Introduction to Monte Carlo methods in statistical physics

miguel.berganza@pr.infn.it

Contents

Abstract

These lecture notes correspond to the 30-hour Ph D course given at the University of Parma in 2016. Not pretending to be self-contained, they can be conceived as a kind of formulary, quite compressed essence of the treated arguments, or as a guide for the lessons.[1]

> Monte Carlo is one of the least efficient [methods]; it should be used only on those intractable problems for which all other numerical methods are even *less* efficient.
>
> A. Sokal in [Sokal(1997)]

> Never make a calculation until you know the answer. Make an estimate before every calculation, try a simple physical argument (symmetry! invariance! conservation!) before every derivation, guess the answer to every paradox and puzzle.
>
> J. A. Wheeler

[1] A set of codes, illustrating many of the topics discussed, can be found in the link [Ibáñez-Berganza(2016)], to which the folder names mentioned in these notes refer. Along the lecture notes, the Examples are thought to be illustrated, sketched or worked out during the lessons, while the more specific Exercises are suggestions of homework for the interested reader. The Exercises often consist in completing an already prepared "skeleton" code, so that the source codes in the mentioned link may be consequently incomplete.

1 Numerical distributions and integration

Cumulative method. To sample a probability distribution (PD) f, of which we know its (invertible) primitive function, F, one samples ξ, uniformly distributed in $[0,1]$ (UD1), and returns $F^{-1}(\xi)$.

Exercise 1. *Develop an algorithm generating a couple of normally distributed variables in polar coordinates using the cumulative method (Box-Muller algorithm, 1985).*

Rejection Sampling. To sample a probability distribution f on a domain $\mathcal{F}$, one chooses an auxiliary PD g over $\mathcal{G} \supset \mathcal{F}$, that one knows how to sample, such that there is a sufficiently large M, $1 < M < \infty$ so that $Mg > f$ everywhere in $\mathcal{F}$. One then extracts $x \in \mathcal{F}$ with probability $g(x)$, and returns x with probability $f(x)/(Mg(x))$. This happens with average probability $1/M$, the efficiency of the algorithm, and with variance $(M-1)/M^2$.

Exercise 2. *To generate uniformly a set of numbers in the d-dim hypersphere, one uniformly generates random vectors in the d-dim ball of radius R, then normalizes them, and returns the resulting normalized vector. Use the rejection algorithm to demonstrate the validity of this algorithm and compute its efficiency $1/M$. Propose a different algorithm outperforming it for large values of d.*

Hit-or-miss Monte Carlo (MC) integration. To integrate a one dimensional function f in $[a,b]$, one chooses $M > \max f$, $m < \min f$, and then: generates ξ_{2j}, ξ_{2j+1}, UD1; computes $x = a + (b-a)\xi_{2j}$ and $y = (M-m)\xi_{2j+1}$; if $y < f(x) - m$ set $R++$; for $j = 1, \ldots, n$. The integral can be estimated as $I \simeq (R/n)(M-m)(b-a) + m(b-a)$. For large n, the standard deviation of the result is given by $s = n^{-1/2}(M-m)(b-a)(R/n - (R/n)^2)^{1/2}$.

Exercise 3. *Check the above proposition on the standard deviation of the hit-or-miss MC integration.*

Exercise 4. *Use the hit-or-miss method to demonstrate the validity of Buffon's 1777 method to compute the number π.*

Crude MC integration. To integrate a one dimensional function f in $[a,b]$, one: generates ξ, UD1; computes $x \equiv a + (b-a)\xi$ and $f_j \equiv f(x)$; for $j = 1, \ldots, n$. The integral of f, I, can be estimated as $I \simeq (b-a)\langle f \rangle_n$ where $\langle f \rangle_n = \sum_{j=1}^n f_j/n$ is the average. In particular, f_j are random numbers with average $\langle f \rangle_\infty = I/(b-a)$ and variance $\sigma^2 = \langle f^2 \rangle_\infty - \langle f \rangle_\infty^2$, the central limit theorem wants the variable $w_n = n^{1/2}(\langle f \rangle_n - \langle f \rangle_\infty)$ to be distributed $(0,\sigma)$-normally for large n, in other words:

$$\langle f \rangle_n = \langle f \rangle_\infty + n^{-1/2}\sigma y \tag{1.1}$$

where y is a standard Gaussian variable. The generalisation to multi-dimensional functions is straightforward, the variance of the estimation is $\sim n^{-1}$, independent of d.

Exercise 5. *Demonstrate the central limit theorem (c.f., for example [Marinari and Parisi(2004)]).*

Importance Sampling. Suppose the (multidimentional) function f to be integrated in the interval $\mathcal{A}$, being very heterogeneous in $\mathcal{A}$, and a probability distribution g on $\mathcal{A}$ such that f/g is less heterogeneous (i.e., $1/g$ is larger in the regions of $\mathcal{A}$ in which f is lower). One, hence, chooses a point x_j in $\mathcal{A}$ with probability $g(x_j)$ and $h_j \equiv f(x_j)/g(x_j)$; for $x = 1, \ldots, n$. The integral of f can be then estimated as $I \simeq (\sum_{j=1}^n h_j)/n$. The error is again $n^{-1/2}\sigma$, but the variance is now $\sigma^2 = \langle h^2 \rangle_\infty - \langle h \rangle_\infty^2$ where $h = f/g$.

Exercise 6. *Consider the integration of the function $I = \int_0^1 \mathrm{d}x \int_0^x \mathrm{d}y\, g(x,y)$, and the following algorithms: 1) generate n couples of UD1 x_j, y_j, $j = 1, \ldots, n$; evaluate $I_1 = (1/n)\sum_j g(x_j, x_j y_j)$. 2) generate n couples of UD1 points x_j, y_j; for each one, if $x_j < y_j$, interchange them ($x_j \equiv y_j$, $y_j \equiv x_j$); compute $I_2 = \sum_j g(x_j, y_j)$. Correct both algorithms so that they estimate I. Which one is more efficient?*

Example 1. *Inefficiency of uniform sampling MC in the canonical ensemble. Recall the canonical ensemble: at inverse temperature β, one is interested in a probability distribution for ϵ, the intensive energy, given by $p_\beta(\epsilon) = \exp[-N\beta\tilde{\Phi}(\beta,\epsilon)]/Z_\beta$, where $\tilde{\Phi} = N(\epsilon - Ts)$ is the free energy functional (and s is the microcanonical entropy), N is the system mass, and Z is the partition function. In saddle-point approximation, it is $Z_\beta = \exp[-N\beta\Phi(\beta)]$, where $\Phi(\beta) = \min_\epsilon \tilde{\Phi}(\epsilon,\beta)$ is the free energy. The probability of finding a configuration with energy ϵ', different from the most probable energy ϵ_β, is, hence, $p_\beta(\epsilon') = \exp[-N\beta(\tilde{\phi}(\beta,\epsilon') - \tilde{\phi}(\beta,\epsilon_\beta))]$, which is exponentially suppressed in N. It follows that a random configuration (as those sampled in an unbiased MC) has exponentially suppressed probability of not having ϵ_0. On their turn, they have exponentially vanishing probability in an ensemble at $\beta > 0$.*

2 Markov-Chain Monte Carlo

Markov Chains. Consider a discrete space $\mathbf{\Sigma}$ of $\mathcal{N}$ configurations (or states, $\boldsymbol{\sigma}_i$, $i = 1, \ldots, \mathcal{N}$)[2]. A Markov Chain is a sequence of configurations such that the conditional probability of having $\boldsymbol{\sigma}^{(t)}$, the configuration at time t depends only on $\boldsymbol{\sigma}^{(t)}$, $\boldsymbol{\sigma}^{(t-1)}$. The transition probabilities can be cast into a matrix p whose element p_{ij} is the transition probability of the i-th to the j-th state, $i, j = 1, \ldots, \mathcal{N}$. The transition matrix is a stochastic matrix, it satisfies: $p_{ij} > 0 \, \forall i, j$ and $\sum_j p_{ij} = 1$. The Markov Chain characterized by p is said *irreducible* if given any two states i, j, one can reach j form i in a finite time, i.e., if there exists n such that $(p^n)_{ij} > 0$. A stronger property is *aperiodicity*: if there exists a n such that $(p^t)_{ij} > 0$ for all i, j, and for all $t > n$ (an irreducible, aperiodic Markov Chain is said *ergodic*).

The matrix p along with the PD for the first element of the chain, $\pi^{(0)}$, define the Markov Chain, and induce a probability measure on the set of n sequences of states, $\boldsymbol{\sigma}_{i_1} \boldsymbol{\sigma}_{i_2} \ldots$, which is $\pi^{(0)}(\boldsymbol{\sigma}_{i_1}) \, p_{i_1 i_2} p_{i_2 i_3} \ldots$ (averages of observables according to such a sequence of states are denoted by $\langle \cdot \rangle_{\pi^{(0)}}$), and the probability of having the j-th state at time t is $= \sum_i (p^t)_{ij} \pi^{(0)}(\boldsymbol{\sigma}_i)$.

Theorem 1. *Discrete, aperiodic, irreducible Markov Chains are such that*

1. *The limit $\pi_j = \lim_{n \to \infty} (p^n)_{ij}$ uniquely exists, independently on i. $\pi_j \equiv \pi(\boldsymbol{\sigma}_j)$ is a PD ($\sum_j \pi_j = 1$), stationary under p:*

$$\pi_j = \sum_i p_{ij} \pi_i \qquad \text{Balance condition} \tag{2.1}$$

2. *The fraction of times the $j-th$ state is visited in a sequence of length n is independent on $\pi^{(0)}$ and equals π_j, in the limit $\lim_{n \to \infty}$.*

3. *If $f \in l^2(\pi)$ (square-integrable with respect to π) and $f_i \equiv f(\boldsymbol{\sigma}_i)$, it is:*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n} f(\boldsymbol{\sigma}^{(t)}) = \sum_{i=1}^{\mathcal{N}} \pi_i f_i \tag{2.2}$$

regardless of $\pi^{(0)}$, the fluctuations for finite n being of order $n^{-1/2}$.

Exercise 7. *Let us define the distance between two distributions $||\boldsymbol{\alpha} - \boldsymbol{\beta}|| = \sum_j |\alpha_j - \beta_j|$. Consider a Markov Chain satisfying the detailed balance condition equation 2.3, and show that the distance between a vector $\boldsymbol{v}$ and the stationary distribution $\boldsymbol{\pi}$ is larger than that between $\boldsymbol{v}^\dagger p$ and $\boldsymbol{\pi}^\dagger$. This proves that the stationary distribution is a fixed point of the matrix p.*

Exercise 8. *Consider a stochastic matrix p with no zero elements and with ϵ being is its lower entry. If M_0, m_0 are the maximum and minimum components of vector $\mathbf{v}$, and M_1, m_1 are the maximum and minimum components of vector $p\mathbf{v}$, then it is $(M_1 - m_1) \leq (1 - 2\epsilon)(M_0 - m_0)$. It follows that if $d_n^{(j)} = M_n^{(j)} - m_n^{(j)}$ is the difference between the maximum an minimum of the j-th column of p^n, it is $d_n^{(j)} < (1 - 2\epsilon)^n$, where we have set $d_1^{(j)} < (1 - 2\epsilon)$. Check that for an aperiodic stochastic matrix p, it is $(p^n)_{ij} = \pi_j + e_{ij}$ where $|e_{ij}| < c \, r^n$, being $c > 0$, $0 < r < 1$ (take $r = (1 - 2\epsilon)^{1/M}$, $c = (1 - 2\epsilon)^{-1}$, M is the first integer such that p^M has no zero elements, and epsilon is the lower entry of p^M) (see [Bhat and Miller(2002)]).*

Exercise 9. *In an aperiodic Markov chain, consider the measured rate of appearance of the j-th state: $\overline{\pi}_j^{(n)} = \sum_{k=1}^{n} \delta_{\boldsymbol{\sigma}^{(k)}, \boldsymbol{\sigma}_j} / n$. Show that the average $\langle \cdot \rangle_\alpha$ of this stochastic number in the MC with initial distribution α is $= \pi_j$ with $\alpha = \pi$. For $\alpha \neq \pi$, it converges to π_j for large n, the bias being of order $1/n$ (hint: use the matrix identity $(1 - X)^{-1} = \sum_{m \geq 0} X^m$).*

The dynamic (or Markov-Chain) MC method consists in choosing a transition matrix P such that its stationary distribution π is the desired one. The theorem before requires for the dynamic MC method to work, that 1) p must be irreducible and 2) that it satisfies the Balance condition. A sufficient condition for balance is *detailed balance*:

[2] we will deal with states composed by N degrees of freedom, $\mathbf{\Sigma} = \mathbf{\Sigma}_1^{\otimes N}$, where $\mathbf{\Sigma}_1$ is the single-particle degree of freedom (a binary spin $\mathcal{S}_1 = \{0, 1\}$, in the case of the Ising model, for example)

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \text{Detailed balance condition} \tag{2.3}$$

The two main limitations of dynamic MC sampling and their associated timescales are the following: 1) when the starting PD $\pi^{(0)}$ is not the stationary distribution, the first measurements produced by the dynamic MC are not distributed according to π, this only happens after an order of steps bounded from above by the *transient or exponential time* τ_{e}. 2) Once the chain is in equilibrium, for times much larger than τ_{e}, measurements of an observable $f \in l^2(\pi)$ from configurations separated by consecutive steps are not independent; the fundamental timescale associated with this problem is the *(auto) correlation or integrated time* $\tau_{\mathrm{i},f}$, the order of consecutive MC steps needed to obtain decorrelated values of the observable f.

