Chapter 6
Large Deviations in Monte Carlo Methods

Andrea Pelissetto and Federico Ricci-Tersenghi

Abstract Numerical studies of statistical systems aim at sampling the Boltzmann-Gibbs distribution defined over the system configuration space. In the large-volume limit, the number of configurations becomes large and the distribution very narrow, so that independent-sampling methods do not work and importance sampling is needed. In this case, the dynamic Monte Carlo (MC) method, which only samples the relevant "equilibrium" configurations, is the appropriate tool.

However, in the presence of ergodicity breaking in the thermodynamic limit (for instance, in systems showing phase coexistence) standard MC simulations are not able to sample efficiently the Boltzmann-Gibbs distribution. Similar problems may arise when sampling rare configurations. We discuss here MC methods that are used to overcome these problems and, more generally, to determine thermodynamic/statistical properties that are controlled by rare configurations, which are indeed the subject of the theory of large deviations.

We first discuss the problem of data reweighting, then we introduce a family of methods that rely on non-Boltzmann-Gibbs probability distributions, umbrella sampling, simulated tempering, and multicanonical methods. Finally, we discuss parallel tempering which is a general multipurpose method for the study of multimodal distributions, both for homogeneous and disordered systems.

6.1 Introduction

Statistical mechanics was developed at the end of the nineteenth century to provide a theoretical framework to thermodynamics. However, the complexity of the formulation made ab initio calculations essentially impossible: only ideal

A. Pelissetto • F. Ricci-Tersenghi (✉)
Dipartimento di Fisica and INFN, Università "Sapienza", Piazzale Aldo Moro 5, Rome, I-00185, Italy
e-mail: andrea.pelissetto@roma1.infn.it; federico.ricci@roma1.infn.it

A. Vulpiani et al. (eds.), *Large Deviations in Physics*, Lecture Notes in Physics 885,
DOI 10.1007/978-3-642-54251-0_6,

(noninteracting) systems could be treated exactly, the two-dimensional Ising model being a notable exception. To understand the behavior of more complex systems, crude approximations and phenomenological models, in most of the cases only motivated by physical intuition, were used. The understanding of statistical systems changed completely in the late 1950s, when computers were first used [1–4]. The first *machine calculations* showed that the behavior of macroscopic systems containing a large number of molecules (of the order of the Avogadro's number, $N_A \approx 6.022 \cdot 10^{23}$) could be reasonably reproduced by relatively small systems with a number of molecules of the order of 10^2–10^3, which could be simulated with the computer facilities of the time. These results, which, for many years, were met with skepticism by the more theoretically-oriented part of the statistical-mechanics community, opened a new era: theoreticians had their own laboratory, in which they could analyze the behavior of different systems under well-controlled *theoretical* conditions. Since then, numerical methods have been extensively used and have provided quantitatively accurate predictions for the behavior of many condensed-matter systems. Similar methods have also been employed in many other fields of science, from high-energy physics (in the 1970s the first lattice QCD simulations were performed) to astrophysics, chemistry, biology, statistics, etc.

The Monte Carlo (MC) method is one of the most powerful techniques for the simulation of statistical systems. Since the Boltzmann-Gibbs distribution is strongly concentrated in configuration space, MC methods implement what is called *importance sampling*: points in configuration space are not generated randomly, but according to the desired probability distribution. In practice, in a MC simulation one only generates typical configurations, i.e. those that most contribute to thermodynamic averages. From a mathematical point of view, a MC algorithm is a Markov chain that (a) is stationary with respect to the Boltzmann-Gibbs distribution and (b) satisfies ergodicity (mathematicians call the latter property irreducibility). If these two conditions are satisfied, time averages converge to configuration averages: hence, by using the MC results one can compute ensemble averages for the system at hand. While condition (a) is usually easy to satisfy—the Metropolis algorithm is a general purpose method to define a Markov chain that satisfies (a)—condition (b) is more subtle. Indeed, in the presence of phase transitions or of quenched disorder, a statistical system may show an infinite number of inequivalent thermodynamic states in the infinite-volume limit, which in turn implies ergodicity breaking for any (physical or MC) local dynamics. For instance, consider the Ising model in a finite volume with some boundary conditions that do not break the up-down symmetry (for instance, the usual periodic boundary conditions). Since the symmetry is exactly preserved, the magnetization per site m is exactly zero. However, if the temperature T is low enough, in any MC simulation of a sufficiently large system one observes that the system magnetizes, i.e. m is equal either to m_0 or to $-m_0$. This result can be easily understood. The correct distribution $P(m)$ of the magnetization has maxima $P_{\max}$ at $\pm m_0$ and a minimum $P_{\min}$ at $m = 0$. The important point is that the ratio $P_{\min}/P_{\max}$ is extremely small, of the order of e^{-aN^p}, $a, p > 0$, where N is the number of system variables. To obtain the correct average one should sample all relevant configurations, i.e., both those that have $m \approx m_0$ and those that have

$m \approx -m_0$. But these two regions of configuration space are separated by a barrier of rare configurations, i.e. that occur with an exponentially small probability and which, therefore, are never sampled—importance sampling MC samples only the typical configuration space. Hence, any simulation gets stuck in one of the two minima, ergodicity is lost, and therefore MC does not provide the correct answer.

In this contribution we wish to discuss MC methods that are used to overcome the problem of ergodicity breaking and, more generally, to determine thermodynamic/statistical properties that are controlled by rare configurations, which are indeed the subject of the theory of large deviations [5]. In this contribution we will first discuss the problem of data reweighting, then we will introduce a family of methods that rely on non-Boltzmann-Gibbs probability distributions, umbrella sampling, simulated tempering, and multicanonical methods. Finally, we will discuss parallel tempering which is a general multipurpose method for the study of multimodal distributions, both for homogeneous and disordered systems.

6.2 Data Reweighting

In this contribution we shall work in the canonical ensemble, considering configurations x distributed according to the Boltzmann-Gibbs probability density

$$\pi_\beta(x) = \frac{e^{-\beta H(x)}}{Z_\beta} ,$$

where $H(x)$ is the energy function and the normalizing constant Z_β is the partition function at inverse temperature β. Note that the energy function H is extensive, i.e., proportional to the number N of system variables; in the thermodynamic limit $N \to \infty$, the distribution $\pi_\beta(x)$ becomes peaked around its maximum. We indicate by $\langle\cdot\rangle_\beta$ the average with respect to $\pi_\beta(x)$.

The dynamic MC method which uses importance sampling can efficiently sample from a distribution which is strongly concentrated in the space of configurations as $\pi_\beta(x)$ is for large values of N. Thus, a MC run at β_0 allows us to compute any interesting thermodynamic quantity at β_0. However, suppose that we are also interested in the behavior at a different temperature β_1: do we need to run a new MC simulation or can we re-use the data collected at β_0? The answer mainly depends on the energy function $H(x)$, on how close β_1 and β_0 are, and, though this is usually much less relevant, on the amount of data collected at β_0.

In this context one useful technique is called *data reweighting* [6–8]. If $A(x)$ is any observable, its average at β_1 can be expressed as

$$\langle A\rangle_{\beta_1} = \frac{\sum_x A(x)e^{-\beta_1 H(x)}}{\sum_x e^{-\beta_1 H(x)}} = \frac{\sum_x A(x)e^{-\Delta\beta H(x)}e^{-\beta_0 H(x)}}{\sum_x e^{-\Delta\beta H(x)}e^{-\beta_0 H(x)}} = \frac{\langle A\, e^{-\Delta\beta H}\rangle_{\beta_0}}{\langle e^{-\Delta\beta H}\rangle_{\beta_0}} , \tag{6.1}$$

where $\Delta\beta = \beta_1 - \beta_0$.

Though in principle this formula solves the problem, in practice it is only useful if the two averages at β_0 can be computed with reasonable accuracy. But this is not obvious. Since H is extensive, the calculation of averages involving $e^{-\Delta\beta H}$ is a large-deviation problem for $N \to \infty$. Therefore, accurate results are only obtained if rare configurations, i.e. configurations that have an exponentially small probability for $N \to \infty$, are correctly sampled. From a physical point of view the origin of the difficulties can be understood quite easily. Problems arise because configurations sampled by the MC at β_0 are not those giving the largest contribution to $\langle A\rangle_{\beta_1}$, since $\pi_{\beta_1}(x)$ and $\pi_{\beta_0}(x)$ are strongly concentrated on different configurations. If we call D_β the set of typical configurations of $\pi_\beta(x)$,[1] then the estimate of $\langle A\rangle_{\beta_1}$ obtained by data reweighting is reliable only if the configurations obtained at β_0 sample well enough D_{β_1}. Usually this requirement is stated by saying that the energy histograms at inverse temperatures β_0 and β_1 should overlap. This statement is qualitatively correct, although of little practical use, given that we do not know the energy histogram at β_1 (this is something we would like to compute from the data measured at β_0).

Data reweighting provides also the answer to a second problem that arises in many different contexts, that of computing free energy differences. In the canonical ensemble one would consider the Helmholtz free energy $F(\beta) = -\beta^{-1} \ln Z_\beta$. Given $F(\beta_0)$, one can compute $F(\beta_1)$ by using

$$\beta_1 F(\beta_1) - \beta_0 F(\beta_0) = -\ln\langle e^{-\Delta\beta H}\rangle_{\beta_0} = \ln\langle e^{\Delta\beta H}\rangle_{\beta_1} \,. \tag{6.2}$$

The same type of averages appear here as in Eq. (6.1) and indeed, this type of computations suffers from the same problems discussed above.

