UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia

Galileo Galilei

Corso di Laurea Magistrale in Fisica

RARE EVENTS SIMULATION IN MODELS FOR ECOLOGY

Relatore: Prof. Amos Maritan
Correlatore: Dr. Marco Formentin

Laureando: Paolo Longhin
1017264

Anno Accademico 2015/2016

Contents

Introduction

The aim of this thesis is to illustrate the technique of importance sampling to simulate rare events and to apply it to a class of ecological models, known as multiplicative models, where rare events are crucial to correctly infer regular patterns of ecosystems.

Ecological systems are among the most studied complex systems as they are a topic of interest for scientists in various research areas besides ecology, ranging from physics, mathematics, to computer science. An ecosystem consists of a large number of interacting players - i.e. individuals belonging to different species. Interactions are between individuals of a species, between species and in general between individuals and the environment, for example atmospheric agents, whose temporal dynamics can be determined by external forces. This interplay among the ecosystem components makes difficult to obtain a deterministic description in terms of variables associated to each component. The microscopic dynamics, regarding the single individuals, often reveals itself as noisy and should then be described by probabilistic rules. All these features make the ecological systems complex and the attempt of modelling such systems leads naturally to consider their components as belonging to large families of identical microscopic units. On a macroscopic scale, self-organization arises from the dynamics of these minimal units, that evolve coupled by interaction terms.

In fact, ecological systems are characterized by the emergence of recurrent dynamical patterns.

One of most frequently observed regularities in ecology is the so called Taylor's Law (TL) [36]. Due to the high complexity of ecological systems the number of individuals of a species, also referred to as population number, can be conveniently represented by a random variable. For this random variable, TL states that variance and mean follow a power law relationship. This statement has an almost universal character. It was first observed by L. R. Taylor in 1961 [32], but since then it has been verified in a variety of systems and within significantly different areas of research, ranging from genetics [24] to finance [15], for example, other than ecology. At present, after more than fifty years since the law was first put forward for consideration, there is no agreement among researchers as to the possible microscopic mechanism giving rise to the statement [36]. The main issue consists in the universal character of the law, nevertheless other aspects of it still wait for an explanation. One of these concerns the behaviour of the power exponent of the law. Most empirical studies report TL with exponent close to 2, somehow irrespective of the details of the ecosystem to which it corresponds [1, 10, 36], whereas theoretical models allow any value for it [5–7, 21]. Thus, is Taylor's law exponent determined by ecological processes or is it a statistical artifact?

In Giometto et al. [17] authors show that limited sampling of sites or replicates, relative to the duration of observation, inevitably leads to an exponent near 2, for a very broad class of underlying processes known as multiplicative processes. Such Markovian processes are widely used to describe the evolution over time of the number of individuals in a given ecosystem. By employing Large Deviation Theory, authors show explicitly how the exponent in TL depends on the number of observations and on the duration of the census time series. Except for astronomically large samples, or times, of observations, the sampled value of the exponent must be close to 2 almost independently of process details. One of the relevant messages of the cited article is that the precise value of the TL's exponent is strongly influenced by rare events that are "invisible" when the process is observed in a limited number of trials. "Rare events" are events that occour with a very low frequency, typically of order 10^{-10} or smaller, requiring respectively 10^{10} or more independent trials to be detected. As example, in the multiplicative Markovian model that will be analysed later, if the random process counts 100 steps approximately 10^{16} independent realizations of it would be needed to reveal its most rare events. Clearly, taking into account such events results in a severe computation strain on the random number generator machine and eventually makes the simulation task impossible to complete. Here is where *importance sampling* comes into play, as a method which decreases the number of trials to make simulation feasible. In doing that, the condition that must be fulfilled is that the precision level of the simulation must not be altered. To overcome the obstacle of the low frequency with which a rare event occurs through a simulation the idea of the importance sampling method is quite simple: to change the probability law of the process to increase the probability of the event. As a consequence the event will occur with a higher frequency and a smaller number of trials will be needed to observe it in the simulation. Large Deviation Theory will have a fundamental role in finding the best probability law to simulate the process.

The thesis work starts by illustrating the theoretical background necessary to deal with rare events and then applies it to the computation of the Taylor's law power exponent for ecological models based on Markovian multiplicative processes.

Plan of the thesis

Chapter 1. The first chapter is dedicated to Large Deviation Theory. This theory treats the problem of calculating probabilities of events in which random processes take values far from what is predicted by the law of large numbers. Such events are characterized by small probabilities - they are then rare events - and will turn out to have the same characteristics of the rare events entering the computation of the Taylor's law exponent. For this reason Large Deviation Theory will be employed all through the remainder of the text, first in defining an efficient tool to estimate low probability events and then in setting the problem regarding the behaviour of the Taylor's law exponent. In order to give the theoretical knowledge needed to understand the following chapters, the main theorems and objects of the theory are presented: the concept of rate function for probability sequences, Cramér's theorem and Gärtner-Ellis theorem.

Chapter 2. As mentioned above, Taylor's law exponent will be obtained by simulating the multiplicative process representing the population number dynamics of the species. This chapter analyzes the problem of designing high-efficiency simulations, that is, simulations that provide the result, with the required precision, with the smallest possible number of repeated trials. In particular, the problem of rare events taking part in simulations is studied. It will be shown, by using Large Deviation Theory, that for the multiplicative process under study a probability distribution for the random variables exists that makes the simulation the most efficient possible. The set of results displayed is generally known as *importance sampling* technique. The first part of the chapter introduces the basic idea of the technique, then its core object, the biasing simulation distribution. The second part is dedicated to find the biasing simulation distribution for the multiplicative process of interest.

Chapter 3. In this chapter the problem regarding the behaviour of the Taylor's law power exponent is presented. In a wide variety of empirical observations, including sampling measurements, ad hoc experiments and simulations, the power exponent appears to be bounded within an almost universal range of values, irrespective of the models used to represent the species (or the system for which Taylor's law holds). Large Deviation Theory is employed to demonstrate that the possible cause of this phenomenon consists in undersampling measurements that are ineffective in detecting the rare events involved in the ecological process under study. From this it will follow that a high-efficiency simulation technique is needed to correctly evaluate the exponent, precisely the importance sampling technique discussed in Chapter 2.

Chapter 4. The importance sampling method is finally exploited to estimate the Taylor's law power exponent. The results, obtained for the class of multiplicative processes, will prove that the new simulation method correctly evaluates the exponent, as it provides estimates that depend on the underlying process and are not restricted to a particular range of values, in agreement with theoretical predictions. This will be a verification that a high-efficiency simulation technique is crucial in providing accurate estimate of the exponent when rare events are involved in the simulation process.

Chapter 1

An Introduction to Large Deviation Theory

In probability theory, Large Deviations Theory deals with processes determined by random variables taking values far from the values predicted by the law of large numbers. In particular, the first aim of Large Deviations Theory (in the following denoted LDT) consists in evaluating the probability that a sum of independent identically distributed random variables deviates from its mean, equal to the mean of each of the random variables. The law of large numbers predicts this probability tends to zero when the number of random variables in the sum grows large, but it doesn't characterize the way the probability decreases. LDT provides this information, first for sums of independent identically distributed random variables, then for a larger family of random variables, under assumptions regarding the random variables themselves or their particular functions.

Beyond mathematics, the role of LDT is recognized as fundamental in disciplines where there is the need to evaluate with the highest accuracy probabilities regarding the dynamics of complex or stochastic systems. As a consequence, LDT is found of utility in different research fields [4, 9, 20].

In defining performance of telecommunication networks LDT is applied to estimate the probability of data loss during transmission. The event of data loss is rare in today communication systems, but it could lead to the failure of the system if it is not properly detected. Here LDT has the role of estimating the probability of a system failure [22, 30]. Another field of application of LDT is finance engineering, where risk management in dealing with loans portfolios involves estimating the probability of large financial losses due to simultaneous loan defaults [18]. The role of LDT as a theory useful in preventing a system failure is known in insurance market also, where the theory is employed to evaluate the probability that a large number of claims is set within a short time window [13, 14].

Finally, in physics LDT finds application in statistical mechanics [12, 34], in problems relating to Brownian motion [31], polymer dynamics [20], percolation [19]. In order to accurately describe such dynamics probabilities of order $10^{-30} \div 10^{-50}$ or lower have to be computed. As to these problems the role of LDT is to provide methods to reduce the computation effort needed to come up with reliable results [30].

Besides applications, in physics LDT has been recognized also as a sound mathematical theory to rigorously formulate statistical mechanics itself [12], to the point that LDT has been seen as a natural mathematical language of statistical mechanics [29, 34]. As example LDT can justify the extremum principles of minimum free energy and maximum entropy [11, 34].

In this chapter an introduction to Large Deviation Theory is given by stating its fundamental theorems. These results, as anticipated in the Introduction, will be necessary in Chapter 2 to implement the importance sampling simulation technique, then in Chapter 3 to define the problem regarding the power exponent of Taylor's law. The work done here to establish the Large Deviation Theory will find its justification in Chapter 4, where the power exponent will be computed by means of the importance sampling technique.

In order to set the starting point and the mathematical context of LDT it is useful to look at the relationship between the theory and the law of large numbers.

1.1 Large Deviation Theory and the Law of Large Numbers

LDT is often presented with examples relating to sums of random variables. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent identically distributed (i.i.d.) real-valued random variables with expectation value, or mean, $\mathbb{E}[X_n] = m$. Assuming $m < \infty$, the law of large numbers (LLN) states that the sum, or *sample average*,

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges to m as $n \to \infty$ almost surely. Then, by a standard result of probability theory, S_n converges to m in probability too:

$$\lim_{n\to\infty} \mathbb{P}(|S_n - m| > \delta) = 0 \qquad \forall\delta > 0.$$

The above expression states nothing about the velocity in n with which the quantity $\mathbb{P}(|S_n - m| > \delta)$ goes to zero. The aim of LDT is to define the behaviour of this probability with respect to n. Figure 1.1 gives an example of how S_n behaves as $n \to \infty$ for sequences of uniformly or Gaussian distributed i.i.d. random variables. In both cases the outcomes converge to the respective means, but with possibly slight differences that LLN is unable to detect.

LDT results come at first in the form:

$$\mathbb{P}(|S_n - m| \geq x) = f(n)e^{-nI(x)}$$

where $f(n)$ is a sequence converging in n to zero more slowly than the inverse of the exponential sequence and $I(x)$ is a function, called the exponential *rate function* for S_n, which has some characteristic properties that will be shown later

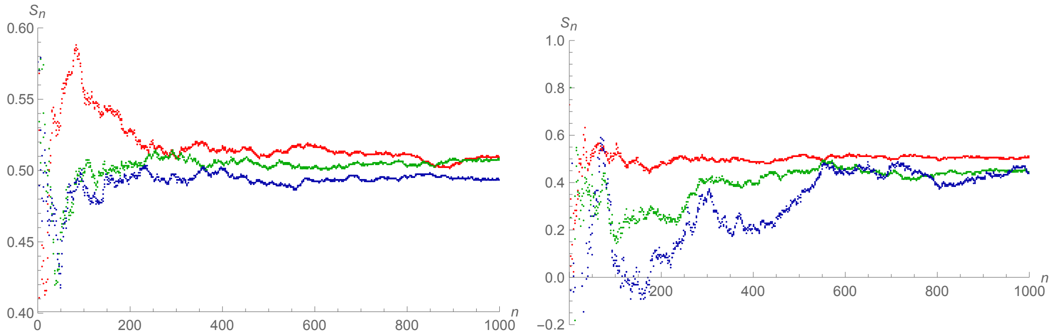


Figure 1.1: On the left is plotted S_n for different realizations of a sequence of i.i.d. r.v.s with uniform distribution in [0,1], on the right the plot refers to a sequence of Gaussian r.v.s with expectation values 1/2 but with different variances ($\sigma^2_{\text{red}} = 1$, $\sigma^2_{\text{green}} = 2$, $\sigma^2_{\text{blue}} = 4$. For all distributions S_n approaches 1/2 when n grows large. Differences are on the velocity, in n, with which this happens.

in the text. It is often difficult to find an expression for the sequence $f(n)$ and this leads LDT to state limit theorems in the simpler form:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(|S_n - m| \geq x) = -nI(x).$$

The new expression is itself a limit for $n \to \infty$ but now the function $I(x)$ gives an explicit information as to the speed with which the sequence $\mathbb{P}(|S_n - m| \geq x)$ approaches zero when n increases: $I(x)$ establishes the dependence of this speed on the point x. Furthermore, LDT gives the exact expression of $I(x)$, according to the particular sequence of random variables considered. With respect to this, LLN states only that, for any sequence of i.i.d. random variables, for $x > 0$ the rate function satisfies $I(x) > 0$.

Although the additional information provided by LDT versus LLN could seem of little importance, it will reveal itself as fundamental to compute probabilities regarding rare events.

According to the problem undertaken, LDT comes with general and ad hoc results. Cramér's theorem is the first mathematical result that laid the foundations for the Theory. Gärtner-Ellis theorem is a refinement of the former, being it applicable to a broader class of sequences of random variables.

1.2 Cramér's theorem

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d random variables taking values in $\mathbb{R}$ and with expectation value $\mathbb{E}[X_1] = m < \infty$. (The probability space of the variables is not relevant.) Cramér's theorem answers the question of determining the probability dynamics of the sequence $(S_n)_{n\in\mathbb{N}}$ where S_n is the sample average, also known as empirical average in physics, defined in section 1.1.:

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The moment generating function of the X_i random variables will be denoted by $M(\cdot)$:

$$M(\theta) = \mathbb{E}[e^{\theta X_1}], \quad \theta \in \mathbb{R}.$$

Definition 1.1. *The function*

$$I(x) = \sup_{\theta}[\theta x - \log M(\theta)]$$

is called the large deviation ***rate function*** *of the sequence* (S_n)*.*

$I(x)$ has some properties, fundamental within LDT.

Proposition 1.2. $I(\cdot)$ *is convex.*

Proof. If $\lambda \in [0,1]$, then $\forall x_1, x_2$,

$$\begin{aligned} I(\lambda x_1 + (1-\lambda)x_2) &= \sup_{\theta}[\theta(\lambda x_1 + (1-\lambda)x_2) - \log M(\theta)] \\ &= \sup_{\theta}[\theta\lambda x_1 - \lambda \log M(\theta) + \theta(1-\lambda)x_2) - (1-\lambda)\log M(\theta)] \\ &\leq \sup_{\theta}[\lambda(\theta x_1 - \log M(\theta))] + \sup_{\theta}[(1-\lambda)(\theta x_2 - \log M(\theta))] \\ &= \lambda I(x_1) + (1-\lambda)I(x_2). \end{aligned}$$

then $I(x)$ is convex. □

Proposition 1.3. *The point of minimum of $I(\cdot)$ is $x_{min} = \mathbb{E}[X_1] = m$ and $I(x_{min} = m) = 0$.*

Proof. $M(0) = 1$, then $I(x) \geq 0x - \log M(0) = 0\ \forall x$. Jensen's inequality gives $M(\theta) \geq exp(\theta m)$, hence $\theta m - \log M(\theta) \leq 0\ \forall \theta$. This implies $I(m) = 0$ and $I(x) \geq I(m)\ \forall x$. □

Proposition 1.4. *For $x > m$*

$$I(x) = \sup_{\theta \geq 0} [\theta x - \log M(\theta)],$$

and $I(\cdot)$ is an increasing function ($x > m$).

For $x < m$

$$I(x) = \sup_{\theta \leq 0} [\theta x - \log M(\theta)],$$

and $I(\cdot)$ is a decreasing function ($x < m$).

Proof. $\forall \theta \in \mathbb{R}$ by Jensen's inequality

$$\log M(\theta) = \log \mathbb{E}[e^{\theta X_1}] \geq \mathbb{E}[\log e^{\theta X_1}] = \theta m.$$

Then for $x \geq m$ and $\forall \theta < 0$

$$\theta x - \log M(\theta) \leq \theta m - \log M(\theta) \leq I(m) = 0,$$

where $I(m) = 0$ from proposition 1.3. Thus in the definition for $I(x)$, for $x > m$ the supremum is realized over positive values of θ. Finally, $\theta x - \log M(\theta)$, as function of x, is increasing, then $I(x)$ is monotone increasing on $[m, \infty]$. The result for $x < m$ is obtained analogously. □

From the above propositions it follows that $I(\cdot)$ can be ∞ for some values of its argument, thus the following definition is useful.

Definition 1.5. *For function $F(\cdot)$ the set $D_F = \{x : F(x) < \infty\}$ is called the effective domain of $F(\cdot)$.*

In what follows the interior of a set D will be denoted by $\tilde{D}$.

Proposition 1.6. *In the interior of the effective domain of* $M(\cdot)$

$$\frac{M'(\theta_x)}{M(\theta_x} = x \quad \Rightarrow \quad I(x) = \theta_x x - \log M(\theta_x) \tag{1.1}$$

Proof. The condition for $M(\cdot)$ of being differentiable in $\tilde{D}$ is assured by a standard result in real analysis (or probability theory). The proposition is demonstrated by considering $g(\theta) = \theta x - \log M(\theta)$. By calculating $g'(\theta)$ and $g''(\theta)$ it is found g is concave and $g'(\theta_x) = 0$, thus $g(\theta_x)$ is the maximum of $g(\theta)$, then (1.1) holds true. □

In the following examples rate functions are derived from absolutely continuous and discrete probability distributions.

Example 1 (Standard Gaussian random variable)**.** For $X_1 \sim \mathcal{N}(0,1)$ the moment generating function is $M(\theta) = \exp(\theta^2/2)$, then

$$I_{\mathcal{N}(0,1)}(x) = \sup_\theta[\theta x - \theta^2/2] = x^2/2.$$

Example 2 (Bernoulli random variable)**.** For $X_1 \sim Be(p)$, where $p \in (0,1)$, $M(\theta) = pe^\theta + 1 - p$, then

$$g'(\theta) = x - \frac{pe^\theta}{pe^\theta + 1 - p}$$

and $g'(\theta_x) = 0$ with

$$\theta_x = \log(\frac{x}{p}) + log\big(\frac{1-p}{1-x}\big)$$

that gives

$$M(\theta_x) = \frac{1-p}{1-x}$$

Thus

$$I_{Be(p)}(x) = \begin{cases} x\log\big(\frac{x}{p}\big) + (1-x)\log\big(\frac{1-x}{1-p}\big) & 0 \le x \le 1 \\ \infty & x < 0,\, x > 1. \end{cases}$$

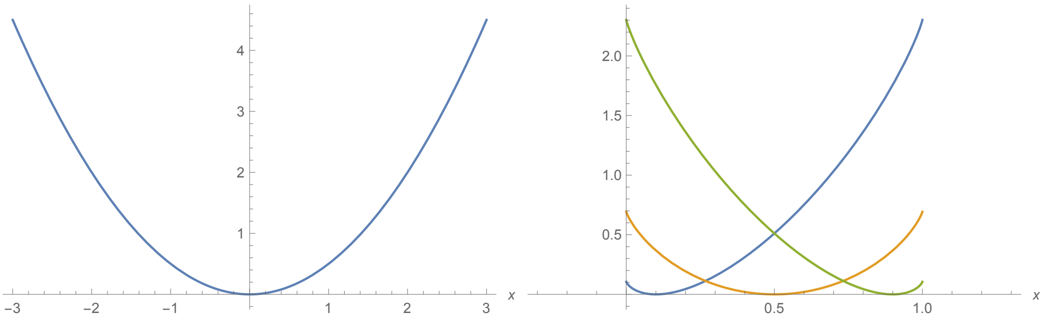


Figure 1.2: On the left is plotted $I_{\mathcal{N}(0,1)}$, on the right $I_{Be(p)}$ for different values of p ($p_{\text{blue}} = 0.1$, $p_{\text{orange}} = 0.5$, $p_{\text{green}} = 0.9$). Outside $[0, 1]$ $I_{Be(p)}$ is ∞.

Having defined the rate function and its properties it is now possible to state Cramér's theorem.

Theorem 1.7 (Cramér)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. $\mathbb{R}$-valued random variables, (S_n) the corresponding sequence of sample averages and $I(\cdot)$ its rate function.*

If $A \subset \mathbb{R}$ is a closed interval then

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \in A) \leq -\inf_{x\in A} I(x). \tag{1.2}$$

If $B \subset \mathbb{R}$ is an open interval and for every $y \in B$ there exists a θ_y for which $I(y) = \theta_y y - \log M(\theta_y)$, then

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \in B) \geq -\inf_{x\in B} I(x).$$

Proof. *Upper bound.* A is of the form $A = [a, b]$. If $m \in (a, b)$, recalling the properties of $I(\cdot)$, $I(m) = 0 \implies \inf_{x\in A} I(x) = 0$. As to the left hand side of the upper bound limit, LLN states $\lim_{n\to\infty} \mathbb{P}(S_n \in [a, b]) = 1 \quad \forall a, b : a < m < b$. Thus the two sides are both zero and the statement is true.

If $m \notin (a,\ b), m \leq a$, for $\theta \geq 0$, indicating with $P_n(x)$ the probability distribution function of S_n,

$$\begin{aligned}\mathbb{P}(S_n \in [a,\ b]) &= \int_{[a,b]} dP_n(x) = \int_{[a,b]} e^{-\theta x} e^{\theta x}\, dP_n(x) \\ &\leq e^{-\theta a} \int_{[a,b]} e^{\theta x}\, dP_n(x) \\ &\leq e^{-\theta a} \int e^{\theta x}\, dP_n(x) = e^{-\theta a} \big(M(\frac{\theta}{n})\big)^n\end{aligned}$$

(the last equality follows from independency and identical distribution of the X_i random variables). Then

$$\frac{1}{n} \log \mathbb{P}(S_n \in [a,\ b]) \leq -\frac{\theta a}{n} + \log M\big(\frac{\theta}{n}\big)$$

and θ can be replaced by $n\theta$ (because $n\theta$ still satisfies $n\theta \geq 0$), giving

$$\frac{1}{n} \log \mathbb{P}(S_n \in [a,\ b]) \leq -[\theta a - \log M(\theta)]$$

that holds true for all $\theta > 0$, thus

$$\begin{aligned}\frac{1}{n} \log \mathbb{P}(S_n \in [a,\ b]) &\leq \inf_{\theta>0} \{-[\theta a - \log M(\theta)]\} \\ &= -\sup_{\theta>0} \{\theta a - \log M(\theta)\} \\ &= -I(a) \\ &= -\inf_{x\in[a,b]} I(x).\end{aligned} \tag{1.3}$$

The upper bound limit of the statement is obtained by taking the limit superior over n on both sides.

If $m \notin (a,\ b), b \leq m$, the result is obtained by replacing X_i with $-X_i$.

Lower bound. Now B is of the form $B = (a,\ b)$. For $y \in (a,\ b)$ there always exists $\delta > 0$ so that $(y-\delta, y+\delta) \subset (a,\ b)$. The statement is proved if

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(S_n \in (y-\delta, y+\delta)\big) \geq -I(y).$$

By hypothesis there is a θ_y that gives $I(y) = \theta_y y - \log M(\theta_y)$. It can be taken $y \geq m$ (the case $y \leq m$ is handled analogously). Then, from the Proposition 2.4 θ_y must be $\theta_y \geq 0$. Now a new random variable X_{θ_y} is adopted, with cumulative distribution function (or probability distribution)

$$P_{\theta_y}(z) = \mathbb{P}(X_{\theta_y} \leq z) = \frac{\int_{-\infty}^{z} exp(\theta_y x)\, dP(x)}{M(\theta_y)}$$

(here P is the distribution function of X_i in the sequence X_n of i.i.d. random variables). X_{θ_y} has expectation value $\mathbb{E}[X_{\theta_y}] = m$. Thus, adopting LLN, for every $\varepsilon \geq 0$

$$\lim_{n\to\infty} \int_{|(x_1+\cdots+x_n)/n-y|<\varepsilon} dP_{\theta_y}(x_1)\cdot \cdots \cdot dP_{\theta_y}(x_n) = 1,$$

while for $\varepsilon < \delta$

$$\begin{aligned}\mathbb{P}\big(S_n \in (y-\delta, y+\delta)\big) &= \int_{|(x_1+\cdots+x_n)/n-y|<\delta} dP(x_1)\cdot \cdots \cdot dP(x_n)\\ &\geq \int_{|(x_1+\cdots+x_n)/n-y|<\varepsilon} dP(x_1)\cdot \cdots \cdot dP(x_n)\\ &\geq e^{(-ny-n\varepsilon)\theta_y} \int_{|(x_1+\cdots+x_n)/n-y|<\varepsilon} e^{\theta_y(x_1+\cdots+x_n)} dP(x_1)\cdots dP(x_n)\\ &\geq e^{(-ny-n\varepsilon)\theta_y} (M(\theta_y))^n \int_{|(x_1+\cdots+x_n)/n-y|<\varepsilon} dP_{\theta_y}(x_1)\cdots dP_{\theta_y}(x_n).\end{aligned}$$

Since the last inequality holds true for every n, it follows

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(S_n \in (y-\delta, y+\delta)\big) &\geq (-y-\varepsilon)\theta_y + \log M(\theta_y)\\ &= -I(y) - \varepsilon\theta_y.\end{aligned}$$

and since $\varepsilon > 0$ is arbitrary the limit $\varepsilon \to 0$ can be taken, leading to the result. □

Chernoff bound. Inequality (1.3) in the first part of the proof shows the upper bound statement (1.2) is satisfied *for all* n, not just for n large. This upper bound is sometimes referred to as the *Chernoff bound*, from the terminology of communications theory, where LDT is applied in computing probabilities of data loss events [3, 4].