The (integrated) correlation time corresponding to f may be defined as:

$$\tau_{\mathrm{i},f} = 1 + 2\sum_{t=1}^{\infty} \rho_f(t); \qquad \rho_f(t) = \frac{C_f(t)}{C_f(0)} \tag{2.4}$$

where C_f is the temporal correlation function of f:

$$C_f(t) = \langle f(\boldsymbol{\sigma}^{(s)}) f(\boldsymbol{\sigma}^{(s+t)} \rangle_\pi - \langle f(\boldsymbol{\sigma}^{(s)}) \rangle_\pi^2 \tag{2.5}$$

(in equilibrium the average $\langle\cdot\rangle_\pi$ is time-translational invariant). One has that the variance of the average of f in the dynamic MC after n steps, supposing that the equilibrium has been attained (i.e., the variance according to $\langle\cdot\rangle_\pi$) is given by (check it!):

$$\lim_{n\to\infty} \mathrm{Var}_\pi \left[\frac{1}{n} \sum_{k=1}^{n} f(\boldsymbol{\sigma}^{(k)}) \right] \simeq \frac{\tau_{\mathrm{i}}}{n} C_f(0) \tag{2.6}$$

i.e., it decays with n as in the unbiased or uncorrelated MC, but it is roughly τ_{i} times larger.

Transient (or exponential correlation) time. It represents the timescale of the slowest mode. It can be defined as the maximum $\tau_{\mathrm{e}} = \sup_f \tau_{\mathrm{e},f}$ over the operator-dependent exponential times: $\tau_{\mathrm{e},f} = \lim_{t\to\infty} \sup(-t)/\ln|\rho_f(t)|$.
τ_{e} is equivalently defined as $R = \exp(-1/\tau_{\mathrm{e}})$ where R is the spectral radius of $P - \Pi$ being $\Pi_{ij} = \pi_j$ (i.e., the spectral radius of P acting on the orthogonal complement of the constant functions). If R_A is the spectral radius of matrix $A \in \mathbb{C}^{n\times n}$, then $R_A < 1$ if and only if $\lim_{k\to\infty} A^k = 0$; it follows that the spectrum of P lies in the unit circle and that aperiodic chains exhibit a single eigenvalue in the complex unit circumference, corresponding to the stationary distribution. R is the modulus of the eigenvalue of P with second larger modulus (the interested reader is invited to check the equivalence of both definitions of τ_{e} in the case of a *reversible*, i.e., that satisfies the detailed balance condition, Markov Chain.

One can see that R governs the convergence to equilibrium of the Markov Chain: consider the difference between the MC average of the observable f, with α as initial PD on $\boldsymbol{\Sigma}$, and the average according to π:

$$|\langle f(\boldsymbol{\sigma}^{(t)}) \rangle_\alpha - \langle f \rangle| = \left| (\boldsymbol{\alpha} - \boldsymbol{\pi})^\dagger \cdot (P - \Pi)^t \cdot \mathbf{f} \right| \tag{2.7}$$

(check it!). By the spectral radius formula this is, for high enough t [3]

$$|\langle f(\boldsymbol{\sigma}^{(t)}) \rangle_\alpha - \langle f \rangle| \leq \left\| (P - \Pi)^t \right\| \left| (\boldsymbol{\alpha} - \boldsymbol{\pi})^\dagger \cdot \mathbf{f} \right| \leq e^{-t/\tau_{\mathrm{e}}} |\langle f \rangle_\alpha - \langle f \rangle| \tag{2.8}$$

i.e., the initial bias decreases exponentially with time, if τ_{e} is finite.

Exercise 10. *Use the spectral decomposition of ρ_f to prove that $\tau_{\mathrm{i}f} \leq \tau_{\mathrm{e}}$ (hint: check [Sokal(1997)]).*

[3] mind that $R_A \leq ||A^k||^{1/k}$.

Role of self-correlation times in numerical studies of phase transitions

- Transient time. One needs τ_e to know how many initial MC measurements have to be discarded before constructing the average of the desierd observable (although in principle it is not needed, since the initial bias decays as n^{-1} after n steps, while the correlation bias dominates as it is $\sim n^{-1/2}$). For most systems of interest, one cannot know τ_e, it diverges or it is overly conservative for the purposes of the MC dynamics. One could estimate it measuring C_f for a given number of observables. Also in this case, however, the estimated value of τ after n steps is reliable only if the true $\tau_e \ll n$.

- Alternatively, one could perform *empirical equilibration checks*, as comparing the stationarity of the relevant observable histogram over different (exponentially larger and larger) intervals of time. Empirical methods may provide wrong answers in the presence of *metastability* or out-of-equilibrium stationarity: the system may present stationarity at times much lower than τ_e. There are no absolute recipes to prevent such a bias. A practical solution is to change the initial distribution $\pi^{(0)}$ (hopefully, exploring the various minima of the relevant thermodynamic potential).

- Correlation time. It is needed to estimate the *efficiency* of the algorithm (defined as ${\tau_i}^{-1}$ divided by the computer time to perform one step of the algorithm), and also to estimate the errors of the desired quantity (of order $(\tau_{i,f}/n)^{1/2}$) or, conversely, the numer n of needed MC steps to achieve a desired accuracy.

Exercise 11. *What? Didn't you say (see (2.8)) that the initial bias decreases exponentially? And now (and in Exercise 9) you say that it decays as* $1/n$ *! Where is the apparent paradox? Convince yourself that the MC bias in the average of a function* f *due to the initial transient is* $\sim n^{-1}$.

Sources of correlations in phase transitions. In substance, we have seen that the efficiency of dynamic MC algorithms is compromised by temporal correlations. In systems undergoing phase transitions, frequent sources of temporal correlations are:

- Critical slowing down ([Hohenberg and Halperin(1977)]). The integrated autocorrelation time is proportional to the z-th power of the correlation length (at intensive thermodynamic variable μ) $\tau \sim \xi(\mu)^z$, z being the *dynamic critical exponent*. In the vicinity of continuous phase transitions τ, hence, it diverges as $\tau \sim |\mu - \mu^*|^{-z\nu}$ in an infinite system, and as $\tau \sim L^z$ in a finite system at μ^* (see Sec. 5). Many *local* MC algorithms exhibit a large dynamical critical exponent, near 2. Cluster methods (see Subsec. 3.3) sensibly reduce them.

- Metastability. In first-order transitions, there is a more severe slowing-down problem, called *exponential critical slowing down*: in phase coexistence, both phases are separated by a surface tension Σ, inducing a thermodynamic potential barrier of heigh of order $\Sigma V^{(d-1)/d}$. The time needed to overcome it is of order $\exp(\Sigma V^{(d-1)/d})$ ([Binder(1987)]).

- Glassyness. Systems with glassy behavior present an exponential growth of the relaxation time [Debenedetti(1996)]. For the continuous transition of spin glasses, the dynamical critical exponent may assume very high values [?]. The *Parallel tempering* algorithm is used to mitigate this effect.

3 Monte Carlo algorithms and phase transitions

3.1 Metropolis and Gibbs sampling (heatbath) algorithms

Metropolis-Hastings algorithm. A general way of constructing a Markov Chain is first proposing a transition from state i-th to j-th defined by the *proposal matrix* $p^{(0)}_{ij}$ (where $p^{(0)}$ is a stochastic irreducible matrix), and accepting it with probability a_{ij}. The transition matrix is hence $p_{ij} = a_{ij}p^{(0)}_{ij}$ for $i \neq j$ and $p_{ii} = p^{(0)}_{ii} + \sum_{j\neq i} p^{(0)}_{ij}(1 - a_{ij})$ (for the correct normalization it is necessary to keep refused configurations). Detailed balance is satisfied if

$$a_{ij} = F\left(\frac{\pi_j p^{(0)}_{ji}}{\pi_i p^{(0)}_{ij}}\right) \tag{3.1}$$

being $F : \mathbb{R}^+ \to [0 : 1]$ satisfying $F(x) = xF(1/x)$. The *Metropolis algorithm* corresponds to:

$$F(x) = \min\{x, 1\} \tag{3.2}$$

If the proposal matrix satisfies detailed balance, all the proposals are accepted. For any symmetric irreducible proposal matrix, the acceptance probabilities depends only on the ratio between the target distribution probabilities:

$$a_{ij} = \min\left\{\frac{\pi_j}{\pi_i}, 1\right\} \tag{3.3}$$

in the canonical ensemble at inverse temperature β, for example, this reads to $a_{ij} = \min\{1, \exp(-\beta N(\epsilon_i - \epsilon_j))\}$ where ϵ_j are the per site energy of the j-th configuration.

Single-particle updating. Consider a system composed by N degrees of freedom $\boldsymbol{\sigma} = \otimes_{m=1}^{N}\sigma^{(m)}$, where $\sigma^{(m)}$ is the m-th particle state. Let $p^{(m)}$ be the transition matrix in which only particle m is updated:

$$p^{(m)}_{ij} > 0 \qquad \sigma^{(n)}_i = \sigma^{(n)}_j \qquad \forall n \neq m \tag{3.4}$$

$$p^{(m)}_{ij} = 0 \qquad \text{otherwise} \tag{3.5}$$

Updating a random sequence of particles, one at once, is called *random-particle updating*, and a sequence of N random particle updating is called a *sweep*, the corresponding transition matrix being $p = (1/N)\sum_{m=1}^{N} p^{(m)}$. If the particles are updated following a given sequence of indices $i_1, \ldots, i_N$, the updating is called *sequential*, the corresponding transition matrix being $p = \prod_{m=1}^{N} p^{(i_m)}$. Still a different scheme is called M- *multi-hit algorithm*, in which one selects one particle, and applies the Metropolis algorithm M times (proposing a new state for particle m and accepting it with matrix a), whose transition matrix corresponds to $p = (1/N)\sum_{m=1}^{N}[p^{(m)}]^M$. If the single-particle transition matrices satisfy detailed balance, so does the random-particle updating matrix, while the sequential matrices satisfy, in general, only the balance condition (which is the required condition for a valid MC).

Exercise 12. *Why not proposing attempts of all the particles at once? Consider the effect of this strategy in the mean square displacement per computer time, in a gas, or in its equivalent in a magnetic system.*

Gibbs sampling (heatbath) algorithm. We define the transition matrix of the *heat bath* algorithm as $p^{(m)}[\boldsymbol{\sigma} \to \boldsymbol{\sigma}'] = \pi^{(m)}(\sigma'^{(m)}|\boldsymbol{\sigma}_{\setminus m})$, equal to the marginal stationary probability distribution of the m-th particle degree of freedom, given the rest of the configuration $\boldsymbol{\sigma}_{\setminus m}$, and new and old configurations being equal except by the m-th particle, $\boldsymbol{\sigma}'_{\setminus m} = \boldsymbol{\sigma}_{\setminus m}$. In other words, the *Gibbs sampling* or *heatbath* algorithm proposes a new state of particle m with its marginal stationary probability, independently of the current state of particle m. Many particles can then be sequentially or randomly updated, as in the precedent paragraph. The MC sweep results to satisfy balance, and is aperiodic.

