Some notes for

"Numerical Methods for Physics"

Claudio Bonati

May 8, 2025

Contents

Introduction

> "No matter how powerful computers become, physicists will always want to study problems that are too difficult for the computers at hand." [1]

In these notes we discuss the topics covered in the following three modules of the "Numerical Methods for Physics" course:

- Introduction to Markov Chain Monte-Carlo and applications in statistical mechanics
- Application of Monte-Carlo methods to the study of the path-integral in quantum mechanics
- Path-integral simulations for quantum field theories

With respect to the material discussed in class (many) more details are present in these notes, mainly to investigate some technical points or to provide complete proofs whose analysis would take too much time, or would be at least partially off topic, during the lectures.

After introducing some general features of the Monte Carlo algorithms, in Part I we discuss quite in detail the approach known as Markov Chain Monte Carlo (MCMC), which is the Monte Carlo technique that is most commonly adopted in nontrivial applications. To put on firm ground the foundations of the MCMC method same basic facts about Markov chains are presented, together with the data analysis techniques needed to reliably estimate (functions of) average values in MCMC simulations, and to asses their statistical accuracy.

Statistical mechanics will be often used to motivate some of the requirements that a good Monte Carlo algorithm has to satisfy, and in Part II the MCMC technique is applied to the study of phase transitions in simple lattice systems. While virtually any problem in (equilibrium) statistical mechanics can be tackled by using Monte Carlo methods, there are several reasons to focus on phase transitions in classical lattice models of ferromagnets. From the algorithmic point of view these models are quite simple to investigate by Monte Carlo methods, and thus constitute an ideal testbed for the application of the techniques introduced in Part I. Given their extreme simplicity, one might expect these models to provide only some very general qualitative information of minor physical interest. This is however not the case for continuous phase transitions: the phenomenon of universality ensures that even the simplest models capture quantitative features (the universal ones) of real world continuous phase transitions. The peculiar behavior that emerges in a system close to a continuous phase transition also presents some challenges for the Monte Carlo method, whose computational efficiency typically decreases as the size of the system is increased (critical slowing down).

In Part III Monte Carlo methods are applied to study quantum mechanical systems, and in particular equilibrium quantum statistical mechanics. The starting point is the Euclidean path-integral technique, by which quantum thermal averages can be rewritten in a way which makes them amenable of being estimated by Monte Carlo methods. Indeed, once a regularization of the path-integral is introduced, the computation of quantum thermal averages becomes formally equivalent to the estimation of thermal averages in a one dimensional classical lattice system. Information on the energy spectrum of the quantum model can be obtained by studying correlators in the corresponding classical statistical system for different Euclidean time separations; using this fact it becomes clear that the process of removing the path-integral regulator is equivalent to the study

of critical phenomena in classical one dimensional systems. Although all the techniques introduced are valid for generic systems, the case of the one dimensional harmonic oscillator is often used to exemplify them in a simple setting in which analytical computations can also be performed.

In Part IV Monte Carlo methods are applied to the numerical investigation of some properties of quantum field theories. Although the general ideas are analogous to those already introduced in Part III, some more difficulties arises, that are discussed in the simplest setting, that of the free bosonic field. Numerical simulations of fermion fields are significantly more challenging than their bosonic counterparts, and some of the difficulties encountered can be easily understood. The fermionic case is used to motivate the introduction of the Hybrid Monte Carlo algorithm for the simulation of non-local actions. Quantum field theories are not only more difficult to simulate than elementary quantum mechanical systems, they also present a richer phenomenology. In order to present a glimpse of this phenomenology, we discuss several aspects of two dimensional lattice gauge theories, which are relatively easy to simulate and for which we have complete analytic control.

This course is thought to be attended in parallel with other courses, more focused on the physics of the systems under investigation, like, e. g., statistical mechanics and quantum field theory courses. For this reason a short summary of the main physical features is provided whenever a deeper physical understanding is needed, e. g., to decide which observable to measure, to plan the simulations or to interpret the numerical results.

The other natural possibility would be to attend this course when already acquainted with the physical side of the problem. It is quite obvious that there are positive aspects also in this second possibility, however one should not underestimate the physical insight that can be gained by numerically simulating a system. Indeed, sometimes, the mathematical subtleties that in a theoretical setting could seem futilely abstruse, or maybe even useless, become quite reasonable after directly verifying what happens by neglecting them. Spontaneous symmetry breaking (especially in gauge field theories) is a typical example of a phenomenon which require some care to be investigated, both from the mathematical point of view and in numerical simulations.

All the numerical results presented have been obtained by using the codes publicly available at

https://github.com/claudio-bonati/NumericalMethods/

and the run times reported refer to a single core Intel(R) Xeon(R) Gold 5218 CPU 2.30GHz, with the code compiled using the GCC compiler (version 9.4.0).

To report typos, oversights, inaccuracies, errors or whatever else, please write to

claudio.bonati@unipi.it

List of abbreviations

b. c.: boundary conditions
FSS: finite size scaling
GCD: greatest common divisor
iid: independent and identical distributed
LGT: lattice gauge theory
MC: Monte Carlo
MCMC: Markov Chain Monte Carlo
pdf: probability distribution function
QCD: quantum chromodynamics
QFT: quantum field theory
QM: quantum mechanics
RG: renormalization group
SSB: spontaneous symmetry breaking

Part I

The Markov Chain Monte-Carlo method

Chapter 1

Basics of Monte Carlo methods

Monte Carlo methods constitute a class of numerical methods which use a stochastic approach to evaluate expressions of the form

$$\langle F \rangle = \int_C F(x)p(x)\mathrm{d}x \ , \tag{1.0.1}$$

where $\mathrm{d}x$ denotes a measure on the set C, $p(x)$ is a probability density function on C (pdf for short), thus

$$p(x) \geq 0 \ , \quad \int_C p(x)\mathrm{d}x = 1 \ , \tag{1.0.2}$$

and $F(x)$ is a function of x. In some cases the quantity to be investigated already has a natural probabilistic interpretation (this is typically the case in statistical mechanics), in other cases some work is needed to rewrite it in the form Eq. (1.0.1), selecting an appropriate ensemble C, an appropriate pdf $p(x)$ and an appropriate function $F(x)$.

Several approaches can be used to evaluate the right hand side of Eq. (1.0.1), and this is the reason for the plural in "Monte Carlo methods": in some cases it is possible to directly sample the pdf, in most of the cases this is however not numerically feasible, and the less direct Markov Chain Monte Carlo approach has to be used; also in this case there is however much freedom on how to construct the appropriate Markov Chain.

Whatever method is used, in the end all Monte Carlo approaches produce "in some way" a sample of N draws $x_1, \ldots, x_N$ from the pdf $p(x)$, from which we get the quantities $F(x_1), \ldots, F(x_N)$, whose sample average $\overline{F}$ is an estimator of $\langle F \rangle$. The values x_i are always identically distributed but non necessarily independent, and a fundamental point is to determine the statistical uncertainty to be associated with $\overline{F}$.

1.1 Sample statistics

In this section we recall some basic facts about sample statistics that will be of fundamental importance in the following, considering only the case of independent and identically distributed (iid for short) samples $\{x_i\}_{i=1,\ldots,N}$. As usual we denote by $\langle F \rangle$ the average of F computed with respect to the pdf $p(x)$, and by $\overline{F}$ the sample average of the quantities $F_i = F(x_i)$. The overline will be used more generally to denote sample estimators.

It is simple to verify that the sample average

$$\overline{F} = \frac{1}{N}\sum_i F_i \tag{1.1.1}$$

is an unbiased estimator of $\langle F \rangle$, i. e., $\langle \overline{F} \rangle = \langle F \rangle$: since the draws x_is are sampled from the same

pdf $p(x)$ we have for each i

$$\langle F_i \rangle = \langle F(x_i) \rangle = \int F(x_i) p(x_i) \mathrm{d}x_i = \langle F \rangle \ , \tag{1.1.2}$$

and by linearity

$$\langle \overline{F} \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle F_i \rangle = \langle F \rangle \ . \tag{1.1.3}$$

To get an unbiased estimator of the variance $\sigma_F^2 = \langle F^2 \rangle - \langle F \rangle^2$ is only slightly more complicated: we have

$$\langle \overline{F^2} - \overline{F}^2 \rangle = \left\langle \frac{1}{N} \sum_i F_i^2 - \left(\frac{1}{N} \sum_i F_i \right)^2 \right\rangle = \frac{1}{N} \sum_i \langle F_i^2 \rangle - \frac{1}{N^2} \sum_{ij} \langle F_i F_j \rangle \ . \tag{1.1.4}$$

Moreover, since $F_i = F(x_i)$ and the x_is are identically distributed, we have $\langle F_i^2 \rangle = \langle F^2 \rangle$, and since the x_is are also independent of each other

$$\langle F_i F_j \rangle = \begin{cases} \langle F^2 \rangle & \text{if } i = j \\ \langle F \rangle^2 & \text{if } i \neq j \end{cases} \ , \tag{1.1.5}$$

hence

$$\begin{aligned} \langle \overline{F^2} - \overline{F}^2 \rangle &= \langle F^2 \rangle - \frac{1}{N^2} \left[N(N-1) \langle F \rangle^2 + N \langle F^2 \rangle \right] = \\ &= \frac{N-1}{N} \left(\langle F^2 \rangle - \langle F \rangle^2 \right) = \frac{N-1}{N} \sigma_F^2 \ . \end{aligned} \tag{1.1.6}$$

An unbiased estimator of σ_F^2 is thus

$$\overline{\sigma}_F^2 = \frac{N}{N-1} \left(\overline{F^2} - \overline{F}^2 \right) \ , \tag{1.1.7}$$

and the bias correcting factor $\frac{N}{N-1}$ is obviously irrelevant in the large sample limit $N \gg 1$.

We can now compute the variance of the stochastic variable defined by the sample average $\overline{F}$. We have (using once again the fact that the x_i are iid)

$$\begin{aligned} \sigma_{\overline{F}}^2 = \langle \overline{F}^2 \rangle - \langle \overline{F} \rangle^2 = \frac{1}{N^2} \left\langle \left(\sum_i F_i \right)^2 \right\rangle - \langle F \rangle^2 = \\ = \frac{1}{N^2} \left[N \langle F^2 \rangle + N(N-1) \langle F \rangle^2 \right] - \langle F \rangle^2 = \frac{1}{N} \left[\langle F^2 \rangle - \langle F \rangle^2 \right] = \frac{1}{N} \sigma_F^2 \ . \end{aligned} \tag{1.1.8}$$

Using the sample estimator of the variance $\overline{\sigma}_F^2$ we immediately obtain the sample estimator of the variance of the sample average:

$$\overline{\sigma}_{\overline{F}}^2 = \frac{1}{N-1} \left(\overline{F^2} - \overline{F}^2 \right) \ . \tag{1.1.9}$$

To appreciate the importance of these results it is useful to recall a simple fact known as Chebyshev's inequality: if X is random variable with finite variance σ_X^2 and average $\langle X \rangle$, the probability of observing a value of X whose distance from $\langle X \rangle$ is larger than $k\sigma_X$ is smaller than $1/k^2$:

$$P(|X - \langle X \rangle| \geq k\sigma_X) \leq \frac{1}{k^2} \tag{1.1.10}$$

From the definition of variance and the positivity of $(X - \langle X \rangle)^2$ we have indeed

$$\begin{aligned} \sigma_X^2 = \int (X - \langle X \rangle)^2 p(X) \mathrm{d}X \geq \int_{|X - \langle X \rangle| \geq k\sigma_X} (X - \langle X \rangle)^2 p(X) \mathrm{d}X \\ \geq k^2 \sigma_X^2 \int_{|X - \langle X \rangle| \geq k\sigma_X} p(X) \mathrm{d}X = k^2 \sigma_X^2 P(|X - \langle X \rangle| \geq k\sigma_X) \ , \end{aligned} \tag{1.1.11}$$

from which Chebyshev's inequality follows. The meaning of the Chebyshev's inequality is that the standard deviation σ_X is a measure of how much a probability distribution is peaked around $\langle X\rangle$. From (1.1.8) we can thus conclude that in the large sample limit $N \to \infty$ it is very unlikely to find a value of the sample average which is far from the true average. This result is nothing but the law of large numbers in its weak form: for any $\epsilon > 0$ the probability of finding a value $\overline{X}$ which differs from $\langle X\rangle$ by more than ϵ goes to zero in the large sample limit $N \to \infty$:

$$\lim_{N\to\infty} P(|\overline{X} - \langle X\rangle| > \epsilon) = 0 \ . \tag{1.1.12}$$

The proof of this result is an immediate consequence of (1.1.8) and Chebyshev's inequality if σ_X^2 is finite, but the result is true also without this assumption (see e. g. [2] §X.2 and [3] §VII.7 or [4] §1.1 and 1.6).

The bound in Chebyshev's inequality (1.1.10) is typically far from optimal and can not be used to precisely assess the uncertainty associated with $\overline{F}$. For distributions with finite variance we have a much more precise statement, the Central Limit Theorem, that will be of fundamental importance in everything that follows: if the quantities $\{X_i\}_{i=1,\dots,N}$ are iid variables with average $\langle X\rangle$ and finite variance σ_X^2, in the large N limit the pdf $\rho(\overline{X})$ of the stochastic variable $\overline{X}$ converges to a Gaussian with average $\langle X\rangle$ and variance[1] σ_X^2/N:

$$\rho(\overline{X}) \to \frac{1}{\sqrt{2\pi\sigma_X^2/N}} \exp\left(-\frac{(\overline{X} - \langle X\rangle)^2}{2\sigma_X^2/N}\right) \ . \tag{1.1.13}$$

A proof of this and of more general statements can be found in [3] §VIII.4 and [4] §5.27, while a proof under quite restrictive hypotheses but with an estimate of the accuracy of the convergence is presented in the appendix of [5].

From the Central Limit Theorem we thus know that, for large enough N, the value $\overline{F}$ has a probability $\approx 68.3\%$ of being closer to $\langle F\rangle$ than $\sigma_{\overline{F}}$, a probability $\approx 95.5\%$ of being closer to $\langle F\rangle$ than $2\sigma_{\overline{F}}$, and a probability $\approx 99.7\%$ of being closer to $\langle F\rangle$ than $3\sigma_{\overline{F}}$. Moreover $\sigma_{\overline{F}}$ can be computed by using its sample estimator $\overline{\sigma}_{\overline{F}}$ in Eq. (1.1.9) and scales $\propto 1/\sqrt{N}$ for large N. The scaling $1/\sqrt{N}$ of statical errors is a consequence of the Central Limit Theorem, is universal in Monte Carlo methods and constitutes their main limitation or advantage, depending on the point of view.

1.2 Integration methods

The results of the previous section can be used to build simple Monte Carlo integrators and estimate their statistical accuracy. We consider for the sake of the simplicity an integral of the form

$$I = \int_0^1 f(x)\mathrm{d}x \ , \tag{1.2.1}$$

where $f(x)$ is a non negative regular function with $0 \le f(x) \le M$ for $x \in [0,1]$, see Fig. (1.1) (left).

Several MC approaches can be devised to estimate I. A simple possibility is to think of I as $\langle f\rangle$, where the average is computed with respect to the uniform pdf $p(x) = 1$ on $[0,1]$. We can thus proceed as follow:

1. generate N numbers $x_i \in [0,1]$ iid with pdf $p(x) = 1$
2. estimate I as $\overline{f} = \frac{1}{N}\sum_{i=1}^N f(x_i)$.

A different possibility is to write $f(x) = \int_0^{f(x)} \mathrm{d}y$ and thus

$$I = \int_0^1 \mathrm{d}x \int_0^{f(x)} \mathrm{d}y = \int_{[0,1]\times[0,M]} F(x,y)\mathrm{d}x\mathrm{d}y = M\int_{[0,1]\times[0,M]} F(x,y)\frac{\mathrm{d}x\mathrm{d}y}{M} \ , \tag{1.2.2}$$

[1] Note the consistency with Eq. (1.1.8).

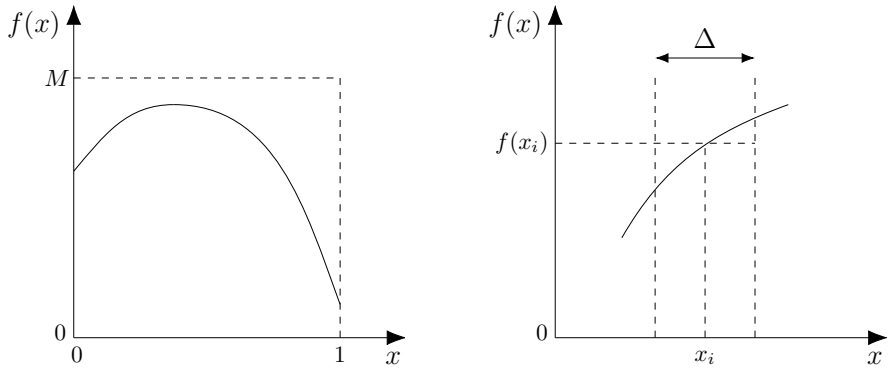

Figure 1.1: (left) The geometry considered for the integration in Sec. 1.2. (right) The basic step of the rectangle integration scheme.

where

$$F(x,y) = \begin{cases} 1 & \text{if } y \le f(x) \\ 0 & \text{else} \end{cases} \quad . \tag{1.2.3}$$

We thus have $I = M\langle F\rangle$, where the average is computed with respect to the uniform pdf $p(x,y) = 1/M$, and $\langle F\rangle$ is just the probability that a randomly chosen point in $[0,1]\times[0,M]$ falls below the curve $f(x)$. To estimate I we can now proceed as follows:

1. generate N points (x_i, y_i) in the rectangle $[0,1]\times[0,M]$ iid with pdf $p(x,y) = 1/M$
2. estimate I as $M\overline{F} = \frac{M}{N}\sum_{i=1}^{N} F(x_i,y_i)$, which is equal to M/N times the number of points below the curve $f(x)$.

The error of the MC estimates of I scales to zero as $1/\sqrt{N}$ in both the approaches, as dictated by the Central Limit Theorem. To understand which of the two method is more efficient we have to estimate the numerical factor multiplying $1/\sqrt{N}$ in the error, i.e. the standard deviation of the single extraction (multiplied by M in the second case). Using the first method we have

$$\sigma_f^2 = \langle f^2\rangle - \langle f\rangle^2 = \int_0^1 f^2(x)\mathrm{d}x - \left(\int_0^1 f(x)\mathrm{d}x\right)^2 \; ; \tag{1.2.4}$$

using the second method we have instead (using $F^2(x,y) = F(x,y)$)

$$\begin{aligned} \sigma_F^2 = \langle F^2\rangle - \langle F\rangle^2 &= \int_{[0,1]\times[0,M]} F(x,y)^2 \frac{\mathrm{d}x\mathrm{d}y}{M} - \left(\int_{[0,1]\times[0,M]} F(x,y)\frac{\mathrm{d}x\mathrm{d}y}{M}\right)^2 = \\ &= \int_{[0,1]\times[0,M]} F(x,y)\frac{\mathrm{d}x\mathrm{d}y}{M} - \left(\int_{[0,1]\times[0,M]} F(x,y)\frac{\mathrm{d}x\mathrm{d}y}{M}\right)^2 = \frac{I}{M} - \left(\frac{I}{M}\right)^2 \; , \end{aligned} \tag{1.2.5}$$

Note that in the second approach $I = M\langle F\rangle$, thus the relevant factor is $M\sigma_F = \sqrt{MI - I^2}$, which is a monotonically increasing function of $M \ge I$. It is thus convenient to chose M as small as possible, hence $M = \max f(x)$.

If we consider for example the case $f(x) = \sqrt{1-x^2}$, in which case $I = \pi/4$, we have (with $M = 1$)

$$\begin{aligned} \sigma_f &= \left(\int_0^1 (1-x^2)\mathrm{d}x - \left(\int_0^1 \sqrt{1-x^2}\mathrm{d}x\right)^2\right)^{1/2} = \left(1 - \frac{1}{3} - \left(\frac{\pi}{4}\right)^2\right)^{1/2} \simeq 0.22 \\ M\sigma_F &= \left(\frac{\pi}{4} - \left(\frac{\pi}{4}\right)^2\right)^{1/2} \simeq 0.41 \; , \end{aligned} \tag{1.2.6}$$

hence the error scales for large N as $\simeq 0.22/\sqrt{N}$ and as $\simeq 0.41/\sqrt{N}$ for the first and the second method, respectively. To achieve a given target precision, the second method thus requires a sample approximately four times larger than that of the first approach.

We can now compare these results with those that can be obtained by using deterministic approaches for the computation of I. The simplest deterministic integration method is the rectangle method (see Fig. (1.1) (right)):

1. divide the unit interval $[0, 1]$ in N intervals of size $\Delta = 1/N$.
2. select x_i in the i-th interval (e.g. $x_i = i/N$ or $x_i = (i + 1/2)/N$, with $i = 0, \ldots, N - 1$)
3. estimate the integral by $I_R = \Delta \sum_i f(x_i)$

The error of this estimate is bounded by

$$|I - I_R| \leq \sum_i \Delta(\max_i f - \min_i f) = \Delta \times (\text{total variation of } f) \ , \tag{1.2.7}$$

where $\max_i f$ denotes the maximum of $f(x)$ on the i-th interval and $\min_i f$ the corresponding minimum. For the case $f(x) = \sqrt{1 - x^2}$ considered above we have (using the fact that f is monotonic)

$$|I - I_R| \leq \Delta(\max f - \min f) = \frac{1}{N} \ . \tag{1.2.8}$$

The scaling with N is thus much more favorable in the rectangle discretization scheme than in the MC approach. Had we used the trapezoidal rule, in which the function is locally approximated by a linear function, we would have obtained an error scaling as $1/N^2$. Using a generic integration algorithm of order k (e.g. using spline interpolation of order k) we get an error which scales as $O(N^{-k})$.

If instead of considering a simple one-dimensional integral we consider a D-dimensional integral on $[0, 1]^D$, things change drastically. Denoting by Δ the linear separation of the grid to be used in a deterministic estimation of the integral, we need to evaluate the integrand function in $1/\Delta^D$ points. If we indicate the typical number of operations to be performed by N, we thus have $N \simeq \Delta^{-D}$, and the error of an integration scheme of order k scales as

$$\Delta^k \simeq N^{-k/D} \ . \tag{1.2.9}$$

On the contrary, the error of any Monte Carlo approach always scales as $1/\sqrt{N}$, independently of the dimensionality. For large enough D Monte Carlo becomes the best choice.

We have thus seen that the scaling of Monte Carlo errors is typically quite bad compared to the scaling of errors that can be obtained by using deterministic approaches. However, there are particular situations in which Monte Carlo methods are the most effective ones, the paradigmatic example being that of integration in spaces of very large dimensionality, which is relevant both for statistical mechanics and path-integration. To summarize [6]:

> Monte Carlo methods should be used only when all alternative methods are worse.

Chapter 2

Sampling a probability distribution function

2.1 Pseudo-random number generators

The output of a standard pseudo-random number generator is typically an integer number in the interval $[0, M)$ (or open or closed interval) with uniform pdf, which becomes a real number with pdf approximately uniform in $[0, 1)$ when dividing by M.

Pseudo-random number generator are usually based on iterative algorithms like $x_{i+1} = f(x_i)$ or $x_{i+k} = f(x_i, \ldots, x_{i+k-1})$, where x_0 (or $x_0, \ldots, x_{k-1}$) is the seed of the generator. It should be clear that the numbers x_i obtained using such an iterative algorithm are neither random nor independent from each other, but for many practical applications everything works "as if" these numbers were truly iid random quantities. Problems that are present in any pseudo-random number generator are

- finite period: a value i_{max} exists such that the sequence x_i repeats itself if $i > i_{max}$
- correlations: x_i clearly depends on the x_j with $j < i$, although this correlation can be quite nontrivial to highlight.

Whether a given random number generator is "good enough" for this cheat to be trustworthy is a nontrivial problem, and several tests are available to verify the quality of the randomness of the sequence x_i. For this reason it is good practice to use pseudo-random number generators that are known to be of high quality, although this is sometimes not sufficient, since what is thought to be a high quality generator is not time independent (see later in this section for an example). Note that, in the context of MC applications, the quality of pseudo-random number generator is typically non correlated with the generator being cryptographically secure.

Simple and very well studied pseudo-random number generators are linear congruential generators [7], in which natural numbers in $[0, m)$ are generated by iterating[1]

$$x_{n+1} = (a x_n + c) \bmod m \ , \tag{2.1.1}$$

where $0 \leq x_0 < m$ is the random seed, $0 < m$ is the modulus, $0 < a < m$ is the multiplier, and $0 \leq c < m$ is the increment. Clearly $0 \leq x_n < m$, thus $y_i = x_i/m$ is a pseudo-random real number in $[0, 1)$, and there are at most m different values that can be obtained by iterating Eq. (2.1.1).

Since x_{n+1} is obtained from x_i in a deterministic way, the sequence of numbers repeats itself once a number x_n is extracted which is equal to x_i for some $i < n$; the period of a linear congruential generator is thus surely not larger then the modulus m. Necessary and sufficient conditions for a linear congruential generator to have period m are provided by the Hull-Dobell theorem (for a proof see, e. g., [8] §3.2.1.2).

[1] we remind the reader that the notation x mod y denotes the remainder of the integer division of x by y.

Theorem 2.1.1 (Hull-Dobell). *A linear congruential generator has period m if and only if the following requirements are satisfied:*

1. *c is relatively prime to m,*
2. *$a-1$ is a multiple of p, for every prime number p dividing m,*
3. *if m is a multiple of 4, then $a-1$ is a multiple of 4*

A combination of parameters which satisfies these constraint is for example $m = 2^b, a = 4n+1$, and $c = 1$. Note however that a large period is not enough for a pseudo-random number generator to be a good one: a linear congruential generator with $a = 1$ and $c = 1$ clearly has period m, with m that can be arbitrarily large, still this is a terrible pseudo-random number generator.

All linear congruential generators with $c = 0$ (often called Lehmer generators) have a known weakness: if we define the numbers $y_k = x_k/m \in [0, 1)$ and we interpret k consecutive $y_i s$ (i.e. $\{y_i, y_{i+1}, \ldots, y_{i+k-1}\}$) as the coordinates of a point in k-dimensional space, then all these points lie in at most $(k!m)^{1/k}$ parallel hyperplanes [9]. Note however that in some cases the actual number of parallel hyperplanes on which these numbers lie is much smaller.

A famous example of such a failure is provided by the RANDU generator, which was the standard IBM pseudo-random generator in the '60s-'70s. This generator is defined by the recursion relation

$$x_{j+1} = (65539 x_j) \bmod 2^{31} \ , \ \text{with } x_0 \text{ odd.} \tag{2.1.2}$$

From the fact that x_0 is odd it immediately follows that x_j is always odd, thus $y_i = x_i/2^{31}$ is a number in $(0, 1)$. This pseudo-random number generator comes with the disclaimer "its very name RANDU is enough to bring dismay into the eyes and stomachs of many computer scientists!" ([8] p. 107), which is motivated by the ridiculously small number of parallel planes on which consecutive triples of numbers lie. According to the previously stated theorem this number is smaller than $(3!2^{31})^{1/3} \simeq 2344$, however the actual number is 15.

To show that the parameters choice used in RANDU is a very bad one we start by noting that $65539 = 2^{16} + 3$, thus

$$x_{j+2} = (2^{16} + 3)x_{j+1} = (2^{16} + 3)^2 x_j \ , \tag{2.1.3}$$

where all equalities hold modulo 2^{31}. Now we use

$$(2^{16} + 3)^2 = 2^{32} + 6 \times 2^{16} + 9 = 2^{32} + 6(2^{16} + 3) - 9 \tag{2.1.4}$$

to rewrite the previous equation as (again all equalities hold modulo 2^{31})

$$x_{j+2} = [6(2^{16} + 3) - 9]x_j = 6x_{j+1} - 9x_j \ . \tag{2.1.5}$$

We thus have $x_{j+2} - 6x_{j+1} + 9x_j = k2^{31}$, where k is an integer number, and finally

$$y_{j+2} - 6y_{j+1} + 9y_j = k \ . \tag{2.1.6}$$

This equation, with integer k, describes a family of parallel planes in $\mathbb{R}^3$, and it is simple to understand that of these planes at most 1+6+9=16 intersect the cube $[0, 1]^3$: 1 plane intersect the $j + 2$ axis, 6 planes intersect the $j + 1$ axis, and 9 planes intersect the j axis. The actual number of planes intersecting the cube $[0, 1]^3$ is in fact 15.

A less spectacular failure, but in some way a much more disturbing one, was reported in [10], where it was shown that a supposedly high quality pseudo-random number generator failed to reproduce the exact solution of the two dimensional Ising model when used in a MC simulation.

Simulations reported in the following of these notes have been performed by using the permuted congruential generator pcg32, in the minimal C implementation available at

https://www.pcg-random.org/download.html

It is good practice to write MC simulation codes in a way that makes it easy to change the pseudo-random number generator; this can be done, e. g., by introducing a wrapper function for the pseudo-random number generator.

2.2 Simple sampling, importance sampling, reweighting

We have seen in the previous section that algorithms are available to generate real pseudo-random numbers in the interval $[0, 1)$, and it is trivial to modify these algorithms to produce numbers in the interval $[0, M)$, with M arbitrary. Using these pseudo-random number generators we can thus sample a constant (eventually multidimensional) pdf, and we have seen in Sec. 1.2 that this is enough to estimate by Monte Carlo methods definite integrals. This approach goes under the name of *simple sampling*.

For many practical uses, and in particular for statistical mechanics applications, simple sampling is however very inefficient. In the large volume limit the Boltzmann distribution gets extremely peaked around the most probable configuration, which is the one with the largest entropy in the microcanonical ensemble or the one with the smallest free energy in the canonical ensemble. By uniformly sampling the configuration space we are thus almost surely selecting configurations which give negligible contribution to the physical result, so we are basically accumulating a lot of noise.

To make this argument more quantitative we can consider the average value

$$\langle O \rangle_p = \int O(x) p(x) \mathrm{d}x \ , \tag{2.2.1}$$

where $O(x)$ is an observable which depends smoothly on x, while $p(x)$ is a probability distribution function that is extremely peaked close to $\bar{x}$, so for example

$$p(x) \simeq \begin{cases} 1/\delta & x \in A \\ 0 & x \notin A \end{cases} \ , \tag{2.2.2}$$

with $\bar{x} \in A$, A a set of measure δ, and we are interested to the case $\delta \to 0$.

In simple sampling we uniformly sample the configuration space, so we use

$$\langle O \rangle_p = V \langle O p \rangle_1 \ , \tag{2.2.3}$$

where V is the total measure of the configuration space (the "volume"), and we denote by $\langle \quad \rangle_1$ the average with respect to the uniform pdf $1/V$. As in Sec. 1.2, to understand the effectiveness of the approach we have to study the standard deviation of the quantity we are averaging, and for simple sampling we get

$$\begin{aligned} & V \left(\int O^2(x) p^2(x) \frac{\mathrm{d}x}{V} - \left[\int O(x) p(x) \frac{\mathrm{d}x}{V} \right]^2 \right)^{1/2} \simeq \\ & \simeq \left(\frac{V}{\delta} O^2(\bar{x}) - O^2(\bar{x}) \right)^{1/2} = O(\bar{x}) \sqrt{\frac{V}{\delta} - 1} \ , \end{aligned} \tag{2.2.4}$$

which is both proportional to the (large) volume and divergent for $\delta \to 0$.

If in a Monte Carlo we instead generate points according to the distribution $p(x)$, the standard distribution which governs the error is for $\delta \to 0$

$$\left(\int O^2(x) p(x) \mathrm{d}x - \left[\int O(x) p(x) \mathrm{d}x \right]^2 \right)^{1/2} \simeq \left(O(\bar{x})^2 - O(\bar{x})^2 \right)^{1/2} = 0 \ . \tag{2.2.5}$$

It is clear that this second approach, known as *importance sampling* is more effective in statistical physics than simple sampling, and to use it we need methods to sample a generic distribution $p(x)$.

In the rest of this chapter we discuss the basic approaches to this problem, which are however typically quite (very) inefficient if the distribution $p(x)$ depends on many variables, as in statistical mechanics. In the next chapter we will discuss this more complicated case, introducing the Markov Chain Monte Carlo approach. Note however that the techniques developed in Secs. (2.3)-(2.4) will turns out to be useful also in the context of Markov Chain Monte Carlo, so it is worth to take them seriously.

$\langle x \rangle$	$\overline{x}$
0	0.0000(10)
0.25	0.2495(11)
0.5	0.4974(15)
0.75	0.7487(23)
1.0	0.9993(35)
1.5	1.492(10)
2.0	1.970(25)
2.5	2.474(69)
3	2.78(19)
4	2.60(32)
5	1.73(34)

Table 2.1: Values of $\overline{x}$ for a Gaussian pdf with average $\langle x \rangle$ and variance 1, obtained by sampling a Gaussian with zero average and variance 1 and reweighting the results. In all the cases 10^6 independent draws have been used.

With a reasoning similar to the one just used it is simple to understand the problems related to the technique commonly referred to as "reweighting". In some cases it is not possible to generate points according to the pdf $p(x)$, for example when $p(x)$ *is not* a pdf because it is not positive definite (we will see one occurrence of this problem when discussing identical fermionic particles in Chap. 12). In these cases one possibility is to generate points according to the pdf $g(x)$ and then use

$$\langle O \rangle_p = \int O(x) p(x) \mathrm{d}x = \int O(x) \frac{p(x)}{g(x)} g(x) \mathrm{d}x = \left\langle O \frac{p}{g} \right\rangle_g . \tag{2.2.6}$$

The variance of the original distribution (i. e. the one obtained by sampling $p(x)$) is

$$\sigma^2_{(p)} = \int O^2(x) p(x) \mathrm{d}x - \left(\int O(x) p(x) \mathrm{d}x \right)^2 \tag{2.2.7}$$

while the variance of the reweighted problem is

$$\sigma^2_{(g)} = \int O^2(x) \frac{p^2(x)}{g(x)} \mathrm{d}x - \left(\int O(x) p(x) \mathrm{d}x \right)^2 . \tag{2.2.8}$$

If $O(x)$ is a smooth function and in some points $p(x)/g(x) \gg 1$ then $\sigma^2_{(g)} \gg \sigma^2_{(p)}$. This means that reweighting works well only for distributions that are at least qualitatively similar, and this problem is usually known as the "overlap problem".

To have an explicit example of the overlap problem we can try to estimate numerically the average of a Gaussian pdf with average $\langle x \rangle$ and variance 1 by sampling a Gaussian pdf with zero average and variance 1, then reweighting the results (as we will see in the next section Gaussian pdf can be easily sampled). The results of this numerical experiment are shown in Tab. (2.1), where the estimate $\overline{x}$ obtained by reweighting a sample of 10^6 independent draws is reported together with the true average $\langle x \rangle$. It is clear that when $\langle x \rangle$ is larger than 1, and the two distributions become significantly different from each other, the reweighting method becomes very inefficient. It is important to explicitly note that, when the original and the reweighted distributions are very different from each other, $\langle x \rangle$ and $\overline{x}$ are not even compatible with each other: huge statistics would be required to even estimate reliably the variance of the average.

2.3 The change of variable method

The simplest method, at least from a theoretical point of view, to generate a non-uniform probability distribution function from a uniform pdf is the change of variable method.

Let us assume that the variable x is a random variable with pdf $p(x)$, that $f(x)$ is a smooth invertible function and let us denote by $\tilde{p}(y)$ the pdf of the random variable $y = f(x)$. Values of x in the interval $[x, x + \mathrm{d}x]$ correspond to values of y between $f(x)$ and $f(x + \mathrm{d}x) \simeq y + \frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x$, thus their probability is the same, thus the transformation law of the probability density functions is (using $\mathrm{d}y = |\mathrm{d}f/\mathrm{d}x|\mathrm{d}x$)

$$p(x)\mathrm{d}x = \tilde{p}(y)\mathrm{d}y \quad , \quad \tilde{p}(y) = \frac{p(x)}{|\mathrm{d}f/\mathrm{d}x|} \ . \tag{2.3.1}$$

In the expression of $\tilde{p}(y)$ there is obviously a slight abuse of notation: this function depends on y but in the right hand side of the equation we left the dependence on y implicit, since $x = f^{-1}(y)$.

Using the general transformation law for pdfs just obtained it is possible to sample nonuniform distributions; the nontrivial part of this task is to find the appropriate change of variable. If x is a random variable with uniform pdf on $[0, 1]$ and $y_0 = f(0)$, then

$$\int_{y_0}^{y} \tilde{p}(y')\mathrm{d}y' = \int_0^x \mathrm{d}x' = x \ , \tag{2.3.2}$$

and we can analytically find the change of variable needed to sample $\tilde{p}(y)$ if

1. we know the primitive of $\tilde{p}(y)$
2. we can invert the primitive of $\tilde{p}(y)$.

The simplest case in which both these requirements are satisfied is that of the uniform distribution function: if $\tilde{p}(y)$ is a uniform distribution function in $[a, a+M]$, we can for example assume $y_0 = a$, then the previous equation becomes $(y-a)/M = x$ and finally $y = a + Mx$. A slightly less trivial example is that of the exponential distribution function. If we want to sample the stochastic variable y in $[0, \infty)$ whose pdf is $\tilde{p}(y) = \mu e^{-\mu y}$, we can assume $y_0 = 0$ and from Eq. (2.3.2) we get

$$x = \int_0^y \mu e^{-\mu y'}\mathrm{d}y' = -e^{-\mu y'}\Big|_0^y = 1 - e^{-\mu y} \ , \tag{2.3.3}$$

from which $y = -\frac{1}{\mu}\log(1-x)$. If we use instead $y_0 = \infty$ we get

$$x = \int_y^\infty \mu e^{-\mu y'}\mathrm{d}y' = -e^{-\mu y'}\Big|_y^\infty = e^{-\mu y} \ , \tag{2.3.4}$$

hence $y = -\frac{1}{\mu}\log(x)$. Both the changes of variables can be used, since they differ only for the order in which one interval is mapped to the other. Indeed we can switch from one to the other using $x \to 1 - x$, which leaves invariant the uniform pdf on $[0, 1]$.

Probably the most famous and used application of the change of variable method is the generation of random numbers distributed with Gaussian pdf. If we need to sample the normal Gaussian pdf $\tilde{p}(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}$ we can not use the simplest strategy, since the primitive of the Gaussian is a non-elementary trascendental function, however we can follow a strategy that is similar to the one adopted to compute Gaussian integrals. If y_1 and y_2 are two independent stochastic variables, both with normal Gaussian pdf, their joint pdf is

$$p(y_1, y_2)\mathrm{d}y_1\mathrm{d}y_2 = \frac{1}{2\pi}e^{-\frac{1}{2}(y_1^2+y_2^2)}\mathrm{d}y_1\mathrm{d}y_2 \ . \tag{2.3.5}$$

Passing to polar coordinates $y_1 = r\cos\phi$, $y_2 = r\sin\phi$ the joint distribution function of the stochastic variables r and ϕ is

$$p(r, \phi)\mathrm{d}r\mathrm{d}\phi = \frac{1}{2\pi}e^{-\frac{1}{2}r^2}r\mathrm{d}r\mathrm{d}\phi = \left(\frac{\mathrm{d}\phi}{2\pi}\right)\left(e^{-\frac{1}{2}r^2}r\mathrm{d}r\right) \ , \tag{2.3.6}$$

hence ϕ and r are stochastically independent, with ϕ uniformly distributed on $[0, 2\pi)$ and r distributed with pdf $\tilde{p}(r) = re^{-\frac{1}{2}r^2}\mathrm{d}r$. Since we know the primitive of this pdf, we can use Eq. (2.3.2) with $r_0 = 0$, to get

$$x = \int_0^r r'e^{-\frac{1}{2}r'^2}\mathrm{d}r' = 1 - e^{-\frac{1}{2}r^2} \ , \tag{2.3.7}$$

Algorithm 1 Box-Muller algorithm to generate two independent normal Gaussian random numbers starting from random numbers distributed with uniform pdf in $(0,1)$.

Require: x, z sampled from uniform pdf in $(0,1)$
$y_1 = \sqrt{-2\log(x)}\cos(2\pi z)$
$y_2 = \sqrt{-2\log(x)}\sin(2\pi z)$

Algorithm 2 Polar form of the Box-Muller algorithm to generate two independent normal Gaussian random numbers starting from random numbers distributed with uniform pdf in $(0,1)$.

Require: r_1, r_2 sampled from uniform pdf in $(0,1)$
repeat
$z_1 = 1 - 2r_1$
$z_2 = 1 - 2r_2$
$S = z_1^2 + z_2^2$
until $0 < S < 1$
$y_1 = \frac{z_1}{\sqrt{S}}\sqrt{-2\log(S)}$
$y_2 = \frac{z_2}{\sqrt{S}}\sqrt{-2\log(S)}$

from which $r = \sqrt{-2\log(1-x)}$. If we use instead $r_0 = \infty$ we get the slightly simpler expression $r = \sqrt{-2\log x}$. We have thus shown that, given two random number $x, z \in (0,1)$ with uniform pdf, the two numbers y_1 and y_2 given by

$$y_1 = \sqrt{-2\log(x)}\cos(2\pi z)\ , \quad y_2 = \sqrt{-2\log(x)}\sin(2\pi z) \tag{2.3.8}$$

are sampled from two independent normal Gaussian distributions. This is the Box-Muller algorithm to generate normal Gaussian random numbers, summarized in Alg. (1).

This basic form of the Box-Muller algorithm is typically (i. e., on standard CPUs) not the most effective one, since the evaluation of the trigonometric functions is quite a slow operation. To increase the computational efficiency of the algorithm it is however possible to completely avoid the use of trigonometric functions: the pdf associated with the uniform probability inside the circle of unit radius is (in polar coordinates)

$$\frac{r\mathrm{d}r\mathrm{d}\phi}{\pi} = \mathrm{d}r^2\frac{\mathrm{d}\phi}{2\pi}\ , \tag{2.3.9}$$

hence by selecting with uniform probability a point inside the unit circle we are effectively selecting an angle ϕ with uniform probability on $[0, 2\pi)$ and the number r^2 with uniform probability on $[0,1)$. To select a point inside the unit circle with uniform pdf we can select a point inside $[-1,1]\times[-1,1]$ with uniform pdf, which is equivalent to generate two numbers z_1, z_2 with uniform pdf in $[-1,1]$, keeping only the selections for which the square distance $S = z_1^2 + z_2^2$ from the origin is smaller than 1. Using the points generated in this way we thus have the following facts

1. $S = z_1^2 + z_2^2$ is uniformly distributed in $[0,1)$
2. the angle ϕ such that $z_1 = \sqrt{S}\cos\phi$, $z_2 = \sqrt{S}\sin\phi$ is uniformly distributed in $[0, 2\pi)$
3. $\cos\phi = z_1/\sqrt{S}$ and $\sin\phi = z_2/\sqrt{S}$.

In this way we obtain the polar form of the Box-Muller algorithm (see Alg. (2)), which requires on average $\frac{4}{\pi} \simeq 1.27$ iteration to exit from the first cycle, but does not use any trigonometric function. The time required to generate 5×10^8 random Gaussian numbers using the polar form of the Box-Muller algorithm is $\simeq 21.58$s, while it is $\simeq 27.30$s using the basic version of the Box-Muller algorithm.

We close this section by explicitly noting that to sample a Gaussian pdf with average μ and standard deviation σ one can use $y = \mu + \sigma x$, where x is a normal Gaussian random variable, as can be easily seen by using Eq. (2.3.1). Several other algorithms which generate normal Gaussian pdf samples are discussed, e. g., in [8] §3.4.1.

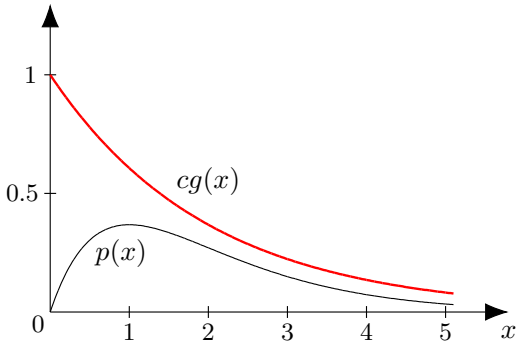

Figure 2.1: von Neumann accept/reject method: example with $p(x) = xe^{-x}$, $g(x) = \frac{1}{2}e^{-x/2}$ and $c = 2$.

Algorithm 3 von Neuman accept reject method to sample the pdf $p(x)$ using samples drawn from the pdf $g(x)$ such that $cg(x) \geq p(x)$.

repeat
  generate x_t with pdf $g(x_t)$
  generate y in $[0, cg(x_t)]$ with uniform pdf
until $p(x_t) < y$

2.4 The von Neumann accept/reject method

This method can be applied whenever we want to sample a pdf $p(x)$ and we know how to sample the pdf $g(x)$ with $cg(x) \geq p(x)$, see Fig. (2.1); note that by integrating the inequality $cg(x) \geq p(x)$ and using the normalization condition for a pdf we immediately get $c \geq 1$.

The strategy to sample $p(x)$ using samples drawn from $g(x)$ is the following:

1. select a value x_t according to the pdf $g(x)$
2. select a number y in $[0, cg(x_t)]$ using the uniform pdf
3. if $y \leq p(x_t)$ the trial number is accepted, else it is rejected and we go back to point 1.

Points 2. and 3. could be stated in a different but equivalent way by saying that we accept x_t with probability $p(x_t)/[cg(x_t)]$.

It is simple to verify that the numbers generated using this algorithm are distributed with pdf $p(x)$, indeed the average probability of accepting the trial state generated in point 1. is given by (remember that $c \geq 1$)

$$\langle P_{acc} \rangle = \int P(\text{selecting } x) P(\text{accepting } x) \mathrm{d}x = \int g(x) \frac{p(x)}{cg(x)} \mathrm{d}x = \frac{1}{c} \, , \tag{2.4.1}$$

and the distribution of the accepted values is

$$\frac{P(\text{selecting } x) P(\text{accepting } x)}{\int P(\text{selecting } y) P(\text{accepting } y) \mathrm{d}y} = \frac{g(x) \frac{p(x)}{cg(x)}}{1/c} = p(x) \, . \tag{2.4.2}$$

Since $1/c$ is the average probability of accepting the trial state, c is the average number of iterations required by the algorithm to accept a trial state, and measures the efficiency of the algorithm: the closer c is to 1 the more efficient the algorithm is.

As a nontrivial example of application of the accept/reject method we discuss how to sample a variable $x \in [-1, 1]$ with pdf $p(x) = A\sqrt{1 - x^2} e^{\gamma x}$, where γ is a parameter and A is a normalization constant whose value is fixed by imposing $\int_{-1}^{1} p(x) \mathrm{d}x = 1$. A possible algorithm to sample this distribution uses the accept/reject method starting from an exponential distribution [11]. The

distribution on $[-1,1]$ with pdf $g(x) = Be^{\gamma x}$, with $B = \gamma/(e^{\gamma} - e^{-\gamma})$, can indeed be easily sampled by the change of variable method: assuming z to be a variable with uniform pdf in $[0,1]$ and using $x(z=0) = -1$ we get

$$B \int_{-1}^{x} e^{\gamma x'} \mathrm{d}x' = z \ , \tag{2.4.3}$$

hence

$$x = \frac{1}{\gamma} \log\left(e^{-\gamma} + \frac{\gamma}{B} z\right) = \frac{1}{\gamma} \log\left(e^{-\gamma} + [e^{\gamma} - e^{-\gamma}]z\right) \ . \tag{2.4.4}$$

To apply the accept/reject method we now have to find a value c such that $cg(x) \geq p(x)$ for all x values in $[-1,1]$. Since $\sqrt{1-x^2} \leq 1$, it is sufficient to use $c = A/B$ and we can thus use the following algorithm

1. generate x_t with pdf $g(x_t)$ using the change of variable method
2. accept x_t with probability $\frac{p(x_t)}{cg(x_t)} = \sqrt{1-x_t^2}$, i.e. generate a random number r in $[0,1]$ with uniform probability and accept x_t if $r < \sqrt{1-x_t^2}$.

It should be intuitively clear that this algorithm becomes inefficient when $\gamma \gg 1$, since in this case $g(x)$ is very peaked close to $x = 1$ but $p(1) = 0$, and it is thus very difficult for the trial state to be accepted.

To be more quantitative we have to estimate A and thus c. We have

$$\frac{1}{A} = \int_{-1}^{1} \sqrt{1-x^2}\, e^{\gamma x} \mathrm{d}x \overset{(1)}{=} \int_{0}^{\pi} \sin^2\theta e^{\gamma \cos\theta} \mathrm{d}\theta \overset{(2)}{=} \frac{2\sqrt{\pi}}{\gamma} \Gamma\left(\frac{3}{2}\right) I_1(\gamma) \overset{(3)}{=} \frac{\pi}{\gamma} I_1(\gamma) \ , \tag{2.4.5}$$

where in the step (1) we used the change of variable $x = \cos\theta$ and in the step (2) we used the integral representation of the modified Bessel functions of first kind (see Eq. 9.6.18 of [12])

$$I_\nu(z) = \frac{\left(\frac{1}{2}z\right)^\nu}{\sqrt{\pi}\Gamma\left(\nu + \frac{1}{2}\right)} \int_0^\pi e^{z\cos\theta} \sin^{2\nu}\theta \mathrm{d}\theta \ , \tag{2.4.6}$$

which is valid for $\Re\nu > -1/2$. Finally in step (3) we used $\Gamma(3/2) = \sqrt{\pi}/2$ (see Eq. 6.1.9 of [12]). For $\gamma \gg 1$ we can use the approximate expression (see Eq. 9.7.1 of [12])

$$I_1(\gamma) \simeq \frac{e^\gamma}{\sqrt{2\pi\gamma}} \ , \tag{2.4.7}$$

hence for $\gamma \gg 1$ we find

$$c = \frac{A}{B} \simeq \sqrt{\frac{2\gamma}{\pi}} \gg 1 \ . \tag{2.4.8}$$

A more efficient algorithm to sample $p(x)$ when $\gamma \gg 1$ is discussed in [13].

Chapter 3

Markov Chain Monte Carlo

3.1 Markov chains: general properties

A Markov chain is a discrete time stochastic process, in which the probability of passing from the state x at time $t = n$ to the state y at time $t = n+1$ depends only on x, y, and n. In the following we consider only stationary chains, in which case the transition probability is independent of time. We denote by Ω the set of all the possible states of the Markov chain, and in the following we will assume Ω to be a finite set; for an analysis of the countably infinite case see, e. g., [2] §XV or [4] §1.8, for the most general case see, e. g., [14] §5.8.

In a stationary Markov chain, we denote by $W_{ab} = P(b \to a)$ the probability for the system to pass from the state b to the state a at any given time[1]. Some obvious properties of the matrix W, which completely characterize the Markov chain, are the following:

1. $0 \leq W_{ab}$,
2. $\sum_a W_{ab} = 1$ for every state b

The second property means that every state b will surely go somewhere in Ω at any step, and can be rephrased by saying that any column of W must sum up to 1. A matrix that satisfies these two requirements is usually called stochastic matrix. It is also convenient to introduce the probability of passing from state b to state a in k steps of the Markov chain, which is given by

$$P(b \to a \text{ in } k \text{ steps}) = \sum_{c_1,\ldots,c_{k-1}} W_{ac_1} W_{c_1 c_2} \cdots W_{c_{k-1} b} = (W^k)_{ab} \ . \tag{3.1.1}$$

We note that it is simple to show that any power of a stochastic matrix is again a stochastic matrix: if W is a stochastic matrix it is immediate to see that the elements of W^n are non negative, and if we assume W^k to be a stochastic matrix we have

$$\sum_i (W^{k+1})_{ij} = \sum_{i\alpha} W_{i\alpha} (W^k)_{\alpha j} = \sum_\alpha (W^k)_{\alpha j} = 1 \ , \tag{3.1.2}$$

hence also W^{k+1} is a stochastic matrix.

A Markov chain is said to be irreducible if for every couple of states $a, b \in \Omega$ a $k \in \mathbb{N}$ exists such that $(W^k)_{ab} > 0$; if this is not the case the Markov chain is said to be reducible. It is possible to represent any Markov chain by a graph: the states are the vertices of the graph, and two vertices b and a are connected by an oriented edge going from b to a if $W_{ab} > 0$. The Markov chain is irreducible if and only if, starting from any given vertex, we can reach any vertex (included the starting one) by traveling along the graph following the oriented edges. If a Markov chain is reducible then (at least) two disjoint subsets A and B of Ω exists such that all the states of A will

[1] Note that in the mathematical literature the different convention $W_{ba} = P(b \to a)$ is typically used.

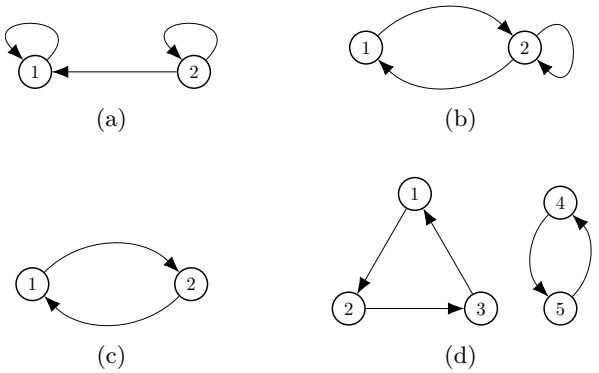

Figure 3.1: Examples of graphs associated with Markov chains.

never reach B during the evolution, hence we can order the states in such a way that the matrix W has the block form

$$W = \left(\begin{array}{c|c} \# & \# \\ \hline 0 & \# \end{array} \right) . \tag{3.1.3}$$

A sufficient condition for a Markov chain to be irreducible is obviously $W_{ab} > 0$ for any a, b.

For any state a of a Markov chain we define the set of its recurrence times by

$$R_a = \{k \in \mathbb{N} \setminus \{0\} | (W^k)_{aa} > 0\} . \tag{3.1.4}$$

The meaning of this definition is the following: if at time $t_0 = n$ the state of the Markov chain is a, then the state at time $t_1 = n + s > t_0$ can be again a only if $s \in R_a$. The period of the state a, denoted by T_a, is the greatest common divisor of R_a:

$$T_a = \mathrm{GCD}(R_a) , \tag{3.1.5}$$

so if k is not a multiple of T_a we surely have $(W^k)_{aa} = 0$; note however that not all the multiples of T_a are necessarily in R_a. If all the states of a Markov chain have period equal to one, then the chain is said to be aperiodic. A sufficient condition for a chain to be aperiodic is $W_{aa} > 0$ for any a, since in this case $1 \in R_a$ and thus $1 = \mathrm{GCD}(R_a)$ for any a.

Let us consider some examples of simple Markov chains.

- The matrix
$$W = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix} \tag{3.1.6}$$
is associated with the graph in Fig. (3.1a), and the corresponding Markov chain is reducible, since there is no way of passing from the state 1 to the state 2 in the evolution. Moreover $R_1 = R_2 = \{1, 2, 3, \ldots\}$, and $T_1 = T_2 = 1$, hence the Markov chain is aperiodic, which follow also from the fact that $W_{ii} > 0$
- The matrix
$$W = \begin{pmatrix} 0 & 1/2 \\ 1 & 1/2 \end{pmatrix} \tag{3.1.7}$$
is associated with the graph in Fig. (3.1b), and the corresponding Markov chain is irreducible, since $W_{12} = 1/2 > 0$ and $W_{21} = 1 > 0$ (alternatively, it is always possible to pass from 1 to 2 and viceversa in the graph). $R_1 = \{2, 3, 4, \ldots\}$ and $R_2 = \{1, 2, 3, \ldots\}$, hence $T_1 = T_2 = 1$ and the Markov chain is aperiodic (although $W_{11} = 0$).
- The matrix
$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.1.8}$$

is associated with the graph in Fig. (3.1c), and the corresponding Markov chain is irreducible, since $W_{12} = 1 > 0$ and $W_{21} = 1 > 0$ (alternatively, it is always possible to pass from 1 to 2 and viceversa in the graph). $R_1 = R_2 = \{2, 4, 6, \ldots\}$ and $T_1 = T_2 = 2$, hence the Markov chain is not aperiodic.

- The matrix

$$W = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{3.1.9}$$

is associated with the graph in Fig. (3.1d), and the corresponding Markov chain is reducible, since the graph is disconnected and there is, e. g., no way of passing from site 1 to site 4 in any number of steps. $R_1 = R_2 = R_3 = \{3, 6, 9, \ldots\}$ and $R_4 = R_5 = \{2, 4, 6, \ldots\}$, hence $T_1 = T_2 = T_3 = 3$, $T_4 = T_5 = 2$, and the Markov chain is not aperiodic.

Theorem 3.1.1. *In an irreducible Markov chain all the states have the same period.*

Proof. Let $a, b \in \Omega$ be states with period T_a and T_b, respectively. Since the Markov chain is irreducible, positive k_1 and k_2 exist such that $(W^{k_1})_{ab} > 0$ and $(W^{k_2})_{ba} > 0$, hence in $\bar{k} = k_1 + k_2$ steps it is possible to start from a, reach b and go back to a. In particular $\bar{k} \in R_a$, hence $\bar{k}$ is divisible by T_a.

We can go from a to a also in other ways: in k_2 steps we go from a to b, then in n steps we go from b to b and, finally, in k_1 steps we go from b to a:

$$a \xrightarrow{k_2} b \xrightarrow{n} b \xrightarrow{k_1} a \ . \tag{3.1.10}$$

Since $\bar{k} + n \in R_a$, $\bar{k} + n$ is divisible by T_a, but we have seen before that $\bar{k}$ is divisible by T_a, hence also n has to be divisible by T_a. Since n is the length of a generic $b \to b$ path, it follows that T_b is divisible by T_a. By switching the roles of a and b we obtain analogously that T_a is divisible by T_b, hence $T_a = T_b$. □

Theorem 3.1.2. *In an irreducible Markov chain of period T it is possible to decompose the configuration space as $\Omega = A_0 \cup \cdots \cup A_{T-1}$, where $A_n \cap A_m = \emptyset$ if $n \neq m$ and if $i \in A_n$ and $W_{ji} > 0$, then $j \in A_{(n+1) \bmod T}$.*

Proof. Let us define the sets A_n, with $n \in \{0, \ldots, T-1\}$, as follows[2]:

$$A_n = \{j \in \Omega \mid \exists k \text{ such that } k \equiv n \bmod T \text{ and } (W^k)_{j1} > 0\} \ . \tag{3.1.11}$$

A_n is thus the set of those states that can be reached, starting from the state 1, in a number of steps that is congruent to n modulo T. Since the Markov chain is irreducible we have $\Omega = \cup_n A_n$, moreover we can show that if $n \neq m$ the intersection $A_n \cap A_m$ is empty. If this were not the case, a j should exist such that $(W^{k_1})_{j1} > 0$, $(W^{k_2})_{j1} > 0$, with $k_1 \not\equiv k_2 \bmod T$; however, since the Markov chain is irreducible, a q exists such that $(W^q)_{1j} > 0$, hence $k_1 + q \in R_1$ and $k_2 + q \in R_1$, hence $k_1 + q$ and $k_2 + q$ are both divisible by T, from which it follows that $k_1 - k_2$ is divisible by T, contradicting $k_1 \not\equiv k_2 \bmod T$.

We have thus shown that the T sets A_n form a disjoint cover of Ω. Let us now assume that $i \in A_n$ and $W_{ji} > 0$. Then, by the definition of A_n, a k exists such that $k \equiv n \bmod T$ and $(W^k)_{i1} > 0$, but then

$$(W^{k+1})_{j1} = \sum_m W_{jm}(W^k)_{m1} \geq W_{ji}(W^k)_{i1} > 0 \ , \tag{3.1.12}$$

hence $j \in A_{(n+1) \bmod T}$ since $(k+1) \equiv (n+1) \bmod T$. □

[2] We remind the reader that the notation $a \equiv b \bmod c$ means that $a - b$ is divisible by c.

Corollary 3.1.1. *If W is the matrix associated with an irreducible Markov chain of period $T > 1$, then the Markov chain with matrix W^T is reducible.*

Proof. Using the decomposition of the previous theorem we immediately see that applying W^T to an element of A_n we can only obtain an element of A_n, hence the corresponding Markov chain is reducible. □

Using the matrix

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.1.13}$$

we get an example of application of the previous corollary: the Markov chain associated with W is irreducible and of period 2. The matrix W^2 is the identity, which corresponds to a reducible Markov chain.

We now recall some elementary facts about greatest common divisors which are needed to prove the following theorem.

Lemma 3.1.1. *If $a \equiv c \mod b$ then $\mathrm{GCD}(a, b) = \mathrm{GCD}(b, c)$.*

Proof. By hypothesis we have $a = c + nb$ for some $n \in \mathbb{Z}$, hence if d divides b and c it also divides a. Moreover, from $c = a - nb$ we see that if d divides a and b it also divides c. Hence

$$\{\text{divisors of } a, b\} = \{\text{divisors of } b, c\} \ , \tag{3.1.14}$$

and in particular $\mathrm{GCD}(a, b) = \mathrm{GCD}(b, c)$. □

Using the previous lemma we get Eulcid's algorithm for the computation of $\mathrm{GCD}(a, b)$. Let us assume that $a > b$, then we can write $a = bq_1 + r_1$, with $0 \leq r_1 < b$, hence $a \equiv r_1 \mod b$ and by Lemma 3.1.1 we have $\mathrm{GCD}(a, b) = \mathrm{GCD}(b, r_1)$. We can now go on by writing $b = r_1 q_2 + r_2$, with $0 \leq r_2 < r_1$, hence $b \equiv r_1 \mod r_2$ and $\mathrm{GCD}(b, r_1) = \mathrm{GCD}(r_1, r_2)$, and so on, until we find $r_k = 0$. In this way we get

$$\mathrm{GCD}(a, b) = \mathrm{GCD}(b, r_1) = \mathrm{GCD}(r_1, r_2) = \cdots = \mathrm{GCD}(r_{k-1}, 0) = r_{k-1} \ . \tag{3.1.15}$$

At each iteration of the Euclid's algorithm the remainder is a linear combination with integer coefficients of a, b: in the first iteration $r_1 = a - bq_1$, in the second iteration $r_2 = b - r_1 q_2 = b - (a - bq_1)q_2$, and using the general relation $r_{n+2} = r_n - q_{n+1} r_{n+1}$ it is immediate to prove the result by induction. From this fact it follows that $\mathrm{GCD}(a, b)$ can be written as a linear combination with integer coefficients of a and b, a fact that is known under the name of Bezout identity.

Using the fact that $\mathrm{GCD}(a, b, c) = \mathrm{GCD}(a, \mathrm{GCD}(b, c))$ it is possible to prove by induction that the Bezout identity can be generalized: given a set $S \subset \mathbb{N}$, the greatest common divisor of S, $\mathrm{GCD}(S)$, can be written as a linear combination with integer coefficients of a finite number r of elements of S, i. e.

$$\mathrm{GCD}(S) = \sum_{i=1}^{r} t_i s_i \ , \quad s_i \in S \ , \quad t_i \in \mathbb{Z} \ . \tag{3.1.16}$$

Lemma 3.1.2. *Let $A \subset \mathbb{N}$ be a set such that $\mathrm{GCD}(A) = 1$ and if $\alpha, \beta \in A$ then $\alpha + \beta \in A$. Then a number N exists such that if $n \in \mathbb{N}$ and $n \geq N$ then $n \in A$.*

Proof. By the Bezout identity we know that we can chose r elements $a_i \in A$ and r integer numbers t_i such that

$$\sum_{i=1}^{r} a_i t_i = 1 \ . \tag{3.1.17}$$

Let us define $\bar{t} = \max |t_i|$ and $\bar{a} = \sum_{i=1}^{r} a_i$. A generic integer number n can then be written in the form $n = k\bar{a} + s$, with $0 \leq s \leq \bar{a}$, and we can rewrite n as follows

$$n = k\bar{a} + s = \sum_{i=1}^{r} k a_i + s = \sum_{i=1}^{r} k a_i + s \sum_{i=1}^{r} a_i t_i = \sum_{i=1}^{r} (k + s t_i) a_i \ . \tag{3.1.18}$$

From this expression we see that, if $k \geq \bar{a}\bar{t}$, the number n is a linear combination with integer and non negative coefficients of the numbers a_i, hence by the properties of A we have $n \in A$ if $n \geq \bar{a}^2 \bar{t}$. □

Theorem 3.1.3. *For an irreducible aperiodic Markov chain a value N exists such that $(W^n)_{ij} > 0$ for every $i, j \in \Omega$ if $n > N$.*

Proof. It is sufficient to show that if $m \geq \bar{m}$ then $(W^m)_{ii} > 0$ for every $i \in \Omega$, since from the fact that the Markov chain is irreducible it follows that for every $i, j \in \Omega$ a k_{ij} exists such that $(W^{k_{ij}})_{ij} > 0$, and hence

$$(W^{m+k_{ij}})_{ij} = \sum_\alpha (W^m)_{i\alpha}(W^{k_{ij}})_{\alpha j} \geq (W^m)_{ii}(W^{k_{ij}})_{ij} > 0 \ . \tag{3.1.19}$$

We can thus choose $N = \bar{m} + \max_{ij} k_{ij}$ (we are obviously using the fact that Ω is a finite set).

Let us now show that for large enough m we have $(W^m)_{ii} > 0$ for every i. For this purpose it is sufficient to show that the set R_i of the return times of $i \in \Omega$ satisfies the hypotheses of the Lemma 3.1.2: if $n, m \in R_i$ then

$$(W^{n+m})_{ii} = \sum_\alpha (W^n)_{i\alpha}(W^m)_{\alpha i} \geq (W^n)_{ii}(W^m)_{ii} > 0 \ , \tag{3.1.20}$$

hence $n + m \in R_i$, moreover $\mathrm{GCD}(R_i) = 1$ since the Markov chain is aperiodic. Using once again the fact that Ω is a finite set we can thus find a $\bar{m}$ such that $(W^m)_{ii} > 0$ for every i if $m \geq \bar{m}$. □

3.2 Markov chains: spectral and ergodic properties

If we consider an ensemble of Markov chains we can introduce the probability p_a to be, at a given time, in the state $a \in \Omega$, and study how this probability depends on the time of the Markov chain. If $p_a^{(k)}$ is the probability of finding the state a at time k, we have the evolution equation

$$p_b^{(k+1)} = \sum_a W_{ba} p_a^{(k)} \ , \tag{3.2.1}$$

and it is meaningful to investigate what happens when $k \to \infty$. In particular, we want to investigate whether a pdf π_a exists such that $\pi_a = \lim_{k\to\infty} p_a^{(k)}$. If such a pdf exists, by performing the limit for $k \to \infty$ in Eq. (3.2.1) we get $\pi_b = \sum_a W_{ba}\pi_a$, hence π_a has to be an eigenvector of W with eigenvalue 1. To study this topic it is thus useful to investigate the spectrum of the matrix W associated with the Markov chain, and we will obtain a particular case of the Perron-Frobenius theorem (for the general case, which is valid for general non negative matrices, see [15] §XIII).

Theorem 3.2.1. *A stochastic matrix W has $\lambda = 1$ as one of its eigenvalues.*

Proof. The condition $\sum_a W_{ab} = 1$ of the stochastic matrix can be rewritten as $\sum_a (W_{ab} - \delta_{ab}) = 0$ for every b, hence the rows of the matrix $W - I$ are linearly dependent, thus $\det(W - I) = 0$ and $\lambda = 1$ is an eigenvalue of W. □

Theorem 3.2.2. *If λ is an eigenvalue of a stochastic matrix then $|\lambda| \leq 1$.*

Proof. Let v_a be the eigenvector corresponding to the eigenvalue λ, hence $\sum_b W_{ab} v_b = \lambda v_a$. Since $W_{ab} \geq 0$ we have

$$|\lambda||v_a| = |\lambda v_a| = \left|\sum_b W_{ab} v_b\right| \leq \sum_b |W_{ab} v_b| = \sum_b W_{ab}|v_b| \ , \tag{3.2.2}$$

and using $\sum_a W_{ab} = 1$ we get

$$|\lambda| \sum_a |v_a| \leq \sum_{ab} W_{ab}|v_b| = \sum_b |v_b| \ , \tag{3.2.3}$$

thus finally $|\lambda| \leq 1$. □

Theorem 3.2.3. *If v_a is an eigenvector with eigenvalue $\lambda \neq 1$ of a stochastic matrix then we have $\sum_a v_a = 0$.*

Proof. From $\lambda v_a = \sum_b W_{ab} v_b$ and $\sum_a W_{ab} = 1$ we get $\lambda \sum_a v_a = \sum_b v_b$, and since $\lambda \neq 1$ we conclude that $\sum_a v_a = 0$. □

Theorem 3.2.4. *If W is the stochastic matrix associated with an irreducible Markov chain and v_a is an eigenvector of W with eigenvalue 1, then all the components of v_a have the same sign (i. e., $v_a > 0$ for every $a \in \Omega$ or $v_a < 0$ for every $a \in \Omega$).*

Proof. Since $W_{ab} \in \mathbb{R}$ we can assume without loss of generality that $v_a \in \mathbb{R}$, moreover it is convenient to introduce the operator M defined by

$$M = \frac{1}{n}(W + W^2 + \cdots + W^n) \ . \tag{3.2.4}$$

Obviously $M_{ij} \geq 0$, and we have seen before that the power of a stochastic matrix is a stochastic matrix, hence also M is a stochastic matrix, and since the Markov chain associated with W is irreducible (and Ω is finite), we can assume n to be large enough for M_{ij} to be strictly positive for any i, j: $M_{ij} \geq \delta > 0$. Since $v_a = \sum_b W_{ba} v_b$ we also have $v_a = \sum_b M_{ab} v_b$.

Let us now introduce the notations

$$v_a^+ = \max\{v_a, 0\} \ , \quad v_a^- = \max\{-v_a, 0\} \ , \quad \alpha = \min\left\{ \sum_i v_i^+, \sum_i v_i^- \right\} \ . \tag{3.2.5}$$

Obviously $v_a = v_a^+ - v_a^-$ and we have

$$(Mv^+)_i = \sum_j M_{ij} v_j^+ \geq \delta \sum_j v_j^+ \geq \alpha\delta \ , \tag{3.2.6}$$

and analogously $(Mv^-)_i \geq \alpha\delta$, so

$$\begin{aligned} \sum_i |v_i| = \sum_i |(Mv)_i| = \sum_i |(Mv^+)_i - (Mv^-)_i| = \sum_i |(Mv^+)_i - \alpha\delta + \alpha\delta - (Mv^-)_i| \leq \\ \leq \sum_i |(Mv^+)_i - \alpha\delta| + \sum_i |(Mv^-)_i - \alpha\delta| = \sum_i (Mv^+)_i + \sum_i (Mv^-)_i - 2N\alpha\delta \ , \end{aligned} \tag{3.2.7}$$

where the last equality follows from the fact $(Mv^\pm)_i \geq \alpha\delta$, and we denoted by N the number of elements of Ω. Using $\sum_i M_{ij} = 1$ we thus get

$$\begin{aligned} \sum_i |v_i| \leq \sum_{ij} M_{ij} v_j^+ + \sum_{ij} M_{ij} v_j^- - 2N\alpha\delta = \\ = \sum_j v_j^+ + \sum_j v_j^- - 2N\alpha\delta = \sum_j |v_j| - 2N\alpha\delta \ , \end{aligned} \tag{3.2.8}$$

from which we conclude that $\alpha = 0$ and we can thus assume (up to a global sign) $v_a \geq 0$ for any $a \in \Omega$. We conclude by noting that

$$v_a = (Mv)_a = \sum_j M_{aj} v_j \geq \delta \sum_j v_j > 0 \tag{3.2.9}$$

since $\delta > 0$, and $\sum_j v_j = 0$ would imply $v_j = 0$ for every $j \in \Omega$, since $v_a \geq 0$. □

Theorem 3.2.5. *If W is the stochastic matrix associated with an irreducible Markov chain the eigenvalue $\lambda = 1$ of W is non degenerate.*

Proof. Let us assume that v and v' are two different eigenvectors of W with eigenvalue 1. By the previous theorem we can assume $v_a > 0$ and $v'_a > 0$ for every $a \in \Omega$, and we can normalize them in such a way that $\sum_a v_a = \sum_a v'_a = 1$. We now introduce $w_a = v_a - v'_a$, which is still another eigenvector of W with eigenvalue 1. By the previous theorem we have $w_a > 0$ for all $a \in \Omega$ or $w_a < 0$ for all $a \in \Omega$, but this is in contradiction with $\sum_a w_a = \sum_a v_a - \sum_a v'_a = 0$. □

The previous two theorems are finite dimensional analogues of the fact that in quantum mechanics the ground state is always non degenerate and its wave function can be chosen to be positive definite, see, e. g., [16] §15.4 for a sketch of the proof, or [17] §3.3.3, [18] §10.5 for more details.

Theorem 3.2.6. *If W is the stochastic matrix associated with an irreducible and aperiodic Markov chain and $\lambda \neq 1$ is an eigenvector of W, then $|\lambda| < 1$.*

Proof. We know from Theorem. 3.2.2 that $|\lambda| \leq 1$ and let us assume that $|\lambda| = 1$, i. e., $\lambda = e^{i\theta}$ for some $\theta \in \mathbb{R}$. If we denote by w_a the eigenvector associated with λ, we can write $w_a = r_a e^{i\theta_a}$, with $r_a \geq 0$ and $\sum_a r_a = 1$, and the eigenvalue equation $\lambda w_a = \sum_b W_{ab} w_b$ becomes

$$r_a e^{i\theta+\theta_a} = \sum_b W_{ab}\, r_b\, e^{i\theta_b} \ . \tag{3.2.10}$$

Multiplying this equation by $e^{-i(\theta+\theta_a)}$ and summing on a we get

$$\sum_{ab} W_{ab}\, r_b\, e^{i(\theta_b-\theta_a-\theta)} = 1 \ . \tag{3.2.11}$$

Since $W_{ab} r_b \geq 0$ and $\sum_{ab} W_{ab} r_b = \sum_b r_b = 1$, the previous equation implies that $e^{i(\theta_b-\theta_a-\theta)} = 1$ for every $a, b \in \Omega$ such that $W_{ab} r_b > 0$. If $r_b = 0$ we can chose arbitrarily the angle θ_b, hence we can assume the stronger condition

$$e^{i(\theta_b-\theta_a-\theta)} = 1 \text{ for every } a, b \text{ such that } W_{ab} > 0 \ . \tag{3.2.12}$$

When used in Eq. (3.2.10) this relation shows that the vector r_a is an eigenvector of W with eigenvalue 1, hence, in particular, $r_a > 0$ for any $a \in \Omega$ by Theorem 3.2.4, since the Markov chain is irreducible. Due to the irreducibility, Eq. (3.2.12) determines all the θ_a values once $\theta_1 = 0$ has been arbitrarily fixed.

For any k such that $(W^k)_{11} > 0$ (i. e., $k \in R_1$, and $R_1 \neq \emptyset$ since the Markov chain is irreducible), k elements $a_1, \ldots, a_k \in \Omega$ exist such that

$$W_{1a_1} W_{a_1 a_2} \cdots W_{a_k 1} > 0 \ , \tag{3.2.13}$$

and Eq. (3.2.12) implies

$$1 = e^{i(\theta_{a_1}-\theta_1-\theta)} e^{i(\theta_{a_2}-\theta_{a_1}-\theta)} \cdots e^{i(\theta_1-\theta_{a_k}-\theta)} = e^{-ik\theta} \ , \tag{3.2.14}$$

hence $k\theta$ is an integer multiple of 2π, and we can assume $\theta = 2\pi\alpha$ for some $\alpha = \frac{n}{d}$, with n and d positive, relatively prime, and $n < d$. Since the previous property is true for any $k \in R_1$, we must have $k_i \alpha \in \mathbb{Z}$ for any $k_i \in R_1$, hence d must be a divisor of any $k_i \in R_1$. Since the chain is aperiodic we have $\mathrm{GCD}(R_1) = 1$, thus $d = 1$ and $\theta = 0$, which gives $\lambda = 1$. □

Summarizing, we have shown that for the stochastic matrix W corresponding to an aperiodic and irreducible Markov chain the following fundamental facts are true

1) all the eigenvalues $\lambda \neq 1$ satisfy $|\lambda| < 1$
2) $\lambda = 1$ is a non degenerate eigenvalue and, with an appropriate choice of sign, all the components of the corresponding eigenvector are strictly positive

These points can be rephrased by saying that any aperiodic and irreducible Markov chain has a unique invariant probability density function, that we will denote by π_a, and π_a is strictly positive for any $a \in \Omega$. These fundamental facts will now be used to discuss the large-k behavior of the quantity $(W^k p)$, where p_a is pdf on Ω.

Sometimes it can be useful to note that the previous implications also work in the opposite direction. If W is the stochastic matrix associated with a Markov chain, and $\lambda = 1$ is a nondegenerate eigenvalue of W, then the Markov chain is irreducible, indeed for a reducible Markov chain

we can find at least two irreducible sub-chains, each one with its own $\lambda = 1$ eigenvalue, and $\lambda = 1$ is thus a degenerate eigenvalue of the original chain. If moreover all the eigenvalues of W different from $\lambda = 1$ satisfy $|\lambda| < 1$, then the Markov chain is also aperiodic. This is true since, by slightly modifying the proof of theorem 3.2.6, it can be shown that if an irreducible Markov chain has period T, then in the spectrum of W the T-th roots of the identity are present.