We wish now to make this qualitative discussion quantitative. For this purpose, let us compute the statistical error on $\langle A\rangle_{\beta_1}$. Since this quantity is expressed as a ratio of two mean values, the variance of the estimator can be obtained by using the general expression

$$\sigma^2_{\rm est} \equiv \mathrm{var}\left(\frac{\frac{1}{n}\sum_i A_i}{\frac{1}{n}\sum_i B_i}\right) = \frac{1}{n}\frac{\langle A\rangle^2}{\langle B\rangle^2}\left\langle \mathscr{O}^2\right\rangle (1 + 2\tau_{\mathscr{O}}) + O(n^{-2}) \,, \tag{6.3}$$

where n is the number of measurements performed,

$$\mathscr{O} = \frac{A}{\langle A\rangle} - \frac{B}{\langle B\rangle} \,, \tag{6.4}$$

and $\tau_{\mathscr{O}}$ is the integrated autocorrelation time associated with $\mathscr{O}$. Equation (6.3) is valid as $n \to \infty$, neglecting corrections of order n^{-2}. In our case the relevant quantity is $\left\langle \mathscr{O}^2\right\rangle$. If we specialize Eq. (6.4) to our case, we obtain

[1] A precise definition of D_β is not necessary for our purposes. For example, we can consider for D_β the smallest set of configurations such that $\sum_{x\in D_\beta} \pi_\beta(x) > 1 - \varepsilon$.

$$\langle \mathscr{O}^2 \rangle_0 = \left\langle \left(\frac{A e^{-\Delta\beta H}}{\langle A e^{-\Delta\beta H} \rangle_0} - \frac{e^{-\Delta\beta H}}{\langle e^{-\Delta\beta H} \rangle_0} \right)^2 \right\rangle_0 = \left\langle \left(\frac{A}{A_1} - 1 \right)^2 \frac{e^{-2\Delta\beta H}}{\langle e^{-\Delta\beta H} \rangle_0^2} \right\rangle_0$$
$$= \frac{Z_0^2}{Z_1^2} \left\langle \left(\frac{A}{A_1} - 1 \right)^2 \right\rangle_2 \langle e^{-2\Delta\beta H} \rangle_0 = \frac{Z_0 Z_2}{Z_1^2} \left\langle \left(\frac{A}{A_1} - 1 \right)^2 \right\rangle_2 .$$

Here we have introduced $\beta_2 = 2\beta_1 - \beta_0$, $\langle \cdot \rangle_{\beta_i}$ has been written as $\langle \cdot \rangle_i$, and $A_1 = \langle A \rangle_1$. In terms of the Helmholtz free energy $F(\beta)$ we have

$$\langle \mathscr{O}^2 \rangle_0 = \left\langle \left(\frac{A}{A_1} - 1 \right)^2 \right\rangle_2 e^{Nf(\beta_0, \beta_1)} ,$$

where

$$Nf(\beta_0, \beta_1) = 2\beta_1 F(\beta_1) - \beta_0 F(\beta_0) - \beta_2 F(\beta_2) .$$

The extensivity of the free energy $F(\beta)$ implies that f is finite for $N \to \infty$. It is easy to show that f is a positive function and increases as $|\beta_0 - \beta_1|$ increases. Indeed, using $E = \partial(\beta F)/\partial\beta$ and $C_V = \partial E/\partial T$ at constant volume (in our language at constant N), we can rewrite

$$Nf(\beta_0, \beta_1) = \int_{\beta_0}^{\beta_1} [E(\beta') - E(\beta' + \beta_1 - \beta_0)] d\beta' =$$
$$= \int_{\beta_0}^{\beta_1} \left[\frac{\beta' - \beta_0}{\beta'^2} C_V(\beta') + \frac{\beta_1 - \beta'}{(\beta' + \beta_1 - \beta_0)^2} C_V(\beta' + \beta_1 - \beta_0) \right] d\beta' .$$

The positivity of the specific heat immediately implies that $f(\beta_0, \beta_1) > 0$. For $|\beta_1 - \beta_0| \ll 1$ we can expand $f(\beta_0, \beta_1)$ in powers of $\beta_1 - \beta_0$, obtaining

$$f(\beta_0, \beta_1) = \frac{c_V}{\beta_0^2} (\beta_1 - \beta_0)^2 , \tag{6.5}$$

where $c_V = C_V/N$ is the specific heat per system variable at $\beta = \beta_0$. Collecting all terms we obtain for the variance of the estimate

$$\frac{\sigma_{\text{est}}^2}{\langle A \rangle_1^2} = \frac{1}{n} \frac{\langle (A - A_1)^2 \rangle_2}{A_1^2} e^{Nf(\beta_0, \beta_1)} (1 + 2\tau_{\mathscr{O}}) + O(n^{-2}) .$$

Since Eq. (6.1) is a ratio, the estimate is also biased. The bias can be easily computed in the case of independent sampling (if correlations are present formulae are more involved, but the conclusions reported below remain unchanged). Using

$$\text{bias}\left(\frac{\frac{1}{n}\sum_i A_i}{\frac{1}{n}\sum_i B_i}\right) = \left\langle \frac{\frac{1}{n}\sum_i A_i}{\frac{1}{n}\sum_i B_i}\right\rangle - \frac{\langle A\rangle}{\langle B\rangle} =$$

$$= \frac{1}{2n}\frac{\langle A\rangle}{\langle B\rangle}\left[\text{var}\mathscr{O} - \frac{\text{var}A}{\langle A\rangle^2} + \frac{\text{var}B}{\langle B\rangle^2}\right] + O(n^{-2}) \,, \qquad (6.6)$$

we easily check that also the bias is proportional to $\exp[Nf(\beta_0, \beta_1)]$.

We can also compute the error $\sigma_{\Delta F}$ on the free-energy difference as computed by using Eq. (6.2). We have

$$\sigma^2_{\Delta F} = \left[\frac{\langle e^{-2\Delta\beta H}\rangle_0}{\langle e^{-\Delta\beta H}\rangle_0^2} - 1\right](1 + 2\tau) \approx \frac{Z_2 Z_0}{Z_1^2}(1 + 2\tau) = e^{Nf(\beta_0,\beta_1)}(1 + 2\tau) \,,$$

where τ is the integrated autocorrelation time associated with $e^{-\Delta\beta H}$. Note that the same exponential factor occurs also here.

The presence of the exponential term sets a bound on the width of the interval in which data reweighting can be performed. Requiring $\sigma_{\text{est}}/\langle A\rangle_1 \ll 1$ we obtain

$$\frac{1}{n}\exp[Nf(\beta_0, \beta_1)] \ll 1 \,,$$

which implies for small values of $\Delta\beta = \beta_0 - \beta_1$ the condition

$$|\Delta\beta| \ll \Delta\beta_{\max} \equiv \beta_0\sqrt{\ln n/(Nc_V)} \,. \qquad (6.7)$$

Notice that this bound depends on the model under study (through the specific heat c_V at β_0) and on the system size, as $N^{-1/2}$, while the dependence on the number of measurements is only logarithmic. The dependence of $\Delta\beta_{\max}$ on $(Nc_V)^{-1/2}$ can be physically explained: energy fluctuations at β_0 are of order $(Nc_V)^{1/2}$ and are thus comparable to the energy difference $E_0 - E_1 \propto Nc_V\Delta\beta_{\max}$, only if $\Delta\beta_{\max}$ scales like $(Nc_V)^{-1/2}$.

The origin of the function f can be better understood by a physical argument which relies on the intuitive idea of the histogram overlaps. Indeed, the probability that a configuration x generated according to $\pi_{\beta_0}(x)$ is in D_{β_1} is given by

$$\sum_{x\in D_{\beta_1}} \pi_{\beta_0}(x) = e^{\beta_0 F(\beta_0)} \sum_{x\in D_{\beta_1}} e^{-\beta_0 H(x)} \simeq$$

$$\simeq e^{\beta_0 F(\beta_0)} e^{S(\beta_1)-\beta_0 E(\beta_1)} \equiv e^{\Delta S - \beta_0 \Delta E} \,, \qquad (6.8)$$

where $S = \beta(E - F)$ is the entropy. To obtain the second equality we have assumed that all configurations in D_{β_1} have the same energy $E(\beta_1)$ and that their number is $e^{S(\beta_1)}$, which is fully justified in the thermodynamic limit. Given that both ΔS and ΔE are extensive, the probability in Eq. (6.8) is exponentially small in N.

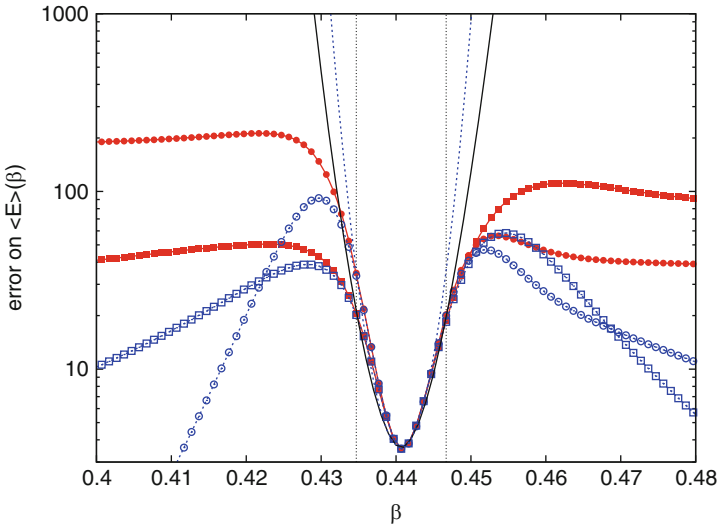

Fig. 6.1 Error on the energy $E(\beta)$ for two different sets of data (we use *squares* and *circles* to distinguish them). We report: (*empty symbols*) the error computed using Eq. (6.3), (*solid symbols*) the error computed using the jackknife method, (*continuous line*) $c\, e^{Nf/2}$ obtained by using Onsager's expression for the free energy, (*dashed line*) $c\, e^{Nf/2}$ using the approximation in Eq. (6.5) and the value of the specific heat at the critical point; c is the error at the critical point. The *vertical dotted lines* give the interval in which we have 100 "good" measures, as defined in the text

The corresponding large deviation (or Cramér) function is given by $\Omega(\beta_0, \beta_1) = \Delta s - \beta_0 \Delta e$, with $s = S/N$ and $e = E/N$. For small $|\beta_1 - \beta_0|$ we have

$$N\Omega(\Delta\beta) = \int_{T_0}^{T_1} dT \left(\frac{C_V}{T} - \frac{C_V}{T_0} \right) \approx -\frac{C_V \Delta\beta^2}{2\beta_0^2} \approx -\frac{1}{2} Nf(\beta_0, \beta_1) \,,$$

where C_V is the specific heat at β_0. The number of "good" measurements for the estimate of $\langle A \rangle_{\beta_1}$ (i.e., those in D_{β_1}) is then $n \exp(-Nf/2)$. The reweighting is reliable if this number is much larger than 1, which again implies condition in Eq. (6.7).

To give an example on how the method works, let us consider the Ising model on a square lattice of size $N = 100^2$ and let us perform a simulation at the critical inverse temperature $\beta_c = \log(\sqrt{2} + 1)/2$. We wish to compute $\langle E \rangle_\beta$ in an interval around the critical point. In Fig. 6.1 we report the statistical error on this quantity obtained by reweighting 10^4 independent measurements. The error computed by using Eq. (6.3) first increases significantly and then decreases exponentially as $|\beta - \beta_c|$ becomes large. This behavior is due to the fact that the reweighted dataset becomes dominated by a single data point and fluctuations within the reweighted dataset disappear. However, this decrease is inconsistent with the exact expression we have derived above—and also with physical intuition—which shows that the

error should always increase as $|\beta - \beta_c|$ increases. The origin of this discrepancy is related to the fact that, as we move out of the critical point, not only does the error on the energy increase, but also the error on the error increases, hence also the error becomes unreliable. It is important to stress that in any case $\sigma_{\rm est}$ cannot be computed by using Eq. (6.3) as soon as the error becomes large. Indeed, that relation is an asymptotic formula valid as long as the neglected corrections (of order n^{-2}) are small. But it is clear that, when the leading term is large, also the corrections become relevant, making the formula unsuitable for the computation of the error. In this case, a more robust method should be used, like the jackknife method [9, 10]. The jackknife error behaves better, but also this method becomes unreliable when the reweighted dataset concentrates on very few data points (those with the largest or smallest energy, depending on the sign of $\Delta\beta$). In practice, the jackknife error converges for large $|\beta - \beta_c|$ to the absolute value of the difference between the two largest (or smallest) energy data points. It is interesting to compare the error determined from the numerical data with the exact result. Hence, in the figure we also report $c\exp(Nf/2)$, where c is the error at the critical point and f has been computed by using Onsager's expression [11, 12] for the free energy. It is clear that the error computed from the MC data becomes immediately unreliable as soon as $|\beta - \beta_c| \gtrsim 0.01$. Indeed, the true error increases quite fast and becomes enormous outside this small interval. For instance, for the extreme case $\beta = 0$ we have $f(\beta_c, 0) = 0.473$ so that $\exp(Nf/2) \sim 10^{1027}$. In Fig. 6.1 we also report the interval in which we have at least $m = 100$ good measures, where m is defined by $n\exp(-Nf/2)$, as discussed above. In this range the jackknife and the asymptotic error estimates agree, as expected. Moreover, in this interval also the quadratic approximation in Eq. (6.5) is quite accurate.