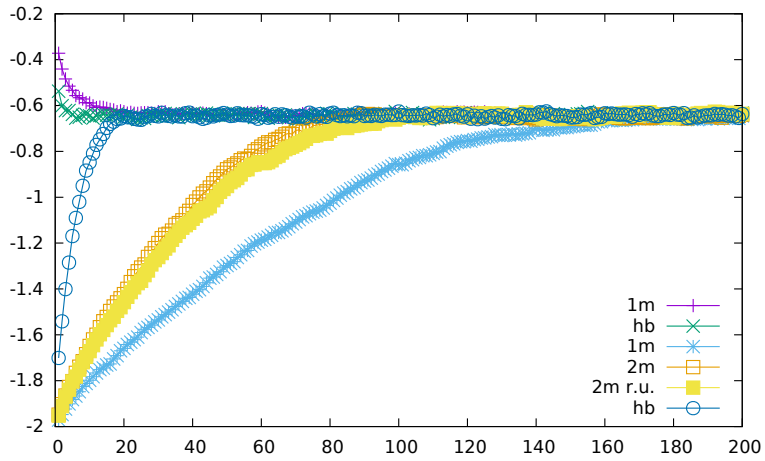

Figure 3.1: Energy vs. number of MC sweeps for the $q = 10$ 2D Potts model in the square lattice with periodic boundary conditions at $\beta = 1.24$, using several MC algorithms (Metropolis, heatbath, 2-hit Metropolis, 2-hit with random sweeps), and starting from ordered and disordered configurations (find the details of the simulation in the folder Potts/beta1.24).

Exercise 13. *The curious reader is invited to discuss the eventual satisfaction of the balance and detailed balance conditions of the algorithm in the case of a magnetic system with the $O(M)$ Hamiltonian $\mathcal{H}[\boldsymbol{\sigma}] = -\sum_i \mathcal{A}_{ij}\vec{\sigma}_i \cdot \vec{\sigma}_j$ (where $\vec{}$ denotes an $\mathcal{S}^{M-1}$ vector). At the level of the single particle, the heat bath algorithm obviously satisfies detailed balance, since it does not depend on the departure state $\boldsymbol{\sigma}$ but only on the target state $\boldsymbol{\sigma}'$. Does the probability $\pi(\boldsymbol{\sigma}')p^{(m)}(\boldsymbol{\sigma}' \to \boldsymbol{\sigma}'')$ equal $\pi(\boldsymbol{\sigma}'')p^{(m)}(\boldsymbol{\sigma}'' \to \boldsymbol{\sigma}')$?*

Example 2. *The q-Potts model Hamiltonian is $\mathcal{H}[\boldsymbol{\sigma}] = -\sum_{1\le i<j\le N} \delta_{\sigma^{(i)}\sigma^{(j)}}\mathcal{A}_{ij}$, where the single particle (spin) space of states is $\sigma_i \in \{1, \dots, q\}$, and where the adjacency matrix $\mathcal{A}$ defines the interaction topology. Consider the Potts model on the square lattice with periodic boundary conditions in the canonical ensemble at inverse temperature β. One can implement the Metropolis and heat-bath algorithms (as done in the codes* potts_main.cpp*) and: 1) test the result, for $q = 2$, with the Onsager solution for the expected value of the magnetization (see* Potts/IsingTest/*); 2) compute the intensive energy $\epsilon = \langle H\rangle/N$ at $\beta = 1.24$, linear length $L = 64$, $q = 10$ using heat-bath, 1-hit and 2-hit Metropolis algorithms, with sequential and random updating (see the folder* Potts/beta1.24/ *and figure 3.1). Which algorithm is faster? What happens with the Metropolis M-multi-hit algorithm in the limit of large M?*

3.2 MC in different ensembles

Example 3. Lennard-Jones fluid. *Consider the Lennard-Jones fluid in the T, V, N ensemble: a collection of N classical particles in a three dimensional box of length $V^{1/d}$ ($d = 3$) with periodic boundary conditions, interacting through the pairwise Lennard-Jones potential (in reduced units):*

$$u(r) = 4\left[r^{-12} - r^{-6}\right]. \tag{3.6}$$

One can use the Metropolis algorithm to compute the pressure as a function of the density $\rho = N/V = v^{-1}$ for a given temperature (see the code LJ_main.cpp*). The pressure can be computed with the virial equation:*

$$P(\beta, \rho) = T\rho + \frac{1}{3V}\langle \sum_{1<i<j<N} \mathbf{f}(\mathbf{r}_{ij}) \cdot \mathbf{r}_{ij}\rangle_{\beta,V} \tag{3.7}$$

(see, for example, [Hansen and McDonald(1990)]), where $\mathbf{f}(\mathbf{r})$ is the force between two particles separated by the vector $\mathbf{r}$. The interaction range is to be cutoff in some way: one can cutoff it at the half box length $L/2$, and use the tail correction for the energy and pressure (see [Frenkel and Smit(2001)]). To check the correctness of the algorithm, one may compare the resuling estimation

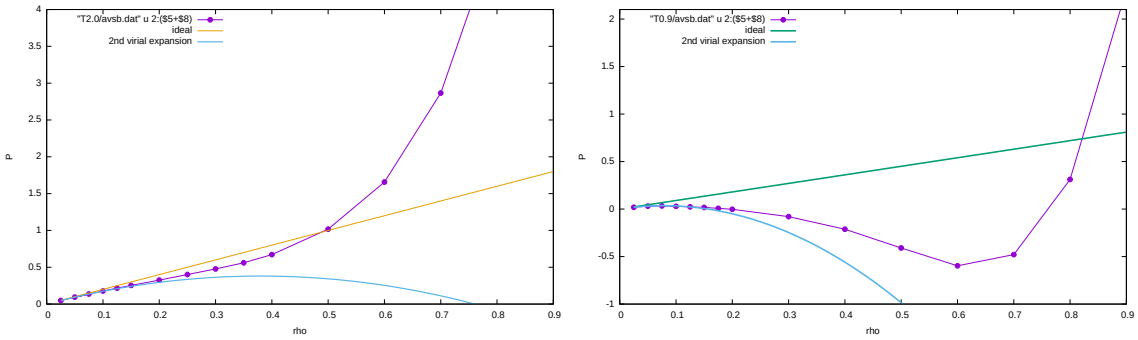

Figure 3.2: P vs ρ for $T = 2.0$ (left) and for $T = 0.9$ (right) of a Lennard-Jones fluid with $N = 130$ in a cubic lattice with periodic boundary conditions according to the Metropolis algorithm (points, see the codes and details of the simulation in the folder LJ/) and to the 1^{st} and 2^{nd}-order Virial expansion.

of the pressure with both its 2^{nd}-order virial expansion for the Lennard-Jones gas (with which it must coincide for sufficiently low ρ or sufficiently large T):

$$\beta P(\beta, V) = \rho + \sum_{i>1} \rho^i B_i(\beta) \qquad B_2(\beta) = -2\pi \int_0^\infty \mathrm{d}r\, r^2 \left[\exp(-\beta u(r)) - 1\right], \tag{3.8}$$

and with the empirical equation of state of [Kolafa and Nezbeda(1994)]. What happens with the pressure when we take low temperatures, below the condensation point (say, $T = 0.9$) in the density range $\rho < 1$, in systems with $N \lesssim 500$ (see figure 3.2), is it as you expected? Do this behavior correspond to equillibrium (the answer may be unexpected, c.f. [Landau et al.(2010)Landau, Lewis, Schuettler, Nussbaumer, Bittner, Neuhaus, and Janke])? Are there metastable effects in this case?

Example 4. Metropolis algorithm in the isobaric ensemble. *Consider a fluid system whose degrees of freedom $\boldsymbol{\sigma}$ are the spatial positions in a d-dimensional space $\boldsymbol{\sigma} = \{\mathbf{r}_i\}_{i=1}^N$.*[4] *In the isobaric (β, P, N) ensemble, the partition function is:*[5]

$$\Theta(\beta, P, N) = \int \mathrm{d}V\, e^{-\beta N[(f(\beta,v)+Pv)]} = \int \mathrm{d}V \int_V \mathrm{d}\mathbf{r}\, e^{-\beta \mathcal{H}[\mathbf{r}] - \beta PV} \tag{3.10}$$

While the probability of having a configuration $\mathbf{r} = \{\mathbf{r}_i\}_{i=1}^N$ enclosed in a volume V is:

$$p_{\beta,P,N}[\mathbf{r}; V] = \frac{1}{\Theta(\beta, P, N)} e^{-\beta \mathcal{H}[\mathbf{r}] - \beta PV} \tag{3.11}$$

We now promote the volume as one extra degree of freedom, so that the coordinates are parametrized as $\mathbf{r}_i = L\mathbf{s}_i$ where $L = V^{1/d}$. The partition function as a function of the degrees of freedom $\mathbf{s}$ reads:

$$\Theta(\beta, P, N) = \int \mathrm{d}V \int_1 \mathrm{d}\mathbf{s}\, V^N\, e^{-\beta \mathcal{H}[\mathbf{s}\cdot L] - \beta PV} \tag{3.12}$$

the probability of having a rescaled configuration $\mathbf{s}$ in a volume V is:

$$\tilde{p}_{\beta,P,N}[\mathbf{s}; V] = \frac{1}{\Theta(\beta, P, N)} e^{-\beta(\mathcal{H}[\mathbf{s}\cdot L] + PV - TN \ln V)} \tag{3.13}$$

One can use this equation to construct a Metropolis algorithm in which both $\mathbf{s}$ and V are changed independently (it is actually more convenient to perform random changes in $\ln V$ –what would be the consequent change in the probability?–).

[4] The integration over the momenta would introduce a further irrelevant factor Λ^{dN}, $\Lambda = (2\pi m T)^{1/2}$ in Θ, where m is the mass of the particles

[5] The partition function for large N is dominated by the most probable value of V:

$$\Theta(\beta, P, N) = e^{-\beta N[g(\beta,P)]} \qquad \text{large } N \tag{3.9}$$

where $g(\beta, P) = \min_v[f(\beta, v) + Pv]$, from which we learn that $P(\beta, v) = -\partial_v f|_{\beta,v}$.

Exercise 14. *Implement a Metropolis algorithm in the β, P, N ensemble as explained in Example 4, apply it to study the Lennard-Jones system of example 3. Does this algorithm solve the problem found at low temperatures? In other words: does one obtain the coexistence curve in this ensemble? Are there metastable effects in this case?*

3.3 Cluster algorithms

Fortuin-Kasteleyn representation. Consider the q-color Potts model, with Hamiltonian $\mathcal{H}[\boldsymbol{\sigma}] = \sum_{(i,j)} J_{ij}(1 - \delta_{\sigma^{(i)},\sigma^{(j)}})$, in a lattice with N nodes in d dimensions, where (i,j) are unordered sets defining bonds of the lattice. The partition function can be written as:

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} \prod_{(ij)} \left[\delta_{\sigma_i,\sigma_j} w_{ij} + (1 - w_{ij})\right] \tag{3.14}$$

where $w_{ij} = 1 - e^{-\beta J_{ij}}$. One can write this in a redundant way as:

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma},\boldsymbol{n}} \prod_{(ij)} \left[\delta_{\sigma_i,\sigma_j} w_{ij} n_{ij} + (1 - w_{ij})(1 - n_{ij})\right] \tag{3.15}$$

where $\boldsymbol{n} = \{n_{ij}\}_{1<i<j<N}$, $n_{ij} = 1, 0$ is a configuration of active (=1) or inactive (=0) bonds. In other words, we can express the partition function in terms of the bond degrees of freedom only (Fortuin-Kasteleyn representation, [Fortuin(1969)]):

$$Z_N(\beta) = \sum_{\boldsymbol{n}} \left[\prod_{(ij)|n_{ij}=1} w_{ij}\right] \left[\prod_{(ij)|n_{ij}=0} (1 - w_{ij})\right] q^{c[\boldsymbol{n}]} \tag{3.16}$$

where $c[\boldsymbol{n}]$ is the number of connected components of the bond configuration $\boldsymbol{n}$ (an isolated site being a connected component) (notice that if $J_{ij} = 1$, it is $Z = \sum_{\boldsymbol{n}} w^{b[\boldsymbol{n}]}(1-w)^{b_0 - b[\boldsymbol{n}]} q^{c(\boldsymbol{n})}$, where b_0 is the total number of bonds in the lattice ($= dN$ in a d-dimensional hypercubic lattice with periodic boundary conditions) and $b[\boldsymbol{n}]$ is the number of bonds of the bond configuration $\boldsymbol{n}$). One can also generalize the model, defining a probability measure in the set of all possible sets of bonds and configurations, treating them independently, based on eq. 3.15:

$$\mu_N(\beta)[\boldsymbol{\sigma}, \boldsymbol{n}] = \frac{1}{Z_N(\beta)} \prod_{(ij)} \left[\delta_{\sigma_i,\sigma_j} w_{ij} n_{ij} + (1 - w_{ij})(1 - n_{ij})\right] \tag{3.17}$$

The marginal probability for the spins (summing μ over the bonds) gives the canonical probability distribution of the Potts model, while the marginal probability of the bonds (the terms in the sum of 3.16) gives the so called *random cluster model*.