Note that we have investigated the spectrum of the stochastic matrix W associated with a Markov chain, but in general stochastic matrices are not diagonalizable (even for irreducible and aperiodic Markov chains). An explicit example is provided by

$$M = \frac{1}{5}\begin{pmatrix} 2 & 2 & 1 \\ 1 & 2 & 1 \\ 2 & 1 & 3 \end{pmatrix} . \tag{3.2.15}$$

It is easily seen that this matrix has eigenvalues 1 and 1/5, with algebraic degeneracy 1 and 2, respectively, but a single eigenvector corresponds to the eigenvalue 1/5 (the vector $\frac{1}{\sqrt{2}}(1, 0, -1)$), hence this matrix is nondiagonalizable, and its Jordan canonical form is

$$M_J = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/5 & 1 \\ 0 & 0 & 1/5 \end{pmatrix} . \tag{3.2.16}$$

To study the large-k behavior of $(W^k p)_a = \sum_b (W^k)_{ab} p_b$, where W is associated with an irreducible and aperiodic Markov chain, let us start by considering the simpler case in which the matrix W can be diagonalized. In this case we can expand the vector p_a on an eigenbasis of W, hence

$$p_a = c_1 \pi_a + \sum_{j>1} c_j v_a^{(j)} , \tag{3.2.17}$$

where π_a is the unique invariant pdf of the Markov chain and $v_a^{(j)}$ is the j-th eigenvector with $j > 1$, associated with an eigenvalue of absolute value smaller than 1. The pdf p_a and the invariant pdf π_a are normalized by $\sum_a p_a = \sum_a \pi_a = 1$, while for the eigenvectors $v_a^{(j)}$ with $j > 0$ we have $\sum_a v_a^{(j)} = 0$ due to Theorem 3.2.3, and we can assume $\sum_a |v_a^{(j)}| = 1$. We thus get

$$1 = \sum_a p_a = c_1 \sum_a \pi_a + \sum_{j>1} \left(c_j \sum_a v_a^{(j)} \right) = c_1 , \tag{3.2.18}$$

and thus

$$p_a = \pi_a + \sum_{j>1} c_j v_a^{(j)} . \tag{3.2.19}$$

Applying W^k to this equation we get

$$(W^k p)_a = \pi_a + \sum_{j>1} c_j \lambda_j^k v_a^{(j)} , \tag{3.2.20}$$

and we can introduce $0 \leq \Lambda = \max_{j>1} |\lambda_j| < 1$ to estimate the convergence rate of $(W^k p)_a$ to π_a as follows

$$\begin{aligned} \sum_a |(W^k p)_a - \pi_a| = \sum_a \left| \sum_{j>1} c_j \lambda_j^k v_a^{(j)} \right| \leq \sum_a \sum_{j>1} |\lambda_j|^k |c_j| |v_a^{(j)}| \leq \\ \leq \Lambda^k \sum_{j>1} |c_j| \sum_a |v_a^{(j)}| = \Lambda^k \sum_{j>1} |c_j| , \end{aligned} \tag{3.2.21}$$

where in the last step we used the normalization $\sum_a |v_a^{(j)}| = 1$. Introducing the notation $A = \sum_{j>1} |c_j|$ we have thus

$$\sum_a |(W^k p)_a - \pi_a| \leq A \Lambda^k = A e^{k \log(\Lambda)} , \tag{3.2.22}$$

which, by introducing the exponential autocorrelation time $\tau_{\rm exp} > 0$ defined by

$$\tau_{\rm exp} = -\frac{1}{\log(\Lambda)} = -\frac{1}{\log \max_{j>1} |\lambda_j|} \ , \tag{3.2.23}$$

can finally be written in the form

$$\sum_a |(W^k p)_a - \pi_a| \le A\, e^{-k/\tau_{\rm exp}} \ . \tag{3.2.24}$$

The quantities $(W^k p)_a$ thus converge exponentially fast in k to π_a, and the typical timescale is set by the largest value of $|\lambda_j|$ smaller than 1.

If the matrix W associated with the irreducible and aperiodic Markov chain is non diagonalizable we need to slightly modify the previous discussion. A possible way to investigate the problem in this case is to use the basis in which W assumes its Jordan canonical form. In this basis W is a block diagonal matrix, with a single unidimensional block with 1 on its diagonal, and blocks with $|\lambda| < 1$, which can be of the following two forms:

$$B_\lambda = \begin{pmatrix} \lambda & 1 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix} \ , \qquad D_\lambda = \begin{pmatrix} \lambda & 0 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix} \ . \tag{3.2.25}$$

It is immediate to verify by induction that

$$B_\lambda^k = \lambda^{k-1} \begin{pmatrix} \lambda & k & 0 & 0 & 0 \\ 0 & \lambda & k & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \lambda & k \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix} \ , \tag{3.2.26}$$

hence $\lim_{k\to\infty} B_\lambda^k \to 0$ and obviously also $\lim_{k\to\infty} D_\lambda^k \to 0$. We thus see that $\lim_{k\to\infty} W^k = P_1$, where P_1 is the projector on the eigenspace corresponding to the eigenvalue $\lambda = 1$. Given any pdf p_a on Ω we thus have $\lim_{k\to\infty}(W^k p)_a = \alpha \pi_a$, and by summing on a we see that $\alpha = 1$. The estimate of the convergence rate of $(W^k p)_a$ to π_a changes in the nondiagonalizable case only (possibly) by logarithmic corrections[3], becoming

$$\sum_a |(W^k p)_a - \pi_a| \le C \Lambda^{k-1} (k + \Lambda) \ , \tag{3.2.27}$$

where Λ has the same meaning as before, hence (using $\Lambda < 1$)

$$\sum_a |(W^k p)_a - \pi_a| \le C(k+1) e^{-(k-1)/\tau_{\rm exp}} \ . \tag{3.2.28}$$

Note that for large k we have asymptotically

$$e^{-k/\tau_{\rm exp}} \le (k+1) e^{-(k-1)/\tau_{\rm exp}} \le e^{-k/(\tau_{\rm exp}+\epsilon)} \tag{3.2.29}$$

for any $\epsilon > 0$, so the nondiagonalizability of W does not significantly affects the asymptotic convergence rate.

[3] This happens if the largest value of $|\lambda_j|$ smaller then 1 corresponds to a non-diagonal Jordan block.

3.3 Sampling a pdf using Markov chains

We have seen in the previous section that in an irreducible and aperiodic Markov chain, given any initial pdf p_a, the late time distribution $(W^k p)_a$ approaches the unique invariant pdf π_a of the Markov chain. In particular, we can start from the completely deterministic initial distribution $p_a = \delta_{ab}$, which means that at time $t = 0$ the state of the Markov chain is b, and generate new states according to the transition probabilities of the Markov chain: the states in Ω will be asymptotically visited, during the evolution, with pdf π_a. This method to sample the pdf π_a is known as the Markov Chain Monte Carlo method (MCMC for short). Note that this method differs in an important aspect from the methods discussed in Chap. 2: in this case the draws are *not* independent.

In the present section we address the following problem: given a probability distribution function π_a, can we build an aperiodic and irreducible Markov chain whose invariant pdf is π_a? We thus want to find a way of constructing an aperiodic and irreducible Markov chain whose associated stochastic matrix W satisfies

$$\pi_a = \sum_b W_{ab} \pi_b \ , \tag{3.3.1}$$

where now π_a is a preassigned pdf, and the unknowns are the matrix elements W_{ab}. In this context the previous equation is usually known as the "balance equation", and it should be clear that, in general, this equation does not uniquely determine the matrix W: a stochastic $N \times N$ matrix has $N^2 - N$ independent elements (since there are N constraints $\sum_a W_{ab} = 1$) and the balance equation adds N constraints, thus leaving $N^2 - 2N$ degrees of freedom.

The balance equation can be rewritten, using $\sum_b W_{ba} = 1$, in the form

$$\sum_b W_{ba} \pi_a = \sum_b W_{ab} \pi_b \tag{3.3.2}$$

and by subtracting $W_{aa}\pi_a$ on both the sides we get

$$\sum_{b \neq a} W_{ba} \pi_a = \sum_{b \neq a} W_{ab} \pi_b \ . \tag{3.3.3}$$

The left hand side of this equation gives the average probability of leaving the state a: if at time t we have a probability π_a of being in the state a, the probability that the state at time $t+1$ is different from a is $\sum_{b \neq a} W_{ba} \pi_a$. The right hand side of the previous equation is instead the average probability of reaching the site a: if we have a probability π_b of being in $b \neq a$ at time t, the probability that the state at time $t+1$ is a is $\sum_{b \neq a} W_{ab} \pi_b$. The balance equation can thus be interpreted as an equilibrium condition between the probabilities of leaving and of reaching the generic state a.

The balance equation is the necessary condition that must be satisfied for π_a to be the invariant pdf of the Markov chain associated with the stochastic matrix W. Since this condition leaves much freedom in the choice of W, it is customary to impose a much stronger requirement, known as the "detailed balance condition":

$$W_{ba} \pi_a = W_{ab} \pi_b \quad \text{for any } a, b \in \Omega \ . \tag{3.3.4}$$

By summing on b the detailed balance condition, and using $\sum_b W_{ba} = 1$, we immediately recover the balance condition. The balance condition ensures that, for any state $a \in \Omega$, the average probability of leaving the state a is the same as the average probability of reaching the state a. The detailed balance condition ensures instead that the average probability of the transition $a \to b$ is the same as the average probability of the transition $b \to a$ for any $a, b \in \Omega$.

Lemma 3.3.1. *If the matrix W is associated with an irreducible Markov chain and satisfies the detailed balance condition, then W is diagonalizable.*

Algorithm 4 Metropolis algorithm to generate a Markov chain which satisfies the detailed balance condition with pdf π_a ($F(x) = \min(1, x)$ or $F(x) = x/(1+x)$).

loop
  a is the present state of the Markov chain
  select b with probability $A_{ba} = A_{ab}$
  draw a random number in $[0, 1)$ with uniform pdf
  if $r \leq F(\pi_b/\pi_a)$ **then**
    the next state of the Markov chain is b
  else
    the next state of the Markov chain is a
  end if
end loop

Proof. if π_a is the invariant distribution of an irreducible Markov chain we have seen in Theorem 3.2.4 that $\pi_a > 0$ for any $a \in \Omega$, hence we can introduce the scalar product

$$(v, u) = \sum_a \pi_a v_a u_a \ , \tag{3.3.5}$$

and we have

$$(v, {}^tWu) = \sum_{ab} \pi_a v_a W_{ba} u_b = \sum_{ab} \pi_b W_{ab} v_a u_b = ({}^tWv, u) \ , \tag{3.3.6}$$

hence ${}^tW$ is Hermitian with respect to the scalar product $(\ ,\ )$, and thus diagonalizable. As a consequence also W is diagonalizable. □

In the following subsections we discuss two algorithms to build a Markov chain which satisfies the detailed balance condition with respect to a given pdf π_a.

3.3.1 The Metropolis(-Hastings) algorithm

The idea of the Metropolis algorithm [19] is somehow similar to that of the von Neumann accept/reject method discussed in Sec. 2.4: we start from a Markov chain with transition matrix A_{ba}, which does not have π_a as invariant pdf, and introduce a correction step to generate a Markov chain for which π_a is an invariant distribution. Note that the final Markov chain is not automatically irreducible and aperiodic; these properties has to be verified *a posteriori*.

The starting point is thus the stochastic matrix A_{ba}, which is used to generate a trial state b starting from the state a at time t, and it is assumed to be a symmetric matrix ($A_{ab} = A_{ba}$). The state b is then accepted or rejected with an acceptance probability of the form $F(\pi_b/\pi_a)$ if $b \neq a$, where $0 \leq F(x) \leq 1$ is a function to be determined, while it is always accepted if $b = a$. If b is accepted, the state at time $t+1$ is b, otherwise the state remains a. The complete transition probabilities are thus

$$\begin{aligned} W_{ba} &= A_{ba} F\left(\frac{\pi_b}{\pi_a}\right) \qquad \text{if } b \neq a \ , \\ W_{aa} &= A_{aa} + \sum_{z \neq a} A_{za} \left(1 - F\left(\frac{\pi_z}{\pi_a}\right)\right) \ . \end{aligned} \tag{3.3.7}$$

Note that the state at time $t+1$ can be a for two different reasons: either the state a has been selected by the Markov chain associated with the matrix A, and thus surely accepted, or a state $z \neq a$ has been selected and rejected. It is immediate to show that W is a stochastic matrix: clearly $W_{ba} \geq 0$, moreover

$$\sum_b W_{ba} = \sum_{b \neq a} A_{ba} F\left(\frac{\pi_b}{\pi_a}\right) + A_{aa} + \sum_{z \neq a} A_{za} \left(1 - F\left(\frac{\pi_z}{\pi_a}\right)\right) = \sum_b A_{ba} = 1 \ . \tag{3.3.8}$$

Algorithm 5 Metropolis-Hastings algorithm to generate a Markov chain which satisfies the detailed balance condition with pdf π_a ($F(x) = \min(1, x)$ or $F(x) = x/(1 + x)$).

loop
  a is the present state of the Markov chain
  select b with probability A_{ba}
  draw a random number in $[0, 1)$ with uniform pdf
  if $r \leq F[(A_{ab}\pi_b)/(A_{ba}\pi_a)]$ **then**
    the next state of the Markov chain is b
  else
    the next state of the Markov chain is a
  end if
end loop

The detailed balance condition $W_{ab}\pi_b = W_{ba}\pi_a$ is trivially satisfied if $b = a$, while for $b \neq a$ it becomes

$$A_{ab}F\left(\frac{\pi_a}{\pi_b}\right)\pi_b = A_{ba}F\left(\frac{\pi_b}{\pi_a}\right)\pi_a \ . \tag{3.3.9}$$

Using the symmetry of A we thus obtain for $F(x)$ the functional equation

$$F(x) = xF(1/x) \ . \tag{3.3.10}$$

This equation has infinite solutions, but the two that are most commonly used are $F_1(x) = \min(1, x)$ and $F_2(x) = \frac{x}{1+x}$. These functions can be easily shown to be solutions of the above functional equation, indeed

$$xF_1\left(\frac{1}{x}\right) = x\min\left(1, \frac{1}{x}\right) = \begin{cases} \text{if } x \geq 1: & x \cdot (1/x) = \min(1, x) = F_1(x) \\ \text{if } x < 1: & x \cdot 1 = \min(1, x) = F_1(x) \end{cases} \ , \tag{3.3.11}$$

and

$$xF_2\left(\frac{1}{x}\right) = x\frac{1/x}{1 + 1/x} = \frac{1}{1 + x} = F_2(x) \ . \tag{3.3.12}$$

Putting everything together we thus obtain the algorithm Alg. (4), and the accept/reject step is often called Metropolis step or Metropolis filter. As already noted, the Metropolis algorithm generates a Markov chain which leaves invariant the pdf π_a, however we also have to check (using the specific form of the matrix A_{ab} and of the function F) that the Markov chain generated in this way is irreducible and aperiodic, in order to be sure that $(W^k p)_a$ converges to π_a for large k values.

Nonsymmetric selection probabilities A_{ba} can also be used, however in this case the previous algorithm has to be slightly modified: the acceptance probability to be used in the accept/reject step becomes

$$F\left(\frac{A_{ab}\pi_b}{A_{ba}\pi_a}\right) \tag{3.3.13}$$

instead of $F(\pi_b/\pi_a)$. In this case the algorithm is called Metropolis-Hastings algorithm [20], and it is summarized in Alg. (5).

It is worth noting a peculiarity of the Metropolis(-Hastings) algorithm: the acceptance probability depends only on the ratio π_b/π_a, hence it is independent of the normalization of the pdf π_a. If this were not the case, this algorithm would be useless in statistical mechanics, since the computation of the normalization of the Gibbs distribution (i. e., the partition function) is as difficult as any other computation.

We now consider a simple example to illustrate the use of the Metropolis algorithm. Let $f(x)$ be a strictly positive ($f(x) > 0$ for any x) and integrable function, like, e. g., a Gaussian, and

define the pdf $\pi(x)$ by

$$\pi(x) = \frac{f(x)}{\int_{-\infty}^{+\infty} f(y)\mathrm{d}y} \ . \tag{3.3.14}$$

If we want to sample the pdf $\pi(x)$ a possible strategy is the following: given an arbitrary x_0 (the initial state of the Markov chain) and a value $\delta > 0$, we can build a Markov chain as follows:

loop
  x_k is the present state of the Markov chain
  select $\bar{x} \in (x_k - \delta, x_k + \delta)$ with uniform pdf
  select $r \in [0, 1)$ with uniform pdf
  if $r \leq \min[1, f(\bar{x})/f(x_k)]$ **then**
    $x_{k+1} = \bar{x}$
  else
    $x_{k+1} = x_k$
  end if
end loop

The selection probability is

$$A_{yx} = \begin{cases} 1/(2\delta) & \text{if } |x - y| < \delta \\ 0 & \text{elsewhere} \end{cases} \ , \tag{3.3.15}$$

and is clearly symmetric. Since $f(x) > 0$ it is possible to reach any point in a finite number of steps, hence the chain is irreducible, moreover it is possible to select $\bar{x} = x_k$, hence the chain is aperiodic[4]. In this way, after a number of iterations that is large with respect to $\tau_{\exp}$, this algorithm asymptotically sample the pdf $\pi(x)$. This is true for any value of the parameter δ, however the numerical efficiency of the algorithm is not independent of δ, as we will discuss in Chap. 4. In particular $\tau_{\exp}$ does depend on δ.

It is possible to slightly improve the algorithm to sample $\pi(x)$ which we have just seen, in order to make it faster on typical CPUs. For this purpose we can substitute the block

select $r \in [0, 1)$ with uniform pdf
if $r \leq \min[1, f(\bar{x})/f(x_k)]$ **then**
  $x_{k+1} = \bar{x}$
else
  $x_{k+1} = x_k$
end if

with the theoretically equivalent

$y = f(\bar{x})/f(x_k)$
if $y \geq 1$ **then**
  $x_{k+1} = \bar{x}$
else
  select $r \in [0, 1)$ with uniform pdf
  if $r \leq \min[1, y]$ **then**
    $x_{k+1} = \bar{x}$
  else
    $x_{k+1} = x_k$
  end if
end if

which is generically faster, since if $y \geq 1$ we do not need to extract a random number, an operation that is typically much slower than an `if-else` control.

[4]These sentences would obviously require more care, since single points have zero measure. From the operative point of view, $\mathbb{R}$ is represented on any physical CPU by a large but finite number of points, so this problem does not exist.

3.3.2 The heat-bath algorithm

We now discuss a different way of generating a Markov chain with preassigned invariant pdf, which can be applied whenever the state of the system is itself a set of several independent numbers which characterize some properties of the system (natural examples are positions and momenta of the particles in classical statistical mechanics). For reason that will become obvious this method is called heat-bath in the physics literature, or Gibbs sampler in mathematics and statistics.

Let us denote the state of the system by the couple (a, α), where a is one of the numbers which characterize the state (e. g. the position of one of the particles) and α collectively denotes all the other numbers needed to uniquely specify the state. The conditional probability of a given α is

$$P(a|\alpha) = \frac{\pi_{(a,\alpha)}}{\sum_{a'} \pi_{(a',\alpha)}} \ , \tag{3.3.16}$$

which is independent of the absolute normalization of the pdf $\pi(a, \alpha)$. The elementary step of the heath-bath algorithm consists in generating the new state (b, β) with probability

$$W_{(b,\beta)(a,\alpha)} = \delta_{\alpha\beta} P(b|\alpha) \ , \tag{3.3.17}$$

hence only the variable a is modified, sampling the conditional probability at fixed α, something that is assumed to be feasible. The name of the algorithm is due to the fact that the part α of the state acts as a heat-bath for the single variable a. Note that the heat-bath algorithm differs from the Metropolis(-Hastings) in one important aspect: there is no rejection.

Let us verify that the transition probability W defined in this way satisfies the detailed balance principle with respect to $\pi_{(a,\alpha)}$. We have indeed

$$\begin{aligned} W_{(b,\beta)(a,\alpha)}\pi_{(a,\alpha)} &= \delta_{\alpha\beta} P(b|\alpha)\pi_{(a,\alpha)} = \delta_{\alpha\beta} \frac{\pi_{(b,\alpha)}\pi_{(a,\alpha)}}{\sum_{b'} \pi_{(b',\alpha)}} \ , \\ W_{(a,\alpha)(b,\beta)}\pi_{(b,\beta)} &= \delta_{\beta\alpha} P(a|\beta)\pi_{(b,\beta)} = \delta_{\beta\alpha} \frac{\pi_{(a,\beta)}\pi_{(b,\beta)}}{\sum_{a'} \pi_{(a',\beta)}} = W_{(b,\beta)(a,\alpha)}\pi_{(a,\alpha)} \ , \end{aligned} \tag{3.3.18}$$

where the last equality is due to the presence of $\delta_{\alpha\beta}$.

The Markov chain generated by the heat-bath algorithm is aperiodic since there is a nontrivial possibility of remaining in the same state[5]. By randomly selecting at each iteration the number a to be updated, the Markov chain also becomes irreducible, and still satisfies the detailed balance condition (see the next subsection for more details on this point).

As a simple example of application of the heat-bath method let us consider a system whose state is a vector of N real numbers $x_1, \ldots, x_N$, and suppose that we want to sample the pdf

$$\pi(x_1, \ldots, x_N) \propto \exp\left(-\prod_i x_i^2\right) \ . \tag{3.3.19}$$

If we denote by $x_1^{(k)}, \ldots, x_N^{(k)}$ the state of the system at the k-th iteration, a MCMC heat-bath algorithm to sample $\pi(x_1, \ldots, x_N)$ is the following

1. select $i \in \{1, \ldots, N\}$ with uniform pdf
2. $x_j^{(k+1)} = x_j^{(k)}$ if $j \neq i$, while $x_i^{(k+1)}$ is generated by using the Box-Muller method (see Sec. 2.3) to sample the Gaussian

$$\sqrt{\frac{\pi}{A}} e^{-Ax^2} \ , \quad A = \prod_{j\neq i} (x_j^{(k)})^2 \ . \tag{3.3.20}$$

[5]Once again, for continuous distribution this would require more care.

3.3.3 Composition of Markov chains

Let us assume to have two different Markov chains, associated with the matrices $W^{(1)}$ and $W^{(2)}$. For any $0 \leq \alpha \leq 1$ we can define the new matrix W by

$$W_{ab} = \alpha W^{(1)}_{ab} + (1-\alpha) W^{(2)}_{ab} \ . \tag{3.3.21}$$

Clearly $W_{ab} \geq 0$, moreover

$$\sum_a W_{ab} = \alpha \sum_a W^{(1)}_{ab} + (1-\alpha) \sum_a W^{(2)}_{ab} = \alpha + 1 - \alpha = 1 \ , \tag{3.3.22}$$

hence W defined in this way is a stochastic matrix, which corresponds to the Markov chain whose elementary step is given by the following two operations

1. select $r \in [0,1)$ with uniform pdf,
2. if $r < \alpha$ apply $W^{(1)}$, else $W^{(2)}$.

The case $\alpha = 1/2$ obviously corresponds to the case in which $W^{(1)}$ and $W^{(2)}$ are selected randomly and with the same probability at each step.

It should be clear that if $0 < \alpha < 1$ and at least one between $W^{(1)}$ and $W^{(2)}$ is an irreducible aperiodic Markov chain, then W is an irreducible aperiodic Markov chain, since we have a finite probability of always selecting $W^{(1)}$ or $W^{(2)}$ in step 2. above. The same is true if, e. g., $W^{(1)}$ is irreducible and $W^{(2)}$ is aperiodic. It is also immediate to verify that if $W^{(1)}$ and $W^{(2)}$ satisfy the balance or the detailed balance condition, then the same is true for W.

Let us consider a different way in which two Markov chain can be composed: we can define W by

$$W_{ab} = (W^{(2)} W^{(1)})_{ab} = \sum_c W^{(2)}_{ac} W^{(1)}_{cb} \ , \tag{3.3.23}$$

and the elementary step of the associated Markov chain is

1. apply $W^{(1)}$,
2. apply $W^{(2)}$.

In this case the two Markov chain are not stochastically "mixed", but executed sequentially.

It is immediate to see that if $W^{(1)}$ and $W^{(2)}$ satisfy the balance condition with respect to the pdf π then also W does the same, however if $W^{(1)}$ and $W^{(2)}$ satisfy the detailed balance condition it is not generically true that W does the same. Indeed we have (in the equality (1) we use the detailed balance condition for $W^{(1)}$)

$$\begin{aligned} W_{ab}\pi_b &= \sum_c W^{(2)}_{ac} W^{(1)}_{cb} \pi_b \overset{(1)}{=} \sum_c W^{(2)}_{ac} W^{(1)}_{bc} \pi_c \ , \\ W_{ba}\pi_a &= \sum_c W^{(2)}_{bc} W^{(1)}_{ca} \pi_a \overset{(1)}{=} \sum_c W^{(2)}_{bc} W^{(1)}_{ac} \pi_c \ , \end{aligned} \tag{3.3.24}$$

and there is in general no reason for the two expression to coincide. Since the condition that is really necessary to ensure the validity of the MCMC algorithm is the balance condition, this is typically not a problem, however it is something to keep in mind if for some reason detailed balance is needed.

Even if $W^{(1)}$ and $W^{(2)}$ are associated with irreducible and aperiodic Markov chains, the composition $W = W^{(2)} W^{(1)}$ can be associated with a reducible Markov chain, as can be explicitly seen in the following example from [21]

$$W^{(1)} = \begin{pmatrix} 0 & 0 & 1/2 \\ 1 & 0 & 0 \\ 0 & 1 & 1/2 \end{pmatrix} \ , \quad W^{(2)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1/2 \\ 1 & 0 & 1/2 \end{pmatrix} \ , \tag{3.3.25}$$

$$W^{(2)} W^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/4 \\ 0 & 1/2 & 3/4 \end{pmatrix} \ . \tag{3.3.26}$$

$W^{(1)}$ and $W^{(2)}$ are irreducible and aperiodic, but $W^{(2)}W^{(1)}$ is clearly reducible. A sufficient, but quite difficult to realize, condition for W to be aperiodic and irreducible is clearly $W_{ab}^{(i)} > 0$ for any $a, b \in \Omega$ and for $i = 1, 2$.

Chapter 4

Data analysis for MCMC

We have seen in Sec. 3.2 that if a stochastic matrix W is associated with an irreducible and aperiodic Markov chain and p_a is any pdf on the state space Ω, we have

$$\sum_{a\in\Omega} |(W^k p)_a - \pi_a| \leq A\, e^{-k/\tau_{\rm exp}} \ , \tag{4.0.1}$$

where π_a is the (unique) invariant pdf of the Markov chain.

If $F : \Omega \to \mathbb{R}$ is a bounded function, and we are interested in computing the average value

$$\langle F\rangle = \sum_{a\in\Omega} F(a)\pi_a \ , \tag{4.0.2}$$

we can estimate $\langle F\rangle$ by using

$$\overline{F} = \frac{1}{N}\sum_{i=1}^{N} F(x_i) \ , \tag{4.0.3}$$

where x_i are the N states obtained by evolving the Markov chain associated to W, starting from a generic initial state x_0. To verify that this is a reliable prescription, let us compute $\langle\overline{F}\rangle_s$, where we denote by $\langle\ \ \rangle_s$ the average on the possible samples, i. e., the possible statistical outcomes of the Markov chain evolution; in $\langle\ \ \rangle_s$ the i-th draw of the sample is thus averaged with weight $(W^i p)_a$. If we introduce the notation $(W^k p)_a = \pi_a + R_a^{(k)}$, and use Eq. (4.0.1) and $|F(a)| \leq M$ for any $a \in \Omega$, we get

$$\begin{aligned} \big|\langle\overline{F}\rangle_s - \langle F\rangle\big| &= \left|\frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\Omega} F(a)(W^i p)_a - \langle F\rangle\right| = \left|\frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\Omega} F(a)R_a^{(j)}\right| \leq \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\Omega} |F(a)|\,|R_a^{(j)}| \leq \frac{M}{N}\sum_{i=1}^{N}\sum_{a\in\Omega} |R_a^{(j)}| \ \leq \frac{AM}{N}\sum_{i=1}^{N} e^{-i/\tau_{\rm exp}} \ . \end{aligned} \tag{4.0.4}$$

Moreover we have

$$\sum_{i=1}^{N} e^{-i/\tau_{\rm exp}} \leq \sum_{i=1}^{\infty} e^{-i/\tau_{\rm exp}} = \frac{e^{-1/\tau_{\rm exp}}}{1-e^{-1/\tau_{\rm exp}}} \ , \tag{4.0.5}$$

hence, finally,

$$\big|\langle\overline{F}\rangle_s - \langle F\rangle\big| \leq \frac{AM}{N}\,\frac{e^{-1/\tau_{\rm exp}}}{1-e^{-1/\tau_{\rm exp}}} \ . \tag{4.0.6}$$

We thus see that $\overline{F}$ is a biased estimator for $\langle F\rangle$, with a bias that vanishes as $1/N$ in the large sample limit.

To speed up the convergence of $\overline{F}$ to $\langle F\rangle$ it is customary, in Monte Carlo simulations, to discard the first $N_{\rm th} \approx$ few $\tau_{\rm exp}$ steps generated by the Markov Chain, which are the ones needed for the system to "thermalize". In this way the previous bound becomes

$$\left|\langle\overline{F}\rangle_s - \langle F\rangle\right| \leq \frac{AM}{N-N_{\rm th}} \sum_{i=N_{\rm th+1}}^{N} e^{-i/\tau_{\rm exp}} \leq \frac{AMe^{-(N_{\rm th}+1)/\tau_{\rm exp}}}{(N-N_{\rm th})(1-e^{-1/\tau_{\rm exp}})} \ . \tag{4.0.7}$$

It is important to note that this thermalization removal procedure is very useful in practice, however it is not needed from the purely theoretical point of view, nor it is really conclusive, since a significantly smaller but nonvanishing $1/N$ bias remains. The fundamental point to stress is however that a bias $O(1/N)$ is negligible with respect to the Monte Carlo statistical error, which approach zero as $O(1/\sqrt{N})$.

The $1/\sqrt{N}$ scaling of the statistical error should at this point sound reasonable, but it can not be obtained from the simplest form of the Central Limit Theorem discussed in Sec. 1.1, since that form assumed the draws to be independent, which is not the case for draws generated by using a Markov chain. The effect of autocorrelation is discussed in the next section.

4.1 Coping with autocorrelations in MCMC

4.1.1 The integrated autocorrelation time(s)

Due to the presence of autocorrelations, we can not use the simple expression Eq. (1.1.8) for the variance $\sigma^2_{\overline{F}}$ of the sample average $\overline{F}$. We have to start again from the basic definition of $\sigma^2_{\overline{F}}$:

$$\begin{aligned}\sigma^2_{\overline{F}} = \left\langle (\overline{F} - \langle F\rangle)^2 \right\rangle_s = \left\langle \left(\frac{1}{N}\sum_{i=1}^{N} F(x_i) - \langle F\rangle \right)^2 \right\rangle_s = \\ = \left\langle \left(\frac{1}{N}\sum_{i=1}^{N} (F(x_i) - \langle F\rangle) \right)^2 \right\rangle_s = \frac{1}{N^2}\sum_{i,j=1}^{N} \langle \delta F_i \, \delta F_j\rangle_s \ ,\end{aligned} \tag{4.1.1}$$

where in the last step we introduced the notation $\delta F_i = F(x_i) - \langle F\rangle$ and the average $\langle\ \ \rangle$ is computed with respect to the invariant pdf of the Markov chain.

Let us introduce $\sigma^2_F = \langle F^2\rangle - \langle F\rangle^2$, which for N large enough coincides with $\overline{\sigma}^2_F = \langle F^2\rangle_s - \langle F\rangle^2_s$. For independent draws we would have

$$\text{(independent draws)} \qquad \langle \delta F_i \, \delta F_j\rangle_s = \sigma^2_F \delta_{ij} \ , \tag{4.1.2}$$

and Eq. (1.1.8) would follow. In the general case it is convenient to introduce the autocorrelation function of F by

$$C_F(i,j) = \frac{\langle \delta F_i \, \delta F_j\rangle_s}{\sigma^2_F} \ , \tag{4.1.3}$$

so we can rewrite $\sigma^2_{\overline{F}}$ in the form

$$\sigma^2_{\overline{F}} = \frac{\sigma^2_F}{N^2}\sum_{i,j=1}^{N} C_F(i,j) \ . \tag{4.1.4}$$

It is now convenient to discuss some properties of the autocorrelation function $C_F(i,j)$ in the post-thermalization regime $i,j \gg \tau_{\rm exp}$, in which we can neglect the exponential corrections to the asymptotic pdf π_a. We have by definition

$$C_F(i,i) = 1 \ , \tag{4.1.5}$$

and from

$$2\delta F_i \delta F_j = (\delta F_i)^2 + (\delta F_j)^2 - (\delta F_i - \delta F_j)^2 = -(\delta F_i)^2 - (\delta F_j)^2 + (\delta F_i + \delta F_j)^2 \tag{4.1.6}$$

we get (using $\langle (\delta F_i)^2 \rangle_s = \langle (\delta F_j)^2 \rangle_s$ for $i, j \gg \tau_{\rm exp}$)

$$-\langle (\delta F_i)^2 \rangle_s \leq \langle \delta F_i \, \delta F_j \rangle_s \leq \langle (\delta F_i)^2 \rangle_s \ , \tag{4.1.7}$$

hence

$$-1 \leq C_F(i,j) \leq 1 \ . \tag{4.1.8}$$

If we denote by z the state of the Markov chain at $t = 0$ and assume $i > j$, the probability of having state a at time $t = j$ and state b at time $t = i$ is,

$$(W^{i-j})_{ba} (W^j)_{az} = (W^{i-j})_{ba} \pi_a + (W^{i-j})_{ba} R_a^{(j)} \simeq (W^{i-j})_{ba} \pi_a \ , \tag{4.1.9}$$

where in the last step we assumed once again $j \gg \tau_{\rm exp}$ and neglected the exponentially small correction due to $R_a^{(j)}$. Using this expression in the autocorrelation function we have

$$C_F(i,j) = \frac{1}{\sigma_F^2} \langle \delta F_i \, \delta F_j \rangle_s = \frac{1}{\sigma_F^2} \sum_{ab} (W^{i-j})_{ba} \pi_a \delta F_a \, \delta F_b = C_F(i+k, j+k) \tag{4.1.10}$$

for any $k \geq 0$. With analogous manipulations, assuming $i, j \gg \tau_{\rm exp}$, we also find

$$C_F(i,j) = C_F(j,i) \ , \tag{4.1.11}$$

which together with the previous identity shows that $C_F(i,j)$ depends only on $|i-j|$. With a clear abuse of notation we can thus write $C_F(i,j) = C_F(|i-j|)$.

Let us now investigate the behavior of $C_F(i,j)$ for large $|i-j|$ (and always $i, j \gg \tau_{\rm exp}$): if as before we denote by z the state of the Markov chain at $t = 0$ and assume $i > j$, the probability of having state a at time $t = j$ and state b at time $t = i$ is

$$(W^{i-j})_{ba} (W^j)_{az} = (W^{i-j})_{ba} \pi_a + (W^{i-j})_{ba} R_a^{(j)} \simeq (W^{i-j})_{ba} \pi_a = \pi_b \pi_a + R_{ba}^{(i-j)} \pi_a \ , \tag{4.1.12}$$

hence

$$\begin{aligned} |\langle \delta F_i \, \delta F_j \rangle_s| &= \left| \sum_{ab} \pi_a \pi_b \delta F_a \, \delta F_b + \sum_{ab} R_{ba}^{(i-j)} \pi_a \delta F_a \, \delta F_b \right| \leq \\ &\leq \sum_{ab} |R_{ba}^{(i-j)} \pi_a \delta F_a \, \delta F_b| = O(e^{-(i-j)/\tau_{\rm exp}}) \ , \end{aligned} \tag{4.1.13}$$

where we used $\sum_a \pi_a \delta F_a = \langle \delta F \rangle = 0$ and the exponential convergence to π_a of $(W^k)_{ab}$. We thus finally have

$$|C_F(i,j)| \leq B e^{-|i-j|/\tau_{\rm exp}} \ . \tag{4.1.14}$$

After this intermezzo on the properties of the autocorrelation function we can go back to our original aim, the computation of $\sigma_{\overline{F}}^2$. We have

$$\begin{aligned} \sigma_{\overline{F}}^2 &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \langle \delta F_i \, \delta F_j \rangle_s = \frac{\sigma_F^2}{N^2} \sum_{i=1}^N \sum_{j=1}^N C_F(i,j) = \frac{\sigma_F^2}{N^2} \sum_{i=1}^N \sum_{j-i} C_F(i,j) \overset{(1)}{\simeq} \\ &\simeq \frac{\sigma_F^2}{N^2} \sum_{i=1}^N \sum_{j-i} C_F(|i-j|) \overset{(2)}{\simeq} \frac{\sigma_F^2}{N^2} \sum_{i=1}^N \sum_{k=-\infty}^{+\infty} C_F(|k|) = \frac{\sigma_F^2}{N} \sum_{k=-\infty}^{+\infty} C_F(|k|) \ , \end{aligned} \tag{4.1.15}$$

where in (1) we neglected $O(\tau_{\rm exp}/N^2)$ terms coming from $1 \leq i, j \lesssim \tau_{\rm exp}$, while in (2) we assumed $N \gg \tau_{\rm exp}$ and neglected terms exponentially small in N. If we now define the integrated autocorrelation time of the observable F by

$$\tau_{\rm int}^{(F)} = \sum_{k=1}^{\infty} C_F(k) \ , \tag{4.1.16}$$

we have finally

$$\sigma_{\overline{F}}^2 = \frac{\sigma_F^2}{N}(1 + 2\tau_{\rm int}^{(F)}) \ . \tag{4.1.17}$$

Pay attention to the fact that slightly different definitions of the integrated autocorrelation time exist in the literature. The moral is that, when autocorrelations are present, the effective sample size is reduced from N to $N/(1 + 2\tau_{\rm int}^{(F)})$.

It is important to stress that $\tau_{\rm exp}$ and $\tau_{\rm int}^{(F)}$ are conceptually two very different objects. On one hand $\tau_{\rm exp}$ is the largest characteristic time of the MCMC evolution, and it is the typical time needed to thermalize the system. On the other hand $\tau_{\rm int}^{(F)}$ depends on the observable F, and it is related to the timescale of the fluctuations of F in the thermalized part of the Markov chain evolution. It is nevertheless possible to show that $\tau_{\rm exp}$ is an upper bound of all the integrated autocorrelation times.

We now show, following [6], that $\tau_{\rm int}^{(F)} \leq \tau_{\rm exp}$ when detailed balance is satisfied. We have seen in Lemma 3.3.1 that if a Markov chain satisfies the detailed balance, then the transpose of its associated stochastic matrix W is Hermitian with respect to the scalar product

$$(u, v) = \sum_a \pi_a u_a v_a \ . \tag{4.1.18}$$

Using Eq. (4.1.10) we can write (assuming $i > j$)

$$\langle \delta F_i \, \delta F_j \rangle_s = \sum_{ab} (W^{i-j})_{ba} \pi_a \delta F_a \, \delta F_b = (\delta F, ({}^tW)^{i-j} \delta F) \ , \tag{4.1.19}$$

and thus

$$\sigma_F^2 = (\delta F, \delta F) \ , \quad C_F(k) = \frac{(\delta F, ({}^tW)^k \delta F)}{(\delta F, \delta F)} \ . \tag{4.1.20}$$

If we denote by $v_a^{(j)}$ the j-th eigenvector of ${}^tW$, from $\langle \delta F_i \rangle_s = 0$ it follows that δF has no component along the eigenvector associated with the eigenvalue 1 (see Theorems 3.2.3-3.2.4), so $\delta F_a = \sum_{j>0} c^{(j)} v_a^{(j)}$ (the $j = 0$ eigenvalue is $\lambda = 1$) and from $\lambda_j \in (-1, 1)$ if $j \neq 0$ we have

$$\sum_{k=1}^{\infty} (\delta F, ({}^tW)^k \delta F) = \sum_{k=1}^{\infty} \sum_{a,j} \pi_a (c^{(j)})^2 \lambda_j^k (v_a^{(j)})^2 = \sum_{a,j} \pi_a (c^{(j)})^2 \frac{\lambda_j}{1 - \lambda_j} (v_a^{(j)})^2 \leq \frac{\Lambda'}{1 - \Lambda'} (\delta F, \delta F) \ , \tag{4.1.21}$$

where $\Lambda' = \max_{j>0} \lambda_j$ and we used the fact that $x/(1 - x)$ is an increasing function on $(-1, 1)$. We thus have (see Eq. (4.1.16))

$$\tau_{\rm int}^{(F)} \leq \frac{\Lambda'}{1 - \Lambda'} \ , \tag{4.1.22}$$

and clearly (see Eq. (3.2.23))

$$\Lambda' \leq \max_{j>0} |\lambda_j| = e^{-1/\tau_{\rm exp}} \ , \tag{4.1.23}$$

hence

$$\tau_{\rm int}^{(F)} \leq \frac{e^{-1/\tau_{\rm exp}}}{1 - e^{-1/\tau_{\rm exp}}} \ . \tag{4.1.24}$$

Moreover the last expression is $\leq \tau_{\rm exp}$ and, when $\tau_{\rm exp} \gg 1$, it approaches $\tau_{\rm exp}$.

We have computed $\sigma_{\overline{F}}^2$, and to conclude this section we have to discuss the statistical distribution of $\overline{F}$. We thus recall one of the possible versions of the Central Limit Theorem for correlated random variables (see, e. g., [4] §5.27, or [14] §8.3 for a different formulation), which can be stated as follows: if $X_1, X_2, \ldots$ is a succession of dependent random variables, whose autocorrelation function $\langle X_i X_{i+k} \rangle - \langle X_i \rangle \langle X_{i+k} \rangle$ vanishes at least $O(k^{-5})$, with $\langle X_i \rangle = 0$ and finite $\langle X_i^{12} \rangle$, then the variance of $S_N = X_1 + \cdots + X_N$ satisfies

$$\frac{1}{N} \sigma_{S_N}^2 \to \sigma^2 = \langle X_1^2 \rangle + 2 \sum_{k=1}^{\infty} \langle X_1 X_{1+k} \rangle \ , \tag{4.1.25}$$

and if $\sigma > 0$ then $S_N/(\sqrt{N}\sigma)$ converges to a normal Gaussian distribution. The outcome of this theorem is thus that in a MCMC simulation, in the large sample limit, $\overline{F}$ is distributed with a Gaussian pdf and variance given by Eq. (4.1.17).

4.1.2 Binning/blocking

It is possible to directly use Eq. (4.1.17) to estimate $\sigma^2_{\overline{F}}$, however there are some subtleties that have to be taken into account when doing this, which are discussed in [22] (see also [23] and, for some background material, [24] §5.3, 6.2). For this reason a more indirect but straightforward procedure is usually adopted, which goes under the names of binning or blocking, likely introduced for the first time in [25], and systematized in [26].

Our aim is to numerically estimate the variance of $\overline{F}$ defined by

$$\overline{F} = \frac{1}{N}\sum_{i=1}^{N} F(x_i) \ , \tag{4.1.26}$$

where the x_is are obtained by evolving a Markov chain. Let k be a positive natural number and let us assume, for the sake of the simplicity, that k divides N; if this is not the case it is sufficient to consider the first[1] $k\lfloor N/k \rfloor$ elements of the sample. We define a new sample composed of N/k elements by averaging blocks of size k as follows:

$$F_i^{(k)} = \frac{1}{k}\big(F(x_{ki+1}) + F(x_{ki+2}) + \cdots + F(x_{ki+k})\big) \ , \quad i = 1, \dots, N/k \ , \tag{4.1.27}$$

and we obviously have $\overline{F} = \overline{F^{(k)}}$, where

$$\overline{F^{(k)}} = \frac{1}{N/k}\sum_{i=1}^{N/k} F_i^{(k)} \ . \tag{4.1.28}$$

If we compute the variance of $\overline{F^{(k)}}$ as if the $F_i^{(k)}$ elements were independent, using Eq. (1.1.9), we get (assuming $N \gg k$)

$$\overline{\sigma}^2_{\overline{F^{(k)}}} = \frac{1}{N/k}\frac{1}{N/k-1}\sum_{i=1}^{N/k}\Big(F_i^{(k)} - \overline{F}\Big)^2 \simeq \frac{k^2}{N^2}\sum_{i=1}^{N/k}\frac{1}{k^2}\big(\delta F_{ki+1} + \cdots + \delta F_{ki+k}\big)^2 \ , \tag{4.1.29}$$

where $\delta F_j = F(x_j) - \overline{F}$. Moreover we have

$$\begin{aligned}\big(\delta F_{ki+1} + \cdots + \delta F_{ki+k}\big)^2 = &\sum_{j=1}^{k}(\delta F_{ki+j})^2 + 2\sum_{j=1}^{k-1}\delta F_{ki+j}\delta F_{ki+j+1} + \\ &+ 2\sum_{j=1}^{k-2}\delta F_{ki+j}\delta F_{ki+j+2} + \cdots + 2\delta F_{ki+1}\delta F_{ki+k} \ ,\end{aligned} \tag{4.1.30}$$

and if k is large enough we can rewrite these sums as sample averages defining the correlation function, hence (to be formally correct we should write $\overline{C_F}$ for the sample estimator of C_F)

$$\big(\delta F_{ki+1} + \cdots + \delta F_{ki+k}\big)^2 = k\overline{\sigma}_F^2 + 2(k-1)\overline{\sigma}_F^2 C_F(1) + 2(k-2)\overline{\sigma}_F^2 C_F(2) + \cdots \ . \tag{4.1.31}$$

Since the correlation function $C_F(j)$ decays exponentially for large j, if k is large enough (in the worst case large with respect to τ_{exp}) we have

$$\big(\delta F_{ki+1} + \cdots + \delta F_{ki+k}\big)^2 \simeq k\overline{\sigma}_F^2\left(1 + 2\sum_{j=1}^{\infty} C_F(j)\right) = k\overline{\sigma}_F^2(1 + 2\tau_{\text{int}}^{(F)}) \ . \tag{4.1.32}$$

Using this expression in Eq. (4.1.29) we finally get, if k is large enough

$$\overline{\sigma}^2_{\overline{F^{(k)}}} = \frac{k^2}{N^2}\sum_{i=1}^{N/k}\frac{1}{k^2}k\overline{\sigma}_F^2(1 + 2\tau_{\text{int}}^{(F)}) = \frac{\overline{\sigma}_F^2}{N}(1 + 2\tau_{\text{int}}^{(F)}) \ , \tag{4.1.33}$$

[1] For $x \in \mathbb{R}$ the floor function $\lfloor x \rfloor$ is the largest $n \in \mathbb{Z}$ such that $n \leq x$.

Algorithm 6 Possible MCMC algorithm to sample a normal Gaussian distribution. The starting point x_0 has been fixed to 5 to clearly visualize the thermalization process.

```
x_0 = 5
loop
    select x̄ ∈ (x_k − δ, x_k + δ) with uniform pdf
    y = −½ x̄² + ½ x_k²
    if y ≥ 0 then
        x_{k+1} = x̄
    else
        select r ∈ [0, 1) with uniform pdf
        if r ≤ min[1, e^y] then
            x_{k+1} = x̄
        else
            x_{k+1} = x_k
        end if
    end if
end loop
```

which coincides with Eq. (4.1.17) found in the previous section.

We thus have a simple operative way of computing $\overline{\sigma}^2_{\overline{F}}$ (i. e. the sample estimate of $\sigma^2_{\overline{F}}$): for several k values define the blocked averages as in Eq. (4.1.27), and compute the *naive* sample variances $\overline{\sigma}^2_{\overline{F^{(k)}}}$, as if the blocked variables were independent. The values $\overline{\sigma}^2_{\overline{F^{(k)}}}$, as a function of k, will saturate for large k at a value that is the correct estimate of $\overline{\sigma}^2_{\overline{F}}$. Note that this method works well when the value of k for which $\overline{\sigma}^2_{\overline{F}}$ saturates is small enough with respect to the sample size N, otherwise the error of $\overline{\sigma}^2_{\overline{F}}$ gets large, making the estimated values oscillate widely as a function of k.

4.1.3 An explicit example

We now present a complete example of MCMC generation and data analysis for the simple case already discussed in Sec. 3.3.1, i. e. for the sampling of a one dimensional distribution. For the sake of the simplicity we consider the case of the normal Gaussian distribution.

A possible MCMC algorithm to sample a normal Gaussian distribution is shown in Alg. (6), and the parameters of this algorithm are the starting point x_0 and the value of δ. We chose $x_0 = 5$ as the starting point, in order to better visualize the thermalization process, since random points extracted from the Gaussian pdf will most likely lie in $[-2, 2]$. For what concern δ we will use several values, in order to investigate how the choice of δ affects the efficiency of the algorithm, measured by the statistical accuracy that can be achieved at fixed CPU time. We thus generated, using the algorithm Alg. (6), 10^8 draws for several values of δ in the range between 0.1 and 50 (which required about 25s of CPU time for each δ).

In Fig. 4.1 the typical behavior of the beginning of a MC history is shown, for $\delta = 1$ and $\delta = 0.2$: both the histories start from $x_0 = 5$, then they drift toward zero (which is the average of the pdf we are sampling) and start to oscillate, with oscillations whose typical amplitude is related to the standard deviation of the invariant pdf (which in the present case is 1). Already looking at this figure it should be clear that data obtained by using $\delta = 1$ are less correlated than data generated using $\delta = 0.2$, hence $\delta = 1$ is numerically more efficient.

In Fig. 4.2 (left) we show the estimated autocorrelation function

$$C_x(n) = \frac{\langle x_i x_{i+n} \rangle_s}{\langle x_i^2 \rangle_s} \tag{4.1.34}$$

of the draws x_n, computed after removing the first 10^6 draws of each sample (in this way we are significantly overestimating the thermalization time, but we had enough statistics not to worry

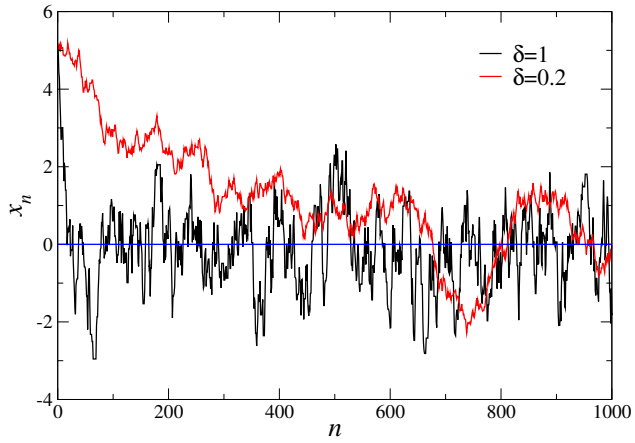

Figure 4.1: Two Monte Carlo histories obtained by performing 1000 loops of the algorithm Alg. (6), for $\delta = 1$ and $\delta = 0.2$.

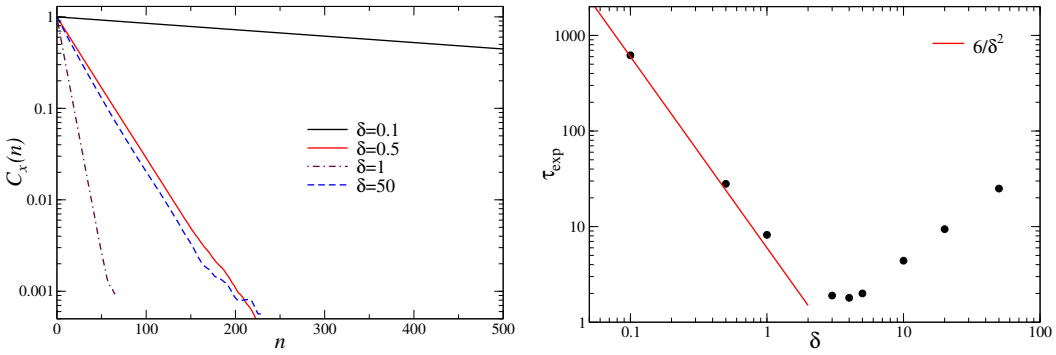

Figure 4.2: (left) Autocorrelation function $C_x(n)$ of the numbers obtained using the algorithm Alg. (6) for several values of the parameter δ. (right) The fitted exponential autocorrelation time as a function of δ.

about it). Autocorrelation functions are well described by a simple exponential behavior starting practically from $n = 0$, and it is thus simple to estimate τ_{exp} by performing a fit. Note however that the values of the autocorrelation function for different time separations have been estimated from the same sample, hence they are correlated. For this reason a simple uncorrelated fit provides a reasonable estimate of τ_{exp} but can not be used to estimate its uncertainty. If a reliable uncertainty is needed a correlated fit has to be used. In Fig. 4.2 (right) we report the exponential autocorrelation time estimated for all the valued of δ simulated. As was already clear from Fig. 4.2 (left) τ_{exp} is very large for small values of δ, it decreases by increasing δ until it reaches a minimum for $\delta \approx 4$ (where $\tau_{\text{exp}} \approx 2$), then it increases again.

This behavior is quite typical and can be easily explained: for $\delta \ll 1$ the trial state $\bar{x}$ is always very close to the previous state x_k (the typical scale of the "distance" being the standard deviation of the pdf we are sampling, in this case 1), so it will be almost always accepted, but a large number of steps will be needed to decorrelate, hence τ_{exp} is large. Since almost every update is accepted, we can approximate the motion of the state by a random walk, and in a random walk the typical distance traveled in a time t is proportional to $\sqrt{t}$. We thus expect τ_{exp} to scale $O(\delta^{-2})$ for $\delta \ll 1$, since $O(\delta^{-2})$ steps are needed to travel an $O(1)$ distance in the configuration space. By increasing δ the acceptance probability decreases, but as far as $\delta \approx 1$ its scaling with δ is still quite mild, however for $\delta \approx 1$ two consecutive draws are almost independent of each other, since their typical distance is of the same order of the standard deviation of the pdf. Hence τ_{exp} reaches a minimum

for $\delta \approx 1$. If we consider the $\delta \gg 1$ limit we find a situation that is the dual of that found for $\delta \ll 1$: two consecutive draws will be practically independent from each other, however it will be very difficult for a draw to be accepted, since it is generated uniformly in (approximately) $(-\delta, \delta)$, and the pdf is concentrated in $(-1, 1)$. The typical acceptance probability will scale as $1/\delta$ and thus we expect $\tau_{\rm exp} = O(\delta)$ for $\delta \gg 1$, since one draw every $O(\delta)$ is accepted. Both these asymptotic behaviors are consistent with data reported in Fig. 4.2 (right).

The acceptance probabilities of the Metropolis accept/reject step for the simulations performed at the different values of δ are the following

δ	50	20	10	5	4	3	1	0.5	0.1
acc. prob.	0.032	0.080	0.160	0.317	0.390	0.492	0.804	0.901	0.980

and a general rule of thumb is that the acceptance probability should be in the range $30\% \lesssim$ acc. prob. $\lesssim 70\%$ for the exponential autocorrelation time to be reasonable. For computationally intensive problems it is however in general convenient to perform a preliminary study of the behavior of $\tau_{\rm exp}$ as a function of the simulation parameters, in order to optimize the resource usage.

For the simple case of MCMC sampling of the normal Gaussian the previous reasoning can be easily made quantitative in the case $\delta \ll 1$ [27, 28]: we have seen that the autocorrelation $C_x(n)$ is exponential practically starting from $n = 0$, and the autocorrelation after one step is (remember that $\sigma_x^2 = 1$)

$$\begin{aligned} C_x(1) = \langle x_i x_{i+1} \rangle_s &= \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \int_{-\delta}^{+\delta} \frac{\mathrm{d}y}{2\delta} x \left[(x+y) P_{acc}(x \to x+y) + x(1 - P_{acc}(x \to x+y)) \right] = \\ &= \frac{1}{2\delta\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{d}x e^{-\frac{1}{2}x^2} \int_{-\delta}^{+\delta} \mathrm{d}y \, x(x + y P_{acc}(x \to x+y)) = \\ &= 1 + \frac{1}{2\delta\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{d}x e^{-\frac{1}{2}x^2} \int_{-\delta}^{+\delta} \mathrm{d}y \, xy P_{acc}(x \to x+y) \ , \end{aligned} \tag{4.1.35}$$

where $P_{acc}(x \to x+y)$ is given by

$$P_{acc}(x \to x+y) = \min \left[1, \exp\left(-\frac{1}{2}(x+y)^2 + \frac{1}{2}x^2 \right) \right] \ . \tag{4.1.36}$$

If we consider the limit $\delta \ll 1$ we can consider only the cases in which x and $x+y$ have the same sign. If they are both positive we can approximate (since $|y| \leq \delta \ll 1$)

$$P_{acc}(x \to x+y) \simeq \begin{cases} 1 & y < 0 \\ 1 - xy & y > 0 \end{cases} \ , \tag{4.1.37}$$

hence

$$\int_{-\delta}^{+\delta} \mathrm{d}y \, xy P_{acc}(x \to x+y) \simeq \int_{-\delta}^{0} xy \mathrm{d}y + \int_{0}^{\delta} xy(1 - xy) \mathrm{d}y = -x^2 \frac{\delta^3}{3} \ . \tag{4.1.38}$$

The same result is obtained also when x and $x+y$ are both negative, thus we obtain

$$C_x(1) \simeq 1 - \frac{1}{2\delta\sqrt{2\pi}} \frac{\delta^3}{3} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} \mathrm{d}x = 1 - \frac{\delta^2}{6} \ , \tag{4.1.39}$$

and using $C_x(n) = e^{-n/\tau_{\rm exp}}$ for $n = 1$ and $\tau_{\rm exp} \gg 1$ we finally get $\tau_{\rm exp} \simeq 6/\delta^2$, which is also shown in Fig. 4.2 (right).

We now consider the numerical evaluation of the moments of the normal Gaussian pdf. In particular we consider for example $\langle x \rangle$, $\langle x^2 \rangle$ and $\langle x^4 \rangle$, whose values are obviously analytically known and are 0, 1, and 3, respectively. The first step for estimating these numbers is the computation of the corresponding sample averages by using the Monte Carlo samples generated (also in this case we discard the first 10^6 draws).

The nontrivial (but fundamental!) part is to estimate also the variance of these sample averages, which requires the use of blocking, due to the autocorrelation of MC data. For several values of the block size k we thus have to build the blocked samples, as in Eq. (4.1.27), using the functions $F(x) = x$, $F(x) = x^2$ and $F(x) = x^4$. Then we have to compute the *naive* (i. e., neglecting autocorrelations) standard deviation of the average of these blocked samples by using Eq. (4.1.29), and study the dependence of the result on the block size.

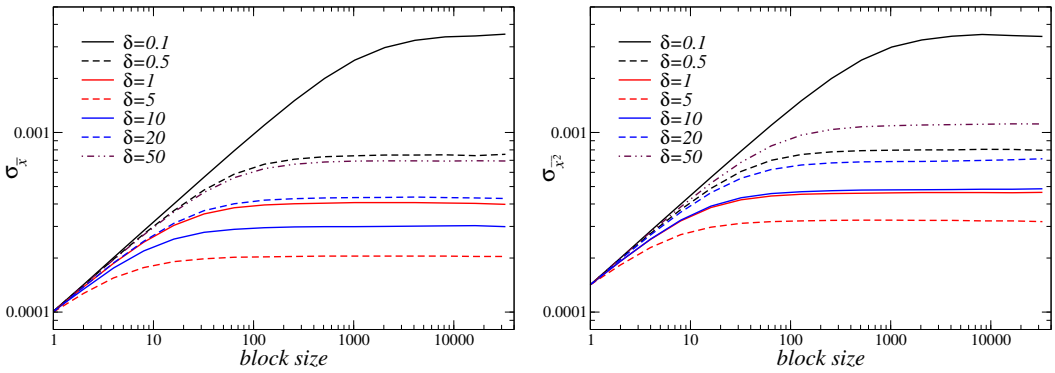

Figure 4.3: (left) Blocking analysis of $\sigma_{\overline{x}}$ for the output of the algorithm Alg. (6) for several values of the parameter δ. (right) Blocking analysis of $\sigma_{\overline{x^2}}$ for the output of the algorithm Alg. (6) for several values of the parameter δ.

The outcomes of this analysis are shown in Fig. (4.3) for some values of δ and for the cases of the first and of the second momentum (the results for the fourth one are completely analogous). In both the cases the standard deviation of the mean of the blocked variables grows as a power-law in the block size when the block size is not large enough, then it saturates and becomes approximately independent of the block size. This plateau value, as discussed in Sec. 4.1.2, is the correct estimation of the error to be associated with the sample average. Note that in the present case the gathered statistic is very large with respect to the exponential autocorrelation time (in the worst case $\tau_{\rm exp}$ is ≈ 620, while the sample size after thermalization is 0.99×10^8), so the curves shown in Fig. (4.3) are very smooth. In more realistic cases oscillations are present, and the plateau is not an horizontal straight line, but rather a line which oscillate randomly around a constant value. The amplitude of these oscillations is related the error to be associated with the standard error of the average.

Using the plateau values we obtain the estimates reported in Tab. (4.1) for the first, second and fourth momenta of the normal Gaussian distribution, which are obviously consistent with theoretical expectations. By looking at these values we can see that, since the gathered statistics are the same for all the cases, the integrated autocorrelation times $\tau_{\rm int}^{(F)}$ have the same behavior of the exponential autocorrelation time $\tau_{\rm exp}$, being larger for very small and very large values of δ. In case an estimate of $\tau_{\rm int}^{(F)}$ is needed, it can be obtained from Eq. (4.1.33): $1 + 2\tau_{\rm int}^{(F)}$ is given by the ratio of two $\sigma^2_{\overline{F^{(k)}}}$ values, one computed using a large block size k (i. e., a block size which corresponds to the plateau) and the other computed for $k = 1$.

δ	$\langle x \rangle$	$\langle x^2 \rangle$	$\langle x^4 \rangle$
50	-0.00056(70)	0.9998(11)	2.9970(67)
20	0.00058(42)	0.99853(68)	2.9923(41)
10	0.00024(29)	0.99948(47)	2.9975(28)
5	0.00040(20)	0.99943(32)	2.9959(20)
4	-0.00006(19)	0.99976(29)	3.0005(19)
3	-0.00016(20)	1.00000(28)	2.9999(20)
1	-0.00008(40)	1.00013(45)	3.0004(32)
0.5	0.00004(75)	1.00008(80)	3.0009(54)
0.1	-0.0030(35)	1.0009(35)	3.002(22)

Table 4.1: Numerical results obtained by using Alg. (6) to extract 10^8 draws.

4.2 Estimating secondary observables

We have considered up to now the so called "primary" observables, i. e., those observables that can be written as average values. There is, however, also another important class of observables, the so called "secondary" observables, which are functions of one or more average values, like e. g.

$$U_4 = \frac{\langle x^4 \rangle}{\langle x^2 \rangle^2} \ . \tag{4.2.1}$$

A natural estimator for this quantity is obviously

$$\overline{U_4} = \frac{\overline{x^4}}{\left(\overline{x^2}\right)^2} \ , \tag{4.2.2}$$

however, when using such an expression, we have to face two different problems. The first problem is related to the presence of a bias in the previous estimator, however it is easily seen that such a bias is $O(1/N)$ and hence subdominant with respect to the statistical errors; for this reason this theoretical problem is practically irrelevant in MC simulations. The second problem is instead more serious, and it is related once again to the estimation of the uncertainty. Using blocking we are taking into account the autocorrelations of data generated using the MCMC approach, however in computing the uncertainty to be associated with Eq. (4.2.2) we face a new problem. Had $\overline{x^4}$ and $\overline{x^2}$ be computed using two independent MCMC we could combine their uncertainties by using standard error propagation. However in standard circumstances both these quantities are estimated by using the same statistical sample, hence their statistical uncertainties are correlated.

Let us start by discussing the first problem. If we are interested in evaluating $F(\langle x \rangle)$, we can estimate the bias of the estimator $F(\overline{x})$ using the following reasoning. The typical fluctuation of $\overline{x}$ around $\langle x \rangle$ is $\sigma_x/\sqrt{N}$, where σ_x is the standard deviation of the variable x and N is the number of (independent) samples used to estimate $\overline{x}$. If N is large enough we can use a Taylor expansion to get

$$\begin{aligned} \langle F(\overline{x}) \rangle &= \langle F(\langle x \rangle) \rangle + \langle F'(\langle x \rangle)(\overline{x} - \langle x \rangle) \rangle + \frac{1}{2} \langle F''(\langle x \rangle)(\overline{x} - \langle x \rangle)^2 \rangle + \cdots \simeq \\ &\simeq F(\langle x \rangle) + \frac{1}{2} F''(\langle x \rangle) \sigma^2_{\overline{x}} = F(\langle x \rangle) + \frac{1}{2} F''(\langle x \rangle) \frac{\sigma^2_x}{N} \ , \end{aligned} \tag{4.2.3}$$

where $\sigma^2_{\overline{x}}$ is the variance of the sample average $\overline{x}$, and in the last step we used Eq. (1.1.8). As anticipated, the bias is $O(1/N)$ and thus negligible, in the large sample limit, with respect to the statistical error $O(1/\sqrt{N})$.

We now discuss the more serious problem of correlations: let A and B be two primary observables and let us suppose that we need to evaluate $F(\langle A \rangle, \langle B \rangle)$ (the discussion can be oviously extended to more general cases). The uncertainty to be associated with $F(\overline{A}, \overline{B})$, is the square root of the variance of the stochastic variable $F(\overline{A}, \overline{B})$, which is defined as usual by

$$\langle F(\overline{A}, \overline{B})^2 \rangle - \langle F(\overline{A}, \overline{B}) \rangle^2 \ . \tag{4.2.4}$$

Proceeding as for the case of the bias, we can approximate

$$F(\overline{A}, \overline{B}) \simeq F + F'_A \delta\overline{A} + F'_B \delta\overline{B} + \frac{1}{2} F''_{AB} \delta\overline{A}\, \delta\overline{B} + \frac{1}{2} F''_{AA} (\delta\overline{A})^2 + \frac{1}{2} F''_{BB} (\delta\overline{B})^2 \ , \tag{4.2.5}$$

where all functions are computed at $\langle A \rangle$, $\langle B \rangle$ and we introduced the notation $\delta\overline{A} = \overline{A} - \langle A \rangle$, and analogously for $\delta\overline{B}$. We thus have

$$\langle F(\overline{A}, \overline{B}) \rangle^2 \simeq F^2 + F\Big(F''_{AB} \langle \delta\overline{A} \delta\overline{B} \rangle + F''_{AA} \langle (\delta\overline{A})^2 \rangle + F''_{BB} \langle (\delta\overline{B})^2 \rangle \Big) \ , \tag{4.2.6}$$

and

$$\begin{aligned} \langle F(\overline{A}, \overline{B})^2 \rangle \simeq F^2 &+ (F'_A)^2 \langle (\delta\overline{A})^2 \rangle + (F'_B)^2 \langle (\delta\overline{B})^2 \rangle + 2 F'_A F'_B \langle \delta\overline{A}\, \delta\overline{B} \rangle + \\ &+ F\Big(F''_{AB} \langle \delta\overline{A} \delta\overline{B} \rangle + F''_{AA} \langle (\delta\overline{A})^2 \rangle + F''_{BB} \langle (\delta\overline{B})^2 \rangle \Big) \ , \end{aligned} \tag{4.2.7}$$

from which finally

$$\langle F(\overline{A},\overline{B})^2\rangle - \langle F(\overline{A},\overline{B})\rangle^2 = (F'_A)^2\langle(\delta\overline{A})^2\rangle + (F'_B)^2\langle(\delta\overline{B})^2\rangle + 2F'_AF'_B\langle\delta\overline{A}\,\delta\overline{B}\rangle\ . \tag{4.2.8}$$

If the fluctuations of $\overline{A}$ and $\overline{B}$ are independent, $\langle\delta\overline{A}\delta\overline{B}\rangle = 0$, we recover the standard formula of the error propagation, however this is the correct expression to be used also when correlations are present.

If we have no information on the covariance $\langle\delta\overline{A}\delta\overline{B}\rangle$ we can only put an upper bound on the true uncertainty: using the Schwartz inequality

$$|\langle\delta\overline{A}\delta\overline{B}\rangle| \le \sqrt{\langle(\delta\overline{A})^2\rangle}\sqrt{\langle(\delta\overline{B})^2\rangle} \tag{4.2.9}$$

we have indeed

$$\langle F(\overline{A},\overline{B})^2\rangle - \langle F(\overline{A},\overline{B})\rangle^2 \le \left(|F'_A|\sqrt{\langle(\delta A)^2\rangle} + |F'_B|\sqrt{\langle(\delta B)^2\rangle}\right)^2\ . \tag{4.2.10}$$

The use of this formula, however, largely overestimates the error in typical cases. Let us consider the example discussed in Sec. 4.1.3 and the secondary observable $\langle x^4\rangle/\langle x^2\rangle^2$ for $\delta = 50$: using data in Tab. (4.1) we get for the error the upper bound ($F(x_1,x_2) = x_1/x_2^2$, and $F'_A = 1$, $F'_B = -6$ when using the average values $x_1 = \langle x^4\rangle = 3$ and $x_2 = \langle x^2\rangle = 1$)

$$\overline{\sigma}_{\overline{U_4}} \le 0.0067 + 6\times 0.0011 = 0.0133\ . \tag{4.2.11}$$

If we wrongly assume that the errors of numerator and denominator are independent we get instead

$$\overline{\sigma}_{\overline{U_4}} \stackrel{?}{=} \sqrt{0.0067^2 + 6^2\times 0.0011^2} \simeq 0.0094\ . \tag{4.2.12}$$

Finally, the true uncertainty, obtained by using the methods discussed in the following two subsections, is

$$\overline{\sigma}_{\overline{U_4}} = 0.0032\ , \tag{4.2.13}$$

and the final estimate is $U_4 = 2.9983(32)$. This happens because the fluctuations of $\overline{x^4}$ and $\overline{x^2}$ are obviously strongly correlated, and in this case, with $2F'_AF'_B = -12$, we can estimate *a posteriori*

$$\langle\delta\overline{A}\delta\overline{B}\rangle \simeq 0.88\sqrt{\langle(\delta\overline{A})^2\rangle}\sqrt{\langle(\delta\overline{B})^2\rangle}\ . \tag{4.2.14}$$

In principle nothing prevents us from using Eq. (4.2.8) to asses the uncertainty of $F(\overline{A},\overline{B})$, since the covariance $\langle\delta\overline{A}\delta\overline{B}\rangle$ can be straightforwardly estimated. The problem with Eq. (4.2.8) is that it requires the computation of a significant number of derivatives and covariances if the function F depends on several primary observables, and its numerical implementation thus becomes quite baroque. To avoid these problems we can use the so called "plug-in estimators", which are defined by an algorithm in which the specific form of F enters only parametrically, without the need of computing the derivatives and covariances appropriate for F. In practice we are trading the man power need to code derivatives and covariances for the CPU power needed to execute these plug-in estimators.

Since our principal aim is the computation of the statistical error to be associated with secondary observables, in the following subsection we initially assume to be able to generate uncorrelated samples. We will then comment on how to take autocorrelations into account.

4.2.1 Bootstrap

We are interested in evaluating a secondary observable F which depends on several primary observables, for example $U_4 = \langle x^4\rangle/\langle x^2\rangle^2$. The sample estimator of this quantity is $\overline{F}$, i. e. the function F evaluated on the sample averages of the primary observables, for example $\overline{U_4} = \overline{x^4}/(\overline{x^2})^2$, and let us assume for the moment that the different draws are statistically independent from each other.

Algorithm 7 Bootstrap estimation of the uncertainty of $U_4 = \langle x^4\rangle/\langle x^2\rangle^2$ for iid draws.

Require: x_i for $i = 1, \dots, N$
  for $r = 1, \dots, R$ **do**
    $S_2 = 0$, $S_4 = 0$
    for $i = 1, \dots, N$ **do**
      generate $j \in \{1, \dots, N\}$ with uniform pdf
      $S_2 \leftarrow S_2 + x_j^2$
      $S_4 \leftarrow S_4 + x_j^4$
    end for
    $\overline{x^2} = S_2/N$
    $\overline{x^4} = S_4/N$
    $\overline{U_4}^{(r)} = \overline{x^4}/\overline{x^2}^2$
  end for
  compute the sample variance of the mean of $\{\overline{U_4}^{(r)}\}_{r=1,\dots,R}$, as in Eq. (4.2.15).

To compute the variance $\sigma^2_{\overline{F}}$ of the estimator $\overline{F}$, in principle, one could use the following strategy: perform R independent Monte-Carlo simulations, generating N draws in each case, and estimate $\sigma^2_{\overline{F}}$ by using the sample variance $\overline{\sigma}^2_{\overline{F}}$ defined by (see Eq. (1.1.7))

$$\overline{\sigma}^2_{\overline{F}} = \frac{R}{R-1}\left[\frac{1}{R}\sum_{j=1}^{R}(\overline{F}^{(j)})^2 - \left(\frac{1}{R}\sum_{j=1}^{R}\overline{F}^{(j)}\right)^2\right], \tag{4.2.15}$$

where $\overline{F}^{(i)}$ is the value of the sample estimator $\overline{F}$ computed by using the i-th sample. This method is in general unfeasible, since to evaluate the uncertainty of the estimator evaluated on a given sample we need to generate many more samples, using an algorithm that is in general nontrivial.

A way to apply Eq. (4.2.15) while minimizing the overhead of generating new samples is to use what is called the plug-in principle, which consists in approximating a probability distribution function with the empirical distribution of a sample of observations drawn from it. In practice: if our sample consists of N independent elements, we can create a bootstrap sample by randomly extracting N draws (with uniform pdf and with replacement) from this sample. The important point to note is that the elements of the bootstrap sample have the same statistical distribution of those of the original one. By resampling in this way the original sample $\{x_i\}_{i=1,\dots,N}$ we can thus generate R bootstrap samples $\{x_i^{(r)}\}_{i=1,\dots,N}$ (the index $r = 1, \dots, R$ identifies the sample), that can be used to evaluate the sample averages of the primary observables and obtain R estimates $\overline{F}^{(r)}$, by which we can evaluate $\overline{\sigma}^2_{\overline{F}}$ using Eq. (4.2.15). It is fundamental that the same bootstrap sample is used to compute *all* the primary observables needed for evaluating $\overline{F}$; correlations are instead lost if we use different bootstrap samples for different primary observables. A simple scheme of a bootstrap computation is reported in Alg. (7), and many more details on the bootstrap and on its statistical basis can be found, e. g., in [29] §10-11 and [30] §5-6-7.

Let us now finally consider the case of a Markov chain, in which different draws are not independent from each other. The simplest way to take into account autocorrelations in the bootstrap method is to divide the sample in N/k blocks (k is the block-size and we are assuming N to be divisible by k), then generate R bootstrap samples by randomly selecting, with uniform pdf and with replacement, N/k blocks each time. As for the case of primary observables discussed in Sec. 4.1.2, the whole procedure has to be repeated for increasing values of the block-size k until saturation is reached.

4.2.2 Jackknife

The idea of the jackknife method is analogous to that of the bootstrap, with the only difference that mock samples are not generated stochastically, but deterministically. Let us once again start by discussing the case of independent draws $x_1, \ldots, x_N$.

Jackknife samples are generated by removing a single drawn from the original sample, so we get N samples of $N-1$ draws, which provide N estimates of the primary observables[2] $\langle g_\alpha(x)\rangle$:

$$g_{\alpha\,(i)} = \frac{1}{N-1}\sum_{j\neq i} g_\alpha(x_i)\ , \quad j = 1, \ldots, N\ , \tag{4.2.16}$$

from which we get N estimates $F_{(i)} = F(g_{\alpha\,(i)})$ of the secondary observable. If we denote by F_J the sample composed by the N estimates $F_{(i)}$, the quantity

$$\overline{F_J^2} - \overline{F_J}^2 = \frac{1}{N}\sum_{i=1}^{N} F_{(i)}^2 - \left(\frac{1}{N}\sum_{i=1}^{N} F_{(i)}\right)^2 \tag{4.2.17}$$

estimates the square fluctuation of $\overline{F}$ induced by changing the sample by removing an element. Since all the elements of the sample enter in a symmetric way in the computation of $\overline{F}$, and the draws are independent from each other, we naively expect

$$\sigma_{\overline{F}}^2 \simeq N\left(\overline{F_J^2} - \overline{F_J}^2\right)\ . \tag{4.2.18}$$

To show that this expectation is indeed true we can rewrite the jackknife estimates $g_{\alpha\,(i)}$ of the primary observables as follows:

$$g_{\alpha\,(i)} = \frac{1}{N-1}\sum_{j\neq i} g_\alpha(x_i) = \langle g_\alpha\rangle + \frac{1}{N-1}\sum_{j\neq i}\delta g_{\alpha\,j}\ , \tag{4.2.19}$$

where we introduced the notation $\delta g_{\alpha\,j} = g_\alpha(x_j) - \langle g_\alpha(x)\rangle$. Since the typical value of $g_{\alpha\,(i)} - \langle g_\alpha\rangle$ is $\sigma_\alpha^2/\sqrt{N}$ we can use the approximation

$$\begin{aligned} F_{(i)} &= F\left(\langle g_\alpha\rangle + \frac{1}{N-1}\sum_{j\neq i}\delta g_{\alpha\,j}\right) \simeq \\ &\simeq F + \sum_\alpha F'_\alpha \frac{1}{N-1}\sum_{j\neq i}\delta g_{\alpha\,j} + \frac{1}{2}\sum_{\alpha\beta} F''_{\alpha\beta}\frac{1}{(N-1)^2}\sum_{j\neq i}\sum_{k\neq i}\delta g_{\alpha\,j}\delta g_{\beta\,k}\ , \end{aligned} \tag{4.2.20}$$

where F and its derivatives are computed in $\langle g_\alpha\rangle$. Analogously we have, using $\overline{g_\alpha} = \langle g_\alpha\rangle + \frac{1}{N}\sum_{i=1}^N \delta g_{\alpha\,i}$,

$$\overline{F} = F(\overline{g_\alpha}) \simeq F + \sum_\alpha F'_\alpha \frac{1}{N}\sum_j \delta g_{\alpha\,j} + \frac{1}{2}\sum_{\alpha\beta} F''_{\alpha\beta}\frac{1}{N^2}\sum_j\sum_k \delta g_{\alpha\,j}\delta g_{\beta\,k}\ . \tag{4.2.21}$$

Using $\langle \delta g_{\alpha\,i}\rangle = 0$ and $\langle \delta g_{\alpha\,j}\delta g_{\beta\,k}\rangle = C_{\alpha\beta}\delta_{jk}$ (where $C_{\alpha\beta}$ is the covariance matrix), we get from the second expression the identities

$$\langle\overline{F}\rangle \simeq F + \frac{1}{2N}\sum_{\alpha\beta} F''_{\alpha\beta}C_{\alpha\beta}\ , \tag{4.2.22}$$

and

$$\langle\overline{F}^2\rangle \simeq F^2 + \frac{1}{N}\sum_{\alpha\beta} F'_\alpha F'_\beta C_{\alpha\beta} + \frac{F}{N}\sum_{\alpha\beta} F''_{\alpha\beta}C_{\alpha\beta}\ , \tag{4.2.23}$$

from which

$$\sigma_{\overline{F}}^2 = \langle\overline{F}^2\rangle - \langle\overline{F}\rangle^2 = \frac{1}{N}\sum_{\alpha\beta} F'_\alpha F'_\beta C_{\alpha\beta}\ , \tag{4.2.24}$$

which is the generalization of Eq. (4.2.8).

If we use instead the expression for $F_{(i)}$ we get

$$\langle F_{(i)}F_{(j)}\rangle \simeq F^2 + \frac{1}{(N-1)^2}\sum_{\alpha\beta} F'_\alpha F'_\beta C_{\alpha\beta}\sum_{k\neq i}\sum_{\ell\neq j}\delta_{k\ell} + \frac{F}{(N-1)^2}\sum_{\alpha\beta} F''_{\alpha\beta}C_{\alpha\beta}\sum_{k\neq i}\sum_{\ell\neq i}\delta_{k\ell}\ , \tag{4.2.25}$$

[2] We denote by greek indices the ones used for labeling the primary observables on which the secondary observable depends. Latin indices will instead be used to label the different draws.

Algorithm 8 Jackknife estimation of the uncertainty of $U_4 = \langle x^4\rangle/\langle x^2\rangle^2$ for iid draws.

