Università degli Studi di Padova

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea Magistrale in Fisica

Extreme values statistics in Brownian motion and other stochastic processes

Relatore:
Prof. Marco Baiesi
Correlatore:
Prof. Satya N. Majumdar[†]

Laureanda: Irene Marzuoli

Anno Accademico 2014/2015

[†]Laboratoire de Physique Théorique e Modèles Statistiques, Université de Paris Sud, Orsay

Contents

Introduction **v**

1 Maximum of a sequence of random variables **1**
1.1 I.I.D. variables 1
1.2 Random walks 4

2 Records: statistics of the number **13**
2.1 I.I.D. variables 14
2.2 Random walks with continuous jump distribution 23
2.3 Lattice Random Walks 27
2.4 An application: climate records 30
2.4.1 Data processing 32
2.4.2 Statistics of records 35

3 Records: statistics of the ages **37**
3.1 I.I.D. variables 38
3.2 Random walks with continuous jump distribution 42
3.3 Lattice Random Walks 46

4 Multiple random walks **49**

5 Conclusions **57**

A Appendices **59**
A.1 Ornstein-Uhlenbeck process 59

Introduction

In everyday life imagination is often stroke by the concept of record: the strongest earthquake ever known, the fastest performance in an athletic race, the longest lasting permanence at the top of a ranking. Similarly, many aspects of practical life are influenced by these extreme events, from the hottest and coldest days in the year to the values of actions on the finance trade market. The interest in extreme events rises when the phenomenons show a component of casual behaviour. Most of the times the casualty just reflects our ignorance of the causes governing the process, but in many cases, and quite remarkably, it is possible to substitute the unknown process with a stochastic one, as long as we are satisfied to obtain the statistical laws to have access to the average behaviours. Nevertheless, this laws provide important and rich information on the phenomenon studied: extremes are usually sensitive to small changes in the underlying process because of their exceptional character, so even their average number or duration keep trace of many aspect of the physics behind them. Because of that, it is useful to study records starting from the simplest known stochastic processes, and then enrich the model according to the features of the natural process considered.

A short list of examples can clarify how different quantities can be thought as exceptional events, and so how very different problems can all be solved, or better understood, using extreme values statistics (see [22]-[26], [28]-[30], and [31] for a recent review). Before considering records in the proper sense, the first quantity to focus on is the maximum or the minimum of a process - in time or in space - which is studied for example in biology: an animal wandering in its habitat can be roughly represented by a Brownian walker [30], for which the maximum distance travelled from the origin define its domain within the collective space; again the actual fitness of the genotype of a certain specie has been modelled as the maximum of a "fitness function" in the space of genotypes, given by all the possible combinations of the four basis of the DNA [28]. In general, as the maximum can only increase with time, several processes characterized by a monotonic function can be studied with the techniques proper of the extreme values analysis. As example, we bring the total magnetization of a ferromagnetic material under an increasing field: it is known that in such conditions the flux of the field inside the sample increases by means of small steps and not in a continuous way, because of the sudden magnetization of small, but not infinitesimal, domains (Barkhausen effect). The function of the flux with respect to the external field thus shows a behaviour similar, for example, to the one of the maximal position of a free random walker. Without extending the analogy further, it is clear that we can use similar tools to analyse the two problems [29].

For processes as the ones listed above, the points in which the maximum takes a new value are the records. While the maximum statistics concerns the value of the variable at every step, the records statistics focusses on the number of the new-values jumps, as

well as the times at which they happen or the intervals between them. It is clear that the number of records in a stochastic process changes with the additional presence of a deterministic drift or the deformation in time of the distribution from which the variables are sampled: this was widely employed in the last years in climatology, both to study the global warming in alternative to the analysis of average temperatures, or, on the contrary, to build estimates of the temperature variability which are not sensible to mean trends [22]-[25]. As anticipated records are very sensible to the underlying process, but to enforce further the precision of these estimates it is often useful to use combinations of up, low records (the maxima and minima), forward and backward ones (the ones registered for the inverted-time sequence), opening questions about the joint statistics of these objects.

As a theory must confront with experimental data, it is non trivial to ask whether the rounding effects and the presence of noise in measures influence the count of records of a sample, adding a non-physical part which changes their statistics. Indeed, it is possible to analyse these effects, analytically or through simulations [16],[17], suggesting that records can be used also in evaluating the fraction of noise present in a measure. At the same time the previous applications of records theory in climatology must consider this rounding effects in analysing the data.

In this work, we propose to draw the extreme values and records statistics landscape for two models of stochastic processes: sequences of independent and identically distributed (i.i.d.) variables, where the correlations between variables are zero; and Markov chains, i.e. the discrete time version of a Brownian motion, both with continuous jumps and on a lattice, which represents an example of strongly correlated variables. We analyse first the statistics of the maximum in the three cases mentioned, looking at the random chain case under different external potentials, then we study the statistics of records: in particular we want to extend the treatment to the joint behaviour of forward and backward records, as the forward behaviour alone has already been studied. In both problems we focus also on the universality of the results as it turns out that many conclusions are independent from the details of the random process. For example, a chain of i.i.d. variables presents a maximum behaviour which strongly depends on the distribution of the entries, but it is possible to group the parental distributions in three classes of universality according to their tails and find, for each of them, a universal scaling function which summarizes the behaviour of the maximum, up to some specific numeric prefactors. Analogously, in the random walk case, all the symmetric distributions are divided in two classes only, according to their variance, which can be finite or not; then, though in an approximate way, all the results are the same within a class. The records problem then broadens the universality: as in their count there is no trace of the height reached by the walker, records turns out to be universal for every distribution in the i.i.d. case (no matter its variance or symmetry), and for every symmetric distribution in the random walk one.

The other interesting aspect of the problem, and the new contribution of this work, is the comparison between i.i.d. and random walks themselves through the joint study of forward and backward records. The two kinds of records are strongly correlated in both cases, but the correlation approaches a constant in the i.i.d. case, and increase with time for random walks. An analysis of the ages of records confirm a stronger interdependence in the random walk case. A quite curious point arisen in this study is a parallelism between forward-backward and upper-lower records in the i.i.d. case, where a variable is a lower records when it is smaller than all the previous ones. This statistical equality does not hold in the random walk case, and it has not been explained yet. In fact, in

the independent case, almost all the computations can be done in a straightforward way, enabling to explore new sides of the problem: the results are often exact but the price to pay is that it is sometimes difficult to give a physical meaning to the quantities which appear in the different passages, and in the example mentioned - the analogy between forward-backward and upper-lower records - the computations does not provide a good explanation for the parallelism. For random walks instead, as the correlation makes the problem very cumbersome, it is customary to simplify it invoking the Markov property of independence between different intervals: this, joint with the Sparre-Andersen theorem explained in section 2.2, makes the problem even easier than the i.i.d. one and sometimes of more immediate understanding. However, we will be often forced to give just asymptotic results where an analytic computation fails: this analysis will be done mainly in the space of generating functions and Laplace transforms, which are non-symmetric transforms and reflect well the asymmetry of the problems studied.

To complete the work we propose a small analysis of a set of real data: they are the European temperatures in the last century, as climatology is one of the field where records are most analysed. To compare data with the simple model studied, we must perform several simplifications, but in the end such analysis provide a further test about the statistical dependence of the data, as we will see that the temperatures of a given day of the year turn out to be independent. Finally we suggest correlated problems: a drift, for example, influence the number of records, as well as a rounding process, which, again, can be performed in several ways, changing the statistics of records. All this problems strongly depend on the variable distribution and correlation. On the contrary, a problem of multiple random walks depends only on the variance of the distribution: we choose to analyse it and extend its study to find the number of backward records. The problem is interesting by itself but it also shows how we can gather all the previous working tools when the the number of records and the level of the variables are important at the same time: here in fact we consider each trajectory with its probability and then apply the analytical results of Sparre-Andersen theorem and of the maximum theory at once.

Chapter 1

Maximum of a sequence of random variables

Consider a set of variables $\{x_i\}$ which represent the relevant quantity of the physical system studied. When the index i represents the time, it is natural to think of them as ordered for increasing i. For a given length of the sequence, i.e. $0 < i \leq N$ (overall time $t = N\Delta t$), the maximum value is the variable $x_j = M$ so that $\{x_j \geq x_i, \forall\, i, j \in]0, N]\}$. If the system can be described by a stochastic process, the variables x_i are sampled from a parent distribution $p(x_i)$ which is, in principle, different for every i. The knowledge of the full joint distribution $p(\{x_i\}) = p(x_1, ..., x_N)$ is needed to compute the distribution of the maximum, but its expression is hardly achievable when the variables are correlated, which is the most general case. On the contrary, if the variables are independent and identically distributed (i.i.d.), they are sampled from the same parent distribution and their joint probability factorizes: $p(\{x_i\}) = p(x_1) \cdots p(x_N)$. We study therefore this case first and then extend the problem to weak correlated variables and finally to strong correlated ones. We will see that while in the uncorrelated case the maximum behaviour shows some universal features, correlated variables strongly depend on the specific model they follow and it is not possible to find common laws.

1.1 I.I.D. variables

We know that for a set of N i.i.d. variables in the large N limit the Central Limit Theorem holds, which states that the distribution of the mean value $\overline{x}$ approaches a Gaussian:

$$P(\overline{x}, N) \xrightarrow{N\to\infty} \sqrt{\frac{N}{2\pi\sigma^2}}\, e^{-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2}\,; \tag{1.1}$$

where μ and σ^2 are the mean and the variance of the parent distribution $p(x)$. The relevant point is the universality with respect to every other particular of $p(x)$. It is interesting to verify if a similar universality holds also for the maximum of the process M. To study its distribution is useful to define the cumulative probability $Q_N(x)$:

$$Q_N(x) = \mathbb{P}[M < x, N] = \mathbb{P}[x_1, ..., x_N \leq x] = \int_{-\infty}^{x} dx_1 \cdots \int_{-\infty}^{x} dx_N P(x_1, ..., x_N)\,. \tag{1.2}$$

As in this case the probability factorizes, the cumulative function simplifies:

$$Q_N(x) = \left[\int_{-\infty}^{x} dy\, p(y)\right]^N = \left[1 - \int_{x}^{\infty} dy\, p(y)\right]^N . \tag{1.3}$$

The second formula shows explicitly the dependence of $Q_N(x)$ from the (upper) tail of the parent distribution, which becomes more and more relevant in the large N limit, as we expect that the maximum itself grows with the length of the sequence. We can group almost all the distributions $p(x)$ in three classes according to the behaviour of their tails and it turns out that for each class a universal form of the cumulative probability emerges in the scaling limit when both N and x are large:

$$Q_N(x) \xrightarrow[z=(x-a_N)/b_N]{x,\, N\to\infty} F\left(\frac{x-a_N}{b_N}\right); \tag{1.4}$$

where the limit is performed at fixed z. The factors a_N, b_N defining the scaling limit are specific of every $p(x)$ and bears the N dependence, but once found them, $F(z)$ assumes three form only (see [2] for a complete treatment, [5] for a recent review). This is known as **Gnedenko's classical law of extremes** [3].

To compute a_N and b_N is useful to notice that in the large N limit the expression of the cumulative distribution can be approximated as:

$$Q_N(x) \xrightarrow{x,N\to\infty} e^{-N\int_x^\infty dy\, p(y)} . \tag{1.5}$$

It is peaked approximately around the value a_N defined by:

$$\int_{a_N}^{\infty} dy\, p(y) = \frac{1}{N}; \tag{1.6}$$

indeed, if we have N variables, which take values in the domain of x according to $p(x)$, a_N is the value of x such that only one out of them is bigger than a_N (on average), so it delimits the domain of the maximum and it is thus taken as its typical value. The determination of b_N is more involving; it is defined as:

$$b_N = \frac{\int_{a_N}^{\infty} dy\, (y-a_N)\, p(y)}{\int_{a_N}^{\infty} dy\, p(y)}; \tag{1.7}$$

and it can be interpreted as the average distance of the maximum from a_N, conditioned by the fact that only one variable (the maximum itself) can stay over a_N.

Power law tail Consider a parent distribution which goes like $p(x) \sim x^{-(\alpha+1)}$ for large x, with $\alpha > 0$. Computing (1.5) in the power law case (integral of the tail $\sim x^{-\alpha}$) gives immediately the scaling factors $a_N = 0$ and $b_N = N^{1/\alpha}$, and the scaling function $F_1(z)$ is:

$$F_1(z) = e^{-z^{-\alpha}}\theta(z) . \tag{1.8}$$

This is the **Fréchet** distribution. The probability of the rescaled variable z ($z \geq 0$) is obtained differentiating $F_1(z)$:

$$f_1(z) = \frac{\alpha}{z^{\alpha+1}}\, e^{-z^{-\alpha}}\theta(z) . \tag{1.9}$$

Faster than power law tail Consider a parent distribution with tails which decay faster than any power law but with unbounded domain: $p(x) \sim e^{-x^\delta}$, $\delta > 0$. Approximating the cumulative probability as above gives:

$$Q_N(x) \sim e^{-e^{-\left(\frac{1}{\delta}\, x^\delta - \ln N\right)}}; \tag{1.10}$$

which is peaked in $a_N = (\delta \ln N)^{1/\delta}$. The approximation $x^\delta \sim x\,(a_N)^{\delta-1}$ in $Q_N(x)$ suggests that the behaviour of the fluctuations is $b_N \sim (\ln N)^{1/\delta-1}$; however the precise computation of b_N must be done case by case by formula (1.7), and it generally requires to express a_N to the second order. The corresponding scaling function is:

$$F_2(z) = e^{-e^{-z}}; \tag{1.11}$$

$$f_2(z) = e^{-z-e^{-z}}; \tag{1.12}$$

for every $z \in \mathbb{R}$. This is the **Gumbel** distribution. The exponential and Gaussian tails distributions belong to this class: for the first is immediate to see that $a_N = \ln N$ and $b_N = 1$, while for a Gaussian with variance σ we have $a_N = \sigma\sqrt{2 \ln N}$ (at the first order) and $b_N = \frac{\sigma}{\sqrt{2 \ln N}}$.

Upper bounded support Consider eventually distributions $p(x)$ which are zero for $x > a$, $a \in \mathbb{R}$, and with upper tail $p(x) \xrightarrow{x \to a} (a - x)^{\beta-1}$, $\beta > 0$. The scaling limit is given by $a_N = a$, $b_N = \beta^{1/\beta}$, with scaling function:

$$F_3(z) = e^{-(-z)^\beta} \quad \text{for } z \leq 0\,; \tag{1.13}$$

$$f_3(z) = e^{-z-e^{-z}} \quad \text{for } z \leq 0\,; \tag{1.14}$$

while $F_3(z) = 1$ for $z > 0$ (which is $x > a$). This is the **Weibull** distribution.

So far, we have detected some universal features of the maximum distribution. The limiting distributions are obtained with a strong use of the large N approximation; however simulations show that the convergence is quite slow, therefore in the experimental context is hard to see them.

Before considering strongly correlated variables, for which the joint probability does not factorize and must be computed for each case, consider a set of weakly correlated variables, i.e. the correlation function decays fast, for example exponentially, within a typical length (or time) η which is much smaller than the sample size (or observed time) N:

$$C_{i,j} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \sim e^{-|i-j|/\eta}\,. \tag{1.15}$$

If $|i - j| \gg \eta$ the two variables are with good approximation uncorrelated. Under the condition $N \gg \eta$, we can consider the sample as built of N/η uncorrelated blocks of size η, each of them represents a set of strongly correlated variables ($C(i, j)$ is relevant on all the block size). Defining the local maximum for each block y_i, the global maximum of the full sample is the biggest among the y_i. As the blocks are uncorrelated, once worked out the distribution of y_i ($q(y_i)$), the global maximum approaches one of the three scaling functions $F_I(z)$ according to the tails of $q(y_i)$ when the number of blocks N/η increases to infinity. Moreover, in several cases a form of the correlation as in (1.15) implies a similar behaviour for the tails of $q(y_i)$, suggesting a Gumbel scaling for the global maximum. At this stage, the only problem left to solve is the maximum distribution for strongly correlated systems, task we must perform case by case.

1.2 Random walks

There are several ways in which the variables of a sequence can be linked and correlated together. The only model of strongly correlated system we are going to analyse is random walks. In this case the variables $\{x_i\}$ represent the position in one dimension of a walker at discrete time steps, starting from a fixed x_0. The walker increases its position according to the equation:

$$\frac{\Delta x_i}{\Delta t} = -\mu \frac{dU(x_i)}{dx_i} + \xi_i; \tag{1.16}$$

where $U(x)$ is a deterministic potential, μ the mobility of the walker, and ξ_i is the realization at step i of a stochastic process. We choose it with zero mean and delta correlated:

$$\langle \xi_i \xi_j \rangle = 2D\, \delta(i-j); \tag{1.17}$$

where D is the diffusion constant. Therefore $2D$ corresponds to the variance of the distribution $p(\xi)$ of ξ which can be, in principle, finite or not. We will see that every other particular of $p(\xi)$ is unimportant in the case of finite variance, while for infinite variance (but still delta correlated process) the behaviour of the tail of the distribution is needed to reach a result. The discrete time formalism can now be replaced by the continuous one in the hypothesis that the time step is very small with respect to the overall time of the sequence.

Finite variance processes: Backward Fokker-Planck approach As above, we define the cumulative probability that the maximum stays under the level $x = M$ ($M > 0$) after N time steps, starting from x_0, as $Q_N(M, x_0)$. An equation for Q is obtained considering a backward approach: a sequence of $N+1$ steps can be split in two parts, the first one, which leads from x_0 to x_1, and the remaining N. Therefore:

$$Q_{N+1}(M, x_0) = \int_{-\infty}^{\infty} d(x_1 - x_0)\, Q_N(M, x_1)\, p(x_1 - x_0)\,; \tag{1.18}$$

where $p(\Delta x)$ is the distribution of the increments. The integration is performed over all the possible increments as, if a jump brings the walker in a forbidden region ($x_1 \geq M$), $Q_N(M, x_1)$ is identically zero, ensuring that only the realizations survived after the first step contributes to $Q_N(M, x_0)$. Equation (1.18) is solvable exactly for some particular $p(\Delta x)$, however in the limit of small time steps it tends to an approximate equation in terms of the continuous time t which is easier to treat and depend only on the variance of the distribution from which the ξ_i are sampled, if it exists.

The small time step limit means that $N \gg 1$ so that the overall length of the sequence $N\Delta t$ is much bigger that Δt itself. The notation is updated to $Q(M, x_0|t)$, $x(t)$ and $\xi(t)$, as now they are all continuous time processes. As Δt is small, the jump between x_0 and x_1 is small too and it is possible to expand the cumulative function in x_1 around x_0:

$$\begin{aligned} Q(M, x_1|t) &= Q(M, x_0|t) + \left(-\mu \frac{dU(x_0)}{dx_0} + \xi(0)\right) \frac{\partial Q(M, x_0|t)}{\partial x_0}\, \Delta t \\ &\quad + \xi^2(0)\, \frac{\partial^2 Q(M, x_0|t)}{\partial x_0^2}\, \frac{\Delta t^2}{2} + o(\Delta t^2). \end{aligned} \tag{1.19}$$

We notice that the right hand side of equation (1.18) is the average over the possible first steps, i.e. over the possible $\xi(0)$, of (1.19). Using the properties of the stochastic process, where $\delta(t)$ is interpreted as $1/\Delta t$, equation (1.18) gives after some rearrangements:

$$\frac{\partial Q}{\partial t} = D\,\frac{\partial^2 Q}{\partial x_0^2} - \mu\,\frac{dU(x)}{dx}\,\frac{\partial Q}{\partial x_0}\,. \tag{1.20}$$

where the argument $(M, x_0|t)$ of the function is understood. This is the **backward Fokker-Planck equation**. The initial and boundary conditions are:

$$Q(M, x_0|t=0) = \theta(M - x_0);\ Q(M, x_0 = M|t) = 0\,;\ Q(M, x_0 \to -\infty|t) = 1\,. \tag{1.21}$$

It is immediate to see that the problem can be extended to the cumulative joint probability of the maximum and minimum imposing that the walker must stay between $-m$ and M, thus replacing the last condition with $Q((M,m), x_0 = -m|t) = 0$. As the probability $P(M, x_0|t)$ that the maximum stays between M and $M + dM$ is the derivative with respect to M of $Q(M, x_0|t)$, the backward Fokker-Planck equation holds also for it.

To transform the partial differential equation into an ordinary one, we work in the space of the **Laplace transform**. We remember its definition as well as the one of its discrete counterpart, the generating function:

$$\mathcal{L}[f(t)] := \tilde{f}(s) := \int_0^\infty f(t)\,e^{-st}dt\,; \quad \tilde{F}(z) := \sum_{N=0}^{\infty} F_N\,z^N. \tag{1.22}$$

Notice that the large time behaviour of $f(t)$ corresponds to the small s one for $\tilde{f}(s)$ as the factor e^{-st} strongly suppresses the integrand for large t except when s is small (roughly $1/t$). Similarly, $z \to 1$ correspond to $N \to \infty$, for which the generating function approaches the Laplace transform itself if $z = e^{-s}$. As the $s \to 0$ limit is often simpler than in the large t one, the transform helps to perform the asymptotic analysis, even if the result must then be inverted.

Taking the Laplace transform of the Fokker-Planck, the time derivative of Q is solved integrating by parts applying the initial condition, so that it becomes an ordinary differential equation of the second order:

$$s\tilde{Q} - 1 = D\,\frac{\partial^2 \tilde{Q}}{\partial x_0^2} - \mu\,\frac{dU(x_0)}{dx_0}\,\frac{\partial \tilde{Q}}{\partial x_0}\,; \tag{1.23}$$

with the argument of $\tilde{Q}$ being $(M, x_0|s)$. The new boundary conditions are:

$$\tilde{Q}(M, x_0 = M|s) = 0\,;\ \tilde{Q}(M, x_0 \to -\infty|s) = \frac{1}{s}\,. \tag{1.24}$$

Equation (1.23) is solved case by case according to the deterministic potential and it is rather simple in several cases but the inversion of the Laplace transform can be troublesome. However, as anticipated, the asymptotic analysis for large t is usually affordable and sometimes the exact expressions for the transform of the average maximum or its variance are accessible in the s space and then trivially inverted, as they bear a simple dependence s^{-n} which leads to $t^{n-1}/\Gamma(n)$.

In the following we present some examples in which, changing the potential, the time dependence of the maximum varies widely.

Simple Random Walk In the case in which the potential $U(x)$ is zero the solution of (1.23) gives:

$$\tilde{Q}(M, x_0|s) = \frac{1}{s} - \frac{1}{s}\, e^{-(M-x_0)\sqrt{s/D}} \quad \text{for } x_0 < M; \tag{1.25}$$

which assumes an undetermined form in the small s limit. It is exactly invertible, giving for the cumulative and the probability of the maximum:

$$Q(M, x_0|t) = 1 - erfc\left(\frac{M-x_0}{2\sqrt{Dt}}\right); \tag{1.26}$$

$$P(M, x_0|t) = \frac{1}{\sqrt{\pi Dt}} e^{-\frac{1}{4Dt}(M-x_0)^2}. \tag{1.27}$$

This leads, when $x_0 = 0$, to $\langle M(t)\rangle = \frac{2\sqrt{Dt}}{\sqrt{\pi}}$ and $\langle M^2(t)\rangle = 2Dt$ (the same as for the position $x(t)$ of the walker), so both the average maximum and its fluctuations go like the square root of time.

Solving with different conditions to obtain the joint probability of the maximum M and the minimum $(-m)$ with $x_0 = 0$, gives the cumulative function:

$$\tilde{Q}((M,m),0|s) = \frac{1}{s} - \frac{1}{s}\,\frac{\sinh\left(\sqrt{\frac{s}{D}}\,M\right) + \sinh\left(\sqrt{\frac{s}{D}}\,m\right)}{\sinh\left(\sqrt{\frac{s}{D}}(M+m)\right)}; \tag{1.28}$$

which tends to (1.25) for $m \to \infty$. The exact inversion exists but is quite complicated. It points out the $\sqrt{t}$ dependence of both M and $-m$. It is possible to compute exactly the correlation between maximum and minimum which gives $\langle M(-m)\rangle = (1 - 2\ln 2)\, t$, independent from the diffusion coefficient.