Cluster algorithms. The Fortuin-Kasteleyn representation has been used to construct a series of algorithms ([Swendsen and Wang(1987)],[Edwards and Sokal(1988)],[Wolff(1989)]) which, for many models, present significantly lower values of the dynamical critical exponent z (references on the applicability of these algorithms for different systems can be found in [Amit and Martín-Mayor(2005)]). Equation 3.17 provide the conditional probability of having a bond configuration, given a spin configuration, and vice versa. In particular, the probability of a bond (i,j) to be active, given the $\boldsymbol{\sigma}$, is: if $\sigma^{(i)} = \sigma^{(j)}$, $n_{ij} = 1$ with probability $1 - e^{-\beta J_{ij}}$; if $\sigma^{(i)} \neq \sigma^{(j)}$, $n_{ij} = 0$. Vice-versa, in the bond (i,j), the spins have uncorrelated random values if $n_{ij} = 0$ and assume a random common value if $n_{ij} = 1$. The Swendsen-Wang cluster algorithm [Swendsen and Wang(1987)] is constructed in the following way: given a spin configuration $\boldsymbol{\sigma}$, one generates a bond configuration with conditional probability given by (3.17); from the resulting bond configuration $\boldsymbol{n}$ one generates a new spin configuration, again with probability given by (3.17), from which one generates a new bond configuration, and so on. The algorithm is a sequence of alternating (first moving bonds only, then moving spins only) heat bath MC changes in the extended system of bonds+spins degrees of freedom. The heatbath MC satisfying the balance condition, after a sufficiently large number of cluster iterations, the transient becomes irrelevant and the resulting visited spin configurations are distributed according to the Potts model probability distribution in the canonical ensemble (which, mind, is the marginal probability (3.17), summing over the bond configurations).

A variant is the Wolff cluster algorithm [Wolff(1989)]: given a spin configuration, one selects a spin randomly, consider the *geometrical cluster* to which it belongs (the set of connected lattice

bonds presenting its same colour); one then generates a FK cluster, or a connected subset of bonds of the geometrical cluster such that the bonds are active ($n_{ij} = 1$) with probability $1 - e^{-\beta J_{ij}}$); one finally selects a new spin configuration by (randomly) "coloring" the resulting FK cluster.

Wolff embedding. One can generalize the Wolff algorithm above to models with continuous degrees of freedom. Consider the $O(M > 1)$ model, one can superimpose an Ising like model, in which every spin has a degree of freedom $s_i = 0, 1$, representing a reflection on the plane perpendicular to a reference vector $\vec{r}$, $\vec{\sigma} \to \vec{\sigma} - 2s_i(\vec{r} \cdot \vec{\sigma})\vec{r}$. One has, in this way, a $q = 2$ effective Potts model with interaction strength $J_{ij} = (\vec{\sigma}_i \cdot \vec{r})(\vec{\sigma}_j \cdot \vec{r})$, for which one can apply the Wolff algorithm [Wolff(1989)].

Exercise 15. Cluster estimators. *(1) Consider the spin-spin correlation function of the ferromagnetic q-color Potts model at inverse temperature β, $\langle \delta_{\sigma^{(i)},\sigma^{(j)}} \rangle_\beta$, and demonstrate that it coincides with $\langle (q-1)\gamma_{ij}/q + 1 \rangle_p$, where γ_{ij} is an observable in the bond configurations, $= 1$ if the sites i, j belong to the same bond cluster, and $= 0$ otherwise, and where the average is over the random cluster model with $p = 1 - e^{-\beta}$ (see [Sokal(1997)] for references at this regard). (2) Calculate the observables energy, $\langle \mathcal{H} \rangle / \mathcal{N}$ and magnetization, $\langle \mathcal{M} \rangle$, $\mathcal{M}[\boldsymbol{\sigma}] = \sum_{i=1}^{N} \delta_{\sigma^{(i)},0}/N$, as a function of bond quantities (so that, using cluster algorithms, one can avoid computing energy and magnetization explicitly).*

Example 5. *One can implement the Wolff algorithm for the Potts model in the square lattice (as done in the folder* PottsWolff*), both recursively and non-recursively. What choice is faster, and how does scale the CPU time with the system size in both cases? Is the Wolff algorithm efficient at undercritical temperatures? Why does one obtain a null average magnetization (within fluctuations) even at undercritical temperatures?*

Exercise 16. Transition temperature of the 2D Potts model by duality. *Consider the q-color Potts model with N sites in the square lattice, with Hamiltonian $\mathcal{H}[\boldsymbol{\sigma}] = -\sum_{(i,j)} \delta_{\sigma^{(i)},\sigma^{(j)}}$, where (i, j) is an unordered set defining a bond of the square lattice. In the Fortuin-Kasteleyn representation, the partition function can be written (check!) as:*

$$Z_N(\beta) = \sum_{b,c} \mathcal{N}_N(b, c) w^b q^c \tag{3.18}$$

where $\mathcal{N}_N(b, c)$ is the number of bond configurations in a square lattice (with the given boundary conditions) with b bonds and c connected components ($c = 1, \ldots, N$ and $b = 0, \ldots, 2N$ with periodic boundary conditions), and where $w = e^\beta - 1$ (notice that every isolated site is to be understood as a component in itself). In the limit of infinite and zero β (order and disorder, respectively), the partition function is $Z_N = qw^{2N}$ and $= q^N$, respectively. One can begin to "disorder" the ordered configuration with one connected component and $b = 2N$, introducing ℓ bonds, $b = 2N - \ell$. Let k be the number of connected components that has been created with the introduction of ℓ bonds, so that $c = k + 1$. Let $\mathcal{N}_N^{(\mathrm{o})}(\ell, k) = \mathcal{N}_N(2N - \ell, k + 1)$. It is immediate to see that (3.18) becomes:

$$Z_N(\beta) = w^{2N} q \sum_{\ell,k} \mathcal{N}_N^{(\mathrm{o})}(\ell, k) \left(\frac{w}{\sqrt{q}}\right)^{-\ell} \left(\frac{1}{\sqrt{q}}\right)^{\ell-2k}. \tag{3.19}$$

Conversely, let us "order" the disordered configuration with $c = N$, $b = 0$ by adding ℓ bonds and obtaining k less components with respect to $N - \ell$: $c = N - \ell + k$, and let $\mathcal{N}_N^{(\mathrm{d})}(\ell, k) = \mathcal{N}_N(\ell, N - \ell + k)$. The partition function in terms of $\mathcal{N}^{(\mathrm{d})}$ is:

$$Z_N(\beta) = q^N \sum_{\ell,k} \mathcal{N}_N^{(\mathrm{d})}(\ell, k) \left(\frac{w}{\sqrt{q}}\right)^{\ell} \left(\frac{1}{\sqrt{q}}\right)^{\ell-2k} \tag{3.20}$$

1) Convince yourself of the following, crucial, result: due to the duality of the square lattice, $\mathcal{N}_N^{(\mathrm{o})} = \mathcal{N}_N^{(\mathrm{d})}$. 2) From the two equations above, calculate the transition temperature of the Potts model in the square lattice (that for which the (large-N) disorder and the order free energies coincide).

4 Error estimation

Jackknife error and bias estimation for uncorrelated quantities. An estimation for the function f of the uncorrelated data $\{x_i\}_{i=1}^n$ is $f(\bar{x})$, where $\bar{x}$ is the average estimation over the data. The bias of this estimation, $\langle f - f(\bar{x})\rangle$ decreases as n^{-1}, where the angles $\langle\cdot\rangle$ denote the average according to the true distribution of f.

The error of this quantity cannot be estimated with the error of the mean of $\{f(x_i)\}_i$, which is very biased. It can be estimated with the Jackknife (JK) method: the strategy is to form $n_b = \lfloor n/b$ *blocks* of length b, $B_j = \{b(j-1)+1, b(j-1)+1, \ldots, bj\}$ and jackknife averages $x_i^{(b)}$ for the data:

$$x_i^{(b)} = \frac{1}{n-b}\sum_{j\notin B_i} x_j, \tag{4.1}$$

and their corresponding function values $f_j^{(b)} = f(x_j^{(b)})$. The jackknife estimator of the average of f and for the variance of the average of f are:

$$f^{(b)} = \frac{1}{n_b}\sum_{j=1}^{n_b} f_j^{(b)} \qquad s^2_{(b)}[f] = \frac{n-b}{n}\sum_{j=1}^{n_b}(f_j^{(b)} - f^{(b)})^2 \tag{4.2}$$

If the data is uncorrelated, and the function f is an unbiased estimator, the JK estimation of the variance with $b = 1$ coincides with the variance of the mean of the $\{f(x_i)\}_{i=1}^n$ (check this!). If the x's are uncorrelated, and the estimation $f(\bar{x})$ is biased, an improved, bias-corrected up to order n^{-2} JK estimator, is (check it!)

$$f_1^{(1)} = f(\bar{x}) + (n-1)(f(\bar{x}) - f^{(1)}) \tag{4.3}$$

Exercise 17. *The second-order JK biased-improved estimator is:*

$$f_2^{(b)} = \frac{1}{n_b}\sum_{j=1}^{n_b}\left[f_j^{(b)} + \frac{1}{n_b-1}\sum_{k\neq j}(f_j^{(b)} - f_{jk}^{(b)})\right] \tag{4.4}$$

where

$$f_{jk}^{(b)} = f(x_{jk}^{(b)}), \qquad x_{jk}^{(b)} = \frac{1}{n-2b}\sum_{m\neq j, m\neq k}^{n_b} x_m \tag{4.5}$$

An unbiased estimator for $f(x) = x^2$ is (check!) $\bar{x}^2 - (-\bar{x}^2 + \overline{x^2})/(n-1)$. Check that this estimator coincides with the second-order JK biased-improved estimator.

Jackknife error estimation for correlated data. If the data are correlated, the error of x cannot be estimated with the equation for the standard deviation of the mean of x, which is an underestimation. However, if b is of the order of the exponential correlation time, the blocks B are independent, and the unbiased estimator for the variance of the mean of the average of x in different blocks:

$$s^2_{(b)}[x] = \frac{1}{n_b(n_b-1)}\sum_{j=1}^{n_b}\left[\left(\frac{1}{b}\sum_{i\in B_j} x_i\right) - \bar{x}\right]^2 \tag{4.6}$$

becomes a correct estimator for the variance of x.

Estimation of $\tau_{\rm i}$ with the JK method. Incidentally, eq. 4.6 provides a method to estimate $\tau_{\rm i}$. According to eq. 2.6, the variance of x is $\tau_{\rm i}$ times larger than if the data were uncorrelated. We can write:

$$s^2_{\rm unc}[x] = \frac{1}{n(n-1)}\sum_{j=1}^{n}(x_j - \bar{x})^2 \tag{4.7}$$

$$\tau_{\rm i} \simeq \frac{s^2_{(b)}[x]}{s^2_{\rm unc}[x]} \qquad \text{large } b \tag{4.8}$$

How to estimate the error of such a $\tau_{\rm i}$ estimation? One can neglect the error of the denominator in 4.8 in front of the error of the numerator. If b is large enough so that the data are uncorrelated, $s^2_{(b)}[x]$ obeys the χ^2-distribution, and its error can be analytically computed given n_b only. In particular, for m normally distributed data y_i, the variable $(m-1)s^2[y]/\sigma_y^2$ is distributed according to the χ^2-distribution, where $s^2[y]$ is the unbiased estimation for the variance and σ_y^2 the true variance. [6] In other words, the variance lies with probability α in the interval $[(m-1)s^2[y]/\chi^2_{\alpha/2} : (m-1)s^2[y]/\chi^2_{1-\alpha/2}]$, where χ^2_α is the quantile, i.e., the value of the variable χ^2 such that the cumulative χ^2-distribution $F(\chi^2_\alpha) = \alpha$.