Minimum rate point. Cramér's theorem shows that the asympotic behaviour ($n \to \infty$) of the probability for the sample average S_n to rest in a set $E \not\ni m$ depends only on one point, the *minimum rate point* of the set, that is the point $y \in E$ defined as $y = \inf_{x \in E} I(x)$. As example, for $\Delta > 0$ fixed but arbitrary it holds that *for every* $r > 1$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(S_n \in (m+\Delta, m+r\Delta)\big) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big((S_n \in (m+\Delta, \infty)\big),$$

which implies, by definition of limit, there exists a $\bar{n}$ so that for every $n \geq \bar{n}$

$$\mathbb{P}\big(S_n \in (m+\Delta, m+r\Delta)\big) = \mathbb{P}\big((S_n \in (m+\Delta, \infty)\big), \quad \forall r > 1.$$

This gives the minimum rate points of a set E the role of governing the rate with which $\mathbb{P}(S_n \in E)$ converges to zero.

Lower and upper bounds. In establishing the lower bound, a new probability distribution $P_{\theta_y}(\cdot)$ has been introduced so that the expectation value of the original random variables changes into the minimum rate point of the considered set: $\mathbb{E}[X_{\theta_y}] = y$, $I_{\theta_y}(y) = \inf_{x \in (a,b)} I_{\theta_y}(x)$ where $I_{\theta_y}(\cdot)$ is the rate function for the sum of the new random variables.

Results expressed in terms of upper bound for closed sets and lower bound for open sets are common within LDT. The two limits coincide for intervals, that are convex sets (Chapt. 2 in [9]). Hence, defining $I(B) = \inf_{x \in B} I(x)$, for intervals Cramér's theorem states

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \in B) = -\inf_{x \in B} I(x) = -I(B).$$

This expression is more often displayed as

$$\mathbb{P}(S_n \in B) \simeq e^{-nI(B)} \quad \text{for} \quad n \to \infty.$$

The last form conveys in the simplest and most intuitive way the improvement brought by LDT to probability theory. For $B \ni m$ proposition 1.3 gives $I(B) = 0$, then $\mathbb{P}(S_n \in B) = 0$. This is no news, since it is also known from LLN. But for $B \not\ni m$ LLN would only say $\mathbb{P}(S_n \in B) \overset{n\to\infty}{\longrightarrow} 0$: LDT describes instead exactly how $\mathbb{P}(S_n \in B)$ behaves when $n \to \infty$, according to $I(B)$. It could be said the rate function $I(\cdot)$ is all that is needed to know the asymptotic behaviour of a probability sequence.

1.3 Gärtner-Ellis theorem

Gärtner-Ellis theorem is a generalization of Cramér's theorem. This theorem makes no direct assumptions on the sequence of r.v.s and focuses instead on the sequence of their moment generating functions. An important consequence is the possibility to state large deviation results for functionals of sequences of r.v.s, including sequences showing dependency, for example Markov chains.

Assumptions. Gärtner-Ellis theorem is given here for a sequence $(Y_n)_{n\in\mathbb{N}}$ of r.v.s with values in $\mathbb{R}^d$. No conditions are requested directly on these Y_n. Assumptions are made for the sequence $(\phi_n)(\cdot)$ of functions defined as

$$\phi_n(\theta) = \frac{1}{n} \log \mathbb{E}[e^{\langle \theta, Y_n \rangle}], \qquad \theta \in \mathbb{R}^d$$

The symbol $\langle \cdot\, , \, \cdot \rangle$ denotes the usual scalar product in $\mathbb{R}^d$: $\langle \theta, Y_n \rangle = \sum_i^d \theta_i Y_{n,i}$.
For $\lambda \in [0,1]$ Holder's inequality gives, $\forall\, \theta_1,\, \theta_2$,

$$\begin{aligned} \phi_n(\lambda\theta_1 + (1-\lambda)\theta_2) &= \frac{1}{n} \log \mathbb{E}\big[(e^{\langle \theta_1, Y_n \rangle})^{\lambda}\, (e^{\langle \theta_2, Y_n \rangle})^{(1-\lambda)}\big] \\ &\le \frac{1}{n} \log\big(\mathbb{E}\big[e^{\langle \theta_1, Y_n \rangle}\big]\big)^{\lambda} + \frac{1}{n} \log\big(\mathbb{E}\big[e^{\langle \theta_2, Y_n \rangle}\big]\big)^{(1-\lambda)} \\ &= \lambda\phi_n(\theta_1) + (1-\lambda)\phi_n(\theta_2), \end{aligned}$$

then (ϕ_n) is a sequence of convex functions.
The following definitions will be used in the assumptions.

Definition 1.8 (Steepness). *A function $f \colon \mathbb{R}^d \to \mathbb{R}$ differentiable on its effective* *domain $\tilde{D}_f$ is called steep if*

$$(x_n) \subset D_f,\, x_n \to x \in \partial D_f \implies \big\|f(x_n)\big\| \to \infty.$$

Definition 1.9 (Essential smoothness)**.** *A convex function f is said essentially smooth if*

$\tilde{D}_f \neq \emptyset$,

f is everywhere differentiable in $\tilde{D}_f$,

f is steep.

Hypothesis in Gärtner-Ellis theorem consists of three assumptions on the (ϕ_n), called standard assumptions when ϕ_n are convex functions. These technical conditions are satisfied in a large number of applications, in particular that of computing rare events probability as will be done in Chapter 4.

Assumption A1. $\phi(\theta) = \lim_{n\to\infty} \phi_n(\theta) \quad \exists \quad \forall\, \theta \in \mathbb{R}^d$, with ∞ regarded both as a valid limit and a possible term in $(\phi_n)_{n\in\mathbb{N}}$. The effective domain of ϕ, D_ϕ, is convex, then $\phi(\theta)$ is itself convex, because it is the limit of a sequence of convex functions on a convex set.

Assumption A2. $0 \in \tilde{D}_\phi$ and $\forall\, \alpha \in \mathbb{R}$ the set $\{\theta \in \mathbb{R}^d : \phi(\theta) \leq \alpha\}$ is closed in $\mathbb{R}^d$.

Assumption A3. ϕ is essentially smooth.

In order to set upper and lower bounds in Gärtner-Ellis theorem a new large deviation rate function is needed.

Definition 1.10 (Gärtner-Ellis rate function)**.** *The function*

$$I(x) = \sup_\theta[\langle\theta,\, x\rangle - \phi(\theta)]$$

is called the large deviation ***rate function*** *of the sequence (Y_n).*

Even if the same notation for Cramér's rate function is adopted, it will be clear from the context which of the two is being used.

If assumptions A1, A2, A3 are met then there exists a point $m \in \mathbb{R}^d$ for which $\nabla\phi(0) = m$ (Remark 3.2.1 in [3]). It must be noted, to avoid misunderstanding, that m here is not directly related to the sequence (Y_n) and it could be not the expectation value of Y_1. This happens because the assumptions do not require the sequence (Y_n) to be of i.i.d. random variables. As example, in cases where (Y_n) shows dependency there would be no evident relationship between m and the statistics of (Y_n). Nevertheless, the point m has the same role as the mean value of the i.i.d. random variables in Cramér's theorem (Remark 3.2.1 in [3]), that is

$$m = \nabla\phi(0) \implies I(x) \geq I(m), \forall x.$$

Gärtner-Ellis theorem provides upper and lower bounds for the probability sequence $\mathbb{P}_n = \mathbb{P}\big(Y_n/n \in K\big)$, depending on K being compact, closed or open set in $\mathbb{R}^d$.

Theorem 1.11 (Gärtner-Ellis). *For every compact subset $K \subset \mathbb{R}^d$*

$$A1 \implies \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in K\big) \leq -\inf_{x\in K} I(x).$$

For every closed subset $C \subset \mathbb{R}^d$

$$A1,\ A2 \implies \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in C\big) \leq -\inf_{x\in C} I(x).$$

For every open subset $B \subset \mathbb{R}^d$

$$A1,\ A3 \implies \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in B\big) \geq -\inf_{x\in B} I(x).$$

Proof. The proof of the theorem comes through technical lemmas. It is reported in section A.1 of the Appendix.. □

Assumptions for Gärtner-Ellis theorem are necessary: if one of them is not satisfied one or more statements of the theorem could fail, as the following examples show.

Example 3. Let (Y_n), $n \in \mathbb{N}$, a sequence of independent r.v.s with distribution $P(Y_n = n) = 1/2 = P(Y_n = -n)$, then $\phi_n(\theta) = (1/n)\log\big((e^{\theta n} + e^{-\theta n})/2\big)$. Since $\lim_n \phi_n(\theta) = |\theta|$, A3 is not met by $\phi(\theta)$. $I(x)$ is null for $x \in [-1,1]$ and ∞ otherwise, while $\mathbb{P}(Y_n/n \in (-1,1)) = 0 \quad \forall n$, then

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in (-1,1)\big) = -\infty \ngeq - \inf_{x\in(-1,1)} I(x) = 0.$$

Here the left hand side does not comply with the theorem because assumption 3 is not satisfied in $\theta = 0$. Nevertheless assumptions A1 and A2 are still satisfied, then the upper bound for the compact set $[-1,1]$ is as given by the theorem: $\mathbb{P}(Y_n/n \in [-1,1]) = 1 \quad \forall n$, then

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in [-1,1]\big) = 0 \leq - \inf_{x\in[-1,1]} I(x) = 0.$$

Example 4. (Heavy tailed r.v. and sample average) Given a sequence (X_n) of i.i.d. r.v.s, $X_n \sim \mathcal{N}(1,1)$, and a r.v. $E \sim \mathit{Exp}(1)$, with X_n independent of E $\forall n$, Gärtner-Ellis theorem can be used to find the asymptotic behaviour of

$$\mathbb{P}_n = \mathbb{P}\big(E + \sum_{i=1}^{n} X_i > nA\big)$$

for some set $A \subset \mathbb{R}$. Writing $Y_n = E + \sum_{i=1}^n X_i$, being X_n independent of E $\forall n$,

$$\begin{aligned}\mathbb{E}\big[\exp(\theta Y_n)\big] &= \mathbb{E}\big[\exp(\theta E)\exp(\theta \sum_{i=1}^{n} X_i)\big] \\ &= \mathbb{E}\big[\exp(\theta E)\big]\mathbb{E}\big[\exp(\theta X_1)\big]^n\end{aligned}$$

then

$$\frac{1}{n}\log \mathbb{E}\big[\exp(\theta Y_n)\big] = \frac{1}{n}\log \mathbb{E}\big[\exp(\theta E)\big] + \log \mathbb{E}\big[\exp(\theta X_1)\big].$$

The expectation values are

$$\mathbb{E}\big[\exp(\theta E)\big] = \begin{cases} \frac{1}{1-\theta} & \theta < 1 \\ \infty & \theta \geq 1 \end{cases}$$

and

$$\mathbb{E}\big[\exp(\theta X_1)\big] = \exp\big(\theta + \frac{\theta^2}{2}\big).$$

Thus $\lim_{n\to\infty} \phi_n(\theta)$ is

$$\phi(\theta) = \begin{cases} \theta + \frac{\theta^2}{2} & \theta < 1 \\ \infty & \theta \geq 1 \end{cases}$$

(because $\phi_n(\theta)|_{\theta\geq 1} = \infty \quad \forall\, n$). The expression of ϕ shows assumption A1 is satisfied by (ϕ_n). As to assumption A2, 0 belongs to $\tilde{D}_\phi$, but the α-level set $\{\theta \in \mathbb{R}\colon \phi(\theta) \leq \alpha\} = \{\theta \in [-1-\sqrt{1+2\alpha}, -1+\sqrt{1+2\alpha}\,] \cap (-\infty, 1)\}$ and for $\alpha > 3/2$ this set is $\{\theta \in [-1-\sqrt{1+2\alpha})\}$ and is not closed (for $\alpha < -1/2$ the square root is not defined, the set is $\emptyset$, then closed). With $(\theta_n) \subset (-\infty, 1) = D_\phi$, $\theta_n \to 1 \in \partial D_\phi$, $|\nabla\phi(\theta_n)| = 1 + \theta_n \to 2 \neq \infty$, then ϕ is not steep and does not comply with assumption A3.

Therefore within Gärtner-Ellis theorem an upper bound for $\mathbb{P}(Y_n > nA)$ is assured for A compact and not for A closed (not compact), nor it can be set a lower bound when A is open. Finally, the rate function is

$$\begin{aligned} I(x) &= \sup_\theta[\theta x - \phi(\theta)] \\ &= \sup_{\theta<1}[\theta x - \theta - \frac{\theta^2}{2}] \\ &= \begin{cases} \frac{(x-1)^2}{2} & x < 2 \\ x - \frac{3}{2} & x \geq 2 \end{cases} \end{aligned}$$

The overall result is that the asymptotic drift of the probability regarding a sum of i.i.d. normal random variables may be significantly altered by adding to it even only one r.v. differently distributed. In this case the new r.v. had an exponential distribution, consequently this result holds for every added random variable that is heavy-tailed (for which $\lim_{x\to\infty} \exp(\lambda x)\mathbb{P}(X > x) = \infty$, $\forall\lambda > 0$). The central limit theorem (CLT) and the law of large numbers wouldn't provide the same information. A careless use of CLT and LLN could lead to conclude that adding the exponential distributed variable does not change $\lim_{n\to\infty} \mathbb{P}_n$. Indeed CLT and LLN statements do not contradict the result of the example, but they do not offer any means to describe how $\mathbb{P}_n$ behaves when $n \to \infty$.

A general Theory of Large Deviations is given in section A.2 of the Appendix, where Varadhan's lemma is also demonstrated.

1.4 Markov processes

It has been said Gärtner-Ellis theorem has to be used in place of Cramér's theorem if the random variables in the sequence are not i.i.d.. The simplest sequence with dependency is a Markov chain. Here the state space is assumed to be finite, for simplicity of the form

$$\chi = \{0, \dots, k\}, \quad k \in \mathbb{N} \quad \text{fixed},$$

and the chain irreducible (any state of the chain can be reached starting from any other state with a strictly positive probability). The transition probabilities will be denoted by

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i),$$

In what follows it is considered the possibility of mapping the variables X_n by a function $f\colon \chi \to \mathbb{R}$. The sample average for $f(X_n)$ is defined as it is usual: $S_n = (1/n)\sum_{i=1}^{n} f(X_n)$.

Let now π be the stationary distribution of the process: $\pi P = \pi$, where P is the chain transition matrix. By hypothesis, the chain is irreducible and its state space is finite, then the ergodic theorem assures that π exists, it is unique and that, for every initial distribution ν, the following convergence holds:

$$\sum_{x\in\chi} |\nu P^n - \pi| \overset{n\to\infty}{\longrightarrow} 0.$$

As a consequence, the ergodic theorem makes it possible to state a law of large numbers for S_n as

$$S_n \xrightarrow{P} \mathbb{E}_\pi[f] = \sum_{j=0}^{k} f(j)\pi(j).$$

The role of Large Deviations Theory in studying Markov chains is then to quantify the rate in n with which S_n approaches $\mathbb{E}_\pi[f]$. Again, it would not be possible to obtain this information with the sole LLN.

To apply LDT the sequence $\phi_n(\theta)$ introduced with the Gärtner-Ellis theorem is now

$$\phi_n(\theta) = \frac{1}{n}\log \mathbb{E}[\exp{(\theta n S_n)}].$$

Denoting with $p(x_1, \dots, x_n)$ the joint probability function for the first n steps of the chain,

$$\begin{aligned}\mathbb{E}[\exp{(\theta n S_n)}] &= \sum_{x_1\in\chi}\sum_{x_2\in\chi}\cdots\sum_{x_n\in\chi} \exp\big(\theta\sum_{i=1}^{n} f(x_i)\big)p(x_1, x_2, \dots, x_n)\\ &= \sum_{x_1=0}^{k}\sum_{x_2=0}^{k}\cdots\sum_{x_n=0}^{k} \exp\big(\theta\sum_{i=1}^{n} f(x_i)\big)p(x_1)\prod_{i=1}^{n-1} p(x_{i+1}|x_i).\end{aligned}$$

Let now $\mathcal{F}$ be the set of all functions mapping χ into $\mathbb{R}$ and be $T_\theta \colon \mathcal{F} \to \mathcal{F}$ the linear operator acting on $g \in \mathcal{F}$ as

$$T_\theta(g)(x) = \sum_{y\in\chi} e^{\theta f(y)} p_{xy}. \tag{1.4}$$

Having defined T_θ, the expectation value becomes

$$\mathbb{E}[\exp{(\theta n S_n)}] = \sum_{x_1=0}^{k} T_\theta^n(1)(x_1)\exp\big(\theta f(x_1)\big)p(x_1),$$

where the $T_\theta^n(1)$ stands for T_θ applied n times to the constant function $g = 1$. Since χ is finite the iteration of T_θ consists in multiplying n times the matrix representing the operator and applying the resulting matrix to the vector with all entries equal to 1. From definition the matrix T_θ is strictly positive ($T_{\theta\, ij} > 0 \;\forall\, i, j$) and Perron-Frobenius theorem assures T_θ has a largest eigenvalue $\lambda(\theta)$:

$$\lambda(\theta) = \sup_{g:\, \|g\| \le 1} \|T_\theta(g)\|$$

corresponding to a unique eigenvector ψ ($\|\cdot\|$ represents the Euclidean norm for vectors). Then it is possible to express T_θ^n at first order in ψ with a rest term negligible with respect to e^n:

$$T_\theta^n(g)(x) = c_g \lambda(\theta)^n \psi(x) + o(n), \qquad \lim_{n\to\infty} o(n)e^{-n} = 0, \quad c_g > 0 \text{ constant},$$

where $\psi(x)$ indicates the component of ψ corresponding to the state x. At this point the introduction of T_θ is justified:

$$\begin{aligned}
\lim_{n\to\infty} \phi_n(\theta) &= \lim_{n\to\infty} \frac{1}{n} \log \sum_{x=0}^{k} T_\theta^n(1)(x) \exp\big(\theta f((x))\big) p(x) \\
&= \lim_{n\to\infty} \frac{1}{n} \log \sum_{x=0}^{k} c_g \lambda(\theta)^n \psi(x) \exp\big(\theta f((x))\big) p(x) \\
&= \log \lambda(\theta) + \lim_{n\to\infty} \frac{1}{n} \log \sum_{x=0}^{k} c_g \psi(x) \exp\big(\theta f((x))\big) p(x) \\
&= \log \lambda(\theta).
\end{aligned}$$

Finally, $\log \lambda(\theta)$, of effective domain $\mathbb{R}$, is closed, convex and steep, then Gärtner-Ellis theorem states the rate function associated to $nS_n = \sum_i^n f(X_i)$ is

$$I(x) = \sup_\theta [\theta x - \log \lambda(\theta)].$$

This theoretical result can be applied to any irreducible Markov chain with finite state space.

Chapter 2

The Rare Event Problem in Simulation Design

Since its first application in physics, in the early 30's by E. Fermi, simulation has become a fundamental part in scientific studies. Thanks to their adaptability, and to the increasing power of computers, simulation techniques are today common to a variety of disciplines, ranging from applied mathematics [2–4], in deriving probability theory results, to finance, where it is used in risk management [18], from physics, where simulations are applied to study complex systems and dynamical processes [16, 30], to natural sciences, when ecosystems parameters are evaluated [17]. The need of employing simulation arises in cases where it is not possible, or straightforward, to obtain results by theory or numerical methods. For example, when systems with a high degree of freedom are under study, equations describing them, be they deterministic or stochastic, could be unsolvable either in analytic or numerical way [3]. In such cases simulations provide an approximate solution expressed in terms of an estimate and the probability with which it approximates the solution within a fixed maximum error.

A system simulation is performed by first assigning a probability distribution to the variables of the system, according to the model used to represent it, and then by carrying out repeated realizations of the variables. The estimate of the searched solution, usually a functional of the variables adopted in the simulation, is obtained by computing sample averages of the outcomes.

However, when low probability events enter the simulation process, the standard simulation could fail in providing the searched result and a more sophisticated simulation architecture is needed. This chapter presents such a new architecture, the *importance sampling* technique, and the theoretical framework that will serve later to apply it to the ecological question.

2.1 Estimating low probability events

The problem of a rare event entering a simulation process consists in finding a method to reduce the computational effort required on the random number generator machine to take the simulation to an end. A prototype of this problem is that of estimating the probability of the rare event itself.

Let E be the rare event and let $\mu = \mathbb{P}(E)$ be the parameter to evaluate. By definition of the expectation value, $\mathbb{E}[\cdot]$, the probability $\mathbb{P}(E)$ equals $\mathbb{E}[\mathbb{1}_E]$ where $\mathbb{1}_E$ stands for the indicator function over E ($\mathbb{1}_E = 1$ if E occurs, $\mathbb{1}_E = 0$ if E does not occur). Thanks to the identity $\mathbb{P}(E) = \mathbb{E}[\mathbb{1}_E]$, in order to get μ a simulation could be performed by repeating a number of independent experiments, or trials, in which the event can happen or not. Any time the event E occurs the corresponding output, $\mathbb{1}_E$, is set to 1, otherwise it is set to 0. The average of the outputs, according to the law of large numbers, converges to $\mathbb{P}(E)$ as the number k of independent trials increases:

$$S_k = \frac{1}{k}\sum_{j=1}^{k}(\mathbb{1}_E)_j \overset{k\to\infty}{\longrightarrow} \mathbb{E}[\mathbb{1}_E] = \mathbb{P}(E),$$

where $(\mathbb{1}_E)_j$ is the outcome in the j-th trial. From this, to obtain $\mathbb{P}(E)$ a simulation should go on indefinitely, to realize the condition $k \to \infty$. This is not possible and an approximate result is then accepted, that will be expressed by a value $\bar{\mu}$, called *estimate* of μ, a maximum error ϵ and the probability, or *confidence*, y_ϵ with which $\bar{\mu}$ lies in the *confidence interval* $[\mu - \epsilon, \mu + \epsilon]$.

The parameters ϵ and y_ϵ define the precision of the estimate. Once a precision level (ϵ, y_ϵ) has been fixed, the simulation is run until the estimate $\bar{\mu}$ satisfies that precision, that is, the simulation is interrupted only when the estimate $\bar{\mu}$ lies about μ within maximum error ϵ with the fixed probability y_ϵ.

Precision requirements may come as the most delicate part of a simulation design. Nevertheless, they control the reliability of the simulation: without a precision constraint the final output is useless, since it is not known the probability with which it matches, within the confidence interval, the parameter to be estimated.

To appreciate the difficulties that may arise in rare event simulation, it is useful to consider, as example, the problem of estimating the mean of a Bernoulli random variable by observing a sequence of i.i.d. Bernoulli random variables of parameter μ:

$$\mathbb{P}(X_1 = 1) = \mu = 1 - \mathbb{P}(X_1 = 0).$$

Since μ is given, the problem is already solved: $\mathbb{E}[X_1] = \mu$ and $Var[X_1] = \mu(1-\mu)$. But supposing not to be able to compute the mean by theory, then it can be estimated by using the sample average

$$\bar{\mu} = \frac{1}{k}\sum_{j=1}^{k} X_j.$$

The variables X_j are i.i.d., thus the expectation value of $\bar{\mu}$ is $\mathbb{E}[\bar{\mu}] = \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[X_j] = \frac{1}{k}k\mathbb{E}[X_1] = \mu$, the parameter to evaluate. As mentioned above, a maximum error ϵ on the estimate $\bar{\mu}$ and a confidence must be set for the simulation to produce a reliable result. In this example, if the estimate $\bar{\mu}$ is requested with a maximum error $\epsilon = 5\%$ with $y_\epsilon = 95\%$ confidence, then $\bar{\mu}$ must satisfy

$$\mathbb{P}\big(|\bar{\mu} - \mu| \leq 0.05\mu\big) = 0.95. \tag{2.1}$$

Since $\bar{\mu}$ depends on k, (2.1) is a condition on the number of independent trials of events $\{X_i = 1\}$. The number $\tilde{k}$ satisfying this condition will be the minimum number of independent trials to execute: any $k \geq \tilde{k}$ could be accepted, because for such k the simulation will provide, according to LLN, an even better estimate of μ.

Let now $\{X_1 = 1\}$ be a rare event: its probability is extremely low. What has to be evaluated is the minimum number $\tilde{k}$ of independent trials of events $\{X_i = 1\}$ necessary to estimate μ with the desired precision. Since, by hypothesis, $\{X_1 = 1\}$ has a low probability, $\mu \simeq 0$ and $Var[X_1] = \mu(1-\mu) \simeq \mu$. Then, because $\bar{\mu}$ is a sum of k i.i.d. random variables, $\frac{1}{k}X_j$, its variance is k times the variance of $(\frac{1}{k}X_1)$:

$$Var[\bar{\mu}] = k\frac{\mu(1-\mu)}{k^2} \simeq \frac{\mu}{k}.$$

Now, the central limit theorem states

$$\frac{\bar{\mu}-\mu}{\sqrt{Var(X_1)}}\sqrt{k} \xrightarrow{L} Y \sim \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ denotes the Gaussian distribution of mean 0 and variance 1 and the right arrow stands for convergence in distribution as $k \to \infty$. This result implies

$$\begin{aligned}
\mathbb{P}\big(|\bar{\mu}-\mu| \leq x\mu\big) &= \mathbb{P}\Big(\Big|\frac{1}{k}\sum_{j=1}^{k}(X_j-\mu)\Big| \leq 0.05\mu\Big) \\
&= \mathbb{P}\Big(\Big|\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\frac{X_j-\mu}{\sqrt{\mu}}\Big| \leq 0.05\sqrt{k\mu}\Big) \\
&\simeq \mathbb{P}\big(|Y| \leq 0.05\sqrt{k\mu}\big),
\end{aligned}$$

where the symbol $\simeq$ means the equality holds for k large. For the standard Gaussian r.v. $\mathbb{P}(|Y| \leq y) = 0.95$ with $y \simeq 2$, then, to comply with (2.1) it must be

$$0.05\sqrt{\tilde{k}\mu} \simeq 2 \iff \tilde{k} \simeq \frac{1600}{\mu}.$$

The last expression quantifies the problem: if the probability of the rare event $\{X_1 = 1\}$ were, as example, $\mu \simeq 10^{-6}$, to reach the desired precision in (2.1) $\tilde{k}$ should be $\simeq 1.6 \times 10^9$. Had the required precision been higher, $\tilde{k}$ would have reached an even greater order of magnitude. These values for $\tilde{k}$ represent an obstacle in simulation design, as they imply a long computation time and, eventually, they could reveal themselves as prohibitive for more complex simulation processes.

