

Chapter 2

Strong Law of Large Numbers and Monte Carlo Methods

Abstract The principles of Monte Carlo methods based on the Strong Law of Large Numbers (SLLN) are detailed. A number of examples are described, some of which correspond to concrete problems in important application fields. This is followed by the discussion and description of various algorithms of simulation, first for uniform random variables, then using these for general random variables. Eventually, the more advanced topic of martingale theory is introduced, and the SLLN is proved using a backward martingale technique and the Kolmogorov zero-one law.

2.1 Strong Law of Large Numbers, Examples of Monte Carlo Methods

The fundamental result for the numerical probability field is the Strong Law of Large Numbers, which will be proved at the end of the chapter.

2.1.1 Strong Law of Large Numbers, Almost Sure Convergence

A fundamental convergence result will now be stated.

Theorem 2.1 (Strong Law of Large Numbers) *Let $(\xi^{(\ell)}, \ell \geq 1)$ be a sequence of independent and identically distributed random variables with values in \mathbb{R}^d . Assume that*

$$\mathbb{E}|\xi^{(1)}| < \infty. \quad (2.1)$$

For $N \geq 1$, denote the empirical mean of $(\xi^{(1)}, \dots, \xi^{(N)})$ by

$$\hat{S}_N := \frac{1}{N} \sum_{\ell=1}^N \xi^{(\ell)}.$$

Then, the Strong Law of Large Numbers holds true:

$$\lim_{N \rightarrow \infty} \hat{S}_N = \mathbb{E}(\xi^{(1)}), \quad \mathbb{P}\text{-a.s.} \quad (2.2)$$

Remark 2.1 The Strong Law of Large Numbers admits a reciprocal statement, which we admit: if $\mathbb{E}|\xi^{(1)}| = \infty$ then the sequence $(\hat{S}_N, N \geq 1)$ diverges \mathbb{P} -a.s.

The Strong Law of Large Numbers can be stated as follows: the sequence of empirical means $(\hat{S}_N(\omega), N \geq 1)$ converges to $\mathbb{E}(\xi^{(1)})$ almost surely in ω . This is a particular case of almost sure convergence, the definition of which we now recall.

Definition 2.1 A property is said to hold almost surely (a.s.) if it holds except on an event of probability zero. The notation \mathbb{P} -a.s. is used to stress the underlying probability measure. In particular, a sequence $(\xi_N, N \geq 0)$ of random variables converges almost surely to a random variable ξ defined on the same probability space if

$$\mathbb{P}\left(\omega \in \Omega : \lim_{N \rightarrow \infty} \xi_N(\omega) = \xi(\omega)\right) = 1.$$

The Strong Law of Large Numbers is at the core of the following **Monte Carlo method**. Let γ be some quantity which must be approximated numerically. Assume that there exists a function f and a family $(X^{(1)}, \dots, X^{(N)})$ of independent and identically distributed random variables, which are easy to simulate on computers¹ and satisfy

$$\mathbb{E}f(X^{(1)}) = \gamma. \quad (2.3)$$

Then, except on an event of probability zero, γ can be approximated as follows.

Algorithm (Monte Carlo method) Draw a sample $(X^{(1)}(\omega), \dots, X^{(N)}(\omega))$, and approximate γ by the empirical mean:

$$\gamma \simeq \hat{S}_N(\omega) := \frac{1}{N} \sum_{\ell=1}^N f(X^{(\ell)}(\omega)).$$

This is a “good” approximation as soon as N is chosen “large enough”. However the SLLN does not make precise the convergence rate of \hat{S}_N . Rigorously proving the SLLN and finding its precise convergence rate is one of our main goals in this chapter and the next one.

To summarize: the Monte Carlo methods in this book consist in:

- exhibiting a **probabilistic representation** of γ of the type (2.3) such that the probability distribution of $X^{(1)}$ can efficiently be simulated,
- and then applying the Strong Law of Large Numbers in order to approximate γ .

¹This means that there exists a low complexity algorithm for generating sequences of independent samples from their common probability distribution.

Versions of the Strong Law of Large Numbers, under various sets of hypotheses can be proved in many ways. In particular, it is unnecessary to assume that the random variables $\xi^{(\ell)}$ are independent or identically distributed.

We will prove Theorem 2.1 in Sect. 2.3. We choose to use martingale techniques because this family of processes plays an important role in the sequel.

In the rest of this section we present some examples of Monte Carlo methods.

2.1.2 Buffon's Needle

Divide the two-dimensional space into vertical strips whose width is 1 cm. Throw at random a needle whose length is also 1 cm. What is the probability that the needle intersects one of the vertical lines?

To answer this question, one needs to make precise the probabilistic model. For instance, we define the random throwing of the needle as follows: the distance X of the center of the needle to the next line at its left side is a random variable with uniform distribution on $[0, 1]$, and the angle θ between the needle and the horizontal axis is a random variable with uniform distribution on $[-\frac{\pi}{2}, \frac{\pi}{2}]$ which is independent of X . The needle intersects a vertical line if

$$X(\omega) \in \left[0, \frac{1}{2} \cos(\theta(\omega))\right] \cup \left[1 - \frac{1}{2} \cos(\theta(\omega)), 1\right].$$

An easy calculation then shows that the desired probability is $\frac{2}{\pi}$.

In 1850, an astronomer from Zürich, R. Wolf, approximated π by the Monte Carlo method using 5000 samples: he set $\xi^{(\ell)}(\omega) = 1$ when the needle intersected a vertical line at the sample ℓ , computed the average

$$\frac{2 \times 5000}{\xi^{(1)} + \dots + \xi^{(5000)}},$$

and obtained 3.1596 as an approximation.

2.1.3 Neutron Transport Simulations

Consider a bounded continuous map λ from $\mathbb{R}^d \times \mathbb{R}^d$ to \mathbb{R}_+ (in neutron transport theory this map is called a scattering diffusion cross-section). In addition, for any (x, y) in $\mathbb{R}^d \times \mathbb{R}^d$ a continuous probability density $\pi^{x,y}$ on \mathbb{R}^d is given.

The random time evolution of the position of a neutron is described by the solution (X_t) of the differential equation

$$\frac{dX_t}{dt} = Y_t, \quad (X_0, Y_0) = (x, y), \quad (2.4)$$

where the velocity (Y_t) is a pure jump process in the following sense: for any ω , the map $t \rightarrow Y_t(\omega)$ is piecewise constant and right continuous; the jump times of (Y_t) and the jump amplitudes are random. For any integer n and pair (x, y) , the value of (Y_t) at the n th jump is a random variable with probability density $\pi^{x,y}$, where (x, y) is the state of (X_t, Y_t) at the time immediately preceding this jump; in other words, for any bounded continuous function f , if S_n is the n th jump time, the conditional expectation of $f(Y_{S_n})$ knowing that the state of (X_t, Y_t) immediately before S_n is (x, y) , is equal to

$$\int_{\mathbb{R}^d} f(z) \pi^{x,y}(z) dz.$$

Denote by T_n the time interval between S_n and S_{n+1} , that is, $T_n := S_{n+1} - S_n$. Knowing that the state of (X_t, Y_t) at time S_n is (x', y') , the distribution function of the random variable T_n is

$$F^{x',y'}(t) = 1 - \exp\left(-\int_0^t \lambda(X_s^{x',y'}, y') ds\right),$$

where $X_s^{x',y'}$ solves (2.4) with $X_0^{x',y'} = x'$ and $Y_s = y'$ for any s . Note that the function $F^{x',y'}$ is independent of n . In addition, for any $i \geq 0$, the random variables T_i and $Y_{S_{i+1}}$ are independent.