Exercise 18. *Demonstrate that, for m normally distributed data y_i, the variable $(m-1)s^2[y]/\sigma_y^2$ is distributed according to the χ^2-distribution (see [Berg(2004)]). Demonstrate eq. 4.9.*

Estimation of $\tau_{\rm i}$ with the correlation function. One could be tempted to estimate $\tau_{\rm i}$ using eq. (2.4), substituting $C_f(t)$ by its numerical estimation:

$$\hat{C}_f(t) = \frac{1}{n-t}\sum_{s=1}^{n-t}\left\{f(\boldsymbol{\sigma}^{(s)}) - [f(\boldsymbol{\sigma}^{(k)})]_k\right\}\left\{f(\boldsymbol{\sigma}^{(s+t)}) - [f(\boldsymbol{\sigma}^{(k)})]_k\right\} \tag{4.10}$$

where $[\cdot]_s = (1/n)\sum_{s=1}^n \cdot$ is the average over n consecutive MC configurations, so that the estimation of $\tau_{\rm i}$ becomes:

$$\hat{\tau}_{\rm i} = 1 + 2\sum_{i=1}^{n}\hat{\rho}_f(i) \tag{4.11}$$

where $\hat{\rho}_f = \hat{C}_f/\hat{C}_f(0)$. Actually, this turns to be a wrong estimate in the sense that the standard deviation of the resulting estimated $\tau_{\rm i}$ does not converge to zero for large n, due to the fact that, in this way, one integrates the statistical fluctuations of $\hat{C}_f$ for large t. A better estimate is to cut off the sum at a value M. The result would be $\hat{\tau}_{\rm i}^{(M)}$. M is to be self-consistently determined as the smallest integer such that $M > c\hat{\tau}_{\rm i}^{(M)}$. c depends on the form of the correlation function, and has to be chosen such that it is not too large (in this case one integrates too undesired fluctuations), and it is not too low (one biased the calculation of $\tau_{\rm i}$, cutting the sum in C_f when its contribution to τ is still significantly). For many correlation functions, the interval $4 < c < 6$ is optimal. A more detailed discussion on this topic can be found in [Sokal(1997)].

Example 6. *One can estimate $\tau_{{\rm i}_\epsilon}$ for the $q = 10$ Potts model on a square lattice of size $L = 40$ with PBC at $\beta = 1.24$, for the Metropolis and heatbath algorithms, as (1): $\hat{\tau}_{{\rm i}_\epsilon}^{(M)}$, defined in the last paragraph (to estimate the error of $\tau_{\rm i}$, you will have to compute the error bars of C_ϵ using the JK method), and (2) using equation 4.8 (see figure 4.2 and the folder* /Potts/JKerror/*).*

Exercise 19. *1) Estimate the efficiency of both algorithms considered in Example 6, in the same circumstances. Does it pay to use heatbath? What do you expect at lower temperatures and lower values of q? 2) Compute the intensive energy $\epsilon = \langle H\rangle/N$ at its (infinite-volume) transition temperature $\beta_q = \ln(1+q^{1/2})$ for $q = 10$ and check that it is compatible with its analytic value $\epsilon_{10} \simeq 1.03179694974411$. 3) Provide an error for the average intensive energy in the circumstances of example 2 ($q = 10$, $\beta = 1.24$, $L = 64$). Your result should be compatible with $\epsilon_{10}(1.24) \simeq -0.643461(25)$. 4) Compare the efficiency of the heatbath and Wolff algorithms for the q-Potts model in the 2D square lattice with $L = 128$ at its transition point, for $q = 2$, 4, 10, 24. What is the explanation for such a behavior?*

Exercise 20. Metastability. *Estimate the intensive energy $\epsilon = \langle H\rangle/N$ of the 2D Potts model with $L = 64$, $q = 20$, at $\beta = \beta_{20} + \delta$, $\delta = 10^{-3}$, starting from a disordered configuration. Do you obtain the equilibrium state? Do you obtain a value of ϵ compatible with the analytical value of the disordered energy at β_{20} ($\epsilon_- \simeq 0.17931557491346$, $\epsilon_+ \simeq 1.37347082958657$)?*

[6] The χ^2-distribution with m degrees of freedom is the one followed by the variable $\chi^2 = \sum_{j=1}^m y_j^2$, where y_j are normally distributed variables. It is

$$f_m(\chi'^2) = mf(m\chi'^2) = a(\Gamma(a))^{-1}e^{-a\chi'^2}(a\chi'^2)^{a-1} \tag{4.9}$$

where $a = m/2$ and $\chi'^2 = \chi^2/m$ is the chi squared variable per degree of freedom.

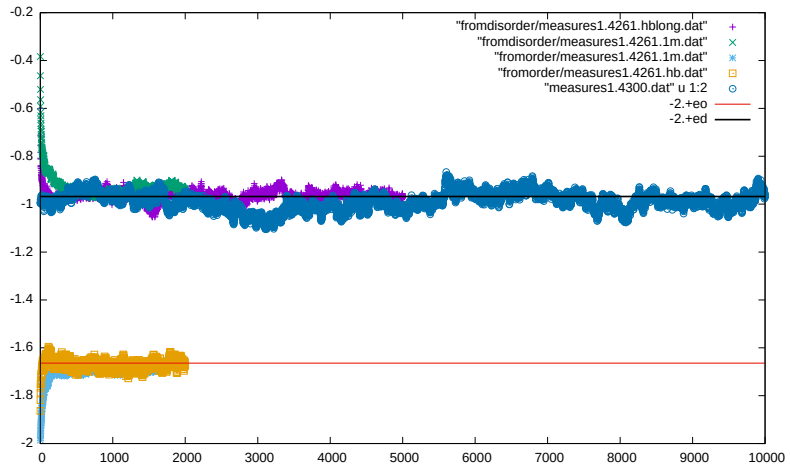

Figure 4.1: Energy versus number of MC sweeps for the $q = 10$ Potts model in a square lattice of linear size $L = 64$ at the transition inverse temperature β_{10}, starting from ordered and from disordered configurations. The blue circle dataset corresponds instead to $\beta_{10} + 5.10^{-3}$. The continuous lines are the exact values for the ordered and disordered energies of the infinite-size square lattice Potts model, see reference [Baxter(1973)].

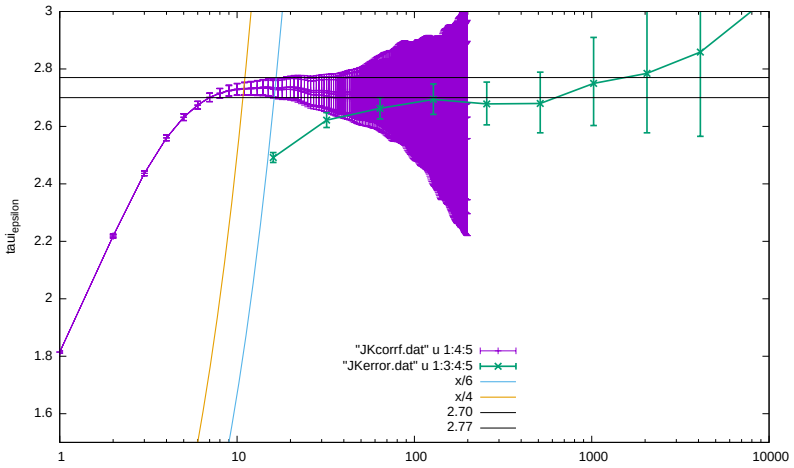

Figure 4.2: Integrated correlation time of the Potts model with parameters given in Exercise 6, heatbath sequential algorithm, as a function of the number of MC sweeps t. The violet points are the integral of the estimated correlation function (and their errors have been calculated with the jackknife method, see the function JKcorrfunction( ) in the script /common/corrfunc.py). The continuous lines are $t/4$ and $t/6$ (= the mentioned cutoff M/c). The green points are result of the Jackknife estimation, equation 4.8 (produced with the script /common/JKerror.py). For the green points, the abscissa indicates the jackknife block size. Notice that the for increasing block sizes, the number of blocks decreases and the (χ^2) error of the error (hence of $\tau_{\rm i}$) increases. Note also that the most precise estimation of $\tau_{\rm i}$ lies within the error bars of the less precise estimations (and not vice-versa), even for large block sizes.

5 Finite-Size Scaling

Reminder of critical phenomena formulae.

$$m \sim t^{\beta} G(r) \sim r^{-d+2-\eta} \exp[-r/\xi] \tag{5.1}$$
$$\xi \sim t^{-\nu} \tag{5.2}$$
$$\chi = -\partial_h \langle m \rangle = N(\langle m^2 \rangle - \langle m \rangle^2) \sim t^{-\gamma} \tag{5.3}$$
$$C = \partial_T \langle \epsilon \rangle = \beta^2 N(\langle \epsilon^2 \rangle - \langle \epsilon \rangle^2) \sim t^{-\alpha} \tag{5.4}$$
$$2 - \eta = \gamma/\nu \tag{5.5}$$
$$d\nu = 2\beta + \gamma \tag{5.6}$$

Deep (or bulk) and FSS regimes. Let us frame the discussion in a magnetic system. The magnetic susceptibility χ is an intensive quantity (c.f. eq. 5.3), since the fluctuations of the magnetization are of order $\sim N^{-1/2}$ (idem for the specifi heat and the energy). For a finite-size system of linear size L, this happens only in the so called *deep (or bulk)* ferromagnetic or paramagnetic regimes, such that $L \gg \xi(\beta, h)$, where ξ is the correlation length. In this circumstance, one can divide the system into boxes of linear size, R, $L \gg R \gg \xi$, and apply the central limit theorem to the block average of the magnetization inside different R-blocks: since they are independent, their variance are of order $\sigma_m^2 = n_b^{-1}\sigma^2$ where $n_b = \lfloor N/b$ is the number of blocks and σ is the standard deviation of the single block magnetization (of order 0 in N), and hence $\chi \sim \mathcal{O}[1]$. It follows that, in the bulk regimes (at fixed size, for large enough L, χ and the rest of the intensive quantities do not depend on L, see figure 5.1). In the deep ferromagnetic regime, the magnetization, hence, is $\mathcal{O}[1]$, while in the deep paramagnetic regime, it is of $\mathcal{O}[N^{-1/2}]$.

The opposite case, when $L \sim \xi$, is called *Finite Size Scaling (FSS) regime*. The central limit theorem cannot be applied any longer, the system is entirely correlated, σ_m^2 does not decrease as N^{-1}. As a consequence, χ is not intensive but it increases with a certain power of N, larger than zero, (see figure 5.1). At the critical point, the correlation length diverges and the bulk regimes cannot be reached even by the infinite-size system: indeed, the susceptibility is no longer of order N^0: it diverges for $N \to \infty$.

The bulk and FSS regimes are distinguished by the condition $\xi \ll L$. The FSS ansatz consists in supposing that the FSS behavior of an L-sized system in a state at β, h is uniquely determined by the ratio $L/\xi(\beta, h)$.

5.1 The Finite Size Scaling ansatz

Supposing an observable diverging at the critical point with the exponent x: $\langle \mathcal{O} \rangle_\infty \sim |t|^{-x}$, where $t = |\beta - \beta_c|/\beta_c$. The Finite Size Scaling (FSS) ansatz for its behavior in a finite lattice of linear size L is:

$$\langle \mathcal{O} \rangle_L(\beta) = L^{x/\nu} f_{\mathcal{O}}(L/\xi_\infty(t)) + \text{ corrections to scaling} \tag{5.7}$$

where ξ_∞ is the correlation length in the thermodynamic limit. In other words:

$$\langle \mathcal{O} \rangle_L(\beta) = L^{x/\nu} \left[\tilde{f}_{\mathcal{O}}(tL^{1/\nu}) + L^{-w} h_{\mathcal{O}}(tL^{1/\nu}) + \cdots \right] \tag{5.8}$$

The FSS hypothesis can be equivalently written (check!):

$$\langle \mathcal{O} \rangle_L(\beta) = L^{x/\nu} \tilde{f}_{\mathcal{O}}(tL^{1/\nu}) + \cdots \qquad \tilde{f}_{\mathcal{O}}(w) = f_{\mathcal{O}}(w^\nu) \tag{5.9}$$
$$\frac{\langle \mathcal{O} \rangle_L(\beta)}{\langle \mathcal{O} \rangle_\infty(\beta)} = \tilde{\tilde{f}}_{\mathcal{O}}(L/\xi_\infty(t)) + \cdots \qquad \tilde{\tilde{f}}_{\mathcal{O}}(w) \propto w^x \tilde{f}_{\mathcal{O}}(w). \tag{5.10}$$

Given the form of the divergence of $\mathcal{O}$, it is $f_{\mathcal{O}}(w) \to w^{-x/\nu}$, or $\tilde{f}_{\mathcal{O}}(w) \to w^{-x}$ for $w \to \infty$. (Figure 5.1 illustrates such a scaling for the susceptibility of the 2D Ising model, for which $\nu = 1/2$ and $\gamma = 7/4$).