It could be argued at this point that every simulation requires the knowledge of the probability distribution of all the random variables involved in the process and that this knowledge in probability theory is sufficient to obtain exact formulas to compute any probability or expectation value of interest. While this is true, the task of providing the result by applying probability theory only may be difficult to perform, as shown by the next argument.

Let (Y_n) be a sequence of i.i.d. random variable in $\mathbb{R}$ with probability distribution function p and expectation value m. In many cases of interest the following probability is to compute:

$$\mu = \mathbb{P}\big(\frac{1}{n}\sum_{i=1}^{n} Y_i > L\big), \tag{2.2}$$

with $L > m$ constant. Here, as a standard result of probability theory, the computation of μ requires the knowledge of the n-fold convolution of p. This calculus becomes rapidly difficult as n increases, both by analytic and numerical means. Moreover, even when the convolution were available, the integral over $\mathbb{R}$ should be computed, and this calculus would also be hard to take to an end. In this case a simulation turns out to be a convenient tool to get the result. However, simulation itself could be ineffective as to the problem of getting the result easily. The direct simulation method would be performed by generating a number k of independent trials of the event $\{\frac{1}{n}\sum_{i=1}^{n} Y_i > L\}$ and the estimate would be

$$\bar{\mu} = \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{\frac{1}{n}\sum_{i=1}^{n} Y_i^j > L\}},$$

where Y_i^j stands for the Y_i variable in the j-th trial of the event $\{\frac{1}{n}\sum_{i=1}^{n} Y_i > L\}$. The limit to the utility of this method is the long running time necessary to come up with a result *and* a specified precision. As seen in Chapter 1, for $L > m$ the probability in (2.2) goes to zero as n increases. This is equivalent to state that the probability of observing the event $\{\frac{1}{n}\sum_{i=1}^{n} Y_i > L\}$ in a simulation with k trials is decreasing with n. In particular, recalling the results of Large Deviation Theory, $0 < \mu \ll 1/n$.

Let E be the event in question, $E = \{\frac{1}{n}\sum_{i=1}^{n} Y_i > L\}$. For $n > k$ the possible situations are:

- the event E does occur. Then

$$\bar{\mu} = \frac{n_E}{k},$$

  where n_E is the number of times E appears through the k trials.

- the event E does not occur. In this case

$$\bar{\mu} = 0.$$

In both situations $\bar{\mu}$ is a wrong estimate of μ, since $0 \not< \bar{\mu} \not\ll 1/n$. The only possibility to reach an adequate result would be to perform the simulation as done at the beginning for the Bernoulli i.i.d. sequence, forcing k to be greater than n. And yet, if the searched probability is very low, the number k of independent trials needed to provide a reliable result would be very large, proportional to the inverse of the probability. As example, probabilities of order $10^{-30} \div 10^{-50}$, found in different scientific and engineering applications, would require k of order

$10^{30} \div 10^{50}$, resulting in a severe computation strain and eventually making the simulation task impossible to complete. The problem consists then in developing a method which decreases the number k to make simulation feasible. In doing that, the condition that must be fulfilled is that the precision level of the simulation must not be altered.

2.1.1 The importance sampling idea

To overcome the obstacle of the low frequency with which a rare event occurs through a simulation, the idea of the *importance sampling* method is to augment artificiously the probability of the event. As a consequence the event will occur with a higher frequency and a smaller number of trials will be needed to observe it in the simulation.

In order to introduce this technique, let $\mu = \mathbb{E}[f(Z)]$ be the parameter to evaluate, where f is a function and Z a random variable of distribution (or probability density) p taking values on the domain of f. The expression returns the expectation value of Z if f is the identity function, while it evaluates to a probability if f is the indicator function over a set E (in a given realization $\mathbb{1}_E(z) = 1$ if $z \in E$, 0 otherwise). Instead of estimating μ directly as $\bar{\mu} = \frac{1}{k}\sum_{j=i}^{k} f(Z_j)$, by generating a sequence (Z_n) of i.i.d. r.v.s all with the same distribution of Z, the importance sampling method employs a different sequence (X_n) of i.i.d. random variables distributed according to a new function b. The role of the new distribution is that of changing the probability with which the event determined by Z occurs, so as to increase its frequency. For this reason b is called **biasing distribution**.
The new estimate of μ is

$$\bar{\mu}_b = \frac{1}{k}\sum_{j=1}^{k} f(X_j)\frac{p(X_j)}{b(X_j)}. \tag{2.3}$$

The new estimator is often referred to as *importance sampling* estimate. It'll be denoted also as $\bar{\mu}_{\mathrm{IS}}$.

Although the variables X_i in (2.3) are governed by b, different from p, $\bar{\mu}_{\mathrm{IS}}$ is still an *unbiased* estimator for μ. Indeed, from the definition, the expectation value of $\bar{\mu}_{\mathrm{IS}}$ is

$$\begin{aligned}
\mathbb{E}[\bar{\mu}_b] &= \mathbb{E}\Big[\frac{1}{k}\sum_{j=1}^{k} f(X_j)\frac{p(X_j)}{b(X_j)}\Big] \\
&= \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}\Big[f(X_j)\frac{p(X_j)}{b(X_j)}\Big] \\
&= \frac{1}{k}\sum_{j=1}^{k}\int f(x)\frac{p(x)}{b(x)}b(x)dx \\
&= \int f(x)p(x)dx \\
&= \mathbb{E}[f(Z)] \\
&= \mu.
\end{aligned} \tag{2.4}$$

This means, by using LLN, that μ_b also will converge to μ as $k \to \infty$. At this point the introduction of the new estimator seems to complicate the calculus for μ, instead of simplifying it, because $\bar{\mu}_{\mathrm{IS}}$ requires to evaluate at each step the ratio $p(\cdot)/b(\cdot)$, and this could neutralize the advantage of using b. The introduction of $\bar{\mu}_{\mathrm{IS}}$ will be of advantage in that, with a convenient probability distribution b, the outputs $f(X_i)\frac{p(X_i)}{b(X_i)}$ come closer to the mean μ throughout the simulation, that is equivalent to a lower dispersion of the outputs about μ. This variance reduction will cause the sample average $\bar{\mu}_b$ to provide the result, within the required precision, with a smaller number of trials. As a consequence, the simulation performed with the new distribution will end in advance with respect to the original one, while still satisfying the precision requirements. In the remainder of the chapter the method to define the new probability distribution is shown.

Remark. There's a domain issue regarding the definition of b in (2.3). The ratio $p(\cdot)/b(\cdot)$, also called the *likelihood ratio*, diverges in points $\tilde{x}$ where $p(\tilde{x}) \neq 0$

and $b(\tilde{x}) = 0$. In the hypothesis these points occur in the simulation, the estimate wouldn't have any physical nor mathematical meaning, since it would be the result of applying a distribution for simulation purpose only. Anyway, this may cause the sum to diverge only if $\tilde{x}$ $f(\tilde{x}) \neq 0$ when $p(\tilde{x})/b(\tilde{x}) = 0$. Then, the support of the biasing distribution must satisfies the following requirement:

$$\text{support}(f \cdot p) \subset \text{support}(f \cdot b).$$

2.1.2 Optimal biasing distributions

The terms involved in the sum (2.3) are independent and identically distributed, thus the variance of $\bar{\mu}_b$ is k times the variance of $\frac{1}{k} f(X_1) \frac{p(X_1)}{b(X_1)}$:

$$\begin{aligned} Var[\bar{\mu}_b] &= k \frac{1}{k^2}\, Var\big[f(X_1) \frac{p(X_1)}{b(X_1)}\big], \\ k\, Var[\bar{\mu}_b] &= \mathbb{E}\big[\big(f(X_1)\frac{p(X_1)}{b(X_1)}\big)^2\big] - \mathbb{E}\big[f(X_1)\frac{p(X_1)}{b(X_1)}\big]^2 \\ &= \int f(x)^2 \frac{p(x)^2}{b(x)^2} b(x)\, dx - \mu^2 \end{aligned} \tag{2.5}$$

where the square of the expectation value in the second line equals μ^2 in the third line because

$$\mathbb{E}\big[f(X_1)\frac{p(X_1)}{b(X_1)}\big] = \int f(x) \frac{p(x)}{b(x)} b(x)\, dx = \mu.$$

Writing $V_b = \int f(x)^2 \frac{p(x)^2}{b(x)}\, dx$ the variance of $\bar{\mu}_b$ becomes

$$k\, Var[\bar{\mu}_b] = V_b - \mu^2. \tag{2.6}$$

Thus, to reduce $Var[\bar{\mu}_b]$, b should minimize V_b. As to V_b, Jensen's inequality gives

$$\begin{aligned} V_b = \mathbb{E}\big[\big(f(X)\frac{p(X)}{b(X)}\big)^2\big] &\geq \Big(\mathbb{E}\big[|f(X)|\frac{p(X)}{b(X)}\big]\Big)^2 \\ &= \Big(\int |f(x)| \frac{p(x)}{b(x)} b(x)\, dx\Big)^2 \\ &= \Big(\int |f(x)| p(x)\, dx\Big)^2. \end{aligned}$$

The above expression is satisfied as equality if and only if X is constant almost surely (a r.v. X is almost surely a constant c if the set on which $X \neq c$ has zero probability measure). This implies V_b is minimized when b makes $|f(x)|p(x)/b(x)$ constant, thus when

$$b(x) = b_{opt}(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx} \tag{2.7}$$

where the integral is to normalize b_{opt} to a probability distribution function and the index *opt* stands for *optimal*. The expression (2.7) for b_{opt} is of no practical utility: the denominator evaluates to μ, the parameter to estimate, that is unknown. Nevertheless, b_{opt} still remains the optimal choice, since it is the probability distribution that best reduces the estimator variance. For this reason it can be a guide in finding a distribution function to put into practical use.

Let f be the indicator function over some set E representing a rare event, $f = \mathbb{1}_E$. First, from (2.7) it follows that the support of b_{opt} is entirely on E. Thus b_{opt} shifts all the probability on the rare event E, with the consequence of increasing the frequency of E throughout the trials. In the second place the behaviour of b_{opt} over the rare event coincides with the behaviour of p, since $b_{opt}(\cdot)|_E = \frac{1}{\mu}p(\cdot)|_E$. Thus b_{opt} maintains the probability structure of p over E. Following these characteristics of b_{opt}, the biasing probability distribution to be used should satisfy three properties:

- b increases the probability of observing the rare event.
- b minimizes the estimator variance.
- b preserves the original probability structure over the rare event.

Remark. The above properties have been asserted in a general form. Indeed, an event has no probability structure. It is the random variable determining the event to be governed by a probability distribution. The probability structure over an event is meant to be the shape of the probability distribution of the variables by which the event is determined.

The search of a biasing probability distribution consists in defining a new function complying with some requirements, among which the properties stated above. There's no assurance that the properties will all be satisfied. Hence, a priority must be defined between the three. Those that cannot be discarded are the first two: the frequency of occurrence of the rare event E must be increased, otherwise the simulation would take even more trials to observe a sufficient number of realizations of E, and the variance of the estimator must be the minimum possible, otherwise the simulation would take more trials to provide a result matching the required precision. As to the third property, there's no advantage in trying to preserve the probability structure over E if E is the rare event in the process, because what has to be observed through the simulation is E itself and not its inside structure. If, on the other hand, a particular collection F_i of subsets of E is of interest, then the procedure to find the biasing distribution could be carried out for each of them, treating each one as a rare event (that will be rarer than E, since $F_i \subset E$).

2.1.3 The simulation-stop criterion

In the previous sections it has been remarked that by adopting an importance sampling technique it is possible to decrease the number of simulation trials while still achieving the required precision. However, there's no standard mechanism which fixes, before the simulation begins, the right number of trials k needed to achieve the fixed precision. Nevertheless, without such a mechanism the simulation would go on indefinitely, because there wouldn't be any k to stop it. This problem is solved by adopting a criterion which uses the results of the simulation itself to define k, then taking the simulation to an end.

Let $\bar{\mu}_b$ the importance sampling estimator of $\mu = \mathbb{E}[f(X)]$, with X a r.v. and f a real function. To simplify notation let $g(x) = f(x)\frac{p(x)}{b(x)}$, so $\bar{\mu}_b = \frac{1}{k}\sum_{i=1}^{k} g(X_i)$. The precision requirement on $\bar{\mu}_b$ is expressed by the condition

$$\mathbb{P}\big(|\bar{\mu}_b - \mu| \leq \epsilon\mu\big) = y_\epsilon, \tag{2.8}$$

where ϵ is the maximum error accepted on μ, and y_ϵ the fixed confidence. The left-hand side of (2.8) can be rewritten with the help of the central limit theorem, as done in 2.1,

$$\begin{aligned}\mathbb{P}\big(|\bar{\mu}_b - \mu| \leq \epsilon\mu\big) &= \mathbb{P}\Big(\Big|\frac{\frac{1}{k}\sum_{j=1}^{k}(g(X_j) - \mu)}{\sqrt{Var[g(X)]}}\sqrt{k}\Big| \leq \frac{\epsilon\mu\sqrt{k}}{\sqrt{Var[g(X)]}}\Big) \\ &\simeq \mathbb{P}\Big(|Y| \leq \frac{\epsilon\mu\sqrt{k}}{\sqrt{Var[g(X)]}}\Big),\end{aligned}$$

where Y is a standard Gaussian random variable. Now, expresion (2.5) gives $Var[g(X)] = k\,Var[\bar{\mu}_b]$ and in the notation of eq. (2.6) $k\,Var[\bar{\mu}_b] = V_b - \mu^2$, then the condition (2.8) becomes

$$\mathbb{P}\Big(|Y| \leq \frac{\epsilon\mu\sqrt{k}}{\sqrt{V_b - \mu^2}}\Big) = y_\epsilon. \tag{2.9}$$

Now, for any $y \in (0,1)$ there's a unique number $t > 0$ such that $\mathbb{P}(|Y| \leq t) = y$. The quantity

$$t(y_\epsilon) = \frac{\epsilon\mu\sqrt{k}}{\sqrt{V_b - \mu^2}}$$

is then fixed by y_ϵ, and not k, according to the distribution of Y, in this case the normal distribution: $t(y_\epsilon)$ is the unique number satisfying $\mathbb{P}(|Y| \leq t(y_\epsilon)) = y_\epsilon$. Hence, condition (2.8), equivalent to

$$\mathbb{P}\big(|Y| \leq t(y_\epsilon)\big) = y_\epsilon,$$

is met when the number k of trials satisfies

$$t(y_\epsilon) = \frac{\epsilon\mu\sqrt{k}}{\sqrt{V_b - \mu^2}},$$

or, solving with respect to k, when

$$k = \Big(\frac{t(y_\epsilon)}{\epsilon}\Big)^2 \Big(\frac{V_b}{\mu^2} - 1\Big). \qquad (2.10)$$

As for the optimal biasing distribution, this expression for k cannot be used: V_b and μ are unkown quantities and μ is what the simulation is done for! (This is the reason why a standard mechanism to find k does not exist.) To fix a solution, k could be obtained by replacing μ with $\bar{\mu}_b$ and V_b with its importance sampling estimator $\bar{V}_b$:

$$\bar{V}_b = \frac{1}{k}\sum_{j=1}^{k}\Big(f(X_j)\frac{p(X_j)}{b(X_j)}\Big)^2.$$

By doing this the number k changes in

$$\tilde{k}(k) = \Big(\frac{t(y_\epsilon)}{\epsilon}\Big)^2 \Big(\frac{\bar{V}_b}{\bar{\mu}_b^2} - 1\Big). \qquad (2.11)$$

Estimators $\bar{\mu}_b$ and $\bar{V}_b$ depend on k, so $\tilde{k}$ is still a function of k: it represents the estimate of k that complies with (2.10) and that is obtained with the best information available, since it is estimated from the outcomes of the simulation itself up to the last k-th trial. Because every $\hat{k} > k$ of (2.10) would be acceptable, having replaced k by $\tilde{k}$ the criterion to stop the simulation becomes:

$$\textit{stop the simulation at} \quad k^* = \min\{k : k \geq \tilde{k}(k)\}. \qquad (2.12)$$

2.2 Importance sampling and Large Deviation Theory

In section 2.1.2 two properties have been identified a biasing distribution must have: it must increase the probability of the rare event and it must minimize the estimator variance. How to achieve this? Large Deviation Theory will provide the answer.

The setting will be:

$(X_{p,n})_{n\in\mathbb{N}}$ is a sequence of random variables, X_n taking values in a space $\mathcal{S}_n$, with distribution p_n.

$(f_n)_{n\in\mathbb{N}}$ is a sequence of measurable functions, $f_n\colon \mathcal{S}_n \to \mathbb{R}^d$.

the parameter of interest is of the form $\mu_n = \mathbb{P}(f_n(X_{p,n})/n \in E)$ with $E \subset \mathbb{R}^d$.

$(X_{b,n})_{n\in\mathbb{N}}$ is the sequence adopted in the importance sampling estimator $\bar{\mu}_{\text{IS}}$, with $X_{b,n}$ of biasing distribution b_n (support$(p_n) \subset$ support(b_n), $\forall n$).

The following assumption is taken on (μ_n):
the sequence (μ_n) satisfies a large deviation principle.
As seen in the previous section, the estimator $\bar{\mu}_{\text{IS},n}$ for the element μ_n is:

$$\bar{\mu}_{\text{IS},n} = \bar{\mu}_n = \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{\frac{f_n(X^j_{b,n})}{n}\in E\}} \frac{dp_n}{db_n}\big(X^j_{b,n}\big), \tag{2.13}$$

where the likelyhood ratio is in the form of a derivative (Radon-Nikodym derivative) when p_n and b_n are continuous distributions (probability density functions). From now on only the importance sampling estimator will be considered, then the symbol $\bar{\mu}_n$ will be used in place of $\bar{\mu}_{\text{IS},n}$ to simplify notation. At this point, it is straighforward to choose b_n such that the probability of the support of the indicator function in (2.13) is increased, by changing conveniently the parameters of p_n. As to decreasing the variance of the estimator, the method is not so evident.

As shown in 2.1.2, $\bar{\mu}_n$ is a sum of k i.i.d. elements, then, for n is fixed, its variance is constant, or $k\,Var[\bar{\mu}_n]$ is constant. Which is, instead, the behaviour of $Var[\bar{\mu}_n]$ as n change? A theorem shows the variance follows a large deviation principle. To demonstrate this, some definitions are in order.
It is first defined the sequence of functions

$$\alpha_n(\theta) = \frac{1}{n}\log\int e^{\langle\theta,\, f_n(x)\rangle}\frac{dp_n}{db_n}(x)dp_n(x), \quad \theta\in\mathbb{R}^d,$$

Each α_n is convex, as ϕ_n in the setting of the Gärtner-Ellis theorem. Then it is introduced the sequence of probability measures (ν_n), with ν_n defined on $\mathcal{S}_n$,

$$\nu_n(A) = e^{-nc_n(0)} \int_A \frac{dp_n}{db_n}(x) dp_n(x),$$

for $A \subset \mathcal{S}_n$. To (ν_n) it is associated the sequence of r.v.s (Z_n), where Z_n is $\mathcal{S}_n$-valued with probability distribution ν_n on $\mathcal{S}_n$. For every n, f_n can be applied to Z_n generating a sequence on $\mathbb{R}^d$.

It is now defined the sequence of log-moment generating functions, for $\theta \in \mathbb{R}^d$

$$\begin{aligned}\beta_n(\theta) &= \frac{1}{n} \log \mathbb{E}\big[\exp\langle\theta,\, f_n(Y_n)\rangle\big] \\ &= \frac{1}{n} \log \int e^{-n\alpha_n(0)} e^{\langle\theta,\, f_n(x)\rangle} \frac{dp_n}{db_n}(x) dp_n(x) \\ &= \alpha_n(\theta) - \alpha_n(0).\end{aligned}$$

Now, if the sequence (α_n) satisfies the standard assumptions A1, A2, A3 stated in 1..., these conditions are also satisfied by the sequence (β_n) and the Gärtner-Ellis theorem thus holds for the sequence $(f_n(Z_n))$ [3, 20].

At this point it is introduced the **variance rate function** R_V:

$$R_V(x) = \sup_{\theta \in \mathbb{R}^d} \big[\langle\theta,\, x\rangle - \alpha(\theta)\big], \quad x \in \mathbb{R}^d, \tag{2.14}$$

where $\alpha(\theta) = \lim_{n\to\infty} \alpha_n(\theta)$. The component of $Var[\bar{\mu}_n]$ that varies with n will be indicated as in (2.6):

$$V_n = \int \mathbb{1}_{\{\frac{f_n(x)}{n} \in E\}} \Big(\frac{dp_n}{db_n}(x)\Big)^2 db_n(x).$$

In what follows, for a set $E \subset \mathbb{R}^d$, $\overset{\circ}{E}$ denotes the interior of E, $\bar{E}$ the closure of E, ∂E the boundary of $\bar{E} \setminus \overset{\circ}{E}$.

Theorem 2.1. *Let $E \subset \mathbb{R}^d$ be a set with $\overset{\circ}{E} \neq \emptyset$, $\bar{E} = \bar{\overset{\circ}{E}}$, over which $0 < \inf_{x\in E} R_V(x) < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log V_n = -\inf_{x\in E} R_V(x).$$

Proof. Thanks to the definition of ν_n,

$$\begin{aligned} V_n &= \int \mathbb{1}_{\{\frac{f_n(x)}{n} \in E\}} \Big(\frac{dp_n}{db_n}(x) \Big)^2 db_n(x) \\ &= e^{n\alpha_n(0)} \int_{\frac{f_n(x)}{n} \in E} d\nu_n(x) \end{aligned}$$

then

$$\frac{1}{n} \log V_n = \alpha_n(0) + \frac{1}{n} \log \int_{\frac{f_n(x)}{n} \in E} d\nu_n(x)$$

and since a large deviation principle holds for the sequence $(f_n(Z_n))$, it follows

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log V_n &= \alpha(0) - \inf_{x\in E} \sup_{\theta\in\mathbb{R}^d} \big[\langle \theta,\, x \rangle - \alpha(\theta) - \alpha(0) \big] \\ &= - \inf_{x\in E} \sup_{\theta\in\mathbb{R}^d} \big[\langle \theta,\, x \rangle - \alpha(\theta) \big] \\ &= - \inf_{x\in E} R_V(x). \end{aligned}$$

where $\lim_n$ is used in place of $\liminf_n$ and $\limsup_n$ because of the conditions put on the set E, that make the upper and lower bounds in Gärtner-Ellis theorem coincide. □

In ultimate analysis theorem 2.1 states that if a large deviation principle holds for the sequence (μ_n) defined by $(f_n(X_n))$, a large deviation principle holds for the sequence (V_n).

2.2.1 An efficiency criterion for simulations

From (2.6), $Var[\mu_n]$ can be written as

$$k\, Var[\bar{\mu}_n] = V_n - \mu_n^2. \tag{2.15}$$

Since for every random variable the variance is greater than or equal to zero, expression (2.15) implies $V_n - \mu_n^2 > 0$ *for all* n. Thus, V_n converges to zero in n with a rate that can only be smaller than the rate with which μ_n^2 approaches zero.

So, if

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \mu_n &= -I \qquad \text{and} \\ \lim_{n\to\infty} \frac{1}{n} \log V_n &= -R, \end{aligned}$$

it must be

$$R \leq 2I. \tag{2.16}$$

At this point, which is the optimal choice for R to realize an efficient importance sampling estimator, that is, to perform the simulation with the smallest number of trials? Recalling eq. (2.10), to comply with a precision (ϵ, y_ϵ) the simulation should run for a number k of trials set by

$$k = \Big(\frac{t(y_\epsilon)}{\epsilon}\Big)^2 \Big(\frac{V_n}{\mu_n^2} - 1\Big)$$

(where $t(y_\epsilon)$ realizes $\mathbb{P}(|Y| \leq t(y_\epsilon)) = y_\epsilon$ for $Y \sim \mathcal{N}(0,1)$). In the above expression V_n and μ_n^2 converge to zero, but with different rates, respectively R and I with $R \leq 2I$. If the inequality (2.16) is met strictly k diverges exponentially in n, because the ratio V_n/μ_n^2, that goes as the exponential of $n(2I - R)$ for n large, diverges. For $R = 2I$ the ratio converges instead to 1.

The optimal choice is then $R = 2I$.

Since R is the rate function for V_n and V_n depends on b_n, the optimal choice for R is a condition on (b_n). This leads to the following *efficiency criterion*:

a sequence (b_n) of biasing simulation distributions is efficient *if it realizes the condition $R = 2I$.*

The need of adopting a criterion on R, instead directly on k, comes because it is not possible to obtain any significant result acting with respect to k: for any sequence (b_n), $\mathrm{Var}[\bar{\mu}_n]$ simply decreases with k and (b_n) does not depend on k, then there's no possibility to force (b_n) to decrease the variance further. On the other side, (b_n) can be chosen so it decreases the variance of the estimator in n.

Indeed, with respect to n, the sequence (b_n) could determine both a decrease or an increase of the variance of the estimator $\bar{\mu}_n$ if n is large.