Require: x_i for $i = 1, \dots, N$
  $S_2 = 0$, $S_4 = 0$
  for $i = 1, \dots, N$ **do**
    $S_2 \leftarrow S_2 + x_i^2$
    $S_4 \leftarrow S_4 + x_i^4$
  end for
  for $i = 1, \dots, N$ **do**
    $(x^2)_{(i)} = (S_2 - x_i^2)/(N-1)$
    $(x^4)_{(i)} = (S_4 - x_i^4)/(N-1)$
    $(U_4)_{(i)} = (x^4)_{(i)}/((x^2)_{(i)})^2$
  end for
  compute $\overline{\sigma}^2_{U_4}$ using Eq. (4.2.32) with $F_{(i)} = (U_4)_{(i)}$.

and from the identities

$$\sum_{k\neq i}\sum_{\ell\neq i}\delta_{k\ell} = \sum_{k\neq i}\left(\sum_{\ell}\delta_{k\ell} - \delta_{ki}\right) = \sum_{k\neq i}(1-\delta_{ki}) = N-1 \tag{4.2.26}$$

and

$$\begin{aligned}\sum_{k\neq i}\sum_{\ell\neq j}\delta_{k\ell} &= \sum_{k\neq i}\left(\sum_{\ell}\delta_{k\ell} - \delta_{kj}\right) = \sum_{k\neq i}(1-\delta_{kj}) = N-1-\sum_{k\neq i}\delta_{kj} \\ &= N-1-\left(\sum_{k}\delta_{kj} - \delta_{ij}\right) = N-2+\delta_{ij}\ ,\end{aligned} \tag{4.2.27}$$

we finally have

$$\langle F_{(i)}F_{(j)}\rangle \simeq F^2 + \frac{N-2+\delta_{ij}}{(N-1)^2}\sum_{\alpha\beta}F'_\alpha F'_\beta C_{\alpha\beta} + \frac{F}{N-1}\sum_{\alpha\beta}F''_{\alpha\beta}C_{\alpha\beta}\ , \tag{4.2.28}$$

and in particular

$$\langle F^2_{(i)}\rangle \simeq F^2 + \frac{1}{N-1}\sum_{\alpha\beta}F'_\alpha F'_\beta C_{\alpha\beta} + \frac{F}{N-1}\sum_{\alpha\beta}F''_{\alpha\beta}C_{\alpha\beta}\ . \tag{4.2.29}$$

We can now evaluate

$$\langle \overline{F^2_J} - \overline{F_J}^2\rangle = \frac{1}{N}\sum_i\langle F^2_{(i)}\rangle - \frac{1}{N^2}\sum_{ij}\langle F_{(i)}F_{(j)}\rangle\ , \tag{4.2.30}$$

which using the previously written expressions becomes

$$\begin{aligned}\langle \overline{F^2_J} - \overline{F_J}^2\rangle &= \left(\frac{1}{N-1} - \frac{N-2}{(N-1)^2} - \frac{1}{N(N-1)^2}\right)\sum_{\alpha\beta}F'_\alpha F'_\beta C_{\alpha\beta} = \\ &= \frac{1}{N(N-1)}\sum_{\alpha\beta}F'_\alpha F'_\beta C_{\alpha\beta} = \frac{1}{N-1}\sigma^2_F\ ,\end{aligned} \tag{4.2.31}$$

where in the last step we used Eq. (4.2.24). We have thus found that a sample estimator of σ^2_F is

$$\overline{\sigma}^2_F = (N-1)\left(\overline{F^2_J} - \overline{F_J}^2\right) = (N-1)\overline{(F_J - \overline{F_J})^2} = \frac{N-1}{N}\sum_i\left(F_{(i)} - \overline{F_J}\right)^2\ . \tag{4.2.32}$$

A summary of the jackknife method to estimate the uncertainty of $B_4 = \langle x^2\rangle/\langle x^2\rangle^2$ is shown in Alg. (8), where it is also shown that to compute all the jackknife samples it is sufficient to scan the original sample only twice. For this reason the jackknife is computationally more efficient than the bootstrap (which requires at least $O(100)$ scans), however to use the jackknife method observables have to be reasonably smooth functions of the sample. If this is not the case jackknife can provide wrong estimates of the variance (larger than the real ones), as it famously happens for the case of the sample median. More details on the jackknife and its relation with bootstrap can be found, e. g., in [29] §10, [30], see also [31].

When autocorrelations are present in the sample, we can take them into account by dividing the sample in N/k blocks of size k (we are assuming N to be divisible by k), then generating

jackknife samples by removing the i-th block instead of the i-th draw. In this case we thus have

$$g_{\alpha\,(i)} = \frac{1}{N-k} \sum_{j \notin i-\text{th block}} g_\alpha(x_j) \tag{4.2.33}$$

and

$$\overline{\sigma}^2_{\overline{F}} = (N/k-1)\left(\overline{F_J^2} - \overline{F_J}^2\right) = (N/k-1)\overline{(F_J - \overline{F_J})^2} = \frac{N-k}{N}\sum_{i=1}^{N/k}\left(F_{(i)} - \overline{F_J}\right)^2 \ . \tag{4.2.34}$$

Part II

Classical statistical mechanics and phase transitions

Chapter 5

The Ising model: physics and simulations

5.1 Basic properties of the Ising model

The Ising model is a classical (i. e., non quantum) lattice model of a monoaxial ferromagnet. The configuration space of this model is obtained by associating to each site of a D-dimensional lattice (denoted by $\boldsymbol{x}$) a variable $s_{\boldsymbol{x}}$ which can only take the values ± 1, and the energy of a configuration is given by

$$E[\{s\}] = -J \sum_{\langle \boldsymbol{x},\boldsymbol{y} \rangle} s_{\boldsymbol{x}} s_{\boldsymbol{y}} - h \sum_{\boldsymbol{x}} s_{\boldsymbol{x}} \ , \tag{5.1.1}$$

where the expression $\sum_{\langle \boldsymbol{x},\boldsymbol{y} \rangle}$ denotes a sum on nearest neighbor sites of the lattice, and J and h are parameters of the model.

To model a monoaxial ferromagnet, the "spins" $s_{\boldsymbol{x}}$ are assumed to be always aligned along a given direction, hence the restriction to $s_{\boldsymbol{x}} = \pm 1$, which only leaves the freedom of a "spin-flip". The first term in Eq. (5.1.1) reminds of a magnetic dipole interaction, and the restriction to nearest neighbor sites is motivated by the fact that such an interaction decays quite rapidly with the distance ($\sim r^{-3}$). The strength of the dipole interaction is parametrized by J, with $J > 0$ corresponding to the ferromagnetic case ($s_{\boldsymbol{x}} s_{\boldsymbol{y}} = 1$ is favoured), while $J < 0$ corresponds to the antiferromagnetic case ($s_{\boldsymbol{x}} s_{\boldsymbol{y}} = -1$ is favored). The second term in Eq. (5.1.1) represents instead the interaction with an external magnetic field of intensity h. While any lattice can be used to define the Ising model (in fact any graph), we will always consider the simplest case of the cubic[1] lattice. It should be clear that this is a very simplistic modeling of a monoaxial ferromagnet, however we will see in the next section that, due to the phenomenon of universality, it is sufficient to quantitatively study what happens close to a continuous phase transition.

For the energy E to be well defined, the lattice has to be finite, hence we need also to specify the boundary conditions (b. c.). The simplest and most used boundary conditions are the periodic ones, which for a cubic lattice of linear size L can be written as

$$s_{\boldsymbol{x}+L\hat{\mu}} = s_{\boldsymbol{x}} \ , \tag{5.1.2}$$

where $\hat{\mu}$ (with $\mu = 1, \ldots, D$) is the unit vector directed along the μ-th direction. These b. c. are often used since they preserve translation symmetry and minimize the effect of boundaries, indeed no boundary is present when using periodic boundary conditions, and the system considered is in fact a torus. Other possible choices that sometimes can be useful are, e. g., anti-periodic boundary conditions ($s_{\boldsymbol{x}+L\hat{\mu}} = -s_{\boldsymbol{x}}$) and open boundary conditions, which correspond to the case of a real

[1] For the sake of the simplicity we will always speak of "cubic" lattice also in $D = 1$, $D = 2$ and $D > 3$, instead of using lattice, square lattice or hyper-cubic lattice, respectively.

finite cubic lattice in $\mathbb{R}^D$. Once a b. c. is adopted on a finite lattice, the statistical physics of the Ising model is encoded in the partition function

$$Z(\beta) = \sum_{\{s\}} \exp(-\beta E[\{s\}]) = \exp(-\beta F(\beta)) \ , \tag{5.1.3}$$

where $\beta = 1/(k_B T)$ is the inverse temperature and F is the free energy. Two dimensionless numbers thus characterize the phase diagram of the Ising model: βJ and βh. To avoid the proliferation of redundant parameters two conventions can be used

1. measure J and h in units of $k_B T$, which is equivalent to fix $\beta = 1$,
2. measure β in units of J, which is equivalent to fix $J = 1$ or $J = -1$ in the ferromagnetic or antiferromagnetic case, respectively.

Here we adopt the second of these possibilities, and we only consider the ferromagnetic Ising model, thus we fix $J = 1$.

The case $h = 0$ is the most interesting one, since it displays what is probably the most important property of the Ising model: for neutral boundary conditions (i. e., b. c. which do not favor a specific spin orientation, just like the periodic, anti-periodic or open b. c.) the energy $E[\{s\}]$ is invariant, for $h = 0$, under a global spin flip:

$$s'_{\boldsymbol{z}} = -s_{\boldsymbol{z}} \text{ for all } \boldsymbol{z} \quad \Longrightarrow \quad E[\{s'\}] = E[\{s\}] \quad \text{if } h = 0 \ . \tag{5.1.4}$$

The $h = 0$ system is thus characterized by a $\mathbb{Z}_2$ discrete symmetry, since by inverting twice the spins we come back to the original configuration.

A consequence the $\mathbb{Z}_2$ symmetry is that if we define the magnetization for unit volume by

$$m[\{s\}] = \frac{1}{L^D} \sum_{\boldsymbol{x}} s_{\boldsymbol{x}} \tag{5.1.5}$$

we always have $\langle m \rangle = 0$ for any inverse temperature β. This can be easily proven as follows

$$\begin{aligned} \langle m \rangle &= \frac{1}{Z(\beta)} \sum_{\{s\}} m[\{s\}] e^{-\beta E[\{s\}]} \overset{(1)}{=} \frac{1}{Z(\beta)} \sum_{\{-s\}} m[\{-s\}] e^{-\beta E[-\{s\}]} \overset{(2)}{=} \\ &= \frac{1}{Z(\beta)} \sum_{\{-s\}} -m[\{s\}] e^{-\beta E[\{s\}]} = -\langle m \rangle \ . \end{aligned} \tag{5.1.6}$$

In the equality denoted by (1) we changed the "mute index" of the sum, summing on the configurations in which every spin has the opposite sign with respect to that in the original sum; in step (2) we used the fact that $E[\{s\}]$ is an even function under the $\mathbb{Z}_2$ symmetry, while $m[\{s\}]$ is odd under a global spin-flip $m[\{-s\}] = -m[\{s\}]$. The simple fact $\langle m \rangle = 0$ seems at first sight to preclude the possibility of a ferromagnetic phase, i. e. of a phase characterized by a spontaneous magnetization. This point, however, needs to be investigated more thoroughly, since symmetries in statistical mechanics (and in quantum field theories) presents a richer phenomenology that in quantum mechanics. In particular, although $\langle m \rangle = 0$ for all β values, the mechanism underlying the vanishing of $\langle m \rangle$ is different at high and at low temperatures (if the space dimensionality is larger than one, $D > 1$), as can be seen by numerically investigating the pdf $P(m)$ of observing a value m of the magnetization in a simulation.

In Fig. (5.1) we report data for $P(m)$, obtained by simulating the two dimensional Ising model on a $L \times L$ lattice with periodic boundary conditions. To obtain these figures we performed 10^7 updates of the whole configuration, using the local Metropolis algorithm that will be discussed in Sec. 5.3. The first 10^5 measures have been discarded for thermalization, and the total simulation time has been ≈ 310 s. In the high temperature phase, Fig. (5.1) (left), by increasing the lattice size the function $P(m)$ gets more and more peaked at $m = 0$, indicating that in the thermodynamic

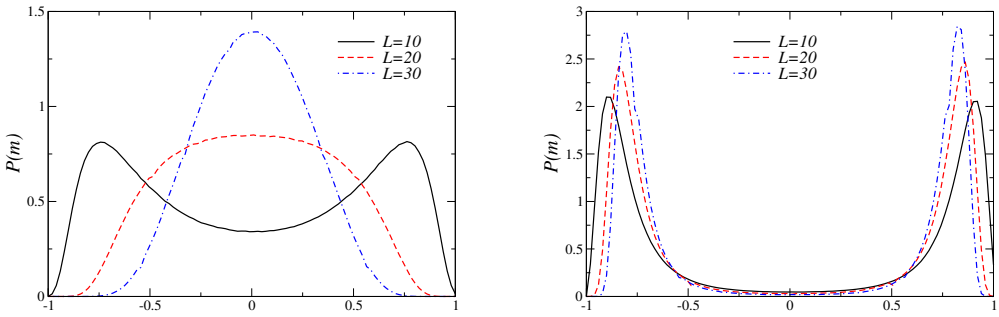

Figure 5.1: The pdf $P(m)$ of observing a value m of the magnetization in a two dimensional $L \times L$ Ising model with periodic b. c. (left) in the high temperature phase, $\beta = 0.4$, (right) in the low temperature phase, $\beta = 0.45$.

limit all the configurations sampled by MCMC have $m \approx 0$. What happens in the low temperature phase, Fig. (5.1) (right), is completely different: the pdf $P(m)$ gets more and more peaked (by increasing the lattice size) close to two values $\pm m_0$, with $m_0 \neq 0$. In this case the exact result $\langle m \rangle = 0$ does not mean that all configurations have $m \approx 0$, but that every configuration has $m \approx \pm m_0$ and the probability of these two cases is the same, thus $\langle m \rangle = 0$ is the consequence of a cancellation between the contributions of different configurations.

By also looking at the time histories of the simulations one can see that, in the low temperature phase, the time required for the system to switch, e. g., from the state $m \approx m_0$ to the state $m \approx -m_0$, grows by increasing the lattice size. This suggests[2] that in the thermodynamic limit the magnetization would always remain frozen at $m = \pm m_0$, and which of the two possibility is chosen depends on the initial condition. The low temperature phase is thus characterized, in the thermodynamic limit, by ergodicity breaking (not all states can be reached) and by an instability with respect to the initial condition.

In order to expose this instability we have to perform the thermodynamic limit more carefully: let us denote by $\langle \ \ \rangle_{L,\beta,h}$ the statistical average carried out using a lattice of linear size L, at the inverse temperature β, and using an external magnetic field h. What we proved before can be written, using this notation, as $\langle m \rangle_{L,\beta,h=0} = 0$ for any β and L; in fact exactly in the same way we can also prove the more general relation $\langle m \rangle_{L,\beta,-h} = -\langle m \rangle_{L,\beta,h}$. It can be shown that (if $D > 1$) a value $\beta_c > 0$ exists such that if $\beta \leq \beta_c$ (high temperature phase)

$$\lim_{h \to 0^+} \lim_{L \to \infty} \langle m \rangle_{L,\beta,h} = \lim_{h \to 0^-} \lim_{L \to \infty} \langle m \rangle_{L,\beta,h} = 0 \ , \tag{5.1.7}$$

while if $\beta > \beta_c$ (low temperature phase) we have

$$0 < m_0(\beta) \equiv \lim_{h \to 0^+} \lim_{L \to \infty} \langle m \rangle_{L,\beta,h} = - \lim_{h \to 0^-} \lim_{L \to \infty} \langle m \rangle_{L,\beta,h} \ . \tag{5.1.8}$$

Note that in the previous equation the order of the limits is essential: at finite L the partition function is analytic in h and β (it is a finite sum of exponentials), hence by using the different order of limits we obtain at *any* temperature

$$\lim_{L \to \infty} \lim_{h \to 0} \langle m \rangle_{L,\beta,h} = \lim_{L \to \infty} \langle m \rangle_{L,\beta,h=0} = \lim_{L \to \infty} 0 = 0 \ . \tag{5.1.9}$$

The function $m_0(\beta)$ defined in Eq. (5.1.8) is the spontaneous magnetization, and the fact that in the low temperature phase $m_0(\beta)$ is non-vanishing shows that the $\mathbb{Z}_2$ symmetry is spontaneously broken at low temperature. When a symmetry is spontaneously broken, average values exists

[2]This is only a suggestion since the details of the MC histories are generically unphysical, depending on the specific algorithm adopted, and only average values are physical (MC evolution is not a real physical evolution). Nevertheless the local Metropolis algorithm is a reasonable approximation of how thermal fluctuations behave in a real system.

which are not invariant under a symmetry of the Hamiltonian: the symmetry group of the state (identified by properly performing the thermodynamic limit) is smaller than the symmetry group of the Hamiltonian. Spontaneous symmetry breaking (SSB for short) is a feature that is present in statistical mechanics and quantum field theory, but not in quantum mechanical systems with a finite number of degrees of freedom, even if their Hilbert space is infinite dimensional. This is basically due to the fact that in quantum mechanical systems with a finite number of degrees of freedom the fundamental state is always non-degenerate (see e. g., [16] §15.4 for a sketch of the proof, or [17] §3.3.3, [18] §10.5 for more details), with a finite gap being present between the fundamental and the first excited state, which makes the system stable under small perturbations. From an algebraic point of view the same conclusion follows from the fact that an essentially unique representation exists of the Heisenberg commutation relations (Stone-von Neumann theorem), see, e. g., [32] §IV.6 for a proof, and [33] for the appearance of inequivalent representations in infinite systems.

The physical interpretation of the previous results is quite simple, as can be seen by thinking how the pdf $P(m)$ of observing a value m of the magnetization in a numerical simulation would change by adding a small external magnetic field h. At high temperature $P(m)$ is peaked at $m = 0$, and the presence of an external magnetic field would simply slightly distort the distribution, which would then be peaked at a value proportional to h. A non-vanishing value for $\langle m \rangle_{L,\beta,h}$ would result, but as far as h is small enough no instability emerges and $\lim_{h\to 0} \lim_{L\to\infty} \langle m \rangle_{L,\beta,h} = 0$. Things drastically changes in the low temperature case: by switching on a magnetic field h, and assuming $h > 0$, the peak at $m \approx m_0$ is enhanced by $e^{\beta h m_0 L^D}$, while the one at $m \approx -m_0$ is suppressed by the factor $e^{-\beta h m_0 L^D}$. Regardless of how small h is, in the thermodynamic limit only one of the two peaks survives, and we get $\lim_{L\to\infty} \langle m \rangle_{L,\beta,h} = m_0 + \chi h$, where χ is a constant related to the slight shift of the peak at m_0 induced by the external field. Thus finally $\lim_{h\to 0^+} \lim_{L\to\infty} \langle m \rangle_{L,\beta,h} = m_0$.

Observables which transform non-trivially under the symmetry group of the Hamiltonian (more precisely, transform as an irreducible representation of the symmetry group) are called order parameters, and their average values vanish in the unbroken phase, while they can be nonzero in the broken phase. The specific values assumed by average values in the broken phase typically depend on the specific way in which the thermodynamic limit is performed, e. g., on the presence of external fields, on the boundary conditions adopted, and so on. The two cases considered in Eq. (5.1.8), which are used to define the spontaneous magnetization $m_0(\beta)$, correspond to the thermodynamic analogue of pure states, but also mixed states exists. A mixed state is obtained, e. g., by using boundary conditions in which a fraction x of the spins on the boundary are fixed to -1, with the remaining fixed at $+1$; in this way one gets in the thermodynamic limit

$$\lim_{L\to\infty} \langle m \rangle_{L,\beta,h=0} = (1-2x) m_0(\beta) \qquad \text{fixed b. c. with } 0 < x < 1 \ . \tag{5.1.10}$$

Rigorous proofs of the existence of a high temperature and a low temperature phases with the previously stated properties can be found, e. g., in [34] §4, [35] §4-5, [36] §3, with the last reference being the most introductory one. That in $D = 1$ no spontaneous magnetization is present at any $\beta > 0$ can be shown or by explicitly solving the $D = 1$ Ising model, [37] §14, or by using a more general reasoning valid for any short range model, [38] §163. Also the two dimensional Ising model can be explicitly solved, see, e. g. [37] §15 or [38] §151 for two different approaches, or [39] for many more details. The exact value $\beta_c = \frac{1}{2}\log(1+\sqrt{2}) \approx 0.440687\ldots$ for $D = 2$ can also be obtained without knowing the explicit solution of the model, by using the low-temperature/high-temperature (self-)duality of the $D = 2$ Ising model, see, e. g., [40]. Several critical properties of the Ising model, obtained by using analytical methods or numerical simulations, are reported in Sec. 7.A.

5.2 Phase transitions and critical phenomena

The points of the phase diagram at which the free energy density $f = F/L^D$ is not an analytic function of the control parameters (temperature, pressure, external magnetic field, ...) are called

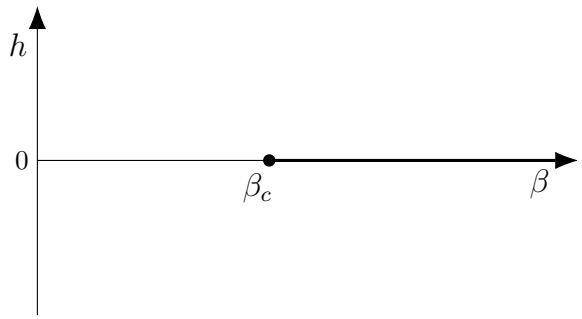

Figure 5.2: The phase diagram of the Ising model when $D > 1$. The black dot at β_c denotes a continuous transition, while the thick line for $\beta > \beta_c$ and $h = 0$ denotes a line of discontinuous transitions.

phase transitions, and this non-analytic behavior can emerge only in the thermodynamic limit $L \to \infty$, since the partition function is always an analytic function of the parameters in a finite volume.

In the modern classification phase transitions can be discontinuous or continuous. At discontinuous transitions different thermodynamic phases coexist and the free energy density has some discontinuous first derivatives; examples of discontinuous (or first order, according to the old Ehrenfest classification) transitions are boiling water and the solid-liquid phase transitions. At continuous transitions there is no phase coexistence and, typically, at least some second derivatives of the free energy density diverge at the transition; examples of continuous transitions are the liquid-vapor critical end-point transition and the Curie transition in ferromagnets. Continuous phase transitions are also often called critical points and, somehow extending the old Ehrenfest definition, second order phase transitions.

From the discussion in the previous section, it follows that the phase diagram of the Ising model in $D > 1$ dimensions is the one sketched in Fig. (5.2): for $h \neq 0$ the free energy density is an analytic function of β and h, and the same is true also for $h = 0$ in the high temperature phase $\beta < \beta_c$. For $\beta > \beta_c$ a line of discontinuous phase transitions is present, associated with the appearance of a spontaneous magnetization which abruptly changes sign when h changes sign. The point $\beta = \beta_c, h = 0$ is the only point in which a continuous phase transition happens.

Close to a continuous transition a peculiar behavior (the so called critical behavior) emerges, in which physical quantities behave as power-law functions of the "distance" from the critical point. Using the Ising model as an example, the "distance" from the critical point is usually parametrized by the so called reduced temperature

$$t = \frac{\beta_c - \beta}{\beta_c} \ , \tag{5.2.1}$$

and by the intensity h of the external magnetic field (note that $t > 0$ corresponds to the high temperature phase $\beta < \beta_c$). The specific heat is defined by

$$\begin{aligned} C =& \frac{1}{L^D}\frac{\partial}{\partial T}U = \frac{1}{L^D}\frac{\partial}{\partial T}\left(-\frac{\partial}{\partial \beta}\log Z(\beta,h)\right) = \frac{\beta}{TL^D}\frac{\partial^2}{\partial \beta^2}\log Z(\beta,h) = \\ &= \frac{\beta}{TL^D}\left(\langle E^2\rangle - \langle E\rangle^2\right) = \frac{\beta}{T}L^D\left(\langle \varepsilon^2\rangle - \langle \varepsilon\rangle^2\right) \ , \end{aligned} \tag{5.2.2}$$

where U is the internal energy and $\varepsilon = E/L^D$ is the energy density. For $h = 0$ and $t \approx 0$ the specific heat behaves, in the thermodynamic limit, as

$$C(\beta, h = 0) \approx \frac{A_\pm}{|t|^\alpha} \ , \tag{5.2.3}$$

where A_+ and A_- are two constants that have to be used for $t > 0$ and $t < 0$, respectively. Analogously, the magnetization $m(\beta, h) = \lim_{L\to\infty}\langle m\rangle_{L,\beta,h}$ has the following behavior for $h = 0^+$

and close to $t \approx 0$ (and $t < 0$ since otherwise $m(\beta, h = 0) = 0$)

$$m(\beta, h = 0^+) \approx B \times (-t)^\beta \ , \tag{5.2.4}$$

where B is a constant, while for $t = 0$ (i. e. on the critical isotherm) and $h \approx 0$ it behaves as

$$m(\beta = \beta_c, h) \approx B_c \times |h|^{1/\delta} \ , \tag{5.2.5}$$

where B_c is another constant. Finally, the magnetic susceptibility is defined by

$$\begin{aligned} \chi = \frac{\partial}{\partial h} \langle m \rangle_{L,\beta,h} = \frac{\partial}{\partial h} \left(\frac{1}{\beta L^D} \frac{\partial}{\partial h} \log Z(\beta, h) \right) = \\ = \frac{\beta}{L^D} \left(\langle M^2 \rangle - \langle M \rangle^2 \right) = \beta L^D \left(\langle m^2 \rangle - \langle m \rangle^2 \right) \ , \end{aligned} \tag{5.2.6}$$

where $M = \sum_{\boldsymbol{x}} s_{\boldsymbol{x}}$ is the total magnetization and $m = M/L^D$ is the magnetization for unit volume. For $h = 0$ and $t \approx 0$ the magnetic susceptibility behaves as

$$\chi(\beta, h = 0) \approx \frac{C_\pm}{|t|^\gamma} \ , \tag{5.2.7}$$

where C_+ and C_- are two constants, to be used once again for $t > 0$ and $t < 0$, respectively. The exponents $\alpha, \beta, \gamma, \delta$ are called critical exponents, and, together with the amplitudes $A_\pm, B, B_c$ and $C_\pm$, characterize the critical behavior.

The power-law critical behavior is not typical only of macroscopic observables, but can be seen also in microscopic ones. To define the "microscopic" critical exponents let us introduce the two point connected correlation function $G(\boldsymbol{x}, \boldsymbol{y})$:

$$G(\boldsymbol{x}, \boldsymbol{y}) = \langle s_{\boldsymbol{x}} s_{\boldsymbol{y}} \rangle - \langle s_{\boldsymbol{x}} \rangle \langle s_{\boldsymbol{y}} \rangle \ . \tag{5.2.8}$$

If $\beta \neq \beta_c$ (if $h = 0$, or for any β if $h \neq 0$) the large distance behavior of this function is given by

$$G(\boldsymbol{x}, \boldsymbol{y}) \propto \frac{1}{|\boldsymbol{x} - \boldsymbol{y}|^{(D-1)/2}} e^{-|\boldsymbol{x}-\boldsymbol{y}|/\xi} \ , \tag{5.2.9}$$

where ξ is the correlation length, and the previous expression is often referred to as the Ornstein-Zernike form[3]. The correlation length ξ thus parametrizes the typical distance at which two spins are correlated, however one does not have to think of ξ as the "size of a bubble", since no bubble at all exists if we are not at a discontinuous phase transition. A fundamental property of continuous phase transitions is the divergence of the correlation length, which leads to the phenomenon of critical opalescence at the critical end-point of the liquid-vapor transition. For the Ising model we have, for $h = 0$ and $t \approx 0$, the critical behavior

$$\xi(\beta, h = 0) \approx f_\pm |t|^{-\nu} \ . \tag{5.2.10}$$

Finally, exactly at the critical point ($t = 0$ and $h = 0$) the large distance behavior of the two point connected correlation function is

$$G(\boldsymbol{x}, \boldsymbol{y}) \propto \frac{1}{|\boldsymbol{x} - \boldsymbol{y}|^{D-2+\eta}} \ , \tag{5.2.11}$$

where the exponent η is typically called anomalous dimension.

All the previously introduced critical exponents are not independent of each other, but are related by several equalities:

$$\alpha + 2\beta + \gamma = 2 \ , \quad \alpha + \beta(1 + \delta) = 2 \ , \quad \gamma = \nu(2 - \eta) \ , \quad 2 - \alpha = D\nu \ . \tag{5.2.12}$$

[3]The Ornstein-Zernike form corresponds to the large distance behavior of the inverse Fourier transform of the scalar propagator $G(\boldsymbol{k}) \propto 1/(k^2 + \xi^{-2})$, as will be shown in Sec. 14.1. This is the reason why ξ is sometimes denoted by ξ_{gap}.

The first three relations are examples of "scaling relations", while the last one (the one explicitly depending on D) is an example of an "hyperscaling relation", and it is true only if $D < 4$. These relations can be proved by assuming a phenomenological scaling form of the free energy density (see, e. g., [41] §11) or, better, by using renormalization group techniques (see, e. g., [42] §3). Scaling relations are typically limit cases of exact thermodynamic inequalities, like, e. g., the Rushbrooke ($\alpha + 2\beta + \gamma \geq 2$) and the Griffiths ($\alpha + \beta(1 + \delta) \geq 2$) inequalities, which follow from the positivity of the specific heat and the convexity of the free energy, respectively (see, e. g., [41] §4).

What makes continuous transitions particularly appealing from the theoretical point of view is the property of universality: critical exponents (and other quantities like, e. g., some ratios of amplitudes) do not depend on the microscopic details of the system considered, but only on some very general properties of the system, like the symmetries, the dimensionality of the system, and the nature of the order parameter. Critical phenomena can thus be classified in universality classes (e. g., the 3D Ising universality class), and this is not only very important from the theoretical point of view, it is also extremely convenient for computational purposes: if we are interested in investigating the critical exponents of a monoaxial ferromagnet, we do need to know all the details of a specific material, we can simply use the Ising model and the results will be the same. Obviously this is not the case if we are interested in nonuniveral quantities, like, e. g., the critical temperature.

To theoretically justify the phenomenon of universality it is often said, somehow colloquially, that close to a critical point the correlation length diverges, and when $\xi \gg 1$ the system "forgets" its microscopic details. This is however quite misleading, since it implicitly suggests that only the "large" length scales, those of the order of ξ, are important to describe the critical state. Reality is more complicated/interesting [43]:

> A classical hydrodynamic wave is characterized by a definite wavelength, and very little motion of the fluid occurs at much shorter wavelengths. It is therefore a relatively trivial matter to introduce continuum forms of density, pressure, etc. for a hydrodynamic wave. However, the critical fluctuations in a magnet for very long wavelengths are not the dominant fluctuations. Instead, fluctuations occur on *all* wavelength scales from the correlation length to the atomic spacing and all these intermediate wavelengths are crucial to the physics of critical phenomena. In particular there is no gap in wavelengths between the wavelengths of fluctuations and the atomic wavelengths. This means it is difficult to determinate which wavelengths of fluctuations to include in a continuum description and which to exclude.

Renormalization group methods have been introduced to cope with this problem, and it is only using this approach that universality becomes natural, see e. g. [42] §3 for an introductory presentation, or [44], or [45] §5 and [46] §25- for a QFT approach.

To show that close to a critical point it is not possibile to simply "neglect" the short distance scales, let us consider, following [42] §1, the connected two point functions in $D = 3$. Using translation invariance we immediately see that $G(\boldsymbol{x}, \boldsymbol{y}) = G(\boldsymbol{x} - \boldsymbol{y})$, and it can be shown (see e. g. [42] §2) that in the mean field approximation $G(\boldsymbol{x}, \boldsymbol{y}) \sim |\boldsymbol{x} - \boldsymbol{y}|^{-1}$ when $h = 0$. If we denote by a the lattice spacing in physical units, by dimensional analysis we thus expect (assuming rotational invariance for the sake of the simplicity)

$$G(r) = \frac{1}{r} g\left(\frac{r}{\xi}, \frac{a}{\xi}\right) \ . \tag{5.2.13}$$

Using the invariance under translations of the average values, we thus have (neglecting an irrelevant multiplicative β factor)

$$\begin{aligned} \chi \propto L^D \left(\langle m^2 \rangle - \langle m \rangle^2\right) = L^D \left(\langle m\, s_{\boldsymbol{y}} \rangle - \langle m \rangle \langle s_{\boldsymbol{y}} \rangle\right) = \sum_{\boldsymbol{x}} \left(\langle s_{\boldsymbol{x}} s_{\boldsymbol{y}} \rangle - \langle s_{\boldsymbol{x}} \rangle \langle s_{\boldsymbol{y}} \rangle\right) = \\ = \sum_{\boldsymbol{x}} G(\boldsymbol{x}, \boldsymbol{y}) \approx \int G(r) \mathrm{d}\boldsymbol{r} \ , \end{aligned} \tag{5.2.14}$$

and, by simply "neglecting" the a/ξ dependence of $g(r)$ close to the critical point, we would get (in $D = 3$)

$$\chi \propto \int G(r) \mathrm{d}\boldsymbol{r} = \int \frac{1}{r} g\left(\frac{r}{\xi}, 0\right) \mathrm{d}\boldsymbol{r} \sim \xi^2 \ . \tag{5.2.15}$$

The correct critical behavior of χ for $h = 0$ and $t \approx 0$ is however (using the definition of γ, the third of Eq. (5.2.12), and the definition of ν)

$$\chi \propto |t|^{-\gamma} = |t|^{-\nu(2-\eta)} \propto \xi^{2-\eta} \ . \tag{5.2.16}$$

The fundamental point is that we can not simply neglect the dependence on a/ξ. From the final result we however see that this dependence is quite simple and, more precisely, the leading behavior for $a/\xi \ll 1$ has to be of the form

$$g\left(\frac{r}{\xi}, \frac{a}{\xi}\right) \sim \left(\frac{a}{\xi}\right)^{\eta} g_1\left(\frac{r}{\xi}\right) \ , \tag{5.2.17}$$

to be consistent both with the correct critical behavior and with dimensional analysis. This is the reason why the critical exponent η is called anomalous dimension: only for $\eta = 0$ we recover the results that could be guessed by using dimensional analysis and a simple "hydrodynamical" separation of scales.

5.3 How to simulate the Ising model

We now discuss how to simulate the Ising model using the MCMC method. The simplest approach is that which makes use of the local Metropolis algorithm, and the elementary step of the Markov chain is given by the following operations:

1. select with uniform pdf a site $\boldsymbol{r}$ of the lattice,
2. define the trial configuration as the configuration in which only the spin in position $\boldsymbol{r}$ is flipped with respect to the original configuration: $s_{\boldsymbol{r}} \to s'_{\boldsymbol{r}} = -s_{\boldsymbol{r}}$,
3. accept the trial configuration with probability $\min(1, e^{-\beta(E'-E)})$, where E is the energy of the initial configuration and E' is the energy of the trial configuration. If the trial configuration is not accepted, keep the old one.

The first two points define the selection probability of the new configuration, which in Sec. 3.3.1 was denoted by A_{ba}. If we denote by C the initial configuration and by C' the trial configuration selected by using the points 1. and 2., the probability of selecting C' given C is the same as the probability of selecting C given C', since in point 1. the lattice point $\boldsymbol{r}$ is selected with uniform pdf and C and C' differ only for the value at a single site. The selection probability is thus symmetric, and according to the general discussion in Sec. 3.3.1 this algorithm satisfies the detailed balance condition with respect to the Gibbs distribution $e^{-\beta E}/Z(\beta)$.

What remains to be proven is that the Markov chain built in this way is irreducible and aperiodic. Let us start from irreducibility: we have to show that in a finite number of steps it is possible to reach any configuration C_2 starting from a generic configuration C_1. Since the position of the spin to be flipped is chosen in point 1. with uniform pdf, there is a nonvanishing probability of selecting in subsequent updates all the points, and only those, whose spins have different orientations in C_1 and C_2. Moreover the spin-flip probability in point 3. never vanishes, so we also have a nonzero probability of accepting all the proposed flips. We thus have a nonvanishing probability of passing from configuration C_1 to configuration C_2 in less than L^D steps, where L is the linear size of the cubic lattice on which the Ising model is defined.

Since we have just seen that the Markov chain is irreducible, and in an irreducible Markov chain all the states have the same period (see theorem 3.1.1), it is sufficient to show that the period of a specific state is 1 to prove that the chain is aperiodic. Let us consider the configuration with all spins equal to $+1$. The energy difference $E' - E$ due to the spin-flip of a randomly chosen site is simply given by $4D$, since any site has $2D$ next neighbor sites, and by flipping the spin $s_{\boldsymbol{r}}$ the quantity $-s_{\boldsymbol{r}} s_{\boldsymbol{x}}$ (where $\boldsymbol{x}$ is a next neighbor site of $\boldsymbol{r}$) change from -1 to 1 (we remind the reader that we conventionally set $J = 1$, see Sec. 5.1). The probability of accepting the spin flip is thus in this case

$$\min(1, e^{-4\beta D}) = e^{-4\beta D} \ , \tag{5.3.1}$$

which is smaller than 1 if $\beta > 0$. We thus have a finite probability of rejecting the proposed updated, and thus $1 \in R_{+1}$, where R_{+1} is set the of the recurrence times of the configuration with all spins equal to $+1$. The period of this state/configuration is thus $\mathrm{GCD}(R_{+1}) = 1$ and the chain is aperiodic.

By iterating the three steps described above we thus obtain an irreducible and aperiodic Markov chain, and the probability of sampling a specific configuration is, for large enough MC time, given by its Gibbs weight. Note that it is not convenient to perform measures (e. g., energy and magnetization measures) after every iteration of the above algorithm, since measures are clearly strongly correlated after a single spin-flip update; measures are usually performed every L^D elementary single spin updates, or integer multiples of this number.

It is often suggested, instead of randomly selecting the point to be updated, to systematically sweep the lattice following a specific order, however for the specific case of the Ising model this does not generically ensure the Markov chain to be irreducible and aperiodic. To provide an explicit example of this fact let us consider the somehow trivial case of a one dimensional lattice with 3 lattice sites and periodic boundary conditions. Let us assume to sweep the lattice starting from the left and going to the right, and consider the configuration $(+1, -1, +1)$. It is immediate to see that by flipping the first spin on the left the energy of the configuration does not change, so the spin-flip is accepted with probability 1 (see point 3. above), and we reach the state $(-1, -1, +1)$. Now we have to update the second spin, but also in this case the energy is unchanged by flipping the spin, and the same happens also for all the subsequent updates. The states sampled by the Markov chain are thus

$$\begin{aligned}(\underline{+1}, -1, +1) \to (-1, \underline{-1}, +1) \to (-1, +1, \underline{+1}) \to (\underline{-1}, +1, -1) \to \\ \to (+1, \underline{+1}, -1) \to (+1, \underline{-1}, -1) \to (+1, -1, \underline{+1}) \ ,\end{aligned} \tag{5.3.2}$$

where the underlined number is the one to be updated. We see that after 6 updates we return back to the initial state, and the states $(+1, +1, +1)$ and $(-1, -1, -1)$ are never reached. The Markov chain is thus reducible, and the state $(+1, -1, +1)$ has period 6. The source of the problem can be traced back to the fact that for the chosen configuration all moves were in fact forced moves, since the energy never changed and all the proposed spin-flips where thus accepted with probability one. Analogous configurations, which generate the same problem, can be found also in less trivial geometries, see [47]. To prevent this type of problem it is sufficient to modify point 2. as follows

2'. with probability $0 < 1 - \epsilon < 1$ define the trial configuration as the configuration in which only the spin in position $\boldsymbol{r}$ is flipped with respect to the original configuration: $s_{\boldsymbol{r}} \to s'_{\boldsymbol{r}} = -s_{\boldsymbol{r}}$. With probability $0 < \epsilon < 1$ the trial configuration is just the old configuration.

Using this prescription it is simple to verify that the Markov chain is irreducible and aperiodic also if we seep the lattice in a deterministic way. If we start from the configuration C_1, to reach the configuration C_2 it is sufficient to select the probability ϵ of not updating the site for all the sites which have the same sign in C_1 and in C_2, and to select the probability $1 - \epsilon$ of updating the site in all the other cases, with all the updates being accepted. This can be unlikely, but surely it has a nonvanishing probability. Furthermore, since there is a nonzero probability of not updating the configuration, aperiodicity is immediate (even for $\beta = 0$). A deterministic lattice sweep is thus legitimate in this case; note however that the balance condition is satisfied using this approach, but the detailed balance condition is not, see the analogous discussion in Sec. 3.3.3.

There are a couple of ways in which we can improve the computational efficiency of the basic Metropolis update. The simplest and more important one is obtained by noting that the energy of a configuration is written as a sum on nearest neighbors lattice sites. As a consequence, to compute the difference of energies $E' - E$ needed in the accept/reject step we do not really have to know the values E' and E (whose computation would require a sum on all the lattice), but only the values of the spins close to the point $\boldsymbol{r}$ where the spin flip is proposed. To make this more explicit we can write the energy of a generic configuration as (we consider just the case $h = 0$, that is the one that will be used in the following)

$$E[\{s\}] = -s_{\boldsymbol{r}} S_{\boldsymbol{r}} + (\text{independent of } s_{\boldsymbol{r}}) \ , \tag{5.3.3}$$

where $S_{\boldsymbol{r}}$ is

$$S_{\boldsymbol{r}} = \sum_{\langle \boldsymbol{x}, \boldsymbol{r} \rangle} s_{\boldsymbol{x}} \ , \tag{5.3.4}$$

Algorithm 9 Metropolis algorithm to simulate the Ising model

loop
  randomly select a site $\boldsymbol{r}$ of the lattice with uniform pdf
  compute $S_{\boldsymbol{r}} = \sum_{\langle \boldsymbol{x},\boldsymbol{r}\rangle} s_{\boldsymbol{x}}$
  if $s_{\boldsymbol{r}} S_{\boldsymbol{r}} \leq 0$ **then**
    flip the spin: $s_{\boldsymbol{r}} \leftarrow -s_{\boldsymbol{r}}$
  else
    draw a random number $w \in [0,1)$ with uniform pdf
    if $w \leq \exp(-2\beta s_{\boldsymbol{r}} S_{\boldsymbol{r}})$ **then**
      flip the spin: $s_{\boldsymbol{r}} \leftarrow -s_{\boldsymbol{r}}$
    end if
  end if
end loop

i. e., the sum of the spins in the next neighbor sites of $\boldsymbol{r}$. The energy difference $E' - E$ associated with the spin flip $s_{\boldsymbol{r}} \to -s_{\boldsymbol{r}}$ is thus equal to

$$E' - E = 2 s_{\boldsymbol{r}} S_{\boldsymbol{r}} \ , \tag{5.3.5}$$

and the corresponding acceptance probability is $\exp(-2\beta s_{\boldsymbol{r}} S_{\boldsymbol{r}})$. The basic form of the Metropolis algorithm to simulate the Ising model can thus be written as in Alg. (9), where we implemented also the optimization already discussed in Sec. 3.3.1: if $E' - E \leq 0$ the update will be surely accepted, and we do not need to compute the exponential and draw a random number.

A further improvement can be obtained by noting that, apart from the unavoidable random number generation, the slowest part of the algorithm is the computation of $\exp(-\beta(E' - E))$: everything else can be done using integer arithmetic. However the difference $E' - E$ can only assume a finite number of values, hence the possible values of $\exp(-\beta(E' - E))$ be computed once for all at the beginning of the simulation. If we define the vector p_k by

$$p_k = e^{-2\beta k} \ , \quad \text{for} \quad k = 1, \ldots, 2D \ , \tag{5.3.6}$$

we can substitute the block

draw a random number $w \in [0,1)$ with uniform pdf
if $w \leq \exp(-2\beta s_{\boldsymbol{r}} S_{\boldsymbol{r}})$ **then**
  flip the spin: $s_{\boldsymbol{r}} \leftarrow -s_{\boldsymbol{r}}$
end if

with the computationally simpler

draw a random number $w \in [0,1)$ with uniform pdf
$k = s_{\boldsymbol{r}} S_{\boldsymbol{r}}$
if $w \leq p_k$ **then**
  flip the spin: $s_{\boldsymbol{r}} \leftarrow -s_{\boldsymbol{r}}$
end if

Note that we can assume $k > 0$ since if $s_{\boldsymbol{r}} S_{\boldsymbol{r}} \leq 0$ we do not even need to draw the random number, see Alg. (9).

We can now discuss the heat-bath algorithm for the Ising model. The starting point is the representation Eq. (5.3.3) of the energy of the model as a function of $s_{\boldsymbol{r}}$, where $\boldsymbol{r}$ is a given site of the lattice. In the heat-bath algorithm the new configuration is generated by sampling the new value of $s_{\boldsymbol{r}}$ using its conditional probability evaluated at fixed $\{s_{\boldsymbol{x}}\}_{\boldsymbol{x} \neq \boldsymbol{r}}$, see Sec. 3.3.2. We thus have to select $s_{\boldsymbol{r}} = +1$ with probability

$$p(s_{\boldsymbol{r}} = +1) = \frac{e^{\beta S_{\boldsymbol{r}}}}{e^{-\beta S_{\boldsymbol{r}}} + e^{\beta S_{\boldsymbol{r}}}} \ , \tag{5.3.7}$$

Algorithm 10 Heat-bath algorithm to simulate the Ising model

loop
  select $\boldsymbol{r}$ using a deterministic sweep or a random choice with uniform pdf
  compute $S_{\boldsymbol{r}} = \sum_{\langle \boldsymbol{x},\boldsymbol{r}\rangle} s_{\boldsymbol{x}}$ and $p(s_{\boldsymbol{r}} = +1) = e^{\beta S_{\boldsymbol{r}}}/(e^{-\beta S_{\boldsymbol{r}}} + e^{\beta S_{\boldsymbol{r}}})$
  draw a random number $w \in [0, 1)$ with uniform pdf
  if $w < p(s_{\boldsymbol{r}} = +1)$ **then**
    $s_{\boldsymbol{r}} \leftarrow +1$
  else
    $s_{\boldsymbol{r}} \leftarrow -1$
  end if
end loop

and $s_{\boldsymbol{r}} = -1$ with probability $p(s_{\boldsymbol{r}} = -1) = 1 - p(s_{\boldsymbol{r}} = +1)$. This is what we have to do once a lattice point $\boldsymbol{r}$ has been selected, but how do we select it? We have two possible choices: we can either select $\boldsymbol{r}$ randomly with uniform pdf or sweep the lattice using a deterministic algorithm, both the methods providing a irreducible and aperiodic Markov chain. This can be proven in a way that is completely analogous to the reasoning presented above when discussing the variant move 2'. of the Metropolis algorithm. As for the Metropolis update, detailed balance is satisfied only if we randomly select the site to be updated, while the balance condition is satisfied anyway. The deterministic sweep is typically computationally more efficient for two different reasons: it does not require to draw a random number, and at least some neighbor sites of the site to be updated have been already updated. Summarizing, we thus obtain Alg. (10). As a further improvement it is possible to pre-compute the possible values of the probabilities $p(s_{\boldsymbol{r}} = +1)$ as a function of $S_{\boldsymbol{r}}$, in a way that is completely analogous to what has been done before for the Metropolis update.

Heat-bath algorithms are typically more efficient than the Metropolis algorithm, since in an heat-bath update the value of the variable to be updated is generated using a process that does not depend on the previous value of the variable. For the specific case of the Ising model, however, since only two values are available, the two algorithms are practically equivalent.

5.4 Finite size scaling and critical slowing-down

As we repeatedly noted before, phase transitions do not happen in finite systems, and the computation of physical quantities related to phase transitions and critical phenomena (critical exponents, critical temperatures, ...) thus require an infinite volume extrapolation. Finite Size Scaling (FSS) is the technique that has been developed to systematically carry out this extrapolation.

Before describing FSS it is convenient to preliminary discuss the two different expressions that can be used to define the susceptibility in the present context, which correspond to different physical conditions. If we define the average magnetization for unit volume as $\langle m \rangle$, where

$$m = \frac{1}{L^D} \sum_{\boldsymbol{x}} s_{\boldsymbol{x}} \ , \tag{5.4.1}$$

we have seen in Sec. 5.1 that we always have $\langle m \rangle = 0$ (we assume periodic b. c.). We have however also seen, in Sec. 5.2, that the dynamics underlying the vanishing of $\langle m \rangle$ is very different in the high and in the low temperature phases.

Let us start by considering the high temperature phase, in which everything is analytic and, in the large volume limit, all configurations have $m \approx 0$. In this case it can be shown that the pdf $P(m)$ of obtaining a configuration with magnetization for unit volume m is well approximated by

$$P_{ht}(m) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{m^2}{2\sigma^2}\right) \ , \tag{5.4.2}$$

where $\sigma^2 = \langle m^2 \rangle$. In [38] §110 this result is obtained by using the microscopic definition of the entropy and considering only slight deviations from equilibrium. We can reach the same conclusion by assuming that the value $m \approx 0$ is "typical" (i. e. $\langle m \rangle = 0$ does not emerge from a cancellation) and using the central limit theorem when $L \gg \xi$, where L is the linear size of the sample and ξ the correlation length. If we add a small external magnetic field we get, remembering Eq. (5.1.1),

$$P_{ht}(m) \propto \exp\left(-\frac{m^2}{2\sigma^2} + \beta h L^D m\right) , \tag{5.4.3}$$

from which we see that $\langle m \rangle_h = h\beta\sigma^2 L^D$. We can then define the susceptibility χ by the relation $\langle m \rangle_h = h\chi$, and we get $\chi = \beta\sigma^2 L^D$, i. e. $\chi = \beta L^D \langle m^2 \rangle$, consistently with the definition used in Sec. 5.2. To summarize we thus have, in the high temperature phase

$$P_{ht}(m) = \sqrt{\frac{\beta L^D}{2\pi\chi}} \exp\left(-\frac{\beta L^D}{2\chi} m^2\right) , \tag{5.4.4}$$

where χ is the magnetic susceptibility. If we compute $\langle |m| \rangle$ using this pdf we get

$$\langle |m| \rangle = \sqrt{\frac{2}{\pi}} \sqrt{\frac{\chi}{\beta L^D}} , \tag{5.4.5}$$

hence $\langle |m| \rangle$ is nonvanishing at finite volume but approach zero as $\sqrt{1/L^D}$ by increasing the lattice size.

Let us now consider the low temperature phase, where $\langle m \rangle = 0$ emerges from the cancellation between the two "typical" values $m = \pm m_0$. Each of these values is associated with a stable thermodynamic state, which is identified by performing the thermodynamic limit with an infinitesimal external magnetic field, as discussed in Sec. 5.2. We can thus expect each of these thermodynamic states to be well approximated by a Gaussian distribution for m, obtaining the pdf

$$P_{lt}(m) = \frac{1}{2}\sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(m-m_0)^2}{2\sigma^2}\right) + \frac{1}{2}\sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(m+m_0)^2}{2\sigma^2}\right) . \tag{5.4.6}$$

Note that, since the two thermodynamic phases are related by the $\mathbb{Z}_2$ symmetry, the variance σ^2 is the same in both the phases. As before $\chi' = \beta\sigma^2 L^D$, where now χ' is the magnetic susceptibility measured in one of the two broken phases, i. e.

$$\chi' = \lim_{h \to 0^\pm} \lim_{L \to \infty} \beta L^D (\langle m^2 \rangle_h - \langle m \rangle_h^2) . \tag{5.4.7}$$

Note that in the low temperature phase we have

$$\lim_{h \to 0^\pm} \lim_{L \to \infty} \langle m \rangle_h = \pm m_0 , \tag{5.4.8}$$

and it is not difficult to show that

$$\langle |m| \rangle = m_0 \operatorname{erf}\left(m_0 \sqrt{\frac{\beta L^D}{2\chi'}}\right) + \sqrt{\frac{2}{\pi}} \sqrt{\frac{\chi'}{\beta L^D}} \exp\left(-m_0^2 \frac{\beta L^D}{2\chi'}\right) , \tag{5.4.9}$$

where erf is the error function defined by

$$\operatorname{erf}(z) = \frac{2}{\pi} \int_0^z e^{-t^2} \mathrm{d}t . \tag{5.4.10}$$

Since for $z \gg 1$ we have (see [12] Eq.7.1.23)

$$\operatorname{erf}(z) \simeq 1 - \frac{1}{\pi z e^{z^2}} , \tag{5.4.11}$$

we obtain

$$\lim_{L\to\infty} \langle |m| \rangle_{h=0} = m_0 = \lim_{h\to 0^+} \lim_{L\to\infty} \langle m \rangle_h \ . \tag{5.4.12}$$

We thus conclude that $\langle |m| \rangle$ can be used as a proxy for the spontaneous magnetization m_0 when performing simulations at $h = 0$: in the high temperature phase $\langle |m| \rangle$ vanishes in the large volume limit, while in the low temperature phase $\langle |m| \rangle$ converges to m_0. For the magnetic susceptibility we should use

$$\chi = \beta L^D \langle m^2 \rangle \ , \quad \chi' = \beta L^D (\langle m^2 \rangle - \langle |m| \rangle^2) \tag{5.4.13}$$

at high and low temperature, respectively. If we use χ' at high temperature we get $\chi' = \chi \left(1 - \frac{2}{\pi}\right)$, while if we use χ in the low temperature phase we get $\chi \simeq \beta L^D m_0$, which diverges in the thermodynamic limit.

From the theoretical point of view χ is a perfectly well defined function also in the low temperature phase (for finite L), and can be used in finite size scaling analyses, however it is not the susceptibility that would be measured by observing the response of a real ferromagnet to an external magnetic field. Since our principal aim in the following will be the study of the critical behavior, we will use χ' both in the high and in the low temperature regime, in order to simplify the analysis. The advantage of χ' with respect to χ is that, in the thermodynamic limit, it diverges only at the critical point, while χ diverges, in the same limit, in all the low temperature phase. Note that that the notation χ, χ' is non standard, so some care is required to understand what "susceptibility" really means.

Let us now come back to Finite Size Scaling (FSS): in the thermodynamic limit we have, close to the critical point

$$\chi' \sim |t|^{-\gamma} = (|t|^{-\nu})^{\gamma/\nu} \sim \xi^{\gamma/\nu} \ . \tag{5.4.14}$$

In a finite and cubic geometry the maximum value that ξ can reasonably reach is L, i. e. the size of the lattice. As a consequence we expect the maximum of χ' to scale, as a function of the lattice size, as $\chi'_{max} \sim L^{\gamma/\nu}$. This simple expectation is confirmed by the result of a more accurate renormalization group analysis, whose outcome is that the behavior for large L of χ' can be parametrized by the form (see [42] §4.4 for an introduction, and [48] for many more details)

$$\chi'(\beta, L) = L^{\gamma/\nu} \chi_1(L/\xi) + \text{corr.} = L^{\gamma/\nu} \chi_2\big[(\beta - \beta_c) L^{1/\nu}\big] + \text{corr.} \ , \tag{5.4.15}$$

where χ_1 and χ_2 are universal scaling functions (which also depend on the boundary conditions adopted and are defined up to multiplicative constants), and the last equality follows from the leading behavior $\xi \sim |t|^{-\nu}$ and the definition of the reduced temperature $t \propto \beta - \beta_c$. The dependence on all the other dynamical scales (smaller than ξ) has been reabsorbed in the dependence on ξ using the renormalization group. In the previous equation the term "corr." stand for finite size corrections, which diverge with L slower than the leading behavior or can even be background analytic terms. In the following analyses we will generally neglect these correction terms, although we will see that they are needed when numerical data are precise enough. Note however that in some cases, e. g. when the leading critical behavior is non-analytic but not divergent their role is essential to explain numerical results (the typical example being that of the specific heat in 3D O(N) models, in which the specific heat exponent α is negative).

The FSS limit, in which Eq. (5.4.15) is valid, is the limit $L \to \infty$ at fixed L/ξ, and it is important to note that, by fixing *any* value of L/ξ, we are approaching the critical point when $L \to \infty$. The standard thermodynamic limit is instead defined by $L \to \infty$ at fixed ξ, and it can be performed only outside criticality. The connection between the two limit is typically hidden in the asymptotic behavior of the scaling function χ_1 for $\xi/L \to 0$.

Eq. (5.4.15) is an example of a FSS relation, and using this equation we can, in principle, determine β_c, $1/\nu$ and γ/ν as follows. Let us consider the typical case in which the function $\chi_2(x)$ has a single maximum at $x = x_0$, with $\chi_2(x_0) = y_0$. We now have to perform simulations for several values of L; for each L value we perform different simulations varying β, until we identify the β value at which the peak is present, which depends on L and will be denoted by

$\beta_{pc}(L)$ (often called the "pseudo-critical" inverse temperature). From Eq. (5.4.15) it follows that $\left[\beta_{pc}(L) - \beta_c\right] L^{1/\nu} = x_0$ if L is large enough that we can neglect the scaling corrections, hence

$$\beta_{pc}(L) = \beta_c + x_0 L^{-1/\nu} \ . \tag{5.4.16}$$

By fitting the numerically obtained values of $\beta_{pc}(L)$ using this expression we can thus estimate both β_c and $1/\nu$. Analogously, we can fit the peak values of χ' using (corrections to this behavior scale typically as $L^{\gamma/\nu-\omega}$, where $\omega > 0$ is a further critical exponent)

$$\chi'_{peak} = y_0 L^{\gamma/\nu} \ , \tag{5.4.17}$$

thus estimating also the value of γ/ν. If β_c and the critical exponents have been correctly estimated, when plotting $\chi'/L^{\gamma/\nu}$ as a function of $(\beta-\beta_c)L^{1/\nu}$ all numerical data should collapse, up to scaling corrections, on a single curve.

Analogous FSS relations hold for any observable which develops a nontrivial critical behavior, and, in particular, from $\langle|m|\rangle \sim (-t)^\beta$ for $t \lesssim 0$ one gets

$$\langle|m|\rangle(\beta, L) = L^{-\beta/\nu} m_2[(\beta - \beta_c)L^{1/\nu}] + \text{corr.} \ , \tag{5.4.18}$$

and from $C \sim |t|^{-\alpha}$ it follows that

$$L^D(\langle\varepsilon^2\rangle - \langle\varepsilon\rangle^2) = L^{\alpha/\nu} C_2[(\beta - \beta_c)L^{1/\nu}] + \text{corr.} \ , \tag{5.4.19}$$

where ε is the energy density E/L^D. To determine β_c using $\langle|m|\rangle$ is nontrivial, while using the specific heat we can adopt, if $\alpha > 0$, a procedure that is completely analogous to the one used when discussing the magnetic susceptibility.

Another commonly used observable in FSS is the Binder cumulant

$$U = \frac{\langle m^4 \rangle}{\langle m^2 \rangle^2} \ . \tag{5.4.20}$$

Note that different notations for this observable exist in the literature (U, U_4, B_4, ...) as well as different normalization constants. The peculiarity of U is that it assumes, in the thermodynamic limit, constant but different values in the high and in the low temperature phases. In the high temperature regime we can use the single Gaussian approximation in Eq. (5.4.2) for the pdf of m, hence we immediately see that $U = 3$ in this phase. In the low temperature phase we have instead to use the double Gaussian approximation in Eq. (5.4.6), from which it follows that, in the large volume limit, U converges to 1. The FSS behavior of the Binder cumulant is particularly simple, since it just depends on the critical exponent ν:

$$U(\beta, L) = U_2[(\beta - \beta_c)L^{1/\nu}] + \text{corr.} \ . \tag{5.4.21}$$

In order for $U(\beta, L)$ to assume, for large L, the correct high and low temperature limits, the function $U_2(x)$ has to satisfy the constraints

$$\lim_{x\to\infty} U_2(x) = 1 \ , \quad \lim_{x\to-\infty} U_2(x) = 3 \ . \tag{5.4.22}$$

The values of $U(\beta, L)$ computed for several β values on different lattices have thus to cross at a point which, up to scaling corrections, coincides with β_c, and the slope of $U(\beta, L)$ at the crossing point is proportional to $L^{1/\nu}$. These properties can be used to estimate β_c and $1/\nu$. The critical value of the Binder parameter[4], often denoted by U^*, is another universal quantity (whose value depends also on the boundary conditions adopted).

[4]While the universal function, e.g., of the susceptibility is defined up to two nonuniversal multiplicative factors (one for the function and one for its argument), the universal function of the Binder cumulant depends just on a single nonuniversal factor, related to the normalization of its argument, since multiplicative terms simplify in the ratio defining the Binder cumulant.

Another very useful quantity, which we however do not discuss, is the so called second moment correlation length ξ_{2nd}, often denoted simply by ξ when there is no possibility of confusion with the infinite volume correlation length, see, e. g., [48] §1.2 or [49]. A peculiarity of this observable is that the values of ξ_{2nd}/L computed on different lattice sizes cross close to a critical point, and the FSS of ξ_{2nd}/L is again of the form Eq. (5.4.21).

Close to a critical point also autocorrelation times diverge, a phenomenon known as critical slowing down, and new critical exponents, related to the "dynamics" of the system, are introduced to describe this behavior:

$$\tau_{\text{exp}} \sim \xi^z \ , \quad \tau_{\text{int}}^{(F)} \sim \xi^{z'} \ . \tag{5.4.23}$$

Note that in these definitions, when local updates are used, the autocorrelation times are defined in unit of complete lattice updates, and not of single site update, with a complete lattice update consisting of L^D single site lattice updates. Although there are theoretical reasons to expect in general $z \neq z'$ (see [6] §2), with $z' \leq z$ (since $\tau_{\text{int}} \leq \tau_{\text{exp}}$, see Sec. 4.1.1) in practice z and z' often turns out to have consistent values. It is important to stress that z and z' are not in general physical quantities, since they characterize the Monte Carlo dynamics, which is completely arbitrary as far as its long time distribution converges to the Gibbs distribution. In particular, different algorithms can exist to simulate the same physical system with very different dynamical critical exponents; an explicit example will be presented in Sec. 6.4. Only in some cases the MC update can be associated with a real physical evolution, and only when this happens the dynamical critical exponents acquire direct physical significance. For example, the dynamics of random thermal fluctuations is analogous to the dynamics generated by the single site Metropolis update, which in this context is known as Glauber or Model A dynamics, see [50].

Although generic dynamical critical exponents (i. e. the dynamical critical exponents of generic updates) are typically of little direct physical relevance, they are extremely important from the algorithmic point of view, since the values of z, z' determine the numerical effectiveness of the MC algorithm. We have indeed seen in Sec. 4.1.1 that the (square) error to be associated with the primary observable $\overline{F}$ is given by

$$\sigma^2_{\overline{F}} = \frac{\sigma_F^2}{N}(1 + 2\tau_{\text{int}}^{(F)}) \ , \tag{5.4.24}$$

where N is the size of the sample, σ_F^2 is the variance of the observable F, and $\tau_{\text{int}}^{(F)}$ the associated integrated autocorrelation time. Let us consider, for example, the case of the magnetization $|m|$: close to a critical point we have

$$\sigma^2_{|m|} = \langle m^2 \rangle - \langle |m| \rangle^2 \sim \chi'/L^D \sim L^{\gamma/\nu - D} \ , \tag{5.4.25}$$

ad if we use $\tau_{\text{int}}^{|m|} \sim L^{z'}$ we get

$$\sigma^2_{\overline{|m|}} \sim \frac{1}{N} L^{z' + \gamma/\nu - D} \ . \tag{5.4.26}$$

If we denote by T_{CPU} the CPU time needed to generate the sample, we have $N \sim T_{\text{CPU}}/L^D$ and finally

$$\sigma^2_{\overline{|m|}} \sim \frac{1}{T_{\text{CPU}}} L^{z' + \gamma/\nu} \ . \tag{5.4.27}$$

To understand the effective numerical significance of this scaling we need some numerical values: typically γ/ν is close to 2, and for local update algorithms (Metropolis or heat-bath), which behaves roughly as a diffusion process[5], we expect $z, z' \approx 2$. If we want the error of $\overline{|m|}$ not to grow when increasing L, we thus have to scale the CPU time approximately as $T_{\text{CPU}} \sim L^4$. If we use instead the cluster algorithm discussed in Sec. 6.4, which is characterized by a extremely small dynamical exponents, it is sufficient to scale the CPU time as L^2.

It is interesting to note that the scaling with L just discussed emerges only close to a critical point. When performing simulations far from phase transitions χ' assumes a finite value in the

[5] We remind the reader that for a diffusion process the mean square distance reached at time t is $\propto \sqrt{t}$, where the proportionality constant depends both on the diffusion constant and on the dimensionality of the system.

thermodynamic limit, as well as the integrated autocorrelation time; from the previous reasoning we thus conclude that the CPU time needed to keep the error of $|m|$ constant does not scale with L. This despite the fact that the number of samples decreases as L^{-D} by increasing L at fixed $T_{\rm CPU}$. This phenomenon goes under the name of (strong) self-averaging, and can be easily explained: far from criticality one has easily $L \gg \xi$, and in this regime different parts of the lattice are effectively independent from each other. If we double the size of L, the number of independent components will be multiplied by 2^D, and for local observables this factor compensates for the smaller number of updates that can be performed at fixed CPU time. Note that locality is essential in the previous reasoning: for quantities that are not written as the average of a local observable, like the specific heat or the magnetic susceptibility, self-averaging simply fails, see [51] §2.3.8.

5.5 An explicit example

In this section we discuss some numerical results obtained for the finite size scaling of the two dimensional Ising model (with vanishing external magnetic field). Our goal is twofold: on one hand we want to show that the leading order FSS relations discussed in the previous section correctly describe numerical data in the large size limit, on the other hand we also want to show that corrections to these relations do exist, which become negligible in the large size limit. For these purposes we reach quite large lattice sizes and use the very efficient single cluster algorithm that will be introduced in Sec. 6.4, in order to keep statistical errors under control also for the largest lattices. Much larger statistics would be required to reach the same accuracy using the local Metropolis or heath-bath updates of Sec. 5.3, as will be discussed in the end of this section when presenting numerical data for the critical slowing down.

Numerical data have been generated using lattice sizes ranging from $L = 20$ up to $L = 160$, performing simulations at 40 different β values for each lattice size. The statistics accumulated is of the order of 10^6 single cluster updates for each simulation point, which correspond to a CPU time ranging from about 30s for each β value on the smallest lattice size to about 80min on the largest lattice size.

In Fig. (5.3) we report data for the main thermodynamic observables as a function of β for the different lattice sizes simulated. Note that, when studying critical phenomena, it is customary to symplify the form of the specific heat and the susceptibilities by removing irrelevant powers of β, which do not affect the critical behavior:

$$C = L^D(\langle \varepsilon^2 \rangle - \langle \varepsilon \rangle^2) \ , \quad \chi = L^D(\langle m^2 \rangle) \ , \quad \chi' = L^D(\langle m^2 \rangle - \langle |m| \rangle^2) \ . \tag{5.5.1}$$

The observed behavior is consistent with theoretical expectations: the average magnetization always vanishes, the average absolute magnetization slowly converges to the spontaneous magnetization by increasing the lattice size, the energy density does not show any divergence, while the specific heat seems to develop a divergence for increasing L. The susceptibility χ diverges, in the thermodynamic limit, in the whole low temperature phase $\beta > \beta_c$, while χ' diverges only at the critical point $\beta = \beta_c$. The Binder cumulant U at finite L smoothly interpolates between 3 and 1, with a crossing point at $\beta = \beta_c$, where the slope diverges for large L.

Of these critical behaviors the only one which requires some more worlds of explanation is that of the specific heat C: the analytically known critical index α of the two dimensional Ising model is $\alpha = 0$, and one could naively think that the specific heat does not diverge at the critical point. This is however not the case: for the two dimensional Ising model $\alpha = 0$ just means that the specific heat diverges for large L slower than any positive power in L. In fact it can be shown that the divergence of C is logarithmic in L, since for $\beta \approx \beta_c$ we have in the thermodynamic limit $C \propto \log|\beta - \beta_c|$ (see, e. g. [45] §2.2.4 or [39] §V).

As a first check that finite size scaling works, we show the collapse plots of the appropriately rescaled observables as a function of $(\beta - \beta_c)L^{1/\nu}$ (see, e.g., Eq. (5.4.15)), using the know critical properties of the two dimensional Ising model,

$$\beta_c = \frac{1}{2}\log(1+\sqrt{2}) \simeq 0.4406867935 \ , \quad \nu = 1 \ , \quad \gamma = 7/4 = 1.75 \ , \quad \beta = 1/8 \ , \tag{5.5.2}$$

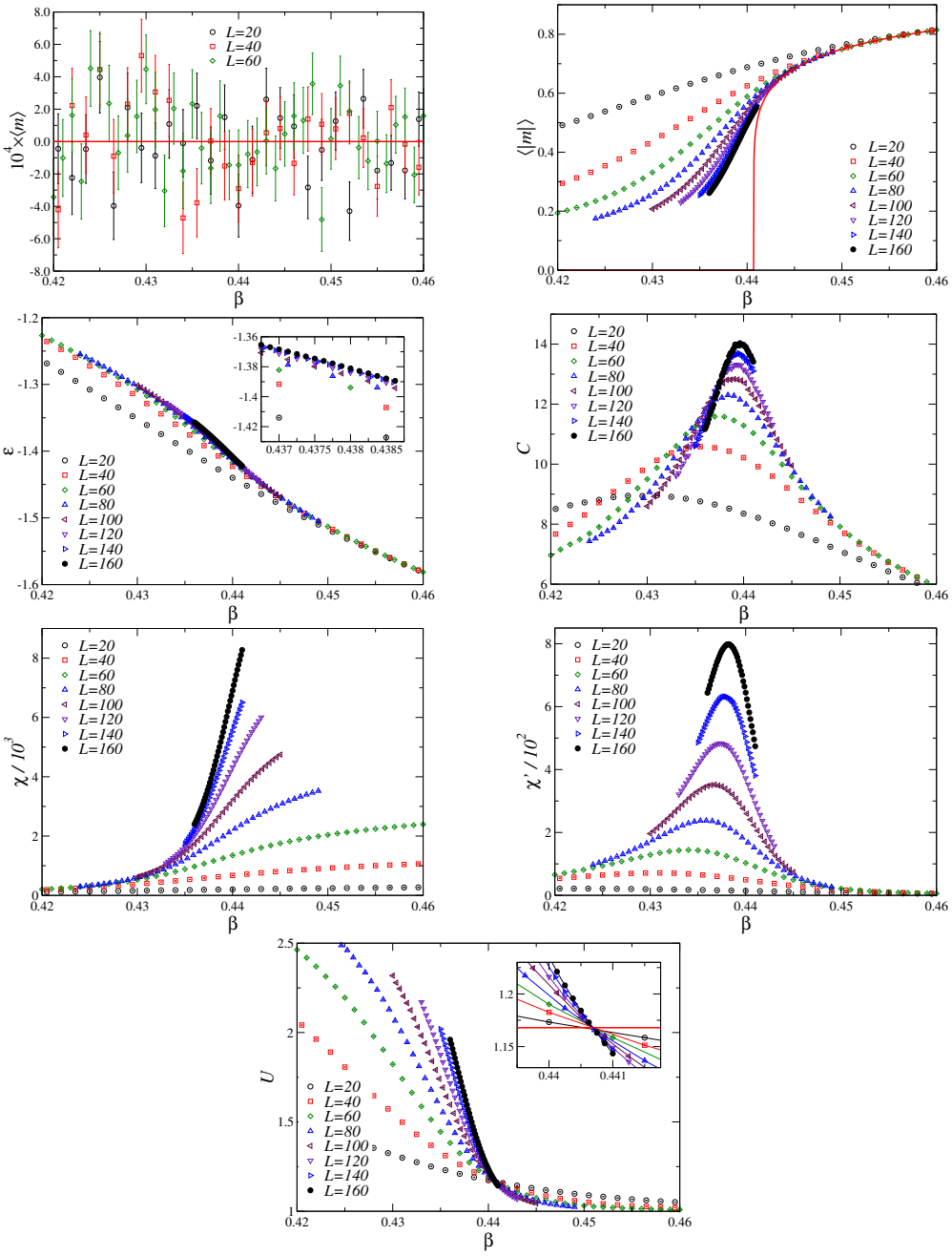

Figure 5.3: Plot of average magnetization, average absolute magnetization, energy density, specific heat, susceptibility χ, subtracted susceptibility χ', and Binder cumulant as a function of β. The solid red line in the plot of $\langle |m| \rangle$ reprents the analytically known spontaneous magnetization for $L \to \infty$, which is given by $m_0(\beta) = (1 - \sinh^{-4}(2\beta))^{1/8}$ (see e. g. [45] §2.2.5 or [39] §X). The solid red line in the inset of the plot of U denotes the analytically computed value of the Binder cumulant at β_c for periodic b. c.: $U_4^* = 1.1679227(4)$, see [52].

see Sec. 7.A and references therein. This check is possible (almost) only in the two dimensional Ising case, for which analytical results exist; however we will soon show how to estimate β_c, ν and γ, so a collapse plot of this type can always be used, to check *a posteriori* the consistency of estimated critical properties. We do not consider the finite size scaling of C due to the subtleties related to the logarithmic divergence previously noted. Numerical results are displayed in Fig. (5.4), and a nice data collapse for the large lattices is clear; corrections to scaling, that have been neglected in the previous section and vanish as L^{-2} (see, e. g., [48] §3.5), can be easily seen from the zooms presented in the insets. Numerical data are thus fully consistent with the analytically known values of β_c, ν, γ and β (critical exponent).

To estimate the values of the critical temperature β_c, and of the critical exponents γ and ν, we can follow the strategy outlined in Fig. (5.4): as a first step we have to estimate, for several lattice sizes L, the temperatures $\beta_{pc}(L)$ at which $\chi'(\beta)$ reaches its maximum value. For this purpose we can fit $\chi'(\beta)$, for fixed L and for β close to the peak position, with a function of the form

$$\chi'(\beta) \approx a(\beta - \beta_{pc}(L))^2 + \chi'_{max}(L) \ , \tag{5.5.3}$$

where a, $\beta_{pc}(L)$ and $\chi'_{max}(L)$ are fit parameters. Close enough to the peak value of $\chi'(\beta)$ this function surely well describes numerical data (it is just a Taylor expansion truncated to second order), but the range of validity of this functional form is not known *a priori*. For this reason we have to try several fit ranges, to identify the ones corresponding to fits with reasonable χ^2/d.o.f values. The residual dependence of the optimal fit parameters on the fit range has to be considered as a systematic error of the fit procedure.

The results obtained for $\beta_{pc}(L)$ and $\chi'_{max}(L)$ are shown in Fig. (5.5). It is clear that $\beta_{pc}(L)$ saturates to a finite limiting value for increasing L values, while χ'_{max} diverges in the same limit. As discussed in the previous section, the values of $\beta_{pc}(L)$ should scale, for large values of L, according to the functional form

$$\beta_{pc}(L) \approx \beta_c + bL^{-1/\nu} \ , \tag{5.5.4}$$

while $\chi'_{max}(L)$ should scale as (for the two dimensional Ising model the exponent ω parametrizing the leading scaling correction is larger then γ/ν).

$$\chi'_{max}(L) \approx c_0 + c_1 L^{\gamma/\nu} \ . \tag{5.5.5}$$

By fitting the data displayed in Fig. (5.5) using these functional forms for several fit ranges, i. e. by systematically removing the smallest lattice sizes (note that the previous functional forms are just leading large L terms), we can estimate the values of β_c, ν and γ/ν (and hence of γ). The results of this analysis are shown in Fig. (5.6), where we report data corresponding to four different fit ranges: $L \geq L_{min}$ with $L_{min} = 20, 40, 60, 80$. The χ^2/d.o.f of all these fits is reasonable, and fit results are quite stable; obviously errorbars increase by reducing the fit range used, and hence the size of the data set. Taking into account the systematics of the fit procedure we report as our final estimates the following values

$$\beta_c = 0.44075(10) \ , \quad 1/\nu = 1.005(25) \ , \quad \gamma/\nu = 1.746(4) \ , \tag{5.5.6}$$

which correspond to the bands shown in Fig. (5.6).

Using just the data of the Binder cumulant close to the crossing, and performing a slightly more sophisticated analysis using Eq. (5.4.21), we get instead

$$\beta_c = 0.44069(3) \ ; \quad \nu = 1.000(10) \ . \tag{5.5.7}$$

We close this section by discussing the critical slowing down in the two dimensional Ising model, comparing the local Metropolis and the cluster updates. For this purpose we performed simulations at fixed $\beta = \beta_c$, using 5×10^7 updates of the whole lattice when adopting the local Metropolis algorithm (i. e., $5 \times 10^7 L^2$ local updates) or 5×10^7 single cluster updates. The scaling with L of the error bars is shown in Fig. (5.7) for two test observables ($\langle |m| \rangle$ and χ), but the same behavior

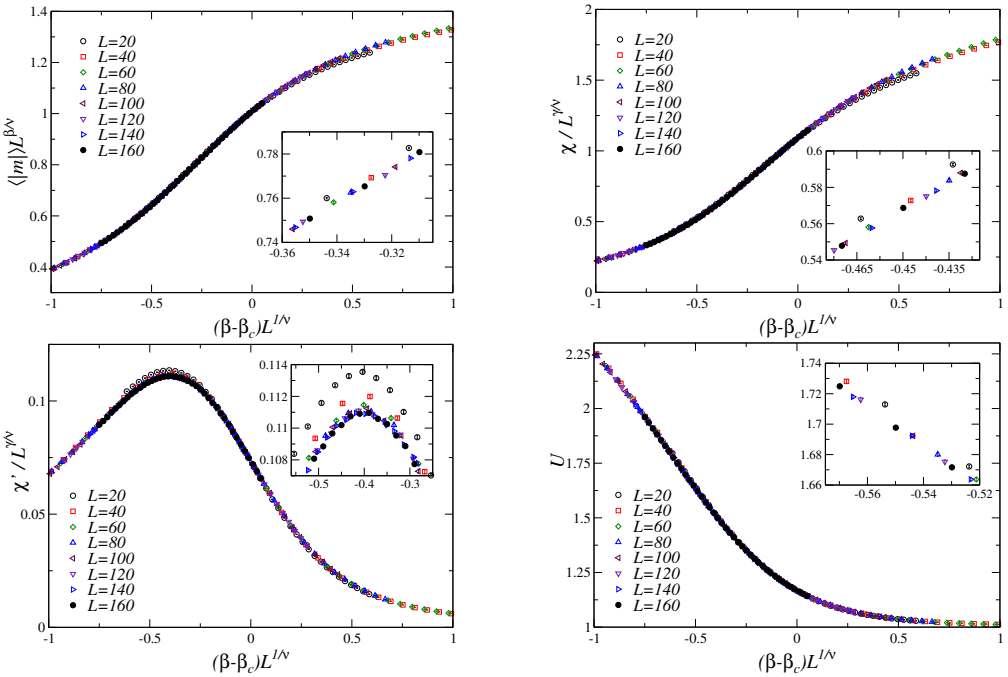

Figure 5.4: FSS of the average absolute magnetization, of the susceptibility, of the substracted susceptibility χ', and of the Binder cumulant U, obtained by using the analytically known values reported in Eq. (5.5.2). Scaling corrections can be clearly seen in the insets.

is seen for all observables: errors scale almost proportionally to L for the local Metropolis update, while they are independent of L for the cluster update. From Eq. (5.4.26) we expect errors to scale proportionally to $L^{(z'+\gamma/\nu-2)/2}$, and we have just seen that $\gamma/\nu = 1.75$, hence we conclude that $z' \approx 2.25$ for the local Metropolis update. For the cluster update we have instead $z' \lesssim 0.25$, since no divergence can be seen for increasing L values in the cluster update data of Fig. (5.7).

