Records statistics beyond the standard model - Theory and applications

Inauguraldissertation
zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Gregor Wergen

aus Köln

Köln 2013

Berichterstatter:	Prof. Dr. Joachim Krug
	Prof. Dr. Alexander Altland
	Prof. Dr. Thomas Guhr (Universität Duisburg-Essen)

Tag der mündlichen Prüfung: 15.01.2013

Diese Dissertation wurde gemäß der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät der Universität zu Köln in kumulativer Form angefertigt. Eine Liste der Veröffentlichungen, die in diese Dissertation eingebunden sind, ist auf Seite ix abgedruckt. Weiterhin findet sich auf Seite 281 eine Zusammenfassung meiner Beteiligungen zu diesen Veröffentlichungen.

Zusammenfassung

Die Statistik von Rekordereignissen ist in den letzten Jahren auf stetig wachsendes Interesse sowohl von Seiten der Wissenschaft als auch der gesamten Öffentlichkeit gestoßen. Nicht nur im Sport und in der Klimatologie wird das Setzen und Brechen neuer Rekorde beobachtet, auch in der Natur und in der Wirtschaft sorgen sie immer wieder für Aufmerksamkeit. Diese kumulative Dissertation ist der Studie von Rekordereignissen gewidmet und fasst eine Reihe von Arbeiten zur Theorie und Anwendung der Rekordstatistik zusammen. Sie besteht, im Wesentlichen, aus fünf Teilen. Im ersten Teil wird die Statistik von Rekorden in unkorrelierten Zufallszahlen aus zeitabhängigen Verteilungen untersucht. Insbesondere wird ein einfaches Modell von Zufallszahlen mit einem linearen Drift im Mittelwert vorgestellt und ausführlich in Hinblick auf Rekorde untersucht. Weiterhin gehen wir auf die Rolle von Diskretheitseffekten in der Rekordstatistik von Messdaten ein. Anschliessend, in Teil zwei, werden diese Arbeiten angewendet um das Auftreten von Rekordtemperaturen und deren Zusammenhang mit dem Klimawandel zu erklären. Aufbauend auf unser einfaches Modell von Zufallszahlen mit linearem Drift wird demonstriert, dass die globale Erwärmung einen signifikanten Einfluß auf die Zahl und die Ausprägung von Hitze- und Kälterekorden hat. Der dritte Teil befasst sich mit Rekorden in korrelierten Prozessen, sogenannten Random Walks. Es werden verschiedene Random-Walk-artige Prozesse untersucht, die Rekordstatistik dieser Prozesse ist erstaunlich vielfältig und interessant. Die vorgestellten Resultate sind wiederum wichtig um die Rekordstatistik von Finanzdaten verstehen zu können, die im vierten Teil diskutiert wird. Random Walks sind ein hilfreiches Mittel um Rekorde in Börsendaten zu beschreiben, wir finden jedoch auch signifikante Abweichungen von diesem Modell und betrachten kompliziertere Prozesse um die Modellierung zu verbessern. Im abschließenden, fünften Teil werden die jüngeren Entwicklungen auf dem Gebiet der Rekordstatistik von zeitabhängigen und korrelierten Zufallszahlen in einem Review zusammengefasst.

Abstract

In recent years, there was a surge of interest in the statistics of record-breaking events, not only from scientists, but also from the general public. In sports and in climatology, but also in nature and in economy, observers are interested in the setting and breaking of new records. This cumulative dissertation is dedicated to the study of record-breaking events. It concludes a series of published and hitherto unpublished articles on theory and applications of record statistics. This work mainly consists of five parts. The first part is about the statistics of records in uncorrelated random variables sampled from time-dependent distributions. In particular, we present a simple model of random variables with a linearly growing mean value and discuss its record statistics thoroughly. Furthermore, the effects of rounding on the occurrence of records in series of independent and identically distributed random variables are considered. Then, in part two, these results are applied to explain and model record temperatures in the context of climatic change. Using our minimal model of random variables with a linear drift, we show that global warming has in fact a significant effect on the occurrence and the values of heat and cold records. The third part focuses on records in correlated processes, in particular random walks. A number of different random walk processes are presented and analyzed. We find that their record statistics are surprisingly interesting and manifold. The results derived in this part are important to understand the occurrence of records in financial data, which will be discussed in the fourth part. There it is demonstrated that random walks are helpful to model records in stock data, nevertheless we find significant deviations from the analytical results and propose an alternative model, which describes the stock data more accurately. The final, fifth part is a general review of the recent developments in the study of record-breaking events in time series of time-dependent and correlated random variables.

Published contributions

The following contributions in this thesis are hitherto published in peer-reviewed journals (in chronological order of their publication):

1. Jasper Franke, Gregor Wergen and Joachim Krug,
   Records and sequences of records from random variables with a linear trend,
   J. Stat. Mech.: Theor. Exp. **P10013** (13 October 2010)
   Chapter 3, pp. 27
2. Gregor Wergen and Joachim Krug,
   Record-breaking temperatures reveal a warming climate,
   EPL **92**, 30008 (29 November 2010)
   Chapter 7, pp. 81
3. Gregor Wergen, Miro Bogner and Joachim Krug,
   Record statistics for biased random walks, with an application to financial data,
   Phys. Rev. E **83**, 051109 (9 May 2011)
   Chapter 9, pp. 117
4. Gregor Wergen, Jasper Franke and Joachim Krug,
   Correlations between record events in sequences of random variables with a linear trend,
   J. Stat. Phys. **144**, 1206 (5 August 2011)
   Chapter 4, pp. 47
5. Jasper Franke, Gregor Wergen and Joachim Krug,
   Correlations of record events as a test for heavy-tailed distributions,
   Phys. Rev. Lett. **108**, 064101 (7 February 2012)
   Chapter 5, pp. 63
6. Gregor Wergen, Satya N. Majumdar and Grégory Schehr,
   Record statistics for multiple random walks,
   Phys. Rev. E **86**, 011119 (18 July 2012)
   Chapter 11, pp. 173
7. Satya N. Majumdar, Grégory Schehr and Gregor Wergen,
   Record statistics and persistence for a random walk with a drift,
   J. Phys. A: Math. Theor. **45**, 355002 (15 August 2012)
   Chapter 10, pp. 129
8. Gregor Wergen, Daniel Volovik, Sidney Redner and Joachim Krug,
   Rounding effects in record statistics,
   Phys. Rev. Lett. **109**, 164102 (19 October 2012)
   Chapter 6, pp. 71

Additional unpublished contributions

Except for the introductory part (chapters 1 and 2), as well as the summary (chapter 14) of this dissertation the following contributions are unpublished at the time of publication of this thesis:

1. Gregor Wergen, Andreas Hense and Joachim Krug,
   Record occurrence and record values in daily and
   monthly temperatures,
   (submitted to Climate Dynamics, arXiv:1210.5416)
   Chapter 8, pp. 93
2. Gregor Wergen,
   Modeling record-breaking stock prices
   (in preparation)
   Chapter 12, pp. 205
3. Gregor Wergen,
   Record statistics beyond the standard model
   - Theory and applications
   (submitted to J. Phys. A.: Math. Theor., arxiv:1211:6005)
   Chapter 13, pp. 231

Contents

Kurzzusammenfassung (deutsch) iii

Abstract v

List of contributions ix

1 Introduction 1
- 1.1 Record statistics of i.i.d. RV's 5
- 1.2 Record statistics of symmetric random walks 12
- 1.3 Recent theoretical progress & applications 14

2 Outline 17
- 2.1 Part I - Records in uncorrelated random variables 17
- 2.2 Part II - Record-breaking temperatures 19
- 2.3 Part III - Record statistics of random walks 20
- 2.4 Part IV - Records in finance 22
- 2.5 Part V - Record statistics - A review 23

I Records in uncorrelated random variables 25

3 Records and sequences of records from random variables with a linear trend 27
- 3.1 Introduction. 28
- 3.2 General theory and an exactly solvable example 30
- 3.3 Record statistics for small drift 32
- 3.4 Record statistics for large drift 39
- 3.5 Conclusions 41

4 Correlations between record events in sequences of random variables with a linear trend 47
- 4.1 Introduction. 48
- 4.2 General theory 49
- 4.3 Explicit examples 54
- 4.4 Conclusion 60

5 Correlations of record events as a test for heavy-tailed distributions 63
- 5.1 Introduction. 64
- 5.2 Record statistics and record correlations 64
- 5.3 Description of the test 66
- 5.4 Fluctuations. 67
- 5.5 Application 67
- 5.6 Summary 68

6 Rounding effects in record statistics 71
6.1 Introduction. 72
6.2 General theory, asymptotic results. 73
6.3 Small-Δ regime 74
6.4 Large-Δ regime 75
6.5 Conclusions 76

II Record-breaking temperatures 79

7 Record-breaking temperatures reveal a warming climate 81
7.1 Introduction. 82
7.2 Theory 83
7.3 European data 83
7.4 American data and discreteness effects 88
7.5 Re-analysis data 89
7.6 Conclusions 89

8 Record occurrence and record values in daily & monthly temperatures 93
8.1 Introduction. 94
8.2 Theory: Record occurrence in the presence of linear drift 96
8.3 General introduction of the data 98
8.4 Occurrence of records in daily and monthly temperatures 101
8.5 Theory: Distributions of record values 103
8.6 Distributions of record values in European temperature recordings . . . 105
8.7 Discussion and conclusion 111

III Record statistics of random walks 115

9 Record statistics for biased random walks, with an application to financial data 117
9.1 Introduction. 118
9.2 Survival probabilities & first passage times 119
9.3 Gaussian random walks with drift 121
9.4 Record statistics of stock prices in the S&P 500 124
9.5 Summary. 126

10 Record statistics and persistence for a random walk with a drift 129
10.1 Introduction. 130
10.2 Record statistics for random walks: Review & new results 132
10.3 Record number distribution via renewal property 139
10.4 Asymptotic behavior of persistence $Q(n)$ for large n 141
10.5 Asymptotic record number distribution $P(R,n)$ for large n. 147
10.6 Extreme statistics of the age of a record 157
10.7 Conclusion 167

11 Record statistics for multiple random walks 173
11.1 Introduction. 174
11.2 The model & the main results 176
11.3 Mean number of records for multiple walkers 178
11.4 The distribution of the number of records for finite σ^2 184
11.5 Numerical simulations 187
11.6 Comparison to stock prices 191
11.7 Conclusion 193

IV Records in finance 203

12 Modeling record-breaking stock prices 205
12.1 Introduction. 206
12.2 Records in i.i.d. RV's & symmetric random walks 207
12.3 Records in biased Gaussian random walks 209
12.4 Records in autoregressive processes 211
12.5 Analysis of single stocks 212
12.6 Record statistics of N stocks 223
12.7 Weekly distributions of records 225
12.8 Summary and conclusions. 226

V Records statistics - A review 229

13 Record statistics beyond the standard model - Theory and applications 231
13.1 Why records? 232
13.2 Classical theory of records 233
13.3 Records in uncorrelated and time-dependent RV's 235
13.4 Discreteness, rounding and ties 241
13.5 Records in correlated processes 246
13.6 Applications. 255
13.7 Summary and outlook 260

14 Summary & Conclusions 265
14.1 Part I - Records in uncorrelated random variables 265
14.2 Part II - Record-breaking temperatures 267
14.3 Part III - Record statistics of random walks 268
14.4 Part IV - Records in finance 269
14.5 Closing remarks & bigger picture 270

Bibliography for chapters 1, 2 & 14 271

Acknowledgements 277

Anhang gemäß der Prüfungsordnung (deutsch) 279

Zusammenfassung meiner Beteiligungen zu den Publikationen 281

Erklärung zur Selbstständigkeit 283

Lebenslauf 285

Chapter 1

Introduction

> I think sometimes I guess you see records,
> say you want to get there and use that as motivation.
> In a way, it's kind of cool if there is a possibility to rewrite history
> and be up there with the greats of Olympic history.
>
> *Michael Phelps, eighteen-fold Olympic Gold medalist*

> The inability to predict outliers
> implies the inability to predict the course of history,
> given the share of these events in the dynamics of events.
>
> *Nicholas Nassim Taleb, The Black Swan:*
> *The Impact of the Highly Improbable*

In our performance-oriented society it is sometimes all about being the best, about achieving more than anybody could ever achieve before. Whether it is a new 100 metre world record, a new record-high skyscraper, or a painting of Edward Munch getting sold for a record-breaking prize, accomplishments, for instance in sports or engineering, but also in business, biology, or in other areas of science are noticed and remembered if they outperform everything that has been accomplished before. On the other hand, events in nature, climatology or also astronomy are considered particularly important and noteworthy if they exceed all previous observations. Scientists and also the general public care a lot about record-breaking storms, heatwaves or earth-quakes. Unfortunately, in this context, records are usually the most dangerous and devastating events.

These are only a few of the reasons why records are of great importance and of general interest. Fig 1.1 shows a few examples of particularly attention-grabbing records that were established by humans, nature or the financial markets. The word 'record' descends from the Latin verb *recordari - to recall, to remind* and everybody has a basic understanding of what a record event is. One has learned that a new record is something relevant that will be remembered. Therefore quite a few people are fascinated by setting and thereupon breaking new and spectacular records, making the famous book 'Guinness World Records'[1] [1] itself a record-holder as the best-selling copyrighted book in history. As a matter of fact, the book is also among the most frequently stolen books from public libraries. From the technical point of view a record is an entry in a series of measurements that exceeds all previous entries. The fact that they fall out of the range that has been covered before makes them interesting.

---

[1] until 2000 known as 'The Guinness Book of Records'

Figure 1.1: Record-breaking events are interesting in many different areas: **Top left:** Death Valley, United States, where, with 56.7° C, the hottest temperature ever recorded was measured in July 1913 (Photo: Jon Sullivan). **Top right:** Felix Baumgartner is an Austrian extreme sportsman who performed a record-breaking jump from a helium balloon in October 2012. He fell with a record-breaking speed of more than 1300km/h from a height of over 39km (Photo: Alexandre Inagaki, Image rights: Creative Commons). **Bottom left:** In 2009, the Jamaican sprinter Usain Bolt set a new world-record over 100m dash in Berlin. He ran this distance in 9.58s and, with that, broke his own world record for the second time (Photo: Erik van Leeuwen, Image rights: GNU Free Documentation License). **Bottom right:** The stock of Apple set various records in recent years, it has an all time high market capitalization of 607.5 billion U.S. Dollars (Image rights: GPL).

On a personal note, for a sports fanatic who always dreamed of setting new world records in his youth[2], it was a great pleasure and also a redemption to have the opportunity to study records from a scientific point of view. My studies on records were initially motivated by an interest in a deeper understanding of the role of global warming on the occurrence of record-breaking temperatures. In recent years, I considered many different aspects and applications of the theory of records. This cumulative dissertation summarizes my work on the subject of record statistics and is, in this course, also supposed to give a survey of recent developments in this field of research. It turned out that records are more than just interesting for the observer. From the mathematical point of view their behavior is particularly rich and, by analyzing records in measured data, there is a lot that one can learn about the underlying dynamical system generating the measurements. Especially in the last decade, researchers made progress in modeling record events in observational data by comparing them to various elementary stochastic processes. In this context, this thesis discusses several stochastic models, which improve our understanding of records in climate and finance, in particular.

---

[2] Apparently the author never succeeded and studied physics instead, which, in turn, ultimately resulted in this thesis.

Throughout this work, we consider time series of random numbers $X_1, X_2, ..., X_n$, which may represent any kind of measurement, such as temperatures, sport results or stock prizes. In such a series, an entry X_n is an upper record if it exceeds all previous entries $X_1, X_2, ..., X_{n-1}$. Analogously, one can also define a lower record, as an entry that is smaller than all previous ones. In the theory of records and especially in this thesis, two quantities are of particular importance: The probability

$$P_n := \mathrm{Prob}\left[X_n > \max\{X_1, X_2, ..., X_{n-1}\}\right] \tag{1.1}$$

that the nth event X_n of a time series is a record and the number of records R_n that occurs until time n. In the following, P_n is often called the record rate.

The classical theory of records in time series of independent and identically distributed (i.i.d.) random variables (RV's) sampled from continuous distributions is well established. In a series of RV's, which were all drawn from the same continuous probability density $f(x)$, the nth entry is a record with probability $P_n = \frac{1}{n}$. The first entry, X_1, is a record with probability $P_1 = 1$, the second one, X_2, with $P_2 = 1/2$ and so on. With this, one can infer that, for large n, the average $\langle R_n \rangle$ of the number of records in such a time series grows logarithmically with n. As we will soon discuss in more detail, this mean record number $\langle R_n \rangle$ is just the sum over the record rate P_n. By now, a lot more is known about this classical model, which is particularly important because of its strong universality. The occurrence of records in time series of i.i.d. RV's is entirely independent from the choice of the underlying distribution.

More recently there has been a surge of interest in the record statistics of processes beyond this standard model. A lot of progress was made towards a better understanding of the statistics of records in time series of time-dependent and correlated random variables. Several research groups studied how the classical results alter if one considers time series of independent RV's sampled from a distribution that changes in time. Additionally, it was possible to compute the record statistics of time series with correlated entries. These efforts were motivated by applications in various different areas of life and science. Records have been studied extensively in climatology, but also in sports, in biology, in physics and in finance.

In this work, various different stochastic processes with both time-dependent and correlated entries are considered. A model of independent RV's with an increasing mean value is introduced and used to predict the occurrence of temperature records in the context of global warming. Furthermore, the record statistics of some simple correlated processes are computed. The findings are useful to describe record-breaking events in financial markets.

This thesis consists of a series of articles that were published, together with various co-workers, in several peer-reviewed journals along with some additional, hitherto unpublished chapters that are either under review or in preparation. The total number of 11 contributions can be arranged in four different groups, which form the four main parts of this thesis and an overall review that shapes a fifth part.

The first part, which contains the chapters 3 to 6, discusses various aspects of the record statistics of uncorrelated random variables. A key role in most of these contributions is played by the so-called Linear Drift Model of record statistics. The main question treated in the chapters 3 and 4 is how the occurrence of records is affected by a linearly increasing mean value. We discuss series of RV's from shifting probability densities $f_k(x) = f_0(x - ck)$ with a common shape $f_0(x)$. It is demonstrated how strongly quantities like the record rate P_n or the correlations between record events are affected by such a constant drift. In chapter 5, we present a possible application of our findings as a record-based test for extremal properties of observational data. While the Linear Drift Model is the dominating topic in this part, there is also a contribution concerning the problem of rounding, which is found in any measurement process. In chapter 6, it is discussed how the statistics of records, the record rate P_n and the mean record number $\langle R_n \rangle$ change if one allows record values to get tied in a discrete setting. We show that rounding up or down of originally continuous RV's can significantly alter the record process.

The second part, with chapters 7 and 8, describes the most important application of the theory developed in part I. There it is demonstrated how the effect of global warming on the statistics of record-breaking temperatures can be quantified by a simple model of uncorrelated RV's sampled from a distribution with a time-dependent mean value. In chapter 7 it is shown that such a minimal model for the record rate P_n is capable of describing the occurrence of record temperatures in historical temperature data from Europe and the United States. It turns out that the frequency in which new hot and cold records occur today is already significantly influenced by the changing climate. This analysis is substantially extended in chapter 8, where also the statistics of the values of record temperatures are modeled and discussed.

Main subject of part III is the record statistics of random walks. We focus on discrete-time random walks where each entry X_i is obtained as the sum of the previous entry X_{i-1} and an i.i.d. RV. In contrast to the models introduced in the first part, the entries of such a random walk are correlated and therefore their record statistics differs systematically from the i.i.d. case. In the chapters 9 and 10 the manifold properties of record events in random walks with a constant drift (bias) are studied in detail. Again, the record rate P_n and the distribution of the record number R_n are computed and discussed. While chapter 9 is focused on biased Gaussian random walks and compares them with record-breaking stock prizes, chapter 10 gives a complete description of the asymptotic record statistics of biased random walks and Lévy flights. Chapter 11 is about the record statistics of the maximum of ensembles of multiple independent random walkers.

The main application of the results derived in part III is found in the financial sciences and is subject of the subsequent part IV (chapter 12). This part consists of a detailed study of record-breaking stock prizes in the American Standard and Poors 500 stock index. It is discussed how well the record-breaking daily stock returns and stock prizes can be described using simple stochastic models like independent random numbers or random walks.

In the fifth part, an attempt of a general review of the recent developments in the field of record statistics is presented. Chapter 13 summarizes both recent theoretical findings and, more briefly, the progress made in the different applications of the theory of records. We introduce and illustrate various elementary stochastic models and discuss them with respect to their record processes. Following this, chapter 14 summarizes and discusses the contributions in this thesis in a few words.

The following section 1.1 gives of a brief summary of the most important 'classical' findings in the history of record statistics of independent and identically distributed random variables. Furthermore, in section 1.2, some important, much newer findings for the record statistics of symmetric random walks are presented, because of their particular importance for the parts III and IV of this thesis. In this introduction, the relevant quantities that will be of interest in the following chapters are defined and some fundamental results are presented, which form the bedrock of most of the recent studies in this area. These new developments are summarized briefly in the remainder of chapter 1 (section 1.3). The most relevant publications of the last decades and their importance for applications also in other areas of research are described and discussed. This introduction can be seen as a starting point for all the individual contributions in this work. In the subsequent chapter 2, the four main parts of this thesis are then introduced in more detail.

1.1 Record statistics of i.i.d. RV's

Most of the classical literature on record statistics discusses the properties of records in time series of independent and identically distributed (i.i.d.) random variables (RV's) (see for instance [2–4]). A comprehensive summary can be found in the book of Arnold et al. [4], which contains a lot of the results presented in this chapter.

Let us consider a time series of i.i.d. RV's $X_1, X_2, ..., X_n$,[3] all sampled from the same continuous distribution with probability density function (pdf) $f(x)$. As already mentioned, an entry X_n in such a time series is an upper record if

$$X_n > \max\{X_1, X_2, ..., X_{n-1}\} \tag{1.2}$$

and X_1 is, by definition, the first record. Of course, one can also consider lower records and X_n is a lower record if $X_n < \min\{X_1, X_2, ..., X_{n-1}\}$. Fig. 1.2 illustrates the evolution of upper and lower records in a time series of i.i.d. RV's.

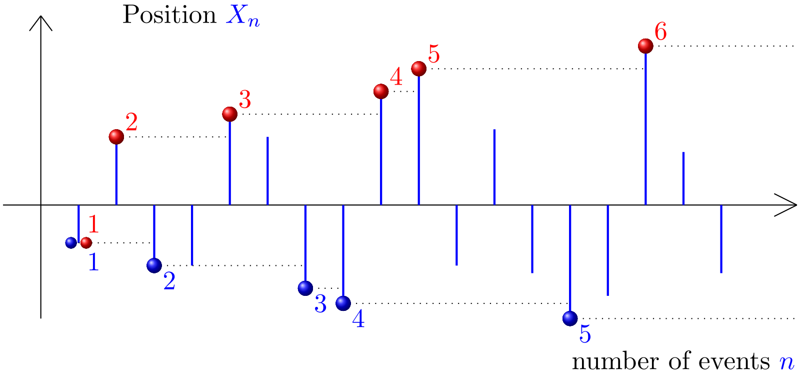

Figure 1.2: Sketch of the record process of i.i.d. RV's. The red (blue) balls mark the upper (lower) records. Here, one has a total number of six upper and five lower records. Of course, the first entry X_1 is both, the first upper and the first lower record.

Probably the most important and most frequently mentioned quantity in this work is the probability P_n that a certain entry n in such a series is a record. We often refer to P_n as the (upper or lower) record rate. For i.i.d. RV's P_n is completely universal for all continuous distributions, which means that it is entirely independent from the choice of the underlying pdf $f(x)$. One finds that

$$P_n = \frac{1}{n}, \tag{1.3}$$

as already mentioned above. This can be seen as follows: Since all n RV's are sampled from the same distribution, every one of them is equally likely to be the largest. Therefore also the last entry X_n is the largest with the same probability. Now, an arbitrary entry is the largest in one of n cases and thus the last one is a record with a probability $P_n = 1/n$.

Despite the simplicity of this so-called *stick-shuffling* argument, it is worth noticing that the record rate can also be derived using a more systematic approach. Since the probability for a RV sampled from a distribution with pdf $f(x)$ to be smaller than some value x is given by the cumulative distribution function (cdf) $F(x) := \int^x dx f(x)$, P_n can be computed by evaluating the integral

$$P_n = \int dx\ f(x)\, F^{n-1}(x)\,. \tag{1.4}$$

---

[3] Occasionally, in this thesis we consider time series of RV's with entries $X_0, X_1 ..., X_n$.

This is just the probability density of a record with value x in the nth step integrated over all possible values x. Partial integration yields

$$P_n = 1 - (n-1) \int \mathrm{d}x \, f(x) \, F^{n-1}(x) = 1 - (n-1) \, P_n, \tag{1.5}$$

which leads directly to $P_n = 1/n$ as in Eq. 1.3.

As a direct corollary of this result, one can also compute the average of the number of records R_n that one expects up to the nth step. This mean record number $\langle R_n \rangle$ is obtained by summing up the record rate from one to n: $\langle R_n \rangle = P_1 + P_2 + ... + P_n$. Using a well known result for the large n behavior of the harmonic numbers $H_n := \sum_{k=1}^{n} 1/k$ [5] one finds that

$$\langle R_n \rangle = \sum_{k=1}^{n} P_k = \sum_{k=1}^{n} \frac{1}{k} \approx \ln n + \gamma, \tag{1.6}$$

where $\gamma \approx 0.577215...$ is the Euler-Mascheroni constant. This result is interesting simply because it diverges. The probability for a new record never goes to zero and there is no upper bound to the total number of records in the i.i.d. case.

The other way around, the fact that the mean record number $\langle R_n \rangle$ matches the nth Harmonic number $H_n := \sum_{k=1}^{n} \frac{1}{k}$ can be used to prove that the Harmonic series $H_\infty = \sum_{k=1}^{\infty} \frac{1}{k}$ diverges [2]. If one assumes that there is never an ultimate record and every record value gets exceeded eventually, the mean record number for $n \to \infty$ must diverge and therefore also the Harmonic series.

1.1.1 Correlations between records from i.i.d. RV's

A very important feature of the record statistics of i.i.d. RV's is the fact that the individual record events are uncorrelated. The probability P_n for a record in the nth entry is completely independent from the probability P_m for a record in another entry $m \neq n$. In other words, the probability $P_{n,m}$ for a record both in the nth and the mth step factorizes:

$$P_{n,m} = \mathrm{Prob}\left[X_n \text{ and } X_m \text{ are both records} \right] = P_n \cdot P_m. \tag{1.7}$$

For two consecutive entries n and $m = n+1$ this can be shown as follows: With an argument similar to the one that leads to Eq. 1.4, the probability $P_{n,n+1}$ for the entries n and $n+1$ to be records in the same time series is given by

$$\begin{aligned} P_{n,n+1} &= \int \mathrm{d}x_{n+1} \, f(x_{n+1}) \int^{x_{n+1}} \mathrm{d}x_n \, f(x_n) \, F^{n-1}(x_n) \\ &= \int \mathrm{d}u \, f(u) \, F^n(u) - (n-1) \int \mathrm{d}u \, f(u) \int^{u} \mathrm{d}v f(v) \, F^{n-1}(v) \\ &= P_{n+1} - (n-1) \, P_{n,n+1}. \end{aligned} \tag{1.8}$$

With this one finds that

$$P_{n,n+1} = P_n P_{n+1} = \frac{1}{n(n+1)}. \tag{1.9}$$

This derivation can be easily generalized to prove that also the probabilities for records in the nth and the $(n+k)$th step are uncorrelated. Additionally, it is straightforward to show that also multi-point correlations vanish. The joint record rate of a set of i integer and pairwise different numbers $n_1, ...n_i$ factorizes:

$$P_{n_1,...,n_i} = P_{n_1} P_{n_2} \cdot ... \cdot P_{n_i}. \tag{1.10}$$

1.1.2 Full distribution of the record number R_n

Knowing that the individual record events are uncorrelated, one can compute the variance and also the full distribution of the record number R_n. For that purpose it is helpful to introduce a set of (uncorrelated) indicator random variables (cf. [4]), $I_1, I_2, ..., I_n$, as follows:

$$I_n := \begin{cases} 1, & \text{if } X_n \text{ is a record,} \\ 0, & \text{else.} \end{cases} \tag{1.11}$$

In this notation, the record rate P_n is just the probability $\mathrm{Prob}\,[I_n = 1] = \langle I_n \rangle$ and, because of Eq. 1.10, the joint probability for a set of entries $X_{n_1}, X_{n_2}, ..., X_{n_i}$ to be records is given by

$$\mathrm{Prob}\,[I_{n_1} = 1, I_{n_2} = 1, ..., I_{n_i} = 1] = \frac{1}{n_1}\frac{1}{n_2} \cdot ... \cdot \frac{1}{n_i}. \tag{1.12}$$

In terms of the indicator random variables, the record number R_n is obtained as the sum $R_n = \sum_{k=1}^n I_k$, and one can reobtain the result for the mean record number $\langle R_n \rangle$ (Eq. 1.6) as follows:

$$\langle R_n \rangle = \sum_{k=1}^{n} \langle I_k \rangle = \sum_{k=1}^{n} \frac{1}{k} \approx \ln n + \gamma \tag{1.13}$$

Thanks to the new notation, one can now also compute the variance $\mathrm{Var}\,(R_n)$ of the mean record number. With $\mathrm{Var}\,(R_n) = \langle (R_n - \langle R_n \rangle)^2 \rangle = \sum_{k=1}^n \mathrm{Var}\,(I_k)$ one finds that

$$\mathrm{Var}\,(R_n) \quad = \quad \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k^2} \right) \approx \ln n + \gamma - \frac{\pi^2}{6}. \tag{1.14}$$

In the second step one has to use $\lim\limits_{n\to\infty} \sum_{k=1}^n 1/k^2 = \pi^2/6$ [5]. Apparently, in the large n limit, the variance of the record number behaves like the mean record number minus a constant: $\mathrm{Var}\,(R_n) = \langle R_n \rangle - \pi^2/6$ and the asymptotic standard deviation of the mean record number grows proportionally to $\sqrt{\ln n}$.

Using the so-called Stirling numbers of the first order (cf. [5]), the full distribution of the record number R_n can be described in terms of its probability generating function $\langle z^{R_n} \rangle$ [6]. With help of the indicator functions, $\langle z^{R_n} \rangle$ can be expressed as

$$\langle z^{R_n} \rangle = \langle z^{\sum_{k=1}^n I_k} \rangle = \prod_{k=1}^{n} \langle z^{I_k} \rangle = \prod_{k=1}^{n} \left(\frac{z}{j} + \left(1 - \frac{1}{j} \right) \right), \tag{1.15}$$

where, in the last step, one uses that z^{I_j} is z with probability $1/j$ and one with a probability of $1 - 1/j$. Therefore, $\langle z^{I_j} \rangle$ is given by $z/j + (1 - 1/j)$. The probability of having $R_n = k$ records in n steps can now be computed using the kth derivative of the generating function $\langle z^{R_n} \rangle$:

$$\mathrm{Prob}\,[R_n = k] = \frac{1}{k!} \frac{\mathrm{d}^\mathrm{k}}{\mathrm{d}z^k} \langle z^{R_n} \rangle |_{z=0} \tag{1.16}$$

The k-fold derivation of $\langle z^{R_n} \rangle$ is quite complicated and we will not discuss it in detail. It was shown in [4] that Eq. 1.16 can be reduced to

$$\mathrm{Prob}\,[R_n = k] = \frac{S_n^k}{n!}, \tag{1.17}$$

where S_n^k is an unsigned Stirling number of the first order [5]. $S_n^k = \left[{n \atop k} \right]$ gives the number of permutations of a set of n elements that has k disjoint cycles [5]. As a simple example, the Stirling number $S_3^2 = \left[{3 \atop 2} \right]$ is given by 3, since there are 3 ways to permutate the set $\{1, 2, 3\}$ into two cycles: $\{(1), (2, 3)\}, \{(1, 2), (3)\}$ and $\{(1, 3), (2)\}$.

Since all indicator functions I_k have a finite variance (it is straightforward to show that $\mathrm{Var}(I_k) \leq \frac{1}{4}$ for all possible values of k), it is a consequence of the central limit theorem that, for $n \to \infty$, the distribution of the record number R_n approaches a Gaussian form. This limiting distribution must have a mean value $\mu \approx \langle R_n \rangle$ a standard deviation of $\sigma \approx \sqrt{\langle R_n \rangle}$. Therefore, one finds that the rescaled mean record number $(R_n - \langle R_n \rangle) / \sqrt{\langle R_n \rangle}$ has a standard normal distribution in the large n limit:

$$\frac{R_n - \ln n}{\sqrt{\ln n}} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, 1). \tag{1.18}$$

1.1.3 Record times

The formalism developed above can also be used to compute the distribution of record times. We define T_k as the time at which the kth record occurs and we are interested in the probability $\mathrm{Prob}[T_k = n]$ that this happens at time n. As discussed in [4] this probability is given by the joint probability for a record at time n together with a record number of $R_{n-1} = k - 1$ at time $n - 1$. Since these two probabilities are independent, we have

$$\mathrm{Prob}[T_k = n] = \frac{1}{n} \mathrm{Prob}[R_{n-1} = k - 1] = \frac{S_{k-1}^{n-1}}{n!}. \tag{1.19}$$

With this, one can show an important result for the mean record times: The expectation value of the record time T_k is divergent for all values $k > 1$:

$$\langle T_k \rangle = \sum_{n=k}^{\infty} n \, \mathrm{Prob}[T_k = n] = \infty. \tag{1.20}$$

Along with this the expectation values of the inter-record times $\Delta_k := T_k - T_{k-1}$ diverge and

$$\langle \Delta_k \rangle = \langle T_k - T_{k-1} \rangle = \infty. \tag{1.21}$$

A full proof of these results is complicated and can for instance be found in [7–9]. It can also be shown that the logarithms $\ln T_k$ of the waiting times approach a Gaussian distribution with mean k and standard deviation $\sqrt{k}$ in the limit of large record numbers k. In other words, the rescaled logarithmic waiting times $(\ln T_k - k) / \sqrt{k}$ approach a standard normal distribution:

$$\frac{\ln T_k - k}{\sqrt{k}} \xrightarrow[k \to \infty]{d} \mathcal{N}(0, 1). \tag{1.22}$$

Interestingly, the logarithmic inter-record times have the same asymptotic distribution [4]:

$$\frac{\ln \Delta_k - k}{\sqrt{k}} \xrightarrow[k \to \infty]{d} \mathcal{N}(0, 1). \tag{1.23}$$

Furthermore, Glick [2] showed that the ratio Δ_k / T_k is uniformly distributed on the interval $(0, 1)$. On average, the time Δ_k between the $(k - 1)$th record and the kth accounts for half of the total waiting time T_k of the kth record.

1.1.4 Record values

While all previously introduced quantities are completely universal for arbitrary series of i.i.d. RV's, this universality gets lost if one considers the values of record-breaking events. The statistical properties of record values, such as their mean values and their distributions,

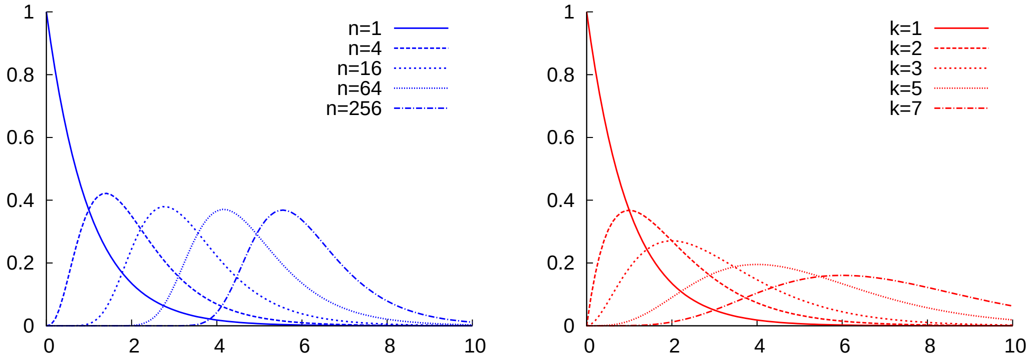

Figure 1.3: Left: The normalized record value distribution of records that occur at fixed times $n = 1, 4, 16, 64, 256$ in a series of i.i.d. RV's from an exponential distribution with $f(x) = e^{-x}$ (and $x > 0$). The distributions were obtained using Eq. 1.25. **Right:** The normalized distributions of records with a fixed record number $k = 1, 2, 3, 5, 7$ (see Eq. 1.30), again for i.i.d. RV's from the exponential distribution.

depend heavily on the shape of the underlying distribution. There are in principle two important types of record values that one can consider: The values of records that occur in a certain entry n in our time series and the values of records with a fixed record number k.

In Fig. 1.3 the distributions of records with fixed n and fixed k for a simple exponential distribution are illustrated with $f(x) = e^{-x}$ (for $x > 0$) and some selected values of n and k.

It is quite simple to give the full distribution of a record that occurs at a fixed time step n in a series of i.i.d. RV's. One can compute the normalized cumulative distribution function (cdf) $Q_n(x)$ of a record with value n from the integral expression of the probability for a record in the nth event (Eq. 1.4). In this context, $Q_n(x)$ is the probability that the nth entry X_n is larger than all previous entries and smaller than x:

$$\begin{aligned} Q_n(x) &:= P_n^{-1}\,\mathrm{Prob}\left[X_n > \max\{X_1, X_2, ..., X_{n-1}\}, X_n < x\right] \\ &= n \int^{x} f(x)\, F^{n-1}(x)\,. \end{aligned} \tag{1.24}$$

Here, $P_n^{-1} = n$ was used to normalize the cdf. Therefore the pdf $f_n(x)$ of a record that occurs in the nth step is simply given by

$$f_n(x) = \frac{\mathrm{d}}{\mathrm{d}x} Q_n(x) = n f(x)\, F^{n-1}(x)\,. \tag{1.25}$$

With this result it is straightforward to compute the mean value as well as arbitrary higher moments of record events with a fixed n. The mean μ_n of a record at time n is

$$\mu_n = n \int \mathrm{d}x\; x f(x)\, F^{n-1}(x)\,. \tag{1.26}$$

For an exponential distribution with $f(x) = \nu^{-1} e^{-x/\nu}$ ($x > 0$ and $\nu > 0$) this leads to

$$\mu_n^{(exp)} = \int_0^\infty \mathrm{d}x\; x\nu^{-1} e^{-\frac{x}{\nu}} \left(1 - e^{-\frac{x}{\nu}}\right)^{n-1} = \nu H_n \approx \nu\left(\ln n + \gamma\right), \tag{1.27}$$

where the integral was evaluated using iterated partial integration. For a Gaussian distribution with

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}, \tag{1.28}$$

and a standard deviation σ it is more complicated to compute the mean μ_n which, in the limit of large n, approaches

$$\mu_n^{(Gauss)} \approx \sqrt{\ln\left(\frac{n^2}{2\pi}\right)}. \tag{1.29}$$

The derivation of the distribution of record values with a fixed record number k is more difficult. As shown previously, a record with record number k can occur at an arbitrary time $n \geq k$. In principle, this would require averaging over all distributions of records with fixed values of n and k. Fortunately, there is an easier method, thanks to a special property of the exponential distribution $f\left(x\right) = e^{-x}$ with $x > 0$. This distribution has the so-called *lack-of-memory* property. This basically means that for any k, the value of the kth record will be given by the value of the $(k-1)$st plus an exponential random variable drawn from the same distribution e^{-x}. With this one can show that a record with record number k has the pdf

$$f_k\left(x\right) = \frac{1}{(k-1)!}x^{-k}e^{-x}, \tag{1.30}$$

which is a Gamma-distribution with shape parameter k [5]. One can use this result to compute the distribution of a record with record number k for an arbitrary continuous underlying distribution. In [4] it is shown that the kth record of such an arbitrary distribution can be given as a function of the kth record from the exponential distribution (see also chapter 13). In fact, the kth record in a series of i.i.d. RV's sampled from the distribution $f\left(x\right)$ has the pdf

$$f_k\left(x\right) = \frac{f\left(x\right)}{(k-1)!}\left(-\ln\left(1-F\left(x\right)\right)\right)^k. \tag{1.31}$$

1.1.5 Limit laws for record values

In the theory of extreme events, it is well known that the distributions of the maximum

$$M_n := \max\{X_1, X_2, ..., X_n\} \tag{1.32}$$

of a given set $X_1, X_2, ..., X_n$ of i.i.d. RV's from a continuous pdf $f\left(x\right)$ converge to one of three possible limiting distributions [10] (for an introduction see for instance [11, 12]). The Fisher-Tippett-Gnedenko Theorem [11, 12] of classical extreme value theory states that if the total number of i.i.d. RV's, n, is going to infinity, the limiting distribution of the maximal value can always be rescaled to one of the following three shapes:

I - Weibull distribution: For RV's with a finite support, the rescaled maximum approaches a (reversed) Weibull distribution. It has the following cdf:

$$F_{\mathrm{I}}\left(x\right) = \begin{cases} e^{-(-x)^{-\kappa}}, & \text{for } x < 0, \\ 1, & \text{for } x \geq 0 \end{cases}, \tag{1.33}$$

with one free parameter $\kappa < 0$.

II - Gumbel distribution: The cdf of the rescaled maximum of a set of RV's from a distribution with an infinite support that decay faster than any power-law (e.g. with an exponential tail) approaches a distribution of the Gumbel form:

$$F_{\mathrm{II}}\left(x\right) = e^{e^{-x}}. \tag{1.34}$$

III - Fréchet distribution: For RV's with infinite support and a tail that decays like a power law, the limiting distribution of the rescaled maximal value is of the Fréchet form:

$$F_{\mathrm{III}}(x) = \begin{cases} 0, & \text{for } x < 0, \\ e^{-x^{-\kappa}}, & \text{for } x \geq 0, \end{cases} \tag{1.35}$$

again with one free parameter $\kappa > 0$.

It turns out that these three universality classes are also of importance for the distributions of record values. In fact, these distributions, both with a fixed record number k, as well as the ones that occur at a fixed time n also approach one of three different limiting distributions.

In the case of the record values with a fixed n, the situation is particularly simple. In Eq. 1.24 we already gave an expression for the normalized cdf of a record that occurs in a fixed entry of the series. The cumulative distribution of the maximum M_n of a series of RV's $X_1, ..., X_n$ has exactly the same shape:

$$\mathrm{Prob}\left[M_n < x\right] = n \int^{x} \mathrm{d}x \; f(x)\, F^{n-1}(x). \tag{1.36}$$

Therefore, the distribution of the record values with a fixed n must approach the same limiting distribution as the maximal value M_n.

In the case of record values with fixed record number k it is more complicated and the record values do not approach the same limiting distributions. However, in 1973, Resnick [13] showed that the distributions of the record values with a record number k obey a limit law of the form:

$$\Phi\left(-\ln\left(-\ln\left(F_i(x)\right)\right)\right), \tag{1.37}$$

where the functions $F_i(x)$ with $i = \mathrm{I}, \mathrm{II}, \mathrm{III}$ are the limiting distributions of the maximal value and $\Phi(x)$ is the cdf of a Gaussian standard normal distribution with

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-t^2/2} \mathrm{d}t. \tag{1.38}$$

With this result Resnick proved that, under proper rescaling, the limiting distribution of a record with record number k approaches one of these three distributions [13]:

I - Negative-log-normal distribution: For RV's of the Weibull class, with a distribution that has a finite support, the (rescaled) record value distribution approaches the following cdf:

$$\Phi_{\mathrm{I}}(x) = \begin{cases} \Phi\left(\ln(-x)^{-\kappa}\right), & \text{for } x < 0, \\ 1, & \text{for } x \geq 0, \end{cases} \tag{1.39}$$

with $\kappa < 0$.

II - Normal distribution: Record values in series of RV's from the Gumbel class are normally distributed and can be rescaled to a standard normal distribution:

$$\Phi_{\mathrm{II}}(x) = \Phi(x). \tag{1.40}$$

III - Log-normal distribution: For RV's of the Fréchet type, from distributions that decay like a power-law, the rescaled limiting distribution is of the following form:

$$\Phi_{\mathrm{III}}\left(x\right)=\begin{cases}0, & \text{for } x<0,\\ \Phi\left(\ln\left(x^{-\kappa}\right)\right), & \text{for } x\geq 0,\end{cases} \tag{1.41}$$

with $\kappa > 0$.

In the following, these findings will be particularly important for our studies of the record statistics of RV's, which are uncorrelated but time-dependent. They allow us to analyze and discuss the records in our models in the context of the three universality classes of extreme value statistics. It will turn out that these classes are helpful to understand and interpret the record statistics of more complicated and time-dependent stochastic processes.

1.2 Record statistics of symmetric random walks

Mostly because of its importance for the parts III and IV of this thesis we will now introduce a second stochastic process and discuss its record statistics. A discrete-time random walk is a simple example for a process of correlated random variables and is defined as a sum of random steps, which are again independently sampled i.i.d. RV's as in the previous section[4]. The entries $X_1, X_2, ..., X_n$ of a discrete-time and continuous-space random walk can be described as follows:

$$X_n = X_{n-1} + \eta_n, \tag{1.42}$$

where the η_i's are i.i.d. RV's drawn from a continuous distribution $f\left(\eta\right)$. X_0 is the origin of the random walk and usually we will set $X_0 = 0$ without loss of generality (see Fig. 1.4).

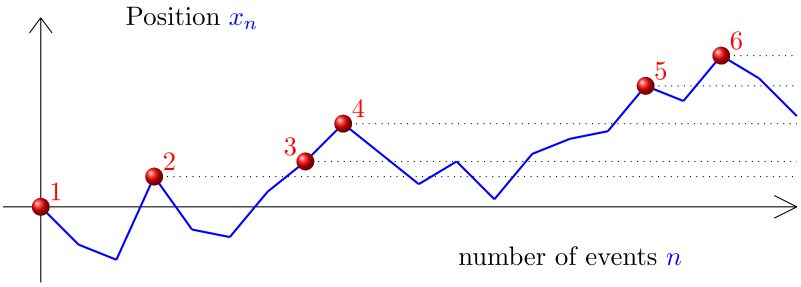

Figure 1.4: Sketch of the record process of a discrete-time random walk. The red balls mark the progression of the upper record.

The simplest scenario one can consider is a symmetric random walk with a fully symmetrical jump distribution $f\left(\eta\right) = f\left(-\eta\right)$. However, for this case already it turned out that it is much more complicated to obtain quantities like the record rate P_n and mean record number $\langle R_n \rangle$ than in the case of i.i.d. RV's. Only a few years ago, in 2008, Majumdar and Ziff [15] computed the full distribution of the record number R_n of a symmetric random walk. Interestingly, this distribution is completely independent from the choice of the symmetric jump distribution $f\left(\eta\right)$. We will now illustrate how to derive these results following the analysis of Majumdar and Ziff [15].

The main idea behind the derivation of the record statistics of the symmetric random walk is to subdivide a process with n steps and R_n (upper) records into a series of $R_n - 1$ first-passage and one survival problem[5]. Fortunately, the random walk has the so-called

---

[4] An introduction can, for instance, be found in the book by Weiss [14].

[5] For a detailed introduction to the theory of these problems see for instance the book by Redner [16].

renewal property, which basically states that the process starting from X_n at a given time n is again a random walk with X_n as a new origin that behaves exactly like the original process. With this it is easy to see that the time l_k between the kth record and the $(k+1)$st record is simply the time required until a random walker starting from the origin (at the time of the kth record) becomes positive for the first time. This duration is also known as the (positive) first-passage time.

Now one can define the (positive) first-passage probability $\phi\left(l\right)$, which is the probability that the walker crosses the origin for the first time after l steps. Similarly the survival probability $q\left(l\right)$ is the probability that a random walk starting at zero stays below (or above) the origin for the first l time steps. It is easy to see that $\phi\left(l\right)$ and $q\left(l\right)$ are connected by $\phi\left(l\right) = q\left(l-1\right) - q\left(l\right)$.

Thanks to a nontrivial theorem by Sparre Andersen [17, 18], it is known that these quantities are entirely universal for symmetric random walks. Sparre Andersen showed that the generating function of the survival probability $\tilde{q}\left(z\right) = \sum_{l=0}^{\infty} q\left(l\right) z^l$ for any random walk with a continuous and symmetric jump distribution, is given by

$$\tilde{q}\left(z\right) = \frac{1}{\sqrt{1-z}}. \tag{1.43}$$

Similarly the first-passage probability $\phi\left(l\right)$ has the generating function $\tilde{\phi}\left(z\right) = 1 - \sqrt{1-z}$. From these results one can obtain

$$q\left(l\right) = \binom{2l}{l} 2^{-2l} \qquad \text{and} \qquad \phi\left(l\right) = \binom{2l}{l} \frac{2^{-2l}}{2l-1}. \tag{1.44}$$

Using these findings one can compute the probability for a random walk with R_n records and inter-record times $l_1, ..., l_{R_n}$ (with $\sum_{k=1}^{R_n} l_k = n$):

$$P\left(l_1, ..., l_{R_n} | n\right) = \phi\left(l_1\right) \cdot \dots \cdot \phi\left(l_{R_{n-1}}\right) q\left(l_{R_n}\right) \delta_{\sum_{k=1}^{R_n} l_k, n}. \tag{1.45}$$

The survival probability $q\left(R_n\right)$ and the δ-function are necessary to exclude the possibility of additional records after the one with record number R_n.

Now, to compute the overall probability of R_n records in n time steps, one has to sum $P\left(l_1, ..., l_{R_n} | n\right)$ over all possible sets $l_1, ..., l_{R_n}$ with $\sum_{k=1}^{R_n} l_k = n$. This summation was done by Majumdar and Ziff [15] using the generating function of the probability $P\left(l_1, ..., l_{R_n} | n\right)$. They showed that the joint probability $P\left(R_n | n\right)$ for R_n records obeys

$$\sum_{n=R_n-1}^{\infty} P\left(R_n | n\right) z^n = \tilde{\phi}\left(z\right)^{R_n-1} \tilde{q}\left(z\right) = \frac{\left(1-\sqrt{1-z}\right)^{R_n-1}}{\sqrt{1-z}}. \tag{1.46}$$

Expanding this result in powers of z, Majumdar and Ziff found the full distribution of the record number R_n. The probability $P\left(R_n | n\right)$ of R_n records in n steps is given by

$$P\left(R_n | n\right) = \binom{2n-R_n+1}{n} 2^{-2n+R_n-1}. \tag{1.47}$$

The mean record number $\langle R_n \rangle$ can be obtained as the first moment of this distribution. One finds that

$$\langle R_n \rangle = \left(2n+1\right) \binom{2n}{n} 2^{-2n} \qquad \text{and} \qquad P_n = \binom{2n}{n} 2^{-2n}. \tag{1.48}$$

In the asymptotic limit of large series length ($n \to \infty$), the distribution $P\left(R_n | n\right)$ approaches a half-Gaussian form:

$$P\left(R_n | n\right) \xrightarrow{n\to\infty} \frac{1}{\sqrt{n\pi}} \exp\left(-\frac{R_n^2}{4n}\right) \tag{1.49}$$

in contrast to the i.i.d. case, where it was found that the asymptotic record number is distributed according to a Gaussian distribution. In this limit, the mean record number $\langle R_n \rangle$ is proportional to $\sqrt{n}$ (and not logarithmically as in the case of i.i.d. RV's). For $n \gg 1$ one obtains

$$\langle R_n \rangle \approx \sqrt{\frac{4n}{\pi}} \qquad \text{and} \qquad P_n \approx \frac{1}{\sqrt{n\pi}}. \tag{1.50}$$

Apparently, the record rate P_n of the discrete-time random walk decays much slower than for i.i.d. RV's with $P_n = 1/n$. Correspondingly, the mean record number $\langle R_n \rangle$ is much larger than the i.i.d. results of $\langle R_n \rangle \propto \ln n$.

1.3 Recent theoretical progress & applications

The results for the record statistics of i.i.d. RV's discussed before have all been derived at least 30 years ago and can, among other findings, be considered as the classical theory of records. Since then, but especially in the last 10 to 15 years, a lot of research was done towards a better understanding of the record statistics of more complicated stochastic processes, in particular time-dependent and correlated RV's. Additionally, the old and new theoretical results found many new applications in various areas of science. While a detailed and comprehensive review of these developments is the subject of chapter 13, I will now briefly summarize the important publications in this field to give the reader a better idea of the current state of research. This introduction is supposed to illustrate the general framework in which the contributions in this thesis were developed.

Temperature records

Certainly the most important application of record statistics in the last years was the study of record events in climatology. The evident and probably mostly anthropogenic climatic change [19] that started to increase the global mean temperature in the second half of the last century made scientists and also the general public particularly interested in extreme weather events like severe storms, heavy rainfall, or record-breaking heatwaves. Over the last 30 years, the study of extremes in climate and also the adaptation to them has become a well established branch of research in climatology (cf. [20–26]). Most likely because of the frequent and attention-grabbing media coverage, researchers, including us, have more recently also started to look at the statistics of climate records.

Even though some earlier publications about record-breaking temperatures exist [27–32], the first comparison between theoretical results in record statistics and temperature measurements was performed a decade ago by Benestad [33, 34]. He analyzed the global monthly mean temperatures and measurements from several Norwegian weather stations. In 2006, he also considered record-breaking precipitation events [35]. Also in this year, Redner and Petersen considered time series of daily temperature measurements for individual calendar days that were recorded in Philadelphia [36]. They made many important theoretical observations, but, due to the fact that they only considered one individual weather station, they could not satisfactorily show a connection between global warming and the occurrence of record temperatures. A few years later, this connection was established by a series of authors. In 2009, Meehl et al. [37] found that, at that time, more than twice as many heat records than cold records were measured in the United States. In an independent study of European weather stations we showed that the rate of daily heat records in Europe is significantly increased [38, 39] (see chapter 7). We also proposed to use the simple Linear Drift Model (LDM) that accurately predicts the effects of global warming on the statistics of record-breaking temperatures. In the following, these findings were confirmed and extended by Newman et al. [40], as well as Elguindi et al. [41] and Rahmstorf et al. [42]. Both Elguindi et al. and Rahmstorf et al. also considered our simple LDM for record temperatures and could, for the most part, confirm its validity. In 2012, we extended this model and analyzed the values of record-breaking temperatures [43] (see chapter 8).

Time-dependent models

The LDM that was employed to model record temperatures proved to be interesting and of importance on its own. It was first introduced by Ballerini and Resnick in 1985 [44, 45], who computed the record rate in the limit of infinite series length. These results were later refined and extended by Borovkov [46]. Motivated by our study of temperature records, we considered this model in 2010 and discussed the behavior of the record rate also for short series length in the context of the three universality classes of extreme value statistics [47] (see chapter 3). During these studies, we made the interesting observation that in time series of RV's with a linear drift, the individual record events are no longer uncorrelated as in the case of i.i.d. RV's. In fact, it turned out that the correlations between records have a complicated and manifold behavior. In some sense, record events can both repel and attract each other, depending on the characteristics of the underlying distribution. These results and a promising application as a test for heavy-tail properties of observational data were published separately [48, 49] (see chapters 4 and 5).

Apart from these studies on the LDM, other possible processes of uncorrelated RV's with non-identical, time-dependent distributions were considered and applied in a different field of research, namely biology. In 2007, Krug [50] considered RV's sampled from broadening distributions and discussed their record statistics. Eliazar and Klafter [51] analyzed a similar problem in 2009 by discussing records in a model of stochastic growth, which was motivated by problems in evolutionary biology. In this context, Krug and Jain studied the connection between biological evolution and records [52] building upon previous work by Kauffman and Levin [53], as well as Sibani et al. [54]. Since then, record and extreme value statistics have become important in the theory of adaption (for an introduction see for instance the article of Orr [55]), especially for adaptive walks on fitness landscapes [56] (see also [57]).

Discreteness & rounding

Very important for experimentalists is certainly the problem of discreteness. Even though the entire classical theory presented in section 1.1 deals with RV's from entirely continuous distributions, one must always be aware of the fact that experimental observations (temperatures, flood heights, sport results,...) can only be measured up to a certain accuracy. They will always be rounded to a certain precision. One way to deal with this problem is to consider the record statistics of discrete distributions, where ties are allowed. This was already done by Vervaat [58] in 1973 and later, for instance, by Prodinger [59], Gouet et al. [60] and Key [61]. Alternatively, Gouet et al. [62] studied the statistics of records that are only counted if they exceed the previous one by a certain constant value $\delta \neq 0$, the so-called δ-records. Eliazar introduced the similar concept of geometric records [63, 64], which only count as such when they exceed a certain multiple of the last record. From an experimental point of view, the probably most realistic model of RV's that are first drawn from a continuous distribution and then discretized by rounding, was first discussed in a recent article of ours [65] (see chapter 6).

Records in correlated processes & finance

The line of research opened up by Majumdar and Ziff in 2008 [15] goes in an entirely different direction. As previously discussed, they used a powerful theorem by Sparre Andersen [6, 17, 18] to study the statistics of record-breaking events in random walks and Lévy flights. They showed that the full distribution of the record number and therefore also the mean record number and the record rate of symmetric, discrete-time random walks with a continuous jump distribution is completely universal for all possible choices of this jump distribution.

This amazing result entailed a series of other publications about the record statistics of random walks. In 2011, Sabhapandit generalized some of the findings of Majumdar and Ziff to random walks with a continuous distribution of waiting-time between the individual steps [66] (see also [57]). In the same year, we considered biased Gaussian random walks with an asymmetric jump distribution and managed to compute some new approximate results for the record rate and mean record number [67]. In this publication we also demonstrated that the record statistics of stock data from the Standard and Poors 500 index [68] can, to some degree, be modeled using biased random walks (see also [69] and chapter 12). The problem of a biased Lévy flight with jumps sampled from a Cauchy distribution was considered by Le Doussal and Wiese [70]. More recently, the complete asymptotic record statistics of biased random walks and Lévy flights were computed together with Majumdar and Schehr [71] (see chapter 10) and we found a surprisingly diverse and manifold universal behavior of the record number distribution. In 2012, we studied ensembles of multiple independent random walkers and Lévy flights and compared the record statistics of their maximal value to ensembles of randomly selected stocks from the Standard and Poors 500 index [72] (see chapter 11). Also building up on the findings of Majumdar and Ziff is the work of Edery et al. [73], who made a first step towards the understanding of records in higher dimensional Markov processes by performing various numerical simulations of the scaling behavior of the record distance of a diffusion process from its origin.

Further theoretical results & applications

Besides these applications in climatology, biology and finance, the theory of records is also useful in various other areas of science. Oliveira et al. found a record process in a model for high-temperature superconductors [74] and, in a similar context, Sibani et al. found a connection between the problem of magnetization and aging in spin-glasses and records [75–77]. Also hydrologists found record events interesting in the past: In 2001, Vogel et al. [78] examined the frequency of record-breaking floods in observations from hundreds of gauging stations in the United States. In a recent study, the interdisciplinary relevance of the theory of records was demonstrated in an analysis of the movement of ants [79]. Of course, the statistics of world records in sports were also studied: Gembris et al. [80, 81] analyzed the progression of world records in athletics and compared them with the record statistics of i.i.d. RV's.

From the mathematical point of view, there has been more progress in many different directions. For instance, researchers have studied record processes with randomly sampled waiting times between the individual entries and, in particular, records in Poisson processes (see for instance [82–84]). In the book by Arnold [4], various more complicated models, introduced in the last decades are discussed, for instance the so-called Pfeiffer model, where the process is constructed in such a way that the individual record values form a Markov chain[6] [4, 85]. In the 1990's, concepts of multivariate records in two- or higher-dimensional time series were developed [4, 86, 87].

---

[6] A stochastic process, where each entry depends only on the previous entry in the series. For an introduction see for instance [6].

Chapter 2

Outline

2.1 Part I - Records in uncorrelated random variables

In this part, we discuss the statistics of record events in uncorrelated random variables (RV's). While the theory of record-breaking events in time series of independent and identically distributed (i.i.d.) RV's was introduced in the first chapter, we will now, for most of part I, consider RV's that are not identically distributed. In particular, we focus on RV's sampled from a time dependent distribution with a constant linear drift. To be more precise, we are interested in series of RV's $X_1, X_2, ..., X_n$, where an arbitrary entry X_k is of the form

$$X_k = Y_k + ck. \tag{2.1}$$

Here, the Y_i's are i.i.d. RV's drawn from a distribution with probability density function (pdf) $f(y)$. Therefore, the pdf's of the X_k's are given by $f(x - ck)$ and the distribution of the RV X_k is shifted in positive x-direction by ck.

This simple model was first introduced and discussed by Ballerini and Resnick [44, 45] in the 1980's. In the following, we will refer to it as the Linear Drift Model (LDM) of record statistics. It plays a key role in chapters 3, 4 and 5. Our initial motivation to consider this model originates from our study of record-breaking temperatures. The LDM is, in some sense, useful as a toy-model for global warming and we used it to analyze the effect of a slowly increasing global temperature mean on the occurrence of heat and cold records. This analysis is subject of part II.

Chapter 3 — Records and sequences of records from random variables with a linear trend

In chapter 3, an article published in collaboration with Jasper Franke and Joachim Krug, we compute and discuss two quantities in particular: The probability for a record event in a given entry (henceforth record rate) in a series of RV's of the LDM and the closely related ordering probability for this model. The ordering probability is the probability that all entries of a time series occur in ascending order so that $X_1 < X_2 < ... < X_n$. This probability proved to be important for some applications in evolutionary biology and in particular in the context of accessible evolutionary pathways on fitness landscapes [56]. In this article, we consider both the regimes of a small and a very large drift c and discuss our findings in the context of the three universality classes of extreme value statistics [10–12], which were already introduced in chapter 1. We find that, while the drift plays a very important role in the Weibull class of distributions with a finite support, it has a decaying effect in the Fréchet class of power-law tailed distributions with a finite first moment. In the Gumbel class the behavior is more complicated and intermediates between these two regimes. We performed numerous Monte-Carlo-type simulations along with our analytical work that confirm these findings.

Chapter 4 — Correlations between record events in sequences of random variables with a linear trend

Chapters 4 and 5 focus on the interesting problem of correlations between record events in the context of our LDM. As shown in chapter 1, the probabilities for records in certain entries in a series of i.i.d. RV's are pairwise uncorrelated. This does not hold in the presence of a linear drift. In fact, we found that record events in our LDM can be correlated both positively and negatively. In some sense, this can be understood as record events, which are either attracting or repelling each other. In chapter 4 which summarizes joint work with Jasper Franke and Joachim Krug, we study the probability $P_{n-1,n}$ for two successive record events in a times series of RV's of the LDM-type to be a record and more importantly we look at the ratio

$$l_{n,n-1} := \frac{P_{n,n-1}}{P_n P_{n-1}} \tag{2.2}$$

between this quantity and the two individual record rates. For all distributions of the Weibull-type as well as distributions of the Gumbel class that decay faster than an exponential distribution, we found that $l_{n,n-1}$ is smaller than one in the $n \to \infty$ limit. Apparently, in this regime, record events repel each other. If the entry $n-1$ is a record, the probability for a new record in the nth event is smaller than the unconditional record rate P_n. The more surprising result is that, for Gumbel-type distributions decaying slower or at least as slow as the exponential distribution and for distributions of the Fréchet class, $l_{n,n-1}$ is greater than one for $n \to \infty$, meaning that after a record in the $(n-1)$th event, there is an increased probability for a second record in the nth event. This can be interpreted as record events which effectively attract each other.

Chapter 5 — Correlations of record events as a test for heavy-tailed distributions

Chapter 5, also work published together with Jasper Franke and Joachim Krug, describes a possible application of these findings. The idea behind this work is that, since significant positive correlations with a ratio $l_{n-1,n} > 1$ are only possible for time series of RV's from a distribution with tails at least as broad as the one of an Exponential distribution, the correlations between record events could be helpful to detect these RV's in experimental data. Distributions with broader than exponential tails are commonly referred to as 'heavy-tailed' and it is a well known problem to find these heavy-tail properties in observational data (see for instance [88–90]), especially in data sets with a small sample size. It turned out that analyzing the record-correlations after artificially adding a linear drift to the observational data provides a simple and efficient test which is able to verify heavy tails even in very small data sets. In chapter 5 we describe this test and present an application to a well known data set of citation records [91].

Chapter 6 — Rounding Effects in Record Statistics

In the final chapter of this part, we consider a different problem, not closely related to the LDM. In this article, published with Joachim Krug and our collaborators Daniel Volovik and Sidney Redner from the Boston University, we study the effects of rounding on the record statistics of i.i.d. RV's. In all practical applications, measurements are rounded to a certain accuracy, which opens up the possibility for ties. Record events cannot only be broken, but also be tied by a new measurement. This can alter the statistics of records depending on whether or not one decides to count those ties as new records. In our studies it turned out that the consequences of this kind of discreteness are quite manifold and depend heavily on the three universality classes of extreme value statistics. Similar to our findings for the LDM, the effects of rounding are strong for distributions of the Weibull class and asymptotically negligible for the Fréchet class. Again, the most interesting behavior is found in the Gumbel class. There, we discovered a particularly noteworthy phenomenon in the regime of very strong discreteness, where almost all records are suppressed because of ties. In this regime record events can become almost predictable on an exponential time scale.

2.2 Part II - Record-breaking temperatures

Building up on the analytical results presented in the first part, part II is focused on an important application of the theory of records, namely the statistics of record temperatures in a climate system with a slowly varying temperature mean. The motivation behind this work was to understand the connection between the statistics of record-breaking temperatures and the phenomenon known as global warming. We observed that in media coverage and in the public opinion, new heat records, such as a warmest April on record, or a record breaking heat wave for several calendar days in the summer, are usually blamed on global warming. Without a deeper understanding of the mathematical reason, people assume that an increased mean value leads to new records. With that observation it was our ambition to better understand and quantify the interplay between climatic change and record temperatures. The two articles in this part are the results of our efforts towards this goal.

Chapter 7 — Record-breaking temperatures reveal a warming climate

In the first contribution, a publication in collaboration with Joachim Krug, we study the occurrence of records in hundreds of European [92] and U.S. [93] weather stations that recorded daily minimum and maximum temperatures over several decades in the last century. Comparing the rates of heat and cold records, as well as the record process in forward and backward time direction, we proved a significant effect of global warming on the occurrence and numbers of records especially in the last 30 years of the 20th century. We found that our simple LDM describes the record statistics of daily temperature measurements for individual calendar days quite accurately. In agreement with our LDM, both the increase in the number of heat and the decrease in the number of cold records is proportional to the ratio of the average speed of warming to the standard deviation of daily temperatures. In 2005, at the end of an observation period of 30 years, we found that in Europe, global warming resulted in a 40% increase of heat records compared to our predictions for a stationary climate. The number of cold records was reduced in the same manner. Considering gridded re-analysis data for a small area in central Europe, we could also predict the spatial and seasonal patterns of record-breaking temperatures using our LDM.

Since the fluctuations of the daily temperatures in the U.S. were a lot larger than in Europe, we could not find a similar effect for the American stations. Here, the discreteness effects described in chapter 6 also played an important role. The U.S. station data was measured in units of 1°F and, because of that, a significant amount of the records in the time series were either suppressed or *fabricated* records, depending on whether or not one counts ties. In the context of global warming, it is important to mention that these rounding effects can, in some sense, disguise the effects of global warming on the occurrence of record temperatures (see also chapter 6).

Chapter 8 — Record occurrence and record values in daily and monthly temperatures

These studies are extended in several ways in the subsequent chapter 8. In this publication together with Joachim Krug and Andreas Hense from the University of Bonn, we did not only consider daily minimum and maximum temperatures, but also monthly mean values.

In contrast to the daily measurements, we found a significant effect of the warming on the occurrence of records in time series of monthly temperature averages from hundreds of weather stations in the U.S.. The fact that we used averaged measurements lead to a significantly reduced standard deviation, which in turn, strongly increased the effect of drift on the record rate. We computed the ratios between the upper and lower record rate after an observation period of 60 years ending in 2009. In the last five years of this period we counted almost three times as many heat records as cold records. In the U.S., global warming had stronger effects on record temperatures in winter and spring, there, with values of around 7.6 and 9.3, these ratios were a lot higher in perfect agreement with our LDM.

The second part of chapter 8 uses more recent analytical results for a detailed analysis of record values in a large gridded data set from the European re-analysis project EOBS [94]. We compared our analytical result for the mean value of a record that occurs at a certain time (year) to the observational data. For that purpose, we performed a rescaling of the measurements to make all available time series comparable. Under this rescaling, we found an interesting seasonal dependence of the heat and cold record values. In the summer months our LDM predicts these values correctly. The heat records are, in the context of global warming, more extreme than predicted by a stationary climate. In the same manner, cold records in summer are not only fewer, but also less cold due to global warming.

In the winter months the situation is different. Here, a LDM with Gaussian RV's is not able to predict the measured record values correctly. In fact in winter, cold records are, in some sense, more extreme than hot records, even though their number is smaller. This interesting finding can be explained by a profound asymmetry of the distribution of daily temperatures. In contrast to daily measurements in summer, daily temperatures in winter, especially in the northern and sub polar regions of Europe, are not Gaussian. Their left, lower tail, which generates the cold records, is much broader than their upper tail, which is responsible for most heat records. This effect makes cold records in winter, in some sense, insensitive to global warming.

2.3 Part III - Record statistics of random walks

In contrast to part I, part III is about the statistics of record-breaking events of the entries of random walks. While in the case of i.i.d. RV's or in the case of the LDM discussed in part I, the individual entries $X_1, X_2, ..., X_n$ are independent from each other, random walks are designed as correlated processes. An entry X_k depends on the previous entry X_{k-1} and a random number. In part III, we consider random walks with entries $X_k = X_{k-1} + \xi_k$, where the ξ_k's are simple i.i.d. RV's sampled from a so-called jump distribution $f(\xi)$. A random walk with a jump distribution $f(\xi)$ without a finite variance is also called a Lévy flight[1]. The behavior of Lévy flights differs from the random walk in many aspects, in particular their mean squared-displacement[2] has a different asymptotic dependence on the walk length.

As mentioned in the introduction, the record statistics of such a process with a symmetric jump distribution $f(\xi) = f(-\xi)$, was not computed until a few years ago. In 2008, Majumdar and Ziff [15] were the first who could derive the record rate, the mean record number and even the full distribution of the record number for a symmetric, discrete-time random walk. These results and their derivation were already described in detail in the previous section 1.2. Majumdar and Ziff showed that the record statistics of symmetric random walks is independent from the choice of the jump distribution $f(\xi)$ and the mean record number has a universal half-Gaussian distribution for $n \gg 1$. The first moment of this distribution is the mean record number $\langle R_n \rangle \approx \sqrt{4n/\pi}$ and with this, the record rate

---

[1] Sometimes in the literature random walks are required to have a finite variance, otherwise they are called Lévy flights.

[2] A measure of the distance of the process from its origin.

decays like $P_n \approx 1/\sqrt{n\pi}$. The contributions in part III are basically generalizations of these results for the symmetric random walk.

The first possible generalization is discussed in chapters 9 and 10. These two publications deal with the interesting and surprisingly manifold effects of a constant bias on the record statistics of the random walk. In both chapters, we considered time series $X_1, X_2, ..., X_n$ with entries

$$X_k = X_{k-1} + \xi_k + c, \tag{2.3}$$

where the ξ_k's are sampled from a symmetric and continuous jump distribution $f(\xi)$ and the bias (or drift) c is an arbitrary real constant.

Chapter 9 — Record statistics for biased random walks, with an application to financial data

In chapter 9, which contains joint work with Miro Bogner and Joachim Krug, we compute the record statistics of biased random walks with a Gaussian jump distribution and a positive drift $c > 0$. The record rate and the mean record number for this special case were calculated, using a generalized version of the theorem by Sparre Andersen [17, 18], as used by Majumdar and Ziff [15]. In this work, we consider the regime of a finite series length n and a small bias $c > 0$ and also the asymptotic regime with $n \to \infty$ and a very large positive c. For smaller $c > 0$ and $n \to \infty$, we demonstrate numerically that the record rate approaches an asymptotically constant value bigger than zero. In the last section of this article, we successfully applied our findings in the small c and finite n regime to stock data from the Standard and Poors 500 stock index [68]. We show that, over time spans of many years, the records in the daily stock prices of this index can be modeled by a biased random walk.

Chapter 10 — Record statistics and persistence for a random walk with a drift

The findings presented in chapter 9 are then substantially generalized and refined in chapter 10. This rather long chapter is an article written in collaboration with Satya N. Majumdar and Grégory Schehr from the Université Paris Sud 11 in Orsay and discusses the record statistics of arbitrary biased random walks with a drift $c \neq 0$ that can be both positive or negative. We found that the asymptotic record statistics of biased random walks split into not less than five different universal regimes. For large series lengths $n \to \infty$, the statistical behavior of the record rate and the record number depends on the sign of the bias and the so called Lévy-index of the jump distribution $f(\xi)$, which is a measure of how fast the tails of this distribution decay. We compute and discuss the asymptotic record number distributions, along with the mean record number, the record rate and also the closely related survival probabilities for all five universal regimes. Additionally, we consider the extreme value statistics of the shortest and longest inter-record times, which also have different asymptotic properties in each of these five regimes.

Chapter 11 — Record Statistics for Multiple Random Walks

Chapter 11 is also an article written with Satya N. Majumdar and Grégory Schehr and discusses another generalization of the single, unbiased random walk. In this work, we consider the record statistics of an ensemble of N uncorrelated random walkers. We examine the distribution of the number of records of the maximum of these N random walks and distinguish between random walks with a jump distribution having a finite variance and Lévy flights, where this is not the case. In the first case, we could compute the full distribution of the record number of the process in the limit of large series length $n \to \infty$. For a large number of walkers the distribution of the record number approaches a Gumbel form[3] with a mean value of $\langle R_{n,N} \rangle \approx 2\sqrt{n \ln N}$.

---

[3] For details on this distribution see chapter 11.

In the case of N independent Lévy flights, we can only compute the asymptotic behavior of the record rate, but not the full distribution of the record number. Surprisingly, we find that the record rate of a large number N of independent Lévy flights becomes completely independent from N. Here, the asymptotic record rate approaches a value which is exactly twice the record rate of a single symmetric random walker. Although we can compute this result analytically, we have no sufficient explanation for this behavior. From numerical simulations we further found that the full asymptotic distribution of the record number of N independent Lévy flights is also completely independent from $N \gg 1$ and from the exact shape of the jump distribution. Even though it was not possible to derive an analytical representation of this hitherto unknown universal distribution, we demonstrated that it closely resembles a Weibull distribution.

Similar to the analysis in chapter 9, our findings for the N independent random walks are compared to stock data from the Standard and Poors 500 index at the end of chapter 11. We find that the mean record number of a subset of N randomly selected stock from this index behaves like the mean record number of N^γ independent Gaussian random walkers with $\gamma \approx 0.655$ for upper and $\gamma \approx 0.605$ for lower records. We tentatively explain this behavior with the correlations between the individual stocks; it seems as if only an effective number of N^γ independent stocks contributes to the record statistics.

2.4 Part IV - Records in finance

Chapter 12 — Modeling record-breaking stock prizes

In this part we discuss our second application of the theory of records on observational data in more detail[4]. Some aspects of the statistics of record-breaking events in stock data were already studied in chapters 9 and 11. In this hitherto unpublished contribution, we present a more thorough analysis of the various different statistical properties of records in daily stock data from the Standard and Poors 500 stock index. We introduce several simple stochastic processes that are useful for comparison with stock data. In addition to the results derived in chapters 9, 10 and 11, we introduce a more complicated model of a so-called autoregressive process and analyze its record statistics numerically.

Following these theoretical considerations, we properly introduce the daily stock data from the Standard and Poors 500 stock index. We describe the so-called Geometric Random Walk Model of stock prices, which was already introduced more than one-hundred years ago by Le Bachelier [95] and discuss how well it reproduces the actual stock data. For that purpose, we compute the distribution of the daily returns, i.e. the changes in the stock prices between two successive trading days.

While in chapters 9 and 11 we only considered the record statistics of the stock prices themselves, here we additionally analyze the record process of the daily returns. The occurrence of records in these returns fluctuates strongly and is dominated by only a few periods of very high market activity, such as the recent financial crisis that started in 2008. Furthermore, lower and upper records are highly correlated. On shorter intervals however, the record statistics of the returns are very similar to the corresponding statistics of i.i.d. RV's.

Subsequently, we discuss the record-breaking events in the stock prices themselves. Our findings are similar to the results presented in the chapters 9 and 11, and on shorter time-scales we observe a significant deviation from the analytical predictions of the Geometric Random Walk Model. Therefore, we compare the stock records with the record statistics of an autoregressive AR(1) process that models these records much more accurately. For the first time, we also compute and discuss the full distribution of the record number of stocks from the Standard and Poors 500 index. Because of their importance for the statistics of

---

[4] The first application was our study of temperature records in part II.

record-breaking events in random walks, we study the first-passage statistics of the stock data and compare them with known analytical results.

In order to give a complete summary of our work on stock data, we present some findings for the record statistics of multiple stocks. As in chapter 11, we 'detrend' and rescale the individual stocks to make them comparable and study the statistics of the maximum of ensembles of N stocks. We discuss both the record rate and the mean record number of these ensembles, in the context of the findings in chapter 11.

This contribution is concluded by an interesting observation which is not explained by the simple models of i.i.d. RV's and geometric random walks. We found that the occurrence of return and stock records strongly depends on the weekday on which the stock prices are recorded. In fact, it turns out that a new record on a Monday can be up to twice as likely as a record on a Friday in the same week.

2.5 Part V - Record statistics - A review

Chapter 13 — Record statistics beyond the standard model - Theory and applications

The intention of chapter 13 is to summarize and interpret the recent progress in the field of record statistics. Various models which were studied in the past, are presented and discussed. We briefly recapitulate the findings for the Linear Drift Model and analyze the related Increasing Variance Model of RV's from broadening probability densities. Subsequently we discuss more general record models, such as δ-records and geometric records, which are records that are only counted if they exceed a fixed barrier above or a fixed multiple of the preceding record value. In this context, we also summarize our findings for the record statistics of rounded RV's. Then we present several models of correlated RV's, such as symmetric and biased discrete-time random walks and also continuous-time random walks with random waiting times between the individual entries. Finally, various applications of the theory of records are described. We recapitulate in a few words how to model record-breaking temperatures and record stock prizes and describe the relevance of records and record processes in physics, biology and sports.

Part I

Records in uncorrelated random variables

3 Records and sequences of records from random variables with a linear trend 27
3.1 Introduction. 28
3.2 General theory and an exactly solvable example 30
3.3 Record statistics for small drift 32
  3.3.1 Record rate 32
    3.3.1.1 Weibull class. 34
    3.3.1.2 Fréchet class. 35
    3.3.1.3 Gumbel class. 35
    3.3.1.4 Relation to the asymptotic record rate $p(c)$. 37
  3.3.2 Ordering probability 38
3.4 Record statistics for large drift 39
3.5 Conclusions 41

4 Correlations between record events in sequences of random variables with a linear trend 47
4.1 Introduction. 48
4.2 General theory 49
  4.2.1 Explicit stochastic independence for two cases 50
  4.2.2 Expansion for small drift 51
  4.2.3 Asymptotics for large N. 53
  4.2.4 Dependence on the distance between record events 54
4.3 Explicit examples 54
  4.3.1 The Weibull class 55
  4.3.2 The Gumbel class 56
  4.3.3 The Fréchet class 58
  4.3.4 Unified picture 59
4.4 Conclusion 60

5 Correlations of record events as a test for heavy-tailed distributions 63
5.1 Introduction. 64
5.2 Record statistics and record correlations 64
5.3 Description of the test 66
5.4 Fluctuations. 67
5.5 Application 67
5.6 Summary. 68

6 Rounding effects in record statistics **71**
6.1 Introduction. 72
6.2 General theory, asymptotic results. 73
6.3 Small-Δ regime 74
6.4 Large-Δ regime 75
6.5 Conclusions 76

Chapter 3

Records and sequences of records from random variables with a linear trend

Jasper Franke, Gregor Wergen and Joachim Krug

Institute for Theoretical Physics, University of Cologne

Published in: Journal of Statistical Mechanics, P10013 (2010)

Abstract: We consider records and sequences of records drawn from discrete time series of the form $X_n = Y_n + cn$, where the Y_n are independent and identically distributed random variables and c is a constant drift. For very small and very large drift velocities, we investigate the asymptotic behavior of the probability $p_n(c)$ of a record occurring in the nth step and the probability $P_N(c)$ that all N entries are records, i.e. that $X_1 < X_2 < ... < X_N$. Our work is motivated by the analysis of temperature time series in climatology, and by the study of mutational pathways in evolutionary biology.

3.1 Introduction

A record is an entry in a discrete time series that is larger (*upper record*) or smaller (*lower record*) than all previous entries. In this sense, a record is an extreme value that is defined relative to all previous values in the time series. Record events are of interest in various areas of life and science such as climatology [1–5] and sports [6, 7], but also in biology [8–10]. A record is usually a rare and remarkable event that will be remembered by observers. Not without good reason the term record originates from the Latin verb *recordari - to recall, to remind.*

The classic results for records drawn from series of independent and identically distributed (i.i.d.) random variables (RV's) are well established, see [11–14] for review. In this work we concentrate on two important quantities in particular. The first one is the probability for a certain entry in a time series to be a record, and the second one is the probability that the entries of a time series are ordered, or in other words, that *all* events are records. For i.i.d. RV's both these quantities are completely universal for all continuous probability density functions. This can be shown by the so called *stick-shuffling* argument: The last one of n identically distributed entries (sticks) in a time series is equally likely to be a record as all other entries, and therefore the probability p_n for the nth event to be a record, henceforth referred to as the *record rate*, is given by

$$p_n = \frac{1}{n}. \tag{3.1}$$

Accordingly the expected mean number of records R_n up to a time n can be obtained by computing the harmonic sum: $R_n = \sum_{i=1}^{n} 1/k \approx \ln(n) + \gamma + O(1/n)$, where $\gamma \approx 0.577215...$ is the Euler-Mascheroni constant. From similar considerations one obtains the statistics of waiting times between record breaking events which turn out to be universal as well. It is equally straightforward to compute the probability for all events in a series of length N to be ordered in size. Since this case is only one of $N!$ possible and equally likely permutations of all N events, the *ordering probability* P_N is given by

$$P_N = \frac{1}{N!}. \tag{3.2}$$

We conclude that the two quantities of interest are related by

$$P_N = \prod_{n=1}^{N} p_n, \tag{3.3}$$

which reflects the fact that record events are independent in the i.i.d. case [11, 13]. We will return to this point below in section 2. In contrast to the properties of record times, the distributions of record values are not completely universal, but their asymptotic behavior falls into three different universality classes that are analogous to the universality classes of extreme value statistics: The *Weibull* class of distributions with finite support, the *Gumbel* class of distributions with exponential-like tails, and the *Fréchet* class of power law tailed distributions [15–17].

Given that the statistics of records for i.i.d. RV's is well understood, it is natural to ask what happens when the underlying time series is correlated, or when the RV's are drawn from a distribution that varies in time. An important example of a correlated random process is the random walk, and the record statistics of this process was recently analyzed by Majumdar and Ziff [18, 19]. The simplest realization of a time-dependent distribution is the *linear drift model* (LDM) first considered by Ballerini and Resnick [20, 21]. In this model the nth entry in the time series is of the form

$$X_n = Y_n + cn, \tag{3.4}$$

where c is a constant and the Y_n are i.i.d. RV's. In this simple scenario the probability density $f_n(x)$ of X_n is of the form $f_n(x) = f(x - cn)$ with a fixed probability density $f(y)$ and the corresponding cumulative distribution function $F(y) = \int_{-\infty}^{y} dy' \, f(y')$, which is the distribution of the i.i.d. part Y_n of X_n. We will usually consider upper records and assume $c > 0$.

The LDM was originally introduced as a model for sports records in improving populations [20], and it has recently appeared in the context of the dynamics of elastic manifolds in random media [22]. An important motivation for the present work comes from the interest in the consequences of global warming for the occurrence of temperature records [3, 4]. In [5, 23] the effect of warming on daily temperature measurements was modeled using a Gaussian probability density with a linear trend, and it was shown that this very simple model is capable of quantitatively describing the statistics of record-breaking temperatures at European and American weather stations. In the climate context the drift speed c is typically small compared to the standard deviation of f, which suggests to consider the behavior of the record rate $p_n(c)$ for small c and finite n. This approach is complementary to previous work on the LDM [20–22, 24], which has mostly been concerned with the asymptotic behavior of the record rate for $n \to \infty$.

Another application that motivates our research comes from the study of adaptive paths in evolutionary biology. In this context, a path is a collection of mutations that change the genotype of an organism into another genotype of higher fitness. Given that mutation rates are small, the evolution of a population usually proceeds one mutation at a time. For a given set of N mutations, there are then $N!$ distinct paths which correspond to the different orders in which the mutations can occur. Since a mutation spreads in the population only if it confers a fitness advantage, a given pathway is *accessible* to adaptive evolution only if the fitness values of the intermediate genotypes increase monotonically along the path, that is, if they are arranged in ascending order [25, 26].

In view of the complexity of real fitness landscapes, the intricate interactions between different mutations are often modeled by assigning fitness values at random to genotypes [10]. One such model, which is closely related to the LDM, was introduced by Aita *et al.* in the context of protein evolution [27]. In this model the fitness X_n of a particular intermediate genotype with n mutations is assumed to consist of an i.i.d. RV Y_n and a systematic part cn, where $c > 0$ if the mutations move the population closer to the global fitness peak, and the value of c (relative to the standard deviation of the Y_n) can be adjusted to tune the ruggedness of the fitness landscape. Taking into account also the initial genotype with no mutations, a total of $N + 1$ genotypes with fitness values $X_0, X_1, ..., X_N$ are encountered along a path. The probability for a path to be accessible in this model is then just $P_{N+1}(c)$, and the expected number of accessible paths is $N!P_{N+1}(c)$. An immediate corollary of (3.2) is that the expected number of accessible paths of length N in a completely random fitness landscape without any average uphill slope ($c = 0$) is $1/(N + 1)$ [28, 29].

Here we consider both the record rate p_n and the ordering probability P_N for the linear drift model. We distinguish between a small drift c that is much smaller than the characteristic width of the distribution (in most cases the standard deviation), and a large drift that is much larger than this width. Both cases are of practical relevance. In section 2 we discuss the general properties of record statistics for systems with linear drift, with particular emphasis on the correlations between record events. In the subsequent section 3 we will present new results for small c. We examine the record statistics for members of the three extreme-value classes individually and find the corresponding asymptotic behaviors. In section 4 we analyze the case of large c. Throughout Monte-Carlo simulations are used to confirm the analytical results. Finally, in section 5 we present a brief summary, discuss related issues and give an outlook on further possible research. Some of the calculational details are relegated to Appendices.

3.2 General theory and an exactly solvable example

The values taken by the $\{X_i\}_{i\in\{1,\dots,n\}}$ are stochastically independent. The probability that all n values are less than a given value x factorizes to

$$\prod_{i=1}^{n}\int_{-\infty}^{x}\mathrm{d}x_i f_i\left(x_i\right)=\prod_{i=1}^{n}F_i(x). \tag{3.5}$$

Here f_i and F_i are, as stated in the introduction, the probability densities and cumulative distribution functions of the X_i. Thus, given the value y_n of the i.i.d. part Y_n of X_n, the probability that all previous RV's $\{X_i\}_{i\in\{1,\dots,n-1\}}$ are smaller than X_n is $\prod_{i=1}^{n-1}\int_{-\infty}^{y_n+ic}\mathrm{d}y_{n-i}f(y_{n-i})$. The probability that a RV X_n drawn from a general time-dependent distribution $F_n(x)$ is a record is therefore given by [30]

$$\mathrm{P}\left[X_n=\max_{i\in\{1,\dots,n\}}\{X_i\}\right]=p_n=\int_{-\infty}^{\infty}\mathrm{d}xf_n\left(x\right)\prod_{i=1}^{n-1}F_i\left(x\right), \tag{3.6}$$

which reduces to

$$p_n(c)=\int_{-\infty}^{\infty}dxf(x)\prod_{i=1}^{n-1}F(x+ci) \tag{3.7}$$

for the LDM. It was shown in [20] that the limiting record rate

$$p(c)\equiv\lim_{n\to\infty}p_n(c)=\int_{-\infty}^{\infty}dxf(x)\prod_{i=1}^{\infty}F(x+ci) \tag{3.8}$$

exists and is nonzero for $c>0$ provided the distribution $f(y)$ of the i.i.d. part in (3.4) has a finite first moment. For $c=0$, (3.7) can be evaluated directly and, with the substitution $u=F(x)$, one obtains

$$p_n(c=0)=\int_{-\infty}^{\infty}\mathrm{d}xf(x)F(x)^{n-1}=\int_{F(-\infty)=0}^{F(\infty)=1}\mathrm{d}uu^{n-1}=\frac{1}{n}, \tag{3.9}$$

independent of F, as already shown in (3.1).

The other quantity under consideration in this article, the ordering probability P_N, can be expressed as

$$\begin{aligned}\mathrm{P}\left[X_1<X_2<\cdots<X_N\right] &= P_N(c)\\ &= \int_{-\infty}^{\infty}\mathrm{d}x_Nf_N(x_N)\dots\int_{-\infty}^{\infty}\mathrm{d}x_1f_1(x_1)\mathbb{1}_{x_1<x_2\cdots<x_N}.\end{aligned}$$

Inserting $f_n(x)=f(x-cn)$ the indicator function $\mathbb{1}_{x_1<x_2<\cdots<x_N}$ can be absorbed in the integral boundaries to yield

$$P_N(c)=\int_{-\infty}^{\infty}\mathrm{d}y_Nf(y_N)\int_{-\infty}^{y_N+c}\mathrm{d}y_{N-1}\dots\int_{-\infty}^{y_2+c}\mathrm{d}y_1f(y_1). \tag{3.10}$$

As for $p_n\left(c\right)$, this equation can be solved for arbitrary F only in the case $c=0$. Using again the substitution $u=F(x)$ in turn in all the N integrals, starting from the inside, one obtains the result already derived in (3.2),

$$
\begin{aligned}
P_N(c=0) &= \int\limits_{-\infty}^{\infty} \mathrm{d}y_N f(y_N) \int\limits_{-\infty}^{y_N} \mathrm{d}y_{N-1} \dots \int\limits_{-\infty}^{y_3} \mathrm{d}y_2 f(y_2) F(y_2) \\
&= \int\limits_{-\infty}^{\infty} \mathrm{d}y_N f(y_N) \int\limits_{-\infty}^{y} \mathrm{d}y_{N-1} \dots \int\limits_{-\infty}^{F(y_3)} \mathrm{d}u u \\
&= \frac{1}{2} \int\limits_{-\infty}^{\infty} \mathrm{d}y_N f(y_N) \int\limits_{-\infty}^{y_N} \mathrm{d}y_{N-1} \dots \int\limits_{-\infty}^{F(y_4)} \mathrm{d}u u^2 \\
&= \dots = \frac{1}{N!}. \qquad (3.11)
\end{aligned}
$$

The reason for re-deriving the two previous results is that here this is done in a way that in principle generalizes to arbitrary c.

For $c > 0$, the exact evaluation of Eqs.(3.7) and (3.10) has proven difficult, but in the case where the Y_n are Gumbel distributed, i.e. $F(y) = \exp\left(-e^{-y}\right)$, one can use the fact that this distribution obeys the relation $F(y+a) = F(y)^{\exp(-a)}$ to explicitly perform the integration in (3.7) [20, 21]. With the abbreviation $\alpha \equiv e^{-c}$ and the substitution $u = F(y)$ one obtains

$$
p_n(c) = \int\limits_{-\infty}^{\infty} \mathrm{d}y f(y) F(y)^{\sum_{i=1}^{n-1} \alpha^i} = \left(\sum_{i=0}^{n-1} \alpha^i \right)^{-1} = \frac{1 - e^{-c}}{1 - e^{-nc}} \qquad (3.12)
$$

by use of the incomplete geometric series. Keeping c fixed, one obtains in the limit $n \to \infty$ the asymptotic record rate $p(c) = 1 - e^{-c}$, while for $c \to 0$ one recovers the i.i.d. result $p_n = 1/n$. For $c < 0$ the record rate is seen to decay exponentially in n, which implies that the expected number of records R_n remains finite for $n \to \infty$. We suspect this to be a general feature of the LDM with $c < 0$, but are not aware of a proof of this fact.

The relation used to evaluate (3.12) for the Gumbel case can also be used in (3.10), as before starting from the innermost integral, which yields

$$
\begin{aligned}
P_N(c) &= \textstyle\int_{-\infty}^{\infty} \mathrm{d}y_N f(y_N) \int_{-\infty}^{y_N+c} \mathrm{d}y_{N-1} \dots \int_{-\infty}^{y_3+c} \mathrm{d}y_2 f(y_2) F(y_2+c) \\
&= \textstyle\int_{-\infty}^{\infty} \mathrm{d}y_N f(y_N) \int_{-\infty}^{y_N+c} \mathrm{d}y_{N-1} \dots \int_{-\infty}^{F(x_3+c)} \mathrm{d}u u^{\alpha} \\
&= \textstyle\frac{1}{\alpha+1} \int_{-\infty}^{\infty} \mathrm{d}y_N f(y_N) \int_{-\infty}^{x_N+c} \mathrm{d}y_{N-1} \dots \int_{-\infty}^{F(x_4+c)} \mathrm{d}u u^{\alpha(\alpha+1)} \\
&= \textstyle\dots = \prod_{l=1}^{N-1} \frac{1}{\sum_{k=0}^{l} \alpha^k}. \qquad (3.13)
\end{aligned}
$$

Summing the geometric series as in (3.12), one obtains

$$
P_N(c) = \left(1 - e^{-c}\right)^N \frac{1}{\prod_{n=1}^{N} \left(1 - e^{-cn}\right)} \equiv \left(1 - e^{-c}\right)^N \mathcal{Z}_N, \qquad (3.14)
$$

where $\mathcal{Z}_N$ is the grand canonical partition function of a system of bosonic particles with energy levels $n = 1, ..., N$ at inverse temperature c. This partition function also occurs as one limit in the integer partition problem (see [31, 32] and references therein).

The product $\prod_{n=1}^{N} \left(1 - \exp(-cn)\right)$ in the denominator is the so-called q-Pochhammer symbol $(q;q)_N$ with $q = e^{-c}$. In the limit $N \to \infty$ with fixed c, one has the asymptotic expression [33]

$$
\lim_{N \to \infty} \prod_{n=1}^{N} \left(1 - e^{-cn}\right) \equiv \left(e^{-c}\right)_\infty \approx \sqrt{\frac{2\pi}{c}} \exp\left(-\frac{\pi}{6c} + \frac{c}{24}\right), \qquad (3.15)
$$

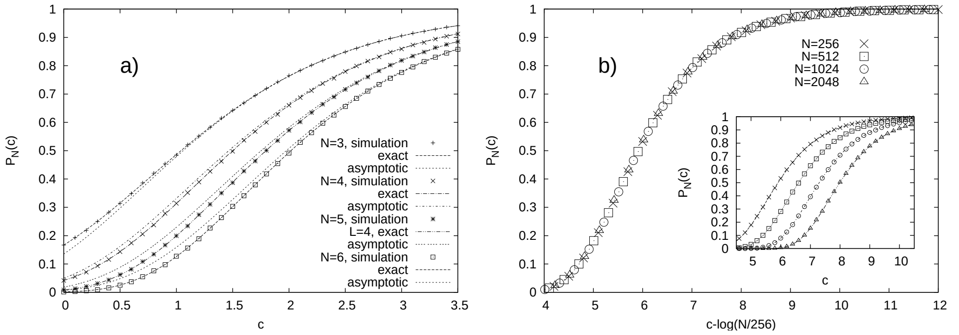

Figure 3.1: Comparison of the exact expression (3.13) and the asymptotic expression (3.17) to numerical simulation. The exact expression is confirmed, and while there is a clear difference between simulations and asymptotic expression for small values of c in **a)**, the approximation holds with good accuracy for large c (inset of **b)**, lines are the asymptotic expressions). The main plot of **b)** demonstrates the scaling between N and c according to (3.17).

and thus, by inserting this into (3.14),

$$P_N(c) \approx \sqrt{\frac{c}{2\pi}} \exp\left(N \ln(1-e^{-c}) + \frac{\pi}{6c} - \frac{c}{24}\right), \quad N \gg 1. \tag{3.16}$$

On the other hand, taking $c \gg 1$ at fixed N, one has $\alpha = \exp(-c) \ll 1$ and thus the geometric series in the denominator of (3.13) can be approximated to first order in $\alpha \equiv \exp(-c)$, as $1/(\sum_{k=0}^{l} \alpha^k) \approx 1 - \alpha + \mathcal{O}(\alpha^2)$. Then (3.13) becomes

$$P_N(c) \approx \exp\left(-(N-1)\alpha\right) = \exp\left(-(N-1)e^{-c}\right), \quad c \gg 1. \tag{3.17}$$

This expression is distinguishable from numerical data only in the region of $c \sim \mathcal{O}(1)$, see figure 3.1.

Comparing the exact expressions Eqs.(3.13) and (3.12), one sees that the relation (3.3) obtained in the i.i.d. case remains valid here. This is a consequence of the mutual stochastic independence of record events in the LDM with Gumbel-distributed i.i.d. part [12, 14, 24]. In fact the Gumbel distribution is uniquely characterized by the mutual independence of record values and record indicator variables (which indicate whether or not a record occurs at time n) [24, 34].

For $c > 0$ and arbitrary distribution F, however, the record events in the LDM are not independent. Numerical studies for several different distributions presented in figure 3.2 show that the records are negatively correlated and seem to repel each other. A more thorough examination of the structure of correlations between record events in this model is currently ongoing research [35]. For the purpose of the present discussion we merely note that record events appear to become asymptotically uncorrelated for large c. This fact will be used to derive some asymptotic results for $P_N(c)$ in section 3.4. First however we consider the case $c \ll 1$.

3.3 Record statistics for small drift

3.3.1 Record rate

In the previous section we gave a general expression for the record rate $p_n\,(c)$ of the LDM. Here, we derive the first order term in a series expansion for $c \ll 1$. If c is very small (3.7)

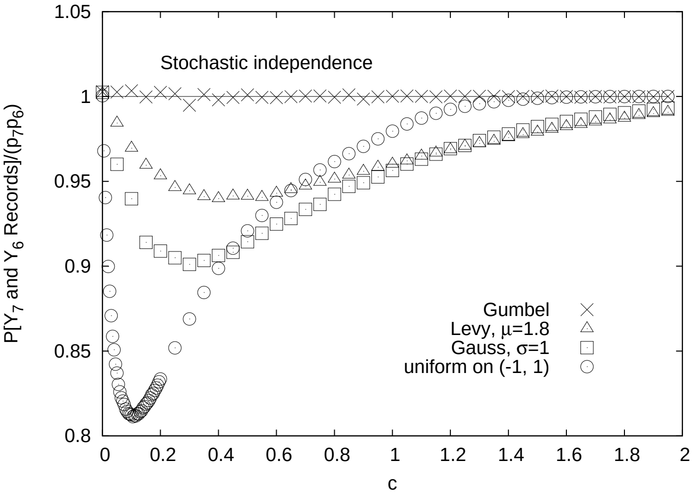

Figure 3.2: Joint probability of two consecutive record events at times $n = 6$ and 7, divided by the product of the corresponding record rates. This ratio is unity if record events are stochastically independent. For $c = 0$, this is the case, just as for Gumbel-distributed i.i.d. parts (crosses). Note that for other probability densities, the record events also seem to become increasingly independent as c grows.

can be simplified as follows:

$$\begin{aligned} p_n(c) &= \int_{-\infty}^{\infty} \mathrm{d}y f(y) \prod_{i=1}^{n-1} F(y+ci) \\ &\approx \int_{-\infty}^{\infty} \mathrm{d}y f(y) \prod_{i=1}^{n-1} \left[F(y) + cif(y)\right] \\ &\approx \int_{-\infty}^{\infty} \mathrm{d}y f(y) F^{n-1}(y) + c\frac{n(n-1)}{2} \int_{-\infty}^{\infty} \mathrm{d}y f^2(y) F^{n-2}(y) \\ &= \frac{1}{n} + cI_n \end{aligned} \tag{3.18}$$

with

$$I_n \equiv \frac{n(n-1)}{2} \int_{-\infty}^{\infty} \mathrm{d}y f^2(y) F^{n-2}(y). \tag{3.19}$$

This expansion is valid provided $f(y)$ is slowly varying between y and $y + ci$, which strictly speaking requires nc to be small compared to the width of the distribution. In the following we will evaluate the first order correction coefficient I_n for several elementary distributions.

Before doing this, we show that our formula for $p_n(c)$ can be generalized with respect to the position of the record in the time-series. Specifically, we consider the probability that the kth event in a time-series of length n with linear drift c is a record. For this purpose we have to consider the following integral instead of (3.6):

$$\mathrm{P}[X_k = \max(X_1, ..., X_n)] = \int_{-\infty}^{\infty} \mathrm{d}y f_n(y) \prod_{i=1, i\neq k}^{n-1} \int_{-\infty}^{y+c(i-k)} \mathrm{d}y_i f_i(y_i). \tag{3.20}$$

Evaluating this integral in the same way as shown above, we obtain the following expression:

$$\begin{aligned} \mathrm{P}[X_k = \max(X_1, ..., X_n)] \quad &\approx \quad \frac{1}{n} + \frac{c}{2}\left(k^2 - k - (n-k)(n-k-1)\right) \times \\ &\times \int\limits_{-\infty}^{\infty} \mathrm{d}y f^2(y) F^{n-2}(y) . \end{aligned} \tag{3.21}$$

Note that for $k = n$ this expression reduces to our approximation (3.18) for $p_n(c)$. Apparently for $c > 0$ this expression assumes its maximum for $k = n$ and its minimum for $k = 1$. The last entry has the largest, and the first entry the smallest chance to be the maximum of the series.

3.3.1.1 Weibull class.

Let us start by considering the Weibull class of extreme value statistics, which contains distributions with finite support. A simple example for a member of the Weibull class is a uniform distribution, which takes the value $\frac{1}{2a}$ between $-a$ and a and 0 outside of this interval. For this case the first order expansion of $p_n(c)$ is given by

$$p_n^{uniform}(c) = \frac{1}{n} + c\frac{n(n-1)}{2} \int\limits_{-\infty}^{\infty} \mathrm{d}y \left(\frac{1}{2a}\right)^2 \left(\frac{y}{2a} + \frac{1}{2}\right)^{n-1} + O\left(c^2\right), \tag{3.22}$$

which can be evaluated to yield

$$p_n^{uniform}(c) \approx \frac{1}{n} + c\frac{n-1}{4a}. \tag{3.23}$$

In this case the correction coefficient I_n increases linearly with the number of events n.

More generally, we consider distributions of the form

$$f(y) = \xi (1-y)^{\xi - 1} \tag{3.24}$$

with $\xi > 0$ and $0 < y \leq 1$. For these distributions we have

$$p_n(c) \approx \frac{1}{n} + c\frac{n(n-1)}{2} \int\limits_{0}^{1} \mathrm{d}y \xi^2 (1-y)^{2\xi - 2} \left(1 - (1-y)^{\xi}\right)^{n-2} . \tag{3.25}$$

The integral is divergent for $\xi < 1/2$, which indicates that $p_n(c)$ is a non-analytic function of c; this case will be considered elsewhere. For $\xi > 1/2$ we use the substitution $(1-y) = z^{1/\xi}$ to express the integral in terms of a Beta-function,

$$p_n(c) \approx \frac{1}{n} + c\xi \frac{n(n-1)}{2} \frac{\Gamma\left(2 - \frac{1}{\xi}\right)\Gamma(n-1)}{\Gamma\left(n + 1 - \frac{1}{\xi}\right)}. \tag{3.26}$$

Using the Stirling approximation for large n one finally arrives at

$$p_n(c) \approx \frac{1}{n} + \frac{c\xi}{2}\Gamma\left(2 - \frac{1}{\xi}\right) n^{\frac{1}{\xi}}, \tag{3.27}$$

which shows that I_n generally increases as a power law in the Weibull class.

3.3.1.2 Fréchet class.

As a representative of the Fréchet class of extreme value statistics we consider a general power-law distribution of the form $f(x) = (1/\mu)\, x^{-\mu-1}$ for $x > 1$ and $\mu > 0$. For distributions of this kind $p_n(c)$ in the small c expansion is given by

$$p_n(c) \approx \frac{1}{n} + c\frac{n(n-1)}{2} \int_1^\infty \mathrm{d}y \mu^2 y^{-2-2\mu} \left(1 - y^{-\mu}\right)^{n-2}. \tag{3.28}$$

Again, the integral is very similar to a Beta-function and it can be transformed into one by elementary means. Doing this we find

$$p_n(c) \approx \frac{1}{n} + c\mu\frac{n(n-1)}{2} \frac{\Gamma\left(2+\frac{1}{\mu}\right)\Gamma(n-1)}{\Gamma\left(n+\frac{1}{\mu}+1\right)}, \tag{3.29}$$

and using again the Stirling approximation we obtain

$$p_n(c) \approx \frac{1}{n} + \frac{c\mu}{2}\Gamma\left(2+\frac{1}{\mu}\right)\frac{1}{n^{1/\mu}}. \tag{3.30}$$

In figure 3.3 we compare this prediction to simulation results.

While in the case of the Weibull class the correction term I_n increases with n, here it decays as a power-law $n^{-1/\mu}$. For $\mu > 1$ the decay is slower than the $1/n$-decay of the record rate in the absence of a drift, which implies that the drift will nevertheless dominate the behavior for long times. This is consistent with the fact that the record rate reaches a nonzero asymptotic limit, as given by (3.8), because $\mu > 1$ implies a finite first moment for the Y_n. On the other hand, for $\mu < 1$ the decay of I_n is faster $1/n$ and the limit on the right hand side of (3.8) vanishes for any c, which implies that the drift is asymptotically irrelevant. The borderline situation $\mu = 1$ has been studied by De Haan and Verkade [36], who find that the asymptotics depends nontrivially on the value of c in this case.

In general the results presented so far show that the effect of the drift on a broad distribution is smaller than on a more narrow distribution. A similar qualitative trend was found in [30] for probability densities with increasing variance.

3.3.1.3 Gumbel class.

The Gumbel class comprises unbounded distributions that decay faster than any power-law. A very simple representative of the Gumbel class is the exponential distribution $f(y) = \nu^{-1} e^{-\frac{y}{\nu}}$. In this case the first order expansion (3.18) assumes the following form:

$$p_n^{exp}(c) \approx \frac{1}{n} + c\frac{n(n-1)}{2} \int_0^\infty \mathrm{d}y \frac{1}{\nu^2} e^{-\frac{2y}{\nu}} \left(1 - e^{-\frac{y}{\nu}}\right)^{n-2}. \tag{3.31}$$

The integral can be solved by two partial integrations and one finds

$$p_n^{exp}(c) \approx \frac{1}{n} + \frac{c}{2\nu}, \tag{3.32}$$

that is, the correction term is independent of n.

The calculation for the Gaussian distribution, arguably the most important member of the Gumbel class, is more complicated. For convenience we consider a Gaussian distribution of unit variance,

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}. \tag{3.33}$$

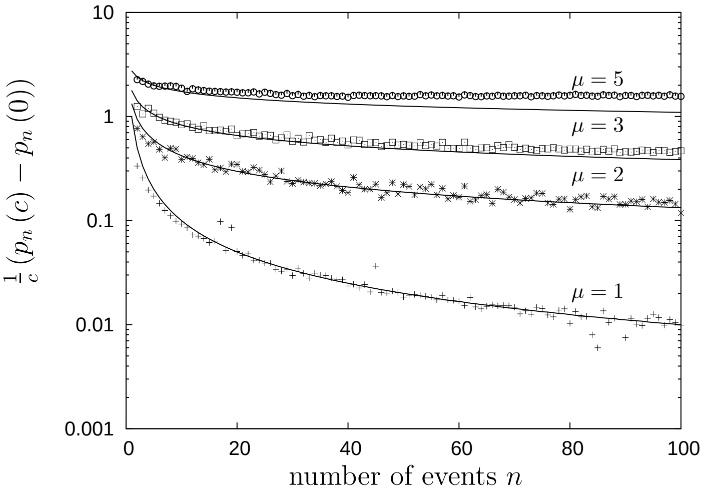

Figure 3.3: Results of Monte-Carlo simulations of the LDM for power law tailed distributions of the Fréchet class. The figure shows the difference between the record rate in the time-independent case for $c = 0$ and the drifting case with drift $c = 0.01$. This difference is given by $\frac{1}{c}\left(p_n\left(c\right)-p_n\left(0\right)\right)$. The dots correspond to simulations with different tail coefficients $\mu = 1, 2, 3, 5$ averaged over 10^6 runs, and the lines show the analytic predictions. The first order approximation is very good for $\mu = 1$ and $\mu = 2$ but becomes less accurate for larger μ.

The integral of interest reads

$$I_n^{gauss} = \frac{n\left(n-1\right)}{2\sqrt{2\pi}^n}\int\limits_{-\infty}^{\infty}\mathrm{d}y e^{-y^2}\left(\int\limits_{-\infty}^{y}\mathrm{d}y' e^{-\frac{-y'^2}{2}}\right)^{n-2}, \tag{3.34}$$

which will be evaluated for large n using the saddle point approximation. With the definition

$$g\left(y\right) := -y^2 + \left(n-2\right)\ln\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{y}\mathrm{d}y' e^{-\frac{y'^2}{2}}\right) \tag{3.35}$$

we have

$$I_n^{gauss} \approx \frac{n\left(n-1\right)}{4\pi}\sqrt{\frac{-2\pi}{\mathrm{d}_y^2 g\left(\tilde{y}\right)}}e^{g(\tilde{y})}, \tag{3.36}$$

where $\tilde{y}$ denotes the saddle-point of the integral. It turns out that the computation of a practicable series-expansion of $g\left(y\right)$ can only be done under some approximations and by using the non-elementary Lambert-W function [37, 38]. In terms of the W-function $\mathrm{W}\left(z\right)$ defined by the relation $\mathrm{W}(z)e^{\mathrm{W}(z)} = z$, we find

$$\tilde{y} = \sqrt{\mathrm{W}\left(\frac{\left(n-2\right)^2}{8\pi}\right)}. \tag{3.37}$$

For large z the Lambert-W function can be approximated by $\mathrm{W}\left(z\right) \approx \ln\left(z\right) - \ln\left(\ln\left(z\right)\right)$, which eventually yields

$$p_n^{gauss}\left(c\right) \approx \frac{1}{n} + c\frac{2\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)}. \tag{3.38}$$

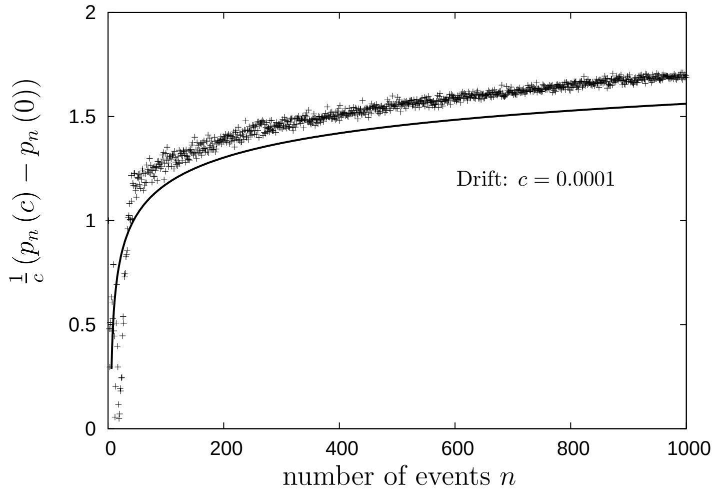

Figure 3.4: Results of Monte-Carlo simulations from 10^9 realizations of the LDM with RV's drawn from a normal distribution with standard deviation $\sigma = 1$. The figure shows the normalized difference between the record rate in the time-independent case and the drifting case, $\frac{1}{c}\left(p_n\left(c\right)-p_n\left(0\right)\right)$. The dots correspond to a simulation with drift velocity $c = 10^{-4}$.

For a detailed derivation of this result see **APPENDIX I**. In figure 3.4 the asymptotic prediction is compared to numerical simulations. The systematic deviations that are visible in this figure can be attributed to strong sub-leading corrections to (3.38), see **APPENDIX I**.

As a more general subset of the Gumbel class we also considered distributions of the form $f\left(y\right) = C_\beta e^{-|y|^\beta}$ with $\beta > 0$ and normalization constant $C_\beta = [2\Gamma(1+1/\beta]^{-1}$. The integral of interest then reads

$$I_n = \frac{n\left(n-1\right)}{2} C_\beta^n \int\limits_{-\infty}^{\infty} \mathrm{d}y e^{-2|y|^\beta} \left(\int\limits_{-\infty}^{y} \mathrm{d}y' e^{-|y'|^\beta} \right)^{n-2}, \tag{3.39}$$

which can again be treated using a saddle-point approximation. Ignoring constant prefactors we find that

$$I_n \propto \ln\left(n\right)^{1-\frac{1}{\beta}} \tag{3.40}$$

for large n, which includes the results for the exponential distribution ($\beta = 1$) and the Gaussian ($\beta = 2$) as special cases. For a detailed derivation of this result see **APPENDIX II**. We conclude that the behavior of the correction coefficient I_n in the Gumbel class is generally intermediate between the power law growth for distributions in the Weibull class, and the power law decay for Fréchet-type distributions. Again, the effect of the drift is stronger for distributions that fall off more rapidly (large β).

3.3.1.4 Relation to the asymptotic record rate $p(c)$.

It is instructive to compare the asymptotics of the correction term I_n derived in the preceding subsections to the behavior of the limiting record rate $p(c)$ for small c, which was studied by Le Doussal and Wiese [22]. Heuristically, the two quantities can be related as follows. We have seen above that, for any choice of $f(y)$ with a finite first moment, the correction term I_n becomes large compared to $1/n$ for large n. This implies that, for any $c > 0$, the first order correction will eventually become comparable to the zero'th order record rate $1/n$. The corresponding time scale n^* can be estimated from

$$n^* I_{n^*} \sim c. \tag{3.41}$$

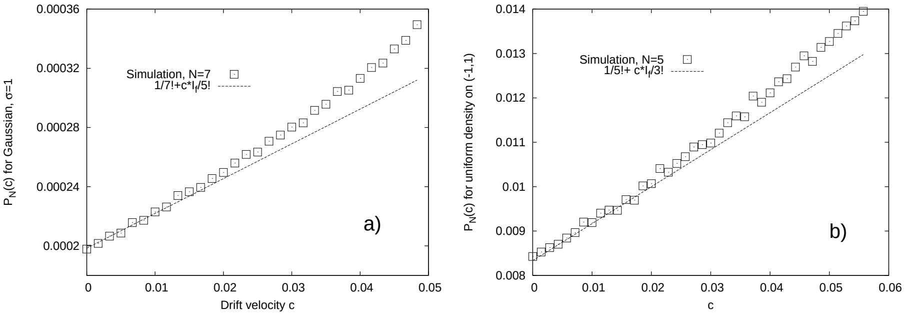

Figure 3.5: Simulations comparing the ordering probability $P_N(c)$ to the first order expansion $P_N(c) = 1/N! + cI_f/(N-2)!$, where $I_f = \int dy\ f(y)^2$, for **a)** Gaussian distribution and $N = 7$, **b)** uniform distribution and $N = 5$.

For times $n > n^*$ the first-order expansion breaks down and the record rate saturates at a nonzero limiting value $p(c)$. Thus we expect that, in order of magnitude, $p(c) \sim 1/n^*(c)$. Using the asymptotic results (3.27,3.30,3.40) together with (3.41) we may then determine the behavior of $p(c)$ for small c. The result

$$p(c) \sim \begin{cases} c^{\xi/(1+\xi)} & \text{Weibull} \\ c^{\mu/(\mu-1)} & \text{Fréchet with } \mu > 1 \\ c|\ln c|^{1-1/\beta} & \text{Gumbel} \end{cases} \tag{3.42}$$

agrees with the analysis of [22] in all cases.

3.3.2 Ordering probability

In this subsection, we derive a first order expansion for the ordering probability $P_N(c)$. Our main result reads

$$P_N(c) = \frac{1}{N!} + c\frac{1}{(N-2)!}\int\limits_{-\infty}^{\infty} \mathrm{d}x f^2(x) + \mathcal{O}(c^2). \tag{3.43}$$

In contrast to the expansion (3.18) for the record rate, one sees that for $P_N(c)$ the distribution $f(x)$ only enters in the form of a non-universal constant but has no influence on the N-dependence of the correction term. Note, however, that similar to the expansion for $p_n(c)$, the correction term diverges when $f^2(x)$ becomes too singular, as is the case for the Weibull-type distribution (3.24) with $\xi < 1/2$.

To prove (3.43), we set up a Taylor expansion of (3.10) in c to first order. With $P_N(0) = 1/N!$ we have

$$\begin{aligned} P_N(c) &= \frac{1}{N!} + c\,\frac{\mathrm{d}}{\mathrm{d}c}P_N(c)\Big|_{c=0} + \mathcal{O}(c^2) \\ &\approx \frac{1}{N!} + \int\limits_{-\infty}^{\infty} \mathrm{d}x_N f(x_N)\,\frac{\mathrm{d}}{\mathrm{d}c}P(N-1,c,x_N)\Big|_{c=0}, \end{aligned}$$

where terms of $\mathcal{O}(c^2)$ and higher have been omitted and

$$\begin{aligned} P(N-1,c,x_N) &\equiv \int_{-\infty}^{x_N+c} \mathrm{d}y_{N-1} f(y_{N-1}) \int_{-\infty}^{y_{N-1}+c} \mathrm{d}y_{N-2} \ldots \int_{-\infty}^{y_2+c} \mathrm{d}y_1 f(y_1) \\ &= \int_{-\infty}^{x_N+c} \mathrm{d}y_{N-1} f(y_{N-1}) P(N-2,c,y_{N-1}). \end{aligned}$$

Clearly, the derivative of $P(N-1,c,x_N)$ obeys the recursion relation

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}c}P(N-1,c,x_N)\right|_{c=0} &= f(x_N)P(N-2,0,x_N)\\ &+\int\limits_{-\infty}^{x_N+c}\mathrm{d}y_{N-1}f(y_{N-1})\left.\frac{\mathrm{d}}{\mathrm{d}c}P(N-2,c,y_{N-1})\right|_{c=0}.\end{aligned}$$

Using the same substitutions as in (3.11), one obtains

$$P(N-2,c=0,x_N)=\frac{1}{(N-2)!}F^{N-2}(x_N)$$

and thus

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}c}P(N-1,c,x_N)\right|_{c=0} &= \frac{f(x_N)}{(N-2)!}F^{N-2}(x_N)\\ &+\int\limits_{-\infty}^{x_N+c}\mathrm{d}y_{N-1}f(y_{N-1})\left.\frac{\mathrm{d}}{\mathrm{d}c}P(N-2,c,y_{N-1})\right|_{c=0}.\end{aligned}$$

Now $P(1,c,x_2)=\int_{-\infty}^{x_2+c}\mathrm{d}y_1 f(y_1)$ and thus $\frac{\mathrm{d}}{\mathrm{d}c}P(1,c,x_2)\big|_{c=0}=f(x_2)$. Putting this into the recursion relation above and integrating over all y_N weighted by $f(y_N)$, we obtain

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}c}P_N(c)\right|_{c=0} &= \frac{1}{(N-2)!}\int\limits_{-\infty}^{\infty}\mathrm{d}y_N f^2(y_N)F^{N-2}(y_N)\\ &+ \frac{1}{(N-3)!}\int\limits_{-\infty}^{\infty}\mathrm{d}y_N f(y_N)\int\limits_{-\infty}^{y_N}\mathrm{d}y_{N-1}f^2(y_{N-1})F^{N-3}(y_{N-1})\\ &+ \ldots+\frac{1}{0!}\int\limits_{-\infty}^{\infty}\mathrm{d}y_N f(y_N)\int\limits_{-\infty}^{y_N}\mathrm{d}y_{N-1}f(y_{N-1})\ldots\\ &\qquad\ldots\int\limits_{-\infty}^{y_2}\mathrm{d}y_1 f^2(y_1),\end{aligned}\tag{3.44}$$

a sum with N terms, the last of which comprises $N-1$ nested integrals. Somewhat miraculously, as shown in **APPENDIX III**, this chain of integrals can be collapsed into the simple closed form advertised in (3.43). Figure 3.5 compares the asymptotic expression for $P_N(c)$ derived here with numerical simulations.

3.4 Record statistics for large drift

In section 3.2 we saw that, although record events in the LDM are generally correlated for $c>0$, the correlations tend to diminish for large c (figure 3.2). This is in some sense expected, as for $c\to\infty$ both $p_n(c)$ and $P_N(c)$ tend to unity, such that the stochastic independence relation (3.3) becomes trivially satisfied. Moreover, numerical studies [23] suggest that the rate of convergence of the record rate to its limiting value $p(c)$ increases with c and for sufficiently large values is to a good accuracy attained from the very beginning. Thus for large c (3.3) can be approximated by

$$P_N(c)\approx p(c)^N=(1-\epsilon(c))^{N-1}\approx e^{-(N-1)\epsilon(c)},\tag{3.45}$$

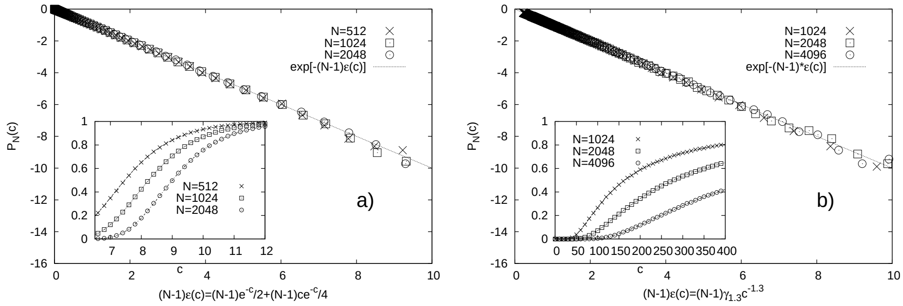

Figure 3.6: Scaling collapse of $P_N(c)$ as suggested by the asymptotic expression (3.45) for **a)** Laplace density $f(x) = e^{-|x|}/2$ and **b)** Lévy-density with $\mu = 1.3$. The ordinate is the corresponding expression from Eqs.(3.48) and (3.49) respectively. Note that the asymptotic expressions get more accurate for larger N. Inset shows direct plots of simulation results (points) versus asymptotic expression (lines).

where $\epsilon(c)$ is the probability that X_n is *not* a record. For large c, only X_{n-1} has an appreciable chance of keeping X_n from being a record. Thus

$$\epsilon(c) \approx \mathrm{P}\left[X_{n-1} > X_n\right] = \int\limits_c^\infty \mathrm{d}x f^{*2}(x). \tag{3.46}$$

Here $f^{*2}(x)$ denotes the twofold convolution of the probability density $f(x)$ of the i.i.d. part of X_n. To quote a few examples:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Rightarrow \epsilon(c) = \frac{1}{2}\mathrm{erfc}(c/2) \approx \frac{1}{c\sqrt{\pi}} e^{-c^2/4} \tag{3.47}$$

$$f(x) = \frac{1}{2} e^{-|x|} \Rightarrow \epsilon(c) = \frac{1}{2} e^{-c} + \frac{c}{4} e^{-c} \tag{3.48}$$

$$f(x) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-ikx+|k|^\mu} \Rightarrow \epsilon(c) = \frac{1}{2\pi} \int\limits_c^\infty \mathrm{d}x \int\limits_{-\infty}^{\infty} \mathrm{d}k e^{-ikx-2|k|^\mu} \approx \gamma_\mu c^{-\mu}, \tag{3.49}$$

with

$$\gamma_\mu = \frac{2\Gamma(1+\mu)\sin\left(\frac{1}{2}\pi\mu\right)}{\pi\mu}. \tag{3.50}$$

The first two of these examples are from the Gumbel class of extreme value statistics, whereas the third example is from the Fréchet class [15, 16]. The asymptotic expression in (3.47) is from [39], while the one in (3.49) can straightforwardly be derived from the known expression for the large-x asymptotics for $f(x)$, see e.g. [40]. Note that the large c asymptotics for the Weibull class is trivial, because both p_n and P_N become identically equal to unity once c exceeds the range of support of $f(y)$. Inserting the expressions (3.47,3.48,3.49) into (3.45) and also considering the asymptotics of the exact expression for $P_N(c)$ derived in (3.17), we see that in the limit of large N and c the behavior of the ordering probability is generally of the approximate form

$$P_N(c) \approx \exp[-N/N^*(c)], \tag{3.51}$$

where $N^*(c) \sim e^c$ for the Gumbel and exponential distributions, $N^*(c) \sim e^{c^2}$ for the Gaussian, and $N^*(c) \sim c^\mu$ for the Lévy distribution.

To verify the approximations made in this section, we performed numerical simulations, see figure 3.6 and figure 3.7. The results indicate that our approach, although quite rough and not necessarily well-controlled, does indeed capture the interesting regime rather well for sufficiently large N and c.

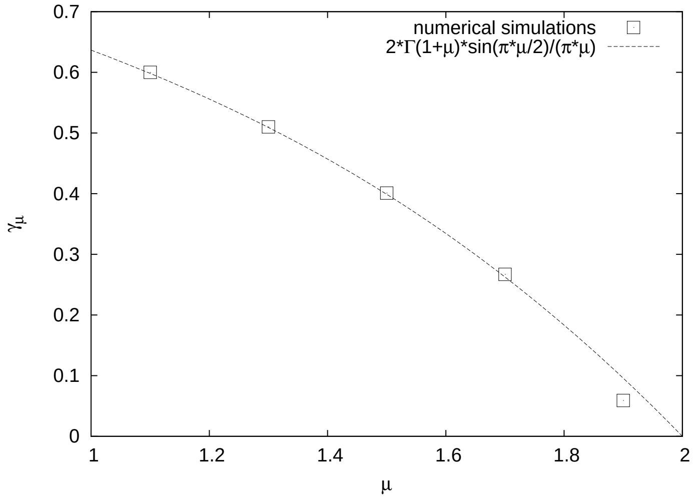

Figure 3.7: Check of the expression for γ_μ from (3.49). For $N = 1024$, the range $0 \le c \le 400$ was numerically explored as for the data shown in figure 3.6. The curves obtained this way were then fitted to the form $\exp\left(-N\gamma_\mu c^{-\mu}\right)$. The value of γ_μ obtained in this way is shown here for various values of μ and compared to the analytic expression (3.50).

3.5 Conclusions

In this article we considered the statistics of records and sequences of records of random variables with a linear trend as described by (3.4). We numerically explored the correlations between record events (cf. figure 3.2) and analytically investigated the record rate $p_n(c)$ and the ordering probability $P_N(c)$ in the limiting regimes of small and large drift velocities, $c \ll 1$ and $c \gg 1$ respectively. For the regime of $c \sim \mathcal{O}(1)$, we have not found a generally applicable method. Thus the behavior of $p_n(c)$ and $P_N(c)$ in this regime remains an open problem.

Specifically, we considered the effect of a small linear drift on distributions of the three extreme-value classes. While this effect is varying even within the individual classes we still found systematic differences between them. For the Fréchet class of distributions with power-law tails we found that the coefficient of the leading order correction to the record rate decays as $I_n \sim n^{-1/\mu}$ for large n. This implies a distinction between distributions with and without a finite first moment: For $\mu > 1$ the correction decays more slowly than the unperturbed record rate $1/n$, which implies that the drift dominates asymptotically and $p_n(c)$ attains a nonzero limit for $n \to \infty$; on the other hand, for $\mu < 1$ the drift is asymptotically irrelevant.

For the considered distributions of the Gumbel class the situation was a bit more complicated. For the exponential distribution we found a constant additive correction to the record rate, while for generalized Gaussian probability densities $f \propto e^{-|x|^\beta}$ the correction term was shown to be of order $\ln(n)^{1-\frac{1}{\beta}}$, which increases (decreases) with n when $\beta > 1$ ($\beta < 1$). For the distributions of the Weibull class, the effect of the drift is the strongest, and the correction term generally increases as a power law in n. Moreover, for highly singular distributions with $\xi < 1/2$ in (3.24), we found indications for a non-analytic behavior of $p_n(c)$ which will be investigated elsewhere. Generally speaking, narrow distributions are very sensitive to drift, while for broad distributions with heavy tails the effect is much weaker. We have also pointed out that the behavior of the first order correction term I_n obtained in this paper precisely matches earlier results for the asymptotic record rate $p(c)$ [22].

For the probability of a sequence of N consecutive records, we find the following: For

$c \ll 1$, the distribution $f(y)$ of the i.i.d. part of X_n enters to leading order in c only as a numerical constant $\int_{-\infty}^{\infty} \mathrm{d}x f^2(x)$, see (3.43), but the N-dependence is completely universal for all distributions for which the integral exists. On the other hand, for $c \gg 1$ and $N \gg 1$, the combination in which c and N enter $P_N(c)$ depends explicitly on the tail of the underlying distribution F. This indicates that somewhere in the regime of intermediate c, there is a crossover in the c-dependence of $P_N(c)$ from a highly universal to a less universal form.

The result (3.43) has important implications in the context of adaptive paths of evolutionary biology: Recalling that the expected number of accessible paths between two genotypes which are N mutations apart is given by $N!P_{N+1}$, we see in the presence of an arbitrarily small drift this quantity *increases* with N as cN. Thus even a weak systematic fitness gradient dramatically increases the accessibility of mutational pathways in the direction of increasing fitness.

Acknowledgements

This work was supported by DFG within SFB 680 *Molecular basis of evolutionary innovations* and the Bonn-Cologne Graduate School for Physics and Astronomy.

APPENDIX I - Computation of I_n for the Gaussian distribution

We begin by computing the saddle point $\tilde{y}$ defined by $\mathrm{d}_y g(\tilde{y}) = 0$, where the function $g(y)$ is given in (3.35). The saddle point satisfies

$$-2\tilde{y} + (n-2)\,\frac{e^{-\frac{\tilde{y}^2}{2}}}{\int_{-\infty}^{\tilde{y}} \mathrm{d}y' e^{-\frac{y'^2}{2}}} = 0. \tag{3.52}$$

For large n this can only be solved by $\tilde{y} \gg 1$, which implies that $\int_{-\infty}^{\tilde{y}} \mathrm{d}y' e^{-\frac{y'^2}{2}} \approx \sqrt{2\pi}$ and reduces (3.52) to

$$\frac{\sqrt{8\pi}\,\tilde{y}}{n-2} = e^{\frac{-\tilde{y}^2}{2}}. \tag{3.53}$$

By taking the square on both sides of (3.53) one finds that the solution is given in terms of the Lambert W-function $\mathrm{W}(z)$ [37, 38] as

$$\tilde{y} = \sqrt{\mathrm{W}\left(\frac{(n-2)^2}{8\pi}\right)} \tag{3.54}$$

(recall that $\mathrm{W}(z)$ is defined implicitly through $\mathrm{W}(z)e^{\mathrm{W}(z)} = z$). Using (3.53) the function g and its second derivative at the saddle point take the form

$$g(\tilde{y}) \approx -\tilde{y}^2 - 2 \tag{3.55}$$

and

$$\mathrm{d}_y^2 g(\tilde{y}) \approx -2(1+\tilde{y}^2). \tag{3.56}$$

It follows that

$$I_n \approx \frac{n\,(n-1)}{4\pi}\sqrt{\frac{-2\pi}{\mathrm{d}_y^2 g\,(\tilde{y})}}\; e^{g(\tilde{y})} \approx \frac{n\,(n-1)}{4\pi e^2}\sqrt{\frac{\pi}{1+\tilde{y}^2}}\; e^{-\tilde{y}^2}. \tag{3.57}$$

Using once more (3.53) to replace $e^{-\tilde{y}^2}$ we obtain

$$I_n \approx \frac{2\sqrt{\pi}}{e^2}\frac{n(n-1)}{(n-2)^2}\frac{\tilde{y}^2}{\sqrt{1+\tilde{y}^2}} \to \frac{2\sqrt{\pi}}{e^2}\tilde{y} \approx \frac{2\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)} \tag{3.58}$$

for large n, where we have used the expansion [37] $\mathrm{W}(z) \approx \ln(z) - \ln(\ln(z))$ to evaluate (3.54). This expansion also shows that the leading corrections to the asymptotic expression (3.58) are of order $\ln(\ln(n^2/8\pi))/\ln(n^2/8\pi)$, which accounts for the relatively large deviations from the numerical results seen in figure 3.4.

APPENDIX II - Generalized Gaussian distributions

Here we consider probability densities of the form

$$f(y) = C_\beta e^{-|x|^\beta} \tag{3.59}$$

with $\beta > 0$ and $C_\beta = [2\Gamma(1+\frac{1}{\beta})]^{-1}$. We want to evaluate the integral (3.39) in the saddle point approximation. Introducing the function

$$g(y) = -2y^\beta + (n-2)\ln\left(C_\beta \int_{-\infty}^{y} \mathrm{d}y' e^{-|y'|^\beta}\right), \tag{3.60}$$

the saddle point equation $\mathrm{d}_y g(y) = 0$ reads, for large n,

$$\frac{2\beta C_\beta}{n-2}\tilde{y}^{\beta-1} = e^{-\tilde{y}^\beta}. \tag{3.61}$$

The solution can again be expressed in terms of the Lambert-W function. Defining $\eta := 1-\frac{1}{\beta}$ we find

$$\tilde{y} \approx \left(\eta \mathrm{W}\left(\eta^{-1}\left(\frac{n-2}{2\beta C_\beta}\right)^{\eta^{-1}}\right)\right)^{\frac{1}{\beta}}. \tag{3.62}$$

Note that this expression is valid both for $\beta > 1$ ($\eta > 0$) and for $\beta < 1$ ($\eta < 0$), but in the latter case the second real branch of $\mathrm{W}(z)$ has to be used [37]. Using the asymptotics of $W(z)$ we obtain

$$\tilde{y} \approx \left(\ln n - \eta \ln(\eta^{-1}\ln n)\right)^{1/\beta} \approx (\ln n)^{1/\beta} \tag{3.63}$$

for large n.

With the help of (3.61) the function g and its second derivative at the saddle point become

$$g(\tilde{y}) \approx -2\tilde{y}^\beta - 2 \approx -2\tilde{y}^\beta \tag{3.64}$$

and

$$\mathrm{d}_y^2 g(\tilde{y}) \approx -2\beta(\beta-1)\tilde{y}^{\beta-2} - 2\beta^2\tilde{y}^{2\beta-2} \approx -2\beta^2\tilde{y}^{2\beta-2} \tag{3.65}$$

for large $\tilde{y}$. Thus, using (3.61), we see that $e^{g(\tilde{y})} \approx e^{-2\tilde{y}^\beta} \sim \tilde{y}^{2\beta-2}/n^2$, and therefore (ignoring all constant prefactors)

$$I_n \sim n^2 \sqrt{\frac{-2\pi}{\mathrm{d}_y^2 g(\tilde{y})}} e^{g(\tilde{y})} \sim \tilde{y}^{\beta-1} \sim (\ln n)^{1-1/\beta}. \tag{3.66}$$

Appendix III - Proof of an expansion

In this appendix, we will provide the details on the expansion of $\int_{-\infty}^{\infty} \mathrm{d}x f^2(x)$ into the terms on the right hand side of (3.44). The starting point is the relation

$$\begin{aligned} F^n(x) \int_{-\infty}^{x} \mathrm{d}y f^2(y) = \quad & n \textstyle\int_{-\infty}^{x} \mathrm{d}y f(y) F^{n-1}(y) \int_{-\infty}^{y} \mathrm{d}z f^2(z) \\ & + \textstyle\int_{-\infty}^{x} \mathrm{d}z f^2(z) F^n(z), \end{aligned} \tag{3.67}$$

which can be proved by applying integration by parts to the first term on the right hand side. With the identities $F^n(\infty) = 1$ and $F^n(-\infty) = 0$, one obtains from (3.67)

$$
\begin{aligned}
\frac{1}{n!}\int\limits_{-\infty}^{\infty} \mathrm{d}z f^2(z) &= \left.\frac{F^n(x)}{n!}\int\limits_{-\infty}^{x} \mathrm{d}z f^2(z)\right|_{x=-\infty}^{\infty} \\
&= \frac{F^n(\infty)}{n!}\int\limits_{-\infty}^{\infty} \mathrm{d}z f^2(z) - \frac{F^n(-\infty)}{n!}\int\limits_{-\infty}^{-\infty} \mathrm{d}x f^2(x) \qquad (3.68) \\
&= \frac{1}{n!}\int\limits_{-\infty}^{\infty} \mathrm{d}z f^2(z) F^n(z) \qquad (3.69) \\
&\quad + \frac{1}{(n-1)!}\int\limits_{-\infty}^{\infty} \mathrm{d}z f(z) F^{n-1}(z) \int\limits_{-\infty}^{z} \mathrm{d}z' f^2(z').
\end{aligned}
$$

The first term of the sum is already identical to the first term in (3.44) for $n \equiv N-2$. Using (3.67) on the inner of the two integrals of the second term of the sum above, one obtains

$$
\begin{aligned}
\int\limits_{-\infty}^{\infty} \mathrm{d}z f(z) F^{n-1}(z) \int\limits_{-\infty}^{z} \mathrm{d}z' f^2(z') &= \int\limits_{-\infty}^{\infty} \mathrm{d}z f(z) \int\limits_{-\infty}^{z} \mathrm{d}z' f^2(z') F^{n-1}(z') \\
&+ (n-1)\int\limits_{-\infty}^{\infty} \mathrm{d}z f(z) \int\limits_{-\infty}^{z} \mathrm{d}z' f(z') F^{n-2}(z') \int\limits_{-\infty}^{z'} \mathrm{d}z'' f^2(z'').
\end{aligned}
$$

Dividing by $(n-1)!$ and putting this back into (3.68) with $n = N-2$, one sees that now the first *two* terms of the sum agree with (3.44). By repeating this procedure on the terms that do not yet match and noting that finally $F^0(z) = 1$, one has expanded $\int_{-\infty}^{\infty} \mathrm{d}x f^2(x)$ into the RHS of (3.44), which concludes the proof of (3.43).

Bibliography

[1] G. W. Bassett, Climatic Change **21**, 303 (1992).

[2] R. E. Benestad, Climate Res. **25**, 3 (2003).

[3] S. Redner and M. R. Petersen, Phys. Rev. E **74**, 061114 (2006).

[4] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).

[5] G. Wergen and J. Krug, EPL **92**, 30008 (2010).

[6] D. Gembris, J. G. Taylor, and D. Suter, Nature **417**, 506 (2002).

[7] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).

[8] J. Krug and K. Jain, Physica A **358**, 1 (2005).

[9] C. Sire, S. N. Majumdar, and D. S. Dean, J. Stat. Mech.: Theor. Exp. **L07001** (2006).

[10] S.-C. Park and J. Krug, J. Stat. Mech.: Theor. Exp. **P04014** (2008).

[11] N. Glick, Amer. Math. Mon. **85**, 2 (1978).

[12] V. B. Nevzorov, Theor. Probab. Appl. **32**, 201 (1987).

[13] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *Records* (New York: Wiley, 1998).

[14] V. B. Nevzorov, *Records: Mathematical Theory* (Providence, RI: American Mathematical Society, 2001).

[15] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics* (Malabar: Krieger, 1987).

[16] L. De Haan and A. Ferreira, *Extreme Value Theory* (New York: Springer, 2006).

[17] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools* (Berlin: Springer, 2000).

[18] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).

[19] S. N. Majumdar, Physica A **389**, 4299 (2010).

[20] R. Ballerini and S. Resnick, J. Appl. Probab. **22**, 487 (1985).

[21] R. Ballerini and S. Resnick, Adv. Appl. Prob. **19**, 801 (1987).

[22] P. Le Doussal and K. J. Wiese, Phys. Rev. E **79**, 051105 (2009).

[23] G. Wergen, "Diploma Thesis," (2009).

[24] K. Borovkov, J. Appl. Probab. **36**, 668 (1999).

[25] D. M. Weinreich, R. A. Watson, and L. Chao, Evolution **59**, 1165 (2005).

[26] D. M. Weinreich *et al.*, Science **312**, 111 (2006).

[27] T. Aita *et al.*, Biopolymers **54**, 64 (2000).

[28] A. Kloezer, "Diploma Thesis," (2008).

[29] J. Franke *et al.*, PLoS Comp. Biol. **7**, e1002134 (2011).

[30] J. Krug, J. Stat. Mech.: Theor. Exp. **P08017** (2007).

[31] A. Comtet, S. N. Majumdar, and S. Ouvry, J. Phys. A: Math. Theor. **40**, 11255 (2007).

[32] A. Comtet *et al.*, J. Stat. Mech.: Theor. Exp. **P10001** (2007).

[33] E. W. Weisstein, "q-Pochhammer Symbol," Wolfram's MathWorld ().

[34] R. Ballerini, Stat. Prob. Lett. **5**, 83 (1987).

[35] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011).

[36] L. De Haan and E. Verkade, J. Appl. Prob. **24**, 62 (1987).

[37] R. M. Corless *et al.*, Adv. Comp. Math. **5**, 329 (1996).

[38] E. W. Weisstein, "Lambert W-Function," Wolfram's MathWorld ().

[39] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (New York: Dover, 1965).

[40] J. P. Bouchaud and A. Georges, Phys. Reports **195**, 127 (1990).

Chapter 4

Correlations between record events in sequences of random variables with a linear trend

Gregor Wergen, Jasper Franke and Joachim Krug

Institute for Theoretical Physics, University of Cologne

Published in: Journal of Statistical Physics 144 (2011) 1206-1222 ⋆

Abstract: The statistics of records in sequences of independent, identically distributed random variables is a classic subject of study. One of the earliest results concerns the stochastic independence of record events. Recently, records statistics beyond the case of i.i.d. random variables have received much attention, but the question of independence of record events has not been addressed systematically. In this paper, we study this question in detail for the case of independent, non-identically distributed random variables, specifically, for random variables with a linearly moving mean. We find a rich pattern of positive and negative correlations, and show how their asymptotics is determined by the universality classes of extreme value statistics.

⋆ A few typos in sections 4.3.1 and 4.3.3 are corrected. See also:
G. Wergen, J. Franke and J. Krug, *Erratum to: Correlations Between Record Events in Sequences of Random Variables with a Linear Trend*, J. Stat. Phys. **145**, 1405 (2011)

4.1 Introduction

The interest in studying record events in a general sense is evident as they by definition tend to be rare events of a special nature. As examples of record-breaking events most readers might think about athletic records like a world record in 100 m dash, or long jump [1, 2]. However records also play an important role in other naturally occurring and man-made phenomena. In particular, climatic records attract much attention and are frequently mentioned in the media in the context of global warming. The statistics of record temperatures, precipitation amounts and floods have been adressed extensively in recent work [3–10]. Also in evolutionary biology record events have a certain relevance as was pointed out in [11–13], and they play an important role in the dynamics of disordered systems like spin glasses and dirty superconductors [14, 15]. Last but not least records can be studied in the financial sciences, e.g. when considering record-breaking values or changes in stock prices [16]. In all these examples not only the rate of record events is of interest but also their correlations: Given that a record flood just occurred, can one relax because it will take a while before the next record flood is to be expected? Based on our previous work in [17], this question will be addressed in the present paper.

Most previous studies of record events and record values [10, 18–20] have been restricted to the case where the underlying random variables (RV's) $\{X_l\}$ are independently drawn from a common distribution $f(x)$ (i.i.d. RV's). In this case, record events can be shown to be stochastically independent, i.e. the knowledge about previous record events does not indicate whether a future record event is more or less likely. However in many applications, the underlying random variables are not i.i.d. For example, in recent work Majumdar and Ziff computed the statistics of records in a symmetric random walk [21] (see also [16, 22, 23]). Here, record events are obviously not independent: In a random walk the probability for a second record to occur immediately after after a previous record is simply 1/2, whereas the unconditional probability for a record in the N'th step is $\approx \frac{1}{\sqrt{\pi N}}$ [21].

In this article we consider correlations between records in sequences of RV's that are independent but not identically distributed. An important example is the statistics of hot or cold records in weather data. In [7] the occurrence of record-breaking temperatures in European and American weather recordings was analyzed by modeling daily temperature measurements as uncorrelated RV's with a linear drift. It was shown that the increase of hot records and the decrease of cold records due to global warming can be described and quantified by this model. We will refer to this particular example of non-i.i.d. RV's, which was originally introduced in the 1980's in the context of athletic records [24, 25], as the Linear Drift Model (LDM).

The precise definition of the LDM is as follows: For a positive constant c, each random variable Y_l has an i.i.d. part X_l with density $f(x)$ and an additive linear drift such that

$$Y_l = X_l + cl. \tag{4.1}$$

In other words, while the shape of the density $f_l(y)$ of the l^{th} RV is invariant, its overall position is shifted by a constant amount c each time a new RV is drawn. As was already noted in [17], the record events under this model are not in general independent (see also [26] for the related case of RV's with changing variance). Below we shall see that the sign and magnitude of correlations depend in a highly non-trivial way on the underlying probability density $f(x)$. To quantify the correlations, consider the joint probability $p_{N,N-1}$ of a record occurring in the $(N-1)^{th}$ as well as in the N^{th} step,

$$p_{N,N-1} \equiv \text{Prob}[Y_N \ \text{ record and } \ Y_{N-1} \ \text{ record}]. \tag{4.2}$$

To normalize this joint probability, we divide by the probabilities p_N and p_{N-1} of records in the N^{th} and $N-1^{th}$ steps, obtaining the key quantity under consideration throughout this paper,

$$l_{N,N-1}(c) \equiv \frac{p_{N,N-1}}{p_N p_{N-1}}. \tag{4.3}$$

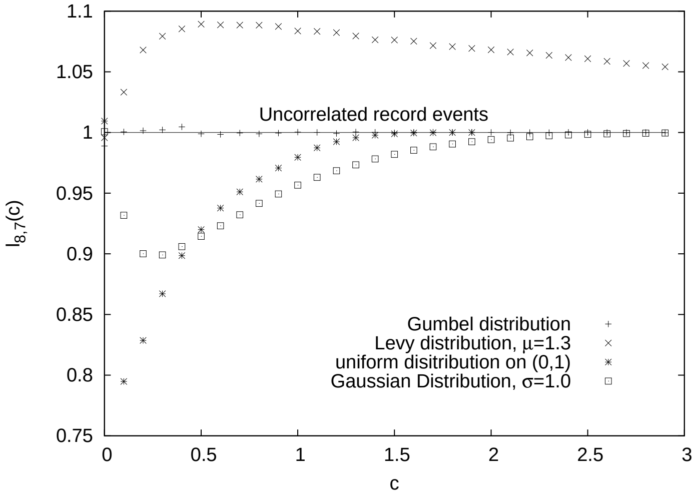

Figure 4.1: Correlations between record events in the 7^{th} and 8^{th} RV for different distributions of the underlying i.i.d. part. For the Gumbel case, the well known stochastic independence of record events is found, while simulations for the other distributions show that generally the records are correlated. Note in particular the curve for the symmetric Lévy-stable distribution with index $\mu = 1.3$, which indicates an attraction between record events.

The simulations shown in Fig. 4.1 illustrate two known results, the fact that record events for i.i.d. RVs are stochastically independent [18–20] (all curves meet at unity for $c \equiv 0$), and that for the LDM with Gumbel distributed i.i.d. part, record events are stochastically independent for all values of c [17, 19, 24, 25]. However the curves for other distributions show that stochastic independence does not hold in general. For some choices of the underlying probability density $f(x)$ record events tend to repel each other ($l_{N,N-1}(c) < 1$), wheras in other cases the correlations are positive, i.e. record events attract each other ($l_{N,N-1}(c) > 1$). In particular this last, somewhat counterintuitive numerical finding triggered our interest in this problem.

In the next section we will derive a general expression for the correlations between record events as quantified by $l_{N,N-1}(c)$. For this we employ methods similar to those used in [17]. In Section 4.3, we provide some explicit examples illustrating the rich and counterintuitive behavior of the correlations. In particular, we discuss the behavior of correlations with respect to the three classes of extreme value statistics [10, 27–29]. We find that while in the Weibull class of distributions with finite support the correlations are generally negative, positive correlations mainly arise when the RV's are in the Frechét class of distributions with power law tails. In the intermediate Gumbel class of distributions with exponential-like tails correlations can be both positive and negative. Finally in Section 4.4, we summarize our results and discuss possible applications and open questions.

4.2 General theory

Consider a sequence $\{X_l\}$ of independent but not necessarily identically distributed RV's, with $f_l(x)$ denoting the probability density of X_l. The probability that the maximum of the first N RV's is smaller than a given value x then factorizes into the probabilities of the RV's individually being less than x,

$$\text{Prob}[\max\{X_1, ..., X_N\} \leq x] = \prod_{l=1}^{N} \int^{x} \mathrm{d}x' f_l(x') = \prod_{l=1}^{N} F_l(x), \tag{4.4}$$

where $F_l(x) = \int^x \mathrm{d}x' f_l(x')$ denotes the cumulative distribution function for the density $f_l(x)$, and the omitted lower (upper) integral boundary is understood to be the lower (upper) bound of the support of f_l respectively. Throughout we assume distribution functions to be continuous.

Using Eq. 4.4 the probability p_N that the N^{th} RV is a record can be written as

$$p_N = \int \mathrm{d}x f_N(x) \prod_{l=1}^{N-1} F_l(x), \tag{4.5}$$

and the joint probability that both X_{N-1} and X_N are records is

$$p_{N,N-1} = \int \mathrm{d}x_N f_N(x_N) \int^{x_N} \mathrm{d}x_{N-1} f_{N-1}(x_{N-1}) \prod_{l=1}^{N-2} F_l(x_{N-1}). \tag{4.6}$$

For the LDM we have $f_l(y) = f(y - cl)$ and $F_l(y) = F(y - cl)$, where $f(x)$ and $F(x)$ are the density and distribution function of the i.i.d. part X_l in (4.1). It then follows that [17]

$$p_N = \int \mathrm{d}y\, f(y - cN) \prod_{l=1}^{N-1} F(y - cl) = \int \mathrm{d}x\, f(x) \prod_{l=1}^{N-1} F(x + cl) \tag{4.7}$$

and similarly

$$p_{N,N-1} = \int \mathrm{d}y\, f(y) \int^{y+c} \mathrm{d}x\, f(x) \prod_{l=1}^{N-2} F(x + cl). \tag{4.8}$$

4.2.1 Explicit stochastic independence for two cases

The expressions (4.6) and (4.8) allow us to verify two known results: Both for the i.i.d. case ($c \equiv 0$) and for the LDM with Gumbel distributed i.i.d. part X_l, record events are independent.

We start with the i.i.d. case. First note that the probability that the N^{th} RV is a record is the probability that it is the largest among N equivalent RV's, so by symmetry $p_N(c = 0) = 1/N$. For the i.i.d. case, the subscripts in (4.6) can be dropped. Using the substitution $u = F(x)$ and $\mathrm{d}u = f(x)\mathrm{d}x$, one obtains

$$p_{N,N-1}(c = 0) = \int_0^1 \mathrm{d}u \int_0^u \mathrm{d}u' u'^{N-2} = \frac{1}{N-1}\frac{1}{N} = p_N p_{N-1}. \tag{4.9}$$

It is possible to insert an arbitrary spacing between the two record events and still obtain this factorization to yield $p_{N,N-j} = p_N p_{N-j}$. Similarly, one can consider more than one record event. The factorization property implies that in the i.i.d. case, record events are stochastically independent [18–20].

Next we consider the LDM with Gumbel distributed i.i.d. part,

$$f(x) = \exp\left(-e^{-x} - x\right) \tag{4.10}$$

and $F(x) = \exp(-e^{-x})$. Using that in this case [27]

$$F(x + a) = \exp\left(-e^{-x-a}\right) = \exp\left(-e^{-x}\right)^{e^{-a}} = F(x)^{e^{-a}}, \tag{4.11}$$

the substitution $u = F(x)$ used above yields [17, 24]

$$p_N(c) = \int \mathrm{d}x f(x) \prod_{l=1}^{N-1} F(x)^{e^{-cl}} = \int_0^1 \mathrm{d}u\; u^{\sum_{l=1}^{N-1} \alpha^l} = \frac{1}{\sum_{l=0}^{N-1} \alpha^l} \tag{4.12}$$

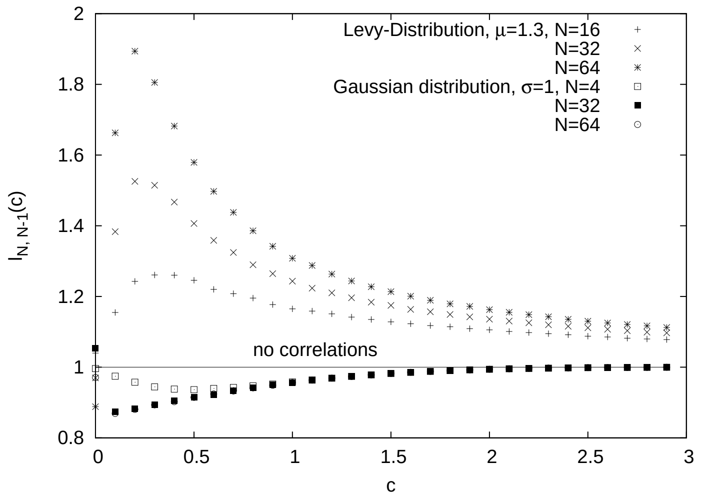

Figure 4.2: Correlations between record events as the number of RV's N increases. For the i.i.d. part drawn from a Gaussian distribution with variance $\sigma = 1$, the correlations are negative and for sufficiently large N tend to a limiting form. For the symmetric Lévy distributed i.i.d. part, the correlations are positive and increase with N.

with $\alpha = e^{-c}$, and

$$\begin{aligned} p_{N,N-1}(c) &= \int_0^1 \mathrm{d}u \int_0^{u^\alpha} \mathrm{d}u' u'^{\sum_{l=1}^{N-2} \alpha^l} = \frac{1}{\sum_{l=0}^{N-2} \alpha^l} \int_0^1 \mathrm{d}u \; u^{\sum_{l=1}^{N-1}} \\ &= \left(\frac{1}{\sum_{l=0}^{N-2} \alpha^l} \right) \left(\frac{1}{\sum_{l=0}^{N-1} \alpha^l} \right) = p_{N-1} p_N . \end{aligned} \tag{4.13}$$

Again one can introduce an arbitrary number of spaces and intermediate records and still have the corresponding joint probability factorize. This implies that for the LDM with Gumbel distributed i.i.d. part, as for arbitrary distributions without drift, record events are stochastically independent.

4.2.2 Expansion for small drift

As shown above, for the i.i.d. case $c \equiv 0$, record events are independent. On the other hand, for very large drift, almost every RV is a record, and thus trivially $l_{N,N-1}(c) \to 1$ for $c \to 1$ (see Fig. 4.1). The behavior of $l_{N,N-1}(c)$ in between these two limits is not as easily characterized. In particular, while for some probability densities the records seem to repel each other for intermediate c ($l_{N,N-1}(c) < 1$), for other probability densities the records seem to *attract* ($l_{N,N-1}(c) > 1$).

Figure 4.2 shows examples for both types of behavior, illustrating in addition how the correlations evolve with increasing N for the cases of Lévy and Gauss distributed i.i.d. parts. Clearly, the correlations are not a finite size effect, but their N-dependence is markedly different in the two cases. For the Gaussian case, the correlations are negative and seem to approach a form independent of N. For the Lévy case, on the other hand, the correlations seem to increase with N. We will argue later that the correlations eventually become N-independent also for heavy-tailed distributions, but this limit is approached much more slowly than in the Gaussian case.

A first step towards understanding the behavior of the correlations is to determine the sign with which the curves in Figs.4.1 and 4.2 depart from the i.i.d. case $c \equiv 0$. To address this question, we compute the record rate $p_N(c)$ and the joint probability $p_{N,N-1}(c)$ in the limit of small c and $c/\sigma \ll N^{-1}$, where σ is the standard deviation or some other measure of the width of the distribution. For the record rate this problem was discussed extensively in [17] (see also [7]). Knowing that $p_N(c=0)=1/N$, a Taylor expansion for small c yields

$$p_N\left(c\right) \approx \frac{1}{N} + c\frac{N\left(N-1\right)}{2} I(N-2),. \tag{4.14}$$

where we have defined the quantity[1]

$$I\left(N\right) = \int \mathrm{d}x f^2\left(x\right) F^N\left(x\right). \tag{4.15}$$

In the same small c regime we can also easily determine the joint probability of occurrence of two consecutive records $p_{N,N-1}\left(c\right)$. To leading order in c, the expansion of Eq.(4.8) yields

$$\begin{aligned} p_{N,N-1}\left(c\right) \quad &\approx \quad \int \mathrm{d}y f\left(y\right) \int^{y} \mathrm{d}x f\left(x\right) F^{N-2}\left(x\right) + c \int \mathrm{d}y f^2\left(y\right) F^{N-2}\left(y\right) \\ & \quad +c\frac{\left(N-1\right)\left(N-2\right)}{2} \int \mathrm{d}y f\left(y\right) \int^{y} \mathrm{d}x f^2\left(x\right) F^{N-3}\left(x\right). \end{aligned} \tag{4.16}$$

The first term can be evaluated by partial integration to $\frac{1}{N(N-1)}$, and another partial integration yields

$$p_{N,N-1}\left(c\right) \quad \approx \quad \frac{1}{N\left(N-1\right)} - c\frac{\left(N-1\right)\left(N-2\right)-2}{2} I\left(N-2\right) \tag{4.17}$$

$$+ \quad c\frac{\left(N-1\right)\left(N-2\right)}{2} I\left(N-3\right). \tag{4.18}$$

In addition to the zeroth order term $\frac{1}{N(N-1)}$, we find a positive and a negative correction term that describe the effect of the drift. Depending on which one of these two terms is larger, positive or negative correlations will result.

A similar expansion can be set up for the normalized correlation $l_{N,N-1}(c)$ defined in Eq.(4.3). Writing

$$l_{N,N-1}(c) = 1 + cJ(N) + \mathcal{O}(c^2), \tag{4.19}$$

a straightforward computation yields

$$\begin{aligned} J(N) \quad &= \quad N(N-1)I(N-2) + \frac{N(N-1)^2(N-2)}{2}\left[I(N-3) - I(N-2)\right] \\ & \quad -\frac{N^2(N-1)}{2} I(N-2) - \frac{(N-1)^2(N-2)}{2} I(N-3) \end{aligned} \tag{4.20}$$

Again, both positive and negative terms occur in (4.20), the relative magnitude of which depends explicitly on the underlying probability density $f(x)$. Thus from (4.20), it is not at all obvious how the counterintuitive positive correlations observed numerically arise, how their occurrence depends on the properties of f and how the clustering of record events behaves as a function of the number N of RV's considered. Before turning to the explicit evaluation of $J(N)$ in Section 4.3, we provide some heuristics on the asymptotics of the correlations in the limit of very large N, and address the behavior of correlations between RV's that are more than one time step apart.

---

[1] Note that this definition differs slightly from that of the related quantity I_n used in [17]. The integral (4.15) also appears in the analysis of the density of near-extreme events [30].

4.2.3 Asymptotics for large N

It is clear from Eqs. (4.7) and (4.8) that the asymptotic behavior of the record rate p_N and the joint probability $p_{N,N-1}$ hinges on the existence of the function

$$G_c(x) \equiv \lim_{N\to\infty} \prod_{j=1}^{N} F(x+cj) \tag{4.21}$$

which was rigorously examined by Ballerini and Resnick. In [24], they prove that the limit (4.21) exists and is nonzero whenever the density $f(x)$ has a finite first moment. For completeness we provide a heuristic version of their argument. Taking the logarithm of Eq. (4.21) one has to consider the convergence of the series

$$\ln\left(G_c(x)\right) = \sum_{j=1}^{\infty} \ln\left(F(x+cj)\right).$$

Now $F(x+cj) = \mathrm{Prob}(X \leq x+cj) = 1 - \mathrm{Prob}(X > x+cj)$ and for any finite x and $c > 0$ there is a $\tilde{j}$ large enough such that $\mathrm{Prob}(X > x + c\tilde{j}) \ll 1$. Therefore

$$\ln\left(G_c(x)\right) \approx s_{\tilde{j}}(x) - \sum_{j=\tilde{j}}^{\infty} \mathrm{Prob}(X > x+cj), \tag{4.22}$$

where $s_{\tilde{j}} = \sum_{j=1}^{\tilde{j}-1} \ln(F(x+cj)$ is a finite sum and the logarithms in the remaining series were approximated by their Taylor expansion around unity. The condition for $f(x)$ to have a finite first moment is that $\mathrm{Prob}(X > x)$ decays faster than $1/x$, which implies convergence of the infinite series in (4.22) and hence of $G_c(x)$ (in the opposite case the series diverges to $-\infty$ and $G_c \equiv 0$). Since $G_c(x)$ is a probability, it is bounded from above by unity and it follows from (4.7) and (4.8) that $p_N(c)$ and $p_{N,N-1}(c)$ have finite, non-zero limits for $N \to \infty$. The same statement then applies to the ratio (4.3).

Although the focus of this paper is on the case of a positive drift which enhances the occurrence of (upper) records, it is instructive to compare the asymptotic behavior for $c > 0$ described above to the case $c < 0$. For negative drift both the record rate p_N and the joint probability $p_{N,N-1}$ vanish for $N \to \infty$, and they do so more rapidly than their i.i.d. counterparts. The asymptotic behavior of the ratio (4.3) is then *a priori* undetermined. We consider a simple example where the quantities of interest can be computed explicitly. Let the probability density of the i.i.d. variables be given by the negative exponential distribution,

$$f(x) = e^{x}, \quad x \leq 0 \tag{4.23}$$

and $f(x) = 0$ elsewhere. Then for $c < 0$ the integrand in (4.7) and (4.8) is of the form

$$\prod_{l=1}^{K} F(x+cl) = \exp\left[Kx + \frac{c}{2}K(K+1)\right], \quad x \leq 0, \tag{4.24}$$

and integration yields

$$p_N(c) = \frac{1}{N}\exp\left[\frac{c}{2}N(N-1)\right], \; p_{N,N-1}(c) = \frac{1}{N(N-1)}\exp\left[\frac{c}{2}N(N-1)\right]. \tag{4.25}$$

For $c < 0$ the record rate is suppressed superexponentially, such that the expected number of records remains finite for $N \to \infty$ (see also [17]). However, forming the ratio (4.3) we see that

$$l_{N,N-1} = \exp\left[-\frac{c}{2}(N-1)(N-2)\right] \tag{4.26}$$

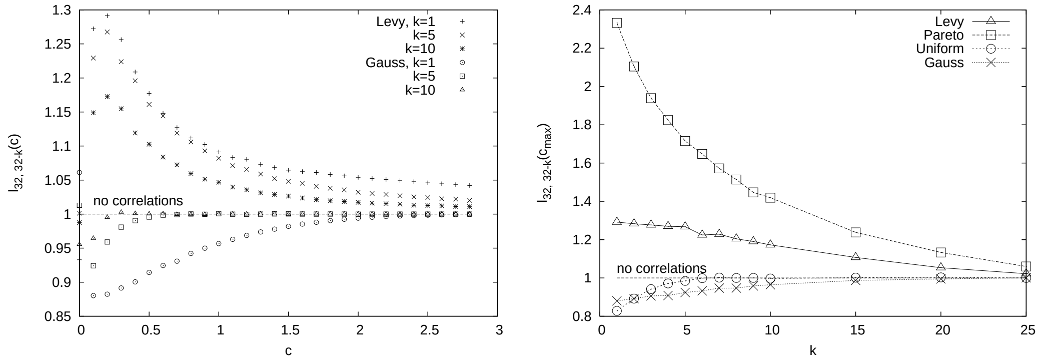

Figure 4.3: Correlations between record events at distance k from time series of length $N = 32$ for Gaussian ($\sigma = 1$), Lévy-stable ($\mu = 1.5$), uniform (on $[0, 1]$) and Pareto ($\mu = 1.5$) distributions. **Left:** Full c dependence of $l_{N,N-k}(c)$. **Right:** At the drift velocity $c_{\max}$ where the correlations are maximal ($c_{\max} = 0.1$ for Gaussian and uniform, $c_{\max} = 0.2$ for the Lévy and $c_{\max} = 0.4$ for the Pareto case), $l_{N,N-k}(c_{\max})$ is shown as function of k to illustrate the decay of correlations.

which, in contrast to the case of positive drift, grows without bound for $N \to \infty$. We will see in Section 4.3 that such a behavior is not restricted to this specific example. The non-existence of the $N \to \infty$ limit for $c < 0$ suggests that the limiting function

$$l^*(c) \equiv \lim_{N\to\infty} l_{N,N-1}(c) \tag{4.27}$$

may not be smooth for $c \to 0$, as will be explicitly demonstrated in Section 4.3.

4.2.4 Dependence on the distance between record events

Before embarking on the study of specific distributions, we briefly comment on the behavior of the correlations as a function of the distance between record events. Denoting by $p_{N,N-k}$ the joint probability for observing a record in the $N-k$th and the Nth event, it was shown in [24] that

$$\lim_{k\to\infty} \lim_{N\to\infty} p_{N,N-k}(c) = \left[\lim_{N\to\infty} p_N(c)\right]^2 \equiv p(c)^2,$$

which implies that record events become uncorrelated for $k \to \infty$. This leads to the expectation that in a finite time series, the correlations between record events at distance k should decay with k, in agreement with the simulations shown in Fig. 4.3. The correlations are seen to be maximal at $k = 1$. In the present paper we therefore focus on the case of nearest neighbor correlations, although extending the computations to $k > 1$ is in principle straightforward.

4.3 Explicit examples

Whereas the statistics of record events in sequences of i.i.d. RV's is completely universal [18–20], in the presence of drift one has to distinguish between distributions belonging to the three different universality classes of extreme value theory, the classes of Weibull, Gumbel and Fréchet [10, 28, 29]. For each of these classes we will analyze the correlations in the LDM for a few exemplary distributions, starting with the Weibull class, and summarize the observed behavior in a unifying scaling picture in Section 4.3.4. The results presented below rely on the study of the record rate $p_N(c)$ in the LDM presented in [17], as well as on the expansions for $p_{N,N-1}(c)$ [Eq. (4.17)] and $l_{N,N-1}(c)$ [Eqs. (4.19,4.20)] derived above in Section 4.2.2.

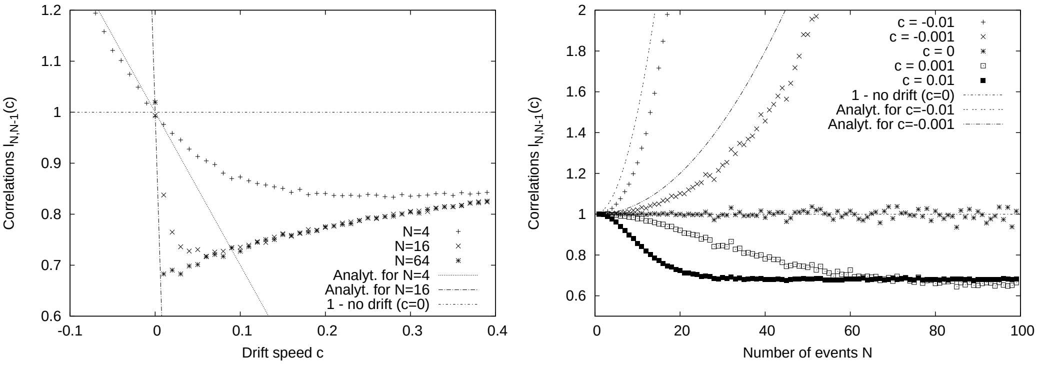

Figure 4.4: Left: Numerical simulations of $l_{N,N-1}(c)$ for a uniform distribution of width $a = 1$. For each c and N we simulated 10^6 time series. The figure shows data for $N = 4, 16$ and 64. At $c = 0$ we find a steep descent of $l_{N,N-1}(c)$. For $N = 4$ and $N = 16$ we show the analytic prediction (4.28) (steep lines). With increasing c, $l_{N,N-1}(c)$ appears to become completely independent of N. **Right:** Simulation results for $l_{N,N-1}$ as a function of N for $c = -0.01, -0.001, 0, 0.001$ and 0.01. We simulated 10^7 series of 100 RV's in each case. For $c < 0$ we find a very steep increase of $l_{N,N-1}$ that is in good agreement with the analytical prediction (4.28). For $c > 0$, $l_{N,N-1}$ quickly saturates at a constant value and our analytical results are not very useful.

4.3.1 The Weibull class

The Weibull class is the class of distributions with bounded support. As a very simple member of the Weibull class we consider a uniform distribution on the interval $[-a, a]$. We compute $l_{N,N-1}(c)$ in the small c regime by making use of (4.19,4.20) and find for $N \gg 1$

$$l_{N,N-1}(c) \approx 1 - \frac{c}{4a} N^2. \tag{4.28}$$

The correlations are negative and their magnitude increases rapidly with N (but note that the leading order approximation is valid only for $cN^2/2a \ll 1$). We performed numerical simulations to check this result with $a = 1$ and $-0.1 < c < 0.4$. Figure 4.4 shows how the descent of $l_{N,N-1}(c)$ for $c > 0$ steepens with increasing N. In the figure we also plot our prediction (4.28) for small values of N, however the steepness of the descent at zero does not allow us to estimate the quality of our approximation reliably. The figure shows that for large enough N and not too small c, $l_{N,N-1}(c)$ eventually becomes independent of N, as expected from the general arguments given in Section 4.2.3. In this case the approach to the $N \to \infty$ limit is particularly rapid, because distributions that are more than $2a/c$ steps apart do not overlap. The sharp initial drop of $l_{N,N-1}(c)$ indicates that the limiting distribution $l^*(c)$ is discontinuous at $c = 0$, which is consistent with the non-existence of a limiting function[2] for $c < 0$.

As a more general subset of the Weibull class we next consider the class of distributions defined by

$$f_\xi(x) = \xi (1 - x)^{\xi - 1} \tag{4.29}$$

with $\xi > 0$ and $0 < x \leq 1$. For $\xi = 1$ this produces a uniform distribution similar to the one used above. In [17] we found that for $\xi > \frac{1}{2}$ the integral in (4.15) is given by

$$I(N) = \xi \frac{\Gamma\left(2 - \frac{1}{\xi}\right)\Gamma(N+1)}{\Gamma\left(N + 3 - \frac{1}{\xi}\right)}. \tag{4.30}$$

---

[2] With regard to its upper tail, the negative exponential distribution (4.23) is equivalent to the uniform distribution, and therefore the argument of Section 4.2.3 applies here as well.

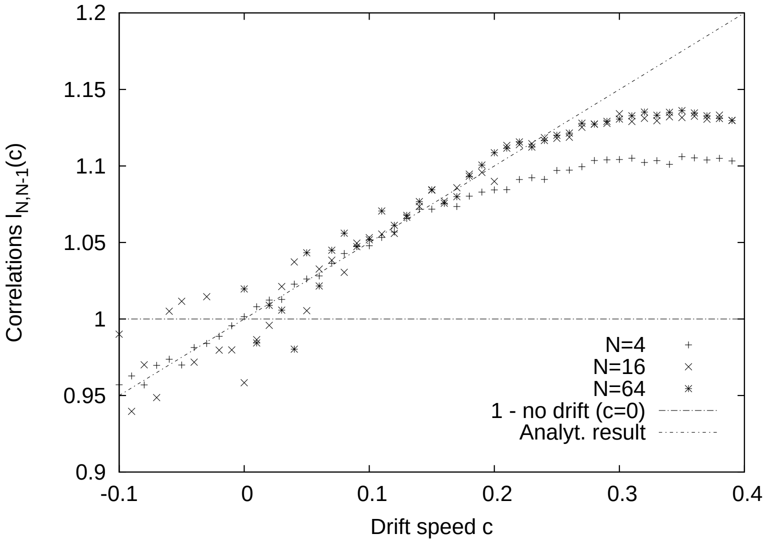

Figure 4.5: Numerical Simulations of $l_{N,N-1}(c)$ for an exponential distributions with unit mean. For each c and N we simulated 10^6 time series. The figure shows results for $N = 4, 16$ and 64 along with our N-independent analytical prediction (4.33).

Using this and assuming $N \gg 1$ we find the following expression for $l_{N,N-1}(c)$:

$$l_{N,N-1}(c) \approx 1 - \frac{cN^3}{2} \frac{\Gamma\left(2 - \frac{1}{\xi}\right)\Gamma(N-1)}{\Gamma\left(N+1-\frac{1}{\xi}\right)} \tag{4.31}$$

and making use of the Stirling approximation we finally obtain

$$l_{N,N-1}(c) \approx 1 - \frac{c}{2}\Gamma\left(2 - \frac{1}{\xi}\right) N^{1+\frac{1}{\xi}}. \tag{4.32}$$

Similar to the uniform case, the correlations between neighboring record events are always negative and their magnitude increases rapidly with increasing N, suggesting a discontinuity of the limiting function $l^*(c)$ at $c = 0$.

4.3.2 The Gumbel class

As a first example in the Gumbel class we consider the exponential distribution with mean a, $f(x) = a^{-1}e^{-\frac{x}{a}}$, $x \geq 0$. In [17] it was found that in this case the effect of the linear drift on the record rate is independent of N to leading order in c. Using this result it is straightforward to obtain

$$l_{N,N-1}(c) \approx 1 + \frac{c}{2a}. \tag{4.33}$$

We compared this result to numerical simulations in Fig. 4.5 and found a very good agreement. Apparently $l_{N,N-1}(c)$ assumes a value independent of N already for very small N.

Another important member of the Gumbel class is the normal distribution with mean zero and standard deviation σ. Here, the computation is usually a bit more complicated; however, as in the previous examples, most of the work was already done in [17]. There we found that for $N \gg 1$ and $cN \ll \sigma$

$$I(N) \approx \frac{c}{N^2\sigma} \frac{4\sqrt{\pi}}{\mathrm{e}^2} \sqrt{\ln\left(\frac{N^2}{8\pi}\right)}. \tag{4.34}$$

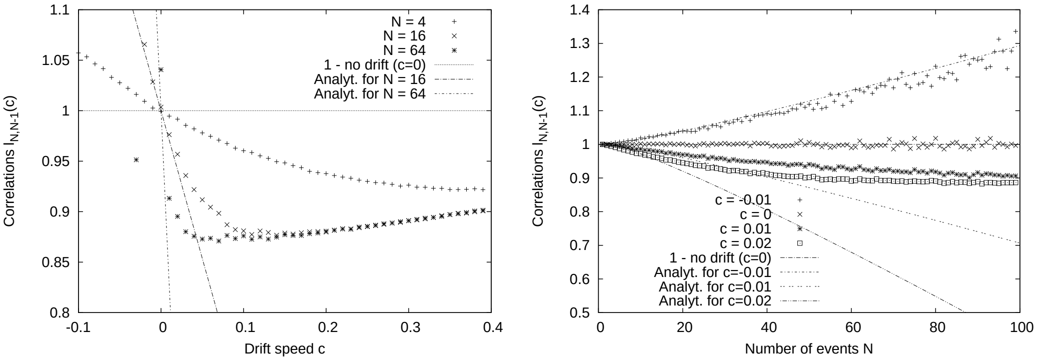

Figure 4.6: Numerical simulations of $l_{N,N-1}(c)$ for a Gaussian distribution of width $\sigma = 1$. **Left figure** shows $l_{N,N-1}(c)$ for $N = 4, 16$ and 64. For each c and N we simulated 10^7 time series. The figure illustrates the steep descent of the correlator at $c = 0$ when c is small. **Right figure** shows $l_{N,N-1}(c)$ for different, fixed values of c together with our analytical results. Here, we analyzed 10^8 series of RV's for each drift rate. Again, we find agreement between the simulations and our analytical computations for small N and c. We manually fitted curves $\propto N\sqrt{\ln\left(N^2/8\pi\right)}$. Interestingly for $c < 0$ the agreement is a lot better and $l_{N,N-1}(c)$ does not saturate at a constant value.

Using (4.20) this result allows us to estimate the effect of the drift on the correlations. We find

$$\left(l_{N,N-1}\left(c\right) - 1\right) \propto -N\frac{c}{\sigma}\sqrt{\ln\left(\frac{N^2}{8\pi}\right)}. \tag{4.35}$$

Unlike the case of the exponential distribution, in the Gaussian case the correlations are negative and depend strongly on N. The behavior is similar to that found in the Weibull case, and matches the result (4.32) (up to logarithimic factors) for $\xi \to \infty$. Numerical results for the Gaussian distribution are shown in Fig. 4.6. Similar to the uniform distribution, $l_{N,N-1}\left(c\right)$ shows a steep descent at $c = 0$ which becomes steeper for increasing N. However, the descent appears to be smoother than in the uniform case. In agreement with the considerations of Section 4.2.3, for $c > 0$ and large N the correlations become independent of N. In contrast, for $c < 0$ the correlations increase monotonically with N and do not appear to saturate. In the case of the Gaussian distribution our analytical prediction (4.35) is only valid in a small regime for small $c > 0$ and relatively small N, but for $c < 0$ the approximation is significantly better.

To understand the marked difference between the exponential and Gaussian distributions, we analyze the general class of Gumbel-type distributions of the form

$$f\left(x\right) = A_\beta \mathrm{e}^{-|x|^\beta}, \tag{4.36}$$

with normalization A_β and $\beta > 0$. In this case it was found in [17] that

$$I\left(N\right) \propto N^{-2}\ln\left(N\right)^{1-\beta^{-1}}. \tag{4.37}$$

With some further computations this leads us to the following behavior of the leading order correction coefficient $J\left(N\right)$:

$$J\left(N\right) \propto -D_1\left(1 - \frac{1}{\beta}\right) N \ln\left(N\right)^{1-\frac{1}{\beta}} + D_2 \ln\left(N\right)^{1-\frac{1}{\beta}}, \tag{4.38}$$

where D_1 and D_2 are positive constants not depending on N. The exact values of D_1 and D_2 are very difficult to compute and we will not consider them in this article. Nevertheless, this

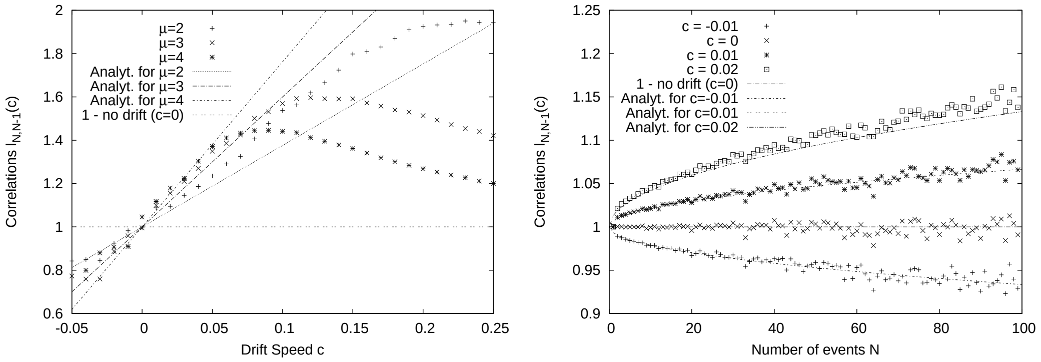

Figure 4.7: Left: Numerical simulations of $l_{N,N-1}(c)$ for Pareto distributions. For each c and μ we simulated 10^6 time series of length $N = 16$. The figure shows simulations for $\mu = 2, 3$ and 4, along with our analytical predictions for small c. At $c = 0$ we find a steep ascent of $l_{N,N-1}(c)$ which gets steeper with increasing μ. Our analytical result from (4.42) is in good agreement with the simulations for small enough c. **Right:** Simulation results for a Pareto distribution with $\mu = 2$ and different values of c. We analyzed 10^8 series of $N = 100$ RV's for each c. Our analytical work predicts a square-root behavior of $l_{N,N-1}(c) - 1$, which is in good agreement with the simulations.

expression nicely reproduces our results for the exponential and the Gaussian distribution and shows how both positive and negative correlations may emerge in the Gumbel class. For $\beta = 1$ the leading first term vanishes and $J(N)$ reduces to a positive constant. For all $\beta \neq 1$ the first term dominates and the correlations grow with N, with a sign determined by $1 - \frac{1}{\beta}$. For values $\beta > 1$ of distributions that decay faster than the standard exponential we find negative correlations between the records, in agreement with the Gaussian example, while for stretched exponential distributions ($\beta < 1$) positive, N-dependent correlations result (see Fig. 4.8 for an example).

In hindsight, the special role of the exponential distribution should not come as a surprise, since the Gumbel distribution (4.10) (for which correlations are completely absent) has an exponential tail. In the sense of extreme value theory, the exponential distribution is close to the Gumbel, and the corresponding records in the LDM are therefore almost uncorrelated (up to small residual correlations which are independent of N).

4.3.3 The Fréchet class

As a subset of the distributions in the Fréchet class of extreme value statistics we consider the well known Pareto distribution

$$f(x) = \mu x^{-\mu-1} \tag{4.39}$$

with $\mu > 1$ and $x \geq 1$. In [17] it was found that in this case

$$I(N) = \mu \frac{\Gamma\left(2 + \frac{1}{\mu}\right)\Gamma(N+1)}{\Gamma\left(N + 3 + \frac{1}{\mu}\right)}. \tag{4.40}$$

This allows us to compute $l_{N,N-1}(c)$ for the Pareto distribution. We find

$$J(N) = \frac{\mu}{2} \frac{\Gamma\left(2 + \frac{1}{\mu}\right)\Gamma(N-1)}{\Gamma\left(N + \frac{1}{\mu} + 1\right)} \frac{N(N-1)\left(N^2 + \mu N + \mu\right)}{\mu N + \mu + 1} \tag{4.41}$$

which leads to the expression

$$l_{N,N-1}(c) \approx 1 + \frac{c}{2}\Gamma\left(2 + \frac{1}{\mu}\right) N^{1-\mu^{-1}} \tag{4.42}$$

for large N. For positive c, we thus expect positive correlations between neighbouring record events for all distributions of Pareto form. These correlations are increasing with N slower than linearly. We expect that, asymptotically, this behavior is universal for all distributions within the Fréchet class[3]. We compare Eq.(4.42) to numerical simulations in Fig. 4.7, finding good agreement for small c, both for the c- and the N-dependence. The agreement improves when the distribution becomes broader for smaller μ. The strength of correlations for strongly heavy-tailed distributions of the Pareto class is remarkable. For example, for a Pareto distribution with coefficient $\mu = 2$ we found correlations $l_{N,N-1} \approx 3$ for large N (see Fig. 4.8).

The correlations displayed in the right panel of Fig.4.7 increase (for $c > 0$) or decrease (for $c < 0$) with N without showing any sign of saturation, although we know from the general considerations of Section 4.2.3 that $l_{N,N-1}$ must approach an N-independent limit for $c > 0$ and $\mu > 1$. As we will see in the next subsection, this reflects the fact that the time scale N^* at which the limit is attained is (for a given value of c) particularly large for distributions of the Fréchet class. We also note that because $l_{N,N-1} < 1$ for $c < 0$, the normalized correlations are confined to lie between 0 and 1 in this case. This implies that the divergence of $l_{N,N-1}(c)$ for $N \to \infty$ and $c < 0$, which was demonstrated in Section 4.2.3 to occur for a specific example in the Weibull class, cannot happen here. We therefore conjecture that, for distributions in the Fréchet class, the limit (4.27) exists also for $c < 0$.

4.3.4 Unified picture

The results of the preceding subsections can be summarized in a simple scaling picture. We first recall from [17] and [22] that the LDM (with $c > 0$) contains a characteristic time scale N^* at which the record rate $p_N(c)$ crosses over from the i.i.d. behavior $p_N \approx \frac{1}{N}$ to the limiting value $p(c) = \lim_{N\to\infty} p_N(c) > 0$. For the different distributions discussed above, this time scale diverges for $c \to 0$ according to

$$\begin{aligned} N^* &\propto c^{-\frac{\xi}{1+\xi}} && \text{Weibull class} && (4.43)\\ N^* &\propto c^{-1}|\ln c|^{\frac{1}{\beta}-1} && \text{Gumbel class} && (4.44)\\ N^* &\propto c^{-\frac{\mu}{\mu-1}} && \text{Fréchet class } (\mu > 1)\,. && (4.45)\end{aligned}$$

Ignoring for the moment the logarithmic factor in the Gumbel case, these behaviors can be further simplified by expressing the different universality classes in terms of the generalized Pareto distribution [31]

$$f(x) = (1 + \kappa x)^{-\frac{\kappa+1}{\kappa}}\,. \tag{4.46}$$

For $\kappa > 0$ this is of Pareto type with $\mu = \frac{1}{\kappa}$, for $\kappa < 0$ it reduces to a Weibull-type distribution similar to (4.29) with $\xi = -\frac{1}{\kappa}$, and the Gumbel class is represented by the exponential distribution which arises from (4.46) for $\kappa \to 0$. Using this representation, the different cases in (45), (46) and (47) reduce to

$$N^* \propto c^{-\nu} \text{ with } \nu = \frac{1}{1-\kappa}. \tag{4.47}$$

Note that $\nu < 1$ in the Weibull class but $\nu > 1$ in the Fréchet class, which explains the slow convergence to the $N \to \infty$ limit in the latter case. Moreover, it was shown in [17] that the integral (4.15) behaves for large N as (see also [30])

$$I(N) \sim N^{\frac{1}{\nu}-3} = N^{-(2+\kappa)}. \tag{4.48}$$

---

[3] In [17] we presented data for a Lévy stable distribution with $\mu = 1.8$ which show negative correlations between record events. We have checked that this behavior is not asymptotic, and that the correlations become positive for larger values of N.

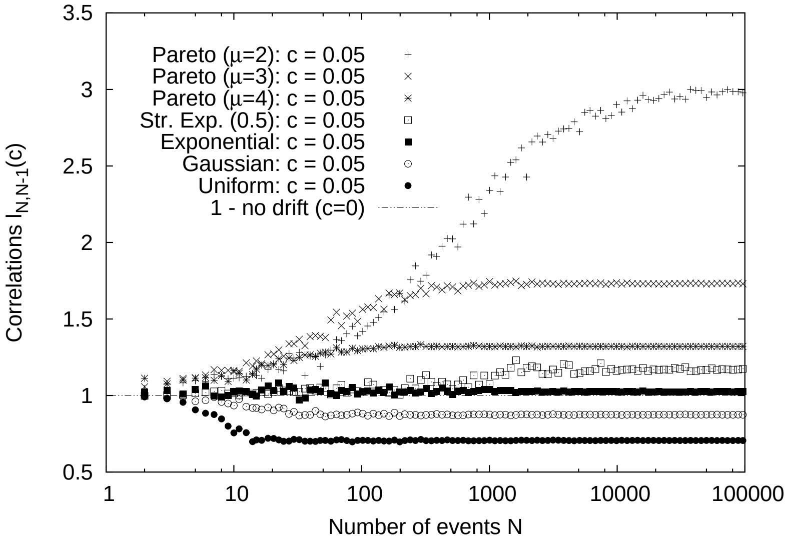

Figure 4.8: Numerical simulations of $l_{N,N-1}(c)$ with $c = 0.05$ for six different distributions. In each case we averaged over 10^6 time series of length $N = 10^5$. The results for $l_{N,N-1}(c)$ were binned logarithmically in order to improve the averaging for large N. For the uniform and the Gaussian distribution $l_{N,N-1}(c)$ is clearly negative for $c > 0$ and it decreases with growing N. At some $N = N^*$ the correlations become independent of N. For the special case of the exponential distribution we find a constant value of $l_{N,N-1}(c)$ which is only slightly larger than unity. Both the stretched exponential (with $\beta = 1/2$) and the Pareto distributions (with $\mu = 2, 3, 4$) show positive correlations. Note the slow convergence for the Pareto distribution with $\mu = 2$.

To relate this to the behavior of the correlations, we note that for large N Eq.(4.20) can be approximately written as

$$J(N) \quad \approx \quad -\frac{1}{2}N^4\frac{\mathrm{d}}{\mathrm{d}N}I(N) - N^3I(N) + \mathcal{O}(N^2I(N)). \tag{4.49}$$

Inserting (4.48) we thus conclude that, to leading order,

$$J(N) \approx \frac{\kappa}{2}N^3I(N) \sim N^{1-\kappa}, \tag{4.50}$$

which correctly reproduces both the sign and the order of magnitude of the correlations derived in Section 4.3 for the Weibull and Fréchet classes: Correlations are positive (negative) for $\kappa > 0$ ($\kappa < 0$), and they scale sublinearly (superlinearly) with N in the two cases. A similar calculation using the refined estimate (4.37) of $I(N)$ for the Gumbel class shows that the correlations are negative (positive) for $\beta > 1$ ($\beta < 1$), in agreement with (4.38).

4.4 Conclusion

In this paper we analyzed the effect of a linear drift on the correlations between records drawn from series of independent RV's, as quantified by the normalized joint probability $l_{N,N-1}(c)$. In Section 2 we derived general expressions (exact and approximate) for $l_{N,N-1}(c)$ and recalled the fact that records are independent both for i.i.d. RV's ($c = 0$) and in the special case where the RV's are drawn from the Gumbel distribution.

Our main analytic results were obtained in Section 3 by way of an expansion in the small c limit, similar to the approach developed previously in [17]. Using this approach we were able to show that the correlations are generally negative ('repulsive') for distributions in the Weibull class, and positive ('attractive') for distributions in the Fréchet class. In the

Gumbel class the sign of the correlations depends on the stretching exponent β in (4.36), with the border between positive and negative correlations being given by the exponential case $\beta = 1$. In contrast to all other cases, for distributions with an exponential tail the correlations are weak and independent of N, which is consistent with the fact that this class of distributions also contains the Gumbel distribution, which has no correlations at all. Simulation results illustrating the different cases are summarized in Fig. 4.8. A special role of the exponential distribution in separating two regimes of qualitatively different behaviors has been noted previously in the related context of near-extreme events [30].

Perhaps the most surprising and counterintuitive outcome of our work is the discovery of strong positive record correlations for distributions with a power law tail[4]. In view of the substantial interest in detecting and explaining heavy-tailed distributions in all areas of science [29, 32], our finding suggests that drift-induced record correlations could be used as a distribution-free test for detecting power laws or streched exponentials in empirical data [33].

An interesting open question concerns the structure of the record correlations in the asymptotic limit $N \to \infty$, where the record rate approaches a nonzero constant and the record process thus becomes stationary. Based on the work of Ballerini and Resnick [24], we have argued in Section 4.2.3 that the limit (4.27) exists for $c > 0$ whenever the underlying distribution has a finite mean, whereas for $c < 0$ it is possible that $l_{N,N-1}(c)$ diverges for $N \to \infty$. The fact that the coefficient of the leading order term in the small c expansion generally diverges with N indicates that the limiting function $l^*(c)$ may be singular for $c \to 0$, and we have presented numerical evidence for the occurrence of a discontinuity at $c = 0$ for the Weibull class. For heavy-tailed distributions in the Fréchet class, the large-N asymptotics is difficult to ascertain numerically because of the slow convergence, but we have argued that in this case $l^*(c)$ may exist also for $c < 0$. Rigorous work addressing these questions along the lines of [24] would be most welcome.

Acknowledgments

This work has been supported by DFG within the *Bonn Cologne Graduate School of Physics and Astronomy*, by Friedrich Ebert Stiftung through a fellowship to GW, and by Studienstiftung des Deutschen Volkes through a fellowship to JF.

Bibliography

[1] D. Gembris, J. G. Taylor, and D. Suter, Nature **417**, 506 (2002).
[2] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).
[3] G. W. Bassett, Climatic Change **21**, 303 (1992).
[4] R. E. Benestad, Climate Res. **25**, 3 (2003).
[5] S. Redner and M. R. Petersen, Phys. Rev. E **74**, 061114 (2006).
[6] G. A. Meehl, C. Tebaldi, G. Walton, and L. McDaniel, Geophys. Res. Lett. **36**, L23701 (2009).
[7] G. Wergen and J. Krug, EPL **92**, 30008 (2010).
[8] W. I. Newman, B. D. Malamud, and D. L. Turcotte, Phys. Rev. E **82**, 066111 (2010).
[9] A. Anderson and A. Kostinski, J. Appl. Meteor. Climat. **49**, 1681 (2010).
[10] L. De Haan and A. Ferreira, *Extreme Value Theory* (New York: Springer, 2006).
[11] P. Sibani, M. Brandt, and P. Alstrø, Int. J. Mod. Phys. B **12**, 361 (1998).
[12] J. Krug and K. Jain, Physica A **358**, 1 (2005).
[13] S.-C. Park and J. Krug, J. Stat. Mech.: Theor. Exp. **P04014** (2008).
[14] L. P. Oliveira *et al.*, Phys. Rev. B **71**, 104526 (2005).

---

[4] We note in this context that in a previous study of records drawn from series of independent RV's with an increasing variance only negative correlations were found [26].

[15] P. Sibani, G. F. Rodriguez, and G. G. Kenning, Phys. Rev. B **74**, 224407 (2006).
[16] G. Wergen, M. Bogner, and J. Krug, Phys. Rev. B **83**, 051109 (2011).
[17] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).
[18] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *Records* (New York: Wiley, 1998).
[19] V. B. Nevzorov, *Records: Mathematical Theory* (Providence, RI: American Mathematical Society, 2001).
[20] N. Glick, Amer. Math. Mon. **85**, 2 (1978).
[21] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).
[22] P. Le Doussal and K. J. Wiese, Phys. Rev. E **79**, 051105 (2009).
[23] S. Sabhapandit, EPL **94**, 20003 (2011).
[24] R. Ballerini and S. Resnick, J. Appl. Probab. **22**, 487 (1985).
[25] R. Ballerini, Stat. Prob. Lett. **5**, 83 (1987).
[26] J. Krug, J. Stat. Mech.: Theor. Exp. **P08017** (2007).
[27] E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications*, Vol. 33 (U.S. Government Printing Office, 1954, 1954).
[28] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics* (Malabar: Krieger, 1987).
[29] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools* (Berlin: Springer, 2000).
[30] S. Sabhapandit and S. N. Majumdar, Phys. Rev. Lett. **98**, 140201 (2007).
[31] J. Pickands III, Ann. Stat. **3**, 119 (1975).
[32] K. Christensen and N. R. Moloney, *Complexity and Criticality* (Imperial College Press, London, 2005).
[33] J. Franke, G. Wergen, and J. Krug, Phys. Rev. Lett. **108**, 064101 (2012).

Chapter 5

Correlations of record events as a test for heavy-tailed distributions

Jasper Franke, Gregor Wergen and Joachim Krug

Institute for Theoretical Physics, University of Cologne

Published in: Phys. Rev. Lett. 108, 064101 (2012)

Abstract: A record is an entry in a time series that is larger or smaller than all previous entries. If the time series consists of independent, identically distributed random variables with a superimposed linear trend, record events are positively (negatively) correlated when the tail of the distribution is heavier (lighter) than exponential. Here we use these correlations to detect heavy-tailed behavior in small sets of independent random variables. The method consists of converting random subsets of the data into time series with a tunable linear drift and computing the resulting record correlations.

5.1 Introduction

Determining the probability distribution underlying a given data set or at least its behavior for large argument is of pivotal importance for predicting the behavior of the system: If the data is drawn from a distribution with heavy tails, one needs to prepare for large events. Of particular relevance is the case when the probability density displays a power law decay, as this implies a drastic enhancement of the probability of extreme events. This is one of the reasons for the persistent interest in the observation and modeling of power law distributions, which have been associated with critical, scale-invariant behavior [1, 2] in diverse contexts ranging from complex networks [3] to paleontology [4], foraging behavior of animals [5, 6], citation distributions [7] and many more [8].

However, when trying to infer the tail behavior of the underlying distribution from a finite data set, one faces the problem that the number of entries of large absolute value is very small. This implies that even though binning the entries by magnitude and plotting them would yield an approximate representation of the probability density, this process becomes inconclusive in particular in the tail of the probability density. Furthermore, in small data sets, extreme outliers can strongly affect the results of methods like maximum likelihood estimators such that leaving out even one of these extreme and possibly spurious data points renders the outcome of the test insignificant. A case in point is the problem of estimating the distribution of fitness effects of beneficial mutations in evolution experiments, which are expected on theoretical ground to conform to one of the universality classes of extreme value theory (EVT) [9]. Because beneficial mutations are rare, the corresponding data sets are typically limited to a few dozen values, and the determination of the tail behavior can be very challenging [10, 11].

In this Letter we present a method for detecting heavy tails in empirical data that works reliably for small data sets (on the order of a few dozen entries) and is robust with respect to removal of extreme entries. The test is based on the statistics of records of subsamples of the data set. Similar to conventional record-based statistical tests [12–14], and in contrast to the bulk of methods available in this field [8], our approach is non-parametric and, hence, does not require any hypothesis about the underlying distribution. Rather than aiming at reliable estimates of the parameters of the distribution (such as the power law exponent), the main purpose of our method is to distinguish between distributions that are heavy-tailed and those that are not.

5.2 Record statistics and record correlations

Given a time series $\{x_1, \ldots, x_N\}$ of random variables (RVs), the n^{th} RV is said to be a record if it exceeds all previous RVs $\{x_j\}_{j<n}$ [13, 15]. For independent, identically distributed (i.i.d.) RVs, it is straightforward to see that the probability p_n for the n^{th} entry to be a record is simply $p_n = 1/n$, because any of the n RVs is equally likely to be the largest. Furthermore, record events are stochastically independent in this case [13, 15] and hence the *joint* probability $p_{n,n-1}$ that both x_{n-1} *and* x_n are records factorizes to $p_{n,n-1} = p_n p_{n-1}$

In a recent surge of interest [16–20], record statistics has been explored beyond the classical situation of i.i.d. RVs, and it has been found that the stochastic independence of record events is largely restricted to the i.i.d. case. In particular, for time series constructed from the linear drift model [19, 21]

$$x_n = cn + \eta_n, \tag{5.1}$$

where $c > 0$ is a constant and $\{\eta_n\}$ a family of i.i.d. RVs with distribution $F(\eta)$ and density $f(\eta)$, correlations between record events were quantified by considering the ratio [22]

$$l_{n,n-1}(c) = \frac{p_{n,n-1}(c)}{p_n(c)p_{n-1}(c)}. \tag{5.2}$$

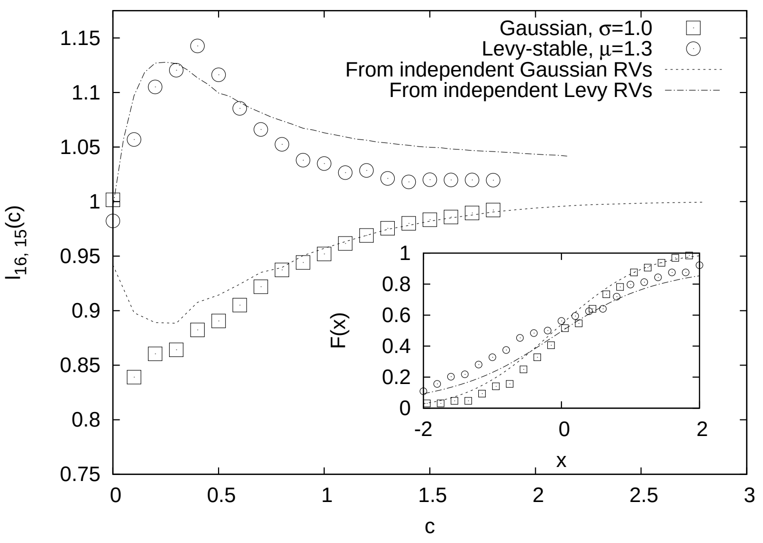

Figure 5.1: A first example of the proposed test, with $N = 64$ i.i.d. RVs drawn from a Gaussian with unit variance (squares) and a symmetric Lévy distribution $L_\mu(x)$ with $\mu = 1.3$ (circles). **Inset:** Comparing the cumulative distribution function $F(x)$ (lines) to its empirical estimate from the 64 data points shows that one distribution is broader than the other but does not allow for a clear distinction between the two data sets. **Main plot:** This difference is however clearly seen under application of the record-based test for subsamples of size $n = 16$. Dotted and dashed-dotted lines show the prediction for $l_{16,15}(c)$ for independent RV's.

For stochastically independent record events, $l_{n,n-1}(c) = 1$ and any positive (negative) deviation from unity can be interpreted as the a sign of attraction (repulsion) between record events. In [22] both cases were found depending on the distribution $F(\eta)$. Specifically, an expansion to first order in c yields $l_{n,n-1}(c) = 1 + cJ(n) + \mathcal{O}(c^2)$ with $J(n) \approx -\frac{1}{2}n^4(I(n) - I(n-1)) - n^3 I(n)$ where

$$I(n) = \int d\eta f^2(\eta) F^n(\eta) \tag{5.3}$$

and clearly $I(n) - I(n-1) < 0$. Thus for large n, there are two competing contributions to $J(n)$ determining the sign of the correlations.

To classify the behavior of the correlations in terms of the EVT classes [1, 23], consider the generalized Pareto distribution [24] $f(\eta) = (1+\kappa\eta)^{-(\kappa+1)/\kappa}$, which reproduces the three classes as $\kappa < 0$ (Weibull), $\kappa > 0$ (Fréchet) and $\kappa = 0$ (Gumbel), respectively. Computing $I(n)$ separately for these three cases [19] it was shown that, up to multiplicative terms in $\log(n)$ or slower, one has $I(n) \sim n^{-(2+\kappa)}$ and therefore [22] $J(n) \approx \frac{\kappa}{2} n^3 I(n)$, showing that the sign of correlations is directly determined by the extreme value index κ [25].

In the Gumbel class ($\kappa = 0$) more refined calculations for the generalized Gaussian densities $f_\beta(x) \sim \exp(-|\eta|^\beta)$ show that correlations are negative for $\beta > 1$ and positive for $\beta < 1$ [22]. The marginal case of a pure exponential distribution also shows positive correlations, but they can be distinguished from the $\beta < 1$ case in magnitude and, more clearly, in their n dependence: While for $\beta < 1$, correlations grow with n up to a limiting value, for $\beta = 1$ they are independent of n. The special, marginal role of the exponential distribution was also encountered in a study of near-extreme events [26], where the integral (5.3) appears in a different context.

To sum up, correlations between record events in time series with a linear drift allow a clear distinction between underlying probability densities that decay like an exponential or faster for large argument, and densities with heavier tails, by looking for positive correlations that grow in n. Using these two criteria, we now present a distribution-free test for heavy tails in data sets of i.i.d. random variables.

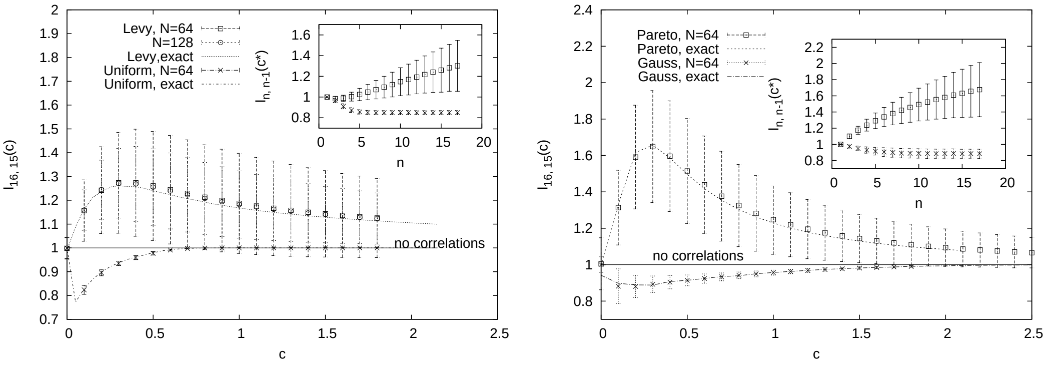

Figure 5.2: Sample-to-sample fluctuations of the HTI $\hat{l}_{n,n-1}$ for different distributions. Lines show mean values of the correlation $l_{16,15}(c)$ obtained from simulations of independent RV's (labeled *exact*), symbols show the mean value of the HTI and error bars indicate the standard deviation of the fluctuations for the symmetric Lévy-stable distribution with tail-parameter $\mu = 1.3$ and uniform distribution on $(0, 1)$ (top), and the Pareto-distribution with $\mu = 2.0$ and standard normal distribution (bottom). The HTI was obtained from simulations with internal statistics $s = 10^4$ (Pareto) or $s = 10^5$ (all other) and averaged over $S = 10^3$ independent data sets. Insets show how the correlations at the value $c^* = 0.25$ where correlations deviate maximally from unity grow as function of n while keeping N fixed.

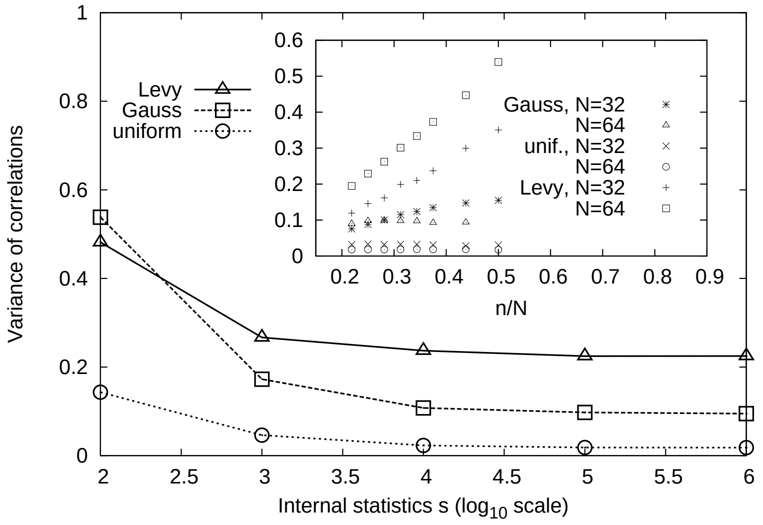

Figure 5.3: Magnitude of sample-to-sample fluctuations for three of the cases considered in Fig.5.2. With increasing internal statistics s, the sample-to-sample fluctuations decrease to a limiting value (main plot). This limiting value increases with n/N (inset), indicating that best results in terms of fluctuations are obtained by considering short subsequences. In the main plot $N = 64$ and $n = 16$.

5.3 Description of the test

Consider a data set with N entries, $x_1, x_2, \ldots, x_N$ that can reasonably be argued to consist of independent samples from the same distribution [27]. Then for each $n < N$, one can pick uniformly at random a subset of n entries and add a linear trend according to the index in the subset (see Eq.(5.1)), thus forming a set of random variables with linear trend. For each n, there are $\binom{N}{n}$ possible subsets [28], which can be used to compute the fraction of times the n^{th} entry is a record $\hat{p}_n(c)$, the corresponding fraction $\hat{p}_{n-1}(c)$ for the $n - 1^{th}$ entry, and the fraction $\hat{p}_{n,n-1}(c)$ of times both entries are records, for each value of a suitably chosen range of c [29]. The number s of subsets used for each value of c will be referred to as 'internal statistics'. Finally, one obtains an estimate for the correlations

$$\hat{l}_{n,n-1}(c) = \frac{\hat{p}_n(c)\hat{p}_{n-1}(c)}{\hat{p}_{n,n-1}(c)}, \tag{5.4}$$

where the hat serves to indicate that we are dealing with one fixed times series of length N and its sub-series, rather than many independent realizations. In the following we refer to $\hat{l}_{n,n-1}(c)$ as the *heavy tail indicator* (HTI).

To see how the test works in practice, consider Fig. 5.1. Two data sets of size $N = 64$ each are presented, one drawn from a standard Gaussian distribution, the other from a symmetric Lévy stable distribution with parameter $\mu = 1.3$. A standard approach to inferring the shape of the distribution is to estimate the cumulative distribution function by rank ordering the data along the x-axis (inset). In the example this shows that one distribution is broader than the other, but does not allow to distinguish between a difference in scale (as for two Gaussians of different standard deviation) and a difference in shape. In contrast, the two data sets come apart quite clearly under application of the test, showing that $\hat{l}_{n,n-1}(c) > 1$ for the Lévy distribution and $\hat{l}_{n,n-1}(c) < 1$ for the Gaussian (main figure).

5.4 Fluctuations

The lines in the main part of Fig. 5.1 show the predicted correlation $l_{n,n-1}(c)$ obtained from simulations of independent RV's. The estimated HTI $\hat{l}_{n,n-1}(c)$ obtained from subsamples of the two finite data sets deviates from these predictions, reflecting the fact that the ensemble of subsamples is *not* independent. The deviations depend on the data set in a random way, compare to Fig. 5.4, and understanding how the magnitude of the deviations depends on the test parameters N, n and s is clearly important for a quantitative assessment of the significance of the test. Figure 5.2 explores these sample-to-sample fluctuations by computing $\hat{l}_{n,n-1}(c)$ for a large number S ('external statistics') of different data sets and recording the mean and the mean squared deviation for different distributions. The fluctuations are large for power law distributions and decrease significantly for representatives of the Gumbel and Weibull classes. The latter implies that it is very unlikely for positive correlations to be produced by chance if the underlying distribution is *not* of heavy tail type; the observation of a HTI exceeding unity can therefore be taken as a strong indication of heavy tailed behavior in the data.

The effect of the internal statistics on the sample-to-sample fluctuations is quantified in Fig.5.3, where their magnitude can be seen to saturate to a limiting value with increasing s. Furthermore the limiting value depends on the ratio n/N: The smaller a subset of the initial data set is used, the more precise the results can be made by using large internal statistics. This behavior underlines a particular strength of our approach, namely that the combinatorially large number of subsequences can be used (up to a point) to reduce fluctuations due to the finite size of the data set. On the other hand, n should not be chosen too small, as the amplitude of correlations generally increases with n [22] (see inset of Fig.5.2). For the examples presented here, we found $n/N = 1/4$ at $N = 64$ to yield the best compromise between these two contradicting requirements, see also Fig.5.3.

5.5 Application

As an application of our approach, we consider the ISI citation data set first analyzed by Redner [7], consisting of citation data for 783339 papers published in 1981 and cited between 1981 and June 1997. Due to the large size of this data set, the existence of a power

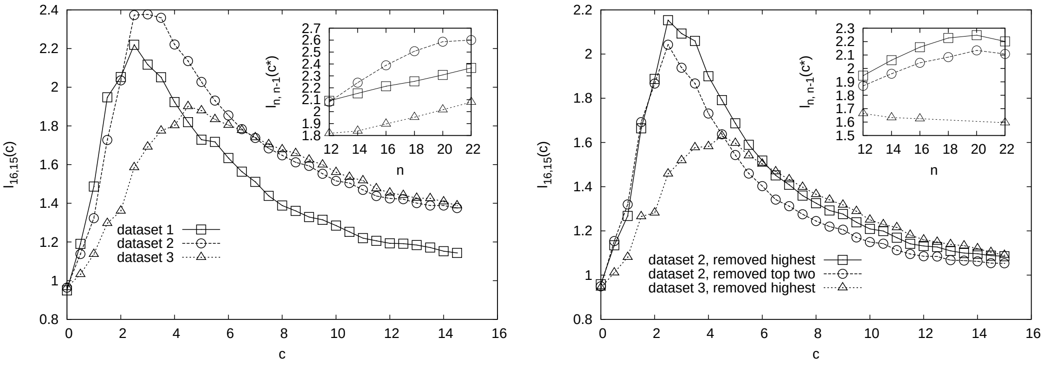

Figure 5.4: **Top:** Three randomly chosen subsets of length $N = 64$ each from the ISI citation data set [7]. The HTI was computed with internal statistics $s = 10^6$ and $n = 16$. The main plot shows attractive correlations in all three cases, the inset verifies growth of these correlations with n. **Bottom:** Removing the largest and even the top two entries of data set 2 does not change the result of the test. In data set 3, which is a somewhat extreme case in that the largest value is more than a factor 10 greater than the second largest, the correlations remain attractive upon removal of the largest entry but the magnitude of correlations no longer increases with n.

law tail with exponent $\mu \approx 2$ is well established [7, 8, 30]. Using our record-based approach, the heavy-tailed property could be recovered by considering small, randomly chosen subsets of only $N = 64$ papers each (Fig. 5.4). Despite the substantial fluctuations between the three subsets, the HTI lies clearly above unity in all cases. The small size of the chosen subsets implies that only a few (if any) data points in the subsets come from the extreme tails of the distribution. The lower panel in Fig.5.4 illustrates the robustness of the test with respect to the removal of putative outliers.

5.6 Summary

In conclusion, in this Letter we propose a record-based distribution-free test for heavy tails that works particularly well for small data sets. It was shown that the test is very versatile and quite robust to the removal of outliers, thus complementing standard methods like maximum likelihood estimates [8]. While record statistics has a long history of yielding distribution free tests [12–14], our approach is conceptually novel in that we make systematic use of the combinatorial proliferation of subsets of the original data set, which are then manipulated by adding a linear drift. We expect our method to be particularly useful in situations where the size of the data set is intrinsically limited, as in the assignement of an EVT universality class to the distribution of beneficial mutations in population genetics [10, 11]. In particular, the test can be used to strengthen the evidence in favor of heavy-tailed behavior in situations where conventional parametric tests have insufficient statistical power. By combining our test with standard approaches such as the maximum likelihood method, the tail parameters can then also be estimated.

Acknowledgments

The authors would like to thank Sid Redner for supplying the ISI data and Ivan Szendro for useful discussions. We acknowledge financial support from Studienstifung des deutschen Volkes (JF), Friedrich-Ebert-Stiftung (GW) and DFG within BCGS and SFB 680.

Bibliography

[1] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools* (Berlin: Springer, 2000).

[2] K. Christensen and N. R. Moloney, *Complexity and Criticality* (Imperial College Press, London, 2005).

[3] A.-L. Barábasi and R. Albert, Science **286**, 509 (1999).

[4] M. E. J. Newman and P. Sibani, Proc. R. Soc. Lond. B **266**, 1593 (1999).

[5] G. M. Viswanathan *et al.*, Nature **381**, 413 (1996).

[6] A. M. Edwards *et al.*, Nature **449**, 1044 (2007).

[7] S. Redner, Eur. Phys. Jour. B **4**, 131 (1998).

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review **51**, 661 (2009).

[9] P. Joyce, D. R. Rokyta, C. J. Beisel, and H. A. Orr, Genetics **180**, 1627 (2008).

[10] D. R. Rokyta *et al.*, J. Mol. Evol. **67**, 368 (2008).

[11] C. R. Miller, P. Joyce, and H. A. Wichman, Genetics **187**, 185 (2011).

[12] F. G. Foster and A. Stuart, J. Roy. Stat. Soc. B **16**, 1 (1954).

[13] N. Glick, Amer. Math. Mon. **85**, 2 (1978).

[14] S. Gulat and W. J. Padgett, *Parametric and Nonparametric Inference from Record-Breaking Data* (Springer, New York, 2003).

[15] V. B. Nevzorov, *Records: Mathematical Theory* (Providence, RI: American Mathematical Society, 2001).

[16] J. Krug, J. Stat. Mech.: Theor. Exp. **P08017** (2007).

[17] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).

[18] I. Eilazar and J. Klafter, Phys. Rev. E **80**, 061117 (2009).

[19] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).

[20] S. Sabhapandit, EPL **94**, 20003 (2011).

[21] R. Ballerini and S. Resnick, J. Appl. Probab. **22**, 487 (1985).

[22] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011).

[23] L. De Haan and A. Ferreira, *Extreme Value Theory* (New York: Springer, 2006).

[24] J. Pickands III, Ann. Stat. **3**, 119 (1975).

[25] *For $\kappa > 1$, corresponding to distributions without a mean, the correction term $J(n)$ is a decreasing function of n and the correlations vanish asymptotically for large n. However for finite n there are positive correlations of substantial magnitude, and the proposed test is still applicable.* ().

[26] S. Sabhapandit and S. N. Majumdar, Phys. Rev. Lett. **98**, 140201 (2007).

[27] *The i.i.d. assumption can be checked using a conventional record-based test, see [13]. Numerical simulations show that weak correlations in the data (e.g., generated by a first order autoregressive process) do not invalidate our method.* ().

[28] *In addition permutations of the subsets can be considered, which are not equivalent in the presence of drift.* ().

[29] *The range of c should be adjusted according to the typical spacing between entries of the time series.* ().

[30] *But note that the power law is limited to the extreme tails of the distribution containing less than 500 papers. In the intermediate regime, the data are better represented by a stretched exponential [7].* ().

Chapter 6

Rounding effects in record statistics

Gregor Wergen[1], **Daniel Volovnik**[2], **Sidney Redner**[2]
and Joachim Krug[1]

[1] *Institute for Theoretical Physics, University of Cologne*
[2] *Center for Polymer Studies and Department of Physics,*
Boston University, USA

Published in: Phys. Rev. Lett. 109, 164012 (2012)

Abstract: We analyze record-breaking events in time series of continuous random variables that are subsequently discretized by rounding to integer multiples of a discretization scale $\Delta > 0$. Rounding leads to ties of an existing record, thereby reducing the number of new records. For an infinite number of random variables that are drawn from distributions with a finite upper limit, the number of discrete records is finite, while for distributions with a thinner than exponential upper tail, fewer discrete records arise compared to continuous variables. In the latter case the record sequence becomes highly regular at long times.

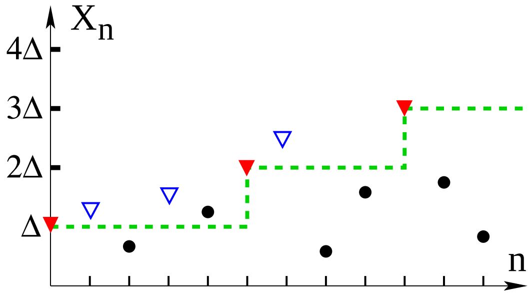

Figure 6.1: (color online) Effect of rounding down records with discretization unit Δ. Inverted triangles indicate records, with those that survive after rounding shown solid. The dashed line shows the evolution of the rounded record value.

6.1 Introduction

The statistics of record-breaking events have been widely studied in many contexts, including sports [1], evolutionary biology [2], the theory of spin glasses [3], and the possible role of global warming in the occurrence of record-breaking temperatures [4–9]. Records are defined as the entries in a time series of measurements that exceed all previous values. While the record statistics of independent, identically distributed (iid) random variables (RVs) that are drawn from continuous distributions are well understood [10, 11], the understanding of records drawn from time-dependent distributions [12–14] and from series of correlated RVs [15, 16] is still developing.

Here we address *discreteness effects* on record statistics. Conventionally, records are recorded from variables that are drawn from a continuous distribution. However, in all practical applications, technical limitations cause observations to be discrete, even if the underlying distribution is continuous. In sports or meteorology, distance, time, temperature, or precipitation measurements are always rounded to a certain accuracy [1, 6, 7], resulting in an effective discrete distribution of RVs. Thus ties of existing records can arise, which alters the probability for a record to occur in any given observation (Fig. 6.1).

For RVs that are explicitly drawn from discrete distributions, the effect of ties strongly affects the number of records [17–21]. For related δ-records and geometric records, where a new record arises only if the current observation exceeds the current record by a fixed constant δ [21, 22] or by a fixed fraction [23], intriguing statistical properties of records were found for the three universality classes of extreme value statistics (EVS) [24]. However, the consequences of measuring *rounded* record values that are drawn from continuous underlying distributions appears not to have been studied previously.

We consider a set of RVs $X_1, ...X_N$ and focus on the probability $P_n \equiv \text{Prob}(X_n > X_1, \ldots, X_{n-1})$ that the n^{th} variable in this series is a record. We denote P_n as the *record rate* and $R_n = \sum_{k=1}^{n} P_k$ as the *record number*. For continuous iid RVs, the universal result is $P_n = \frac{1}{n}$ (see, e.g., [10, 11]). Thus for $n \gg 1$, $R_n \approx \ln n + \gamma$, with $\gamma \approx 0.577...$ the Euler constant. We assume that. the RVs X_i are discretized in units of a minimal scale Δ. That is, each X_i gets rounded to a value of $X_i^\Delta = k\Delta$. We may consider (i) *rounding down*, with $k = \lfloor X_i/\Delta \rfloor$ and $\lfloor X \rfloor$ the floor function, which gives the largest integer smaller than X, or (ii) *rounding to the nearest lattice point*, with $k = \lfloor X_i/\Delta + \Delta/2 \rfloor$. Because asymptotic results do not depend on the rounding protocol, we will discuss only rounding down. We define the *strong* record rate

$$P_n^\Delta \equiv \text{Prob}\big(X_n^\Delta > X_1^\Delta, \ldots, X_{n-1}^\Delta\big), \tag{6.1}$$

in which ties caused by the discretization are *not counted* as new records. Thus not only X_n, but also the rounded value X_n^Δ has to be larger than all previous RVs for a new record to occur (Fig. 6.1).

6.2 General theory, asymptotic results

For iid RVs X_i drawn from a distribution with probability density $f(x)$ and cumulative distribution $F(x) = \int^x dy\, f(y)$, the record rate is obtained from $P_n = \int dx\, f(x) F^{n-1}(x)$ [11]. For any continuous density $f(x)$, this integral gives the universal behavior mentioned above, $P_n = \frac{1}{n}$. However, if the measurement X_i is rounded down to X_i^Δ, the integral for P_n breaks into the sum

$$\begin{aligned} P_n^\Delta &= \sum_k \left[\int\limits_{k\Delta}^{(k+1)\Delta} dx\, f(x) \right] F^{n-1}(k\Delta)\,, \\ &= \sum_k \big[F((k{+}1)\Delta) - F(k\Delta)\big] F^{n-1}(k\Delta)\,. \end{aligned} \tag{6.2}$$

This gives the strong record rate from continuous RVs that are rounded down to the closest integer multiple of Δ. We emphasize that in the practically more relevant case where record values are rounded either up or down to the closest integer multiple of Δ, the record rate has the same statistical properties as those from only rounding down. We now give asymptotic results for P_n^Δ for the three basic classes of EVS [24]: Weibull (distributions with a finite upper limit), Gumbel (unbounded upper tail decaying faster than any power law), and Fréchet (power-law upper tail). Our asymptotic approximations for the discrete record rate P_n^Δ for these classes of EVS agree well with numerical results.

Weibull class: For illustration, we start with the uniform distribution: $f(x) = 1$ for $x \in [0,1]$ and 0 otherwise. For discretization scale $\Delta = \frac{1}{L}$, with integer-valued $L > 1$, Eq. (6.2) reduces to:

$$P_n^\Delta = \sum_{k=1}^{\frac{1}{\Delta}-1} \Delta\, (k\Delta)^{n-1} = \Delta^n H_{\frac{1}{\Delta}-1,n-1}, \tag{6.3}$$

where $H_{m,n}$ is the m^{th} harmonic number of power n. At some point in the time series of RVs, a record with a rounded value $1-\Delta$ occurs; this is necessarily the *last record.* For a fine discretization scale, $\Delta \ll 1$, the sum in (6.3) can be replaced by an integral to give $P_n^\Delta \approx \frac{1}{n}(1-\Delta)^n$. Thus for any $\Delta > 0$, P_n^Δ no longer decays as $\frac{1}{n}$, but instead approaches zero exponentially with n — rounding strongly depresses the asymptotic record rate for the uniform distribution.

A more general example of the Weibull EVS class is $f(x) = \xi(1-x)^{\xi-1}$, with $\xi > 0$ and $x \in [0,1]$. By expanding Eq. (6.2) to second order for $\Delta \ll 1$, we find

$$\begin{aligned} P_n^\Delta &\approx \int\limits_1^{\frac{1}{\Delta}-1} dk \left[(1{-}k\Delta)^\xi - (1{-}(k{+}1)\Delta)^\xi\right] \\ &\qquad\qquad \times \left[1 - (1{-}k\Delta)^\xi\right]^{n-1}, \\ &\approx \begin{cases} \frac{1}{n}\left[1 - n\Delta^\xi - \frac{\Delta\xi}{2}\Gamma\big(2{-}\frac{1}{\xi}\big) n^{1/\xi}\right], & n\Delta^\xi \ll 1, \\ \frac{1}{n}\exp(-n\Delta^\xi), & n\Delta^\xi \gg 1. \end{cases} \end{aligned} \tag{6.4}$$

Since the underlying distribution has a bounded support, the total number of records is again finite. The results in (6.4) reproduce those found for the uniform distribution.

Gumbel class: As a basic example, we treat the exponential distribution $f(x) = e^{-x}$. For $n \gg 1$ we replace the sum in Eq. (6.2) by an integral and find

$$P_n^\Delta \approx \sum_{k=1}^{\infty} e^{-k\Delta}(1{-}e^{-k\Delta})^n \approx \frac{1}{n\Delta}(1{-}e^{-\Delta}) \tag{6.5}$$

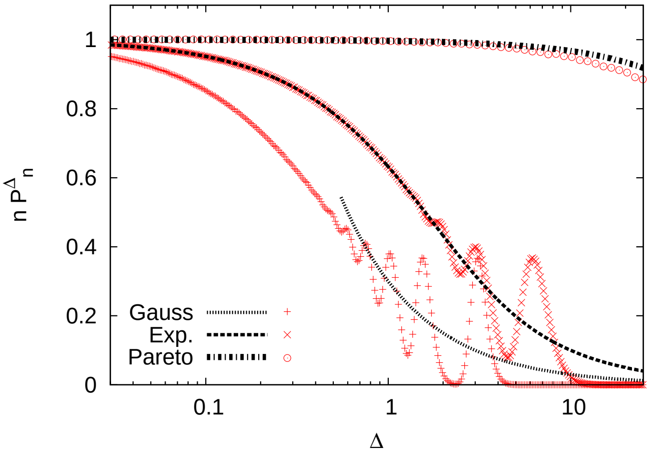

Figure 6.2: (color online) Scaled record rate nP_n^Δ for $n = 1000$ for the Gaussian, exponential, and Pareto (with $\mu = 1.2$) distributions. Without rounding, $P_n = \frac{1}{n}$. Simulations (symbols) are averaged over 10^6 time series and over $975 \leq n \leq 1025$ to smooth the data. Analytical predictions (curves) are shown for comparison. For the origin of the peaks for the Gaussian and exponential distributions, see the text following Eq. (6.14).

for arbitrary $\Delta \geq 0$, in agreement with findings for the geometric distribution in Ref. [18] and with our simulations (Fig. 6.2). For $\Delta \ll 1$, (6.5) reduces to $P_n^\Delta \approx \frac{1}{n}\left(1 - \frac{\Delta}{2}\right)$, while for $\Delta \gg 1$, $P_n^\Delta \approx \frac{1}{n\Delta}$. In contrast to the Weibull class, the P_n^Δ asymptotically decays as $\frac{1}{n}$ for arbitrary Δ.

For the Gaussian distribution $f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$, with unit standard deviation, we find that as $n \to \infty$

$$P_n^\Delta \approx \frac{1}{2}\int dx \left[\operatorname{erfc}\left(\frac{k\Delta}{\sqrt{2}}\right) - \operatorname{erfc}\left(\frac{(k+1)\Delta}{\sqrt{2}}\right)\right] F(x)^{n-1}\,,$$
$$\approx \frac{1}{\Delta}\int \mathrm{d}x\, \frac{1}{\sqrt{2\pi}}\,\frac{1}{x}\, e^{-x^2}\, F(x)^{n-1}. \tag{6.6}$$

For $n \to \infty$ we evaluate this integral by the Laplace method by expanding the integrand about $x^* = \ln(n^2/2\pi)$, where x^* is the mean value of the n^{th} record. After some calculation, we obtain

$$P_n^\Delta \approx \frac{1}{n\Delta}\left[\sqrt{\ln\left(\frac{n^2}{2\pi}\right)}\right]^{-1}. \tag{6.7}$$

Thus the record rate decays slightly faster than $\frac{1}{n}$ (Fig. 6.2). Correspondingly, $R_n^\Delta \propto \Delta^{-1}(\ln n)^{1/2}$, which diverges weakly as $n \to \infty$.

Fréchet class: A representative for this class is the Pareto distribution $f(x) = \mu x^{-\mu-1}$, with $x > 1$ and $\mu > 0$. Using again Eq. (6.2), the asymptotic record rate P_n^Δ is

$$P_n^\Delta \approx \tfrac{1}{n}\left[1 - \tfrac{\Delta}{2}\,\mu\,\Gamma\!\left(2 + \tfrac{1}{\mu}\right) n^{-1/\mu}\right]. \tag{6.8}$$

In contrast to the two previous classes, the effect of the rounding is negligible, as $P_n^\Delta \to P_n$ for $n \to \infty$ and arbitrary Δ (Fig. 6.2).

6.3 Small-Δ regime

We now focus on the effects of rounding when the discretization scale is small ($\Delta \ll 1$) for fixed n. Here we find a useful analogy between the effect of a linear drift in RVs [13] and the effect of rounding, and we adapt methods developed for the former problem to help elucidate rounding effects. For small Δ the general expression (6.2) for P_n^Δ simplifies to

$$\begin{aligned} P_n^\Delta &= \sum_k \left[\int\limits_{k\Delta}^{(k+1)\Delta} dx\, f(x) \right] F^{n-1}(k\Delta)\,, \\ &= \int dx\, f(x) F^{n-1}(\lfloor x \rfloor_\Delta)\,, \\ &\approx \tfrac{1}{n} - n \int dx\, (x - \lfloor x \rfloor_\Delta) f^2(x) F^{n-2}(x)\,. \end{aligned} \tag{6.9}$$

Here $\lfloor x \rfloor_\Delta$ is defined as the largest integer multiple of Δ that is smaller than x. Thus, in the second line, $k\Delta = \lfloor x \rfloor_\Delta$ for $k\Delta \le x < (k+1)\,\Delta$, which obviates writing the sum. In the last step, we expand to first order in the quantity $x - \lfloor x \rfloor_\Delta$ and employ the crude assumption that, on average, $x - \lfloor x \rfloor_\Delta \approx \frac{\Delta}{2}$ to give

$$P_n^\Delta \approx \tfrac{1}{n} \left(1 - \tfrac{\Delta}{2} n^2 \mathcal{I}_n\right), \tag{6.10}$$

where $\mathcal{I}_n \equiv \int dx\, f^2(x) F^{n-2}(x)$. The approximation underlying (6.10) is valid if $n^2 \Delta \mathcal{I}_n \ll 1$. The quantity $\mathcal{I}_n$ appears in record statistics that arise from continuous RVs with a linear drift [13], whose behavior is known for a wide range of distributions. In the following we use the results from [13] to determine P_n^Δ in the small-Δ regime.

Weibull and Fréchet classes: For the distribution $f(x) = \xi(1-x)^{\xi-1}$ introduced above, the approximation given by Eq. (6.10) is useful for $\xi > 1$ and we find, for $n\Delta^\xi \ll 1$,

$$P_n^\Delta \approx \tfrac{1}{n} \left[1 - \tfrac{\Delta \xi}{2} \Gamma\big(2 - \tfrac{1}{\xi}\big) n^{1/\xi}\right], \tag{6.11}$$

which, for $n\Delta^\xi \ll 1$ and $\xi > 1$, agrees with the result derived from our general approach in Eq. (6.4). Similarly, for the Pareto distribution we recover Eq. (6.8).

Gumbel class: For the exponential distribution, we find $P_n^\Delta \approx \frac{1}{n}\left(1 - \frac{\Delta}{2}\right)$, which agrees with the small-Δ behavior of Eq. (6.5). For the Gaussian distribution, the small-Δ approximation allows us to obtain a new expression for the record rate when $\sqrt{\ln n} \ll \Delta^{-1}$,

$$P_n^\Delta \approx \frac{1}{n} \left[1 - \frac{2\Delta\sqrt{\pi}}{e^2} \sqrt{\ln\left(\frac{n^2}{8\pi}\right)}\right]. \tag{6.12}$$

The regime $\sqrt{\ln n} \ll \Delta^{-1}$ is not accessible through the general approach and this range is particularly important for applications, such as in climatology [7]. For $n \gg 1$ and $\Delta \ll 1$, Eq. (6.12) reproduces the numerical simulation values for P_n^Δ very accurately (Fig. 6.3).

6.4 Large-Δ regime

For Gumbel-class distributions that decay at least exponentially fast near the upper limit, we can provide an alternative description for the record number R_n^Δ. For these distributions, it is known that the average spacings between the record events do not increase in time for large n [11]. Therefore, we may choose a sufficiently large value of Δ that almost all records are suppressed because of ties. It then follows that all discrete values $k\Delta$ (with $k \ge 0$) will eventually be record values and R_n^Δ is just the sum over the probabilities that a record has already occurred for a certain value $k\Delta$. The corresponding probabilities $\Pi_n(k)$ for record value $k\Delta$ are given by $\Pi_n(k) \approx 1 - F(k\Delta)^{n-1}$, which leads to

$$R_n^\Delta \approx \sum_{k=0} \Pi_n(k) \approx 1 + \sum_{k=1}^{\infty} \left[1 - F(k\Delta)^{n-1}\right]. \tag{6.13}$$

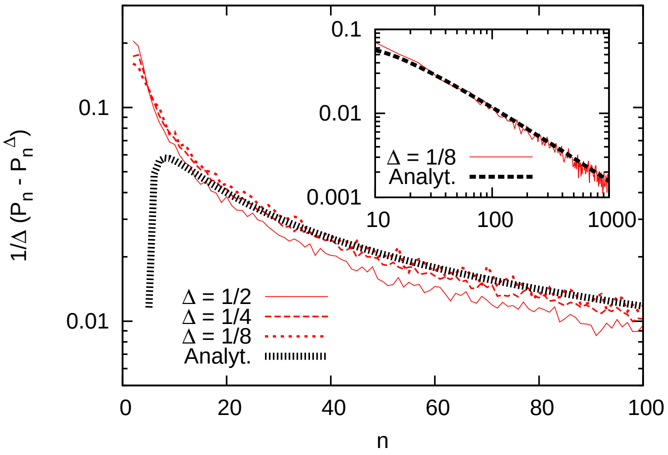

Figure 6.3: (color online) Simulations of P_n^Δ for Gaussian RVs in the regime $\sqrt{\ln n} \ll \frac{1}{\Delta}$. Thin curves are $\frac{1}{\Delta}\left(P_n - P_n^\Delta\right)$ for $\Delta = \frac{1}{2}, \frac{1}{4}$ and $\frac{1}{8}$ and $n \in [0, 100]$. For each Δ, 10^6 time series were simulated. The thick dashed curve depicts the analytical prediction Eq. (6.15). Inset shows the same analysis for $\Delta = \frac{1}{8}$ with $n \in [1, 1000]$.

For elementary Gumbel distributions, interesting properties emerge from $\Pi_n(k)$. For a small n and large $k\Delta$, it is obvious that $\Pi_n(k) \approx 0$. Conversely, for large n and arbitrary $k\Delta$ eventually $\Pi_n(k) \approx 1$, since $F(k\Delta) < 1$ for finite $k\Delta$.

We now estimate the regime where $\Pi_n(k)$ switches between 0 and 1; this condition also determines the point where the mean record number switches from $k-1$ to k. Since $\Pi_n(k)$ will never be exactly 0 or 1, we seek the time n, where $\Pi_n(k)$ is either smaller than ϵ ($n = n_-$) or larger than $1-\epsilon$ ($n = n_+$) for small $\epsilon \ll 1$. By elementary means we find

$$n_- < \frac{\ln \epsilon}{\ln\left[F(k\Delta)\right]}, \qquad n_+ > \frac{\epsilon}{-\ln\left[F(k\Delta)\right]}. \tag{6.14}$$

Evidently $\Pi_n(k)$ switches between 0 and 1 when n is between n_- and n_+, where n_- and n_+ are both proportional to $[\ln(F(k\Delta))]^{-1}$. For the exponential distribution, for example, we find that $n_- = \epsilon\, e^{k\Delta}$ and $n_+ = \ln(1/\epsilon)\, e^{k\Delta}$, so the k^{th} record will occur at a time proportional to $e^{k\Delta}$, leading to a mean record number of $R_n^\Delta \approx \frac{1}{\Delta} \ln n$. In the large $k\Delta$ regime, records occur in an ordered fashion and are well separated from each other. The $(k+1)^{\text{st}}$ record occurs at time $e^{(k+1)\Delta}$, which for $\Delta \gg 1$, is much later than the time of the k^{th} record. Thus the mean record number undergoes a step-like periodicity when plotted against e^n. For the Gaussian distribution, the same approach now predicts that $\Pi_n(k)$ switches for $n \approx \sqrt{2\pi} k\Delta\ e^{k^2\Delta^2/2}$ (Fig. 6.4). For large $k\Delta$ and large n, the mean record number becomes

$$R_n^\Delta \approx \sum_{k=0} \Pi_n(k) \approx \frac{1}{\Delta}\sqrt{\ln\left(\frac{n^2}{2\pi}\right)}, \tag{6.15}$$

which was already obtained with the general approach above and confirms the validity of the form for R_n^Δ given in Eq. (6.13). The step periodicity in R_n^Δ is the source of the observed peaks (Fig. 6.2) in the record rate P_n^Δ as a function of Δ for exponential and Gaussian distributions.

6.5 Conclusions

We determined how rounding down continuous random variables affects the statistics of records. Our results directly apply to the practical situation where continuous variables are

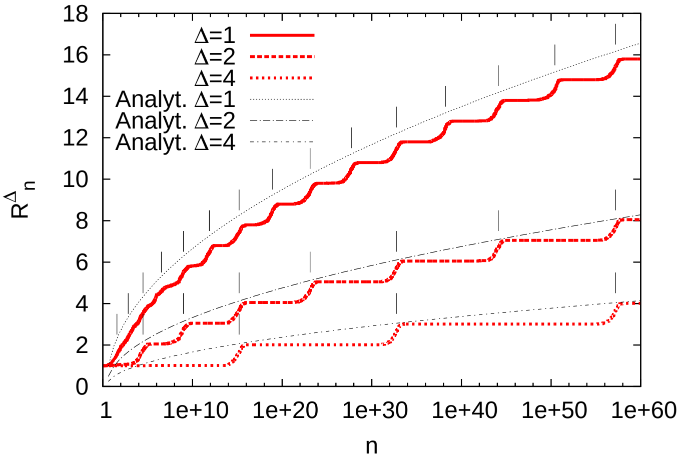

Figure 6.4: (color online) Record number R_n^Δ for Gaussian RVs for $\Delta = 1, 2, 4$. Data (red) are based on 100 realizations with a maximal $n = 10^{60}$. For $n > 10^6$ we used an algorithm that directly simulates record events by sampling both the distribution and the waiting time of the $(k + 1)^{\text{st}}$ record from the value of the k^{th} record. Thin lines (black) show the asymptotic behavior predicted by Eq. (6.15). The vertical lines show the steps predicted by $n \approx \sqrt{2\pi}\, k\, \Delta\, e^{(k\Delta)^2/2}$.

rounded either up or down to the closest integer multiple of a fixed discretization scale Δ.

For distributions with bounded support, rounding leads to an exponential decay of the record rate, P_n^Δ, and an asymptotically finite record number. In contrast, for power-law distributions, the effect of rounding becomes negligible for $n \to \infty$ and $P_n^\Delta \to \frac{1}{n}$ independent of Δ. In the intermediate Gumbel class, the behavior is more subtle. For the exponential distribution, P_n^Δ decays as $\frac{1}{n}$ with a Δ-dependent prefactor, while for the general distribution $f(x) \propto \exp(-|x|^\beta)$ with $\beta > 1$, the record rate decays as $n^{-1} \ln(n)^{1/\beta - 1}$.

For underlying distributions that decay at least exponentially, the record sequence becomes ordered at long times, in marked contrast to independent record events in for continuous iid RVs [10, 11]. While correlations between record events have been previously observed for RVs that are drawn from drifting [14] or broadening [12] distributions, the effect of rounding is much stronger and renders record events predictable on a time scale that grows exponentially (or faster) with record number.

To illustrate that rounding effects have an observationally significant influence on records, we analyzed 50 years of daily temperatures from 361 U.S. weather stations [25] along the lines of [7]. The measurements were reported in integer units of $\Delta = 1°$F and we considered all 361 times 365 time series for the individual calendar days with an average standard deviation of $\sigma \approx 8.9°$F. Only 75% of the weak upper (ties allowed) and 78% of the weak lower records were also strong records (no ties), in good agreement with the value of 79% predicted by our analytical result in Eq. (6.12). In this example the effect of ties on the record rate has a similar magnitude as that of the small warming trend in the data (cf. [5–7]). Thus rounding effects should be carefully accounted for if one wishes to use record statistics to detect secular trends in data, such as global warming.

Acknowledgements

GW acknowledges financial support from Friedrich-Ebert-Stiftung and BCGS as well as the kind hospitality of the Center for Polymer Studies in the early stages of this work. DV and SR thank NSF grant DMR-0906504 for partial financial support of this research.

Bibliography

[1] D. Gembris, J. G. Taylor, and D. Suter, Nature **417**, 506 (2002).

[2] J. Krug and K. Jain, Physica A **358**, 1 (2005).

[3] L. P. Oliveira, Phys. Rev. B **71**, 104526 (2005).

[4] G. W. Bassett, Climatic Change **21**, 303 (1992).

[5] S. Redner and M. R. Petersen, Phys. Rev. E **74**, 061114 (2006).

[6] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).

[7] G. Wergen and J. Krug, EPL **92**, 30008 (2010).

[8] W. I. Newman, B. D. Malamud, and D. L. Turcotte, Phys. Rev. E **82**, 066111 (2010).

[9] S. Rahmstorf and D. Coumou, Proc. Natl. Acad. Sci. USA **108**, 17905 (2011).

[10] N. Glick, Amer. Math. Mon. **85**, 2 (1978).

[11] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *Records* (New York: Wiley, 1998).

[12] J. Krug, J. Stat. Mech.: Theor. Exp. **P08017** (2007).

[13] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).

[14] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011).

[15] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).

[16] G. Wergen, M. Bogner, and J. Krug, Phys. Rev. B **83**, 051109 (2011).

[17] W. Vervaat, Stoch. Proc. Appl. **1**, 317 (1973).

[18] H. Prodinger, Discrete Math. **153**, 253 (1996).

[19] R. Gouet, F. J. Lopez, and G. Sanz, Adv. Appl. Prob. **37**, 781 (2005).

[20] E. S. Key, J. Theor. Probab. **18**, 99 (2005).

[21] R. Gouet, F. J. Lopez, and G. Sanz, Bernoulli **13**, 754 (2007).

[22] N. Balakrishnan, K. Balasubramanian, and N. Panchapakesan, J. Appl. Statist. Sci. **4**, 123 (1996).

[23] R. Gouet, F. J. Lopez, and G. Sanz, J. Stat. Mech.: Theor. Exp. **P01005** (2012).

[24] E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications*, Vol. 33 (U.S. Government Printing Office, 1954, 1954).

[25] C. Williams Jr. *et al.*, *Historical Climatology Network Daily Temperature, Precipitation, and Snow Data*, Tech. Rep. (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, 2006) oRNL/CDIAC-118, NDP-070.

[26] S. Sabhapandit, EPL **94**, 20003 (2011).

[27] V. B. Nevzorov, *Records: Mathematical Theory* (Providence, RI: American Mathematical Society, 2001).

[28] J. Franke, G. Wergen, and J. Krug, Phys. Rev. Lett. **108**, 064101 (2012).

[29] P. Sibani, G. F. Rodriguez, and G. G. Kenning, Phys. Rev. B **74**, 224407 (2006).

[30] L. De Haan and A. Ferreira, *Extreme Value Theory* (New York: Springer, 2006).

[31] I. Eliazar, Physica A **348**, 181 (2005).

[32] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).

[33] R. E. Benestad, Climate Res. **25**, 3 (2003).

Part II

Record-breaking temperatures

7 Record-breaking temperatures reveal a warming climate 81
7.1 Introduction . . . 82
7.2 Theory . . . 83
7.3 European data . . . 83
7.4 American data and discreteness effects . . . 88
7.5 Re-analysis data . . . 89
7.6 Conclusions . . . 89

8 Record occurrence and record values in daily & monthly temperatures 93
8.1 Introduction . . . 94
8.2 Theory: Record occurrence in the presence of linear drift . . . 96
  8.2.1 Record statistics of iid. RV's . . . 96
  8.2.2 Linear Drift . . . 96
8.3 General introduction of the data . . . 98
  8.3.1 US monthly station data . . . 99
  8.3.2 European daily data . . . 100
8.4 Occurrence of records in daily and monthly temperatures . . . 101
  8.4.1 US monthly station data . . . 101
  8.4.2 European daily re-analysis data . . . 102
8.5 Theory: Distributions of record values . . . 103
  8.5.1 Records in the n'th step . . . 103
  8.5.2 k'th records . . . 105
8.6 Distributions of record values in European temperature recordings . . . 105
  8.6.1 Mean values of records in a given year . . . 107
  8.6.2 Distribution of records in a certain year . . . 109
  8.6.3 Records with a given record number . . . 110
  8.6.4 Record values in the far north . . . 111
8.7 Discussion and conclusion . . . 111

Chapter 7

Record-breaking temperatures reveal a warming climate

Gregor Wergen and Joachim Krug

Institute for Theoretical Physics, University of Cologne

Published in: EPL 92 (2010) 30008

Abstract: We present a mathematical analysis of records drawn from independent random variables with a drifting mean. To leading order the change in the record rate is proportional to the ratio of the drift velocity to the standard deviation of the underlying distribution. We apply the theory to time series of daily temperatures for given calendar days, obtained from historical climate recordings of European and American weather stations as well as re-analysis data. We conclude that the change in the mean temperature has increased the rate of record breaking events in a moderate but significant way: For the European station data covering the time period 1976-2005, we find that about 5 of the 17 high temperature records observed on average in 2005 can be attributed to the warming climate.

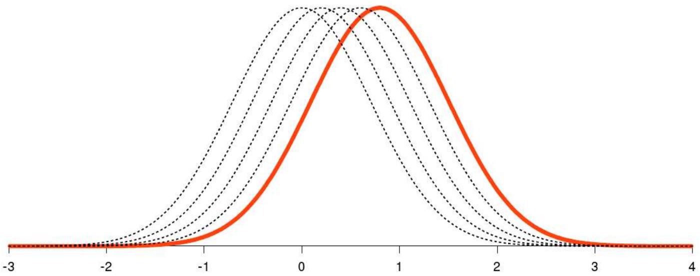

Figure 7.1: Schematic of the evolution of the daily temperature distribution under linear drift of the mean.

7.1 Introduction

In current media coverage the occurrence of record-breaking temperatures and other extreme weather conditions is often associated with global climate change. However, record breaking events occur at a certain rate in any stationary random process. In mathematical terms, a record is an entry in a time series that is larger (upper record) or smaller (lower record) than all previous entries [1–3]. If the entries are independent and identically distributed random variables drawn from a continuous probability distribution, the probability P_n to observe a new record after n steps, hereafter referred to as the *record rate*, is simply $P_n = 1/n$, because all n values are equally likely to be the largest. Applying this result to maximal temperatures measured at a specific calendar day over a time span of n years, it follows that the expected number of records per year is $365/n$, i.e. about 12 records for an observation period of 30 years. Remarkably, this prediction is entirely independent of the underlying probability distribution, which may even differ for different calendar days.

Despite considerable current interest in extreme climate events [4–14], the subject of climate records has received relatively little attention. It is intuitively obvious that an increase in the mean temperature will lead to an increased occurrence of high temperature records, but attempts to detect this effect in observational data have long remained inconclusive [15–18]. Only very recently an empirical study of temperature data from the US found a significant effect of warming on the *relative* occurrence of hot and cold records [19].

Here we present a detailed analysis of several large data sets of temperature measurements from both American and European weather stations, as well as re-analysis data[1]. We find that the observed increase in the number of high temperature records (and the corresponding decrease in the low records) is well described by a minimal model which assumes that the distribution of temperatures measured on a given calendar day is a Gaussian with constant standard deviation σ and a mean that increases linearly in time at rate v (see Fig.7.1). This model is consistent with previous findings [18, 20–23] and it is supported by our own analysis of the available data sets [24], see Fig.7.2 for an example. While changes in temperature *variability* have also been argued to be important in the generation of extreme temperature events [5, 7], we have failed to detect a clear systematic trend in σ in the data [Fig.7.2 (**b**)]. Moreover, the increase in the mean supersedes a possible effect on σ, in the sense that the former leads to an asymptotically constant record rate [25–28] whereas the latter only increases the record rate from $1/n$ to $(\ln n)/n$ [29]. For these reasons we restrict ourselves to the simplest setting of a temperature distribution of constant shape and linearly increasing mean. Although temperature fluctuations are well known to display long-term correlations [30, 31], the assumption that the daily temperatures are not correlated is justified because individual measurements in a time-series are always one year apart (see [18] and the quantitative discussion below).

---

[1] The re-analysis approach combines meteorological observations from a variety of sources with advanced data assimilation techniques in order to create a continuous stream of observables on a three-dimensional grid, see [20] for details.

7.2 Theory

We begin by deriving an approximate analytic expression for the increase in the record rate P_n caused by a linear drift of the mean. In general, the record rate for a sequence of independent but not identically distributed random variables x_n is given by [29]

$$P_n = \int_{-\infty}^{\infty} f_n(x)dx \prod_{k=1}^{n-1} \left(\int_{-\infty}^{x} dx_k f_k(x_k) \right) \tag{7.1}$$

where $f_n(x)$ denotes the probability density at time step n. Here we consider a drifting distribution of constant shape, which implies $f_n(x) = f(x - vn)$ with a common density $f(x)$. This reduces (7.1) to

$$P_n = \int_{-\infty}^{\infty} f(x)dx \prod_{k=1}^{n-1} \left(\int_{-\infty}^{x+vk} dx_k f(x_k) \right). \tag{7.2}$$

An explicit evaluation of (7.2) is possible for special choices of $f(x)$, but in general it is only known that P_n converges to a nonzero limit $P^* = \lim_{n\to\infty} P_n$ when $v > 0$ [25–28]. In the climate context the drift speed is expected to be small compared to the standard deviation of the distribution. We therefore expand (7.1) to linear order in v, which yields

$$P_n \approx \frac{1}{n} + \frac{vn\,(n-1)}{2} \int_{-\infty}^{\infty} dx f^2\,(x)\, F^{n-2}\,(x) \tag{7.3}$$

where $F\,(x)$ is the cumulative distribution function of $f\,(x)$. In [28] the integral in the second term is evaluated for various elementary distributions. For distributions with a power law tail one finds that the correction term decreases for large n. On the other hand, for distributions that decay faster than exponential, the correction term generally increases with n. In the Gaussian case of interest here the integral can be evaluated in closed form only for $n = 2$ and 3, with the result

$$P_2 \approx \frac{1}{2} + \frac{v}{2\sqrt{\pi}\sigma}, \quad P_3 \approx \frac{1}{3} + \frac{3v}{4\sqrt{\pi}\sigma}. \tag{7.4}$$

Using a saddle point approximation and the properties of the Lambert W-functions [32] to extract the behavior for large n one arrives at the asymptotic expression [28]

$$P_n \approx \frac{1}{n} + \frac{v}{\sigma}\frac{2\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)}, \tag{7.5}$$

which is accurate for $n \geq 7$. For $n = 4, 5, 6$ the integral can be evaluated numerically. For a typical value of $v/\sigma \approx 0.01$ and a time span of 30 years, (7.5) implies an increase of the record rate from $1/30 \approx 0.033$ to 0.042, or an increase in the expected number of record events per year from 12 to 15. In the following this prediction will be compared to empirical temperature data.

7.3 European data

The most comprehensive analysis was carried out for temperature data obtained within the ECAD project, which comprises a total of 752 European stations [22]. The data consist of daily recordings for the minimum, mean and maximum temperature, as well as precipitation and snowfall. These stations recorded over time-spans of varying length between less than

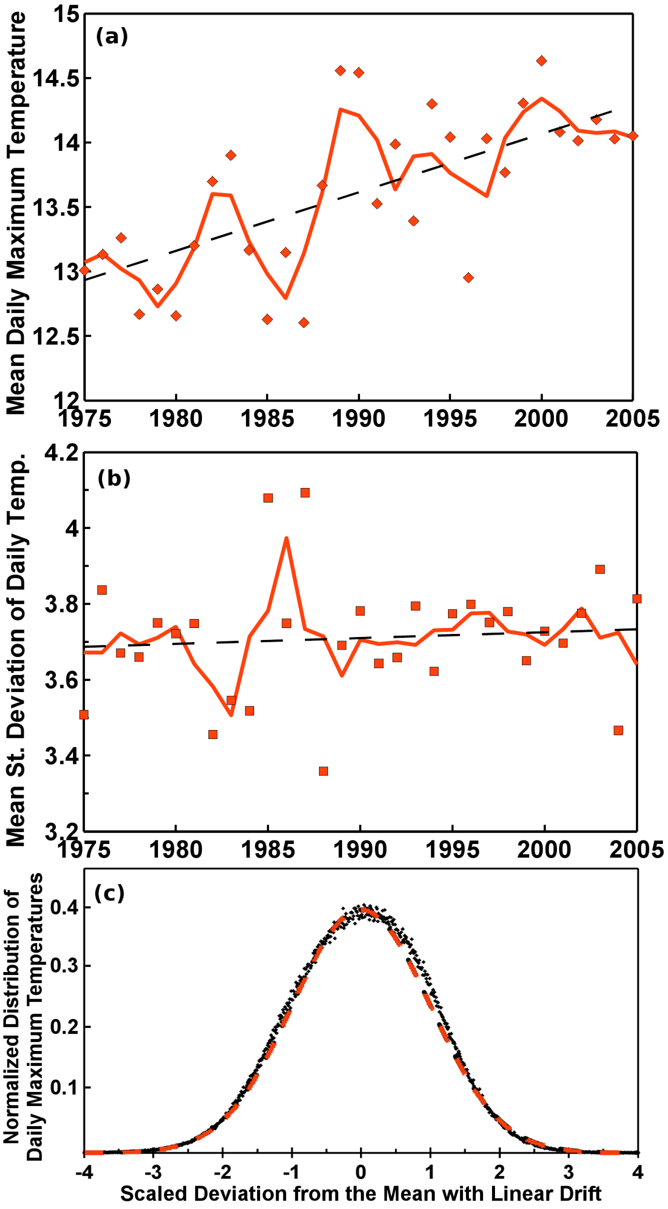

Figure 7.2: This figure summarizes the behavior of the distribution of daily maximum temperatures for data set EII. **(a)** Mean daily maximum temperature. Daily maximum temperatures were averaged over all stations and the entire calendar year. Diamonds show the average of dTmax for individual years and the full line is a sliding 3-year average. The regression line (dashed) shows a clear increase over the last 30 years. **(b)** Standard deviation of daily maximal temperatures. To estimate the standard deviation, first a linear regression of dTmax was carried out for each station and each calendar day. The standard deviation for a given year was then computed by averaging the squared deviation of dTmax from the linear fit over all stations and all calendar years. Full line is a 3-year average and the dashed line the result of a linear regression. We find no systematic trend in the standard deviation. **(c)** Distribution of daily temperatures on individual calendar days. The measurements were detrended and normalized for all time series individually and then accumulated. The dashed line is the probability density of a standard normal distribution.

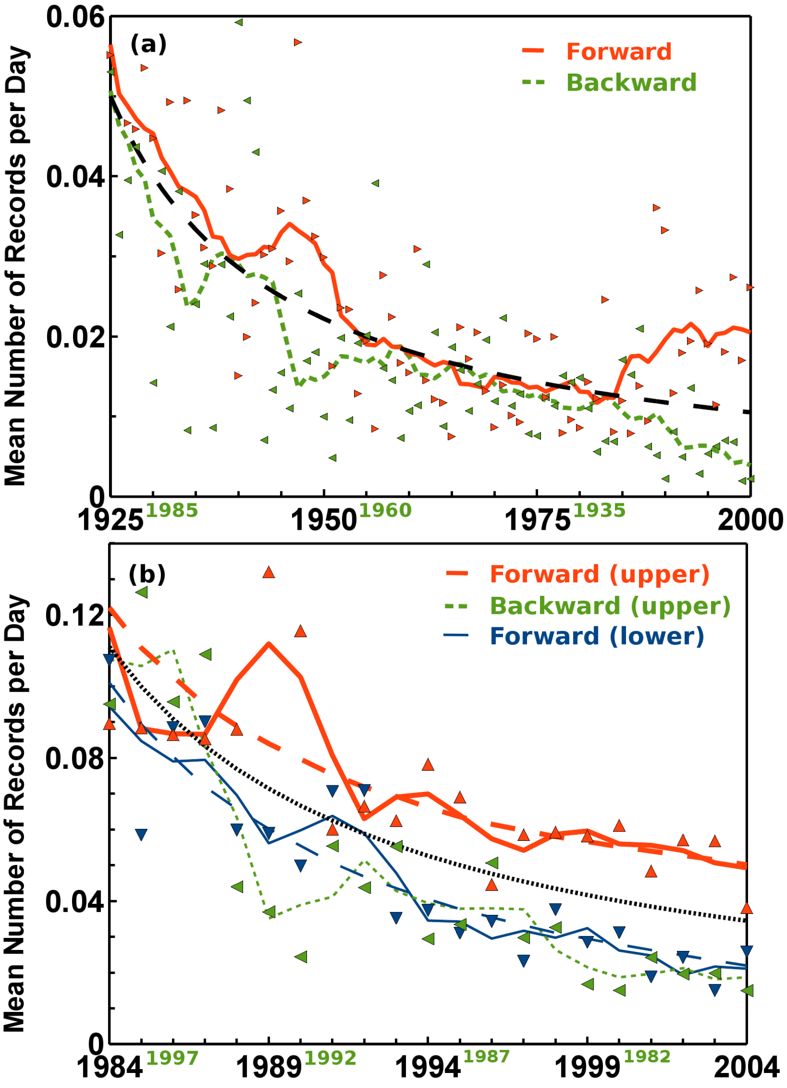

Figure 7.3: **(a)** Record frequencies for data set EI (1906-2005). Symbols show the average number of records per calendar day in forward (▷) and backward (◁) time direction, averaged over 365 calendar days and 43 stations. Full and short-dashed bold lines were obtained by an additional sliding 9-year average. Long-dashed bold line shows the prediction $P_n = 1/n$ for a stationary climate. **(b)** Record frequencies for data set EII (1976-2005). Symbols show the average number of records per calendar day for upper (△) and lower (▽) records in forward direction, and for upper records in backward direction (◁). Full lines were obtained by an additional sliding 3-year average. Dotted line shows the prediction $P_n = 1/n$ for a stationary climate, and long-dashed lines show the model predictions.

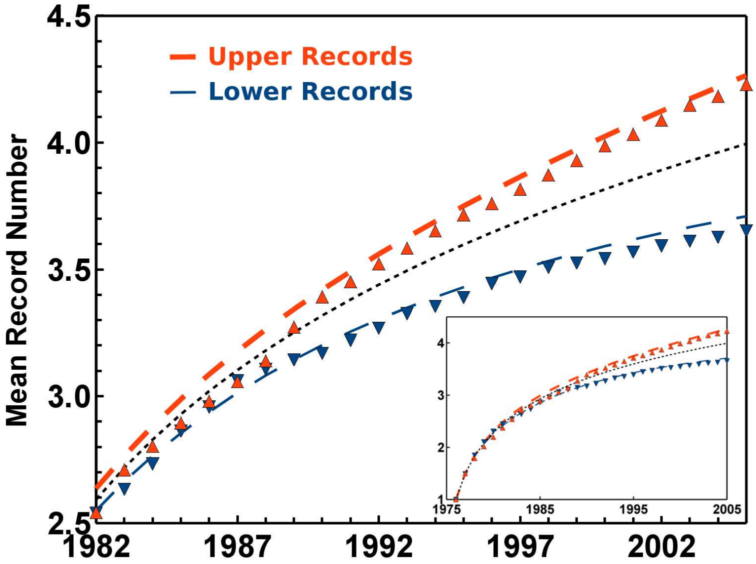

Figure 7.4: Mean record number at European stations (1976-2005). Symbols show the average number of upper (△) and lower (▽) records observed since 1976 at a given calendar year in the forward time analysis. Dotted line shows the prediction for a stationary climate, and dashed lines show the prediction for a constant rate of warming. Inset shows the results for the entire time-span from 1976 to 2005.

10 and more than 200 years. Defective and missing entries were marked in the data sets. We restricted our study to time-series of daily minimum (dTmin) and maximum temperatures (dTmax) which were to at least 95% reliable. This resulted in two sets of station data, one (set EI) consisting of 43 stations that recorded over the 100 year period 1906-2005, and a second (set EII) containing 187 stations that recorded over the 30 year period 1976-2005. Each station recorded 365 time series, and the results presented below constitute averages over all calendar days and all stations within the respective data set.

Taken together, we thus had roughly 15,000 time series in data set EI and 68,000 time series in data set EII at our disposal. However, it is important to note that the effective number of independent time series is much smaller. The number of independent series is limited by correlations both in space and in time. The correlations in space result from the fact that the time series from neighboring stations are strongly correlated if they are less than 1000 km apart [33]. As the spatial distribution of the stations over Europe was relatively homogeneous, we estimate an effective number of 12-15 independent stations for the European data. Furthermore, although daily temperature measurements in subsequent years can be assumed to be uncorrelated, time series recorded at individual calendar days that are close to each other are correlated as well. Based on the analysis of [31] we estimate that these correlations extend over a duration of approximately 10 days, which implies that the number of independent calendar days is around 36. We therefore conclude that our analysis of the European data effectively comprises about 400-500 independent time series.

Figure 7.2 summarizes the analysis of the distribution of dTmax for data set EII. The mean maximal temperature is found to increase at rate $v = 0.047 \pm 0.003$°C/yr, while the standard deviation is essentially constant with a mean value of $\sigma = 3.4 \pm 0.3$°C. The detrended temperature fluctuations are Gaussian to a good approximation. A corresponding analysis for data set EI yields a warming rate of $v = 0.0085 \pm 0.003$°C/yr and the same standard deviation as for data set EII.

To directly test for correlations between daily temperatures, we computed the average two-point correlation for subsequent years after subtracting the drift and normalizing. For both data sets the correlations were found to fluctuate around a small average value of order

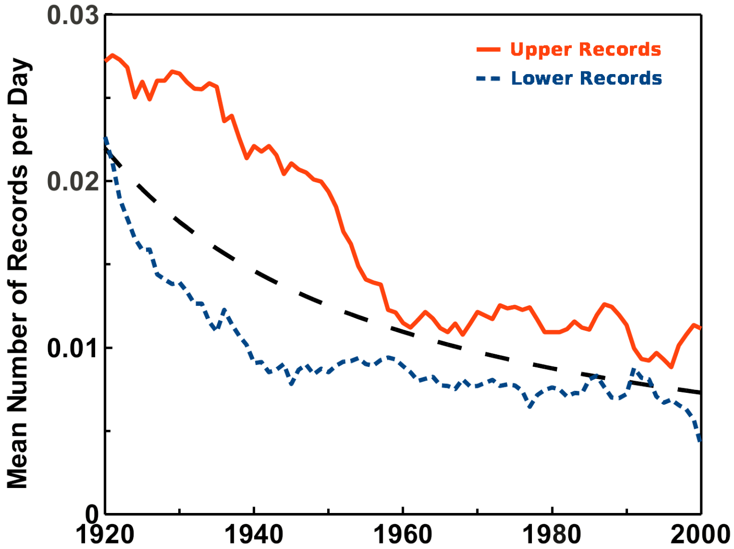

Figure 7.5: Record frequencies for data set AI (1881-2005). Full line shows the average number of upper records per calendar day after a 9-year sliding average, short-dashed line shows the corresponding frequency for lower records, and long-dashed line shows the prediction $P_n = (1 - d/\sigma)/n$ for a stationary climate with discrete measurements.

± 0.01 with a standard deviation of order 0.1. These values are consistent with a power law decay of the form found in [30, 31].

In Figure 7.3**(a)** we show results of the analysis of temperature records in data set EI. The figure depicts the measured daily record frequency for upper records of dTmax, obtained both from a forward analysis (where a record is the highest value of dTmax since 1906) and from a backward analysis (where years are counted backwards in time and records are defined with respect to the temperature in 2005). According to the prediction (7.5), the forward and backward record rates should lie symmetrically around the record rate $1/n$ of the stationary climate, which is consistent with the displayed data. Throughout the analyzed time span (with the exception of a short period around 1960 in which the climate was effectively cooling) the forward record frequency lies above the backward record frequency. This shows that the increase in the mean temperature significantly affects the statistics of records. The effect is particularly pronounced during the last two decades, where warming has been most significant (see the discussion of data set EII below). For the year 2005, the measured forward record frequency is about twice as large as expected for a stationary climate. Using the mean warming rate estimated over the entire 100 yr time period, only an enhancement of 40% is predicted by Eq.(7.5). This shows that the assumption of a constant rate of warming is not a good approximation for data set EI.

Figure 7.3**(b)** displays the corresponding results for data set EII. Since the rate of temperature increase was relatively constant during this time period, we find good quantitative agreement between the data and the model predictions. The agreement is even more striking for the mean record number displayed in Figure 7.4. In a stationary climate the expected number of records observed over n years is

$$R_n = \sum_{k=1}^{n} \frac{1}{k} \approx \ln n + \gamma \tag{7.6}$$

where $\gamma \approx 0.5772156...$ is the Euler-Mascheroni constant. For a 30 year period this amounts to an expected record number of 3.98, which is to be compared to the observed number 4.24 for the upper records, and 3.66 for the lower records of dTmax. Together Figures 7.3 and 7.4 provide a strong validation of our model. Using our estimate $v/\sigma = 0.014$ for data

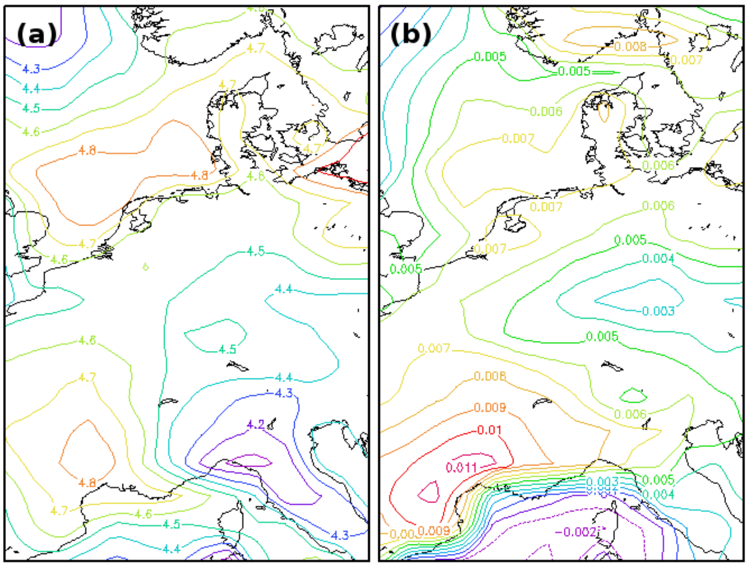

Figure 7.6: Spatial distribution of record number and normalized warming rate in central Europe based on re-analysis data (1957-2000). **(a)** Contour map of the number of records, computed from the 365 time series of daily high temperatures for each point on a rectangular grid of $14 \times 18 = 252$ stations. The expected number of records in a stationary climate is 4.36. **(b)** Contour map of the spatial distribution of the rate of warming, normalized by the standard deviation.

set EII, Eq.(7.5) predicts that the increase in mean temperature has increased the rate of record occurrence by about 40% over the time period from 1976-2005, which implies an additional 5 out of 17 records per year in 2005.

Similar analyses were carried out for upper and lower records of dTmin. We find that the mean record number of dTmin behaves similar to the number of records of dTmax, with 4.32 upper records and only 3.66 lower records. In the backward time analysis we found 3.71 upper records for dTmax and only 3.62 for dTmin. The number of lower records was increased in the backward time analysis, which is in agreement with the results for the upper records. In summary, the number of lower records has decreased in the same manner as the number of upper records has increased (see Fig.7.3**(b)**).

7.4 American data and discreteness effects

The American data sets were extracted from a total of 1062 stations [21]. Requiring again a reliability of at least 95%, we were left with 10 stations that recorded over the 125 year time span 1881-2005 (data set AI) and 207 stations that recorded over the 30 year time span 1976-2005 (data set AII). While the 10 stations of data set AI can be assumed to be independent, the number of effectively independent time series in data set AII is again much smaller.

The result of the record analysis was similar to that performed on the European data sets, with two important differences. First, owing to the continental character of the American climate, the standard deviation σ is considerably larger than in Europe, which, according to Eq.(7.5), implies a weaker effect on the record rate. For example, for data set AII we estimate a warming rate of $v = 0.025 \pm 0.002$°C/yr and a standard deviation of $\sigma = 4.9 \pm 0.1$°C, which yields a ratio v/σ that is only one third of the value for data set EII.

Second, the American data were rounded to full degrees Fahrenheit, whereas the European data were measured in tenths of degrees Celsius. As a consequence, the probability of ties is significant in the American data but negligible for the European data sets. Here we count only *strong* records, which are broken only by a value that exceeds the current record. To account for these discreteness effects one computes the probability that a current record is tied in the nth event. For a small unit of discretization $d \ll \sigma$, this probability is given by $P_n^{\text{tie}} \approx d/(\sigma n)$. This leads to the probability for a record event with discretization as $P_n^d \approx (1 - d/\sigma)/n$, and summing over n the reduction of the number of strong records due to ties in a stationary climate is well described by [24]

$$R_n^d \approx (\ln n + \gamma)(1 - d/\sigma) + 2d/\sigma. \tag{7.7}$$

For the American data sets $d = 5/9°\text{C} = 0.5555..°\text{C}$, which reduces the number of strong records per day expected in a stationary climate over a 30 year period from 3.98 to 3.75. In comparison, the observed number of records in data set AII is equal to 3.86 in the forward analysis and 3.66 in the backward analysis. Again, warming has significantly increased the number of records, but the effect is less pronounced than for the European stations. The evolution of record frequencies in data set AI is shown in Fig.7.5. Note that a failure to account for the discreteness effect would lead to an apparent asymmetry between high and low records relative to the stationary case. Such an asymmetry was observed in the analysis of American temperature data in [19], where it was suggested that warming primarily reduces the number of low temperature records, while the effect on high records is less pronounced.

7.5 Re-analysis data

Taken together, the results presented so far show that the increased occurrence of temperature records can be linked quantitatively to the ratio v/σ of warming rate and temperature variability. Using the ERA-40 Re-Analysis data [20], we were able to extend this analysis to the spatial distribution of the record rate. The data consist of daily temperature series over the period 1957-2000 for 252 geographic locations in central Europe arranged on a regular grid, covering an area of about 3×10^6 km^2. For each location the number of upper records of the daily maximal temperature was determined, and the results are shown in the form of a “record map” in Fig. 7.6**(a)**. The comparison with a corresponding map of local values of the ratio v/σ in Fig.7.6**(b)** shows similar patterns, supporting our conclusion that v/σ is a good (if not perfect) predictor for the increased occurrence of records. Interestingly, the two most pronounced islands of high record occurrence ($R_n > 4.8$) in Fig. 7.6**(a)** are attributed to different mechanisms. One, in southern France, reflects the exceptionally high rate of warming v in this region, whereas the other, over the North Sea, is a consequence of a low temperature variability σ.

An analysis of the seasonal variability of the record events in the (spatially averaged) re-analysis data leads to a similar result. We compared the seasonal distribution of the difference between the mean record numbers in forward and backward time analysis to that of the ratio v/σ, and found a close match between the two (Fig.7.7). While the standard deviation shows a clear seasonal pattern with a pronounced maximum in winter, the seasonal variability of the warming rate is rather complicated. As a consequence, a simple seasonal pattern in the rate of record occurrence could not be identified.

7.6 Conclusions

In summary, by combining a simple mathematical model with extensive data analysis, we have conclusively established that the current rise in mean temperature significantly affects the rate of occurrence of new temperature records. While the majority of the high temperature records observed in Europe at the end of the 30 year period from 1976-2005 would have occurred even in a stationary climate, the effect of warming is substantial, leading to an additional 5 out of 17 records per year.

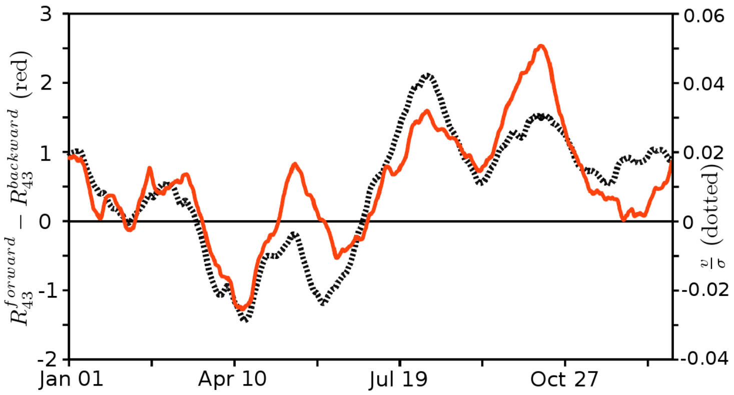

Figure 7.7: Seasonal distribution of the excess number of temperature records compared to the seasonal distribution of v/σ in central Europe based on re-analysis data (1957-2000). The warming rate v and the standard deviation σ for a given calendar day was computed as described above in the caption of Fig.7.2. The warming rate for a given calendar day is the average over all stations of the slope of the corresponding linear regression, and the standard deviation is the averaged squared deviation from the linear trend over all stations and years. Full line represents the difference between the mean number of high temperature records in the forward time analysis and the backward time analysis for the entire 43 year period (left axis). Dotted line gives the seasonal distribution of v/σ (right axis). Both lines were obtained by performing a sliding 30-day average.

The key parameter governing the effect of warming on the occurrence of records is the ratio v/σ, and to leading order the change in record rate is linear in this parameter. It is instructive to explore the future frequency of record-breaking events under the assumption that v/σ will remain constant. The expression (7.5) then predicts that the enhancement of the record frequency (compared to the expectation $P_n = 1/n$ in a stationary climate) will continue to increase, up to the point where the expansion underlying (7.5) breaks down when the two terms become of comparable magnitude at a time roughly of order $n^* \sim \sigma/v$. Beyond this time the record rate saturates at a constant value P^*. Using our estimate $v/\sigma \approx 0.014$ for data set EII, we find that $P^* \approx 0.033$ for the Gaussian distribution. This implies that, towards the end of this century, daily high temperatures exceeding all values measured since 1976 will continue to occur in Europe on about 12 days of the year; at the same time the occurrence of low temperature records will essentially cease.

Acknowledgements

We acknowledge fruitful interactions with Andreas Hense and Sid Redner. This work was supported by DFG within the Bonn Cologne Graduate School of Physics and Astronomy. J.K. is grateful to the Hebrew University of Jerusalem and the Lady Davis Fellowship Trust for their kind hospitality during the completion of this work.

Bibliography

[1] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *Records* (New York: Wiley, 1998).
[2] N. Glick, Amer. Math. Monthly **85**, 2 (1978).
[3] B. Schmittmann and R. K. P. Zia, Am. J. Phys. **67**, 1269 (1999).
[4] R. C. Balling, J. A. Skindlov, and D. H. Phillips, J. Climate **3**, 1491 (1990).
[5] R. W. Katz and B. G. Brown, Climate Change **21**, 289 (1992).
[6] D. R. Easterling, Science **289**, 2068 (2000).
[7] C. Schär, Nature **427**, 332 (2004).

[8] P. A. Stott, D. A. Stone, and M. R. Allen, Nature **432**, 610 (2004).

[9] S. Sabhapandit and S. N. Majumdar, Phys. Rev. Lett. **98**, 140201 (2007).

[10] N. Nicholls and L. Alexander, Progress in Physical Geography **31**, 77 (2007).

[11] G. A. Meehl, J. M. Arblaster, and C. Tebaldi, Geophys. Res. Lett. **34**, L19709 (2007).

[12] S. J. Brown, J. Caesar, and C. A. T. Ferro, J. Geophys. Res. **113**, D05115 (2008).

[13] J. Cattiaux, R. Vautard, and P. Yiou, Geophys. Res. Lett. **36**, L06713 (2009).

[14] J. Ballester, F. Giorgi, and X. Rodó, Climatic Change **98**, 277 (2010).

[15] G. W. Bassett, Climatic Change **21**, 303 (1992).

[16] R. E. Benestad, Climate Res. **25**, 3 (2003).

[17] R. E. Benestad, Global Planet. Change **44**, 11 (2004).

[18] S. Redner and M. R. Petersen, Phys. Rev. E **74**, 061114 (2006).

[19] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).

[20] S. M. Uppala *et al.*, (2005).

[21] C. N. Williams Jr. *et al.*, *United States historical climatology network daily temperature, precipitation, and snow data* (Tech. Rep., Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, 2006).

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

[23] S. Solomon *et al.*, *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2007).

[24] G. Wergen, “Diploma Thesis,” (2009).

[25] R. Ballerini and S. Resnick, J. Appl. Probab. **22**, 487 (1985).

[26] R. Ballerini, Stat. Prob. Lett. **5**, 83 (1987).

[27] K. Borovkov, J. Appl. Probab. **36**, 668 (1999).

[28] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).

[29] J. Krug, J. Stat. Mech.: Theor. Exp. **P08017** (2007).

[30] E. Koscielny-Bunde *et al.*, Phys. Rev. Lett. **81**, 729 (1998).

[31] J. F. Eichner *et al.*, Phys. Rev. E **68**, 046133 (2003).

[32] R. M. Corless *et al.*, Adv. Comp. Math. **5**, 329 (1996).

[33] M. E. Mann and J. Park, Geophys. Res. Lett. **20**, 1055 (1993).

Chapter 8

Record occurrence and record values in daily & monthly temperatures

Gregor Wergen[1], **Andreas Hense**[2] **and Joachim Krug**[1]

[1] *Institute for Theoretical Physics, University of Cologne*
[2] *Meteorological Institute, University of Bonn*

submitted to Climate Dynamics (arXiv:1210.5416)

Abstract: We analyze the occurrence and the values of record-breaking temperatures in daily and monthly temperature observations. Our aim is to better understand and quantify the statistics of temperature records in the context of global warming. Similar to earlier work we employ a simple mathematical model of independent and identically distributed random variables with a linearly growing expectation value. This model proved to be useful in predicting the increase (decrease) in upper (lower) temperature records in a warming climate. Using both station and re-analysis data from Europe and the United States we further investigate the statistics of temperature records and the validity of this model. The most important new contribution in this article is an analysis of the statistics of record values for our simple model and European reanalysis data. We estimate how much the mean values and the distributions of record temperatures are affected by the large scale warming trend. In this context we consider both the values of records that occur at a certain time and the values of records that have a certain record number in the series of record events. We compare the observational data both to simple analytical computations and numerical simulations. We find that it is more difficult to describe the values of record breaking temperatures within the framework of our linear drift model. Observations from the summer months fit well into the model with Gaussian random variables under the observed linear warming, in the sense that record breaking temperatures are more extreme in the summer. In winter however a significant asymmetry of the daily temperature distribution hides the effect of the slow warming trends. Therefore very extreme cold records are still possible in winter. This effect is even more pronounced if one considers only data from subpolar regions.

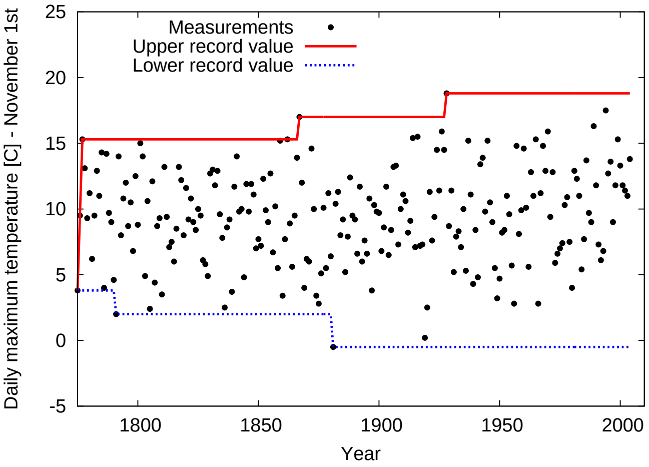

Figure 8.1: Daily maximum temperature measurements for November 1st in Prag for the time span from 1772 to 2010. The figure illustrates the progression of the upper and lower record values. The full red line gives the progression of the highest temperature in the time-series and the dotted blue line the progression of the lowest one. In each case the number of steps is the respective upper or lower record number. The statistics of those record progressions, the number and the height of the steps is the subject of this paper.

8.1 Introduction

In the context of global warming, record-breaking temperatures have received considerable attention recently. In newspapers and in television one frequently hears of hottest summer days, extreme heat streaks, or record breaking storms. Extreme and record breaking weather events are not only interesting for the observer but can also have a big impact on agriculture, economy and human life. If one considers record-breaking events in climatology, these should of course always be seen in the context of global warming. The crucial question is: How much are the climate records we encounter today affected by the evident climatic change [1] of the last decades?

While the study of extremal events in general is very important in climatology [2–7], record temperatures have not received much attention until recently. Even though it is intuitively clear that increasing temperatures should result in a higher than average number of hot day records, this effect was not studied and detected in observational data for a long time. However, there has been some research on the statistics of temperature records in the last years. Redner and Peterson [8] analyzed the statistics of record temperatures for daily temperature measurements from Philadelphia over a 126 year time-span. Due to the fact that they only used data from a single station, their analysis did not establish a significant connection between the increase of global mean temperatures and local record-breaking events, but made important theoretical contributions to the subject. There is also some earlier work by Benestad [9], who analyzed the occurrence of records for stations in Scandinavia.

Meehl et al [10] found a significant effect of slowly changing temperatures on the occurrence of records for daily temperatures at weather stations in the United states. In particular they demonstrated that in an analysis starting from 1970 the rate at which upper records occurred declined more slowly than the rate of lower records. In 2010 two of us obtained similar results in an independent study [11]. We performed an extensive analysis of European and American station data and combined it with a simple mathematical model [12] to quantify the effect of slow temperature changes upon the records of daily temperatures. For the European station data covering the time period 1976-2005, we found that on average 5 of the 17 high temperature records recorded at one station in one year can be attributed to the observed slow increase of temperature.

Our findings and our analytical model were discussed and confirmed by Elguindi et al [13] using gridded data from regional climate models. They also made predictions for the spatial distribution of record temperatures in Europe for the future based on model data from the ENSEMBLES project. Newman et al [14] analyzed record breaking temperatures at a very high resolution from the Mauna Loa Observatory on Hawaii for the time-span from 1977 to 2006. They also presented evidence for slowly increasing temperatures in the occurrence of records. However, in their data they found that while the rate of cold records is significantly decreased, the number of hot records remained unchanged. Rahmstorf and Coumou [15] considered monthly mean temperatures from a weather station in Moscow and could show that the number of hot records in these mean values increased significantly. They also discussed the effect of climatic change on the occurrence of global-mean temperature records based to a large extent on numerical experiments.

The purpose of this paper is to provide a detailed analysis of the statistical properties of record-breaking temperatures also from a theoretical point of view. The main idea behind this analysis is illustrated in Fig. 8.1. We consider the progression of the records and record values in time series of temperature measurements for individual calendar days and months. This way, the daily measurement are always one year apart from each other and, to a good approximation, their statistics can therefore be compared to uncorrelated random variables. We use both station and gridded re-analysis data of daily and monthly mean and maximum temperatures to get a more complete picture in space and time.

In particular we will consider monthly mean temperatures at single stations for the continental United States and gridded temperature data for Europe. In 2010 we already considered daily station data from the United States and had difficulties to quantify the effect of slow changes in temperature with our model [11]. In this paper we will find that if one analyses monthly mean values both increase of the upper record number and the decrease of the lower record number is much more pronounced than in the daily data. The reason for this is the strongly reduced level of variability in the monthly averages. For the European daily data we already found a strong effect of slowly increasing temperatures since the increase of the mean temperature was stronger and the standard deviation was smaller in this data set. Therefore and because of the high density and homogeneity of the gridded temperature data [16] we decided to analyze the statistics of record values of these European data.

At this point we will give a brief outline of the article: Before we introduce and discuss the different data sets, we will give an overview and some new results on the record statistics of independent and identically distributed (iid) random variables (RV's) with a linear drift. This Linear Drift Model (LDM) discussed by Franke et al [12] is very important for our understanding of temperature records. The results for the occurrence of records in the LDM have been published before, but we briefly describe them to make the article self contained.

In section 3 we introduce the data sets that were mentioned above. We describe the statistical properties of the measurements and their time-dependence and discuss how well the observational data fit the LDM. For that purpose we analyzed the time-dependence of the mean and the standard deviation of the recordings. In the following section 4 we will then first consider the occurrence of records in the different data sets, in particular the record rate in the daily and monthly temperature recordings. We analyzed the record rate in the United States with respect to the different seasons to find out when the effect of slowly evolving temperatures on the record statistics is strongest. In particular we will discuss the ratio between the number of upper records and the number of lower records and compare it to the predictions from our analytical model.

Section 5 is again about theoretical aspects. Here we will discuss the statistics of record values within the LDM using both an analytical approach and numerical simulations. We quantify the effect of the linear drift on the mean value and the standard deviation of record values that occur at a certain time step and of record values that have a certain record number in the series of record events. The aim is to understand if the record events we have to expect in the presence of slowly increasing temperatures are more extreme or more variable than without any climatic change.

In the subsequent section 6 we will then compare the findings of section 5 to the observational data we already introduced and discussed in sections 3 and 4. In section 6 we will consider the behaviour of the mean values of record events as well as their full distributions. For that purpose we will introduce a simple rescaling of the observational data to account for seasonal and spatial variations in the standard deviation of the time series. Then we will discuss both the values of records that occur in a given year as well as the values of records that have a certain record number.

Finally in section 7 we summarize and evaluate our findings and discuss them in the context of ongoing research in the field of temperature records.

8.2 Theory: Record occurrence in the presence of linear drift

8.2.1 Record statistics of iid. RV's

We consider time series of uncorrelated random variables (RV) X_k from continuous probability densities $f_{X_k}(x_k), k \in \{1, 2 \ldots, n\}$. As mentioned earlier, an upper (lower) record in the nth step is an entry X_n that is larger (smaller) than all previous entries X_k with $k < n$. The basic properties of record events in such time series can for instance be found in [17–19]. In the special case of identically and independently distributed (iid) RV's from a single probability density $f_X(x)$ the probability that the nth entry is a record is simply given by $P_n = 1/n$ (cf. [17, 19]). This holds, because of the symmetry of the problem, both for upper and lower records. From now on we will call the probability P_n that the nth entry in the series is a record the record rate. In the iid case the mean number of records R_n up to the nth step can be obtained by summing over the record rate and for large n we find:

$$R_n = \sum_{k=1}^{n} P_k = \sum_{k=1}^{n} \frac{1}{k} \approx \ln(n) + \gamma. \tag{8.1}$$

Here, $\gamma \approx 0.577215...$ is the Euler-Mascheroni constant [17, 18]. In this case, one can prove that record events are uncorrelated [20] and, if one goes to a logarithmic time scale one finds that they form a Poisson process [21]. Another important feature of this result for records of iid RV is that P_n and R_n are completely independent of the shape of the underlying probability density $f_X(x)$.

8.2.2 Linear Drift

In general, when the RV's X_k are not identically distributed, it is more difficult to compute the record rate P_n. For independent RV's X_k from a series of arbitrary continuous distributions $f_{X_k}(x_k)$ the upper record rate is given by the following integral [17, 18]:

$$P_n = \int \mathrm{d}x_n \, f_{X_n}(x_n) \prod_{k=1}^{n-1} F_{X_k}(x_n) \tag{8.2}$$

where $F_{X_k}(x_n)$ is the probability distribution function of $f_{X_k}(x_k)$ with

$$F_{X_k}(x_n) = \int^{x_n} \mathrm{d}x \, f_{X_k}(x). \tag{8.3}$$

Here, the integrand is just the probability that the nth entry has a value of x_n times the probability that all previous entries are smaller, which is represented by the product.

This probability is then integrated over all possible values for a record in the nth step x_n. Analogous to this the lower record rate $P_n^\star$ is given by:

$$P_n^\star = \int \mathrm{d}x_n \, f_{X_n}(x_n) \prod_{k=1}^{n-1} \left(1 - F_{X_k}(x_n)\right). \tag{8.4}$$

The LDM was first considered by Ballerini and Resnick [22, 23] and later also by Borovkov [24]. In this model we consider iid RV's Y_k with a linear drift of the following form:

$$X_k = Y_k + ck, \tag{8.5}$$

with a constant c. In this case $f_{X_k}(x)$ is simply given by $f_{X_k}(x) = f(x - ck)$ with fixed $f(x)$. The underlying distributions have all the same shape, but the mean value increases with a constant speed c. This model was used before to better understand the statistics of athletic records [22, 25], but we showed that is also capable of describing the occurrence of records in daily temperature recordings [11, 15]. By considering (8.2) one finds that for any constant drift c the record rate in the LDM is of the following form:

$$P_n(c) = \int \mathrm{d}x \, f(x) \prod_{k=1}^{\infty} F(x + ck). \tag{8.6}$$

Most interesting for us is the statistics of records for a drift velocity c much smaller than the width of the probability distribution $f(x)$. In most cases this width is just the standard deviation σ of the probability distribution. Performing a series expansion of (8.6) in the regime of $c \ll n/\sigma$ one finds the following approximation for $P_n(c)$ [11]:

$$P_n(c) \approx \frac{1}{n} + \frac{c}{\sigma}\frac{n(n-1)}{2} \int \mathrm{d}y \, f^2(y) \, F^{n-2}(y). \tag{8.7}$$

[12] evaluated this expression for distributions from all three classes of extreme value statistics [19, 26–28]. The dependence of the record rate $P_n(c)$ on the drift c is systematically different between the three classes, but also within them one can find differences between different individual probability distributions. For the Weibull class of probability distributions with a finite support the effect of the linear drift on $P_n(c)$ was found to be strong and increasing with n. For the distributions of the Fréchet class with power-law tails it decays with growing n and vanishes for $n \to \infty$ [12]. The behavior of $P_n(c)$ in the Gumbel class of distributions with an infinite support that decay faster than a power law is intermediate between these two cases. For a simple exponential distribution with $f(x) = 1/\nu e^{-x/\nu}$ the effect of the drift on the record rate is independent of n and we have $P_n(c) \approx \frac{1}{n} + c\nu/2\sigma$.

Most interesting for our applications is $P_n(c)$ for a Gaussian distribution of the following form:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{8.8}$$

Here, μ is the mean value of the probability distribution and σ its standard deviation. The approximate evaluation of Eq. (8.7) for large enough n with $c \ll n/\sigma$ and a Gaussian distribution with standard deviation σ yields:

$$P_n(c) \approx \frac{1}{n} + \frac{c}{\sigma}\frac{2\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)}. \tag{8.9}$$

[12] compared this approximation to numerical simulations and good agreement was found for $n > 7$ and $c \ll n/\sigma$. The result was applied to the statistics of record-breaking temperatures for the first time by [11], who showed that the regime of $c/\sigma \ll 1$ is appropriate for the modeling of daily temperatures in a climate with moderate warming like that occurring

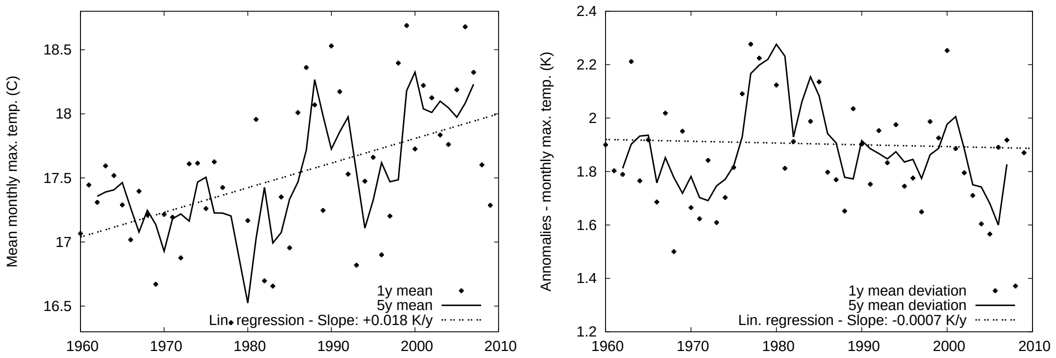

Figure 8.2: Annual mean temperature (left) and standard deviation (right) of monthly mean temperatures from contiguous US stations 1960-2010. Dots are one year averages, full lines give a five year running mean. The anomalies defining the standard deviations of the mean temperature were computed by subtracting the mean and linear trend 1960-2010 individually for each month. Annual mean values increase over the last 50 years with an average rate of 0.018 K°/y, whereas the standard deviation of the monthly anomalies remained more or less constant.

	$\overline{c}$	$\overline{\sigma}$
Winter	0.043 ± 0.015 K°/y	2.45 ± 0.012 K°
Spring	0.023 ± 0.008 K°/y	1.68 ± 0.006 K°
Summer	0.004 ± 0.007 K°/y	1.70 ± 0.006 K°
Fall	0.007 ± 0.005 K°/y	1.56 ± 0.005 K°
	$\overline{c/\sigma}$	
Winter	$0.017 \pm 0.006\, y^{-1}$	
Spring	$0.014 \pm 0.005\, y^{-1}$	
Summer	$0.002 \pm 0.004\, y^{-1}$	
Fall	$0.004 \pm 0.003\, y^{-1}$	

Table 8.1: Averaged drift $\overline{c}$, averaged monthly standard deviation $\overline{\sigma}$ and averaged normalized drift $\overline{c/\sigma}$ for the contiguous US stations 1960-2010. While in winter and spring there is a significant and strong drift in the data, we find only little or no effects in summer and fall.

in the last decades. This result will also be employed for comparison with observational data in section 4 of this article.

Another complication that can arise in the context of record statistics of historical temperature recordings, is the problem of rounding. The fact that, for technical reasons, temperature measurements can only be recorded up to a certain degree of accuracy opens up the possibility for ties. Usually the observations are rounded down (or up) to a certain value kd, where d is a discretization length (e.g. 0.1 K° if temperatures are stored up to the first digit) and k an integer number. [29] recently explored the interesting and manifold consequences of this rounding for the statistics of records in time series of iid RV's in the context of the three universality classes of extreme value statistics.

8.3 General introduction of the data

We focus our analysis on two different sets of data from Europe and the contiguous United States. The purpose of this section is to analyze the distributional properties of these data sets. We want to know if the LDM discussed in the previous section is applicable to the observational data. In particular we will examine if a Normal or Gaussian probability density function with a slowly changing expectation value is a reasonable approximation.

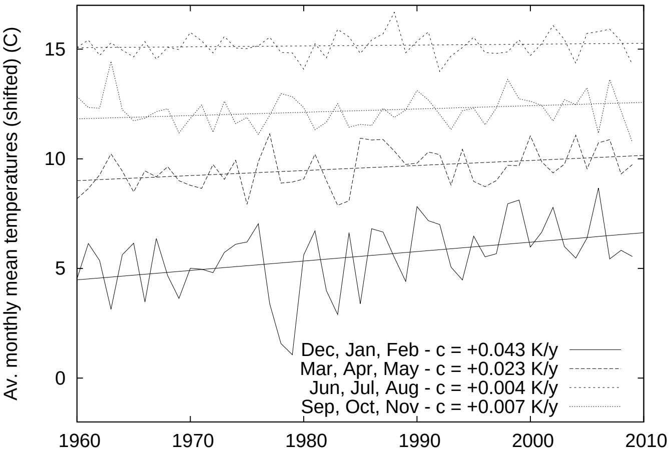

Figure 8.3: Time series of seasonal averages of monthly mean temperatures from contiguous US stations 1960-2010. For clarity, three year moving averages are plotted. Inset shows the estimated trends over the period 1960-2010. Winter months exhibit the strongest warming with $c = 0.043$ K°$/y$, during spring and fall the warming is moderate with $c = 0.023$ K°$/y$ and $c = 0.007$ K°$/y$ and in the summer months we find only a very small warming of $c = 0.004$ K°$/y$. However for the normalized drift and therefore the record-statistics also the standard deviation is important, which is significantly smaller in the summer as well (see text).

Therefore we are interested in the linear trend c of the daily and monthly temperature averages and the standard deviation σ around that transient mean value $\mu + ck$. As outlined above the most important quantity in our analysis is the normalized drift c/σ.

8.3.1 US monthly station data

We obtained monthly mean temperatures from 1217 weather stations of the contiguous US [30]. The data cover the period 1895 and 2010, but but not all stations are complete. Therefore we decided to analyse the 50 year period 1960 to 2010 (data set USM) which is the period with least missing data. This way we could consider 1217×365 time series. Based on the discussion in [11] we estimate that the number of effectively independent time series in this data set is much smaller than the total number of series. The number of independent stations is limited because of correlations both in space and in time. Following [11] we estimate not more than 20-25 independent stations and around 36 independent calendar days, leading to a total number of around 700-900 independent time series.

As one can see in Fig. 8.2 the monthly averages show a clear increase in the mean value with a drift of $c = 0.019 \pm 0.005$ K°$/y$. In contrast, the standard deviation of the monthly averages remains constant around $\sigma = 1.88 \pm 0.03$ K°, which is smaller than the standard deviation of the daily measurements of around 5 K° for the same period [11]. Using a linear regression analysis we tried to detect a trend in the standard deviation, but a only found a insignificant trend of -0.0007 ± 0.0021 K°. Assuming the constant σ, we obtain a normalized drift $c/\sigma \approx 0.010 \pm 0.003\ y^{-1}$, which is almost as large as the c/σ for the European daily data in [11].

The seasonal dependence of the normalized drift c/σ is presented in Fig. 8.3. During winter the US mean temperature increased most strongly, while the trend for summer is much smaller. However, important for the record statistics is the ratio c/σ. The normalized trends are listed in Tab. 8.1. Based on the theoretical results we expect a strong effect of the slow temperature increases on the record rates in winter and spring in contrast to summer and fall if all prerequisites of the theory are fulfilled. This is due to the strong difference in normalized drift (Tab. 8.1, third column).

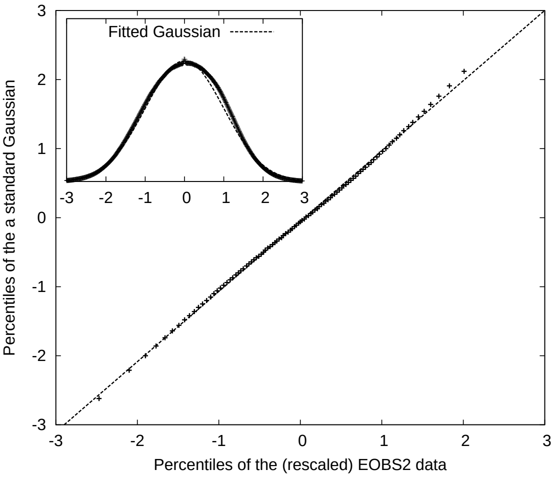

Figure 8.4: Q-Q plot of the linearly detrended and normalized daily maximum temperatures in the EOBS data (EOBS2) with RV's sampled from a Gaussian distribution with standard deviation unity. The percentiles collapse on a line showing that the rescaled measurements are Gaussian to a good approximation. The inset shows the estimated probability density of the data compared to a Gaussian (dashed line).

	Rec. ratio - US stations	Ratio - LDM
Winter	7.6	7.1 ± 5.6
Spring	9.4	4.8 ± 3.7
Summer	2.4	1.2 ± 0.7
Fall	1.6	1.5 ± 0.4

Table 8.2: Ratio between the number of upper and lower records after $n = 50$ years in the monthly data from the US stations along with the corresponding predictions from the LDM. The values for the normalized drift $\overline{c/\sigma}$ were taken from Tab. 8.1.

8.3.2 European daily data

The second data set we obtained are taken from the E-OBS project [16]. The E-OBS data set provides daily minimum, mean and maximum temperatures on a 0.25 degree regular grid for most of Europe and also parts of northern Africa starting from 1950 to the end of 2010. We analyzed the daily maximum temperature data set for two time periods, one between 1950 and 2010 (data set EOBS1) and a second one between 1980 and 2010 (data set EOBS2). Again, it is important to mention, that despite the fact that we analyzed a much larger number of time series, only around 300-400 series were effectively independent. This number is smaller in Europe due to the smaller size of the considered area. The time series analysis of the mean of the daily maximum temperature and the standard deviation were not significantly different from the station data considered in [11]. During the first 30 years of EOBS1 the mean temperature shows no prominent trend. During the second period, between 1980 and 2010 (EOBS2), we find a clear warming trend of about $c = 0.042\ \mathrm{K}^{\circ}/y$. In data set EOBS2, the standard deviation of the daily temperatures around the linear trend remained constant at a value of about 3.4 K° resulting in a normalized drift of around $c/\sigma \approx 0.012 \pm 0.003\ y^{-1}$, which is slightly smaller than that obtained from European station data [11].

In Fig. 8.4 we present a quantile-quantile (Q-Q) analysis of the distribution of the detrended and normalized temperature measurements in data set EOBS2. We subtracted a linear trend, which was obtained by a regression analysis, from the time-series at each grid point and then normalized them by dividing with the standard deviation of the mea-

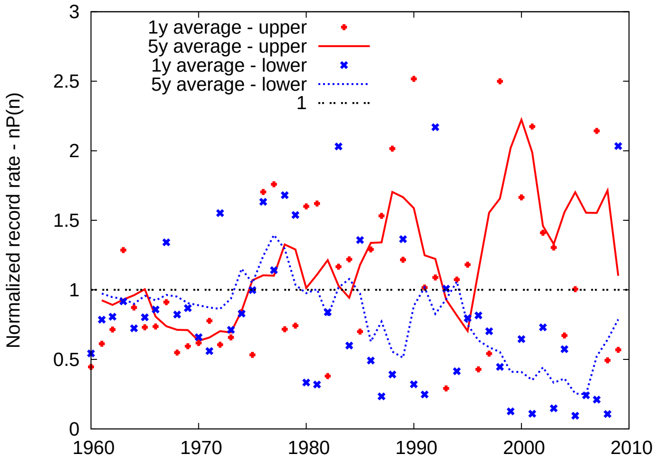

Figure 8.5: Normalized record rate nP_n of monthly mean temperature data from US stations. The dots represent the annual averages of nP_n and the lines represent a five-year running average. The line at one is the behaviour expected in the case of a stationary climate with realizations from an iid process. Before 1980 it is hard to distinguish between the upper and lower record rate. After 1980 the upper record rate increases and the lower record rate decreases.

surements in the individual series. Then the percentiles of the measurements were plotted against the corresponding percentiles of a Gaussian distribution with mean zero and standard deviation unity. Apparently the measurements are Gaussian to good approximation and the data therefore fits our model of a Gaussian with a linear drift. The inset shows an estimation of the probability density function based on a simple histogram estimator with a bin width of 0.01 K°.

8.4 Occurrence of records in daily and monthly temperatures

8.4.1 US monthly station data

[11] showed that the record rate P_n of daily US station temperatures does not display a significant effect of the slowly evolving mean temperature. When one considers monthly data this changes basically due to the reduced variance. In Fig. 8.5 we show the normalized record rate nP_n for monthly means. Apparently, for the years after 1980, the number of high temperature (upper) records is above the stationary case of $nP_n = 1$ and the number of low temperature (lower) records is decreased. Compared to the null-hypothesis of a stationary climate, in the years between 2000 and 2010, we found around 1.52 times the number of hot temperature records and only half (0.48) the cold records one would expected. Therefore, the number of upper records was about 3.13 times the number of lower records.

These results are in good agreement with the Gaussian LDM we described above. For the normalized drift $c/\sigma \approx 0.01\ y^{-1}$ obtained from our data, the LDM predicts 1.44 ± 0.13 times more upper records than expected in a stationary model and only 0.50 ± 0.15 the number of lower records. The ratio between upper and lower records in the last 10 entries should be 2.88 ± 1.12, which is also in agreement with the observational data. Interestingly, without the very cold year of 2009, these numbers become more extreme. Ignoring the 2009 data, the ratio between the number of upper and lower records in the last ten years is around 4.9.

Additionally we can consider the record rate for the different seasons. We computed the ratios between upper and lower records for the four seasons and compared them to

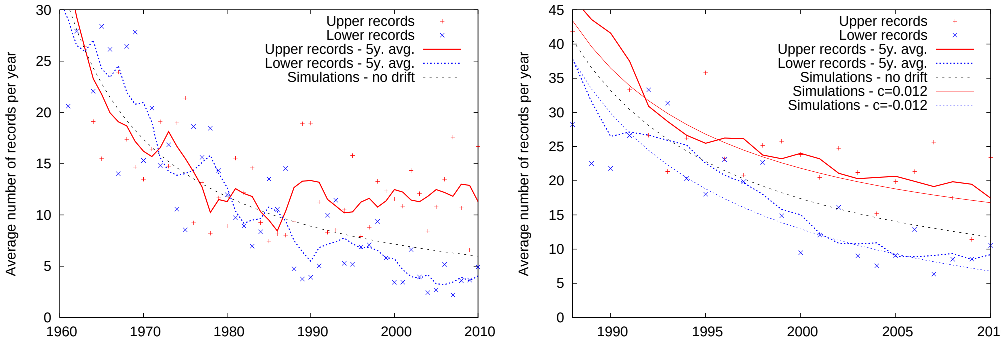

Figure 8.6: Left figure: Records per station and year in the daily data set EOBS1 (1950-2010). Red represents the occurrence of upper records and blue the occurrence of lower records. The crosses give the annual average and the lines a five-year running average. The black dashed line is the $365/n$ behavior we would have expected in the case of a stationary climate. Until 1980 the upper and lower record rate are hard to distinguish and do not vary much from the stationary behavior. After 1980 the rate of upper records ceases to decrease and approaches a constant value of about 12 records per year, while the rate of lower records decreases faster than the $365/n$ line. **Right figure:** The same analysis for EOBS2 (1980-2010). Here the upper record rate is significantly increased compared to the no-warming model and the lower record rate. Red and blue dashed lines are the analytical predictions with a normalized annual mean trend of $c/\sigma = 0.012y^{-1}$ estimated from the data.

the estimates of c/σ and the resulting analytical predictions from the LDM with Gaussian RV's. The results of this analysis are shown in Tab. 8.2. As expected mostly the winter and spring months experience a strong effect of slow temperature increase on the record statistics. In spring a heat record during the last ten years of the observational period was almost ten times as likely as a cold record. In winter this factor is almost eight. However, also in summer and fall the moderate warming lead to a significant effect in the statistics of records. Interestingly in spring and summer the ratio between the upper and lower record rate is larger than predicted by the Gaussian - LDM, but given the large fluctuations in the data, this can very well be a coincidence. It is also interesting to notice, that both in winter and in spring less then 20 % of the cold records that we would have expected in the case of a stationary climate actually occurred.

8.4.2 European daily re-analysis data

We analyzed the record rates for the two time spans chosen for EOBS1 and EOBS2 (1950-2010 and 1980-2010) and compared them to the analytical predictions from our linear drift model. The results can be found in Fig. 8.6. Considering the data set EOBS1 we find that in the first 30 years of the observation period, there was no significant effect of slowly increasing temperatures on the record statistics. Between 1950 and 1980, both the rate of upper and lower records behaved roughly like the record rate of iid RV's. After 1985 the upper record rate is significantly larger than the lower record rate, with the lower record rate significantly decreasing. For the years after 2000 the ratio between upper and lower records exceeds two and even approaches a value of three at the end of the observational period.

In the data set EOBS2, the averaged upper record rate was higher than the lower record rate for almost the entire time span between 1980 and 2010. At the end of the observational period there were twice as many upper records as there were lower records. Here, the predictions from our linear drift model are very accurate in predicting the effect of the warming on the occurrence of both upper and lower records.

In Fig. 8.7 we considered the number of records that occurred over a prolonged time span.

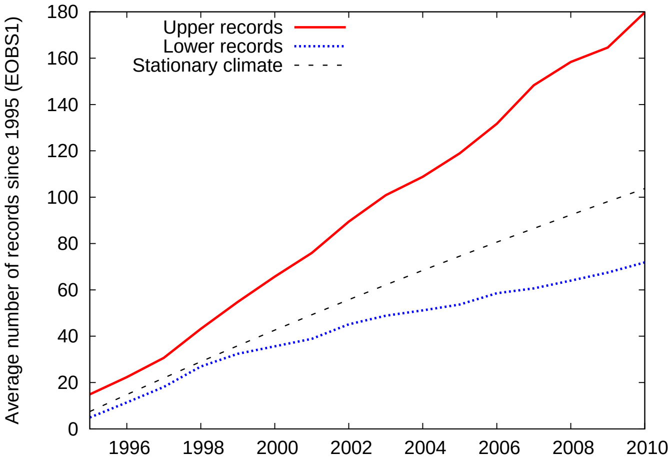

Figure 8.7: Excess number of records that occurred since the year 1995 per grid point per year in data set EOBS1. The record temperatures were recorded from the beginning of the time series. For this figure we only summed up records that occurred after the beginning of 1995. The records that contribute to this figure are the ones that exceeded the record that was valid at the end of 1994.

We analyzed the upper and lower record temperatures that occurred in data set EOBS1 and started summing up the record rate beginning with the year 1995. Even though this figure is difficult to compare with our LDM, since we do not assume a linear drift over the entire time span from 1950 to 2010, it shows how strong the effect of warming on the record occurrence was in the last decades. Towards the end of the considered period, one finds that upper records occurred on average more than 2.5 times more often than lower records.

8.5 Theory: Distributions of record values

After discussing the occurrence of record events in the previous sections, we now turn to the record values themselves. There are in principle two ways to study the effect of a slowly changing mean value on the record values. One approach is to consider the probability density function (pdf) of record values with a fixed record number k, which can however happen at an arbitrary time n. The other is to study pdf's of record events occurring at a fixed time step n. We do not know any simple way of computing the pdf's of the kth records in the presence of linear drift analytically. We will discuss the iid case and present results of some numerical simulations. If we consider instead the probability densities of record values for records occurring at a fixed time n we can use the methods described above to compute a small c approximation in the framework of the LDM.

8.5.1 Records in the n'th step

With the general expression for the record rate $P_n(c)$ in the LDM we gave in section 2, we can also obtain an expression for the probability distribution or cumulative distribution (cdf)

$$Q_n(c,x) = \text{Prob}\left[X_n \text{ is rec. } \& \ X_n < x\right] \tag{8.10}$$

of a record that occurs in a certain time step n. This is given by:

$$Q_n(c,x) = \frac{1}{P_n(c)} \int^{x} \mathrm{d}y \; f_X(y) \prod_{k=0}^{n-1} F_X(y+kc) \, . \tag{8.11}$$

The prefactor $P_n(c)^{-1}$ is necessary for the normalization. The corresponding pdf is given by:

$$p_n(c,x) = \frac{1}{P_n(c)} f_X(x) \prod_{k=0}^{n-1} F_X(x+kc). \tag{8.12}$$

For the iid case ($c=0$) these expressions reduce to $Q_n(0,c) = n\int^x \mathrm{d}y\ f_X(y) F_X^n(y)$ and therefore a pdf $p_n(0,x) = nf_X(x) F_X^n(x)$, which is well known as the pdf of the maximum of n iid RV's. Most interesting and important for our analysis of temperature record values is the mean value $\mu_n(c)$ of a record that occurs at time n. This is given by the first moment of the pdf $p_n(c,x)$:

$$\mu_n(c) = \frac{1}{P_n(c)} \int \mathrm{d}y\ y f_X(y) \prod_{k=0}^{n-1} F_X(y+kc). \tag{8.13}$$

Similar to the case of the record rate $P_n(c)$ we can compute a series expansion for this expression in the regime of $cn \ll \sigma$. Doing this we find

$$\mu_n(c) \approx \left(1 - \frac{c}{2} n^3 I_n^{(0)}\right) \mu_n(0) + \frac{c}{2} n^3 I_n^{(1)}, \tag{8.14}$$

where we defined

$$I_n^{(j)} := \int \mathrm{d}y\ y^j f_X^2(y) F_X^{n-1}(y) \tag{8.15}$$

and

$$\mu_n(0) = n \int \mathrm{d}y\ y f_X(y) F_X^{n-1}(y). \tag{8.16}$$

Apparently $I_n^{(0)}$ is the integral that appears in our result for the record rate $P_n(c)$ [Eq. (8.7)]. Furthermore $\mu_n(0)$ is the mean value of a record that occurs at time n in the case of iid RV's. Since we have an exact expression for the pdf $p_n(c,x)$, we can of course also compute higher moments and, for instance, the variance of a record that occurs in the LDM at a certain time n. For the variance we find:

$$\begin{aligned} \sigma_n^2(c) &\approx \sigma_n^2(0)\left(1 - \frac{c}{2} n^3 I_n^{(0)}\right) \\ &\quad + \frac{c}{2} n^3 \left(\mu_n(0)\left(I_n^{(0)} - I_n^{(0)}\right) + I_n^{(2)}\right) \end{aligned} \tag{8.17}$$

From now on we will focus on the mean value $\mu_n(c)$. Again, as in the work of [12] it is possible to compute $\mu_n(c)$ for instances of all three classes of extreme value theory and we find a systematic classification of the behavior with respect to these classes [31]. Here, we will focus on the Gaussian density, because it is the one we need for our comparison of the LDM to observational data. For the same Gaussian pdf as in section 2, we find that:

$$\mu_c(c) \approx \mu_n(0) + n \frac{c}{\sigma} \frac{2\sqrt{\pi}}{e^2} \ln(4). \tag{8.18}$$

Interestingly, in contrast to the case of the record rate $P_n(c)$, the effect of the drift on $\mu_n(c)$ up to first order in c is linear in n. In the limit of $cn \gg \sigma$ we expect that $\mu_n(c) - \mu_n(0)$ is again linear in n. It is easy to see that for $n \to \infty$ we get $\mu_n(c) - \mu_n(0) \approx cn$. In this regime, the drift dominates the behaviour and the mean record value $\mu_n(c)$ is given by the linearly growing mean of $f(x - cn)$ plus a sublinear contribution.

We compared these analytical findings for the Gaussian density to numerical simulations. The results can be found in Fig. 8.8. Both in the small and in the large n regime the analytical predictions describe the behaviour of $\mu_n(c) - \mu_n(0)$ very accurately. For the chosen drift rate of $c = 0.01$ the intermediate regime of $cn \propto \sigma$, where both descriptions given above fail seems to be very small.

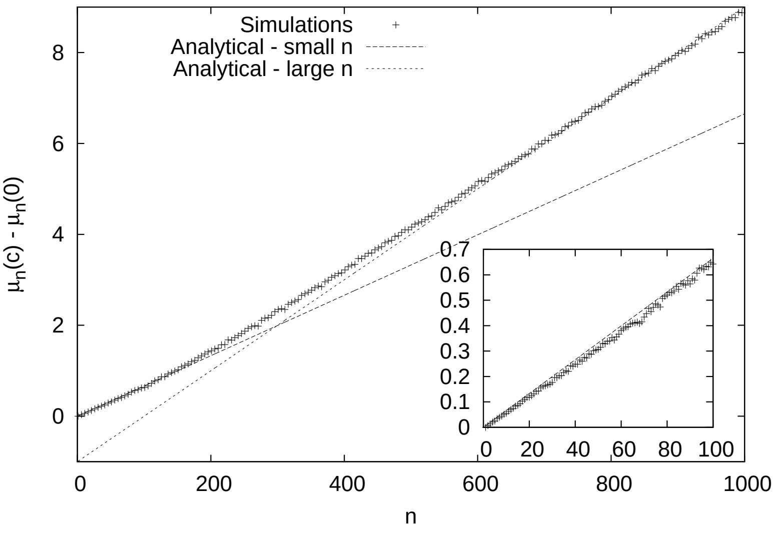

Figure 8.8: Effect of the drift on the mean value $\mu_n(c) - \mu_n(0)$ for Gaussian RV's. The crosses are results from numerical simulations. For $c = 0$ and $c = 0.01$ we performed 10^6 runs each with $n = 1000$ RV's and averaged the differences between the mean record values. The lines are the analytical results for small and large n. The inset shows a comparison with the small n result for n between 0 and 100.

8.5.2 k'th records

In the case of the record values of a record that occur at an arbitrary time n, but with a fixed record number k, one can easily give the full distribution of record values in the iid case. Here, we will briefly discuss the findings presented in [17]. There it was shown that the pdf $p_k(x)$ of a record that occurs with record number k can be written as follows:

$$p_k(x) = \frac{f_X(x)}{(k-1)!} \left(-\ln\left(1 - F_X(x)\right)\right)^{k-1}. \tag{8.19}$$

This result is basically a consequence of the lack-of-memory property of the exponential distribution, i.e. the fact that a new record from an exponential distribution, independent from the time of its occurrence, will always be an exponential RV plus the value of the last record [17]. This leads to the results that for an exponential distribution the kth record has a Gamma distribution $p_k^{exp}(x) = \Gamma[k-1, x]$. The general result given above can then be obtain through a simple mapping.

Unfortunately, for RV's with a time-dependence like a linear drift, the lack-of-memory property is lost and we do not know how to obtain a simple expression for the $p_k(x)$ for the LDM. In Fig. 8.9 we show the normalized densities of kth records for a few small record numbers k. Apparently, already for a small drift of $c = 0.01$, there is a significant effect on the record value pdf for larger k. Also the width of the pdf increases as an effect of the drift. It seems that especially the right tails of the distributions become broader. In the context of temperature records these events in the tails are particularly interesting.

8.6 Distributions of record values in European temperature recordings

Based on the analytical work from the previous section we can now consider the values of records in observational data. For this analysis we focused on the European reanalysis data and in particular EOBS2.

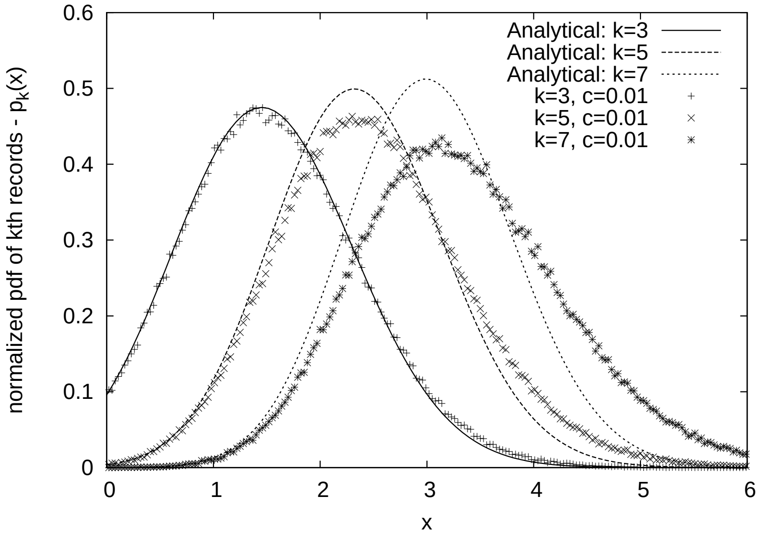

Figure 8.9: Normalized distributions of records with record values k occurring at an arbitrary time n for Gaussian RV's in the iid case and with a constant drift of $c = 0.01$. Shown are the pdf's for $k = 3, 5$ and 7. The analytical results for the iid case (Eq. (8.19)) are plotted as lines, the crosses are the numerical results with drift. For each k and c we averaged over 10^6 realizations of a time series of length 10^6. The finite length of the time series does not have a significant effect on the distributions of records with the given values of $k \ll \ln 10^6 + \gamma$.

Here, the situation is a bit more complicated than in the case of the record rate $P_n(c)$. While the record rate in the LDM only depends on the normalized drift c/σ, the values of the records depend on the standard deviation σ itself. The standard deviations of daily temperatures vary both spatially and seasonally and therefore it is difficult to compare the values of record breaking temperatures without some additional assumptions. To make the time series in EOBS2 more comparable we performed a rescaling of the data. This was done as follows:

The LDM assumes that an individual series of temperatures measurements $T_1, ..., T_n$ measured in n subsequent years is given by

$$T_k = \mu_0 + ck + \sigma\xi_k, \tag{8.20}$$

where the ξ_k are iid RV's from a Gaussian distribution with standard deviation one. We subtract the intercept μ_0 and divide by the standard deviation σ to obtain the following time series

$$\tilde{T}_k = \frac{1}{\sigma}\left(T_k - \mu_0\right). \tag{8.21}$$

It is easy to see that these rescaled measurements have standard deviation unity around a normalized linear drift $\tilde{c} \equiv c/\sigma$. This kind of rescaling was done for all time series in the respective data sets individually so that we obtained comparable series of rescaled measurements. This way the data is most-suitable for comparison with a Gaussian LDM. If the observations were perfectly uncorrelated Gaussian RV's with an arbitrary but fixed standard deviation the record values of the time series after this rescaling would look exactly like the record values from a Gaussian LDM with standard deviation one. These rescaled temperatures should then obey the following LDM:

$$\tilde{T}_k = \frac{c}{\sigma}k + \xi_k = \tilde{c}k + \xi_k. \tag{8.22}$$

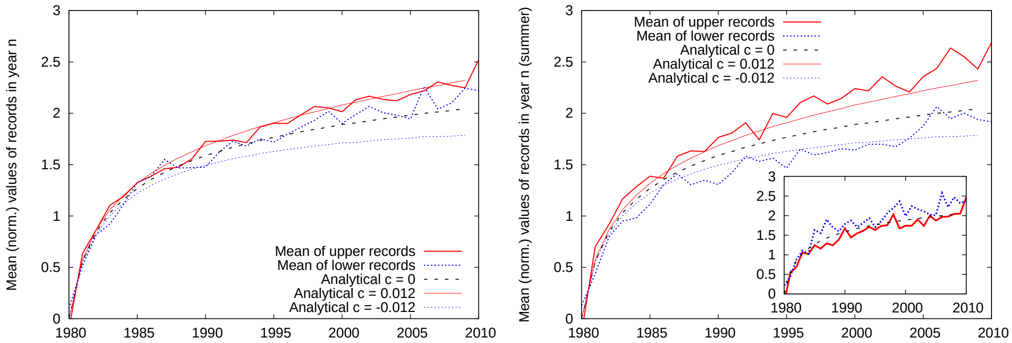

Figure 8.10: Left figure: Time series of the rescaled mean value of temperature records that occurred in data set EOBS2. The full line gives the behaviour of the upper record value and the dotted line gives the inverse (negative) behaviour of the lower record. The values of the lower records were multiplied with -1 to make them comparable to the upper records. The rescaling is described in more detail in the text. The figure also contains the analytic results for no drift ($\tilde{c} = 0$) and the LDM with a normalized drift of $\tilde{c} = 0.012\ y^{-1}$ that was estimated from the observations. **Right figure:** The same plot, but only for the two summer months of July and August. The analytic results for $\tilde{c} = 0$ and the LDM case $\tilde{c} = 0.012\ y^{-1}$ are added. The inset shows the same analysis for the two winter months January and February.

It is important to notice that the ordering of the measurements and in particular the statistics of records is not altered by this procedure because

$$T_n = \max\left[T_1, ..., T_n\right] \Rightarrow \tilde{T}_n = \max\left[\tilde{T}_1, ..., \tilde{T}_n\right]. \tag{8.23}$$

So if and only if T_n was a record in the original series, $\tilde{T}_n$ will also be a record in the rescaled series. The record rate P_n and the record number R_n are therefore completely invariant under this rescaling and only the values of the records will change according to Eq. (8.21).

8.6.1 Mean values of records in a given year

In Fig. 8.10 we analyze the mean values of a record that occurred in a certain year for the 30 years of observation in data set EOBS2. The rescaled data was first analyzed for each time series individually and then averaged over all grid points and calendar days. The upper figure gives the behavior of the mean value for the entire calendar year. The figure also shows the analytic results for the iid case and for the LDM with a normalized drift of $\tilde{c} = 0.012$, which is determined from the observations. The behaviour of the mean upper and lower record values appear to be very similar. Both curves seem to have exactly the same shape and both lie slightly above the null hypothesis of a stationary climate with Gaussian daily temperatures. These results are not in agreement with the analytic results given by the LDM and do not show any apparent effect of slow temperature increase on the statistics of record values.

The lower half of Fig. 8.10 shows the same analysis but only for two months in summer (July and August) as well as for two months in winter (January and February) in the inset. In summer, we find a much better agreement of the observations with a Gaussian LDM. The mean of the upper records increases much faster than the negative mean of the lower records, i.e., the upper records are more extreme than the lower records. Here, the difference between upper and lower records is in good agreement with the difference predicted from the LDM. In the two winter months (inset), the situation is completely different. Here, the negative mean of the lower records increases faster than the mean of the upper records. It seems, that despite a significant positive trend in the mean values of the daily temperatures, the values of the lower records in winter are still more extreme than those related to the upper records.

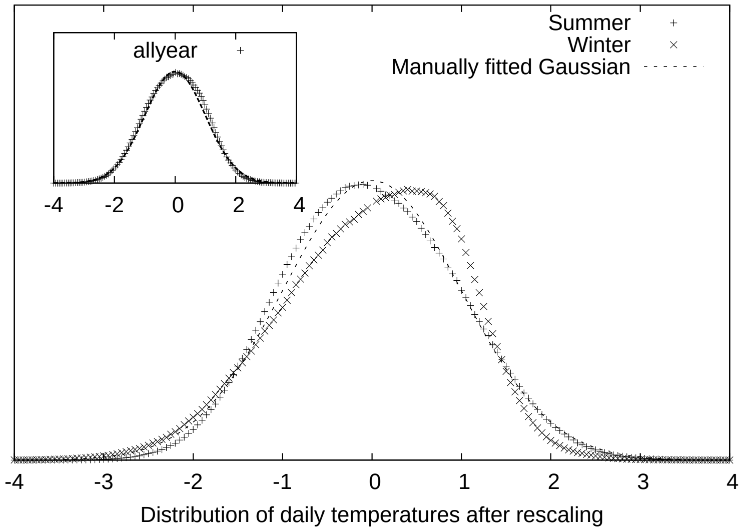

Figure 8.11: Estimated probability density functions of the daily temperature measurements in data set EOBS2 after rescaling to standard deviation unity. Estimation is based on a histogram. The main figure gives the pdf's for the two summer months July and August, as well as the two winter months January and February. The dashed line is a standard normal distribution. The pdf for the summer months is in good agreement with the Gaussian. The pdf for winter deviates significantly from a Gaussian. The inset shows the same analysis for the entire calendar year. Here, the rescaled distribution of daily temperature is again in good agreement with the Gaussian.

With these findings it is clear how to explain the inconsistency of the top half of Fig. 8.10 with the analytical results for the LDM: A strong discrepancy of the behavior in the winter months averages out the effect of a slow positive increase on the record values in summer and leads to the fact that, when averaged over the entire calendar year, the upper and lower record values behave more or less in the same way.

To understand the anomaly in the winter months, we consider the seasonal variability of the pdf's of the daily temperature gridded values. In Fig. 8.11 we show the distributions of daily maximum temperatures in EOBS2 for two months in summer (July and August) and two months in winter (January and February). The distributions were obtained after rescaling of the daily temperatures, as described above, so that they fitted a LDM with standard deviation one. The inset in Fig. 8.11 shows the analysis for the entire calendar year. Apparently the distribution for the winter months in data set EOBS2 is not Gaussian and has a much broader left tail. We believe that this asymmetry of the distribution is responsible for the anomalous behaviour of the mean record values in winter.

In conclusion we find that while the ratio between rescaled upper and negative lower record values in summer is larger than one, it is less than one in winter, because of an asymmetric distribution of daily temperatures. Nevertheless, it might still be possible to describe the record values of the observations with a LDM, but we have shown that the underlying LDM for the winter months can not be based on a a symmetric Gaussian distribution.

To further explore this asymmetry we performed numerical simulations based on the empirical pdf's of daily temperatures in EOBS2 for summer, winter and the entire year. In Fig. 8.12 we show the ratio between upper and lower record values in the data and compare them to the results from simulations with the distribution obtained from the data as well as the predictions from the Gaussian LDM.

For the entire calendar year we find that the ratio remains close to one for the entire observation period, in contrast to a Gaussian LDM. However, if we consider the summer and winter months separately and compare the ratios obtained from the data with the

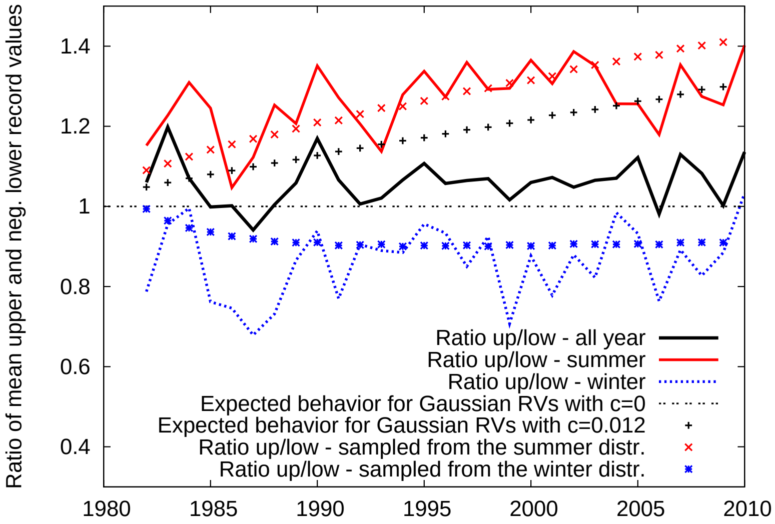

Figure 8.12: Ratio between rescaled upper and negative lower record values in EOBS2 for July and August (red line), January and February (blue line) and the entire calendar year (black line). The black crosses give the behavior of this ratio for a Gaussian LDM with a drift of $\tilde{c} = 0.012\ y^{-1}$. The red and blue crosses give the development of the ratio for the summer and winter month when sampled from the respective observed distributions of daily temperatures in summer and winter. The thin black dashed line gives the behavior one would expect in the case of an iid Gaussian stationary climate.

ratios one obtains from a LDM with RV's sampled from the respective distributions of daily summer and winter temperatures, we find a good agreement with these non-Gaussian LDM's. In summer the ratio is strongly positive with upper records being generally further away from the increasing mean value than lower records. In winter, due to the skewness of the distribution, the situation is reversed and lower record values are shifting to be more extreme.

8.6.2 Distribution of records in a certain year

We also analyzed the full empirical pdf's of daily temperature records in the rescaled temperature data. Figure 8.13 shows the pdf's of temperature records for the entire calendar year (upper figure) and, again, the two considered months in winter and summer (lower figure) that occurred in the last five years of the observation period of EOBS2. We did not normalize the distributions for illustrative purposes. In the upper figure we compared the observational distributions with numerical results from the Gaussian LDM with a normalized drift of $\tilde{c} = 0.012\ y^{-1}$.

We find that while the rescaled upper record values are in good agreement with the Gaussian LDM, the distribution of the lower record value is significantly broader than expected from the simulations. The pdf's of the upper and lower record values seem to have more or less the same mean value, which is in good agreement with our findings in the previous section and Fig. 8.10. The pdf's in this figure are not normalized, the total area under the curves corresponds to the number of records that occurred in the last five years. As a result the upper figure shows that there were many more upper records than lower records in the data, but the shapes of the pdf's of upper and lower records look very similar.

The pdf's in the lower figure are also in good agreement with the mean values in Fig. 8.10. The pdf's for the winter months show that in winter the lower records were more extreme with a mean value further away from the mean of the daily maximum temperatures.

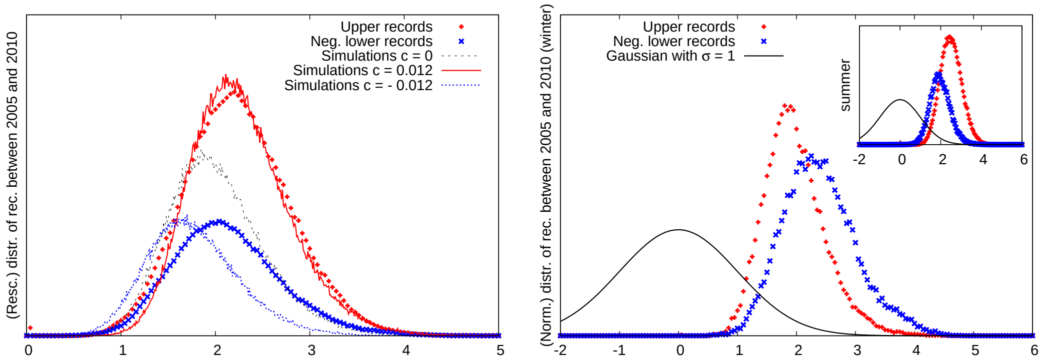

Figure 8.13: Left figure: Estimated pdf's of record values in the rescaled temperature data from EOBS2 for the entire calendar year. The red crosses give the behavior of the upper records and the blue crosses the behavior of the negative cold records. The red and blue lines give the results from numerical simulations with a Gaussian LDM and a drift of $c = 0.012$. The black line is the distribution one expects in a stationary climate ($c = 0$) and Gaussian daily temperature measurements. All distributions are not normalized, the fact that the area under the curve from the upper records is much larger than the area under the corresponding curve from the lower record is a consequence of there being much more upper records than lower records. Note that, for these rescaled temperatures, the origin corresponds to the mean of daily temperatures in the initial year. **Right figure:** The same analysis for the two considered months in winter (January and February) as well as, in the inset, the two months in summer (July and August). Numerical results are not given in this figure. For illustrative purposes we also plotted a Standard Normal distribution (with standard deviation $\sigma = 1$), which in the framework of a Gaussian LDM, represents the rescaled distribution of daily temperatures in the first year. The area under the Gaussian equals the average of the areas under the shown probability densities of the upper and lower records.

Also the width of the pdf of the lower records is larger than in the case of the upper records, so the values of lower records have a large inherent uncertainty. While there are almost no upper records more than 3σ away from the average behavior, there is a large number of lower records that exceeded this barrier, some of them were even beyond 4σ, which, given the large standard deviation of daily temperatures in winter (up to 8 K°), is a huge fluctuation. The inset in the lower figure gives the same plot for the summer. Here, in agreement with a Gaussian LDM, upper records are more extreme than lower records.

8.6.3 Records with a given record number

We also analyzed the statistics of record temperatures in the data set EOBS2 with a given record number k. For this purpose we performed the same rescaling as described above in the context of records with a fixed time of occurrence n. The results of the analysis for two months in winter, in summer and over the entire calendar year in data set EOBS2 can be found in Fig. 8.14. The figure shows the estimated mean value of records in the time series of the rescaled entries $\tilde{T}_i$ (Eq. (8.21)) plotted against the record number k.

This analysis of the mean record value with fixed record number is in good agreement to our findings in the above analysis of the mean record values with fixed occurrence time n. The analysis for the entire calendar year shows that the behaviour of the mean is again comparable for the upper and lower records. If we consider the two winter months January and February (left inset), we find again that cold records are further away from the mean value than the upper records and are therefore more extreme. In July and August (right inset) it is exactly the other way around. Here the LDM based on symmetric, Gaussian RV's works very well and mean record values show the expected asymmetry with more extreme upper records due to the positive trend.

In Fig. 8.15 we also show the estimated pdf's of record values for some selected record

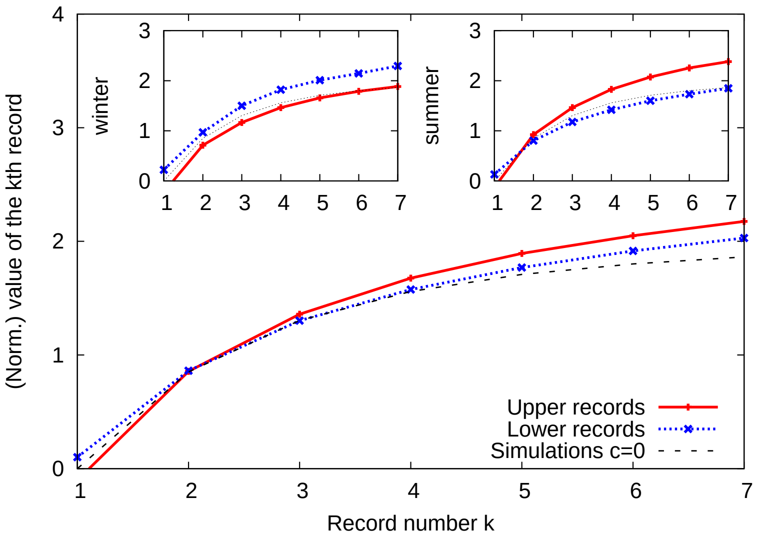

Figure 8.14: Mean values of record-breaking temperatures with a fixed record number k in EOBS2 data using the same rescaling as for the record values with fixed time of occurrence n. The red lines show the time series of the upper record values depending on k, the blue lines the corresponding series of the lower records. The black dashed line shows the growth of the mean of records with record number k for Gaussian RV's without drift. The insets show the same analysis for January and February (left inset) and July and August (right inset).

numbers k and the entire calendar year. Again, the density functions not normalized, so the area under the curves corresponds to the total number of records with a given record number. Apparently, even though in all cases the number of upper records was higher than the number of lower records, the shapes of the pdf's look very similar.

8.6.4 Record values in the far north

A reason for the asymmetry of the daily temperature distribution in winter may be that some parts of Europe, especially in the north, are covered by snow during the winter months. Snow cover decouples the atmosphere from the soil through its isolating effects. Further a snow surface is a very efficient black body radiator. Therefore snow covered surfaces tend to produce very low near surface atmospheric temperatures especially under conditions of small insolation [32].

In Fig. 8.16 we show the pdf's of records that occurred in the last five years of the EOBS2 data only for grid points north of $60°N$. Apparently, here the effect of the asymmetry in the daily temperature pdf's on the record values is even stronger. The estimated density functions of the lower records is centered more than one standard deviation further away from the mean than the density functions of the upper records. The inset shows the pdf of the daily temperatures for this region, which is much more heavily skewed than the pdf for the entire data set of EOBS2 (compare to Fig. 8.11).

8.7 Discussion and conclusion

In this article we analyzed both the probability of occurrence and the probability density function of record breaking temperatures values in Europe and the United States. After presenting some analytical findings on the Linear Drift Model (LDM) in section 2, we analyzed the occurrence of records in daily and monthly data in sections 3 and 4. In agreement with earlier work by [11] we found a significant effect of the observed slow

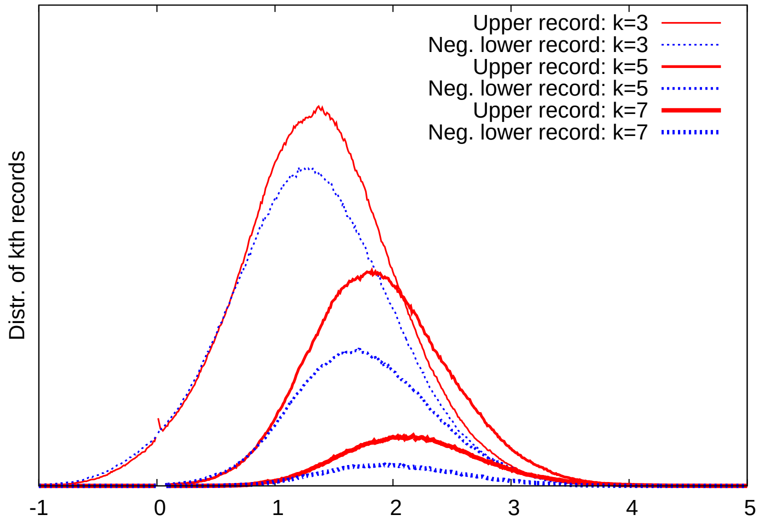

Figure 8.15: Estimated pdf's of record values with a given record number k of the rescaled daily maximum temperatures from EOBS2. The red lines give the pdf's of upper records with record numbers $k = 3, 5$ and 7. The blue lines give the pdf's of the negative lower records with the same record numbers k. Note that, as in Fig. 8.13, the density functions are not normalized.

positive trend of temperature on the number of upper and lower records. This effect can be described by the LDM up to some accuracy. The effect of increasing temperatures on the monthly mean temperatures is clearly stronger than the effect on the daily measurements because of the smaller standard deviation of those averages. During certain seasons in the United States we found up to nine times more upper records than lower ones at the end of the 51 years of observation 1960-2010. This also explains the findings of Rahmstorf and Coumou [15], who studied monthly and annual mean temperatures and found the strongest effect of warming for the annual global-mean temperature. The global mean has a very small standard deviation of $\sigma = 0.088$ K° which implies that already a very small drift can measurably increase the upper record rate.

In section 5 we presented new results on the mean value of a record that occurs at a certain time in the LDM. We were able to obtain analytical results of the effect of a linear drift on the mean value of a record in case of Gaussian RV's in the important regime of $cn \ll \sigma$. In section 6 we compared these results to the gridded data set EOBS2. When analyzing the entire calendar year, a Gaussian LDM fails to describe the effect of slowly increasing temperatures on the mean record values. The reason for this failure is a pronounced asymmetry of the daily temperature probability density function in winter. In contrast the record values during the summer months can be well described by a Gaussian LDM.

When we use a non-Gaussian LDM with RV's sampled from the daily temperature probability density of the considered winter months, this model is also capable of describing temperature record values in winter. As a general result the lower records in winter are significantly more extreme than upper records because of this asymmetry. This leads to the unintuitive results that lower records of near surface temperature in the case of slowly increasing temperatures occur with a reduced probability but once they occur they are more extreme than their upper counterpart due to the pronounced asymmetry of the daily temperature values in winter. When we consider grid points in the northern parts of Europe this effect is even stronger. Here lower records are on average more than one standard deviation further away from the mean than upper records and the estimated density of lower records is also clearly broader than that of the upper records.

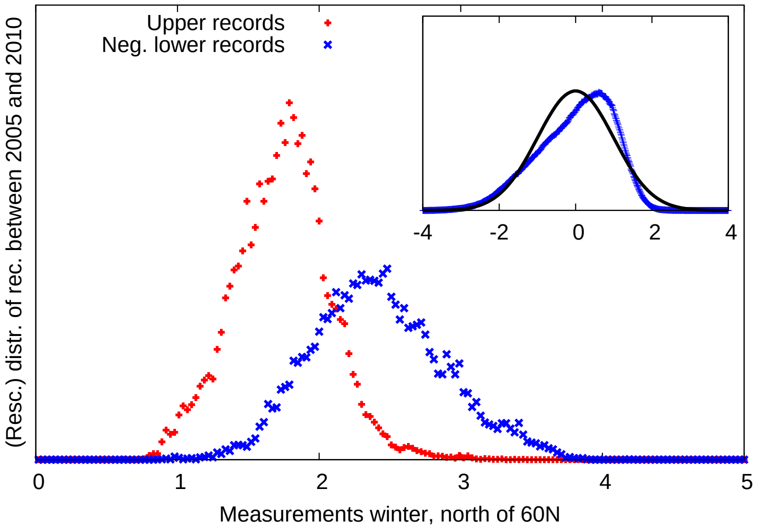

Figure 8.16: Estimated pdf of record values in the rescaled temperature data from EOBS2 for the months of January and February in the regions north of $60°N$. The red crosses give the behaviour of the upper records and the blue crosses the behaviour of the lower records. The inset shows the estimated pdf of rescaled daily temperatures in these two months and in the region north of $60°N$. Note that, as in Fig. 8.13, the density functions are not normalized.

It might be interesting to further explore the effect of specific weather conditions especially in northern Europe, but also in other regions, on the statistics of record values. In particular it remains an open question, how strongly the occurrence of very extreme lower records in winter correlates with snow coverage and other meteorological events, such as winter blocking highs. It is however clear that the fact, that we still encounter quite extreme cold streaks in winter, particularly in northern regions, is not in contradiction with global warming and only a consequence of the very skew distribution of daily temperatures in winter.

Acknowledgements

GW is grateful for financial support provided by DFG within the Bonn Cologne Graduate School of Physics and Astronomy as well as the Friedrich-Ebert-Stiftung. We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES and the data providers in the ECA&D project.

Bibliography

[1] S. Solomon et al (Editors), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2007).

[2] D. R. Easterling et al, Science **277**, 364 (1997).

[3] P. A. Stott, D. A. Stone, and M. R. Allen, Nature **432**, 610 (2004).

[4] S. J. Brown, J. Caesar, and C. A. T. Ferro, J. Geophys. Res. **113**, D05115 (2008).

[5] J. Cattiaux, R. Vautard, and P. Yiou, Geophys. Res. Lett. **36**, L06713 (2009).

[6] G. A. Meehl, J. M. Arblaster, and C. Tebaldi, Geophys. Res. Lett. **34**, L19709 (2007).

[7] S.-K. Min, Clim. Dynam. **32**, 95 (2009).

[8] S. Redner and M. R. Peterson, Phys. Rev. E **74**, 061114 (2006).

[9] R. E. Benestad, Climate Res. **25**, 3 (2003).
[10] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).
[11] G. Wergen and J. Krug, EPL **92**, 30008 (2010).
[12] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).
[13] N. Elguindi, S. A. Rauscher, and F. Giorgi, Climatic Change (2012).
[14] W. I. Newman, B. D. Malamud, and D. L. Turcotte, Phys. Rev. E **82**, 066111 (2010).
[15] S. Rahmstorf and D. Coumou, P. Natl. Acad. Sci. USA **108**, 17905 (2011).
[16] M. R. Haylock *et al.*, J. Geophys. Res.-Atmos. **113**, D20119 (2008).
[17] B. C. Arnold, N. Balakrishnan, and H. N. Nagraja, *Records*, 1st ed. (Wiley-Interscience, 1998).
[18] N. Glick, Am. Math. Mon. **85**, 2 (1978).
[19] V. B. Nevzorov, *Records: Mathematical Theory* (American Mathematical Society, 2000).
[20] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011).
[21] P. Sibani and P. B. Littlewood, Phys. Rev. Lett. **71**, 1482 (1993).
[22] R. Ballerini and S. Resnick, J Appl. Probab. **22**, 487 (1985).
[23] R. Ballerini and S. Resnick, Adv. Appl. Probab. **19**, 801 (1987).
[24] K. Borovkov, J Appl. Probab. **36**, 668 (1999).
[25] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).
[26] J. Galambos, J. Lechner, and E. Simiu, *Extreme Value Theory and Applications: Vol 1* (Kluwer Academic Publ, 1994).
[27] L. De Haan and A. Ferreira, *Extreme Value Theory: An Introduction* (Springer, 2006).
[28] E. J. Gumbel, Natl. Bur. St. Appl. Math. Ser. **33** (1954).
[29] G. Wergen *et al.*, Phys. Rev. Lett. **109**, 164102 (2012).
[30] M. J. Menne, C. N. Williams Jr., and R. S. Vose, *United States historical climatology network daily temperature, precipitation, and snow data* (National Climatic Data Center, National Oceanic and Atmospheric Administration, 2010).
[31] G. Wergen, “unpublished,” (2011).
[32] R. Geiger, R. H. Aron, and P. Todhunter, *The Climate Near the Ground* (Vieweg + Teubner, 1995).
[33] J. Krug, J. Stat. Mech.: Theor. Exp. **P07001** (2007).
[34] S. M. Uppala et al, Q. J. R. Meteorol. Soc. **131**, 2961 (2005).
[35] C. N. Williams Jr. *et al.*, Technical Report, Carbon Dioxide InformationAnalysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, (2006).
[36] Project Team ECAD, Technical Report, Royal Netherlands Meteorological Institute, KNMI, (2011).
[37] M. J. Menne, C. N. Williams, Jr., and R. S. Vose, United States Historical Climatology Network (USHCN) Version 2 Serial Monthly Dataset. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee. (2010).
[38] R. Corless, Adv. Comput. Math. **5**, 329 (1996).
[39] H. Prodinger, Discrete Math. **254**, 459 (2002).
[40] G. Wergen, “Diploma Thesis,” (2009).

Part III

Record statistics of random walks

9 Record statistics for biased random walks, with an application to financial data 117
9.1 Introduction. 118
9.2 Survival probabilities & first passage times 119
9.3 Gaussian random walks with drift 121
  9.3.1 Record rate for small c/σ and $n \ll (\sigma/c)^2$ 121
  9.3.2 Asymptotic record rate for large n 123
9.4 Record statistics of stock prices in the S&P 500 124
9.5 Summary. 126

10 Record statistics and persistence for a random walk with a drift 129
10.1 Introduction. 130
10.2 Record statistics for random walks: Review & new results 132
10.3 Record number distribution via renewal property 139
10.4 Asymptotic behavior of persistence $Q(n)$ for large n 141
  10.4.1 Regime I: $0 < \mu < 1$ and c arbitrary 143
  10.4.2 Regime II: $\mu = 1$ and c arbitrary 144
  10.4.3 Regime III: $1 < \mu < 2$ and $c > 0$ 145
  10.4.4 Regime IV: $\mu = 2$ and $c > 0$ 146
  10.4.5 Regime V: $1 < \mu \leq 2$ and $c < 0$ 147
10.5 Asymptotic record number distribution $P(R, n)$ for large n. 147
  10.5.1 Regime I: $0 < \mu < 1$ and c arbitrary 148
  10.5.2 Regime II: $\mu = 1$ and c arbitrary 149
  10.5.3 Regime III: $1 < \mu < 2$ and $c > 0$ 151
  10.5.4 Regime IV: $\mu = 2$ and $c > 0$ 153
  10.5.5 Regime V: $1 < \mu \leq 2$ and $c < 0$ 156
10.6 Extreme statistics of the age of a record 157
  10.6.1 Age of the longest lasting record $l_{\max,n}$ 157
    10.6.1.1 Regime I: $0 < \mu < 1$, c arbitrary: 158
    10.6.1.2 Regime II: $\mu = 1$ and c arbitrary:. 159
    10.6.1.3 Regime III: $1 < \mu < 2$ and $c > 0$: 160
    10.6.1.4 Regime IV: $\mu = 2$ and $c > 0$: 160
    10.6.1.5 Regime V: $1 < \mu \leq 2$ and $c < 0$: 162
  10.6.2 Age of shortest lasting record $l_{\min,n}$ 163
    10.6.2.1 Regime I: $0 < \mu < 1$ and c arbitrary. 164
    10.6.2.2 Regime III: $1 \leq \mu < 2$ and $c > 0$ 165
    10.6.2.3 Regime II: $\mu = 1$ and c arbitrary 165
    10.6.2.4 Regime IV: $\mu = 2$ and $c > 0$. 166
    10.6.2.5 Regime V: $1 < \mu \leq 2$ and $c < 0$ 166
10.7 Conclusion 167

11 Record statistics for multiple random walks **173**
11.1 Introduction. . . . 174
11.2 The model & the main results 176
11.3 Mean number of records for multiple walkers 178
11.4 The distribution of the number of records for finite σ^2 184
11.5 Numerical simulations 187
11.5.1 Statistics of the record numbers 187
11.5.2 Temporal evolution of two stochastic processes 189
11.5.3 Correlations between record events 191
11.6 Comparison to stock prices 191
11.7 Conclusion 193

Chapter 9

Record statistics for biased random walks, with an application to financial data

Gregor Wergen, Miro Bogner and Joachim Krug

Institute for Theoretical Physics, University of Cologne

Published in: Phys. Rev. E 83, 051109 (2011)

Abstract: We consider the occurrence of record-breaking events in random walks with asymmetric jump distributions. The statistics of records in symmetric random walks was previously analyzed by Majumdar and Ziff [1] and is well understood. Unlike the case of symmetric jump distributions, in the asymmetric case the statistics of records depends on the choice of the jump distribution. We compute the record rate $P_n(c)$, defined as the probability for the nth value to be larger than all previous values, for a Gaussian jump distribution with standard deviation σ that is shifted by a constant drift c. For small drift, in the sense of $c/\sigma \ll n^{-1/2}$, the correction to $P_n(c)$ grows proportional to $\arctan(\sqrt{n})$ and saturates at the value $\frac{c}{\sqrt{2}\sigma}$. For large n the record rate approaches a constant, which is approximately given by $1-\left(\sigma/\sqrt{2\pi}c\right)\exp\left(-c^2/2\sigma^2\right)$ for $c/\sigma \gg 1$. These asymptotic results carry over to other continuous jump distributions with finite variance. As an application, we compare our analytical results to the record statistics of 366 daily stock prices from the Standard & Poors 500 index. The biased random walk accounts quantitatively for the increase in the number of upper records due to the overall trend in the stock prices, and after detrending the number of upper records is in good agreement with the symmetric random walk. However the number of lower records in the detrended data is significantly reduced by a mechanism that remains to be identified.

9.1 Introduction

The random walk is a paradigmatic model of statistical physics, which combines utmost conceptual simplicity with a surprising richness of emergent behaviors [2, 3]. Among the many interesting features of random walks, recent research has focused in particular on its extremal properties, exploring quantities such as the height and position of the globally maximal excursion of a one-dimensional walk of a given number of steps, and the statistics of *records* of this process [4]. Here a record is defined as an entry in a discrete, real valued time series that is larger (upper record) or smaller (lower record) than all previous entries. While the mathematical theory of records is well developed for time series of independent, identically distributed random variables [5–7], little has been known about the record statistics of correlated processes. It is therefore remarkable that records of a large class of one-dimensional random walks can be characterized in considerable detail, as was shown in recent work by Majumdar and Ziff (MZ) [1]. Specifically, they considered the random process defined by

$$X_n = X_{n-1} + \xi_n, \tag{9.1}$$

where $X_0 = 0$ (say) and the step sizes ξ_n are independent, identically distributed random variables drawn from a probability density $\phi(\xi)$ that is required to be continuous and symmetric, but is otherwise arbitrary. We say that an upper record occurs at time n if $X_n = \max\{X_0, ..., X_n\}$. Based on the Sparre Andersen theorem for the survival probability of the random walk [8–11], MZ show that the probability $\Pi(m,n)$ for the nth event to be the mth record is given by

$$\Pi(m,n) = \binom{2n-m+1}{m} 2^{-2n+m-1} \tag{9.2}$$

for $m \leq n+1$. The first moment of this distribution with respect to m yields the mean number of records after n steps, which equals $m_n \approx \frac{2}{\sqrt{\pi}}\sqrt{n}$ for large n, and the probability P_n for the nth event to be a record (henceforth referred to as the *record rate*) decays like $P_n \approx \frac{1}{\sqrt{\pi n}}$. In the present paper we aim to generalize these results to random walks with asymmetric jump distributions. In the first part of the paper (Sections 9.2 and 9.3) we study records generated by random walks with a symmetric jump distribution that have an additional constant *drift* c, such that (9.1) generalizes to

$$X_n = X_{n-1} + \xi_n + c \tag{9.3}$$

with a symmetric jump distribution $\phi(\xi)$. For the special case of a Cauchy distribution this problem was considered previously in [12]. Here, we derive approximate results for the case of a Gaussian jump distribution that apply also more generally to distributions with a finite variance.

Similar to our earlier work [13, 14] on the related problem of records from independent random variables with drift [12, 15], our strategy will be to analyze the limiting cases of small and large drift, respectively, as quantified by the ratio c/σ of the drift speed to the standard deviation σ of the jump distribution $\phi(\xi)$. For the Gaussian random walk we find that in the limit of $\frac{c}{\sigma} \ll \frac{1}{\sqrt{n}}$ the mean number of records and the record rate are given by

$$m_n(c) \approx \frac{2\sqrt{n}}{\sqrt{\pi}} + \frac{c}{\sigma}\frac{\sqrt{2}}{\pi}\left(n \arctan\left(\sqrt{n}\right) - \sqrt{n}\right), \tag{9.4}$$

$$P_n(c) \approx \frac{1}{\sqrt{\pi n}} + \frac{c}{\sigma}\frac{\sqrt{2}}{\pi}\arctan\left(\sqrt{n}\right). \tag{9.5}$$

In the limit of $\frac{c}{\sigma} \gg \frac{1}{\sqrt{n}}$ the record rate $P_n(c)$ approaches a constant value. If in addition $\frac{c}{\sigma} \gg 1$, this constant is given approximately by

$$\lim_{n\to\infty} P_n \approx 1 - \frac{c}{\sqrt{2\pi}\sigma} e^{-\frac{c^2}{2\sigma^2}}. \tag{9.6}$$

In Section 9.4 we apply our results to fluctuations in stock prices, arguably one of the most important (and ancient) applications of random walk theory [16–18]. The basic model of a stock price is the geometric random walk $S_n = e^{X_n}$ with an upward bias reflecting long-term economic growth. Our analysis of record events in the Standard & Poors 500 index shows a corresponding surplus of upper record events, which is consistent with the theoretical expectation. However, an asymmetry between upper and lower records remains even when the bias has been (approximately) removed [19], a feature that may be related to the gain-loss asymmetry reported in previous analyses of stock market fluctuations [20–23]. We conclude with a summary and a discussion of some open problems.

9.2 Survival probabilities & first passage times

The record statistics of a random walk can be analyzed by considering the generating functions of the survival and first passage probabilities of the process [1, 4, 12]. In [1] it was shown that the generating function of $\Pi(m,n)$ is of the form

$$\sum_{n=m-1}^{\infty} \Pi(m,n)\, z^n = \tilde{f}_-^{m-1}(z)\, \tilde{q}_-(z), \tag{9.7}$$

where $\tilde{q}_\pm(z)$ is the generating function of the positive (negative) survival probability $q_\pm(n)$ of the random walk. The latter is defined as the probability that the process stays above (below) the origin up to the nth step. Similarly $\tilde{f}_\pm(z)$ is the generating function of the positive (negative) first-passage probability $f_\pm(n)$ of the random walk, with $f_\pm(n) = q_\pm(n-1) - q_\pm(n)$. In the case of the symmetric random walk considered in [1] we have $q_-(n) = q_+(n) = q(n)$ and $f_-(n) = f_+(n) = f(n)$ and both $q(n)$ and $f(n)$ are completely universal for all continuous jump distributions.

Since we want to study asymmetric random walks, we need distinguish between positive and negative survival probabilities and first passage times, and consider the functions $q_\pm(n)$ and $f_\pm(n)$. As in [1] a theorem by Sparre Andersen will play a key role in our considerations. In [8, 11] it was shown that

$$\tilde{q}_\pm(z) = \sum_{n=0}^{\infty} q_\pm(n)\, z^n = \exp\left(\sum_{n=1}^{\infty} \frac{p_\pm(n)}{n} z^n\right), \tag{9.8}$$

where $p_\pm(n)$ is the probability for the walker to be above or below the origin at the nth step. This quantity can be easily computed from $p_\pm = \int_0^\infty G(\pm x, n)\,\mathrm{d}x$, where $G(x,n)$ is the positional probability density of a random walk of n steps that started at the origin. Details on the computation of $G(\pm x, n)$ and $p_\pm(n)$ can be found in [4] and [10]. In the case of a symmetric random walk we simply have $p_\pm(n) = \frac{1}{2}$ independent of n and we find that in this case $\tilde{q}_\pm(z) = (1-z)^{-\frac{1}{2}}$ and $q_\pm(n) = \binom{2n}{n} 2^{-2n}$ [1]. These results eventually lead to Eq. (9.2) [1].

In the case of an asymmetric random walk the situation gets a bit more complicated. We compute $q_\pm(n)$ and its generating function for a Gaussian random walk with drift c. Here, the jump distribution of the symmetric random variable ξ in (9.3) is of the form $\phi(\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\xi^2}{2\sigma^2}}$ with standard deviation σ. It is easy to show that the probability density $G(x,n)$ of the random walk after n steps is given by $G(x,n) = \frac{1}{\sqrt{2\pi n}} \exp\left(-\frac{(x-nc)^2}{2\sigma^2 n}\right)$ and $p_\pm(n) = \frac{1}{2}\left(1 \pm \operatorname{erf}\left(\sqrt{\frac{n}{2}}\frac{c}{\sigma}\right)\right)$. We start with the case of a small linear drift with $c \ll \frac{\sigma}{\sqrt{n}}$ such that $p_\pm(n) \approx \frac{1}{2} \pm \sqrt{\frac{n}{2\pi}}\frac{c}{\sigma}$. Now we can employ Eq. (9.8). Expanding up to first order in c we find

$$\tilde{q}_\pm(z) \approx \frac{1}{\sqrt{1-z}} \left(1 \pm \frac{c}{\sqrt{2\pi}\sigma} \sum_{n=1}^{\infty} \frac{z^n}{\sqrt{n}}\right). \tag{9.9}$$

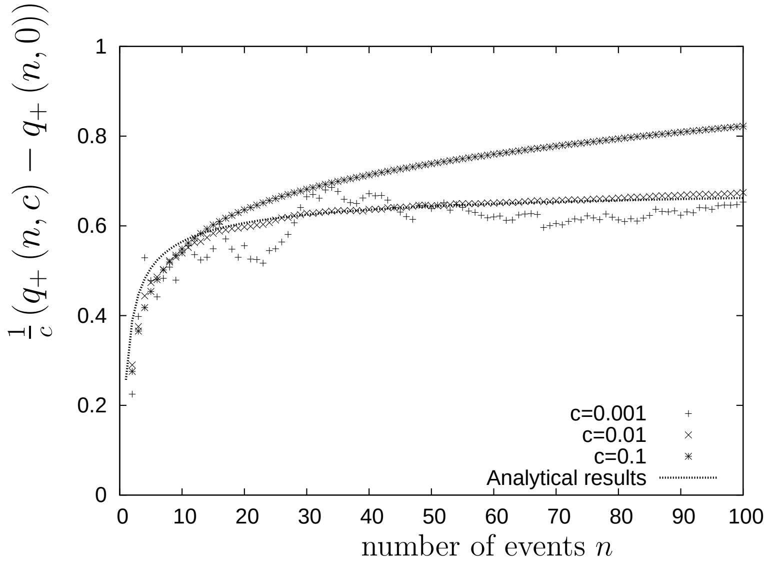

Figure 9.1: Relative effect of the drift on the positive survival probability $q_+(n,c)$ of a Gaussian random walk with $\sigma = 1$. The effect of the drift is represented by $\frac{1}{c}\left(q_+(n,c) - q_+(n,0)\right)$ for different drift speeds c. We simulated 10^7 realization of a random walk with $n = 100$ steps for each drift speed. The dotted line represents the analytical results obtained in Eq. (9.13). For small drift $c = 0.001$ and $c = 0.01$ we find good agreement with this approximation.

With $\sqrt{1-z}^{-1} = \sum_{n=0}^{\infty}\binom{2n}{n}2^{-2n}z^n$ and making use of the Cauchy formula for products of infinite sums we obtain the following expression for $\tilde{q}_\pm(z)$:

$$\tilde{q}_\pm(z) \approx \frac{1}{\sqrt{1-z}} \pm \frac{c}{\sqrt{2\pi}\sigma}\sum_{n=0}^{\infty}\sum_{k=0}^{n}\binom{2k}{k}\frac{2^{-2k}z^{n+1}}{\sqrt{n-k+1}}. \tag{9.10}$$

The binomial coefficient can be approximated by $\binom{2k}{k} \approx 4^k/\sqrt{\pi k}$ and with this we can approximate the sum over k by an integral,

$$\sum_{k=0}^{n}\frac{1}{\sqrt{\pi k}}\frac{z^{n+1}}{\sqrt{n-k+1}} \approx \frac{1}{\sqrt{\pi}}\int_0^n \frac{\mathrm{d}k z^{n+1}}{\sqrt{k}\sqrt{n-k+1}} \approx \sqrt{\pi}z^{n+1}. \tag{9.11}$$

We thus obtain a simple result for the generating function of the survival probability,

$$\tilde{q}_\pm(z) \approx \frac{1}{\sqrt{1-z}} \pm \frac{c}{\sqrt{2}\sigma}\sum_{n=1}^{\infty} z^n, \tag{9.12}$$

and finally the following expression for the survival probability $q_\pm(n)$ under a small linear drift:

$$q_\pm(n) \approx \frac{1}{\sqrt{\pi n}} \pm \frac{c}{\sqrt{2}\sigma}. \tag{9.13}$$

The first term on the right hand side is the result for the symmetric random walk discussed in [1], which is now supplemented by a correction linear in $\frac{c}{\sigma}$. Although this particular result for $q_\pm(n)$ will not be needed in our derivation of the record statistics, we found it useful to test it against numerical simulations. The results are shown in Fig. 9.1. For small c, Eq.(9.13) is in good agreement with the simulations.

For the sake of completeness we also provide the small c expansion of $\tilde{f}_\pm(z)$ that will become important later. With $\tilde{f}_\pm(z) = 1 - (1-z)\,q_\pm(z)$ we find

$$\tilde{f}_\pm(z) \approx 1 - \sqrt{1-z}\left(1 \pm \frac{c}{\sqrt{2\pi}\sigma}\sum_{n=1}^{\infty}\frac{z^n}{\sqrt{n}}\right). \tag{9.14}$$

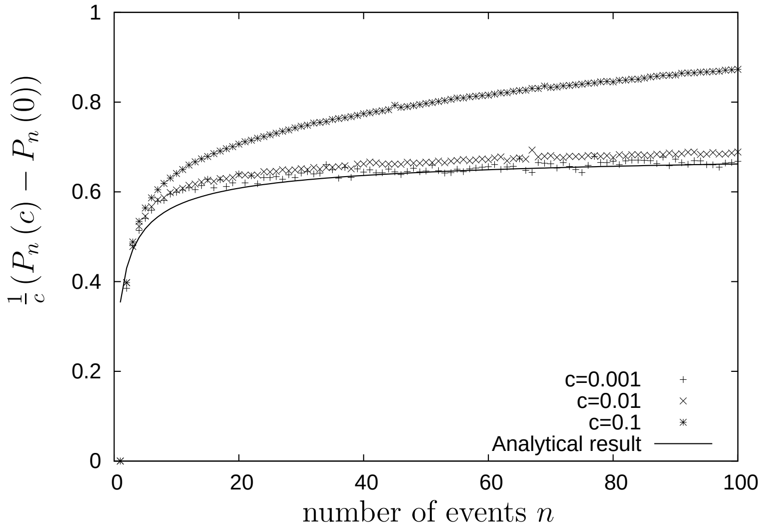

Figure 9.2: Relative effect of the drift on the record rate $P_n\left(c\right)$ of a random walk with a Gaussian jump distribution ($\sigma=1$). The effect is represented by $\frac{1}{c}\left(P_n\left(c\right)-P_n\left(0\right)\right)$ for different drift speeds c. Again, we simulated 10^7 realizations of a random walk with $n=100$ steps for each drift speed. The line represents the analytical results obtained in Eq. (9.5). For small drift speeds $c=0.001$ and $c=0.01$ we find good agreement with the approximation, but for $c=0.1$ the approximation is no longer accurate.

From this we obtain, by methods very similar to those used above to derive $q_{\pm}\left(n\right)$, the result

$$f_{\pm}\left(n\right)\approx\frac{1}{2\sqrt{\pi}}n^{-\frac{3}{2}}\pm\frac{c}{\sqrt{2\pi}\sigma}n^{-\frac{1}{2}}. \tag{9.15}$$

Next we consider the case of large drift, $\frac{c}{\sigma}\gg1$. Here we will only discuss $q_{-}\left(n\right)$, as this is the quantity needed for the computation of the record rate; $q_{+}\left(n\right)$ has a different behavior in this regime. In the limit of $\frac{c}{\sigma}\gg1$ we find that $p_{-}\left(n\right)\approx\sigma\left(2\pi nc^2\right)^{-1/2}e^{-\frac{c^2n}{2\sigma^2}}$. Using this we find that for large n, $\tilde{q}_{-}\left(z\right)$ and $q_{-}\left(n\right)$ are of the form

$$\begin{aligned}\tilde{q}_{-}\left(z\right) &\approx 1+\sum_{n=1}^{\infty}\frac{\sigma}{c\sqrt{2\pi n^3}}e^{-\frac{c^2n}{2\sigma^2}}z^n, & (9.16)\\ q_{-}\left(n\right) &\approx \frac{\sigma}{c\sqrt{2\pi n^3}}e^{-\frac{c^2n}{2\sigma^2}}. & (9.17)\end{aligned}$$

These particular results were already reported in [12]. At this point it is important to notice that all results concerning the first-passage and survival probabilities in the large n limit are easily transferable to other jump distributions as long as these have a finite variance. Because of the central limit theorem, $G\left(\pm x,n\right)$ and therefore $p_{\pm}\left(n\right)$ will approach the same expressions for large n as were derived here for the Gaussian jump distribution.

9.3 Gaussian random walks with drift

9.3.1 Record rate for small c/σ and $n\ll\left(\sigma/c\right)^2$

With the small c expansions in Eqs. (9.9) and (9.14) we have all ingredients needed to derive the record statistics for a Gaussian random walk with a small linear drift. We start by computing the mean number of records m_n expected up to the nth step. For

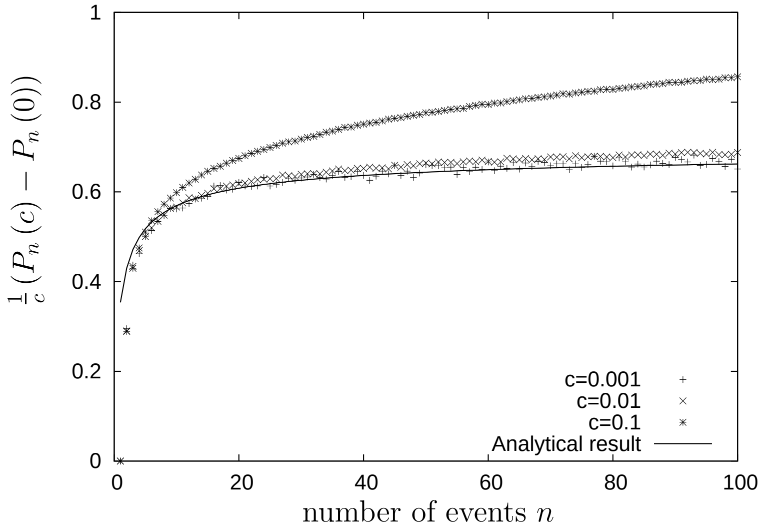

Figure 9.3: Relative effect of the drift on the record rate $P_n\left(c\right)$ for a random walk with a uniform jump distribution with standard deviation $\sigma = 1$. The parameters of the simulation are the same as in Fig. 9.2. Even though the expression (9.5) was derived for a Gaussian jump distribution, it is in a good agreement with the numerical results for small c.

the generating function $\tilde{m}(z) = \sum_{n=0}^{\infty} m_n z^n$ of this quantity it was found in [12] that $\tilde{m}\left(z\right) = 1/\left(\left(1-z\right)^2 \tilde{q}_-\left(z\right)\right)$, a result that can be obtained by computing the first moment of Eq. (9.7). We can now evaluate this expression making use of the generating function for $q_-\left(n\right)$ given in Eq.(9.9). In the limit of small $\frac{c}{\sigma}$ this yields

$$\tilde{m}\left(z\right) \approx \frac{1}{\sqrt{1-z}^3}\left(1+\frac{c}{\sqrt{2\pi}\sigma}\sum_{n=1}^{\infty}\frac{z^n}{\sqrt{n}}\right). \tag{9.18}$$

Using the series expansion of $\sqrt{1-z}^{-3}$ and employing once again the Cauchy formula for infinite sums and the Stirling approximation, we find

$$\tilde{m}\left(z\right) \approx \frac{1}{\sqrt{1-z}^3} + \frac{\sqrt{2}c}{\pi\sigma}\sum_{n=1}^{\infty} z^n \sum_{k=0}^{n-1}\frac{\sqrt{k}}{\sqrt{n-k+1}}. \tag{9.19}$$

If n is not too small, the sum over k can be replaced by an integral and we finally obtain an approximate expression for the generating function of m_n,

$$\tilde{m}\left(z\right) \approx \frac{1}{\sqrt{1-z}^3} + \frac{\sqrt{2}c}{\pi\sigma}\sum_{n=1}^{\infty} z^n \left(n \arctan\left(\sqrt{n}\right) - \sqrt{n}\right). \tag{9.20}$$

The mean number of records of the random walk with a small linear drift c is therefore approximately given by

$$m_n \approx \binom{2n}{n}\frac{2n+1}{2^{2n}} + \frac{\sqrt{2}c}{\pi\sigma}\left(n \arctan\left(\sqrt{n}\right) - \sqrt{n}\right). \tag{9.21}$$

Making use of the Stirling approximation this yields the previously announced expression (9.4) for $m_n(c)$ and, by taking a derivative with respect to n, the record rate $P_n\left(c\right)$ in the large n limit as given in Eq.(9.5). The leading order correction of the record rate due to the drift is seen to increase with $\arctan\left(\sqrt{n}\right)$ and for larger n it approaches a constant value.

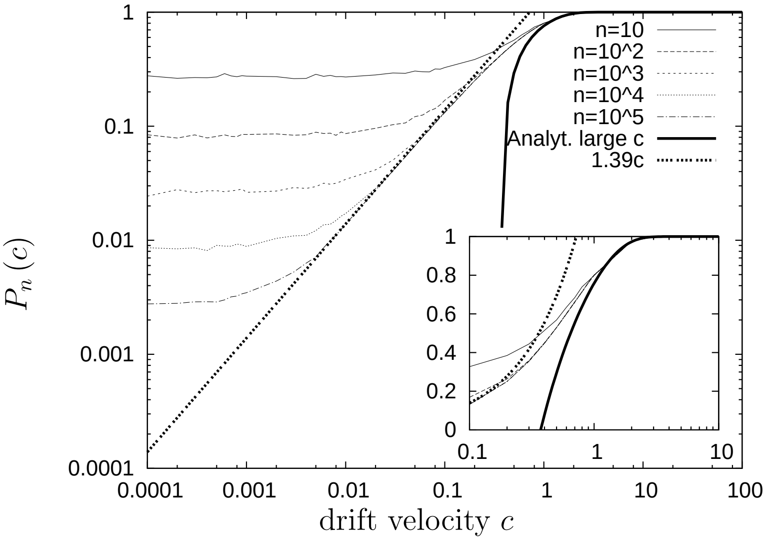

Figure 9.4: Record rate for a biased Gaussian random walk with standard deviation $\sigma = 1$. The figure illustrates the convergence of $P_n(c)$ to the asymptotically constant record rate $P(c)$ for $n \to \infty$. The inset shows that the large drift result (9.24) becomes accurate for $c/\sigma > 1$, and the bold dotted line shows that $P(c) \approx 1.39\frac{c}{\sigma}$ for $c \to 0$.

For large n (but still in the regime $\frac{c}{\sigma} \ll \frac{1}{\sqrt{n}}$) we find the simple result

$$P_n(c) \approx \frac{1}{\sqrt{\pi n}} + \frac{c}{\sqrt{2}\sigma}. \tag{9.22}$$

We compared Eq.(9.5) to simulations and found good agreement in the regime $\frac{c}{\sigma} \ll \frac{1}{\sqrt{n}}$ (Fig. 9.2). We also compared this result with numerical simulations of the record rate for random walks with step sizes drawn from a uniform distribution (Fig. 9.3). The results for the Gaussian and the uniform distribution are very similar to each other already for small n, reflecting the convergence expected from the central limit theorem.

9.3.2 Asymptotic record rate for large n

Next we consider the limit of strong drift, $\frac{c}{\sigma} \gg 1$. Applying the same method as above and making use of our result (9.16) for $\tilde{q}_-(z)$ in the regime of large c/σ, we find that the number of records increases linearly with time according to

$$m_n(c) \approx n\left(1 - \frac{\sigma}{\sqrt{2\pi}c}e^{-\frac{c^2}{2\sigma^2}}\right). \tag{9.23}$$

Correspondingly the record rate P_n is independent of n in this case. In fact, simulations show that the record rate approaches a finite, nonzero limit $P(c) \equiv \lim_{n\to\infty} P_n(c)$ for $n \to \infty$ for any positive value of the drift (Fig. 9.4). This can be understood, on the basis of the general relation (9.7) between the distribution of record events and the negative first passage probability, to be a consequence of the fact that the negative mean first passage time of a random walk with positive drift is finite [9, 10]; roughly speaking, one expects that the asymptotic record rate $P(c)$ is proportional to the inverse of the negative mean first passage time. The result (9.23) implies that the asymptotic record rate behaves as

$$P(c) \approx 1 - \frac{\sigma}{\sqrt{2\pi}c}e^{-\frac{c^2}{2\sigma^2}} \tag{9.24}$$

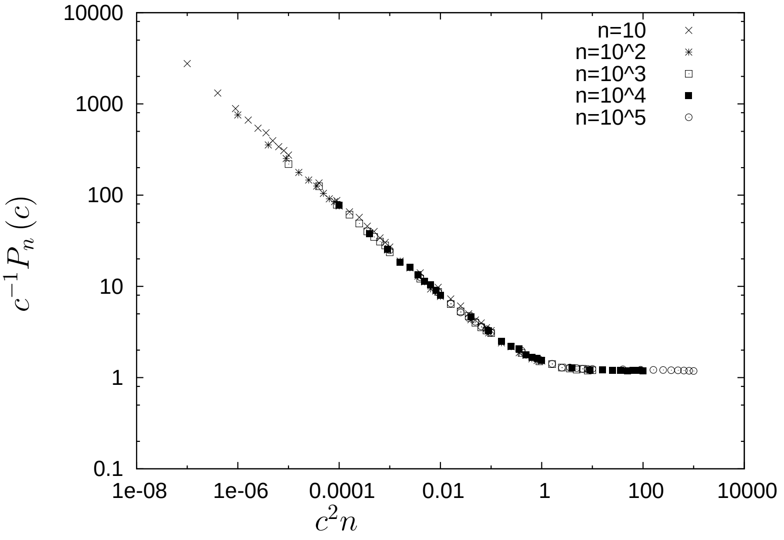

Figure 9.5: Illustration of the conjectured scaling collapse (9.26) of the record rate $P_n(c)$ for Gaussian random walks with $\sigma = 1$ and various drift speeds $c \leq 0.1$.

for large c/σ (see inset of Fig.9.4). Furthermore, since the negative mean first passage time diverges as c^{-1} for $c \to 0$ [9, 10], the asymptotic record rate should behave as $P(c) \sim c$ for small c. This is confirmed by the simulations, which indicate that $P(c) \approx 1.39\ (c/\sigma)$ for $c/\sigma \ll 1$.

The time scale $n^*(c)$ at which the saturation of the record rate occurs can be estimated by comparing the two terms in Eq.(9.22), which shows that

$$n^* \sim \left(\frac{\sigma}{c}\right)^2 \tag{9.25}$$

for small c. Not surprisingly, this is also the time scale at which the drift begins to dominate the mean square displacement of the random walk. Together with the linear behavior of the asymptotic record rate, this suggests the scaling form

$$P_n(c) = \frac{c}{\sigma} g((c/\sigma)^2 n) \tag{9.26}$$

for small c/σ and arbitrary n, where the limiting behaviors of the scaling function are $g(x \to 0) \approx \frac{1}{\sqrt{\pi x}}$ and $g(x \to \infty) \approx 1.39$. This relation is well fulfilled by the numerical data shown in Fig.9.5.

9.4 Record statistics of stock prices in the S&P 500

A prominent application of the random walk process can be found in the financial sciences. Originally introduced by Bachelier in 1900 [16], the geometric random walk is the standard model used to describe the evolution of stock prices. In the application of this model to actual data, trends are always an issue, which in the simplest case are described by a linear drift in the logarithm of the stock price. In this section we present an empirical analysis of record events in historical stock prices taken from the Standard & Poors 500 index, and compare the results to the theoretical predictions derived above. The observational data we used consist of daily recordings of 366 stocks that were contained in the index from January 1990 to March 2009, resulting in 366 time series of length $n = 5000$ [24]. We first analyzed the recordings without any detrending and then considered detrended data in which a fitted linear trend was subtracted from the logarithms of the stock prices.

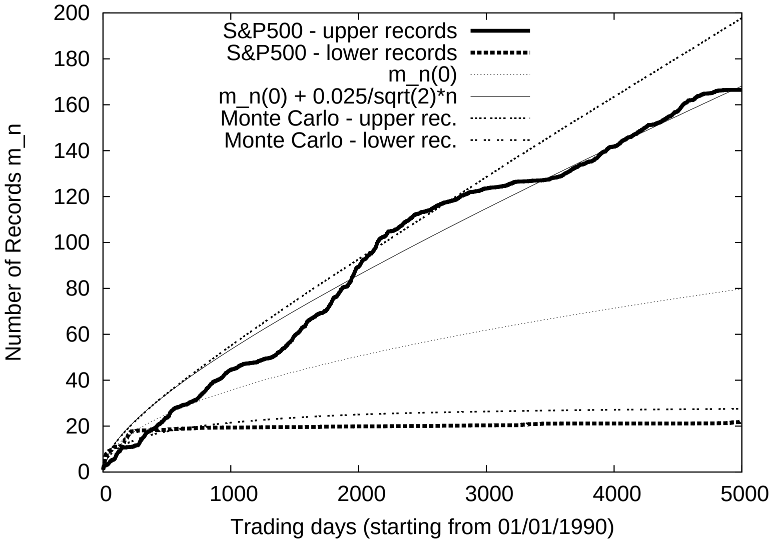

Figure 9.6: Mean number of records averaged over 366 stocks from the S&P 500 index, computed from daily values for the time period 1.1.1990 - 31.3.2009. Full thick line shows the number of upper records, dotted thick line shows the number of lower records in the data set. The expected number of records $m_n(0) = \frac{2}{\sqrt{\pi}}\sqrt{n}$ for a symmetric random walk is shown by the thin dotted line. Also shown are the predictions of the biased random walk model with effective normalized drift $c/\sigma = 0.025$ obtained from Monte Carlo simulations as well as from the approximate expression $m_n(c) = m_n(0) + \frac{c}{\sqrt{2}\sigma}n$ (thin full line).

In the raw stock data the number of upper records after $n = 5000$ trading days is considerably larger than the expected number of $2\sqrt{5000/\pi} \approx 79.79$ for a symmetric random walk. At the end of the observation period, we found an average number of 166.56 upper records in the stocks, but only 22.33 lower records. The rate of upper records was roughly constant over the entire period, whereas the rate of lower records was almost zero already after 300 days. Apparently a positive trend had a very strong effect on the record statistics of the analyzed stocks. To quantify the trend, we performed a linear regression analysis on the logarithms of the individual stock prices, determining the drift c_i and the standard deviation of increments σ_i for each stock $i = 1, ..., 366$. The normalized drift c_i/σ_i was then averaged over all stocks, yielding the estimate $\langle c_i/\sigma_i \rangle \approx 0.025$. At $n = 5000$ we are thus well outside the regime in which the pertubative result (9.4) should be valid. Still, inserting the estimated normalized drift $c/\sigma = 0.025$ into (9.4) we obtain a record number of 166.59, in very close agreement with the observed value. The comparison with Monte Carlo simulations of biased random walks with the same drift shows that this accuracy is actually fortuitous, but the description of the stock market data by the biased random walk model is nevertheless quite reasonable (Fig. 9.6).

Next we detrended the data by subtracting the fitted linear trend from the logarithmic stock prices, and counted the number of records in the detrended time series. We found an average number of 75.79 upper records after 5000 steps, in close agreement with the result for a symmetric random walk. However, the number of lower records was only 53.65, which is significantly smaller than expected. This residual asymmetry between upper and lower records persists if, instead of subtracting an overall linear trend, the data are detrended by normalizing each stock by the index [19]. To further explore this phenomenon we split the time series into 50 shorter series each lasting 100 trading days. We detrended each of the shorter time series individually by subtracting a linear trend, counted the number of upper and lower records, and then averaged the record numbers over the whole ensemble of 50×366 series of length 100. The results are shown in Fig. 9.7. It appears that while the number of upper records is in a very good agreement with the symmetric random walk

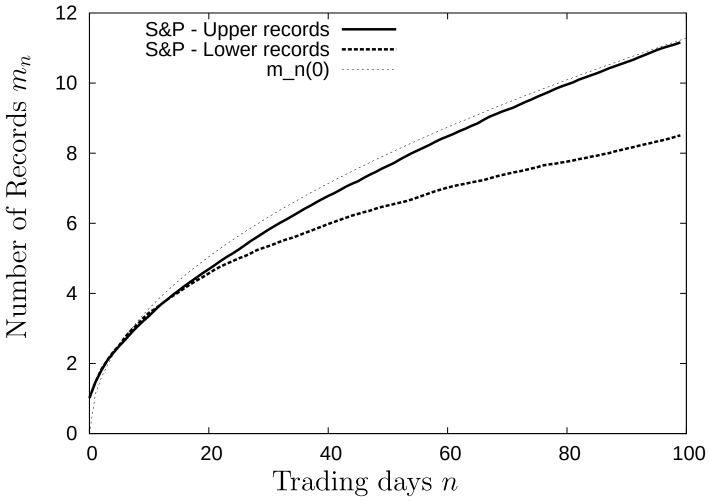

Figure 9.7: Mean number of records in subsequences of the time series taken from the S&P 500 index. The entire data set of 5000 consecutive daily values was split into 50 subsequences of length 100. For each of the subsequences a linear detrending of the logarithm of the daily values was performed and the upper and lower record numbers were determined from the detrended data. The results, averaged over all stocks and all subsequences, are given by the thick black line (upper records) and the thick dashed line (lower records). The thin dashed line shows the analytical prediction for a symmetric random walk $m_n(0) = 2\sqrt{n/\pi}$. The number of upper records is in good agreement with $m_n(0)$, but the number of lower records is significantly reduced.

model, the number of lower records is still suppressed. This effect was found for different choices of the lengths of the time series and appears to be independent of this choice.

Qualitatively, a reduced number of lower records indicates that the positive first-passage times are increased compared to the corresponding negative first passage times. An asymmetry between first passage times to a prescribed (positive or negative) return level has in fact been observed in previous analyses of stock market data, and is known as the *gain-loss asymmetry* [20–23]. However, this phenomenon differs in several important respects from the one reported here. First, in most (though not all [23]) cases the sign of the asymmetry is opposite to that suggested by the asymmetry in the record statistics, in that first passage times for crossing a prescribed level from below are larger than for crossings from above [20, 22]. Second, the asymmetry vanishes when the prescribed return level tends to zero, which is the relevant limit for the analysis of records. Finally, in contrast to the asymmetry between upper and lower records reported here, the gain-loss asymmetry is a property of entire stock indices which does not occur in individual stocks [21, 22]. Indeed, a preliminary analysis of first-passage times to the origin in the detrended S&P 500 data shows an asymmetry between positive and negative excursions only when the starting point of the excursion is conditioned to be a record event [25]. An explanation of the observed residual asymmetry between upper and lower records must therefore be left to future work.

9.5 Summary

In conclusion, using the methods introduced in [1] and a more general form of the Sparre Andersen Theorem [4, 8], we were able to describe the effect of a linear drift on the record statistics of a Gaussian random walk in two regimes. For short times $n \ll \left(\frac{\sigma}{c}\right)^2$ we find that the correction to the record rate $P_n(c) - P_n(0)$ increases proportional $\arctan(n)$ and then saturates at a value of $\frac{c}{\sqrt{2}\sigma}$. On the other hand, for large n the record rate saturates at a constant limiting value $P(c)$, which is linear in c for $c/\sigma \ll 1$ and approaches unity for large c/σ according to Eq.(9.24). The transition between the two regimes is described by the scaling form (9.26).

We applied our results to the statistics of records in 366 stocks contained in the S&P 500 index from 1990 to 2009. We found that, after detrending, the number of upper records in the stocks is basically identical to that predicted for the symmetric random walk. The fact that the number of lower records appears to be systematically decreased is interesting and needs to be examined more thouroughly in the future. On the theoretical side, a possible topic for future research is the record statistics of asymmetric random walks with a more complicated asymmetry than just a constant drift. The issue of asymmetric random walks with discrete jump distributions is also still open for further investigations.

Acknowledgements

We thank I. Szendro and J. Franke for helpful discussions and support, and gratefully acknowledge the Department of Business Administration and Finance at the University of Cologne for providing access to the stock data from the S&P 500. Financial support was provided by DFG within the Bonn Cologne Graduate School of Physics and Astronomy.

Bibliography

[1] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).

[2] P. L. Krapivsky, S. Redner, and E. Ben-Naim, *A kinetic view of statistical physics* (Cambridge University Press, 2010).

[3] G. H. Weiss, *Aspects and applications of the random walk* (North-Holland, 1994).

[4] S. N. Majumdar, Physica A **389**, 4299 (2010).

[5] N. Glick, Amer. Math. Monthly **85**, 2 (1978).

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

[7] V. B. Nevzorov, *Records: Mathematical Theory* (Providence, RI: American Mathematical Society, 2001).

[8] E. Sparre Andersen, Math. Scand. **1**, 263 (1953).

[9] W. Feller, *An introduction to probability theory and its applications* (Wiley, New York, 1968).

[10] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, 2001).

[11] M. Bauer, C. Godreche, and J. Luck, J. Stat. Phys. **96**, 963 (1999).

[12] P. Le Doussal and K. J. Wiese, Phys. Rev. E **79**, 051105 (2009).

[13] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).

[14] G. Wergen and J. Krug, EPL **92**, 30008 (2010).

[15] R. Ballerini, Stat. Probab. Lett. **5**, 83 (1987).

[16] L. Bachelier, Ann. Sci. Ecole Norm. S. **17**, 21 (1900).

[17] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000).

[18] J. Voit, *The Statistical Mechanics of Financial Markets* (Springer Berlin, 2001).

[19] M. Bogner, "Rekordstatistik in Finanzdaten, unpublished thesis," (2009).

[20] M. H. Jensen, A. Johansen, and I. Simonsen, Physica A **324**, 338 (2003).

[21] A. Johansen, I. Simonsen, and M. H. Jensen, Physica A **370**, 64 (2006).

[22] I. Simonsen *et al.*, Eur. Phys. J. B **57**, 153 (2007).

[23] K. Karpio, M. A. Zaluska-Kotur, and A. Orlowski, Physica A **375**, 599 (2007).

[24] "Thomson Datastream Advance 4.0 SP4," (2003).

[25] G. Wergen, "unpublished," (2011).

Chapter 10

Record statistics and persistence for a random walk with a drift

Satya N. Majumdar[1], **Grégory Schehr**[1] **and Gregor Wergen**[2]

[1] *Laboratoire de Physique Théorique et Modèles Statistiques, Université Paris Sud 11 and CNRS, Orsay, France*
[2] *Institute for Theoretical Physics, University of Cologne*

Published in: Journal of Physics A: Mathematical and Theoretical, 45 (2012) 355002

Abstract: We study the statistics of records of a one-dimensional random walk of n steps, starting from the origin, and in presence of a constant bias c. At each time-step the walker makes a random jump of length η drawn from a continuous distribution $f(\eta)$ which is symmetric around a constant drift c. We focus in particular on the case were $f(\eta)$ is a symmetric stable law with a Lévy index $0 < \mu \leq 2$. The record statistics depends crucially on the persistence probability which, as we show here, exhibits different behaviors depending on the sign of c and the value of the parameter μ. Hence, in the limit of a large number of steps n, the record statistics is sensitive to these parameters (c and μ) of the jump distribution. We compute the asymptotic mean record number $\langle R_n \rangle$ after n steps as well as its full distribution $P(R, n)$. We also compute the statistics of the ages of the longest and the shortest lasting record. