Some notes for

"Numerical Methods for Physics"

Claudio Bonati

April 12, 2024

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 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 is 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 typo, oversights, inaccuracies, errors or whatever else, please write to

claudio.bonati@unipi.it

List of abbreviations

GCD: greatest common divisor
iid: independent and identical distributed
MC: Monte Carlo
MCMC: Markov Chain Monte Carlo
pdf: probability distribution function
QFT: quantum field theory

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)

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

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 to $\overline{F}$. 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).

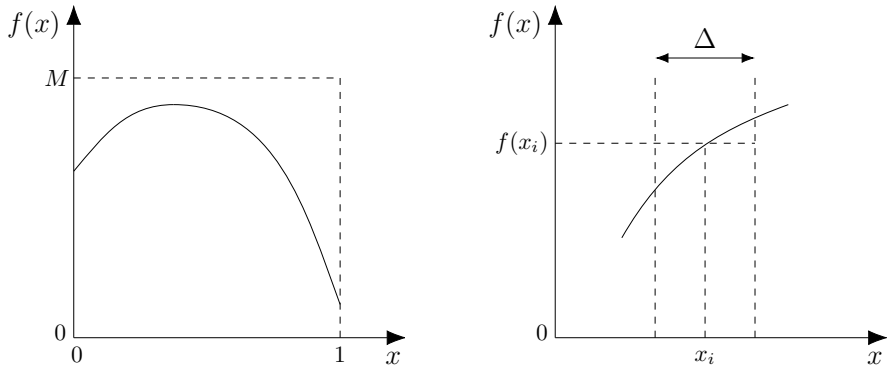

Figure 1.1: (left) The geometry considered for the integration in Sec. 1.2. (right) The basic step of the rectangle integration scheme.

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 ensamble or the one with the smallest free energy in the canonical ensamble. 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 $\langle \ \ \rangle_1$ 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 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 $[x, \mathrm{d}x]$ 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
$x = 1 - 2r_1$
$y = 1 - 2r_2$
$S = x^2 + y^2$
until $0 < S < 1$
$y_1 = \frac{x}{\sqrt{S}}\sqrt{-2\log(S)}$
$y_2 = \frac{y}{\sqrt{S}}\sqrt{-2\log(S)}$

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 to 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 x, y with uniform pdf in $[-1,1]$, keeping only the selections for which the square distance $S = x^2 + y^2$ from the origin is smaller than 1. Using the points generated in this way we thus have the following facts

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

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.

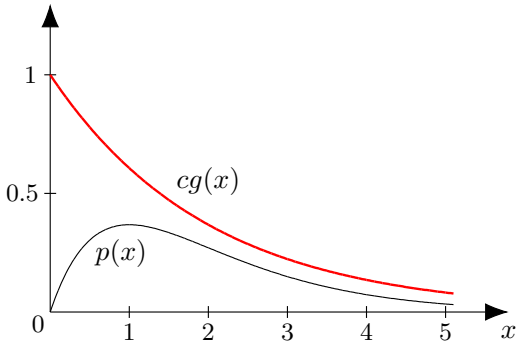

Figure 2.1: von Neumann accept/reject method: example with $p(x) = xe^{-x}$, $g(x) = \frac{1}{2}e^{-x/2}$ and $c = 2$.

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 imediately 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

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

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.

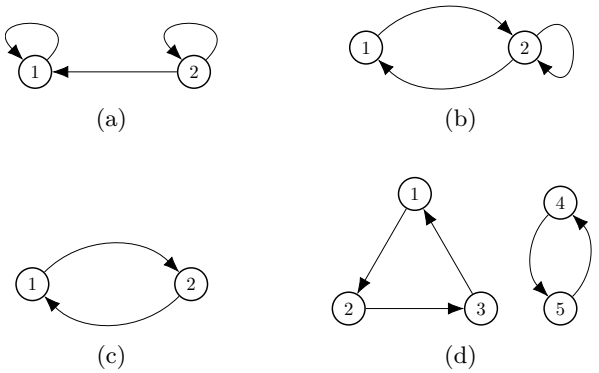

Figure 3.1: Examples of graphs associated to Markov chains.