From a more general point of view, the concept at the basis of the efficiency criterion is that the variance reduction for a particular estimator $\bar{\mu}_m$ of interest can be achieved by first identifying the sequence (b_n) that realizes the best variance reduction for the *sequence* of estimators $(\bar{\mu}_n)$. Then, the estimator $\bar{\mu}_m$ is obtained by selecting, from the sequence (b_n), the particular b_m.
In doing so, the problem of finding the optimal biasing distribution is reduced to the problem of maximizing the variance rate function R, instead of minimizing directly the variance for the particular estimator $\bar{\mu}_m$ of interest. This is a simplification, in that the task of maximizing R in most cases turns out to be simpler than finding the b_n that minimize the variance of $\bar{\mu}_m$. This happens because the latter method is a functional minimization problem, generally more complicated than the former.

2.2.2 The theory behind the technique

In the following the assumptions are:

- the log-moment generating functions β_n associated to the $\mathbb{R}^d$-valued random variables $f_n(X_{p,n})$ satisfy the standard assumptions A1, A2, A3.

- the set $E \subset \mathbb{R}^d$ satisfies: $\overset{\circ}{E} \neq \emptyset$, $\bar{E} = \bar{\overset{\circ}{E}}$, $0 < I(E) < \infty$, where $I(E) = \inf_{x \in E} I(x)$.

An element $t \in E$ is called a *minimum rate point* of E if $I(t) = I(E)$.
A point $t \in E$ is called a *dominating point* of E if it is the unique element satisfying:

- $t \in \partial E$,

- $\exists!\, \theta_t \in \mathbb{R}^d$: $\nabla \beta(\theta_t) = t$,

- $E \subset \mathcal{H}(t) = \{x \colon \langle \theta_t,\, x - t\rangle \geq 0\}$.

In the hypothesis E has a dominating point t, it holds

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n = -I(t) = -\sup_{\theta}[\langle \theta,\, x\rangle - \beta(\theta)] = -[\langle \theta,\, t\rangle - \beta(\theta)],$$

where $\lim_n$ has been used because from the assumptions $\lim_n(\cdot) = \liminf_n(\cdot)$. Thus μ_n^2 has rate $2I(t)$. As to V_n, it must be

$$\liminf_{n\to\infty} \frac{1}{n} V_n \geq -2I(t).$$

The inequality is met as equality with the choice $db_n = (\mathbb{1}_{\{\frac{f_n(X_{p,n})}{n} \in E\}})/\mu_n dp_n$, but, as seen in section 2.1, this is of no utility.

Let now b_n be exponential shifts of p_n:

$$db_n(x) = \frac{\exp \langle \psi,\, f_n(x)\rangle}{\int \exp \langle \psi,\, f_n(y)\rangle \, dp_n(y)} dp_n(x) = \frac{\exp \langle \psi,\, f_n(x)\rangle}{\exp\big(n\beta_n(\psi)\big)} dp_n(x), \tag{2.17}$$

with $\psi \in \mathbb{R}^d$. For (b_n) the associated (α_n) is

$$\begin{aligned}\alpha_n(\theta) &= \frac{1}{n} \log \int \exp \langle \theta,\, f_n(x)\rangle \frac{dp_n}{db_n}(x)\, dp_n(x)\\ &= \beta_n(\theta - \psi) + \beta_n(\psi)\end{aligned}$$

then

$$\begin{aligned}\alpha(\theta) &= \lim_{n\to\infty} \big(\beta_n(\theta - \psi) + \beta_n(\psi)\big)\\ &= \beta(\theta - \psi) + \beta(\psi),\end{aligned}$$

Thus, for $[-\psi,\, \psi] \subset \overset{\circ}{E}$, assumptions A1, A2, A3 hold for α also.

Theorem 2.1 says the rate function for V_n is

$$\begin{aligned}R_V(x) &= \sup_{\theta}[\langle \theta,\, x\rangle - \alpha(\theta)]\\ &= \sup_{\theta}[\langle \theta,\, x\rangle - \beta(\theta - \psi) - \beta(\psi)]\\ &= \sup_{\theta}[\langle \theta,\, x\rangle - \beta(\theta - \psi)] + [\langle \psi,\, x\rangle - \beta(\psi)]\\ &= I(x) + [\langle \psi,\, x\rangle - \beta(\psi)].\end{aligned}$$

If it is selected $\psi = \theta_t$, $R_V(x)$ becomes

$$R_V(x) = I(x) + [\langle \theta_t,\, x\rangle - \beta(\theta_t)],$$

that gives, in t, $R_V(t) = 2I(t)$. Here t is the dominating point of E, so that $\forall x \in E$, $\langle \theta_t,\, x - t\rangle \geq 0$ and $I(x) \geq I(t)$. Hence

$$R_V(X) - R_V(t) = I(x) - I(t) + \langle \theta_t,\, x - t\rangle \geq 0,$$

from which it follows

$$R_V(E) = \inf_{x\in E} R_V(x) = R_V(t) = 2I(t) = 2I(E).$$

The last expression shows the sequence (b_n) defined in (2.17) is efficient (and this can be employed in the simulation). This proves the following theorem.

Theorem 2.2. *If the set E has a dominating point η, the sequence (b_n) defined as*

$$db_n(x) = \exp[-n\beta_n(\theta_t)]\exp[\langle \theta_t,\, f_n(x)\rangle]db_n(x)$$

is efficient.

The problem of reducing the estimator variance to design an efficient simulation to estimate $\mu_n = \mathbb{P}(f_n(X_{p,n})/n \in E)$ is now solved: b_n is the probability distribution to be used. As example, to estimate

$$\mathbb{P}\big(f_{100}(X_{p,100})/100 \in E\big)$$

the importance sampling estimator will be

$$\bar{\mu}_{100} = \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{\frac{f_{100}(X^j_{b,100})}{100}\in E\}} \frac{dp_{100}}{db_{100}}\big(X^j_{b,100}\big),$$

where each $X^j_{b,100}$ in the j-th trial is generated with distribution b_{100}.

2.2.3 Importance sampling technique for Markov chains

The sequence of efficient biasing distribution has been obtained through the Gärtner-Ellis theorem, which does not require the sequence (X_n) to be i.i.d.. Indeed, the i.i.d. condition has never been assumed, only the standard assumptions for the sequence (β_n) have played a role in getting the result. As a consequence it is possible to derive an efficient sequence of biasing distributions for Markov chains also.

Let $(\tilde{A}_i)$ a sequence of random variables representing samples from a Markov chain $\tilde{A}_n$.

The hypothesis are

- $\tilde{A}_n$ has a finite state space $\chi = \{0, 1, \dots, k\}$, $k \in \mathbb{N}$ fixed.

- $\tilde{A}_n$ is homogeneous, irreducible and aperiodic.

- f is a fixed function, $f \colon \chi \to \mathbb{R}$.

The transition probabilities are denoted by $p_{xy} = \mathbb{P}(\tilde{A}_{n+1} = y | \tilde{A}_n = x)$. The quantity of interest is

$$\mu_n = \mathbb{P}\Big(\sum_{i=1}^{n} f(\tilde{A}_i) > nt\Big),$$

for which the importance sampling estimator is

$$\bar{\mu}_n = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{\sum_{i=1}^{n} f(A_i^j) > nt\}} \frac{p(A_1^j, \dots, A_n^j)}{b(A_1^j, \dots, A_n^j)},$$

where $p(x_1, \dots, x_n)$ and $b(x_1, \dots, x_n)$ are respectively the original and the biasing joint distributions and A_n is the chain generated according to the transition probabilities b_{ij}. With respect to the notations of the previous section, $\mathcal{S}_n = \chi^n$, $X_{p,n} = (\tilde{A}_1, \dots, \tilde{A}_n)$, $X_{b,n} = (A_1, \dots, A_n)$, $d = 1$ and $f_n(x_1, \dots, x_n) = \sum_{i=1}^{n} f(x_i)$.

Let $\mathcal{F}$ be now the set of all functions $g\colon \chi \to \mathbb{R}$. Following a method parallel to that adopted for Markov chains in Chapter 1, let $W_\theta\colon \mathcal{F} \to \mathcal{F}$ be the operator

$$W_{b,\theta}(g)(x) = \sum_{y\in\chi} e^{\theta f(y)} g(y) \frac{p_{xy}^2}{b_{xy}}.$$

As T_θ in (1.4), $W_{b,\theta}$ is positive, thus, by the Perron-Froebenius theorem, it has a unique largest positive eigenvalue $\nu_b(\theta)$. The element $\alpha_n(\theta)$ is then

$$\begin{aligned}\alpha_n(\theta) &= \frac{1}{n}\log\Big[\sum_{y_1\in\chi}\cdots\sum_{y_n\in\chi}\frac{p(y_1,\ldots,y_n)^2}{b(y_1,\ldots,y_n)}\exp\big(\theta\sum_{i=1}^n f(y_i)\big)\Big]\\ &= \frac{1}{n}\log\Big[\sum_{y_1,\ldots,y_n\in\chi}\frac{p(y_1)^2\prod_{i=1}^{n-1}p_{y_i,y_{i+1}}^2}{b(y_1)\prod_{i=1}^{n-1}b_{y_i,y_{i+1}}}\exp\big(\theta\sum_{i=1}^n f(y_i)\big)\Big]\\ &= \frac{1}{n}\log\Big[\sum_{y_1\in\chi}W_{b,\theta}^n(1)(y_1)\frac{p(y_1)^2}{b(y_1)}\exp\big(\theta f(y_1)\big)\Big]\end{aligned}$$

and by the same argument used with T_θ, taking the limit for $n \to \infty$ gives

$$\alpha_b(\theta) = \lim_{n\to\infty}\alpha_n(\theta) = \log\nu_b(\theta).$$

For the component V_n of the estimator variance ($k\,Var[\bar{\mu}_n] = V_n - \mu_n^2$), theorem 2.1 gives

$$\lim_{n\to\infty}\frac{1}{n}\log V_n = -\inf_{x\in(t,\infty)} RV(x) = R_V(t) = -\sup_\theta\big[\theta t - \log\nu_b(\theta)\big].$$

The rate function for μ_n is instead

$$\begin{aligned}\lim_{n\to\infty}\log\mu_n = -I(t) &= \sup_\theta\big[\theta t - \log\lambda(\theta)\big]\\ &= \theta_t t - \log\lambda(\theta_t)\big],\end{aligned}$$

where $\lambda(\theta)$ is the largest eigenvalue of T_θ and θ_t is the root of the equation
$t = \lambda'(\theta)/\lambda(\theta)$, seen in section 1.4.

Following what has been done in the previous section, the sequence of simulation distributions is chosen among the family of exponential shifts of p_{ij}:

$$b_{xy} = p_{xy}\exp\big(\theta_t f(y)\big)\frac{\psi_{\theta_t}(y)}{\lambda(\theta_t)\psi_{\theta_t}(x)}, \tag{2.18}$$

with ψ_θ the eigenvector of T_θ relative to $\lambda(\theta)$ and $\psi_{\theta_t}(x)$ its component corresponding to the state x. For b_{xy} as in (2.18) the operator $W_{b,\theta}$ becomes

$$W_{b,\theta}(g)(x) = \lambda(\theta_t)\psi_{\theta_t}(x) \sum_{y \in \chi} \exp\big((\theta - \theta_t) f(y)\big) g(y) \frac{p_{xy}}{\psi_{\theta_t}(y)}.$$

The largest eigenvalue of $W_{b,\theta}$ is $\nu_b(\theta) = \lambda(\theta_t)\lambda(\theta - \theta_t)$ and its corresponding eigenvector is $\xi_\theta(y) = \psi_{\theta_t}(y)\psi_{\theta-\theta_t}(y)$. Then, with b as in (2.18), the variance rate function is

$$\begin{aligned} R_V(t) &= \sup_\theta \big[\theta t - \log \nu_b(\theta)\big] \\ &= \sup_\theta \big[\theta t - \log \lambda(\theta_t) - \log \lambda(\theta - \theta_t)\big] \\ &= \sup_\theta \big[\theta_t t - \log \lambda_b(\theta_t) + (\theta - \theta_t)t - \log \lambda(\theta - \theta_t)\big] \qquad (2.19) \\ &= \sup_\theta \big[I(t) + (\theta - \theta_t)t - \log \lambda(\theta - \theta_t)\big] \\ &= 2I(t) \end{aligned}$$

and this shows the choice (2.18) is efficient.

Chapter 3

Taylor's Law from Multiplicative Models: the Role of Rare Events

The state of a species belonging to a given ecosystem can be described by adopting different variables and parameters. One of these, useful in ecology to keep control of the state of a species within an ecosystem, is the degree of aggregation, or simply aggregation, defined as the tendency of individuals of the species to aggregate in groups instead of keeping random distance or constant distance between them [1, 32]. According to this qualitative definition the condition of individuals of keeping random distance among them corresponds to a low degree of aggregation, while the condition of constant distance corresponds to a regular pattern of individuals and then to zero degree of aggregation. Taylor's law states that variance and mean of the index representing the aggregation of a species are governed by a power law [10, 32, 36].

In this chapter the problem of evaluating the power exponent is analyzed. It will be demonstrated that in order to correctly estimate the exponent a high-efficiency simulation method is needed.

3.1 Origins of Taylor's law

The variable adopted in ecology to represent the aggregation of a species is the population density d of the species:

$$d = N/S,$$

where N is the number of individuals per unit of area S.

Individuals of a species are subjected to interactions, among themselves or between them and other species of the system, thus d is a dynamic variable. Moreover the number of interactions and their nature make difficult to predict the time evolution of d in a deterministic way, giving d the character of random variable. However at this point it is not possible to define for d the expectation value $\mathbb{E}[d]$ and the variance $Var[d]$ as for a random variable in probability theory, because no probability distribution has been assigned to it. Indeed from measurements of d a model and possibly a probability distribution are sought after to predict the dynamics of the system. Being d a random variable, statistics suggests to take as estimators of $\mathbb{E}[d]$ and $Var[d]$ respectively

$$m = m_k = \frac{1}{k}\sum_{j=1}^{k} d_j$$

and

$$\sigma^2 = \sigma_k^2 = \frac{1}{k}\sum_{j=1}^{k}(d_k - m)^2,$$

where k is the number of independent measurements and d_j denotes the value of d in the j-th independent measurement. The subscript k indicates that m and σ^2 depend on the number of measurements. For random variables with finite expectation value LLN states

$$\lim_{k\to\infty} m_k = \mathbb{E}[d].$$

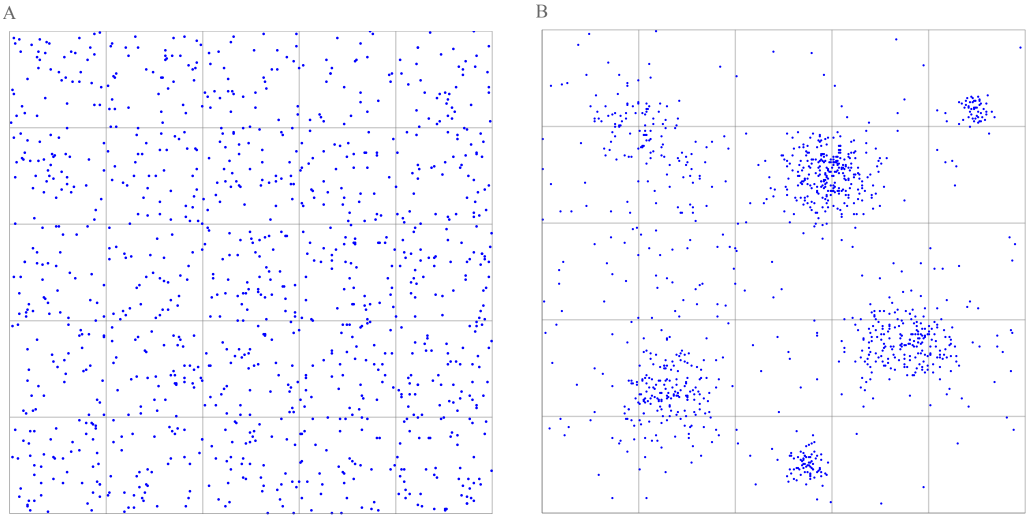


Figure 3.1: The dots represent the individuals of a population over a sampled area. The grid shows the sample units. In A the individuals distribute in a random way, thus the difference among the numbers of individuals per unit area is small and the population number variance is small. In B aggregation is evident, corresponding to a higher variance. A zero variance would instead correspond to a regular distribution pattern over the grid.

Population densities from real ecological systems are always finite, then m_k is a good estimator, at least asymptotically, for $\mathbb{E}[d]$. As to σ^2, it evaluates the fluctuations of the values of d around its mean m: for a set of measurements a small value of σ^2, with respect to m, indicates d_j are close to m for each j, while a high value of σ^2 indicates d_j are distant from the mean m. A small value of σ^2 then corresponds to a low degree of aggregation, a large value to a high degree of aggregation [32]. Figure 3.1 displays two examples of aggregation.

The population structure, or state, of a species could depend on its size, but a species index as, for example, the index of aggregation, should be independent of it, so that it could be used to identify the species or its population structure independently of the size.

Analysing 24 papers (dated 1936 to 1960) reporting population densities of different species ranging from worm larvae to shellfish on seashore [32], the English entomologist L. R. Taylor found their variance σ^2 could be related to their mean m by a power law:

$$\sigma^2 = am^b.$$

Taylor interpreted the a parameter as a computing factor, or a parameter depending on the size of the sampling unit, thus having no physical meaning, and he stated the b parameter was the index of aggregation of the species. This interpretation is supported by the scale invariant property of b: if the sampling unit is multiplied by a constant factor c and under this operation d changes in cd, then the mean m changes in

$$m_c = \mathbb{E}[d_c] = \mathbb{E}[cd] = c\mathbb{E}[d] = cm,$$

where c comes out of the expectation value due to the linearity of $\mathbb{E}[\cdot]$. The variance σ^2 changes instead in

$$\sigma_c^2 = \mathbb{E}[(d_c - m_c)^2] = \mathbb{E}[c^2(d-m)^2] = c^2\mathbb{E}[(d-m)^2] = c^2\sigma^2.$$

Thus, if

$$\sigma^2 = am^b$$

it follows

$$\sigma_c^2 = c^2\sigma^2 = c^2am^b = a_c m_c^b$$

with $a_c = c^{2-b}a$. This shows b can be adopted as an index of aggregation because it is independent on the mean of the population density and it can vary only if some change in the population structure occours. An example is given in figure 3.2.

Since it appeared in the 1961 article by Taylor, it became evident this law could have a universality character, because it reasonably fitted data regarding a variety of living species [10, 36]. Other data were obtained from an increasing number of

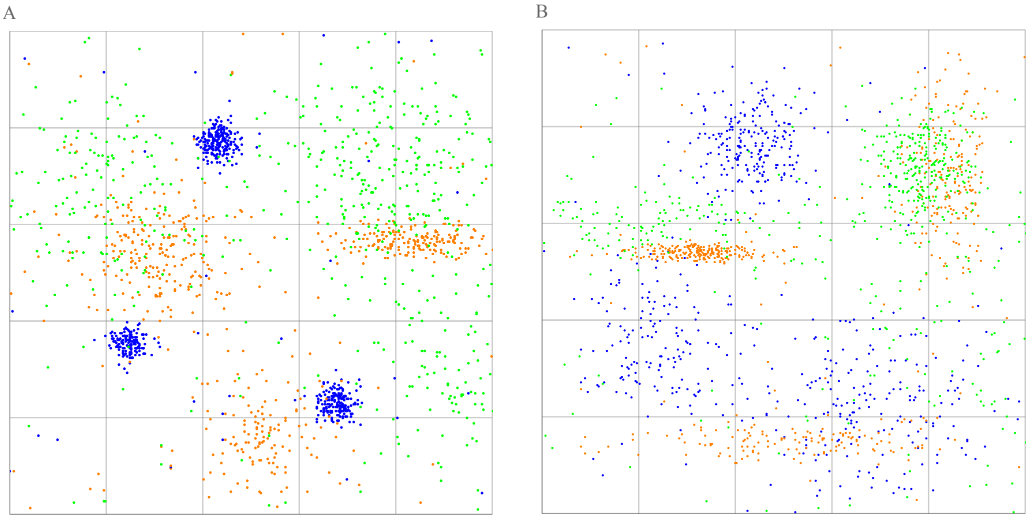


Figure 3.2: Interactions between individuals belonging to different species, above in different colors, give rise to a variety of distribution patterns corresponding generally to different degrees of aggregation. Here, A and B photograph two states of the same ecosystem. If changes occour in the populations, the distribution patterns are very likely to change also, causing the aggregation indices to vary.

ecological systems and the law seemed to apply well in describing aggregation of species very different between them and living in areas with no common characteristics [1]. In addition, other data were brought showing some ecological systems followed Taylor's law with respect to time, that is, it seemed it was possible to relate variance and mean of population densities by a power law when they were computed over time but on the same site [10, 26]. This was an enhancement of the validity of the law, suggesting the existence of a universal mechanism governing ecological systems both in time and space. An even more interesting property of the law was its appearance in contexts other than the ecological one, including physics, life science, finance [10]. It is reported in [24] that even in human genome it is possible to find a clustering phenomenon well described by a power law for variance and mean. In [24] the number of genes per unit of physical length in a human chromosome was measured and it was shown its variance and mean followed a power law relationship. Taylor's power law is also found in epidemiology where it seems to well describe diseases diffusion [23]. Other examples refer of clustering or aggregation phenomenons in economics, where Taylor's law models relevant financial fluctuations [10, 15].

3.2 Models and issues

In order to explain Taylor's power law (TL) models were proposed which were specific for the ecosystems at first studied, but as the almost universal character of the law became clear more general mechanisms were put at trial and methods from statistical mechanics to statistics and probability theory were adopted [5, 7, 8, 10, 28].

Two questions constitute the main issues about the law:

- Which is the law's origin?
- Which is the possible range of values of the b exponent?

Which is the law's origin? As to the first question there is at present no agreement among researchers. This is due to the ubiquitous character of the law that may be explained by two ways:

- the systems for which the law holds share similar dynamical features, or a common physical mechanism [10, 33],

- the power law - like behaviour of variance against mean of a system observable is due entirely to a probability distribution not related to any physical common mechanism, or having no physical meaning [25, 35].

Which is the possible range of values of the b exponent? In the attempt to understand the law the behaviour of the b exponent plays a crucial role. In ecological

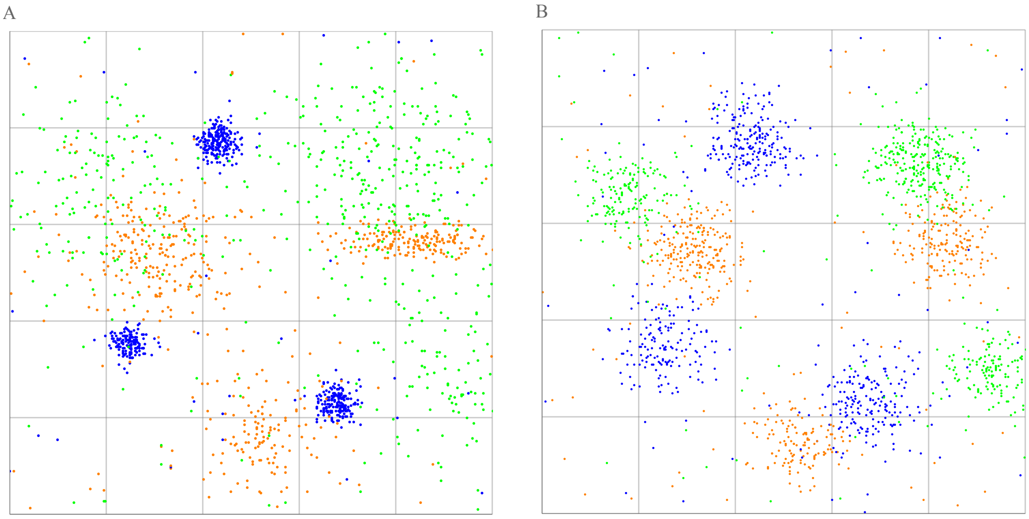


Figure 3.3: In A what is expected for an ecological system is shown: individuals of diverse species, marked in varied colors, should follow different distribution patterns, characterized by different degrees of aggregation, besides other parameters. In B is pictured what is instead observed: individuals belonging to different species seem to distribute according to the same pattern, irrespective of the species. This surprising feature is also found outside ecology [10].

settings it should be representative of species. Although some of the species of a given ecosystem could share the same value for b, there's no evident reason for which b should be the same for a variety of species within a particular and among different ecosystems. This is predicted by theoretical studies showing that population dynamics can be reasonably described by models known as population growth models and that Taylor's law appear with possibly any real value of b [5–8, 21]. Further studies about multiplicative growth models, a subset of the population growth models, have shown b can vary abruptly when even small changes in the interactions occur [5, 6, 21].

In contrast to these theoretical predictions empirical studies show b is bounded and takes values within the interval $[1, 2]$ and more frequently $b \simeq 2$ [1, 10]. Figure 3.3 offers a simplified representation of the problem.

As to the estimate of b, its value should be obtained from the probability distribution governing the variable. In the case of ecological systems, that are complex systems, to determine the real probability distribution even of only one variable is almost impossible. This is why theoretical models are used: the value of b is determined by the probability distribution used in the model that is assumed to best reproduce the dynamics of the variable. Once such a theoretical model is developed a test is needed to validate it. The issue arises at this point: while theoretical models (for a given variable) predict for b values in a wide range, tests report bounded values and mostly $b \simeq 2$. This is a crucial and delicate issue: if consensus on Taylor's law origin was achieved, theory should justify also why b is bounded, why $b \simeq 2$, why this behaviour is almost universally observed.

A possible explanation is given in "*Sample and population exponents of generalized Taylor's law*" by *Giometto et al.* [17].

3.3 Taylor's law exponents and the Large Deviation Theory

In the cited article the authors propose that the range of values observed for b could be a statistical artifact. In particular, the limited range observed could be a consequence of undersampling measurements, that is, measurements that are ineffective in detecting rare events, of major importance for computing the correct value of m and σ^2, and thus b. This section reviews the mathematical arguments and steps adopted in the article to reach the conclusion.