There are operators, as the specific heat, for which different forms of scaling apply (see figure 5.2):

$$C_L(\beta) = \ln(L) \tilde{f}_C(tL^{1/\nu}) + \text{ corrections to scaling} \tag{5.11}$$

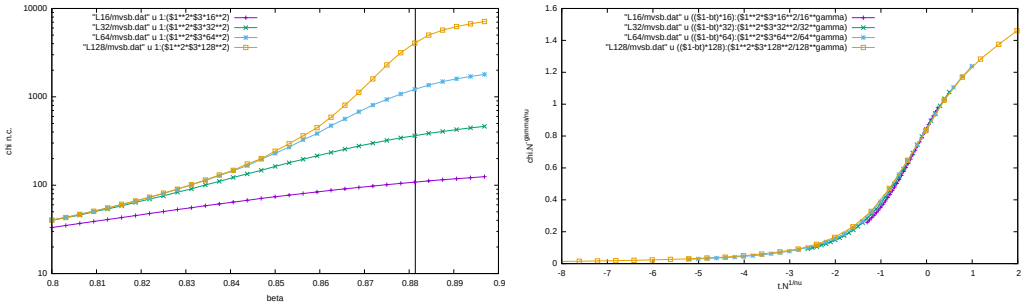

Figure 5.1: Left: Non-connected susceptibility, $\chi_{\rm nc} = N\langle m^2\rangle\beta^2$ vs. β for the 2D Ising model with periodic boundary conditions and several linear sizes, $L = 2^i$, $i = 4, \ldots, 7$. Right: $\chi_{\rm nc} L^{-7/4}$ vs tL. The vertical line signals the Onsager exact critical temperature. The codes and simulation data can be found in /PottsCUDA/graphPT_FSS/simulation1/. The observables (susceptibility, specific heat, binder cumulant...) have been produced with the python scripts in /common/dataAnalysis/.

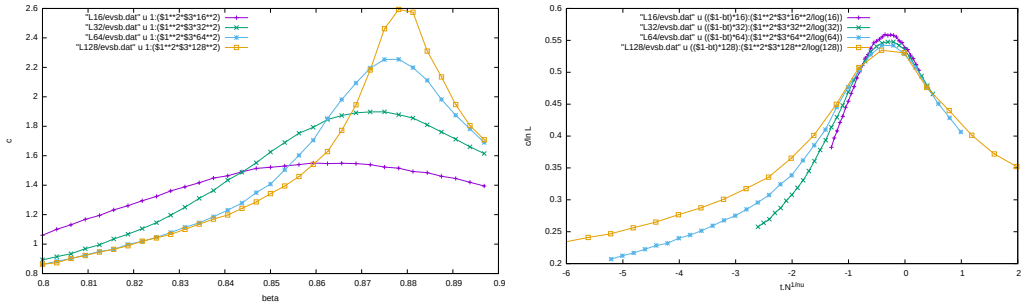

Figure 5.2: Idem as in figure 5.1 for the specific heat. The scaling is $c/\ln L$ vs tL.

the scaling function presents a maximum at the *apparent critical (inverse) temperature*:

$$\beta_{{\rm c},L} = \beta_{\rm c}(1 + w^* L^{-1/\nu}) \tag{5.12}$$

where w^* is such that $\tilde{f}'(w^*) = 0$.

5.2 Origin of FSS

The FSS ansatz can be actually proved using real-space renormalization group techniques. Under a real space renormalization group transformations, in the presence of large correlation lengths, under a scale transformation, $L \to bL$, the free energy density of the scaled and the original system present the scaling:

$$f_{bL}(t, h) = b^{-d} f_L(b^{1/\nu} t, b^\lambda h) \tag{5.13}$$

$$\lambda = (d + 2 - \eta)/2 \tag{5.14}$$

as can bee seen by a decimation argument, *à la* Kadanoff, or using phenomenological scaling arguments (see [Chaikin and Lubensky(2000)]). Based on this scaling form, and proposing a particular real-space renormalization group transformation, one can deduce the FSS ansatz, including corrections to scaling, the scaling of the order parameter probability density, equation 5.27, and the alternative form for the specific heat (see [Suzuki(1977), Pelissetto and Vicari(2002), Amit and Martín-Mayor(2005)]).

5.3 Determination of critical quantities with FSS

Scaling of the correlation length. We consider a lattice proxy for the (exponential) correlation length, $\xi_{\rm e} = \lim_{r\to\infty}(-r)/\ln(G(r))$, the finite-lattice correlation length:

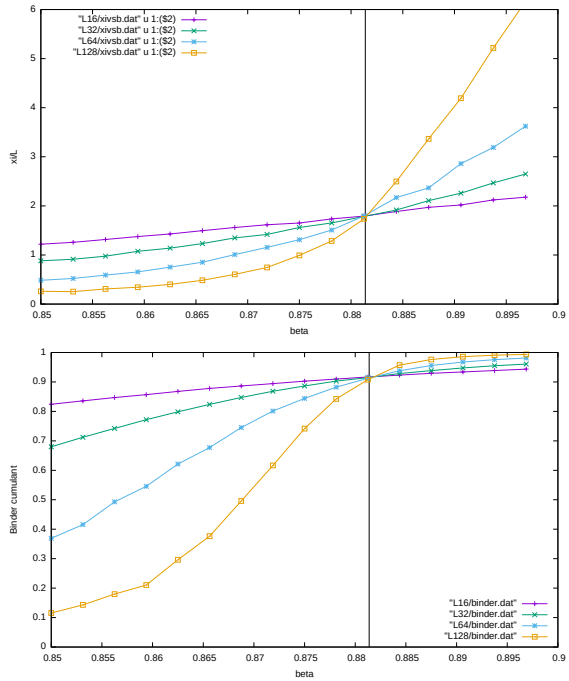

Figure 5.3: Left: ξ/L vs. β for the same system of figure 5.1. The vertical line indicates the Onsager inverse temperature $\beta_2 = \ln(1+2^{1/2})$. Right: Binder cumulant, g_4.

$$\xi_L = \frac{1}{2\sin(|\mathbf{k}_0|)}\left[\frac{s(\mathbf{0})}{s(\mathbf{k}_0)} - 1\right]^{1/2} \tag{5.15}$$

where $\mathbf{k}_0$ is (one of the) the minimum momentum(a) of the lattice, and s is the discrete Fourier transform of the correlation function G. The continuum limit of ξ_L, in the deep paramagnetic regime $tL^{1/\nu} \ll -1$, is the second-moment correlation length:

$$\xi^{(2)} = -s(\mathbf{k})^{-1}\partial_{k^2} s(\mathbf{k})|_{\mathbf{k}=\mathbf{0}} \tag{5.16}$$

which is of the same order of ξ_{e}, and coincides with the correlation length of a free scalar theory.

In the FSS regime, the quantity ξ_L obeys a scaling (see Eq. 5.8):

$$\xi_L(\beta) = L\tilde{f}_\xi(tL^{1/\nu})\left[1 + L^{-w}h_{\mathcal{O}}(tL^{1/\nu}) + \cdots\right] \tag{5.17}$$

(see [Amit and Martín-Mayor(2005)] for a deeper discussion on this point). See an illustration of the scaling of ξ_L for the Ising model in Figure 5.3.

The quotient method: eliminating the temperature Using Eqs. 5.17 and 5.8, one can write:

$$\langle\mathcal{O}\rangle_L(\beta) = L^{x/\nu}\left[\hat{f}_{\mathcal{O}}(\xi_L(t)/L) + L^{-w}\hat{h}_{\mathcal{O}}(\xi_L(t)/L) + \cdots\right] \tag{5.18}$$

this formulation poses several advantages (the non-necessity of knowing the critical point, the reduction of statistical errors, and the lower corrections to scaling).

The quotient method for the determination of critical temperatures and exponents consist in performing MC simulations at different pairs of lattice size values. Given a pair L_1, L_2, one defines the quotient:

$$\mathcal{Q}_{\mathcal{O}}(\beta, L_1, L_2) = \frac{\langle\mathcal{O}\rangle_{L_2}(\beta)}{\langle\mathcal{O}\rangle_{L_1}(\beta)} \tag{5.19}$$

One now defines $L_2 = nL_1$ being n an integer, and defines the temperature $\beta_n(L_1)$ at which $\mathcal{Q}_\xi(\beta_n(L_1), L_1, nL_1) = n$. For a different observable, the quotient

$$\mathcal{Q}_{\mathcal{O}}(\beta, L_1, nL_1) = s^{x/\nu} \frac{\left[\hat{f}_{\mathcal{O}}(\xi_{nL_1}/(nL_1)) + (nL_1)^{-w}\hat{h}_{\mathcal{O}}(\xi_{nL_1}/(nL_1)) + \cdots\right]}{\left[\hat{f}_{\mathcal{O}}(\xi_{L_1}/L_1) + L_1^{-w}\hat{h}_{\mathcal{O}}(\xi_{L_1}/L_1) + \cdots\right]} \tag{5.20}$$

evaluated at the temperature $\beta_n(L_1)$ becomes:

$$\mathcal{Q}_{\mathcal{O}}(\beta_n(L_1), L_1, nL_1) = n^{x/\nu} + AL_1^{-w} \tag{5.21}$$

being A a constant.

There are typically two independent critical exponents, say ν and η. The may be determined applying the quotient method to $\langle M^2\rangle$ and $\partial_\beta \xi$:

$$x_M^2 = (2 - \eta - d)\nu \qquad x_{\partial_\beta \xi} = 1 + \nu \tag{5.22}$$

The dimensionless quantities X (as ξ or g_4) scale in a different way:

$$X(\beta(L_1)) = X^* + A_X L_1^{-w} \tag{5.23}$$

while the inverse temperatures convergence to the critical point is:

$$\beta(L_1) - \beta_c \propto \frac{1 - n^{-w}}{n^{1/\nu} - 1} L^{-w-1/\nu} \tag{5.24}$$

The interested reader is invited to try to work out equations 5.21, 5.24 and 5.23 using Eq. 5.17 and supposing that the scaling functions are smooth.

Example 7. *Use the quotient method to estimate the exponent γ of the 2D Ising model in the square lattice, using the data and the scripts generated in the folder (*/PottsCUDA/graphPT_FSS/simulation2/*). For simplicity, one can suppose to know a priori the true value of the exponent $\nu = 1/2$ (but to ignore, of course, the critical temperature). Given the quality of the data, one should obtain something compatible with $\gamma = 1.750(5)$.*

Nightingale's phenomenological renormalization group. Suppose one starts from a system with size L_1 at a temperature β_1, and computes the inverse temperature, β_2, of a system of size $L_2 = nL_1$ such that $\xi_{L_1}(\beta_1)/L_1 = \xi_{L_2}(\beta_2)/L_2$. One then iterates this procedure, starting from L_i, β_i and computing β_{i+1} at $L_{i+1} = nL_i$ (or, alternatively, $L_{i+1} = L_i + 1$). The solutions will satisfy, neglecting corrections to scaling, $(\beta_{i+1} - \beta_c)/(\beta_i - \beta_c) = n^{-1/\nu_i}$. The fixed point of such an equation for large i is $\beta_i = \beta_c$, $\nu_i = \nu$

The Binder cumulant. In some circumstances it is convenient to use the *Binder cumulant* as a *scaling variable*, i. e., in the place of ξ_L/L:

$$g_4 = \frac{3}{2} - \frac{1}{2}\frac{\langle m^4\rangle}{\langle m^2\rangle^2} \tag{5.25}$$

It is immediate to check that g_4 converges to one and zero in the deep paramagnetic and ferromagnetic phases respectively [7].

The fact that it is a scaling variable comes from the extension of the FSS ansatz to the m probability distribution. For $N \gg \xi_\infty$ it is (deep ferromagnetic regime):

$$p(m,t) = \frac{N}{2(2\pi)^{1/2}\chi_t}\left[e^{-(m-|\mu_t|)^2 N/\chi_t} + e^{-(m+|\mu_t|)^2 N/\chi_t}\right] \tag{5.26}$$

in the FSS region the FSS ansatz is:

$$p_N(m,t) = L^{\beta/\nu} g(L/\xi, mL^{\beta/\nu}) = L^{\beta/\nu}\tilde{g}(tL^{1/\nu}, mL^{\beta/\nu}) \tag{5.27}$$

so that, for $t = 0$, the expectation value of $\langle m^4\rangle/\langle m^2\rangle^2$ (check it! by performing the average $\langle\cdot\rangle$ using an integration over the variable $\tilde{m} = L^{\beta/\nu}m$) does not depend on L.