A similar conclusion can be reached by studying the autocorrelation function of the magnetization using local Metropolis updates: results for the autocorrelation function (again at $\beta = \beta_c$) for different lattice sizes are shown in Fig. (5.8) (left panel). By fitting with an exponentially decreasing function these data we can extract the exponential autocorrelation times $\tau_{\rm exp}(L)$, which are shown in Fig. (5.8) (right panel) together with a fit of the form $\tau_{\rm exp} \approx aL^z$. The optimal fit value for the exponent z is $z \approx 2.1$. The precise determination $z = 2.1667(5)$ of the dynamic critical exponent of the two dimensional Ising model with local Metropolis update has been obtained in Ref. [53], and this result is in the same ball-park of our rough estimates.

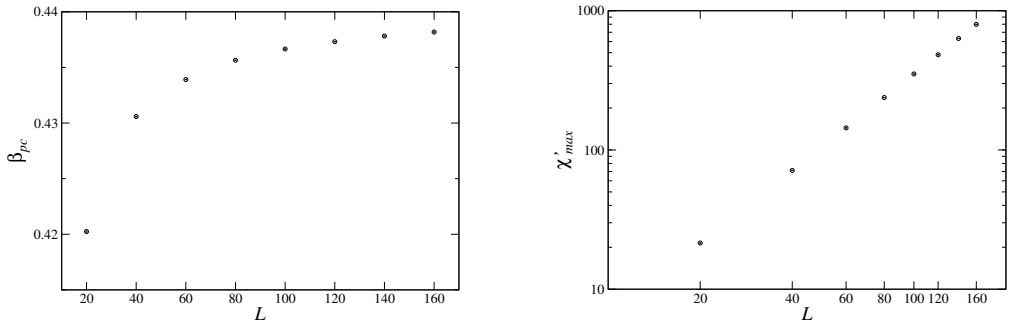

Figure 5.5: Plot of the values $\beta_{pc}(L)$ and $\chi'_{max}(L)$ obtained by fitting $\chi'(\beta)$, for each value of L and close to the maximum, using the functional form in Eq. (5.5.3).

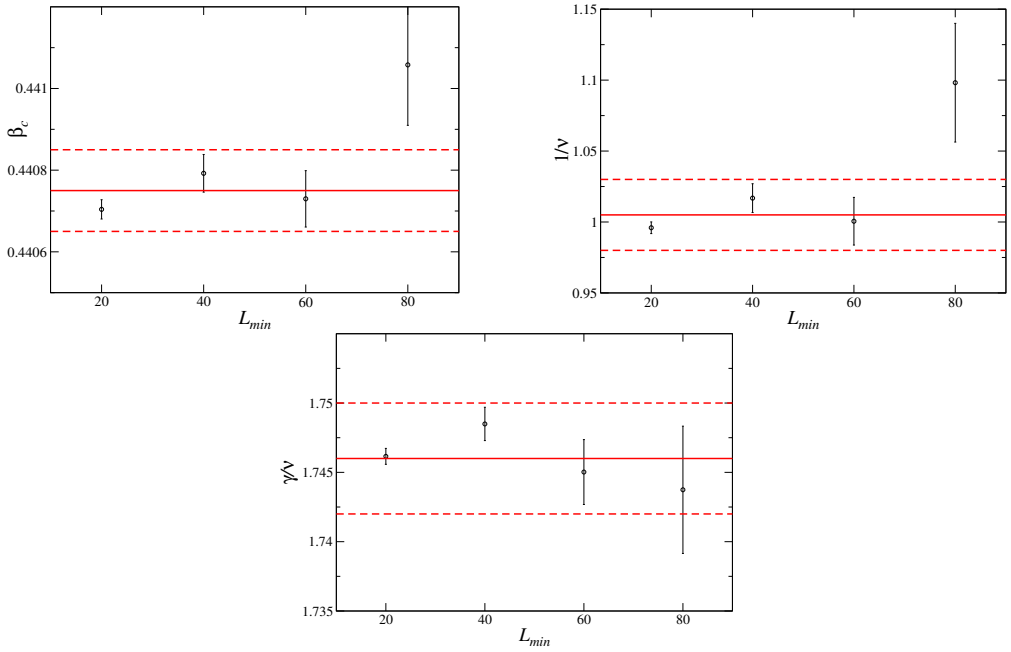

Figure 5.6: Values of β_c, $1/\nu$ and γ/ν obtained by fitting data shown in Fig. (5.5), using the fit ranges $L \geq L_{min}$. Horizontal bands denote the final values (with error) reported in the main text, obtained by taking into account the systematic errors of the fit procedure.

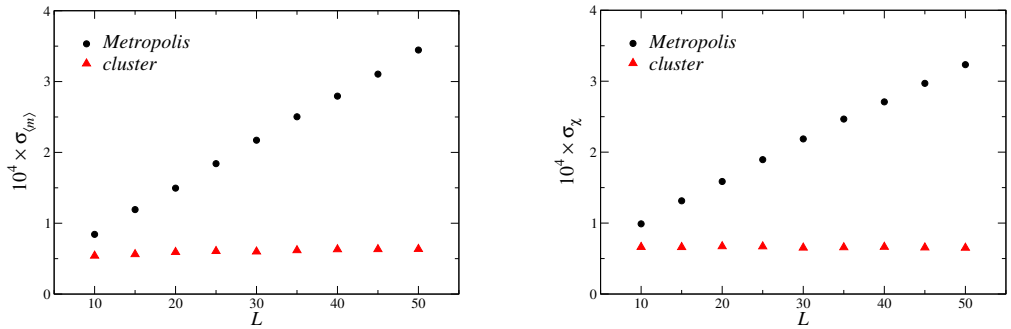

Figure 5.7: Scaling of the errors of $\langle |m| \rangle$ and χ at $\beta = \beta_c$, as a function of L, for the local Metropolis and the cluster updates. The statistics is kept constant when increasing L.

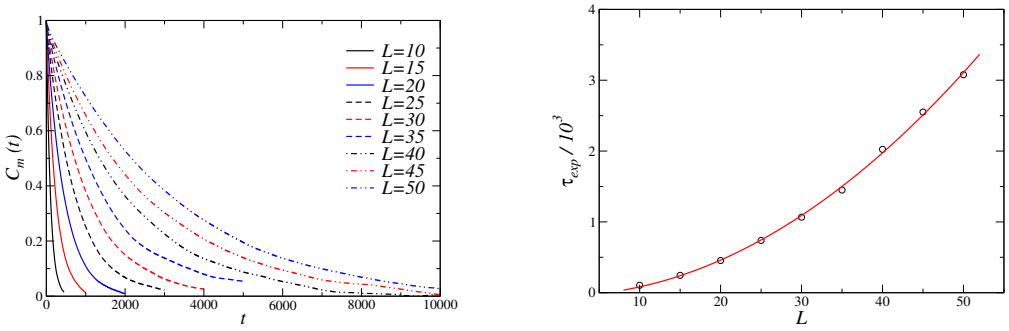

Figure 5.8: (left) Autocorrelation function of the magnetization when using local Metropolis updates. (right) Exponential autocorrelation times extracted from the autocorrelation function of the magnetization when using local Metropolis update. The solid line is a fit of the form aL^z and the optimal fit parameter is $\zeta \approx 2.1$.

Chapter 6

Other models and algorithms

6.1 Potts models

The q-states Potts model differs from the Ising model in that the variables associated with each lattice site can assume q values instead of the 2 values of the Ising model: $s_{\boldsymbol{x}} = 0, 1, \ldots, q-1$. The energy of a configuration in the Potts model is given by

$$E[\{s_{\boldsymbol{x}}\}] = -J \sum_{\langle \boldsymbol{x},\boldsymbol{y} \rangle} \delta_{s_{\boldsymbol{x}},s_{\boldsymbol{y}}} - h \sum_{\boldsymbol{x}} \delta_{s_{\boldsymbol{x}},0} \ , \tag{6.1.1}$$

hence there is a contribution $-J$ for each couple of nearest neighbor sites with the same "orientation", and a contribution $-h$ for each site whose variable has the 0 value. The model is thus ferromagnetic if $J > 0$ and anti-ferromagnetic if $J < 0$, while h acts as an external magnetic field for the variables whose value is 0. Note that the choice of coupling the magnetic field only to sites $\boldsymbol{x}$ with vanishing $s_{\boldsymbol{x}}$ is completely arbitrary: the use of $-h\delta_{s_{\boldsymbol{x}},\alpha}$ with any value $0 \leq \alpha \leq q-1$ is equally legitimate. As we did for the Ising model, we consider only the ferromagnetic case, and by measuring the temperature in units of J we can formally fix $J = 1$ (see Sec. 5.1). To completely define the model we also have to specify the boundary conditions, and, if not otherwise specified, we will always assume periodic b. c.

When using periodic b. c. (or, more generally, b. c. which do not favor any state), for $h = 0$ the energy of a configuration just depends on the number of nearest neighbor sites whose site variables have the same value. As a consequence, if denote by V_i (with $i = 0, \ldots, q-1$) the set of the sites with $s_{\boldsymbol{x}} = i$, and we perform the transformation

$$\text{if } \boldsymbol{y} \in V_i \text{ then } s_{\boldsymbol{y}} = \sigma(i) \ , \tag{6.1.2}$$

where the function $i \to \sigma(i)$ is a permutation of the integers $0, \ldots, q-1$, the energy does not change. The q-states Potts model is thus invariant under the symmetric group S_q, i. e. the group of permutations of q objects. Note that $S_2 = \mathbb{Z}_2$ and the 2-states Potts model can be exactly mapped to an Ising model: for $q = 2$ we have indeed $s_{\boldsymbol{x}} = 0, 1$, and we can define $I_{\boldsymbol{x}} = 2(s_{\boldsymbol{x}} - 1/2) = \pm 1$. It is then immediate to verify that in this case

$$\delta_{s_{\boldsymbol{x}},s_{\boldsymbol{y}}} = \frac{I_{\boldsymbol{x}} I_{\boldsymbol{y}} + 1}{2} \quad (\text{for } q = 2) \ , \tag{6.1.3}$$

hence the $q = 2$ Potts model with couplings J and h is equivalent to the Ising model with couplings $J/2$ and $-h$.

When $h = 0$, the S_q symmetry of the q-states Potts model is spontaneously broken to S_{q-1} in the low temperature phase $\beta < \beta_c$. For $D = 2$ it can be shown that $\beta_c = \log(1 + \sqrt{q})$ and that the transition is continuous for $q = 2, 3, 4$ and discontinuous for $q \geq 5$, with the latent heat being a monotonously increasing function of $q \geq 5$ (see, e. g., [54], [55] §12). In $D = 3$ the transition

is continuous only for $q = 2$. When $h > 0$ a single state is favored, and the symmetry S_q is explicitly broken to its subgroup S_{q-1}, which does not get spontaneously broken for any value of β. If instead $h < 0$ a single state is disfavoured, and the symmetry is once again explicitly broken to S_{q-1}; however in this case this residual symmetry is spontaneously broken to S_{q-2} (if $q > 2$, obviously) in the low temperature phase, see, e. g., [56] for the $D = 3$, $q = 3$ case.

Two different order parameters can be introduced for the Potts model: a real and a complex one. Let us start from the real one:

$$m_1 = \frac{1}{L^D} \sum_{\boldsymbol{x}} \frac{q\delta_{s_{\boldsymbol{x}},0} - 1}{q-1} \ . \tag{6.1.4}$$

Using this definition $\langle m_1 \rangle = 0$ if all the states have the same probability of occurring in the simulation, and $\langle m_1 \rangle \neq 0$ otherwise (more precisely: if the state 0 has a probability of occurring which is not equal to the average probability of the other $q-1$ states). As for the magnetic coupling in Eq. (6.1.1), the use of the 0 state as reference state in the definition of m_1 is completely arbitrary. The complex order parameter is more symmetric, since it does not use a reference state, and it is defined by

$$m_2 = \frac{1}{L^D} \sum_{\boldsymbol{x}} \exp\left(i \frac{2\pi}{q} s_{\boldsymbol{x}} \right) \ . \tag{6.1.5}$$

Also in this case $\langle m_2 \rangle$ vanishes if all the states are equiprobable and it is nonzero otherwise.

On a finite lattice with periodic boundary conditions we always have, for $h = 0$, $\langle m_1 \rangle_L = \langle m_2 \rangle_L = 0$, as can be shown by adapting the proof of the analogous identity $\langle m \rangle_L = 0$ presented in Sec. 5.1 for the Ising model (instead of a spin flip we have to use a permutation of the states). This means that the numerical estimates of $\langle m_1 \rangle_L$ and $\langle m_2 \rangle_L$ have to be compatible with zero if the simulation time is long enough. Since the order parameter m_1 uses a reference state in its definition, it is not completely trivial to introduce for this order parameter a proxy of the spontaneous magnetization, which plays a role analogous to $\langle |m| \rangle$ for the Ising model. For the order parameter m_2 we can instead use $\langle |m_2| \rangle$, where now $|\ \ |$ denotes the absolute value of a complex number. If we use m_1 we can thus only study the susceptibility $L^D \langle m_1^2 \rangle$ (as in Sec. 5.5 we neglect irrelevant powers of β), which diverges in the whole low temperature phase; if we use instead m_2 we can once again introduce both the susceptibility $\chi = L^D \langle |m_2|^2 \rangle$ and the subtracted susceptibility $\chi' = L^D (\langle |m_2|^2 \rangle - \langle |m_2| \rangle^2)$, and χ' diverges only at β_c. Let us stress once again that both χ and χ' can be used in a FSS analysis, the only advantage of χ' being that it simplifies the estimation of β_c (and ν).

Let us now discuss the numerical simulation of the Potts models. Following the same line of thought of Sec. 5.3 it is immediate to see that the following algorithm produces an irreducible and aperiodic Markov chain, which satisfies the detailed balance with respect to the Gibbs distribution:

1. select with uniform pdf a site $\boldsymbol{r}$ of the lattice,
2. define the trial configuration as the configuration in which only the value $s_{\boldsymbol{r}}$ is changed; in particular define the trial state $s'_{\boldsymbol{r}}$ by sampling with uniform pdf the $q-1$ states different from $s_{\boldsymbol{r}}$,
3. accept the trial configuration with probability $\min(1, e^{-\beta(E'-E)})$, where E is the energy of the initial configuration and E' is the energy of the trial configuration. If the trial configuration is not accepted, keep the old one.

Step 2. is the analogous of the spin flip in the Ising model, and the selection probability of $s'_{\boldsymbol{r}}$ has to be uniform in order to guarantee the symmetry of the selection matrix, and thus the possibility of using the Metropolis algorithm instead of the Metropolis-Hastings one. If we denote by r a random number in $[0, 1)$ with uniform pdf, step 2. can be implemented as follow:

$$s'_{\boldsymbol{r}} = \lfloor s_{\boldsymbol{r}} + 1 + (q-1)r \rfloor \bmod q \ , \tag{6.1.6}$$

indeed the argument x of the floor function $\lfloor\ \ \rfloor$ satisfies $s_{\boldsymbol{r}} + 1 \leq x < s_{\boldsymbol{r}} + q$, hence $s_{\boldsymbol{r}} + 1 \leq \lfloor x \rfloor \leq s_{\boldsymbol{r}} + q - 1$, with all the integer values in this range being equiprobable.

An algorithm associated with an irreducible and aperiodic Markov chain, which satisfies the balance condition but not the detailed balance condition is instead the following (compare with the analogous discussion in Sec. 5.3)

1'. sweep the lattice in a deterministic way, selecting site $\boldsymbol{r}$
2'. define the trial configuration as the configuration in which only the value $s_{\boldsymbol{r}}$ is changed; in particular define the trial state $s'_{\boldsymbol{r}}$ by sampling with uniform pdf all the q states,

followed by the usual accept/reject step 3., which does not change. Point 2. can obviously be implemented by using $s'_{\boldsymbol{r}} = \lfloor qr \rfloor$, where $r \in [0,1)$ is a random number with uniform pdf.

In Sec. 5.3 it was shown by an explicit example that the condition 2'. (or analogous ones in which a nontrivial possibility of selecting the same value is introduced) has to be necessarily adopted when using a deterministic sweep of the lattice, since otherwise the associated Markov chain is not irreducible nor aperiodic. It is a peculiarity of the models with $q > 2$ (not necessarily Potts models) that an irreducible and aperiodic Markov chain is obtaining also by using

1'. sweep the lattice in a deterministic way, selecting site $\boldsymbol{r}$
2. define the trial configuration as the configuration in which only the value $s_{\boldsymbol{r}}$ is changed; in particular define the trial state $s'_{\boldsymbol{r}}$ by sampling with uniform pdf the $q-1$ states different from $s_{\boldsymbol{r}}$,

once again followed by the step 3.

The proof of this fact is not completely trivial, and we approximately follow the discussion presented in [47]. We denote by $W^{(1)}, \dots, W^{(L^D)}$ the stochastic matrices associated with a single spin update, and by $W = W^{(\sigma(1))} \cdots W^{(\sigma(L^D))}$ the stochastic matrix associated with a deterministic sweep of the whole lattice, where $i \to \sigma(i)$ is a permutation which specifies the way in which the lattice is swept. Since the matrices $W^{(k)}$ are associated with the update of a single site, since all the values of the site variable different from the original one are equiprobable, and since the acceptance probability never vanishes, the restriction of the matrix $W^{(k)}$ to the single site variable on which it acts nontrivially is a $q \times q$ matrix $a^{(k)}_{ij}$ whose elements satisfy $a^{(k)}_{ij} > 0$ if $i \neq j$ and $a^{(k)}_{ii} = 0$. It is then immediate to show that all the entries of the matrix $(a^{(k)})^2$ are positive definite if[1] $q > 2$. From this fact we can draw two consequences: it can be easily shown that the stochastic matrix

$$W' = [W^{(\sigma(1))}]^2 \cdots [W^{(\sigma(L^D))}]^2 \tag{6.1.7}$$

represents an irreducible and aperiodic Markov chain, moreover it follows from the theorems of Sec. 3.2 that $(a^{(k)})^2$ has a single eigenvalue $\lambda = 1$, with all the other eigenvalues satisfying $|\lambda| < 1$. In particular, if λ is an eigenvalue of $a^{(k)}$ with $|\lambda| = 1$, then $\lambda = 1$. Since $W^{(k)}$ can be written in a block diagonal form, in which the two diagonal blocks are a $q^{L^D-1} \times q^{L^D-1}$ identity matrix and $a^{(k)}$, the same property is true also for $W^{(k)}$: if λ is an eigenvalue of $W^{(k)}$ with $|\lambda| = 1$, then $\lambda = 1$.

Let us now suppose that v is an eigenvector of W, and the eigenvalue satisfies $|\lambda| = 1$. From

$$W^{(\sigma(1))} W^{(\sigma(2))} \cdots W^{(\sigma(L^D))} v = \lambda v \tag{6.1.8}$$

it follows that $W^{(\sigma(2))} \cdots W^{(\sigma(L^D))} v$ is an eigenvector of $W^{(\sigma(1))}$ with eigenvalue λ, which satisfies $|\lambda| = 1$, hence, due to the previous discussion, we must have $\lambda = 1$. Moreover, by induction, it easily follows that v is an eigenvector of any $W^{(k)}$ with eigenvalue 1.

Let us finally assume that two linearly independent eigenvectors of W exist, denoted by v and w, both associated with the eigenvalue $\lambda = 1$. From what we have just seen v and w have to be also eigenvectors of all the $W^{(k)}$ matrices, with eigenvalue 1. Then they should also be two linearly independent eigenvectors of W' with eigenvalue 1, which is however impossible: W' is associated with an irreducible and aperiodic Markov chain, hence the eigenspace of $\lambda = 1$ is unidimensional.

We have thus proved the following properties of the spectrum of the stochastic matrix W: the eigenvalue $\lambda = 1$ is nondegenerate, and all the eigenvalues with $\lambda \neq 1$ satisfy $|\lambda| < 1$. As noted on pag. 27 ("Sometimes it can be useful to note...") this implies that the Markov chain associated with W is irreducible and aperiodic.

Analogously to the Ising model case, to compute the difference of energies $E' - E$ entering the Metropolis step we do not need to sum up the contributions of all the sites, we just need to study the nearest neighbors of the site $\boldsymbol{r}$ we want to update. In particular, it is convenient to define the q quantities $N^{(\boldsymbol{r})}_i$:

$$N^{(\boldsymbol{r})}_i = \sum_{\langle \boldsymbol{x}, \boldsymbol{r} \rangle} \delta_{s_{\boldsymbol{x}}, i} \ , \tag{6.1.9}$$

[1] This is the only point in which this fundamental assumption is used. This sentence is not true for $q = 2$, as the Ising case shows.

i. e. the number of nearest neighbors of the site $\boldsymbol{r}$ whose site variable is equal to i. Using this definition we immediately have

$$\exp\Big(-\beta(E'-E)\Big) = \exp\Big(\beta J(N^{(\boldsymbol{r})}_{s'_{\boldsymbol{r}}} - N^{(\boldsymbol{r})}_{s_{\boldsymbol{r}}})\Big)\ , \tag{6.1.10}$$

moreover these numbers can be computed once at the beginning of the simulation and stored in an array.

Using the variables $N^{(\boldsymbol{r})}_i$ it also simple to write an heat-bath update algorithm for the q-states Potts model: it is sufficient to select $s_{\boldsymbol{r}} = \alpha$ with probability

$$p_\alpha = \frac{\exp\Big(\beta J N^{(\boldsymbol{r})}_\alpha\Big)}{\sum_{k=0}^{q-1}\exp\Big(\beta J N^{(\boldsymbol{r})}_k\Big)}\ , \tag{6.1.11}$$

which can be done by introducing $Q_i = \sum_{\alpha\le i} p_\alpha$ (for $i = 0,\dots,q-1$), generating a random number $r \in [0,1)$ with uniform pdf, and selecting the smallest i such that $r \le Q_i$. Note that the largest the value of q is, the more efficient the heat-bath algorithm is with respect to the Metropolis one.

6.1.1 FSS at discontinuous transitions

Finite size scaling at discontinuous transitions is in some cases formally similar to the one that is present at continuous phase transitions, but the physics underlaying the two cases is completely different: at a discontinuous transition the correlation length is *not* divergent, and RG arguments can not be directly applied. Strictly speaking there is not even universality in discontinuous transitions, since the observed FSS behavior strongly depends on the boundary conditions adopted, see, e. g., [57].

If we consider periodic b. c. (which do not favor any state and do not induce domain-walls), we can use for the pdf of the energy density a double Gaussian approximation analogous to the one used for the magnetization density in the low temperature phase of the Ising model in Sec. 5.4. This does not happen by chance: in the low temperature phase of the Ising model a discontinuous transition is present when going from $h = 0^-$ to $h = 0^+$. Note however that in the Ising case the two Gaussian distributions correspond to two different phases related by the $\mathbb{Z}_2$ symmetry, hence the height and width of the two Gaussian have to be same; this is generically not the case for the phases coexisting at a temperature driven discontinuous transition. Coexistence of the two phases at the discontinuous transition only implies the statistical weights of the two phases to be the same, hence the areas of the two Gaussian distributions have to be the same [58]. Using this double Gaussian approximation it is possible to obtain the scaling $\chi'_{max} \propto L^D$ and $C_{max} \propto L^D$ of the maxima of χ' and C as a function of the lattice size L, as well as the scaling $\beta_{pc}(L) - \beta_c \propto 1/L^D$ of the pseudo-critical temperature (see, e. g., [58], or [59] and references therein for a rigorous discussion). When using periodic b. c., FSS at discontinuous transition is thus analogous to the one observed at continuous transitions, with effective critical exponents $\alpha = \gamma = 1$ and $\nu = 1/D$.

An analogue of the critical slowing down is present also at first order phase transitions, although in this case the precise form of the scaling depends on the boundary conditions adopted. For periodic b. c. the autocorrelation time grows with the lattice size exponentially in L^D when using local update algorithms, and this timescale can be interpreted as the typical time needed to explore the different coexisting phases. This time is the equivalent, in the quantum context, of the inverse of the tunneling probability, which is indeed exponential in the volume. The reason for this behavior is quite clear in the two Gaussian approximation: to switch from one Gaussian to the other by means of local moves we have to cross a region where the probability density is exponentially small in the volume. Note however that this behavior (unlike the case of continuous phase transitions) strongly depends on the boundary conditions adopted: if open b. c. are used it is simpler for "bubbles" of the other phase to enter from the boundary, and the critical slowing down is less severe, at the expense of larger finite volume corrections. To cope with the exponential critical slowing down that is present when using periodic b. c. one possibility is the use of the multicanonical algorithm [60]:

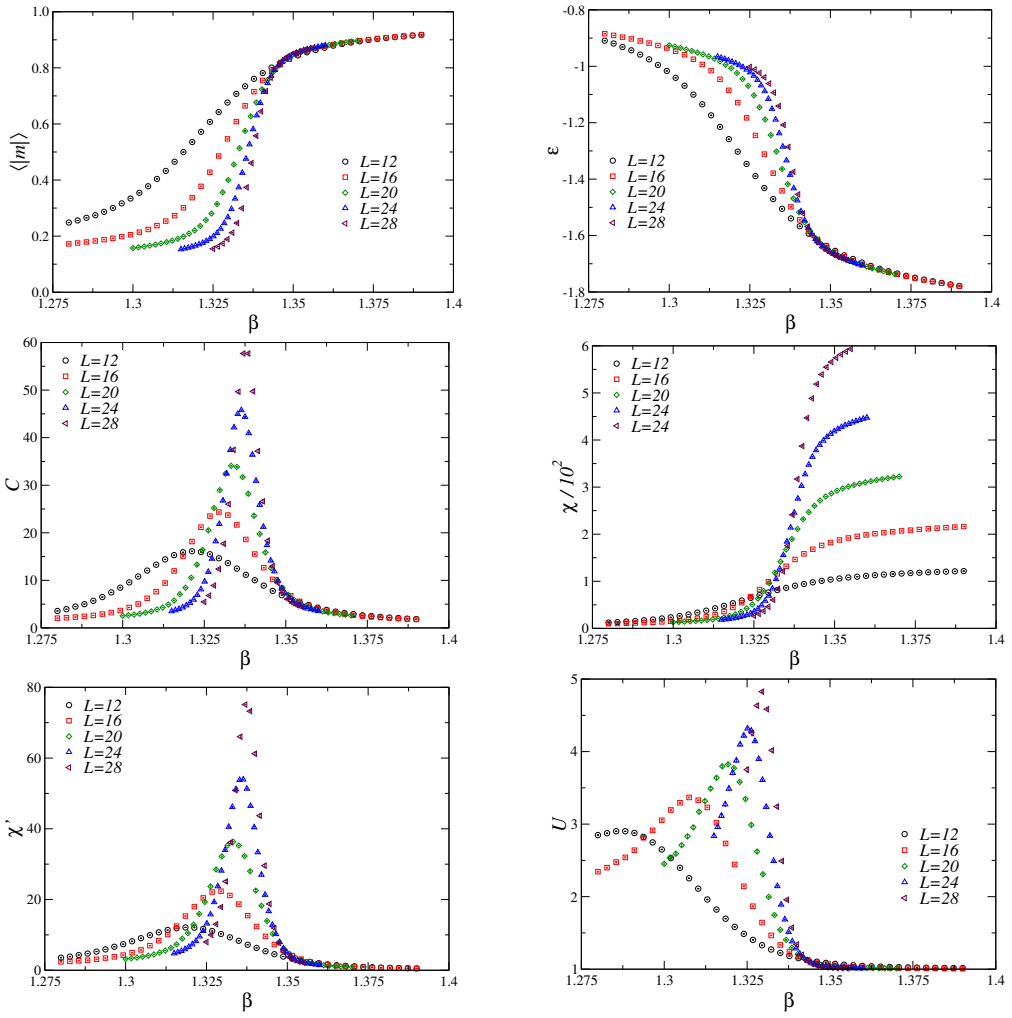

Figure 6.1: Plot of average absolute magnetization, energy density, specific heat, susceptibility χ, subtracted susceptibility χ', and Binder cumulant as a function of β for the two dimensional 8-states Potts model.

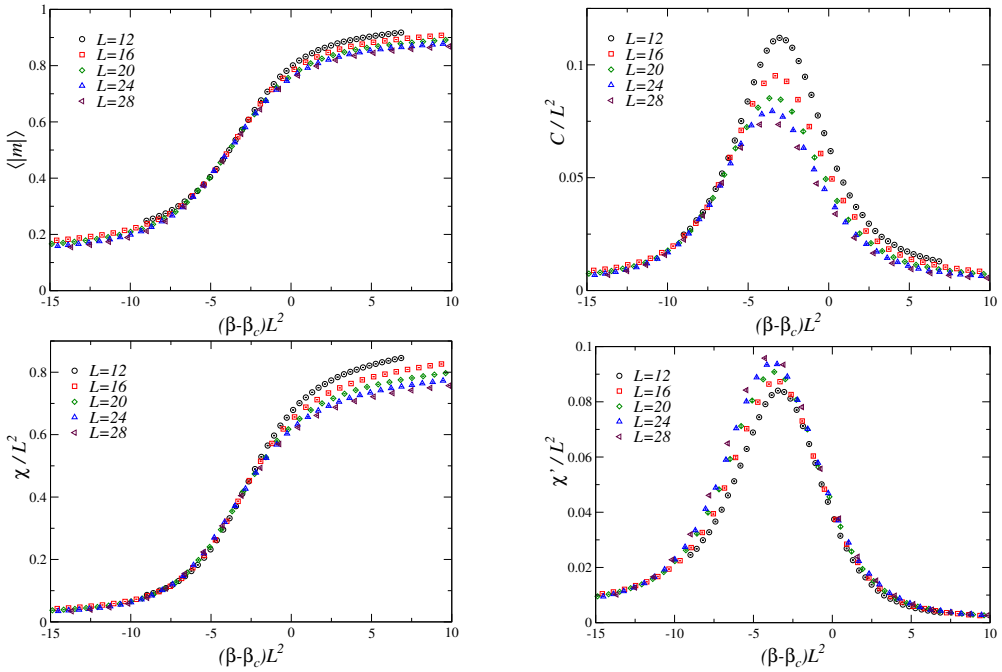

Figure 6.2: FSS of the average absolute magnetization, the specific heat, the susceptibility and the subtracted susceptibility. The expected theoretical values $\beta_c = \log(1+\sqrt{8})$, $\nu = 1/2$, $\gamma = 1$ and $\beta = 0$ have been used.

an auxiliary probability distribution is sampled, in which no obstacles prevent to move from one coexisting phase the other, and the results are finally reweighted to estimate expectation values with respect to the original Gibbs distribution[2]. For more informations of discontinuous transitions and their simulation see, e. g., [57] and [61].

We close this section by presenting some numerical results for the 8-states two dimensional Potts model, which displays a discontinuous phase transition at $\beta_c = \log(1+\sqrt{8}) \approx 1.34245$. These results have been obtained by using 2×10^7 complete lattice sweeps of the heat-bath algorithm for $L = 12, 16, 20$, and 5×10^7 complete sweeps for $L = 24, 28$. Simulation time for a single β value goes from approximately 5.5 min on $L = 12$ to about 70 min for $L = 28$. The complex order parameter m_2 has been used, and the β factors in C, χ, and χ' have been neglected, as done in Sec. 5.5.

In Fig. (6.1) raw data are presented, while in Fig. (6.2) data have been rescaled according to the expected behavior at discontinuous transitions for two-dimensional models with periodic boundary conditions. The discontinuous nature of the transition is quite clear (the Binder cumulant U diverges), however significant scaling corrections are evident. To obtain result with similar accuracy on larger lattices, without using more sophisticated simulation algorithms, would however require a significantly larger simulation time, due to the exponential critical slowing down.

6.2 Clock models

The configuration space of the clock models is the same of the Potts models: to each site $\boldsymbol{x}$ of the lattice a variable $s_{\boldsymbol{x}}$ is associated, which takes value in $\{0, \ldots, q-1\}$, however the energy of a

[2]Note that in this case the overlap problem discussed in Sec. 2.2 is absent: the sampled distribution is broader than the original one.

configuration is now

$$E = -J \sum_{\langle \boldsymbol{x},\boldsymbol{y}\rangle} \cos\left(\frac{2\pi}{q}(s_{\boldsymbol{x}} - s_{\boldsymbol{y}})\right) - h \sum_{\boldsymbol{x}} \cos\left(\frac{2\pi}{q} s_{\boldsymbol{x}}\right) \ . \tag{6.2.1}$$

Clock models can be rewritten also using the complex site variable $C_{\boldsymbol{x}} = \exp\left(i\frac{2\pi}{q}s_{\boldsymbol{x}}\right)$, indeed it is immediate to verify that the energy of a configuration is equal to

$$E = -J \sum_{\langle \boldsymbol{x},\boldsymbol{y}\rangle} \Re\left(C^*_{\boldsymbol{x}} C_{\boldsymbol{y}}\right) - h \sum_{\boldsymbol{x}} \Re\left(C_{\boldsymbol{x}}\right) \ . \tag{6.2.2}$$

Note that sometimes, especially in the early references, clock models are called planar Potts models.

In the Potts models the interaction energy of two nearest neighbor sites can only take two values (for any q), depending whether the variables associated with the two sites are equal or not; in clock models, instead, more possibilities exist, at least for $q > 3$. For $q = 2$ it is simple to show that the clock model is just the Ising model, while for $q = 3$ the function $\cos\left(\frac{2\pi}{q}(s_{\boldsymbol{x}} - s_{\boldsymbol{y}})\right)$ can only assume two values (1 if $s_{\boldsymbol{x}} = s_{\boldsymbol{y}}$ and $-1/2$ if $s_{\boldsymbol{x}} \neq s_{\boldsymbol{y}}$), hence we have in this case

$$\cos\left(\frac{2\pi}{q}(s_{\boldsymbol{x}} - s_{\boldsymbol{y}})\right) = \delta_{s_{\boldsymbol{x}},s_{\boldsymbol{y}}} - \frac{1}{2}(1 - \delta_{s_{\boldsymbol{x}},s_{\boldsymbol{y}}}) = -\frac{1}{2} + \frac{3}{2}\delta_{s_{\boldsymbol{x}},s_{\boldsymbol{x}}} \quad (q = 3) \ , \tag{6.2.3}$$

and the $q = 3$ clock model is thus equivalent to the $q = 3$ Potts model with $J \to \frac{3}{2}J$. For $q > 3$ the symmetry group of the clock models is instead smaller than that of the Potts models: for $h = 0$ it is easy to verify that the energy is invariant under the transformation

$$s_{\boldsymbol{x}} \to s'_{\boldsymbol{x}} = (s_{\boldsymbol{x}} + \alpha) \bmod q \tag{6.2.4}$$

where α is a constant integer number. As a consequence, the invariance group of the clock model is (for $q \neq 3$) $\mathbb{Z}_q$, the group of integers modulo q, which for $q > 2$ is a proper subgroup of S_q.

It is not difficult to show that clock models with $q = 4$ and $q = 2$ (the Ising model) are related to each other [62]. Using the complex formulation, the partition function of the $q = 4$ model can be written (considering the case $h = 0$ for the sake of the simplicity) as

$$Z_{q=4}(\beta) = \sum_{\{C_{\boldsymbol{x}}\}} e^{-\beta E} = \sum_{\{C_{\boldsymbol{x}}\}} \prod_{\langle \boldsymbol{x},\boldsymbol{y}\rangle} e^{\beta J \Re(C^*_{\boldsymbol{x}} C_{\boldsymbol{y}})} \ , \tag{6.2.5}$$

moreover, using $C_{\boldsymbol{x}} \in \{\pm 1, \pm i\}$, it is easily seen that the numerical values of the exponential $e^{\beta J \Re(C^*_{\boldsymbol{x}} C_{\boldsymbol{y}})}$, with their degeneracies, can only be

$$e^{\beta J \Re(C^*_{\boldsymbol{x}} C_{\boldsymbol{y}})} = \begin{cases} e^{\beta J} & \deg = 4 \\ e^{-\beta J} & \deg = 4 \\ 1 & \deg = 8 \end{cases} \ . \tag{6.2.6}$$

The square of the partition function of the model with $q = 2$ can instead be written in the form

$$Z_{q=2}(\beta)^2 = \sum_{\{D'_{\boldsymbol{x}}\},\{D_{\boldsymbol{x}}\}} \sum_{\langle \boldsymbol{x},\boldsymbol{y}\rangle} e^{\beta J\left[D_{\boldsymbol{x}} D_{\boldsymbol{y}} + D'_{\boldsymbol{x}} D'_{\boldsymbol{y}}\right]} \ , \tag{6.2.7}$$

where $D_{\boldsymbol{x}}, D'_{\boldsymbol{x}} = \pm 1$, and it is simple to verify that

$$e^{\beta J\left[D_{\boldsymbol{x}} D_{\boldsymbol{y}} + D'_{\boldsymbol{x}} D'_{\boldsymbol{y}}\right]} = \begin{cases} e^{2\beta J} & \deg = 4 \\ e^{-2\beta J} & \deg = 4 \\ 1 & \deg = 8 \end{cases} \ . \tag{6.2.8}$$

Since the values and the degeneracies of the exponentials are the same in the two cases (but for a factor of two in the exponent) and the total number of configuration of $C_{\boldsymbol{x}}$ is equal to the total number of configurations $D_{\boldsymbol{x}}, D'_{\boldsymbol{x}}$, we conclude that

$$Z_{q=4}(\beta) = \Big(Z_{q=2}(\beta/2)\Big)^2 \ . \tag{6.2.9}$$

In the low temperature phase, the $\mathbb{Z}_q$ symmetry that is present when $h = 0$ gets spontaneously broken to $\mathbb{Z}_{q-1}$, and an order parameter for this SSB is identical to the complex order parameter m_2 of the Potts models:

$$m = \frac{1}{L^D} \sum_{\boldsymbol{x}} C_{\boldsymbol{x}} \ . \tag{6.2.10}$$

If $h > 0$ the symmetry is explicitly broken to $\mathbb{Z}_2$ and no phase transition is present; this is the case also if $h < 0$ and q is even, while if $h < 0$ and q is odd the residual $\mathbb{Z}_2$ symmetry can be spontaneously broken in the low temperature phase.

In $D = 2$ the $h = 0$ transition between the low and the high temperature phases is always continuous, however only the cases $q = 2, 4$ (Ising universality class) and $q = 3$ (3-states Potts model) correspond to "standard" continuous phase transitions. For $q \geq 5$ the more exotic case of the Berezinskii-Kosterlitz-Thouless (BKT) transition appears, an infinite order phase transition according to the old Ehrenfest classification, see [63] for an early investigation. In $D = 3$ the transition is discontinuous for $q = 3$, it is continuous in the Ising universality class for $q = 2, 4$, and it is continuous in the O(2) universality class for $q \geq 5$; this is an example of symmetry enlargement at a second order phase transition, which requires a renormalization group framework to be understood (in short: the terms of the continuous effective action which are invariant under $\mathbb{Z}_q$ but not under O(2) are irrelevant if $q \geq 5$).

To simulate the clock models we can use a Metropolis algorithm completely analogous to the one used for the Potts models. However, especially in the low temperature phase, it can be more convenient not to select the trial state with uniform pdf between the q different possibilities, but use instead a trial state that is "close enough" to the previous state, in order for the acceptance probability not to be too small. The minimal possibility is obviously

$$s_{\boldsymbol{x}} \to s'_{\boldsymbol{x}} = \big(s_{\boldsymbol{x}} \pm 1\big) \bmod q \ , \tag{6.2.11}$$

where the two signs are selected with equal probability. Note however that also in this case the acceptance probability will become negligibly small if β is large enough. As for all the models whose variables only assume a finite number of values, it is also straightforward to use an heat-bath update, computing (and storing in an array) at the beginning of the simulation all the required conditional probabilities.

6.3 O(N) models and microcanonical updates

In O(N) models, to each site $\boldsymbol{r}$ of the lattice we associate a real, unit-length, vector in the N-dimensional real space ($\boldsymbol{s}_{\boldsymbol{r}} \in \mathbb{R}^N$, $|\boldsymbol{s}_{\boldsymbol{r}}| = 1$), and each configuration has energy

$$E = -J \sum_{\langle \boldsymbol{x}, \boldsymbol{y} \rangle} \boldsymbol{s}_{\boldsymbol{x}} \cdot \boldsymbol{s}_{\boldsymbol{y}} - \sum_{\boldsymbol{x}} \boldsymbol{h} \cdot \boldsymbol{s}_{\boldsymbol{x}} \ . \tag{6.3.1}$$

For $\boldsymbol{h} = 0$ the model is invariant (assuming as usual periodic b. c.) under the transformation $\boldsymbol{s}_{\boldsymbol{x}} \to \boldsymbol{s}'_{\boldsymbol{x}} = M \boldsymbol{s}_{\boldsymbol{x}}$, where M is an orthogonal $N \times N$ matrix, hence the name of the model. The magnetic field is a vector in $\mathbb{R}^N$, however, since the first term of the energy in Eq. (6.3.1) is invariant under O(N) transformations, we can safely assume $\boldsymbol{h}$ to be directed along the 1-axis. The $N = 1$ case of the O(N) models is just the Ising model, the $N = 2$ case is often called XY model, while the $N = 3$ case is often called Heisenberg model.

Algorithm 11 Algorithm to sample with uniform pdf the unit sphere in $\mathbb{R}^N$.

Require: r_i (with $i = 1, \dots, N$) sampled from uniform pdf in $[0, 1)$
  repeat
    $x_i = 1 - 2r_i$
    $S = \sum_i x_i^2$
  until $0 < S < 1$
  $y_i = \frac{x_i}{\sqrt{S}}$

Algorithm 12 Algorithm to sample with uniform pdf the unit sphere in $\mathbb{R}^N$.

Require: g_i (with $i = 1, \dots, N$) sampled from normal Gaussian distribution
  repeat
    $x_i = g_i$
    $S = \sum_i x_i^2$
  until $0 < S$
  $y_i = \frac{x_i}{\sqrt{S}}$

In $D > 2$ the O(N) symmetry is broken to O($N - 1$) in the low temperature phase, and an order parameter is naturally

$$\boldsymbol{m} = \frac{1}{L^D} \sum_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{x}} \ . \tag{6.3.2}$$

As for the Ising model it is possible to introduce a proxy of the spontaneous magnetization by using $|\boldsymbol{m}|$, where now $|\boldsymbol{m}|^2 = \boldsymbol{m} \cdot \boldsymbol{m}$, hence we can define also in this case two different susceptibilities (neglecting irrelevant β factors): $\chi = L^D \langle |\boldsymbol{m}| \rangle$ and $\chi' = L^D (\langle |\boldsymbol{m}|^2 \rangle - \langle |\boldsymbol{m}| \rangle^2)$. Some critical properties of the three dimensional O(N) models with $N = 2$ and $N = 3$ are summarized in Sec. 7.A. The $D = 2$ case is more involved, since in two dimensional models continuous symmetries can not be spontaneously broken; this is the content of the so called Coleman-Mermin-Wagner theorem, see, e. g. [45] §4.A or [42] §6.1 for an heuristic argument. For $N = 2$ a finite temperature transition exists in $D = 2$, of the Berezinskii-Kosterlitz-Thouless type, see, e. g.,[42] §6.2-6.4 for an elementary presentation or [45] §4.2 for some more details. When $N > 2$ there is no finite temperature transition in $D = 2$, but a transition exists at $T = 0$, in which a correlation length diverging exponentially in $1/T$ appears, see [42] §6.5 for an elementary presentation and [46] §14-15 for more details.

To simulate the O(N) models using single site updates it is convenient to introduce the sum

$$\boldsymbol{S}_{\boldsymbol{r}} = \sum_{\langle \boldsymbol{x}, \boldsymbol{r} \rangle} \boldsymbol{s}_{\boldsymbol{x}} \ , \tag{6.3.3}$$

where $\boldsymbol{r}$ is the site to be updated. Using this vector we can write the energy of a configuration (for $h = 0$) in the form

$$E = -J \boldsymbol{s}_{\boldsymbol{r}} \cdot \boldsymbol{S}_{\boldsymbol{r}} + (\text{independent of } \boldsymbol{s}_{\boldsymbol{r}}) \ , \tag{6.3.4}$$

which is completely analogous to the form used for the Ising model, see Eq. (5.3.3). The energy difference required to accept/reject the update $\boldsymbol{s}_{\boldsymbol{r}} \to \boldsymbol{s}'_{\boldsymbol{r}}$ using a local Metropolis algorithm is thus simply

$$E' - E = -J(\boldsymbol{s}'_{\boldsymbol{r}} - \boldsymbol{s}_{\boldsymbol{r}}) \cdot \boldsymbol{S}_{\boldsymbol{r}} \ . \tag{6.3.5}$$

To generate the trial state $\boldsymbol{s}'_{\boldsymbol{r}}$ to be used in the local Metropolis update we have several possibilities. The simplest possibility is to generate $\boldsymbol{s}'_{\boldsymbol{r}}$ with uniform pdf in S^N (the unit sphere in $\mathbb{R}^N$). In this case we can for example use the algorithm in Alg. (11), which start from a uniform sampling of the hypercube, or the variant in Alg. (12), which is more efficient if N is large (since when $N \gg 1$ the volume of the sphere is much smaller than 2^N). A better possibility, especially

for large N or in the low temperature phase, is to generate $\boldsymbol{s}'_{\boldsymbol{x}}$ by slightly changing $\boldsymbol{s}_{\boldsymbol{x}}$, for example by selecting with uniform pdf two indices $i, j \in \{1, \dots, N\}$ (with $i \neq j$, hence $N \geq 2$) and using

$$\begin{aligned} (\boldsymbol{s}'_{\boldsymbol{r}})_k &= (\boldsymbol{s}_{\boldsymbol{r}})_k \quad \text{if } k \neq i, j \\ (\boldsymbol{s}'_{\boldsymbol{r}})_i &= \quad (\boldsymbol{s}_{\boldsymbol{r}})_i \cos\theta + (\boldsymbol{s}_{\boldsymbol{r}})_j \sin\theta \ , \\ (\boldsymbol{s}'_{\boldsymbol{r}})_j &= -(\boldsymbol{s}_{\boldsymbol{r}})_i \sin\theta + (\boldsymbol{s}_{\boldsymbol{r}})_j \cos\theta \end{aligned} \tag{6.3.6}$$

where θ is generated with uniform pdf in the range $[-\alpha, \alpha]$, and α is fixed at the beginning of the simulation in order to have a reasonable acceptance probability. Note that for a rotation and its inverse to be equiprobable (requirement needed for the use of the Metropolis algorithm) the selection of i, j has to be performed with uniform pdf, and the range of the angle θ has to be symmetric with respect to the origin.

A complete algorithm to simulate the O(N) model is thus the following

1. select the lattice site $\boldsymbol{r}$ to be updated (randomly, with uniform pdf, or by a deterministic sweep of the lattice)
2. define the trial configuration as the configuration in which only the value $\boldsymbol{s}_{\boldsymbol{r}}$ is changed; the trial site variable $\boldsymbol{s}'_{\boldsymbol{r}}$ can be generated by using Eq. (6.3.6) or Alg. (11), Alg. (12).
3. accept the trial configuration with probability $\min(1, e^{-\beta(E'-E)})$, where E is the energy of the initial configuration and E' is the energy of the trial configuration. If the trial configuration is not accepted, keep the old one.

This algorithm satisfies the detailed balance principle if in point 1. we use the random selection, while only the balance equation is satisfied if a deterministic sweep is adopted. Since the evaluation of $E' - E$ is straightforward once $\boldsymbol{S}_{\boldsymbol{r}}$ has been computed, it can be convenient to repeat steps 2. and 3. several times (depending on the values of N and α) before updating a different site.

Note that the transformation in Eq. (6.3.6) leaves invariant the constraint $\boldsymbol{s}_{\boldsymbol{r}} \cdot \boldsymbol{s}_{\boldsymbol{r}} = 1$ in exact algebra, however this is no more the case on a real CPU, where rounding errors are present (see, e. g., [64] §2.4 for more details on the floating point arithmetic). The violation of the unit-length constraint is almost certainly negligible after a single update, but the accumulation of rounding errors can introduce a bias in the simulation. For this reason it is necessary, after a fixed number of updates, to project back the variables on S^N, using for example $\boldsymbol{s}_{\boldsymbol{r}} \leftarrow \boldsymbol{s}_{\boldsymbol{r}}/|\boldsymbol{s}_{\boldsymbol{r}}|$.

Instead of using a local Metropolis update it is also possible to use a local heat-bath update. To implement such an update we need to sample (for $h = 0$) the conditional probability distribution

$$P(\boldsymbol{s}_{\boldsymbol{r}}) \propto \delta(\boldsymbol{s}_{\boldsymbol{r}} \cdot \boldsymbol{s}_{\boldsymbol{r}} - 1) \exp\left(\beta J \boldsymbol{s}_{\boldsymbol{r}} \cdot \boldsymbol{S}_{\boldsymbol{r}}\right) \ . \tag{6.3.7}$$

For this purpose it is convenient to decompose $\boldsymbol{s}_{\boldsymbol{r}}$ as a sum of a longitudinal component $s_\parallel$ and a transverse component $\boldsymbol{s}_\perp$, with respect to $\boldsymbol{S}_{\boldsymbol{r}}$, i. e.

$$\boldsymbol{s}_{\boldsymbol{r}} = s_\parallel \frac{\boldsymbol{S}_{\boldsymbol{r}}}{|\boldsymbol{S}_{\boldsymbol{r}}|} + \boldsymbol{s}_\perp \ , \quad \text{with} \quad \boldsymbol{s}_\perp \cdot \boldsymbol{S}_{\boldsymbol{r}} = 0 \ . \tag{6.3.8}$$

To generate $\boldsymbol{s}_{\boldsymbol{r}}$ using this decomposition we can first generate $s_\parallel$ with distribution

$$\tilde{P}(s_\parallel) = \int P(\boldsymbol{s}_{\boldsymbol{r}}) \mathrm{d}\boldsymbol{s}_\perp \ , \tag{6.3.9}$$

and then generate the remaining components using a uniform pdf on the $(N-1)$-sphere of radius $\sqrt{1 - s_\parallel^2}$ (some trigonometry is obviously needed to give these components the proper orientation in the N-dimensional space). The probability $\tilde{P}(s_\parallel)$ is given by (using a coordinate system in

which $\hat{x}_N$ is directed along $\boldsymbol{S_r}$)

$$\begin{aligned}\tilde{P}(s_\parallel) &\propto \exp\left(\beta J s_\parallel |\boldsymbol{S_r}|\right) \int \mathrm{d}x_1 \cdots \int \mathrm{d}x_{N-1} \delta(x_1^2 + \cdots + x_{N-1}^2 + s_\parallel^2 - 1) \propto \\ &\propto \exp\left(\beta J s_\parallel |\boldsymbol{S_r}|\right) \int_0^\infty r^{N-2} \delta(r^2 + s_\parallel^2 - 1) \mathrm{d}r \propto \\ &\propto \exp\left(\beta J s_\parallel |\boldsymbol{S_r}|\right) \int_0^\infty y^{\frac{N-3}{2}} \delta(y + s_\parallel^2 - 1) \mathrm{d}y = \exp\left(\beta J s_\parallel |\boldsymbol{S_r}|\right) (1 - s_\parallel^2)^{\frac{N-3}{2}} \ ,\end{aligned} \tag{6.3.10}$$

where in the third step we used the change of variable $y = r^2$. The case $N = 3$ is thus particularly simple, since it is enough to sample an exponential distribution with $s_\parallel \in [-1, 1]$, while for $N \neq 3$ an accept/reject von Neumann algorithm is generally required to sample the distribution $\tilde{P}(s_\parallel)$. Note, however, that the local heat-bath update is not particularly more efficient than the local Metropolis update, especially when used together with the microcanonical update that we are now going to introduce.

The basic idea of the micorcanonical method (sometimes also called overrelaxation) is to generate, using a deterministic procedure, a trial state $\boldsymbol{s}'_{\boldsymbol{r}}$ which is surely accepted by the Metropolis test, and it is as far as possible from the previous state $\boldsymbol{s_r}$. If we remind that the energy of a configuration is given by

$$E = -J \boldsymbol{s_r} \cdot \boldsymbol{S_r} + (\text{independent of } \boldsymbol{s_r}) \ , \tag{6.3.11}$$

it is simple to generate a trial state $\boldsymbol{s}'_{\boldsymbol{r}}$ such that $E' \leq E$: everything which does not increase the relative angle between $\boldsymbol{s_r}$ and $\boldsymbol{S_r}$ will do the job (we are obviously assuming $N \geq 2$, so that a continuous angle can be used). If we want $\boldsymbol{s}'_{\boldsymbol{r}}$ to be as far as possible from $\boldsymbol{s_r}$ the simplest possibility is to change sign to the transverse (with respect to $\boldsymbol{S_r}$) components of $\boldsymbol{s_r}$: while in the heat-bath update the most probable outcome is directed along $\boldsymbol{S_r}$, in the microcanonical update we overshoot the minimum of the energy, generating a mirror image with respect to $\boldsymbol{S_r}$ of the original state[3]. In formulae

$$\boldsymbol{s}'_{\boldsymbol{r}} = \frac{(\boldsymbol{s_r} \cdot \boldsymbol{S_r})\boldsymbol{S_r}}{|\boldsymbol{S_r}|^2} - \left(\boldsymbol{s_r} - \frac{(\boldsymbol{s_r} \cdot \boldsymbol{S_r})\boldsymbol{S_r}}{|\boldsymbol{S_r}|^2}\right) = 2\frac{(\boldsymbol{s_r} \cdot \boldsymbol{S_r})\boldsymbol{S_r}}{|\boldsymbol{S_r}|^2} - \boldsymbol{s_r} \ . \tag{6.3.12}$$

To avoid numerical errors it is convenient, when implementing this expression, to check that the value of $|\boldsymbol{S_r}|$ is not too small, e. g. it has be larger than 10^{-13}. If this is not the case the update is aborted or a random vector is used for $\boldsymbol{s}'_{\boldsymbol{r}}$.

It is simple to show that by applying twice the transformation Eq. (6.3.12) we come back to the original configuration, i. e. $\boldsymbol{s_r} \to \boldsymbol{s}'_{\boldsymbol{r}} \to \boldsymbol{s_r}$. This property ensures the symmetry of the selection matrix, and thus the possibility of using a Metropolis test; moreover, since by construction we have $\boldsymbol{s_r} \cdot \boldsymbol{S_r} = \boldsymbol{s}'_{\boldsymbol{r}} \cdot \boldsymbol{S_r}$, the Metropolis test is always passed and the new configuration always accepted. Note however that the simple condition of being energy preserving is not a sufficient condition for the update algorithm to be accepted with unit probability: if the symmetry condition is not satisfied the Metropolis-Hastings algorithm has to be applied, and the acceptance probability does not depend just on $E' - E$.

It is important to note that the microcanonical update does not generate an irreducible Markov chain: since this update conserves the energy it can not connect two configurations with different energies. Nevertheless the single site microcanonical update satisfies the detailed balance principle, so it can be used together with local Metropolis or heat-bath updates to reduce the autocorrelation time, and thus speed-up the simulation. Following the discussion of Sec. 3.3.3, a simple possibility is to use a stochastic mixture of the different updates: after the extraction of a random number $r \in [0, 1)$ with uniform pdf, a complete update of the lattice using the local Metropolis or heat-bath algorithm is performed if $r < \epsilon$, while a complete update of the lattice using the single site microcanonical update is performed if $r \geq \epsilon$. The parameter ϵ has to be fixed at the beginning

[3]The name overrelaxation originates from the similarity of this idea with the one used in the iterative solution method for large linear systems known as Successive Overrelaxation Method (SOR), see, e. g. [65] §3.3.

of the simulation, and it is usually convenient to have $\epsilon \lesssim 0.2$, i. e. to perform significantly more microcanonical updates than Metropolis or heath-bath updates. Microcanoncial updates, being of deterministic nature, are indeed associated with dynamical critical exponents close to one, $z \approx 1$, while Metropolis and heat-bath updates, which behaves like diffusion processes, are associated with dynamical critical exponents $z \approx 2$.

6.4 The cluster update for the Ising model

The cluster update (more precisely the "single" cluster update) that we are now going to introduce is a nonlocal update scheme, in which a possibly large fraction of the sites change its value. For the sake of the simplicity we discuss its application to the Ising model, but with minor modifications it can be applied also to O(N) models, see [66].

The fundamental ingredient of the algorithm is the recipe to build the cluster. Roughly speaking, the idea is to build a cluster of sites all having the same orientation (the same value of the site variable), adding neighboring sites to the cluster with a probability P_{add} that is a parameter of the algorithm. More precisely, at step 1 we initialize the cluster by using a randomly chosen site $\boldsymbol{r}$ of the lattice, and we denote by C_n the set of sites added to the cluster at step n (hence $C_1 = \{\boldsymbol{r}\}$). At step $n+1$, for each site $\boldsymbol{x}$ in C_n we add to the cluster, with probability P_{add}, the nearest neighbors of $\boldsymbol{x}$ having the same value of the site variable and that are not already in the cluster. This process ends when an iteration $\bar{n}$ is reached at which no new sites are added to the cluster (i. e. $C_{\bar{n}}$ is the empty set). It can happen that, at the step $n+1$, a site $\boldsymbol{y}$ exists which has the same orientation of the site $\boldsymbol{r}$ and it is nearest neighbor of several sites in C_n; in this case the site $\boldsymbol{y}$ has several chances of being added to the lattice, coming from different sites in C_n, and it important to try them all.

It is important to stress that the way in which we described the algorithm to build the cluster is not the most general one: what is really fundamental is that, once a site is added to the cluster, all its nearest neighbor sites are (sooner or later) tested for joining the cluster, and that all the possibilities of adding a site (e. g. coming from different neighboring sites) are examined. These properties are automatically satisfied by the process described above, but this is by no means the only way of satisfying them, the order in which the different sites are added to the cluster being irrelevant.

A simple way to check if a given site has been already added to the cluster is to use an auxiliary lattice; we can for example use a lattice with all the lattice variables initialized to zero, and set the value at site $\boldsymbol{x}$ to one when the site $\boldsymbol{x}$ is added to the cluster. In this way the sentence "the site $\boldsymbol{x}$ is not in the cluster" translates as "the variable at site $\boldsymbol{x}$ of the auxiliary lattice is zero". Using this trick it is simple to build the cluster with a recursive algorithm like that shown in Alg. (13). This algorithm can be easily implemented in low level programming languages like C, Fortran or C++, which allow for recursive functions to be defined, however such a recursive algorithm can lead to a stack overflow for large lattices and low temperatures. In these cases the cluster to be built is

Algorithm 13 Function to recursively build the cluster starting from its first site

```
function BUILD(site x)
    for all y nearest neighbor of x do
        if s_y = s_x and y is not already in the cluster then
            if random number in [0,1) with uniform pdf < P_add then
                add y to the cluster
                call BUILD(y)
            end if
        end if
    end for
end function
```

Algorithm 14 Algorithm to build the cluster starting from its first site without using recursion

Require: array `cluster` to store the cluster sites
  $\boldsymbol{r}$ fist site of the cluster, `cluster`[0]=$\boldsymbol{r}$
  $n_{old} = 0$, $n_{new} = \ell_c = 1$
  while $n_{new} > n_{old}$ **do**
    for p values in $n_{old} \leq p < n_{new}$ **do**
      for all $\boldsymbol{x}$ nearest neighbor of `cluster`[p] **do**
        if $\boldsymbol{x}$ is not in the cluster and $s_{\boldsymbol{x}} = s_{\boldsymbol{r}}$ **then**
          if random number in $[0, 1)$ with uniform pdf $< P_{add}$ **then**
            add $\boldsymbol{x}$ to the cluster
            `cluster`[ℓ_c] = $\boldsymbol{x}$
            $\ell_c \leftarrow \ell_c + 1$
          end if
        end if
      end for
    end for
    $n_{old} = n_{new}$
    $n_{new} = \ell_c$
  end while

very large, and its size could exceed that of the buffer used to store the recursion instructions in the CPU, resulting in a segmentation fault at execution time.

A non recursive algorithm to build the cluster is shown in Alg. (14). Such an algorithm is clearly less straightforward than the recursive one, but its logic is nevertheless quite simple, and implements the building strategy described above. The auxiliary structures needed are the array `cluster`, used to store the cluster sites, the cluster length ℓ_c, and the two integers n_{old} and n_{new}, which are used to keep track of the sites recently added to the cluster: n_{old} is the cluster size at the beginning of the previous step, while n_{new} is the cluster size at the beginning of present step; $n_{new} - n_{step}$ is thus the number of sites that have been added to the cluster in the previous step. In the first step the cluster is composed of only the site $\boldsymbol{r}$, $n_{new} = \ell_c = 1$ and $n_{old} = 0$ (this is obviously conventional, but it is useful to enter the "while" loop in its first occurrence). The terminating condition $n_{new} \leq n_{old}$ just means that in the previous step no sites have been added to the cluster. If this is not the case we sweep the newly added sites, which are the one corresponding to `cluster`[p] for $n_{old} \leq p < n_{new}$, and for each of these sites we test their nearest neighbors. If any of these nearest neighbors is added to the cluster, its position is appended to the `cluster` array, and we update the cluster size ℓ_c. After all nearest neighbors of the sites previously added to the cluster have been tested, the values of n_{old} and n_{new} are updated and the process starts again, until the termination condition $n_{old} = n_{new}$ is reached.

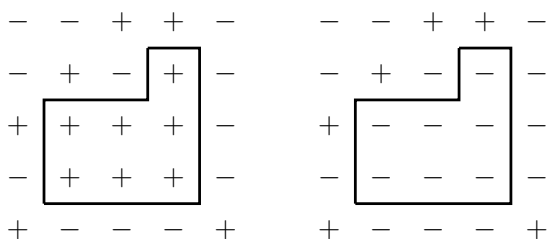

Figure 6.3: Example of cluster update: in the left panel the initial configuration is represented, and the region enclosed by the solid line is the cluster (obtained by using the algorithm described in the main text). In the right panel we show the trial configuration, obtained by flipping all the spins of the cluster.

Once a cluster has been identified, the trial configuration is generated by flipping all the spins of the cluster, as shown in Fig. (6.3) for the two dimensional case, and we now have to discuss the probability for this trial configuration to be accepted. Let us start by noting that the simple Metropolis algorithm is *not* applicable in the present case: the probability of selecting a cluster in the original configuration is not the same as the probability of selecting the same cluster in the trial configuration.

To verify this fact, let us denote by n_e the number of bonds in the original configuration connecting cluster sites to sites outside the cluster having the same orientation ($n_e = 3$ in Fig. (6.3)), and by n_d the number of bonds in the original configuration connecting cluster sites to sites outside the cluster having different orientation ($n_d = 9$ in Fig. (6.3)). Using the same notation of Sec. 3.3.1, let us denote by A_{ba} the probability of selecting a given cluster in the configuration a: we have

$$A_{ba} = P_{in}(1 - P_{add})^{n_e} \ , \tag{6.4.1}$$

where P_{in} is the total probability coming from the sites added to the cluster, and $(1 - P_{add})^{n_e}$ is needed since n_s attempt of adding some more sites to the cluster have been rejected. The specific form of P_{in} depends, e. g. on which is the fist site that has been added to the cluster, but we will not need to know it. After the cluster has been flipped, the probability of selecting again the same cluster (starting from the same initial site) is

$$A_{ab} = P_{in}(1 - P_{add})^{n_d} \ , \tag{6.4.2}$$

since now the sites that we have to reject to get the same cluster boundaries are the ones connected by n_d bonds, which in the original configuration connected sites with different orientations (that could not be added to the cluster) while now they connect sites with the same orientations and could be added to the cluster.

We thus generically have $A_{ba} \neq A_{ba}$ and the Metropolis-Hastings algorithm must be used, which has acceptance probability

$$\min\left(1, \frac{A_{ab}\pi_b}{A_{ba}\pi_a}\right) \ , \tag{6.4.3}$$

where π_a and π_b are Gibbs weights. In the specific case of the single cluster update we have

$$\frac{A_{ab}\pi_b}{A_{ba}\pi_a} = \frac{(1 - P_{add})^{n_d}\pi_b}{(1 - P_{add})^{n_e}\pi_a} \ . \tag{6.4.4}$$

The energy of the configuration b (the one in which the cluster has been flipped) differs from the energy of the configuration a only because of the different orientations of the spins at the boundary of the cluster (this should remind of the Peierls argument, see, e. g., [37] §14.3), hence

$$\frac{\pi_b}{\pi_a} = \frac{e^{\beta J(n_d - n_e)}}{e^{\beta J(n_e - n_d)}} = e^{2\beta J(n_d - n_e)} \ , \tag{6.4.5}$$

and finally

$$\frac{A_{ab}\pi_b}{A_{ba}\pi_a} = (1 - P_{add})^{n_d - n_e} e^{2\beta J(n_d - n_e)} = \left((1 - P_{add})e^{2\beta J}\right)^{n_d - n_e} \ . \tag{6.4.6}$$

We thus see that if we chose P_{add} in such a way that $(1 - P_{add})e^{2\beta J} = 1$, i. e.

$$P_{add} = 1 - e^{-2\beta J} \ , \tag{6.4.7}$$

then the cluster flip is always accepted, and this is the value that (obviously) is used in simulations.

The Markov chain built as previously described satisfies by construction the detailed balance with respect to the Gibbs weight, and before using the single cluster update in a simulation we just need to show that this Markov chain is irreducible and aperiodic. Irreducibility is quite simple to show: since we have a nonvanishing probability of always creating clusters composed of a single site, and the starting seed of the cluster is selected with uniform pdf on the lattice, the proof of the

irreducibility is the same as for the local Metropolis update. Aperiodicity requires just a little more care, as cluster flips are accepted with probability 1. Since all the states of an irreducible Markov chain have the same period, it is enough to show that a specific configuration has period equal to one. Let us consider the completely polarized configuration and study its recurrence times. One possible recurrence is the following

1. site $\boldsymbol{r}$ is selected to start the cluster,
2. the cluster composed of the only site $\boldsymbol{r}$ is selected,
3. site $\boldsymbol{r}$ is flipped,
4. site $\boldsymbol{r}$ is selected as cluster,
5. site $\boldsymbol{r}$ is flipped.

Note that in step 4. (unlike in step 2.) the cluster is surely coincident with the site $\boldsymbol{r}$, since the starting configuration was completely polarized, and after the spin flip (point 3.) the site $\boldsymbol{r}$ has orientation different from that of all the other sites of the lattice. We have thus seen that 2 is a recurrence time for the completely polarized configuration. A possible path of length 3 that starts from (and arrives to) the completely polarized configuration is

1. site $\boldsymbol{r}$ is selected to start the cluster,
2. the cluster composed of the only site $\boldsymbol{r}$ is selected,
3. site $\boldsymbol{r}$ is flipped,
4. site $\boldsymbol{x}$ (which is a nearerst neighbor of $\boldsymbol{r}$) is selected to start the cluster,
5. the cluster composed of the only site $\boldsymbol{x}$ is selected,
6. site $\boldsymbol{x}$ is flipped,
7. site $\boldsymbol{r}$ is selected to start the cluster,
8. the cluster composed sites $\boldsymbol{r}$ and $\boldsymbol{x}$ is selected,
9. sites $\boldsymbol{r}$ and $\boldsymbol{x}$ are flipped.

Since we have shown that 2 and 3 are possible recurrence times for the totally polarized configuration, and $\mathrm{GCD}\{2,3\} = 1$, the Markov chain is aperiodic.

We close this section by signaling that when using the cluster update we not only have very small autocorrelation times close to second order phase transitions, but we also have the possibility of using "improved" estimators for some observables, like the magnetic susceptibility. Improved estimators (see also Sec. 18.5) are quantities which have the same statistical averages of the standard observables but smaller variances, hence smaller statistical errors, see [67].

Chapter 7

Appendices to Part II

7.A Critical properties of some commonly used models

Ising model in $D = 2$

$$\beta_c = \frac{1}{2}\log(1+\sqrt{2}) \simeq 0.44068679350977151262\ ; \quad \nu = 1\ ; \quad \gamma = 7/4\ ; \quad \beta = 1/8$$

See, e. g. [45] §2 for a quick presentation or [39] for many more details. The Binder cumulant is defined by $U = \frac{\langle m^4\rangle}{\langle m^2\rangle^2}$, and its critical values is (with periodic boundary conditions) $U_4^* = 1.1679227(4)$, see [52].

Ising model in $D = 3$

	β_c
[68]	0.221654626(5)

	ν
[49]	0.63002(10)
[68]	0.629912(86)
[69]	0.6299710(40)
[70]	0.62997097(12)

	γ
[49]	1.23719(21)
[68]	1.23708(33)
[69]	1.2370752(79)
[70]	1.23707549(27)

$U_4 = \frac{\langle m^4\rangle}{\langle m^2\rangle^2}$

	U_4^* (pbc)
[49]	1.6036(1)
[68]	1.60356(15)

XY/O(2) model in $D = 3$

	β_c
[71]	0.454169(4)
[72]	0.4541652(5)(6)
[73]	0.45416474(10)(7)
[74]	0.45416476(10)

	ν
[72]	0.6717(1)
[73]	0.67169(7)
[75]	0.67175(10)

	γ
[72]	1.3178(2)
[73]	1.31778(15)
[75]	1.31786(20)

$U_4 = \frac{\langle (\boldsymbol{m}^2)^2\rangle}{\langle \boldsymbol{m}^2\rangle^2}$

	U_4^* (pbc)
[72]	1.2431(1)(1)
[73]	1.24296(8)

Heisenberg/O(3) model in $D = 3$

	β_c
[76]	0.6930(1)
[71]	0.693001(10)

	ν
[77]	0.7112(5)
[78]	0.7116(10)
[79]	0.71164(10)
[80]	0.71168(41)

	γ
[77]	1.3960(9)
[78]	1.3963(20)
[79]	1.39635(20)
[80]	1.39641(81)

$$U_4 = \frac{\langle (\boldsymbol{m}^2)^2 \rangle}{\langle \boldsymbol{m}^2 \rangle^2}$$

	U_4^* (pbc)
[77]	1.1394(1)(2)
[78]	1.1394(3)

7.B Benchmark for the two dimensional Ising model

Notation:

$$e = \frac{E}{L^2} = -\frac{1}{L^2} \sum_{\langle \boldsymbol{x}, \boldsymbol{y} \rangle} s_{\boldsymbol{x}} s_{\boldsymbol{y}} \ , \quad m = \frac{1}{L^2} \sum_{\boldsymbol{x}} s_{\boldsymbol{x}} \ , \quad Z = \sum_{\{s\}} e^{-\beta E} \tag{7.B.1}$$

β	$\langle e \rangle$	$\langle e^2 \rangle - \langle e \rangle^2$	$\langle \lvert m \rvert \rangle$	$\langle m^2 \rangle - \langle \lvert m \rvert \rangle^2$	$\langle m^2 \rangle$	$\langle m^4 \rangle / \langle m^2 \rangle^2$
0.000000	0.000000	0.125000	0.196381	0.023935	0.062500	2.875000
0.040000	-0.080345	0.126629	0.214653	0.027778	0.073854	2.828513
0.080000	-0.162853	0.131884	0.236667	0.032676	0.088687	2.764772
0.120000	-0.250180	0.141970	0.263841	0.039009	0.108621	2.675390
0.160000	-0.346083	0.159164	0.298240	0.047234	0.136181	2.549494
0.200000	-0.456135	0.186791	0.342766	0.057711	0.175199	2.376421
0.240000	-0.588063	0.227799	0.401043	0.070127	0.230963	2.153343
0.280000	-0.750199	0.279986	0.476252	0.082324	0.309140	1.895476
0.320000	-0.945770	0.328503	0.568067	0.089330	0.412030	1.637480
0.360000	-1.164138	0.346650	0.668913	0.085439	0.532884	1.417313
0.400000	-1.379116	0.317389	0.764712	0.070041	0.654826	1.255417
0.440000	-1.562847	0.253310	0.842716	0.049605	0.759774	1.149801
0.480000	-1.702039	0.182647	0.898543	0.031509	0.838888	1.086344
0.520000	-1.799371	0.124279	0.935246	0.018805	0.893490	1.049937
0.560000	-1.864728	0.082608	0.958399	0.010970	0.929498	1.029392
0.600000	-1.908070	0.054784	0.972867	0.006427	0.952898	1.017739
0.640000	-1.936904	0.036620	0.981994	0.003841	0.968152	1.011005
0.680000	-1.956281	0.024762	0.987853	0.002357	0.978211	1.007009
0.720000	-1.969454	0.016939	0.991690	0.001486	0.984936	1.004570
0.760000	-1.978511	0.011708	0.994250	0.000961	0.989494	1.003040
0.800000	-1.984798	0.008163	0.995984	0.000636	0.992620	1.002055

Table 7.1: Values computed by enumeration on the lattice 4^2 with periodic boundary conditions

Part III

The study of path-integrals in quantum mechanics

Chapter 8

*Quantum statistical mechanics and path-integrals

Chapter 9

*MCMC in quantum mechanics: thermodynamics

Chapter 10

*MCMC in quantum mechanics: spectrum

Chapter 11

*Path-integrals with nontrivial topology

Chapter 12

*Identical particles

Part IV

The study of path-integrals in quantum field theories

Chapter 13

Statistical quantum field theory and path-integrals

In this and in the following chapters we will mainly use the free scalar field as an example to develop the theory, but almost nothing would change by adding also an interaction term in most of the cases.

13.1 Path-integral formulation of the free scalar field

The Lagrangian of a free scalar field in D space-time dimensions is (in Minkowski metric and with natural units $\hbar = c = 1$)

$$L = \int \mathrm{d}^{D-1}x\, \mathcal{L}\ , \quad \mathcal{L} = \frac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - \frac{1}{2}m^2\phi^2\ , \tag{13.1.1}$$

and the equation of motion is easily shown to be $\partial_\mu\partial^\mu\phi + m^2\phi = 0$. The conjugate momentum of the variable $\phi(t,\boldsymbol{x})$ is

$$\pi_\phi(t,\boldsymbol{x}) = \frac{\delta L}{\delta\partial_t\phi(t,\boldsymbol{x})} = \partial_t\phi(t,\boldsymbol{x})\ , \tag{13.1.2}$$

and the Hamiltonian is thus

$$H = \int \mathrm{d}^{D-1}x\, \mathcal{H}\ , \quad \mathcal{H} = \frac{1}{2}\pi_\phi^2 + \frac{1}{2}(\boldsymbol{\nabla}\phi)^2 + \frac{1}{2}m^2\phi^2\ . \tag{13.1.3}$$

In order to refer back to the QM case discussed in Sec. 8 let us introduce a lattice spacing a for the $D-1$ spatial directions, in such a way that (we always assume the volume to be finite, for concreteness)

$$L \simeq \frac{1}{2}\sum_{\boldsymbol{x}} a^{D-1}(\partial_t\phi_{\boldsymbol{x}})^2 - \frac{1}{2}\sum_{\boldsymbol{x}}\sum_{\mu>0} a^{D-1}\left(\frac{\phi_{\boldsymbol{x}+a\hat{\mu}}-\phi_{\boldsymbol{x}}}{a}\right)^2 - \frac{1}{2}\sum_{\boldsymbol{x}} a^{D-1}m^2\phi_{\boldsymbol{x}}^2\ , \tag{13.1.4}$$

where $\phi_{\boldsymbol{x}}$ stands for $\phi(t,\boldsymbol{x})$ (time is implied), $\hat{\mu}$ is the versor of the μ-th direction (with time corresponding to $\mu = 0$), and we used the discretization $\partial_\mu\phi(t,\boldsymbol{x}) \simeq \big(\phi_{\boldsymbol{x}+a\hat{\mu}} - \phi_{\boldsymbol{x}}\big)/a$ for $\mu > 0$. This is now the Lagrangian of a system with a finite number of degrees of freedom, the conjugate momentum of $\phi_{\boldsymbol{x}}$ is

$$p_{\boldsymbol{x}} = \frac{\partial L}{\partial\partial_t\phi_{\boldsymbol{x}}} = a^{D-1}\partial_t\phi_{\boldsymbol{x}}\ , \tag{13.1.5}$$

which corresponds to $p_{\boldsymbol{x}} = a^{D-1}\pi_\phi(\boldsymbol{x})$, and the discretized Hamiltonian is

$$H = \frac{1}{2}\sum_{\boldsymbol{x}}\frac{p_{\boldsymbol{x}}^2}{a^{D-1}} + \frac{1}{2}\sum_{\boldsymbol{x}}\sum_{\mu>0} a^{D-1}\left(\frac{\phi_{\boldsymbol{x}+a\hat{\mu}}-\phi_{\boldsymbol{x}}}{a}\right)^2 + \frac{1}{2}\sum_{\boldsymbol{x}} a^{D-1}m^2\phi_{\boldsymbol{x}}^2\ . \tag{13.1.6}$$

To write a path-integral formulation of the partition function of the system we introduce the states $|\phi_{\boldsymbol{x}}\rangle$, eigenstates of the discrete field operator $\phi_{\boldsymbol{x}}(t=0)$ (note that $[\phi_{\boldsymbol{x}}, \phi_{\boldsymbol{y}}] = 0$, hence the operators corresponding to all the positions can be diagonalized simultaneously). The starting point is obviously

$$Z(\beta) = \mathrm{Tr}\, e^{-\beta H} = \sum \left\{ \prod_{\boldsymbol{x}} \langle \phi_{\boldsymbol{x}}| \right\} e^{-\beta H} \left\{ \prod_{\boldsymbol{x}} |\phi_{\boldsymbol{x}}\rangle \right\} \ , \tag{13.1.7}$$

where the sum extends on all the states $\prod_{\boldsymbol{x}} |\phi_{\boldsymbol{x}}\rangle$ of the system. As in the QM case we write $e^{-\beta H} = (e^{-\beta H/N})^N$, with N a large natural number, and we assume for the sake of the simplicity that $a = \beta/N$, in such a way that the temporal and the spatial lattice spacings are equal. We then insert $N-1$ resolutions of the identity

$$\int \left\{ \prod_{\boldsymbol{x}} \mathrm{d}\phi_{\boldsymbol{x}} \right\} \prod_{x} |\phi_{\boldsymbol{x}}\rangle\langle\phi_{\boldsymbol{x}}| = 1 \tag{13.1.8}$$

inside the trace, and to keep track of the integration variables we introduce the notation $|\phi^{(t)}_{\boldsymbol{x}}\rangle$ for the integration variables at time $t = 0, a, \ldots, a(N-1)$, where $t=0$ variables are those used in Eq. (13.1.7). In this way we obtain a product of terms of the form

$$\begin{aligned} \left\{ \prod_{\boldsymbol{x}} \langle \phi^{(t+a)}_{\boldsymbol{x}}| \right\} e^{-aH} \left\{ \prod_{\boldsymbol{x}} |\phi^{(t)}_{\boldsymbol{x}}\rangle \right\} \simeq \left\{ \prod_{\boldsymbol{x}} \langle \phi^{(t+a)}_{\boldsymbol{x}}| \right\} e^{-aT} e^{-aV} \left\{ \prod_{\boldsymbol{x}} |\phi^{(t)}_{\boldsymbol{x}}\rangle \right\} = \\ = e^{-aV(\phi^{(t)}_{\boldsymbol{x}})} \left\{ \prod_{\boldsymbol{x}} \langle \phi^{(t+a)}_{\boldsymbol{x}}| \right\} \exp\left(- \sum_{\boldsymbol{x}} \frac{p^2_{\boldsymbol{x}}}{2a^{D-2}} \right) \left\{ \prod_{\boldsymbol{x}} |\phi^{(t)}_{\boldsymbol{x}}\rangle \right\} \ , \end{aligned} \tag{13.1.9}$$

where V is the potential energy term, and T is the kinetic energy term. We can now introduce the momentum eigenvectors $|p\rangle$ and $|k\rangle$:

$$\begin{aligned} \langle \phi^{(t+a)}_{\boldsymbol{x}} | e^{-p^2_{\boldsymbol{x}}/(2a^{D-2})} | \phi^{(t)}_{\boldsymbol{x}} \rangle = \int \mathrm{d}p \mathrm{d}k \langle \phi^{(t+a)}_{\boldsymbol{x}} | p \rangle \langle p | e^{-p^2_{\boldsymbol{x}}/(2a^{D-2})} | k \rangle \langle k | \phi^{(t)}_{\boldsymbol{x}} \rangle = \\ = \int \frac{e^{ip\phi^{(t+a)}_{\boldsymbol{x}}}}{\sqrt{2\pi}} e^{-p^2/(2a^{D-2})} \delta(p-k) \frac{e^{-ik\phi^{(t)}_{\boldsymbol{x}}}}{\sqrt{2\pi}} \mathrm{d}p \mathrm{d}k = \\ = \frac{1}{2\pi} \sqrt{2\pi a^{D-2}} \exp\left[-\frac{a^D}{2} \left(\frac{\phi^{(t+a)}_{\boldsymbol{x}} - \phi^{(t)}_{\boldsymbol{x}}}{a} \right)^2 \right] \ , \end{aligned} \tag{13.1.10}$$

where in the last step we obviously used $\int_{-\infty}^{+\infty} e^{-\alpha t^2 + \beta t} \mathrm{d}t = \sqrt{\frac{\pi}{\alpha}} e^{\beta^2/(4\alpha)}$. This result is completely analogous to the one obtained in QM (see Sec. 8), but for an important detail: for $D > 1$ the value of $\phi^{(t+a)}_{\boldsymbol{x}}$ does not generically converge to $\phi^{(t)}_{\boldsymbol{x}}$ when $a \to 0$, and QFT configurations are much more singular than QM configurations.