In what follows the theoreticalcal value of b (that should represent the value of b for the species considered) will be denoted simply with b and will be called the *population exponent*, the experimental value will be denoted with b_S and will be called the *sample exponent*, to stress it is computed via sampling.

3.3.1 The multiplicative random process in Markovian environment

The aim is to describe the evolution of the power exponent b for a species *with respect to time*. The population number, or density, of the species to which b is related, is then represented by a random variable N that is function of time t. The model adopted to describe the dynamics of $N(t)$ is the multiplicative growth model in Markovian environment:

$$N(t) = \prod_{n=1}^{t} A_n \tag{3.1}$$

where

$N(t)$ is the population number, or density, of the species over a fixed area, depending on time t,

$N_0 > 0$ is its initial value (initial t is 0),

A_n is a homogeneous Markov chain with

state space $\chi = \{r, s\}$, $\quad r \neq s, \quad r, s > 0,$

symmetric transition matrix Γ, $\quad \Gamma_{xy} > 0 \ \forall x, y \in \chi$:

$$\Gamma = \begin{pmatrix} 1-\gamma & \gamma \\ \gamma & 1-\gamma \end{pmatrix}, \quad \gamma \in (0,1).$$

The chain A_n is assumed to be at equilibrium, then the distribution $\pi(x)$ of the initial state A_1 is the stationary one. Since Γ is symmetric and $\Gamma_{xy} > 0 \ \forall x, y \in \chi$ it follows that

$$\pi(x) = 1/2, \quad x \in \chi, \quad \forall \gamma \in (0,1).$$

3.3.2 Sample and population exponents

The estimate of $b(t)$ requires the knowledge of expectation value and variance of $N(t)$. The definition of $N(t)$ implies that $b(t)$ depends also on the state space χ

and the transition probability γ, but these quantities are fixed in the model, then they will be regarded as parameters, not as independent variables. The random variable $N(t)$ is a product of r and s up to time t, so it does not depend on the order of r and s and it can be written as

$$N(t) = N_0 r^{tL_t(r)} s^{tL_t(s)} \tag{3.2}$$

with $L_t(z)$, for $z \in \{r, s\}$, defined by

$$L_t(z) = \frac{1}{t} \sum_{n=1}^{t} \delta_{A_n,z}$$

(δ is the Kronecker's delta). Then $L_t(r)$ is the fraction of times r appears in a given Markov chain and it is a discrete random variable valued in $[0,1]$. $L_t(s)$ has analogous meaning.

The connection between $Var[N(t)]$ and $\mathbb{E}[N(t)]$ is looked for when t is large, because for small t the chain is easily predicted studying its distribution law. In mathematical terms "t large" is simplified with $t \to \infty$ and the relationship to establish is between $\lim_{t\to\infty} Var[N(t)]$ and $\lim_{t\to\infty} \mathbb{E}[N(t)]$. The problem here is of the kind encountered in section 2.1: the probability distribution of $N(t)$ is known, but for t large the calculus of the statistics is anyway not practicable by analytical or numerocal methods.

Since $Var[N(t)]$ and $\mathbb{E}[N(t)]$ are both strictly positive and the logarithmic function is bijective on its domain, then $\log Var[N(t)]$ and $\log \mathbb{E}[N(t)]$ can be used. As to the first quantity, it is demonstrated that positivity of Γ and $r \neq s$ imply [5]

$$\lim_{t\to\infty} \frac{1}{t} \log Var[N(t)] = \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}[N(t)^2]. \tag{3.3}$$

Then, now, the link to set is between $\lim_{t\to\infty} \log \mathbb{E}[N(t)^2]$ and $\lim_{t\to\infty} \mathbb{E}[N(t)]$. This relationship can be obtained in the framework of Large Deviation Theory. In this context Gärtner-Ellis theorem states $L_t(r)$ obeys [20]:

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(L_t(r) \in [x, x+dx]) = -I_\Gamma(x) \tag{3.4}$$

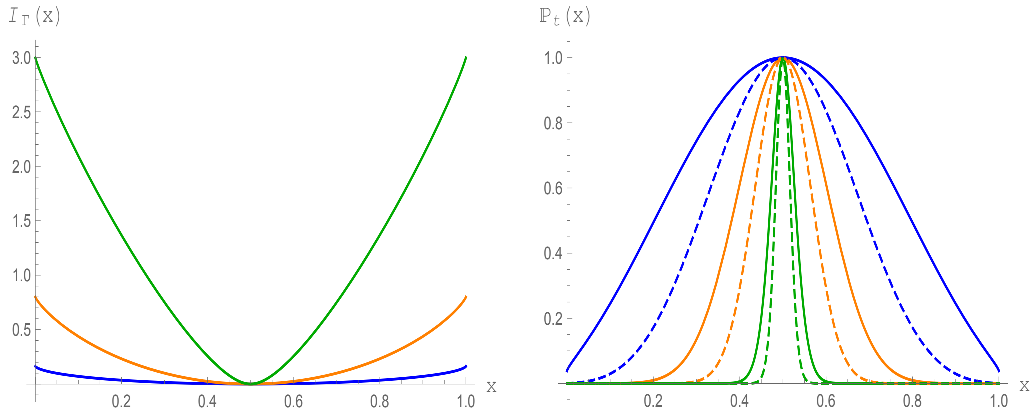


Figure 3.4: The rate function is here plotted for $\gamma_{\text{blue}} = 0.15$, $\gamma_{\text{orange}} = 0.55$ and $\gamma_{\text{green}} = 0.95$. On the right, with the same color code, $\mathbb{P}(L_t(r) \in [x, x + dx])$ is plotted for times $t_{\text{line}} = 20$ and $t_{\text{dashed}} = 50$. As x approaches 0 or 1 the probability for the state r to appear in a realization of the Markov chain A_n with frequency x decreases rapidly to zero with higher rate for larger t.

for $x \in [0, 1]$ and dx an infinitesimal interval, with rate function

$$I_\Gamma(x) = \sup_{u>0} \Big[x \log\big(\frac{u_1}{(\Gamma u)_1}\big) + (1-x)\log\big(\frac{u_2}{(\Gamma u)_2}\big)\Big] \tag{3.5}$$

where u is a vector in $\mathbb{R}^2$ with $u_1, u_2 > 0$. The dependency of I_Γ on the transition matrix is made explicit in the following form:

$$I_\Gamma(x) = (x-1)\log\big[1 - \gamma\big(\frac{2x(\gamma-1)}{C_\gamma(x) - 2\gamma x} + 1\big)\big] - x\log\big[1 + \frac{\gamma(2x - C_\gamma(x))}{2x(\gamma-1)}\big], \tag{3.6}$$

with

$$C_\gamma(x) = \gamma + \sqrt{\gamma^2 + x(x-1)(8\gamma - 4)}.$$

Figure 3.4 shows I_Γ for different values of γ. As seen in Chapter 1, every rate function I is convex and $I(x) \geq 0 \quad \forall x \in \mathbb{R}$. Here I_Γ has its minimum in $x = 1/2$, $I_\Gamma(1/2) = 0$, independently of Γ.

Finally, once (3.4) is obeyed, Large Deviation Theory gives, by means of Varadhan's lemma [20],

$$\lim_{t\to\infty} t^{-1} \log \mathbb{E}[N(t)^k] = \sup_{x\in[0,1]} \left[kG(x) - I_\Gamma(x)\right], \tag{3.7}$$

with $G(x) = x \log r + (1-x) \log s$ and $k \in \mathbb{N}$. From (3.3) and LDT result (3.7), it is now possible to give an expression for Taylor's law exponent b:

$$b(\gamma) = \frac{\sup_{x\in[0,1]} \left[2G(x) - I_\Gamma(x)\right]}{\sup_{x\in[0,1]} \left[G(x) - I_\Gamma(x)\right]}. \tag{3.8}$$

Here it has been remarked the b dependence on Γ.

Generalized exponents

Taylor's law may be extended to set a connection between moments higher than the first and the second, in which case it is called the *generalized Taylor's law*:

$$\mathbb{E}[N^k(t)] = a_{jk}\mathbb{E}[N^j(t)]^{b_{jk}}. \tag{3.9}$$

As to this form, the goal is to predict the behaviour of the generalized exponents b_{jk}. By repeating the passages adopted to derive $b(\gamma)$, in the hyphotesis $t \to \infty$, result (3.7) gives

$$b_{jk}(\gamma) = \frac{\lim_{t\to\infty} t^{-1} \log \mathbb{E}[N(t)^k]}{\lim_{t\to\infty} t^{-1} \log \mathbb{E}[N(t)^j]} = \frac{\sup_{x\in[0,1]} \left[kG(x) - I_\Gamma(x)\right]}{\sup_{x\in[0,1]} \left[jG(x) - I_\Gamma(x)\right]}. \tag{3.10}$$

The population exponents $b(\gamma)$, equivalent to $b_{12}(\gamma)$, and $b_{jk}(\gamma)$ can have discontinuities for some critical values γ_c, as reported in figure 3.5 in black continuous line, depending on the state space χ. This should be a property of the species to which $N(t)$ is related (because γ is the transition probability that should reproduce the dynamics of $N(t)$ through the chain A_n).

From now on the focus will be on $b(\gamma)$, having $b_{jk}(\gamma)$ identical structure.

3.3.3 The role of rare events

Now comes the most important step to get the result: $b(\gamma)$ in (3.8) will be obtained, in a sampling of $N(t)$, only if in the sampling *all* the values $x \in [0,1]$ are observed. The value $b(\gamma)$ in (3.8) that *should* appear in Taylor's law, is correctly estimated, in a sampling, by b_S only if in the sampling that gives b_S *all* the values $x \in [0,1]$ come out, because only in this case the *suprema* in (3.8) are correctly computed. In the expression for b the argument x is the fraction of times r appears in a generic realization, or trial, of the Markov chain A_n up to t, determining $N(t)$ in that generic realization. Thus b_S can coincide with b only if the sampling of $N(t)$ is made of a number of trials of the chain A_n sufficient to see the state r appearing 0 times, 1 time, $\ldots$, t times in the chain up to time t with the proper frequencies through the trials. Indeed only in this case the sampling detects all the possible values for x in $[0,1]$ needed to evaluate the suprema in (3.8).

The obstacle is that the probability for r to appear, in a trial of the chain, with extreme frequencies, $x \simeq 0$ or $x \simeq 1$, is extremely low, then a very large number of trials is needed to observe this rare event. In other words, the probability of missing these extreme values of x in the sampling is very high and this corresponds to compute the *suprema* in (3.8) over a sub-interval $(x_-, x_+) \subset [0,1]$, getting $b_S \neq b$ (and $b_{Sjk} \neq b_{jk}$).

On the right of figure 3.5 it is reported the gap between population b and sample b_S exponents as a function of time for fixed χ and γ. Here the sampling is performed via a direct simulation, as explained in 3.3.4. The number of trials of the sampling -the size of the simulation- changes the time up to which the estimate b_S of b is still acceptable: a higher number of trials provides b_S close to b for a longer time. Nevertheless, for finite number of trials, all the simulations fail in estimating b as t increases.

Up to now it has been demonstrated that a possible cause to the discrepancy between population and sample exponents is the sampling inefficiency: for large t the sampling of N carried out by a direct simulation method cannot detect all the events needed in (3.8) to compute the population exponents.

By the point of view of probability theory this happens because, for large t the

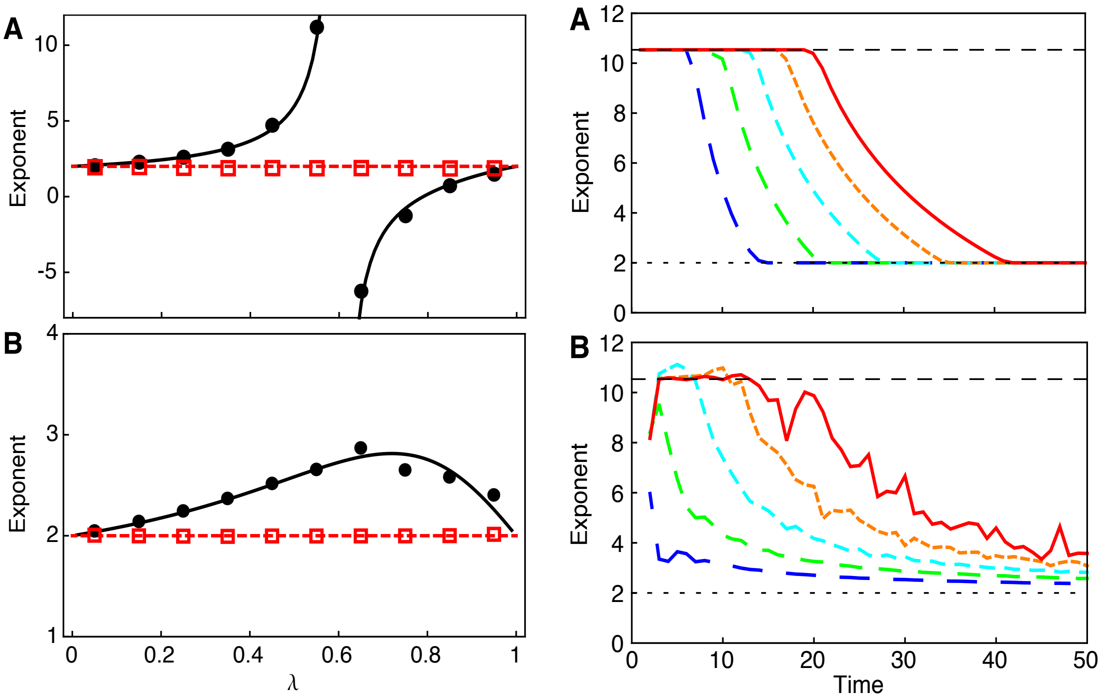


Figure 3.5: Results from [17]. On the left it is shown the behaviour of $b(\gamma)$, in A for $\chi = \{2, 1/4\}$, in B for $\chi = \{4, 1/2\}$. As predicted, in A $b(\gamma)$ shows a discontinuity. Theoretical results, expression (3.8), are in black lines, results from simulation are in black dots for $t = 10$, $R = 10^6$, red squares for $t = 400$, $R = 10^4$. Black dots meet condition $t \ll \log R$ and, according to discussion of 3.3.3, they well reproduce $b(\gamma)$, while red squares, being obtained with $t \ll \log R$, comply with condition (3.18), reproducing a wrong estimate. On the right it is displayed $b(t)$ for $\chi = \{2, 1/4\}$ and with fixed $\gamma = 0.55$, close to the critical γ_c. In A theoretical results are displayed from (3.15). The upper dashed horizontal line is $b(\gamma = 0.55)$ from (3.8), the dotted lower line is instead $b_R(\gamma = 0.55)$ in (3.15). Different colors refer to $R = 10^n$ trials of the chain A_n, going from $n_{\text{blue}} = 2$ to $n_{\text{red}} = 6$ and results are averaged over $10^8/R$ simulations (the blue curve over 10^5 simulations). As predicted, for t large a simulation with finite size R anyway fails in providing an accurate estimate for b.

direct simulation method fails in revealing all the possible paths of the Markov chain A_n up to the time t. In the remainder of the section Large Deviation Theory is again used to find the number R of independent trials required to correctly estimate b for a fixed t, with a direct simulation method.

Since the difference between b_S and b originates from the difficulty of detecting the events of r appearing with extreme frequencies in the sampling, the probability to be studied is $\mathbb{P}(L_t(r) > x)$ when $x \simeq 1$ and $x \simeq 0$.
In order to do this, independent identically distributed random variables $X^i(t) = L_t^i(r)$, $i = 1, \dots, R$, are introduced, where $L_t^i(r)$ is the frequency of r in the i-th trial of the chain A_n and R the number of independent trials of the Markov chain A_n realized in the sampling. The random variable x_+ defined as

$$x_+ = \max\{X^1(t), \dots, X^R(t)\}$$

can then be interpreted as the typical maximum frequency with which the state r is observed in a chain A_n. Because the Markov chains A_n are independently replicated and x_+ is computed over R of these chains, it follows, for R large,

$$\mathbb{P}(L_t(r) > x_+) = \frac{1}{R}. \tag{3.11}$$

Analogously the typical minimum frequency with which r is observed in a chain is

$$x_- = \min\{X^1(t), \dots, X^R(t)\}$$

and it will be

$$\mathbb{P}(L_t(r) < x_-) = \frac{1}{R}. \tag{3.12}$$

Now Large Deviation Theory comes into play to estimate R for t large. Recalling the condition $t \to \infty$, Varadhan's lemma gives [17, 20]

$$R \simeq \exp\left[tI_\Gamma(x_\pm)\right], \tag{3.13}$$

For what has been said above, R in (3.13) is (an estimate of) the number of independent realizations of the chain A_n needed to compute $\mathbb{P}(L_t(r) > x)$ for extreme x $(x \simeq 0, x \simeq 1)$. It is now possible to get $x_\pm$ by taking the logarithm of both sides and expanding I_Γ in Taylor series around x_{min}:

$$x_\pm = \frac{1}{2} \pm \sqrt{\Big(\frac{1-\gamma}{2\gamma}\Big)\frac{\log R}{t}}. \tag{3.14}$$

Finally the sample exponents are

$$b_S(\gamma, t, R) = b_R(\gamma, t) \simeq \frac{\sup_{x\in[x_-,x_+]}\left[2G(x) - I_\Gamma(x)\right]}{\sup_{x\in[x_-,x_+]}\left[G(x) - I_\Gamma(x)\right]}, \tag{3.15}$$

and

$$b_{Sjk}(\gamma, t, R) = b_{Rjk}(\gamma, t) \simeq \frac{\sup_{x\in[x_-,x_+]}\left[kG(x) - I_\Gamma(x)\right]}{\sup_{x\in[x_-,x_+]}\left[jG(x) - I_\Gamma(x)\right]} \tag{3.16}$$

where for notation clarity S is substituted by R, being R the "size" of S.

From (3.14) it follows that for fixed R the arguments in (3.15) are computed over the interval $[x_-, x_+]$ which is centred on $x = x_{min}$ and becomes smaller as t increases.

Because of $I_\Gamma(x = x_{min}) = 0$, for a finite number R of realizations of the process $N(t)$, the weight of I_Γ in the arguments of (3.15) goes to 0 as t increases and therefore

$$\lim_{t\to\infty} b_R(\gamma, t) = 2.$$

In particular after just a time t' such that $\log R = o(t')$, the estimate of b is $b_R(\gamma, t) \simeq 2$, whatever b is! According to (3.14) to get instead an estimate b_R near b for a time t the order of magnitude of R is $R = e^t$.

On the contrary, b_R is close to the population exponent b when the arguments of the suprema in (3.8) occcur for $x' \in [x_-, x_+]$ (in this case, in the region $t \to \infty$ where (3.8) describes exactly b, $b_R = b$).

For example, if the suprema are reached in $x' > 1/2$, then, from $x_\pm = \frac{1}{2} \pm \sqrt{\left(\frac{1-\gamma}{2\gamma}\right)\frac{\log R}{t}}$, b_R well approximates b if $x' < x_+$, equivalently if

$$t < \left(\frac{1-\gamma}{2\gamma}\right)(x' - 1/2)^{-2} \log R, \tag{3.17}$$

while $b_R \to 2$ if

$$t > \left(\frac{1-\gamma}{2\gamma}\right)(x' - 1/2)^{-2} \log R. \tag{3.18}$$

implying an estimate b_R close to b requires a number R of trials of order $R = e^t$. If the values r and s cause a discontinuity for $b(\gamma)$ at a critical γ_c, then the limits given for t define the regions in which measures of b_R show the same discontinuity. Analogous results hold for b_{Rjk}.

3.3.4 The direct simulation procedure

The process $N(t)$ is defined by a Markov chain whose transition probabilities are known, then the sampling of $N(t)$ will be obtained through simulation.

Procedure for $b_\gamma(t)$. Here γ and $\chi = \{r, s\}$ are fixed. To get $b(\cdot)$ for times t' up to t the procedure consists of the steps:

1. simulate R chains A_n up to $n = t$,

2. at each time $n \leq t$ compute the expectation value and variance of $N(\cdot)$ as

$$\mathbb{E}[N(\cdot)] = \frac{1}{R}\sum_{j=1}^{R} N(\cdot)_j,$$

$$Var[N(\cdot)] = \frac{1}{R}\sum_{j=1}^{R} N^2(\cdot) - \mathbb{E}[N(\cdot)]^2,$$

   by using (3.1),

3. compute the linear interpolation for the points $\log \mathbb{E}[N(\cdot)]$ versus $\log Var[N(\cdot)]$ up to t'.

Table 3.1: Scheme of the procedure to compute $b_\gamma(t)$.

	times						
trial	1	2	$\dots$	t'	$\dots$	$t-1$	t
1	$A_1(1)$	$A_1(2)$	$\dots$	$A_1(t')$	$\dots$	$A_1(t-1)$	$A_1(t)$
2	$A_2(1)$	$A_2(2)$	$\dots$	$A_2(t')$	$\dots$	$A_2(t-1)$	$A_2(t)$
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
R	$A_R(1)$	$A_R(2)$	$\dots$	$A_R(t')$	$\dots$	$A_R(t-1)$	$A_R(t)$
	$\downarrow$	$\downarrow$		$\downarrow$		$\downarrow$	$\downarrow$
	$\mathbb{E}[N(1)]$	$\mathbb{E}[N(2)]$	$\dots$	$\mathbb{E}[N(t')]$	$\dots$	$\mathbb{E}[N(t-1)]$	$\mathbb{E}[N(t)]$
	$Var[N(1)]$	$Var[N(2)]$	$\dots$	$Var[N(t')]$	$\dots$	$Var[N(t-1)]$	$Var[N(t)]$
				$\downarrow$		$\downarrow$	$\downarrow$
				$b(t')$	$\cdots$	$b(t-1)$	$b(t)$

The exponent $b(t')$ is the angular coefficient of the linear interpolation. The criterion to select the time interval for the interpolation is the following:

a time interval can be used if, increasing it by a time unit, the angular coefficient of the interpolation does not change.

A scheme of the procedure is reported in table 3.1.

Procedure for $b_t(\gamma)$. The procedure is the same followed for $b_\gamma(t)$, but here t is fixed and the parameter γ varies in $(0, 1)$. The computation used different state spaces to verify the existence of critical values γ_c, as predicted in sections 3.3.2 and 3.3.3. Results reported in figure 3.5 refer to $\chi = \{2, 1/4\}$ and $\chi = \{4, 1/2\}$. The t parameter and the number R of independent trials have been chosen to stress in the most clear way the different behaviour of $b(\gamma, t)$, according to the discussion in 3.3.3, and to make the simulation accomplishable: $t = 10$ with $R = 10^6$ (satisfying (3.17)) and $t = 400$ with $R = 10^4$ (satisfying (3.18)).

3.3.5 How to measure b_R?

In the article by Giometto et al. [17] two important results have been established:

- the behaviour of the *observed* Taylor's law exponent ($b_R \simeq 2$ almost independently of process characteristics) may be the result of limited sampling efforts;
- to get an estimate b_R close to b a number $R \simeq e^t$ of independent replicates of the process is needed.

It is now clear that an accurate estimate of b or b_{jk} is not always attainable: for t large the condition $R \simeq e^t$ can make the sampling impossible to perform. As example, for the multiplicative process adopted, if $\gamma = 0.5$ and $t = 100$, condition (3.17) requires a number $R \simeq 10^{13}$ of independent trials to observe the rare event $\{x_+ = 0.9\}$. To accomplish this task the random number generator should provide $R \times t \simeq 10^{15}$ independent random numbers r or s according to the transition matrix of the chain, since every chain, in this example, consists of 100 steps. Even when t is one order of magnitude smaller, for example in the range $20 \div 50$, the computation cannot be completed within short time intervals. An estimate of the running time that clinches the matter is given at the end of Chapter 4, it's somewhere in the vicinity of years ...

This argument is the ultimate reason why a high-efficiency simulation method is needed. The next Chapter is dedicated to design the best sampling for b, finally showing the utility of Large Deviation Theory.

Chapter 4

Importance Sampling Estimate of Taylor's Law Exponent

This chapter is dedicated to estimating the power exponent b of Taylor's law. The problem of computing b_R - estimate of b - is that exposed in Chapter 2: to identify a simulation method that provides expectation value and variance of a random variable with a fixed precision and with a number R of independent trials far lesser than the size of a direct simulation. In Chapter 2 the analysis of rare events by means of the Large Deviation Theory lead to the importance sampling technique as a high-efficiency simulation method. This new technique is implemented here to finally compute b. Results will show that a highly efficient simulation based on importance sampling method reveals the predicted value of the exponent, confirming what has been reached in Chapters 2 and 3.

4.1 Rare events in the Markovian multiplicative model

The quantities to be estimated are expectation value and variance of the random variable

$$N(t) = N_0 \prod_{n=1}^{t} A_n,$$

where A_n is the Markov chain as defined in 3.3.1.

The procedure based on the direct simulation method (DS), explained in 3.3.4, evaluates at each time t $\mathbb{E}[N(t)]$ and $Var[N(t)]$ *directly*, and treats all the events entering the process $N(t)$ in the same manner. Such a simple approach, schematized in table 3.1, generates R independent Markov chains A_n and, for each time step, computes the statistics at the end of the simulation: it is ineffective in revealing the rare events hidden in $N(t)$.

The importance sampling technique (IS), on the contrary, focuses on single sets: it estimates rare events probabilities by means of new probability distributions -the biasing distributions studied in 2.1.1- which are specific of each rare event. The procedure based on IS method, then, requires isolating the single sets that make $N(t)$, in particular the low probability events.

Now, in detailed way, we find these important sets, the rare events of $N(t)$.