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 to 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 to 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 to 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 to 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, if 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 to 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 to 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 $a \equiv c \bmod b$ 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 $a \equiv r_1 \bmod b$ 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 $b \equiv r_1 \bmod r_2$ 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 to 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 $|\lambda| < 1$. □

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 to 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 to 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 to 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 to 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 to 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.

Note that we have investigated the spectrum of the stochastic matrix W associated to 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 degeneration 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 to 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 to an eigenvalue of absolute value smaller than 1. The pdf p_a and the invariant pdf π_a are normalized by $\sum_a v_a = \sum_a \pi_a = 1$, 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

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

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 $\tau_{\exp} > 0$ defined by

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

can finally be written in the form

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

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 to 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

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

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

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

Note that for large k we have asymptotically

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

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

3.3 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

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

$p_a = \delta_{ab}$, 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 to 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 to an irreducible Markov chain and satisfies the detailed balance condition, then W is diagonalizable.*

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. □

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
    select a random number in [0, 1) with uniform pdf
    if r ≤ F(π_b/π_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
```

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 to 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

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

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

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

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

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}
  select 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

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.

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

[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.

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

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

3.3.3 Composition of Markov chains

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

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

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

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

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

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

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 to irreducible and aperiodic Markov chains, the composition $W = W^{(2)}W^{(1)}$ can be associated to 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 $W^{(i)}_{ab} > 0$ 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 to 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

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

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

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

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

$$\sigma^2_{\overline{F}} = \frac{\sigma^2_F}{N}(1 + 2\tau^{(F)}_{\rm int}) \ . \tag{4.1.17}$$

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 $N/(1 + 2\tau^{(F)}_{\rm int})$.