Figure 5.3 shows an illustration of the scale invariance of the Binder cumulant in the Ising model.

[7] Recall that the 2-nd 4-th moments of a Gaussian distribution are: $\mu^2 + \sigma^2$ and $\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$, respectively

6 Elements of Markov-Chain Monte Carlo in Bayesian inference

6.1 Brief reminders

Bayesian estimators. We remind Bayes theorem:

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\pi(x)} \tag{6.1}$$

where θ are the hypothesis, x are the data, $f(\theta|x)$ is the posterior probability, $f(x|\theta) = \mathbb{L}(\theta; x)$ is the data likelihood probability, $\pi(\theta)$ is the prior probability of hypothesis θ and $\pi(x) = \sum_\theta f(x|\theta)\pi(\theta)$ is the marginal likelihood or the evidence.

Given the data, a *Bayesian estimator* for the hypothesis, $\hat{\theta}$, is a value of the hypothesis minimizing the expectation $\langle R(\theta, \theta')\rangle_{f(\theta|x)}$ over the posterior of a given function R called Bayes risk. The Bayesian estimator corresponding to the mean square error as the Bayes risk is the average over the posterior: $\hat{\theta}(x) = \sum_\theta \theta f(\theta|x)$.

An alternative estimation is the Maximum A Posteriori (MAP) estimator, or $\hat{\theta} = \arg\max_\theta f(\theta|x)$.

The Expectation-Maximization algorithm is a standard iterative method to provide a Maximum likelihood or a MAP Bayesian estimator. When the direct maximization over the posterior function is not generally possible, but it may become feasible when additional, or *missing values*, $\mathbf{z}$ are known. The corresponding *complete likelihood* is

$$f_\mathrm{c}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta})\, \pi(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) \tag{6.2}$$

$$\mathbb{L}_\mathrm{c}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \mathbb{L}(\boldsymbol{\theta}; \mathbf{x}) + \ln \pi(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) \tag{6.3}$$

$$f(\mathbf{x}|\boldsymbol{\theta}) = \sum_\mathbf{z} f_\mathrm{c}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \tag{6.4}$$

The EM algorithm consists in (randomly) initialising $\boldsymbol{\theta}^{(0)}$ and for $t = 0, \ldots, t_\mathrm{m}$, repeating the following two steps:

- Expectation. One computes the expectation value of the complete log-likelihood with respect to the values of $\boldsymbol{\theta}$ at the precedent step:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) = \langle \ln [f_\mathrm{c}(\mathbf{x}, \cdot\,|\boldsymbol{\theta})]\rangle_{\pi(\cdot|\mathbf{x}, \boldsymbol{\theta}^{(t-1)})} \tag{6.5}$$

- Maximization. At the current iteration, the hypothesis are set to the value maximising Q:

$$\boldsymbol{\theta}^{(t)} \equiv \arg\max_\theta Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) \tag{6.6}$$

The correctness of the algorithm can be seen by taking the expectation value of (6.3) with respect to π, $\sum_z \cdots \pi(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t-1)})$, hence using Gibbs inequality,[8] and noticing that if $\ln f(\mathbf{x}|\boldsymbol{\theta}) - \ln f(\mathbf{x}, \boldsymbol{\theta}^{(t-1)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$, so that every step of the EM algorithm maximises *the original likelihood* $f(\mathbf{x}|\boldsymbol{\theta})$. There is no guarantee, however, that this algorithm converges to the absolute maximum of $f(\mathbf{x}|\boldsymbol{\theta})$.

6.2 Algorithms for inferring in mixtures of probability distributions

Mixtures of probability distributions. Consider n data $\mathbf{x} = \{x_i\}_{i=1}^n$ generated with a mixture of K probability distributions, each data generated from the distribution with parameters θ_j with probability p_j, being $\sum_{j=1}^K p_j = 1$, $\mathbf{p} = \{p_j\}_{j=1}^K$, $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^K$. The likelihood can be written as:

$$\mathbb{L}(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^n \left[\sum_{j=1}^k p_j f(x_i|\theta_j)\right]. \tag{6.7}$$

[8] $\sum_i p_i \ln p_i \leq \sum_i p_i \ln q_i$, where p and q are probability distributions, the equality is obtained if $p = q$.

We will consider the simple case of a mixtures of Gaussians: $\mathcal{N}(\theta_i, 1)$, i.e., $f(x_i|\theta_j) = (2\pi)^{-1/2} \exp(-(x_i - \theta_j)^2/2)$. From now on we will consider that the hypothesis to be inferred from the data $\mathbf{x}$ are the average of the distribution and the prior probability corresponding to j, $\theta_j = (\mu_j, p_j)$.

Although the likelihood (6.7) can be evaluated in $\mathcal{O}[Kn]$, there are K^n terms in the sum, so that the direct evaluation of Bayesian estimators is not feasible. A solution is the application of the EM algorithm.

EM algorithm for the Gaussian mixture. Let us define $\mathcal{Z} = \{1, \dots, K\}^{\otimes n}$. The likelihood (6.7) can be written as:

$$\mathbb{L}(\boldsymbol{\theta}; \mathbf{x}) = \sum_{\{z_i\} \in \mathcal{Z}} \prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i}) \tag{6.8}$$

according to (6.4), it is:

$$\ln f_{\mathrm{c}}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[p_{z_i} f(x_i|\theta_{z_i}) \right]. \tag{6.9}$$

E step: the expectation value $\langle \ln \mathbb{L}_{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) \rangle_{\pi(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})}$ of the EM algorithm is given by

$$\sum_{i=1}^{n} \sum_{z_i=1}^{K} \left[\ln p_{z_i} + \ln f(x_i|\theta_{z_i}) \right] \pi_1(z_i|x_i, \theta_{z_i}) \tag{6.10}$$

when the conditional probability of $\mathbf{z}$ has been factorized, $\pi(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = \prod_i \pi_1(z_i|x_i, \theta_{z_i})$. We now define the shorthand: $\pi_1(j|x_i, \theta_j) = w_{ij}$, $i = 1, \dots, n$, $j = 1, \dots, K$, in terms of which the expectation of the EM algorithm is:

$$\langle \ln \mathbb{L}_{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) \rangle_{\pi(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})} = \sum_{i=1}^{n} \sum_{z_i=1}^{K} w_{ij} \left[\ln p_{z_i} + \ln f(x_i|\theta_{z_i}) \right] \tag{6.11}$$

What about the w's? They are obtained from (6.2) and from the definition of the complete likelihood in the mixture context, Eq. 6.9 (notice that it also responds to Bayes rule):

$$w_1(j|x_i, \theta_j) = \frac{p_j^{(t-1)} f(x_i|\theta_j^{(t-1)})}{\sum_{m=1}^{K} p_m^{(t-1)} f(x_i|\theta_m^{(t-1)})} \tag{6.12}$$

M step: maximising with respect to θ_j, i.e., with respect to μ_j and to p_j results in the equations:

$$p_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ij} \tag{6.13}$$

$$\theta_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}}. \tag{6.14}$$

Intuitive motivation for the EM algorithm A Bayes estimator for the $\boldsymbol{\theta}^{(t)}$ given the $\mathbf{x}$ and the $\mathbf{z}$ (M step) is feasible (since $\ln f_c$ has in this case a linear (no longer K^n) number of terms)! An estimator of $\pi(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$ is feasible from $(\mathbf{x}, \boldsymbol{\theta}^{(t-1)})$ (E step, see Eq. 6.10).

Metropolis algorithm An alternative strategy is to sample $\boldsymbol{\theta}$ from a Markov-Chain MC with the Metropolis algorithm, such that the target distribution is $f(\mathbf{x}|\boldsymbol{\theta})$. For the case of the Gaussian mixture, the algorithm reads: one choses random initial conditions $\boldsymbol{\theta}^{(0)}$, then:

1. at the t-th iteration, one performs an attempt $\tilde{\mu}_j = \mu_j^{(t)} + \xi$ where $\xi \sim \mathcal{N}(0, \eta)$ being η a parameter (to be optimized). The constraint parameters $p_j^{(t)}$ can be updated as $\ln \tilde{p}_j = \ln p_j^{(t-1)} + \zeta$ being $\zeta \sim \mathcal{N}(0, \eta^2)$ (see Exercise 2), eventually evaluating this trial with a further prior probability $\pi(\tilde{\boldsymbol{\theta}})$.

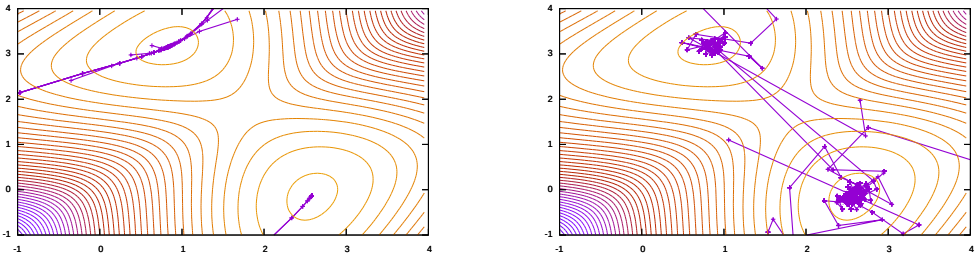

Figure 6.1: Successive values of the parameters $\theta_{1,2}$ found by 16 realizations of the EM algorithm (left) and of the Metropolis algorithm (right), for a mixture of 2 Gaussians, and a likelihood (represented by contour iso-likelihood lines) corresponding to $n = 500$, points (see the details, the algorithm scripts and the data in /BayesianMixture/Metropolis/). The probabilities are supposed to be known. The EM algorithm converges to each of the two maxima, depending on the initial conditions (only the absolute maximum corresponds to (but do not coincide with) the true parameters used to generate the data, $\mu_1 = 2.5$, $\mu_2 = 0$). The Metropolis algorithm samples both maxima in every run but, asymptotically, the absolute maxima is infinitely more sampled.

2. Application of the Metropolis rule: with probability: $r = f(\mathbf{x}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})/[f(\mathbf{x}|\boldsymbol{\theta}^{(t)})\pi(\boldsymbol{\theta}^{(t)})]$ accept the trial, $\boldsymbol{\theta}^{(t+1)} \equiv \tilde{\boldsymbol{\theta}}$; $t++$, go to 1.

Example 8. *One can implement the Metropolis algorithm for a mixture of two Gaussians (as done in the function* MetropolisThetaAndProbs() *inside the script* gaussianMixture.py *that you can find in the folder* /BayesianMixture/Metropolis/*). The Metropolis algorithm manages to find accurate estimators for* $\mathbf{p}$ *and* $\boldsymbol{\theta}$ *(as illustrated for a sample of* $n = 500$ *data in the mentioned folder). When fixing the probability* p_1 *to the true one, the EM algorithm for* $\boldsymbol{\theta}$ *(see the function* EMalgorithm()*, eventually commenting out the line searching in the* p*'s converges) finds alternatively the true or the "fake" maximum of the likelihood, depending on the initial conditions, while the Metropolis algorithm for a sufficiently large number of steps* n *and/or different runs with different initial conditions, samples with arbitrarily large probability the maximum corresponding to the "true" value of the parameters* $\boldsymbol{\theta}$ *(see figure 6.1). Is there an optimal value of the 'trial amplitude' for the changes in* $\boldsymbol{\theta}$ *(the variable* theta*, in the script)?*

Exercise 21. Gibbs sampling (heatbath) algorithm.