Putting all the pieces together, and neglecting proportionality factors, we obtain

$$\begin{aligned} Z(\beta) \propto \sum_{\phi^{(0)}_{\boldsymbol{x}} = \phi^{(\beta)}_{\boldsymbol{x}}} \left\{ \prod_{t,\boldsymbol{x}} \mathrm{d}\phi^{(t)}_{\boldsymbol{x}} \right\} \exp\left\{ - \sum_{t,\boldsymbol{x}} a^D \left[\frac{1}{2} \left(\frac{\phi^{(t+a)}_{\boldsymbol{x}} - \phi^{(t)}_{\boldsymbol{x}}}{a} \right)^2 + \right.\right. \\ \left.\left. + \frac{1}{2} \sum_{\mu > 0} \left(\frac{\phi^{(t)}_{\boldsymbol{x}+a\hat{\mu}} - \phi^{(t)}_{\boldsymbol{x}}}{a} \right)^2 + \frac{1}{2} m^2 (\phi^{(t)}_{\boldsymbol{x}})^2 \right] \right\} \ , \end{aligned} \tag{13.1.11}$$

where the sum extends on all configurations which satisfy the constraint $\phi^{(0)}_{\boldsymbol{x}} = \phi^{(\beta)}_{\boldsymbol{x}}$ for all $\boldsymbol{x}$ values. By formally sending $a \to 0$ we obtain the path-integral expression

$$Z(\beta) = \int_{\phi(0,\boldsymbol{x}) = \phi(\beta,\boldsymbol{x})} [\mathcal{D}\phi(t,\boldsymbol{x})] e^{-S_E[\phi]} \ , \tag{13.1.12}$$

where a (divergent) numerical factor has been absorbed in the definition of the integration measure $[\mathcal{D}\phi]$, and the Euclidean action $S_E[\phi]$ is given by

$$S_E[\phi] = \int_0^\beta \mathrm{d}t \int \mathrm{d}^{D-1}x \left(\frac{1}{2}\dot{\phi}^2 + \frac{1}{2}(\boldsymbol{\nabla}\phi)^2 + \frac{1}{2}m^2\phi^2 \right) \ . \tag{13.1.13}$$

The Euclidean action S_E is formally obtained from the real-time action $S = \int \mathrm{d}t\, L$ by performing the substitution $t \to -i\tau$, which transforms $iS \to -S_E$.

It should be clear that for an interacting field with potential $V(\phi)$ the only modification that is needed in Eq. (13.1.13) is $\frac{1}{2}m^2\phi^2 \to V(\phi)$ in the Euclidean action.

13.2 Discretization of the scalar field

Exactly as in Sec. 8, the thermal average of an observable $O[\phi]$ depending on the field operator ϕ, but independent of the momenta π_ϕ, can be rewritten in the form

$$\langle O[\phi] \rangle = \frac{\mathrm{Tr}(Oe^{-\beta H})}{\mathrm{Tr}(e^{-\beta H})} = \frac{\int_{\phi(0,\boldsymbol{x})=\phi(\beta,\boldsymbol{x})} [\mathcal{D}\phi(t,\boldsymbol{x})] O[\phi] e^{-S_E[\phi]}}{\int_{\phi(0,\boldsymbol{x})=\phi(\beta,\boldsymbol{x})} [\mathcal{D}\phi(t,\boldsymbol{x})] e^{-S_E[\phi]}} \ , \tag{13.2.1}$$

and we can interpret $\frac{e^{-S_E[\phi]}[\mathcal{D}\phi]}{Z}$ as a probability density function. In order to perform a MC estimation of $\langle O \rangle$ we have to discretize the Euclidean action, approximating it with a sum depending on a finite number of variables.

We denote by a the lattice spacing in the temporal and spatial directions (we are thus using a isotropic discretization), and we denote by $\phi_{\boldsymbol{n}}$ the value of $\phi(t,\boldsymbol{x})$ (note that $\boldsymbol{n}$ is a point of the D-dimensional space-time lattice and it entries are dimensionless: $(t,\boldsymbol{x}) \simeq a\boldsymbol{n}$); as already done in Sec. 13.1 we approximate derivatives by their so called forward lattice counterparts:

$$\partial_\mu \phi(t,\boldsymbol{x}) \to \partial_\mu^{(F)} \phi_{\boldsymbol{n}} = \frac{\phi_{\boldsymbol{n}+\hat{\mu}} - \phi_{\boldsymbol{n}}}{a} \ . \tag{13.2.2}$$

In order to write the lattice action in dimensionless form it is usual, in QFT, to rescale all the quantities by the appropriate powers of the lattice spacing a, and dimensionless quantities will be denoted by a $\hat{}$ symbol. Since in D-dimensional space we have $[\phi] = (D-2)/2$, we introduce $\hat{\phi}_{\boldsymbol{n}} = a^{(D-2)/2}\phi_{\boldsymbol{n}}$ and $\hat{m} = am$. With these notations the discrete Euclidean lattice action can be written in the form

$$\begin{aligned} S_L &= \frac{1}{2}\sum_{\boldsymbol{n}} a^D \left\{ \frac{1}{a^D}\hat{m}^2\hat{\phi}_{\boldsymbol{n}}^2 + \sum_{\mu\geq 0} \frac{1}{a^D}(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}} - \hat{\phi}_{\boldsymbol{x}})^2 \right\} = \\ &= \frac{1}{2}\sum_{\boldsymbol{n}} \left\{ (\hat{m}^2 + 2D)\hat{\phi}_{\boldsymbol{n}}^2 - 2\sum_{\mu\geq 0} \hat{\phi}_{\boldsymbol{n}}\, \hat{\phi}_{\boldsymbol{n}+\hat{\mu}} \right\} = \frac{1}{2}\sum_{\boldsymbol{n},\boldsymbol{j}} \hat{\phi}_{\boldsymbol{n}} K_{\boldsymbol{n}\,\boldsymbol{j}} \hat{\phi}_{\boldsymbol{j}} \ , \end{aligned} \tag{13.2.3}$$

where the symmetric kernel K is defined by

$$K_{\boldsymbol{n}\,\boldsymbol{j}} = \hat{m}^2\delta_{\boldsymbol{n},\boldsymbol{j}} - \sum_{\mu\geq 0} (\delta_{\boldsymbol{n}+\hat{\mu},\boldsymbol{j}} + \delta_{\boldsymbol{n}-\hat{\mu},\boldsymbol{j}} - 2\delta_{\boldsymbol{n},\boldsymbol{j}}) \ , \tag{13.2.4}$$

and the partition function is given by

$$Z(\beta) = \int \left(\prod_{\boldsymbol{n}} \mathrm{d}\hat{\phi}_{\boldsymbol{n}} \right) e^{-S_L} \ . \tag{13.2.5}$$

To approach the continuum limit we have to send $a \to 0$ at fixed physical mass m, hence in term of the dimensionless coupling entering S_L the continuum limit corresponds to $\hat{m} \to 0$. Note that for

interacting theories the relation between $\hat{m}$ and m is nontrivial; we will come back to the definition of the continuum limit for a generic interacting theory after having discussed the spectrum of QFTs, in Sec. 14.2.

To convince ourselves that Eq. (13.2.3) is a proper discretization of the Euclidean action, we can study the behavior of the lattice propagator in the $a \to 0$ limit, to verify that the correct continuum expression is recovered. Given a function $f_{\boldsymbol{n}}$ defined on the sites of a lattice with periodic boundary conditions, we can define its lattice Fourier transform by

$$\tilde{f}_{\boldsymbol{p}} = \sum_{\boldsymbol{n}} e^{-i\boldsymbol{p}\cdot\boldsymbol{n}} f_{\boldsymbol{n}} \ , \tag{13.2.6}$$

where the components of the dimensionless momentum $\boldsymbol{p}$ can only take the values $p_\mu = \frac{2\pi}{N_\mu} b_\mu$, $N_\mu = L_\mu/a$ is the number of sites of the lattice in the μ-th direction, and $b_\mu = 0, 1, \dots N_\mu - 1$. In the solid-state terminology $\boldsymbol{p}$ is a vector in the first Brillouin zone of the reciprocal lattice. The inverse lattice Fourier transform is given by

$$f_{\boldsymbol{n}} = \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} e^{i\boldsymbol{p}\cdot\boldsymbol{n}} \tilde{f}_{\boldsymbol{p}} \ , \tag{13.2.7}$$

where the sum extends on all momenta in the first Brillouin zone, and $\hat{V} = \prod_\mu N_\mu$ is the number of sites of the lattice. To show that Eq. (13.2.7) is the inverse of Eq. (13.2.6) we can proceed as follows: if $\boldsymbol{k}$ is a vector in the first Brillouin zone, and $\boldsymbol{d}$ a generic lattice point, by using translation invariance we have

$$e^{i\boldsymbol{k}\cdot\boldsymbol{d}} \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot(\boldsymbol{r}+\boldsymbol{d})} = \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \ . \tag{13.2.8}$$

Since $\boldsymbol{d}$ is a generic lattice point, we can have $e^{i\boldsymbol{k}\cdot\boldsymbol{d}} = 1$ only if $\boldsymbol{k} = 0$, hence

$$\sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \hat{V}\delta_{\boldsymbol{k},0} \ , \tag{13.2.9}$$

and analogously

$$\sum_{\boldsymbol{p}} e^{i\boldsymbol{p}\cdot\boldsymbol{r}} = \hat{V}\delta_{\boldsymbol{r},0} \ . \tag{13.2.10}$$

Using these identities we have

$$\frac{1}{\hat{V}} \sum_{\boldsymbol{p}} e^{i\boldsymbol{p}\cdot\boldsymbol{n}} \tilde{f}_{\boldsymbol{p}} = \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} e^{i\boldsymbol{p}\cdot\boldsymbol{n}} \sum_{\boldsymbol{m}} e^{-i\boldsymbol{p}\cdot\boldsymbol{m}} f_{\boldsymbol{m}} \ , = \sum_{\boldsymbol{m}} \delta_{\boldsymbol{n}-\boldsymbol{m},0} f_{\boldsymbol{m}} = f_{\boldsymbol{n}} \ , \tag{13.2.11}$$

which proves Eq. (13.2.7).

Using Eq. (13.2.10) to rewrite the δs in the definition of $K_{\boldsymbol{n}\,\boldsymbol{j}}$ in Eq. (13.2.4) we have

$$\begin{aligned} K_{\boldsymbol{n}\,\boldsymbol{j}} &= \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} \left\{ \hat{m}^2 - \sum_{\mu\geq 0} \left(e^{ip_\mu} + e^{-ip_\mu} - 2 \right) \right\} e^{i\boldsymbol{p}\cdot(\boldsymbol{n}-\boldsymbol{j})} = \\ &= \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} \left\{ \hat{m}^2 - 2 \sum_{\mu\geq 0} (2\cos(p_\mu) - 2) \right\} e^{i\boldsymbol{p}\cdot(\boldsymbol{n}-\boldsymbol{j})} = \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} \left\{ \hat{m}^2 + \sum_{\mu\geq 0} 4\sin^2\left(\frac{p_\mu}{2}\right) \right\} e^{i\boldsymbol{p}\cdot(\boldsymbol{n}-\boldsymbol{j})} \ . \end{aligned} \tag{13.2.12}$$

where we used $\cos(\alpha) = 1 - 2\sin^2(\alpha/2)$ in the last step. It is now simple to verify that

$$(K^{-1})_{\boldsymbol{n}\,\boldsymbol{j}} = \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} \frac{1}{\hat{m}^2 + \sum_{\mu\geq 0} 4\sin^2\left(\frac{p_\mu}{2}\right)} e^{i\boldsymbol{p}\cdot(\boldsymbol{n}-\boldsymbol{j})} \ , \tag{13.2.13}$$

indeed, defining $\mathcal{K}(\boldsymbol{p}) = \hat{m}^2 + \sum_{\mu\geq 0} 4\sin^2\left(\frac{p_\mu}{2}\right)$, we have (using Eqs. (13.2.9)-(13.2.10))

$$\begin{aligned} \sum_{\boldsymbol{n}} K_{\boldsymbol{a}\,\boldsymbol{n}} (K^{-1})_{\boldsymbol{n}\,\boldsymbol{b}} &= \sum_{\boldsymbol{n}} \frac{1}{\hat{V}} \sum_{\boldsymbol{p}_1} \mathcal{K}(\boldsymbol{p}_1) e^{i\boldsymbol{p}_1\cdot(\boldsymbol{n}-\boldsymbol{a})} \frac{1}{\hat{V}} \sum_{\boldsymbol{p}_2} \frac{1}{\mathcal{K}(\boldsymbol{p}_2)} e^{i\boldsymbol{p}_2\cdot(\boldsymbol{b}-\boldsymbol{n})} = \\ &= \frac{1}{\hat{V}} \sum_{\boldsymbol{p}_1,\boldsymbol{p}_2} \mathcal{K}(\boldsymbol{p}_1) \frac{1}{\mathcal{K}(\boldsymbol{p}_2)} \delta_{\boldsymbol{p}_1,\boldsymbol{p}_2} e^{i\boldsymbol{p}_1\cdot(\boldsymbol{b}-\boldsymbol{a})} = \delta_{\boldsymbol{a},\boldsymbol{b}} \ . \end{aligned} \tag{13.2.14}$$

We now have to rewrite the lattice propagator in physical units, and see what happens in the limit $a \to 0$. In this limit we have $\int \mathrm{d}^D x \mathrm{d}^D y \simeq a^{2D} \sum_{\boldsymbol{n},\boldsymbol{j}}$, hence from

$$\sum_{\boldsymbol{n},\boldsymbol{j}} \hat{\phi}_{\boldsymbol{n}} K_{\boldsymbol{n}\,\boldsymbol{j}} \hat{\phi}_{\boldsymbol{j}} = \sum_{\boldsymbol{n},\boldsymbol{j}} a^{D-2} \phi_{\boldsymbol{n}} K_{\boldsymbol{n}\,\boldsymbol{j}} \phi_{\boldsymbol{j}} = \sum_{\boldsymbol{n},\boldsymbol{j}} a^{2D} \phi_{\boldsymbol{n}} \frac{K_{\boldsymbol{n}\,\boldsymbol{j}}}{a^{D+2}} \phi_{\boldsymbol{j}} \tag{13.2.15}$$

we see that $\frac{K_{\boldsymbol{n}\,\boldsymbol{j}}}{a^{D+2}}$ is the kernel with the correct dimensionality, which should converge to $K(\boldsymbol{x},\boldsymbol{y})$ (we denote here by $\boldsymbol{x}$ a point of the Euclidean space-time). The continuum propagator $K^{-1}(\boldsymbol{y},\boldsymbol{z})$ is defined by the equation

$$\int \mathrm{d}^D y K(\boldsymbol{x},\boldsymbol{y}) K^{-1}(\boldsymbol{y},\boldsymbol{z}) = \delta(\boldsymbol{x}-\boldsymbol{y}) \ , \tag{13.2.16}$$

and close to the continuum we have $\delta_{\boldsymbol{a},\boldsymbol{b}} \simeq a^D \delta(\boldsymbol{x}-\boldsymbol{z})$. The lattice propagator with the correct dimensionality is thus

$$\begin{aligned}(K^{-1})_{\boldsymbol{n}\,\boldsymbol{j}} &= \frac{1}{\hat{V}}\frac{1}{a^{D-2}}\sum_{\boldsymbol{p}}\frac{1}{\hat{m}^2+\sum_{\mu\geq 0}4\sin^2\left(\frac{p_\mu}{2}\right)}e^{i\boldsymbol{p}\cdot(\boldsymbol{n}-\boldsymbol{j})} = \\ &= \frac{1}{\hat{V}}\sum_{\boldsymbol{p}}\frac{1}{a^D}\frac{1}{m^2+\sum_{\mu\geq 0}\frac{4}{a^2}\sin^2\left(\frac{aq_\mu}{2}\right)}e^{ia\boldsymbol{q}\cdot(\boldsymbol{n}-\boldsymbol{j})}\ ,\end{aligned} \tag{13.2.17}$$

where in the last equality we introduced the dimensionfull (D-)momentum $\boldsymbol{q}=\boldsymbol{p}/a$. In the limit $a\to 0$ the sum on the first Brillouin zone becomes

$$\frac{1}{\hat{V}}\sum_{\boldsymbol{p}} \to \int_{-\pi}^{\pi}\frac{\mathrm{d}^D p}{(2\pi)^D} = a^D\int_{-\pi/a}^{\pi/a}\frac{\mathrm{d}^D q}{(2\pi)^D}\ , \tag{13.2.18}$$

where the integration region of the last integral is the cube $(-\pi/a,\pi/a)^D$. Hence

$$(K^{-1})_{\boldsymbol{n}\,\boldsymbol{j}} \to \int_{-\pi/a}^{\pi/a}\frac{\mathrm{d}^D q}{(2\pi)^D}\frac{1}{m^2+\sum_{\mu\geq 0}\frac{4}{a^2}\sin^2\left(\frac{aq_\mu}{2}\right)}e^{i\boldsymbol{q}\cdot(\boldsymbol{x}-\boldsymbol{y})}\ . \tag{13.2.19}$$

To conclude it is sufficient to note that in the first Brillouin zone $\sin(aq_\mu/2)$ vanishes only for $q_\mu=0$, hence in the limit $a\to 0$ we can expand using $|aq_\mu|\ll\pi$ (since otherwise the denominator would diverge), and finally obtain the standard continuum result

$$K^{-1}(\boldsymbol{x},\boldsymbol{y}) = \int\frac{\mathrm{d}^D q}{(2\pi)^D}\frac{1}{m^2+\boldsymbol{q}^2}e^{i\boldsymbol{q}\cdot(\boldsymbol{x}-\boldsymbol{y})}\ . \tag{13.2.20}$$

It is not difficult to show that not all the possible discretizations of the free scalar action produce the correct continuum limit. In particular, let us investigate what happens by using the lattice symmetric discretization of the derivative instead of the lattice forward discretization:

$$\partial_\mu\phi(t,\boldsymbol{x}) \to \partial_\mu^{(S)}\phi_{\boldsymbol{n}} = \frac{\phi_{\boldsymbol{n}+\hat{\mu}}-\phi_{\boldsymbol{n}-\hat{\boldsymbol{\mu}}}}{2a}\ . \tag{13.2.21}$$

Using the symmetric discretization it is immediate to see that the lattice action takes the form

$$\begin{aligned}S_L &= \frac{1}{2}\sum_{\boldsymbol{n}}\left\{\hat{m}^2\hat{\phi}_{\boldsymbol{n}}^2+\frac{1}{4}\sum_{\mu\geq 0}\left(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}}-\hat{\phi}_{\boldsymbol{n}-\hat{\mu}}\right)^2\right\} = \\ &= \frac{1}{2}\sum_{\boldsymbol{n}}\left\{\left(\hat{m}^2+\frac{D}{2}\right)\hat{\phi}_{\boldsymbol{n}}^2-\frac{1}{2}\sum_{\mu\geq 0}\hat{\phi}_{\boldsymbol{n}+\hat{\mu}}\hat{\phi}_{\boldsymbol{n}-\hat{\mu}}\right\}\ ,\end{aligned} \tag{13.2.22}$$

and this action has a peculiar property: it is the sum of 2^D independent contributions. Let us discuss, for the sake of the simplicity, the $D=2$ case, but everything can be clearly extended to the general case. The $D=2$ lattice can be decomposed as the union of 4 sub-lattices, identified by the parity of the components of the lattice sites. For example, the sub-lattice (e,e) is the one in which all the lattice sites have even components, while the points of the sub-lattice (e,o) have the first component which is even and the second that is odd. While this decomposition is obviously independent of the discretization adopted, the peculiarity of the discretization in Eq. (13.2.22) is that the variables of the different sub-lattices are decoupled from each other. This is probably more clearly seen if we use the identity (valid for any value of μ)

$$\sum_{\boldsymbol{n}}\left(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}}-\hat{\phi}_{\boldsymbol{n}-\hat{\mu}}\right)^2 = \sum_{\boldsymbol{n}}\left(\hat{\phi}_{\boldsymbol{n}+2\hat{\mu}}-\hat{\phi}_{\boldsymbol{n}}\right)^2 \tag{13.2.23}$$

to rewrite S_L in the equivalent form

$$S_L = \frac{1}{2}\sum_{\boldsymbol{n}}\left\{\hat{m}^2\hat{\phi}_{\boldsymbol{n}}^2+\frac{1}{4}\sum_{\mu\geq 0}\left(\hat{\phi}_{\boldsymbol{n}+2\hat{\mu}}-\hat{\phi}_{\boldsymbol{n}}\right)^2\right\}\ , \tag{13.2.24}$$

from which it should be clear that only sites with the same parity interact with each other. Since the action is the sum of four independent and identical components, the partition function is the

fourth-power of the sub-lattice partition function:

$$Z(\beta) = Z^4_{(e,e)}(\beta) \ , \quad Z_{(e,e)}(\beta) = \int \prod_{\boldsymbol{n}\in(e,e)} \mathrm{d}\hat{\phi}_{\boldsymbol{n}} e^{-S_{(e,e)}} \ . \tag{13.2.25}$$

Note now that the action in Eq. (13.2.24), restricted to one of the sub-lattices, has exactly the same form of Eq. (13.2.3), since in the sub-lattice the distance between two next-neighboring sites is $2a$ instead of a: the lattice action written using the symmetric lattice derivative is thus equivalent, in each sub-space, to the discretization carried out using the forward lattice derivative. For this reason we expect the action Eq. (13.2.24) to describe, in the continuum, 4 (in general 2^D) independent scalar particles (known as "doublers"), instead of just one.

While for the scalar field case the symmetric discretization can appear quite pathological and unnatural, a similar problem is present in all fermionic discretizations which preserves chiral symmetry, see, e. g., [81] §4, [82] §4 [83] §13.

The existence of doublers can be seen also by performing the continuum limit of the lattice propagator: the action in Eq. (13.2.22) can be written in the form $\frac{1}{2}\sum_{\boldsymbol{n},\boldsymbol{j}} \hat{\phi}_{\boldsymbol{n}} K_{\boldsymbol{n}\,\boldsymbol{j}} \hat{\phi}_{\boldsymbol{j}}$, with the kernel

$$K_{\boldsymbol{n}\,\boldsymbol{j}} = \hat{m}^2 \delta_{\boldsymbol{n},\boldsymbol{j}} + \frac{1}{4}\sum_{\mu} \left(2\delta_{\boldsymbol{n},\boldsymbol{j}} - \delta_{\boldsymbol{n}+2\hat{\mu},\boldsymbol{j}} - \delta_{\boldsymbol{n}-2\hat{\mu},\boldsymbol{j}}\right) \ . \tag{13.2.26}$$

This can be written in Fourier transform (by using Eqs. (13.2.9)-(13.2.10)) as

$$\begin{aligned} K_{\boldsymbol{n}\,\boldsymbol{j}} &= \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} \left\{ \hat{m}^2 + \frac{1}{4}\sum_{\mu} \left(2 - e^{i2p_\mu} - e^{-i2p_\mu}\right) \right\} e^{i\boldsymbol{p}\cdot(\boldsymbol{n}-\boldsymbol{j})} = \\ &= \frac{1}{\hat{V}} \sum_{\boldsymbol{p}} \left\{ \hat{m}^2 + \sum_{\mu} \sin^2 p_\mu \right\} e^{i\boldsymbol{p}\cdot(\boldsymbol{n}-\boldsymbol{j})} \ , \end{aligned} \tag{13.2.27}$$

where we used $2 - 2\cos(2\alpha) = 4\sin^2\alpha$. By proceeding exactly as done in the case of the forward discretization we obtain for the propagator in the $a \to 0$ limit the expression

$$\int_{-\pi/a}^{\pi/a} \frac{\mathrm{d}^D q}{(2\pi)^D} \frac{1}{m^2 + \frac{1}{a^2}\sum_\mu \sin^2(aq_\mu)} e^{i\boldsymbol{q}(\boldsymbol{x}-\boldsymbol{y})} \ . \tag{13.2.28}$$

The function $\sin(aq_\mu)$ vanishes not only for $q_\mu = 0$ but also at the boundaries of the first Brillouin zone (note that $q_\mu = +\pi/a$ and $-\pi/a$ are in fact the same value, due to periodic boundary conditions), and in total there are 2^D zeros. Developing $\sin(aq_\mu)$ close to one of these zeros we get the propagator of a scalar field in the continuum, thus the symmetric discretization generates 2^D independent scalar fields in the continuum.

13.3 Simulation of the lattice scalar field

To simulate the theory with Euclidean lattice action Eq. (13.2.3) it is convenient to rewrite the action in the form

$$S_L = \frac{1}{2}\left\{(\hat{m}^2 + 2D)\hat{\phi}^2_{\boldsymbol{n}} - 2\hat{\phi}_{\boldsymbol{n}} S_{\boldsymbol{n}}\right\} + \text{independent of } \hat{\phi}_{\boldsymbol{n}} \ , \tag{13.3.1}$$

where we introduced the notation

$$S_{\boldsymbol{n}} = \sum_{\mu} \left(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}} + \hat{\phi}_{\boldsymbol{n}-\hat{\mu}}\right) \tag{13.3.2}$$

for the sum of the scalar variables in the next-neighbors of the site $\boldsymbol{n}$.

The more general update scheme that can be used is the simple Metropolis algorithm: after having selected the lattice site $\boldsymbol{r}$ we generate the trial variable $\hat{\phi}^t_{\boldsymbol{r}}$ by using

$$\hat{\phi}^t_{\boldsymbol{r}} = \hat{\phi}_{\boldsymbol{r}} + \Delta(2\,\mathrm{rand} - 1) \ , \tag{13.3.3}$$

where "rand" is a random number with uniform pdf in $[0, 1)$, and Δ is a fixed number. Such a trial state is accepted or rejected with probability $\exp(-\Delta S_L)$, where $\Delta S_L = S_L[\hat{\phi}^t] - S_L[\hat{\phi}]$, and

the parameter Δ is selected in such a way that the average acceptance probability is neither too small not too close to 1. This algorithm is clearly irreducible, since any value of $\hat{\phi}_{\boldsymbol{n}}$ can be reached in a finite number of steps, and it is aperiodic since there is the possibility that the state is not changed[1]. The site $\boldsymbol{r}$ can be selected in two different ways: by randomly selecting a site of the lattice with uniform pdf, or by performing a deterministic sweep of all the lattice sites. In the first case the algorithm satisfies the detailed balance condition, while only the balance condition is fulfilled in the second case, see Sec. 3.3.3 and the discussion in Sec. 6.1. The Metropolis update can obviously be adopted also when studying interacting theories: we only have to use the interacting action instead of the free action when computing ΔS.

When studying the free theory it is also easy to implement the heat-bath and microcanonical updates. The conditional pdf of the variable $\hat{\phi}_{\boldsymbol{n}}$ is indeed proportional to

$$\exp\left(\hat{\phi}_{\boldsymbol{n}} S_{\boldsymbol{n}} - \frac{\hat{m}^2 + 2D}{2}\hat{\phi}_{\boldsymbol{n}}^2\right) \propto \exp\left\{-\frac{\hat{m}^2 + 2D}{2}\left(\hat{\phi}_{\boldsymbol{n}} - \frac{S_{\boldsymbol{n}}}{\hat{m}^2 + 2D}\right)^2\right\} , \tag{13.3.4}$$

i. e. a Gaussian pdf with average $\mu = \frac{S_{\boldsymbol{n}}}{\hat{m}^2+2D}$ and standard deviation $\sigma = \frac{1}{\sqrt{\hat{m}^2+2D}}$, and such a distribution can be easily sampled using the Box-Muller algorithm (see Sec. 2.3). The microcanonical update is obtained by selecting for the site variable $\hat{\phi}_{\boldsymbol{n}}$ a value which is obtained by reflecting with respect to μ the original value:

$$\hat{\phi}_{\boldsymbol{n}} \to 2\frac{S_{\boldsymbol{n}}}{\hat{m}^2 + 2D} - \hat{\phi}_{\boldsymbol{n}} \ . \tag{13.3.5}$$

It is immediate to verify that such a transformation is involutive and does not change the Euclidean action, hence it is a legitimate microcanonical step (see the analogous discussion in Sec. 6.3). It is possible to extend the heat-bath and microcanonical algorithms also to the case of interacting theories, however in the interacting case a von Neumann accept/reject step (see Sec. 2.4) is typically required in the heat-bath to sample the conditional pdf, and the numerical solution of a nonlinear equation is needed to obtain the value to be used in the microcanonical step.

When we introduced Eq. (13.2.3), approximating the continuum derivative with its forward lattice form, we neglected $O(a^2)$ terms, so it is natural to expect average values to converge to their continuum limits with $O(a^2)$ corrections. This is indeed what happens, apart from the renormalizations required by some observables, that will be discussed in Chap. 15.

[1] This sentence would obviously require more care, since single points have zero measure. From the operative point of view, $\mathbb{R}$ is represented on any physical CPU by a large but finite number of points, so this problem does not exist in practice.

Chapter 14

MCMC in quantum field theory: spectrum

14.1 Spectrum computation

Let us briefly recall what was found in quantum mechanics assuming the presence only of a discrete spectrum (see Sec. 10): given an Hermitian operator O, if we define $O(\tau) = e^{H\tau} O e^{-H\tau}$ (which is the analytic continuation under $t \to -i\tau$ of the Heisenberg representation $O(t) = e^{iHt} O e^{-iHt}$), we have in the zero temperature limit[1]

$$\begin{aligned} &\langle O(\tau)O(0)\rangle - \langle O\rangle^2 = \langle 0|e^{H\tau} O e^{-H\tau} O|0\rangle - \langle 0|O|0\rangle^2 = \\ &= \sum_n \langle 0|e^{H\tau} O e^{-H\tau}|n\rangle\langle n|O|0\rangle - \langle 0|O|0\rangle^2 = \sum_n e^{-(E_n - E_0)\tau} |\langle n|O|0\rangle|^2 - \langle 0|O|0\rangle^2 = \\ &= \sum_{n>0} e^{-(E_n - E_0)\tau} |\langle n|O|0\rangle|^2 \quad \stackrel{\tau\to\infty}{\longrightarrow} \quad e^{-(E_{\bar n} - E_0)\tau} |\langle \bar n|O|0\rangle|^2 \ , \end{aligned} \tag{14.1.1}$$

where in the last step we introduced the notation

$$\bar n = \min\{n \in \mathbb{N} \text{ such that } \langle n|O|0\rangle \neq 0\} \ , \tag{14.1.2}$$

and the large (Euclidean) time limit generally means $(E_{\bar n+1} - E_0)\tau \gg (E_{\bar n} - E_0)\tau$. From the large time behavior of the O correlator we can thus estimate the energy gap between the $\bar n$-th state and the fundamental state.

In a quantum field theory the single particle states have energies $\sqrt{m^2 + p^2}$, hence the spectrum is always continuum, moreover also multiparticle states exist, which also have continuous spectra. For these reasons some complications arise in the QFT case with respect to the QM case.

Let us start by studying the single particle case, considering for the sake of the simplicity the example of a free scalar Hermitian field: we have in Minkowski space-time

$$\phi(t, \boldsymbol{x}) = \int \frac{\mathrm{d}^{D-1} p}{(2\pi)^{D-1}} \frac{1}{\sqrt{2E_{\boldsymbol{p}}}} \left(a_{\boldsymbol{p}} e^{-i(E_{\boldsymbol{p}} t - \boldsymbol{p}\cdot\boldsymbol{x})} + a^\dagger_{\boldsymbol{p}} e^{+i(E_{\boldsymbol{p}} t - \boldsymbol{p}\cdot\boldsymbol{x})} \right) \ , \tag{14.1.3}$$

where

$$[a_{\boldsymbol{p}}, a^\dagger_{\boldsymbol{k}}] = (2\pi)^{D-1} \delta(\boldsymbol{p} - \boldsymbol{k}) \ . \tag{14.1.4}$$

Simulation are carried out in a finite volume system of linear extent L, and the state with the smallest nonvanishing momentum (assuming periodic b. c.) has energy $\sqrt{m^2 + \boldsymbol{p}^2_{min}}$ with $\boldsymbol{p}_{min} =$

[1] We assume, as usual, the states to be ordered in such a way that $E_0 < E_1 < E_2 < \cdots$.

$(2\pi/L, 0, \dots, 0)$. The energy of the first excited state above the vacuum ($E^{(0)} = 0$, conventionally) is thus $E^{(1)} = m$ (corresponding to $\boldsymbol{p} = 0$), and the energy of the second excited state is

$$E^{(2)} = \sqrt{m^2 + \frac{4\pi^2}{L^2}} \simeq m\left(1 + \frac{2\pi^2}{m^2 L^2}\right) . \tag{14.1.5}$$

To correctly estimate the gap $E^{(1)} - E^{(0)} = m$ between the fundamental and the first excited state using Eq. (14.1.1), we need to use values of τ such that $\exp[-(E^{(2)} - E^{(0)})\tau] \ll \exp[-(E^{(1)} - E^{(0)})\tau]$, hence $\tau \gg \frac{mL}{2\pi^2} L$. If we want finite volume effect to be negligible we need to use $mL \gg 1$, hence the previous relation imply that we have to study the correlator for times $\tau \gg L \gg 1/m$. This is however unfeasible, since the correlator approaches zero with a timescale $\approx 1/m$, and the physical signal would be completely hidden by statistical errors.

The presence of the continuous spectrum (or of almost-degenerate states in a finite volume) generates power-law corrections in the large distance behavior of Euclidean correlators. A simple example is provided by the free Euclidean propagator

$$G(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}) \phi(\boldsymbol{y}) \rangle = \int \frac{\mathrm{d}^D k}{(2\pi)^D} \frac{e^{i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}}{m^2 + k^2} , \tag{14.1.6}$$

which has the large distance behavior[2]

$$G(\boldsymbol{x}, \boldsymbol{y}) \propto \frac{1}{|\boldsymbol{x} - \boldsymbol{y}|^{(D-1)/2}} e^{-m|\boldsymbol{x}-\boldsymbol{y}|} . \tag{14.1.7}$$

Note that $D = 1$ corresponds to QM, and in this case the behavior is exponential, as in Eq. (14.1.1).

Let us show that Eq. (14.1.7) is indeed the asymptotic behavior of Eq. (14.1.6): we introduce the Schwinger proper time t:

$$G(\boldsymbol{x}, \boldsymbol{y}) = \int \frac{\mathrm{d}^D k}{(2\pi)^D} \frac{e^{i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}}{m^2 + k^2} = \frac{1}{(2\pi)^D} \int \mathrm{d}^D k \int_0^\infty \mathrm{d}t\, e^{i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y}) - t(m^2 + \boldsymbol{k}^2)} , \tag{14.1.8}$$

and by using $\int \mathrm{d}z e^{-az^2 - bz} = \sqrt{\frac{\pi}{a}} e^{b^2/(4a)}$ for each component of $\boldsymbol{k}$ we get

$$G(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{(2\pi)^D} \int_0^\infty \mathrm{d}t \left(\frac{\pi}{t}\right)^{D/2} e^{-\frac{1}{4t}|\boldsymbol{x}-\boldsymbol{y}|^2 - tm^2} = \frac{\pi^{D/2}}{(2\pi)^D} \int_0^\infty \mathrm{d}t\, e^{f(t)} , \tag{14.1.9}$$

where we defined

$$f(t) = -tm^2 - \frac{1}{4t}|\boldsymbol{x} - \boldsymbol{y}|^2 - \frac{D}{2}\log(t) . \tag{14.1.10}$$

When $|\boldsymbol{x} - \boldsymbol{y}| \gg 1/m$ we can estimate the asymptotic behavior of this integral by using the Laplace method, see, e. g., [84] §2.4 or [85] §6.4. When $|\boldsymbol{x} - \boldsymbol{y}| \gg 1/m$ the equation $f'(t) = 0$ has positive solution $\bar{t} \simeq \frac{1}{2m}|\boldsymbol{x} - \boldsymbol{y}|$, moreover

$$f(\bar{t}) = -m|\boldsymbol{x} - \boldsymbol{y}| - \frac{D}{2}\log\frac{|\boldsymbol{x} - \boldsymbol{y}|}{2m} , \quad f''(\bar{t}) \simeq -\frac{4m^3}{|\boldsymbol{x} - \boldsymbol{y}|} , \tag{14.1.11}$$

hence

$$\begin{aligned} G(\boldsymbol{x}, \boldsymbol{y}) &\simeq \frac{\pi^{D/2}}{(2\pi)^D} \int_{-\infty}^\infty \mathrm{d}t\, e^{f(\bar{t}) + \frac{1}{2}f''(\bar{t})(t-\bar{t})^2} = \frac{\pi^{D/2}}{(2\pi)^D} e^{f(\bar{t})} \sqrt{\frac{2\pi}{|f''(\bar{t})|}} \simeq \\ &\simeq \frac{1}{2m}\left(\frac{m}{2\pi|\boldsymbol{x} - \boldsymbol{y}|}\right)^{\frac{D-1}{2}} e^{-m|\boldsymbol{x}-\boldsymbol{y}|} . \end{aligned} \tag{14.1.12}$$

To avoid contaminations from the single particle continuum states it is customary to project field operators on fixed momentum states: given an operator $O(t, \boldsymbol{x})$ we define the operator $O_{\boldsymbol{k}}(t)$ by means of a Fourier transform on spatial variables:

$$O_{\boldsymbol{k}}(t) = \int \mathrm{d}^{D-1} x\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}} O(t, \boldsymbol{x}) . \tag{14.1.13}$$

Since $O(t, \boldsymbol{x}) = e^{i\boldsymbol{x}\cdot\boldsymbol{p}} O(t, \boldsymbol{0}) e^{-i\boldsymbol{x}\cdot\boldsymbol{p}}$, we have $\langle 0|O(t, \boldsymbol{x})|\boldsymbol{p}\rangle = e^{-i\boldsymbol{p}\cdot\boldsymbol{x}} \langle 0|O(t, \boldsymbol{0})|\boldsymbol{p}\rangle$, and the matrix element $\langle 0|O_{\boldsymbol{k}}(t)|\boldsymbol{p}\rangle$ is proportional to $\delta(\boldsymbol{k} - \boldsymbol{p})$. Hence all the almost degenerate single particle

[2] This large distance behavior is just the Ornstein-Zernike form of non-critical correlators with $\xi = 1/m$, see Eq. (5.2.9).

states with $\boldsymbol{k} \approx \boldsymbol{p}$ do not contribute to the correlators of $O_{\boldsymbol{k}}$. Note that this is true whenever translation invariance holds, not only for free fields.

Projecting the free field operator ϕ on fixed momentum states we get (in Minkowski space-time)

$$\begin{aligned} O_{\boldsymbol{k}}(t) &= \int \mathrm{d}^{D-1}x e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\phi(t,\boldsymbol{x}) = \\ &= \int \mathrm{d}^{D-1}x e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \int \frac{\mathrm{d}^{D-1}p}{(2\pi)^{D-1}} \frac{1}{\sqrt{2E_{\boldsymbol{p}}}} \left(a_{\boldsymbol{p}} e^{-i(E_{\boldsymbol{p}}t - \boldsymbol{p}\cdot\boldsymbol{x})} + a^\dagger_{\boldsymbol{p}} e^{+i(E_{\boldsymbol{p}}t - \boldsymbol{p}\cdot\boldsymbol{x})} \right) = \\ &= \frac{1}{\sqrt{2E_{\boldsymbol{k}}}} \left(a_{-\boldsymbol{k}} e^{-iE_{\boldsymbol{k}}t} + a^\dagger_{\boldsymbol{k}} e^{+iE_{\boldsymbol{k}}t} \right) \ , \end{aligned} \tag{14.1.14}$$

and it is simple to verify that no power-law corrections are present in the large time behavior of the Euclidean temporal correlator of these operators: using the fact that in the Euclidean time we have $O(\tau) = e^{H\tau}O(0)e^{-H\tau}$, we get for $\tau > 0$

$$\begin{aligned} &\langle O^\dagger_{\boldsymbol{q}}(\tau) O_{\boldsymbol{p}}(0)\rangle = \langle e^{H\tau} O^\dagger_{\boldsymbol{q}}(0) e^{-H\tau} O_{\boldsymbol{p}}(0)\rangle = \\ &= \frac{1}{2\sqrt{E_{\boldsymbol{p}}E_{\boldsymbol{q}}}} \langle 0 | a_{\boldsymbol{q}} e^{-H\tau} a^\dagger_{\boldsymbol{p}} | 0\rangle = (2\pi)^{D-1}\delta(\boldsymbol{q}-\boldsymbol{p}) \frac{1}{2E_{\boldsymbol{p}}} e^{-E_{\boldsymbol{p}}\tau} \ , \end{aligned} \tag{14.1.15}$$

where in the last step we used the fact that $a^\dagger_{\boldsymbol{p}}|0\rangle$ is an eigenstate of the free Hamiltonian H with eigenvalue $E_{\boldsymbol{p}}$, and the canonical commutation relations. The same conclusion can be reached by using the explicit form of the propagator:

$$\begin{aligned} \langle O^\dagger_{\boldsymbol{q}}(\tau) O_{\boldsymbol{p}}(0)\rangle &= \left\langle \int \mathrm{d}^{D-1}x\, e^{-i\boldsymbol{q}\cdot\boldsymbol{x}} \phi(\tau,\boldsymbol{x}) \int \mathrm{d}^{D-1}y\, e^{i\boldsymbol{p}\cdot\boldsymbol{y}} \phi(0,\boldsymbol{y}) \right\rangle = \\ &= \int \mathrm{d}^{D-1}x \int \mathrm{d}^{D-1}y\, e^{i\boldsymbol{p}\cdot\boldsymbol{y} - i\boldsymbol{q}\cdot\boldsymbol{x}} \int \frac{\mathrm{d}^D k}{(2\pi)^D} \frac{e^{ik_0\tau} e^{i\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}}{m^2 + k_0^2 + \boldsymbol{k}^2} = \\ &= \int \frac{\mathrm{d}^D k}{(2\pi)^D} (2\pi)^{D-1}\delta(\boldsymbol{q}-\boldsymbol{k}) (2\pi)^{D-1}\delta(\boldsymbol{p}-\boldsymbol{k}) \frac{e^{ik_0\tau}}{m^2 + k_0^2 + \boldsymbol{k}^2} = \\ &= (2\pi)^{D-1}\delta(\boldsymbol{q}-\boldsymbol{p}) \int \frac{\mathrm{d}k_0}{(2\pi)} \frac{e^{ik_0\tau}}{m^2 + k_0^2 + \boldsymbol{p}^2} = (2\pi)^{D-1}\delta(\boldsymbol{p}-\boldsymbol{q}) \frac{e^{-E_{\boldsymbol{p}}\tau}}{2E_{\boldsymbol{p}}} \ , \end{aligned} \tag{14.1.16}$$

where the last integral has been computed using the residue theorem, closing the integral in the upper half-plane since $\tau > 0$.

On the lattice the integral entering the spatial Fourier transform at fixed time becomes a sum on the lattice points of a given time-slice, and only the vectors of the first Brillouin zone are legitimate momenta. On a finite lattice the spatial δ function is smoothed as

$$(2\pi)^{D-1}\delta(\boldsymbol{p}-\boldsymbol{q}) \to \delta_{\boldsymbol{p},\boldsymbol{q}} \hat{V}_s \ , \quad \hat{V}_s = \prod_{\mu>0} N_\mu \ , \tag{14.1.17}$$

where N_μ is the lattice extent in the μ-th direction (time corresponds to $\mu = 0$), and $\hat{V}_s$ is the number of sites in any lattice time-slice. For this reason the momentum projected operators are usually normalized on the lattice as follows

$$O_{\boldsymbol{p}}(n_0) = \frac{1}{\sqrt{\hat{V}_s}} \sum_{\boldsymbol{n} \in\, n_0 \,\mathrm{t.s.}} e^{i\boldsymbol{p}\cdot\boldsymbol{n}} O(n_0, \boldsymbol{n}) \ , \tag{14.1.18}$$

where the sum extends on all the points of the n_0-th time-slice of the lattice, and the $(D-1)$-dimensional vector $\boldsymbol{n}$ denotes the spatial components of each of these points. The correlators of these fields are often called "wall-wall" correlators, since $O_{\boldsymbol{p}}(n_0)$ is computed by summing on the wall at fixed n_0.

We have seen that momentum projected operators remove the almost degenerate contributions of unwanted single particle states, however multiparticle contributions cannot be removed in this way: for example, in the free case the state $a^\dagger_{\boldsymbol{k}-\boldsymbol{p}} a^\dagger_{\boldsymbol{p}}|0\rangle$ has momentum $\boldsymbol{k}$ but energy

$$E^{(2p)}_{\boldsymbol{k}-\boldsymbol{p},\boldsymbol{p}} = \sqrt{m^2 + (\boldsymbol{k}-\boldsymbol{p})^2} + \sqrt{m^2 + \boldsymbol{p}^2} \,, \tag{14.1.19}$$

depending on the value of $\boldsymbol{p}$. In free field theories these states can not be reached from the ground state by using just one insertion of the field operator, but this is not the case in interacting quantum field theories (see, e. g., [46] §6.9 or [86] §10.7 for the Källen-Lehmann representation of the interacting propagator).

Note that this problem is not particularly serious if we are interested just in studying the mass-gap, since the two particle states have energy $\geq 2m$, hence a finite energy gap separates them from the single particle zero-momentum states of energy m. In QM an analogous situation is encountered when studying the bound states of a potential which also supports scattering states. Two particle states are instead problematic if we are interested in studying single particle states with $\boldsymbol{p}^2 \geq 3m^2$ (e. g. to verify the form of the dispersion relation), since in this case we need to study energies above the two-particle threshold, and almost degenerate two-particle states contribute to the matrix elements.

For the study of single particle states below the two-particle threshold, all the techniques that have been introduced in Sec. 10 to study the large time behavior of two point functions in QM can be used to study the large time behavior of correlation function of fixed momentum operators.

14.2 How to perform the continuum limit

When discussing the discretization of the free scalar field in Sec. 13.2 we noted that the continuum limit is approached as $\hat{m} \to 0$. This is quite trivial, since in the free case we know from the beginning the mass value m, and $\hat{m} = am$. Moreover, as in QM, the approach to the continuum limit can be seen as the approach to a critical point: the large time behavior of the appropriate correlator is proportional to $e^{-m\tau}$, which on the lattice becomes $e^{-\hat{m}n_0}$, where $n_0 = \tau/a$ is the euclidean time in lattice units (i. e. the number of lattice slides between the two walls of the wall-wall correlator). In the limit $a \to 0$ at fixed m, we have $\hat{m} \to 0$ and thus the lattice correlation length $\hat{\xi} = 1/\hat{m}$ diverges. Let us explicitly note that the lattice size has to be increased while approaching the continuum limit: if we are interested in infinte volume quantities we have to use $N_\mu/\hat{\xi} = L_\mu m \gg 1$ (N_μ is the number of sites in the μ-th direction, $L_\mu = aN_\mu$), but even if we are interested in finite volume QFT we have to keep $L_\mu m$ fixed, which means that $N_\mu \to \infty$ as $\hat{m} \to 0$ (this limit corresponds to the FSS limit of Sec. 5.4).

The same conclusions can be reached in an interacting theory: by measuring the correlation function of appropriate (momentum projected) interpolating operator we can estimate the dimensionless mass $\hat{m}$ (i. e. the mass in lattice units) of a given state, which can for example be the mass in lattice units of a glueball state with given spin, parity and charge conjugation properties in Yang-Mills theory. By changing the simulation parameters we can change the value of $\hat{m}$, and in particular we have to find a set of parameters for which $\hat{m} \to 0$, which correspond to a critical point of the lattice model. We then interpret $\hat{m} \to 0$ as $a \to 0$ at fixed physical mass. With respect to the free case the difficulty is that $\hat{m}$ is not a parameter that can be directly tuned, since it can only be estimated a posteriori, after the simulation. Note that this is the case also if a mass parameter $\hat{m}_0$ enters the lattice action, as for the interacting scalar field: the parameter $\hat{m}_0$ is the equivalent of the bare mass in continuum QFT, which is related in a nontrivial way to the physical mass.

The procedure to extract physical information form lattice simulations is thus the following: we measure the mass in lattice unit of a given state, let it be $\hat{m}_1$, and perform several simulations to approach the limit $\hat{m}_1 \to 0$. In the meantime we also measure the mass in lattice unit of a second state, let it be $\hat{m}_2$. If $\hat{m}_2/\hat{m}_1 \to \alpha$, with $0 < \alpha < \infty$, when $\hat{m}_1 \to 0$ then we predict that in the continuum limit $m_2/m_1 = \alpha$. It should be clear that in this way we can only predict

ratio of physical masses. In particular, a standard procedure is to fix the value of m_1 to its known physical value (obviously this can be done only when simulating a theory with a direct experimental counterpart), measure $\hat{m}_1$ in a simulation, and extract the lattice spacing $a = \hat{m}_1/m_1$ in physical units. This procedure is usually called scale setting, and can be interpreted as a nonperturbative renormalization of the lattice regularized QFT. Once the value in physical units of a is known we can convert all the lattice masses to physical masses. This is obviously equivalent to compute mass ratios with respect to a given reference mass.

When several parameters enter the theory, like the quark masses, several lattice quantities (e. g. meson masses) have to be measured to set all the parameters, and the continuum limit has to be performed keeping constant the ratios of these masses. This requirement identifies the so called "lines of constant physics". If, for example, we study QCD in the isospin symmetric limit, we have to fix the lattice spacing and the lattice masses $\hat{m}_u = \hat{m}_d$ and $\hat{m}_s$ of the up, down and strange quarks. The lattice spacing a is typically determined by using observables related to the static potential between color sources, and once a has been fixed, the quantities $\hat{m}_u$ and $\hat{m}_s$ entering the action have to be tuned to reproduce, e. g., the physical values of the π^+ and K^+ masses. Once these 3 numbers (lattice spacing and bare quark masses) have been fixed, we can predict the mass of the other mesons or baryons. The whole procedure has then to be repeated by decreasing the lattice spacing, in order to extract the continuum limit.

Let us note explicitly that we have so far implicitly assumed that

- the continuum limit of a given lattice model exists,
- we know which QFT emerges when approaching the continuum limit of a give lattice model,

but life is not always that easy. Let us assume that we are interested in studying the nonperturbative properties of a specific QFT. We have first of all to discretize the Euclidean action of this QFT, then to numerically simulate it, looking for critical points. Several possibilities can happen:

1. no critical point is found,
2. a single critical point is found,
3. several critical points are found.

In case 1) we can not obtain continuum physical results. This can be due to a pathology of the discretization: the QFT is well defined, but the lattice model is not "close enough" to it; more precisely, we are not in the attraction basin of the RG fixed point associated to the QFT we are interested in (and in fact of any QFT at all). But this could also be due to a pathology of the QFT, which is not well defined beyond perturbation theory. In case 2) we can define a continuum limit, but we have to understand whether the emerging QFT is really the one we are interested in, since it can happen that we are in the attraction basin of a different QFT. This can be done by comparing the numerical results with some nonpertubative predictions available in particular limits (e. g. large-N results), or by comparing directly with experimental data, when they are available. In case 3) we can define several continuum limits, and we have to understand which (if any) is the one corresponding to the QFT we are interested in. Some nontrivial examples of the problems encountered when discretizing a QFT are discussed in [87], for the case of three dimensional multiflavor scalar QED.

In free field theories the continuum limit is approached with $O(a^2)$ corrections, just like the case of quantum mechanics. This can be understood by using the same strategy used in QM (which is viable since no renormalizations are present), or by directly checking the behavior of the lattice propagator (see Sec. 13.2). For generic interacting field theories the approach to the continuum is governed by a nontrivial exponent, just like the finite size scaling correction in Sec. 5.4, since the continuum limit correspond to a continuus phase transition. We recover the scaling $O(a^2)$ (up to possible logarithmic corrections) in asymptotically free theories, like four dimensional Yang-Mills theory and QCD.

Chapter 15

MCMC in quantum field theory: thermodynamics

Since the lattices that will be used in the following are typically cubic, with the only possible exception of the temporal direction (related to the inverse temperature, see Sec. 13.1), we introduce the notation N_t to denote the number of lattice sites in the temporal direction (dentoted by $\mu = 0$), while N_s denotes the number of lattice sites in any of the spatial directions (denoted by $\mu > 0$).

15.1 Anisotropic discretization

Let us start by discussing the case of the free scalar field, we will later comment on the interacting case, and the changes that are needed to study it. The internal energy can be computed from the partition function using the relation

$$U = -\frac{\partial}{\partial \beta} \log Z \ , \tag{15.1.1}$$

where the spatial volume V_s and other physical quantities (like, e. g., the masses) have to be kept fixed while taking the derivative. Since $\beta = aN_t = N_t \hat{m}/m$ and m has to be kept fixed, we could think of using

$$\frac{\partial}{\partial \beta} = \frac{\partial \hat{m}}{\partial \beta} \frac{\partial}{\partial \hat{m}} = \frac{m}{N_t} \frac{\partial}{\partial \hat{m}} \ . \tag{15.1.2}$$

Since $\hat{m}$ directly enters the lattice Euclidean action in Eq. (13.2.3) it would then be straightforward to estimate U. This is however not possible: $\hat{m}$ is related to the lattice spacing a, and when changing the value of $\hat{m}$ (and hence of a) we are changing both the temperature and the volume.

To correct for this fact we can introduce an anisotropic discretization, by using the lattice spacing a_t along the temporal direction and the lattice spacing a along the spatial direction. The ratio $\xi = a_t/a$ quantifies the anisotropy, and clearly $\beta = N_t a_t = \xi N_t a$. Using the anisotropic discretization it is immediate to see that

$$\int \mathrm{d}^D x \to \sum_{\boldsymbol{n}} a_t a^{D-1} = \sum_{\boldsymbol{n}} \xi a^D \ , \quad \partial_t f \to \frac{f_{\boldsymbol{n}+\hat{0}} - f_{\boldsymbol{n}}}{a_t} = \frac{1}{\xi} \frac{f_{\boldsymbol{n}+\hat{0}} - f_{\boldsymbol{n}}}{a} \ , \tag{15.1.3}$$

hence the lattice action can be written in the form

$$S_L = \sum_{\boldsymbol{n}} \frac{1}{2} \left\{ \xi \hat{m}^2 \hat{\phi}_{\boldsymbol{n}}^2 + \xi \sum_{\mu>0} \left(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}} - \hat{\phi}_{\boldsymbol{n}} \right)^2 + \frac{1}{\xi} \left(\hat{\phi}_{\boldsymbol{n}+\hat{0}} - \hat{\phi}_{\boldsymbol{n}} \right)^2 \right\} \ . \tag{15.1.4}$$

We can now use

$$\frac{\partial}{\partial \beta} = \frac{\partial \xi}{\partial \beta} \frac{\partial}{\partial \xi} = \frac{1}{a N_t} \frac{\partial}{\partial \xi} \tag{15.1.5}$$

to compute the internal energy, obtaining

$$U = \frac{1}{2aN_t}\left\langle \sum_{\boldsymbol{n}} \left(\hat{m}^2 \hat{\phi}^2_{\boldsymbol{n}} + \sum_{\mu>0} \left(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}} - \hat{\phi}_{\boldsymbol{n}} \right)^2 - \frac{1}{\xi^2} \left(\hat{\phi}_{\boldsymbol{n}+\hat{0}} - \hat{\phi}_{\boldsymbol{n}} \right)^2 \right) \right\rangle_{\xi} , \tag{15.1.6}$$

where we used $\langle\ \rangle_\xi$ to denote the average value computed by using the anisotropic action. Note however that once the correct expression has been obtained, we can put $\xi = 1$ and use the isotropic action to generate the configurations. From here on we thus assume $\xi = 1$.

It is more convenient to compute, instead of the internal energy U, the internal energy density $\varepsilon = U/V_s$ normalized by T^D, which is a dimensionless quantity. Using

$$\frac{1}{T^D V_s} = \frac{N_t^D a^D}{N_s^{D-1} a^{D-1}} = \frac{N_t^D a}{N_s^{D-1}} \tag{15.1.7}$$

we can rewrite ε/T^D in the form

$$\frac{\varepsilon}{T^D} = \frac{N_t^D}{2} \langle O_1 + O_2 - O_3 \rangle \ , \tag{15.1.8}$$

where we introduced the definitions

$$\begin{gathered} O_1 = \frac{1}{N_t N_s^{D-1}} \sum_{\boldsymbol{n}} \hat{m}^2 \hat{\phi}^2_{\boldsymbol{n}} \ , \quad O_2 = \frac{1}{N_t N_s^{D-1}} \sum_{\boldsymbol{n}} \sum_{\mu>0} \left(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}} - \hat{\phi}_{\boldsymbol{n}} \right)^2 , \\ O_3 = \frac{1}{N_t N_s^{D-1}} \sum_{\boldsymbol{n}} \left(\hat{\phi}_{\boldsymbol{n}+\hat{0}} - \hat{\phi}_{\boldsymbol{n}} \right)^2 . \end{gathered} \tag{15.1.9}$$

Note that O_1, O_2 and O_3 are just lattice averages of local operators, since $N_t N_s^{D-1}$ is the number of lattices sites, and the "temporal" term O_3 enters ε/T^D with a minus sign (see Eq. (15.1.8)), which is the analogous in QFT of what happened in Chap. 9 to the kinetic term.

This result is still not the end of the story: the average value $\langle O_1 + O_2 - O_3 \rangle$ in Eq. (15.1.8) generates a divergence in the continuum limit, and an additive contribution is needed to cancel this divergence, exactly as happened in QM. For the free scalar it would be possible to subtract analytically this divergence, which originates from the integrals over momenta performed in Sec. 13.1, however this is not possible in interacting theories. To get a finite result in the general case we have to take the difference between two values of $\langle O_1 + O_2 - O_3 \rangle$, computed for two different temperatures. Ultraviolet divergences in QFT are indeed independent of the temperature (see [88] §11). It is fundamental to stress that, for this subtraction to be effective, we have to change the temperature keeping the lattice spacing (and hence $\hat{m}$ in the free case, or the bare couplings in the interacting case) constant: we thus must change the temperature by varying the dimensionless lattice size (i. e. the number of sites) along the temporal direction.

In practice, it is convenient to use $T \approx 0$ simulations to perform this subtraction, in such a way that the renormalized energy density is normalized to zero at vanishing temperature. Since the temperature is the inverse of the lattice extent (in physical units) along the temporal direction, to perform the $T \approx 0$ subtraction we have to use a lattice with

$$\frac{m}{T} = \overline{N}_t \hat{m} \gg 1 \ . \tag{15.1.10}$$

The final expression for the renormalized energy density is thus

$$\left.\frac{\varepsilon}{T^D}\right|_R = \frac{1}{2} N_t^D \left(\langle O_1 + O_2 - O_3 \rangle_{N_t} - \langle O_1 + O_2 - O_3 \rangle_{\overline{N}_t} \right) , \tag{15.1.11}$$

where the subscripts in $\langle\ \rangle_{N_t}$ and $\langle\ \rangle_{\overline{N}_t}$ denote the dimensionless temporal extent of the lattice used to estimate these average values.

The method discussed in this section can be applied also to interacting theories: everything goes on exactly as in the free case, but for a significant detail. We have to distinguish the bare anisotropy ξ_B (the one entering the lattice action) from the physical anisotropy ξ, with the relation $\partial\xi/\partial\beta = 1/(aN_t)$ which is valid for the physical ξ. In practice, we thus need to perform simulations specifically targeted at determining $\partial\xi_B/\partial\xi|_{\xi_B=1}$ for each value of the lattice spacing used, see, e. g., [89] for an early reference. In the work [90], where this method of estimating thermodynamic observables has been introduced, perturbation theory was used to evaluate $\partial\xi_B/\partial\xi|_{\xi_B=1}$.

15.2 Thermodynamic integration

We have discussed in the previous section that to directly estimate the internal energy we have introduce an anisotropic coupling, since otherwise it is not possible to rewrite the derivative with respect to T as a derivative with respect to the temporal lattice spacing.

Let us now see what happens if we instead consider the derivative of $\log Z$ with respect to the lattice spacing using the isotropic discretization:

$$-\frac{\partial}{\partial a}\log Z = -\left.\frac{\partial \log Z}{\partial \beta}\right|_{V_s}\frac{\partial\beta}{\partial a} - \left.\frac{\partial \log Z}{\partial V_s}\right|_{\beta}\frac{\partial V_s}{\partial a} \ . \tag{15.2.1}$$

In the first term we recognize the internal energy Eq. (15.1.1), while to rewrite the second term we have to use the fact that $\log Z = -\beta F$, where F is the free energy, and

$$\mathrm{d}F = -S\mathrm{d}T - P\mathrm{d}V_s \ , \tag{15.2.2}$$

where S and P are the entropy and the pressure, respectively. We thus have

$$-\left.\frac{\partial \log Z}{\partial V_s}\right|_{\beta} = -\beta P \ . \tag{15.2.3}$$

Using these relations, together with $\beta = aN_t$ and $V_s = a^{D-1}N_s^{D-1}$, we get

$$\begin{aligned} -\frac{\partial}{\partial a}\log Z &= UN_t - \beta P(D-1)N_s^{D-1}a^{D-2} = \\ &= N_t\Big(U - (D-1)PV_s\Big) = N_tV_s\Big(\varepsilon - (D-1)P\Big) \ , \end{aligned} \tag{15.2.4}$$

where we introduced the internal energy density $\varepsilon = U/V_s$. The quantity $\varepsilon - (D-1)P$ is the trace of the energy-momentum tensor and it is sometimes improperly called trace anomaly (it is really the trace anomaly only in massless theories). Multiplying the last equation by the lattice spacing we thus have (using $aN_t = \beta$)

$$-\frac{T}{V_s}a\frac{\partial}{\partial a}\log Z = \varepsilon - (D-1)P \ , \tag{15.2.5}$$

and dividing by T^D we get (using $1/(T^{D-1}V_s) = (N_t/N_s)^{D-1}$)

$$-a\frac{\partial}{\partial a}\left[\left(\frac{N_t}{N_s}\right)^{D-1}\log Z\right] = \frac{\varepsilon - (D-1)P}{T^D} \ . \tag{15.2.6}$$

As we will show in a moment the left hand side of this equation can be easily computed in lattice simulations, since in the free case we can use $a\partial/\partial a = \hat{m}\partial/\partial\hat{m}$ (at constant m), however the right hand side is typically not what one is interested in when studying thermodynamics.

To obtain a more standard thermodynamic observable it is convenient to rewrite the left hand side of the previous equation in a different way:

$$\left(\frac{N_t}{N_s}\right)^{D-1}\log Z = \frac{1}{T^{D-1}V_s}\log Z = -\frac{1}{T^DV_s}F = \frac{P}{T^D} \ , \tag{15.2.7}$$

indeed $\log Z = -\beta F$, and introducing the (intensive, i. e. V_s independent) free energy density $f = F/V_s$ we have

$$P = -\frac{\partial F}{\partial V_s} = -\frac{\partial}{\partial V_s}\Big(f\, V_s\Big) = -f \ . \tag{15.2.8}$$

Thus

$$-a\frac{\partial}{\partial a}\left[\left(\frac{N_t}{N_s}\right)^{D-1}\log Z\right] = -a\frac{\partial}{\partial a}\left(\frac{P}{T^D}\right) \ , \tag{15.2.9}$$

and since P/T^D is an intensive quantity we have (using $T = 1/(N_t a)$ hence $\partial T/\partial a = -T/a$)

$$-a\frac{\partial}{\partial a}\left(\frac{P}{T^D}\right) = -a\left(\frac{\partial T}{\partial a}\,\frac{\partial}{\partial T}\Big|_{V_s} + \frac{\partial V_s}{\partial a}\,\frac{\partial}{\partial V_s}\Big|_{T}\right)\frac{P}{T^D} = T\frac{\partial}{\partial T}\left(\frac{P}{T^D}\right) \ . \tag{15.2.10}$$

We thus finally get

$$-a\frac{\partial}{\partial a}\left[\left(\frac{N_t}{N_s}\right)^{D-1}\log Z\right] = T\frac{\partial}{\partial T}\left(\frac{P}{T^D}\right) \ , \tag{15.2.11}$$

which together with Eq. (15.2.6) gives as a byproduct the relation

$$T\frac{\partial}{\partial T}\left(\frac{P}{T^D}\right) = \frac{\varepsilon - (D-1)P}{T^D} \ . \tag{15.2.12}$$

The relation in Eq. (15.2.12) can obviously be proved also without using the lattice discretization. We have indeed

$$T\frac{\partial}{\partial T}\left(\frac{P}{T^D}\right) = -D\frac{P}{T^D} + \frac{1}{T^{D-1}}\frac{\partial P}{\partial T} \ , \tag{15.2.13}$$

moreover, see Eq. (15.2.2), $P = -\frac{\partial F}{\partial V_s}$ and $S = -\frac{\partial F}{\partial T}$, hence

$$\frac{\partial P}{\partial T} = -\frac{\partial}{\partial T}\frac{\partial F}{\partial V_s} = \frac{\partial S}{\partial V_s} = s \ , \tag{15.2.14}$$

where $s = S/V_s$ is the entropy density. From $F = U - TS$ we have $f = \varepsilon - Ts$ and we saw above that $f = -P$, thus $s = (\varepsilon + P)/T$, and

$$\frac{\partial P}{\partial T} = \frac{\varepsilon + P}{T} \ . \tag{15.2.15}$$

Using this relation in the first equation we finally have

$$T\frac{\partial}{\partial T}\left(\frac{P}{T^D}\right) = -D\frac{P}{T^D} + \frac{1}{T^{D-1}}\frac{\varepsilon + P}{T} = \frac{\varepsilon - (D-1)P}{T^D} \ . \tag{15.2.16}$$

Let us now discuss how we can use these relations to compute thermodynamic observables on the lattice. In the free case we can use, at fixed m, the relation $a\partial/\partial a = \hat{m}\partial/\partial\hat{m}$ to rewrite (using the lattice action in Eq. (13.2.3))

$$\begin{aligned} -a\frac{\partial}{\partial a}\left[\left(\frac{N_t}{N_s}\right)^{D-1}\log Z\right] &= -\left(\frac{N_t}{N_s}\right)^{D-1}\hat{m}\frac{\partial}{\partial\hat{m}}\log Z = \\ = \left(\frac{N_t}{N_s}\right)^{D-1}\hat{m}\left\langle\hat{m}\sum_{\boldsymbol{n}}\hat{\phi}^2_{\boldsymbol{n}}\right\rangle &= N_t^D\left\langle\frac{\hat{m}^2}{N_t N_s^{D-1}}\sum_{\boldsymbol{n}}\hat{\phi}^2_{\boldsymbol{n}}\right\rangle \ , \end{aligned} \tag{15.2.17}$$

which using the notation of the previous section is just $N_t^D\langle O_1\rangle$, see Eq. (15.1.9). Using this relation we can compute (but for the divergences to be discussed in a moment) the quantity $(\varepsilon - (D-1)P)/T^D$ for several temperatures, using Eq. (15.2.6). We can then exploit Eq. (15.2.12) to obtain the pressure by a numerical integration:

$$\frac{P(T)}{T^D} - \frac{P(T_0)}{T_0^D} = \int_{T_0}^{T}\mathrm{d}\mathcal{T}\,\frac{1}{\mathcal{T}}\frac{\varepsilon - (D-1)P}{\mathcal{T}^D} \ . \tag{15.2.18}$$

Also using this strategy we have to subtract divergent additive contributions from the average values of local operators, and it is convenient to use as reference temperature $T \approx 0$ in the subtractions. We thus define

$$\left.\frac{\varepsilon - (D-1)P}{T^D}\right|_R = N_t^D(\langle O_1\rangle_{N_t} - \langle O_1\rangle_{\overline{N}_t}) \ , \tag{15.2.19}$$

where $\overline{N}_t\hat{m} \gg 1$, and $\langle\ \rangle_X$ denotes average values computed by using a lattice with X sites in the temporal direction. We thus finally obtain

$$\left.\frac{P(T)}{T^D}\right|_R = \int_0^T \mathrm{d}\mathcal{T}\,\frac{1}{\mathcal{T}}\,\left.\frac{\varepsilon - (D-1)P}{\mathcal{T}^D}\right|_R \ . \tag{15.2.20}$$

The form of this equations should make self-evident the origin of the name thermodynamic integration. Since the physical mass m is always constant, this expression can be written in the equivalent form

$$\left.\frac{P(T)}{T^D}\right|_R = \int_0^{T/m} \mathrm{d}(\mathcal{T}/m)\,\frac{m}{\mathcal{T}}\,\left.\frac{\varepsilon - (D-1)P}{\mathcal{T}^D}\right|_R \ , \tag{15.2.21}$$

which is particularly convenient from the numerical point of view in the free case, since $T/m = 1/(N_t\hat{m})$ and $\hat{m}$ is an external parameter. Once the quantities $(\varepsilon-(D-1)P)/T^D$ and P/T^D have been computed, it is then obviously possible to estimate ε/T^D.

This method can be applied also to interacting theories [91], and the only difference is in the point where we traded $a\frac{\partial}{\partial a}$ with $\hat{m}\frac{\partial}{\partial\hat{m}}$. In an interacting theory $\hat{m}$ is not a free parameter, so it is convenient to use instead the bare coupling of the lattice theory (that we denote by γ_B) in the chain rule:

$$a\frac{\partial}{\partial a} = a\frac{\partial\gamma_B}{\partial a}\frac{\partial}{\partial\gamma_B} \ . \tag{15.2.22}$$

To use this expression (and to rewrite $\mathrm{d}\mathcal{T}$ in the integral determining the pressure) we need to know the dependence of the lattice spacing on the bare coupling γ_B. This relation is needed also for scale setting, see Sec. 14.2, and it is typically well studied, although it obviously depends on the discretization details.

15.3 Continuum results for the free scalar case

We are now going to derive continuum results for thermodynamic quantities in the free scalar case, which is the only case in which computations can be performed in almost closed form.

The starting point is the partition function, which we write as

$$Z = \mathcal{N}\int_{\phi(0,\boldsymbol{x})=\phi(\beta,\boldsymbol{x})}[\mathcal{D}\phi]\exp\left(-\frac{1}{2}\int_0^\beta \mathrm{d}t\int \mathrm{d}^{D-1}x\,\phi(-\boldsymbol{\nabla}^2+m^2)\phi\right) \ , \tag{15.3.1}$$

where we explicitly show the proportionality factor $\mathcal{N}$ that in Sec. 13.1 was hidden in the definition of the integration measure. Since Z is dimensionless and ϕ has dimension $[\phi] = \frac{D-2}{2}$, the proportionality factor is dimenionfull, and this is the reason for exposing it in the present discussion. In order to make the proportionality factor dimensionless, let us rescale the variables using the inverse temperature β:

$$\hat{x} = x/\beta \ , \quad \hat{m} = m\beta \ , \quad \hat{t} = t/\beta \ , \quad \hat{\phi} = \phi\beta^{\frac{D-2}{2}} \ . \tag{15.3.2}$$

Using these dimensionless variables we have

$$Z = \hat{\mathcal{N}}\int_{\hat{\phi}(0,\hat{\boldsymbol{x}})=\hat{\phi}(1,\hat{\boldsymbol{x}})}[\mathcal{D}\hat{\phi}]\exp\left(-\frac{1}{2}\int_0^1 \mathrm{d}\hat{t}\int \mathrm{d}^{D-1}\hat{x}\,\hat{\phi}(-\hat{\boldsymbol{\nabla}}^2+\hat{m}^2)\hat{\phi}\right) \ , \tag{15.3.3}$$

where $\hat{\mathcal{N}}$ is now dimensionless (and from Sec. 13.1 it follows that it is a function just of $\beta/a = 1/N_t$).