It is important to stress that $\tau_{\rm exp}$ and $\tau^{(F)}_{\rm int}$ 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 $\tau^{(F)}_{\rm int}$ 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 $\tau^{(F)}_{\rm int} \leq \tau_{\rm exp}$ 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

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

If we denote by $v^{(j)}_a$ 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 to the eigenvalue 1 (see Theorems 3.2.3-3.2.4), so $\delta F_a = \sum_{j>0} c^{(j)} v^{(j)}_a$ (the $j = 0$ eigenvalue is $\lambda = 1$) and from $\lambda_j \in (-1, 1)$ if $j \neq 0$ we have

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

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))

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

and clearly (see Eq. (3.2.23))

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

hence

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

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

We have computed $\sigma^2_{\overline{F}}$, 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 $O(k^{-5})$, 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

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

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.

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:

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

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 sum 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)

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

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 $\tau_{\rm exp}$) we have

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

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

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

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

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 = exp(−½ x̄² + ½ x_k²)
    if y ≥ 1 then
        x_{k+1} = x̄
    else
        select r ∈ [0, 1) with uniform pdf
        if r ≤ min[1, 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 $\overline{\sigma}^2_{\overline{F}}$ get 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

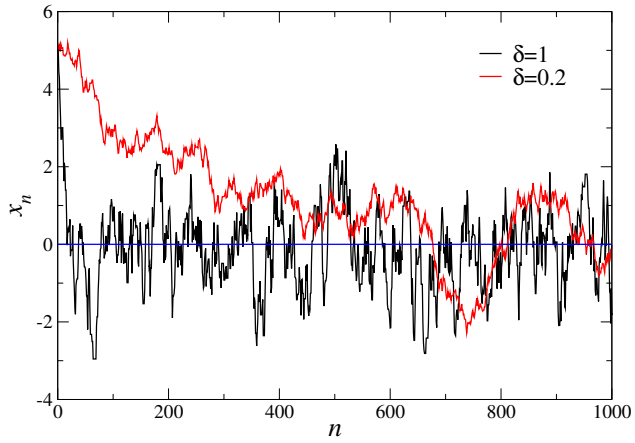

Figure 4.1: Two Monte Carlo histories obtained by performing 1000 loops of the algorithm Alg. (6), for $\delta = 1$ and $\delta = 0.2$.

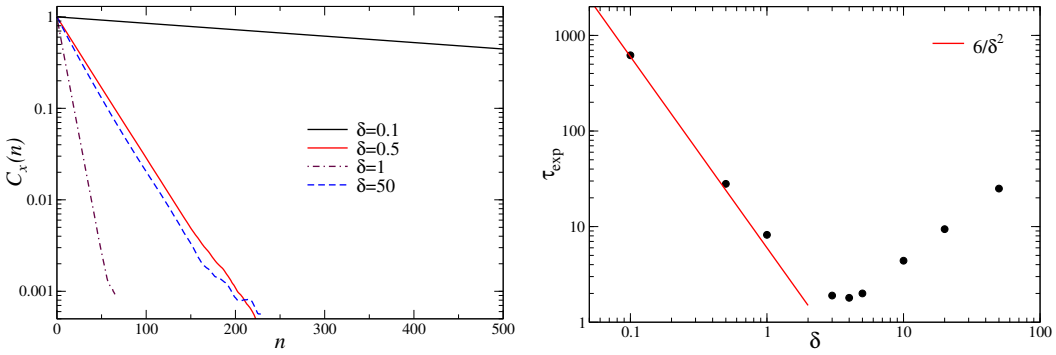

Figure 4.2: (left) Autocorrelation function $C_x(n)$ of the numbers obtained using the algorithm Alg. (6) for several values of the parameter δ. (right) The fitted exponential autocorrelation time as a function of δ.

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 $\delta \ll 1$ [25]: 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 $\sigma_x^2 = 1$)

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

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

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

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 $|y| \leq \delta \ll 1$)

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

hence

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

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.

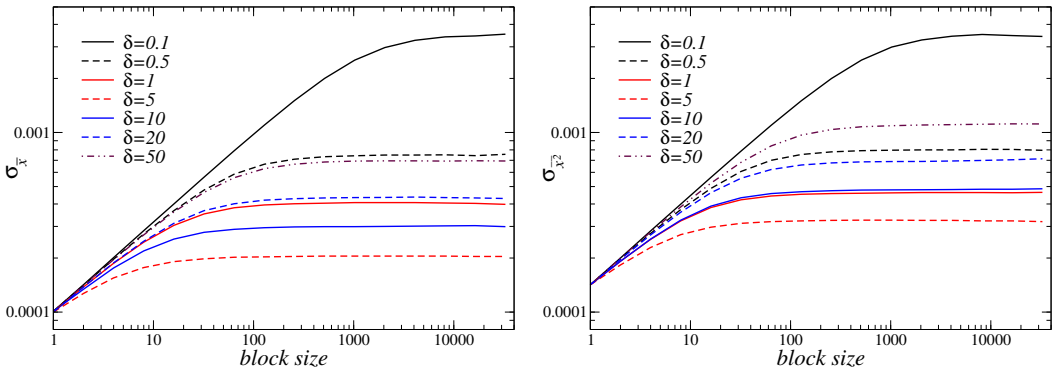

Figure 4.3: (left) Blocking analysis of $\sigma_{\overline{x}}$ for the output of the algorithm Alg. (6) for several values of the parameter δ. (right) Blocking analysis of $\sigma_{\overline{x^2}}$ for the output of the algorithm Alg. (6) for several values of the parameter δ.

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 to 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 to 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 a 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

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

where $\sigma_{\overline{x}}^2$ 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 to $F(\overline{A}, \overline{B})$, 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

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

from which finally

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

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

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

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*

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

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 to 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 $i = 1, \ldots, N$
  for $r = 1, \ldots, R$ **do**
    $S_2 = 0$, $S_4 = 0$
    for $i = 1, \ldots, N$ **do**