6.3 Multiple Histogram Method

Given that the reweighting method can cover only a limited temperature range of width $\Delta\beta_{\rm max}$ around the temperature β_0 where the original data were collected, one could improve it by running new simulations at β_1, with $|\beta_1 - \beta_0| > \Delta\beta_{\rm max}$, but such that, combining all measured data, the entire range (β_0, β_1) is covered. More generally, suppose one has performed MC simulations at R different inverse temperatures $\{\beta_i\}_{i=1,\ldots,R}$. What is the best way to combine these R datasets to estimate average values $\langle A\rangle_\beta$ at any β?

The most naive method would consist in performing a weighted average of the reweighted data. To explain the shortcomings of this approach, let us assume $R = 2$ and, for instance, let us consider a value of β between β_1 and β_2 which is closer to β_1 than to β_2. A formally correct strategy to compute an average $\langle A\rangle_\beta$ could be the following. We first use the data at β_1 to obtain an estimate A_1 with error σ_1 and then the data at β_2 to obtain an estimate A_2 with error σ_2. Since β is not close to β_2, A_2 has a somewhat large error; but, what is worse, also the error estimate σ_2 has a somewhat large error. Hence, σ_2 as estimated from the data could be largely

underestimated, as we have seen in Sect. 6.2. Finally, one could combine the two estimates as

$$A_{12} = \frac{A_1\sigma_1^{-2} + A_2\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \ .$$

But, if σ_2 is largely underestimated, we would give too much weight to A_2, adding essentially noise and not signal to A_1. In these cases A_{12} would be a worst estimate than A_1.

A much better method has been proposed by Ferrenberg and Swendsen [13].[2] Before presenting the method let us define a few fundamental quantities. For simplicity, let us assume that the system is discrete so that the energy takes discrete values. Then, we introduce the density of states $\rho(E)$ which is defined such that

$$Z_\beta = \sum_E \rho(E) e^{-\beta E} \ ,$$

and the energy histogram variable $h(E_0, \beta)$ defined by

$$h(E_0, \beta) = \langle \delta_{E,E_0} \rangle_\beta = \frac{1}{Z_\beta} \sum_E \rho(E) \delta_{E,E_0} e^{-\beta E} = \frac{1}{Z_\beta} \rho(E_0) e^{-\beta E_0} \ . \tag{6.9}$$

The latter quantity has the important property

$$\mathrm{var}\,[h(E_0, \beta)] = \langle \delta^2_{E,E_0} \rangle_\beta - \langle \delta_{E,E_0} \rangle^2_\beta = h(E_0, \beta)[1 - h(E_0, \beta)] \approx h(E_0, \beta) \ ,$$

where we have used the obvious property $\delta^2_{E,E_0} = \delta_{E,E_0}$ and, in the last step, that $h(E_0, \beta) \ll 1$.

We can now define the method. Suppose we have taken n_i independent measurements[3] at β_i and let us denote with $N_i(E)$ the number of measures with energy E. The ratio $N_i(E)/n_i$ is an estimator of the histogram variable $h(E, \beta_i)$. Using Eq. (6.9) we can estimate $\rho(E)$ using the data at β_i as

$$\rho_i(E) \approx n_i^{-1} N_i(E) e^{\beta_i E} Z_i \ ,$$

where Z_i, the partition function at β_i, has still to be determined. The variance of the estimator of $\rho_i(E)$ can be easily computed if one assumes that Z_i is known exactly. Indeed, with this assumption

[2] It is interesting to observe that, for $R = 2$, the multiple histogram method is equivalent to Bennett's acceptance ratio method [14] which was developed for liquid systems.

[3] In the case the measures are correlated with an autocorrelation time τ_i, then an effective $\tilde{n}_i = n_i/(2\tau_i + 1)$ should be used in all following formulae.

$$\sigma_i^2(E) = n_i^{-1}\mathrm{var}\,[\rho(E)] = n_i^{-1} e^{2\beta_i E} Z_i^2 \mathrm{var}\,[h(E,\beta_i)] = \\ = n_i^{-1} e^{2\beta_i E} Z_i^2 h(E,\beta_i) = n_i^{-1} e^{\beta_i E} Z_i \rho(E)\ .$$

In the usual error analysis one would replace $\rho(E)$ in the r.h.s. with its estimator ρ_i. Since we know that this estimator may be very imprecise—it provides an accurate estimate of $\rho(E)$ only if E is a typical energy at inverse temperature β_i—we do not do it here. This is a crucial point in the method and it is the one that guarantees the robustness of the results. It is also important to stress that σ_i is not the "true" error, since Z_i is also a random variable which has to be determined. However, we will only use σ_i to write down a weighted average of the estimators $\rho_i(E)$. For this purpose, it is not necessary that the weights are correct variances or estimates thereof.[4] A robust estimate of the density of states using all R datasets is given by a weighted average, where each estimate $\rho_i(E)$ enters with a weight proportional to $1/\sigma_i^2(E)$:

$$\rho(E) = \sum_{i=1}^{R} \rho_i(E) \frac{1/\sigma_i^2(E)}{\sum_{j=1}^{R} 1/\sigma_j^2(E)} = \frac{\sum_{i=1}^{R} N_i(E)}{\sum_{j=1}^{R} n_j\, e^{-\beta_j E}\, Z_j^{-1}}\ . \tag{6.10}$$

At this point it is important to stress two important differences between this method and the naive method presented at the beginning. First, observe that for any given E, the only runs that contribute to the determination of $\rho(E)$ are those for which $N_i(E) \neq 0$. This means that we are using the data at β_i only where they are relevant. Moreover, the estimate of the error σ_i is robust, since it follows from an exact identity for the histogram variable.

Equation (6.10) still depends on the unknown partition functions Z_i. They can be determined in a self-consistent way by noticing that

$$Z_k = \sum_E \rho(E) e^{-\beta_k E} = \sum_E \frac{\sum_{i=1}^{R} N_i(E)}{\sum_{j=1}^{R} n_j\, e^{(\beta_k-\beta_j)E}\, Z_j^{-1}}\ . \tag{6.11}$$

which can be rewritten as

$$\sum_E \frac{\sum_{i=1}^{R} N_i(E)}{\sum_{j=1}^{R} n_j\, e^{(\beta_k-\beta_j)E}\, (Z_k/Z_j)} = 1\ .$$

[4] We remind the reader of a few basic facts. If A_i are different estimates of the same quantity, i.e., they all satisfy $\langle A_i\rangle = a$, any weighted average $A_{\rm wt} = \sum w_i A_i$, $\sum_i w_i = 1$, is correct in the sense that $\langle A_{\rm wt}\rangle = a$. Usually, one takes $w_i = k\sigma_i^{-2}$ (k is the normalization factor) because this gives the optimal estimator, that is the one with the least error. Here, however, robustness and not optimality is the main issue.

The consistency condition gives us R equations for the partition function ratios Z_i/Z_j. Since the number of independent ratios is $R-1$, one would expect only $R-1$ independent equations and, indeed, the R equations are linearly dependent:

$$\sum_{k=1}^{R} n_k \sum_E \frac{\sum_{i=1}^{R} N_i(E)}{\sum_{j=1}^{R} n_j\, e^{(\beta_k-\beta_j)E}\,(Z_k/Z_j)} = \sum_E \sum_{i=1}^{R} N_i(E) = \sum_{i=1}^{R} n_i \ .$$

To solve the problem one proceeds iteratively. We define $\hat{Z}_k = Z_k/Z_1$ and rewrite Eq. (6.11) as

$$\hat{Z}_k = \sum_E \frac{\sum_{i=1}^{R} N_i(E)}{\sum_{j=1}^{R} n_j\, e^{(\beta_k-\beta_j)E}\,\hat{Z}_j^{-1}} \ .$$

A first estimate of $\hat{Z}_k$ can be obtained by using the data reweighting method presented before.[5] The first estimate the $\hat{Z}_i$'s is plugged on the r.h.s. and the l.h.s. provides a new estimate, which is used again in the r.h.s. to get a third estimate and so on. Since we are only able to compute the ratios of the partition functions, we do not obtain at the end $\rho(E)$ but rather $\rho(E)/Z_i$ for all values of i. However, this is enough to compute averages of functions of the energy since

$$\langle g(E)\rangle_{\beta_i} = \sum_E g(E) e^{-\beta_i E}[\rho(E)/Z_i] \ .$$

or ratios of partition functions

$$\frac{Z_\beta}{Z_i} = \sum_E e^{-\beta E}[\rho(E)/Z_i] \ .$$

The procedure we have presented can be generalized to allow us to compute averages of generic observables $A(x)$. In this case, the basic quantities are the joint histogram with respect to E and A

$$h(E_0, A_0, \beta) = \langle \delta_{E,E_0}\delta_{A,A_0}\rangle_\beta \ ,$$

its estimator $N_i(E_0, A_0)/n_i$, and the density of states $\rho(E_0, A_0)$ which counts the number of states such that $E = E_0$ and $A = A_0$. Repeating the same steps as before, we end up with

$$\rho(E, A) = \frac{\sum_{i=1}^{R} N_i(E, A)}{\sum_{j=1}^{R} n_j\, e^{-\beta_j E}\, Z_j^{-1}} \ .$$

[5] If the inverse temperatures β_i are ordered, one could determine Z_i/Z_{i-1} by using the reweighting method and then $\hat{Z}_i = (Z_i/Z_{i-1})(Z_{i-1}/Z_{i-2})\ldots Z_2/Z_1$.

Once $\rho(E, A)/Z_i$ is known, any average involving E and A can be directly computed.