The stochastic process (X_t, Y_t) is constructed by recursively solving (2.4) on each time interval $[S_n, S_{n+1}[$ with $Y_t = Y_{S_n}$. The pair (X_t, Y_t) is a homogeneous Markov process, and called a transport process. When λ and π do not depend on the space variable x , (Y_t) is called a pure jump process and describes the motion of particles in a homogeneous environment.

To simplify the notation we now limit ourselves to the case $d = 1$. Let g be a function from \mathbb{R}^2 to \mathbb{R} . Suppose that there exists a function $u(t, x, y)$ of class $\mathcal{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^2)$, bounded with bounded derivatives of all orders, and such that

$$\begin{aligned} \frac{\partial u}{\partial t}(t, x, y) &= y \frac{\partial u}{\partial x}(t, x, y) - \lambda(x, y)u(t, x, y) + \lambda(x, y) \int_{\mathbb{R}} u(t, x, z) \pi^{x,y}(z) dz, \\ t &> 0, \quad x \in \mathbb{R}, \quad y \in \mathbb{R}, \end{aligned} \tag{2.5}$$

$$u(0, x, y) = g(x, y).$$

One can show that

$$u(t, x, y) = \mathbb{E}_{x,y} g(X_t, Y_t), \tag{2.6}$$

where $\mathbb{E}_{x,y}$ denotes the conditional expectation knowing that the position and velocity at time 0 respectively are x and y .

The Monte Carlo method to approximate $u(t, x, y)$ consists in simulating large number of trajectories of the process (X_t, Y_t) . The above construction of the process provides an algorithm of simulation of each trajectory.

This topic will be further developed in Sect. 6.3.4.

2.1.4 Stochastic Numerical Methods for Partial Differential Equations

The probabilistic representation (2.6) for the integro-differential equation (2.5) allowed us to construct a Monte Carlo method. This methodology can be extended to numerous linear and non-linear partial differential equations, provided that their solutions satisfy representations of the type

$$u(t, x) = \mathbb{E}\Psi(Z(t, x)),$$

where $(Z(t, x))$ is a family of suitable random variables.

Let us give an elementary example. Let ν be a strictly positive number and $u(t, x)$ be the solution of the heat equation

$$\frac{\partial u}{\partial t}(t, x) = \nu \Delta u(t, x), \quad \forall (t, x) \in]0, T] \times \mathbb{R}^d,$$

whose initial condition $u(0, \cdot) = u_0(\cdot)$ is assumed, say, to be continuous and bounded. By using the analytical expression of the Gaussian density of W_t one readily checks that

$$\forall (t, x) \in [0, T] \times \mathbb{R}^d, \quad u(t, x) = \mathbb{E}u_0(x + \sqrt{2\nu}W_t),$$

where (W_t) is an \mathbb{R}^d valued standard Brownian motion (thus, for any t , the components of the random vector W_t are independent and Gaussian, have zero mean and variance equal to t). Therefore one can approximate $u(t, x)$ by

$$\frac{1}{N} \sum_{\ell=1}^N u_0(x + \sqrt{2\nu t} \xi^{(\ell)}(\omega)),$$

where the $\{\xi^{(\ell)}\}$ are \mathbb{R}^d valued independent Gaussian vectors with zero mean and unit covariance matrix.

The linear parabolic partial differential equations related to European option prices in classical diffusion models are examples of equations whose solutions admit probabilistic representations. However these representations, which are called Feynman–Kac’s formulas, involve processes which are much more complex than Brownian motions, that is, the solutions of stochastic differential equations (see Chap. 7).

Another example is the Poisson equation in \mathbb{R}^d

$$Lu(x) := \operatorname{div}(a(x)\nabla u(x)) = f(x),$$

where $a(x)$ is a real-valued function. This equation arises in various fields, e.g., in Geophysics and in Molecular Dynamics. When the function $a(x)$ is smooth, under

suitable other hypotheses, one can prove the following equality which is analogous to (2.6):

$$u(x) = \int_0^\infty \mathbb{E}_x \left(f(X_t) - \int f(\xi) \mu(d\xi) \right) dt, \quad (2.7)$$

where (X_t) is the solution to a certain stochastic differential equation, \mathbb{E}_x denotes the conditional expectation knowing that X_0 is equal to x , and μ is the limit probability law, when t tends to infinity, of the law of X_t . The stochastic numerical method combines the standard Monte Carlo method and long time simulations of (X_t) : this leads to important numerical difficulties which are current subjects of research.

In addition, one often needs to consider discontinuous functions $a(x)$. For example, in Geophysics, the discontinuities of $a(x)$ reflect the soil heterogeneity. In such cases, the formula (2.7) does not involve the solution of a classical stochastic differential equation and, when the state space is multi-dimensional, the construction of *easy-to-simulate* processes (X_t) satisfying (2.7) is being investigated by many authors.

Stochastic numerical methods are being developed for various Partial Differential Equations, including non-linear ones such as Boltzmann equations, Vlasov equations, Navier–Stokes equations, Burgers equation, variational inequalities, etc. This difficult subject is out of the scope of this monograph.

We conclude this subsection by emphasizing three advantages of the numerical resolution of partial differential equations by stochastic methods: it not only allows one to solve problems in large dimension, but also:

- Monte Carlo methods allow to compute solutions whose gradient is locally very large: whereas deterministic methods require thin grids in the areas where the gradient of the solution is large, stochastic particles methods are grid-free and concentrate the simulated particles, and therefore the numerical information, in these areas.
- Most often, Monte Carlo methods are simpler and faster to code than deterministic methods, and the computer programs for stochastic numerical methods are easier to modify and adapt.
- Monte Carlo methods are naturally propitious for parallel or grid computing.

2.2 Simulation Algorithms for Simple Probability Distributions

Before introducing the more advanced material necessary for the proof of the SLLN, we pursue the subject of stochastic simulation, which will play a key role in the actual implementation of the Monte Carlo methods developed in the sequel.

We describe various methods to simulate samples, first from the uniform distribution on $[0, 1]$, and then from classical probability distributions on \mathbb{R} or \mathbb{R}^d .

For more insight in this topic, we recommend for instance the books of Devroye [10] and Asmussen and Glynn [4].

2.2.1 Uniform Distributions

The following theorem, which has many variants (see, e.g., Kuipers and Niederreiter [29]) allows one to produce sequences (u_n) on $[0, 1]$ which are uniformly distributed in the following sense:

$$\forall 0 \leq a \leq b \leq 1, \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N \mathbb{1}_{(a,b)}(u_j) = b - a.$$

Theorem 2.2 *Let θ be a positive irrational number. The sequence*

$$u_n = n\theta \pmod{1}$$

is dense everywhere in $[0, 1]$ and is uniformly distributed in $[0, 1]$.

The sequences (u_n) in the above theorem have poor statistical properties: in particular, two consecutive terms in the sequence are strongly correlated. In addition, irrational numbers cannot be represented exactly in computers.

Many algorithms have been designed to generate sequences of “pseudo-random” numbers with statistical properties close to those of sequences of samples of independent and uniformly distributed random variables. The most frequently used pseudo-random number generators are congruential methods.