      generate $j \in \{1, \ldots, N\}$ 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 $\{\overline{U_4}^{(r)}\}_{r=1,\ldots,R}$, 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 $\{x_i\}_{i=1,\ldots,N}$ we can thus generate R bootstrap samples $\{x_i^{(r)}\}_{i=1,\ldots,N}$ (the index $r = 1, \ldots, R$ 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 [26] §10-11 and [27] §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

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

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 $\sigma^2_\alpha/\sqrt{N}$ 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

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

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 labling 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

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

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), howerver 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 [26] §10, [27], see also [28].

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

Statistical mechanics and phase transitions

Chapter 5

*The Ising model: physics and simulations

Chapter 6

*Other models and algorithms

Part III

The study of path-integrals in quantum mechanics

Chapter 7

*Quantum statistical mechanics and path-integrals

Chapter 8

*MCMC in quantum mechanics: thermodynamics

Chapter 9

*MCMC in quantum mechanics: spectrum

Chapter 10

*Path-integrals with nontrivial topology

Chapter 11

*Identical particles

Part IV

The study of path-integrals in quantum field theories

Chapter 12

*Statistical quantum field theory and path-integrals

Chapter 13

*MCMC in quantum field theory: thermodynamics

Chapter 14

*MCMC in quantum field theory: spectrum

Chapter 15

*The Hybryd Monte Carlo algorithm

Chapter 16

*Gauge field theories

Chapter 17

*Two dimensional gauge field theories

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.

[2] W. Feller. *An Introduction to Probability Theory and Its Applications, volume 1.* John Wiley & Sons (1968).

[3] W. Feller. *An Introduction to Probability Theory and Its Applications, volume 2.* John Wiley & Sons (1970).

[4] P. Billingsley. *Probability and Measure.* John Wiley & Sons (1995).

[5] A. I. Khinchin. *Mathematical foundations of statistical mechanics.* Dover Publications (1949).

[6] A. D. Sokal. "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms". *In* C. DeWitt-Morette, P. Cartier and A. Folacci (Editors), "Functional Integration. Basics and applications", Springer (1997).

[7] D. H. Lehmer. "Mathematical methods in large-scale computing units". *In* H. H. Aiken (Editor), "The Annals of the Computation Laboratory of Harvard University", volume XXVI. Harvard University Press (1951). Proceedings of a Second Symposium on Large-Scale Digital Calculating Machinery (1949).

[8] D. E. Knuth. *The Art of Computer Programming, vol. 2 (Seminumerical Algorithms).* Addison-Wesley (1998).

[9] G. Marsaglia. "Random numbers fall mainly in the planes". Proc. Natl. Acad. Sci. USA, **61**, (1968) 25.

[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.

[11] M. Creutz. "Monte Carlo Study of Quantized SU(2) Gauge Theory". Phys. Rev. D, **21**, (1980) 2308.

[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).

[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.

[14] R. Durrett. *Probability. Theory and Examples.* Cambridge University Press (2018).

[15] F. R. Gantmacher. *The theory of matrices, volume 2.* American Mathematical Society (2000).

[16] S. Sternberg. *A Methematical Companion to Quantum Mechanics.* Dover Publications (2019).

[17] F. A. Berezin and M. A. Shubin. *The Schrödinger Equation.* Kluwer Academic Publishers (1991).

[18] G. Teschl. *Mathematical Methods in Quantum Mechanics With Applications to Schrödinger Operators.* American Mathematical Society (2009).

[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.

[20] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". Biometrika, **57**, (1970) 97.

[21] G. O. Roberts and J. S. Rosenthal. "Markov-chain Monte Carlo: Some pratical implications of theoretical results". Canad. J. Statist., **26**, (1998) 5.

[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.

[23] U. Wolff. "Monte Carlo errors with less errors". Comput. Phys. Commun., **156**, (2004) 143. [Erratum: Comput.Phys.Commun. 176, 383 (2007)], `hep-lat/0306017`.

[24] M. B. Priestley. *Spectral Analysis and Time Series. Volume 1. Univariate Series.* Academic Press (1981).

[25] M. D'Elia. "Appunti del Corso di Metodi Numerici della Fisica Teorica, Parte I" (2016).

[26] B. Efron and T. Hastie. *Computer Age Statistical Inference.* Cambridge University Press (2016).

[27] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrian and Applied Mathematics (1982).

[28] R. G. Miller. "The jackknife – a review". Biometrika, **61**, (1974) 1.