It is worth noticing that the use of histograms in the multiple histogram method (which is in general information degrading) is not strictly necessary if one is able to save the full configurations or, at least, the measurements $A_{i,t}$ and $E_{i,t}$ at each MC time t. Indeed, the consistency equations can be rewritten as

$$\hat{Z}_k = \sum_{i,t} \frac{1}{\sum_{j=1}^{R} n_j \, e^{(\beta_k - \beta_j) E_{i,t}} \, \hat{Z}_j^{-1}} ,$$

where $E_{i,t}$ is the t-th energy measurement at β_i, while the average of any quantity at any inverse temperature β can be computed as

$$\langle A \rangle_\beta = \hat{Z}_\beta^{-1} \sum_{i,t} \frac{A_{i,t}}{\sum_j n_j \, e^{(\beta - \beta_j) E_{i,t}} \, \hat{Z}_j^{-1}} ,$$

with

$$\hat{Z}_\beta = \sum_{i,t} \frac{1}{\sum_j n_j \, e^{(\beta - \beta_j) E_{i,t}} \, \hat{Z}_j^{-1}} .$$

Remember that each term entering the sums in the denominators is exponential in N. Much care needs to be taken in doing these sums, since the summations involve terms of very different sizes, and even a single term can exceed the range of floating-point numbers. The suggestion is to work with the logarithms of these terms.

6.4 Umbrella Sampling and Simulated Tempering

6.4.1 Umbrella Sampling

In the previous sections we have shown how to use several runs at $\beta_1 < \ldots < \beta_R$ to compute averages for any β in the interval $[\beta_1, \beta_R]$ and to compute free energy differences. The umbrella sampling (US) method was introduced by Torrie and Valleau [15] to perform the same tasks by means of a single simulation. The idea consists in performing MC simulations with a non-Boltzmann-Gibbs distribution function of the form

$$\pi(x) = \frac{1}{Z_\pi} \sum_{i=1}^{R} \alpha_i e^{-\beta_i H(x)} , \tag{6.12}$$

where i runs over the R different temperatures, Z_π is the normalizing factor, and α_i are positive constants that should be carefully chosen as described below. By sampling the distribution in Eq. (6.12) one aims at sampling in a single run the configurations that are typical for all β_i's and, as a consequence, all configuration space which is relevant for the computation of $\langle A \rangle_\beta$ with $\beta_1 \leq \beta \leq \beta_R$. In order for the method to work two requirements should be satisfied. First of all, the temperatures should be finely spaced, so that typical configurations at inverse temperature β_i overlap with those at $\beta_{i\pm1}$. If this does not occur, the system is unable to move in configuration space and does not visit the typical configuration domain of all β_i's. This condition is the same that occurs in the application of the data reweighting method. Using the results presented in Sect. 6.2 and, in particular, Eq. (6.7), we can conclude that $|\beta_i - \beta_{i+1}|$ should scale as $(c_V N)^{-1/2}$: if the system size is increased, temperatures should be closer. A second important condition fixes the coefficients α_i or, more precisely, their ratios. We require that the typical configuration domains at each β_i have approximately the same probability under π. Using the notations of Sect. 6.2, the probability of the typical domain D_{β_k} is given by

$$\sum_{x \in D_{\beta_k}} \pi(x) = \frac{1}{Z_\pi} \sum_i \alpha_i \sum_{x \in D_{\beta_k}} e^{-\beta_i H(x)} \approx \frac{1}{Z_\pi} \alpha_k Z_k \ .$$

Therefore, we require

$$\frac{1}{Z_\pi} \alpha_i Z_i = \frac{1}{Z_\pi} \alpha_j Z_j \qquad \Rightarrow \qquad \frac{\alpha_i}{\alpha_j} = \frac{Z_j}{Z_i} = e^{\beta_i F(\beta_i) - \beta_j F(\beta_j)} \ . \tag{6.13}$$

Hence the ratios α_i / α_j must be related to the free-energy differences. This is a shortcoming of the method, since these differences are exactly one of the quantities one wishes to compute from the simulation. However, the algorithm is correct, though not optimal, for any choice of the α_i's, so that it is enough to have a very rough estimate of the free-energy differences to run a US simulation. Note that we only fix the ratios of the α_i's: this is not a limitation since one can always set, say, $\alpha_1 = 1$, by redefining Z_π. Once the US simulation has been performed, one can compute averages with respect to the Boltzmann-Gibbs distribution by using

$$\langle A \rangle_\beta = \frac{\langle A \, e^{-\beta H} (\sum_i \alpha_i e^{-\beta_i H})^{-1} \rangle_\pi}{\langle e^{-\beta H} (\sum_i \alpha_i e^{-\beta_i H})^{-1} \rangle_\pi} \ . \tag{6.14}$$

6.4.2 Simulated Tempering

As the US method, also the simulated tempering (ST) method [16, 17] aims at sampling the configurations that are typical at a set of inverse temperatures $\beta_1 < \ldots < \beta_R$ and, indeed, it represents a stochastic version of the US method. In the ST

case, one enlarges the configuration space by adding an index i which runs from 1 to R. Hence, a configuration in the ST simulation is a pair (x, i). Configurations are sampled with probability ($\alpha_i > 0$)

$$\Pi(x, i) = \frac{\alpha_i e^{-\beta_i H}}{Z_i} . \tag{6.15}$$

As in the US method, the temperatures and the coefficients α_i should be carefully chosen, using the same criteria we discussed in the US case. In particular, also the ST method requires an a priori determination of the free energy differences. As we discuss in Sect. 6.4.3, the ST and the US method are essentially equivalent, although the ST has a practical advantage: it is trivial to modify a standard MC code into a ST code (we discuss in Sect. 6.4.3 how to implement ST), while significant more work is needed to implement the US method.

6.4.3 *Equivalence of Simulated Tempering and Umbrella Sampling*

Madras and Piccioni [18] have analyzed the US and ST methods and shown their equivalence under very general conditions, that are usually satisfied in practical applications. We will present here their results trying to avoid all mathematical details. Let us first extend the US method to a general family of probabilities. Consider a state space S and a family of probability functions $\pi_i(x)$, $i = 1, \ldots R$, defined on S. We assume the state space S to be discrete, to avoid mathematical subtleties, but the arguments can be easily extended to the continuous case. In physical terms S is the space of the configurations, while π_i are the Boltzmann-Gibbs distributions $e^{-\beta_i H}/Z_i$. A general *umbrella probability* is given by

$$\pi(x) = \sum_i a_i \pi_i(x) \qquad\qquad \sum_i a_i = 1, \qquad a_i > 0 .$$

The coefficients a_i are related to the coefficients α_i defined before by $a_i = \alpha_i Z_i / Z_\pi$, while the optimality condition in Eq. (6.13), which is not assumed to be satisfied in the following, becomes $a_i/a_j = 1$.

By means of a single MC simulation (i.e., by considering a Markov chain that has π as stationary distribution) one generates a set of points $x_1, \ldots, x_n$ in S. If $A(x)$ is a function defined on S, the sample average converges to the average with respect to π as $n \to \infty$:

$$\frac{1}{n} \sum_{k=1}^{n} A(x_k) \to \sum_x \pi(x) A(x) = \langle A \rangle_\pi .$$

One can also obtain averages with respect to any of the probabilities π_i, by simply reweighting the data. Equation (6.14) can be rewritten as

$$\langle A\rangle_i = \sum_x \pi_i(x)A(x) = \frac{\langle \pi_i A/\pi\rangle_\pi}{\langle \pi_i/\pi\rangle_\pi} \ . \tag{6.16}$$

Let us now formulate the ST method in the same framework. The idea is to enlarge the state space S to

$$S' = S \times \{1 \ldots, R\}$$

(we shall often call the additional index a *label*) and consider the probability $\Pi(x,i) = a_i \pi_i(x)$ on S'.

We first show that the ST and the US methods generate equally distributed points in S. Suppose that we use a general MC algorithm on S' (mathematically, a Markov chain that has Π as stationary distribution) to generate data (x_1, i_1), ..., (x_n, i_n). If $A(x)$ is a function defined on S, the sample mean converges to Π-averages as $n \to \infty$:

$$\frac{1}{n}\sum_{k=1}^{n} A(x_k) \to \sum_{x,i} A(x)\Pi(x,i) = \sum_{x,i} A(x)a_i\pi_i(x) = \sum_x A(x)\pi(x) \ .$$

Roughly speaking, this means that, if we start the MC in equilibrium, $x_1, \ldots x_n$ are distributed according to the umbrella sampling distribution, as if they had been obtained by a MC US simulation.

The fact that the US and the ST methods generate data with the same distribution probability does not imply that dynamics are equivalent in the two methods and one may wonder whether, by enlarging the state space, one can define algorithms that can speed up significantly simulations. After all, there is a well-known example in which this strategy works very nicely: the Swendsen-Wang (or cluster) algorithm [19] for the Ising model is indeed obtained [20] by enlarging the configuration space of the Ising spins $\{s_i\}$ to $\{s_i\} \times \{b_{\langle ij\rangle}\}$, where $b_{\langle ij\rangle}$ are the bond occupation variables. For the case of the US and ST methods, this issue has been investigated in Ref. [18], for the case in which each system is updated by means of the Metropolis algorithm.

Let us first define the specific update considered in Ref. [18] for ST. This is not the most general one, but it corresponds to the update used in practical implementations. If (x,i) is the present configuration, an iteration consists first in updating the label i, followed by an update of the configuration x. Labels are updated using the conditional probability of the labels at fixed x: a new label j is chosen with probability $a_j\pi_j(x)/\sum_k[a_k\pi_k(x)] = a_j\pi_j(x)/\pi(x)$. Then, a new configuration $y \in S$ is chosen by using a MC method appropriate for the system with probability π_j, i.e. the system is updated with a Markov chain $T_j(x,y)$ which is stationary with respect to π_j (we remind the reader that this corresponds to the condition $\sum_x \pi_j(x)T_j(x,y) = \pi_j(y)$, a formula which will be often used in the following). The transition matrix is therefore

$$P(x,i;y,j) = \frac{a_j \pi_j(x)}{\pi(x)} T_j(x,y) \ . \tag{6.17}$$

Note that one often uses the Metropolis algorithm to update the labels, by proposing, for instance, $i \to i \pm 1$. This choice is certainly (slightly) more efficient, but should not change the general conclusions: the label dynamics should not be the relevant part of the algorithm. The Markov process in Eq. (6.17) induces a Markov process on S whose transition matrix is obtained by summing $P(x,i;y,j)$ over j:

$$P_s(x,y) = \sum_j P(x,i;y,j) = \frac{1}{\pi(x)} \sum_j a_j \pi_j(x) T_j(x,y) \ . \tag{6.18}$$

Such a process has $\pi(x)$ as equilibrium distribution, since

$$\sum_x \pi(x) P_s(x,y) = \sum_j a_j \sum_x \pi_j(x) T_j(x,y) = \sum_j a_j \pi_j(y) = \pi(y) \ .$$

We will finally show that under very general conditions, if the T_j are Metropolis updates, also P_s is a Metropolis update:

(a) Assume that the probabilities $\pi_i(x)$ satisfy the following condition: for any pair $x, y \in S$ we have either $\pi_i(x)/\pi_i(y) < 1$ for all i's or $\pi_i(x)/\pi_i(y) \geq 1$ for all i's. This is obviously satisfied for Boltzmann-Gibbs distributions. Given x and y one computes the energies $E(x)$ and $E(y)$. If $E(x) > E(y)$ then $e^{-\beta_i E(x)}/e^{-\beta_i E(y)} < 1$ for all $\beta_i > 0$. If the energies satisfy the opposite inequality, also the ratio of the Boltzmann-Gibbs factors satisfies the opposite inequality for all $\beta_i > 0$.