Algorithm (Congruential method) Chose a triple (a, m, v_0) of integers, and compute inductively the successive samples u_k from the formula

$$v_k = av_{k-1} = a^k v_0 \pmod{m}, \quad u_k = \frac{v_k}{m}.$$

In practice, often $m = 2^\alpha$ is chosen, where α is the number of bits of the computer: then the congruence calculation reduces to truncating a bit sequence. The following simple statement indicates relevant choices of the parameters in order to maximize the periodicity of the method.²

Proposition 2.1 *If $m \geq 2^\alpha$ and $\alpha \geq 4$, then the period of the congruential method is less than $\frac{m}{4}$, and this upper bound is attained when v_0 is odd and $a = 3 \pmod{8}$ or $a = 5 \pmod{8}$.*

The preceding choices do not suffice to generate sequences with good statistical properties, that is, which statistically behave as sequences of independent samples from the uniform distribution. An example of a poor generator is

$$v_{n+1} = (2^{16} + 3)v_n \pmod{2^{32}}.$$

²This result is originally due to M. Greenberger, “Notes on a new pseudo-random number generator”, J. Assoc. Comput. Mach. **8**, 163–167 (1961).

For a survey on random number generators and a discussion on statistical tests issues, see, e.g., L'Ecuyer [33] or Gentle [19] and references therein. For theoretical issues, we refer to Niederreiter [40].

Initialization of the Samples

Monte Carlo simulations require very long sampling sequences. The root v_0 of the generator must be chosen once only, before the very first trial.

Given a root, a good generator will produce a sampling sequence with good statistical properties; however, two different sequences issued from different roots may be correlated.

We do not recommend the use of automatic initializations, e.g., by means of the computer internal clock. Being able to choose the same root in several runs of a simulation program may be useful to correct programming errors.

A Natural Question

Under which conditions is a simulation method of the uniform distribution satisfying? This is a critical issue without a universal answer.

In practice, one tests the uniform distribution hypothesis and the independence hypothesis of sampling sequences by using classical statistical procedures such as the Kolmogorov–Smirnov test, the χ^2 test, etc.

For an extended discussion on this subject and for analyses of efficient statistical tests, see Asmussen and Glynn [4], L'Ecuyer and Hellekaleke [34] and L'Ecuyer and Simard [35], for example.

Modern generators are often non-linear: for a survey on this issue, see for instance Niederreiter and Shparlinski [41].

2.2.2 Discrete Distributions

A probability distribution on a discrete set $\{x_1, x_2, \dots\}$ is given by the corresponding probability weights p_1, p_2, \dots , and a random variable X has this distribution if

$$\mathbb{P}(X = x_1) = p_1, \quad \mathbb{P}(X = x_2) = p_2, \quad \dots$$

Such a random variable X can be simulated by the following procedure:

Algorithm (Discrete distribution) To obtain a sample x from a discrete distribution giving weight p_i to x_i for $i \geq 1$: draw a sample u from the uniform distribution on $[0, 1]$, and set $x = x_n$ for the $n \geq 1$ satisfying $\sum_{i=1}^{n-1} p_i < u \leq \sum_{i=1}^n p_i$.

2.2.3 Gaussian Distributions

There are various simulation methods for the standard $\mathcal{N}(0, 1)$ Gaussian distribution, with density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ on \mathbb{R} .

An approximate simulation method uses the Central Limit Theorem, which will be recalled in Chap. 3:

Algorithm (Gaussian distribution, approximate) To obtain an approximate sample x of the $\mathcal{N}(0, 1)$ distribution, draw N independent samples u_1, \dots, u_N of the uniform distribution on $[0, 1]$, and compute

$$x = \sqrt{\frac{12}{N}} \left(\sum_{i=1}^N u_i - \frac{N}{2} \right).$$

In practice, the choice $N = 12$ is often made: empirical studies show that the corresponding sampling sequences have satisfying statistical properties, and the preceding formula simplifies into $\sum_{i=1}^{12} u_i - 6$.

The Box–Muller Method

This is an exact simulation technique, which moreover seems more efficient than the preceding algorithm in terms of computational time.

Two independent samples of the $\mathcal{N}(0, 1)$ Gaussian distribution are obtained by using the following result (due to Wiener):

Proposition 2.2 *Let U and V be independent random variables uniformly distributed on $[0, 1]$. Then the random variables X and Y defined by*

$$X = \sqrt{-2 \log U} \sin(2\pi V), \quad Y = \sqrt{-2 \log U} \cos(2\pi V),$$

have a $\mathcal{N}(0, 1)$ Gaussian distribution and are independent.

From this we deduce the following simulation algorithm.

Algorithm (Gaussian distribution, Box–Muller) To obtain two independent samples x and y from the $\mathcal{N}(0, 1)$ Gaussian distribution: draw two independent samples u and v from the uniform distribution on $[0, 1]$, and compute

$$x = \sqrt{-2 \log u} \sin(2\pi v), \quad y = \sqrt{-2 \log u} \cos(2\pi v).$$

In order to simulate a random variable X with $\mathcal{N}(m, \sigma^2)$ Gaussian distribution, observe that if Y is a $\mathcal{N}(0, 1)$ Gaussian variable then $X := \sigma Y + m$ has the appropriate distribution. More generally, the following allows to simulate a Gaussian random vector $X = (X^1, \dots, X^d)$ with mean vector m and covariance matrix C .

Algorithm (Gaussian vector distribution) To obtain a sample x from the $\mathcal{N}(m, C)$ distribution, where m is in \mathbb{R}^d and C is a symmetric non-negative $d \times d$ matrix: compute

$$\sigma := (\sigma_{ij})_{1 \leq i, j \leq d} = \begin{cases} \sigma_{i1} := \frac{C_{i1}}{\sqrt{C_{11}}}, & 1 \leq i \leq d, \\ \sigma_{ii} := \sqrt{C_{ii}} - \sum_{j=1}^{i-1} |C_{ij}|^2, & 1 < i \leq d, \\ \sigma_{ij} := \frac{C_{ij} - \sum_{k=1}^{j-1} \sigma_{ik} \sigma_{jk}}{\sigma_{jj}}, & 1 < j < i \leq d, \\ \sigma_{ij} := 0, & i < j \leq d, \end{cases}$$

draw a vector of independent samples $y = (y^1, \dots, y^d)$ from the $\mathcal{N}(0, 1)$ Gaussian distribution, and compute $x = \sigma y + m$.

2.2.4 Cumulative Distribution Function Inversion, Exponential Distributions

Monte Carlo methods for neutron transport partial differential equations, network models, neurons firing train models, etc., require to sample the exponential distribution $\mathcal{E}(\lambda)$ with parameter $\lambda > 0$, with density $\lambda e^{-\lambda x} \mathbb{1}_{\{x \geq 0\}}$ on \mathbb{R} (actually, \mathbb{R}_+).

This issue will be seen to be crucial in the sequel, and we will solve it using a general result for simulation of real random variables: the cumulative distribution function inversion method.

