellent approximation set n_a (number of attempts) $= n$ (number of successful chains). However, for $d = 2$ we have to distinguish between the two. With (3.16) we now only have to calculate $\langle W_N \rangle^2$ and $\langle W_N^2 \rangle$ in order to get the fluctuations. Using (3.11*c*) and (3.12) we have to calculate $\langle W_N(\{r_i\}) \rangle^2$ and $\langle W_N^2(\{r_i\}) \rangle$ which yields for $n_a = n$ ($d = 3$)

$$\begin{aligned} \langle W_N^2 \rangle &= \exp(2\bar{x} + 4s^2) \\ \langle W_N \rangle^2 &= \exp(2\bar{x} + 3s^2) \end{aligned} \tag{3.17}$$

where $\bar{x}$ and s^2 are defined in (3.11) for the distribution of the logarithms but with the normalisation $\int \mathrm{d}W h(W) = 1$.

Therefore the relative real width of the weight distribution is given by

$$\frac{\delta W_N}{\langle W_N \rangle} = \frac{(\langle W_N^2 \rangle - \langle W_N \rangle^2)^{1/2}}{\langle W_N \rangle} = \{\exp[(s^2) - 1]\}^{1/2} \simeq \exp(\tfrac{1}{2}s^2). \tag{3.18}$$

This results in an error ΔW for the partition function analogous to (3.15),

$$\Delta W / \langle W \rangle = (1/\sqrt{n}) \exp(\tfrac{1}{2}s^2). \tag{3.19}$$

On the other hand, s^2 is directly related to the specific heat of the biased chains. For the actual simulation this gives a very simple way of estimating the accuracy. In the framework of our mean-field-like argument (equations (3.9) and (3.10)) we identified, to first order, $\ln W_N(\{r_i\})$ with the number of contacts along the chain. With this we identify, via equation (3.11*a*),

$$\langle \ln^2 W \rangle - \langle \ln W \rangle^2 = (1/q_1)^2 (\langle (mN)^2 \rangle - \langle mN \rangle^2) \tag{3.20a}$$

and

$$\langle \ln W \rangle = -(1/q_1) \langle mN \rangle. \tag{3.20b}$$

Using the data of figure 7 the agreement is reasonably good. For the FCC lattice we thus find for large N ($N > 200$, figure 6(b))

$$\langle s^2 \rangle = (9.5N - 630)/q_1^2 \qquad N \to \infty. \tag{3.21}$$

For figure 8 we calculated $\Delta W/W$ due to equations (3.20a) and (3.21) for each chain length indicated, as well as the error in Z_N, directly as mentioned in the figure caption. For $N > 40$ our theory and the sample data coincide. As one expects for short chains, our assumption of the functional form of the weight distribution is no longer valid. This is clearly shown by figure 8. The error of the RR sample increases exponentially with N while that of the SS sample remains constant. However, one should be aware of the fact that the corresponding error increase of the SS method is significantly larger if one counts the SS attempts rather than the successfully completed chains. Qualitatively, this occurs because each generated RR chain gives information proportional to its weight for the chain set investigated, while information is lost if a chain must be stopped in the SS procedure.

If one now accepts as a criterion that a biased sample of n_{RR} chains should have the same relative error as a sample of n_{SS} unbiased chains, (3.19) and (3.15) yield

$$n_{RR}/n_{SS} = \exp(s^2). \tag{3.22}$$

For the FCC lattice we find from figure 8, for $N = 100$, $n_{RR}/n_{SS} = 28$. For $N = 240$ we use (3.22) and find $n_{RR}/n_{SS} = 8.3 \times 10^5$, in good agreement with the interpretation of figure 3. There we found a sample of 3×10^5 chains not sufficient for this chain length.

Note that the present criterion is much deeper in a physical sense than the first approach of equation (3.6). Here we make use of the contact distribution of the biased sample and relate the quality of our sample to the mean value and the width of this contact distribution.

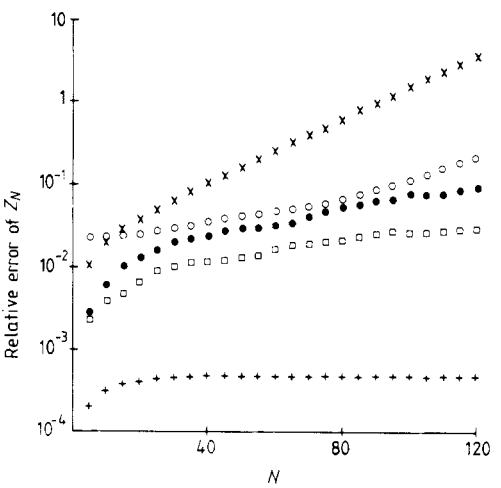

Figure 8. Plot of sample qualities containing 2048 chains against N. ×, 2048 SS attempts; ○, predicted error; ●, measured error; □, measured EBS error (see § 4); +, error of 2048 SS chains. The experimental error bars were determined by taking 64 samples each containing 2048 chains. For $N \geq 100$ we get an increasing deviation between the predicted and the measured error. Here the systematic errors of the RR method mentioned in § 2 become important. As 2048 chains are not enough for $N \geq 100$ we get an apparent lower weight with less fluctuation.

It should be noted that in the present form the argument only holds for $n_a = n$, which means that all the attempts to produce a biased chain are successful. If this is not the case, namely $n_a > n$, we have to start again at (3.17) and modify our arguments slightly. We rewrite (3.16) as

$$\langle W_N \rangle = \left(\frac{n_a}{n}\right) \frac{1}{n_a} \sum_{i=1}^{n_a} W_N(\{r_i\}). \tag{3.23}$$

We then can use the same arguments as before but have to incorporate the 'probability of existence' n_a/n in our argument, similar to the case of the unbiased sample. Doing this, equation (3.19) becomes

$$\left(\frac{\Delta w}{\langle w \rangle}\right)^2 = \left(\frac{1}{\sqrt{n_a}} \exp(\tfrac{1}{2}s^2)\right)^2 + \frac{1}{n_a}\left(\frac{n_a}{n} - 1\right). \tag{3.24}$$

For sampling with a soft bias [6, 23, 24] this smoothly crosses over to unbiased sampling for $s^2 \to 0$. As equation (3.24) illustrates, the dominant contribution comes from the exponential part.

3.2. Measurable quantities

In the previous section we looked at the partition function generated by a biased sampling procedure. This is a very special case, because one is looking for a quantity relative to the exactly known NRRW results. As measurable quantities we here define the number of contacts E or the radius of gyration R_G^2 and so on. These quantities cannot be directly related to the exact NRRW data simply by looking at probabilities. Here we are looking at the physical system itself, which means that we must relate our results to the set of SAW rather than of NRRW. As we will see below, this has consequences for the way we have to argue in order to find the error bars. For the partition function the natural way to correct for the bias is given by the coordination number q_1 of the corresponding NRRW. This is not necessarily the case here. Using (2.5*b*) the normalisation of the weights is arbitrary. Any normalisation of our correction factor cancels in (2.5*b*). However, here (2.3*c*) should be replaced by

$$P_{RR}^i w^i = 1/\bar{f}_{SS}(N) \tag{3.25}$$

with $\bar{f}_{SS}(N)$ being the mean number of vacant sites if one is building up a SAW by simple sampling. As $1/\bar{f}_{SS}(N)$ is the probability of an SS walk to proceed to a vacant site while growing, we thus have taken into account that the new reference sample is the set of SAW. $\bar{f}_{SS}(N)$ can simply be related to the contact number $m_{SS}(N)$ of the SS walk by

$$\bar{f}_{SS}(N) = q_1 - m_{SS}(N) \tag{3.26}$$

where $m_{SS}(N)$ asymptotically is a constant.

At first glance one would expect $\bar{f}_{SS}(N) = q_{eff}(N)$. Figure 9, where $q_{eff}(N)$ and $\bar{f}_{SS}(N)$ are plotted against $1/N$, shows that these two quantities differ significantly. For the FCC lattice $q_{eff}(N)$ extrapolates to $q_{eff} = 10.035$ in very good agreement with earlier investigations, while $\bar{f}_{SS}(N)$ extrapolates to $\bar{f}_{SS} = 10.29$. What was surprising to us is this strong difference between q_{eff} and $\bar{f}_{SS}$. For $d = 2$, as figure 10 illustrates, one can expect such a difference. For $d = 3$ such cages, however, are very rare. As we will see below, the quantities we quote give reliable estimates for the error bars.

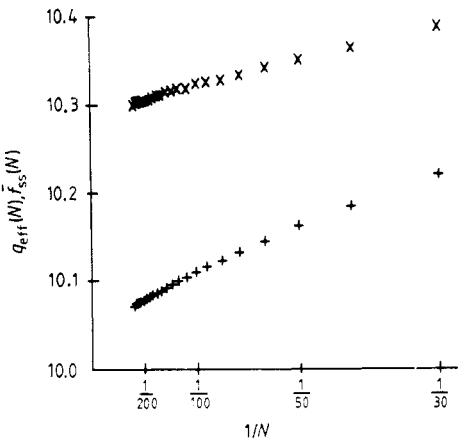

Figure 9. $q_{eff}(N)$ (+) and $\bar{f}_{SS}(N)$ (×) against N^{-1} (2^{17} chains created by EBS; see § 4). Note that the two quantities differ significantly.

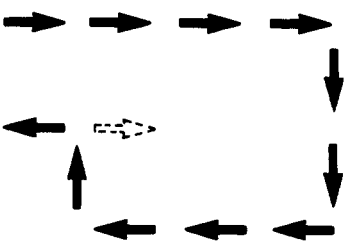

Figure 10. 2D illustration of one possible source for the difference between q_{eff} and $\bar{f}_{SS}$. For f_{SS} the open site leading into the dead end counts for the infinitely long chain while for q_{eff} this does not count.

After this discussion we can proceed to calculate the error bars of the biased sample. We want to calculate a physical quantity $X(N)$. The distribution function of $X(N)$ should have a relative width $\delta X = \langle (X^2 - \langle X \rangle^2)^{1/2} / \langle X \rangle$. Both the mean value $\langle X \rangle$ and the width δX might depend on N. For R_G^2, δX approaches a constant for large N while for the contacts δX should be proportional to $N^{-1/2}$. Note that here distribution always means the corrected distribution or the (same) directly sampled ss distribution. The relative error $\Delta X(N)$ of a sample of n_{SS} chains is then given for ss by

$$\Delta X_{SS} = n_{SS}^{-1/2} \delta X. \tag{3.27a}$$

With suitable normalisation of the weights for biased sampling we get as lower limit for the relative error

$$\Delta X_{BS} = \left(\sum_{i=1}^{n} W'_N(\{r_i\}) \right)^{-1/2} \delta X \tag{3.27b}$$

with

$$W'_N(\{r\}) = \prod_{j=1}^{N} f_j / \bar{f}_{SS} = \left(\frac{q_1}{f_{SS}} \right)^N \prod_{j=1}^{N} \frac{f_j}{q_1} = \left(\frac{q_1}{f_{SS}} \right)^N W_N(\{r\}) \tag{3.28}$$

where f_j is the number of free sites to proceed to for link j. Figure 11 gives the calculated and the 'measured' error bars for a biased sample of 2048 chains for R_G^2.

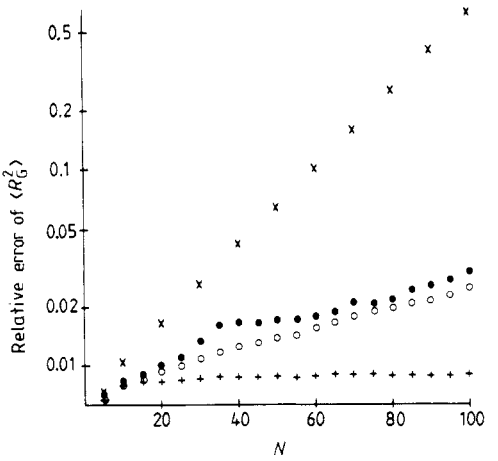

Figure 11. Error bars of the radius of gyration. Symbols as in figure 8. The procedure for determining the experimental errors is completely analogous to that of figure 8.

If we now have n_{SS} simple sampling chains we need, with (3.28), on average

$$n_{BS} = n_{SS} \langle W'_N \rangle^{-1} \tag{3.29}$$

biased chains to arrive at the same relative error. For the FCC lattice we find

$$\ln (n_{BS}/n_{SS}) = N \left[\frac{\langle m \rangle}{q_1} - \ln \frac{q_1}{\bar{f}_{SS}} \right] = 3.78 \times 10^{-2} N - 1.5 \tag{3.30}$$

giving a ratio of 21 for $N = 120$ and 1.9×10^{-3} for $N = 240$. Note that (3.30) holds for any quantity besides the partition function.

At this point we should remember that our arguments are in principle mean-field-like, especially if we consider the strong fluctuations in the corrected R_G distribution of figure 5. However, as the analysis shows, these fluctuations are of minor importance for our case.

It is important to note that the relative accuracy of the partition function is much worse compared to the 'measurable quantities'. The reason is the need of taking the NRRW set as a reference system for the partition function.

4. Extensions and modifications

4.1. Extended biased sampling (EBS)

There are different ways to improve the RR method by looking more than one step ahead. Meirovitch [7] proposed his 'scanning future step' algorithm. He wants to continue a given walk by n additional steps and then looks for all ways to perform n steps. He then takes one of these possible choices. For $n = 1$ this reduces to the biased sampling described above. In practice, n usually is restricted to $n \leqslant 3$, because otherwise the enumeration of the 'future steps' is too time consuming.

Here we want to propose a more direct approach to the 'scanning future steps' algorithm. We still want to perform a one-step growth of the walk. Nevertheless, the new step has some knowledge about the future further out. In this paper we constrain ourselves to the two-step extension. Further generalisations are straightforward.

Consider a walk of N steps. In order to add the $(N+1)$th step we first look for empty sites to proceed to. Then we count for each open site the number of open sites for the $(N+2)$th step. We then proceed to one of the open sites with a probability proportional to the number of empty sites for step $N+2$.

Figure 12 gives an example for a walk on the square lattice. Let q_n be the weight normalisation constant, which is q_1 for the partition function and $\bar{f}_{SS}$ for 'measurable quantities' as defined in § 3. The probability p_i of step i is then

$$p_i = k_t \left(\sum_{j=1}^{m_i} k_j \right)^{-1}. \tag{4.1}$$

m_i is the number of jump sites for the following step while k_j is the number of empty neighbours for step $i+1$ if the system proceeds in direction j. The weight of a generated walk is then defined by

$$W_N(\{r_i\}) = \prod_{i=1}^{N} \frac{p_i^{-1}}{q_n}. \tag{4.2}$$

Here a mapping of weights onto contacts, analogous to equation (3.9), is not possible. However, it is evident that EBS gives more accurate results as a huge number of low-weighted configurations with many contacts is avoided (figures 6, 8 and 13). The weighting of both RR and EBS leads of course to the same partition function Z_N. This

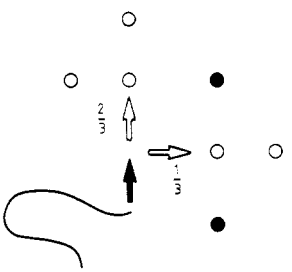

Figure 12. Extended biased sampling. The site having two free neighbours is preferred. This leads to less collapsed configurations.

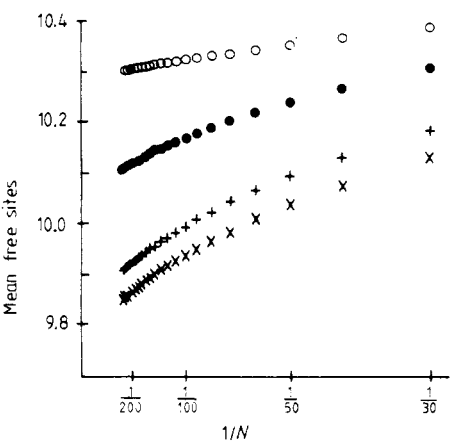

Figure 13. Mean free sites $f(N)$ for BS (×), EBS (+), SS (○) and $q_{\text{eff}}(N)$ (●).

means that for measurable quantities the lower error limit is the same. EBS of comes closer to this limit.