The eigenfunctions of the differential operator $-\hat{\boldsymbol{\nabla}}^2 + \hat{m}^2$, with periodic b. c. along all directions, are

$$\exp\left(2\pi i n \hat{t}\right) \exp\left(i \frac{2\pi}{\hat{L}} \boldsymbol{h} \cdot \hat{\boldsymbol{x}}\right) \ , \quad \hat{L} = L/\beta \ , \quad n, h_j \in \mathbb{Z} \ , \tag{15.3.4}$$

and the corresponding eigenvalues are

$$(2\pi n)^2 + (2\pi)^2 \frac{\boldsymbol{h}^2}{\hat{L}^2} + \hat{m}^2 \ . \tag{15.3.5}$$

We formally have

$$Z \propto \left[\det\left(-\hat{\boldsymbol{\nabla}}^2 + \hat{m}^2\right)\right]^{-1/2} \ , \tag{15.3.6}$$

where the proportionality factor is dimensionless, hence (going back to dimensionfull quantities)

$$\log Z = -\frac{1}{2} \sum_n \sum_{\boldsymbol{h}} \log \left\{ (2\pi n)^2 + \left(\frac{2\pi}{L}\right)^2 \beta^2 \boldsymbol{h}^2 + \beta^2 m^2 \right\} + \text{const} \ . \tag{15.3.7}$$

The "+const" term will be neglected in the following since it is independent of β, hence it does not change the internal energy $U = -\frac{\partial}{\partial \beta} \log Z$. If we introduce the notations

$$\boldsymbol{k} = \frac{2\pi}{L} \boldsymbol{h} \ , \quad E(\boldsymbol{k}) = \sqrt{m^2 + \boldsymbol{k}^2} \ , \tag{15.3.8}$$

we can thus write

$$\log Z = -\frac{1}{2} \sum_n \sum_{\boldsymbol{k}} \log \left\{ (2\pi n)^2 + \beta^2 E^2(\boldsymbol{k}) \right\} \ . \tag{15.3.9}$$

Following [92] §2.3, we now rewrite the previous expression using

$$\log\left[(2\pi n)^2 + \beta^2 E^2\right] = \int_1^{\beta^2 E^2} \frac{\mathrm{d}\theta^2}{\theta^2 + (2\pi n)^2} + \log[1 + (2\pi n)^2] \ , \tag{15.3.10}$$

and the identity

$$\sum_{n=-\infty}^{+\infty} \frac{1}{n^2 + (\theta/2\pi)^2} = \frac{2\pi^2}{\theta} \left(1 + \frac{2}{e^\theta - 1}\right) \ . \tag{15.3.11}$$

To prove the previous identity we can use the Poisson summation formula (see Chap. 11)

$$\sum_{n=-\infty}^{+\infty} f(n) = \sum_{k=-\infty}^{+\infty} e^{-2\pi i k x} f(x) \mathrm{d}x \tag{15.3.12}$$

with $f(x) = 1/[x^2 + (\theta/2\pi)^2]$. If we define

$$I_k = \int_{-\infty}^{+\infty} \frac{1}{x^2 + (\theta/2\pi)^2} e^{-2\pi i k x} \mathrm{d}x \ , \tag{15.3.13}$$

for $k \geq 0$ we can close the integration contour in the lower half-plane, and use the residue theorem with the pole in $-i\theta/(2\pi)$, hence

$$I_{k \geq 0} = -2\pi i \frac{1}{-2i\theta/(2\pi)} e^{-\theta k} = \frac{2\pi}{\theta} e^{-\theta k} \ . \tag{15.3.14}$$

In an analogous way we obtain, for $k < 0$, the expression

$$I_{k<0} = \frac{2\pi}{\theta} e^{\theta k} \ . \tag{15.3.15}$$

We thus have

$$\begin{aligned} \sum_{k=-\infty}^{+\infty} I_k &= \frac{2\pi^2}{\theta} + \frac{2\pi^2}{\theta} \sum_{k=1}^{\infty} e^{-k\theta} + \frac{2\pi^2}{\theta} \sum_{k=-\infty}^{-1} e^{k\theta} = \frac{4\pi^2}{\theta} \sum_{k=0}^{\infty} e^{-k\theta} - \frac{2\pi^2}{\theta} = \\ &= \frac{4\pi^2}{\theta} \frac{1}{1 - e^{-\theta}} - \frac{2\pi^2}{\theta} = \frac{2\pi^2}{\theta} \left(\frac{2e^\theta}{e^\theta - 1} - 1\right) = \frac{2\pi^2}{\theta} \left(\frac{e^\theta + 1}{e^\theta - 1}\right) = \frac{2\pi^2}{\theta} \left(1 + \frac{2}{e^\theta - 1}\right) \ . \end{aligned} \tag{15.3.16}$$

The logarithm of the partition function can be written as

$$\begin{aligned}
\log Z &\overset{(1)}{=} -\frac{1}{2}\sum_n \sum_{\boldsymbol{k}} \int_1^{\beta^2 E^2(\boldsymbol{k})} \frac{\mathrm{d}\theta^2}{\theta^2 + (2\pi n)^2} = \\
&= -\frac{1}{2}\sum_{\boldsymbol{k}} \left(\frac{1}{2\pi}\right)^2 \int_1^{\beta^2 E^2(\boldsymbol{k})} \sum_n \frac{1}{n^2 + (\theta/2\pi)^2} \mathrm{d}\theta^2 = \\
&\overset{(2)}{=} -\sum_{\boldsymbol{k}} \left(\frac{1}{2\pi}\right)^2 \int_1^{\beta E(\boldsymbol{k})} \frac{2\pi^2}{\theta} \left(\frac{2}{e^\theta - 1} + 1\right) \theta \mathrm{d}\theta = \\
&= -\sum_{\boldsymbol{k}} \int_1^{\beta E(\boldsymbol{k})} \left(\frac{1}{2} + \frac{1}{e^\theta - 1}\right) \mathrm{d}\theta \overset{(3)}{=} -\sum_{\boldsymbol{k}} \left\{\frac{1}{2}\beta E(\boldsymbol{k}) + \log\left(1 - e^{-\beta E(\boldsymbol{k})}\right)\right\} ,
\end{aligned} \tag{15.3.17}$$

where in the step (1) we neglected a β-independent additive term, coming from the second term in the right hand side of Eq. (15.3.10), in step (2) we used Eq. (15.3.11), and in step (3) we used the fact that a primitive of $1/(e^\theta - 1)$ is $\log(1 - e^{-\theta})$ (and neglected further β-independent terms). In the large spatial volume limit we have

$$\sum_{\boldsymbol{k}} \to V_s \int \frac{\mathrm{d}^{D-1}k}{(2\pi)^{D-1}} , \tag{15.3.18}$$

hence

$$F(\beta) = V_s \int \frac{\mathrm{d}^{D-1}k}{(2\pi)^{D-1}} \left\{\frac{1}{2}E(\boldsymbol{k}) + \frac{1}{\beta}\log\left(1 - e^{-\beta E(\boldsymbol{k})}\right)\right\} , \tag{15.3.19}$$

From $\log Z$ we can also compute the internal energy, obtaining

$$U(\beta) = V_s \int \frac{\mathrm{d}^{D-1}k}{(2\pi)^{D-1}} \left\{\frac{1}{2}E(\boldsymbol{k}) + \frac{E(\boldsymbol{k})}{e^{\beta E(\boldsymbol{k})} - 1}\right\} . \tag{15.3.20}$$

The first term of F clearly generates a divergence, so we have to introduce a renormalized free energy, which is defined by subtracting the zero temperature divergent contribution:

$$F_R(\beta) = F(\beta) - F(\beta = \infty) = V_s \int \frac{\mathrm{d}^{D-1}k}{(2\pi)^{D-1}} \frac{1}{\beta} \log\left(1 - e^{-\beta E(\boldsymbol{k})}\right) . \tag{15.3.21}$$

In the same way we obtain for the internal energy

$$U_R(\beta) = U(\beta) - U(\beta = \infty) = V_s \int \frac{\mathrm{d}^{D-1}k}{(2\pi)^{D-1}} \frac{E(\boldsymbol{k})}{e^{\beta E(\boldsymbol{k})} - 1} . \tag{15.3.22}$$

From now on we will consider the particular case $D = 2$, and we have thus for the renormalized free energy density

$$f_R = \frac{1}{2\pi\beta} \int_{-\infty}^{+\infty} \mathrm{d}k \, \log\left(1 - e^{-\beta E(k)}\right) = \frac{1}{\pi\beta} \int_0^{+\infty} \mathrm{d}k \, \log\left(1 - e^{-\beta E(k)}\right) , \tag{15.3.23}$$

and for the renormalized internal energy density

$$\varepsilon_R = \frac{1}{\pi} \int_0^{+\infty} \mathrm{d}k \, \frac{E(k)}{e^{\beta E(k)} - 1} . \tag{15.3.24}$$

Note that $f_R < 0$, consistently with the relation $P = -f$ (see Sec. 15.2). To evaluate numerically these integrals, it is convenient to perform a change of variable in the integration, in order to

factorize the dependence on β and leave in the integral the dependence on the dimensionless quantity $\alpha = m\beta$. Using $z = \beta k$ we have

$$\begin{aligned} f_R(\beta) &= \frac{1}{\pi\beta^2}\int_0^\infty \mathrm{d}z \log\left(1 - e^{-\sqrt{z^2+\alpha^2}}\right) \\ \varepsilon_R(\beta) &= \frac{1}{\pi\beta^2}\int_0^\infty \mathrm{d}z \frac{\sqrt{z^2+\alpha^2}}{e^{\sqrt{z^2+\alpha^2}} - 1} \ , \end{aligned} \tag{15.3.25}$$

and for example for $T = m$ we find

$$\begin{aligned} f_R(T = m) &\simeq -0.2194658931\, T^2 \ , \\ \varepsilon_R(T = m) &\simeq 0.40612349888\, T^2 \ . \end{aligned} \tag{15.3.26}$$

In the high temperature limit $\alpha \ll 1$ these expressions can be simplified by approximating $\alpha \simeq 0$, hence

$$\begin{aligned} f_R &\simeq \frac{1}{\pi\beta^2}\int_0^\infty \mathrm{d}k\, \log(1 - e^{-k}) = -\frac{1}{\pi\beta^2}\sum_{n=1}^\infty \frac{1}{n}\int_0^\infty e^{-nk}\mathrm{d}k = \\ &= -\frac{1}{\pi\beta^2}\sum_{n=1}^\infty \frac{1}{n^2} = -\frac{\pi}{6}T^2 \ , \end{aligned} \tag{15.3.27}$$

where we used $\log(1-x) = -\sum_{n=1}^\infty x^n/n$ and the fact that the sum in the second line is equal to $\pi^2/6$.

This can be proved by applying the Parseval identity to the Fourier series of x on $(-\pi, \pi)$: since

$$\int_{-\pi}^{\pi} x \sin(nx) = -2\pi\frac{(-1)^n}{n} \ , \tag{15.3.28}$$

we have

$$x = \sum_{n=1}^\infty \frac{2(-1)^{n+1}\sqrt{\pi}}{n}\frac{\sin(nx)}{\sqrt{\pi}} \ , \quad \int_{-\pi}^{\pi} x^2\mathrm{d}x = \sum_{n=1}^\infty \left|\frac{2(-1)^{n+1}\sqrt{\pi}}{n}\right|^2 \ , \tag{15.3.29}$$

from which the desired result imediately follows.

Analogously, for the internal energy in the high temperature limit we have ($\alpha \simeq 0$)

$$\varepsilon_R \simeq \frac{1}{\pi\beta^2}\int_0^\infty \frac{z}{e^z - 1}\mathrm{d}z = \frac{1}{\pi\beta^2}\Gamma(2)\zeta(2) = \frac{\pi}{6}T^2 \tag{15.3.30}$$

We have, following [38] §58, the relations

$$\begin{aligned} \int_0^\infty \frac{z^{x-1}}{e^z - 1}\mathrm{d}z &= \int_0^\infty z^{x-1}\frac{e^{-z}}{1 - e^{-z}}\mathrm{d}z = \int_0^\infty z^{x-1}e^{-z}\sum_{n=0}^\infty e^{-nz}\mathrm{d}z = \sum_{n=0}^\infty \int_0^\infty z^{x-1}e^{-(n+1)z}\mathrm{d}z = \\ &= \sum_{n=0}^\infty \frac{1}{(n+1)^x}\int_0^\infty \xi^{x-1}e^{-\xi}\mathrm{d}\xi = \Gamma(x)\sum_{n=0}^\infty \frac{1}{(n+1)^x} = \Gamma(x)\sum_{n=1}^\infty \frac{1}{n^x} = \Gamma(x)\zeta(x) \ , \end{aligned} \tag{15.3.31}$$

where we introduced the Euler Γ and the Riemann ζ functions

$$\Gamma(x) = \int_0^\infty \xi^{x-1}e^{-\xi}\mathrm{d}\xi \ , \quad \zeta(x) = \sum_{n=1}^\infty \frac{1}{n^x} \ . \tag{15.3.32}$$

It is immediate to prove by induction that if $n \in \mathbb{N}$ then $\Gamma(n+1) = n!$, moreover $\zeta(2) = \pi^2/6$, as shown before.

To find asymptotic expansions in the low temperature regime $\alpha = m\beta \gg 1$ is significantly more complicated, and it is convenient to rewrite f_R and ε_R using known special functions. Let us start from f_R: using $\log(1-x) = -\sum_{n=1}^\infty x^n/n$ we have

$$f_R = \frac{1}{\pi\beta^2}\int_0^\infty \mathrm{d}z\, \log\left(1 - e^{-\sqrt{z^2+\alpha^2}}\right) = -\frac{1}{\pi\beta^2}\sum_{n=1}^\infty \frac{1}{n}\int_0^\infty e^{-n\sqrt{z^2+\alpha^2}}\mathrm{d}z \ , \tag{15.3.33}$$

and using the change of variable $z = \alpha\sinh x$ we get

$$\int_0^\infty e^{-n\sqrt{z^2+\alpha^2}}\mathrm{d}z = \alpha\int_0^\infty \cosh(x)e^{-\alpha n\cosh(x)}\mathrm{d}x = \alpha K_1(\alpha n) \ , \tag{15.3.34}$$

where we used the following identity for the modified Bessel functions of second kind (see [12] §9.6.24)

$$K_\nu(z) = \int_0^\infty \cosh(\nu t) e^{-z\cosh t} \mathrm{d}t \ . \tag{15.3.35}$$

We thus have

$$f_R = -\frac{m}{\pi\beta} \sum_{n=1}^{\infty} \frac{1}{n} K_1(\alpha n) \ , \tag{15.3.36}$$

and using the asymptotic expansion for large argument of the modified Bessel functions of second kind (see [12] §9.7.2)

$$K_\nu(z) \simeq \sqrt{\frac{\pi}{2z}} e^{-z} \ , \tag{15.3.37}$$

we get in the low tempertature regime $\alpha = m/T \gg 1$

$$f_R \simeq -\frac{\sqrt{m}\, T^{3/2}}{\sqrt{2\pi}} e^{-m/T} \ . \tag{15.3.38}$$

Using the asymptotic expansion for small argument of the modified Bessel functions of second kind (see [12] §9.6.8-9)

$$K_0(z) \simeq -\log z \ , \quad K_\nu(z) \simeq \frac{1}{2}\Gamma(\nu)\left(\frac{z}{2}\right)^{-\nu} \ , \quad \nu > 0 \ , \tag{15.3.39}$$

it is also simple to find the high-temperature expansion.

To obtain the low temperature behavior of the renormalized energy density is only slightly more complicated:

$$\begin{aligned} \varepsilon_R &= \frac{1}{\pi\beta^2} \int_0^\infty \mathrm{d}z \, \frac{\sqrt{z^2+\alpha^2}}{e^{\sqrt{z^2+\alpha^2}} - 1} = \frac{1}{\pi\beta^2} \int_0^\infty \mathrm{d}z \, \frac{\sqrt{z^2+\alpha^2}}{e^{\sqrt{z^2+\alpha^2}} \left(1 - e^{-\sqrt{z^2+\alpha^2}}\right)} = \\ &= \frac{1}{\pi\beta^2} \int_0^\infty \mathrm{d}z \, \frac{\sqrt{z^2+\alpha^2}}{e^{\sqrt{z^2+\alpha^2}}} \sum_{n=0}^{\infty} e^{-n\sqrt{z^2+\alpha^2}} = \frac{1}{\pi\beta^2} \sum_{n=1}^{\infty} \int_0^\infty \mathrm{d}z \, \sqrt{z^2+\alpha^2} e^{-n\sqrt{z^2+\alpha^2}} \end{aligned} \tag{15.3.40}$$

Using also in this case the change of variable $z = \alpha \sinh x$, and $\cosh^2 x = \frac{1}{2}\cosh(2x) + \frac{1}{2}$, we have

$$\begin{aligned} &\int_0^\infty \mathrm{d}z \, \sqrt{z^2+\alpha^2} e^{-n\sqrt{z^2+\alpha^2}} = \alpha^2 \int_0^\infty \mathrm{d}x \, \cosh^2(x) e^{-n\alpha\cosh x} = \\ &= \frac{\alpha^2}{2} \int_0^\infty \mathrm{d}x \left(1 + \cosh(2x)\right) e^{-n\alpha\cosh x} = \frac{\alpha^2}{2}\Big(K_0(n\alpha) + K_2(n\alpha)\Big) \ . \end{aligned} \tag{15.3.41}$$

We thus have

$$\varepsilon_R = \frac{m^2}{2\pi} \sum_{n=1}^{\infty} \Big(K_0(n\alpha) + K_2(n\alpha)\Big) \ , \tag{15.3.42}$$

and in the low temperature regime $\alpha = m/T \gg 1$

$$\varepsilon_R \simeq \frac{\sqrt{T}\, m^{3/2}}{\sqrt{2\pi}} e^{-m/T} \ . \tag{15.3.43}$$

Also in this case, using the asymptotic behavior of the modified Bessel functions of second kind for $\alpha \ll 1$, it is possible to obtain the high-temperature expansion of ε_R.

15.4 Numerical examples for the two dimensional free scalar field

In this section we present some numerical results for the thermodynamic of the two dimensional free scalar field, in order to show how the previously explained techniques work in practice.

In the case of the free scalar field with lattice action Eq. (13.2.3) we have 3 free parameters to set before starting the simulations: $\hat{m}$, N_t and N_s (in fact we will need also $\overline{N}_t$, the temporal extent to be used to perform the zero temperature subtraction). Since $\hat{m} = am$, m is the physical mass of the field, and physical correlators decay with typical length-scale $1/m$ (see Sec. 14.1), $\hat{m}$ has to be "small", in order for the lattice spacing to be much smaller than the typical length-scale of the system. This is however not the only constraint: we also need $1/m$ to be "small" with respect to the physical side of the lattice aN_s, in order not to have large finite volume effects. We thus need

$$\frac{1}{N_s} \ll \hat{m} \ll 1 \ . \tag{15.4.1}$$

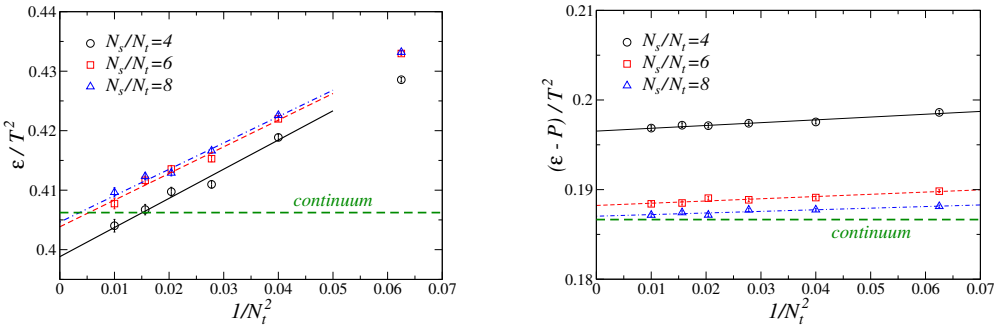

Figure 15.1: Results obtained for ε/T^2 and $(\varepsilon - P)/T^2$ at $T = m$, using the lattice action Eq. (13.2.3) and the numerical setup described in the main text. Data at fixed N_s/N_t have been extrapolated to the continuum limit ($N_t \to \infty$) by using a linear ansatz in $1/N_t^2$. The horizontal dashed green line denotes the continuum values computed by using Eq. (15.3.26).

Note that this is strictly true only at zero temperature, since at finite temperature we have screening, and the typical length-scale is not $1/m$, but something which scale as $1/T$ at high temperature (see e. g. [92] §6). However, when performing the renormalization discussed in Secs. (15.1)-(15.2), we also need zero temperature simulations, so the previous condition has anyway to be satisfied (this is one drawback of the choice of the $T \approx 0$ point to perform the subtractions). The value of the temperature in units of the mass m depends on the two variables $\hat{m}$ and N_t by (remember that $aN_t = 1/T$, see Sec. 13.1):

$$\frac{m}{T} = \hat{m} N_t \ . \tag{15.4.2}$$

We can thus change the ratio m/T or by changing the value $\hat{m}$ at fixed N_t (always paying attention to Eq. (15.4.1)), or by changing the value of N_t at fixed $\hat{m}$.

Let us consider for example the case of the temperature $T = m$. To impose this condition we have to fix $\hat{m} = 1/N_t$, and to approach the continuum limit we have increase the value of N_t (so that $\hat{m} \to 0$ at fixed m/T). For each value of N_t we have to chose a value of N_s large enough to be close to the termodynamic limit; in fact we have to extrapolate the large N_s limit. We thus consider

$$\frac{1}{\hat{m}} = N_t = 4, 5, 6, 7, 8, 10 \ , \tag{15.4.3}$$

and for each of these values we performed simulations with

$$N_s/N_t = 4, 6, 8 \ , \tag{15.4.4}$$

in such a way that $\hat{m} N_s = 4, 6, 8$, see Eq. (15.4.1). To perform the "zero" temperature subtraction, we use simulations adopting the same values $1/\hat{m} = 4, 5, 6, 7, 8, 10$, $\hat{m} N_s = 4, 6, 8$, and $\overline{N}_t = N_s$ (note that the convergence to the zero temperature limit is exponential in the ratio $m/T = \hat{m}\overline{N}_t$, see Sec. 15.3). For each of these cases we performed 5×10^7 updates (20% heatbath and 80% microcanonical), measuring O_1, O_2 and O_3 (see Eq. (15.1.9)) after every update. Execution times for a single data point go from ≈ 2 minutes (for the 4×16 lattice) to ≈ 120 minutes (for the 80×80 lattice). Using the results of Sec. 15.1 we then computed ε/T^2, and using the results of Sec. 15.2 we computed $(\varepsilon - P)/T^2$.

Numerical results obtained using this setup are shown in Fig. (15.1). Linear fits in $1/N_t^2 \propto a^2$ have been performed for data corresponding to fixed values of N_s/N_t: in this way we are extrapolating to the continuum limit results obtained at fixed physical volume (fixed $mL = N_s \hat{m}$), which then have to be extrapolated to the infinite volume limit. It is clear that data are approaching the continuum values computed in Eq. (15.3.26) (remember that $P = -f_R$) when increasing the value of N_s/N_t (the approach to the infinite volume limit is typically exponentially fast in N_s/N_t).

It is interesting to see what happens if we use, instead of the discretization obtained by using the forward derivative Eq. (13.2.3), the one obtained by using the symmetric discretization of the

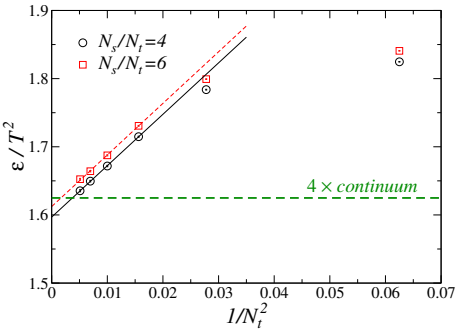

Figure 15.2: Results obtained for ε/T^2 and $(\varepsilon - P)/T^2$ at $T = m$, using the lattice action Eq. (13.2.22) and the numerical setup described in the main text. Data at fixed N_s/N_t have been extrapolated to the continuum limit ($N_t \to \infty$) by using a linear ansatz in $1/N_t^2$. The horizontal dashed green line denotes four times the continuum values computed by using Eq. (15.3.26).

derivative, see Eq. (13.2.22), which was previously shown to describe 2^D free scalar fields in the continuum limit. The expressions deduced in Sec. 15.1 can be easily adapted to this case, obtaining the final Eq. (15.1.8) but with

$$O_1 = \frac{1}{N_t N_s^{D-1}} \sum_{\boldsymbol{n}} \hat{m}^2 \hat{\phi}_{\boldsymbol{n}}^2 \ , \quad O_2 = \frac{1}{N_t N_s^{D-1}} \sum_{\boldsymbol{n}} \sum_{\mu>0} \frac{1}{4} \left(\hat{\phi}_{\boldsymbol{n}+\hat{\mu}} - \hat{\phi}_{\boldsymbol{n}-\hat{\mu}} \right)^2 \ ,$$
$$O_3 = \frac{1}{N_t N_s^{D-1}} \sum_{\boldsymbol{n}} \frac{1}{4} \left(\hat{\phi}_{\boldsymbol{n}+\hat{0}} - \hat{\phi}_{\boldsymbol{n}-\hat{0}} \right)^2 \ . \tag{15.4.5}$$

Data obtained for ε/T^2 by using $\frac{1}{\hat{m}} = N_t = 4, 6, 8, 10, 12, 14$ and $N_s/N_t = 4, 6$ are shown in Fig. (15.2). It is clear that the values of ε/T^2 obtained in this case are much larger than the corresponding ones we have seen when using the forward discretization (see Fig. (15.1)), and they seem to converge to four times the continuum result for a single scalar field, consistently with the analysis of Sec. 13.2. Notice also that, with respect to the forward discretization case, lattice artifacts are significantly larger, and their linear behavior in $1/N^2$ sets in for larger values of N_t. This is consistent with the analysis carried out in Sec. 13.2: the symmetric discretization is equivalent, in each of the 4 independent sublattices, to the forward discretization with a lattice spacing that is two times larger.

In order to study thermodynamical properties in an extended range of temperatures it is convenient to perform simulation at fixed N_t, changing the temperature by varying the parameter $\hat{m}$, alway keeping in mind Eq. (15.4.1). We now present data obtained using lattices with $N_s/N_t = 5$, $N_t = 4, 6, 8, 10$ and $\frac{2}{N_s} \lesssim \hat{m} \lesssim 1$, which taking into account $m/T = \hat{m} N_t$ means $\frac{2N_t}{N_s} \lesssim \frac{m}{T} \lesssim N_t$, i. e.

$$\frac{1}{N_t} \lesssim \frac{T}{m} \lesssim \frac{N_s}{2N_t} = 2.5 \ . \tag{15.4.6}$$

Note that to reach low temperatures we need large N_t values, while to reach large temperatures we need large values of N_s/N_t. Also in this case subtractions have performed by using lattices with $N_t = N_s$, and simulations with $N_t = 10$ and $N_s/N_t = 10$ have been used to check for finite volume effects. In all the cases we gathered a statistics of 5×10^7 updates of the whole lattice (20% heatbath and 80% microcanonical), with simulation times ranging from ≈ 2 minutes for the 4×16 lattices to ≈ 190 minutes for the 100×100 lattices.

Results obtained by using the anisotropic discretization method, i. e. Eq. (15.1.11), are displayed in Fig. (15.3): from the left panel we see that $N_t = 10$ is large enough to be close to the continuum, while from the right panel we see that there is only a very mild dependence on the volume size, which is more significant for larger values of T/m. Note the very slow convergence of ε/T^2 to the asymptotic limit $\pi/6$ computed in Sec. 15.3. This does not come as a surprise, since

from the exact formula for ε/T^2, written as a sum of modified Bessel functions of second kind, it is simple to see that logarithmic corrections to the leading asymptotic behavior are present.

In Fig. (15.3) we report the trace of the energy momentum tensor normalized by T^2, computed by using Eq. (15.2.19) (the systematic error induced by the choice of the numerical integration method has been verified to be negligible starting from second order methods). Also in this case the dependence on the lattice spacing is quite mild (left panel), but finite volume effects are much more significant than for ε_R/T^2, as can be seen from the right panel of Fig. (15.3). The trace of the energy momentum tensor has then been integrated using Eq. (15.2.21) to obtain P/T^2 and, by addition, ε/T^2. A comparison of the results obtained by using the anisotropic discretization and the thermodynamic integration methods is shown in Fig. (15.5): good agreement between the results of the two methods is found for $N_s/N_t = 10$ data.

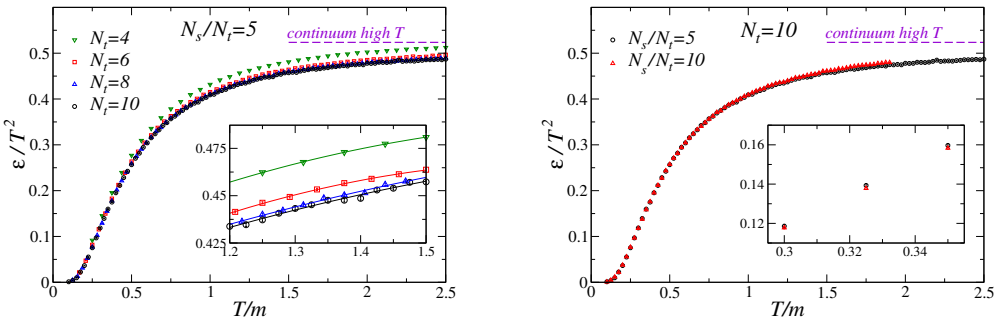

Figure 15.3: Behavior of ε/T^2 as a function of T/m computed using Eq. (15.1.11). (Left) Results for $N_s/N_t = 5$ and serveral values of N_t. The inset shows a zoom to better appreciate the convergence of the results to the continuum limit. (Right) comparison of data obtained by using $N_s/N_t = 5$ and $N_s/N_t = 10$ for $N_t = 10$. The agreement between the two data sets is very good for $T/m \lesssim 1$ (see also the inset), while deviations appear for larger values of the temperature.

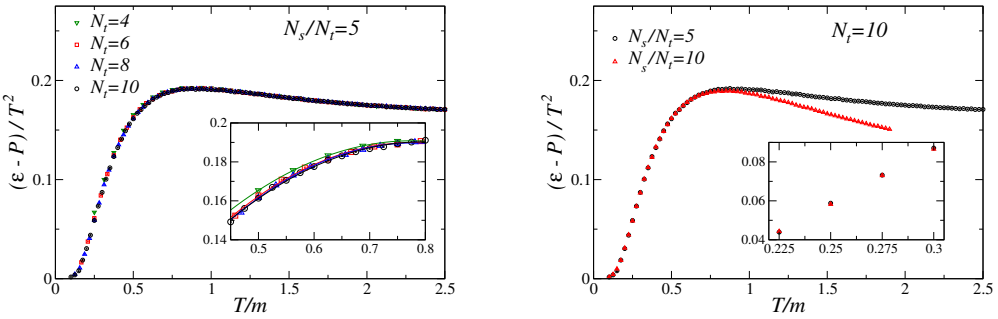

Figure 15.4: Behavior of $(\varepsilon - P)/T^2$ as a function of computed T/m using Eq. (15.2.19). (Left) Results for $N_s/N_t = 5$ and serveral values of N_t. The inset shows a zoom to better appreciate the convergence of the results to the continuum limit. (Right) comparison of data obtained by using $N_s/N_t = 5$ and $N_s/N_t = 10$ for $N_t = 10$. The agreement between the two data sets is very good for $T/m \lesssim 1$ (see also the inset), while deviations appear for larger values of the temperature.

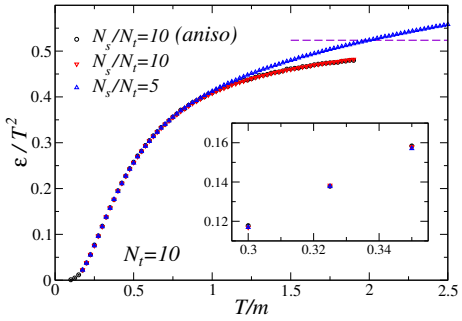

Figure 15.5: Behavior of ε/T^2 as a function of T/m. Comparison of the results obtained for $N_s/N_t = 5$ and $N_s/N_t = 10$ using the thermodynamic integration method (see Eqs. (15.2.19)-(15.2.21)), with the results obtained using Eq. (15.1.11) for $N_s/N_t = 10$. The two computation methods nicely agree with each other for $T/m \lesssim 2$ when $N_s/N_t = 10$.

Chapter 16

The Hybryd Monte Carlo algorithm

In this chapter we are going to introduce a new variant of the Monte Carlo method, the Hybrid Monte Carlo (HMC) algorithm, which can be used whenever we have to sample continuous variables. Despite its generality, this algorithm is really useful only when it is computationally difficult to perform local updates; in the case of lattice systems of physical interest this happens when the action is non-local. We thus start by discussing fermionic field theories, in order to clarify how non-local lattice actions can emerge, and then discuss the HMC algorithm in the simpler case of scalar field theories.

16.1 Why we need HMC: the fermionic case

In the path-integral formulation fermionic fields are associated to Grassmann variables (see, e. g. [46, 86, 93]), i. e. anticommuting variables, and this prevent us from using standard techniques to sample the fermionic fields. Nevertheless several interesting fermionic actions are quadratic in the fermionic fields, or can be rewritten as quadratic in the fermionic fields by introducing auxiliary scalar fields. In these cases the path-integration on the fermionic fields can be exactly carried out by using

$$\int [\mathcal{D}\psi \mathcal{D}\bar{\psi}] e^{\bar{\psi} M \psi} = \det M \ . \tag{16.1.1}$$

Note that, for the previous expression to be well defined, we are assuming to work in the context of the lattice regularized theory. We are however neglecting all the difficulties related to the lattice discretization of fermionic fields outlined in Sec. 13.2, since they are irrelevant for the problem we want to discuss in this section.

If we consider as an example the case of Quantum Chromodynamics (QCD) with two degenerate fermionic fields, the partition function is given by (up to irrelevant proportionality factor)

$$\begin{aligned} Z &= \int [\mathcal{D}A_\mu] \prod_{i=u,d} [\mathcal{D}\psi_i \mathcal{D}\bar{\psi}_i] \exp\left\{ -S_g[A] - \bar{\psi}_u D[A] \psi_u - \bar{\psi}_d D[A] \psi_d \right\} = \\ &= \int [\mathcal{D}A_\mu] \Big(\det D[A] \Big)^2 e^{-S_g[A]} = \int [\mathcal{D}A_\mu] \exp\left\{ -S_g[A] + \log\Big((\det D[A])^2 \Big) \right\} \ , \end{aligned} \tag{16.1.2}$$

where we denote by A_μ the gauge fields and by $S_g[A]$ the gauge action, whose specific form we do not need to known now. The important point to note is that, although $S_g[A]$ is a simple local functional of the gauge fields (as it will be shown Chap. 17), the integration of the fermion fields generates a very nonlocal action for the gauge fields. This makes all the sampling techniques that have been used so far extremely inefficient: even if we generate a trial configuration by changing

the value of the gauge field just in a single lattice site, to evaluate the acceptance probability we have to perform a computation which involves all the variables of the lattice. To generate a trial configuration by randomly changing the values of the gauge field on all the lattice sites does not help either, since the acceptance probability would typically be ridiculously small. This is the prototypical case in which the HMC is useful, since the aim of this algorithm is to generate a new trial configuration in which all variables change, but they do not change randomly: they change in such a way as to keep the acceptance probability reasonably close to one.

The form of the action in Eq. (16.1.2) is however not the one that is normally used in numerical simulations. It is customary to introduce the so-called pseudo-fermionic fields to rewrite Eq. (16.1.2) in a different form. Using the fact that

$$\int [\mathcal{D}\phi \mathcal{D}\phi^*] e^{-\phi^\dagger N \phi} \propto \frac{1}{\det N} \ , \tag{16.1.3}$$

where ϕ is a complex scalar field (the pseudo-fermionic field), we have indeed (assuming $\det D[A] \in \mathbb{R}$)

$$\begin{aligned} Z &= \int [\mathcal{D}A_\mu] \Big(\det D[A] \Big)^2 e^{-S_g[A]} = \\ &= \int [\mathcal{D}A_\mu][\mathcal{D}\phi\mathcal{D}\phi^*] \exp\left\{ -S_g[A] - \phi^\dagger (D^\dagger[A] D[A])^{-1} \phi \right\} \ . \end{aligned} \tag{16.1.4}$$

The action is still non local, due to the term $(D^\dagger D)^{-1}$, but this form is easier to sample: the field ϕ can indeed be generated (for fixed A_μ) using an heat-bath algorithm. If we define $R = (D^\dagger)^{-1}\phi$ we have (using $(D^\dagger D)^{-1} = D^{-1}(D^\dagger)^{-1}$)

$$\frac{e^{-\phi^\dagger (D^\dagger D)^{-1} \phi} [\mathcal{D}\phi\mathcal{D}\phi^*]}{\int e^{-\phi^\dagger (D^\dagger D)^{-1} \phi} [\mathcal{D}\phi\mathcal{D}\phi^*]} = \frac{e^{-R^\dagger R} [\mathcal{D}R\mathcal{D}R^*]}{\int e^{-R^\dagger R} [\mathcal{D}R\mathcal{D}R^*]} \ , \tag{16.1.5}$$

hence we can sample R using the Box-Muller algorithm (see Sec. 2.3), and then reconstruct $\phi = D^\dagger R$.

We still have the problem of sampling A_μ at fixed ϕ. The idea of the HMC algorithm is to add to the action a new term, written by using the conjugate momenta of A_μ, and to generate a trial configuration by numerically integrating the Hamiltonian equations of motion. In the computation of the force entering the equations of motion for the conjugate momenta we need to evaluate $\eta = (D^\dagger D)^{-1}\phi$, i. e. to solve a very large sparse (and typically not so well conditioned) linear system. This is typically done by using iterative Krylov solvers (see [94] §8.8 for a quick introduction, or [95] for many more details), and this is the main bottleneck in performing simulations with fermion fields, especially in the light mass limit. The algorithm we have just described is the so called Φ algorithm [96], which is still, but for some minor changes, the standard algorithm used to simulate QCD, see e. g. [94, 97] for more details.

When writing Eqs. (16.1.2)-(16.1.4) we have assumed $\det D[A] \in \mathbb{R}$, since otherwise it is not possible to use a Monte Carlo approach at all, the weight not being positive definite (note that when using an odd number of flavors we need the stronger requirement $\det D[A] > 0$.). This condition is however not satisfied at nonvanishing baryon density, see e. g. [98, 99], and this is the reason why we know so little of the QCD phase diagram at finite density.

When performing simulations with fermionic fields also the computation of observables can present some problems: most observables can be written in the form

$$\frac{\partial}{\partial \alpha} \log Z \ , \tag{16.1.6}$$

where α is some control parameter entering the fermion matrix, like, e. g., the fermion mass. To write explicitly these observables we can use the so called Jacobi's formula for the derivative of the determinant (we denote the derivative with respect to α by $'$)

$$(\det M)' = \det M \, \mathrm{tr}(M' M^{-1}) \ , \tag{16.1.7}$$

which can be proved for diagonalizable matrices and extended by continuity to the general case. We have indeed

$$(\det M)' = \lambda_1' \lambda_2 \cdots \lambda_N + \lambda_1 \lambda_2' \cdots \lambda_N + \cdots + \lambda_1 \lambda_2 \cdots \lambda_N' = (\det M) \left(\frac{\lambda_1'}{\lambda_1} + \cdots + \frac{\lambda_N'}{\lambda_N} \right) \ , \tag{16.1.8}$$

and if $M = U^{-1}DU$, with D a diagonal matrix, we have

$$\begin{aligned}\mathrm{tr}(M'M^{-1}) &= \mathrm{tr}\left[\Big((U^{-1})'DU + U^{-1}D'U + U^{-1}DU'\Big)U^{-1}D^{-1}U\right] = \\ &= \mathrm{tr}\Big[(U^{-1})'U + U^{-1}U'\Big] + \mathrm{tr}(D'D^{-1}) = \mathrm{tr}(D'D^{-1}) \ ,\end{aligned} \tag{16.1.9}$$

where in the last step we used $\mathrm{tr}[(U^{-1}U)'] = \mathrm{tr}(1') = 0$, moreover $\mathrm{tr}(D'D^{-1}) = \sum_i \lambda_i'/\lambda_i$.

To estimate some observables we thus need to evaluate $\mathrm{tr}(M'M^{-1})$, which is extremely demanding from the computational point of view. We can however use the following trick: if η_i (where $i = 1, \ldots, N$, and N is the size of M) are independent random variables such that

$$[\eta_i^* \eta_j] = \delta_{ij} \ , \tag{16.1.10}$$

where we denoted by $[\ ]$ the average with respect to the distribution of the η_i, then we can write

$$\mathrm{tr}(M'M^{-1}) = \sum_i (M'M^{-1})_{ii} = \sum_{ijk} [\eta_i^* (M')_{ij} M^{-1}_{jk} \eta_k] \ . \tag{16.1.11}$$

The average value can then be estimated by using the sample mean obtained by generating K random sets $\{\eta_i\}_{i=1,\ldots,N}$, and we once again just need to solve (large sparse) linear systems. These estimators are known as noisy estimators.

Since the sample average is an unbiased estimator of the true average, it is not necessary (although it can sometimes be computationally convenient) to use very large values of K, however some care is needed to avoid introducing biases in nonlinear observables: for example to estimate $\Big(\mathrm{tr}(M'M^{-1})\Big)^2$ we have to use

$$\Big(\mathrm{tr}(M'M^{-1})\Big)^2 = \sum_{ijk} [\eta_i^* (M')_{ij} M^{-1}_{jk} \eta_k] \ \sum_{ijk} [\zeta_i^* (M')_{ij} M^{-1}_{jk} \zeta_k] \ , \tag{16.1.12}$$

where η_i and ζ_i are independent random variables. Moreover it is convenient to use random variables taking values in ± 1, since in this way the error is minimized, see, e. g., [100] App. B.

16.2 The HMC algorithm for a single bosonic variable

Let us now discuss the details of the Hybrid Monte Carlo algorithm [101] considering for the sake of the simplicity the sampling of a single variable, since everything can be trivially generalized to more complicated cases. Our aim is thus to sample the pdf

$$P_S(q)\mathrm{d}q \propto e^{-S(q)}\mathrm{d}q \ . \tag{16.2.1}$$

The main idea of the HMC algorithm is to introduce the additional variable p, which is intepreted as the conjugate momentum of the variable q, thus building the "fake" Hamiltonian $H = \frac{1}{2}p^2 + S(q)$. Note that this operation is legitimate as far as we are interested in computing average values which depends just on q, since obviously

$$\langle f(q) \rangle = \frac{\int f(q) e^{-S(q)} \mathrm{d}q}{\int e^{-S(q)} \mathrm{d}q} = \frac{\int f(q) e^{-H} \mathrm{d}q\mathrm{d}p}{\int e^{-H} \mathrm{d}q\mathrm{d}p} \ . \tag{16.2.2}$$

In this way our configuration consists now of a couple of variables, q and p, and the process to generate a new configuration is the following:

1. generate the momentum with pdf $P_G(p) \propto e^{-p^2/2}$
2. solve the Hamiltonian equations of motion with initial values (q, p) for a fixed time t, obtaining $(q(t), p(t))$, and use this as trial configuration. This corresponds to select the trial configuration (q', p') with pdf

$$P_C\Big((q,p) \to (q',p')\Big) = \delta\Big((q',p') - (q(t),p(t))\Big) \tag{16.2.3}$$

3. accept the trial configuration with probability

$$P_A\Big((q,p) \to (q',p')\Big) = \min(1, e^{-\delta H}) \ , \quad \delta H = H(q',p') - H(q,p) \ . \tag{16.2.4}$$

The transition probability of going from q to q' is thus

$$P(q \to q') = \int \mathrm{d}p P_G(p) P_C\Big((q,p) \to (q',p')\Big) P_A\Big((q,p) \to (q',p')\Big) \ , \tag{16.2.5}$$

and the Makov chain generated by this transition probability is irreducible, due to point 1), whenever starting from q we can chose the momentum p in such a way as to reach any q' at time t. Aperiocity follows from the fact that we can select p in such a way that $q' = q$ (with the usual caveat concerning continuous pdf, which should be investigated in a more precise way). We are now going to show that this transition probability also satisfies the detailed balance

$$P_S(q) P(q \to q') \mathrm{d}q = P_S(q') P(q' \to q) \mathrm{d}q' \tag{16.2.6}$$

provided that

a. the evolution is reversible: in a fixed time t the configuration (q,p) evolves in (q',p') if and only if the configuration $(q',-p')$ evolves in $(q,-p)$ in a time t
b. the evolution preserves the measure of the phase space: $\mathrm{d}q\,\mathrm{d}p = \mathrm{d}q'\mathrm{d}p'$.

It is indeed simple to verify that

$$e^{-H(q,p)} \min(1, e^{-\delta H}) = e^{-H(q',p')} \min(e^{\delta H}, 1) \ , \tag{16.2.7}$$

which can be rewritten as

$$\begin{aligned} P_S(q) P_G(p) P_A\Big((q,p) \to (q',p')\Big) &= P_S(q') P_G(p') P_A\Big((q',p') \to (q,p)\Big) = \\ &= P_S(q') P_G(-p') P_A\Big((q',-p') \to (q,-p)\Big) \ , \end{aligned} \tag{16.2.8}$$

where in the last step we just used the fact that p^2 is even. From the invariance of the phase space measure we have

$$\begin{aligned} P_S(q) P_G(p) P_A\Big((q,p) \to (q',p')\Big) \mathrm{d}q\,\mathrm{d}p = \\ = P_S(q') P_G(-p') P_A\Big((q',-p') \to (q,-p)\Big) \mathrm{d}q'\mathrm{d}p' \ , \end{aligned} \tag{16.2.9}$$

and using the reversibility of the evolution, $P_C\Big((q,p) \to (q',p')\Big) = P_C\Big((q',-p') \to (q,-p)\Big)$, we thus have

$$\begin{aligned} P_S(q) P_G(p) P_A\Big((q,p) \to (q',p')\Big) P_C\Big((q,p) \to (q',p')\Big) \mathrm{d}q\,\mathrm{d}p = \\ = P_S(q') P_G(-p') P_A\Big((q',-p') \to (q,-p)\Big) P_C\Big((q',-p') \to (q,-p)\Big) \mathrm{d}q'\mathrm{d}p' \ , \end{aligned} \tag{16.2.10}$$

which becomes the detailed balance equation by integrating/marginalizing the momentum [1].

Conditions a) and b) are obviously satisfied by the exact solution of the equations of motion, at least if $S(\phi)$ is sufficiently regular, as follows from the existence and uniqueness theorem for ordinary differential equations, and the Liouville theorem of analytical mechanics. If we use numerical integration schemes of the equations of motion which do not satisfy these requirements, an extrapolation to vanishing integration time-step of the simulation results is required, as in non-equilibrium molecular dynamics simulations. If instead we adopt integration schemes which exactly (up to round-off errors, obviously) satisfies conditions a) and b), the HMC algorithm is stochastically exact already for finite integration time-steps. If the integration step is too coarse, the acceptance probability typically becomes very small, and the algorithm is stochastically exact but inefficient.

[1] To be more precise, we need to know that for each q, q' and p' a value p exists (unique if $S(q)$ is a sufficiently regular function) such that (q,p) evolves at time t in (q',p').

Let us see how to build integration schemes for the Hamiltonian flow of the Hamiltonian $H(q,p) = T(p) + S(q)$ which satisfy the conditions a) and b). If we denote by $U(\tau)$ the evolution operator up to time τ in the phase space, defined by $U(\tau)f(q_0,p_0) = f(q(\tau),p(\tau))$ with $q(0) = q_0$ and $p(0) = p_0$, we can formally write (using the chain rule)

$$U(\tau) = \exp\left(\tau\frac{\mathrm{d}}{\mathrm{d}t}\right) = \exp\left(\tau\dot{q}\frac{\partial}{\partial q} + \tau\dot{p}\frac{\partial}{\partial p}\right) = \exp\left(\tau T'(p)\frac{\partial}{\partial q} - \tau S'(q)\frac{\partial}{\partial p}\right) \ , \tag{16.2.11}$$

where in the last step the Hamiltonian equations of motion

$$\dot{p} = -\frac{\partial H}{\partial q} \ , \quad \dot{q} = \frac{\partial H}{\partial p} \tag{16.2.12}$$

have been used. It is now convenient to introduce the differential operators

$$Q = T'(p)\frac{\partial}{\partial q} \ , \quad P = -S'(q)\frac{\partial}{\partial P} \ , \tag{16.2.13}$$

in such a way that $U(\tau) = \exp\{\tau(P+Q)\}$. These operators satisfy

$$\exp(\tau Q)f(q,p) = f(q+\tau T'(p),p) \ , \quad \exp(\tau P)f(q,p) = f(q,p-\tau S'(q)) \ , \tag{16.2.14}$$

and it is immediate to show that $\exp(\tau Q)$ and $\exp(\tau P)$ preserve the measure of the phase space: we have for example

$$\left|\frac{\partial\big(e^{\tau Q}(q,p)\big)}{\partial(q,p)}\right| = \begin{vmatrix} 1 & \tau T''(p) \\ 0 & 1 \end{vmatrix} = 1 \ . \tag{16.2.15}$$

By expanding $U(\tau)$ as a product of terms of the form e^{aQ} and e^{bP} we thus obtain integration schemes which preserves the measure of the phase space, which are known as symplectic integrators (for some interesting properties of these integration schemes see [102]).

The simplest symplectic integrator is (using the Baker-Campbell-Hausdorff formula)

$$\big(e^{\delta\tau\,Q}e^{\delta\tau\,P}\big)^{\tau/\delta\tau} = \Big(e^{\delta\tau\,(Q+P)+O(\delta\tau^2)}\Big)^{\tau/\delta} = e^{\tau(Q+P)+O(\delta\tau)} = U(\tau)\Big(1+O(\delta\tau)\Big) \ , \tag{16.2.16}$$

which acts as

$$e^{\delta\tau\,Q}e^{\delta\tau\,P}(q_0,p_0) = e^{\delta\tau\,Q}\Big(q_0,\underbrace{p_0-\delta\tau S'(q_0)}_{=p_1}\Big) = (q_0+\delta\tau T'(p_1),p_1) \equiv (q_1,p_1) \ , \tag{16.2.17}$$

or, more explicitly,

$$\begin{cases} p(\tau+\delta\tau) = p(\tau) - \delta\tau S'\big(q(\tau)\big) \ , \\ q(\tau+\delta\tau) = q(\tau) + \delta\tau T'\big(p(\tau+\delta\tau)\big) \ . \end{cases} \tag{16.2.18}$$

This is just the symplectic version of the standard Euler integrator, which can be easily seen not to be symplectic: from

$$\begin{cases} p(\tau+\delta\tau) = p(\tau) - \delta\tau S'\big(q(\tau)\big) \ , \\ q(\tau+\delta\tau) = q(\tau) + \delta\tau T'\big(p(\tau)\big) \ , \end{cases} \tag{16.2.19}$$

it indeed follows that

$$\left|\frac{\partial\Big(q(\tau+\delta\tau),p(\tau+\delta(\tau))\Big)}{\partial\Big(q(\tau),p(\tau)\Big)}\right| = \begin{vmatrix} 1 & \delta\tau T''(p(\tau)) \\ -\delta\tau S''(q(\tau)) & 1 \end{vmatrix} \neq 1 \ . \tag{16.2.20}$$

The symplectic Euler algorithm is however not reversible. To build a symmetric (i.e. reversible) symplectic integrator we can start from

$$V(\delta\tau) = e^{\frac{1}{2}\delta\tau\,P}e^{\delta\tau\,Q}e^{\frac{1}{2}\delta\tau\,P} \ , \tag{16.2.21}$$

indeed it is immediate to see that $V(\delta\tau)V(-\delta\tau) = 1$. Using again the Baker-Campbell-Hausdorff formula $\log(e^{tX}e^{tY}) = tX + tY + \frac{t^2}{2}[X,Y] + \frac{t^3}{12}([X,[X,Y]] - [Y,[X,Y]]) + O(t^4)$ we have

$$\begin{aligned} V(\delta\tau)^{\tau/\delta\tau} &= \left(\exp\left\{(P+Q)\delta\tau - \frac{1}{24}\Big([P,[P,Q]] + 2[Q,[P,Q]]\Big)\delta\tau^3 + O(\delta\tau^5)\right\}\right)^{\tau/\delta\tau} = \\ &= \exp\left\{(P+Q)\tau - \frac{\tau}{24}\Big([P,[P,Q]] + 2[Q,[P,Q]]\Big)\delta\tau^2 + \cdots\right\} = U(\tau) + O(\delta\tau^2)\ . \end{aligned} \tag{16.2.22}$$

This integration scheme is known as PQP (a QPQ version also exists) leapfrog or Verlet algorithm and it can be rewritten in the form

$$\begin{cases} p(\tau + \delta\tau/2) = p(\tau) - \frac{\delta\tau}{2} S'\big(q(\tau)\big)\ , \\ q(\tau + \delta\tau) = q(\tau) + \delta\tau\, T'\big(p(\tau + \delta\tau/2)\big)\ , \\ p(\tau + \delta\tau) = p(\tau + \delta\tau/2) - \frac{\delta\tau}{2} S'\big(q(\tau + \delta\tau)\big)\ . \end{cases} \tag{16.2.23}$$

It is not difficult to recursively build higher-order symmetric symplectic integrators, i. e. symmetric symplectic integrators with error $O(\delta\tau^n)$ with $n > 2$, see [103], however the use of these higher-order integration schemes is typically not particularly convenient (at least in QCD simulations). Other methods to reduce the size of the integration errors (and hence increase the integration time-step and the computational efficiency of the HMC algorithm) are discussed in [104], with focus on classical mechanics, and in [105, 102], with focus on QCD.

In several cases it is possible to write the potential term of the Hamiltonian as the sum of two terms, of which one is computationally simple and the other is computationally difficult (the typical case being that of fermionic simulations of gauge theories). In these cases it is convenient to use multi-step integrators, which perform a different number of integration steps in the "simple" and in the "difficult" part of the Hamiltoninan, see, e. g. [106, 107]. A general summary of the techniques adopted in the numerical simulation of fermionic systems can be found in [97].

Chapter 17

Gauge field theories

17.1 Generalities on group representations

Before discussing gauge theories, it is convenient to recall some facts about group representations; many more details can be found, e. g., in [108, 109, 110]. A unitary representation of rank n of a group is a mapping (continuous if the group is a continuous group) from the group to the unitary $n \times n$ complex matrices, $g \to G(g)$, characterized by the properties

$$G(g_1)G(g_2) = G(g_1 g_2) \ , \quad G(\mathrm{id}) = 1 \ , \tag{17.1.1}$$

where g_1 and g_2 are generic group elements and id is the identity of the group (hence, in particular, $G(g^{-1}) = G(g)^{-1} = G(g)^\dagger$). A representation is called reducible if a proper subspace of $\mathbb{C}^n$ exists which is left invariant by the action of $G(g)$ for all the elements of the group; if such a proper subspace does not exist the representation is said to be irreducible. If a representation is reducible we can chose a basis of $\mathbb{C}^n$ such that, in this basis, the matrix $G(g)$ (for any g) has the following block form

$$G(g) = \left(\begin{array}{c|c} X & Y \\ \hline 0 & Z \end{array}\right) \ , \tag{17.1.2}$$

moreover unitary representations are in fact completely reducible: the matrix $G(g)$ can be written in diagonal block form by a proper choice of the basis[1]. Irreducible representations can thus be considered as the building blocks of general unitary representations. Shur's lemma describes a peculiar and useful property of irreducible representations: if a $n \times n$ matrix M satisfies $[M, G(g)] = 0$ for any g, then M is proportional to the identity. In particular, irreducible representations of Abelian groups exist only for $n = 1$, since in the Abelian case $G(g_1)$ commutes with $G(g_2)$ for any g_1, g_2.

If the group is continuous, and r real numbers are needed to identify one of its elements (r is the dimensionality of the group), we can introduce a parametrization θ^a of the group, with $a = 1, \ldots, r$ and $\theta^a = 0$ corresponding to id, such that $G(g) = \exp(i\theta^a T_a)$, where the T_a matrices are the generators of the given representation. For $g \simeq \mathrm{id}$ we have in particular $|\theta^a| \ll 1$ and $G(g) \simeq 1 + i\theta^a T_a$. The generators T_a span the group algebra, which is the tangent space to the group manifold at the identity of the group. Since $G(g)$ is a unitary matrix, the generators T_a are Hermitian, moreover, if $\det G(g) = 1$, then $\mathrm{Tr}\, T_a = 0$ (since for an Hermitian matrix M it is easily seen that $\det e^{iM} = e^{i\mathrm{Tr}\, M}$). Since $G([g_1, g_2]) = [G(g_1), G(g_2)]$, where g_1 and g_2 are generic group elements, we have in particular, if g_1 and g_2 are close to the identity,

$$[1 + i\theta^a T_a, 1 + i\psi^b T_b] \simeq 1 + i\xi^c T_c \ . \tag{17.1.3}$$

[1] This follows from the fact that if a subspace is invariant under the group action, then also its orthogonal complement is invariant.

From this equation it follows that the commutator of two generators, $[T_a, T_b]$, can be written as a linear combination of generators:

$$[T_a, T_b] = i f^c_{ab} T_c \ . \tag{17.1.4}$$

In fact also the expression of ξ^c as a function of θ^a and ψ^b follows, but we will not need its precise form; note that ξ^c is of the second order, but we do not need to keep track of the second order terms in θ^a or in ψ^b in Eq. (17.1.3), since they commute with the identity. The coefficients f^c_{ab} are the structure constants, which are real numbers (since $[T_a, T_b]^\dagger = -[T_a, T_b]$), and obviously satisfy $f^c_{ab} = -f^c_{ba}$. Note that, using the Campbell-Baker-Hausdorff formula (see, e. g., [111]) and the structure constants, we can compute all the terms of the expansion of $\log(e^{i\theta^a T_a} e^{i\psi^b T_b})$, hence the structure constants completely characterize the group multiplication rule.

For compact and semisimple[2] groups it is possible to choose the generators in such a way that they satisfy the relation $\mathrm{Tr}\,(T_a T_b) = C\delta_{ab}$, where $C > 0$. Using this choice of generators (something that will be always assumed in the following), it is immediate to see that the structure constants can be written as

$$f^c_{ab} = -i\frac{1}{C}\mathrm{Tr}(T_c[T_a, T_b]) \ . \tag{17.1.5}$$

Using the cyclicity of the trace it is then simple to show that f^c_{ab} is completely antisymmetric in all indices. Since the indices enter now on equal footing, it is more standard to write just f_{abc} instead of f^c_{ab}.

A simple example which should be familiar from quantum mechanics is that of the continuous group[3] SU(2): elements of the group SU(2) can be parametrized by three angles, SU(2) irreducibile representations are characterized by the spin s, a spin s representation acts on fields with $n = 2s+1$ complex components, and $G(g)$ is the $(2s+1)\times(2s+1)$ Wigner rotation matrix $D^{(s)}_{i'i}$. The so-called fundamental representation is the one directly related to the definition of the group as the 2×2 group of unitary matrices with unit determinant: using the standard relation

$$e^{i\alpha \boldsymbol{n}\cdot\boldsymbol{\sigma}/2} = \cos(\alpha/2) + i\boldsymbol{n}\cdot\boldsymbol{\sigma}\sin(\alpha/2) \ , \tag{17.1.6}$$

with $\boldsymbol{n}^2 = 1$, it is simple to verify that any SU(2) matrix can be written in the form $e^{i\alpha\boldsymbol{n}\cdot\boldsymbol{\sigma}/2}$ for some $\boldsymbol{n}$ and some $\alpha \in [0, 4\pi)$, either by direct computation or by using the infinitesimal form and the connectivity of the group. This means that we can use as generators of the fundamental representation the matrices $T_a = \sigma_a/2$, which satisfy $\mathrm{Tr}(T_a T_b) = \frac{1}{2}\delta_{ab}$ and

$$\left[\frac{\sigma_a}{2}, \frac{\sigma_b}{2}\right] = i\epsilon_{abc}\frac{\sigma_c}{2} \ , \tag{17.1.7}$$

hence the structure constants of SU(2) are $f_{abc} = \epsilon_{abc}$. For SU(N) with $N > 2$ no simple parametrization of the group elements exists, but it is simple to understand that the algebra associated with the fundamental representation is that of the (N^2-1)-dimensional space of $N \times N$ traceless Hermitian matrices. Generators of the fundamental representation are typically normalized according to $\mathrm{Tr}(T_a T_b) = \frac{1}{2}\delta_{ab}$ just like in SU(2).

If we denote by T_a the generators of the fundamental representation of SU(N), the set $\{T_a, \frac{1}{\sqrt{2N}} I\}$ costitutes a basis for the (complex) vector space of $N \times N$ complex matrices, which is orthogonal with respect to the scalar product $\langle M|N\rangle = \mathrm{ReTr}(M^\dagger N)$. If we now consider the matrix $M^{(ik)}$, with matrix elements $(M^{(ik)})_{lm} = \delta_{im}\delta_{lk}$, and expand it on this basis we get the so called Fiez identity

$$\delta_{im}\delta_{lk} = \frac{1}{N}\delta_{ik}\delta_{lm} + 2(T_a)_{ik}(T_a)_{lm} \ , \tag{17.1.8}$$

[2] A continuous group is (often) called semisimple if its algebra has no proper invariant Abelian subalgebras. Note however that, despite the fact that the theory of continuous group of transformations dates back to the late 19th century, the definition of "semisimple" continuous group (and even of "simple" continuous group) is not always the same in the mathematical literature.

[3] To be precise, in quantum mechanics the group SO(3) is typically used, but half-integer spin cases do not satisfy $G(\mathrm{id}) = I$, hence they are not representations of SO(3) (this is the reason for $\ell \in \mathbb{N}$), and are often called two-valued representations. From the mathematical point of view a better characterization of these representations is the following: they correspond to prjoective representations (i. e. representations up to a phase) of SO(3), which can be lifted to proper representation of the group SU(2). SU(2) is indeed the covering of SU(3), i. e. the simply connected group with the same structure constants of SO(3).

and contracting this identity with δ_{kl} we get

$$T_a T_a = \frac{N^2-1}{2N} \ . \tag{17.1.9}$$

The operator $T_a T_a$ is the quadratic Casimir operator, and plays in SU(N) a role analogous to that of $\boldsymbol{J}^2$ in SU(2).

A case which is technically even simpler is that of the U(1) group: elements of the U(1) group can be parametrized by a single angle $\varphi \in [0, 2\pi)$, irreducible representations of the group U(1) are unidimensional (due to Shur's lemma), are characterized by an integer number $q \in \mathbb{Z}$, and their action is just the multiplication by the complex number $G(g) = e^{iq\varphi}$.

From the Jacobi identity

$$[T_a, [T_b, T_c]] + [T_b, [T_c, T_a]] + [T_c, [T_a, T_b]] = 0 \tag{17.1.10}$$

it imediately follow that

$$f_{ade} f_{bcd} + f_{bde} f_{cad} + f_{cde} f_{abd} = 0 \ . \tag{17.1.11}$$

If we introduce the matrices $T_a^{(adj)}$ by

$$(T_a^{(adj)})_{bc} = i f_{bac} \ , \tag{17.1.12}$$

it is simple to verify that the Jacobi identity for the structure constants can be rewritten (using the antisymmetry of the structure constants) as

$$(T_a^{(adj)})_{bd} (T_c^{(adj)})_{de} - (T_c^{(adj)})_{bd} (T_a^{(adj)})_{de} = i f_{acd} (T_d^{(adj)})_{be} \ , \tag{17.1.13}$$

i. e.

$$[T_a^{(adj)}, T_b^{(adj)}] = i f_{abc} T_c^{(adj)} \ . \tag{17.1.14}$$

The matrices $T_a^{(adj)}$ are the generators of the adjoint representation. Note that for SU(2) the generators $T_a^{(adj)}$ are nothing but the generators of the spin 1 representation (which is the fundamental representation of SO(3)).

The action of the adjoint representation can be visualized as follows: if T_a are r the generators of the fundamental representation of SU(N), and x is a N-dimensional complex vector, we can define the r real numbers y_a by $y_a = x^\dagger T_a x$. Under the action of the group we have $x \to G(g)x$, where $G(g) = e^{i\theta^a T_a}$, hence $y_a \to ({}^g y)_a = x^\dagger G^\dagger(g) T_a G(g) x$, which can be written as the linear combination $V_{ab}(g) y_b$ of the original variables, with the real matrix elements $V_{ab}(g)$ defined by

$$G^\dagger(g) T_a G(g) = V_{ab}(g) T_b \ . \tag{17.1.15}$$

Using the normalization of the generators of the fundamental representation we have explicitly

$$V_{ab}(g) = 2\mathrm{Tr}(G^\dagger(g) T_a G(g) T_b) \ . \tag{17.1.16}$$

From the fact that $G(g)$ is a representation of the group, hence $G(g_1 g_2) = G(g_1) G(g_2)$, the equality $V_{ab}(g_1) V_{bc}(g_2) = V_{ac}(g_1 g_2)$ easily follows, hence $V(g)$ is also a representation. Moreover the transformation $T_a \to G^\dagger(g) T_a G(g)$ is unitary with respect to the scalar product obtained extending by linearity $\langle T_a | T_b \rangle = \mathrm{ReTr}(T_a T_b)$, from which the orthogonality of $V(g)$ follows (a result that can also be obtained by using the explicit form of $V_{ab}(g)$ and the Fierz identity). For $g \approx \mathrm{Id}$ we have

$$G^\dagger(g) T_a G(g) \simeq (1 - i\theta^c T_c) T_a (1 + i\theta^c T_c) = T_a + i\theta^c [T_a, T_c] = T_a + i\theta^c i f_{acb} T_b \ , \tag{17.1.17}$$

hence $V_{ab}(g) \simeq \delta_{ab} + i f_{acb} i\theta^c = \delta_{ab} + i\theta^c (T_c^{(adj)})_{ab}$, and V_{ab} is the adjoint representation matrix.

17.2 Continuum gauge theories

Let us assume $\phi(x)$ to be a field[4] with several components, which transforms according to a given irreducible representation of a continuous group of transformations which commutes with the Poincarè group (a so called "internal" group):

$$\phi(x) \to {}^g\phi(x) = G(g)\phi(x) \ . \tag{17.2.1}$$

[4] We assume for the sake of the simplicity this field to be a scalar one, but the presence of Lorentz indices is irrelevant for what follows.

In this equation g denotes a constant (i. e. independent of x) element of the group of transformations, while $G(g)$ is the matrix associated with g by the given representation (in the following we will often write just G instead of $G(g)$). For concreteness, and since this is the most common case in applications, we will assume $\phi(x)$ to be a complex field with n components, and $G(g)$ to be a unitary $n \times n$ matrix.

Since we assumed g to be independent of x in Eq. (17.2.1) (and G to be unitary), it is immediate to verify that expressions like, e. g.,

$$\phi^\dagger(x) \cdot \phi(x) \ , \quad [\partial_\mu \phi(x)]^\dagger \cdot \partial^\mu \phi(x) \ , \tag{17.2.2}$$

are invariant under the transformation $\phi \to G(g)\phi$. This is no more the case if we consider local transformations, i. e. trasformations for which g depends on the point of application: under the action of a local transformation $\phi^\dagger(x) \cdot \phi(x)$ is still invariant, but $[\partial_\mu \phi(x)]^\dagger \cdot \partial^\mu \phi(x)$ is not. Let us consider how $\partial_\mu \phi$ changes under a local transformation (the dependence of g, and thus of $G(g)$, on x is implied):

$$\partial_\mu \phi \to \partial_\mu {}^g\phi = \partial_\mu (G\phi) = G(\partial_\mu \phi) + (\partial_\mu G)\phi = G\Big(\partial_\mu \phi + G^\dagger (\partial_\mu G)\phi\Big) \ . \tag{17.2.3}$$

In order to remove the non-homogeneous term from the previous equation, and promote the global symmetry to a local (gauge) symmetry, let us introduce the gauge field A_μ (represented by a $n \times n$ complex matrix) and the covariant derivative

$$D_\mu = \partial_\mu + ieA_\mu \ , \tag{17.2.4}$$

where e is the coupling constant. Note that the coupling constant is typically denoted by g in the non-Abelian case, but g could be confused with the group element. We obviously have

$$D_\mu \phi \to {}^g(D_\mu \phi) = (\partial_\mu + ie\, {}^g A_\mu) G\phi = G\Big(\partial_\mu + G^\dagger(\partial_\mu G) + ieG^\dagger {}^g A_\mu G\Big)\phi \ , \tag{17.2.5}$$

and if we impose ${}^g(D_\mu \phi) = G D_\mu \phi$ we obtain the relation

$$G^\dagger(\partial_\mu G) + ieG^\dagger {}^g A_\mu G = ieA_\mu \tag{17.2.6}$$

and thus the transformation law of the gauge field

$${}^g A_\mu = GA_\mu G^\dagger + \frac{i}{e}(\partial_\mu G)G^\dagger = GA_\mu G^\dagger - \frac{i}{e}G(\partial_\mu G^\dagger) \ , \tag{17.2.7}$$

where in the last equality we used

$$0 = \partial_\mu 1 = \partial_\mu (GG^\dagger) = (\partial_\mu G)G^\dagger + G\partial_\mu G^\dagger \ . \tag{17.2.8}$$

The derivative of the exponential of a non-constant matrix $M(\alpha)$ can be written as (see, e. g., [111])

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} e^{M(\alpha)} = \int_0^1 e^{(1-t)M(\alpha)} \frac{\mathrm{d}M(\alpha)}{\mathrm{d}\alpha} e^{tM(\alpha)} \mathrm{d}\alpha \ , \tag{17.2.9}$$

which can be intuitively understood by writing e^M as $e^{M/N} \cdots e^{M/N}$ (where N terms are present), using the fact that $e^{M/N}$ and $\frac{1}{N}\frac{\mathrm{d}M}{\mathrm{d}\alpha}$ commute to leading order in $1/N$, and rewriting the $N \to \infty$ limit as an integral. By using this identity we can rewrite the non-homogeneous term $i(\partial_\mu G)G^\dagger$ which appears in the transformation of A_μ, where $G(g(x)) = e^{i\theta^a(x)T_a}$, as

$$\begin{aligned} i(\partial_\mu G)G^\dagger &= i \int_0^1 e^{(1-t)i\theta^b T_b} i(\partial_\mu \theta^a) T_a e^{ti\theta^b T_b} \mathrm{d}t G^\dagger = \\ &= -(\partial_\mu \theta^a) \int_0^1 e^{i(1-t)\theta^b T_b} T_a e^{-i(1-t)\theta^b T_b} \mathrm{d}t \ . \end{aligned} \tag{17.2.10}$$

For any $G(g)$ the matrix $G^\dagger(g)T_aG(g)$ is in the group algebra, since an element g_ϵ of the group exists whose representation is $e^{i\epsilon T_a}$ and

$$G(g^{-1}g_\epsilon g) = G^\dagger(g)G(g_\epsilon)G(g) \simeq G^\dagger(g)(1+i\epsilon T_a)G(g) = 1 + i\epsilon G^\dagger(g)T_aG(g) \ , \tag{17.2.11}$$

hence $i(\partial_\mu G)G^\dagger$ is in the group algebra and, for the same reason, if A_μ is in the algebra the quantity $GA_\mu G^\dagger$ is also in the algebra. It is thus consistent to assume the matrix A_μ to live in the group algebra: $A_\mu = A^a_\mu T_a$. Moreover, from the transformation rule of A_μ (see Eq. (17.2.7)), we see that under global transformations (i. e., with constant g and thus constant $G(g)$) the matrix A_μ transforms in the adjoint representation of the group.

Note that we used (and we will use in the following) the matrix notation $A_\mu = A^a_\mu T_a$ to write formulas in a compact form, but the true gauge fields are the components A^a_μ (N^2-1 components, for SU(N)). This can be understood by considering a model in which two different matter fields are present, transforming in two different representations, and interacting with the same gauge fields. If we denote by $T^{(1)}_a$ and $T^{(2)}_a$ the generators of these representations, the covariant derivatives for the two matter fields are

$$D^{(i)}_\mu = \partial_\mu + ieA^a_\mu T^{(i)}_a \ , \tag{17.2.12}$$

where $i = 1, 2$ and the same fields A^a_μ enter both the covariant derivatives.

The field strength $F_{\mu\nu}$ is defined, in the non-Abelian case, by

$$F_{\mu\nu} = -\frac{i}{e}[D_\mu, D_\nu] \ . \tag{17.2.13}$$

To verify that this definition makes sense let us compute $F_{\mu\nu}\phi$:

$$\begin{aligned} -\frac{i}{e}[D_\mu, D_\nu]\phi &= -\frac{i}{e}\left\{(\partial_\mu + ieA_\mu)(\partial_\nu + ieA_\nu)\phi - (\partial_\nu + ieA_\nu)(\partial_\mu + ieA_\mu)\phi\right\} = \\ &= -\frac{i}{e}\left\{ieA_\mu\partial_\nu\phi + ie\partial_\mu(A_\nu\phi) - e^2A_\mu A_\nu\phi - ieA_\nu\partial_\mu\phi - ie\partial_\nu(A_\mu\phi) + e^2A_\nu A_\mu\phi\right\} = \\ &= -\frac{i}{e}\left\{ie(\partial_\mu A_\nu - \partial_\nu A_\mu)\phi - e^2[A_\mu, A_\nu]\phi\right\} = (\partial_\mu A_\nu - \partial_\nu A_\mu + ie[A_\mu, A_\nu])\phi \ , \end{aligned} \tag{17.2.14}$$

from which we see that $F_{\mu\nu}$ is not a differential operator, and explicitly

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + ie[A_\mu, A_\nu] \ . \tag{17.2.15}$$

Just like A_μ, also $F_{\mu\nu}$ lives in the algebra of the group, and if we introduce its components using $F_{\mu\nu} = F^c_{\mu\nu}T_c$ we get (from the definition of the structure constants f_{abc})

$$F^c_{\mu\nu} = \partial_\mu A^c_\nu - \partial_\nu A^c_\mu - ef_{abc}A^a_\mu A^b_\nu \ . \tag{17.2.16}$$

Under a gauge transformation we have

$$F_{\mu\nu}\phi \to {}^gF_{\mu\nu}{}^g\phi = -\frac{i}{e}[{}^gD_\mu, {}^gD_\nu]^g\phi = -\frac{i}{e}G[D_\mu, D_\nu]\phi = -\frac{i}{e}G[D_\mu, D_\nu]G^{\dagger g}\phi \ , \tag{17.2.17}$$

hence

$${}^gF_{\mu\nu} = GF_{\mu\nu}G^\dagger \ , \tag{17.2.18}$$

and $F_{\mu\nu}$ transforms in the adjoint representation of the gauge group. Only in the Abelian case the field strength $F_{\mu\nu}$ is gauge invariant.

The Euclidean action of the SU(N) gauge theory (with generators normalized according to $\mathrm{Tr}(T_aT_b) = \frac{1}{2}\delta_{ab}$) in D dimensions is

$$S_E = \frac{1}{4}\int F^a_{\mu\nu}(x)F^a_{\mu\nu}(x)\mathrm{d}^Dx = \frac{1}{2}\int \mathrm{Tr}\Big[F_{\mu\nu}(x)F_{\mu\nu}(x)\Big]\mathrm{d}^Dx \ , \tag{17.2.19}$$

which is the natural generalization of the U(1) Abelian case

$$S_E = \frac{1}{4}\int F_{\mu\nu}(x)F_{\mu\nu}(x)\mathrm{d}^D x \ . \tag{17.2.20}$$

Note that in the non-Abelian case $F_{\mu\nu}F_{\mu\nu}$ is not gauge invariant, but $\mathrm{Tr}\Big[F_{\mu\nu}F_{\mu\nu}\Big]$ is, as imediately follows from Eq. (17.2.18). Since the Euclidean action is dimensionless (in natural units), the field strength $F_{\mu\nu}$ has mass dimension $D/2$, from which we obtain $[A_\mu] = \frac{D}{2} - 1$ and

$$[e] = \frac{D}{2} - 2[A_\mu] = \frac{D}{2} - 2\left(\frac{D}{2} - 1\right) = \frac{4-D}{2} \ . \tag{17.2.21}$$

The quantum field theory described by the action S_E introduced above is typically called Yang-Mills theory, or pure gauge theory, to distinguish it from QCD or QCD-like theories, in which matter fields are coupled to the gauge fields. Note that, in the non-Abelian case, Yang-Mills theories are not free theories, since the field strength $F_{\mu\nu}$ also contains a term which is quadratic in the gauge fields.

To formulate gauge theories on the lattice the concept of parallel transport along a curve will turn out to be useful. We can define the parallel transport along the curve C from x to y as

$$U_{C_{y\leftarrow x}} = \mathrm{P}\exp\left(-ie\int_x^y A_\mu(z)\mathrm{d}z_\mu\right) \ , \tag{17.2.22}$$

where P exp denotes the path-ordered exponential (points closer to x along C stay on the right), which is a simple extention of the usual time-ordered exponential encountered in time-dependent perturbation theory. Let us note that, since A_μ is an element of the algebra of the group, the parallel transport lives in the same representation of $\phi(x)$, moreover if C is a path from x to y and C' is a path from y to z, we have

$$U_{C'_{z\leftarrow y}}U_{C_{y\leftarrow x}} = U_{C''_{z\leftarrow x}} \ , \tag{17.2.23}$$

where C'' is the path which goes from x to z obtained by joining the paths C and C'.

If we parameterize the path from x to y by the function $z(t)$, with $z(0) = x$ and $z(1) = y$, the parallel transport from x to y along C is the solution (computed at $t = 1$) of the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}U(t) = -ieA_\mu(z(t))\dot{z}_\mu(t)U(t) \tag{17.2.24}$$

with initial condition $U(0) = 1$. This can be shown by rewriting this initial problem in the integral form

$$U(t) = 1 - ie\int_0^t A_\mu(z(\tau))\dot{z}_\mu(\tau)U(\tau)\mathrm{d}\tau \ , \tag{17.2.25}$$

and solving it by iteration:

$$\begin{aligned}
U^{(0)}(t) &= 1 \ , \\
U^{(1)}(t) &= 1 - ie\int_0^t A_\mu(z(\tau))\dot{z}_\mu(\tau)\mathrm{d}\tau \ , \\
U^{(2)}(t) &= 1 - ie\int_0^t A_\mu(z(\tau))\dot{z}_\mu(\tau)U^{(1)}(\tau)\mathrm{d}\tau = \\
&= 1 - ie\int_0^t A_\mu(z(\tau))\dot{z}_\mu(\tau)\mathrm{d}\tau + (ie)^2\int_0^t \mathrm{d}\tau\int_0^\tau \mathrm{d}\xi A_\mu(z(\tau))\dot{z}_\mu(\tau)A_\nu(z(\xi))\dot{z}_\nu(\xi) = \\
&= 1 - ie\int_0^t A_\mu(z(\tau))\dot{z}_\mu(\tau)\mathrm{d}\tau + \frac{(ie)^2}{2}\int_0^t \mathrm{d}\tau\int_0^t \mathrm{d}\xi P\Big[A_\mu(z(\tau))A_\nu(z(\xi))\Big]\dot{z}_\mu(\tau)\dot{z}_\nu(\xi) \ ,
\end{aligned} \tag{17.2.26}$$

and so on, where the path-ordered product is defined by

$$P\Big[A_\mu(z(\tau))A_\nu(z(\xi))\Big] = \begin{cases} A_\mu(z(\tau))A_\nu(z(\xi)) & \text{if } \tau > \xi \\ A_\mu(z(\xi))A_\nu(z(\tau)) & \text{if } \xi > \tau \end{cases} . \tag{17.2.27}$$

The gauge transformed parallel transport ${}^gU$ can be obtained by using ${}^gA_\mu$ instead of A_μ in the definitions Eq. (17.2.22) or Eq. (17.2.24), so

$$\frac{\mathrm{d}}{\mathrm{d}t}{}^gU(t) = -ie\Big[G(z(t))A_\mu(z(t))G^\dagger(z(t)) + \frac{i}{e}\partial_\mu G(z(t))G^\dagger(z(t))\Big]\dot{z}_\mu(t){}^gU(t) , \tag{17.2.28}$$

and it is simple to verify that exactly the same differential equation is satisfied by the quantity $V(t) = G(z(t))U(t)G^\dagger(z(0))$. Using Eq. (17.2.24) we have indeed

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}V(t) &= \partial_\mu G(z(t))\dot{z}_\mu(t)U(t)G^\dagger(z(0)) + G(z(t))\dot{U}(t)G^\dagger(z(0)) = \\ &= -ie\left(\frac{i}{e}\partial_\mu G(z(t))U(t)G^\dagger(z(0)) + G(z(t))A_\mu(z(t))U(t)G^\dagger(z(0))\right)\dot{z}_\mu(t) = \\ &= -ie\left(\frac{i}{e}\partial_\mu G(z(t))G^\dagger(z(t)) + G(z(t))A_\mu(z(t))G^\dagger(z(t))\right)\dot{z}_\mu(t)V(t) . \end{aligned} \tag{17.2.29}$$

Moreover we have by definition ${}^gU(0) = 1$ and $V(0) = G(z(0))U(0)G^\dagger(z(0)) = 1$. ${}^gU(t)$ and $V(t)$ thus satisfy the same differential equation with the same initial condition, hence they are equal and

$${}^gU(t) = G(z(t))U(t)G^\dagger(z(0)) , \tag{17.2.30}$$

and in particular, using $t = 1$, we have

$${}^gU_{C_{y\leftarrow x}} = G(y)U_{C_{y\leftarrow x}}G^\dagger(x) . \tag{17.2.31}$$

A more elementary way of reaching the same conclusion is to consider the infinitesimal parallel transport $U_{C_{x+dx\leftarrow x}} \simeq 1 - ieA_\mu(x)\mathrm{d}x_\mu$ and proceed as follows (no sum on μ is present):

$$\begin{aligned} {}^gU_{C_{x+dx\leftarrow x}} &\simeq 1 - ie{}^gA_\mu(x)\mathrm{d}x_\mu = 1 - ieG(x)A_\mu(x)G^\dagger(x)\mathrm{d}x_\mu + (\partial_\mu G(x))G^\dagger(x)\mathrm{d}x_\mu \simeq \\ &\simeq 1 - ieG(x+\mathrm{d}x)A_\mu(x)G^\dagger(x)\mathrm{d}x_\mu + \frac{G(x+\mathrm{d}x) - G(x)}{\mathrm{d}x_\mu}G^\dagger(x)\mathrm{d}x_\mu \simeq \\ &= G(x+\mathrm{d}x)G^\dagger(x) - ieG(x+\mathrm{d}x)A_\mu(x)G^\dagger(x)\mathrm{d}x_\mu = \\ &= G(x+\mathrm{d}x)\left(1 - ieA_\mu(x)\mathrm{d}x_\mu\right)G^\dagger(x) \simeq G(x+\mathrm{d}x)U_{C_{x+dx\leftarrow x}}G^\dagger(x) . \end{aligned} \tag{17.2.32}$$

Under global transformations, with $G(x)$ independent of x, parallel transports thus transform according to the adjoint representation of the group.