Definition 2.2 The cumulative distribution function (c.d.f.) of a probability distribution \mathbb{P} on \mathbb{R} , or of a real random variable X with distribution P , is the function

$$F : x \in \mathbb{R} \mapsto F(x) = P((-\infty, x]) = \mathbb{P}(X \leq x).$$

The (left-continuous) inverse of a c.d.f. F is the function

$$F^{\leftarrow} : u \in [0, 1] \mapsto F^{\leftarrow}(u) = \inf\{y \in \mathbb{R} : F(y) \geq u\}.$$

Note that if F is a bijection then $F^{\leftarrow} = F^{-1}$, and that the c.d.f. F characterizes the probability distribution P .

Theorem 2.3 Let P be a probability distribution on \mathbb{R} with c.d.f. F , and F^{\leftarrow} be its inverse. If U is a uniformly random variable on $[0, 1]$, then the random variable $X := F^{\leftarrow}(U)$ has c.d.f. F and hence distribution P .

Exercise 2.1 Prove this result.

An important application of the c.d.f. inversion method is the simulation of exponential $\mathcal{E}(\lambda)$ random variables.

Algorithm (Exponential distribution) To obtain a sample x from the $\mathcal{E}(\lambda)$ exponential distribution, $\lambda > 0$: draw a sample u from the uniform distribution on $[0, 1]$, and compute

$$x = -\frac{1}{\lambda} \log(u).$$

Another similar application is for Cauchy random variables.

Algorithm (Cauchy distribution) To obtain a sample x from the Cauchy distribution with density function $\frac{\sigma}{\pi(x^2 + \sigma^2)}$ on \mathbb{R} , for $\sigma > 0$: draw a sample u from the uniform distribution on $[0, 1]$, and compute $x = \sigma \tan(\pi u)$.

The actual implementation of the c.d.f. inversion method requires an explicit representation of the function F^{\leftarrow} (i.e., of F^{-1} when F is a bijection).

In practice, an alternative procedure consists in the numerical resolution of the equation $F(x) = u$ for any sampled value u of U , but the numerical cost may be high. The Newton–Raphson method is an example of this procedure.

Algorithm (Newton–Raphson method) To obtain a sample x from a strictly positive density f on \mathbb{R} , with c.d.f. $F(\cdot) = \int_0^\cdot f(y) dy$: draw a sample u from the uniform distribution on $[0, 1]$, set $x_0 = u$, compute

$$x_{k+1} = x_k - \frac{F(x_k) - u}{f(x_k)}, \quad k \geq 0$$

up to the step ℓ at which $|x_{\ell+1} - x_\ell|$ is less than a prescribed threshold, and set $x = x_{\ell+1}$.

2.2.5 Rejection Method

The rejection method is often used to simulate a random vector with density f on \mathbb{R}^d . It consists in choosing a density g on \mathbb{R}^d such that:

- the random vectors with density g are easy to simulate (for instance, g is a Gaussian density),
- there exists $\varepsilon > 0$ such that $h(x) := \varepsilon \frac{f(x)}{g(x)} \leq 1$ for all x .

Then we proceed as follows.

Algorithm (Rejection method) To obtain a sample x from the density f :

- draw independent samples y from the density g and u from the uniform distribution on $[0, 1]$, then
- – if $u \leq h(y) := \varepsilon \frac{f(y)}{g(y)}$, accept the sample y for x (i.e., set $x = y$)
- else, reject it and start over again,

(repeat until successful).

Note that a random number of rejections occur before a sample is accepted. This number of trials depends on the sampling sequence and on the chosen density g . This observation is important in practice: the rejection method is efficient if the acceptance rate is high.

The justification of the rejection method, and the control on the number of iterations, is based on what follows.

Proposition 2.3 *Let the random variables Y_1, Y_2, \dots have density g and U_1, U_2, \dots be uniform on $[0, 1]$, and all be independent. Let the rank and value of the first accepted sample be given by the random variables*

$$M := \inf \left\{ k \geq 1 : U_k \leq \varepsilon \frac{f(Y_k)}{g(Y_k)} \right\}, \quad X := Y_M.$$

Then M is a.s. finite, $\mathbb{P}(M = k) = \varepsilon(1 - \varepsilon)^{k-1}$ for $k \geq 1$ (geometric distribution), and X has density f and is independent of M . In particular $\mathbb{E}(M) = 1/\varepsilon$.

Proof By definition, for any $k \geq 2$ and open set A in \mathbb{R}^d ,

$$\begin{aligned} \mathbb{P}(M = k, Y_M \in A) &= \mathbb{P}(M = k, Y_k \in A) \\ &= \prod_{\ell=1}^{k-1} \mathbb{P}\left(U_\ell > \varepsilon \frac{f(Y_\ell)}{g(Y_\ell)}\right) \mathbb{P}\left(U_k \leq \varepsilon \frac{f(Y_k)}{g(Y_k)}, Y_k \in A\right) \end{aligned}$$

and, since f is a probability density,

$$\mathbb{P}\left(U_1 \leq \varepsilon \frac{f(Y)}{g(Y)}\right) = \int \varepsilon \frac{f(y)}{g(y)} g(y) dy = \varepsilon \int f(y) dy = \varepsilon.$$

Therefore, for all $k \geq 1$,

$$\mathbb{P}(M = k, Y_M \in A) = (1 - \varepsilon)^{k-1} \varepsilon \int_A \frac{f(y)}{g(y)} g(y) dy = \varepsilon(1 - \varepsilon)^{k-1} \int_A f(y) dy.$$

Choosing $A = \Omega$ leads to the distribution of M and shows that M is a.s. finite; in addition, the product form of the right-hand side shows that M and Y_M are independent. Finally, we also deduce

$$\mathbb{P}(Y_M \in A) = \int_A f(y) dy,$$

which shows that Y_M has density f . □

The preceding result and its proof can easily be extended, for instance as in the following proposition.

Proposition 2.4 *Let (S, \mathcal{S}) be a measurable space. Let ν and μ be two probability measures on this space. Suppose that μ is absolutely continuous w.r.t. ν and*

$$\exists \varepsilon > 0, \quad h(x) := \varepsilon \frac{d\mu}{d\nu}(x) \leq 1, \quad \nu\text{-a.s.}$$

Let (Y_n, I_n) be a sequence of independent and identically distributed random variables taking values in $S \times \{0, 1\}$. Suppose that the common probability distribution of the Y_i is ν and that

$$\mathbb{P}(I_1 = 1 \mid Y_1) = h(Y_1), \quad \text{a.s.}$$

Set

$$M := \inf\{k \geq 1 : I_k = 1\}, \quad X := Y_M.$$

Then M is a.s. finite, $\mathbb{E}(M) < \infty$, and the probability distribution of X is μ .

2.3 Discrete-Time Martingales, Proof of the SLLN

In this section we introduce the important notion of martingale processes, which plays a key role in the sequel. We then prove the Strong Law of Large Numbers using a technique which relies on the convergence of a backward martingale.

Our presentation is inspired by Jacod and Protter [24]. For a rigorous construction of the abstract conditional expectation and a systematic study of discrete-time martingales, an excellent reference book is Williams [47].

2.3.1 Reminders on Conditional Expectation

A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is given. Let X be an integrable \mathbb{R}^n -valued random variable:

$$X \in L^1(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^n).$$

For a sub- σ -field \mathcal{G} of \mathcal{F} , the conditional expectation $\mathbb{E}(X \mid \mathcal{G})$ of X knowing \mathcal{G} is defined as (a representative of) the class of random variables in $L^1(\Omega, \mathcal{G}, \mathbb{P}; \mathbb{R}^n)$ satisfying the characteristic property

$$\mathbb{E}(\mathbb{E}(X \mid \mathcal{G})Z) = \mathbb{E}(XZ), \quad \forall Z \in L^\infty(\Omega, \mathcal{G}, \mathbb{P}; \mathbb{R}), \quad (2.8)$$

where it is in fact enough to consider all Z of the form $\mathbb{1}_A$ for A in \mathcal{G} .