Random Walk with constant drift The case in which $U(x) = -ax$, with a constant, is intuitively simple as we expect that for large t the maximum increases linearly with time if the induced drift is positive (positive a), or approaches a constant if it is negative (negative a). Calling $c = a\mu$, the solution of the Fokker-Planck equation in the s space is:

$$\tilde{Q}(M, x_0|s) = \frac{1}{s} - \frac{1}{s}\, e^{-\frac{M-x_0}{2D}\left(\sqrt{c^2+4Ds} - c\right)} \quad \text{for } x_0 < M. \tag{1.29}$$

When c is negative, in the small s limit the parenthesis does not vanish, and the leading divergence is $1/s$. Inverting it is trivial and gives:

$$Q(M, x_0|t) = 1 - e^{-(M-x_0)\frac{|c|}{D}}; \tag{1.30}$$

$$P(M, x_0|t) = \frac{|c|}{D} e^{-(M-x_0)\frac{|c|}{D}}; \tag{1.31}$$

which is constant in time as expected: the constant maximum is the typical scale length of the problem $D/|c|$. Looking for the correct dependence also in the transient period we must solve the full problem. Moreover a full analytic treatment is required when c is positive as $\tilde{Q}$ is undetermined for s going to zero.

Besides the trivial term $1/s$, which inverted gives 1, the second part of expression (1.29) is the product of $1/s$ and an exponential function of the shifted variable $s + c^2/4D$. Using the convolution and translation theorems of the Laplace transform we have:

$$Q(M, x_0|t) = 1 - \int_0^t du\; e^{-\frac{(M-x_0-cu)^2}{4Du}}\; \frac{M-x_0}{2\sqrt{\pi D u^3}}\,. \tag{1.32}$$

Fixing $x_0 = 0$ to simplify the computations, the integral is solvable exactly considering the change of variables y_i as follows:

$$y_{1,2} = \pm \frac{M \pm cu}{\sqrt{4Du}}; \qquad dy_{1,2} = \mp \frac{M \mp cu}{4\sqrt{Du^3}}. \tag{1.33}$$

The fraction in the integrand can be split to obtain $-dy_1 + dy_2$, giving:

$$Q(M,0|t) = \frac{1}{2}\left[erfc\left(-\frac{M-ct}{\sqrt{4Dt}}\right) - e^{M\frac{c}{D}} erfc\left(\frac{M+ct}{\sqrt{4Dt}}\right)\right]; \tag{1.34}$$

$$P(M,0|t) = \frac{1}{\sqrt{\pi Dt}} e^{-\frac{(M-ct)^2}{4Dt}} - \frac{c}{2D}\, e^{M\frac{c}{D}} erfc\left(\frac{M+ct}{\sqrt{4Dt}}\right); \tag{1.35}$$

$$\langle M(t)\rangle = \frac{ct}{2}\left(1 + erf\left(\frac{c\sqrt{t}}{2\sqrt{D}}\right)\right) + \sqrt{\frac{Dt}{\pi}} e^{-\frac{2D}{c^2 t}} + \frac{D}{c} erf\left(\frac{c\sqrt{t}}{2\sqrt{D}}\right). \tag{1.36}$$

The probability has a Gaussian component peaked in ct normalized with $\sqrt{t}$, and a correction which decays to zero if the drift is positive, while approaches formula (1.31), constant in time, if c is negative. Accordingly, the limit for $t \to \infty$ of the average maximum tends to $\langle M(t)\rangle \sim ct$ for positive c, while for negative c a careful computation gives $D/|c|$ as already found.

Potential $\propto |x|$ Consider a potential $U(x) = -a|x|$ and, as above, $c = a\mu$. It has a discontinuity in its derivative for $x = 0$ and it breaks the translational invariance. When c (a) is positive, the particle is repelled from the origin: if it starts in $x_0 = 0$, it chooses one of the two sides with the same probability, so in half of the cases the maximum increases linearly with time and in the other half it approaches a constant. On the contrary, when c (a) is negative, the particle is confined and the maximum will grow slowly, in a way not intuitively predictable.

The Fokker-Planck equation is solved on each side, giving two functions $\tilde{Q}_>$, $\tilde{Q}_<$ with four constants to fix overall. We impose the continuity in $x_0 = 0$ of $\tilde{Q}$ (for physical reasons) and of its derivative, as well as the usual conditions in $x_0 = M$ (which apply on $\tilde{Q}_>$) and $x_0 \to -\infty$ (on $\tilde{Q}_<$). The continuity of the derivative of $\tilde{Q}$ with respect to x_0 can be proved integrating the Fokker-Planck itself with the potential in exam in $[-\epsilon, \epsilon]$ with $\epsilon \to 0$. As expected, the solutions do not show translational invariance. After the determination of the constants, it is possible to set $x_0 = 0$, and the cumulative function becomes:

$$\tilde{Q}(M,0|s) = \frac{1}{s} - \frac{1}{s}\,\frac{T(s)\, e^{\frac{M}{2D}(c - T(s))}}{c\, e^{-\frac{M}{D}T(s)} - c - T(s)}; \tag{1.37}$$

where $T(s) = \sqrt{c^2 + 4Ds}$. As it is not trivial to invert such function, we use an expansion for small s. According to the sign of c, the leading terms are different.

Where c is positive, we expect M big in one half of the cases, so we expand accordingly, then develop $T(s)$ to the second order and perform the limit $s \to 0$. The cumulative function becomes:

$$\tilde{Q}(M,0|s)\Big|_{c>0} \xrightarrow[M\to\infty]{s\to 0} \frac{1}{s} - \sqrt{\frac{M}{c}}\, e^{-\frac{M}{c}s}\, \frac{1}{2s\sqrt{M/c}}\, e^{+\frac{D}{c^2}\left(\sqrt{\frac{M}{c}}s\right)^2}. \tag{1.38}$$

Equation (1.38) is not full simplified, but this form is useful in the following analysis. A first order expansion would be insufficient as it leaves only the pole in zero, which leads to the normalization term 1/2: we do not recover 1 because (1.38) is obtained in the big M limit, but on average only one half of the particles escapes on the positive side. To solve the time dependence of Q we make an ansatz on the scaling behaviour of M and analyse the resulting cumulative function. Calling $Q(M,0|t) = 1-q(z)$, with z the scaling combination, we make the hypothesis:

$$M \sim c\,t - m\,c\,t^{\alpha}\,; \quad q(z)|_{c>0} = q\left(\frac{t - M/c}{t^{\alpha}}\right); \tag{1.39}$$

with m of order one and α to be found. Rewriting Q with this scaling form, the corresponding scaling for $\tilde{Q}$ (obtained in the large M limit) matches (1.38) if $\int_{-\infty}^{\infty} dz\, e^{-\lambda z} q(z) = e^{+\frac{D}{c^2}\lambda^2}$ and $\alpha = 1/2$. Therefore the maximum grows linearly with time for one half of the particules and its fluctuations grow like its square root. Indeed the simulations gives as average maximum one half of ct with $\sim\sqrt{t}$ fluctuations, accounting also for the particles escaped in the negative region. A function $q(z)$ which meets the above definition is guessed by analogy with the linear drift:

$$q(z)|_{c>0} = \frac{1}{4}\, erfc\left(\frac{M - ct}{2\sqrt{DM/c}}\right). \tag{1.40}$$

For negative c, it is sufficient to expand $T(s)$ at the first order:

$$\tilde{Q}(M,0|s)\Big|_{c<0} \xrightarrow{s\to 0} \frac{1}{s} - \frac{1}{s}\,\frac{c^2\, e^{\frac{M}{D}c}}{c^2\, e^{\frac{M}{D}c} + 2Ds}\,. \tag{1.41}$$

The inverse Laplace transform is computed calculating the residues in the two poles; for $s = 0$ the residues of the two term erase, but the second pole gives:

$$Q(M,0|t)|_{c<0} \xrightarrow{s\to 0} e^{-e^{\frac{Mc}{D} + \ln\left(\frac{c^2}{2D}t\right)}}; \tag{1.42}$$

which is a Gumbel distribution for the variable $z = \frac{M|c|}{D} - \ln\left(\frac{c^2}{2D}t\right)$. This shows not only that the maximum grows as the logarithm of time, but also that the fluctuations are constant in time and of order one.

It is interesting to compare the results for this potential when c is negative, not found in the precedent literature, with the well known ones for the Ornstein-Uhlenbeck process in which the potential takes the form $U(x) = ax^2$. Both trap the particle, and even if the $|x|$ potential is sharper at the beginning, for x bigger than one in modulus it has a weaker effect compared to the quadratic one. As a consequence, for the latter the maximum grows like $\sqrt{\ln t}$, which is suppressed in comparison with the $\ln t$ growth of the other case (consequently we guess that a potential $\propto x^{\alpha}$ gives a behaviour $M \sim (\ln t)^{1/\alpha}$). It is not convenient to analyse the Ornstein-Uhlenbeck process with the Laplace inversion method as the solution of the backward Fokker-Planck equation in the s space exists is hard to invert, nor it is easy to compute the average maximum. In appendix A.1 we show briefly the classical way of solving it, which proceeds through a forward Fokker-Planck approach working directly on the probability $P(M, x_0|t)$ [1].

Bessel process Consider now a Brownian motion in d-dimensions: we want to figure out the statistics of the maximum distance from the origin, in the simple case without external potential. If the process is:

$$\frac{d\overline{r}}{dt} = \overline{\xi}(t) \tag{1.43}$$

and $\overline{r}$ follows a probability $P(\overline{r}, t)$, we want to work out the behavior of $|r|$ and its probability $p(|r|, t)$. $P(\overline{r}, t)$ follows the d-dimensional (forward) Fokker-Planck equation:

$$\frac{\partial P(\overline{r}, t)}{\partial t} = D\ \nabla^2 P(\overline{r}, t). \tag{1.44}$$

Considering that $P(\overline{r}, t)\, d\overline{r} = P(\overline{r}, t)\, r^{d-1}\, dr\, d\theta\, d\phi$, and $P(\overline{r}, t) = P(r, t)$, independent from the angles thanks to isotropy, it is possible to define a new object $p(r, t) = P(r, t)\, r^{d-1}$ which is the probability of being in the region $[r, r + dr]$, at whatever angle, up to a renormalization of 4π. It follows:

$$\frac{\partial p(r, t)}{\partial t} = D\left(\frac{\partial^2 p(r, t)}{\partial r^2} - \frac{\partial}{\partial r}\left(\frac{d-1}{r} p(r, t)\right)\right); \tag{1.45}$$

which is a (forward) Fokker-Planck equation with a potential $-z\frac{dU(r)}{dr} = D\frac{d-1}{r}$, meaning that $|r|$ follows the Langevin equation:

$$\frac{dr}{dt} = \frac{D(d-1)}{r} + \xi(t) \tag{1.46}$$

and $Q(M, r_0|t)$, the cumulative probability of the maximum of the process started at r_0, follows the correspondent backward Fokker-Planck:

$$\frac{\partial Q}{\partial t} = D\,\frac{\partial^2 Q}{\partial r^2} + \frac{D(d-1)}{r}\,\frac{\partial Q}{\partial r}. \tag{1.47}$$

The boundary conditions in the radial case are slightly different from the linear case as r_0 has positive support: as usual $\tilde{Q}(M, r_0 = M|t) = 0$, then in $r_0 = 0$ we impose that Q and $\tilde{Q}$ do not diverge as there is no physical reason for a singularity if the motion starts in the origin rather than in any other point (remember that Q is linked with $p(r, t) = P(r, t)\, r^{d-1}$, and even if $P(r, t)$ can be divergent, $p(r, t)$ must not). The solution is then given in terms of the modified Bessel functions of the first kind $I_n(x)$:

$$\tilde{Q}(M, r_0|s) = \frac{1}{s} - \frac{1}{s}\left(\frac{r}{M}\right)^{-\nu} \frac{I_\nu\left(r\sqrt{\frac{s}{D}}\right)}{I_\nu\left(M\sqrt{\frac{s}{D}}\right)}; \tag{1.48}$$

with $\nu = \frac{d-2}{2}$. The inverse Laplace transform can be evaluate exactly computing the residues in zero and in the other (infinite) simple poles using the properties of the Bessel functions. However it is more illuminating to compute directly the expression for the average maximum in the s space. As the k-th moment of the maximum M (starting at r_0) can be expressed in terms of $Q(M, r_0|t)$ if one integrates by parts its definition $\int dM M^k P(M, r_0|t)$, the correspondent Laplace transform is easily obtained from $\tilde{Q}(M, r_0|s)$ and results:

$$\mathcal{L}[\langle M^k(t)\rangle_{r_0}] = \frac{k}{s}\, C_{k,d}\left[r_0^{\frac{2-d}{2}}\, I_{\frac{d-2}{2}}\left(\sqrt{\frac{s}{D}}\, r_0\right)\right]\left[\frac{D}{s}\right]^{\frac{1}{2}\left(k+\frac{d}{2}-1\right)}; \tag{1.49}$$

where

$$C_{k,d} = \int_0^\infty dx\, x^{k+\frac{d}{2}-2} \left(I_{\frac{d-2}{2}}(x)\right)^{-1} . \tag{1.50}$$

Further simplifications occur starting at the origin $r_0 = 0$. In this case, taking $r_0 \to 0$:

$$\mathcal{L}[\langle M^k(t)\rangle_0] = 2^{\frac{2-d}{2}} \frac{k\, C_{k,d}}{\Gamma(d/2)} \frac{1}{s} \left(\frac{D}{s}\right)^{k/2} . \tag{1.51}$$

The simple power s dependence in the right hand side of (1.51) can be inverted exactly for all t, giving:

$$\langle M^k(t)\rangle_0 = A_{k,d}\, (D\, t)^{k/2}\,; \quad \text{with} \;\; A_{k,d} = \frac{2^{2-d/2}\, C_{k,d}}{\Gamma(k/2)\,\Gamma(d/2)}. \tag{1.52}$$

In $d = 3$ the explicit expression can be derived using $I_{1/2}(x) = \sqrt{\frac{2}{\pi\, x}}\, \sinh(x)$ in $C_{k,3}$ and then putting $k = 1$:

$$\langle M(t)\rangle_0^{3D} = \frac{\pi^{3/2}}{2} \sqrt{D\, t}\,. \tag{1.53}$$

In $d = 1$, using $I_{-1/2}(x) = \sqrt{\frac{2}{\pi\, x}}\, \cosh(x)$ we get instead $\langle M(t)\rangle_0^{1D} = \sqrt{\pi D t}$, to be compared with the maximum of the position itself $\langle M_x(t)\rangle_0 = \frac{2}{\sqrt{\pi}}\sqrt{Dt}$ which bears the same time dependence, but of course a smaller prefactor.

Exact cases In our analysis of the maximum distribution, the first step was to set equation (1.18) for the probability $Q_N(M, x_0)$. The approach developed has its basis in the expansion of Q when N is big. However, we anticipated that there are two cases in which the equation is solvable exactly without recurring to the expansion. They are example of simple random walk with a particular jump distribution. Even if in the large limit it is just recovered the already derived result, it is interesting to define these situations. The first case is the exponential jump distribution, which is solvable as it holds the relation:

$$f''(x) = \frac{1}{D}\,(1 - \delta(x))\, f(x)\,; \quad f(x) = \frac{e^{-|x|/\sqrt{D}}}{\sqrt{4D}}; \tag{1.54}$$

with $\langle x^2\rangle = 2D$ as required and the convention that $f''(0) = 0$. Deriving two times with respect to x_0 equation (1.18) leads to:

$$Q''_{N+1}(M, x_0) = \frac{1}{D}\,(Q_{N+1}(M, x_0) - Q_N(M, x_0)) \quad \text{for } x_0 < M. \tag{1.55}$$

As we are working in a discrete time framework, we take the generating function of the above equation. Setting the initial condition $Q(M, x_0|0) = \theta(M - x_0)$, the problem recast in a differential equation in the variable x_0 only:

$$\tilde{Q}''(M, x_0|z) = \frac{1}{D}\left((1 - z)\tilde{Q}(M, x_0|z) - z\right) \quad \text{for } x_0 < M; \tag{1.56}$$

with boundary conditions $\tilde{Q}(M, M|z) = 0$ and $\tilde{Q}(M, -\infty|z) = 0$. As in the general case, it is immediate to extend the problem to the joint probability of the maximum M and the minimum $-m$ changing the second boundary condition so that $\tilde{Q}((M, m), x_0|z) = 0$.

The solutions for these problems are exactly the ones found through the expansion of Q (1.25 and 1.28) but for the fact that $\sqrt{s}$ is replaced with $\sqrt{1-z}$. Clearly it is not trivial to invert them, i.e. to find an analytical form for the coefficient of z^N of their expansion around $z = 0$. Thus, we perform again an asymptotic analysis: as N big corresponds to $z \to 1$, with the change of variable $z = e^{-s}$ we recover the limit $s \to 0$. Expanding z, it is clear that the solution found with the previous method (the Fokker-Planck approximated approach) is the first order expansion of the exact solution for the exponential distribution.

The second case exactly solvable is the symmetric lattice random walk, where the particle performs jumps of fixed length (for example 1) to the left or right with equal probability:

$$f(x) = \frac{1}{2}\left[\delta(x-1) + \delta(x+1)\right]. \tag{1.57}$$

The support of the positions x_i is therefore the set of integers. Inserting $f(x)$ in equation (1.18):

$$Q_{N+1}(M, x_0) = \frac{1}{2}\left[Q_N(M, x_0 - 1) + Q_N(M, x_0 + 1)\right] \quad \forall\, n \geq 1\,; \tag{1.58}$$

the initial condition in $N = 0$ and the boundary ones are unvaried. The generating function of the equation above gives:

$$\tilde{Q}(M, x_0|z) - 1 = \frac{z}{2}\left[\tilde{Q}(M, x_0 - 1|z) + \tilde{Q}(M, x_0 + 1|z)\right]. \tag{1.59}$$

where the explicit term comes from the initial condition. This equation is solved, after a proper shift $\frac{1}{1-z}$ to eliminate the -1 term, looking for solution of the form λ^{x_0}. Applying the boundary conditions and then setting $x_0 = 0$ it results:

$$\tilde{Q}(M|z) := \tilde{Q}(M, 0|z) = \frac{1}{1-z}\left[1 - \lambda^M\right]; \quad \lambda = \frac{1 - \sqrt{1-z^2}}{z}\,; \tag{1.60}$$

and the analogous for the joint cumulative probability of the maximum and the minimum ($x_0 = 0$), obtained if $\tilde{Q}$ vanishes in M and $-m$, is:

$$\tilde{Q}((M, m)|z) = \frac{1}{1-z}\left[1 - \frac{\lambda^{-M} + \lambda^{-m}}{1 + \lambda^{-M-m}}\right]. \tag{1.61}$$

It is not surprising that equations (1.60 and 1.61) expanded at the first order ($\lambda \sim e^{-\sqrt{2s}} \sim 1 - \sqrt{2s}$) still give the same result as the backward Fokker-Planck method, when the diffusion coefficient is set to 1/2. For the maximum however, it is possible to compute the exact average value: the probability of M is given by $P_N(M) = Q_N(M+1) - Q_N(M)$, where $x_0 = 0$ was omitted (we use the forward derivative as $Q_N(M, x_0)$ is the probability that the walker stays below M without reaching it). Using the above formula and the normalization of $P_N(M)$:

$$\sum_{N=0}^{\infty} \langle M_N \rangle z^N = \sum_{M=0}^{\infty} M\left(\tilde{Q}(M+1|z) - \tilde{Q}(M|z)\right) = \frac{\sqrt{1-z} + \sqrt{1+z}}{2(1-z)^{3/2}} - \frac{1}{1-z}\,. \tag{1.62}$$

The exact inversion is performed expanding each term in power of z, multiplying and then summing the coefficients of the same power z^N. After some calculations:

$$\langle M_N \rangle = \frac{1}{2}\left[1 + (-1)^{N+1}\frac{\Gamma(N - \frac{1}{2})}{2\sqrt{\pi}\,\Gamma(N+1)}\,{}_2F_1\left(\frac{3}{2}, -N, \frac{3}{2} - N, -1\right)\right] - 1; \tag{1.63}$$

where ${}_2F_1$ is the hypergeometric function. In the large N limit, we have the expected behaviour $\sqrt{2N/\pi}$ (corresponding to $2\sqrt{Dt/\pi}$ where $D = 1/2$ and $t \to N$). We anticipate that the same result is obtained in a different framework: for the lattice random walk, as the walker can increase its position only of one step every time, the number of records (i.e. how many times the maximum value is updated) corresponds to the maximum itself. Therefore, despite the fact that the formalism of records uses slightly different objects than $Q_N(M, x_0)$, the same expression will be recovered.

Non finite variance processes In the previous cases we focussed on different potentials or particular distributions but this last ones had always a finite variance. To treat a random process where the distribution of the variables has, for example, power law tails, we need different tools. In particular, we can not build a Fokker-Planck equation, so the treatment starts from ab initio consideration on the survival probability and the result is achieved working in the space of Laplace transform and using the Pollaczek-Spitzer formula (in the following chapter we will introduce the Sparre-Andersen theorem which is a particular case of it). We are not going to show this procedure, and rather pass to the next topic, where even such distinction based on the variance of the process becomes unimportant. However we cite [7] the result for a simple random walk (no potential) with Lévy distribution of the variables $f(\xi)$:

$$f(\xi) \xrightarrow{\xi\to\infty} |\xi|^{-1-\mu}\,; \tag{1.64}$$

$$\langle M(t)\rangle = \frac{a\mu}{\pi}\,\Gamma\left(1-\frac{1}{\mu}\right)\,t^{1/\mu}\,; \tag{1.65}$$

where a is the typical microscopic length.

Chapter 2

Records: statistics of the number

In this and in the next sections we will explore an other aspect of extreme statistics: records. As in the previous part, we consider a sequence of N discrete random variables $\{x_i\} = \{x_1, ..., x_N\}$. The variable x_i is a (upper) record if $x_i > x_j \ \forall \ j < i$, namely if it is the only maximum of the sequence $\{x_1, ..., x_i\}$. According to the definition, it is necessary to surpass and not only reach the current maximum level to be a record, so ties (x_i equal to a previous x_j) are not counted as records. If the spectrum of the variables is continuous, two of them have probability zero of being exactly equal, but in lattice models with discrete spectrum this choice of the definition modifies, at least quantitatively, the results. The usual convention sets the first element of the sequence as the first record; moreover, even if we wrote it as x_1, is usual to call it x_0 for random walk models.

Calling forward records the ones of the original series $\{x_1, ..., x_N\}$, we define backward records the ones of the reverted time series $\{x_N, ..., x_1\}$: in this work we want to explore their joint statistics as it has been scarcely studied, especially in the physical field, while the forward records alone are well understood. In absence of ties, the maximum of $\{x_i\}$ is both the last forward and the last backward record: it divides the sequence in two sub-sequences with the maximum as a common point, one hosting all the forward records of $\{x_i\}$, the second all the backward ones (see Figure 2.1). Sharing the last point induces a correlation between them and consequently between the two kind of records.

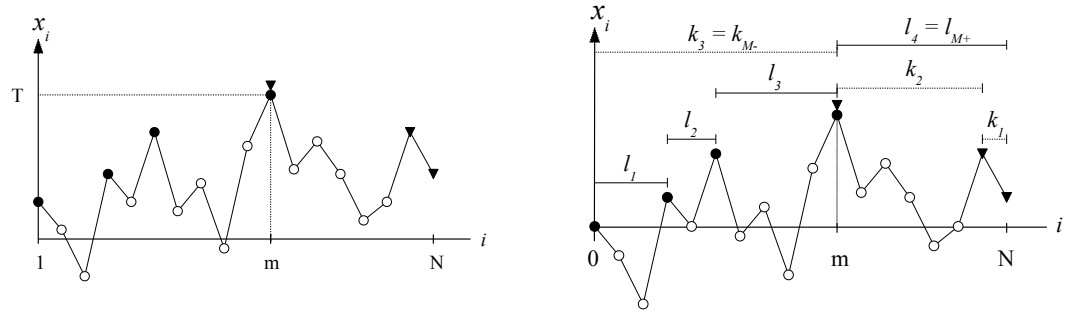

Figure 2.1: Left. Sequence of N=18 i.i.d. variables, with maximum (m, T) in $m = 11$. There are 4 forward records (full circles) in the subsequence $\{x_1, ..., x_{11}\}$ and 3 backward ones (full triangles) in $\{x_{18}, ..., x_{11}\}$. Right: the same sequence but in the R.W. framework in which the first step, labeled zero, is put at $x_0 = 0$ and we wrote the ages of the records (see section 2.2).

Another interesting aspect of records concerns their ages, namely the number of steps each of them keeps the role of maximum of the sequence, before being surpassed. In particular, is it possible to understand the behaviour of the longest among the ages, and what is its link with the maximum itself? Again, the forward case has already been studied, but the forward-backward one is new and provides further information on the correlation which links the variables of the series.

Indeed, as for the maximum, the statistics of the number and ages of records depends on the behaviour of the random variables x_i, in particular on their correlation. Again we consider independent and identically distributed random variables, and random walks as an example of strongly correlated ones. In both cases, for a big class of distributions from which the stochastic component of the process is sampled, the details of the distribution itself turn out to be unimportant to understand the behaviour of the number and ages of records: this is possible as the value of x_i is not analysed. In particular, the universality holds for every continuous (in its support) distribution in the i.i.d. case and for every symmetric continuous jump probability for random walks, even if the reasons of the universality are completely different in the two situations. As a further case of correlated variables, it is shown the analysis of symmetric random walk on a lattice. The discrete jump distribution does not belong to the universality class above mentioned, so the results are different from the continuous case.

As anticipated, the number of forward records has been widely studied in the past: for i.i.d. variables and random walks with continuous jump distribution a complete review is provided in [5], which has been followed in organizing this work; while [6] provides further results for the lattice random walk. Ages have been studied in [5], also, and an in-depth analysis for the random walks is provided in [12]. The joint distribution of the number of forward and backward records has been studied incidentally for the i.i.d. case in [8] as it presents an analogy with the statistic of upper and lower records, for which a wider literature exists ([10], [11]). We do not know at present studies for that joint distribution in the random walk case.

The chapter presents the analysis of the statistics of the number of records for the different types of variables treated. For all the situations, once defined the proper tools to analyse records, we first apply them to find the well known forward statistics and then explore the joint forward-backward statistics. We propose also an application to climate analysis, one of the fields that employs widely records statistics. The ages of records will be explored in the next chapter with the same structure as for number analysis in the present one, while the last chapter suggests some situations (the presence of a drift, a rounding process or noise) which affected the number of records, and proposes a brief study of the number of global records for multiple random walks.

2.1 I.I.D. variables

I.I.D.: forward records Consider $p(x)$ the probability distribution from which every x_i is sampled (independently one from the other). We make no particular assumption on this distribution, but its being continuous on its support. For every site i we define a variable σ_i which takes the value 1 if x_i is a record, 0 otherwise. This allows to express

the number of records in the sequence of length N as:

$$R_N = \sum_{i=1}^{N} \sigma_i \,. \tag{2.1}$$

The mean value of σ_i, through which $\langle R_N \rangle$ can be computed, is given by:

$$\langle \sigma_i \rangle = \int_{d_-}^{d_+} p(x) \left[\int_{d_-}^{x} p(y) dy \right]^{i-1} dx \,; \tag{2.2}$$

where d_-, d_+ are the extreme of the domain of x. In fact $\langle \sigma_i \rangle$ is the probability that x_i is a record, namely that all the x_j with $1 \leq j < i$ are under x_i. Thanks to the independence of the variables, for every fixed record level x_i, it factorises in the product of the probabilities that x_j is smaller than x_i, for every j. As the record level is free, we integrate the product over x_i getting (2.2). The integral is solved with a change of variable, which erases the dependence from the distribution. As all the following results are obtained starting from $\langle \sigma_i \rangle$ (or successive momenta), it is clear that they are universal, so that the particular distribution plays no role in the statistics of the number of records:

$$u = \int_{d_-}^{x} p(y)\, dy \,; \quad \langle \sigma_i \rangle = \int_0^1 u^{i-1}\, du = \frac{1}{i} \,. \tag{2.3}$$

This formalism hold as long as the cumulative function of $p(x)$ is continuous (thus discrete variables are not included). The result in (2.3) points out that every position between 1 and i can host the maximum with equal probability; moreover, as expected, the σ_i are uncorrelated:

$$\langle \sigma_i \sigma_j \rangle \overset{i \neq j}{=} \int_{d_-}^{d_+} dx\, p(x) \int_{d_-}^{d_+} dy\, p(y) \left[\int_{d_-}^{x} dx'\, p(x') \right]^{i-1} \left[\int_{d_-}^{y} dy'\, p(y') \right]^{j-1} = \frac{1}{i}\,\frac{1}{j} \,; \tag{2.4}$$

while $\langle \sigma_i^2 \rangle = \langle \sigma_i \rangle$, as σ_i takes only the values 0 and 1. The mean value of the number of records and its variance are:

$$\langle R_N \rangle = \sum_{i=1}^{N} \frac{1}{i} = H_N \sim \ln N + \gamma \,; \tag{2.5}$$

$$\langle R_N^2 \rangle - \langle R_N \rangle^2 = \sum_{i=1}^{N} \left(\frac{1}{i} - \frac{1}{i^2} \right) \sim \ln N + \gamma - \frac{\pi^2}{6} \,; \tag{2.6}$$

where H_N is the harmonic number and γ the Euler constant. Both (2.5) and (2.6) go like $\ln N$ for large N, so the fluctuations of the mean value (going like the square root of the variance) are much smaller than the mean value itself.

To have complete information on the behaviour of R_N we look at the probability $P(M|N)$ of having M records in N steps (probability that $R_N = M$), defining its generating function. Thanks to the independence between the x_i it can be calculated analytically:

$$Q_N(x) = \sum_{M=1}^{N} P(M|N) x^M = \langle x^{R_N} \rangle = \prod_{i=1}^{N} \langle x^{\sigma_i} \rangle = \prod_{i=1}^{N} \left(\frac{x-1}{i} + 1 \right) . \tag{2.7}$$

In the last passage we use the fact that σ_i is 1 with probability $1/i$, 0 otherwise. Recognizing that the right hand side is:

$$\frac{1}{N!}\,x(x+1)\cdots(x+N-1) = \sum_{M=1}^{\infty} \frac{1}{N!} \begin{bmatrix} N \\ M \end{bmatrix} x^M; \tag{2.8}$$

the probability is $P(M|N) = \frac{1}{N!}\begin{bmatrix} N \\ M \end{bmatrix}$, with the expression in the bracket indicating an unsigned Stirling number of the first kind.

As in other cases it is not simple to derive the exact expression for P, we would like to have a method to understand at least the large N behaviour. This limit is equivalent to consider x going to 1 for which the sum is more and more dominated by the high index terms, so high N. In general we expect that the distribution tends to a Gaussian with the average and variance already found, according to the central limit theorem. To prove it, consider that for large N the sum defining the generating function can be replaced by an integral: making the change of variable $x = e^{-s}$, $s \to 0$, the generating function defined in (2.7) tends to a Laplace transform. Expanding to the second order in s the logarithm of the right hand side of (2.7) and then inverting the Laplace transform we obtain, as expected:

$$P(M|N) \sim \frac{1}{\sqrt{2\pi\left(H_N - H_N^{(2)}\right)}} \exp\left(-\frac{(M-H_N)^2}{2\left(H_N - H_N^{(2)}\right)}\right). \tag{2.9}$$

where $H_N^{(2)}$ is a generalized harmonic number in power 2.

In anticipation of the forward-backward case, notice that the exact form of the generating function can be used to get the expression of $\langle R_N \rangle$ directly. In particular, using the fact that $Q_N(x) = \langle x^{R_N} \rangle$, we have:

$$\langle R_N \rangle = \frac{d}{dx}\, Q_N(x)\bigg|_{x=1}. \tag{2.10}$$

Instead of computing it directly, we consider its generating function with respect to N. Defining the generating function for $Q_N(x)$ as $\tilde{Q}(x,z)$, equation (2.10) becomes:

$$\sum_{N=0}^{\infty} \langle R_N \rangle z^N = \frac{\partial}{\partial x} \sum_{N=0}^{\infty} Q_N(x)\, z^N \bigg|_{x=1} = \frac{\partial}{\partial x}\, \tilde{Q}(x,z)\bigg|_{x=1}. \tag{2.11}$$

Once worked out the expression for $\tilde{Q}(x,z)$, $\langle R_N \rangle$ is the coefficient of the power z^N in the expansion of the right hand side of (2.11). To compute $\tilde{Q}(x,z)$ we note the recursion $Q_N(x) = Q_{N-1}(x)\left(1 + \frac{x-1}{N}\right)$, with $Q_1(x) = x$ and, by coherence, $Q_0(x) = 1$. Summing both sides of the recursion from $N = 1$ to infinity, we have:

$$\tilde{Q}(x,z) - 1 = z\,\tilde{Q}(x,z) + (x-1) \sum_{N=1}^{\infty} \frac{1}{N}\, Q_{N-1}(x)\, z^N. \tag{2.12}$$

Differentiating it with respect to z we obtain a partial differential equation which is easily integrable (the derivative of the sum in the last term becomes just $\tilde{Q}(x,z)$). Applying the boundary condition $\tilde{Q}(x, z=0) = Q_0(x) = 1$, the solution gives:

$$\tilde{Q}(x,z) = (1-z)^{-x}; \tag{2.13}$$

and so we recover for the average number of records the result in (2.6):

$$\frac{\partial}{\partial x}\,\tilde{Q}(x,z)|_{x=1} = -\frac{\ln(1-z)}{1-z} = \sum_{N=0}^{\infty} H_N\, z^N. \tag{2.14}$$

Again in anticipation of the forward-backward case, we want to obtain the function $Q_N(x)$ in a different way: consider the absolute maximum of the sequence of i.i.d., say its level is T and happens at the step m (as in Figure 2.1). What is the generating function of the probability of the $R_N - 1$ records before m when (m, T) is fixed? The points are no longer distributed according to $p(x)$ but to the conditioned (normalized) probability:

$$q_T(x) = \frac{p(x)\,\Theta(T-x)}{\int_{-\infty}^{T} dx'\, p(x)}\,. \tag{2.15}$$

They still retain their independence, so for $i < m$ it still holds $\langle\sigma_i\rangle = \frac{1}{i}$, but for $i > m$ $\langle\sigma_i\rangle = 0$. The conditioned generating function is therefore $Q_N(x)|_m = Q_{m-1}(x)$, with $Q_a(x)$ as in (2.7). The unconditioned one is the sum of $Q_N(x)|_m$ over all m, multiplied by the probability that the maximum stays in m (which is $\frac{1}{N}$ because the variables are independent) and by x as now the generating function comprehends also the maximum, which is always a record (so the factor x^{σ_m} is always x):

$$Q_N(x) = \sum_{m=1}^{N} \frac{x}{N} \prod_{i=1}^{m-1} \left(\frac{x-1}{i} + 1\right). \tag{2.16}$$

The sum over the level of the maximum T is trivial and gives 1 as the probability is normalized. It easy to prove the equivalence between (2.16) and (2.7) by induction or by generating function: transforming (2.16) with respect to N and taking the derivative in z, we have:

$$\frac{\partial}{\partial\, z} \sum_{N=0}^{\infty} Q_N(x)\, z^N = x \sum_{N=1}^{\infty} \sum_{m=1}^{N} z^{N-1} \prod_{i=1}^{m-1} \left(\frac{x+1}{i} + 1\right) = \tag{2.17}$$

$$= x \sum_{p=0}^{\infty} z^p \sum_{m=1}^{\infty} z^{m-1} \prod_{i=1}^{m-1} \left(\frac{x+1}{i} + 1\right) = \frac{x}{(1-z)^{x+1}}\,; \tag{2.18}$$

where, in the last passage, the sum in m gives the generating function with respect to $m-1$ of $Q_{m-1}(x)$, found in (2.13); and the sum in p gives $(1-z)^{-1}$. This expression is equal to the derivative in z of $\tilde{Q}(x,z)$ in (2.13), obtained from the first definition of $Q_N(x)$ (2.7). As the generating functions of the two definition of $Q_N(x)$ have the same derivative and the same value at $x = 1$ (for example), the generating and the functions themselves are identically equal.

I.I.D.: forward and backward records The original part of the present work consists in extending the formalism to forward and backward records. Do perform this task, we call their numbers respectively R_N^+ and R_N^- and define the variables σ_i^+ and σ_i^- in analogy with σ_i. It is immediate to see that:

$$\langle\sigma_i^+\rangle = \frac{1}{i}\,; \quad \langle\sigma_i^-\rangle = \frac{1}{N-i+1}\,; \tag{2.19}$$

as the backward quantities have the same statistics as the forward ones but on the reverted time series. The correlations $\langle \sigma_i^+ \sigma_j^+ \rangle$ and $\langle \sigma_i^- \sigma_j^- \rangle$ come from the forward case, so they factorize for ever $i \neq j$. For the correlations between forward and backward quantities $\langle \sigma_i^+ \sigma_j^- \rangle$, i.e. the probability that site i is a forward record and j a backward one, we distinguish three cases. If $i = j$:

$$\langle \sigma_i^+ \sigma_i^- \rangle = \int_{d_-}^{d_+} dx\, p(x) \left[\int_{d_-}^{x} dx'\, p(x') \right]^{i-1} \left[\int_{d_-}^{x} dy'\, p(y') \right]^{N-i} = \frac{1}{N} \neq \langle \sigma_i^+ \rangle \langle \sigma_i^- \rangle \,; \quad (2.20)$$

indeed x_i is both a forward and a backward record if it is the sequence maximum, which can be in every site with equal probability. If $i > j$, $\langle \sigma_i^+ \sigma_j^- \rangle$ is identically zero: if x_i is a forward record, then all the previous variables are smaller and $x_j,\ j < i$, cannot be a record coming from the right (a backward one). Finally, if $i < j$ we have not correlation as the records levels x_i and x_j are different:

$$\langle \sigma_i^+ \sigma_j^- \rangle = \int_{d_-}^{d_+} dy_i\, p(y_i) \int_{d_-}^{d_+} dy_j\, p(y_j) \left[\int_{d_-}^{y_i} dx'\, p(x') \right]^{i-1} \left[\int_{d_-}^{y_j} dy'\, p(y') \right]^{N-j} = \langle \sigma_i^+ \rangle \langle \sigma_j^- \rangle \,. \quad (2.21)$$

As a consequence of the above relations, also R_N^+ and R_N^- are correlated, and their covariance is explicitly calculated as:

$$\begin{aligned} \langle R_N^+ R_N^- \rangle - \langle R_N^+ \rangle \langle R_N^- \rangle &= \sum_{i=1}^{N} \left(\frac{1}{N} + \sum_{j>i} \langle \sigma_i^+ \sigma_j^- \rangle \right) - \sum_{i,j=1}^{N} \langle \sigma_i^+ \rangle \langle \sigma_j^- \rangle = \\ &= 1 + \sum_{i=1}^{N} \sum_{j>i} \frac{1}{i(N-j+1)} - \left(\sum_{i=1}^{N} \frac{1}{i} \right)^2 . \end{aligned} \quad (2.22)$$

The first of the two sums is of difficult computation, but the problem is solvable using the derivatives of the generating function of the probability, as already anticipated in the forward case.

The generating function of the probability $P(M^+, M^-|N)$ of having M^+ forward records and M^- backward ones in a sequence long N can be written, in analogy with the forward case, as:

$$Q_N(x,y) = \sum_{M^+,M^-=1}^{N} P(M^+, M^-|N)\, x^{M^+} y^{M-} = \langle x^{R_N^+} y^{R_N^-} \rangle = \langle x^{\sum_{i=1}^{N} \sigma_i^+} y^{\sum_{j=1}^{N} \sigma_j^-} \rangle \,; \quad (2.23)$$

but the correlations between the σ_i^+ and σ_j^- prevent to develop it further. Instead, we can adapt the expression at fixed maximum in position m with height T: there are $M^+ - 1$ forward records before the maximum, in a sequence of length $m-1$, and $M^- - 1$ backward ones after it, in a $N - m$ sequence (in Figure 2.1 this correspond to 3 forward records in the 10 steps before the maximum and 2 backward ones in the 7 steps after it). As (m, T) is fixed, the two sequences are uncorrelated so the joint probability of the number of records factorizes in a forward and a backward component, as well as $Q_N(x,y)|_m$. To obtain the generating function at free maximum we sum over m with a factor $\frac{xy}{N}$: $1/N$ for the probability that the maximum stays in m, and the new degree of freedom - the maximum

itself - contributes with xy as it is always both forward and backward record. Then:

$$Q_N(x,y) = \frac{xy}{N} \sum_{m=1}^{N} Q_{m-1}(x)\, Q_{N-m}(y)\,; \tag{2.24}$$

with $Q_a(u)$ as in (2.7). We call $\tilde{Q}(x,y,z)$ the generating function of $Q_N(x,y)$ with respect to N and using (2.24) we can write:

$$\frac{\partial}{\partial z}\,\tilde{Q}(x,y,z) = \sum_{N=1}^{\infty} N\, Q_N(x,y)\, z^{N-1} = xy \sum_{N=1}^{\infty}\sum_{m=1}^{N} z^{N-1}\, Q_{m-1}(x)\, Q_{N-m}(y)\,. \tag{2.25}$$

The right hand side presents a convolution structure giving $xy\,\tilde{Q}(x,z)\,\tilde{Q}(y,z)$ where $\tilde{Q}(u,z)$ is meant in the forward sense (2.13). Replacing its explicit form in the differential equation (2.25) with boundary condition $\tilde{Q}(x,y,z=0)=1$ the solution is:

$$\tilde{Q}(x,y,z) = 1 + \frac{xy}{1-x-y}\left(1-(1-z)^{1-x-y}\right)\,. \tag{2.26}$$

To compute the correlation between the number of forward and backward records we calculate:

$$\sum_{N=0}^{\infty} \langle R_N^+\, R_N^-\rangle\, z^N = \frac{\partial^2}{\partial x\,\partial y}\,\tilde{Q}(x,y,z)\,\Big|_{x,y=1} = \frac{z}{1-z} + \frac{\ln^2(1-z)}{1-z}\,. \tag{2.27}$$

The right hand side must be expanded in power of z to invert it. The first term is $\sum_{N=1}^{\infty} z^N$, so the coefficient is 1 (a part the formal case $N=0$, for which it is 0). To evaluate the second term, we expand each factor and, after some computations, group them in a single sum over z. We show the passages for this case, and skip them in the future analogous cases:

$$\sum_{i=0}^{\infty} z^i \cdot \sum_{j=1}^{\infty} \frac{z^j}{j} \cdot \sum_{k=1}^{\infty} \frac{z^k}{k} = \sum_{N=2}^{\infty}\left[\sum_{n=2}^{N} \frac{2H_{n-1}}{n}\right] z^N\,. \tag{2.28}$$

The sum in the coefficient of z^N (called s_N in the next lines) is solved using the following formula, which is the discrete counterpart of the integration by parts:

$$\sum_{n=1}^{N} a_n b_n = a_N B_N - \sum_{n=1}^{N-1} (a_{n+1}-a_n) B_n; \quad B_n := \sum_{j=1}^{n} b_n\,. \tag{2.29}$$

Applying it, after some rearrangements we obtain a recursion for s_N:

$$s_N = \sum_{n=1}^{N-1} \frac{2H_n}{n+1} = 2\left((H_N)^2 - \sum_{n=1}^{N} \frac{1}{n^2} - \sum_{n=1}^{N-1} \frac{H_n}{n+1}\right); \tag{2.30}$$

which gives the solution $s_N = (H_N)^2 - H_N^{(2)}$. Recalling the coefficient 1 from the first term in (2.27) and subtracting $\langle R_N^+\rangle\langle R_N^-\rangle = H_N^2$, the covariance between forward and backward records is:

$$\langle R_N^+\, R_N^-\rangle - \langle R_N^+\rangle\langle R_N^-\rangle = 1 - \sum_{n=1}^{N} \frac{1}{n^2} = 1 - \frac{\pi^2}{6} + \frac{1}{N} + O\left(\frac{1}{N^2}\right)\,; \tag{2.31}$$

where we used $\sum_{n=1}^{N} \frac{1}{n^2} \sim \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$, and evaluated the correction ($\sum_{n=N+1}^{\infty} \frac{1}{n^2}$) at the first order transforming the sum into an integral as N is large.

The asymptotic form of the probability $P(M^+, M^-|N)$ can be obtained expanding (2.24) directly, but the sum over m makes the task hard. However, it is possible to recast the generating function in a new form, which depends only on $x + y$:

$$Q_N(x, y) = \frac{xy}{N} \prod_{i=1}^{N-1} \left(1 + \frac{x + y - 1}{i}\right) . \tag{2.32}$$

The equality between (2.24) and (2.32) is proved as they have the same generating functions: taking apart the xy/N coefficient, the sum over N of (2.24) gives the convolution of two forward generating functions resulting in $z\,(1-z)^{-x-y}$, which is the generating function of the product in (2.32). The physical reason for which this simplification occurs is still unclear. Nevertheless, it allows us to perform the expansion using, as usual, $x \sim e^{-s}$ ad $y \sim e^{-t}$. Replacing the product with the exponential of a sum, it is easy to develop it to the second order and the inverse Laplace transform is then straightly performed if we express the object in function of $s+t$ and $s-t$. The final result can be expressed in terms of two rescaled variables f and b, which are the normalised fluctuations of the two types of records from their average values:

$$f := \frac{M^+ - H_N}{\sqrt{C_N}}; \quad b := \frac{M^- - H_N}{\sqrt{C_N}}; \tag{2.33}$$

$$P(f, b|N) = \frac{1}{2\pi} \frac{1}{\sqrt{1 - (\Omega_N)^2}} \exp\left(-\frac{(f+b)^2}{4\,(1+\Omega_N)} - \frac{(f-b)^2}{4\,(1-\Omega_N)}\right); \tag{2.34}$$

$$\sim \frac{1}{2\pi} \exp\left(-\frac{f^2 + b^2}{2} + \Omega_N\, fb\right) + O\left(\Omega_N\right); \tag{2.35}$$

with:

$$C_N := H_N - H_N^{(2)}; \quad \Omega_N := \frac{1 - H_N^{(2)}}{C_N} \sim (\ln N)^{-1}.$$

The probability $P(M^+, M^-|N)$ of the original variables is obtained from the above formulas substituting (2.33) and multiplying by the Jacobian C_N^{-1}. Using (2.34) to calculate the correlation between R_N^+ and R_N^- we get back $1 - H_N^{(2)}$, which is correct for every N. In the large N limit, we can express the **connected part of the joint probability** of f and b as a **universal scaling function** times a scaling factor going like $(\ln N)^{-1}$:

$$\begin{aligned} C(x, y|N) &:= P(f, b|N) - P(f|N)P(b|N) \sim \Omega_N\, F(f, b)\,; \\ F(f, b) &:= \left(1 - \frac{\pi^2}{6}\right) \frac{fb}{2\pi}\, e^{-\frac{f^2+b^2}{2}}. \end{aligned} \tag{2.36}$$

Once more we have the proof that the correlation between the variables becomes less and less important but in a slow way, being influenced by the logarithm of the sequence length and not by N itself.

An interesting quantity is the difference Δ between the number of forward and backward records ($\Delta = M^+ - M^-$). By symmetry the average of Δ is zero and the direct computation of its variance gives $2(H_N - 1)$. We can look at its distribution integrating

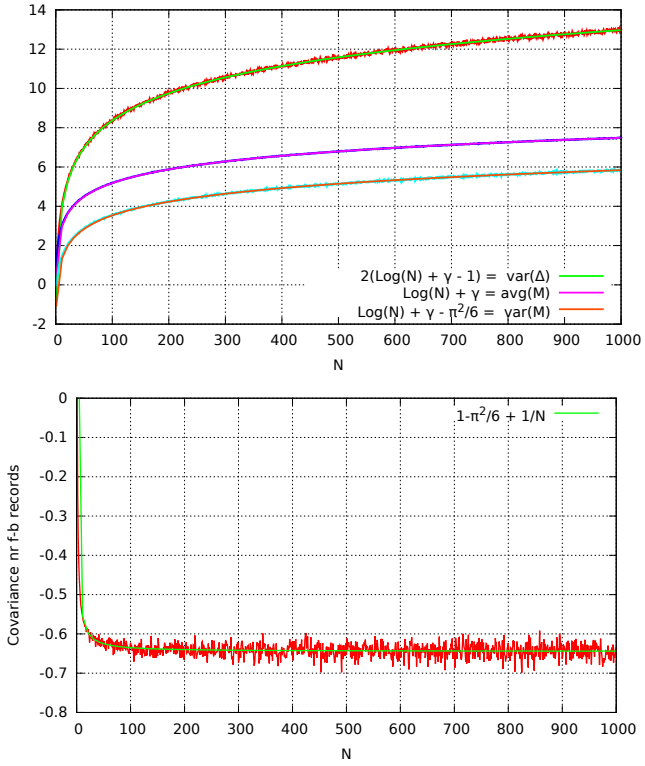

Figure 2.2: In the first plot, from the higher line, the variance of Δ, the mean value of the number of forward maxima, the correlation of forward maxima. For every quantity it is shown the theoretical curve, superimposed to the one obtained by simulation of Gaussian i.i.d.. In the second plot, the correlation between the number of forward and backward records, and the theoretical curve of equation (2.22).

(2.34) on the sum $f + b$:

$$\begin{aligned} P(\Delta|N) &= \frac{1}{\sqrt{C_N}} \int_0^\infty d(f+b)\; P(f,b|N)|_{f-b=\Delta/\sqrt{C_N}} = \\ &= \frac{1}{\sqrt{4\pi(H_N-1)}} e^{-\frac{\Delta^2}{4(H_N-1)}}\;; \end{aligned} \tag{2.37}$$

so the distribution is a Gaussian with the expected variance: the double of the variances of M^+ and M^- which dominates at large N, as expected if they were uncorrelated, plus a correction.

Simulations We verified the analytical results for the statistic of the number of records simulating a sequence of i.i.d. for different jump distributions: flat in $[-1/2; 1/2]$, Gaussian, exponential. For each case we used 10^5 samples of length between 1 and 10^3 steps, calculating for every length the average number of forward and backward records, their variance and covariance and the variance of Δ, founding a good agreement with the theory (Figure (2.2) shows the result for the Gaussian distribution).

Moreover, to check the scaling of the joint probability function, we simulated series of 10^3, 10^4 and $2 \cdot 10^5$ steps, using 10^6 samples for each of them, and checked that the

three-dimensional histograms of the connected component of the probability collapsed on the same theoretical curve once rescaled with the logarithm of the number of steps.

I.I.D.: statistics of upper and lower records For many purposes, it is interesting to look at the same time at the number of upper (forward) records and lower (forward) records where x_i is a lower record of the sequence $\{x_i\}$ if $i = 1$, by convention, and if $x_i < x_j \ \forall \ j < i$ (in Figure 2.1 the steps $i = 1, 2, 3$ are lower records). We consider here this problem as, in the i.i.d. variables case, it presents a strong analogy with the forward-backward (upper) records problem, the physical reason behind it being still unclear. In a previous work [9] it was briefly suggested a similar correspondence, but without developing the full probability of forward and backward records. Moreover all the records were treated with different mathematical instruments.

Defining σ_i^u, σ_i^l for each site i so that σ_i^u is 1 if i is a upper record, 0 otherwise and σ_i^l the analogous for lower records, we have $\langle \sigma_i^u \rangle = \langle \sigma_j^l \rangle = 1/i$ (the usual statistics if taken singularly) and:

$$\begin{aligned} \langle \sigma_i^u \sigma_j^l \rangle &= \langle \sigma_i^u \rangle \langle \sigma_j^l \rangle \quad \text{if} \ \ i \neq j \,; \\ &= 0 \qquad\qquad\ \ \text{if} \ \ i = j \neq 1 \,; \\ &= 1 \qquad\qquad\ \ \text{if} \ \ i = j = 1 \,. \end{aligned} \tag{2.38}$$

Indeed a site cannot be an upper and a lower record at the same time except the case $i = 1$, which is both by definition. Different sites instead are uncorrelated. Applying (2.38) and taking $\langle \sigma_i^u \sigma_j^u \rangle$, $\langle \sigma_i^l \sigma_j^l \rangle$ from the forward case, the correlation between the number of upper and lower records in N steps (R_N^u and R_N^l respectively) is:

$$\langle R_N^u R_N^l \rangle - \langle R_N^u \rangle \langle R_N^l \rangle = \sum_{i,j=1}^{N} \Big(\langle \sigma_i^u \sigma_j^l \rangle - \langle \sigma_i^u \rangle \langle \sigma_j^l \rangle \Big) = - \sum_{i=2}^{N} \langle \sigma_i^u \rangle \langle \sigma_i^l \rangle = 1 - H_N^{(2)}; \tag{2.39}$$

the same as between the number of forward and backward records. As the only correlations are between σ_i^u and σ_i^l, the generating function for the joint probability of having U upper and L lower records given N is:

$$S(x, y|N) = \sum_{U,L=1}^{N} P(U, L|N)\, x^U y^L = \langle x^{\sum_{i=1}^N \sigma_i^u}\, y^{\sum_{i=1}^N \sigma_i^l} \rangle = \prod_{i=1}^{N} \langle x_i^{\sigma_i^u}\, y_i^{\sigma_i^l} \rangle \,; \tag{2.40}$$

$$\begin{aligned} \langle x_i^{\sigma_i^u}\, y_i^{\sigma_i^l} \rangle &= P_i(0,0) + x\, P_i(1,0) + y\, P_i(0,1) + xy\, P_i(1,1) = \\ &= 1 - \frac{2}{i} + \frac{x+y}{i} \qquad \text{if} \ \ i \neq 1; \\ &\quad\ xy \,; \qquad\qquad\qquad\ \ \text{if} \ \ i = 1 \,; \end{aligned} \tag{2.41}$$

where we defined the probability $P_i(a, b)$ that $\sigma_i^u = a$ and $\sigma_i^l = b$. So the generating function of $P(U, L|N)$ becomes:

$$S(x, y|N) = xy \prod_{i=2}^{N} \left(1 + \frac{x + y - 2}{i} \right). \tag{2.42}$$

If we expand it to the second order for $x \sim e^{-s}$, $y \sim e^{-t}$, s, t small, as in the previous cases, we recover the same approximate form as the one obtained from $Q_N(x, y)$ (2.32) in the

forward-backward case. To check that the two probabilities are identical at every order, we look at the generating function of $S(x, y|N)$ with respect to N, called $\tilde{S}(x, y, z)$. As in the forward case, it holds a recursion relation for every $N > 1$: $S(x, y|N) = S(x, y|N - 1)\left(1 + \frac{x+y-2}{N}\right)$, while $S(x, y|0) = 0$ and $S(x, y|1) = xy$. Using it:

$$\sum_{N=2}^{\infty} S(x, y|N)\, z^N = \tilde{S} - 1 - xy\, z = z(\tilde{S} - 1) + (x + y - 2) \sum_{N=2}^{\infty} \frac{1}{N}\, S(x, y|N - 1)\, z^N\,; \quad (2.43)$$

where the middle term is the direct computation of the sum and the right hand side uses the recursion. The argument of $\tilde{S}$ is understood. Taking the derivative with respect to z and rearranging the expression we write a differential equation in z:

$$(1 - z)\, \frac{\partial \tilde{S}}{\partial z} = (x + y - 1)\tilde{S} + xy + (1 - x - y)\,. \quad (2.44)$$

Instead of solving it explicitly, we try the solution $\tilde{S}(x, y, z) = \tilde{Q}(x, y, z)$, with $\tilde{Q}$ as in (2.26). Indeed it solves (2.44), leading to the conclusion that the joint statistics of the number of forward (upper) and backward (upper) records is the same as the joint statistics of the number of (forward) upper and lower ones: **upper backward records and lower forward ones have a statistically identical relation with the upper forward ones** even if they are not the same object nor take necessarily the same value for a single sample.

Simulations We repeated the simulation as above (10^5 samples of maximum length 10^3 steps for flat, Gaussian and exponential distribution) computing the number of upper and lower records, finding again good agreement with the predicted mean values, variances and correlation between them.

2.2 Random walks with continuous jump distribution

As in many physical situations we deal with sequences of correlated variables, we want to repeat the previous analysis for a system where the correlation among the x_i is strong, and compare the results to the ones for the independent case. The random walk represents a good example of this sequences, and as it is used as a model for many physical systems, it is interesting to define the statistics of the number of records in this case. The random walk variables are defined such that:

$$x_i = x_{i-1} + \xi_i\,; \quad (2.45)$$

where ξ_i are i.i.d. variables taken from a distribution $\phi(\xi)$. In this section, we consider distributions $\phi(\xi)$ with a continuous cumulative function, i.e. the walker occupies positions in the space on a continuous range, and the lattice case is not included. The starting level x_0 is set at zero for simplicity and, by convention, it counts as step number zero (and not one as in i.i.d. variables), namely $\{x_i\} = \{x_0, ..., x_N\}$. Records are defined as in the previous case, but in the random walk problem it is useful to define also the ages of records as the number of steps they survive before a new variable surpasses them: if x_i and x_j are two successive records, then the age of x_i is $l_i = j - i$ (see the right plot of Figure 2.1). The age of the last record is set by convention so that $\sum_{i=1}^{R_N} l_i = N$.

As the variables are not any more independent, objects like the σ_i used before depend on the history, and so are useless as they have not a fixed average value. The quantities we need are the survival probability that a random walk, starting in x_0, stays below this level for l time steps:

$$q_{\text{-}}(l) = \mathbb{P}\left[x_k < x_0, \forall\ 1 \leq k \leq l\right]\ ; \tag{2.46}$$

and the first passage probability that the random walk crosses the starting level x_0 between the step $(l-1)$ and l, which comes from the survival probability:

$$f_{\text{-}}(l) = q_{\text{-}}(l-1) - q_{\text{-}}(l)\,. \tag{2.47}$$

By translational invariance $q_{\text{-}}(l)$ does not depend on the initial point x_0. For the class of jump distributions considered (with continuous cumulative function) the **generalized Sparre-Andersen theorem** provides a formula for the generating function of $q_{\text{-}}(z)$:

$$\tilde{q}_{\text{-}}(z) = \sum_{k=0}^{\infty} q_{\text{-}}(k) z^k = \exp\left[\sum_{k=1}^{\infty} \frac{z^k}{k}\, \mathbb{P}(x_k < 0)\right]. \tag{2.48}$$

To explicit it, we must specify just the probability of being under the level zero according to $\phi(\xi)$. For symmetric continuous distributions it is $\frac{1}{2}$, so that:

$$\tilde{q}_{\text{-}}(z) = \tilde{q}(z) = \frac{1}{\sqrt{1-z}};\quad q_{\text{-}}(l) = q(l) = \frac{1}{2^{2l}}\binom{2l}{l}\,. \tag{2.49}$$

Thus, given this symmetry, the result does not depend on further details of the jump distribution: in particular it has no consequence if the second momentum of $\phi(\xi)$ is finite or not, and the following analysis includes Gaussian and flat statistics as well as Lévy-Flights. Moreover we abandon the subscript minus as the probability of staying below or above the initial level is the same. In this case, the first passage probability takes the form:

$$\tilde{f}(z) = \sum_{l=1}^{\infty} f(k) z^l = 1 - (1-z)\,\tilde{q}(z) = 1 - \sqrt{1-z}\,. \tag{2.50}$$

The probabilities $q(l)$ and $f(l)$ alone enlighten the statistics of records as long as we make the further assumption that the intervals l_i in which the sequence is naturally split by records are statistically independent thanks to the **Markov chain property** (except for the constraint on their sum, which has to be N).

R.W.: forward records As usual, we investigate first the known case of forward records to get familiar with the formalism, skipping some computation which have already been performed elsewhere. Under the assumption that intervals are independent, the probability of having M records with ages $\{\vec{l}\} := \{l_1, .., l_M\}$ in N steps is:

$$P(\{\vec{l}\}, M|N) = f(l_1) f(l_2) \cdots f(l_{M-1}) q(l_M)\, \delta_{\sum_{i=1}^{M} l_i, N}\,; \tag{2.51}$$

in fact every records survives for l_i steps and after the maximum the walker must stay below it for l_M steps. The probability of having M records in N steps is the sum over all ages of $P(\{\vec{l}\}, M|N)$. It can be obtained exactly from its generating function, using

(2.51) and applying the delta function to distribute z^N among the sums on the ages (and get rid of N). Finally it is possible to write:

$$\sum_{N=M-1}^{\infty} P(M|N)\, z^N = \tilde{f}(z)^{M-1}\, \tilde{q}(z); \tag{2.52}$$

$$P(M|N) = \binom{2N-M+1}{N} 2^{-2N+M-1}, \quad M \leq N+1\,; \tag{2.53}$$

where the inversion has been performed finding the z^N-coefficient as usual (see [6] for further details). Equation (2.52) holds in general for every $\phi(\xi)$, while (2.53) uses the Sparre-Andersen result for symmetric ones. The average number of records $\langle R_N \rangle$ and its mean square value can be computed directly or through its generating function:

$$\sum_{N=0}^{\infty} \langle R_N \rangle\, z^N = \sum_{M=1}^{\infty} M\, \tilde{f}^{M-1}\, \tilde{q} = \frac{\tilde{q}}{(1-\tilde{f}\,)^2} \overset{S.A.}{=} \frac{1}{(1-z)^{3/2}}\,; \tag{2.54}$$

$$\sum_{N=0}^{\infty} \langle R_N^2 \rangle\, z^N = \cdots = \frac{2\,\tilde{q}\,\tilde{f}}{(1-\tilde{f}\,)^3} + \frac{\tilde{q}}{(1-\tilde{f}\,)^2} \overset{S.A.}{=} \frac{2}{(1-z)^2} - \frac{1}{(1-z)^{3/2}}; \tag{2.55}$$

and the argument z for $\tilde{q}$ and $\tilde{f}$ is understood. Extracting the coefficient of the z^N power gives:

$$\langle R_N \rangle = (2N+1)\binom{2N}{N} 2^{-2N} \sim \frac{2}{\sqrt{\pi}}\sqrt{N} + O\left(\frac{1}{\sqrt{N}}\right); \tag{2.56}$$

$$\langle R_N^2 \rangle - \langle R_N \rangle^2 = 2\left(1-\frac{2}{\pi}\right) N + O(\sqrt{N}). \tag{2.57}$$

The fluctuations of R_N grow as the mean value itself, while in the i.i.d. variables case they were much smaller. It is thus expected that the probability approaches asymptotically a function of $(M/\sqrt{N})$, and indeed, using a Stirling expansion on (2.53), it appears a scaling form:

$$P(M|N) \sim \frac{1}{\sqrt{N}}\, g\left(\frac{M}{\sqrt{N}}\right); \quad g(x) = \frac{1}{\sqrt{\pi}}\, e^{-\frac{x^2}{4}}\,. \tag{2.58}$$

Once more, we recover a Gaussian tail for the asymptotic probability, but here the natural distribution is a half Gaussian: the mean value is correctly $\sim \sqrt{N}$, but the mode is zero (see Figure 2.3).

R.W.: forward and backward records In analogy with the forward case, the analysis moves from the joint probability of having in N steps M^+ forward records with ages l_i and M^- backward ones with ages k_j (numbered backwardwise, as in Figure 2.1). If the maximum value is reached only once in the sequence (no ties for it), which happens with probability one for a continuous jump distributions, we have:

$$P(\{\vec{l},\vec{k}\}, M^+, M^-|N) = f(l_1)\cdots f(l_{M^+-1})\, f(k_1)\cdots f(k_{M^--1})\, \delta_{\sum_{i=1}^{M^+-1} l_i + \sum_{j=1}^{M^--1} k_j, N}\,. \tag{2.59}$$

The probability $P(M^+, M^-|N)$ is the sum of (2.59) over all the ages and its generating function with respect to N is easily obtained applying the δ function as in the forward

case, getting a function of $M^+ + M^-$ only:

$$\sum_{N=0}^{\infty} P(M^+, M^-|N)\, z^N = \tilde{f}(z)^{M^+ + M^- - 2} = (1 - \sqrt{1-z})^{M^+ + M^- - 2}\,. \tag{2.60}$$

Remember that in the independent case the analysis was performed transforming in M^+ and M^- but here that process would have been much more difficult; moreover we see that this single transformation is sufficient to reach, in this case, a simple form for the probability. From the generating function of the joint probability we can compute exactly the generating function of $\langle R_N^+ R_N^- \rangle$:

$$\begin{aligned} \sum_{N=0}^{\infty} \langle R_N^+ R_N^- \rangle\, z^N &= \sum_{M^+=1}^{N} \sum_{M^-=1}^{N} M^+ M^-\, \tilde{f}(z)^{M^+ + M^- - 2} = \\ &= \frac{1}{(1-\tilde{f}(z))^4} \overset{S.A.}{=} \frac{1}{(1-z)^2} = \sum_{N=1}^{\infty} N\, z^N\,. \end{aligned} \tag{2.61}$$

Therefore, knowing $\langle R_N^+ \rangle$ and $\langle R_N^- \rangle$ from the forward case, the covariance results:

$$\langle R_N^+ R_N^- \rangle - \langle R_N^+ \rangle \langle R_N^- \rangle = \left(1 - \frac{4}{\pi}\right) N\,; \tag{2.62}$$

which is negative, as expected, and grows with N, while for i.i.d. variables it approaches a constant.

Further information is given by the asymptotic expansion of the probability for large N, as the exact inversion is hard to perform: expanding the right hand side of (2.60) at the first order in $z = e^{-s}$ and doing its inverse Laplace transform, it results:

$$P(M^+, M^-|N) \sim \frac{M^+ + M^-}{2\sqrt{\pi\, N^3}}\, e^{-\frac{(M^+ + M^-)^2}{4N}}\,; \tag{2.63}$$

where we skip the constant terms when compared to M^+, M^-. As it is a **function of the sum of the number of the two records only**, the correlation $\langle R_N^+ R_N^- \rangle$ computed from (2.63) takes the exact value 1. Therefore the joint probability takes a scaling form which depends on one significant quantity only:

$$P(M^+, M^-|N) \sim \frac{1}{N}\, h\left(\frac{M^+ + M^-}{\sqrt{N}}\right); \quad h(x) = \frac{x}{2\sqrt{\pi}}\, e^{-\frac{x^2}{4}}\,. \tag{2.64}$$

In contrast with the forward case, the mode is no more zero, but the tail is again exponentially dumped (even if not exactly as a Gaussian).

Simulations To check the results obtained and the scaling behaviour of the probability we simulated sequences of random walks with flat, Gaussian and exponential jump distribution: as usual, for the number of records we used 10^5 samples of maximum length 10^3 steps and calculated, for each step, the average number of forward records, backward ones and their covariance. The results agree with the expected theory. For the full distribution of records, we used series of 10^4, 10^5 and 10^6 steps, using 10^5 samples for each of them and computing the distributions of the rescaled variables in the forward and forward-backward case. The points fit the universal scaling functions found in the large N limit. We report the plots of these two distributions for Gaussian jumps in Figure (2.3).

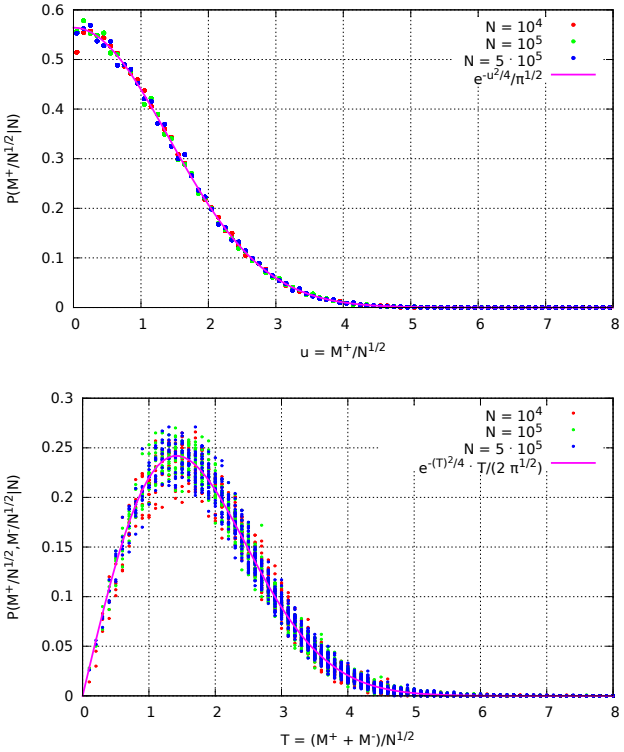

Figure 2.3: In the first plot, the distribution of the rescaled variable $M/\sqrt{N}$ in the forward case. In the second one, the distribution of the rescaled variables of forward and backward records in function of their sum, which appears to be the relevant quantity. For both graphics, are shown the results of the simulations of three different lengths with Gaussian jump distribution, and the theoretical curve.

2.3 Lattice Random Walks

An other interesting and non trivial case of random walk is the symmetric one-dimensional lattice problem: the walker jumps of a fixed quantity to its left or right with equal probability 1/2. Setting the size of the jump at one (which is unimportant in the count of records), the maximum of the position of the walker at every step can be stationary or increase of 1. As an increase marks a new records, if the walker starts from zero it turns out that the value of the maximum is also the number of records minus one (as the position zero has a level zero but is counted as a record). In the large time limit we can solve the problem looking for the statistics of the maximum for a random walk with continuous jump distribution with a proper diffusion coefficient, which turns out to be 1/2; however, it is possible to solve it exactly finding the discrete survival probability and using it in analogy with the continuous case. Here the fact that ties are neglected in the count of records has a strong influence: as the range of possible x_i is discrete, the probability that the walker comes back to a position already visited is finite. To be consistent, the first passage condition is true if the walker surpasses the starting value and not merely equals it.

We define $q_+(x_0, l)$ the probability that the walker, starting at x_0, stays above or at zero up to step l. The starting point x_0 is kept explicit in order to built a backward

equation; it will be set to zero in the end. The survival probability defined as in the continuous case is recovered as $q(l) := q_+(0,l) = q_-(0,l)$, where the last is the probability of staying below or at zero for l steps, equal to $q_+(0,l)$ by symmetry. Given a realization of the walk of $N+1$ time steps, we proceed as done in the treatment of the maximum of a lattice random walk, and split in two parts: the first step and the remaining N steps. Therefore we obtain an equation analogous to (1.18):

$$q(x_0, N+1) = \int_{-\infty}^{\infty} d(x_1 - x_0)\, q(x_1, N)\, f(x_1 - x_0)\,; \tag{2.65}$$

where $f(x)$ is the jump distribution. In the symmetric lattice case the jump distribution is:

$$f(x) = \frac{1}{2}\,\left[\delta(x-1) + \delta(x+1)\right]\,. \tag{2.66}$$

which gives the equation:

$$\begin{aligned} &q(x_0, N+1) = \frac{1}{2}\,\left[q(x_0-1, N) + q(x_0+1, N)\right] \quad \forall\ N \geq 1\,; \\ &q(x_0, 0) = 1 \ \text{ if } x_0 \geq 0\,; \quad q(-1, N) = 0 \ \ \forall\ N\,; \end{aligned} \tag{2.67}$$

where we wrote also the initial and boundary conditions. The solution is achieved following the steps used for computing $Q(M,z)$ (the cumulative probability of the maximum) for the lattice random walk, but for the different boundary conditions. Applying the present ones leads to:

$$\tilde{q}(z) = \frac{\sqrt{1+z} - \sqrt{1-z}}{z\sqrt{1-z}}\,; \quad \tilde{f}(z) = \frac{1}{z}\left(1 - \sqrt{1-z^2}\right). \tag{2.68}$$

Where the first passage probability is as usual $\tilde{f}(z) = 1 - (1-z)\,\tilde{q}(z)$. We invert $\tilde{q}(z)$ extracting the coefficient of the power z^N (as usual, we expand each term and then group them to have one sum in z), getting:

$$q(N) = \left(2\left\lfloor \frac{N-1}{2} \right\rfloor - 1\right) \frac{\Gamma\left(\left\lfloor \frac{N-1}{2} \right\rfloor - \frac{1}{2}\right)}{2\sqrt{\pi}\,\Gamma\left(\left\lfloor \frac{N-1}{2} \right\rfloor + 1\right)}\,; \tag{2.69}$$

where $\lfloor a \rfloor$ define the floor function. Asymptotically, it decays like $\sim \sqrt{2/(\pi N)}$. With these tools we can analyse the statistics of records for the lattice random walk.

Lattice R.W.: forward records In the forward case there is a complete analogy between the lattice and the continuous problem, therefore, keeping the same notation, the probability $P(\{\vec{l}\}, M|N)$ has the same expression as in (2.51) and the generating function of $P(M|N)$ is $[\tilde{f}(z)]^{M-1}\tilde{q}(z)$. The average number of records is computed as usual through the generating function, but injecting in (2.68) the lattice survival and first passage probability:

$$\sum_{N=0}^{\infty} \langle R_N \rangle z^N = \frac{\tilde{q}(z)}{(1-\tilde{f}(z))^2} = \frac{\sqrt{1+z} + \sqrt{1-z}}{2(1-z)^{3/2}}\,. \tag{2.70}$$

The inversion gives:

$$\langle R_N \rangle = \frac{1}{2}\left[1 + \frac{(-1)^{N+1}\,\Gamma\left(N - \frac{1}{2}\right)}{2\sqrt{\pi}\,\Gamma(N+1)}\ {}_2F_1\left(\frac{3}{2}, -N, \frac{3}{2} - N, -1\right)\right]; \tag{2.71}$$

where ${}_2F_1$ is the hypergeometric function. As anticipated, this is exactly the same result (up to a +1 term) obtained when looking at the maximum of a lattice random walk, but here it is recovered through a different reasoning. In the large N limit, $\langle R_N \rangle \sim \sqrt{2N/\pi}$ which is $1/\sqrt{2}$ of the result of the continuous case. Indeed inverting the asymptotic expansion at the first order in z of the probability gives:

$$P(M|N) \sim \frac{1}{\sqrt{N/2}}\, g\left(\frac{M}{\sqrt{N/2}}\right); \tag{2.72}$$

with $g(x)$ the continuous case scaling function (2.58). This asymptotic result on lattice provides an interesting bridge between the number of records of a random walk with continuous jump distribution and the value of the maximum in the same situation: indeed, it suggests a physical explanation for the one to one correspondence which sets for large N between these two objects, up to a proper rescaling based on the diffusion length (in particular a rescaled time Dt makes the maximum statistics equal to the one of records in a N step sequence).

Lattice R.W.: forward and backward records For the forward-backward case we must pay attention to **ties**. The formula for the joint probability of M^+ forward records with ages l_i and M^- backward ones with ages k_j found in the continuous case is valid only if the last forward record happens at the same time of the last backward one, i.e. the maximum level is reached only once by the walker. In the continuous case the probability of visiting again a position is infinitesimal, so this description holds; on the contrary on a lattice ties are very frequent, and the maximum level can be reached more than once, at different steps m_i: the smallest among them (m_1) determines the last forward record and the biggest (m_2) the last backward record. Calling $m := m_2 - m_1$ the gap between them, we define $g(m)$ the probability that the walker returns at the initial level after m steps and stays below it in the meanwhile. The correct form of the joint probability of forward and backward records is therefore:

$$P(\{\vec{l},\vec{k}\}, M^+, M^-|N) = f(l_1)\cdots f(l_{M^+-1}) f(k_1)\cdots f(k_{M^--1})\, g(m)\, \delta_{\sum_{i=1}^{M^+} l_i + \sum_{j=1}^{M^-} k_j + m, N}\,. \tag{2.73}$$

To compute $g(m)$ we notice that $f(m) = g(m-1)/2$: the probability of surpassing the initial level at step m is the probability of being in it at the previous time multiplied by the one of doing a +1 jump which is 1/2. Taking the generating function of $f(m)$, and knowing that $f(0) = 0$:

$$\tilde{f}(z) = \sum_{m=0}^{\infty} f(m)\, z^m = \sum_{m=1}^{\infty} f(m)\, z^m = \sum_{m=1}^{\infty} \frac{1}{2}\, g(m-1)\, z^m = \frac{z}{2}\, \tilde{g}(z)\,. \tag{2.74}$$

Remembering the expression of $\tilde{f}(z)$, we see that $\tilde{g}(z) = 2\tilde{f}(z)/z$ is a function of z^2 so it has only even power of z because the walker can return to the original level only with an even number of steps. The generating function of $P(M^+, M^-|N)$ (obtained summing (2.73) over all the ages) is therefore:

$$\tilde{P}(M^+, M^-|z) = \frac{2}{z}\, [\tilde{f}(z)]^{M^+ + M^- - 1}\,; \tag{2.75}$$

again, it can not be inverted analytically, while there is an exact formula for the covariance between forward and backward records. Skipping some calculations:

$$\sum_{N=0}^{\infty} \langle R_N^+ R_N^- \rangle = \frac{2\tilde{f}(z)}{z} \frac{1}{(1-\tilde{f}(z))^{-4}} = \frac{1+\sqrt{1-z^2}}{2(1-z)^2}; \tag{2.76}$$

$$\langle R_N^+ R_N^- \rangle = \frac{1}{2}\left[N+1+(-1)^{r_N+1}(2r_N-1)(2r_N+2r_{N+1}+1)\binom{1/2}{r_N}\right]; \tag{2.77}$$

where $r_N = \lfloor N/2 \rfloor$, the floor function of half N. The second part is subdominant and in the large N limit $\langle R_N^+ R_N^- \rangle \sim N/2 + \sqrt{2N/\pi}$. The leading term is one half of that in the continuous case as forward and backward records have a $\sqrt{2}$ scaling factor each. Coherently the expansion at the first order in z of the probability (2.75) and a successive inversion gives:

$$P(M^+, M^-|N) = \frac{2}{N} h\left(\frac{M^+ + M^-}{\sqrt{N/2}}\right); \tag{2.78}$$

where $h(x)$ is the scaling function of the probability of the number of forward and backward records in the continuous case, as defined in (2.64).

Simulations We performed simulations with 10^5 samples of 10^3 steps, calculating the relevant quantities, and found good agreement with the theory for every time step. Moreover, as in the continuous case, we simulates series of 10^4, 10^5 and 10^6 steps, using 10^5 samples for each of them and computed the statistics of the number of records, which collapse in the curves $g(x)$ (forward case) and $h(x)$ (forward-backward) after the rescaling proper of this case, similarly to Figure (2.3).

Enforcing correlation As we specified at the beginning, there are many different models of strong correlated variables. It is interesting to have an idea of how a different kind of correlation can modify the model we found, in order to address new efforts of research towards non trivial situations. We can try to enforce the correlation between subsequent steps of the random walk considering for example a short range persistence: if a step is positive, the following one has a probability slightly bigger than one half to be positive too, and slightly smaller for a negative step. A simple simulation shows that in this case the number of records (both forward and backward) is bigger than in the simple case, but keeps the square root dependence on N. If instead we simulate a long range correlation, for example enforcing the probability that step i is positive or decreasing it according to the sign of the i^{th} Ising variable on a 1D lattice with antiferromagnetic long range interaction (spins tend to a long range alignment), we see that the number of forward records approaches a linear N dependence with some fluctuations which retains the long range character, while backward ones are decreased. Therefore a non local change such as this breaks the universality changing the whole behaviour of records and it would be interesting to work out the correlation length for which the switch of the N dependence happens.

2.4 An application: climate records

In this section, we would like to give the general features of one of the most relevant applications of statistics of records: the understanding of climate behaviour.

The global climate is a very complex system, determined locally by thousands of degrees of freedom. In the last ten years the problem of global warming has raised the issue of estimating quantitatively whether the average temperature is raising or not on our planet. The lack of data from several zones in the world, the local fluctuations and the short window of data available (only in the last century temperatures have been systematically recorded) prevent to rely only on the simple temperatures average to evaluate the trend. The search of quantities sensible to trends in mean brought in the past years to a wider study of the record temperatures, both in the sense of high and low records, or forward and backward ones. These tools have proved themselves powerful in a wider range of problems than the pure mean trend one.

Before proceeding with the sample analysis, we highlight briefly how records can provide information in problems concerning climate (the quantitative computations can be found in [18]-[23]) and also what problems arise when dealing with measured data. First of all notice that a change of the distribution of the variables with time destroys the universality which characterises the records of a sequence in a stationary framework. Therefore the attempts to find the distribution of records in a changing environment have always been done choosing a specific distribution for the temperatures. Very often we look at changes on years time scale, so a suitable sample of data is the temperature at a fixed day of the year, in different years. They are Gaussian distributed around the typical value of the temperature in that day of the year, as emerges from several analysis (see for example [23]). The change of the distribution in time due to the warming is small, thus it is possible to recognise the Gaussian shape at first approximation.

It is intuitive that a **constant increase in the mean** of the temperatures increases the number of forward upper records and decreases the forward lower ones, while for the backward it holds the opposite. It can be seen that for a Gaussian distribution a mean value growing like vi, where i is the time step and v a constant, at first order in small v increases the record rate as $\langle \sigma_i \rangle = 1/i + C\, v\sqrt{\ln i}$ with C constant, until saturation is reached if v is positive (and the number of records approaches $R_N \propto N$), or a rate 0 if v is negative ($R_N \to const$) [19][22]. A **change in the variance** of the distribution gives instead smaller effects: if it increases as qi, q positive, the rate becomes $\langle \sigma_i \rangle \propto \ln i/i$ (and $R_N \propto (\ln N)^2$) [18].

Another important issue comes from the **finite precision** with which data are recorded: it decreases the number of records more and more with the growth of the discrete step Δ. As for a Gaussian distribution the effects of ties and linear growth in mean gives corrections at the number of records which bear the same N-dependence, given Δ, for some values of v the two contributions are comparable, and are summed or erased according to the type of record studied (ties always decrease the number, while the trend has different effects). This is the case of the present global warming rate when full degree rounding is chosen.

Following the above considerations, to evaluate the mean trend a good object is the difference between upper and lower records, rather than the comparison with a stationary theory, as the ties effect is erased on average. In this context also backward records have been analysed [22], but they can be more effectively employed in the study of climate variability: looking for the annual one, the point is whether temperatures are going to smooth their behaviour, approaching the seasonal oscillation closer and closer or it is expected an increase of big fluctuations. It is thus required an estimate insensible to mean shift, and possibly of easier study than the variance of the sample itself. Anderson and

Kostinski [25] built and index α including backward records to estimate the variability, which is both insensible to ties (as it is made by differences of records) and to first order effects of the mean trend (as it is a sum of upper and lower corresponding quantities): $\alpha = U^+ - U^- + L^+ - L^-$, where U and L denotes upper and lower, while the apexes are for the forward and backward direction. Indeed, by simulations, they proved that the influence of the variance of temperature distribution is the dominant one, and employ this object to find a diminishing temperature variability on the globe.

We want now to test the results on the statistics of forward and backward records on a real sample of data. To compute their joint probability in an evolving framework is a task beyond our present treatment, but we can take data and subtract the deterministic behaviour we suspect they have, in order to recover a purely stochastic process with stationary distribution and verify the statistics of this new sample.

2.4.1 Data processing

We choose the average daily temperature in eighteen European cities, recorded for 100 years (from 1901 to 2000). The data, recorded in tenths of degrees Celsius, are provided by the *European Climate Assessment and Dataset* database (ECAD) [27]. A set of temperatures during the years shows mainly two features: every year the seasonal trend follows approximately a sinusoid, and over a large gap of years it is possible to see the increasing of the temperatures due to global warming and local factors. As anticipated, the temperatures at a fixed day of the year, recorded at different years, are distributed at first approximation as a Gaussian, and we expect them to be uncorrelated once subtracted the global warming. Therefore, taking as samples the sequences of detrended fixed-day temperatures in function of the year, they are supposed to show an i.i.d. behaviour. To compare the correlation inter and intra samples is useful to shift also all the sequences of their mean values, as explained in the following.

For a meteorological station, we define $T(d, y)$, the temperature at a given day d and year y. In our analysis we consider 100 years (N_y) of 365 days each (N_d), and 18 stations. For every station first of all we compute the mean annual temperature $A(y)$ to have a simple estimate of the warming trend:

$$A(y) = \frac{1}{N_d} \sum_{d=1}^{N_d} T(d, y)\,; \tag{2.79}$$

in Figure (2.4), is shown the behaviour of $A(y)$ for one of the stations: it is highly oscillatory, but we can recognize a slow increase of the temperatures with the time. Subtracting the respective average annual temperature from all the data we have the new set $z(d, y) = T(d, y) - A(y)$: fixing a day, every sample of z-temperatures is made of oscillations around a different value $G(d)$ (the typical deviation of the temperature at that day of the year from the typical average annual one). This value is:

$$G(d) = \frac{1}{N_y} \sum_{y}^{N_y} z(d, y)\,. \tag{2.80}$$

We finally define the quantity $x(d, y) = z(d, y) - G(d)$, which is centred around zero both looking at the day dependence or at the year dependence and is suitable to compare easily the samples between them.

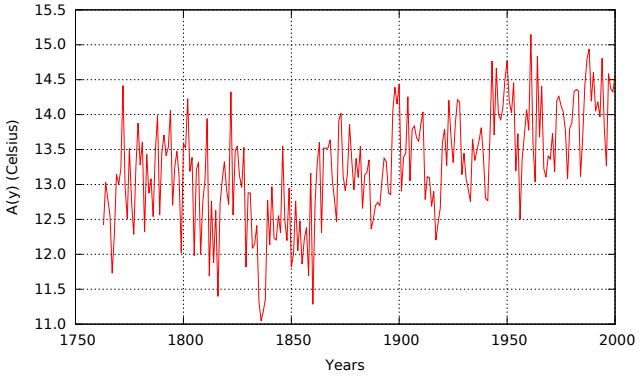

Figure 2.4: Station of Milano: the mean temperature for every year (averaged over 365 days).

Correlation between the data To see if the variables in a sample are independent, we can compute directly the correlation between them, i.e. between $x(d, y)$ at different years. If then their records will follow the i.i.d. statistics, it will be a further proof of their independence. The normalized correlation is:

$$D(i,j) = \frac{\frac{1}{N_d}\sum_{d=1}^{N_d} x(d,i)\,x(d,j)}{\langle \frac{1}{N_d}\sum_{d=1}^{N_d} x(d,y)^2 \rangle}\,, \quad i > j\,; \tag{2.81}$$

where the brackets means an average over the years (this is already the connected component of the correlation as the average over the days of $x(d, y)$ is zero). The normalization permits to have values of order one for $i = j$. To have an object easier to visualize, and to reduce the noise, we consider the correlation in function of the years gap Δy defined as:

$$D(\Delta y) = \frac{1}{N_y - \Delta y} \sum_{y=1}^{N_y - \Delta y} D(y, y + \Delta y) \tag{2.82}$$

where we sum over the initial year as long as $y + \Delta y$ still belongs to our set of data (this means that for large Δy the average is made on a few data only, according to $N_y - \Delta y$). With the chosen normalization, $D(0)$ is identically one. The first plot in Figure (2.5) represents $D(\Delta y)$ for one of the stations (Milano, which has a recorded period of 238 years, plot entirely) and it shows no correlation as soon as $\Delta y \neq 0$; the values for big Δy oscillate because of the poor statistic.

Then, is important to check the correlation between samples, which, in our case, is the correlation between subsequent days of the years. We defined the normalized correlation between two different days as:

$$C(\mu,\sigma) = \frac{\frac{1}{N_y-1}\sum_{y=1}^{N_y-1} x(\mu,y)\,x(\sigma,y)}{\frac{1}{N_y}\sum_{y=1}^{N_y} x(\mu,y)^2}\,, \quad \sigma > \mu\,; \tag{2.83}$$

as done by Redner and Petersen in [23]. With this normalization $C(\mu, \mu)$ is identically 1. As $\sigma > \mu$, for some μ (at the end of the year) σ can refer to the following year (the correlation decreases fast enough we can keep $\sigma - \mu < N_d$), that is why the sum end at $N_y - 1$. Again, we look at the more useful object of the average annual correlation in

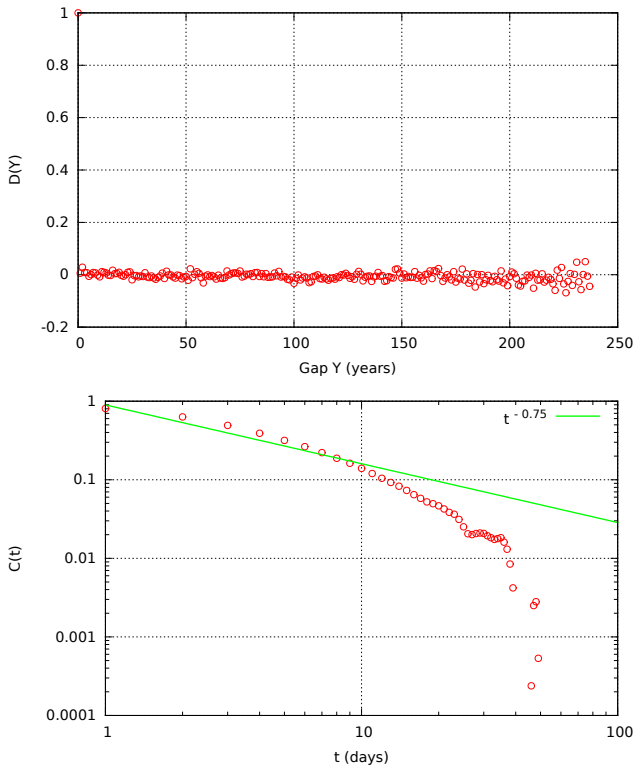

Figure 2.5: Station of Milano: above, the correlation between data recorded at the same day of the years in function of the years gap and averaged over the 365 days; below, the correlation between data referring subsequent days in function of the days gap, averaged over the years.

function of the days gap t:

$$C(t) = \frac{1}{N_d} \sum_{d=1}^{N_d} C(d, d+t)\,. \tag{2.84}$$

C(t), starting from the value 1, remains non zero until day 40. The first part of its behaviour can be described as a power law decay $C(t) \sim t^{-\gamma}$, as shown in the second plot of Figure (2.5), and it is possible to perform a fit over the first 10-20 days to find γ. According to the station, γ varies between 0.6 and 0.9: this values are around one half the corresponding ones found in [23], but have good agreement with other results ([26]). In particular in this last work it is found an average value of 0.7 for continental meteorological stations, and 0.8 for island ones. Considering the different data in literature and the variability between the stations, we look at our result as an esteem of the number of days in which the correlation is important (around 10).

The spatial correlation between the stations is not studied here. However, in [22], which uses the same data set here employed, it is found that the number of independent stations in Europe is around 15, as a distance inferior to 1000 km implies strong correlations on the data. On that basis, we repeated our calculations increasing the number of stations used from 3 to 64, finding that results do not improve considering more than 20 stations, according to their effective number. The final pool of chosen stations is quite well spread in Europe and was selected on the completeness of data in the time window considered.

2.4.2 Statistics of records

We can now analyse the sequences of $x(d, y)$ to compute the number of records for each day of the year. First of all we compute the average of the number and variance of records, and of the covariance between forward and backward ones summing over all the days and all the stations, so considering 365×18 samples.

To check if the correlation between following days influences the results, we repeat the average summing not over all the 365 sequences for every station but skipping some days between two subsequent ones (considering for example a gap of a week, there are 52 effective samples for each station). Gap periods between 2 and 20 days were considered, reminding that $C(t)$ goes to zero in about ten days. There is no significant improvement of the results rising the gap, as the noise increases too, due to the smaller number of data. Consequently we use the results from the 365 days average process.

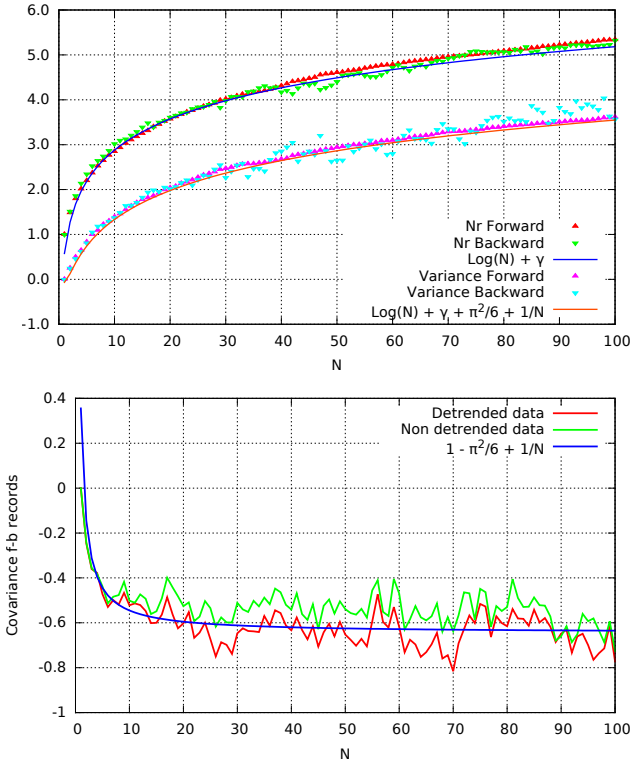

Figure 2.6: Above, the number and covariance of the forward and backward records of the detrended data; below, the connected component of the covariance between them for detrended and undetrended samples. The data are computed from all the 365×18 samples. Theoretical curves are superimposed.

We report in the first plot of Figure (2.6) the number of forward and of backward records in function of the time, and the connected part of their variance: $\langle R_N^i \rangle$, $\langle R_N^i R_N^i \rangle - \langle R_N^i \rangle \langle R_N^i \rangle$ with N the years passed from the beginning of the sequence. The agreement with the theory is good; forward records show a smoother behaviour as increasing the length N just add a step to the previous sequence, while in the backward case the starting point is shifted by one step. The second plot in Figure (2.6) shows the correlation $\langle R_N^+ R_N^- \rangle - \langle R_N^+ \rangle \langle R_N^- \rangle$: the computed quantity from the sample oscillates around the correct value $1 - \pi^2/6 + 1/N$. The plot shows also the same quantity computed for the

original samples, without subtracting the linear trend: it is clear a small shift with respect to the detrended curve which makes the results quite incompatible with the theoretical value predicted for i.i.d. variables. Moreover, for the un-detrended samples, the number of forward records grows a bit faster and the one of backward records a bit slower with respect to the detrended ones (not shown in the plot).

In conclusion, the average daily temperature in Europe in the last century, at a fixed day of the year, follows the statistics of a set of i.i.d. variables, once detrended from the global warming. In particular we showed that the records follow the expected i.i.d. behaviour and the correlation between forward and backward ones approaches the predicted constant value. Thus a simple subtraction of the annual average temperature can describe the deterministic part of temperatures behaviour, while the stochastic component is made of uncorrelated fluctuations.

Chapter 3

Records: statistics of the ages

In the previous section the focus was on the number of records of a given sequence long N. The other non trivial aspect of records statistics concerns how many steps a record survives, the so called age: the average age is simply given by the length of the sequence divided by the average number of records, but we can ask for the statistics of the longest age or the probability that the longest lasting record is exactly the maximum. The concept of age was already employed to obtain the statistic of the number of records for random walks and we recall here the definition. Given the sequence $\{x_i\}$, and two successive records x_i, x_j with $i < j$, the age of x_i is $l_n = j - i$ where n is the index of the record x_i (first record of the series, second one...). As the maximum is never surpassed, we must set a convention for its age, and this is usually done in slightly different ways according to the model in exam. In random walks we saw it is set to $l_M = N - m$, where m is the step at which the maximum happens and, for a sequence long N, $M = R_N$ is the number of (forward) records. As the first point of the sequence has by convention the index 0, i.e. $\{x_i\} = \{x_0, ..., x_N\}$, it holds $\sum_{i=1}^{M} l_i = N$. In the i.i.d. case instead $l_M = N - m + 1$; but here the first step has index 1, i.e. $\{x_i\} = \{x_1, ..., x_N\}$, so that the relation $\sum_{i=1}^{M} l_i = N$ is still valid. The ages of the backward records are taken naturally looking at the inverted time sequence, starting from $i = N$ and proceeding backward, as shown in Figure 2.1.

It must be noticed that l_M has not the same physical meaning as the other l_i: it is not defined by a record breaking but depends on the observed length N and, once fixed the other ages, it is fixed as the ages sum to N. Nonetheless, for forward records, it is useful to explore the ages statistics both including l_M or excluding it (an index I will refer to the first case and II to the second); while for forward-backward records, non trivial quantities are obtained only without considering the forward and backward ages of the maximum. Including them would give trivial results as, if the maximum level is reached only once, there is the further constraint that the forward age of the maximum is the sum of all the backward ages a part the one of the maximum (and plus one in the i.i.d. case) and viceversa. In this chapter, we are going to study extreme object for the forward cases I and II and the forward-backward case for i.i.d. variables and random walks with continuous jump distribution. For the lattice random walks we ask for some specific questions concerning the role of ties in this situation.

The statistics for the forward case is already well known [5], [12]. Remarkably, the mathematical treatment of the problem can be easily adapted to study other intervals problems, such as the ages of the zeros of a stochastic process ([13], [14]), which show strong analogies with the ages of records (qualitatively and even quantitatively). No

previous studies are known for the forward backward case instead.

3.1 I.I.D. variables

I.I.D.: longest ages of forward records We want to build for the i.i.d. model the joint probability of having M records in a N-sequence, each with an age l_i. We denote $\{\vec{l}, l_M\} = \{l_1, l_2, ..., l_M\}$, while $\vec{l} = \{l_1, l_2, ..., l_{M-1}\}$. As in the statistics of the number of records, it is useful to define a change of variable:

$$u_k = \int_{d_-}^{x_k} p(y)dy; \tag{3.1}$$

which is the probability of staying under the level x_k for one step. We will see that, as for the numbers, the specific distribution $p(x)$ will have no role in the statistics of ages. A record with level x_k survives l steps if the walker remains under x_k so far, but as every point - and so every step - is independent, the probability of the event factorizes giving $(u_k)^l$. This applies for every record of the sequence therefore the joint probability of the ages becomes:

$$\begin{aligned} P^I(\{\vec{l}, l_M\}, M|N) &= \int_{d_-}^{d_+} dy_M\, p(y_M) \left[\int_{d_-}^{y_M} p(x)\, dx\right]^{l_M-1} \\ &\quad \times \prod_{k=1}^{M-1} \int_{d_-}^{y_{k+1}} dy_k\, p(y_k) \left[\int_{d_-}^{y_k} p(x)\, dx\right]^{l_k-1} \delta_{\sum_{k=1}^M l_k, N} = \quad (3.2) \\ &= \int_0^1 du_M\, (u_M)^{\,l_M-1} \prod_{k=1}^{M-1} \int_0^{u_{k+1}} du_k\, (u_k)^{\,l_k-1}\, \delta_{\sum_{i=1}^M l_i, N} = (3.3) \\ &= \frac{\delta_{\sum_{i=1}^M l_i, N}}{l_1(l_1+l_2)\cdots(l_1+l_2+\cdots+l_M)}\,; \quad (3.4) \end{aligned}$$

where we integrated on the record levels, provided each of them is smaller than the successive ones. Again the change of variables gives an expression independent from the distribution, ensuring the universality of the results. The analogous object for the set $\vec{l}$ is the sum over l_M of (3.4):

$$P^{II}(\{\vec{l}\}, M|N) = \frac{\Theta\left(\sum_{i=1}^{M-1} l_i - (N-1)\right)}{l_1 \cdots (l_1 + \cdots + l_{M-1})N}; \tag{3.5}$$

and for $M = 1$ is defined as $P^{II}(\{\cdot\}, 1|N) = 1/N$ which is the probability that the first variable is the maximum.

The first interesting quantity we look at is the longest age: if the average age scales for large N like $\sim N/\ln N$, we will see that the longest one displays a behaviour linear in N. The results are clearly different if we look for the longest age among the set $\{\vec{l}, l_M\}$ or $\{\vec{l}\}$. The two possibilities will be explored, and we started from the first case (I) in which l_M is considered.

To get the distribution of $l_{max,N}$, consider the cumulative function $F(l|N) = \mathbb{P}(l_{max,N} \leq l)$. For case I, F^I is given by (3.3) summed over the number of records M up to N, and

over all the ages up to l. Once more, the task is easier taking the generating function. Using $F^I(l|\,0)=1$, and the delta function to distribute z^N:

$$\sum_{N=0}^{\infty} F^I(l|N)\, z^N = 1 + \sum_{M=1}^{\infty} \int_0^1 du_M\, f(u_M) \prod_{k=1}^{M-1} \int_0^{u_{k+1}} du_k\, f(u_k)\,; \tag{3.6}$$

$$f(u_k) = z \sum_{m=1}^{l} (zu_k)^{m-1}\,. \tag{3.7}$$

The function $f(u_k)$ is easily integrable: $g(u)=\int_0^u f(u')\,du' = \sum_{m=1}^{l} \frac{(zu)^m}{m}$. It follows:

$$\sum_{N=0}^{\infty} F^I(l|N)\, z^N = 1 + \sum_{M=1}^{\infty} \frac{(g(1))^M}{M!} = \exp\left(\sum_{k=1}^{l} \frac{z^k}{k}\right)\,. \tag{3.8}$$

The mean value of the longest age can be expressed as $\langle l_{max,N}\rangle^I = \sum_{l=1}^{\infty}(1-F^I(l|N))$, so that:

$$\sum_{N=0}^{\infty} \langle l_{max,N}\rangle^I z^N = \frac{1}{1-z} \sum_{l=1}^{\infty} \left[1 - \exp\left(-\sum_{k=l+1}^{\infty} \frac{z^k}{k}\right)\right]\,. \tag{3.9}$$

The analysis of this expression for large N is done taking the limit $z=e^{-s}\to 1$ where sums can be replaced by integrals: a rescale of the variables ($x=sl$; $y=sm$) gives a s^{-2} dependence which means $\langle l_{max,N}\rangle^I \sim c_I N$; the constant c_I is given by:

$$c_I = \int_0^{\infty} dx\,(1 - e^{-\int_x^{\infty} dy \frac{e^{-y}}{y}}) = 0.62432... \tag{3.10}$$

The calculations are more complicated for case II: as the sum of the ages in $\vec{l}$ has not a fixed value, the sum over N in the generating function can not be easily eliminated. However, instead of using $P^{II}(\vec{l},M|N)$, it is possible to recover $\langle l_{max,N}\rangle^{II}$ developing the formalism for the forward and backward case as in the following paragraph and integrating it over the backward ages.

I.I.D.: longest ages of forward and backward records Given M^+ forward record and M^- backward ones in a N-sequence, we define the forward and backward ages respectively:

$$\vec{l}=\{l_1,...,l_{M^+-1}\}\,;\quad \vec{k}=\{k_1,...,k_{M^--1}\}\,. \tag{3.11}$$

$$\sum_{i=1}^{M^+-1} l_i + \sum_{j=1}^{M^--1} k_j = N-1 \tag{3.12}$$

As anticipated, the analysis focusses on these sets only, where the ages of the maximum are not included. In the following, when speaking of both forward and backward ages, this is the understood convention.

Consider first the probability of $\{\vec{l},\vec{k}\}$ at fixed maximum in position m: it factorizes in the probabilities of each of the two sets in the respective sub-sequences, which are independent. The unconditioned distribution is given summing it over m, which has probability $1/N$ of hosting the maximum:

$$P(\{\vec{l},\vec{k}\},M^+,M^-|N) = \frac{1}{N}\sum_{m=1}^{N} P^I(\vec{l},M^+-1|m-1)\,P^I(\vec{k},M^--1|N-m)\,. \tag{3.13}$$

In the right hand side we use formula (3.3) for the probability (case I) as in the single sub-sequence also the ages l_{M^+-1} or l_{M^--1} is considered. We explicit it and also introduce a fictive variable v to get rid of the denominator N and express it as an integral. This variable corresponds to the level of the maximum. Calling for simplicity $\mu^+ = M^+ - 1$ and $\mu^- = M^- - 1$:

$$P(\{\vec{l},\vec{k}\}, M^+, M^-|N) = \int_0^1 dv \int_0^v du_{\mu^+} u_{\mu^+}^{l_{\mu^+}-1} \prod_{i=1}^{\mu^+-1} \int_0^{u_{i+1}} du_i u_i^{l_i-1} \times \int_0^v du_{\mu^-} u_{\mu^-}^{k_{\mu^-}-1} \prod_{j=1}^{\mu^--1} \int_0^{u_{j+1}} du_j u_j^{k_j-1} \delta_{\sum_{i=1}^{\mu^+} l_i + \sum_{j=1}^{\mu^-} k_j, N-1}. \quad (3.14)$$

Once again it is independent from the distribution. As in the previous case, it is interesting to calculate the longest age $l_{abs,N}$ among the two sets $\{\vec{l},\vec{k}\}$. First of all we compute the cumulative function that all the forward ages are smaller than l and the backward ones than k, summing (3.14) over l_i until l, k_i until k and over the two numbers of records until N. Considering the generating function, the computations are similar to what already done in the forward case except for the last degree of freedom v:

$$\sum_{N=0}^{\infty} F(l,k|N)\, z^N = 1 + \int_0^1 dv\, \frac{z}{(1-zv)^2}\, e^{-\overline{g_l}(v) - \overline{g_k}(v)}; \quad (3.15)$$

$$g_p(v) = \sum_{m=1}^{p} \frac{(zv)^m}{m}; \quad \overline{g_p}(v) = \sum_{m=p+1}^{\infty} \frac{(zv)^m}{m}; \quad (3.16)$$

the term 1 comes from $N = 0$ (in the forward case it was absorbed to recover the full expression of the exponential). We must pay attention to the case $\mu^+ = \mu^- = 0$: it contributes to F with z for $N = 1$, while for $N > 1$ it represent a "flat case", which has probability zero (in a continuous model). However it is re-expressed in terms of a v integral and absorbed in the exponential. In the limit $l, k \to \infty$, (3.15) gives $\frac{1}{1-z}$ which means, correctly, $F(\infty,\infty|N) = 1$. The average of the absolute longest age (among forward and backward ones) is the sum over l of $1 - F(l,l|N)$; its generating function is:

$$\sum_{N=0}^{\infty} \langle l_{abs} \rangle\, z^N = \sum_{l=1}^{\infty} \int_0^1 dv\, \frac{z}{(1-zv)^2} \left(1 - e^{-2\sum_{m=l+1}^{\infty} \frac{(zv)^m}{m}}\right). \quad (3.17)$$

We consider the continuous limit replacing the sums with integrals, and perform the changes of variables $x = A\, s\, l$; $y = A\, s\, m$ with the definition $A = 1 - \frac{\ln v}{s}$ (and $z = e^{-s} \to 1$). Keeping the inferior extreme of integration in x at zero (as in the continuous limit l starts from zero) gives the result $\langle l_{abs,N} \rangle \sim c_{abs} N$, with:

$$c_{abs} = \frac{1}{2} \int_0^{\infty} dx \left(1 - e^{-2\int_x^{\infty} dy \frac{e^{-y}}{y}}\right) = 0.475639.... \quad (3.18)$$

The general moment of the absolute longest age goes like $\langle l_{abs,N}^n \rangle \sim L_n N^n$:

$$L_n = \frac{1}{(n+1)!} \int_0^{\infty} dx\, x^{n-1} \left(1 - e^{-2\int_x^{\infty} dy \frac{e^{-y}}{y}}\right). \quad (3.19)$$

Coming back to the forward case II, we ask for the longest age in the $\vec{l}$ set. Its average value is given reproducing the calculations for $\langle l_{abs,N}\rangle$ but using $F(l,\infty|N)$, so "integrating" on the backward ages. The result is:

$$\langle l_{max}\rangle^{II} \sim c_{II}N = \frac{c_I}{2} = 0.31216...N\,; \tag{3.20}$$

but there is no simple physical explanation for the relation between the two cases.

In the forward-backward case is also interesting the correlation between the longest forward age (in the set $\vec{l}$) $\langle l^f_{max,N}\rangle^{II}$ and the longest backward one (in $\vec{k}$) $\langle l^b_{max,N}\rangle^{II}$. We adapt the formula of the discrete derivative:

$$\langle l^f_{max,N}\, l^b_{max,N}\rangle \;=\; \sum_{l=1}^{\infty}\sum_{k=1}^{\infty}(1+F(l,k|N)-F(l,\infty|N)-F(\infty,k|N))\,; \tag{3.21}$$

with $F(l,k|N)$ from (3.15). The analysis at large N gives $\langle l^f_{max,N}\, l^b_{max,N}\rangle \sim c_{fb}N^2$, with $c_{fb}=c_I/6$ so that the covariance between the two object is negative and takes asymptotically the value $-c_IN^2/12$.

I.I.D.: last age behaviour Another interesting quantity is the probability that the last age is longer than all the previous ones (independently from the number of records). We analyse this object for the three cases treated. Starting from the forward case I, the probability that l_M is bigger than every other age in $\vec{l}$ (the maximum is reached early in the sequence) is:

$$Q^I(l_M) = P(l_M > l_1,...,l_{M-1}|N) = \sum_{l_M=1}^{N}\sum_{l_1=1}^{l_M}\cdots\sum_{l_{M-1}=1}^{l_M}\sum_{M=1}^{\infty}P^I(\{\vec{l},l_M\},M|N). \tag{3.22}$$

As usual the large N behaviour is extract by generating function; using (3.3) in (3.22):

$$\sum_{N=0}^{\infty}Q^I(l_M)\,z^N \;=\; 1+\int_0^1 du_M\, z\sum_{l_M=1}^{\infty}(zu_M)^{l_M-1}\exp\left(\sum_{m=1}^{l_M}\frac{(zu_M)^m}{m}\right)\,; \tag{3.23}$$

Passages similar to the ones that brought to (3.18), now with $A=1-\frac{\ln u_M}{s}$, gives:

$$\sum_{N=0}^{\infty}Q^I(l_M)\,z^N \;=\; 1+\frac{1}{s}\int_0^1 du_M\int_0^{\infty}dx\,\frac{e^{-x}}{A\,u_M}\,\frac{e^{-\int_x^{\infty}dy\frac{e^{-y}}{y}}}{1-e^{-As}}\;\stackrel{s\to 0}{\longrightarrow}\;\frac{c_I}{s}\,; \tag{3.24}$$

where the integral in x gives c_I and the one in u_M gives one when solved for $e^{-s}\sim 1-s$. Therefore the probability that the last age is the longest in the large N limit is c_I, equal to the average longest age divided by N or, roughly speaking, to its derivative in N. Indeed, adding a step at the end of the sequence, the probability that the longest age increases of one ($\Delta l_{max,N}$=1) is exactly Q^I, as it happens when the last age is already the longest. With probability $(1-Q^I)$, instead, the longest age remains constant, so the average $\Delta l_{max,N}$ gives Q^I.

To find the result without considering the last age, so the probability that $l_{\mu^+} > l_1,...,l_{\mu^+-1}$ (isolated maximum with respect to the previous records), we consider the

forward-backward formalism and sum over the backward ages up to infinity. Using the correspondent probability (3.14) the leading term is:

$$\sum_{N=0}^{\infty} Q^{II}(l_{\mu^+})\, z^N = \frac{1}{s}\int_0^1 dv\, \frac{e^{-s}}{1-e^{-s}\,v}\, \int_0^v du_{\mu^+} \int_0^\infty dx\, \frac{e^{-x}}{A\, u_{\mu^+}}\, \frac{e^{-\int_x^\infty dy \frac{e^{-y}}{y}}}{1-e^{-As}} \overset{s\to 0}{\longrightarrow} \frac{c_I}{s}; \tag{3.25}$$

and the subdominant ones from $N = 0, 1$ are constant or logarithmic divergences. In the integral in v, we recognize a factor similar to the one in (3.24) and a prefactor coming from the integration of the backward part. $Q^{II}(l_{\mu^+})$ turns out to be equal to $Q^I(l_M)$: this can be explained considering that the variables are independent, so every age of the set can be the longest with equal probability.

Finally in the forward-backward case we look at the probability that l_{μ^+} is the longest in the $\vec{l}$ set and k_{μ^-} the longest in $\vec{k}$ at the same time (records gather at the two extremes of the sequence, but for the maximum which is in the middle). Neglecting the constant and s^{-1} terms from $N = 0$ and $N = 1$, the leading one is:

$$\sum_{N=0}^{\infty} Q(l_{\mu^+}, k_{\mu^-})\, z^N = \frac{1}{s^2}\int_0^1 dv\, z \left(\int_0^v du_{\mu^+} \int_0^\infty dx\, \frac{e^{-x}}{A\, u_{\mu^+}}\, \frac{e^{-\int_x^\infty dy \frac{e^{-y}}{y}}}{1-e^{-As}} \right)^2 \overset{s\to 0}{\longrightarrow} \frac{(c_I)^2}{s}\,; \tag{3.26}$$

The result points out that for i.i.d. variables, considering the two sequences on the left and on the right of the maximum, the statistic of the internal order of the ages is uncorrelated, so it is not affected by the fact they have their last point in common and must stay under it.

Simulations We verified the analytical results for the statistic of the ages of records simulating a sequence of i.i.d. variables with flat, Gaussian and exponential distribution. For the average longest ages behaviour we used 10^4 samples of 10^3 steps, calculating for every step the average forward longest age, with convention I and II; the absolute longest age in the set $\{\vec{l}, \vec{k}\}$; the correlation between $\langle l^f_{max,N}\rangle^{II}$ and $\langle l^b_{max,N}\rangle^{II}$. There is excellent agreement with the theory for every step N. For the probability that the last age is the longest we used 10^4 samples of $2 \cdot 10^3$ steps: we computed the fraction of trajectories for which the last age was the longest, for each case we treated. They all reach the correct constant, but with different relaxation times: the forward case including the last age relaxes in about 200 steps, faster than the case without it (about 400 steps). Finally the forward backward case reaches the asymptotic value around step 800. This last three quantities are shown in Figure (3.1).

3.2 Random walks with continuous jump distribution

R.W.: longest age In the precedent section we already built the joint probability of the ages and number of records for the random walk; for the forward and forward backward cases they are respectively:

$$P^I(\{\vec{l}, l_M\}, M|N) = f(l_1)f(l_2)\cdots f(l_{M-1})\, q(l_M)\, \delta_{\sum_{i=1}^M l_i, N}\,; \tag{3.27}$$

$$P(\{\vec{l}, \vec{k}\}, M^+, M^-|N) = f(l_1)\cdots f(l_{\mu^+})\, f(k_1)\cdots f(k_{\mu^-})\, \delta_{\sum_{i=1}^{\mu^+} l_i + \sum_{j=1}^{\mu^-} k_j, N}\,. \tag{3.28}$$

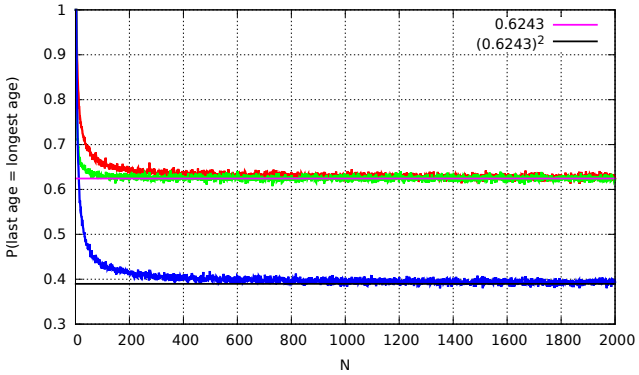

Figure 3.1: The probability that the last age is the longest in the forward case excluding the age of the maximum (red), including it (green), and, below, the forward backward case (blue). For every quantity it is shown the theoretical curve, superimposed to the one obtained by simulation for Gaussian jump distribution.

with $\mu^{\pm} = M^{\pm} - 1$; $P^{II}(\vec{l}, M|N)$ (the forward case without the maximum age) is obtained integrating (3.27) over l_M.

We start studying the behaviour at large N of the longest age in the set $\{\vec{l}, l_M\}$. As in the i.i.d. case we calculate $F^I(l|N)$, the cumulative probability that all the ages, including the last one, are smaller than l. It is computed summing (3.27) over all l_i up to l and over the number of records M up to $N+1$. Applying the delta function to solve the sum in N, the generating function of $F^I(l|N)$ is:

$$\sum_{N=0}^{\infty} F^I(l|N)\, z^N = \sum_{M=1}^{\infty} \left[\sum_{m=1}^{l} f(k)\, z^m\right]^{M-1} \sum_{m=1}^{l} q(m)\, z^m = \frac{\sum_{m=1}^{l} q(m)\, z^m}{1 - \sum_{m=1}^{l} f(m)\, z^m}\,. \quad (3.29)$$

By definition $f(j) = q(j-1) - q(j)$ and $q(0) = 1$ so, using $z = e^{-s}$, the denominator can be written as:

$$q(l)\, e^{-sl} + (1 - e^{-s}) \sum_{m=0}^{l-1} q(m)\, e^{-sm}. \quad (3.30)$$

In the small s limit the sum in (3.30) becomes an integral; after the change of variable $y = sm$ (s small) the persistence $q(y/s)$ can be approximated as $1/\sqrt{\pi y/s}$ (large y/s). The integral gives therefore $erf\sqrt{sl}/\sqrt{s}$. The same result holds for the numerator in (3.29). We know that $\langle l_{max,N}\rangle^I = \sum_{l=1}^{\infty}(1 - F^I(l|N))$, and applying the above results with $1 - e^{-s} \sim s$, its generating function is:

$$\begin{aligned}\sum_{N=0}^{\infty} \langle l_{max,N}\rangle^I\, z^N &= \sum_{l=1}^{\infty}\left[\frac{1}{s} - \frac{erf\sqrt{sl}}{\sqrt{s}\, q(l) e^{-sl} + s\, erf\sqrt{sl}}\right] \\ &\overset{y=sl}{=} \frac{1}{s^2}\int_0^{\infty} dy\, \frac{1}{1 + \sqrt{\pi\, y}\, e^y\, erf\sqrt{y}} = \frac{1}{s^2}\, 0.6265...\,; \end{aligned} \quad (3.31)$$

where again the asymptotic behaviour of $q(x)$ was used. Therefore $\langle l_{max,N}\rangle^I \sim b_I N$, with $b_I = 0.6265...$ which is a little bigger than the correspondent result $c_I = 0.6243...$ for the i.i.d. variables.

The average of the absolute longest age of the set $\{\vec{l}, \vec{k}\}$ is $\langle l_{abs,N}\rangle = \sum_{l=1}^{\infty}(1 - F(l, l|N))$ with $F(l, k|N)$ the cumulative probability that all the forward ages are smaller than l and

the backward ones than k. It is obtained from (3.28) summing l_i up to l, k_i up to k, and μ^+, μ^- up to N. The generating function is:

$$\sum_{N=0}^{\infty} F(l,k|N)\, z^N = \sum_{\mu^+,\mu^-=0}^{\infty} \left[\sum_{m=1}^{l} f(m)\, z^m\right]^{\mu^+} \left[\sum_{n=1}^{k} f(n)\, z^n\right]^{\mu^-} = \\ = \frac{1}{\left(1-\sum_{m=1}^{l} f(m)\, z^m\right)\left(1-\sum_{n=1}^{k} f(n)\, z^n\right)}\,. \tag{3.32}$$

Passages similar to the ones done before lead for the absolute longest age to:

$$\sum_{N=0}^{\infty} \langle l_{abs,N}\rangle\, z^N = \frac{1}{s^2}\int_0^{\infty} dy\, \left[1-\frac{\pi\, y\, e^{2y}}{(1+\sqrt{\pi\, y}\, e^y\, erf\sqrt{y})^2}\right] = \frac{1}{s^2}\, 0.4035...\,; \tag{3.33}$$

and again we confront this coefficient $b_{abs} = 0.4035$ with $c_{abs} = 0.4756...$ for the i.i.d. case. In this case the coefficient is smaller.

Finally, the average longest forward age excluding the one of the maximum is obtained from $\langle l_{max,N}\rangle^{II} = \sum_{l=1}^{\infty}(1-F(l,\infty|N))$. The limit $k\to\infty$ in (3.32) gives a factor $(1-\tilde{f}(z))^{-1} \sim 1/\sqrt{s}$, and the result is:

$$\sum_{N=0}^{\infty} \langle l_{max,N}\rangle^{II}\, z^N = \frac{1}{s^2}\int_0^{\infty} dy\, \left[1-\frac{\sqrt{\pi\, y}\, e^y}{1+\sqrt{\pi\, y}\, e^y\, erf\sqrt{y}}\right] = \frac{1}{s^2}\, 0.2417...\,; \tag{3.34}$$

which has no a simple relation with b_I, while the correspondent c_{II} in the i.i.d. case was simply $c_I/2$. This suggest that in the random walk case records are close one to the other until the maximum, as a consequence of the persistence of the walk itself.

R.W.: last age behaviour As for i.i.d. variables, the other object we look at is the probability that the last forward age is the longest, defining it with the two conventions, and the joint probability that l_{μ^+} is the longest in $\vec{l}$ and k_{μ^-} the longest in $\vec{k}$.

In forward case I, the probability $Q^I(l_M)$ that l_M is bigger than all the others is computed from (3.27) summing over all possible M and l_M, and summing the other ages up to l_M. We have:

$$\sum_{N=0}^{\infty} Q^I(l_M)\, z^N = \sum_{l_M=1}^{\infty} \frac{q(l_M)\, z^{l_M}}{1-\sum_{m=1}^{l_M} f(m)\, z^m}\,. \tag{3.35}$$

Doing the same passages as before in the denominator, the expression becomes:

$$\sum_{N=0}^{\infty} Q^I(l_M)\, z^N = \frac{1}{s}\int_0^{\infty} dy\, \frac{1}{1+\sqrt{\pi\, y}\, e^y\, erf\sqrt{y}} = \frac{1}{s}\, b_I\,. \tag{3.36}$$

As explained in the i.i.d. case, $Q^I(l_M)$ is just the derivative of $\langle l_{max,N}\rangle^I$. The analogous calculation for the last age in the set $\vec{l}$ (excluding the age of the maximum) uses the forward backward probability (3.28): $Q^{II}(l_{\mu^+})$ comes from its sum over μ^+, μ^- up to N, l_{μ^+} and all k_i up to infinity, while the other l are summed up to l_{μ^+}. Moreover we set a convention for the case in which the maximum appears in the first position (step 0): we choose to consider it among the ones to be counted in $Q^{II}(l_{\mu^+})$, as in this way the probability is

normalized. Retaining it or not gives a difference detectable in the simulations: indeed, we are going to see that $Q^{II}(l_{R^+})$ goes to zero with N, so we calculate not only the zero order of its expression (unaffected by the marginal situation) but also the first correction, which is of the same order of the term given by the particular case said. Before computing it, we underline the difference between random walk and i.i.d. variables: in the first, as x_i is given by $x_{i-1} + \xi_i$, usually the walker does not change its position of a big quantity in one jump and we expect that a record is followed soon by another one. Roughly, this is why the maximum follows the previous record after few steps, so that the age between them has little probability of being the longest, while in the i.i.d. case the probability was the same for each age thanks to the independence of variables. The expression for $Q^{II}(l_{\mu^+})$ is:

$$\sum_{N=0}^{\infty} Q^{II}(l_{\mu^+})\, z^N = \frac{1}{1-\tilde{f}(z)} \left[1 + \sum_{l_{\mu^+}=1}^{\infty} \frac{f(l_{\mu^+})\, z^{l_{\mu^+}}}{1 - \sum_{m=1}^{l_{\mu^+}} f(m)\, z^m} \right] . \tag{3.37}$$

As said, we want to perform a more precise analysis than before. The denominator of the non trivial term in the parenthesis, which we call $S(z)$, tends to $q(l_{\mu^+})$ when $z \to 1$. Indeed, as $f(l) = q(l-1) - q(l)$:

$$1 - \sum_{m=1}^{l_{\mu^+}} f(m)\, z^m = 1 - \tilde{f}(z) + \sum_{m=l_{\mu^+}+1}^{\infty} f(m)\, z^m = \sqrt{1-z} + q(l_{\mu^+})\, z^{l_{\mu^+}} ; \tag{3.38}$$

but replacing it with $q(l_{\mu^+})$ the sum has an exact expression:

$$J(z) = \sum_{i=1}^{\infty} \frac{f(i)}{q(i)}\, z^i = \sum_{i=1}^{\infty} \frac{z^i}{2i-1} = \frac{1}{2}\sqrt{z}\, \ln\left[\frac{(1+\sqrt{z})^2}{1-z} \right] \sim \ln 2 - \frac{1}{2}\ln(1-z) . \tag{3.39}$$

The correction $S(z) - J(z)$ can now be computed in an approximate way substituting the sum with integrals in the limit $z \to 1$. Considering $q(x) \sim 1/\sqrt{\pi x}$, $f(x) \sim 1/(2x\sqrt{\pi x})$ and computing the denominator in $S(z)$ through (3.30):

$$S(z) - J(z) = \int_0^{\infty} dy\, \frac{e^{-y}}{2y} \left[\frac{e^y}{1 + \sqrt{\pi\, y}\, e^y\, erf\sqrt{y}} - 1 \right] = \frac{1}{2}(\gamma - \ln \pi) . \tag{3.40}$$

The last two equations together injected in (3.37) gives the generating function of the probability, the inverse of which decays non trivially with N:

$$\sum_{N=0}^{\infty} Q^{II}(l_{\mu^+})\, z^N = \frac{1}{2\sqrt{s}}\, (2 + \gamma - \ln \pi + 2\ln 2 - \ln s); \tag{3.41}$$

$$Q^{II}(l_{\mu^+}) = \frac{\ln N + a_0}{2\sqrt{\pi\, N}} ; \quad a_0 = 2(1 + \gamma + 2\ln 2) - \ln \pi \sim 4.78229\ldots . \tag{3.42}$$

The forward backward case can be inferred from the previous one: in (3.37) the term $(1-\tilde{f})^{-1}$ comes from the integration over all the backward ages; considering instead their sum up to k_{μ^-} it gives another $1 + S(z)$ factor, so we have:

$$\sum_{N=0}^{\infty} Q(l_{\mu^+}, k_{\mu^-})\, z^N = 1 + 2S(z) + S^2(z) ; \tag{3.43}$$

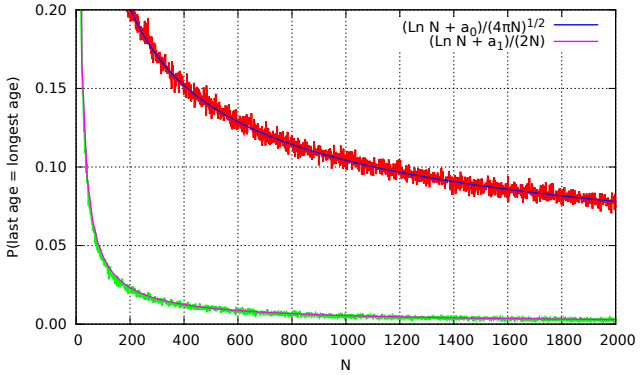

Figure 3.2: The probability that the last age is the longest, in the forward case excluding the age of the maximum and in the forward backward one. For every quantity it is shown the theoretical curve, superimposed to the one obtained by simulation for Gaussian jump distribution.

where 1 comes from $N = 0$ and $2S(z)$ from the cases $\mu^+ = 0$ or $\mu^- = 0$. This last term is not negligible in our analysis until the first correction. Knowing the expression for $S(z)$, we can invert (3.43):

$$Q(l_{\mu^+}, k_{\mu^-}) = \frac{\ln N + a_1}{2N}\,; \quad a_1 = 3\gamma - \ln\pi + 2\ln 2 + 2 \sim 3.97321\ldots\,. \tag{3.44}$$

It goes to zero more rapidly than the forward case, but less fast than its square: for i.i.d. variables $Q(l_{\mu^+}, k_{\mu^-}) = (Q^{II}(l_{\mu^+}))^2$, but here this relation does not hold as the variables are strongly correlated and so are the sets $\vec{l}$ and $\vec{k}$ and the internal order of the ages.

Simulations We simulated 10^5 samples of 10^3 steps each and calculated for every step the longest age in the three cases considered, using flat, Gaussian and exponential jump distributions. The simulations agree with the analytical results. For the probability that the last age is the longest, we used the same number of samples and 2000 steps, computing the fraction of trajectories for which the last age was the longest. We found good agreement with the theory, in particular for $Q^{II}(l_{\mu^+})$ and $Q(l_{\mu^+}, k_{\mu^-})$ the first order corrections are clearly detectable. Figure (3.2) shows a focus on these last two probabilities.

3.3 Lattice Random Walks

It was already pointed out how the lattice random walk approaches the behaviour of a continuous one with diffusion coefficient 1/2 in the large time limit, and how the statistics of the number of records takes the same form as in the continuous case up to a prefactor $\sqrt{2}$ for each degree of freedom. Rather than ask the same previous questions about the longest age, it is interesting to analyse better the possibility that a tie occurs, looking at the probability that the maximum level is reached more than once and, if it is the case, the probability that the age of the gap m which divide the "first" and the "last" maximum is longer than all the records ages, object which is not known in literature. Remember that the probability of having M^+ and M^- records with the respective sets of ages $\{\vec{l}, \vec{k}\}$, not

counting the age of the maximum, is (2.73):

$$P(\{\vec{l},\vec{k}\},M^+,M^-|N) = f(l_1)\cdots f(l_{M^+-1})f(k_1)\cdots f(k_{M^--1})\,g(m)\,\delta_{\sum_{i=1}^{M^+} l_i+\sum_{j=1}^{M^-} k_j+m,N}\,. \tag{3.45}$$

If the maximum is reached only once, it means $m = 0$. The probability of having more than one maximum is obtained summing (3.45) over all the ages and number of records, and over all m except 0. The generating function of $P(m = 0)$ then becomes:

$$\begin{aligned}\sum_{N=0}^{\infty} P(m=0) &= \sum_{M^+=1}^{\infty}\sum_{M^+=1}^{\infty}\sum_{m=1}^{\infty} g(m)\,z^m\left(\sum_{p=1}^{\infty} f(p)\,z^p\right)^{M^++M^--2} = \\ &= \frac{\tilde{g}(z)-1}{(1-\tilde{f}(z))^2} = \frac{1}{1-z} - \frac{1}{(1-\tilde{f}(z))^2}\,. \end{aligned}\tag{3.46}$$

In the first passage the factor z^N has been divided between the sums to recover the expression of $\tilde{f}(z)$ as usual, then we used the fact that $g(0) = 1$ and that the probability is normalized (the generating function of 1 is $(1-z)^{-1}$). Remember that $f(0) = 0$ so the sum over p gives $\tilde{f}(z)$ even if it does not start from 0. The inversion gives the probability in terms of gamma functions, as usual for the discrete case, but in the large N limit it reaches the constant $1/2$: in half of the case the maximum in unique. However, as we are going to see, the fraction of cases in which two maxima happens very distant is small: the probability that m is longer than all the ages of the records (both forward and backward) is given by the sum of (3.45) over all the ages up to m and over all m, M^+, M^-. It gives:

$$\sum_{N=0}^{\infty} P(m > l_1,...,l_{M^+-1},k_1,...,k_{M^-}|N) = \sum_{m=0}^{\infty}\frac{g(m)\,z^m}{(1-\sum_{p=1}^{m} f(p)\,z^p)^2}\,. \tag{3.47}$$

For an asymptotic analysis is sufficient to consider the large l expansion of the first passage probability (2.68): $f(l) \sim \sqrt{2/(\pi l)}$, which has the $\sqrt{2}$ factor of difference with respect to the continuous case. The function $g(m)$ is different from zero only for m even. Writing $m = 2r$, in the large r limit:

$$g(2r) = 2\,f(2r+1) \sim 2\,\frac{d}{dr}\left(\sqrt{\frac{2}{\pi 2r}}\right) \sim \frac{1}{\sqrt{\pi r^3}}\,; \tag{3.48}$$

so rearranging the denominator as usual (but with the correct $\sqrt{2}$ factors) it holds ($z = e^{-s}$, $y = 2rs$):

$$\sum_{N=0}^{\infty} P(m > \cdots|N) \xrightarrow{s\to 0} \sqrt{\frac{\pi}{2s}}\int_0^{\infty} dy\,\frac{e^{-y}}{\sqrt{y}}\,\frac{1}{(e^{-y}+\sqrt{\pi y}erf\sqrt{y})^2} = \sqrt{\frac{2}{s}}\,; \tag{3.49}$$

which means $P(m > \cdots|N) \to \sqrt{\frac{2}{\pi N}}$ for $N \to \infty$. Therefore for large N the probability that the maximum is reached in very distant steps is weak: if it was the contrary, the forward and backward records would have been less and less correlated, while we saw in the previous section that the correlation between their number approaches a constant as in the continuous case.

Chapter 4

Multiple random walks

Besides the ones already treated, there are many aspects of the statistics of records that we can explore. As this work focus on the joint statistics of forward and backward records, it is interesting to look at situations in which the relation between the two kinds of records is not trivial. The presence of a drift for example, changes the correlation between forward and backward records, which approaches one when the drift becomes infinite, but a positive drift has on forward records the same effect as a negative one on backward ones, thanks to symmetry. The statistics of the forward records alone in presence of a drift has already been studied: it is not a trivial problem as **the universality breaks** and there are different behaviours according to the variance of the jump distribution and the sign of the drift but asymptotic analytical results are available both for i.i.d. variables [19] and random walks [20],[21]. The first order effect is a linear increase (or decrease) of records with time, but the development of these result for the joint statistics of forward and backward records is difficult as they are based on the asymptotic expansion of the probability that a site is a record for N big: if a site i is big, the corresponding backward index $N - i + 1$ is small, invalidating the expansion in this case. As anticipated, the universality is broken also when considering a **rounding effect**: if every variable is rounded to a value multiple of a fixed step Δx, some ties occurs and some records can be suppressed as they exceed the previous record value of an amount less than a full step. It is shown ([16]-[17]) that in this case the number of records is suppressed in different ways according to the class of universality of the distribution (bounded support, power law tails or faster than power law) as it is important how the percentiles of the distribution are arranged.

The fact that for the examples above only the large time behaviour of the probability of being a record can be worked out makes difficult to find the correct expression for the correlation between forward and backward records and, as it provides no great information (in both situations records becomes uncorrelated as shown from simulations), we decide to focus on another problem.

Contrary to the previous cases, the symmetry between forward and backward records is broken in a non specular way when considering a number P of Brownian walkers starting from 0, with trajectories x_i^α, where i stands for the discrete time step and α is the label of the walker, $1 \leq \alpha \leq P$. We will call the overall length of the studied sequence with a small n. The records of a single walk respect the usual statistics but now we are interested in global records, which we define as follow. Consider, at a fixed time i, the maximum position of all the random walkers:

$$x_i^{max} = max[x_i^1, x_i^2, ..., x_i^P]; \tag{4.1}$$

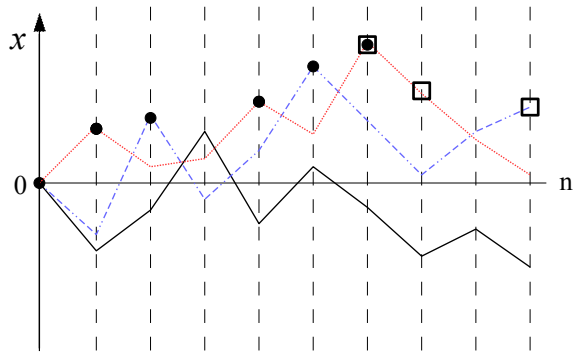

Figure 4.1: A realization of three random walks, all starting in zero and spreading in the space. Forward records are marked by full circles, backward ones by empty squares.

then a global record happens if x_i^{max} is bigger than all the previous maximum positions ($x_i^{max} > x_j^{max}$, $\forall\ 0 \leq j < i$). In other words we build the sequence of the maximal position among the walkers at every time and count the records in this sequence. We notice that the maximal positions at different time steps can come from different walkers as shown in Figure (4.1), which presents a realization for three walkers. Moreover, as this is a probabilistic process, we expect that the walkers spread in the space, causing an increased number of forward records as long as P grows. On the contrary, looking at the inverse sequence, all the paths converge toward zero, so we expect the number of backward records to decrease with P. In [15], it is studied the P dependence of the number of global forward records at time n, which we call $R^+_{n,P}$. It is shown that the universality is broken, according to the variance of the jump distribution, i.e. whether it is finite or not: indeed in the case of Lévy distributions, records are dominated by the occasional big jumps and this mechanism is not much influenced by the presence of many walkers while in the finite variance case the more the walkers are, the more the positions spread in space.

Following [15], the best approach to this problem is an ab-initio one, where we use the propagator of the random walk process and the persistence (survival) probability. The propagator $G(x, x_0, m)$ (the Green function of the single random walk) is the probability that a walker reaches the infinitesimal interval $[x, x + dx]$ after m time steps, starting at x_0 at time 0. In this situation we need it as we must compare the levels reached by the walkers, while in the single walk case the level was unimportant. As previously, the persistence $q_m(x)$ is the probability of staying under the level x, starting from 0, in the time interval from 0 to m. If x is zero, we know from the Sparre-Andersen theorem that $q_m(0) \sim 1/\sqrt{\pi\, m}$ for large m, whatever the jump distribution is. For a generic x instead we must specify whether the jump distribution has finite variance or not, as the objects differs in the two cases. The propagator differs as well. Before specifying them in the two cases, however, we can write a general formula for the number of global backward records using these two quantities:

$$\langle R^-_{n,P} \rangle = \sum_{m=1}^{n} r^-_{n,P}(m)\,; \tag{4.2}$$

$$r^-_{n,P}(m) = P\, q_{n-m}(0) \int_{-\infty}^{\infty} dx\, G(x, 0, m) \left[\int_0^{+\infty} dt\, G(x - t, 0, m)\, q_{n-m}(t) \right]^{P-1}. \tag{4.3}$$

The rate $r^-_{n,P}(m)$ is the probability that at time step m there is a backward record. This happens if a walker which has reached a level x in m starting from zero (without any constraint), remains under it for all the following time steps until n, and the other walkers stay below x as well. We integrate over the maximum position x and over the positions $x-t$ of the other walkers at time m (which must be however below x). The factor P arises from the fact that every walker can assume the maximum position with equal probability. We must use this "forward approach" keeping trace of the evolution of the walkers from the first step as there is the constraint that they all start in zero. It is useful to compare this expression with the analogous one for the number of forward records, worked out in [15]:

$$r^+_{n,P}(m) = P \int_0^\infty dx\, G_+(x,0,m)[q_m(x)]^{P-1}; \tag{4.4}$$

where $G_+(x,0,m)$ is the probability of reaching x in m step for the first time (starting from zero), and $q_m(x)$ as before. We now report the result on the asymptotic behaviours of the relevant quantities (persistence probability and propagator) without deriving them, and focus on the computation of the backward records rate.

Finite variance jump distribution We consider first the case where the jump distribution $f(x)$ has a finite variance σ^2, which means its Fourier transform for small k can be written as $\tilde{f}(k) \approx 1 - \sigma^2 k^2/2$. In the continuous time approximation, possible in the large time limit, the propagator follows a diffusion equation with initial condition $G(x,x_0,0) = \delta(x-x_0)$ and diffusion coefficient $2D = \sigma^2$, so its behaviour is easily worked out:

$$G(x,x_0,m) \xrightarrow{m\to\infty} \frac{1}{\sqrt{2\pi m\sigma^2}}\, e^{-\frac{(x-x_0)^2}{2m\sigma^2}}\,. \tag{4.5}$$

Renaming the scaling argument $z := x/\sqrt{2m\sigma^2}$, in [15] is shown that the persistence probability and the constrained propagator $G_+(x,0,m)$ behave for large x and large m but fixed z like:

$$q_m(x) \quad \xrightarrow[z\ fixed]{x,m\to\infty} \quad erf\left(\frac{x}{\sqrt{2m\sigma^2}}\right); \tag{4.6}$$

$$G_+(x,0,m) \quad \xrightarrow[z\ fixed]{x,m\to\infty} \quad \frac{z}{\sqrt{2\sigma^2\, m}}\, \frac{d}{dz}\, erf(z)\Big|_{z=\frac{x}{\sqrt{2m\sigma^2}}}\,. \tag{4.7}$$

For large m (and so x) it is immediate to replace these asymptotic behaviours in (4.4) and, through the scaling change of variable, we obtain in the large n limit:

$$\langle R^+_{n,P}\rangle = \int_0^n dm\, \frac{1}{\sqrt{m}} \int_0^\infty dz\, z\, \frac{d}{dz}\, [erf(z)]^P\,. \tag{4.8}$$

The first integral gives $2\sqrt{n}$ and the second can be interpreted as the average value of the maximum of P independent and identically distributed variables with parental distribution $p(z) = \frac{1}{\pi}\, e^{-z^2}$, which is $\sqrt{\ln P}$ for large P (see 1.1). Thus:

$$\frac{\langle R^+_{n,P}\rangle}{\sqrt{n}} \xrightarrow[n\to\infty]{P\to\infty} 2\sqrt{\ln P}\,; \tag{4.9}$$

the result shows the typical square root dependence on the time and a slower growth with the number of the walkers. Note also that it is independent from the value of the

variance σ^2. The corrections to the first order are strong (a term in $-\frac{\ln(\ln P)}{2\sqrt{\ln P}}$ followed by $O[(\ln P)^{-1/2}]$, see [15]) with the third order still clearly detectable for simulations of 1000 samples (as the ones we performed).

In the backward case, we expect an analogous square root time dependence; moreover we assume that the behaviour is dominated by large m but such that $n-m$ is still large. On that basis, we can replace (4.5) and (4.6) in (4.3) and rescale the variables as follow:

$$x' = \frac{x}{\sqrt{2m\sigma^2}}\,;\; t' = \frac{t}{\sqrt{2m\sigma^2}}\,;\; m = y\,n \;\; (0 \le y \le 1). \tag{4.10}$$

With these changes the average number of backward records becomes:

$$\frac{\langle R^-_{n,P}\rangle}{\sqrt{n}} = P\int_0^1 dy\,\frac{1}{\sqrt{\pi(1-y)}}\int_{-\infty}^{\infty} dx'\,\frac{e^{-x'^2}}{\sqrt{\pi}}\,G_P(x)\,; \tag{4.11}$$

$$G_P(x) := [F(x)]^{P-1} := \left[\int_0^{\infty} dt'\,\frac{e^{-(x'-t')^2}}{\sqrt{\pi}}\,erf\left(t'\sqrt{\frac{y}{1-y}}\right)\right]^{P-1}. \tag{4.12}$$

We want now to study this multiple integral when P is large. In this limit $G_P(x)$ approaches a step function: the change from zero to one happens around a value x^*, in which $G_P(x)$ has a flex point ($G''_P(x^*)=0$ and $G'_P(x)$ has a maximum). If one computes the derivatives of $G_P(x)$, it holds:

$$G''_P(x^*) = F''(x^*)\,[F(x^*)] + (P-2)\,[F'(x^*)]^2 \sim F''(x^*)\,[F(x^*)] + P\,[F'(x^*)]^2 = 0 \tag{4.13}$$

where the last equality uses the fact that P is big. As in x^* it holds $F(x^*) \sim 1 - \frac{\alpha}{P} \sim 1$ for a certain α, the equation can be approximated as:

$$\frac{[F'(x^*)]^2}{F''(x^*)} = -\frac{1}{P}\,. \tag{4.14}$$

After computing the first derivative in x of $F(x)$, it is possible to solve the integral in t' by parts, which leads to the following expressions:

$$F'(x) = \sqrt{\frac{y}{\pi}}\,e^{-y\,x^2}\,\left[1 + erf\left(x\sqrt{1-y}\right)\right]\,; \tag{4.15}$$

$$F''(x) = \sqrt{\frac{y}{\pi}}\,2e^{-y\,x^2}\,\left[-xy\left(1 + erf\left(x\sqrt{1-y}\right)\right) + \sqrt{\frac{1-y}{\pi}}\,e^{-(1-y)x^2}\right]. \tag{4.16}$$

As we guess that the main dependence comes from $e^{-y\,(x^*)^2} \sim 1/P$, which means $x^* \sim \sqrt{\ln P/y}$, in the second derivative we can neglect the first term. Therefore (4.14) becomes:

$$e^{-y\,x^{*2}} = \frac{2y\,x^*}{P}\,\sqrt{\frac{\pi}{y}}\,\frac{1}{1 + erf\left(x^*\sqrt{1-y}\right)}\,. \tag{4.17}$$

Back to (4.11), as $G_P(x)$ approaches a step function, we can approximate the integral in x' as follows:

$$P\int_{-\infty}^{\infty} dx'\,\frac{e^{-x'^2}}{\sqrt{\pi}}\,G_P(x) \sim P\int_{x^*}^{\infty} dx'\,\frac{e^{-x'^2}}{\sqrt{\pi}} \sim \frac{P}{\sqrt{\pi}\,x^*}\,e^{-x^{*2}}\,; \tag{4.18}$$

where the last equality holds for $x^* \to \infty$ and it is easily proven computing the limit of the ratio of the two expression by Hopital theorem. Using the expression of $e^{-x^{*2}}$ just found, (4.18) becomes:

$$\frac{1}{\sqrt{\pi}} \left(x^* e^{-\ln P}\right)^{\frac{1-y}{y}} \left[\frac{2\sqrt{\pi y}}{1+erf\left(x^*\sqrt{1-y}\right)}\right]^{\frac{1}{y}} . \tag{4.19}$$

This term must be included in the y integral in (4.11). It is immediate to see that the main y dependence comes from $1-y$. Therefore we approximate $y \sim 1$ whenever it appears alone, and $x^* \sim \sqrt{\ln P}$; moreover x^* is subleading with respect to $e^{-\ln P}$, so:

$$\frac{\langle R^-_{n,P}\rangle}{\sqrt{n}} = \int_0^1 dy \, \frac{1}{\sqrt{\pi(1-y)}} \, \frac{e^{-\ln P(1-y)}}{\sqrt{\pi}} \left[\frac{2\sqrt{\pi}}{1+erf\left(\sqrt{\ln P(1-y)}\right)}\right] ; \tag{4.20}$$

a change of variable $u = \ln P(1-y)$ gives the final result:

$$\frac{\langle R^-_{n,P}\rangle}{\sqrt{n}} = \frac{2}{\sqrt{\ln P}} \left(\ln(1+erf(\sqrt{\ln P})\right) \sim \frac{2\,\ln 2}{\sqrt{\ln P}} . \tag{4.21}$$

Also in this case there are corrections to the first order, going like $\ln(\ln P)/\ln P$ but, as they decade with the number of walkers, they are less important than in the forward situation. In conclusion, as the walkers spread in space, on average the number of forward records increases with the number of the walkers and the number of backward ones decreases, but always a "slow" way, with a logarithmic dependences from P.

Jump distribution with divergent variance The second class of jump distributions is the one where is not defined a variance: $\int_{-\infty}^{+\infty} dx\, x^2 f(x) \to \infty$. The Lévy-Flights distribution for example shows power law tails, i.e. for $x \to \infty$ the jump probability behaves like $f(x) \sim |x|^{-\mu-1}$, with $0 < \mu < 2$, therefore $\sigma^2 = \langle x^2\rangle$ goes to infinity. As we can not build a diffusion equation, in this case the propagator (the Green function) is obtained solving a recursion relation similar to (1.18) going in the Fourier space: the transform of the distribution is $\tilde{f}(x) \approx 1 - |ak|^\mu$, where a is a microscopic length. The well known result is:

$$G(x, x_0, m) \xrightarrow{m\to\infty} \frac{1}{a\, m^{1/\mu}} \, \Phi_\mu \left(\frac{(x-x_0)}{a\, m^{1/\mu}}\right) ; \tag{4.22}$$

where $\Phi_\mu(z)$ is the Lévy stable function of index μ, defined through its Fourier transform:

$$\Phi_\mu(z) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \, e^{-|k|^\mu - ikz} . \tag{4.23}$$

For large z, it behaves like $|z|^{-\mu-1}$. Clearly, for these distributions the correct scaling argument is $z := x/(a\, m^{1/\mu})$. The behaviour of the persistence probability and the constrained propagator for large x and m but fixed z is [15]:

$$q_m(x) \xrightarrow[z\ fixed]{x,m\to\infty} 1 - \frac{B_\mu}{(x\, a^{-1}\, m^{-1/\mu})^\mu} ; \tag{4.24}$$

$$G_+(x, 0, m) \xrightarrow[z\ fixed]{x,m\to\infty} \frac{2\mu B_\mu}{\sqrt{\pi}\, z^{1+\mu}} . \tag{4.25}$$

The coefficients B_μ are different for distributions with definite mean value or unbounded one:

$$\begin{aligned} B_\mu &= \frac{a^\mu}{\pi\,\Gamma(1-\mu)} \int_0^\infty du\, \frac{u^\mu}{1+u^2}\,, \qquad 0<\mu<1\,; \\ B_\mu &= \frac{2a^\mu}{\pi\,\Gamma(2-\mu)} \int_0^\infty du\, \frac{u^\mu}{(1+u^2)^2}\,, \quad 1<\mu<2\,. \end{aligned} \tag{4.26}$$

Replacing these behaviours in (4.4) and rescaling the variables, we obtain for the number of forward records in the large n limit an expression similar to the one for finite σ^2:

$$\langle R^+_{n,P}\rangle = \int_0^n dm\, \frac{1}{\sqrt{m}} \int_0^\infty dz\, \frac{A_\mu}{\mu\, B_\mu}\, \frac{d}{dz}\, \left[1-\frac{B_\mu}{z^\mu}\right]^P\,. \tag{4.27}$$

The integral in z is immediate to solve, as $\frac{A_\mu}{\mu\, B_\mu} = \frac{2}{\sqrt{\pi}}$, so that:

$$\frac{\langle R^+_{n,P}\rangle}{\sqrt{n}} \xrightarrow[n\to\infty]{P\to\infty} \frac{4}{\sqrt{\pi}}\,. \tag{4.28}$$

independent from the number of walkers and from the index μ. The simulations shows that for different Lévy coefficients this value is indeed reached but with different relaxation times: in a few steps for $\mu\sim 1$, many more for $\mu\to 2$.

For the backward case, it is harder than for finite σ^2 distributions to work out the exact asymptotic dependence as we have an explicit form for the Lévy functions $\Phi_\mu(z)$ only for some μ, while in the majority of the cases we just know the large z behaviour. Moreover, even when the propagator is known (for $\mu=1$, for example, $\Phi_1(z)$ becomes a Cauchy distribution up to some constant), the previous approach does not hold, as the persistence probability is not integrable in zero, and we need a more accurate analysis to work out the correct coefficients. As the task is beyond the present treatment, we prefer to perform just a rough scaling analysis to look for the P dependence of the number of backward records and then we will try to understand from the simulations whether the unknown coefficient depend on μ or not.

Rescaling the variables in the proper way and introducing y as above, the number of backward records becomes:

$$x' = \frac{x}{a\, m^{1/\mu}}\,;\; t' = \frac{t}{a\, m^{1/\mu}}\,;\; m = y\, n \quad (0\le y\le 1)\,; \tag{4.29}$$

$$\begin{aligned} \frac{\langle R^-_{n,P}\rangle}{\sqrt{n}} = P \int_0^1 dy\, \frac{1}{\sqrt{\pi(1-y)}} \int_{-\infty}^\infty dx'\, \Phi_\mu(x') \\ \times \left[\int_0^\infty dt'\, \Phi_\mu(x'-t')\, \left(1-\frac{y}{1-y}\,\frac{B_\mu}{t'^\mu}\right)\right]^{P-1}\,. \end{aligned} \tag{4.30}$$

We suppose then that even with power low tails jump distributions the function $G_P(x')$ becomes a step function in the large P limit and that $F(x')$ (the function in the square brackets) can be approximated as $1-\alpha/P$ around the point $x'=x^*$ where $G_P(x')$ switches from 0 to 1, with $\alpha=\alpha(y,\mu)$ an unknown function. As the integrand in t' reaches its maximum around $t'\sim x'$, we approximate $F(x')\sim 1-\frac{y}{1-y}\,\frac{B_\mu}{x'^\mu}$ considering $\Phi_\mu(0)$ of order one. Thus $F(x^*)$ must satisfy:

$$F(x^*) \sim 1-\frac{y}{1-y}\,\frac{B_\mu}{x^{*\mu}} = 1-\frac{\alpha(y,\mu)}{P}\,; \tag{4.31}$$

which gives $x^* \sim \beta(y,\mu)\, P^{\frac{1}{\mu}}$ where $\beta(y,\mu)$ absorbs all the dependences from the coefficient μ and the variable y. The important information is the P dependence: indeed, if we perform the integral on x' treating $G_P(x')$ as a step function it becomes:

$$\int_{-\infty}^{\infty} dx'\, \Phi_\mu(x')\, G_P(x') \sim \int_{x^*}^{\infty} dx'\, \Phi_\mu(x') \sim -\gamma(y,\mu)\, x^{-\mu}\Big|_{x^*}^{\infty} \sim \gamma(y,\mu)\,; \tag{4.32}$$

with $\gamma(y,\mu)$ an unknown function. The P dependence has disappeared. Simulations show that the coefficient is indeed universal for many μ (Figure (4.2)). In particular a rough estimation gives the number 0.71, smaller than $4/\pi \sim 1.3$ found in the forward case. In conclusion, for power law tail jump distributions, the number of forward and backward records is respectively raised and diminished with respect to the single walk case, as expected. However asymptotically, their are both independent from the number of walkers, thus the asymmetry concerns only the numerical prefactors. This is attributed to the fact that in this situations records are not governed by small jumps, linked to the gradual spread of the walkers, but by the occasional big jumps (as a consequence of having "fat" tails in the jump distribution).

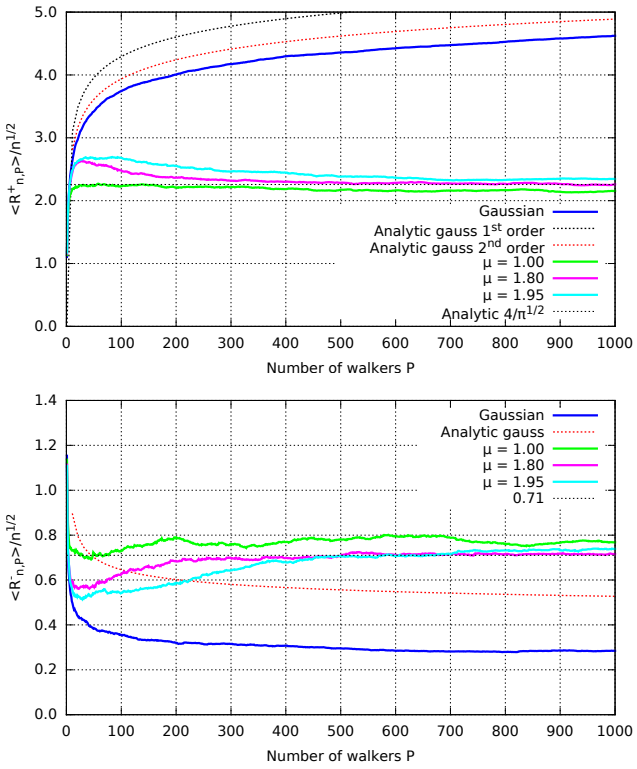

Figure 4.2: Above, the forward records, rescaled by the square root of the length of the sequence, in function of the number of walkers; below, the backward ones. For Gaussian random walk is shown the theoretical behaviour to the first and to the second order for forward records, only the first for backward ones. For Lévy walks we show the behaviour for different indexes μ.

Simulations We performed simulations for multiple random walks with a number of walkers from 1 to 1000 and computed the number of forward and backward records for

a fixed time length of 1000 steps in the finite σ^2 case (where we used Gaussian jump distribution), and 500 steps length for power law jump distribution. In this last case, we performed simulations for Lévy indexes $\mu = 1, 1.5, 1.8$ and 1.95 to check their influence. As anticipated in the respective sections, the results agree with the theory, but for finite σ^2 the corrections are important, and for the Lévy case the coefficient μ affects strongly the transient regime (as well as the particular form of the jump distribution, indeed we found different transient times with respect to the ones found for the forward case in [15]). In Figure (4.2) we show the plot of the number of forward and of backward records, rescaled with the square root of the sequence length, for Gaussian jump distribution and Lévy distributions for different indexes, together with the analytic results where available.

Chapter 5

Conclusions

In this work we wanted to sketch the state of art of some extreme value problems and provide some new results. In particular we analysed the maximum and the records of a given sequence, exploring the meaning and statistics of backward records in situations where the forward behaviour was already well understood. We treated sequences of identically and independent distributed random variables and, as a possible example of strongly correlated variables, random walks variables.

The first important aspect to retain is the universality which emerges in the different situations: we saw that for the independent variables the problem of the maximum of the sequence leads to distinguish three universality classes in which we can group a large number of distributions according to the behaviour of their tails, the only relevant information in this context. When we shift to the records problem, the universality for these variables becomes even wider, as for every distribution with continuous cumulative function the same results hold.

For random walks instead, when looking at the maximum, we must distinguish distributions with finite second moment or not. The existence of a second moment allows to write a backward Fokker-Planck equation, which can be easily adapted to the presence of an arbitrary external potential. We did not treat the maximum for a Lévy-Flights distribution of the stochastic increments but it is clear that, as σ^2 does not exist, another approach must be followed and different results are obtained as cited at the end of chapter 1. In particular, the important parameter is not the broadness of the distribution (difficult to define) but how its tails go to zero. Also for random walks we recover a wider universality in the records problem where every symmetric distribution with continuous cumulative function leads to the same statistics (the symmetry is an additional requirement with respect to the independent case). This universality has complete different causes from the ones of the i.i.d. case and derives from the Sparre-Andersen theorem. Discrete distributions are not included in the class, but we can compute explicitly the results if the jump function is simple enough. Note that some records problems present a new break of the universality when the level of the record becomes important, as the level itself was important in the maximum problem: we saw an example in the multiple walks situation, where it arises a difference according to the boundedness of the second moment of the distribution.

Without citing again the specific results for each section, we remember the new results of this work: we found an approximated universal scaling form for the joint probability of forward and backward records and computed exactly their correlation both in the i.i.d.

case and for random walks. We explored extreme values ages concerning the two types of records jointly and used the forward-backward formalism to find the longest age in specifical situations (i.e. in the set of ages without considering the one of the maximum), task already performed with different tools in the random walk case, but not known in literature for the i.i.d. one. Similarly we worked out the probability for a certain age to be the longest, giving the results for a wider number of cases with respect to the ones already known. Finally, we worked out the behaviour of backward records for a multiple random walk.

A part from the universality, there are other key points to highlight in relation to what we found: first of all the statistical equivalence between backward upper records and forward lower ones in relation to forward upper ones for independent variables, which has not been explained yet. This relation does not hold for random walks. Then, for random walks, the connection between maximum value and number of records, which is pointed out from the ± 1 jump distribution case. This last one gives the same record statistics as for the continuous distributions a part a scaling prefactor and, at the same time, the problem can be seen asymptotically as a maximum problem for a random walk with diffusion coefficient $1/2$, linking the two objects. Finally, it is interesting to see how ages of records provide information about the independence (or dependence) of successive records through some extreme values objects as the longest age.

The only example we treated, the daily temperatures in several European cities in the last century, confirms that temperatures recorded at a fixed day of the year are distributed around an average value which increases with time, but the annual mean value temperature is sufficient to describe this warming trend, so that the detrended data are independently distributed around the average value. The independence can be proven in a straighter way simply computing their correlation, which is zero on average. Therefore, the simple model of independent and identically distributed random variables is not only a "toy model" but can be employed in climate analysis, where combinations of records provide estimates for different relevant (and more complicated) quantities.

Many problems are still open: besides those here treated through an asymptotic or incomplete analysis which can maybe be afforded analytically using different methods, it would be interesting to extend the study of joint forward backward statistic to biased random walks or independent variables taken from changing distribution (so not any more identically distributed). Again, it would be useful to see the effects of rounding on the correlation between the two types of records, keeping in mind that a rounding process breaks the universality: can we recognize some general features in the forward-backward correlation? Are there any oscillatory transient behaviours (as some simulations suggest)?

In this sense, the statistics of extreme values is always rich and non trivial, suggesting that the pursuit of the studies in this field can provide new understanding of stochastic processes, new links between different approaches and physical problems, and valid tools to analyse real every-day life problems.

Appendix A

Appendices

A.1 Ornstein-Uhlenbeck process

Here we show the main passages to compute the probability of the maximum for the Ornstein-Uhlenbeck process [1]. As anticipated, this solution proceeds through a forward Fokker-Planck approach.

Consider the potential $U(x) = ax^2$ with a positive. Renaming $c = 2a\mu$, the equation for the increments and its formal solution in the continuous time limit are:

$$\frac{\Delta x_i}{\Delta t} = -c\, x_i + \eta_i \quad \xrightarrow{\Delta t \to 0} \quad \frac{dx}{dt} = -c\, x(t) + \eta(t)\,; \tag{A.1}$$

$$x(t) = x_0 e^{-ct} + \int_0^t ds\, e^{-c(t-s)} \eta(s)\,. \tag{A.2}$$

The correlation between the positions at different times is easily computed remembering that $\eta(t)$ has zero mean and is delta correlated:

$$C(t_1, t_2) = \langle x(t_1) x(t_2) \rangle = \frac{D}{c} \left[e^{-c|t_1 - t_2|} - e^{-c(t_1 + t_2)} \right]. \tag{A.3}$$

Therefore in the case of positive c the correlation decays exponentially with time difference when t_1, t_2 are both bigger than $1/c$. It means that looking at times bigger than this "coarse graining" scale, the system is weakly correlated, and, as explained in the end of section 1.1 the problem becomes an i.i.d. variables maximum problem. In particular we expect a Gumbel distribution, as the decay of the correlation is exponential.

Consider the conditioned probability $P(x|t)|_z$ that the particle, starting at $x_0 = 0$ for simplicity, arrives in x at time t while staying below the level z. It follows the forward Fokker-Planck equation as the corresponding unconditioned probability, but in the domain $x < z$:

$$\frac{\partial P}{\partial t} = D \frac{\partial^2 P}{\partial x^2} + c \frac{\partial P}{\partial x}\,; \tag{A.4}$$

with the conditions $P(x|0)|_z = \delta(x)$ and, for every time t, $P(x \to -\infty|t)|_z = 0$; $P(z|t)|_z = 0$. To solve it, we can expand the probability in eigenfunctions with eigenvalues λ to be set with the boundary conditions. Setting for simplicity $D = 1/2$:

$$P(x|t)|_z = \sum_\lambda a_\lambda e^{-\lambda t - cx^2/2} D_{\lambda/c}(-\sqrt{2c}\, x)\,; \tag{A.5}$$

where $D_p(y)$ is the parabolic cylinder function which satisfy the second order ordinary differential equation: $D_p''(y) = (p + 1/2 - z^2/4)D_p(y)$. Keeping a generic diffusion coefficient gives an additional term in $D_p'(y)$ in the equation, for which the solutions are much more complicated but with no more information on the physical aspects of the problem. To choose the correct eigenvalues in the solution (A.5) we must apply the boundary condition in z, which implies $D_{\lambda/c}(z) = 0$. The solution is found in the large z limit, which corresponds to large time t. In this situation the smallest eigenvalue λ_0 dominates the solution, and it is found through an asymptotic analysis of the cylindric functions:

$$\lambda_0(z) \xrightarrow{z\to\infty} \frac{2}{\sqrt{\pi}}\, c^{3/2} z\, e^{-c\,z^2}\,. \tag{A.6}$$

Thus, as the cumulative probability for the maximum M decays as the probability itself, for large t and z we have (with the notation used in the previous sections):

$$Q(M,0|t) \sim e^{-\lambda_0(M)t} \sim e^{-c\,z^2+\ln\left(\frac{2\,c^{3/2}}{\sqrt{\pi}}\,z\,t\right)} \tag{A.7}$$

$$\to F_2\left(\sqrt{4c\,\ln t}\,\left(M - \sqrt{\frac{\ln t}{c}}\right)\right)\,; \tag{A.8}$$

where the shift (the coefficient a_N of section 1.1) is immediate to obtain, telling that the maximum goes like the square root of the logarithm of the time, so its growth is much suppressed; the width of the fluctuations instead must be computed more carefully using formula (1.7) and it decreases as $\sim 1/\sqrt{\ln t}$.

Bibliography

[1] S.N. Majumdar and A. Pal, *Extreme value statistics of correlated random variables*, arXiv:1406.6768v3 (14 pp.) (2015).

[2] B.C. Arnold, N. Balakrishnan and H.N. Nagraja, *Records*, 1st ed. Wiley-Interscience (1998).

[3] B.V. Gnedenko, Annals of Mathematics **44**, 2613 (1981).

[4] S.N. Majumdar, *Universal First-passage Properties of Discrete-time Random Walks and Lévy Flights on a Line: Statistics of the Global Maximum and Records*, arXiv:0912.2586v3 (2010).

[5] G. Schehr and S.N. Majumdar, *Exact record statistics of random walks via first-passage ideas*, in "First-Passage Phenomena and Their Applications", Eds. R. Metzler, G. Oshanin, S. Redner; World Scientific (2013).

[6] S.N. Majumdar and R.M. Ziff, *Universal record statistics of random walks and Lévy Flights*, Phys. Rev. Lett. **101**, 050601 (2008).

[7] A. Comtet and S.N. Majumdar, *Precise asymptotic for a Random Walker's Maximum*, arXiv:cond-mat/0506195v2 (2005).

[8] D.E. Barton and C.L. Mallows, *The randomization bases of the problem of the amalgamation of weighted means*, Journal of the Royal Statistical Society B, vol. 23 No. 2 (1961).

[9] D.E. Barton and C.L. Mallows, *Some aspects of the random sequence*, Ann. Math. Statist. **36**, No. 1 (1965).

[10] F.G. Foster and A. Stuart, *Distribution-free tests in time-series based on the breaking of records*, Journal of the Royal Statistical Society B, vol. 16 No. 2 (1954).

[11] W. Katzenbeisser, *On the joint distribution of the number of upper and lower records and the number of inversion in a random sequence*, Adv. Appl. Prob. **22**, 957-960 (1990).

[12] C. Godrèche, S.N. Majumdar and G. Schehr, *Universal statistics of longest lasting records of random walk and Lévy-Flights*, J. Phys. A: Math. Theor. **47**, 255001 (2014).

[13] C. Godrèche, S.N. Majumdar and G. Schehr, *Statistics of the longest interval in renewal processes*, J. Stat. Mech. P03014 (2015).

[14] C. Godrèche, S.N. Majumdar and G. Schehr, *The longest excursion of stochastic processes in nonequilibrium systems*, Phys. Rew. Lett. **102**, 240602 (2009).

[15] G. Wergen, S.N. Majumdar, G. Schehr, *Record Statistics for Multiple Random Walks*, Phys. Rev. E **86**, 011119 (2012).

[16] G. Wergen, D. Volovik, S. Redner and J. Krug, *Rounding Effects in Record Statistics*, Phys. Rev. Lett. **109**, 164102 (2012).

[17] Y. Edery, A. Kostinski, S.N. Majumdar and B. Berkowitz, *Record-breaking statistics for random walks in the presence of measurement error and noise*, Phys. Rev. Lett. **110**, 180602 (2013).

[18] J. Krug, *Records in a changing world*, J. Stat. Mech. P07001 (2007).

[19] J. Franke, G. Wergen, J. Krug, *Records and sequences of records from random variables with a linear trend*, J. Stat. Mech. P10013 (2010).

[20] G. Wergen, M. Bogner, J. Krug, *Record statistics for biased random walks, with an application to financial data*, Phys. Rev. E **83**, 051109 (2011).

[21] S. N. Majumdar, G. Schehr, G. Wergen, *Record statistics and persistence for a random walk with a drift*, J. Phys. A: Math. Theor. **45**, 355002 (2012).

[22] G. Wergen and J. Krug, *Record-breaking temperatures reveal a warming climate*, EPL **92**, 30008 (2010).

[23] S. Redner and R. Peterson, *Role of global warming on the statistics of record-breaking temperatures*, Phys. Rev. E **74**, 061114 (2006).

[24] A. Anderson and A. Kostinski, *Evolution and Distribution of Record-Breaking High and Low Monthly Mean Temperature*, J. Appl. Meteor. Climatol. **50**, 1859-1871 (2011).

[25] A. Anderson and A. Kostinski, *Reversible Record Breaking and Variability: Temperature Distributions across the Globe*, J. Appl. Meteor. Climatol. **49**, 1681-1691 (2010).

[26] J.F. Eichner, E. Koscielny-Bunde, A. Bunde, S. Havlin and H.-J. Schellnhuber, *Power-law persistence and trends in the atmosphere: a detailed study of long temperature records* Phys. Rev. E **68**, 046133 (2003).

[27] Project Team ECAD, *European climate assessment and dataset* (Tech. Rep., Royal Netherlands Meteorological Institute KNMI) (2008).

Data from http://www.ecad.eu/dailydata/predefinedseries.php/

[28] J. Krug and K. Jain, *Breaking records in the evolutionary race*, Physica A **358**, 1 (2005).

[29] B. Alessandro, C. Beatrice, G. Bertotti and A. Montorsi, *Domain wall dynamics and Barkhausen effect in metallic ferromagnetic materials. I. Theory*, J. Appl. Phys. **68**, 2901-2907 (1990).

[30] J.R. Furling, S.N. Majumdar and A. Comtet, *Convex Hull of N Planar Brownian Motions: Exact Results and an Application to Ecology*, Phys. Rev. Lett. **103**, 140602 (2009).

[31] G. Wergen, *Records in stochastic processes - Theory and applications*, J. Phys. A: Math. Theo. **46**, 223001 (2013).