Note that using the parallel transport the covariant directional derivative along the direction n_ν can be written as

$$n_\mu D_\mu\phi(x) = \lim_{\delta\to 0}\frac{U_{C_{x\leftarrow x+n\delta}}\phi(x+n\delta) - \phi(x)}{\delta} , \tag{17.2.33}$$

and this definition is meaningful since the gauge transformation rule Eq. (17.3.3) ensures that both the terms on the right hand side transform in the same way under local gauge transformations:

$$\begin{aligned} {}^g\Big[U_{C_{x\leftarrow y}}\phi(y)\Big] &= {}^gU_{C_{x\leftarrow y}}{}^g\phi(y) = G(x)U_{C_{x\leftarrow y}}G^\dagger(y)G(y)\phi(y) = \\ &= G(x)U_{C_{x\leftarrow y}}\phi(y) . \end{aligned} \tag{17.2.34}$$

17.3 Lattice gauge theories: basics

If we define, as usual, the lattice fields $\phi_{\boldsymbol{n}}$ on the lattice sites, it is immediate to see that the forward lattice derivative

$$\partial_\mu^{(F)}\phi_{\boldsymbol{n}} = \frac{1}{a}(\phi_{\boldsymbol{n}+\hat{\mu}} - \phi_{\boldsymbol{n}}) \tag{17.3.1}$$

is not gauge covariant, just like its continuum counterpart $\partial_\mu\phi(x)$. To write a lattice covariant derivative we can introduce the lattice gauge fields

$$U_\mu(\boldsymbol{n}) = U_{\boldsymbol{n}+\mu\leftarrow\boldsymbol{n}} \tag{17.3.2}$$

associated with the parallel transports along the positive directions of the links (i. e. $\mu \geq 0$). The transformation rule Eq. (17.2.31) then gives

$${}^gU_\mu(\boldsymbol{n}) = G(\boldsymbol{n}+\hat{\mu})U_\mu(\boldsymbol{n})G^\dagger(\boldsymbol{n}) \ , \tag{17.3.3}$$

and it is immediate to verify that

$$\frac{1}{a}\Big(U_\mu^\dagger(\boldsymbol{n})\phi_{\boldsymbol{n}+\hat{\mu}} - \phi_{\boldsymbol{n}}\Big) \tag{17.3.4}$$

is gauge covariant.

If we assume to know $A_\mu(x)$ in the continuum, we have

$$U_\mu(\boldsymbol{n}) = \mathrm{P}\exp\left(-ie\int_0^1 A_\mu(z(t))\dot{z}_\mu(t)\mathrm{d}t\right) \ . \tag{17.3.5}$$

We can chose $\boldsymbol{z}(t) = \boldsymbol{n} + t\hat{\mu}$ and develop $A_\mu(z(t))$ in Taylor series around $t = 1/2$, to get

$$U_\mu(\boldsymbol{n}) = \exp\left(-ieaA_\mu(\boldsymbol{n}+\hat{\mu}/2) + O(a^3)\right) \ . \tag{17.3.6}$$

Note however that in the lattice setup the fundamental variable is $U_\mu(\boldsymbol{n})$: unlike continuum gauge fields, lattice gauge fields live in the group representation, and not in its algebra.

Using gauge and matter fields it is easy to write gauge invariant expressions: for example

$$\phi^\dagger_{\boldsymbol{m}}\left(\prod_{\boldsymbol{m}\leftarrow\boldsymbol{n}} U_\mu(\boldsymbol{i})\right)\phi_{\boldsymbol{n}} \ , \tag{17.3.7}$$

is gauge invariant, where the product stands for the lattice path-ordered product along a path connecting $\boldsymbol{n}$ with $\boldsymbol{m}$. If we consider just the path-ordered product of gauge variables along a path we have

$${}^g\left(\prod_{\boldsymbol{m}\leftarrow\boldsymbol{n}} U_\mu(\boldsymbol{i})\right) = G(\boldsymbol{m})\left(\prod_{\boldsymbol{m}\leftarrow\boldsymbol{n}} U_\mu(\boldsymbol{i})\right)G^\dagger(\boldsymbol{n}) \ , \tag{17.3.8}$$

and to get a gauge invariant quantity from this expression we have to restrict to the case $\boldsymbol{m} = \boldsymbol{n}$ and take the trace:

$$\mathrm{Tr}\left(\prod_{\boldsymbol{n}\leftarrow\boldsymbol{n}} U_\mu(\boldsymbol{i})\right) \ , \tag{17.3.9}$$

where the path-ordered product extends on a closed path, is gauge invariant. The simplest closed nontrivial path is that obtained by starting from $\boldsymbol{n}$ and then moving forward in direction μ, moving forward in direction ν (with $\mu \neq \nu$), moving backward in direction μ and finally moving backward in direction ν. The gauge invariant quantity associated with this path is typically called the plaquette:

$$\begin{aligned} P_{\mu\nu}(\boldsymbol{n}) &= \mathrm{ReTr}\Pi_{\mu\nu}(\boldsymbol{n}) \ , \\ \Pi_{\mu\nu}(\boldsymbol{n}) &= U_{\boldsymbol{n}\leftarrow\boldsymbol{n}+\hat{\nu}}U_{\boldsymbol{n}+\hat{\nu}\leftarrow\boldsymbol{n}+\hat{\mu}+\hat{\nu}}U_{\boldsymbol{n}+\hat{\mu}+\hat{\nu}\leftarrow\boldsymbol{n}+\hat{\mu}}U_{\boldsymbol{n}+\hat{\mu}\leftarrow\boldsymbol{n}} = \\ &= U_\nu^\dagger(\boldsymbol{n})U_\mu^\dagger(\boldsymbol{n}+\hat{\nu})U_\nu(\boldsymbol{n}+\hat{\mu})U_\mu(\boldsymbol{n}) \ . \end{aligned} \tag{17.3.10}$$

From $[\mathrm{Tr}(M)]^* = \mathrm{Tr}[M^\dagger]$ it is simple to show that $P_{\mu\nu}(\boldsymbol{n}) = P_{\nu\mu}(\boldsymbol{n})$; note however that often $P_{\mu\nu}$ is defined with Tr instead of the ReTr used in Eq. (17.3.10), in which case one gets instead $P_{\mu\nu}(\boldsymbol{n})^* = P_{\nu\mu}(\boldsymbol{n})$. Often the product of liks $\Pi_{\mu\nu}$ in Eq. (17.3.10) is also called plaquette.

By using Eq. (17.3.6) and the Campbell-Baker-Hausdorff formula to the first nontrivial order ($e^M e^N = e^{M+N+\frac{1}{2}[M,N]}$) it is not difficult to show that the plaquette (for fixed gauge fields and $a \to 0$, the so called *naive* continuum limit) is related to the fields strength. We have indeed (no sum on repeated indices)

$$\begin{aligned} U_\nu(\boldsymbol{n}+\hat{\mu})U_\mu(\boldsymbol{n}) = \exp\Big(& -ieaA_\nu(\boldsymbol{n}+\hat{\mu}+\hat{\nu}/2) - iaeA_\mu(\boldsymbol{n}+\hat{\mu}/2) - \\ & - \frac{e^2a^2}{2}\left[A_\nu(\boldsymbol{n}+\hat{\mu}+\hat{\nu}/2), A_\mu(\boldsymbol{n}+\hat{\mu}/2)\right] + O(a^3)\Big) = \\ = \exp\left(-iea\left\{A_\mu + A_\nu\right\} - iea^2\left\{\frac{1}{2}\partial_\mu A_\mu + \frac{1}{2}\partial_\nu A_\nu + \partial_\mu A_\nu\right\} - \frac{e^2a^2}{2}[A_\nu, A_\mu] + O(a^3)\right) & , \end{aligned} \tag{17.3.11}$$

where in the final expression all quantities are computed in $\boldsymbol{n}$. In the same way we get

$$\begin{aligned} U_\nu^\dagger(\boldsymbol{n})U_\mu^\dagger(\boldsymbol{n}+\hat{\nu}) = \exp\Bigg(iea\left\{A_\mu + A_\nu\right\} + iea^2\left\{\partial_\nu A_\mu + \frac{1}{2}\partial_\mu A_\mu + \frac{1}{2}\partial_\nu A_\nu\right\} - \\ - \frac{e^2a^2}{2}[A_\nu, A_\mu] + O(a^3)\Bigg) \quad , \end{aligned} \tag{17.3.12}$$

and finally

$$\begin{aligned} \Pi_{\mu\nu}(\boldsymbol{n}) &= \exp\Big(- iea^2\big\{\partial_\mu A_\nu - \partial_\nu A_\mu + ie[A_\mu, A_\nu]\big\} + O(a^3)\Big) = \\ &= \exp\Big(- iea^2 F_{\mu\nu}(\boldsymbol{n}) + O(a^3)\Big) \quad . \end{aligned} \tag{17.3.13}$$

We are now ready to introduce the lattice action for gauge fields. In the U(1) case we have

$$P_{\mu\nu}(\boldsymbol{n}) = \cos\Big(ea^2 F_{\mu\nu}(\boldsymbol{n}) + O(a^3)\Big) \quad , \tag{17.3.14}$$

and in the naive $a \to 0$ limit,

$$1 - P_{\mu\nu} \simeq \frac{1}{2}e^2a^4F_{\mu\nu}^2 \quad . \tag{17.3.15}$$

We can write the sum of all the plaquettes as

$$\sum_{\text{plaq}} = \frac{1}{2}\sum_{\boldsymbol{n},\mu\neq\nu} \quad , \tag{17.3.16}$$

hence the quantity (note that $P_{\mu\mu} = 1$ for each μ)

$$\begin{aligned} \beta\sum_{\text{plaq}}(1 - P_{\mu\nu}) \simeq \frac{\beta}{2}\sum_{\boldsymbol{n},\mu\nu}\frac{1}{2}e^2a^4F_{\mu\nu}^2 = \beta e^2 a^{4-D}\sum_{\boldsymbol{n},\mu\nu}\frac{1}{4}a^D F_{\mu\nu}^2 \simeq \\ \simeq \beta e^2 a^{4-D}\sum_{\mu\nu}\int \mathrm{d}^D x\frac{1}{4}F_{\mu\nu}^2 \end{aligned} \tag{17.3.17}$$

reproduces the Euclidean continuum action if we use for the dimensionless parameter β the value[5]

$$\beta = \frac{1}{e^2a^{4-D}} \quad . \tag{17.3.18}$$

[5]Note that in performing the $a \to 0$ limit we have neglected the fluctuations of the fields, and this computation is equivalent to a tree-level perturbative computation (hence the name of *naive* continuum limit).

In the non-Abelian case we have to remember that the SU(N) generators satisfy $\mathrm{Tr}T_a = 0$ and $\mathrm{Tr}(T_aT_b) = \frac{1}{2}\delta_{ab}$, hence

$$\begin{aligned} 1 - \frac{1}{N}P_{\mu\nu} = 1 - \frac{1}{N}\mathrm{ReTr}\exp\Big(- iea^2 F_{\mu\nu}(\boldsymbol{n}) + O(a^3)\Big) \simeq \\ \simeq \frac{e^2a^4}{2N}\mathrm{ReTr}(F_{\mu\nu}^2) = \frac{e^2a^4}{4N}\sum_a (F_{\mu\nu}^a)^2 \ . \end{aligned} \tag{17.3.19}$$

In the naive continuum limit we thus have, proceeding as in the U(1) case, the relation

$$\beta\sum_{\mathrm{plaq}}\left(1 - \frac{1}{N}P_{\mu\nu}\right) \simeq \beta\frac{e^2a^{4-D}}{2N}\sum_{\mu\nu}\int \mathrm{d}^D x \frac{1}{4}\sum_a (F_{\mu\nu}^a)^2 \ , \tag{17.3.20}$$

and this expression reproduces the Euclidean continuum action if the dimensionless β parameter is equal to

$$\beta = \frac{2N}{e^2a^{4-D}} \ . \tag{17.3.21}$$

The action

$$\beta\sum_{\mathrm{plaq}}\left(1 - \frac{1}{N}P_{\mu\nu}\right) \tag{17.3.22}$$

is known as Wilson action [112]. In the following we will use exclusively the Wilson form of the lattice action, however it is important to stress that the form of the lattice action is by no means unique: to perform the continuum limit we have to approach a continuous phase transition, and for two different lattice actions to describe the same continuum physics it is enough that they display continuous transitions in the same universality class (for other lattice actions see, e. g., [82] §3.2.9, [113] §9)

When scalar fields are coupled to the gauge fields, it is not difficult to show in a similar way that the quantity

$$\begin{aligned} \big(U_\mu^\dagger(\boldsymbol{n})\phi_{\boldsymbol{n}+\mu} - \phi_{\boldsymbol{n}}\big)^\dagger \cdot \big(U_\mu^\dagger(\boldsymbol{n})\phi_{\boldsymbol{n}+\mu} - \phi_{\boldsymbol{n}}\big) = \\ = \phi_{\boldsymbol{n}+\mu}^\dagger \cdot \phi_{\boldsymbol{n}+\mu}^\dagger - 2\mathrm{Re}\Big(\phi_{\boldsymbol{n}+\mu}^\dagger U_\mu(\boldsymbol{n})\phi_{\boldsymbol{n}}\Big) + \phi_{\boldsymbol{n}}^\dagger \cdot \phi_{\boldsymbol{n}} \end{aligned} \tag{17.3.23}$$

reduces in the naive continuum limit to $|D_\mu\phi|^2$.

Important comment on notations: the "operator ordering" used above is probably the most natural one when building the lattice theory from continuum parallel transports, however from the purely lattice point of view a different convention can be (and often is) adopted. The definition of the plaquette

$$P_{\mu\nu}(\boldsymbol{n}) = \mathrm{ReTr}\Big(U_\mu(\boldsymbol{n})U_\nu(\boldsymbol{n}+\hat{\mu})U_\mu^\dagger(\boldsymbol{n}+\hat{\nu})U_\nu^\dagger(\boldsymbol{n})\Big) \tag{17.3.24}$$

is absolutely legitimate if we use the gauge transformation rule

$$U_\mu(\boldsymbol{n}) = G(\boldsymbol{n})U_\mu(\boldsymbol{n})G^\dagger(\boldsymbol{n}+\hat{\mu}) \tag{17.3.25}$$

instead of Eq. (17.3.3), corresponding to the definition $U_\mu(\boldsymbol{n}) = U_{\boldsymbol{n}\leftarrow\boldsymbol{n}+\mu}$ instead of $U_\mu(\boldsymbol{n}) = U_{\boldsymbol{n}+\mu\leftarrow\boldsymbol{n}}$. Using this convention the quantity $\phi_{\boldsymbol{n}}U_\mu(\boldsymbol{n})\phi_{\boldsymbol{n}+\mu}^\dagger$ is gauge invariant, while using Eq. (17.3.3) it is $\phi_{\boldsymbol{n}+\mu}^\dagger U_\mu(\boldsymbol{n})\phi_{\boldsymbol{n}}$ that is gauge invariant.

To close this section we still have to define the integration measure to be used for lattice gauge variables. As noted above, the fundamental lattice variables are $U_\mu(\boldsymbol{n})$, which are element of the (unitary representation of the) compact groups SU(N) or U(1), depending on the case considered (see later for further comments on the U(1) case). For all compact groups a "special" measure exists, the so called Haar measure, which is the only invariant measure of the group [108, 109, 110]: this means that for any g_0 we have

$$\mathrm{d}g = \mathrm{d}(g_0 g) = \mathrm{d}(g g_0) \ , \tag{17.3.26}$$

and this measure is typically normalized in such a way that $\int_G \mathrm{d}g = 1$. Some specific examples of the Haar measure will be used when discussing the heat-bath algorithm, but for now it is enough to note that the use of this integration measure presents two important advantages:

- it is consistent with gauge invariance:

$$\begin{aligned}
&\int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n})\right) e^{-S[U]} O[U] \overset{(1)}{=} \int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}{}^g U_\mu(\boldsymbol{n})\right) e^{-S[{}^g U]} O[{}^g U] \overset{(2)}{=} \\
&= \int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}\big(G(\boldsymbol{n}+\mu) U_\mu(\boldsymbol{n}) G^\dagger(\boldsymbol{n})\big)\right) e^{-S[U]} O[U] \overset{(3)}{=} \\
&= \int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n})\right) e^{-S[U]} O[U] \ ,
\end{aligned} \tag{17.3.27}$$

  where in step (1) we renamed $U \to {}^g U$, in step (2) we used the gauge invariance of the action functional $S[U]$ and of the observable O, and in step (3) we used the left and right invariance of the Haar measure.
- it allows the use of the Metropolis (and not Metropolis-Hastings) algorithm, since it is uniform on the group as a consequence of left and right invariance.

Let us explicitly note that, when the variables $U_\mu(\boldsymbol{n})$ are elements of a compact group, on a finite space time lattice all average values are well defined from the mathematical point of view, since the integration manifold has finite measure. When using instead a non-compact group several problems can arise:

- two different invariant measures can exist, one that is left invariant and one that is right invariant (see [114] §2.2 for a simple explicit example)
- even in a finite space time lattice, not all average values of gauge invariant quantities are well defined, unless peculiar boundary conditions are used (see, e. g., [87] for a non-compact U(1) formulation in which fundamental variables live in $\mathbb{R}$).

It is finally important to note that, since in the lattice formulation the fundamental variables $U_\mu(\boldsymbol{n})$ are elements of the group and not of the algebra, it is possible to study on the lattice also gauge theories with discrete gauge groups (like, e. g., the cyclic or the dihedral groups), which do not have a direct continuum counterpart. For finite groups the equivalent of the Haar measure is simply the sum on all the group elements (normalized by the order of the group, i. e., the number of group elements). For finite gauge groups the Wilson action is typically written in the form

$$S_E = \beta \sum_{\text{plaq}} \left(1 - \frac{1}{n} P_{\mu\nu}(\boldsymbol{n})\right) \ , \tag{17.3.28}$$

where n is the rank of the group representation, and $P_{\mu\nu}$ was defined in Eq. (17.3.10).

17.4 Lattice gauge theories: general properties

As we noted before, when we consider lattice gauge theories with compact gauge group on a finite space-time lattice all average values are mathematically well defined, and thus there is no need of introducing a gauge fixing (unless you want to use lattice perturbation theory, see, e. g., [81, 82]). We will however see that, in some cases, it is in fact useful to fix a gauge and reduce the number of degrees of freedom, e. g. to analytically solve two dimensional gauge theories. On the lattice, the most natural way of fixing a gauge is to use the gauge freedom

$$U_\mu(\boldsymbol{n}) \to G(\boldsymbol{n}+\hat{\mu}) U_\mu(\boldsymbol{n}) G^\dagger(\boldsymbol{n}) \tag{17.4.1}$$

to set to the identity some lattice gauge variables, e. g., using $G(\boldsymbol{n}) = U_\mu(\boldsymbol{n})$. In particular, it is simple to understand that we can fix to the identity all the links of a lattice tree (i. e. a set of

links in which no closed loops are present). It should instead be clear that we can not generically fix to the identity all the links of a closed loop, since the quantity

$$\mathrm{Tr}\left(\prod_{\text{loop}} U_\mu(\boldsymbol{n})\right) \tag{17.4.2}$$

is gauge invariant. In this regard it is important to note that, on a finite space-time lattice with periodic boundary conditions, we also have loops which wind around the lattice; it is thus not possible, e. g., to fix $U_0(\boldsymbol{n}) = 1$ on all sites $\boldsymbol{n}$.

It is simple to show that, if $f[U]$ is a functional which satisfies

$$\int \mathrm{d}g f[{}^g U] = 0 \ , \tag{17.4.3}$$

where g is a global (i. e. independent of $\boldsymbol{n}$) gauge transformation, then $\langle f \rangle = 0$. Indeed

$$\begin{aligned} \langle f \rangle = \int \mathrm{d}g \langle f \rangle = \frac{\int \mathrm{d}g \int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n})\right) f[U] e^{-S[U]}}{\int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n})\right) e^{-S[U]}} \overset{(1)}{=} \\ = \frac{\int \mathrm{d}g \int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}{}^g U_\mu(\boldsymbol{n})\right) f[{}^g U] e^{-S[{}^g U]}}{\int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n})\right) e^{-S[U]}} \overset{(2)}{=} \frac{\int \mathrm{d}g \int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n})\right) e^{-S[U]} \int \mathrm{d}g f[{}^g U]}{\int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n})\right) e^{-S[U]}} = 0 \ , \end{aligned} \tag{17.4.4}$$

where in step (1) we changed $U_\mu(\boldsymbol{n}) \to {}^g U_\mu(\boldsymbol{n})$, in step (2) we used the gauge invariance of the action and the properties of the Haar measure, and in the last step we finally used the hypothesis Eq. (17.4.3). Eq. (17.4.3) can appear difficult to verify, but it is sufficient that $f[U]$ transforms according to an irreducible representation of the gauge group for Eq. (17.4.3) to be satisfied: if $f[{}^g U] = R(g) f[U]$, then

$$\int \mathrm{d}g f[{}^g U] = \int \mathrm{d}g R(g) f[U] \ . \tag{17.4.5}$$

Using the invariance of the Haar measure we have, for any g_0,

$$\int \mathrm{d}g R(g) = \int \mathrm{d}(g_0 g) R(g_0 g) = R(g_0) \int \mathrm{d}g R(g) \tag{17.4.6}$$

hence $\int \mathrm{d}g R(g)$ is proportional to the projector on an invariant subspace of the representation. Since the representation is irreducible, this projector has to vanishes (the case $\int \mathrm{d}g R(g) = 1$ corresponds to the trivial representation, $R(g) = 1$ for any g), hence we conclude that Eq. (17.4.3) is satisfied. This fact shows, in practice, that the class of nontrivial local observables coincides with the class of gauge invariant observables. In particular, since under a global transformation the gauge variables $U_\mu(\boldsymbol{n})$ (as all parallel transports) transform according to the adjoint representation of the group, it follows that $\langle U_\mu(\boldsymbol{n}) \rangle = 0$ and $\langle [\text{any products of } U_\mu(\boldsymbol{n})] \rangle = 0$.

The previous result is the equivalent, in the present context, of the relation $\langle m \rangle = 0$, valid for the Ising model when $h = 0$ and periodic boundary conditions are used, which follows from the global $\mathbb{Z}_2$ invariance, see Sec. 5.1. In gauge theories a much stronger result holds, known as Elitzur theorem, which roughly states the impossibility of spontaneously breaking local gauge symmetries. A more precise statement is the following: let $f[U]$ be a functional which depends on a finite number (independent of the lattice size) of compact gauge variables, which under local gauge transformations satisfies

$$\int f[{}^g U] \prod_{\boldsymbol{n}} \mathrm{d}g(\boldsymbol{n}) = 0 \ . \tag{17.4.7}$$

If we explicitly break the gauge symmetry by adding to the Euclidean action a term of the form ($h_\mu(\boldsymbol{n})$ is a sort of external magnetic field)

$$\sum_{\boldsymbol{n},\mu} \mathrm{Re}\mathrm{Tr}\Big(h_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n})\Big) \ , \tag{17.4.8}$$

then we have

$$\lim_{h\to 0}\lim_{V\to\infty}\langle f\rangle_{V,h}=0\ . \tag{17.4.9}$$

Note that the order of the limits is the same which, in the case of the Ising model, was used to expose the spontaneous breaking of the global $\mathbb{Z}_2$ symmetry.

To prove Elitzur theorem we will follow the presentation in [45] §6.1.3. The starting point is the definition of $\langle f\rangle_{V,h}$:

$$\langle f\rangle_{V,h}=\frac{1}{Z_{V,h}}\int e^{-S[U]-\sum_{\boldsymbol{n},\mu}\mathrm{ReTr}(h_\mu(\boldsymbol{n})U_\mu(\boldsymbol{n}))}f[U]\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\ , \tag{17.4.10}$$

where

$$Z_{V,h}=\int e^{-S[U]-\sum_{\boldsymbol{n},\mu}\mathrm{ReTr}(h_\mu(\boldsymbol{n})U_\mu(\boldsymbol{n}))}\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\ . \tag{17.4.11}$$

To simplify the notation some indices will be implied in the following expressions, using the shorthand

$$h\cdot U=\sum_{\boldsymbol{n},\mu}\mathrm{ReTr}\Big(h_\mu(\boldsymbol{n})U_\mu(\boldsymbol{n})\Big)\ . \tag{17.4.12}$$

Let us denote collectively by U' the gauge variables in the support of the functional $f[U]$, and by U'' all the remaining ones (we analogously denote by h' and h'' the associated gauge breaking fields). By renaming in the numerator of $\langle f\rangle_{V,h}$ the U' variables, using $U'\to {}^gU'$, we get

$$\langle f\rangle_{V,h}=\frac{1}{Z_{V,h}}\int e^{-S[U]-h'\cdot{}^gU'-h''\cdot U''}f[{}^gU']\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\ , \tag{17.4.13}$$

where we exploited the invariance of the action under local gauge transformation[6] and the invariance properties of the Haar measure. Averaging on g we get

$$\langle f\rangle_{V,h}=\frac{1}{Z_{V,h}}\int e^{-S[U]-h''\cdot U''}\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\int f[{}^gU']e^{-h'\cdot{}^gU'}\prod_{\boldsymbol{n}}\mathrm{d}g(\boldsymbol{n})\ . \tag{17.4.14}$$

We can now note that, if the h' variables are small enough, we have

$$\Big|e^{-h'\cdot{}^gU'}-1\Big|\le\epsilon \tag{17.4.15}$$

uniformly in V (the total number of lattice sites), where ϵ is an arbitrarily small positive number. The previous relation holds true since the set of all U' is bounded, which is a consequence of the fact that the gauge group is compact and the support of the functional $f[U]$ consists of a finite (independent of V) number of gauge variables. We thus have, using Eq. (17.4.7) and the normalization of the Haar measure, the inequality

$$\begin{aligned}|\langle f\rangle_{V,h}|&=\frac{1}{Z_{V,h}}\left|\int e^{-S[U]-h''\cdot U''}\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\int f[{}^gU']\Big(e^{-h'\cdot{}^gU'}-1\Big)\prod_{\boldsymbol{n}}\mathrm{d}g(\boldsymbol{n})\right|\le\\&\le\epsilon\frac{\max f}{Z_{V,h}}\int e^{-S[U]-h''\cdot U''}\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\equiv\epsilon\frac{\max f}{Z_{V,h}}Z_{V,h'=0,h''}\ ,\end{aligned} \tag{17.4.16}$$

where we denoted by $Z_{V,h'=0,h''}$ the partition function computed by fixing $h'=0$. Carrying out analogous manipulations for the partition function we have

$$\begin{aligned}Z_{V,h}&=\int e^{-S[U]-h''\cdot U''}\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\int e^{-h'\cdot{}^gU'}\prod_{\boldsymbol{n}}\mathrm{d}g(\boldsymbol{n})=\\&=Z_{V,h'=0,h''}+\int e^{-S[U]-h''\cdot U''}\prod_{\boldsymbol{n},\mu}\mathrm{d}U_\mu(\boldsymbol{n})\int\Big(e^{-h'\cdot{}^gU'}-1\Big)\prod_{\boldsymbol{n}}\mathrm{d}g(\boldsymbol{n})\ ,\end{aligned} \tag{17.4.17}$$

[6] Note that this step can not be performed for global transformations.

hence

$$\left| Z_{V,h} - Z_{V,h'=0,h''} \right| \leq \epsilon Z_{V,h'=0,h''} \ , \tag{17.4.18}$$

Using this bound we find

$$|\langle f \rangle_{V,h}| \leq \frac{\epsilon}{1-\epsilon} \max f \ , \tag{17.4.19}$$

with ϵ independent of V, and finally

$$\lim_{h \to 0} \lim_{V \to \infty} \langle f \rangle_{V,h} = 0 \ . \tag{17.4.20}$$

Elitzur theorem can be extended, with practically no changes, to the case in which matter fields are also present. It is important to stress that this result *is not* inconsistent with the existence of the Higgs mechanism: Elitzur theorem states that the Higgs mechanism, in a nonperturbatively regularized and gauge-invariant setting, can not be related to the existence of a nonvanishing expectation value for local observables. Although this could seem at odds with the standard presentations of the Higgs mechanism, in fact it is not: in the standard discussions of the Higgs mechanism (see, e. g., [115] §21) a gauge-fixed theory is used, and local observables in a gauge-fixed theory typically correspond to non-local observables in the gauge invariant theory (see, e. g., [116] for the analogous case of charged gauge-invariant states in electrodynamics). The Higgs mechanism affects the spectrum of the theory, and we have seen in Sec. 14.1 that the spectrum of a theory can be determined by studying the large distance behavior of local observables correlations, not by studying local observables themselves. For further discussions of the Higgs mechanism in (mainly lattice) gauge theories see, e. g., [117, 118, 119], and [33] §C.II for a textbook presentation.

Elitzur theorem prevents the possibility of characterizing the phases of Yang-Mills theories by means of a local order parameter: no symmetry breaking can be used to characterize the phases, since no gauge symmetry breaking can happen and there are no other (internal) symmetries available beyond the gauge ones. This is no more true if matter fields are present: for example, if a N-component scalar field is coupled to a U(1) gauge field in such a way that the global symmetry is SU(N), the local operator $Q_{ij} = \phi_i^* \phi_j$ is gauge invariant and transforms according to the adjoint representation of SU(N). We can thus characterize different phases by the way in which the SU(N) symmetry is realized (see, e. g., [87], or [120] for the non-Abelian gauge case).

In Yang-Mills theories, i. e. gauge field theories without matter fields, different phases can nevertheless exist, characterized by the different behaviors of non-local observables. The simplest non-local observable is the Wilson loop, which is a generalization of the plaquette: a Wilson loop of size $w_t \times w_s$ is defined by

$$W(w_t, w_s) = \langle \mathrm{Tr} \left(\prod_C U_\mu(\boldsymbol{n}) \right) \rangle \ , \tag{17.4.21}$$

where the path-ordered product of gauge variables $\prod_C U_\mu(\boldsymbol{n})$ is carried out along a rectangular contour C of sides w_t and w_s. The importance of the Wilson loop lies in its relation with the so-called static potential: it can be shown (see, e. g., [81] §7 or [121]) that the static potential between two infinitely massive color[7] sources, at distance w_s from each other, is given by

$$V(w_s) = - \lim_{w_t \to \infty} \frac{1}{w_t} \log W(w_t, w_s) \ . \tag{17.4.22}$$

The different behaviors of the Wilson loop for large values of w_t and w_s can thus be related to different large distance behaviors of the static potential.

In some cases it happens that, for $w_t w_s \gg 1$, the Wilson loop behaves as

$$W(w_t, w_s) \propto e^{-\sigma w_t w_s} \ , \tag{17.4.23}$$

[7] The "color" terminology is typical of quantum chromodynamics. In the more general context of generic gauge theories color just refers to the degrees of freedom on which the gauge group acts.

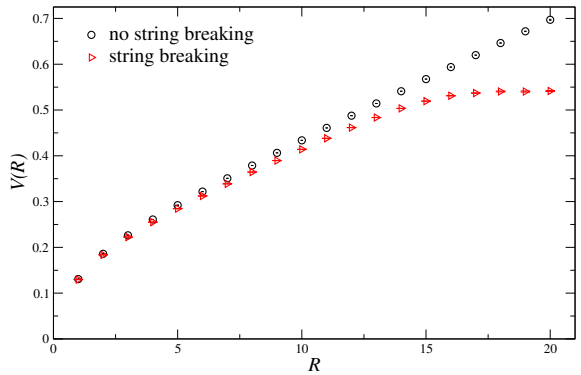

Figure 17.1: Static potential in three dimensional SU(2) gauge models (in lattice units): "no string breaking" points refer to the Yang-Mills theory, while "string breaking" points refer to a model in which scalar matter fields are coupled to the Yang-Mills theory. Adapted from [122].

a relation which is known as area-law. If this happens, the large distance behavior of the static potential is

$$\lim_{r\to\infty} V(r) \simeq \sigma r \ , \tag{17.4.24}$$

and the theory is said to be confining: an infinite amount of energy is needed to separate two static color sources. It may instead happen that the leading large distance behavior of the Wilson loop is described by the so-called perimeter-law

$$W(w_t, w_s) \propto e^{-\alpha(w_t+w_s)} \ , \tag{17.4.25}$$

in which case $\lim_{r\to\infty} V(r) = \alpha < \infty$, and the theory is not confining. It can be rigorously shown (see [123, 124]) that large Wilson loops can neither approach zero faster then Eq. (17.4.23) nor slower than Eq. (17.4.25). The parameter σ entering the area-law is called string tension, and it is a (non-local) order parameter for the confinement/deconfinement transition.

Note that in theories with matter fields transforming in the fundamental representation of the gauge group Wilson loops never obey the area-law: when the distance between the two color sources is increased beyond a critical value, a "two meson" state becomes energetically favorable with respect to the two separated unscreened color sources. At that point the linear grow of the static potential abruptly stops, a phenomenon know as string breaking, see Fig. (17.1).

It can be shown (see [45] §6.3) that any Yang-Mills theory (hence without matter fields) with compact gauge group and nontrivial center[8] confines for sufficiently small β values. Confinement is however not necessarily present for generic values of β, the most famous case being probably that of the three dimensional $\mathbb{Z}_2$ lattice gauge theory. It can be shown that this theory is dual to the three dimensional Ising model (see [125, 40], and also [45] §6.1), it displays a continuous phase transition at

$$\beta_c = -\frac{1}{2} \ln \tanh \beta_c^{\text{Ising}} \simeq 0.761413\ldots \tag{17.4.26}$$

(see [68] for β_c^{Ising}), which is of the 3D Ising universality class[9]. This model is confining for $\beta < \beta_c$ (consistently with the fact that $\mathbb{Z}_2$ is a compact group with nontrivial center) and not confining for $\beta > \beta_c$ [125]. A similar but less popular case is that of 4D (compact) U(1) Yang-Mills theory: this model is confining for small β values, however for large values of β this is no more true [126], and a (discontinuous) deconfinement phase transition happens for $\beta \approx 1.011$ (see, e. g., [127]). In

[8] We remind the reader that the center of the group is the set of elements (in fact the subgroup) that commute with every element of the group. The center is said to be trivial when it coincides with the identity element.

[9] To say that the transition is of the 3D Ising universality class is not completely appropriate, since in the $\mathbb{Z}_2$ gauge model all the magnetic/$\mathbb{Z}_2$-odd sector is missing. A more precise statement is that the singularity of the free energy is of the same form as that of the three dimensional Ising model without magnetic field.

several cases (e. g. in 3D and 4D SU(N) Yang-Mills theories) confinement is numerically observed for any $\beta > 0$, but a rigorous proof of this fact is still lacking.

In confining gauge theories, the string tension or, more generally, other properties of the static potential (e. g., the Sommer parameter, see [128] for the 4D SU(3) case), are commonly used to set the physical scale of lattice simulations, see the discussion in Sec. 14.2.

Chapter 18

Numerical simulation of lattice gauge theories

Let us remind that the Wilson action is

$$S_E = \beta \sum_{\text{plaq}} \left(1 - \frac{1}{N} \text{ReTr} \Pi_{\mu\nu}(\boldsymbol{n}) \right) \ , \tag{18.0.1}$$

where

$$\Pi_{\mu\nu}(\boldsymbol{n}) = U^\dagger_\nu(\boldsymbol{n}) U^\dagger_\mu(\boldsymbol{n}+\hat{\nu}) U_\nu(\boldsymbol{n}+\hat{\mu}) U_\mu(\boldsymbol{n}) \ . \tag{18.0.2}$$

Note that, written in this way, S_E takes the same form for the SU(N) gauge theory, for the U(1) gauge theory ($N = 1$ has to be used), and for the finite group cases (N is just the dimension of the representation). If our aim is to update the gauge variable $U_\mu(\boldsymbol{n})$, it is convenient to note that this variable enters only $2(D-1)$ plaquettes (in D space-time dimensions), and we can write

$$\begin{aligned} S_E &= -\frac{\beta}{N} \sum_{\nu \neq \mu} \text{ReTr} \Pi_{\mu\nu}(\boldsymbol{n}) - \frac{\beta}{N} \sum_{\nu \neq \mu} \text{ReTr} \Pi_{\mu\nu}(\boldsymbol{n} - \hat{\nu}) + \text{independent of } U_\mu(\boldsymbol{n}) = \\ &= -\frac{\beta}{N} \text{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] + \text{independent of } U_\mu(\boldsymbol{n}) \ , \end{aligned} \tag{18.0.3}$$

where we used $\text{ReTr} \Pi_{\mu\nu} = \text{ReTr} \Pi_{\nu\mu}$ and introduced

$$S_\mu(\boldsymbol{n}) = \sum_{\nu \neq \mu} \Big(U^\dagger_\nu(\boldsymbol{n}) U^\dagger_\mu(\boldsymbol{n}+\hat{\nu}) U_\nu(\boldsymbol{n}+\hat{\mu}) + U_\nu(\boldsymbol{n}-\hat{\nu}) U^\dagger_\mu(\boldsymbol{n}-\hat{\nu}) U^\dagger_\nu(\boldsymbol{n}-\hat{\nu}+\hat{\mu}) \Big) \ , \tag{18.0.4}$$

which is known as the sum of the "staples", for obvious geometrical reasons, see Fig. (18.1).

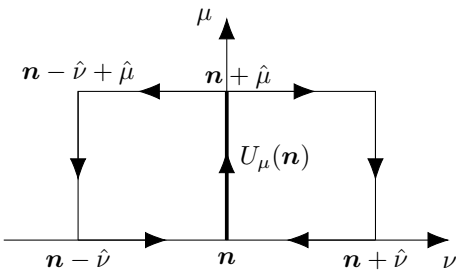

Figure 18.1: Graphical representation of the sum $S_\mu(\boldsymbol{n})$ of the staples associated with the gauge variable $U_\mu(\boldsymbol{n})$.

18.1 Metropolis update

A possible update scheme for lattice gauge theories consists in sweeping through the lattice (or randomly selecting sites and directions) and proposing the trial update

$$U_\mu(\boldsymbol{n}) \to \text{random matrix of the group} \ , \tag{18.1.1}$$

which is then accepted or rejected using a Metropolis test. The difference of Euclidean action ΔS_E associated with the trial update, needed in the Metropolist step, can obviously be computed using Eq. (18.0.3).

This update scheme presents however two important drawbacks. The first (less important) one is that the acceptance probability could be very small. The most important drawback is that, for the previous update scheme to be stochastically exact, the selection probability of the random matrices has to be uniform on the group, since otherwise the selection probability would not be symmetric and the Hastings correction would be needed, see Sec. 3.3.1. It is however generally nontrivial to generate matrices distributed uniformly on a group. This is typically possible only for finite groups (in which case it is sufficient select a random element of the group) or groups with very simple algebraic characterizations (like SU(2)). A simple way of of forcing the selection symmetry in the general case is the following:

1. generate a random matrix R of the group (not necessarily with uniform distribution)
2. generate the random number r in $[0, 1)$ with uniform pdf
3. use the trial update

$$U_\mu(\boldsymbol{n}) \to \begin{cases} RU_\mu(\boldsymbol{n}) & \text{if } r < 1/2 \\ R^\dagger U_\mu(\boldsymbol{n}) & \text{if } r \geq 1/2 \end{cases} \ . \tag{18.1.2}$$

Point (3) ensures that the trial selections $U_\mu(\boldsymbol{n}) \to V_\mu(\boldsymbol{n})$ and $V_\mu(\boldsymbol{n}) \to U_\mu(\boldsymbol{n})$ have the same probability, obviously assuming that the algorithm used to generate the random matrices does not change during the update.

It is important to note that gauge variables, after some update sweeps, have to be projected back on the group, in order to avoid the accumulation of rounding errors, analogously to the case of O(N) vector models discussed in Sec. 6.3. In the U(1) case it is sufficient to use

$$U_\mu(\boldsymbol{n}) \to \frac{U_\mu(\boldsymbol{n})}{|U_\mu(\boldsymbol{n})|} \ , \tag{18.1.3}$$

while for SU(N) matrices we have two possible alternatives. The Grahm-Schmidt algorithm is numerically quite unstable (see, e. g., [129] §5.2) but it is typically sufficient to correct rounding errors for small matrices. A more stable possibility is to find (e. g. using a pseudo-heat-bath update with $\beta = \infty$, see Sec. 18.3) the matrix $V_\mu(\boldsymbol{n}) \in \mathrm{SU}(N)$ which minimize

$$-\mathrm{ReTr}\Big(V_\mu(\boldsymbol{n}) U_\mu^\dagger(\boldsymbol{n})\Big) \ , \tag{18.1.4}$$

and then substitute $V_\mu(\boldsymbol{n}) \to U_\mu(\boldsymbol{n})$.

It is now convenient to analyze separately the cases of some gauge groups which are particularly useful in applications, for which specific techniques can be adopted to increase the efficiency of the update algorithm.

U(1) case

In this case we can generate R using $R = e^{i\theta}$, where θ is a random number with uniform pdf in $(-\epsilon, \epsilon)$. Note that, since the distribution of θ is symmetric with respect to zero, R and R^{-1} have the same probability of being chosen, and we do not have to force the symmetry using the steps (2) and (3) of the general algorithm described above. The parameter ϵ can be chosen at the beginning of the simulation in such a way to have a reasonable acceptance probability.

Since the complex exponential function is typically quite slow, on some hardware it can be more efficient to use instead the stereographic projection of $\mathbb{R}$ on the half-circumference to generate the random complex number R:

$$R = \frac{1 + i\theta}{\sqrt{1 + \theta^2}} \ . \tag{18.1.5}$$

This distribution is not uniform on the group, but it is still symmetric, and it is immediate to verify that $R(-\theta) = R^{-1}(\theta)$, which is enough for the selection probability to be symmetric if θ is selected with uniform pdf on a symmetric interval.

SU(2) case

It is convenient to use the parametrization of the SU(2) group

$$R = r_0 + i\boldsymbol{\sigma} \cdot \boldsymbol{r} \ , \tag{18.1.6}$$

where $r_\mu \in \mathbb{R}$ for $\mu = 0, \dots, 3$, and $\sum_\mu r_\mu^2 = 1$. We indeed have (using $\{\sigma_j, \sigma_k\} = 0$ and $\sigma_j^2 = 1$)

$$R^\dagger R = (r_0 - i\sigma_j r_j)(r_0 + i\sigma_k r_k) = r_0^2 + r_j r_k \sigma_j \sigma_k = \sum_\mu r_\mu^2 \ , \tag{18.1.7}$$

moreover from the explicit expression

$$R = \begin{pmatrix} r_0 + ir_3 & r_2 + ir_1 \\ -r_2 + ir_1 & r_0 - ir_3 \end{pmatrix} \tag{18.1.8}$$

it is immediate to verify that $\det R = \sum_\mu r_\mu^2$. Using this parametrization we thus see that the group SU(2) can be parametrized by the four dimensional sphere of unit radius S^3. To generate matrices uniformly distributed on SU(2) we can thus use the parametrization in Eq. (18.1.6) and Alg. (11) or Alg. (12) (see Sec. 6.3). If instead we are interested in generating a random matrix close to the identity, we can generate three real numbers a_i with uniform pdf in $(0, 1)$, then compute $b_i = 1 - 2a_i$ (which are uniformly distributed in $(-1, 1)$), and if $\sum_i b_i^2$ is nonvanishing[1] use

$$r_0 = \sqrt{1 - \epsilon} \ , \quad r_i = \frac{\sqrt{\epsilon}\, b_i}{\sqrt{\sum_{i=1}^3 b_i^2}} \ . \tag{18.1.9}$$

As in the U(1) case it is indeed simple to verify that $R(-\boldsymbol{b}) = R^\dagger(\boldsymbol{b})$, which ensures the symmetry of the selection probability.

SU(N) case

In this case it is possible to obtain a random SU(N) matrix close to the identity by multiplying SU(N) immersions of SU(2) random matrices, using a strategy analogous to the one put forward in [130] for the heat-bath. Let us consider for the sake of the simplicity the SU(3) case. Three "natural" ways exist to immerse a SU(2) matrix M in SU(3):

$$\begin{gathered} R^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & M_{11} & M_{12} \\ 0 & M_{21} & M_{22} \end{pmatrix} \ , \quad R^{(2)} = \begin{pmatrix} M_{11} & 0 & M_{12} \\ 0 & 1 & 0 \\ M_{21} & 0 & M_{22} \end{pmatrix} \ , \\ R^{(3)} = \begin{pmatrix} M_{11} & M_{12} & 0 \\ M_{21} & M_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix} \ , \end{gathered} \tag{18.1.10}$$

[1] More precisely: if it is larger than a fixed target accuracy.

and it is simple to understand that using these three SU(2) subgroups we can cover the whole SU(3). The generalization to SU(N) just need more SU(2)s, with $N(N-1)/2$ natural possibilities for the immersion.

If M is a SU(2) matrix that is generated in such a way that M and $M^\dagger$ are equiprobable, we can for example use $R = R^{(i)}$, where i is a random number in $\{1,2,3\}$ with uniform pdf, or $R = R^{i_1} R^{i_2} R^{i_3}$, where i_1, i_2, i_3 is a random permutation of $\{1,2,3\}$. These choices ensure the symmetry of the selection probability. If instead we use $R = R^{(1)} R^{(2)} R^{(3)}$, we need to chose R or $R^\dagger$ with equal probability to have a symmetric selection probability. A different possibility is to use sequentially $R^{(1)}$, $R^{(2)}$, and $R^{(3)}$, performing a Metropolis test after each multiplication; in this case detailed balance is not satisfied but balance is, see the discussion in Sec. 3.3.3.

18.2 Microcanonical update

If we want to update the gauge variable $U_\mu(\boldsymbol{n})$, and we are able to generate a new gauge variable U' in such a way that

- the Euclidean actions of the original and of the updated configurations are the same,
- the selection probability of the process $U_\mu(\boldsymbol{n}) \to U'$ is the same of the selection probability of the process $U' \to U_\mu(\boldsymbol{n})$,

the update $U_\mu(\boldsymbol{n}) \to U'$ is a legitimate Metropolis step, which is always accepted (see Sec. 6.3). A practical way of ensuring the symmetry of the selection probability is to use a deterministic procedure to generate U' starting from $U_\mu(\boldsymbol{n})$ which produces $U_\mu(\boldsymbol{n})$ if we start from U'.

U(1) case

If the absolute value of the sum of the staples $S_\mu(\boldsymbol{n})$ is nonvanishing, it is sufficient to use

$$U' = U^*_\mu(\boldsymbol{n}) \left(\frac{S^*_\mu(\boldsymbol{n})}{|S_\mu(\boldsymbol{n})|} \right)^2 . \tag{18.2.1}$$

We have indeed

$$S'_E = -\beta \mathrm{Re}\big(U' S_\mu(\boldsymbol{n})\big) + \text{ind. of } U' = -\beta \mathrm{Re}\big(U^*_\mu(\boldsymbol{n}) S^*_\mu(\boldsymbol{n})\big) + \text{ind. of } U_\mu(\boldsymbol{n}) = S_E \ , \tag{18.2.2}$$

and

$$(U')^* \left(\frac{S^*_\mu(\boldsymbol{n})}{|S_\mu(\boldsymbol{n})|} \right)^2 = U_\mu(\boldsymbol{n}) \left(\frac{S_\mu(\boldsymbol{n})}{|S_\mu(\boldsymbol{n})|} \right)^2 \left(\frac{S^*_\mu(\boldsymbol{n})}{|S_\mu(\boldsymbol{n})|} \right)^2 = U_\mu(\boldsymbol{n}) \ . \tag{18.2.3}$$

SU(2) case

A peculiarity of the SU(2) group is that the sum of SU(2) matrices is proportional to a SU(2) matrix, with a real and positive proportionality factor. This fact follows immediately from the parametrization in Eq. (18.1.6): a sum of SU(2) matrices can indeed be written as

$$a_0 + i\boldsymbol{\sigma} \cdot \boldsymbol{a} = \sqrt{a_0^2 + \boldsymbol{a}^2} \frac{a_0 + i\boldsymbol{\sigma} \cdot \boldsymbol{a}}{\sqrt{a_0^2 + \boldsymbol{a}^2}} \ , \tag{18.2.4}$$

where $\sqrt{a_0^2 + \boldsymbol{a}^2} \in \mathbb{R}^+$ and the remaining matrix is in SU(2). In particular we can write $S_\mu(\boldsymbol{n}) = \alpha V$, with $\alpha \in \mathbb{R}$ and $V \in$ SU(2), and we can easily verify that

$$U' = V^\dagger U^\dagger_\mu(\boldsymbol{n}) V^\dagger \tag{18.2.5}$$

is a legitimate microcanonical update [131]. We have indeed

$$\begin{aligned} \mathrm{ReTr}\big(U' S_\mu(\boldsymbol{n})\big) &= \mathrm{ReTr}\big(V^\dagger U^\dagger_\mu(\boldsymbol{n}) V^\dagger \alpha V\big) = \alpha \mathrm{ReTr}\big(V^\dagger U^\dagger_\mu(\boldsymbol{n})\big) = \\ &= \alpha \mathrm{ReTr}\big(U_\mu(\boldsymbol{n}) V\big) = \mathrm{ReTr}\big(U_\mu(\boldsymbol{n}) S_\mu(\boldsymbol{n})\big) \ , \end{aligned} \tag{18.2.6}$$

and

$$V^\dagger (U')^\dagger V^\dagger = V^\dagger V U_\mu(\boldsymbol{n}) V V^\dagger = U_\mu(\boldsymbol{n}) \ . \tag{18.2.7}$$

SU(N) case

We can also in this case work on SU(2) subgropus [130]. Let us consider for the sake of the simplicity just the case in which we consider the SU(2) subgroup corresponding to the first two rows and columns, since the same strategy can be applied also in all the other cases. Following the discussion in [13], we start by writing in a block-form the matrix $U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})$:

$$U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n}) = \left(\begin{array}{c|c} w & a \\ \hline c & b \end{array}\right) \ , \tag{18.2.8}$$

where w is a complex (in general not unitary) 2×2 matrix, and we are interested in performing the update $U_\mu(\boldsymbol{n}) \to RU_\mu(\boldsymbol{n})$, where

$$R = \left(\begin{array}{c|c} h & 0 \\ \hline 0 & 1 \end{array}\right) \ , \tag{18.2.9}$$

and $h \in \mathrm{SU}(2)$ is a matrix to be determined. The Euclidean action of the initial configuration is (neglecting terms independent of $U_\mu(\boldsymbol{n})$),

$$S_E = -\frac{\beta}{N}\mathrm{ReTr}\big(U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})\big) = -\frac{\beta}{N}\mathrm{ReTr}(w) - \frac{\beta}{N}\mathrm{ReTr}(b) \ , \tag{18.2.10}$$

while the Euclidean action after the update $U_\mu(\boldsymbol{n}) \to RU_\mu(\boldsymbol{n})$ is

$$S'_E = -\frac{\beta}{N}\mathrm{ReTr}\big(RU_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})\big) = -\frac{\beta}{N}\mathrm{ReTr}(hw) - \frac{\beta}{N}\mathrm{ReTr}(b) \ . \tag{18.2.11}$$

The matrix w can be written in the form $w = w_0 + i\boldsymbol{\sigma}\cdot\boldsymbol{w}$, where the coefficient w_μ are in general complex numbers, while for h we have $h = h_0 + i\boldsymbol{\sigma}\cdot\boldsymbol{h}$ with $h_\mu \in \mathbb{R}$ and $\sum_\mu h_\mu^2 = 1$. We thus have (using $\sigma_j\sigma_k = i\epsilon_{jkl}\sigma_l + \delta_{jk}$, $\mathrm{Tr}\sigma_j = 0$, and $h_\mu \in \mathbb{R}$)

$$\begin{aligned} &\mathrm{ReTr}(hw) = \mathrm{ReTr}\Big((h_0 + i\sigma_j h_j)(w_0 + i\sigma_k w_k)\Big) = \\ &= \mathrm{ReTr}\Big(h_0 w_0 - \boldsymbol{h}\cdot\boldsymbol{w}\Big) = \mathrm{ReTr}\Big(h_0\mathrm{Re}(w_0) - \boldsymbol{h}\cdot\mathrm{Re}(\boldsymbol{w})\Big) \ . \end{aligned} \tag{18.2.12}$$

If we introduce the notation

$$u = \mathrm{Re}(w_0) + i\boldsymbol{\sigma}\cdot\mathrm{Re}(\boldsymbol{w}) \ , \tag{18.2.13}$$

it is clear that we can write $u = \alpha V$, where $V \in \mathrm{SU}(2)$ and $\alpha \in \mathbb{R}$, hence

$$S'_E = -\frac{\alpha\beta}{N}\mathrm{ReTr}(hV) - \frac{\beta}{N}\mathrm{ReTr}(b) \ . \tag{18.2.14}$$

We are now ready to show that the choice

$$h = (V^\dagger)^2 \tag{18.2.15}$$

is the correct choice to be used in a microcanonical update. We have indeed

$$\alpha\mathrm{ReTr}(hV) = \alpha\mathrm{ReTr}(V^\dagger) = \alpha\mathrm{ReTr}(V) = \mathrm{ReTr}(u) = \mathrm{ReTr}(w) \ , \tag{18.2.16}$$

hence $S'_E = S_E$. To verify the reversibility we have to note that $h = \frac{1}{\alpha^2}(u^\dagger)^2$, hence (using $\mathrm{Re}(\boldsymbol{w}) \times \mathrm{Re}(\boldsymbol{w}) = 0$)

$$\begin{aligned} h &= \frac{1}{\sum_\mu(\mathrm{Re}w_\mu)^2}(\mathrm{Re}w_0 - i\boldsymbol{\sigma}\cdot\mathrm{Re}\boldsymbol{w})(\mathrm{Re}w_0 - i\boldsymbol{\sigma}\cdot\mathrm{Re}\boldsymbol{w}) = \\ &= \frac{1}{\sum_\mu(\mathrm{Re}w_\mu)^2}\left((\mathrm{Re}w_0)^2 - (\mathrm{Re}\boldsymbol{w})^2 - 2i(\mathrm{Re}w_0)\boldsymbol{\sigma}\cdot\mathrm{Re}\boldsymbol{w}\right) \ . \end{aligned} \tag{18.2.17}$$

If we denote by w' the equivalent of w after the update, $w' = hw$, a tedious but straightforward computation gives

$$w' = \frac{1}{\sum_\mu (\mathrm{Re}w_\mu)^2}\Big\{ w_0\big[(\mathrm{Re}w_0)^2 - (\mathrm{Re}\boldsymbol{w})^2\big] + 2\mathrm{Re}(w_0)\boldsymbol{w}\cdot\mathrm{Re}\boldsymbol{w} + \\ + i\big[(\mathrm{Re}w_0)^2 - (\mathrm{Re}\boldsymbol{w})^2\big]\boldsymbol{\sigma}\cdot\boldsymbol{w} - 2iw_0\mathrm{Re}(w_0)\boldsymbol{\sigma}\cdot\mathrm{Re}(\boldsymbol{w}) - 2i\boldsymbol{\sigma}\cdot\big(\mathrm{Re}(w_0)\boldsymbol{w}\times\mathrm{Re}\boldsymbol{w}\big)\Big\}\ , \tag{18.2.18}$$

and for the equivalent u' of u after the update we get $u' = u'_0 + i\boldsymbol{\sigma}\cdot\boldsymbol{u}'$, with $u'_0 = \mathrm{Re}w'_0$ and $\boldsymbol{u}' = -\mathrm{Re}\boldsymbol{w}'$. Explicitly

$$u' = \mathrm{Re}w_0 - i\boldsymbol{\sigma}\cdot\mathrm{Re}\boldsymbol{w} = u^\dagger\ , \tag{18.2.19}$$

and hence $V' = V^\dagger$. We thus have $h' = [(V')^\dagger]^2 = V^2$ and $h'h = 1$. This update is thus reversible, and this ensures that the selection probability is symmetric.

18.3 Heat-bath update

In the heat-bath update scheme, see Sec. 3.3.2, we update a gauge variable by sampling the conditional probability of $U_\mu(\boldsymbol{n})$ when all other gauge variables are kept fixed. In particular, also the sum of the staples $S_\mu(\boldsymbol{n})$ (see Eq. (18.0.4)) is to be considered as fixed.

U(1) case

The conditional probability of the gauge variable $U_\mu(\boldsymbol{n})$ when all the rest of the lattice is kept fixed is

$$P(U)\mathrm{d}U \propto \exp\big\{\beta\mathrm{Re}\big(US_\mu(\boldsymbol{n})\big)\big\}\,\mathrm{d}U\ , \tag{18.3.1}$$

where $\mathrm{d}U$ is the Haar measure on U(1). If we introduce $V \in U(1)$ by the relation $V = S_\mu(\boldsymbol{n})/|S_\mu(\boldsymbol{n})|$ we have thus

$$P(U)\mathrm{d}U \propto \exp\big\{\beta|S_\mu(\boldsymbol{n})|\mathrm{Re}\big(UV\big)\big\}\,\mathrm{d}U\ , \tag{18.3.2}$$

and if we define $u = UV$ and $\alpha = \beta|S_\mu(\boldsymbol{n})|$ we get, using the invariance of the Haar measure,

$$P(u)\mathrm{d}u \propto \exp\big(\alpha\mathrm{Re}u\big)\mathrm{d}u\ . \tag{18.3.3}$$

If we finally parametrize the $U(1)$ variable u by $u = e^{i\theta}$, we can write the Haar measure on U(1) as $\mathrm{d}u = \frac{1}{2\pi}\mathrm{d}\theta$, where $\mathrm{d}\theta$ is the usual Lebesgue measure on $\theta \in [-\pi,\pi)$. We thus have to sample the probability distribution

$$P(\theta)\mathrm{d}\theta \propto e^{\alpha\cos\theta}\mathrm{d}\theta\ . \tag{18.3.4}$$

An algorithm to sample this distribution, which uses a change of variable as a first step and von Neumann accept/reject step to correct the result, is described in Ref. [132].

If $\beta \gtrsim 1$ it is also possible to use a special Metropolis-Hastings algorithm, whose results are in practice impossible to distinguish from those obtained by using the heat-bath algorithm. Since for $\alpha \gg 1$ we have approximately

$$\exp(\alpha\cos\theta) \simeq \exp\left[\alpha\left(1 - \frac{1}{2}\theta^2\right)\right]\ , \tag{18.3.5}$$

to update $U_\mu(\boldsymbol{n}) = V^* e^{i\theta_{old}}$ we can start by generating $\theta_{new} \in [-\pi,\pi)$ with pdf proportional to $\exp(-\frac{\alpha}{2}\theta^2_{new})$. This can be done by using the Box-Muller, in the "original" version on $(-\infty,\infty)$ (see Sec. 2.3) rejecting draws with $|\theta| > \pi$, or modifying it to directly sample Gaussian variables in $(-\pi,\pi)$. Once θ_{new} is generated, we have to accept or reject the update $U_\mu(\boldsymbol{n}) \to U' = V^* e^{i\theta_{new}}$ with probability (see Sec. 3.3.1)

$$\frac{A_{\theta_{old}}P(U')}{A_{\theta_{new}}P(U_\mu(\boldsymbol{n}))} = \exp\Big\{-\frac{\alpha}{2}(\theta^2_{old} - \theta^2_{new}) + \alpha\Big(\cos\theta_{new} - \cos\theta_{old}\Big)\Big\}\ , \tag{18.3.6}$$

which for large α is very close to 1. With respect to the general case discussed in Sec. 3.3.1, in the present case the selection probability for θ_{new} is independent of its previous value, hence we used the shorthand A_b for the probability of selecting b, instead of A_{ba}, which denoted the probability of selecting b starting from a in Sec. 3.3.1.

SU(2) case

To write a heat-bath update for SU(2) Yang-Mills theory is once again convenient to use the parametrization in Eq. (18.1.6) of a SU(2) matrix: $U = u_0 + i\boldsymbol{u}\cdot\boldsymbol{\sigma}$, with $u_\mu \in \mathbb{R}$ and $\sum_\mu u_\mu^2 = 1$. The fact that group elements are associated with points of the unit sphere in four dimension suggests the invariant measure of the group to be proportional to

$$\delta\left(\sum_\mu u_\mu^2 - 1\right)\prod_\mu \mathrm{d}u_\mu \ . \tag{18.3.7}$$

This fact can be explicitly verified by considering another SU(2) matrix $M = m_0 + i\boldsymbol{m}\cdot\boldsymbol{\sigma}$, with $m_\mu \in \mathbb{R}$ and $\sum_\mu m_\mu^2 = 1$. Using $\sigma_j\sigma_k = \delta_{jk} + i\epsilon_{jkl}\sigma_l$ we have indeed

$$U' = MU = (m_0 + i\boldsymbol{m}\cdot\boldsymbol{\sigma})(u_0 + i\boldsymbol{u}\cdot\boldsymbol{\sigma}) = m_0u_0 - \boldsymbol{m}\cdot\boldsymbol{u} + i(u_0m_l + m_0u_l - \epsilon_{ikl}m_ju_k)\sigma_l \ , \tag{18.3.8}$$

hence $U' = u_0' + i\boldsymbol{u}'\cdot\boldsymbol{\sigma}$, with

$$u_0' = m_0u_0 - \boldsymbol{m}\cdot\boldsymbol{u} \ , \quad \boldsymbol{u}' = u_0\boldsymbol{m} + m_0\boldsymbol{u} - \boldsymbol{m}\times\boldsymbol{u} \ . \tag{18.3.9}$$

Using these relations it is immediate to see that

$$\frac{\partial u'}{\partial u} = \begin{pmatrix} m_0 & -m_1 & -m_2 & -m_3 \\ m_1 & m_0 & m_3 & -m_2 \\ m_2 & -m_3 & m_0 & m_1 \\ m_3 & m_2 & -m_1 & m_0 \end{pmatrix} \ , \quad \det\left(\frac{\partial u'}{\partial u}\right) = (m_0^2 + m_1^2 + m_2^2 + m_3^2)^2 = 1 \ , \tag{18.3.10}$$

hence

$$\delta\left(\sum_\mu u_\mu^2 - 1\right)\prod_\mu \mathrm{d}u_\mu = \delta\left(\sum_\mu {u'}_\mu^2 - 1\right)\prod_\mu \mathrm{d}u_\mu' \ . \tag{18.3.11}$$

The conditional probability of the gauge variable $U_\mu(\boldsymbol{n})$ when all the rest of the lattice is kept fixed is

$$P(U)\mathrm{d}U \propto \exp\left\{\frac{\beta}{2}\mathrm{ReTr}\Big(US_\mu(\boldsymbol{n})\Big)\right\} \tag{18.3.12}$$

As already noted, in SU(2) the sum of the staples is proportional to a SU(2) matrix, and the proportionality constant is a real and positive number: $S_\mu(\boldsymbol{n}) = \alpha V$, $V \in \mathrm{SU}(2)$, $\alpha = \sqrt{\det S_\mu(\boldsymbol{n})} \in \mathbb{R}$. We thus have

$$P(U)\mathrm{d}U \propto \exp\left\{\frac{\alpha\beta}{2}\mathrm{ReTr}\Big(UV\Big)\right\} \ , \tag{18.3.13}$$

and if we define $u = UV$ we get, using the invariance of the Haar measure,

$$P(u)\mathrm{d}u \propto e^{\alpha\beta u_0}\delta\left(\sum_\mu u_\mu^2 - 1\right)\prod_\mu \mathrm{d}u_\mu \ . \tag{18.3.14}$$

If we now separate in the integral the 0-th component from the other three components, introduce polar coordinates in the three dimensional integration, and use the transformation properties of the δ distribution to write $\delta(|\boldsymbol{u}|^2 + u_0^2 - 1)$ as a function of $|\boldsymbol{u}|$, we find

$$P(u)\mathrm{d}u \propto e^{\alpha\beta u_0}\frac{\delta\Big(|\boldsymbol{u}| - \sqrt{1-u_0^2}\Big)}{2\sqrt{1-u_0^2}}|\boldsymbol{u}|^2\mathrm{d}|\boldsymbol{u}|\mathrm{d}u_0\mathrm{d}\Omega_2 \ , \tag{18.3.15}$$

and by integrating on $|\boldsymbol{u}|$ we get

$$P(u)\mathrm{d}u \propto e^{\alpha\beta u_0}\sqrt{1-u_0^2}\,\mathrm{d}u_0\mathrm{d}\Omega_2 \ , \tag{18.3.16}$$

where $u_0 \in [-1,1]$ and $\mathrm{d}\Omega_2$ is the infinitesimal solid angle in three dimensions. We thus have to identify a random direction on the sphere, which can be easily done using Alg. (11) (see Sec. 6.3), and generate $u_0 \in [-1,1]$ with pdf

$$p(u_0)\mathrm{d}u_0 \propto e^{\alpha\beta u_0}\sqrt{1-u_0^2}\,\mathrm{d}u_0 \ . \tag{18.3.17}$$

An algorithm to sample this probability distribution has been introduced in [11], and has already been discussed in Sec. 2.4. A more efficient algorithm to sample $p(u_0)\mathrm{d}u_0$ when $\alpha\beta \gg 1$ is discussed in [13].

Note that from Eq. (18.3.13) it easily follows that in the $\beta \to \infty$ limit (in which the update corresponds to a minimization of the Euclidean action) $U = V^\dagger$ is the only possible outcome.

SU(N) case

A heat-bath for the SU(3) Yang-Mills theory exists (see [133]), however this algorithm is quite complex, and its generalization to SU(N) models with $N > 3$ is still more complex. In these cases it is convenient to use the so called pseudo-heat-bath algorithm introduced in [130]. The fundamental idea of this algorithm is to select an immersion of SU(2) in SU(N), and use the immersion of a SU(2) matrix, that will be denoted by R (see Eq. (18.1.10) for the SU(3) case), to perform the update $U_\mu(\boldsymbol{n}) \to U' = RU_\mu(\boldsymbol{n})$; the SU(2)-like matrix R is drawn in such a way that U' is sampled from the conditional probability of the gauge variable $U_\mu(\boldsymbol{n})$ when all the rest of the lattice is kept fixed. Since the Haar measure of SU(N), with $N > 2$, is obviously invariant under SU(2) transformations, in this way we are effectively sampling the links accorting to a heat-bath algorithm restricted to a specific subgroup of the gauge group. This can be done by using Eq. (18.2.14) and the SU(2) heat-bath algorith. By using enough different immersions of SU(2) in SU(N), see Sec. 18.1, we can then sample the whole SU(N) group.

The pseudo-heat-bath update can also be used to project-back on SU(N) gauge variables, in order to prevent the accumulation of rounding errors. Let us assume that we have to project on $SU(N)$ the matrix M. A possible projection strategy is to find the matrix $P \in$ SU(N) which maximizes $\mathrm{ReTr}(P^\dagger M)$. If we define $U = P^\dagger$, to find U (and thus P) is equivalent to perform an heath-bath update at $\beta = \infty$ of the link U, which is initially equal to 1, with sum of staples M. We thus have to iterate deterministic $\beta = \infty$ SU(2) heat-bath on several SU(2) subgroups of SU(N), until a terminating condition like $\|U_{(n+1)} - U_{(n)}\| < \epsilon$ is met, where $U_{(n)}$ is the result of the n-th iteration.

18.4 Hybrid Monte Carlo update

The HMC algorithm discussed in Sec. 16.2 can obviously be adopted only for continuous gauge groups, and it is convenient to start discussing the U(1) case, which is simpler than the SU(N) case. For what concerns the HMC, the SU(2) case is not particularly simpler than the general SU(N) case, the only simplification being that some matrix functions can be immediately written in a compact way by using the Pauli matrices. Note that for Yang-Mills theories, in which Metropolis, heat-bath and microcanonical updates are available, it is typically not convenient to use the HMC algorithm, unless very peculiar representations are used. As previously discussed, the HMC algorithm is instead a forced choice when fermions are present.

U(1) case

If we write the gauge variables as $U_\mu(\boldsymbol{n}) = e^{i\theta_\mu(\boldsymbol{n})}$, we can associate the conjugate momentum $p_\mu(\boldsymbol{n})$ with the variable $\theta_\mu(\boldsymbol{n})$, and write the Hamiltonian

$$H = \frac{1}{2}\sum_{\boldsymbol{n},\mu} p_\mu^2(\boldsymbol{n}) - \beta \sum_{\boldsymbol{n},\mu>\nu} \mathrm{Re}\Big(\Pi_{\mu\nu}(\boldsymbol{n})\Big) \ . \tag{18.4.1}$$

The first equation of motion is

$$\dot{\theta}_\mu(\boldsymbol{n}) = \frac{\partial H}{\partial p_\mu(\boldsymbol{n})} = p_\mu(\boldsymbol{n}) \ , \tag{18.4.2}$$

from which we get the elementary gauge variable evolution $e^{ip_\mu(\boldsymbol{n})\mathrm{d}\tau}U_\mu(\boldsymbol{n})$. The other equation is (no sum on μ)

$$\dot{p}_\mu(\boldsymbol{n}) = -\frac{\partial H}{\partial \theta_\mu(\boldsymbol{n})} = \beta\frac{\partial}{\partial\theta_\mu(\boldsymbol{n})}\mathrm{Re}\Big(U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})\Big) = \beta\mathrm{Re}\Big(iU_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})\Big) \ , \tag{18.4.3}$$

which using the notation $V = U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{b})$ can be written as

$$\dot{p}_\mu(\boldsymbol{n}) = \beta\mathrm{Re}(iV) = -\beta\mathrm{Im}V = -\beta\frac{V-V^*}{2i} \ , \tag{18.4.4}$$

and finally

$$i\dot{p}_\mu(\boldsymbol{n}) = -\beta\frac{U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n}) - U_\mu^*(\boldsymbol{n})S_\mu^*(\boldsymbol{n})}{2} \ . \tag{18.4.5}$$

The elementary integration steps of $U_\mu(\boldsymbol{n})$ and $p_\mu(\boldsymbol{n})$ have obviously to be combined using a symmetric symplectic integrator, as discussed in Sec. 16.2.

SU(N) case

In the SU(N) case we have to understand how to define momenta, and the simplest possibility is to decide that the momenta are the generator of the left-evolution of gauge variables, hence the elementary evolution of $U_\mu(\boldsymbol{n})$ is

$$U_\mu(\boldsymbol{n}) \to \exp\Big(ip_\mu^a(\boldsymbol{n})T_a\mathrm{d}\tau\Big)U_\mu(\boldsymbol{n}) \ , \tag{18.4.6}$$

and the Hermitian matrix associated with the momenta is $P_\mu(\boldsymbol{n}) = p_\mu^a(\boldsymbol{n})T_a$, where T_a are the SU(N) generators. Following the same logic it is convenient to introduce the left derivative in position $\boldsymbol{n}$, μ of a functional of the gauge fields:

$$\partial_{\boldsymbol{n},\mu,a} f[U] = \frac{\mathrm{d}}{\mathrm{d}\epsilon} f\Big[U_\mu(\boldsymbol{n}) \to e^{i\epsilon T^a}U_\mu(\boldsymbol{n})\Big]\Big|_{\epsilon=0} \ . \tag{18.4.7}$$

This derivative has to be used in the equations of motions associated with the Hamiltonian

$$H = \frac{1}{2}\sum_{\boldsymbol{n},\mu,a}(p_\mu^a)^2 - \frac{\beta}{N}\sum_{\boldsymbol{n},\mu>\nu}\mathrm{ReTr}\Pi_{\mu\nu}(\boldsymbol{n}) \ . \tag{18.4.8}$$

For a precise discussion of these definitions, which requires some differential geometry, see [134].

The exponential of the momenta can be easily computed in SU(2), using the properties of the Pauli matrices, while for $N > 2$ one simple possibility is to use the Taylor series of the exponential, truncated to a given order and then projected on SU(N). This procedure can be carried out since the integration time step $\mathrm{d}\tau$ is small, and a small number of terms in the Taylor expansion will likely be sufficient. Note however that this computation has to be performed with good accuracy, since otherwise the reversibility of the integration algorithm would be compromised. A

different possibility is to use the Cayley-Hamilton theorem to resum the Taylor series of the matrix exponential, see [135] for the SU(3) case.

The equation of motion of the momentum $p_\mu^a(\boldsymbol{n})$ is

$$\dot{p}_\mu^a(\boldsymbol{n}) = -\partial_{\boldsymbol{n},\mu,a} H = \frac{\beta}{N}\partial_{\boldsymbol{n},\mu,a}\mathrm{ReTr}\Big(U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})\Big) = \frac{\beta}{N}\mathrm{ReTr}\Big(iT_a U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})\Big) \ . \tag{18.4.9}$$

For any matrix M we have, using $(\mathrm{Tr}A)^* = \mathrm{Tr}(A^\dagger)$ and $T_a = T_a^\dagger$, the identity

$$\mathrm{ReTr}(T_a M) = \mathrm{ReTr}\left(T_a \frac{M + M^\dagger}{2}\right) \ . \tag{18.4.10}$$

If we now introduce the notation $V = (M + M^\dagger)/2$, the matrix V is obviously Hermitian, hence we can write it in the form $V_0 + \sum_a V_a T_a$, with $v_\mu \in \mathbb{R}$ and $\mu = 0, \ldots, N^2 - 1$. Using $\mathrm{Tr}T_a = 0$ and $\mathrm{Tr}(T_a T_b) = \frac{1}{2}\delta_{ab}$ we thus have

$$\sum_a T_a \mathrm{ReTr}(T_a M) = \sum_a T_a \mathrm{ReTr}(T_a V) = \sum_a \frac{1}{2} V_a T_a = \frac{1}{2}\left(V - \frac{1}{N}\mathrm{Tr}V\right) \ . \tag{18.4.11}$$

Using $M = iU_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})$, and the notation

$$[W]_{TA} = \frac{W - W^\dagger}{2} - \frac{1}{N}\mathrm{Tr}\left(\frac{W - W^\dagger}{2}\right) \tag{18.4.12}$$

for the traceless anti-Hermitian part of W, the equation of motion of the matrix $P_\mu(\boldsymbol{n})$ can finally be written in the form

$$\dot{P}_\mu(\boldsymbol{n}) = \sum_a \dot{p}_\mu^a(\boldsymbol{n}) T_a = i\frac{\beta}{2N}[U_\mu(\boldsymbol{n})S_\mu(\boldsymbol{n})]_{TA} \ , \tag{18.4.13}$$

in complete analogy with the U(1) case.

18.5 Error reduction techniques

We have seen in Sec. 17.4 that Wilson loops are important observables in gauge theories, since their large size behavior is related to the potential energy of two static sources, and in particular to the confining properties of the theory. To study large Wilson loops is however a challenging numerical task, especially in confined phases: the average values of Wilson loops quickly approach zero as the size of the loops is increased, and it becomes increasingly difficult to obtain a result that is not compatible with zero within statistical errors. It is thus important to use error reduction techniques, and in particular improved observables. An improved observable A_{imp} is an observable which has the same expectation value of A but a smaller variance; as a consequence it is convenient to look at A_{imp} instead of A in numerical simulations. A simple and general way of building improved observables for Wilson loops is the multihit method introduced in [136]. The basic idea of the method is to use local averages of links to reduce the noise of the Wilson loop.

Before discussing Wilson loops in lattice gauge theory, let us study a simple example to understand why this strategy is effective. Let us imagine to be interested in estimating by Monte Carlo methods the expectation value of the product of L independent random variables x_i, which for the sake of the simplicity we assume to be identically distributed and with zero average. Clearly the expectation value $\langle x_1 \cdots x_L \rangle$ vanishes, but we want to find the most effective way of estimating numerically this expectation value, i. e., the procedure which minimize the statistical error for fixed computation time. If we denote by σ^2 the variance of x_i, it is immediate to see that the variance of $x_1 \cdots x_L$ is σ^{2L}, hence if we draw $N \gg 1$ samples from $x_1, \ldots, x_L$, the statistical error of

$$\overline{x_1 \cdots x_L} = \frac{1}{N}\sum_{i=1}^{N} x_1^{(i)} \cdots x_L^{(i)} \ , \tag{18.5.1}$$

where $i = 1, \dots N$ labels the draw, is

$$\sigma_A = \frac{\sigma^L}{\sqrt{N}} \ . \tag{18.5.2}$$

We now want to verify that $\overline{x_1} \cdots \overline{x_L}$ is an improved estimator for the same quantity, where

$$\overline{x_k} = \frac{1}{N} \sum_i x_k^{(i)} \ . \tag{18.5.3}$$

The quantities $\overline{x_k}$ have zero average and variance σ^2/N, hence the statistical error of $\overline{x_1} \cdots \overline{x_L}$, using the same number of draws used before, is

$$\sigma_B = \left(\frac{\sigma}{\sqrt{N}} \right)^L = \frac{\sigma_A}{N^{(L-1)/2}} = \sigma_A \exp\left(-\frac{L-1}{2} \log(N) \right) \ . \tag{18.5.4}$$

We thus see that the error obtained by using the second method (the one with the "local averages") is smaller than the *naive* one by a quantity which is exponential in L. Moreover, at fixed L, also the scaling with N is much more favorable. What makes this example trivial is the fact that all the variables are independent from each other, but in numerical simulation gauge fields associated with different points and directions are not independent from each other, hence more care is required. The locality of the Wilson action will be the fundamental ingredient to generalize the local averaging procedure to lattice gauge theories.