It can be proved that any two random variables in this class are equal except on a null probability set. The fact that statements involving conditional expectations hold true \mathbb{P} -a.s. is often left implicit.

If Y is an \mathbb{R}^k -valued random variable, the conditional expectation of X knowing Y is defined by

$$\mathbb{E}(X | Y) := \mathbb{E}(X | \sigma(Y)) \in L^1(\Omega, \sigma(Y), \mathbb{P}; \mathbb{R}^n),$$

where $\sigma(Y)$ is the σ -field generated by Y , constituted of all events $\{Y \in A\}$ for sets A in $\mathcal{B}(\mathbb{R}^k)$. Doob's lemma shows that any $\sigma(Y)$ -measurable random variable is of the form $f(Y)$ for some Borel function f . Therefore, $\mathbb{E}(X | Y)$ can be characterized as the \mathbb{P} -a.s. unique integrable random variable of the form $f(Y)$ for some Borel function f satisfying

$$\mathbb{E}(f(Y)g(Y)) = \mathbb{E}(Xg(Y)) \quad \text{for every real bounded Borel function } g.$$

One can then set $\mathbb{E}(X | Y = y) := f(y)$. Note that all suitable functions f are identical except on sets A such that $\mathbb{P}(Y \in A) = 0$. As expected, if $\mathbb{P}(Y = y) \neq 0$ then

$$\mathbb{E}(X | Y = y) = \frac{\mathbb{E}(X \mathbb{1}_{\{Y=y\}})}{\mathbb{P}(Y = y)}.$$

If $A \in \mathcal{F}$ then $\mathbb{P}(A | \mathcal{G})$ and $\mathbb{P}(A | Y)$ and $\mathbb{P}(A | Y = y)$ are respectively used as notations for $\mathbb{E}(\mathbb{1}_A | \mathcal{G})$ and $\mathbb{E}(\mathbb{1}_A | Y)$ and $\mathbb{E}(\mathbb{1}_A | Y = y)$.

We now recall some important properties of conditional expectation, which hold \mathbb{P} -a.s. Let \mathcal{G} and \mathcal{H} be two sub- σ -fields of \mathcal{F} .

- Most properties of expectation carry over to conditional expectation: linearity, positivity, Jensen's and Hölder inequalities, monotone and dominated convergence, Fatou's lemma, etc.
- One may “take out what is known”: if Y is a \mathcal{G} -measurable random variable such that XY is integrable, then

$$\mathbb{E}(XY | \mathcal{G}) = Y\mathbb{E}(X | \mathcal{G}). \quad (2.9)$$

- The “tower property” holds:

$$\text{if } \mathcal{G} \subset \mathcal{H} \quad \text{then } \mathbb{E}(\mathbb{E}(X | \mathcal{H}) | \mathcal{G}) = \mathbb{E}(X | \mathcal{G}). \quad (2.10)$$

In particular $\mathbb{E}(\mathbb{E}(X | \mathcal{H})) = \mathbb{E}(X)$ (take $\mathcal{G} = \{\emptyset, \Omega\}$).

- Knowing facts independent of our purpose does not help:

$$\text{if } \mathcal{H} \text{ is independent of } \sigma(X, \mathcal{G}), \quad \text{then } \mathbb{E}(X | \sigma(\mathcal{G}, \mathcal{H})) = \mathbb{E}(X | \mathcal{G}), \quad (2.11)$$

where $\sigma(X, \mathcal{G})$ [resp. $\sigma(\mathcal{G}, \mathcal{H})$] denotes the smallest σ -field containing $\sigma(X)$ and \mathcal{G} [resp. \mathcal{G} and \mathcal{H}].

Exercise 2.2 Prove the statements in this subsection using the characteristic property of conditional expectation (2.8).

2.3.2 Martingales and Sub-martingales, Backward Martingales

Definition 2.3 Let $(\mathcal{F}_n, n \in \mathbb{N})$ be a filtration, i.e., a collection of sub- σ -fields of \mathcal{F} such that $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all n . Let $(M_n, n \in \mathbb{N})$ be a discrete-time process which is (\mathcal{F}_n) -adapted and integrable, i.e., such that $M_n \in L^1(\Omega, \mathcal{F}_n, \mathbb{P})$ for all n .

Then, the process $(M_n, n \in \mathbb{N})$ is called a martingale if

$$M_n = \mathbb{E}(M_{n+1} \mid \mathcal{F}_n), \quad \mathbb{P}\text{-a.s.}, \quad (2.12)$$

and a sub-martingale if

$$M_n \leq \mathbb{E}(M_{n+1} \mid \mathcal{F}_n), \quad \mathbb{P}\text{-a.s.} \quad (2.13)$$

The related notion of super-martingale is obtained by changing the sense of the inequality, so that $(M_n, n \in \mathbb{N})$ is a super-martingale if and only if $(-M_n, n \in \mathbb{N})$ is a sub-martingale. Note that $(M_n, n \in \mathbb{N})$ is a martingale if and only if it is both a sub-martingale and a super-martingale. Therefore, we concentrate on sub-martingales for most statements, except to make explicit some reinforcements for martingales.

Exercise 2.3 Let $(M_n, n \in \mathbb{N})$ be a sub-martingale [resp. a martingale], and ϕ be an increasing convex function [resp. a convex function] such that $\mathbb{E}|\phi(M_n)| < \infty$ for every n . Prove that $(\phi(M_n), n \in \mathbb{N})$ is again a sub-martingale.

One often needs to consider martingales and sub-martingales at random times, such as the first time at which the process hits a given threshold. Certain specific random times, called *stopping times*, have very useful properties.

Definition 2.4 A random variable T taking values in $\mathbb{N} \cup \{+\infty\}$ is a stopping time for the filtration (\mathcal{F}_n) if $\{T \leq n\} \in \mathcal{F}_n$ for every n in \mathbb{N} .

If T is a stopping time, then \mathcal{F}_T is defined as the σ -field constituted of all events A in $\mathcal{F}_\infty := \sigma(\cup_{n \in \mathbb{N}} \mathcal{F}_n)$ such that $A \cap \{T \leq n\} \in \mathcal{F}_n$ for every n in \mathbb{N} .

Our next result generalizes the sub-martingale property.