(b) The Metropolis update consists in two steps: a proposal in which a new configuration y is proposed, and an acceptance step. We assume that the proposal does not depend on the label i. For the Boltzmann-Gibbs distribution, this means that, given configuration x, we propose a new configuration y with a method which does not depend on temperature. Moreover—most practical algorithms satisfy this condition—we require the proposal matrix to be symmetric: the probability of proposing y given x is the same as that of proposing x given y.

For the Metropolis update, if $K(x,y)$ is the proposal matrix, we have [21]

$$T_i(x,y) = K(x,y) \min\left(1, \frac{\pi_i(y)}{\pi_i(x)}\right) \qquad x \neq y \ ,$$

$$T_i(x,x) = 1 - \sum_{y \neq x} T_i(x,y) \ .$$

Inserting this expression in Eq. (6.18), we obtain for $x \neq y$

$$P_s(x,y) = \frac{1}{\pi(x)} \sum_j a_j \pi_j(x) K(x,y) \min\left(1, \frac{\pi_j(y)}{\pi_j(x)}\right).$$

Now, assume that $\pi_i(y)/\pi_i(x) > 1$ for all i (we use here assumption (a)). In this case we have also $\pi(y)/\pi(x) > 1$ and

$$P_s(x,y) = \frac{1}{\pi(x)} \sum_j a_j \pi_j(x) K(x,y) = K(x,y) \ .$$

In the opposite case we have instead

$$P_s(x,y) = \frac{1}{\pi(x)} \sum_j a_j \pi_j(x) K(x,y) \frac{\pi_j(y)}{\pi_j(x)} = K(x,y) \frac{\pi(y)}{\pi(x)} \ .$$

Hence

$$P_s(x,y) = K(x,y) \min\left(1, \frac{\pi(y)}{\pi(x)}\right) \ .$$

But this is the transition matrix of a Metropolis update with respect to the probability $\pi(x)$. Hence, for the Metropolis case there is a complete equivalence between the US and the ST methods. Madras and Piccioni [18] have also considered the case in which condition (a) is not satisfied, proving that in this case ST is no better than the US method (they prove that the probability of null transitions in the US method is equal or smaller than that in the ST).

Finally, let us compare how averages are computed in the US and in ST methods. To compute averages with respect to π_i in the US method one uses formula (6.16). This formula also holds for the ST:

$$\langle A \rangle_i = \frac{\langle \pi_i A/\pi \rangle}{\langle \pi_i/\pi \rangle} \ , \tag{6.19}$$

where averages $\langle \cdot \rangle$ without any subscript refer to the ST measure $\Pi(x,i)$. However, in ST simulations, one usually considers

$$\langle A \rangle_i = \frac{\langle A\, I_i \rangle}{\langle I_i \rangle} \ , \tag{6.20}$$

where $I_i(x,j) = \delta_{ij}$ for every point $(x,j) \in S'$. That is, in Eq. (6.20) only data at β_i are used for estimating $\langle A \rangle_i$. The two expressions (6.19) and (6.20) are clearly different, but not that unrelated. Indeed, one could also determine $\langle A \rangle_i$ by reweighting the data measured at β_k:

$$\langle A \rangle_i = \frac{\langle A \pi_i I_k/\pi_k \rangle}{\langle \pi_i I_k/\pi_k \rangle} = A_{i,k} \ .$$

The average in Eq. (6.20) corresponds to $A_{i,i}$. Let us now show that the estimator in Eq. (6.19) is roughly a weighted average of the $A_{i,k}$. Let us define

$$\frac{1}{b_i} = \left\langle \frac{\pi_i}{\pi} \right\rangle .$$

Then, note that in systems of physical interest the supports of the distributions π_i are mostly disjoint: if x is such that $\pi_i(x)$ is significantly larger than zero, then $\pi_k(x)$ is very small for all $k \neq i$. In physical terms it means that, if we have a configuration that is typical at temperature β_i, such a configuration will be not be typical at all other temperatures. If this holds, then we can approximate

$$\frac{1}{\pi(x)} \approx \sum_k \frac{I_k}{a_k \pi_k(x)} , \tag{6.21}$$

so that Eq. (6.19) can be rewritten as

$$\langle A \rangle_i \approx \sum_k \frac{b_i}{a_k} \left\langle A \frac{I_k \pi_i}{\pi_k} \right\rangle .$$

Hence, the estimator in Eq. (6.19) is essentially equivalent to the following weighted average of the $A_{i,k}$:

$$\langle A \rangle_i \approx \sum_k \frac{b_i}{a_k} \left\langle \frac{\pi_i I_k}{\pi_k} \right\rangle A_{i,k} . \tag{6.22}$$

Using Eq. (6.19) to estimate $\langle A \rangle_i$, one is taking into account not only the data with label i, but all data by means of a proper reweighting as shown by Eq. (6.22). Of course, Eq. (6.22) is not quantitatively correct, since in practical implementations there must be configurations that are typical for two distributions (otherwise, the algorithm would not work): for them the approximation made in Eq. (6.21) fails. However, the argument gives a direct physical interpretation of Eq. (6.19) as some kind of, though not exact, reweighting of the data. Note that, when reweighting is used, there is always the technical problem of determining the weights of the average (see Sect. 6.3). No such problem arises here: everything is fixed in Eq. (6.19).

6.5 Generalizing the Umbrella Method: Multicanonical Sampling

Umbrella sampling, like simulated or parallel tempering, provides a way to sample in the same run different probability distributions along a *connected configuration path* $\mathscr{P}$, i.e., a connected subset of the configuration space S such that, if $x \in \mathscr{P}$,

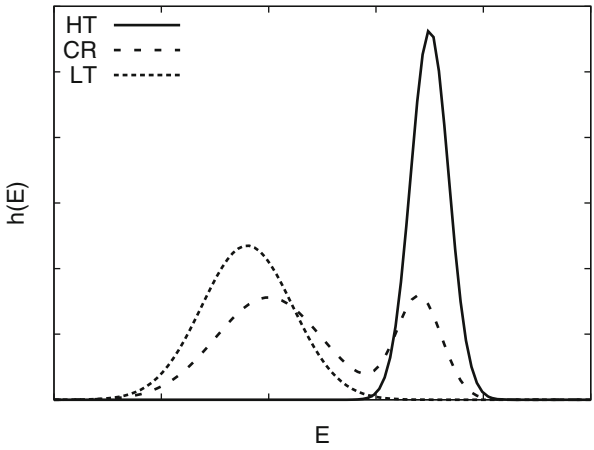

Fig. 6.2 A sketch of the energy distributions in the Potts model for $q > 4$ (first-order transition): the high-temperature (*HT*) and low-temperature (*LT*) distributions are unimodal, while at the critical temperature (*CR*) the distribution is bimodal

x is a typical configuration of at least one of the probabilities $\pi_i(x)$: in our previous notations $\mathscr{P}$ should be connected and contained in $\cup_{i=1}^{R} D_{\beta_i}$. In the presence of first-order phase transitions this path may not exist, hence the above methods cannot be applied. As an example let us consider the q-state Potts model on a square lattice. The model is defined in terms of spins (sometimes called colors) s_i defined at the sites of the lattice. Each s_i can assume q integer values between 1 and q. The Hamiltonian is given by

$$H_q(\{\sigma\}) = -\sum_{\langle ij \rangle} \delta_{s_i, s_j} \ ,$$

where the sum is over all nearest-neighbor lattice pairs $\langle ij \rangle$, $\delta_{s,s} = 1$ and $\delta_{s,t} = 0$ if $s \neq t$. As probability distribution we consider the usual Boltzmann-Gibbs distribution

$$\pi \propto e^{-\beta H_q} \ .$$

This model shows two different phases depending on β. For $\beta = \beta_c$ a phase transition occurs. For $q > 4$ such a transition is of first order and the energy has a bimodal distribution at $\beta = \beta_c$. A sketch of the energy distributions close to the transition point is reported in Fig. 6.2. Typical high-temperature (HT) distributions are unimodal and overlap with only one of the peaks appearing at the critical point, that with the highest energy. Analogously, low-temperature (LT) distributions are also unimodal; they only overlap with the low-energy peak of the critical-point distribution. This particular behavior of the energy distributions implies that any US or ST (these considerations also apply to the parallel tempering method which will be discussed in the next section) algorithm with local updates of the spins cannot move rapidly between LT and HT typical configurations. For the mean-field case (Potts model on a complete graph) and the Metropolis algorithm, this is indeed a rigorous theorem [22]: the exponential autocorrelation time of Metropolis ST algorithms increases exponentially with the size of the system. The origin of

the phenomenon is easily understood qualitatively. Suppose we use any of the previously mentioned methods and consider a set of temperatures such that β_1 is in the HT phase and β_R is in the LT phase. Start the simulation in the HT phase. Provided that the temperatures are finely spaced, one would eventually reach the critical point. Since the configuration has been obtained by cooling a HT configuration, it has an energy that belongs to the HT peak. Because of the bimodal nature of the energy distribution at β_c, local updates at β_c would only generate new configurations with energy belonging to the HT peak. Hence, any attempt to further reduce the temperature would fail, since the configuration would never be a typical LT configuration. Hence, LT configurations would never be visited. This argument is quite general and shows that US and ST, when used in combination with local algorithms, only work when the configuration path does not go through first-order transitions. A second-order transition should not be a limitation, since at the transition distributions are broader but usually still unimodal.[6]

To solve the problem one might consider an enlarged parameter space that allows one to go from the LT phase to the HT phase without intersecting the first-order transition point. In the Potts model this could be obtained by adding, for instance, a magnetic field, but this should in any case be done carefully, to be sure that all low-temperature degenerate states are equally visited. In practice, these extensions are usually not efficient.

We now discuss a family of methods that generalize the umbrella sampling method and are appropriate for the study of first-order transitions. They also work with a nonphysical distribution function $\pi(x)$ which is constructed in such a way to allow good sampling of both phases. Sometimes that are called *multicanonical* algorithms following Berg and Neuhaus [25, 26] that applied these methods to the study of first-order transitions.