Let us write a Wilson loop in the form

$$W(w_t, w_s) = \langle \mathrm{Tr} \left(\prod_C U_\nu(\boldsymbol{m}) \right) \rangle = \langle \mathrm{Tr} \Big(U_\mu(\boldsymbol{n}) R[\check{U}] \Big) \rangle \ , \tag{18.5.5}$$

where $R[\check{U}]$ is the path-ordered product of all the links of the Wilson loop different from $U_\mu(\boldsymbol{n})$, and we collectively denote by $\check{U}$ all the lattice gauge variables different from $U_\mu(\boldsymbol{n})$. We can now exploit the locality of the Wilson action to write

$$S_E[U] = -\frac{\beta}{N} \mathrm{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] + \check{S}_E[\check{U}] \ , \tag{18.5.6}$$

where $\hat{S}_E$ is the part of the Euclidean action independent of $U_\mu(\boldsymbol{n})$. Note that also the sum of the staples $S_\mu(\boldsymbol{n})$ depends on $\check{U}$. Putting everything together we have

$$\langle \mathrm{Tr} \Big(U_\mu(\boldsymbol{n}) R[\check{U}] \Big) \rangle = \frac{\int \mathrm{d}U_\mu(\boldsymbol{n}) \mathrm{d}\check{U} \ \mathrm{Tr} \Big(U_\mu(\boldsymbol{n}) R[\check{U}] \Big) \exp\left(\frac{\beta}{N} \mathrm{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] - \check{S}_E[\check{U}] \right)}{\int \left(\prod_{\boldsymbol{m},\nu} \mathrm{d}U_\nu(\boldsymbol{m}) \right) \exp\left(S_E[U] \right)} \ ,$$

If we now define

$$\overline{U_\mu(\boldsymbol{n})} = \frac{\int \mathrm{d}U_\mu(\boldsymbol{n}) \, U_\mu(\boldsymbol{n}) \exp\left(\frac{\beta}{N} \mathrm{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] \right)}{\int \mathrm{d}U_\mu(\boldsymbol{n}) \exp\left(\frac{\beta}{N} \mathrm{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] \right)} \tag{18.5.7}$$

it is immediate to see that (since $\overline{U_\mu(\boldsymbol{n})}$ does not depend on $U_\mu(\boldsymbol{n})$)

$$\begin{aligned}
&\int \mathrm{d}U_\mu(\boldsymbol{n}) \mathrm{d}\check{U} \ \mathrm{Tr} \Big(\overline{U_\mu(\boldsymbol{n})} R[\check{U}] \Big) \exp\left(\frac{\beta}{N} \mathrm{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] - \check{S}_E[\check{U}] \right) = \\
&= \int \mathrm{d}\check{U} \ \mathrm{Tr} \Big(\overline{U_\mu(\boldsymbol{n})} R[\check{U}] \Big) \exp\left(-\check{S}_E[\check{U}] \right) \int \mathrm{d}U_\mu(\boldsymbol{n}) \exp\left(\frac{\beta}{N} \mathrm{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] \right) = \\
&= \int \mathrm{d}U_\mu(\boldsymbol{n}) \mathrm{d}\check{U} \ \mathrm{Tr} \Big(U_\mu(\boldsymbol{n}) R[\check{U}] \Big) \exp\left(\frac{\beta}{N} \mathrm{ReTr} \Big[S_\mu(\boldsymbol{n}) U_\mu(\boldsymbol{n}) \Big] - \check{S}_E[\check{U}] \right) \ ,
\end{aligned} \tag{18.5.8}$$

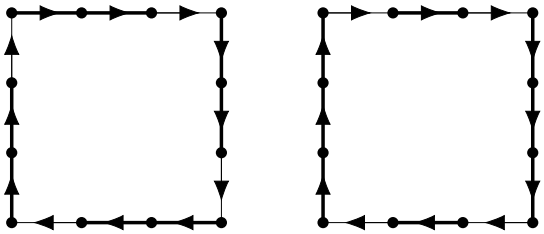

Figure 18.2: Possible choices of gauge variables which can be simultaneously averaged (thick lines) while computing a 3×3 Wilson loop.

hence

$$\langle \mathrm{Tr}\Big(U_\mu(\boldsymbol{n}) R[\check{U}]\Big)\rangle = \langle \mathrm{Tr}\Big(\overline{U_\mu(\boldsymbol{n})}\, R[\check{U}]\Big)\rangle \ . \tag{18.5.9}$$

This shows that by using the local average of a single gauge variable we obtain an estimator which has the same expectation value of the original one. That this estimator is indeed an improved estimator should be intuitively clear given the example with independent variables discussed before.

This local averaging procedure can be applied to different gauge variables as far as all local averages can be performed independently, i. e., as far as one of the two variables does not enter the sum of the staples of the other variable. The choice of which gauge variables to average is generally not unique, and in Fig. (18.2) two different possibilities are shown for the case of a 3×3 Wilson loop. Explicit expressions for the local averages exists [137], however it is typically more convenient to estimate them by using Metropolis, heat-bath or microcanonical updates:

$$\overline{U_\mu(\boldsymbol{n})} = \frac{1}{M} \sum_{i=1}^{M} U_\mu^{(i)}(\boldsymbol{n}) \ , \tag{18.5.10}$$

where $U_\mu^{(i)}(\boldsymbol{n})$ is the result of the i-th update at fixed sum of staples. This procedure can be used since the sample average is an unbiased estimator of the true average.

In the multihit technique only single gauge variables are locally averaged, but it is also possible to use local averages of the product of gauge links, like $\overline{U_\nu(\boldsymbol{n}+\hat{\mu})U_\mu(\boldsymbol{n})}$, and different averages can be recursively nested in order to increase the effectiveness of the error reduction, using e. g.

$$\overline{\overline{U_\mu(\boldsymbol{n}+\hat{\mu})}\ \overline{U_\mu(\boldsymbol{n})}} \ , \tag{18.5.11}$$

where different neighbor variables have to be kept fixed when averaging the single links or a product of links. This general point of view is discussed in [138], where the so-called multilevel algorithms have been introduced.

Chapter 19

Two dimensional U(1) gauge theory

19.1 θ-dependence

Two dimensional gauge theories are peculiar gauge theories, in which several computations can be carried out analytically or almost analytically. It is in particular possible to have complete control on several nonperturbative phenomena, like, e. g., confinement. In two dimensional U(N) theories θ-dependence is also present, which is the QFT analogue of the phenomenon discussed for QM in Chap. 11. Note that, while a nontrivial topology of the configuration space was fundamental for the existence of this phenomenon in QM, what really matters in QFT is the topology of the gauge group.

In this section we present a simple semiclassical argument for the existence of θ-dependence in two (space-time) dimensional U(1) gauge theory, that will be confirmed by the analytical solution of the lattice model in the next section. A canonical quantization approach to this system can be found, e. g., in [139, 140], which has the advantage of directly showing the analogy with the case of the QM particle on a circumference. The standard semiclassical approach to θ-dependence in four dimensional SU(N) theories is discussed in [141] §7, [115] §23, [46] §41, while the canonical approach has been introduced and developed in [142, 143], see [83] §4-5 for a textbook presentation.

The semiclassical approach starts from identifying the configurations with finite Euclidean action. For the Euclidean action to be finite, we have to require $F_{\mu\nu}(\boldsymbol{x}) \to 0$ for $|\boldsymbol{x}| \to \infty$. This means that, for $|\boldsymbol{x}| \to \infty$, the gauge field $A_\mu(\boldsymbol{x})$ reduces to a trivial gauge field ($A_\mu(\boldsymbol{x}) = 0$) up to gauge transformations. Using Eq. (17.2.7) in the Abelian case, with $G(g(\boldsymbol{x})) = e^{i\Lambda(\boldsymbol{x})}$, we have

$$ {}^{g}A_\mu = A_\mu + \frac{i}{e}(\partial_\mu G)G^\dagger = A_\mu - \frac{1}{e}\partial_\mu \Lambda \ , \tag{19.1.1} $$

hence finite action configurations correspond to

$$ \lim_{|\boldsymbol{x}|\to\infty} A_\mu(\boldsymbol{x}) \to \frac{1}{e}\partial_\mu \Lambda \ . \tag{19.1.2} $$

If we now require the gauge transformation $G(g(\boldsymbol{x})) = e^{i\Lambda(\boldsymbol{x})}$ to be globally well defined, we obtain

$$ Q = \frac{e}{2\pi}\oint_{|\boldsymbol{x}|\to\infty} A_\mu \mathrm{d}x^\mu = \frac{1}{2\pi}\oint_{|\boldsymbol{x}|\to\infty} \partial_\mu \Lambda \mathrm{d}x^\mu = \frac{1}{2\pi}\Big(\Lambda(2\pi) - \Lambda(0)\Big) = n \in \mathbb{Z} \ . \tag{19.1.3} $$

Q is the topological charge of the configuration, which can take nontrivial values since $U(1) \sim S^1$ and $\pi_1(S^1) = \mathbb{Z}$. Using the Stokes theorem we can rewrite Q as follows

$$ Q = \frac{e}{2\pi}\oint_{|\boldsymbol{x}|\to\infty} A_\mu \mathrm{d}x^\mu = \frac{e}{2\pi}\int \mathrm{d}^2x \epsilon_{\mu\nu}\partial_\mu A_\nu = \frac{e}{4\pi}\int \mathrm{d}^2x \epsilon_{\mu\nu}F_{\mu\nu} = \frac{e}{2\pi}\int \mathrm{d}^2x F_{01} \ , \tag{19.1.4} $$

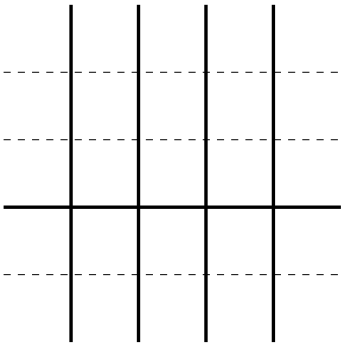

Figure 19.1: Complete axial gauge fixing: links that can be fixed to 1 in an *infinite* two dimensional lattice are depicted by thick lines.

and the quantity

$$q(\boldsymbol{x}) = \frac{e}{2\pi} F_{01}(\boldsymbol{x}) \tag{19.1.5}$$

is the topological charge density.

It is simple to see that adding $-\theta q(\boldsymbol{x})$ to the real time action does not change the equations of motion (since $\partial_\mu \frac{\partial q}{\partial \partial_\mu A_\nu} = 0$ and $\partial q / \partial A_\nu = 0$), however this change corresponds in the Euclidean formulation to adding $i\theta Q$ to the statistical weight of configurations, which generates nontrivial changes in the free energy density.

Since in the U(1) lattice gauge theory the plaquette is related (in the *naive* continuum limit) to the field strength by (see Sec. 17.3)

$$\Pi_{\mu\nu}(\boldsymbol{n}) = \exp\left(-iea^2 F_{\mu\nu}(\boldsymbol{n}) + O(a^3)\right) \tag{19.1.6}$$

we have $\arg \Pi_{\mu\nu} \approx -ea^2 F_{\mu\nu}$, and it is thus meaningful to define on the lattice

$$Q = -\frac{1}{2\pi} \sum_{\boldsymbol{n}} \arg \Pi_{01}(\boldsymbol{n}) \ . \tag{19.1.7}$$

Note that we could have also used $\mathrm{Im}\, \Pi_{\mu\nu} \approx -ea^2 F_{\mu\nu}$ to define Q, but the definition in Eq. (19.1.7) has the important property of being integer valued already at finite lattice spacing, as follows from

$$\prod_{\boldsymbol{n}} \Pi_{01}(\boldsymbol{n}) = 1 \tag{19.1.8}$$

by taking the logarithm. The previous equation holds true in Abelian theories when using periodic boundary conditions, since each link enters (in two space-time dimensions) in just two neighboring plaquettes, once with a complex conjugation. Definitions of the topological charge which are integer already at finite lattice spacing are usually called geometric definitions, as opposed to the so-called field-theoretic definitions. Geometric definitions of the topological charge in 4d SU(N) gauge theories exist but are far less trivial than in two dimensional cases, see e.g. [144], and not commonly used.

19.2 Analytical solution

We now discuss the analytical solution (and θ-dependence) of the two dimensional lattice U(1) gauge theory.

To solve the U(1) LGT with Wilson action it is convenient to use the lattice analogue of the axial gauge fixing, and we will consider the simplest case of an infinite lattice (more properly: a lattice large enough that we can neglect the effect of boundary conditions), beginning with the

$\theta = 0$ case. As discussed in Sec. 17.4 it is possible to fix to 1 all the links of a maximal tree, and it is easy to understand that a maximal tree for the two dimensional infinite square lattice is the one depicted in Fig. (19.1). We can thus fix to 1 all links in the temporal direction (which in our conventions corresponds to $\mu = 0$), and all the links in direction 1 of a single timeslice, that we conventionally choose as the one corresponding to $t = 0$:

$$U_0(\boldsymbol{n}) = 1 \ , \quad U_1(0, n_1) = 1 \ . \tag{19.2.1}$$

Using this gauge fixing the partition function factorizes as the product of the "vertical stripes" partition functions (note that we are using the simplified form $-\beta \sum P_{\mu\nu}(\boldsymbol{n})$ of the Wilson action, neglecting an irrelevant additive constant):

$$\begin{aligned} Z(\beta) &= \int \left(\prod_{\boldsymbol{n},\mu} \mathrm{d}U_\mu(\boldsymbol{n}) \right) e^{\beta \sum_{\boldsymbol{n}} P_{01}(\boldsymbol{n})} = \\ &= \left(\int \left(\prod_t \mathrm{d}U_1(t) \right) e^{\beta \sum_t \mathrm{Re}[U_1(t) U_1^\dagger(t-1)]} \right)^{N_s} \ , \end{aligned} \tag{19.2.2}$$

where N_s is the lattice extent in the $\mu = 1$ direction, assumed to be large enough to neglect bounday effects. We can now use the invariant properties of the Haar measure (and the remaining constraint $U_1(t = 0) = 1$) to further simplify this expression: if we introduce the U(1) variables

$$V_1 = U_1(t = 1) \ , \quad V_j = U_1(t = j) U_1^\dagger(t = j - 1) \text{ for } j > 2 \ , \tag{19.2.3}$$

and analogous definitions for $j < 0$, we have

$$\int \left(\prod_t \mathrm{d}U_1(t) \right) e^{\beta \sum_t \mathrm{Re}[U_1(t) U_1^\dagger(t-1)]} = \int \left(\prod_t \mathrm{d}V_t \right) e^{\beta \sum_t \mathrm{Re}[V_t]} = \left(\int \mathrm{d}V e^{\beta \mathrm{Re}[V]} \right)^{N_t} \ . \tag{19.2.4}$$

We have thus found a single plaquette model, and reduced the evaluation of the partition function to a single link Haar integration:

$$Z(\beta) = \left(\int \mathrm{d}V e^{\beta \mathrm{Re}[V]} \right)^{N_t N_s} \ . \tag{19.2.5}$$

Note that, so far, we have not used any specific property of U(1), so this expression of the partition function is valid for generic groups, both continuous and discrete, Abelian and nonAbelian.

In the specific case of the U(1) group we can use the parametrization $V = e^{i\phi}$ and the Haar measure $\mathrm{d}V = \frac{1}{2\pi} \mathrm{d}\phi$ to get

$$\int \mathrm{d}V e^{\beta \mathrm{Re}[V]} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{\beta \cos\phi} \mathrm{d}\phi = I_0(\beta) \ , \tag{19.2.6}$$

where we used (see [12] Eq. 9.6.19)

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\phi} e^{\beta \cos\phi} \mathrm{d}\phi = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(n\phi) e^{\beta \cos\phi} \mathrm{d}\phi = I_n(\beta) \ , \tag{19.2.7}$$

with I_n the modified Bessel function of first kind of integer order, $n \in \mathbb{Z}$. The derivatives of this function can be computed by using the simple relation (see [12] Eq. 9.6.28)

$$\begin{aligned} \frac{\partial}{\partial\beta} I_n(\beta) &= \frac{1}{2\pi} \frac{\partial}{\partial\beta} \int_{-\pi}^{\pi} e^{in\phi} e^{\beta \cos\phi} \mathrm{d}\phi = \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i(n+1)\phi} + e^{i(n-1)\phi}}{2} e^{\beta \cos\phi} \mathrm{d}\phi = \frac{1}{2} \Big(I_{n+1}(\beta) + I_{n-1}(\beta) \Big) \ . \end{aligned} \tag{19.2.8}$$

Using this formula we can for example compute the average value of the plaquette for generic values of β:

$$\langle \cos\phi \rangle = \frac{1}{N_s N_t} \frac{\partial}{\partial \beta} \log Z = \frac{\partial_\beta I_0(\beta)}{I_0(\beta)} = \frac{I_1(\beta)}{I_0(\beta)} \ , \tag{19.2.9}$$

where we used $I_1(\beta) = I_{-1}(\beta)$. Similar computations can be carried out also in the U(N) case, in which, instead of Bessel functions, $N \times N$ determinants of Bessel functions appear, see, e. g., [145, 146].

The computation of the average plaquette shows that we have a good analytic control of this model, but is not particularly useful from the physical point of view. We now switch to the computation of the string tension, which follows basically the same strategy. Since the gauge group is Abelian, a Wilson loop operator of size $w_t \times w_s$ can be written as a product of $w_t \times w_s$ plaquettes, and using the same steps used before we get

$$\begin{aligned} W(w_t, w_s) &= \left(\frac{1}{\int_{-\pi}^{\pi} e^{\beta \cos\phi} \mathrm{d}\phi} \int_{-\pi}^{\pi} \cos(\phi) e^{\beta \cos\phi} \mathrm{d}\phi \right)^{w_t w_s} = \\ &= \left(\frac{I_1(\beta)}{I_0(\beta)} \right)^{w_t w_s} = \exp\Big(- w_t w_s \log[I_0(\beta)/I_1(\beta)] \Big) \ , \end{aligned} \tag{19.2.10}$$

from which we get the string tension in lattice units as a function of the coupling β (note that $I_0(\beta)/I_1(\beta) > 1$)

$$\hat{\sigma}(\beta) = \log[I_0(\beta)/I_1(\beta)] \ . \tag{19.2.11}$$

We have thus shown that the two dimensional lattice U(1) gauge theory is confining for all the values of the coupling β, and in fact we explicitly computed the dimensionless string tension. An analogous computation can also be carried out in two dimensional U(N) models, and in particular in the $N \to \infty$ limit, see [147].

Since $\hat{\sigma} = a^2 \sigma$, the critical value β_c that we have to approach to extract the continuum limit can be identified by requiring that $\lim_{\beta \to \beta_c} \hat{\sigma}(\beta) = 0$, as discussed in Sec. 14.2. Using the asymptotic expansion of the modified Bessel functions of first kind (see, e. g., [12] Eq. 9.7.1) it is simple to see that $\beta_c = \infty$, indeed for $\beta \gg 1$ we have

$$\hat{\sigma}(\beta) = \frac{1}{2\beta} + \frac{1}{4\beta^2} + \frac{11}{48\beta^3} + O(\beta^{-4}) \ . \tag{19.2.12}$$

This fact could have been guessed in a more elementary way by remembering that in the naive continuum limit the lattice coupling β is related to the continuous coupling by (see Sec. 17.3)

$$\beta = \frac{1}{e^2 a^{4-D}} \ , \tag{19.2.13}$$

hence in $D = 2$ the limit $a \to 0$ with fixed e^2 corresponds to $\beta \to \infty$.

Let us now discuss θ-dependence: the Wilson action in the presence of a θ term is given by (see Eq. (19.1.7))

$$S_\theta = -\beta \sum_{\boldsymbol{n}} P_{01}(\boldsymbol{n}) + i\theta Q = -\beta \sum_{\boldsymbol{n}} \mathrm{Re}\, \Pi_{01}(\boldsymbol{n}) - i \frac{\theta}{2\pi} \sum_{\boldsymbol{n}} \arg \Pi_{01}(\boldsymbol{n}) \ , \tag{19.2.14}$$

and the evaluation of the partition function when $\theta \neq 0$ follows exactly the same steps of the $\theta = 0$ case, obtaining

$$Z(\beta, \theta) = \left(\mathcal{I}_{\frac{\theta}{2\pi}}(\beta) \right)^{N_t N_s} \ , \tag{19.2.15}$$

where we defined

$$\mathcal{I}_\nu(\beta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\nu\phi} e^{\beta \cos\phi} \mathrm{d}\phi \ . \tag{19.2.16}$$

Note that $\mathcal{I}_\nu(\beta)$ *is not* a modified Bessel function of first kind of non-integer order, as can be seen by comparing with [12] Eq. 9.6.20. Using similar manipulations we can compute the expectation value of Wilson loops at nonvanishing θ

$$\begin{aligned} W(w_t, w_s) &= \left(\frac{1}{\int_{-\pi}^{\pi} e^{\beta\cos\phi \mathrm{d}\phi + i\frac{\theta}{2\pi}\phi}} \int_{-\pi}^{\pi} \cos(\phi) e^{\beta\cos\phi + i\frac{\theta}{2\pi}\phi} \mathrm{d}\phi \right)^{w_t w_s} = \\ &= \left(\frac{\mathcal{I}_{\frac{\theta}{2\pi}+1}(\beta) + \mathcal{I}_{\frac{\theta}{2\pi}-1}(\beta)}{2\mathcal{I}_{\frac{\theta}{2\pi}}(\beta)} \right)^{w_t w_s} \end{aligned} \tag{19.2.17}$$

and the string tension

$$\hat{\sigma}(\beta, \theta) = -\log\left(\frac{\mathcal{I}_{\frac{\theta}{2\pi}+1}(\beta) + \mathcal{I}_{\frac{\theta}{2\pi}-1}(\beta)}{2\mathcal{I}_{\frac{\theta}{2\pi}}(\beta)} \right) . \tag{19.2.18}$$

It is not difficult to verify that $\lim_{\beta\to\infty} \frac{\hat{\sigma}(\beta,\theta)}{\hat{\sigma}(\beta)} = 1$, hence there is no dependence of the string tension on θ in the continuum limit. This is true also for two dimensional U(N) theories, but *not* for four dimensional SU(N) theories, see [148].

Using the definition of the topological susceptibility (compare with Chap. 11)

$$\chi(\beta) = \frac{\partial^2}{\partial\theta^2} f(\beta,\theta)\Big|_{\theta=0} = -\frac{1}{V\beta_T} \frac{\partial^2}{\partial\theta^2} \log Z(\beta,\theta)\Big|_{\theta=0} , \tag{19.2.19}$$

(with $\beta_T = 1/T$) we obtain for the topological susceptibility in lattice units the expression

$$\hat{\chi}(\beta) = a^2\chi(\beta) = -\frac{1}{I_0(\beta)} \frac{\partial^2}{\partial\theta^2} \mathcal{I}_{\frac{\theta}{2\pi}}(\beta)\Big|_{\theta=0} = \frac{1}{(2\pi)^3 I_0(\beta)} \int_{-\pi}^{\pi} \phi^2 e^{\beta\cos\phi} \mathrm{d}\phi , \tag{19.2.20}$$

where we used

$$\begin{aligned} \frac{\partial}{\partial\theta} \mathcal{I}_{\frac{\theta}{2\pi}}(\beta)\Big|_{\theta=0} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{i\phi}{2\pi} e^{\beta\cos\phi} \mathrm{d}\phi = 0 , \\ \frac{\partial^2}{\partial\theta^2} \mathcal{I}_{\frac{\theta}{2\pi}}(\beta)\Big|_{\theta=0} &= -\frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{\phi}{2\pi}\right)^2 e^{\beta\cos\phi} \mathrm{d}\phi . \end{aligned} \tag{19.2.21}$$

As already noted for the partition function and the string tension, using conceptually similar but technically more involved manipulations it is possible to obtain the topological susceptibility of two dimensional U(N) gauge theories, see [149]. Using the Laplace method, see, e. g., [84] §2.4 or [85] §6.4, it is simple to estimate in the large β limit (i. e. approaching the continuum limit) the integral in the expression of $\hat{\chi}(\beta)$, obtaining

$$\int_{\pi}^{\pi} \phi^2 e^{\beta\cos\phi} \mathrm{d}\phi \simeq \int_{-\infty}^{\infty} \phi^2 e^{\beta(1-\phi^2/2)} \mathrm{d}\phi = \frac{\sqrt{2\pi}}{\beta^{3/2}} e^{\beta} , \tag{19.2.22}$$

Together with $I_0(\beta) \simeq e^\beta/\sqrt{2\pi\beta}$, see, e. g. [12] Eq. 7.7.1, this expression gives

$$\hat{\chi}(\beta) \simeq \frac{1}{4\pi^2\beta} . \tag{19.2.23}$$

It is straightforward but a little tedious to find also the subleading corrections:

$$\hat{\chi}(\beta) \simeq \frac{1}{4\pi^2\beta} \left(1 + \frac{1}{2\beta} + \frac{13}{24\beta^2} + O(\beta^{-3}) \right) . \tag{19.2.24}$$

Combining Eq. (19.2.11) and Eq. (19.2.24) we get

$$\frac{\hat{\chi}(\beta)}{\hat{\sigma}(\beta)} = \frac{1}{2\pi^2} \left(1 + \frac{1}{12\beta^2} + O(\beta^{-3}) \right) \tag{19.2.25}$$

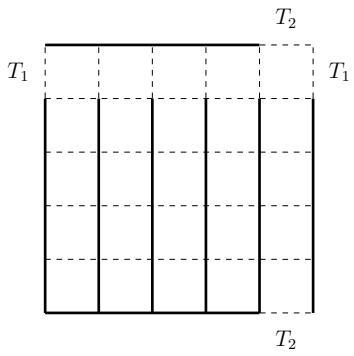

Figure 19.2: Complete axial gauge fixing: links that can be fixed to 1 in a two dimensional lattice with periodic boundary conditions are depicted by thick lines.

and hence the continuum result (note that χ and σ has the same mass dimension in two dimensional models)

$$\frac{\chi}{\sigma} = \lim_{\beta\to\infty} \frac{\hat{\chi}(\beta)}{\hat{\sigma}(\beta)} = \frac{1}{2\pi^2} \ . \tag{19.2.26}$$

The same continuum result can be obtained also by using the non-integer "field-theoretic" definition of the topological charge

$$Q_{FT} = -\frac{1}{2\pi} \sum_{\boldsymbol{n}} \mathrm{Im}\Pi_{01}(\boldsymbol{n}) \ , \tag{19.2.27}$$

since in this case one gets

$$\hat{\chi}_{FT}(\beta) = \frac{1}{(2\pi)^3 I_0(\beta)} \int_{-\pi}^{\pi} \sin^2\phi e^{\beta\cos\phi} \mathrm{d}\phi \ , \tag{19.2.28}$$

which using Eq. (19.2.7) can be rewritten as

$$\hat{\chi}_{FT}(\beta) = \frac{1}{(2\pi)^2 I_0(\beta)} \frac{I_0(\beta) - I_2(\beta)}{2} \ . \tag{19.2.29}$$

Using [12] Eq. 7.7.1 it is then possible to obtain

$$\hat{\chi}_{FT}(\beta) = \frac{1}{4\pi^2\beta} \left(1 - \frac{1}{2\beta} - \frac{1}{8\beta^2} + O(\beta^{-3})\right) \ , \tag{19.2.30}$$

whose leading term in the large-β limit (i. e. approaching the continuum limit) coincides with the one found before using the geometric definition. Note that the coincidence of the results obtained using the two discretizations is by no means trivial, and it is in fact false for more complex theories, for which the field-theoretic definition requires nontrivial renormalizations, see, e. g., [144]. It is also useful to explicitly note that despite the fact that for the two dimensional U(1) model

$$\lim_{\beta\to\infty} \frac{\hat{\chi}(\beta)}{\hat{\sigma}(\beta)} = \lim_{\beta\to\infty} \frac{\hat{\chi}_{FT}(\beta)}{\hat{\sigma}(\beta)} = \frac{1}{2\pi^2} \ , \tag{19.2.31}$$

scaling corrections are different in $\hat{\chi}$ and $\hat{\chi}_{FT}$: using the geometric definition of the topological charge $\hat{\chi}(\beta)/\hat{\sigma}(\beta)$ has corrections $O(\beta^{-2}) \sim O(a^4)$ (see above), while using the field-theoretic definition $\hat{\chi}_{FT}(\beta)/\hat{\sigma}(\beta)$ corrections $O(\beta^{-1}) \sim O(a^2)$ are present.

Let us now consider the finite lattice case. A maximal tree that can be used to fix the temporal gauge on a finite two dimensional lattice with periodic boundary conditions is shown in Fig. (19.2): we can fix to one $(N_t-1)N_s + N_s - 1 = N_tN_s - 1$ links, thus $N_tN_s + 1$ link integrals remain after gauge fixing. In an Abelian

theory it is simple to see that the product W of all the plaquettes is equal to one (see Eq. (19.1.8) above), hence only N_tN_s-1 plaquettes are independent of each other. By performing a change of variable in the Haar integrals, introducing plaquette variables, we end up with an integral on all the plaquettes, with the constraint $\delta(W,1)$, and on two remaining links. In the Abelian cases the two remaining link integrals are totally irrelevant, while they are fundamental in the nonAbelian case. Indeed, in the nonAbelian case it can be shown that the product W of all the plaquettes (carried out in a specific order) is equal to $T_1T_2T_1^\dagger T_2^\dagger$, where T_1 and T_2 are the links defined in Fig. (19.2), see [150] for details.

Let us consider for the sake of the simplicity just the Abelian U(1) case. Whenever the action can be written as $S=\sum_{\boldsymbol{n}}\mathcal{S}(\Pi_{01}(\boldsymbol{n}))$ we have

$$Z=\int\delta(W,1)\prod_{\boldsymbol{n}}\left(e^{-\mathcal{S}(\Pi_{01}(\boldsymbol{n}))}\mathrm{d}\Pi_{01}(\boldsymbol{n})\right)\ . \tag{19.2.32}$$

If we consider the parametrization $e^{i\phi}$ of U(1), the Haar measure is just $\frac{\mathrm{d}\phi}{2\pi}$. From the Fourier series expansion of periodic functions on $[-\pi,\pi]$ (i. e. functions defined on U(1)) we have

$$f(\psi)=\sum_k a_k e^{ik\psi}\ ,\quad a_k=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(\psi)e^{-ik\psi}\mathrm{d}\psi \tag{19.2.33}$$

and thus[1] (for a formal proof using generalized functions see, e. g., [151] II.7)

$$\delta_G(\psi=\varphi\bmod 2\pi)=\sum_k e^{ik(\psi-\varphi)}\ , \tag{19.2.34}$$

where δ_G is the δ function on the group, defined with respect to the Haar measure. Using this identity we can finally write the partition function as

$$Z=\sum_k\int_{-\pi}^{\pi}\prod_{\boldsymbol{n}}\left(e^{-\mathcal{S}(\phi(\boldsymbol{n}))+ik\phi}\frac{\mathrm{d}\phi(\boldsymbol{n})}{2\pi}\right)=\sum_k\left(\int_{-\pi}^{\pi}e^{-\mathcal{S}(\phi)+ik\phi}\frac{\mathrm{d}\phi}{2\pi}\right)^{N_sN_t}\ . \tag{19.2.35}$$

If we now use the Wilson action with the geometric discretization of the topological charge we have (see Eq. (19.2.16))

$$\int_{-\pi}^{\pi}e^{-\mathcal{S}(\phi)+ik\phi}\frac{\mathrm{d}\phi}{2\pi}=\int_{-\pi}^{\pi}e^{\beta\cos\phi+i\frac{\theta}{2\pi}\phi+ik\phi}\frac{\mathrm{d}\phi}{2\pi}=\mathcal{I}_{k+\frac{\theta}{2\pi}}(\beta) \tag{19.2.36}$$

and thus

$$Z(\beta,\theta)=\sum_k\left(\mathcal{I}_{k+\frac{\theta}{2\pi}}(\beta)\right)^{N_sN_t}\ . \tag{19.2.37}$$

For $\theta=0$, using Eq. (19.2.8), it is now immediate to see that the average plaquette is given by

$$\langle\cos\phi\rangle=\frac{\sum_k I_k(\beta)^{N_tN_s-1}\left(I_{k+1}(\beta)+I_{k-1}(\beta)\right)}{2\sum_k I_k(\beta)^{N_tN_s}}\ . \tag{19.2.38}$$

Since $I_n(\beta)=I_{-n}(\beta)$, and $|I_n(\beta)/I_0(\beta)|<1$ for $\beta>0$ and $n>0$, in the thermodynamic limit $N_tN_s\gg 1$ we recover Eq. (19.2.9). For the topological susceptibility in lattice units one gets analogously the expression

$$\begin{aligned}\left(\sum_k I_k(\beta)^{N_tN_s}\right)\hat{\chi}(\beta)=&\frac{1}{(2\pi)^3}\sum_k I_k(\beta)^{N_tN_s-1}\int_{-\pi}^{+\pi}\phi^2\cos(k\phi)e^{\beta\cos\phi}\mathrm{d}\phi-\\&-\frac{N_tN_s-1}{(2\pi)^4}\sum_k I_k(\beta)^{N_tN_s-2}\left[\int_{-\pi}^{+\pi}\phi\sin(k\phi)e^{\beta\cos\phi}\mathrm{d}\phi\right]^2\ .\end{aligned} \tag{19.2.39}$$

When $N_tN_s\gg 1$, the I_0 terms dominate, moreover the second term in the right hand side vanishes for $n=0$, and we recover the expression found before in the thermodynamic limit.

19.3 Numerical results

Let us now discuss some numerical results obtained by simulating the two dimensional U(1) model. The first observable studied is the static potential, see Sec. 17.4, which is computed by using

$$V(w_s)=-\lim_{w_t\to\infty}\frac{1}{w_t}\log W(w_t,w_s)\ . \tag{19.3.1}$$

We have seen in the previous section that $W(w_t,w_s)=e^{-\hat{\sigma}w_sw_t}$, hence Wilson loops exactly obey the area-law. With this we mean that the area-law does not just describe the large area behavior

[1] From a group-theoretic point of view this formula represents the character expansion for the case of U(1). Character expansion has to be used also in the nonAbelian cases to rewrite $\delta(W,T_1T_2T_1^\dagger T_2^\dagger)$, see [150] (see, e. g., [146] for background material).

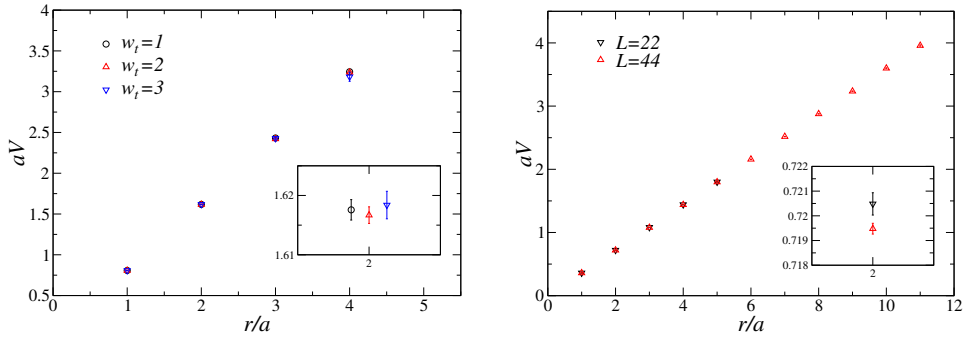

Figure 19.3: Static potential (in lattice units) in two dimensional U(1) gauge theory. (left) Results obtained by using $\beta = 1$ on a 16^2 lattice, for different values of the temporal extents of the Wilson loop. The inset shows a zoom of $r = 2a$ data, slightly shifted horizontally to increase their readability. (right) Results obtained by using $\beta = 2$ on 22^2 and 44^2 lattices, measuring Wilson loops with temporal extent $w_t = 1$.

of the Wilson loops (as in any confining gauge theory), but exactly parametrizes $W(w_t, w_s)$ for *any* value of w_t and w_s. From this it follows that we could compute $\hat{\sigma}$ just from the average plaquette, and that $V(r) = \sigma r$, without any correction to the linear behavior (at least as far as no finite volume effects are present).

As a first step we explicitly check that the static potential $V(w_s)$ computed by using Wilson loops of size $w_s \times w_t$ does not in fact depends on w_t. In Fig. (19.3) (left) we report the results obtained for the static potential by using $\beta = 1$ on a 16^2 lattice, measuring Wilson loops $W(w_t, w_s)$ with $1 \leq w_t, w_s \leq 4$. We collected a statistic of about 10^6 total lattice updates (20% Metropolis updates, 80% microcanonical updates), performing Wilson loop measures every 50 lattice updates and using (whenever possible) the multihit algorithm described in Sec. 18.5; link averages in multihit are estimated using 10 Metropolis updates and 10 microcanonical updates, and the total simulation time was ≈ 3 minutes. The same update scheme was used also for the other simulations that will be discussed below. Results presented in Fig. (19.3) (left) show that the theoretical expectations are perfectly reproduced by numerical data, both regarding the independence of the results from w_t (see in particular the inset) and regarding the absence of corrections to the linear behavior of the static potential.

To carry out a complete investigation of the static potential we then need to check for the presence of finite volume effects. For this reason we carried out simulations for $\beta = 1$ on two lattices of different extent: 16^2 and 32^2. The results obtained show that there is only a very small dependence on the lattice size. The final step is the continuum limit: to perform the continuum limit we have to consider $\beta \to \infty$, however it is important to note that, when changing the value of β, also the lattice extent has to be varied, in order to keep the lattice size approximately constant in physical units. Since $a \propto 1/\sqrt{\beta}$, we have approximately $16a(\beta = 1) \simeq 22a(\beta = 2) \simeq 32a(\beta = 4)$. In Fig. (19.3) (right) we report the results of simulations carried out for $\beta = 2$ using the lattices 24^2 and 44^2: finite size systematic effects can be seen also in this case (see inset), of approximately the same size of the statistical uncertainties.

For each value of β it is possible to fit the string tension, and to compare the fit results with the theoretical values. This is however not completely trivial, since the values of the static potential at different distances are generically correlated with each other, as they are estimated using Wilson loops evaluated on the same configurations. We thus have to perform a correlated fit, or use independent simulations to compute the static potential at different distances. Once the string tension (or any other dimensionfull quantity) has been measured, we can use it to set the scale, i. e. convert other quantities in physical units, in order to investigate the approach to the continuum limit.

To investigate the continuum limit of the static potential we can plot $V(r)/\sqrt{\sigma} = aV/\sqrt{\hat{\sigma}(\beta)}$

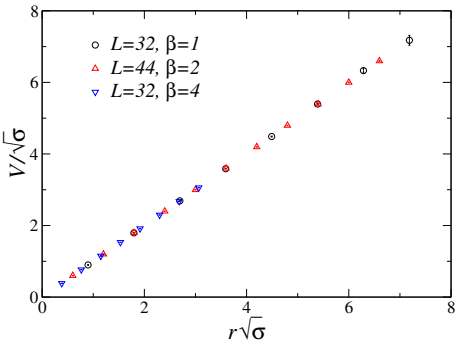

Figure 19.4: Approach to the continuum limit of the static potential in two dimensional U(1) gauge theory.

as a function of $r\sqrt{\sigma} = w_s\sqrt{\hat{\sigma}(\beta)}$: in the absence of scaling corrections all points obtained using simulations at different values of the lattice spacing should collapse on the same curve. For the case of the two dimensional U(1) theory this curve is in fact a straight line, since the static potential is exactly linear, as can be seen in Fig. (19.4).

For comparison, we now discuss some results obtained in the three dimensional U(1) model at $\beta = 1.7$, using a 42^3 lattice. The same update scheme adopted for the two dimensional model was used also in this case, collecting 10^5 lattice updates. The time required for the simulation (measuring Wilson loops with $1 \leq w_t, w_s \leq 10$ every 50 updates), has been of ≈ 5 days. The results obtained are presented in Fig. (19.5), from which it is clear that in this case an extrapolation is required to estimate the static potential, a consequence of the fact that Wilson loops cannot be parametrized by the simple expression $e^{-\hat{\sigma} w_t w_s}$. We thus performed, for each value of w_s, an extrapolation using the ansatz

$$-\frac{1}{w_t}\log W(w_t, w_s) \simeq a + b/w_t \ , \tag{19.3.2}$$

neglecting correlations between data as a first approximation. The results of these extrapolations are shown in the right panel of Fig. (19.5): the theory is clearly confining, but the static potential is not exactly linear as was in the two dimensional case. A functional form which describes the static potential fairly well is the so called Cornell potential

$$V(r) = \sigma r + \frac{\alpha}{r} + c \ , \tag{19.3.3}$$

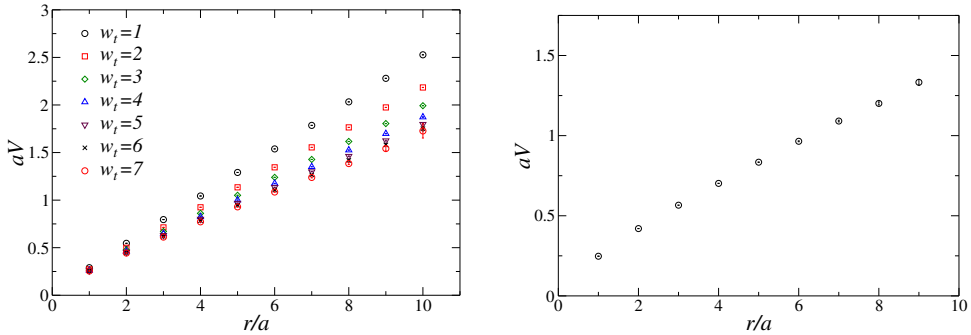

Figure 19.5: Static pontential in three dimensional U(1) gauge theory, computed at $\beta = 1.7$ using a 42^3 lattice. (left) dependence of $-\frac{1}{w_t}\log W(w_t, w_s)$ on w_t (right) static potential obtained by performing the extrapolation $w_t \to \infty$ of the data shown in the left panel.

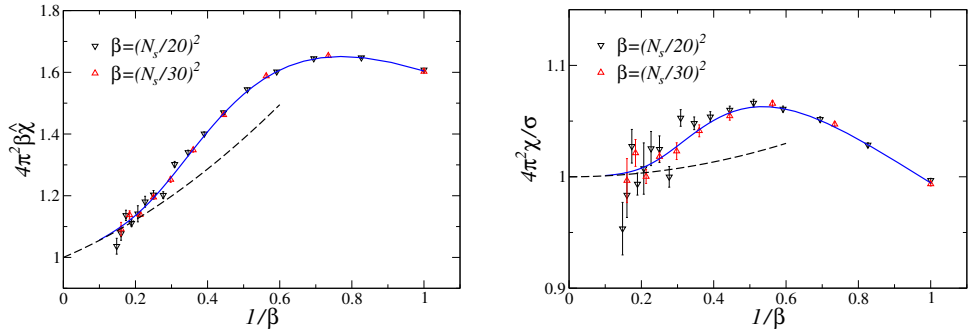

Figure 19.6: (left) Approach to the continuum limit of the quantity $4\pi^2\beta\hat{\chi}$, for two different values of the physical volume. The solid line corresponds to the analytical result Eq. (19.2.20), while the dashed line represents the development of the analytical result up to $O(\beta^{-2})$. (right) Approach to the continuum limit of the quantity $2\pi^2\chi/\sigma$, for two different values of the physical volume. The solid line corresponds to the analytical result obtained by using Eq. (19.2.11) and Eq. (19.2.20), while the dashed line represents the development of the analytical result up to $O(\beta^{-2})$.

introduced in [152] to describe charmonium bound states. From the large distance behavior of the static potential we estimate $\hat{\sigma} \approx 0.123(2)$ (the reason for $\approx$ is that we neglected correlations when performing the fits needed for the extrapolations, so the final error is likely underestimated). Using a simulation algorithm specifically designed for the three dimensional U(1) gauge model, the string tension at $\beta = 1.7$ was estimated to be $\hat{\sigma} = 0.122764(2)$ in [153], a value which is consistent with the value we found (and obviously more accurate).

Let us finally discuss the topological susceptibility of the two dimensional U(1) model. For this purpose we performed several simulations at different values of the lattice spacing (i. e. of the lattice coupling β) at fixed physical volume, for two different values of the physical volume. Since $a \propto 1/\sqrt{\beta}$, to approach the continuum limit at constant physical volume we can perform simulations for different values of the lattice extent N_s (with $N_t = N_s$), choosing β in such a way that

$$a(\beta)N_s \propto \frac{N_s}{\sqrt{\beta}} \tag{19.3.4}$$

is constant. We used $\beta = (N_s/20)^2$ for $N_s \in [20, 52]$, and to check for the presence of finite volume corrections we also used $\beta = (N_s/30)^2$ for $N_s \in [30, 75]$. In both the cases 10^6 updates (20% Metropolis updates, 80% microcanonical updates) have been collected for most of the lattice sizes. Exactly as in the QM model discussed in Chap. 11, also in this model an exponential critical slowing down is present, and for this reason 10^7 updates have been used for the four largest lattice sizes.

In Fig. (19.6)(left) we report the results obtained for $4\pi^2\beta\hat{\chi}$, which nicely agree with the analytical results, and in particular approach 1 for $\beta \to \infty$. In Fig. (19.6)(right) we instead report results obtained for the dimensionless ratio $2\pi^2\chi/\sigma = 2\pi^2\hat{\chi}/\hat{\sigma}$ (using for $\hat{\sigma}$ the analytical result in Eq. (19.2.11)). Note that the vertical scale is different in the two panel, and that lattice artifacts for the quantity $4\pi^2\beta\hat{\chi}$ are larger than those of the quantity $2\pi^2\hat{\chi}/\hat{\sigma}$, consistently with the fact that $4\pi^2\beta\hat{\chi}$ has $O(1/\beta) \propto a^2$ corrections, see Eq. (19.2.24), while $2\pi^2\hat{\chi}/\hat{\sigma}$ only has $O(1/\beta^2) \propto a^4$ corrections, see Eq. (19.2.25).

For the two dimensional U(1) (in fact also for $U(N)$ models with $N > 1$) model it is not difficult to implement an algorithm which completely remove the critical slowing down related to topological modes [154]. The basic idea is the following: since the topological carge is odd under complex conjugation of the links, if we select a portion of the lattice (e. g. the region inside a square) and apply the transformation $U_\mu(\boldsymbol{n}) \to U_\mu(\boldsymbol{n})^*$ to all the links inside this region, such a transformation will likely change the value of the topological charge, at least if the region has a linear size which is comparable with the correlation length. It is however unlikely for such a transformation to be accepted by a Metropolis step, since the change of action will typically be large. Note that only links on the

boundary of the region to be updated contributes to the difference of action ΔS, which can be written in the form

$$\Delta S = -\beta \sum_{\substack{\text{boundary} \\ \text{links}}} \mathrm{Re}\Big[\big(U^*_\mu(\boldsymbol{n}) - U_\mu(\boldsymbol{n})\big) \big({S^{(i)}_\mu}^*(\boldsymbol{n}) + S^{(o)}_\mu(\boldsymbol{n})\big) \Big] \,, \tag{19.3.5}$$

where we denoted by $S^{(i)}_\mu(\boldsymbol{n})$ the components of the sum of the staples which lay inside the region to be updated, and by $S^{(e)}_\mu(\boldsymbol{n})$ the components of the sum of the staples which lay outside the same region. In the particular case of two dimensional models, it is however possible to set to the identity all the links on the boundary, up to a single link, before proposing the update. In this way $U^*_\mu(\boldsymbol{n}) - U_\mu(\boldsymbol{n}) = 0$ but for a single boundary link, and the values of ΔS corresponding to this nonlocal update are analogous to those associated to a single link update.

Chapter 20

Appendices to Part IV

20.A Benchmark for the two dimensional free scalar theory

The lattice action obtained by using the forward discretization of the derivative is (see Eq. (13.2.3) with $D = 2$)

$$\begin{aligned} S_L &= \frac{1}{2}\sum_{mbfn}\left\{(\hat{m}^2+4)\hat{\phi}_{\boldsymbol{n}}^2 - 2\sum_{\mu=0}^{1}\hat{\phi}_{\boldsymbol{n}}\hat{\phi}_{\boldsymbol{n}+\hat{\mu}}\right\} = \frac{1}{2}\sum_{\boldsymbol{n},\boldsymbol{j}}\hat{\phi}_{\boldsymbol{n}}K_{\boldsymbol{n}\,\boldsymbol{j}}\,\hat{\phi}_{\boldsymbol{j}}\ , \\ K_{\boldsymbol{n}\,\boldsymbol{j}} &= (\hat{m}^2+4)\delta_{\boldsymbol{n},\boldsymbol{k}} - \sum_{\mu=0}^{1}(\delta_{\boldsymbol{n}+\hat{\mu},\boldsymbol{j}} + \delta_{\boldsymbol{n}-\hat{\mu},\boldsymbol{j}})\ . \end{aligned} \tag{20.A.1}$$

If we use periodic b. c. and the ordering $(0,0),(0,1),(1,0),(1,1)$ for the sites of a 2×2 lattice, the matrix K becomes

$$K = \begin{pmatrix} \hat{m}^2+4 & -2 & -2 & 0 \\ -2 & \hat{m}^2+4 & 0 & -2 \\ -2 & 0 & \hat{m}^2+4 & -2 \\ 0 & -2 & -2 & \hat{m}^2+4 \end{pmatrix}\ , \tag{20.A.2}$$

and

$$Z(\hat{m}) = \int\left(\prod_{\boldsymbol{n}}\mathrm{d}\hat{\phi}_{\boldsymbol{n}}\right)e^{-S_L} = \frac{(2\pi)^2}{\sqrt{\det K}} = \frac{4\pi^2}{\sqrt{\hat{m}^8+16\hat{m}^6+80\hat{m}^4+128\hat{m}^2}}\ . \tag{20.A.3}$$

As a consequence

$$X = \frac{1}{4}\left\langle\sum_{\boldsymbol{n}}\hat{m}^2\hat{\phi}_{\boldsymbol{n}}^2\right\rangle = -\frac{1}{4}\hat{m}\frac{\partial}{\partial\hat{m}}\log Z(\hat{m}) = \frac{\hat{m}^4+8\hat{m}^2+8}{\hat{m}^4+12\hat{m}^2+32}\ . \tag{20.A.4}$$

Note that (since $N_t N_s^{D-1} = 4$) X is just the average value of O_1 defined in Eq. (15.1.9).

$\hat{m}$	X	X (MC result)
0.5	0.286987...	0.28731(35)
1	0.377777...	0.37756(34)
1.5	0.484878...	0.48467(36)

Table 20.1: Values computed on the lattice 2^2 with periodic boundary conditions using 4×10^7 single site updates, 20% heatbath and 80% overrelaxation (execution time $\approx$ 10s for each case).

20.B Benchmark for the two dimensional U(1) LGT

Using the Wilson action

$$S = -\beta \sum_{\boldsymbol{n}} P_{01}(\boldsymbol{n}) \ , \tag{20.B.1}$$

where $P_{01}(\boldsymbol{n}) = \mathrm{Re}\Pi_{01}(\boldsymbol{n})$ and $\Pi_{01}(\boldsymbol{n})$ is the plaquette operator in position $\boldsymbol{n}$, the average plaquette on a $N_t \times N_s$ lattice with periodic boundary conditions is given by (I_n is the modied Bessel function of first kind of order n)

$$\langle P_{01} \rangle = \frac{\sum_{n=-\infty}^{+\infty} I_n(\beta)^{N_t N_s - 1}(I_{n+1}(\beta) + I_{n-1}(\beta))}{2\sum_{n=-\infty}^{+\infty} I_n(\beta)^{N_t N_s}} \ , \tag{20.B.2}$$

which in the thermodynamic limit reduces to

$$\langle P_{01} \rangle^{(t.l.)} = \frac{I_1(\beta)}{I_0(\beta)} \ . \tag{20.B.3}$$

The topological susceptibility in lattice units $\hat{\chi}(\beta) = \langle Q^2 \rangle / (N_t N_s)$ (where Q is defined in Eq. (19.1.7)) is given by

$$\begin{aligned} \left(\sum_{n=-\infty}^{+\infty} I_n(\beta)^{N_t N_s} \right) \hat{\chi}(\beta) = \frac{1}{(2\pi)^3} \sum_{n=-\infty}^{+\infty} I_n(\beta)^{N_t N_s - 1} \int_{-\pi}^{+\pi} \phi^2 \cos(n\phi) e^{\beta \cos\phi} \mathrm{d}\phi - \\ - \frac{N_t N_s - 1}{(2\pi)^4} \sum_{n=-\infty}^{+\infty} I_n(\beta)^{N_t N_s - 2} \left[\int_{-\pi}^{+\pi} \phi \sin(n\phi) e^{\beta \cos\phi} \mathrm{d}\phi \right]^2 \ , \end{aligned} \tag{20.B.4}$$

which in the termodynamic limit reduces to

$$\hat{\chi}^{(t.l.)}(\beta) = \frac{1}{(2\pi)^3} \frac{1}{I_0(\beta)} \int_{-\pi}^{+\pi} \phi^2 e^{\beta \cos\phi} \mathrm{d}\phi \ . \tag{20.B.5}$$

$N_t \times N_s$	β	$\langle P_{01} \rangle$	$\langle P_{01} \rangle$ (MC)	$\langle P_{01} \rangle^{(t.l.)}$	$\hat{\chi}$	$\hat{\chi}$ (MC)	$\hat{\chi}^{(t.l)}$
2×2	1.0	0.505197	0.50558(63)	0.44639	0.0212675	0.02125(10)	0.0406362
2×2	2.0	0.779561	0.77941(29)	0.69777	0.00132364	0.001337(28)	0.019364
5×5	1.0	0.44639	0.44626(21)	0.44639	0.0406362	0.040596(81)	0.0406362

Table 20.2: Values computed using 10^7 update of the whole lattice, 20% heatbath and 80% over-relaxation (execution time $\approx$ 15s, $\approx$ 15s and $\approx$ 60s, respectively).

Bibliography

[1] S. Caracciolo, R. G. Edwards, S. J. Ferreira, A. Pelissetto, and A. D. Sokal, "Extrapolating Monte Carlo Simulations to Infinite Volume: Finite-Size Scaling at $\xi/L \gg 1$," *Phys. Rev. Lett.* **74** (1995) 2969. 4

[2] W. Feller, *An Introduction to Probability Theory and Its Applications, volume 1.* John Wiley & Sons, 1968. 10, 21

[3] W. Feller, *An Introduction to Probability Theory and Its Applications, volume 2.* John Wiley & Sons, 1970. 10

[4] P. Billingsley, *Probability and Measure.* John Wiley & Sons, 1995. 10, 21, 40

[5] A. I. Khinchin, *Mathematical foundations of statistical mechanics.* Dover Publications, 1949. 10

[6] A. D. Sokal, "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms," in *Functional Integration. Basics and applications*, C. DeWitt-Morette, P. Cartier, and A. Folacci, eds. Springer, 1997. 12, 40, 67

[7] D. H. Lehmer, "Mathematical methods in large-scale computing units," in *The Annals of the Computation Laboratory of Harvard University*, H. H. Aiken, ed., vol. XXVI. Harvard University Press, 1951. Proceedings of a Second Symposium on Large-Scale Digital Calculating Machinery (1949). 13

[8] D. E. Knuth, *The Art of Computer Programming, vol. 2 (Seminumerical Algorithms).* Addison-Wesley, 1998. 13, 14, 18

[9] G. Marsaglia, "Random numbers fall mainly in the planes," *Proc. Natl. Acad. Sci. USA* **61** (1968) 25. 14

[10] A. M. Ferrenberg, D. P. Landau, and Y. J. Wong, "Monte Carlo simulations: Hidden errors from "good" random number generators," *Phys. Rev. Lett.* **69** (1992) 3382. 14

[11] M. Creutz, "Monte Carlo Study of Quantized SU(2) Gauge Theory," *Phys. Rev. D* **21** (1980) 2308. 19, 152

[12] M. Abramowithz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards Applied Mathematics Series, 1972. 20, 64, 118, 159, 160, 161, 162

[13] A. D. Kennedy and B. J. Pendleton, "Improved Heat Bath Method for Monte Carlo Calculations in Lattice Gauge Theories," *Phys. Lett. B* **156** (1985) 393. 20, 149, 152

[14] R. Durrett, *Probability. Theory and Examples.* Cambridge University Press, 2018. 21, 40

[15] F. R. Gantmacher, *The theory of matrices, volume 2.* American Mathematical Society, 2000. 25

[16] S. Sternberg, *A Methematical Companion to Quantum Mechanics*. Dover Publications, 2019. 27, 56

[17] F. A. Berezin and M. A. Shubin, *The Schrödinger Equation*. Kluwer Academic Publishers, 1991. 27, 56

[18] G. Teschl, *Mathematical Methods in Quantum Mechanics With Applications to Schrödinger Operators*. American Mathematical Society, 2009. 27, 56

[19] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.* **21** (1953) 1087. 31

[20] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika* **57** (1970) 97. 32

[21] G. O. Roberts and J. S. Rosenthal, "Markov-chain Monte Carlo: Some pratical implications of theoretical results," *Canad. J. Statist.* **26** (1998) 5. 35

[22] N. Madras and A. D. Sokal, "The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk," *J. Stat. Phys.* **50** (1988) 109. 41

[23] U. Wolff, "Monte Carlo errors with less errors," *Comput. Phys. Commun.* **156** (2004) 143, arXiv:hep-lat/0306017. [Erratum: Comput.Phys.Commun. 176, 383 (2007)]. 41

[24] M. B. Priestley, *Spectral Analysis and Time Series. Volume 1. Univariate Series*. Academic Press, 1981. 41

[25] K. Wilson, "Monte-Carlo Calculations for the Lattce Gauge Theory," in *Recent Developments in Gauge Theories*, G. 't Hooft, C. Itzykson, A. Jaffe, H. Lehmann, P. K. Mitter, I. M. Singer, and R. Stora, eds. Plenum Press, 1980. 41

[26] C. Whitmer, "Over-relaxation methods for Monte Carlo simulations of quadratic and multiquadratic actions," *Phys. Rev. D* **29** (1984) 306. 41

[27] M. D'Elia, "Appunti del Corso di Metodi Numerici della Fisica Teorica, Parte I." 2016. 44

[28] M. D'Elia, K. Langfeld, and B. Lucini, *Stochastic Methods in Scientific Computing. From Foundations to Advanced Techniques*. CRC press, 2024. 44

[29] B. Efron and T. Hastie, *Computer Age Statistical Inference*. Cambridge University Press, 2016. 48, 50

[30] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrian and Applied Mathematics, 1982. 48, 50

[31] R. G. Miller, "The jackknife – a review," *Biometrika* **61** (1974) 1. 50

[32] E. Prugovečki, *Quantum Mechanics in Hilbert Space*. Dover Publication, 2006. 56

[33] F. Strocchi, *Elements of Quantum Mechanics of Infinite Systems*. World Scientific, 1985. 56, 142

[34] D. Ruelle, *Statistical Mechanics. Rigorous Results*. World Scientific, 1999. 56

[35] J. Glimm and A. Jaffe, *Quantum Physics. A Functional Integral Point of View*. Springer, 1987. 56

[36] S. Friedli and Y. Velenik, *Statistical Mechanics of Lattice Systems. A Concrete Mathematical Introduction*. Cambridge University Press, 2018. 56

[37] K. Huang, *Statistical Mechanics*. John Wiley & Sons, 1987. 56, 87

[38] L. D. Landau and E. M. Lifshitz, *Statistical Physics*. Pergamon Press, 1980. 56, 64, 117

[39] B. M. McCoy and T. T. Wu, *The Two-Dimensional Ising Model*. Dover Publications, 2014. 56, 68, 69, 89

[40] R. Savit, "Duality in Field Theory and Statistical Systems," *Rev. Mod. Phys.* **52** (1980) 453. 56, 143

[41] H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena*. Clarendon Press, 1971. 59

[42] J. Cardy, *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, 1996. 59, 65, 82

[43] K. G. Wilson, "The renormalization group: Critical phenomena and the Kondo problem," *Rev. Mod. Phys.* **47** (1975) 773. 59

[44] F. J. Wegner, "The Critical State, General Aspects," in *Phase Transitions and Critical Phenomena. Volume 6*, C. Domb and M. S. Green, eds. Academic Press, 1976. 59

[45] C. Itzykson and J. M. Drouffe, *Statistical field theory*. Cambridge University Press, 1989. 59, 68, 69, 82, 89, 141, 143

[46] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena*. Clarendon Press, 2002. 59, 82, 108, 123, 157

[47] C. R. Hwang and S. J. Sheu, "A Remark on the Ergodicity of Systematic Sweep in Stochastic Relaxation," in *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, P. Barone, A. Frigessi, and M. Piccioni, eds. Springer, 1992. 61, 76

[48] A. Pelissetto and E. Vicari, "Critical phenomena and renormalization group theory," *Phys. Rept.* **368** (2002) 549, cond-mat/0012164. 65, 67, 70

[49] M. Hasenbusch, "Finite size scaling study of lattice models in the three-dimensional Ising universality class," *Phys. Rev. B* **82** (2010) 174433, arXiv:1004.4486 [cond-mat.stat-mech]. 67, 89

[50] P. C. Hohenberg and B. I. Halperin, "Theory of Dynamic Critical Phenomena," *Rev. Mod. Phys.* **49** (1977) 435. 67

[51] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics. An Introduction*. Springer, 2002. 68

[52] J. Salas and A. D. Sokal, "Universal amplitude ratios in the critical two-dimensional Ising model on a torus," *J. Statist. Phys.* **98** (2000) 551, arXiv:cond-mat/9904038. 69, 89

[53] M. P. Nightingale and H. W. J. Blöte, "Monte Carlo computation of correlation times of independent relaxation modes at criticality," *Phys. Rev. B* **62** (2000) 1089. 71

[54] F. Y. Wu, "The Potts model," *Rev. Mod. Phys.* **54** (1982) 235. 74

[55] R. J. Baxter, *Exactly Solver Models in Statistical Mechanics*. Dover Publications, 2007. 74

[56] C. Bonati and M. D'Elia, "The three-dimensional, three state Potts model in a negative external field," *Phys. Rev. D* **82** (2010) 114515, arXiv:1010.3639 [hep-lat]. 75

[57] A. Pelissetto and E. Vicari, "Scaling behaviors at quantum and classical first-order transitions," in *50 years of the renormalization group, dedicated to the memory of Michael. E. Fisher*, A. Aharony, O. Entin-Wohlman, D. Huse, and L. Radzihovsky, eds. World Scientific, 2024. arXiv:2302.08238 [cond-mat.stat-mech]. 77, 79

[58] J. Lee and J. M. Kosterlitz, "Finite-size scaling and Monte Carlo simulations of first-order phase transitions," *Phys. Rev. B* **43** (1991) 3265. 77

[59] C. Borgs and R. Kotecky, "Finite-size effects at asymmetric first-order phase transitions," *Phys. Rev. Lett.* **68** (1992) 1734. 77

[60] B. A. Berg and T. Neuhaus, "Multicanonical ensemble: A new approach to simulate first order phase transitions," *Phys. Rev. Lett.* **68** (1992) 9, arXiv:hep-lat/9202004. 77

[61] W. Janke, "First-Order Phase Transitions," in *Computer Simulations of Surfaces and Interfaces*, B. Dunweg, D. P. Landau, and A. I. Milchev, eds. Springer, 2003. 79

[62] M. Suzuki, "Solution of Potts Model for Phase Transition," *Prog. Theor. Phys.* **37** (1967) 770. 80

[63] J. V. Jose, L. P. Kadanoff, S. Kirkpatrick, and D. R. Nelson, "Renormalization, vortices, and symmetry-breaking perturbations in the two-dimensional planar model," *Phys. Rev. B* **16** (1977) 1217. 81

[64] R. E. Bryant and D. R. O'Hallaron, *Computer Systems. A programmer's perspective.* Pearson, 2011. 83

[65] D. M. Young, *Iterative Solution of Large Linear Systems.* Dover Publications, 2003. 84

[66] U. Wolff, "Collective Monte Carlo Updating for Spin Systems," *Phys. Rev. Lett.* **62** (1989) 361. 85

[67] U. Wolff, "Asymptotic Freedom and Mass Generation in the O(3) Nonlinear σ Model," *Nucl. Phys. B* **334** (1990) 581. 88

[68] A. M. Ferrenberg, J. Xu, and D. P. Landau, "Pushing the limits of Monte Carlo simulations for the three-dimensional Ising model," *Phys. Rev. E* **97** (2018) 043301, arXiv:1806.03558 [physics.comp-ph]. 89, 143

[69] F. Kos, D. Poland, D. Simmons-Duffin, and A. Vichi, "Precision Islands in the Ising and $O(N)$ Models," *JHEP* **08** (2016) 036, arXiv:1603.04436 [hep-th]. 89

[70] C.-H. Chang, V. Dommes, R. S. Erramilli, A. Homrich, P. Kravchuk, A. Liu, M. S. Mitchell, D. Poland, and D. Simmons-Duffin, "Bootstrapping the 3d Ising Stress Tensor," arXiv:2411.15300 [hep-th]. 89

[71] H. G. Ballesteros, L. A. Fernandez, V. Martin-Mayor, and A. Munoz Sudupe, "Finite size effects on measures of critical exponents in d=3 O(N) models," *Phys. Lett. B* **387** (1996) 125, arXiv:cond-mat/9606203. 89, 90

[72] M. Campostrini, M. Hasenbusch, A. Pelissetto, and E. Vicari, "The Critical exponents of the superfluid transition in He-4," *Phys. Rev. B* **74** (2006) 144506, arXiv:cond-mat/0605083. 89

[73] M. Hasenbusch, "Monte Carlo study of an improved clock model in three dimensions," *Phys. Rev. B* **100** (2019) 224517, arXiv:1910.05916 [cond-mat.stat-mech]. 89

[74] W. Xu, Y. Sun, J.-P. Lv, and Y. Deng, "High-precision Monte Carlo study of several models in the three-dimensional U(1) universality class," *Phys. Rev. B* **100** (2019) 064525, arXiv:1908.10990 [cond-mat.stat-mech]. 89

[75] S. M. Chester, W. Landry, J. Liu, D. Poland, D. Simmons-Duffin, N. Su, and A. Vichi, "Carving out OPE space and precise $O(2)$ model critical exponents," *JHEP* **06** (2020) 142, arXiv:1912.03324 [hep-th]. 89

[76] C. Holm and W. Janke, "Critical exponents of the classical 3-D Heisenberg model: A Single cluster Monte Carlo study," *Phys. Rev. B* **48** (1993) 936, arXiv:hep-lat/9301002. 90

[77] M. Campostrini, M. Hasenbusch, A. Pelissetto, P. Rossi, and E. Vicari, "Critical exponents and equation of state of the three-dimensional Heisenberg universality class," *Phys. Rev. B* **65** (2002) 144520, arXiv:cond-mat/0110336. 90

[78] M. Hasenbusch and E. Vicari, "Anisotropic perturbations in three-dimensional O(N)-symmetric vector models," *Phys. Rev. B* **84** (2011) 125136, arXiv:1108.0491 [cond-mat.stat-mech]. 90

[79] M. Hasenbusch, "Monte Carlo study of a generalized icosahedral model on the simple cubic lattice," *Phys. Rev. B* **102** (2020) 024406, arXiv:2005.04448 [cond-mat.stat-mech]. 90

[80] S. M. Chester, W. Landry, J. Liu, D. Poland, D. Simmons-Duffin, N. Su, and A. Vichi, "Bootstrapping Heisenberg magnets and their cubic instability," *Phys. Rev. D* **104** (2021) 105013, arXiv:2011.14647 [hep-th]. 90

[81] H. J. Rothe, *Lattice Gauge Theories. An Introduction*. World Scientific, 2005. 103, 139, 142

[82] I. Montvay and G. Munster, *Quantum Fields on a Lattice*. Cambridge University Press, 1994. 103, 138, 139

[83] A. Smilga, *Lectures on Quantum Chromodynamics*. World Scientific, 2001. 103, 157

[84] A. Erdelyi, *Asymptotic Expansions*. Dover Publications, 1956. 106, 161

[85] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*. McGraw-Hill Book Company, 1978. 106, 161

[86] S. Weinberg, *The Quantum Theory of Fields. Volume I. Foundations*. Cambridge University Press, 1995. 108, 123

[87] C. Bonati, A. Pelissetto, and E. Vicari, "Lattice Abelian-Higgs model with noncompact gauge fields," *Phys. Rev. B* **103** (2021) 085104, arXiv:2010.06311 [cond-mat.stat-mech]. 109, 139, 142

[88] J. C. Collins, *Renormalization*. Cambridge University Press, 1989. 111

[89] T. R. Klassen, "The Anisotropic Wilson gauge action," *Nucl. Phys. B* **533** (1998) 557, arXiv:hep-lat/9803010. 112

[90] J. Engels, F. Karsch, H. Satz, and I. Montvay, "Gauge Field Thermodynamics for the SU(2) Yang-Mills System," *Nucl. Phys. B* **205** (1982) 545. 112

[91] J. Engels, J. Fingberg, F. Karsch, D. Miller, and M. Weber, "Nonperturbative thermodynamics of SU(N) gauge theories," *Phys. Lett. B* **252** (1990) 625. 114

[92] J. I. Kapusta and C. Gale, *Finite Temperature Field Theory. Principles and Applications*. Cambridge University Press, 2006. 115, 119

[93] A. Altland and B. Simons, *Condensed Matter Field Theory*. Cambridge University Press, 2010. 123

[94] T. DeGrand and C. DeTar, *Lattice Methods for Quantum Chromodynamics*. World Scientific, 2006. 124

[95] H. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*. Cambridge University Press, 2003. 124

[96] S. A. Gottlieb, W. Liu, D. Toussaint, R. L. Renken, and R. L. Sugar, "Hybrid Molecular Dynamics Algorithms for the Numerical Simulation of Quantum Chromodynamics," *Phys. Rev. D* **35** (1987) 2531. 124

[97] A. D. Kennedy, "Algorithms for dynamical fermions," arXiv:hep-lat/0607038. 124, 128

[98] G. Aarts, "Introductory lectures on lattice QCD at nonzero baryon number," *J. Phys. Conf. Ser.* **706** (2016) 022004, arXiv:1512.05145 [hep-lat]. 124

[99] O. Philipsen, "Lattice QCD at non-zero temperature and baryon density," in *Les Houches Summer School: Session 93: Modern perspectives in lattice QCD: Quantum field theory and high performance computing*, p. 273. 9, 2010. arXiv:1009.4089 [hep-lat]. 124

[100] M. Mesiti, *The QCD Phase Diagram at Imaginary Chemical Potential*. PhD thesis, University of Pisa, 2017. Available at https://etd.adm.unipi.it/theses/available/etd-05152017-121104/. 125

[101] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Phys. Lett. B* **195** (1987) 216. 125

[102] A. D. Kennedy, P. J. Silva, and M. A. Clark, "Shadow Hamiltonians, Poisson Brackets, and Gauge Theories," *Phys. Rev. D* **87** (2013) 034511, arXiv:1210.6600 [hep-lat]. 127, 128

[103] M. Creutz and A. Gocksch, "Higher Order Hybrid Monte Carlo Algorithms," *Phys. Rev. Lett.* **63** (1989) 9. 128

[104] I. P. Omelyan, I. M. Mryglod, and R. Folk, "Symplectic analytically integrable decomposition algorithms: classification, derivation, and application to molecular dynamics, quantum and celestial mechanics simulations," *Comput. Phys. Commun.* **151** (2003) 272. 128

[105] T. Takaishi and P. de Forcrand, "Testing and tuning new symplectic integrators for hybrid Monte Carlo algorithm in lattice QCD," *Phys. Rev. E* **73** (2006) 036706, arXiv:hep-lat/0505020. 128

[106] J. C. Sexton and D. H. Weingarten, "Hamiltonian evolution for the hybrid Monte Carlo algorithm," *Nucl. Phys. B* **380** (1992) 665. 128

[107] C. Urbach, K. Jansen, A. Shindler, and U. Wenger, "HMC algorithm with multiple time scale integration and mass preconditioning," *Comput. Phys. Commun.* **174** (2006) 87, arXiv:hep-lat/0506011. 128

[108] W.-K. Tung, *Group theory in physics*. World Scientific, 1985. 129, 138

[109] M. Hamermesh, *Group theory and its application to physical problems*. Dover Publications, 1989. 129, 138

[110] B. Simon, *Representations of Finite and Compact Groups*. American Mathematical Society, 1996. 129, 138

[111] R. M. Wilcox, "Exponential Operators and Parameter Differentiation in Quantum Physics," *J. Math. Phys.* **8** (1967) 962. 130, 132

[112] K. G. Wilson, "Confinement of Quarks," *Phys. Rev. D* **10** (1974) 2445. 138

[113] C. Gattringer and C. B. Lang, *Quantum Chromodynamic on the Lattice. An Introductiory Presentation*. Springer, 2010. 138

[114] T. Muta, *Foundations of Quantum Chromodynamics. An Introduction to Perturbative Methods in Gauge Theories*. World Scientific, 2010. 139

[115] S. Weinberg, *The Quantum Theory of Fields. Volume II. Modern Applications*. Cambridge University Press, 1996. 142, 157

[116] P. A. M. Dirac, "Gauge invariant formulation of quantum electrodynamics," *Can. J. Phys.* **33** (1955) 650. 142

[117] E. H. Fradkin and S. H. Shenker, "Phase Diagrams of Lattice Gauge Theories with Higgs Fields," *Phys. Rev. D* **19** (1979) 3682. 142

[118] S. Dimopoulos, S. Raby, and L. Susskind, "Light Composite Fermions," *Nucl. Phys. B* **173** (1980) 208. 142

[119] J. Frohlich, G. Morchio, and F. Strocchi, "Higgs phenomenon without symmetry breaking order parameter," *Nucl. Phys. B* **190** (1981) 553. 142

[120] C. Bonati, A. Pelissetto, and E. Vicari, "Three-dimensional lattice multiflavor scalar chromodynamics: interplay between global and gauge symmetries," *Phys. Rev. D* **101** (2020) 034505, arXiv:2001.01132 [cond-mat.stat-mech]. 142

[121] L. S. Brown and W. I. Weisberger, "Remarks on the Static Potential in Quantum Chromodynamics," *Phys. Rev. D* **20** (1979) 3239. 142

[122] C. Bonati, M. Caselle, and S. Morlacchi, "The Unreasonable effectiveness of effective string theory: The case of the 3D SU(2) Higgs model," *Phys. Rev. D* **104** (2021) 054501, arXiv:2106.08784 [hep-lat]. 143

[123] E. Seiler, "Upper Bound on the Color Confining Potential," *Phys. Rev. D* **18** (1978) 482. 143

[124] B. Simon and L. G. Yaffe, "Rigorous perimeter law upper bound on wilson loops," *Phys. Lett. B* **115** (1982) 145. 143

[125] F. J. Wegner, "Duality in Generalized Ising Models and Phase Transitions Without Local Order Parameters," *J. Math. Phys.* **12** (1971) 2259. 143

[126] A. H. Guth, "Existence Proof of a Nonconfining Phase in Four-Dimensional U(1) Lattice Gauge Theory," *Phys. Rev. D* **21** (1980) 2291. 143

[127] G. Arnold, B. Bunk, T. Lippert, and K. Schilling, "Compact QED under scrutiny: It's first order," *Nucl. Phys. B Proc. Suppl.* **119** (2003) 864, arXiv:hep-lat/0210010. 143

[128] S. Necco and R. Sommer, "The $N_f = 0$ heavy quark potential from short to intermediate distances," *Nucl. Phys. B* **622** (2002) 328, arXiv:hep-lat/0108008. 144

[129] G. H. Golub and C. F. Van Loan, *Matrix Computations*. The Johns Hopkins University Press, 1996. 146

[130] N. Cabibbo and E. Marinari, "A New Method for Updating SU(N) Matrices in Computer Simulations of Gauge Theories," *Phys. Lett. B* **119** (1982) 387. 147, 149, 152

[131] M. Creutz, "Overrelaxation and Monte Carlo Simulation," *Phys. Rev. D* **36** (1987) 515. 148

[132] K. J. M. Moriarty, "Monte Carlo Study of Compact U(1) Four-dimensional Lattice Gauge Theory," *Phys. Rev. D* **25** (1982) 2185. 150

[133] E. Pietarinen, "String Tension in SU(3) Lattice Gauge Theory," *Nucl. Phys. B* **190** (1981) 349. 152

[134] A. D. Kennedy and P. Rossi, "Classical mechanics on group manifolds and applications to hybrid Monte Carlo," *Nucl. Phys. B* **327** (1989) 782. 153

[135] C. Morningstar and M. J. Peardon, "Analytic smearing of SU(3) link variables in lattice QCD," *Phys. Rev. D* **69** (2004) 054501, arXiv:hep-lat/0311018. 154

[136] G. Parisi, R. Petronzio, and F. Rapuano, "A Measurement of the String Tension Near the Continuum Limit," *Phys. Lett. B* **128** (1983) 418–420. 154

[137] R. Brower, P. Rossi, and C.-I. Tan, "The External Field Problem for QCD," *Nucl. Phys. B* **190** (1981) 699. 156

[138] M. Luscher and P. Weisz, "Locality and exponential error reduction in numerical lattice gauge theory," *JHEP* **09** (2001) 010, arXiv:hep-lat/0108014. 156

[139] N. S. Manton, "The Schwinger Model and Its Axial Anomaly," *Annals Phys.* **159** (1985) 220. 157

[140] C. Cao, M. van Caspel, and A. R. Zhitnitsky, "Topological Casimir effect in Maxwell Electrodynamics on a Compact Manifold," *Phys. Rev. D* **87** no. 10, (2013) 105012, arXiv:1301.1706 [hep-th]. 157

[141] S. Coleman, *Aspects of symmetry*. Cambridge University Press, 1988. 157

[142] R. Jackiw, "Introduction to the Yang-Mills Quantum Theory," *Rev. Mod. Phys.* **52** (1980) 661. 157

[143] R. Jackiw, "Topological investigations of quantized gauge theories," in *Current algebra and anomalies*, S. B. Treiman, R. Jackiw, B. Zumino, and E. Witten, eds. World Scientific, 1985. 157

[144] E. Vicari and H. Panagopoulos, "Theta dependence of SU(N) gauge theories in the presence of a topological term," *Phys. Rept.* **470** (2009) 93, arXiv:0803.1593 [hep-th]. 158, 162

[145] I. Bars and F. Green, "Complete Integration of U(N) Lattice Gauge Theory in a Large N Limit," *Phys. Rev. D* **20** (1979) 3311. 160

[146] J.-M. Drouffe and J.-B. Zuber, "Strong Coupling and Mean Field Methods in Lattice Gauge Theories," *Phys. Rept.* **102** (1983) 1. 160, 163

[147] D. J. Gross and E. Witten, "Possible Third Order Phase Transition in the Large N Lattice Gauge Theory," *Phys. Rev. D* **21** (1980) 446. 160

[148] C. Bonanno, C. Bonati, M. Papace, and D. Vadacchino, "The θ-dependence of the Yang-Mills spectrum from analytic continuation," *JHEP* **05** (2024) 163, arXiv:2402.03096 [hep-lat]. 161

[149] C. Bonati and P. Rossi, "Topological susceptibility of two-dimensional $U(N)$ gauge theories," *Phys. Rev. D* **99** no. 5, (2019) 054503, arXiv:1901.09830 [hep-lat]. 161

[150] J. Kiskis, R. Narayanan, and D. Sigdel, "Correlation between Polyakov loops oriented in two different directions in SU(N) gauge theory on a two dimensional torus," *Phys. Rev. D* **89** no. 8, (2014) 085031, arXiv:1403.1770 [hep-th]. 163

[151] V. S. Vladimirov, *Generalized Functions in Mathematical Physics*. Mir Publisher, 1979. 163

[152] E. Eichten, K. Gottfried, T. Kinoshita, J. B. Kogut, K. D. Lane, and T.-M. Yan, "The Spectrum of Charmonium," *Phys. Rev. Lett.* **34** (1975) 369–372. [Erratum: Phys.Rev.Lett. 36, 1276 (1976)]. 166

[153] M. Caselle, M. Panero, R. Pellegrini, and D. Vadacchino, "A different kind of string," *JHEP* **01** (2015) 105, arXiv:1406.5127 [hep-lat]. 166

[154] G. Iannelli, "A topology inversion cluster algorithm for the u(1) pure gauge theory in two dimensions," Master's thesis, University of Pisa, 2018. 166