Theorem 2.4 (Optional sampling) *Let $(M_n, n \in \mathbb{N})$ be a sub-martingale. If S and T are two stopping times satisfying $S(\omega) \leq T(\omega) \leq K$, ω -a.s., for some deterministic integer K , then*

$$\mathbb{E}(M_S) \leq \mathbb{E}(M_T \mid \mathcal{F}_S), \quad \mathbb{E}(M_S) \leq \mathbb{E}(M_T).$$

Proof Observe that

$$M_S = \sum_{j=0}^K M_j \mathbb{1}_{\{S=j\}} \mathbb{1}_{\{T \geq j\}}.$$

Since $\{T \geq K + 1\} = \emptyset$, expressing $M_j \mathbb{1}_{\{T \geq j\}}$ as a telescopic sum yields

$$\begin{aligned} M_S &= \sum_{j=0}^K \sum_{k=j}^K (M_k \mathbb{1}_{\{T \geq k\}} - M_{k+1} \mathbb{1}_{\{T \geq k+1\}}) \mathbb{1}_{\{S=j\}} \\ &= \sum_{j=0}^K \sum_{k=j}^K M_k \mathbb{1}_{\{T=k\}} \mathbb{1}_{\{S=j\}} + \sum_{j=0}^K \sum_{k=j}^K (M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\}}. \end{aligned}$$

Moreover,

$$\sum_{j=0}^K \sum_{k=j}^K M_k \mathbb{1}_{\{T=k\}} \mathbb{1}_{\{S=j\}} = M_T \sum_{j=0}^K \sum_{k=j}^K \mathbb{1}_{\{T=k\}} \mathbb{1}_{\{S=j\}} = M_T$$

and hence

$$M_S = M_T + \sum_{j=0}^K \sum_{k=j}^K (M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\}}.$$

Now, if A is in \mathcal{F}_S then

$$(M_S - M_T) \mathbb{1}_A = \sum_{j=0}^K \sum_{k=j}^K (M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\} \cap A}.$$

Definition 2.4 implies that $\{T \geq k + 1\} = \{T \leq k\}^c \in \mathcal{F}_k$ and

$$\{S = j\} \cap A = \{S \leq j\} \cap A - \{S \leq j - 1\} \cap A \in \mathcal{F}_j \subset \mathcal{F}_k$$

so that taking expectations, (2.13) yields

$$\mathbb{E}((M_S - M_T) \mathbb{1}_A) = \sum_{j=0}^K \sum_{k=j}^K \mathbb{E}((M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\} \cap A}) \leq 0.$$

Since A is arbitrary, the characteristic property (2.8) allows to conclude the first inequality. The second follows by taking $A = \Omega$. \square

Definition 2.5 Let real numbers $a < b$ be fixed. For $n \geq 1$, the number of upcrossings of $[a, b]$ between times 0 and n by a sequence of real numbers $(M_k, k \in \mathbb{N})$ is defined as

$$U_n := \max\{j \geq 0 : T_j \leq n\}$$

where the $(T_j)_{j \geq 0}$ are recursively defined as follows: $T_0 = 0$ and

$$\tau_{j+1} = \inf\{k \geq T_j : M_k \leq a\}, \quad T_{j+1} = \inf\{k > \tau_{j+1} : M_k \geq b\}.$$

Theorem 2.5 *In the framework of Definition 2.5, if $(M_k, k \in \mathbb{N})$ is a sub-martingale, then*

$$\mathbb{E}(U_n) \leq \frac{1}{b-a} \mathbb{E}((M_n - a)^+).$$

Proof Set $Y_n := (M_n - a)^+$. As $\tau_{n+1} > n$, one has $Y_{\tau_{n+1} \wedge n} = Y_n$, which can be expressed as the telescopic sum

$$\begin{aligned} Y_n &= Y_{\tau_1 \wedge n} + \sum_{i=1}^n (Y_{\tau_{i+1} \wedge n} - Y_{\tau_i \wedge n}) \\ &= Y_{\tau_1 \wedge n} + \sum_{i=1}^n (Y_{\tau_{i+1} \wedge n} - Y_{T_i \wedge n}) + \sum_{i=1}^n (Y_{T_i \wedge n} - Y_{\tau_i \wedge n}). \end{aligned}$$

Observe that

$$Y_{\tau_1 \wedge n} \geq 0, \quad \sum_{i=1}^n (Y_{T_i \wedge n} - Y_{\tau_i \wedge n}) \geq (b-a)U_n,$$

and thus

$$Y_n \geq \sum_{i=1}^n (Y_{\tau_{i+1} \wedge n} - Y_{T_i \wedge n}) + (b-a)U_n.$$

The result of Question 2 of Exercise 2.3 applied to the increasing convex function $\phi(x) = (x - a)^+$ yields that $(Y_n, n \in \mathbb{N})$ is a sub-martingale. Taking expectations and applying Theorem 2.4 to the stopping times $T_i \wedge n \leq \tau_{i+1} \wedge n$ then yields that

$$\mathbb{E}(Y_n) \geq (b-a)\mathbb{E}(U_n),$$

from which the result follows. \square

The notion of filtration expresses the practical fact that information is increasing with time: the σ -fields satisfy $\mathcal{F}_N \subset \mathcal{F}_{N+1}$. In the proof of Theorem 2.1, we will consider the σ -fields $\sigma(\hat{S}_N, \hat{S}_{N+1}, \dots)$, which decrease w.r.t. time $N \geq 1$, but however define a filtration when time is run backwards. We thus are led to introduce the following definition.

Definition 2.6 Let $(\mathcal{G}_{-N}, N \geq 1)$ be a family of sub- σ -fields satisfying

$$\mathcal{G}_{-(N+1)} \subset \mathcal{G}_{-N}, \quad N \geq 1. \quad (2.14)$$

A process $(M_{-N}, N \geq 1)$ is a (\mathcal{G}_{-N}) -backward martingale if the M_{-N} are integrable, \mathcal{G}_{-N} -measurable, and

$$M_{-(N+1)} = \mathbb{E}(M_{-N} | \mathcal{G}_{-(N+1)}), \quad \mathbb{P}\text{-a.s.}$$

Theorem 2.5 can then be easily adapted as follows. We leave the details as an exercise for the interested reader.

Theorem 2.6 *Let $(\mathcal{G}_{-N}, N \geq 1)$ be a family of sub- σ -fields satisfying (2.14), and $(M_{-N}, N \geq 1)$ be a (\mathcal{G}_{-N}) -backward martingale. Let $a < b$ be fixed, and for $N \geq 1$ let U_{-N} denote the number of upcrossings of the interval $[a, b]$ between times 0 and $N - 1$ by the sequence*

$$(\bar{M}_k, k \in \mathbb{N}) = (M_{-N}, \dots, M_{-1}, M_{-1}, \dots)$$

(see Definition 2.5). Then

$$\mathbb{E}(U_{-N}) \leq \frac{1}{b-a} \mathbb{E}((M_{-1} - a)^+).$$

2.3.3 Proof of the Strong Law of Large Numbers

Now, Theorem 2.1 will be proved. Recall that $(\xi^{(\ell)}, \ell \geq 1)$ is a sequence of independent identically distributed random variables such that $\mathbb{E}|\xi^{(1)}| < \infty$, and that $\hat{S}_N := \frac{1}{N} \sum_{\ell=1}^N \xi^{(\ell)}$ for $N \geq 1$.

Step 1

The key idea consists in considering

$$\mathcal{G}_{-N} := \sigma(\hat{S}_N, \hat{S}_{N+1}, \dots), \quad N \geq 1,$$

which clearly satisfies (2.14), and observing that $(\hat{S}_N, N \geq 1) = (M_{-N}, N \geq 1)$ for an appropriately defined (\mathcal{G}_{-N}) -backward martingale $(M_{-N}, N \geq 1)$.