To do this, it is convenient to express $N(t)$ by using $L_t(z)$ defined in 3.3.2: since $L_t(r) + L_t(s) = 1$ the expression of $N(t)$ becomes

$$\begin{aligned} N(t) = N_0 r^{tL_t(r)} s^{t(1-L_t(r))} &= N_0 s^t \Big(\frac{r}{s}\Big)^{tL_t(r)} \\ &= N_0 s^t \Big(\frac{r}{s}\Big)^{\sum_{i=1}^{t} \delta_{A_i,r}}. \end{aligned} \tag{4.1}$$

The expectation value of $N(t)$ can thus be written as

$$\begin{aligned} \mathbb{E}[N(t)] = N_0 s^t \mathbb{E}\big[\Big(\frac{r}{s}\Big)^{tL_t(r)}\big] &= N_0 s^t \sum_{m=0}^{t} \Big(\frac{r}{s}\Big)^m \mathbb{P}(tL_t(r) = m) \\ = N_0 s^t \sum_{m=0}^{t} \Big(\frac{r}{s}\Big)^m \mathbb{P}\big(\sum_{i=1}^{t} \delta_{A_i,r} = m\big) \end{aligned} \tag{4.2}$$

and the variance as

$$\begin{aligned} Var[N(t)] = \mathbb{E}[N(t)^2] - \mathbb{E}[N(t)]^2 &= N_0^2 s^{2t} \sum_{m=0}^{t} \Big[\Big(\frac{r}{s}\Big)^{2m} \mathbb{P}(tL_t(r) = m) \Big] - \mathbb{E}[N(t)]^2 \\ &= N_0^2 s^{2t} \sum_{m=0}^{t} \Big[\Big(\frac{r}{s}\Big)^{2m} \mathbb{P}(\sum_{i=1}^{t} \delta_{A_i,r} = m) \Big] - \mathbb{E}[N(t)]^2. \end{aligned} \tag{4.3}$$

The above expressions show that the statistics of $N(t)$ depend on the sets

$$\Big\{ \frac{1}{t} \sum_{i=1}^{t} \delta_{A_i,r} = x \Big\}, \quad x \in \{0, \frac{1}{t}, \frac{2}{t}, \dots, \frac{t-1}{t}, 1\}. \tag{4.4}$$

As discussed in Chapter 3, the direct simulation method fails in estimating b for large t because it is ineffective in detecting the events in (4.4) in which the random variable

$$L_t(r) = \frac{1}{t} \sum_{i=1}^{t} \delta_{A_i,r}$$

takes values far from its expectation value μ:

$$\Big\{ \frac{1}{t} \sum_{i=1}^{t} \delta_{A_i,r} \ll \mu \Big\} \quad \text{and} \quad \Big\{ \frac{1}{t} \sum_{i=1}^{t} \delta_{A_i,r} \gg \mu \Big\}. \tag{4.5}$$

The direct simulation cannot reveal events in which the state r appears in the chain A_n up to t (large) with frequency $x \simeq 0$ or $x \simeq 1$:

$$\Big\{ \frac{1}{t} \sum_{i=1}^{t} \delta_{A_i,r} = x \Big\} \quad \text{with} \quad 0 \leq x \ll \mu \quad \text{or} \quad \mu \ll x \leq 1. \tag{4.6}$$

The probability characterizing these events has been computed by LDT: the large deviation principle (3.4) and the expression (3.5) of the rate function imply the probability is extremely low. Thus, the sets defined in (4.5) or (4.6) are the rare events that the new simulation procedure must detect and whose probability it must estimate.

4.1.1 The importance sampling procedure

The procedure to compute $b(t)$, *for t fixed* will be the following:

1) for $t' \leq t$, estimate $\mathbb{P}(\sum_{i=1}^{t'} \delta_{A_i,r} > m)$ for $m = 0, 1, \ldots, t'$ by adopting the importance sampling technique,

2) compute $\mathbb{P}(\sum_{i=1}^{t'} \delta_{A_i,r} = m)$ for $m = 1, \ldots, t'$ as
$\mathbb{P}(\sum_{i=1}^{t'} \delta_{A_i,r} = m) = \mathbb{P}(\sum_{i=1}^{t'} \delta_{A_i,r} > m-1) - \mathbb{P}(\sum_{i=1}^{t'} \delta_{A_i,r} > m)$,

3) obtain the curve $\log \mathbb{E}[N(t')]$ versus $\log \mathit{Var}[N(t')]$ through (4.2) and (4.3) respectively and $b(t)$ as the angular coefficient of its linear interpolation.

Remark.

- According to LDT results given in (2.2), probabilities can be estimated for sets admitting a dominating point. Then the events considered in step 1) are those in which $\sum_{i=1}^{t'} \delta_{A_i,r}$ is *greater* than m.

- While for a generic Markov chain the probability has to be estimated for every possible value $m = 0, 1, \ldots, t'$, the simmetry of the transition matrix of the chain considered in Chapter 3 allows to compute the probabilities for the values $m = t'/2, \ldots, t'$ only.

- The procedure described above must be repeated for every t of interest, since for different values of t the sets $\{\sum_{i=1}^{t'} \delta_{A_n,r} = m\}$ are different sets with different probabilities.

- The power exponent $b(t)$ can be obtained by an interpolation procedure only, because there is no relationship between the two parameters a and b in the Taylor's law that makes it possible to determine $b(t)$ with the only point $(\mathbb{E}[N(t)], \mathit{Var}[N(t)])$.

4.2 Computation of the power exponent

The importance sampling technique requires the knowledge of the biasing probability distribution $b(A_1, \dots, A_n)$ for $n = t'/2, \dots, t'$. From now on the distribution b will be denoted with p_b. The work to identify p_b has been done in 2.2.3. In the notation of that section

$$\chi = \{r, s\}, \quad k = 1,$$

$$f \colon \chi \to \mathbb{R}, \quad f(A_i) = \delta_{A_i,r}.$$

The biasing p_b is given by expression (2.18):

$$p_{b,xy} = p_{xy} \exp\bigl(\theta_t f(y)\bigr) \frac{\psi_{\theta_t}(y)}{\lambda(\theta_t)\psi_{\theta_t}(x)}, \tag{4.7}$$

where $\lambda(\theta)$ and ψ_θ are respectively the maximum eigenvalue and its corresponding eigenvector of the operator T_θ defined in (1.4) and θ_t is solution of

$$t = \frac{\lambda'(\theta)}{\lambda(\theta)}. \tag{4.8}$$

For the chain A_n considered here the operator is

$$T_\theta = \begin{bmatrix} p_{rr} \exp\theta & p_{rs} \\ p_{sr} \exp\theta & p_{ss} \end{bmatrix} = \begin{bmatrix} (1-\gamma)\exp\theta & \gamma \\ \gamma \exp\theta & 1-\gamma \end{bmatrix} \tag{4.9}$$

with maximum eigenvalue

$$\lambda(\theta) = \frac{(1-\gamma)(1+\exp\theta) + \sqrt{(1-\gamma)^2(1+\exp\theta)^2 - 4\exp\theta(1-2\lambda)}}{2} \tag{4.10}$$

and corresponding eigenvector

$$\psi_\theta = \begin{pmatrix} 1 \\ \frac{\gamma \exp\theta}{\gamma - 1 + \lambda(\theta)} \end{pmatrix}. \tag{4.11}$$

The biasing distribution is represented by the transition matrix Γ_b of elements $\Gamma_{b,xy} = p_{b,xy}$ as given in (4.7):

$$\Gamma_b(\theta_t) = \frac{1}{\lambda(\theta_t)} \begin{pmatrix} (1-\gamma)\exp\theta_t & \gamma\frac{\psi_{\theta_t}(s)}{\psi_{\theta_t}(r)} \\ \gamma\frac{\psi_{\theta_t}(r)}{\psi_{\theta_t}(s)} \exp\theta_t & 1-\gamma \end{pmatrix}. \tag{4.12}$$

In the above expression $\psi_{\theta_t}(z)$ is the component of the ψ_{θ_t} eigenvector that corresponds to the state $z \in \chi$, then, from (4.11), $\psi_{\theta_t}(r) = 1$, $\psi_{\theta_t}(s) = \gamma \exp \theta_t / (\gamma - 1 + \lambda(\theta_t))$.

4.2.1 The importance sampling estimators

Estimators of the probabilities $\mathbb{P}(\sum_{i=1}^n \delta_{A_i,r} > m)$ of step 1) of the procedure to get $b(t)$ are

$$\begin{aligned}
\bar{\mu}_n &= \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{\sum_{i=1}^n \delta_{\tilde{A}_i,r} > m\}} \frac{p(\tilde{A}_1^j, \ldots, \tilde{A}_n^j)}{p_b(\tilde{A}_1^j, \ldots, \tilde{A}_n^j)} \\
&= \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{\sum_{i=1}^n \delta_{\tilde{A}_i,r} > m\}} \frac{p(\tilde{A}_1^j) \prod_{i=1}^{n-1} p_{\tilde{A}_i^j \tilde{A}_{i+1}^j}}{p_b(\tilde{A}_1^j) \prod_{i=1}^{n-1} p_{b,\tilde{A}_i^j \tilde{A}_{i+1}^j}} \\
&= \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{\sum_{i=1}^n \delta_{\tilde{A}_i,r} > m\}} \frac{p(\tilde{A}_1^j) \prod_{i=1}^{n-1} p_{\tilde{A}_i^j \tilde{A}_{i+1}^j}}{p_b(\tilde{A}_1^j) \prod_{i=1}^{n-1} \big(p_{\tilde{A}_i^j \tilde{A}_{i+1}^j} \exp{(\theta_t \delta_{\tilde{A}_{i+1}^j,r})} \frac{\psi_{\theta_t}(\tilde{A}_{i+1}^j)}{\psi_{\theta_t}(\tilde{A}_i^j)\lambda(\theta_t)}\big)}
\end{aligned}$$

and, thanks to the products in the last ratio,

$$\bar{\mu}_n = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{\sum_{i=1}^n \delta_{\tilde{A}_i,r} > m\}} \frac{p(\tilde{A}_1^j)}{p_b(\tilde{A}_1^j)} \frac{\psi_{\theta_t}(\tilde{A}_1^j)}{\psi_{\theta_t}(\tilde{A}_n^j)} \frac{\lambda^{n-1}(\theta_t)}{\exp\big(\theta_t \sum_{i=1}^{n-1} \delta_{\tilde{A}_{i+1}^j,r}\big)} \tag{4.13}$$

where the index j stands for the j-th trial of the new Markov chain $\tilde{A}_n$ evolving in each trial with transition matrix Γ_b. The expression (4.13) requires the knowledge of $p(\tilde{A}_1)$ and $p_b(\tilde{A}_1)$, for each j-th trial. How to choose these values? By hyphotesis the chain A_n starts at equilibrium, that is, the distribution π of A_1 is the stationary one (being Γ symmetric $\pi(r) = \pi(s) = 1/2$). The element A_1 can be thought as the element of A_n following a starting point A_0 which, itself, is chosen with the stationary distribution π. The random variable $\tilde{A}_1$ is instead the first state of the chain $\tilde{A}_n$ that is governed by the new transition matrix Γ_b and that cannot be a priori assumed to be at equilibrium. Nevertheless, $\tilde{A}_n$ is a realization of the chain A_n modified according to Γ_b from element $\tilde{A}_1$ only: $\tilde{A}_1$ follows an element $\tilde{A}_0$

whose distribution is π, the same stationary distribution of the chain A_n. Indeed p and p_b have for $\tilde{A}_1^j$ the meaning:

$$\begin{aligned} p(\tilde{A}_1^j) &= p_{\tilde{A}_0^j \tilde{A}_1^j} \\ p_b(\tilde{A}_1^j) &= p_{b,\tilde{A}_0^j \tilde{A}_1^j}, \end{aligned}$$

where, for every j-th trial, $\tilde{A}_0^j$ is chosen according to the stationary distribution $\pi(\cdot) = 1/2$. By inserting the last expressions into (4.13), with p_b as in (4.7), the estimators are

$$\bar{\mu}_n = \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{\sum_{i=1}^{n} \delta_{\tilde{A}_i,r} > m\}} \frac{\lambda^n(\theta_t)}{\exp\left(\theta_t \sum_{i=1}^{n} \delta_{\tilde{A}_i,r}\right)} \frac{\psi_{\theta_t}(\tilde{A}_0^j)}{\psi_{\theta_t}(\tilde{A}_n^j)}. \tag{4.14}$$

The simulation codes that provided the outputs of this Chapter are listed in B.1.

4.3 Results

The organization is: first it is reported the case where χ causes a discontinuity in $b(\gamma)$, then the case of no discontinuity (denoted respectively with χDis and χNoD). In the former case $b(t)$ and $b(\gamma)$ are both analyzed, in the latter, less critical, it is displayed only the behaviour of $b(\gamma)$. The displays will be easily compared with those of figure 3.5 in Chapter 3.

Error analysis.

The error bars shown in the figures have been computed as $[b-\epsilon b, b+\epsilon b]$ with b the predicted value of the exponent and ϵ the maximum error in the simulations. The correct error bars, computed with the theory of error propagation (as in Chapters 10 and 11 in [27]), would be quite larger than the one here adopted. Then, the error bars reported in all the figures stand for precision requirements even stricter than those that have been chosen for the simulations.

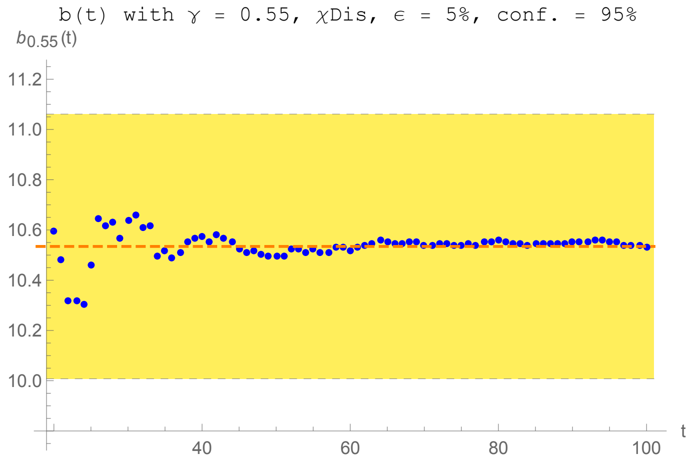
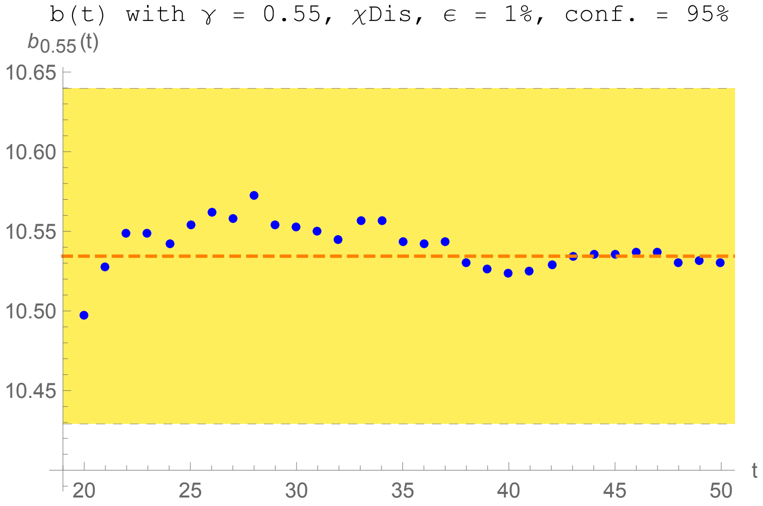


Figure 4.1: Behaviour of $b(t)$ for fixed $\gamma = 0.55$.

Discontinuity case: $b(t)$.

Figure 4.1 reports values of $b(t)$ for $\chi = \{2, 1/4\}$ with fixed $\gamma = 0.55$. The simulation has been performed requiring a maximum error $\epsilon = 5\%$ and a 95% confidence. The bright orange line is the theoretical value ($b = 10.5344$ for $\gamma = 0.55$). The light orange region is the 5% error bar $[b - \epsilon b, b + \epsilon b]$ within which results should be found. In the following figures results refer to higher precision levels. All graphics reported above show the IS method provides estimate of $b(t)$ within the desired maximum error. In particular results reproduce the theoretical value for large t better than for early t. This feature can be explained one more time by Large Deviations Theory. For small times the events entering the process are not as rare as for large times and the biasing probability distributions are not appreciably different from the original ones. For this reason results show for small t the instability already seen with the direct simulation mehod. As to the behaviour of $b(t)$ reported in figure 4.2, values obtained by requiring 95% confidence level seem to better reproduce the theoretical value with respect to the values of the second graphic. This is due only to the particular random paths followed by the



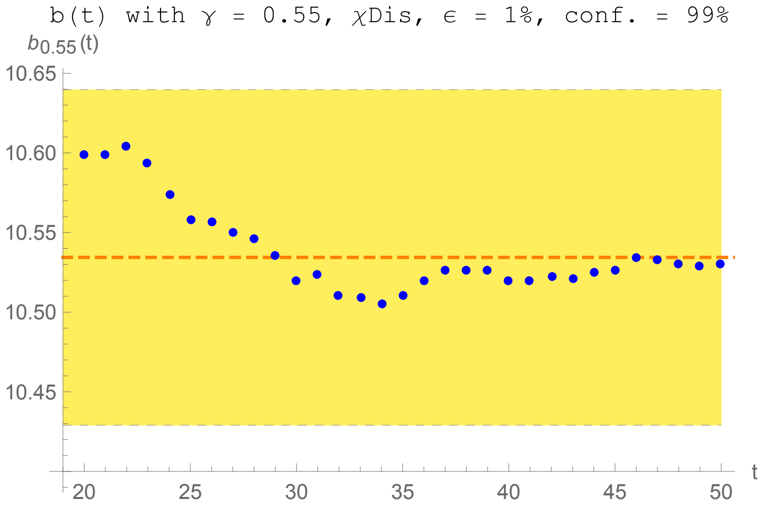


Figure 4.2: The error bar is $[b - \epsilon b, b + \epsilon b]$ with $\epsilon = 1\%$ in both graphics, but the reported values have been obtained requiring different confidence level.

process in the simulations. Being the maximum error requirement the same, the only difference that could be found between the two sets of outcomes is the presence of some values outside the error bar, more in the first graphic than in the second.

Figure 4.3 reports the numbers $k(m)$ of independent trials of the Markov chain A_n realized in the simulation to estimate $\mathbb{P}(\sum_i^t \delta_{A_i,r} > m)$, where $t/2 \leq m \leq t$, according to the criterion (2.12) defined in 2.1.3. Results shown correspond to graphic in figure 4.1, for times t multiple of 10. A common pattern is shared by each curve, for t equal to 10 through 100. This feature is due to the combined choice of the IS biasing probability distribution and the simulation stop criterion: for small m the events entering the process have high probabilities relatively to the events at m large and the estimate of their probability requires a low number of trials to satisfy the stop criterion (2.12). As m increases the events probabilities decrease and a higher number of trials are needed to meet the stop criterion. But the number $k(m)$ doesn't increase for every m. For m approaching t the biasing probability distribution makes the rare event more likely and simultaneously the estimate $\bar{\mu}_b^2$ in (2.11) decreases to very low values. Since the estimator variance $\bar{V}_b$ is reduced but not zero, the ratio $\bar{V}_b/\bar{\mu}_b^2$ becomes large and the stop criterion is met sooner than for m in the middle of $[t/2, t]$. The behaviour of $k(m)$ in estimating $b(t)$ is similar for different precision levels, as shown in figure 4.4.

Discontinuity case: $b(\gamma)$.

The curve $b(\gamma)$ has been reproduced for $t = 10$ and $t = 40$. Results are shown in figures 4.5 and 4.6. The error bar $[b(\gamma) - \epsilon b(\gamma), b(\gamma) + \epsilon b(\gamma)]$, due to the behaviour of the theoretical $b(\gamma)$, diverges for $\gamma \to \gamma_c$ and reduces to zero for $\gamma \simeq 0.7$.
Because of the performance of the IS simulation method for small t, discussed before, the precision requirements are not satisfied for $t = 10$. This, anyway, doesn't imply the IS method has failed: the requirements were set on the estimators

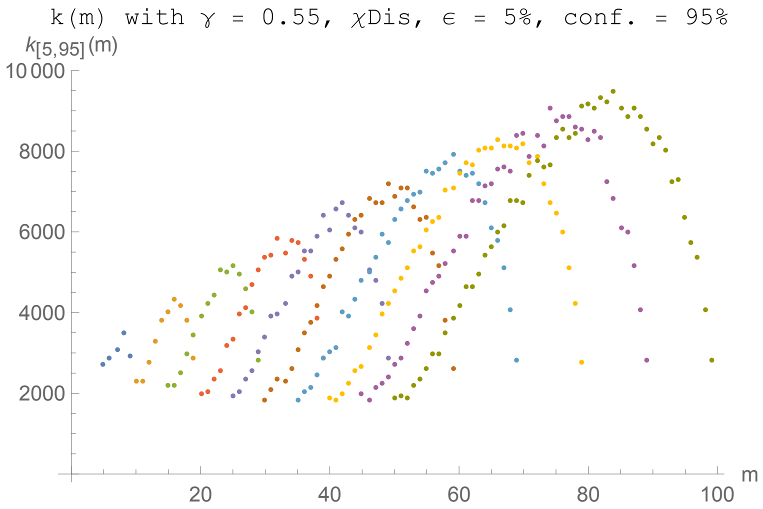
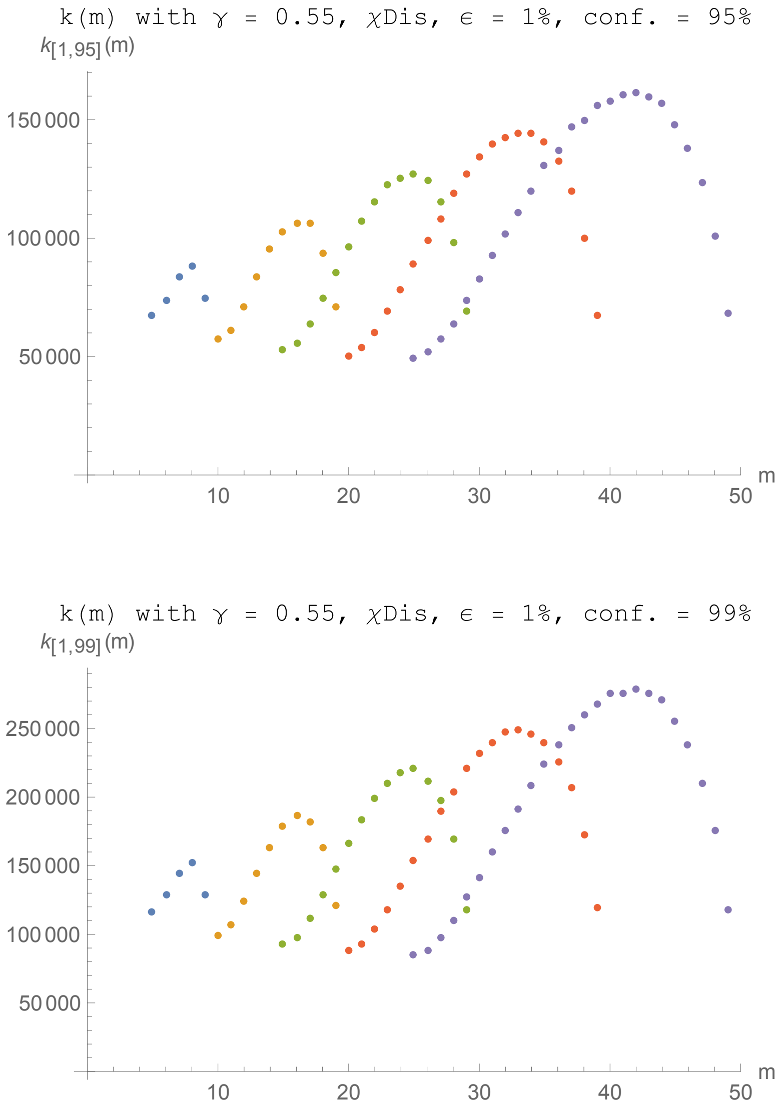


Figure 4.3: $k(m)$ = #trials to estimate $\mathbb{P}(\sum_i^t \delta_{A_i,r} > m)$. Different colors refer to t multiple of 10, going 10 to 100. $k(m)$ shows identical features independently of t.

of the partial probabilities $\mathbb{P}(\sum_i^t \delta_{A_i,r} > m)$ that had to be multiplied by $(r/s)^m$ and this term, for $\chi = \{2, 1/4\}$ and $m \simeq t$, rises to the point of eventually producing a final value of $b(\gamma)$ quite different from the theoretical value. Results for $t = 40$, instead, comply with the required precision for every γ, again confirming the IS method is reliable for increasing times.

No discontinuity case: $b(\gamma)$**.**

In figures 4.7 and 4.8 results of $b(\gamma)$ are reported when $\chi = \{4, 1/2\}$, for which the theoretical $b(\gamma)$ shows no discontinuity. Also in absence of discontinuity the IS simulation method is more reliable for large than for small times. Nevertheless, for $t = 10$ the experimental curve is anyway far from being a straight orizontal line, showing the IS method has been able to reveal that b varies for different γ.



k(m) with γ = 0.55, χDis, ϵ = 1%, conf. = 95%

k(m) with γ = 0.55, χDis, ϵ = 1%, conf. = 99%

Figure 4.4: To meet higher precision levels $k(m)$ increases significantly, to the point that a maximum error $\epsilon = 1\%$ requires a number of trials more that five times higher with respect to $\epsilon = 5\%$ for the same confidence of 95%, as expected.

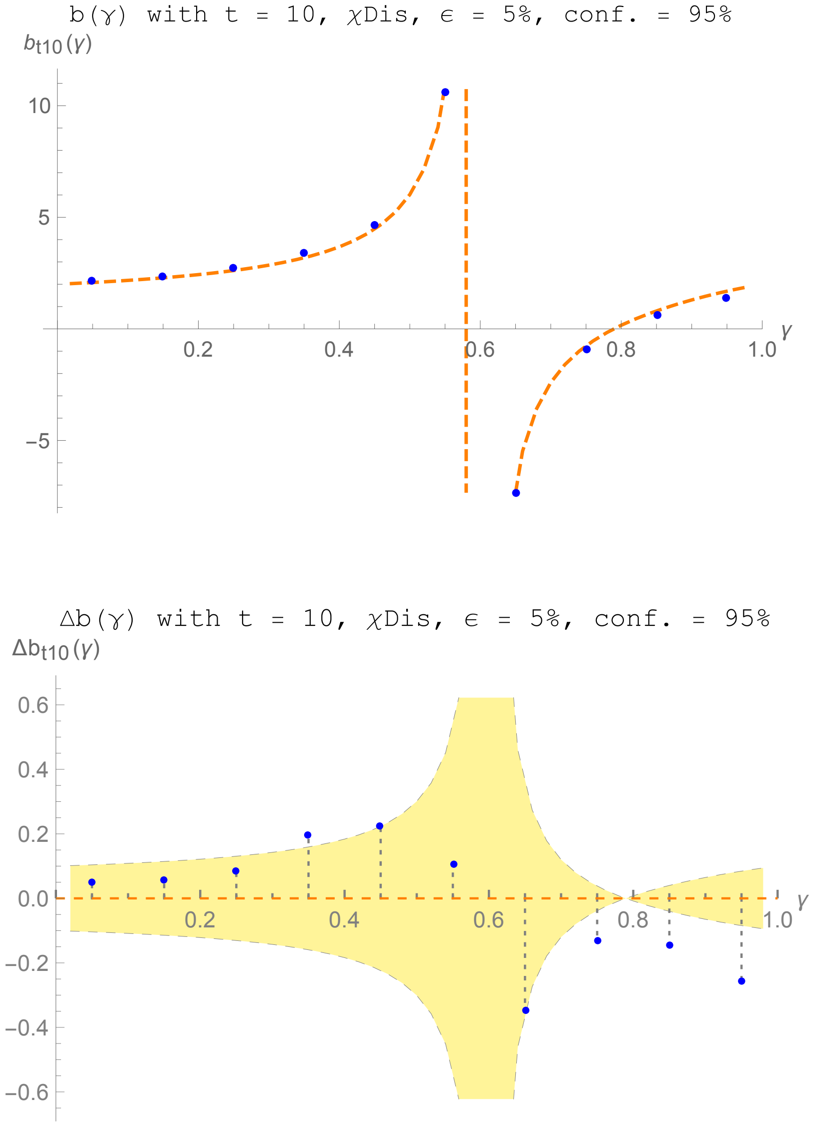


Figure 4.5: Theoretical values are in orange dashed line, simulation values in blue circles. The vertical line is the asymptoth $\gamma = \gamma_c$ where $b(\gamma)$ diverges. The light orange region in the graphic at the bottom is the error bar within which results should be found. This region has contours varying with γ, since $b(\gamma)$ is not constant.

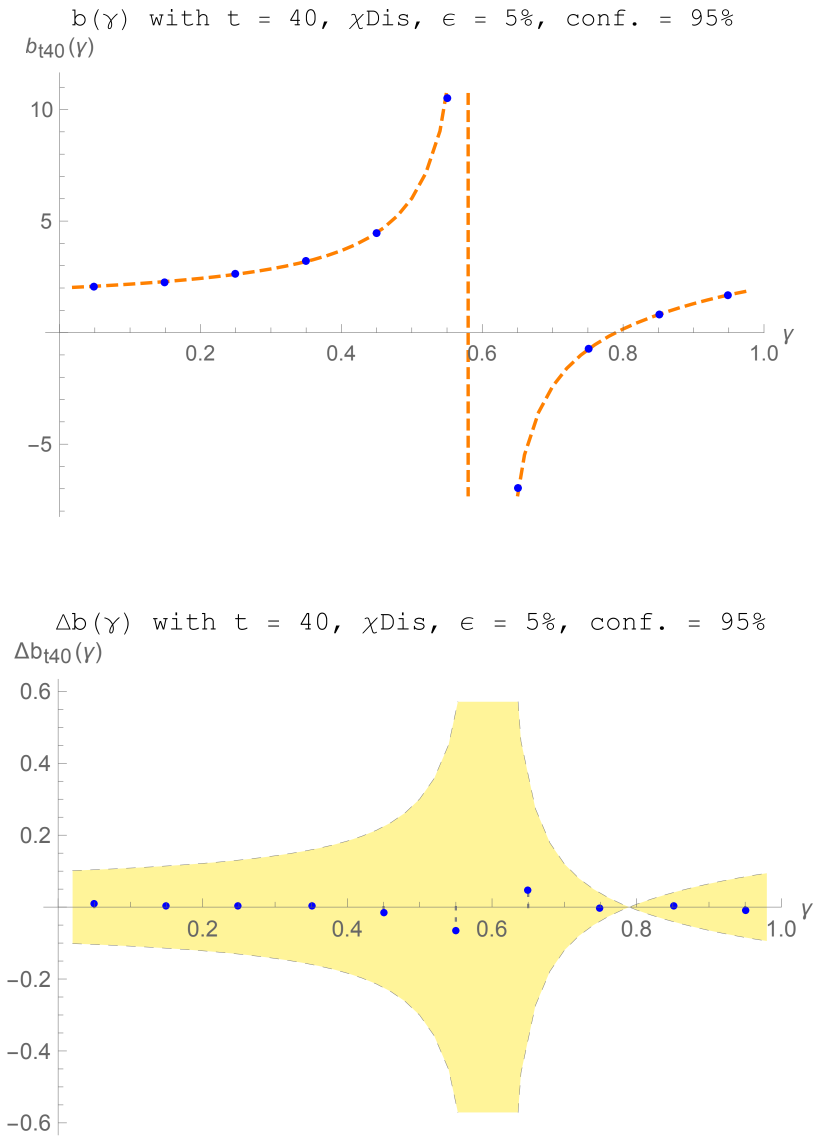


Figure 4.6: For larger t IS simulation method provides better estimates of $b(\gamma)$ (theoretical curve given in orange dashed line). Values obtained for $t = 40$ are not only inside the error bar, but they are almost on the theoretical curve, with appreciable errors only for γ close to γ_c (where $b(\gamma)$ diverges).

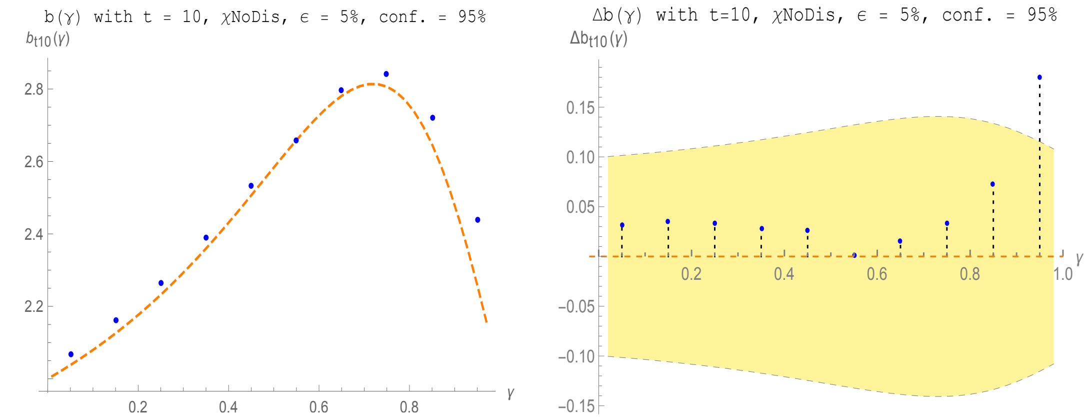


Figure 4.7: For $\chi = \{4, 1/2\}$ the error bar never goes to zero but is nevertheless close to it and, for $t = 10$, one point falls outside the maximum error region.

4.4 Importance sampling versus direct simulation method

Results displayed for $b(t)$ and $b(\gamma)$ verify that the importance sampling method in simulation has been a successful tool in estimating the power exponent of Taylor's law.

The major differences between the IS and the DS methods are:

- the IS simulation method selects, among all the admissible probability distributions, the most efficient. The DS method simply doesn't consider the efficiency problem.

- the IS simulation method, based on LDT, provides an estimate that complies with the precision requirements with a number of trials far smaller than the number necessary to the DS method. Evidently, by comparing the numbers k displayed in figures 4.3 and 4.4 to $R \simeq 10^{13}$ (given in 3.3.5), the IS method outclasses the second in terms of efficiency.

The most relevant consequence is the difference between the computation times.

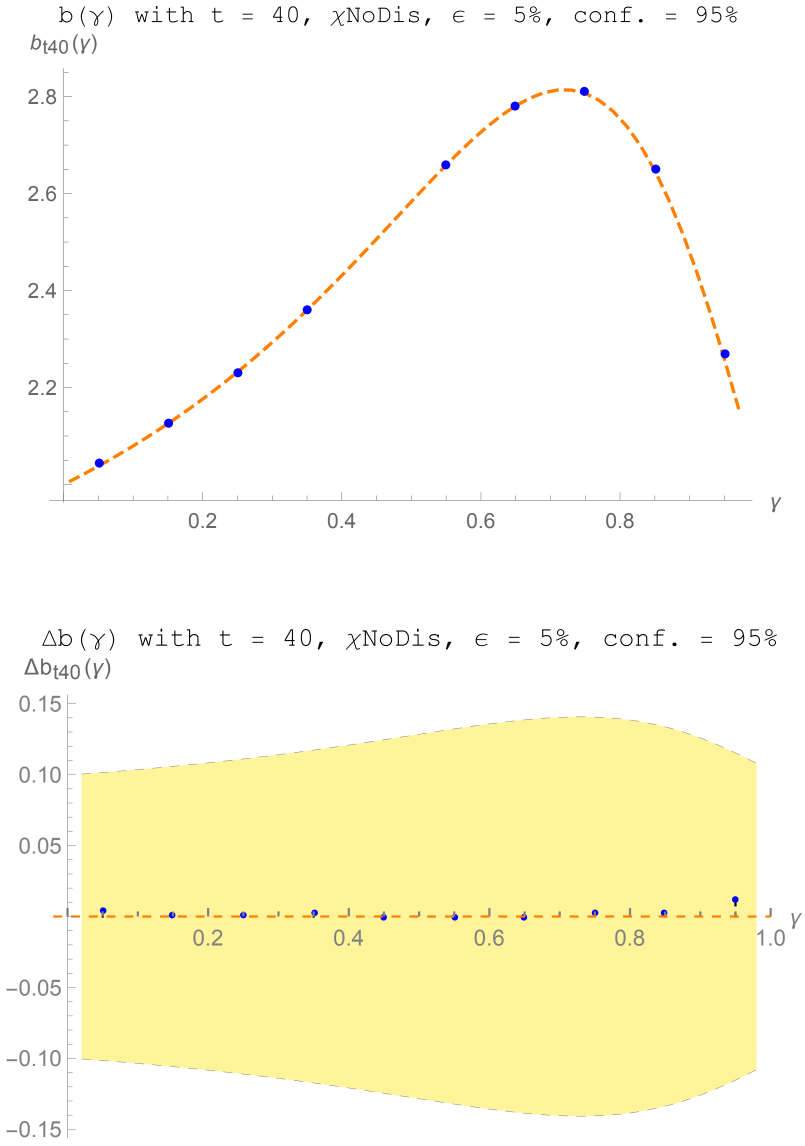


Figure 4.8: As for the discontinuity case, for $t = 40$ all the results provided by the IS simulation method fall within the error bar $[b - \epsilon b, b + \epsilon b]$.

A reduced number of trials implies a reduced computation time. In order to compare the two methods, taking as example the curve $b(t)$ in the discontinuity case for $\gamma = 0.55$, while the direct simulation method took some hours to generate the curve up to $t = 50$, but providing completely wrong estimates for every $t \geq 20$, the IS simulation method took about half an hour to generate the curve up to $t = 100$ with results satisfying the precision requirements $\epsilon = 5\%$, $\mathit{confidence} = 95\%$, that is, providing what had been asked (the computation time, or run-time, for this task is reported in figure 4.9). For stricter precision level the computation time increased significantly, but still being far and far smaller than the corresponding time needed by the DS method and still satisfying the required precision.

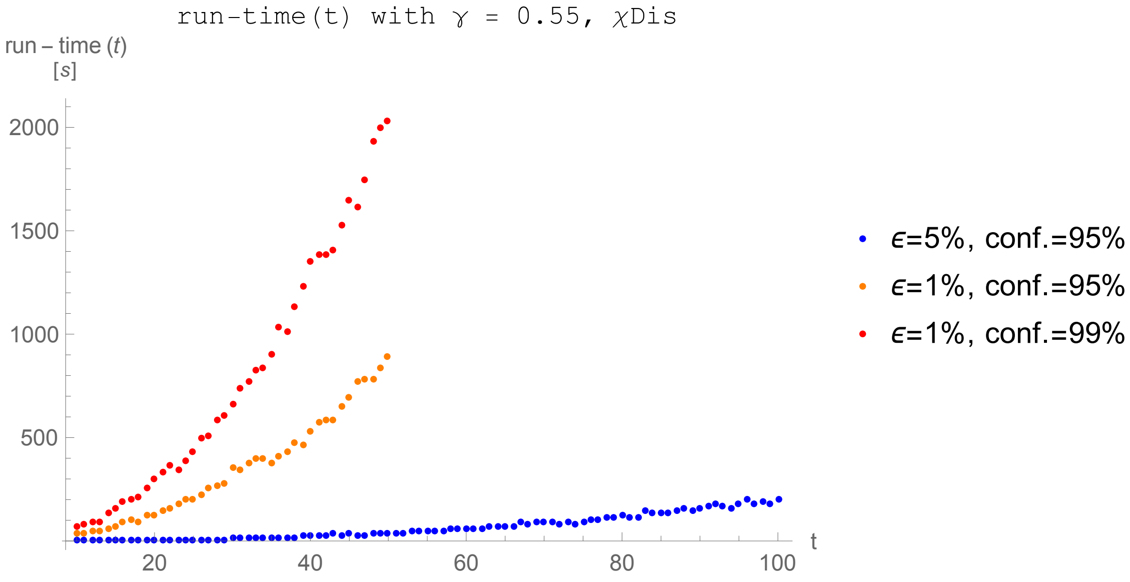


Figure 4.9: Different precision levels required different run-times, expressed in seconds [s]. Time steps t have no time unit, since they do not necessarily represent a physical time.

To compare numbers, in figure 4.3 typical k are at most of order 10^4 (to be summed to get $b(t)$), the order of magnitude of R is instead 10^{13}: the IS running time is 10^8 times smaller than the DS one. Having required the IS simulation half an hour, the DS method would have taken, at least, $1/2 \times 10^8$ *hours* … quite a longer time! Clearly, the DS simulation would never come to an end and the IS method is the only possible choice.

The purpose of designing a high-efficiency simulation that could detect the rare events hidden in the random multiplicative process studied here has been reached.

Conclusions

In this thesis we have devised an algorithm, based on importance sampling method, to perform simulations of random processes representing ecological systems. We have selected the class of Markovian multiplicative models for their relevant role in describing the emergence of Taylor's law.

Taylor's law for ecological systems states that variance and mean of population abundance are related by a power-law relationship. This empirical law has been verified in a wide variety of research fields to the point of suggesting that a context independent mechanism may be responsible for it. As it often happens in the study of complex systems - and those ecological are complex systems - computational methods turn out to be indispensable to infer key features of the ecological model of interest, in our case to determine the range of values of the power exponent appearing in Taylor's law.

In Chapters 2 and 3, proceeding from the previous work by Giometto et al. [17], we have shown that standard direct simulation methods (DS) provide incorrect estimates of the power exponent and that the possible cause for this is the incapacity of such sampling techniques in revealing the rare events entering the random process of the model under study. To tackle the problem of estimating statistics of rare events - events occurring with extremely low probabilities - a new sampling technique, the importance sampling (IS), has been studied in Chapter 2.

The name ”importance sampling” refers to a wide family of methods employed to decrease the number of independent trials in a simulation, while providing estimates in agreement with a required precision. In Chapters 1 and 2 we have analyzed rare event probabilities and selected the most efficient probability distribution to detect, in the simulation, the rare events of the Markovian multiplicative random process. The results of Chapter 4 demonstrate the IS simulation method outclasses the DS method for the former provides estimates in agreement with the required precision by performing a number of trials of the process far lesser than the number needed by the the latter. The most evident advantage of the IS method is the reduction of computation time: to reach the same precision the DS method would have needed a computation time about 10^8 times larger than that needed by the IS method! The results, then, have confirmed what was predicted in [17]: the anomalous behaviour of the power exponent in Taylor's law can be a consequence of ineffective sampling measurements.

In order to implement the IS technique for the multiplicative model, we have studied the basic theorems and principles of Large Deviation Theory and learned how to employ them. LDT has played a crucial and leading role: it served to describe rare event probabilities and to find the most efficient probability distribution (the *biasing distribution*) for the model.

The simulation technique exposed in this work may be profitably applied in any context where multiplicative random processes, Markovian also, are adopted.

The design of efficient simulations is still today a topic of active research, that founds itself on a rigorous mathematical theory (as seen, LDT and the principles of importance sampling). With this work we hope to have brought a useful example that shows the possible consequences of using a standard simulation method and the advantages of adopting an importance sampling method, advantages that could be of precious help in different areas of research.

Appendix A

Large Deviation Theory

Here proofs of the Gärtner-Ellis theorem and Varadhan's lemma are reported. As given in [20], the results come through lemmas. The demonstrations, quite technical, are an example of the typical way adopted in LDT to get results: to set upper and lower bounds, respectively for compact and closed sets, and then to extend the bounds to open sets.

A.1 Proof of Gärtner-Ellis theorem

Lemma A.1 (Exponential overbound)**.** *Given a $\mathbb{R}$-valued r.v. Z, then for $t \geq 0$ and $z \in \mathbb{R}$, $\mathbb{P}(Z > z) \leq \mathbb{E}[\exp(tZ)]\exp(-tz)$.*

Proof. With $f(z)$ the distribution function of Z, $\mathbb{P}(Z > z) = \int_z^\infty df(x)$. For all $t \geq 0$, for $x \in [z, \infty)$, $\exp[t(x-z)] \geq 1$, then

$$\begin{aligned}\mathbb{P}\big(Z > z\big) &\leq \int_z^\infty e^{t(x-z)} df(x) \\ &\leq \int_\infty^\infty e^{t(x-z)} df(x) \\ &= \mathbb{E}\big[e^{tZ}\big] e^{-tZ}.\end{aligned}$$

□

Lemma A.2. *For $a \in \mathbb{R}$, $\theta_1, \ldots, \theta_m \in \mathbb{R}^d$ it is defined the half-space*

$H_\theta(a) = \{x \in \mathbb{R}^d \colon \langle x,\, \theta\rangle - \phi(\theta) \le a\}$. *With* $C = \cap_{i=1}^m H_{\theta_i}(a)$, *then*

$$A1 \implies \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \notin C\big) \le -a.$$

For a subset $E \subset \mathbb{R}^d$ $\inf_{x\in E} I(x)$ will be replaced by the simpler notation $I(E)$ and L_a will denote the a-level set of $I(x)$: $L_a = \{x \colon I(x) \le a\}$, $a \in \mathbb{R}$.

Proof. (Gärtner-Ellis theorem, upper bound for compact sets.) A1 is assumed. With $\epsilon > 0$ fixed,

$$\begin{aligned} K &\subset L^c_{I(K)-\epsilon} \\ &= \{x \colon I(x) > I(K) - \epsilon\} \\ &= \{x \colon \sup_\theta[\langle \theta, x\rangle - \phi(\theta)] > I(K) - \epsilon\} \\ &= \cup_\theta \{x \colon \langle \theta, x\rangle - \phi(\theta) > I(K) - \epsilon\} \end{aligned}$$

and the last set is an open cover of a compact set. For this compact set Heine-Borel theorem states there exists a finite subcover, then there exists a finite sequence $\theta_1, \ldots, \theta_m$ with which

$$\begin{aligned} K &\subset \cup_{i=1}^m \{x \colon \langle \theta_i, x\rangle - \phi(\theta_i) > I(K) - \epsilon\} \\ &= \cup_{i=1}^m H^c_{\theta_i}(I(K) - \epsilon) \\ &= (\cap_{i=1}^m H^c_{\theta_i}(I(K) - \epsilon))^c. \end{aligned}$$

Lemma 2.16 now leads to

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in K\big) \le -[I(K) - \epsilon], \quad \forall\, \epsilon > 0$$

and the upper bound for the compact set K is obtained by taking $\epsilon \to 0$. □

Lemma A.3. $\forall\, a \in \mathbb{R}$,

$$A1,\ A2 \implies L_a \textit{ is compact.}$$

Lemma A.4. *With* $L^\delta_a = \{x \colon \|x - y\| < \delta \textit{ for some } y \in L_a\}$, *then*

$$A1,\ A2 \implies \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \notin L^\delta_a\big) \le -a.$$

Proof. (Gärtner-Ellis theorem, upper bound for closed sets.) A1 and A2 are assumed and $C \subset \mathbb{R}$ is closed. A positive a chosen, L_a^δ is bounded, thus $L_a^\delta \cap A$ is compact. Since $P(Y_n/n \in C) = P(Y_n/n \in C \cap L_a^\delta) + P(Y_n/n \in C \cap (L_a^\delta)^c)$, from

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in C \cap L_a^\delta\big) \leq -I(C)$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in C \cap (L_a^\delta)^c\big) \leq -a$$

it follows

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in C\big) \leq -\max\{I(C), a\}.$$

The last inequality holds true for every $a > 0$, then the upper bound for closed sets is proved by taking $\lim_{a\to\infty}$. □

Lemma A.5. *F_n denotes the distribution function of Y_n and a new sequence of r.v.s $Y_n^{(\theta)}$ is introduced with distribution function given by*

$$dF_n^{(\theta)}(x) = \frac{dF_n(x)\exp(\langle\theta, x\rangle)}{\int \exp(\langle\theta, x\rangle)dF_n(x)} = \frac{dF_n(x)\exp(\langle\theta, x\rangle)}{\exp(n\phi_n(\theta))}.$$

Defined $B_\delta(v) = \{x\colon \|x - v\| < \delta\}$, $\delta > 0$, if $v \in \nabla\phi(D_\phi)$ and θ_v is the solves $\nabla\phi(\theta) = v$, then

$$\lim_{n\to\infty} \mathbb{P}\big(\frac{Y_n^{(\theta_v)}}{n} \notin B_\delta(v)\big) = 0.$$

Lemma A.6. *If $v = \nabla\phi(\theta)$ for a $\theta = \theta_v$, then*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in B_\delta(v)\big) \geq -I(v) - \delta\|\theta_v\|.$$

Proof. (Gärtner-Ellis theorem, lower bound for open sets.) First an open $B \subset \nabla\phi(D_\phi)$ is considered. With $v \in B$ fixed but arbitrary, then

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in B\big) &\geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in B_\delta(v)\big) \\ &\geq -I(v) - \delta|\theta_v|,\end{aligned}$$

for a nonempty set of small $\delta > 0$. Since v is arbitrary the lower bound statement for open $B \subset \nabla\phi(D_\phi)$ is proved applying $\lim_{\delta\to 0}$.

If B is an arbitrary open set, the statement for the lower bound is again true if, for any $v \in B$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\frac{Y_n}{n} \in B\big) \geq -I(v). \tag{A.1}$$

Theory on convex functions states $\tilde{D}_I \subset \nabla\phi(\tilde{D}_\phi)$. Thus for $v \in B$ three cases may happen.

If $v \notin D_I$ (A.1) follows because $-I(v) = -\infty$.

If $v \in \tilde{D}_I$, the proof is as given for $B \subset \nabla\phi(D_\phi)$, since the condition on v is the same.

If $v \in D_I \setminus \tilde{D}_I$, for every ball B_v centred in v it is possible to choose a $v' \in B_v \cap \tilde{D}_I$ that satisfies $I(v') \leq I(v)$.

This shows that in minimizing the rate function over B the points in ∂D_I do not play any role and only the points in $\tilde{D}_I$ must be considered. Then, the lower bound for open sets is established. □

A.2 General theory

Cramér's and Gärtner-Ellis theorems define upper and lower bounds for a sequence of probabilities, regarding the sequence of sample averages S_n or random variables Y_n when the random variables meet certain conditions. The knowledge of these two theorems makes it possible to formulate the general theory of large deviations on the base of a more general definition of the rate function and by introducing the large deviation principle. The content of this section will serve only as a background knowledge for Chapter 3, where results are given in terms of the large deviation principle and Varadhan's lemma.

Definition (Polish space). A Polish space is a complete separable metric space.

Definition A.7. *Let $\mathcal{P}$ be a Polish space with distance $d\colon \mathcal{P} \times \mathcal{P} \to [0, \infty)$. A function $f\colon \mathcal{P} \to [-\infty, \infty]$ is called lower semi-continuous if any of the following equivalent conditions are satisfied:*

(i) $\liminf_{n\to\infty} f(x_n) \geq f(x) \quad \forall\,(x_n), x\colon x_n \to x \in \mathcal{P}$.

(ii) $\lim_{\epsilon\downarrow 0} \inf_{y\in B_\epsilon(x)} f(y) = f(x)$ *with* $B_\epsilon(x) = \{y \in \mathcal{P}\colon d(x,y) < \epsilon\}$,

(iii) *the level sets* $f^{-1}\big([-\infty, c]\big)$ *are closed for all* $c \in \mathbb{R}$.

Definition A.8 (Rate function)**.** *A function $I\colon \mathcal{P} \to [0, \infty]$ is called **rate function** if*

(RF1) $I \not\equiv \infty$.

(RF2) I is lower semi-continuous.

(RF3) I has compact level sets.

For a set $S \subset \mathcal{P}$ it will be be denoted $I(s) = \inf_{x\in S}(x)$, the closure of S by $cl(S)$, the interior by $int(S)$.

Definition A.9 (LARGE DEVIATION PRINCIPLE)**.** *A sequence of probability measures $(\mathbb{P}_n)$ on $\mathcal{P}$ is said to satisfy the **large deviation principle (LDP)** with rate n and rate function I if*

(LDP1) I is a rate function as given in Definition A.8.

(LDP2) $\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(C) \leq -I(C) \quad \forall\, C \subset \mathcal{P}$ *closed.*

(LDP3) $\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(A) \geq -I(A) \quad \forall\, A \subset \mathcal{P}$ *open.*

Theorem A.10 (Uniqueness of I)**.** *If a sequence of probabilities $(\mathbb{P}_n)$ satisfies a LDP, then its associated rate function is unique.*

Proof. Let I and J be two rate functions for $(\mathbb{P}_n)$ and $x \in \mathcal{P}$ be fixed. Defined in $\mathcal{P}$ the sequence $B_N = B_{1/N}(x)$ of open balls with radius $1/N$, $N \in \mathbb{N}$, then RF1 and RF2 imply

$$\begin{aligned}-I(x) &\overset{(a)}{\leq} -I(B_{N+1} \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(B_{N+1}) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n\big(cl(B_{N+1})\big) \leq -J\big(cl(B_{N+1})\big) \overset{(b)}{\leq} -J(B_N),\end{aligned}$$

where (a) holds because $x \in B_{N+1}$, (b) because $B_N \supset cl(B_{N+1})$. Taking $N \to \infty$, being J lower semi-continuous, it follows that $\lim_{N\to\infty} J(B_N) = J(x)$. Hence $I(x) \geq J(x)$. The opposite follows by interchanging I and J. □

A.2.1 Varadhan's lemma

The first important theorem in the general theory of large deviations is due to Varadhan. It is a generalization of the Laplace's method of integration.

Theorem A.11 (Varadhan's lemma)**.** *Given a sequence $(\mathbb{P}_n)$ satisfying the LDP on $\mathcal{P}$ with rate n and rate function I and $F\colon \mathcal{P} \to \mathbb{R}$ a continuous function bounded from above, then*

$$\lim \frac{1}{n} \log \int_{\mathcal{P}} e^{nF(x)} P_n(dx) = \sup_{x\in\mathcal{P}} \big[F(x) - I(x)\big].$$

where P_n denotes the distribution or the probability density function relating to $\mathbb{P}_n$.

In what follows, for two sequences of positive numbers (a_n) and (b_n) the symbol $\approx$ will denote *logarithmic equivalence*:

$$a_n \approx b_n \quad \overset{\text{def.}}{\Longleftrightarrow} \quad \lim_{n\to\infty} \frac{1}{n}(\log a_n - \log b_n) = 0. \tag{A.2}$$

By definition of logarithmic equivalence, for two sequences (c_n), (d_n) of positive numbers a *largest exponent dominates* principle holds true:

$$c_n + d_n \approx c_n \vee d_n, \tag{A.3}$$

that can be extended to a finite but arbitrary set of positive sequences.

Proof. (Varadhan's lemma.) It is defined the sequence of set functions (J_n):

$$J_n(S) = \int_S e^{nF(x)} P_n(dx), \quad \text{for } S \subset \mathcal{P} \text{ a Borel set,}$$

and

$$\alpha = \sup_{x \in \mathcal{P}} F(x), \quad \beta = \sup_{x \in \mathcal{P}} \big[F(x) - I(x)\big].$$

Suprema α and β satisfy $-\infty < b \leq a < \infty$, because $I \geq 0$ and F is continuous and bounded from above.

Upper bound. The space $\mathcal{P}$ is partitioned by F^{-1} by the following set definitions:

$D = F^{-1}([\beta, \alpha])$,

$D^N_k = F^{-1}([d^N_{k-1}, d^N_k]), \quad k = 1, \dots, N, \quad N \in \mathbb{N}$, with $d^N_k = \beta + \frac{k}{N}(\alpha - \beta)$ for $k = 0, 1, \dots, N$.

For sets D and D^N_k it holds that $D = \bigcup_{k=1}^N D^N_k$.

Since F is continuous all D^N_k are closed sets, then the LDP gives

$$\limsup_{n \to \infty} \frac{1}{n} \log P_n(D^N_k) \leq_I (D^N_k) \quad \forall k.$$

For every $k = 0, 1, \dots, N$ and every N, restriction of F on D^N_k is bounded: $F(x)\big|_{D^N_k} \leq d^N_k$, then relationship (A.3) gives

$$\limsup_{n \to \infty} \frac{1}{n} \log J_n(D) \leq \max_{1 \leq k \leq N} \big[d^N_k - I(D^N_k)\big].$$

Now the inequality

$$d^N_k \leq \inf_{x \in D^N_k} F(x) + \frac{1}{N}(\alpha - \beta)$$

can be used to get a better bound on the lim sup:

$$\begin{aligned}
\limsup_{n \to \infty} \frac{1}{n} \log J_n(D) &\leq \max_{1 \leq k \leq N} \Big[\inf_{x \in D^N_k} F(x) - \inf_{x \in D^N_k} I(x)\Big] + \frac{1}{N}(\alpha - \beta) \\
&\leq \max_{1 \leq k \leq N} \sup_{x \in D^N_k} [F(x) - I(x)] + \frac{1}{N}(\alpha - \beta) \\
&= \sup_{x \in C} [F(x) - I(x)] + \frac{1}{N}(\alpha - \beta) \\
&\leq \beta + \frac{1}{N}(\alpha - \beta).
\end{aligned}$$

The upper bound for $J_n(D)$ is then proved by taking the limit $N \to \infty$. For the sequence $(J_n(\mathcal{P} \setminus D))$ it holds

$$J_n(\mathcal{P} \setminus D) \leq \exp n\beta,$$

then, by applying (A.3), it follows

$$\limsup_{n\to\infty} \frac{1}{n} \log J_n(\mathcal{P}) \leq \beta$$

and the upper bound is proved.

Lower bound. Since F is continuous, for $x \in \mathcal{P}$ and $\delta > 0$ fixed but arbitrary the set

$$B_{x,\delta} = \{z \in \mathcal{P} \colon F(z) > F(x) - \delta\}$$

is open (and a neighbourhood of x). From *LDP3* the sequence (P_n) satisfies

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n(B_{x,\delta}) \geq -I(B_{x,\delta}).$$

Properties of I and the definition of $B_{x,\delta}$ imply that $I(B_{x,\delta}) \leq I(x)$ and this can be used in the last inequality to obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log J_n(B_{x,\delta}) \geq F(x) - \delta - I(x).$$

On $B_{x,\delta}$ the sequence (J_n) satisfies $J_n(\mathcal{P}) \geq J_n(B_{x,\delta})$. By letting $\delta \downarrow 0$ and by applying $\sup_{x\in\mathcal{P}}$ it follows:

$$\liminf_{n\to\infty} \frac{1}{n} \log J_n(\mathcal{P}) \geq \beta,$$

that proves the lower bound for the statement. □

Appendix B

Simulation Codes

The computation of $b(t)$ and $b(\gamma)$ requires the solution of transcendental equation (4.8), the simulation of Markov chains A_n for different transition probabilities, the calculus of linear interpolation coefficients. To carry out these steps the symbolic programming language Wolfram Mathematica® has been used.

B.1 Code for $b(\gamma)$

Here it is reported the code that implements the IS simulation as devised in Chapter 4. The exponent $b(\gamma)$ is computed for $t = 40$. Explanations are given in comment format `(* ... *)`.

Code for $b(\gamma)$ with fixed $t = 40$, $\epsilon = 5\%$, *confidence* $= 95\%$, discontinuity case ($\chi = \chi$Dis

```
r=2;     (* state r *)
s=1/4;   (* state s *)
N0=1;    (* N(t=0)  *)

\[Gamma]min=0.05;        (* minimum \[Gamma]       *)
\[Gamma]max=0.95;        (* maximum \[Gamma]       *)
d\[Gamma]=0.10;          (* increment of \[Gamma] *)

tmax=40;                 (* time for b *)

\[Epsilon]=0.05;(* maximum error on partial probabilities *)
quantil\[Epsilon]=1.960;
```

```
matrixT[\[Theta]_]:=    (* operator T_\[Theta] *)
{{(1-\[Gamma]) Exp[\[Theta]],\[Gamma]},
{\[Gamma] Exp[\[Theta]],1-\[Gamma]}};\[Lambda][\[Theta]_]:=
(1/2)((1-\[Gamma])(1+Exp[\[Theta]])+Sqrt[((1-\[Gamma])^2)
((1+Exp[\[Theta]])^2)+(4(-1+2 \[Gamma])Exp[\[Theta]])]);
Print[\[Lambda][\[Theta]]];

For[            (* loop on \[Gamma],transition probability *)
\[Gamma]=\[Gamma]min,\[Gamma]<=\[Gamma]max,
\[Gamma]=\[Gamma]+d\[Gamma],
Print[\[Gamma]];
time[\[Gamma]]=         (* computation time control *)
Timing[
For[                    (* loop on t, time steps *)
t=tmax/2,t<=tmax,t++,
If[                     (* case: even time *)
EvenQ[t]==True,
For[                    (* loop on n = t/2, ..., t *)
n=0,n<=(t/2)-1,n++,
root\[Theta]T=x/.Solve[ (* solution \[Theta]T *)
\[Lambda]'[x]==(1/2+n (1/t)) \[Lambda][x],x,Reals];
\[Theta]T=root\[Theta]T[[1]];
\[Lambda]\[Theta]T=     (* max. eigenvalue of T_\[Theta]  *)
Eigenvalues[matrixT[\[Theta]T]][[1]];
\[Lambda]\[Theta]Tmin=  (* min. eigenvalue of T_\[Theta]  *)
Eigenvalues[matrixT[\[Theta]T]][[2]];
\[Psi]\[Theta]Tr=       (* r component of max.eigenvector *)
Eigenvectors[matrixT[\[Theta]T]][[1,1]];
\[Psi]\[Theta]Ts=       (* s component of min.eigenvector *)
Eigenvectors[matrixT[\[Theta]T]][[1,2]];
\[Pi]q=         (* biasing transition probability matrix  *)
{{((1-\[Gamma])/\[Lambda]\[Theta]T) Exp[\[Theta]T],
(\[Gamma]/\[Lambda]\[Theta]T)
(\[Psi]\[Theta]Ts/\[Psi]\[Theta]Tr)},
{(\[Gamma]/\[Lambda]\[Theta]T)
(\[Psi]\[Theta]Tr/\[Psi]\[Theta]Ts) Exp[\[Theta]T],
(1-\[Gamma])/\[Lambda]\[Theta]T}};
\[Pi]qrr=\[Pi]q[[1,1]]; (* components: bias. transit. prob.*)
\[Pi]qrs=\[Pi]q[[1,2]];
\[Pi]qsr=\[Pi]q[[2,1]];
\[Pi]qss=\[Pi]q[[2,2]];
jsum\[Mu]=0;    (* initialization sum for expectation value*)
```

```
jsumVar=0;      (* initialization sum for variance *)
safetyk=Infinity;       (* safety number of trials *)
ksure=100;      (* initial safety number of trials *)
For[            (* loop on j, trials of Markov chain A_n *)
j=1,j<=ksure,j++,
\[Psi]0=0;      (* initializations \[Psi] for bias. prob.*)
\[Psi]t=0;
a0=RandomChoice[{r,s}];    (* initial state of chain A_n *)
If[a0==r,       (* initial \[Psi] in chain A_n *)
\[Psi]0=\[Psi]\[Theta]Tr,\[Psi]0=\[Psi]\[Theta]Ts];
state=a0;       (* start of j-th chain A_n *)
\[Delta]=0;     (* Kronecker delta *)
sum\[Delta]=0;  (* initial sum of Kronecker delta *)
indFunct=0;     (* indicator function *)
For[            (* loop on i, steps for chain A_n *)
i=1,i<=t,i++,
a=RandomReal[]; (* generation of states of A_n *)
If[state==r,If[a<\[Pi]qrr,state=r,state=s],
If[a<\[Pi]qsr,state=r,state=s]];
If[state==r,\[Delta]=1,\[Delta]=0];
sum\[Delta]=sum\[Delta]+\[Delta]];
If[sum\[Delta]>t (0.5+n (1/t)),indFunct=1,indFunct=0];
If[state==r,    (* final \[Psi] in expres. of bias. prob.*)
\[Psi]t=\[Psi]\[Theta]Tr,\[Psi]t=\[Psi]\[Theta]Ts];
jsum\[Mu]=                  (* sum for expectation value *)
jsum\[Mu]+indFunct (((\[Lambda]\[Theta]T^t)
\[Psi]0)/(Exp[\[Theta]T sum\[Delta]] \[Psi]t));
jsumVar=                             (* sum for variance *)
jsumVar+indFunct (((\[Lambda]\[Theta]T^t)
\[Psi]0)/(Exp[\[Theta]T sum\[Delta]] \[Psi]t))^2;
\[Mu]=(1/j) jsum\[Mu];              (* expectation value *)
var=(1/j) jsumVar;                           (* variance *)
If[\[Mu]!=0,          (* decision to stop the simulation *)
safetyk=((quantil\[Epsilon]/\[Epsilon])^2)
((var/(\[Mu]^2))-1),safetyk=Infinity];
ksure=Max[100,safetyk]];   (* update of number of trials *)
indexL=t ((1/2)+(n/t));
pLarger[indexL]=\[Mu]];  (* partial probability P(...>m) *)
pLarger[t]=0;   (* partial probability P(...>t) set to 0 *)

For[    (* computation of partial probabilities P(...=m) *)
indexP=t/2,indexP<=t,indexP++,      (* for t/2<=index<=t *)
If[
indexP==t/2,p[indexP]=1-2 pLarger[indexP],
p[indexP]=pLarger[indexP-1]-pLarger[indexP]]];
For[                          (* for 0<=index<=t/2 *)
indexPmirror=0,indexPmirror<=(t/2)-1,indexPmirror++,
p[indexPmirror]=p[t-indexPmirror]];

estEN[t]=    (* computation of expectation value E[N(t)] *)
N0 (s^t) Sum[((r/s)^m) p[m],{m,0,t}];
estEN2[t]=(N0^2) (s^(2 t)) Sum[((r/s)^(2 m)) p[m],{m,0,t}];
estVarN[t]=          (* computation of variance Var[N(t)] *)
estEN2[t]-((estEN[t])^2);
Print[t,"␣␣␣",Log[estEN[t]],"␣␣␣",Log[estVarN[t]]]
,
For[               (* repeat procedure for case: odd times *)
n=1/2,n<=(t/2)-1,n++,
root\[Theta]T=x/.Solve[
\[Lambda]'[x]==((1/2)+(n/t)) \[Lambda][x],x,Reals];
\[Theta]T=root\[Theta]T[[1]];
\[Lambda]\[Theta]T=Eigenvalues[matrixT[\[Theta]T]][[1]];
\[Lambda]\[Theta]Tmin=Eigenvalues[matrixT[\[Theta]T]][[2]];
\[Psi]\[Theta]Tr=Eigenvectors[matrixT[\[Theta]T]][[1,1]];
\[Psi]\[Theta]Ts=Eigenvectors[matrixT[\[Theta]T]][[1,2]];
\[Pi]q={{((1-\[Gamma])/\[Lambda]\[Theta]T) Exp[\[Theta]T],
(\[Gamma]/\[Lambda]\[Theta]T)
(\[Psi]\[Theta]Ts/\[Psi]\[Theta]Tr)},
{(\[Gamma]/\[Lambda]\[Theta]T)
(\[Psi]\[Theta]Tr/\[Psi]\[Theta]Ts) Exp[\[Theta]T],
(1-\[Gamma])/\[Lambda]\[Theta]T}};
\[Pi]qrr=\[Pi]q[[1,1]];
\[Pi]qrs=\[Pi]q[[1,2]];
\[Pi]qsr=\[Pi]q[[2,1]];
\[Pi]qss=\[Pi]q[[2,2]];
jsum\[Mu]=0;
jsumVar=0;
safetyk=Infinity;
ksure=100;

For[j=1,j<=ksure,j++,
\[Psi]0=0;
\[Psi]t=0;
a0=RandomChoice[{r,s}];
If[a0==r,\[Psi]0=\[Psi]\[Theta]Tr,\[Psi]0=\[Psi]\[Theta]Ts];
state=a0;
```

```
\[Delta]=0;
sum\[Delta]=0;
indFunct=0;

For[i=1,i<=t,i++,
a=RandomReal[];
If[state==r,If[a<\[Pi]qrr,state=r,state=s],
If[a<\[Pi]qsr,state=r,state=s]];
If[state==r,\[Delta]=1,\[Delta]=0];
sum\[Delta]=sum\[Delta]+\[Delta]];
If[sum\[Delta]>t(0.5+n (1/t)),indFunct=1,indFunct=0];
If[state==r,\[Psi]t=\[Psi]\[Theta]Tr,\[Psi]t=\[Psi]\[Theta]Ts];
jsum\[Mu]=
jsum\[Mu]+indFunct (((\[Lambda]\[Theta]T^t)
\[Psi]0)/(Exp[\[Theta]T sum\[Delta]] \[Psi]t));
jsumVar=
jsumVar+indFunct (((\[Lambda]\[Theta]T^t)
\[Psi]0)/(Exp[\[Theta]T sum\[Delta]] \[Psi]t))^2;
\[Mu]=(1/j) jsum\[Mu];
var=(1/j) jsumVar;
If[\[Mu]!=0,
safetyk=((quantil\[Epsilon]/\[Epsilon])^2)((var/(\[Mu]^2))-1),
safetyk=Infinity];
ksure=Max[100,safetyk]];
indexL=t((1/2)+(n/t));
pLarger[indexL]=\[Mu]];
pLarger[t]=0;

For[indexP=(t+1)/2,indexP<=t,indexP++,
If[indexP==(t+1)/2,p[indexP]=(1/2)-pLarger[indexP],
p[indexP]=pLarger[indexP-1]-pLarger[indexP]]];
For[indexPmirror=0,indexPmirror<=(t-1)/2,indexPmirror++,
p[indexPmirror]=p[t-indexPmirror]];

estEN[t]=N0 (s^t) Sum[((r/s)^m) p[m],{m,0,t}];
estEN2[t]=(N0^2) (s^(2 t)) Sum[((r/s)^(2 m)) p[m],{m,0,t}];
estVarN[t]=estEN2[t]-((estEN[t])^2);
Print[t,"␣␣␣",Log[estEN[t]],"␣␣␣",Log[estVarN[t]]]
]
]];
           (* table of LogE[N(t)] vs LogVar[N(t)] *)
bofgammaLogLogDist40e5[\[Gamma]]=
Table[{Log[estEN[z]],Log[estVarN[z]]},{z,tmax/2,tmax}];
```

```
fit=              (* linear interpolation of Log... vs Log... *)
Fit[bofgammaLogLogDist40e5[\[Gamma]],{1,x},x];
bofgammaDist40e5[\[Gamma]]=          (* b(\[Gamma]) for t=tmax *)
Coefficient[fit,x,1];
Print[\[Gamma],"␣␣␣␣␣",bofgammaDist40e5[\[Gamma]],
"␣␣␣␣",time[\[Gamma]][[1]]]];
(* print results: b(\[Gamma] and computation time *)
```

B.2 Code for $b(t)$

The code to evaluate $b(t)$ for fixed γ is simply obtained from the previous one by removing the for-loop

```
For[\[Gamma]=\[Gamma]min,
\[Gamma]<=\[Gamma]max,
\[Gamma]=\[Gamma]+d\[Gamma],
 ... ] .
```

Results of figures 4.1 and 4.3 have been obtained by setting `tmax = 100`, results of figures 4.2 and 4.4 with `tmax = 50` and by changing the precision requirements to `\[Epsilon] = 0.01` and `quantil\[Epsilon] = 2.5758`.

Bibliography

[1] R. M. Anderson et al. “Variability in the abundance of animal and plant species”. In: *Nature* 296 (1982), pp. 245–248.

[2] S. Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis.* Stochastic modelling and applied probability. New York: Springer Science+Business Media, 2007.

[3] J. A. Bucklew. *Introduction to Rare Event Simulation.* Springer series in statistics. New York: Springer-Verlag, 2004.

[4] J. A. Bucklew. *Large Deviations Techniques in Decision, Simulation and Estimation.* New York: John Wiley, 1990.

[5] J. E. Cohen. “Stochastic population dynamics in a Markovian environment implies Taylor’s power law of fluctuation scaling”. In: *Theor. Popul. Biology* 93 (2014), pp. 30–37.

[6] J. E. Cohen. “Taylor’s law and abrupt biotic change in a smoothly changing environment”. In: *Theor. Ecol.* 2014.7 (2013), pp. 77–86.

[7] J. E. Cohen. “Taylor’s power law of fluctuation scaling and the growth-rate theorem”. In: *Theor. Popul. Biology* 88 (2013), pp. 94–100.

[8] J. E. Cohen, M. Xu, and Schuster W. S. F. “Stochastic multiplicative population growth predicts and interprets Taylor’s power law of fluctuation scaling”. In: *Proc. R. Soc. B* 280 (2013).

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Berlin, Heidelberg: Springer, 2009.

[10] Z. Eisler, I. Bartos, and J. Kertész. “Fluctuation scaling in complex systems: Taylor’s law and beyond”. In: *Adv. Phys.* 57.85 (2008).

[11] R. S. Ellis. “An overview of the theory of large deviations and applications to statistical mechanics”. In: *Scand. Actuar. J.* 1 (1995), pp. 97–142.

[12] R. S. Ellis. *Entropy, Large Deviations and Statistical Mechanics.* New York: Springer, 1985.

[13] P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance.* Berlin, Heidelberg: Springer-Verlag, 1997.

[14] P. Embrechts and N. Veraverbeke. “Estimates for the probability of ruin with special emphasis on the possibility of large claims”. In: *Insurance: Mathematics and Economics* 1 (1982), pp. 55–72.

[15] X. Gabaix et al. “A theory of power-law distributions in financial market fluctuations”. In: *Nature* 423 (2003), pp. 267–270.

[16] C. Giardina et al. *Simulating rare events in dynamical processes.* research paper. Universities of Modena e Reggio emilia, Sorbonne (Paris), ESPCI, 2011.

[17] A. Giometto et al. “Sample and population exponents of generalized Taylor’s law”. In: *PNAS* 112.25 (2015), pp. 7755–7760.

[18] P. Glasserman. *Monte Carlo Methods in Financial Engineering.* New York: Springer-Verlag, 2004.

[19] G. Grimmett. “Large deviations in subadditive processes and first-passage percolation”. In: *Particle Systems, Random Media and Large Deviations.* Ed. by R. Durrett. Contemporary Mathematics 41. American Mathematical Society, 1984, pp. 175–194.

[20] F. den Hollander. *Large Deviations.* Ed. by Fields Institute Monographs. American Mathematical Society, 2000.

[21] J. Jiang et al. “Population age and initial density in a patchy environment affect the occurrence of abrupt transitions in a birth-and-death model of Taylor’s law”. In: *Ecological Modelling* 289 (2014), pp. 59–65.

[22] S. Juneja and P. Shahabuddin. “Rare-event simulation techniques: an introduction and recent advances”. In: (2005).

[23] M. Keeling and B. Grenfell. “Stochastic dynamics and a power law for measles variability”. In: *Phil. Trans, R. Soc. London B* (1999), pp. 769–776.

[24] W. S. Kendal. “A scale invariant clustering of genes on human chromosome 7”. In: *BMC Evolutionary Biology* (2004).

[25] W. S. Kendal and B. Jørgensen. “Taylor’s power law and fluctuation scaling explained by a central-limit-like convergence”. In: *Phys. Rev. E* 83.066115 (2011).

[26] A. M. Kilpatrick and A. R. Ives. “Species interactions can explain Taylor’s power law for ecological time series”. In: *Nature* 422 (2003), pp. 65–68.

[27] M. Loreti. *Teoria degli errori e fondamenti di statistica.* Padova: Decibel editrice, 1998.

[28] P. A. Marquet et al. “Scaling and power-laws in ecological systems”. In: *The Journal of Experimental Biology* 208 (2005), pp. 1749–1769.

[29] Y. Oono. “Large deviation and statistical physics”. In: *Progr. Theoret. Phys.* Suppl. 99 (1989), pp. 165–205.

[30] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis, Queues, Communications and Computing.* London: Chapman and Hall, 1995.

[31] A.-S. Sznitman. *Brownian Motion, Obstacles and Random Media.* Berlin: Springer, 1998.

[32] L. R. Taylor. “Aggregation, variance and the mean”. In: *Nature* 189.4766 (1961), pp. 732–735.

[33] L.R. Taylor and R. A. J. Taylor. “Aggregation, migration and population dynamics”. In: *Nature* 265 (1977), pp. 415–421.

[34] H. Touchette. “The large deviation approach to statistical mechanics”. In: *Physics Reports* 478 (2009), pp. 1–69.

[35] X. Xiao, K. J. Locey, and E. P. White. “A process-independent explanation for the general form of Taylor’s law”. In: *The American Naturalist* 186 (2015), E51–E60.

[36] M. Xu. *Taylor’s power law: before and after 50 years of scientific scrutiny.* Available at arxiv.org/abs/1505.02033v2, 2016.