Note that for $d=3$ and the partition function again it is sufficient to consider all attempts of building a chain as successful for all practical N. For $d=2$ again the effect of cages has to be incorporated, as mentioned in § 3. However, as should be mentioned, EBS does not sample exactly the same configurations as BS. For the last step a path leading into termination of the walk at the next step is not sampled. This is not severe for $d=3$, but for $d=2$ one should be aware of some deviations from the standard sample for short chains. In this case, one should modify EBS slightly taking $p_i=(k_i+1)/\Sigma_j^{m_i}(k_j+1)$ instead of (4.1).

4.2. Soft biased sampling (SBS)

SBS is a one-step procedure as is the standard BS. However, to overcome the huge deviations from BS distributions and SS distributions in some cases it is useful to allow for more or less probable directions. This is done even if the less probable one leads to termination of the chain. This method has proven to be quite useful for some surface and interface problems. To determine the accuracy of the 'measurable quantities' we follow exactly the same procedure described above. Because of the huge number of terminated chains for analysis of the partition function we need to consider the modified treatment of (3.23) and (3.24).

4.3. Biased sampling with temperature

BS turns out to be especially useful for investigations of the collapse transition of polymer chains. A first attempt in this direction has already been made by Mazur and McCrackin [14]. As discussed at the beginning of this paper one can introduce a temperature by giving a walk with m nearest-neighbour contacts a Boltzmann weight $\exp(m/T)$ in the case of attraction between the bonds. The Boltzmann constant is set to unity. Because BS prefers configurations with many contacts the bias can be interpreted as an effective attraction between monomers. As one can directly see from the arguments above, the effective temperature T_{BS} of uncorrected biased sample simulation is given by

$$\bar{f}_{SS}(T_{BS})=\bar{f}_{BS} \tag{4.3}$$

where we use notation analogous to (3.29). For the FCC lattice $T_{BS}=8.5$ which is slightly above the theta temperature. It is well known that biased sampling chains are very much near the theta temperature [25, 26], but as given above and seen earlier numerically [22] and analytically [27] such chains still belong to the SAW universality class.

Now SS chains give a much worse result compared to BS.

5. Conclusion

In the present paper we have developed a criterion to calculate the accuracy of biased sampling compared to direct simple sampling simulations. In most cases BS is much worse compared to SS. However, because of the almost 100% success rate of biased sampling there exists a wide region of chain lengths where BS is of technical advantage

even for standard SAW. For increasing chain length N BS becomes exponentially bad. Typical ways out, such as Meirovitch's 'scanning future step' algorithm or extended biased sampling (EBS) proposed here do not overcome this problem. These methods shift the limit of applicability to larger N. For the FCC lattice and the EBS with two steps looking ahead the limiting chain length is increased by a factor of two. For most practical purposes this is sufficient. This allows us to sample relatively long chains ($N \leqslant 240$) with very high accuracy. For much longer chains one has to use dynamic methods which do not allow us to calculate γ and q_{eff} [3] or one has to use the constant fugacity approach [11]. Note that both methods do not allow the simultaneous analysis of different chain lengths and/or temperatures. They are therefore naturally very time consuming and will only be of advantage for very long chains. We think the above discussion shows that there was an urgent need for a more detailed theoretical analysis of biased sampling procedures. The philosophy of our treatment can of course be translated to the umbrella sampling techniques [28] as well. As far as we know no such treatment for the umbrella sampling techniques has been performed up to now. In addition, as we saw above, for investigating the θ properties ($d = 3$) BS seems to be the ideal approach. However, for $d = 2$ the situation is much worse. Standard BS is not able to sample the θ properties or SAW properties [29]. For this case non-standard methods, such as BS with temperature, etc, have to be employed.

To conclude this paper, we repeat the recipes developed in the previous section. In order to estimate whether a BS approach can be used for the investigation of SAW properties, we simply have to calculate the average weight due to (3.28) and (3.29). This directly gives the number of chains n_{BS} we need to arrive at the accuracy of a sample of n_{SS} simple sampling chains. By modifying $\bar{f}_{\text{SS}}$ one can adjust this criterion to various temperatures. This holds for the measurable quantities as defined in § 3.2. For the partition function the only difference is that we have to normalise the weights by q_1 instead of $\bar{f}_{\text{SS}}$. Then the logarithmic relative width of the weight distribution directly gives an accuracy criterion, due to (3.20). Thus we have derived a direct way of calculating the accuracy of biased sampling procedures. The formulae allow us to extrapolate to arbitrary chain lengths. The above arguments give tools to estimate from very cheap simulations whether an extensive project using long chains generated by a kind of BS algorithm can lead to reasonable results or not.

Acknowledgments

This work was supported in part by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 41. The authors thank K Binder, B Dünweg, D W Heerman and J W Lyklema for stimulating the helpful comments and discussions.

References

[1] McKenzie D S 1976 *Phys. Rep.* **27** 37
[2] Domb C 1969 *Adv. Chem. Phys.* **15** 229
[3] Baumgärtner A 1984 *Applications of the Monte Carlo Method in Statistical Physics* (*Topics in Current Physics* **36**) ed K Binder (Berlin: Springer) p 145
[4] de Gennes P G 1979 *Scaling Concepts in Polymers Physics* (Ithaca, NY: Cornell University Press)
[5] Rosenbluth M and Rosenbluth A 1955 *J. Chem. Phys.* **23** 356
[6] Kremer K 1983 *PhD thesis* Cologne

[7] Meirovitch H 1982 *J. Phys. A: Math. Gen.* **15** L735
[8] Alexandrowicz Z 1969 *J. Chem. Phys.* **51** 561
[9] Baumgärtner A 1984 *Ann. Rev. Phys. Chem.* **35** 419
[10] Redner S and Reynolds P J 1981 *J. Phys. A: Math. Gen.* **14** 2679
[11] Berretti A and Sokal A D 1985 *J. Stat. Phys.* **40** 483
[12] Binder K (ed) 1983 *Monte Carlo Methods in Statistical Physics* (*Topics in Current Physics* **7**) (Berlin: Springer)
[13] Domb C 1974 *Polymer* **15** 259
[14] Mazur J and McCrackin F L 1968 *J. Chem. Phys.* **49** 648
McCrackin F L, Mazur J and Guttmann C M 1973 *Macromol.* **6** 859
[15] Kremer K, Baumgärtner A and Binder K 1982 *J. Phys. A: Math. Gen.* **15** 2879
[16] Lipson J E G, Whittington S G, Wilkinson M, Martin J L and Gaunt D J 1985 *J. Phys. A: Math. Gen.* **18** L469
Wilkinson, M K, Gaunt, D J, Lipson J E G and Whittington S G 1986 *J. Phys. A: Math. Gen.* **19** 789; 1986 *Macromol.* **19** 1241
[17] Barett A J and Tremain D L 1987 *Macromol.* **20** 1687
[18] Kolinski A and Sikorski A 1982 *J. Polym. Sci.* **20** 3147
[19] Mazur J and McCrackin F L 1977 *Macromol.* **10** 326
[20] Smith N C and Fleming R J 1975 *J. Phys. A: Math. Gen.* **8** 929
[21] Nienhuis B 1982 *Phys. Rev. Lett.* **49** 1062
[22] Lyklema J W and Kremer K 1986 *J. Phys. A: Math. Gen.* **19** 279
Kremer K and Lyklema J W 1985 *Phys. Rev. Lett.* **55** 2091
[23] Eisenriegler E, Kremer K and Binder K 1982 *J. Chem. Phys.* **77** 6292
[24] Kremer K 1985 *J. Chem. Phys.* **83** 5882
[25] Majid J, Jan N, Coniglio A and Stanley H E 1984 *Phys. Rev. Lett.* **52** 1257; 1985 *Phys. Rev. Lett.* **55** 2092
[26] Guttman C 1986 *Macromol.* **19** 833
[27] Pietronero L 1985 *Phys. Rev. Lett.* **55** 2025
[28] Mezei M 1987 *J. Comput. Phys.* **68** 237
[29] Lyklema J W and Kremer K 1986 unpublished