Let us consider again the Potts model and suppose that one is at the transition point β_c, or at least very close to it. For $q > 4$ (the case we are considering now) the distribution of the energy $h(E)$ is bimodal, with two maxima at $E_1 < E_2$. If $h_i = h(E_i)$ is the value of the distribution at the maximum i, one defines the multicanonical distribution $\pi(x)$ as follows

$$\begin{aligned} \pi(x) &= \frac{e^{-\beta H}}{Z h_1} && E(x) \le E_1 \,, \\ &= \frac{e^{-\beta H}}{Z h(E)} && E_1 < E(x) < E_2 \,, \\ &= \frac{e^{-\beta H}}{Z h_2} && E_2 \le E(x) \,, \end{aligned}$$

[6] There are instances of second-order transitions which show bimodal distributions in finite volume [23, 24]: however, in these cases the two peaks get closer and the gap decreases as the volume increases. ST should work efficiently in these instances. Note, however, that the algorithm may not work in some disordered systems, even if the transition is of second order. One example is the random field Ising model.

where Z is the partition function for π. If we now compute the distribution of the energy in the new ensemble, we find

$$\begin{aligned} h_\pi(E) &= \alpha\, h(E)/h_1 && E(x) \le E_1\ , \\ &= \alpha && E_1 < E(x) < E_2\ , \\ &= \alpha\, h(E)/h_2 && E_2 \le E(x)\ , \end{aligned}$$

where α is a normalization constant. The probability is now flat for $E_1 < E(x) < E_2$ and thus local algorithms should have no problem in going from one phase to the other.

The main problem of the method stays in the fact the β_c as well as $h(E)$ are not known beforehand. In practical implementations one may work as follows. First, one roughly determines the position of the transition point. This can be obtained by running a hysteresis cycle. One thermalizes a configuration at a value of β which is deep in the HT phase and measures its energy. Then, one slightly increases β, thermalizes the configuration at this new temperature and recomputes the energy. One keeps repeating these steps until the configuration "jumps" in the LT phase: this is signalled by a big decrease of the energy. Let us call $\beta_{\max}$ this value of β. Then, one begins a series of runs in which β is decreased until the configuration (for $\beta = \beta_{\min}$) jumps back in the HT phase. The cycle allows one to infer that β_c lies in the interval $[\beta_{\min}, \beta_{\max}]$. In the absence of any other information we can just take the midpoint as the value of β at which the multicanonical simulation is performed. Note that it is not needed that such value be an accurate estimate of β_c. It is only crucial that at this value of β the distribution is bimodal, i.e. that there is a significant overlap with both phases.

Once the value of β at which the simulation should be performed has been chosen, one must determine $\pi(x)$. This can be done recursively. We will illustrate the procedure with an example, considering the liquid-gas transition in a fluid. Here the number N of molecules present in the system plays the role of order parameter in the transition (it is the analogue of E in the Potts model), while the grand canonical distribution $\pi_0 = e^{-\beta H + \beta\mu N}/(N!Z)$ plays the role of the Boltzmann-Gibbs distribution. The gas and liquid phases are the analog of the HT and LT Potts phases. The iterative procedure starts by performing two runs: one run starts from a gas configuration, while the other run starts from a liquid (dense) configuration. For each of the two runs (discarding the equilibration transient) we measure the histograms $h_{0G}(N)$ and $h_{0L}(N)$ of N, see Fig. 6.3 (top, left). We observe two clearly separated peaks centered around $N \approx 30$ and $N \approx 200$. Then, we choose an interval $I = [N_{\min}, N_{\max}]$ that contains the two peaks. In the present case, we choose $N_{\min} = 0$ and $N_{\max} = 220$. Then, we modify the updating step so that N always belongs to the interval I. This is a crucial modification to have a stable recursion; of course, this restriction should be eliminated at the end, once $\pi(x)$ has been determined.

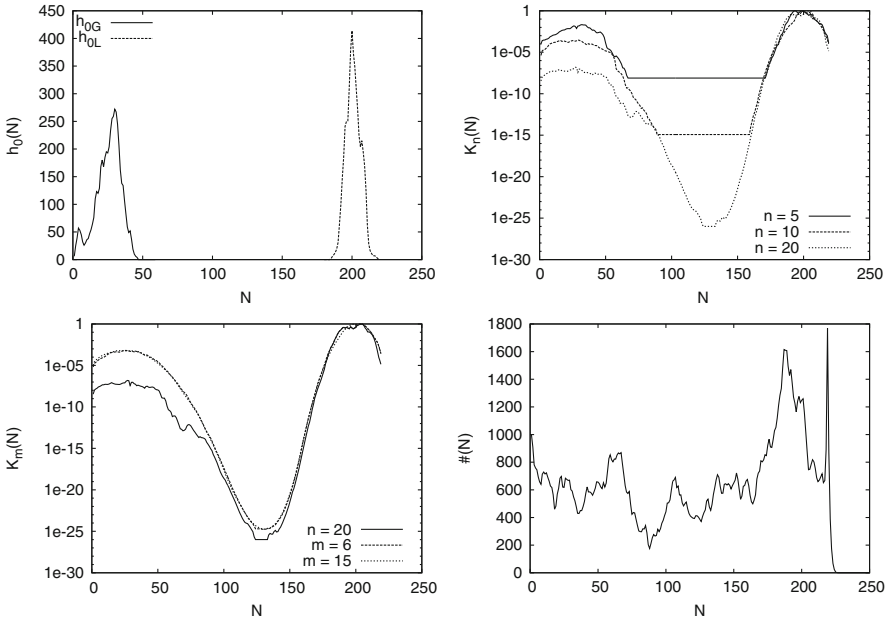

Fig. 6.3 *Top*, *left*: we report the distributions $h_{0G}(N)$ and $h_{0L}(N)$ of N at the beginning of the iterative procedure. *Top*, *right*: function $K_n(N)$ after several iterations ($n = 5, 10, 20$) during the first part of the procedure. *Bottom*, *left*: function $K_n(N)$ at the end of the first part of the procedure ($n = 20$) and after other m iterations following Eq. (6.23). *Bottom*, *right*: histogram of N obtained in a MC simulation using the final umbrella distribution $\pi(x)$

The recursion method determines at each step a function $K_n(N)$ and uses $\pi_{n+1} \propto \pi_0/K_n(N)$ as the distribution function for the next MC simulation. The function $K_n(N)$ should be such that the new distribution π_{n+1} is as flat as possible in the range $[N_{\min}, N_{\max}]$. Let us first determine the zeroth-order approximation $K_0(N)$. If $M_{0G} = \max h_{0G}(N)$ and $M_{0L} = \max h_{0L}(N)$, we define $H_{0G}(N) = h_{0G}(N)/M_{0G}$ and $H_{0L}(N) = h_{0L}(N)/M_{0L}$. Then, we set

$$
\begin{aligned}
K_0(N) &= \varepsilon && \text{if } H_{0G}(N) \le \varepsilon \text{ and } H_{0L}(N) \le \varepsilon\,, \\
K_0(N) &= H_{0G}(N) && \text{if } H_{0G}(N) > \varepsilon \text{ and } H_{0L}(N) \le \varepsilon\,, \\
K_0(N) &= H_{0L}(N) && \text{if } H_{0G}(N) \le \varepsilon \text{ and } H_{0L}(N) > \varepsilon\,.
\end{aligned}
$$

We have introduced a lower cutoff ε on the histograms to discard noisy data (the longer the runs, the smaller ε can be). In the example we use $\varepsilon = 1/200$, hence we use all data except those for which $h_{0G}(N) \le \varepsilon\, M_{OG} \simeq 1$ and $h_{0L}(N) \le \varepsilon\, M_{OL} \simeq 2$. Once $K_0(N)$ is defined, we perform two runs using $\pi_1 \propto \pi_0/K_0(N)$ and again determine the distributions $h_{1G}(N)$ and $h_{1L}(N)$. The successive approximations $K_n(N)$ are obtained as

$$
\begin{aligned}
K_n(N) &= K_{n-1}(N)\,\varepsilon && \text{if } H_{nG}(N) \le \varepsilon \text{ and } H_{nL}(N) \le \varepsilon\,, \\
K_n(N) &= K_{n-1}(N)\,H_{nG}(N) && \text{if } H_{nG}(N) > \varepsilon \text{ and } H_{nL}(N) \le \varepsilon\,, \\
K_n(N) &= K_{n-1}(N)\,H_{nL}(N) && \text{if } H_{nG}(N) \le \varepsilon \text{ and } H_{nL}(N) > \varepsilon\,,
\end{aligned}
$$

where $H_n(N) = h_n(N)/M_n$ and M_n is the maximum of $h_n(N)$. The procedure is repeated several times until the distributions $h_{nG}(N)$ and $h_{nL}(N)$ overlap, i.e., there is an $\bar{N}$ such that $H_{nG}(\bar{N}) > \varepsilon$ and $H_{nL}(\bar{N}) > \varepsilon$. In the example 20 iterations are needed. In Fig. 6.3 (top, right) we show $K_n(N)$ for $n = 5, 10$, and 20. To allow a better comparison, we have multiplied the functions by a constant (irrelevant in the definition of π_n) so that the maximum of $K_n(N)$ is always 1. Note how the double-peak structure emerges as the number of iterations is increased, in spite of the fact that there are 26 orders of magnitude between maximum and minimum. From a practical point of view the procedure can be improved and speeded up in several ways. First, one can smooth the histograms to eliminate noise. Second, after a few iterations one can try to guess $K(N)$: one can fit the peaks with Gaussians and restart the iterations from the fitted function. Third, one can perform a different number of iterations in the two phases if the efficiency of the algorithm is phase dependent. Finally, note that thermalization is needed only in the first run. Then, one can restart the simulation from the last configurations generated in the previous iteration.

Once the gas and liquid distributions overlap, there is no longer need of two different simulations. One performs a single run m times, determines the distribution $h_m(N)$, its maximum M_m, defines $H_m(N) = h_m(N)/M_m$, and updates K_m as follows:

$$
\begin{aligned}
K_m(N) &= K_{m-1}(N)\,\varepsilon && \text{if } H_m(N) \le \varepsilon\,, \\
K_m(N) &= K_{m-1}(N)\,H_m(N) && \text{if } H_m(N) > \varepsilon\,.
\end{aligned}
\tag{6.23}
$$

In this second part of the procedure it is usually a good idea to increase both ε and the number of iterations, to increase the precision on $h_m(N)$. The obtained $K_m(N)$ for the specific example are reported in Fig. 6.3 (bottom, left). After $m = 6$ iterations following Eq. (6.23), the function $K_m(N)$ reaches its asymptotic form. Note that this iterative procedure is quite stable: if we increase the number of iterations, $K_m(N)$ does not change (see the curve for $m = 15$ in the figure). Once $K(N)$ has been determined, we can eliminate the restrictions on N, setting $K(N) = K(N_{\max})$ for $N > N_{\max}$ and $K(N) = K(N_{\min})$ for $N < N_{\min}$. In Fig. 6.3 we report the histogram of N obtained by using the final $\pi \propto \pi_0/K_m(N)$. All values of N are visited and in particular we are sampling in both phases. We can thus use the final $\pi(x)$ to analyze in detail the behavior at coexistence.