The heatbath method explained in section 3.1 can be mutatis mutandis applied to this case. We exemplify it for a simple (with fixed variance) Gaussian mixture. The interested reader is invited to work out the following algorithm, recovering and implementing it, eventually consulting [Marin et al.(2005)Marin, Mengersen, and Robert]. One first choses an initial condition, $\boldsymbol{\theta}^{(0)}$*. Afterwards:*

1. *one generates* z_i $i = 1, \ldots, n$ *from its probability distribution:*

$$\mathbb{P}(z_i = j) \propto p_j f(x_i, \theta_j) \tag{6.15}$$

2. *one computes average magnetization and average* x*'s:* $n_j^{(t)} = \sum_i \delta_{z_i,j}$, $s_j^{(t)} = \sum_i \delta_{z_i,j} x_i$, *and they are used to compute novel probabilities and averages:*

$$\mu_j^{(t)} \sim \mathcal{N}\left(\frac{\lambda\delta + s_j^{(t)}}{\lambda + n_j^{(t)}}, \frac{1}{\lambda + n_j^{(t)}}\right) \tag{6.16}$$

where the prior on the choose of μ_j *is* $\mathcal{N}(\delta, 1/\lambda)$, $\delta \in \mathbb{R}$, $\lambda > 0$.

Exercise 22. Reversible Jump MC (Green (1995)) *When dealing with a mixture of an unknown number of Gaussians* M*, it is possible to perform a Metropolis MC in which also* M*, besides*

$\mathbf{p}$ *and* $\boldsymbol{\theta}$*, is inferred. The algorithm, called Reversible Jump MC, proposes trials between configurations* ($\mathbf{p}$,$\boldsymbol{\theta}$) *belonging to spaces* $\mathcal{E}_M$ *with with different* M*'s. The balance condition to be satisfied requires the transition from one space to the other to be a bijection, and this amounts in a nontrivial factor in the acceptance probability, depending on the Jacobian of the transformation* $\mathcal{E}_M \to \mathcal{E}_{M'}$*. The interested reader is invited to consult the details in reference [Marin and Robert(2007)] and, eventually, to implement and test the Reversible Jump algorithm (perhaps using the machinery already present in* /BayesianMixture/Metropolis/*).*

7 Acknowledgements

I acknowledge Fabrizio Antenucci, Ludovica B. Romano, Francesco Di Renzo and Javier Rodríguez-Laguna for their advices, and Cristiano Viappiani for his support as Chair of the Doctoral Studies Committee in Physics of the University of Parma.

8 Bibliographic guide, possible completions and References

For the realization of the course, the lecture notes and the code examples, we have mainly followed: [Pelissetto(1993b)] for the initial static MC part; [Sokal(1997)],[Pelissetto(1993b)],[Bhat and Miller(2002)],[Janke(2008)] for the dynamic MC; [Berg(2004)] for the error estimation; [Frenkel and Smit(2001)] for the Metropolis algorithm in different ensembles; [Pelissetto and Vicari(2002)],[Amit and Martín-Mayor(2005)], [Chaikin and Lubensky(2000)] for the theory of finite-size scaling; [Marin et al.(2005)Marin, Mengersen, and Robert, Marin and Robert(2007)] for the Bayesian inference part (see also the more general references [Andrieu et al.(2003)Andrieu, De Freitas, Doucet, and Jordan, Rubinstein and Kroese(2011)]). A useful review reference treating scaling in first-order transitions is [Binder(1997)].

Should this course be repeated for an audience composed by physicists, one could add a section about Quantum Monte Carlo, with the Ising spin chain [Landau and Binder(2014)] or the path integral MC for the study of condensed Helium [Ceperley(1995)], as examples. A further possible proposal is the presentation of the reweighting (see references in [Amit and Martín-Mayor(2005)]) and the tethered MC methods ([Martin-Mayor et al.(2011)Martin-Mayor, Seoane, and Yllanes]). Furthermore, one could present a numerical analysis of the replica symmetry breaking transition of a finite-dimensional spin glass model, as an illustration of the utility of the Parallel Tempering algorithm ([Amit and Martín-Mayor(2005)]).

References

[Amit and Martín-Mayor(2005)] Amit, D. J. and V. Martín-Mayor, 2005: *Field Theory, the Renormalization Group, and Critical Phenomena: Graphs to Computers (3RD Edition)*. World Scientific Press, doi:10.1142/5715.

[Andrieu et al.(2003)Andrieu, De Freitas, Doucet, and Jordan] Andrieu, C., N. De Freitas, A. Doucet, and M. I. Jordan, 2003: An introduction to mcmc for machine learning. *Machine learning*, **50 (1-2)**, 5–43.

[Baxter(1973)] Baxter, R. J., 1973: Potts model at the critical temperature. *Journal of Physics C: Solid State Physics*, **6 (23)**, L445, URL http://stacks.iop.org/0022-3719/6/i=23/a=005.

[Berg(2004)] Berg, B. A., 2004: *Markov Chain Monte carlo simulations and their statistical analysis*. World Scientific.

[Bhat and Miller(2002)] Bhat, U. N. and G. K. Miller, 2002: *Elements of applied stochastic processes*. J. Wiley.

[Binder(1987)] Binder, K., 1987: Theory of first-order phase transitions. *Reports on Progress in Physics*, **50 (7)**, 783+, doi:10.1088/0034-4885/50/7/001, URL http://dx.doi.org/10.1088/0034-4885/50/7/001.

[Binder(1997)] Binder, K., 1997: Applications of monte carlo methods to statistical physics. *Reports on Progress in Physics*, **60 (5)**, 487, URL http://stacks.iop.org/0034-4885/60/i=5/a=001.

[Ceperley(1995)] Ceperley, D. M., 1995: Path integrals in the theory of condensed helium. *Rev. Mod. Phys.*, **67**, 279–355, doi:10.1103/RevModPhys.67.279, URL http://link.aps.org/doi/10.1103/RevModPhys.67.279.

[Chaikin and Lubensky(2000)] Chaikin, P. M. and T. C. Lubensky, 2000: *Principles of condensed matter physics*, Vol. 1. Cambridge Univ Press.

[Debenedetti(1996)] Debenedetti, P., 1996: *Metastable Liquids: Concepts and Principles*. Physical chemistry : science and engineering, Princeton University Press, URL http://books.google.it/books?id=tzvvsltE6Y8C.

[Edwards and Sokal(1988)] Edwards, R. G. and A. D. Sokal, 1988: Generalization of the fortuin-kasteleyn-swendsen-wang representation and monte carlo algorithm. *Phys. Rev. D*, **38**, 2009–2012, doi:10.1103/PhysRevD.38.2009, URL http://link.aps.org/doi/10.1103/PhysRevD.38.2009.

[Fortuin(1969)] Fortuin, C., 1969: Physica (utrecht) 57, 536 (1972); pw kasteleyn and cm fortuin. *J. Phys. Soc. Jpn. Suppl*, **26 (11)**.

[Frenkel and Smit(2001)] Frenkel, D. and B. Smit, 2001: *Understanding molecular simulation: from algorithms to applications*, Vol. 1. Academic press.

[Hansen and McDonald(1990)] Hansen, J.-P. and I. R. McDonald, 1990: *Theory of simple liquids*. Elsevier.

[Hohenberg and Halperin(1977)] Hohenberg, P. C. and B. I. Halperin, 1977: Theory of dynamic critical phenomena. *Rev. Mod. Phys.*, **49**, 435–479, doi:10.1103/RevModPhys.49.435, URL http://link.aps.org/doi/10.1103/RevModPhys.49.435.

[Ibáñez-Berganza(2016)] Ibáñez-Berganza, M., 2016: Introduction to monte carlo method in statistical physics. URL http://www.pr.infn.it/home/miguel.berganza/pages/teaching.html.

[Janke(2008)] Janke, W., 2008: *Monte Carlo Methods in Classical Statistical Physics*, 79–140. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-540-74686-7_4, URL http://dx.doi.org/10.1007/978-3-540-74686-7_4.

[Kolafa and Nezbeda(1994)] Kolafa, J. and I. Nezbeda, 1994: The lennard-jones fluid: An accurate analytic and theoretically-based equation of state. *Fluid Phase Equilibria*, **100**, 1–34.

[Landau et al.(2010)Landau, Lewis, Schuettler, Nussbaumer, Bittner, Neuhaus, and Janke] Landau, D., S. Lewis, H.-B. Schuettler, A. Nussbaumer, E. Bittner, T. Neuhaus, and W. Janke, 2010: Computer simulation studies in condensed matter physics xx, csp-2007 universality of the evaporation/condensation transition. *Physics Procedia*, **7**, 52 – 62, doi:http://dx.doi.org/10.1016/j.phpro.2010.09.044, URL http://www.sciencedirect.com/science/article/pii/S1875389210006590.

[Landau and Binder(2014)] Landau, D. P. and K. Binder, 2014: *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press.

[Marin and Robert(2007)] Marin, J. and C. Robert, 2007: *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer Texts in Statistics, Springer New York, URL https://books.google.it/books?id=5xwuouehKQoC.

[Marin et al.(2005)Marin, Mengersen, and Robert] Marin, J.-M., K. Mengersen, and C. P. Robert, 2005: Bayesian modelling and inference on mixtures of distributions. *Bayesian ThinkingModeling and Computation*, D. Dey and C. Rao, Eds., Elsevier, Handbook of Statistics, Vol. 25, 459 – 507, doi:http://dx.doi.org/10.1016/S0169-7161(05)25016-2, URL http://www.sciencedirect.com/science/article/pii/S0169716105250162.

[Marinari and Parisi(2004)] Marinari, E. and G. Parisi, 2004: Trattatello di probabilità. URL http://www.phys.uniroma1.it/DipWeb/web_disp/d3/dispense/marinari-parisi--prob.pdf.

[Martin-Mayor et al.(2011)Martin-Mayor, Seoane, and Yllanes] Martin-Mayor, V., B. Seoane, and D. Yllanes, 2011: Tethered monte carlo: Managing rugged free-energy landscapes with a helmholtz-potential formalism. *Journal of Statistical Physics*, **144 (3)**, 554–596, doi:10.1007/s10955-011-0261-4, URL http://dx.doi.org/10.1007/s10955-011-0261-4.

[Pelissetto(1993a)] Pelissetto, 1993a: *Elementary particles, quantum fields and statistical mechanics: lectures given at the Summer school in theoretical physics, Parma 1991-1993 : Seminario nazionale di fisica teorica*. Centro grafico dell'Universita di Parma, URL https://books.google.it/books?id=WX-YoAEACAAJ.

[Pelissetto(1993b)] Pelissetto, A., 1993b: Introduction to the monte carlo method. *Summer School in Theoretical Physics and Bonini, M. and Marchesini, G. and Onofri, E.*, URL http://www.fis.unipr.it/~direnzo/ModSim/pelissetto.pdf.

[Pelissetto and Vicari(2002)] Pelissetto, A. and E. Vicari, 2002: Critical phenomena and renormalization-group theory. *Physics Reports*, **368 (6)**, 549 – 727, doi:http://dx.doi.org/10.1016/S0370-1573(02)00219-3, URL http://www.sciencedirect.com/science/article/pii/S0370157302002193.

[Rubinstein and Kroese(2011)] Rubinstein, R. Y. and D. P. Kroese, 2011: *Simulation and the Monte Carlo method*, Vol. 707. John Wiley & Sons.

[Sokal(1997)] Sokal, A., 1997: *Functional Integration: Basics and Applications*, chap. Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms, 131–192. Springer US, Boston, MA, doi:10.1007/978-1-4899-0319-8_6, URL http://dx.doi.org/10.1007/978-1-4899-0319-8_6.

[Suzuki(1977)] Suzuki, M., 1977: Static and dynamic finite-size scaling theory based on the renormalization group approach. *Progress of Theoretical Physics*, **58 (4)**, 1142–1150, doi:10.1143/PTP.58.1142, URL http://ptp.oxfordjournals.org/content/58/4/1142.abstract, http://ptp.oxfordjournals.org/content/58/4/1142.full.pdf+html.

[Swendsen and Wang(1987)] Swendsen, R. H. and J.-S. Wang, 1987: Nonuniversal critical dynamics in monte carlo simulations. *Phys. Rev. Lett.*, **58**, 86–88, doi:10.1103/PhysRevLett.58.86, URL http://link.aps.org/doi/10.1103/PhysRevLett.58.86.

[Weigel et al.(2002)Weigel, Janke, and Hu] Weigel, M., W. Janke, and C.-K. Hu, 2002: Random-cluster multihistogram sampling for the q -state potts model. *Phys. Rev. E*, **65**, 036 109, doi:10.1103/PhysRevE.65.036109, URL http://link.aps.org/doi/10.1103/PhysRevE.65.036109.

[Wolff(1989)] Wolff, U., 1989: Collective monte carlo updating for spin systems. *Phys. Rev. Lett.*, **62**, 361–364, doi:10.1103/PhysRevLett.62.361, URL http://link.aps.org/doi/10.1103/PhysRevLett.62.361.