Indeed, from (2.10) it follows that

$$M_{-N} := \mathbb{E}(\xi^{(1)} | \mathcal{G}_{-N}) \tag{2.15}$$

defines a (\mathcal{G}_{-N}) -backward martingale $(M_{-N}, N \geq 1)$. In addition, by symmetry,

$$\forall 1 \leq \ell \leq N, \quad M_{-N} = \mathbb{E}(\xi^{(\ell)} | \mathcal{G}_{-N}),$$

and therefore, since \hat{S}_N is \mathcal{G}_{-N} measurable,

$$M_{-N} = \frac{1}{N} \sum_{\ell=1}^N \mathbb{E}(\xi^{(\ell)} | \mathcal{G}_{-N}) = \mathbb{E}(\hat{S}_N | \mathcal{G}_{-N}) = \hat{S}_N.$$

Note that in particular $M_{-1} = \xi^{(1)}$.

Step 2

We prove that backward martingales such as $(M_{-N}, N \geq 1)$ converge a.s. to an integrable random variable $M_{-\infty}$.

Fix $a < b$ and use the notation of Theorem 2.6. For every ω in Ω , the increasing sequence $(U_{-N}(\omega), N \geq 1)$ has a limit $U_{-\infty}(\omega)$ in $\mathbb{R} \cup \{\infty\}$. The monotone convergence theorem and Theorem 2.6 yield

$$\mathbb{E}(U_{-\infty}) = \lim_{N \rightarrow \infty} \mathbb{E}(U_{-N}) \leq \frac{1}{b-a} \mathbb{E}((M_{-1} - a)^+) < \infty.$$

Thus the random variable $U_{-\infty}$ is finite a.s., and the sequence $(M_{-N}, N \geq 1)$ almost surely crosses $[a, b]$ a finite number of times, which implies that

$$\mathbb{P}\left(\liminf_{N \rightarrow \infty} M_{-N} < a < b < \limsup_{N \rightarrow \infty} M_{-N}\right) = 0.$$

This being true for all rational $a < b$, necessarily

$$\mathbb{P}\left(\liminf_{N \rightarrow \infty} M_{-N} < \limsup_{N \rightarrow \infty} M_{-N}\right) = 0$$

and thus, in $\mathbb{R} \cup \{-\infty, \infty\}$,

$$\lim_{N \rightarrow \infty} M_{-N} = M_{-\infty} := \liminf_{N \rightarrow \infty} M_{-N}, \quad \mathbb{P}\text{-a.s.}$$

Using (2.15) and (2.10) yields

$$\mathbb{E}(|M_{-N}|) = \mathbb{E}(|\mathbb{E}(\xi^{(1)} | \mathcal{G}_{-N})|) \leq \mathbb{E}(\mathbb{E}(|\xi^{(1)}| | \mathcal{G}_{-N})) = \mathbb{E}(|\xi^{(1)}|)$$

and Fatou's lemma yields

$$\mathbb{E}(|M_{-\infty}|) \leq \liminf_{N \rightarrow \infty} \mathbb{E}(|M_{-N}|) \leq \mathbb{E}(|\xi^{(1)}|) < \infty.$$

Step 3

We show that backward martingales such as $(M_{-N}, N \geq 1)$ converge in $L^1(\Omega)$.

This readily follows from a.s. convergence and (2.15) and classic uniform integrability results (see Problem 2.3), but we will prove it using less advanced notions.

For any $N \geq 1$ and $C \in \mathbb{R}_+$, it holds that

$$\begin{aligned} \mathbb{E}(|M_{-\infty} - M_{-N}|) &\leq \mathbb{E}(|M_{-\infty} - M_{-N}| \mathbb{1}_{\{|M_{-N}| \leq C\}}) \\ &\quad + \mathbb{E}((|M_{-\infty}| + |M_{-N}|) \mathbb{1}_{\{|M_{-N}| > C\}}). \end{aligned}$$

In view of (2.15), (2.9) and (2.10),

$$\begin{aligned}
|M_{-N}| \mathbb{1}_{\{|M_{-N}| > C\}} &= |\mathbb{E}(\xi^{(1)} | \mathcal{G}_{-N}) \mathbb{1}_{\{|M_{-N}| > C\}}| = |\mathbb{E}(\xi^{(1)} \mathbb{1}_{\{|M_{-N}| > C\}} | \mathcal{G}_{-N})| \\
&\leq \mathbb{E}(|\xi^{(1)}| \mathbb{1}_{\{|M_{-N}| > C\}} | \mathcal{G}_{-N})
\end{aligned}$$

and hence

$$\mathbb{E}(|M_{-N}| \mathbb{1}_{\{|M_{-N}| > C\}}) \leq \mathbb{E}(\mathbb{E}(|\xi^{(1)}| \mathbb{1}_{\{|M_{-N}| > C\}} | \mathcal{G}_{-N})) = \mathbb{E}(|\xi^{(1)}| \mathbb{1}_{\{|M_{-N}| > C\}}).$$

Thus, for $Y := |M_{-\infty}| + |\xi^{(1)}|$ it holds that

$$\mathbb{E}(|M_{-\infty} - M_{-N}|) \leq \mathbb{E}(|M_{-\infty} - M_{-N}| \mathbb{1}_{\{|M_{-N}| \leq C\}}) + \mathbb{E}(Y \mathbb{1}_{\{|M_{-N}| > C\}}). \quad (2.16)$$

Since $M_{-\infty}$ is integrable, by the Dominated Convergence Theorem,

$$\lim_{N \rightarrow \infty} \mathbb{E}(|M_{-\infty} - M_{-N}| \mathbb{1}_{\{|M_{-N}| \leq C\}}) = 0.$$

In addition, for any $B \in \mathbb{R}_+$,

$$\mathbb{E}(Y \mathbb{1}_{\{|M_{-N}| > C\}}) \leq \mathbb{E}(Y \mathbb{1}_{\{Y > B\}}) + B \mathbb{P}(|M_{-N}| > C)$$

and by dominated convergence, since Y is integrable, for any $\varepsilon > 0$ one can choose B and then C large enough to have

$$\mathbb{E}(Y \mathbb{1}_{\{Y > B\}}) < \varepsilon, \quad B \mathbb{P}(|M_{-\infty}| > C) < \varepsilon,$$

and moreover

$$\lim_{N \rightarrow \infty} B \mathbb{P}(|M_{-N}| > C) = B \mathbb{P}(|M_{-\infty}| > C) < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrarily chosen, (2.16) and the results that follow it yield

$$\lim_{N \rightarrow \infty} \mathbb{E}(|M_{-\infty} - M_{-N}|) = 0.$$

Note that this $L^1(\Omega)$ convergence and $\mathbb{E}(M_{-N}) = \mathbb{E}(\mathbb{E}(\xi^{(1)} | \mathcal{G}_{-N})) = \mathbb{E}(\xi^{(1)})$, see (2.15) and (2.10), yield that

$$\mathbb{E}(M_{-\infty}) = \lim_{N \rightarrow \infty} \mathbb{E}(M_{-N}) = \mathbb{E}(\xi^{(1)}). \quad (2.17)$$

Step 4

For every $k \geq 1$,

$$M_{-\infty} = \lim_{N \rightarrow \infty} \hat{S}_N := \lim_{N \rightarrow \infty} \frac{\xi^{(1)} + \dots + \xi^{(N)}}{N} = \lim_{N \rightarrow \infty} \frac{\xi^{(k)} + \dots + \xi^{(N)}}{N}, \quad \mathbb{P}\text{-a.s.},$$

so that $M_{-\infty}$ is measurable w.r.t. the tail σ -field $\bigcap_{k \geq 1} \sigma(\xi^{(k)}, \xi^{(k+1)}, \dots)$.

We admit the following classic result, which is an ingredient in most proofs of the Strong Law of Large Numbers (see, e.g., Williams [47] for a proof using martingales).

Theorem 2.7 (Kolmogorov zero-one law) *Let $(\xi^{(\ell)}, \ell \geq 1)$ be a sequence of independent random variables, and $\mathcal{T} := \bigcap_{k \geq 1} \sigma(\xi^{(k)}, \xi^{(k+1)}, \dots)$ be its tail σ -field. Then, any \mathcal{T} -measurable random variable is a.s. constant.*

Applying this theorem to $M_{-\infty}$ yields that this random variable is a.s. constant. It must then be a.s. equal to its expectation $\mathbb{E}(\xi^{(1)})$, see (2.17).

This concludes the proof of the SLLN.

2.4 Problems

2.1 (A Poisson Distribution Simulation Method) Recall that the Laplace transform of a non-negative random variable X , or of its law \mathbb{P}^X , is given by $\mathbb{E}(e^{-\theta X}) = \int e^{-\theta x} \mathbb{P}^X(dx)$ for all $\theta \geq 0$ such that the expectation is finite.

1. Compute the Laplace transform of the distribution function of the exponential probability law with parameter $\lambda > 0$.
2. Let $(X_k, k \geq 1)$ be a family of independent random variables with the same exponential probability distribution with parameter $\lambda > 0$. Set

$$S_N := \sum_{i=1}^N X_k.$$

Compute the Laplace transform of the distribution function of S_N . Deduce that the probability density function of S_N is

$$p_N(x) := \frac{\lambda^N}{(N-1)!} x^{N-1} e^{-\lambda x} \mathbb{1}_{x \geq 0}.$$

3. Let M be the smallest integer N such that $S_{N+1} > \lambda$. Show that this random variable has a Poisson distribution.
4. Propose a simulation method of the Poisson distribution which requires samples only of the uniform distribution on $[0, 1]$.

2.2 (Lyapunov Exponent of Linear Random Recursive Sequences) Let a and b be two real numbers. Let $1 > h > 0$ be a time discretization step. For any integer p set

$$\bar{X}_{p+1}^h(x) = \left(1 + b\sqrt{h}G_{p+1} + \left(a + \frac{b^2}{2} \right) h \right) \bar{X}_p^h(x),$$

where the G_p are mutually independent and centered Gaussian random variables with unit variance, and where $\bar{X}_0^h(x) = x$ a.s.

1. Check that the functions $|\log(|x|)| \exp(-x^2)$ and $(\log(|x|))^2 \exp(-x^2)$ are integrable over \mathbb{R} .

Hint: Use that $(\log(x))^2$ is the derivative of $x(\log(x))^2 - 2x \log(x) + 2x$.

2. Show that

$$\exists \bar{\lambda}^h \in \mathbb{R} \text{ for all } x \in \mathbb{R}^d - \{0\}, \quad \bar{\lambda}^h = \lim_{N \rightarrow +\infty} \frac{1}{Nh} \log |\bar{X}_N^h(x)|, \quad \text{a.s.}$$

3. Show that, for all x in $\mathbb{R} - \{0\}$,

$$\exists C_h \in \mathbb{R}, \quad \frac{1}{N^2} \mathbb{E}[\log |\bar{X}_N^h(x)|]^2 < C_h \quad \text{for all } N \in \mathbb{N} - \{0\}.$$

Problem 2.3 shows that the preceding inequality implies

$$\forall x \in \mathbb{R} - \{0\}, \quad \bar{\lambda}^h = \lim_{N \rightarrow +\infty} \frac{1}{Nh} \mathbb{E} \log |\bar{X}_N^h(x)|.$$

Deduce from this result that

$$\bar{\lambda}^h = \frac{1}{h} \mathbb{E} \log \left| 1 + b\sqrt{h}G_1 + \left(a + \frac{b^2}{2}\right)h \right|$$

for all h small enough and all N .

4. Let

$$Y := b\sqrt{h}G_1 + \left(a + \frac{b^2}{2}\right)h.$$

Prove that

$$\begin{aligned} \mathbb{E} \log |1 + Y| &= \mathbb{E} \left[Y - \frac{Y^2}{2} + \frac{Y^3}{3} \right] + \mathbb{E} \left[\mathbb{1}_{|Y| < 1} \left(\log(1 + Y) - Y + \frac{Y^2}{2} - \frac{Y^3}{3} \right) \right] \\ &\quad + \mathbb{E} \left[\mathbb{1}_{|Y| \geq 1} \left(\log |1 + Y| - Y + \frac{Y^2}{2} - \frac{Y^3}{3} \right) \right]. \end{aligned}$$

Deduce that $\bar{\lambda}^h = a + \mathcal{O}(h)$.

2.3 (Uniformly Integrable Random Variables (★)) Consider a sequence (X_n) of random variables with finite expectations which are uniformly integrable, that is,

$$\lim_{C \rightarrow \infty} \sup_{n \geq 0} \mathbb{E}[|X_n| \mathbb{1}_{|X_n| \geq C}] = 0. \quad (2.18)$$

1. Show that (2.18) is equivalent to the conjunction of the two following ones:

$$\sup_n (\mathbb{E}|X_n|) < \infty, \quad (2.19)$$

$$\forall \varepsilon > 0, \exists \delta(\varepsilon), \forall A \in \mathcal{F}, \quad \mathbb{P}(A) \leq \delta(\varepsilon) \Rightarrow \sup_n \mathbb{E}(|X_n| \mathbb{1}_A) < \varepsilon. \quad (2.20)$$

2. Let (X_n) be a uniformly integrable sequence, and let X be an integrable random variable. Show that the sequence $(|X_n - X|)$ is uniformly integrable.
Hint: Observe that $|X_n - X| \leq |X_n| + |X|$, and apply the preceding question.
3. Now assume in addition that (X_n) has an almost sure limit X . Show that

$$\mathbb{E}|X| < \infty.$$

Hint: Start with observing that

$$\mathbb{E}(|X_n|) = \mathbb{E}(|X_n| \mathbb{1}_{|X_n| \leq C}) + \mathbb{E}(|X_n| \mathbb{1}_{|X_n| \geq C}); \quad (2.21)$$

then use Fatou's lemma and Lebesgue's Dominated Convergence Theorem.

4. Deduce from Questions 1 and 2 that

$$\lim_{n \rightarrow \infty} \mathbb{E}|X_n - X| = 0. \quad (2.22)$$

Hint: Observe that, for all positive ε ,

$$\mathbb{E}|X_n - X| \leq \mathbb{E}[|X_n - X| \mathbb{1}_{|X_n - X| \geq \varepsilon}] + \varepsilon.$$

5. Write an alternative proof to Step 3 in Sect. 2.3.3 for

$$\lim_{N \rightarrow \infty} \mathbb{E}(|M_{-\infty} - M_{-N}|) = 0.$$

Hint: Show that $\sup_{n \geq 0} \mathbb{E}(X_n)^2 < \infty$ implies uniform integrability.

Stochastic Simulation and Monte Carlo Methods
Mathematical Foundations of Stochastic Simulation

Graham, C.; Talay, D.

2013, XVI, 260 p., Hardcover

ISBN: 978-3-642-39362-4