It is important to stress that this procedure correctly works for first-order transitions with two single minima and for which the relevant order parameter is known, but cannot be applied to study the LT phase of disordered systems, like spin glasses.

6.6 Parallel Tempering

6.6.1 General Considerations

In the previous section we have discussed multicanonical sampling, which is appropriate for the study of first-order transitions. In that case, sampling correctly all free-energy minima requires the system to visit also the barrier region, where the probability distribution is extremely small, of the order of e^{-aN^p}, where N is the system size. In the presence of second-order phase transitions, the behavior is quite different, since the different free-energy minima characterizing the ordered phase merge at the critical point, giving rise to a single thermodynamic state. Hence, if one wishes to visit all ordered states, there is no need to go over the barriers. For instance, consider a thermal second-order transition, as it occurs in the Ising model. To sample the LT magnetized phases, one can adopt an algorithm in which temperature is varied. Starting from a LT configuration, one can rise the temperature till that of the critical point, where all minima merge, then move into the HT phase, where a single thermodynamic state exists. If the system spends enough time in the HT phase, it loses memory of the thermodynamic LT phase it was coming from. Hence, when temperature is decreased again, it may well fall into a different LT thermodynamic state. This simple argument should convince the reader that algorithms that allow temperature changes are powerful tools for the study of the ordered phases in the presence of second-order phase transitions. ST was indeed devised with this motivation in mind [17]. However, as we discussed in Sect. 6.4.2, ST has a serious shortcoming: a ST simulation requires some free-energy differences to be determined before starting the simulation; moreover, the efficiency of the simulation depends on the accuracy with which these quantities are determined. These problems can be avoided by using a variant of ST, the parallel tempering (PT) method, which is, at present, the most efficient general-purpose algorithm for studying models undergoing second-order phase transitions. The PT method works well even in very complex models, like spin glasses, that have a very large number of LT local minima. It is also very useful in systems which, even in the absence of phase transitions, cannot be simulated efficiently due to the presence of geometric constraints, like complex molecules in dense systems, or in the presence of boundaries, or in porous systems, just to name a few examples. In computer science and statistics, PT is often used in connection with multimodal distributions.

PT has a quite interesting history. It was first introduced in the computer-science/statistics community by Geyer in 1991 [27], as an efficient method to sample multimodal probability distributions and it was named Metropolis-coupled Markov chain Monte Carlo. The work of Geyer stirred a lot of interest in the statistical physics community working on polymer physics and PT was carefully analyzed and compared with US by Tesi et al. [28]. Independently, in 1996 the PT algorithm was introduced by Hukushima and Nemoto [29] with the name of replica-exchange algorithm, and found widespread application in spin-glass simulations. At the same time, thanks to the works of Hansmann [30] the algorithm found its

way in the chemical physics and biophysics community, as a more efficient and simpler alternative to US and multicanonical algorithms (for a list of applications in this field, see the review by Earl and Deem [31]). At present the name "parallel tempering" is apparently the most widely used name in the physics community, while mathematicians prefer to indicate it as "swapping algorithm".

The PT algorithm is a simple generalization of ST. The state space S' is formed by R replicas of the original state space S: $S' = S \times \ldots \times S$. On S' one takes as probability

$$\Pi(x_1, \ldots, x_R) = \pi_1(x_1)\pi_2(x_2) \ldots \pi_R(x_R) \ .$$

In the standard case, $\pi_1(x), \ldots, \pi_R(x_R)$ are the Boltzmann-Gibbs distributions at R different values of the inverse temperatures $\beta_1 < \ldots < \beta_R$. The algorithm usually works as follows:

(a) If $(x_1, \ldots, x_R)$ is the present configuration, one updates each x_i using any MC algorithm that leaves $\pi_i(x)$ invariant.
(b) One proposes a swapping move[7]:

$$(x_1, \ldots, x_i, x_{i+1}, \ldots x_R) \to (x_1, \ldots, x_{i+1}, x_i, \ldots x_R) \ ,$$

which is accepted with probability

$$p_{\text{swap}} = \min\left(1, \frac{\pi_{i+1}(x_i)\pi_i(x_{i+1})}{\pi_{i+1}(x_{i+1})\pi_i(x_i)}\right) = \min\left(1, e^{(\beta_{i+1}-\beta_i)(E_{i+1}-E_i)}\right) \ .$$

It is immediate to verify that the algorithm satisfies the stationarity condition with respect to Π, though it may not necessarily satisfy detailed balance (this depends on how i and $i+1$ are chosen).

As in the US or ST case, in order to perform a PT simulation, one must decide the number R of inverse temperatures and their values. We note that one of the two conditions we discussed in the case of US and ST should hold also here: temperatures should be close enough, so that the typical configuration domains at nearby temperatures overlap. If this does not occur, no swap is accepted. For an efficient simulation it is important to discuss how close temperatures should be. This will be discussed in Sect. 6.6.3.

Whenever a PT run is performed, it is important to make checks to verify that the algorithm is working correctly. The simplest quantity to measure is the swapping rate $a_{i,i+1}$ between adjacent temperatures, that is the fraction of accepted swaps. The algorithm works efficiently only if, for all i, $a_{i,i+1}$ is not too small. As we discuss in Sect. 6.6.3, the optimal value for $a_{i,i+1}$ lies between 0.2 and 0.3, but larger, or

[7] In principle the swapping can be attempted among any pair of replicas, but only for nearby replicas the swap has a reasonable probability of being accepted.

slightly smaller values, although not optimal, are still acceptable. A reasonable swapping rate is, however, not enough to guarantee that the algorithm is working correctly. Indeed, there are situations in which the swapping rates take the desired values, but the PT simulation is inefficient. This typically occurs when there is a "bottleneck" at a certain temperature β_K (usually it is the closest to the critical temperature). In this case $a_{K-1,K}$ and $a_{K,K+1}$ are both reasonable, but the algorithm is unable to move a LT configuration to the other, HT side. In this case, HT replicas mix very slowly with the LT replicas, so that the dynamics, which is based on the idea that LT replicas rapidly move into the HT phase, becomes very slow. To identify bottlenecks, it is not enough to compute the swapping acceptances. One should measure quantities that take into account how temperature changes for each individual replica. Often one considers the average round-trip time, i.e., the time for a replica to start from the lowest temperature, reach the highest one, and finally go back to the lowest one. If the swapping procedure is working efficiently, the round-trip time should be comparable to the return time of a random walker moving among temperatures with the swapping rates actually measured in the simulation. On the contrary, if the swapping procedure has a bottleneck, then the round-trip time becomes large and is essentially controlled by the time it takes for a replica to go through the bottleneck.[8]

As in all MC simulations, also in PT simulations one should thermalize the system before measuring. Two checks should be performed: first, one should check that equilibrium has been attained at all temperatures. Note that it is not enough to check convergence at the lowest temperature. For instance, in PT simulations of non-disordered systems that go through a second-order phase transition, the slowest mode is controlled by the behavior at the critical point, not at the lowest-temperature point (see the discussion in Sect. 6.6.2). Second, the thermalization time should be larger than the time needed to go through any bottleneck present in the model: typically a few round-trip times suffice.

6.6.2 Some General Rigorous Results

The PT algorithm has been studied in detail by mathematicians which have proved theorems [32–34] confirming the general arguments given at the beginning of Sect. 6.6.1. These theorems give bounds on the spectral gap λ of the Markov chain associated with the PT algorithm. In physical terms λ is related to the exponential autocorrelation time $\tau_{\rm exp} = -1/\ln(1-\lambda)$, which gives the number of iterations needed to generate an independent configuration. An efficient algorithm requires λ to be significantly different from zero.

[8] If the PT method is applied to a system undergoing a first-order transition, the swapping procedure would be highly inefficient, because HT replicas would hardly swap with LT replicas. The two sets of replicas would remain practically non-interacting.

To establish the notation, let $P_k(x, y)$ be a Markov chain defined on the state space S which leaves invariant π_k: $\sum_x \pi_k(x) P_k(x, y) = \pi_k(y)$. The basic idea used in the theorems is the state-space decomposition of Madras and Randall [35]. If (a) the state space is decomposable as $S = \cup_l A_l$ such that all π_k are unimodal in each A_l,[9] (b) swaps occur with sufficient frequency along a configuration path that connects all sets A_l, and (c) P_1 is a fast update on S, then the size of the spectral gap is essentially controlled by the spectral gap of the restrictions P_{kl} of P_k on A_l. In other words, PT is, at most, as fast as the slowest of the P_{kl} [34]. To clarify this result, let us consider the Ising model and a PT simulation with β_1 in the HT phase and β_R in the LT phase. Suppose we use the Metropolis algorithm to update the configurations at each temperature. In the LT phase the Metropolis algorithm is of course inefficient (it cannot go through the barriers). However, if we partition $S = M_+ \cup M_-$, where M_+ and M_- are the positive and negative magnetization configurations, respectively, the restrictions P_{k+} and P_{k-} of P_k to M_+ and M_- are efficient algorithms that sample correctly each free-energy minimum. With this decomposition, the slowest dynamics occurs at the critical point, which represents the bottleneck of the simulation. Hence, the theorem essentially states that the autocorrelation time of the PT simulation is of the order of the autocorrelation time of the algorithm at the critical point, which is also the typical time it takes for a HT configurations to become a LT one and viceversa. Note that the improvement is enormous. We are able to sample the LT phase with autocorrelations that increase polynomially as N^z when the system size N goes to infinity ($z \approx 2$ for the Ising model with Metropolis update) and not exponentially in $N^{1-1/d}$, where d is the space dimension (for the two-dimensional Ising model one can prove $\tau \approx e^{aN^{1/2}}$ for a standard MC simulation [36]).

6.6.3 Optimal Choice of Temperatures

Let us now discuss how to choose optimal temperatures in a PT simulation. First of all, the set of temperatures must extend enough in the HT phase in order to allow replicas at the highest temperatures to decorrelate fast. More precisely, we would like the autocorrelation time τ at the highest temperature to be smaller than the typical time a replica spends in the HT regime, so that, when a replica goes back to the LT phase, it has completely forgotten the previous free-energy minimum. This condition fixes the highest temperature, while the lowest temperature is usually determined by the problem we wish to study (e.g., critical properties of the model or the nature of its LT phase).

Once $\beta_{\min}$ and $\beta_{\max}$ have been fixed, what is the best sequence for the remaining temperatures? Under the hypothesis that there are no bottlenecks and thus the

[9] The condition of unimodality is not required in the proofs of the theorems. However, the theorems have physically interesting consequences only if a unimodal decomposition is possible.

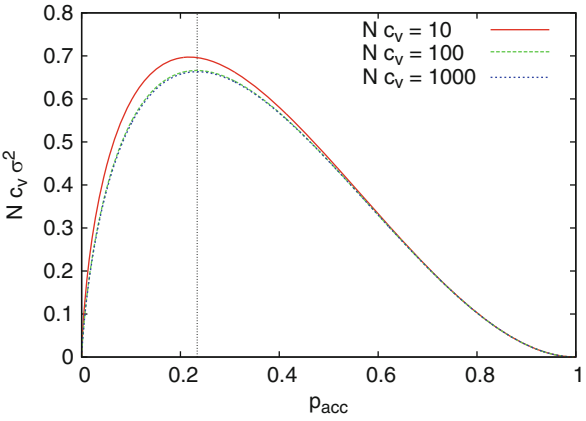

Fig. 6.4 Graph of $Nc_V\sigma^2$ with $\sigma^2 = (\ln r)^2\, p_{\rm acc}$ versus $p_{\rm acc}$ for three different values of Nc_V. The results corresponding to $Nc_V = 100$ and 1,000 cannot be distinguished as they are one on top of the other. The *vertical dotted line* corresponds to $p_{\rm acc} = 0.23$

round-trip time is mainly determined by the swapping rate, the optimal solution is to keep swapping rates constant in the whole temperature range, so that the diffusion of the replicas in temperature space is maximal. The optimal value of the swapping rate depends on the system under study, but in general one has to avoid too small values (replicas almost do not swap) and also too large ones (in this case a smaller number of temperatures would be enough). With the random-walk picture in mind, in order to obtain the largest diffusion rate in temperature space (more precisely in the variable $\ln\beta$), one would like to maximize $\sigma^2 = \ln^2(\beta_i/\beta_{i+1})\, p_{\rm acc}(\beta_i, \beta_{i+1})$, where $p_{\rm acc}(\beta_i, \beta_{i+1})$ is the average acceptance rate for the swap between β_i and β_{i+1}. If the specific heat is constant, the average acceptance rate is well represented for $N \to \infty$ by the formula [37–39]

$$p_{\rm acc}(\beta_i, \beta_{i+1}) = \mathrm{erfc}\left[\frac{1-r}{1+r}(Nc_V)^{1/2}\right],$$

where $r = \beta_i/\beta_{i+1} < 1$ (we assume $\beta_{i+1} > \beta_i$). If we require a constant acceptance rate, r should be constant, hence temperatures should increase geometrically, i.e. $\beta_{i+1} = r\,\beta_i$. The optimal value for r can be found by maximizing σ^2. It turns out, see Fig. 6.4, that the average acceptance rate for the optimal r is very close to 0.23, with essentially no dependence on Nc_V [38–40]. This gives rise to the so-called 0.23 rule, according to which temperatures should be spaced in such a way to guarantee a 0.23 average acceptance rate. Note also that $Nc_V\sigma^2$ is essentially a universal function of $p_{\rm acc}$, see Fig. 6.4, that converges very quickly to its large Nc_V limit

$$Nc_V\sigma^2 = 4\, p_{\rm acc}\left[\mathrm{erfc}^{-1}(p_{\rm acc})\right]^2.$$

This function has a maximum of height 0.6629 at $p_{\rm acc} = 0.2338$. Hence

$$(\ln r_{\text{opt}})^2 = \frac{0.6629}{0.2338}\frac{1}{Nc_V}\,, \qquad r_{\text{opt}} = 1 - \frac{1.684}{\sqrt{Nc_V}}\,.$$

As already found in Sect. 6.2 when discussing data reweighting, also in this case $\Delta\beta \propto (1 - r_{\text{opt}}) \propto (Nc_V)^{-1/2}$. If the specific heat is not constant, $\Delta\beta \propto (Nc_V)^{-1/2}$ should still hold, hence temperatures should be denser where the specific heat is larger.

6.6.4 Improving Parallel Tempering

Sometimes, even with an optimal choice of the temperatures, the PT simulation may show up a bottleneck in temperature, unexpectedly. The problem is that the analytical computation of the swapping rate is made under the hypothesis that each configuration at inverse temperature β is generated according to $\pi_\beta(x)$ with no memory of its past trajectory; this assumption is valid if the time Δt between two consecutive swapping attempts is larger than the autocorrelation time τ_β at β. On the contrary, if $\tau_\beta > \Delta t$, then a replica is likely to swap back to the temperature it came from, since its energy is still correlated with its old temperature. This phenomenon of swapping forward and then immediately backward is exactly what makes diffusion in temperature space much slower.

Recently there have been some proposals to overcome this problem and improve the PT method. In Ref. [41] a method called feedback-optimized PT has been proposed, which iteratively readjusts the temperatures in order to minimize the average round-trip time. The outcome of this procedure is an increase of the density of temperatures (and thus of the swapping rate) where the autocorrelation time τ is larger. In some sense this solution can be viewed as a brute-force one, because forces replicas to spend more time where τ is larger by adding temperatures there. A more elegant solution has been proposed in Ref. [42] and it consists in adapting the time Δt between consecutive swapping attempts to the autocorrelation time τ. Indeed, results for the 2D Ising model show that, by taking $\Delta t \sim \tau$, the resulting time series are nearly uncorrelated and replicas make an unbiased diffusion among temperatures; unfortunately this choice makes the simulation too long, so the final suggestion is to have the ratio $\Delta t/\tau$ more or less fixed to a small number.

Let us finish this overview of the PT method with a comment on its use for disordered systems. Certainly the numerical study of disordered models (e.g., spin glasses) has benefited very much from the PT algorithm in the last decades. Nonetheless, it is important to recall that models with strong quenched disorder show impressive sample-to-sample fluctuations. As a consequence, the optimizations illustrated above should be performed separately on each different sample: indeed, we would expect a very different scheduling of temperatures and swapping times for a strongly frustrated sample with respect to a weakly frustrated one. Since this sample-by-sample optimization is not easy to do, in practice one usually fixes a common scheduling of temperatures and times for all samples, based on average

properties (e.g., on the sample-averaged specific heat). However, thermalization checks and autocorrelation-time analyses should be performed on each sample separately, allowing the simulation to run longer for the slower samples [43].

6.7 Conclusions

In this contribution we present several numerical methods which are used to compute large-deviation observables, that is quantities that require a proper sampling of rare configurations. First, we discuss the problem of data reweighting and the optimal multiple-histogram method [6, 13]. Then, we introduce a family of algorithms that rely on non-Boltzmann-Gibbs distributions and which are able to sample the typical configurations corresponding to a large temperature interval. We present the umbrella sampling [15] and the simulated tempering method [16, 17] and show that these two algorithms are equivalent [18] if configurational updates are performed by using the Metropolis method. The main difficulty in the implementation of the US and ST methods is the determination of the constants α_i that parametrize the probability distribution, see Eqs. (6.12) and (6.15). This problem can be overcome by using the PT algorithm [27–29], which is at present the most efficient algorithm to sample the low-temperature phase of systems undergoing a second-order phase transition, even in the presence of quenched disorder—hence, it can be applied successfully to, e.g., spin glasses. None of these methods can be employed directly in the presence of first-order phase transitions. Multicanonical methods, in which the non-Boltzmann-Gibbs distribution is determined recursively, can be used instead [25, 26].

References

1. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. **21**, 1087 (1953)
2. M.N. Rosenbluth, A.W. Rosenbluth, J. Chem. Phys. **22**, 881 (1954)
3. B.J. Alder, S.P. Frankel, V.A. Lewinson, J. Chem. Phys. **23**, 417 (1955)
4. B.J. Alder, T.E. Wainwright, J. Chem. Phys. **27**, 1208 (1957)
5. S.R.S. Varadhan, Ann. Prob. **36**, 397 (2008)
6. A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988); Erratum ibid. **63**, 1658 (1989)
7. R.H. Swendsen, A.M. Ferrenberg, in *Computer Studies in Condensed Matter Physics II*, ed. by D.P. Landau, K.K. Mon, H.B. Schüttler (Springer, Berlin, 1990), pp. 179–183
8. A.M. Ferrenberg, D.P. Landau, R.H. Swendsen, Phys. Rev. E **51**, 5092 (1995)
9. B. Efron, Ann. Stat. **7**, 1 (1979)
10. B. Efron, Biometrika **68**, 589 (1981)
11. L. Onsager, Phys. Rev. **65**, 117 (1944)
12. B.M. McCoy, T.T. Wu, *The Two-Dimensional Ising Model* (Harvard University Press, Cambridge, 1973)

13. A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989)
14. C.H. Bennett, J. Comp. Phys. **22**, 245 (1976)
15. G.M. Torrie, J.P. Valleau, J. Comp. Phys. **23**, 187 (1977)
16. A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov, P.N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1991)
17. E. Marinari, G. Parisi, Europhys. Lett. **19**, 451 (1992)
18. N. Madras, M. Piccioni, Ann. Appl. Prob. **9**, 1202 (1999)
19. R.H. Swendsen, J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987)
20. R.G. Edwards, A.D. Sokal, Phys. Rev. D **38**, 2009 (1988)
21. W.K. Hastings, Biometrika **57**, 97 (1970)
22. N. Bhatnagar, D. Randall, in *Proceeding of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans (ACM, New York, 2004), pp. 478–487
23. M. Fukugita, H. Mino, M. Okawa, A. Ukawa, J. Phys. A **23**, L561 (1990)
24. J.F. McCarthy, Phys. Rev. B **41**, 9530 (1990)
25. B. Berg, T. Neuhaus, Phys. Lett. B **267**, 249 (1991)
26. B. Berg, T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992)
27. C.J. Geyer, in *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, ed. by E.M. Keramidas, S.M. Kaufman (American Statistical Association, New York, 1991), pp. 156–163
28. M.C. Tesi, E.J. Janse van Rensburg, E. Orlandini, S.G. Whittington, J. Stat. Phys. **82**, 155 (1996)
29. K. Hukushima, K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996)
30. U.H.E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997)
31. D.J. Earl, M.W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005)
32. N. Madras, Z. Zheng, Random Struct. Alg. **22**, 66 (2003)
33. Z. Zheng, Stoch. Proc. Appl. **104**, 131 (2003)
34. D.B. Woodard, S.C. Schmidler, M. Huber, Ann. Appl. Prob. **19**, 617 (2009)
35. N. Madras, D. Randall, Ann. Appl. Prob. **12**, 581 (2002)
36. F. Martinelli, E. Olivieri, R. Schonmann, Comm. Math. Phys. **165**, 33 (1994)
37. D.A. Kofke, J. Chem. Phys. **117**, 6911 (2002)
38. A. Kone, D.A. Kofke, J. Chem. Phys. **122**, 206101 (2005)
39. C. Predescu, M. Predescu, C.V. Ciobanu, J. Chem. Phys. **120**, 4119 (2004)
40. D. Sabo, M. Meuwly, D.L. Freeman, J.D. Doll, J. Chem. Phys. **128**, 174109 (2008)
41. H.G. Katzgraber, S. Trebst, D.A. Huse, M. Troyer, J. Stat. Mech. P03018 (2006)
42. E. Bittner, A. Nußbaumer, W. Janke, Phys. Rev. Lett. **101**, 130603 (2008)
43. R. Alvarez Baños et al. (Janus collaboration), J. Stat. Mech. P06026 (2010)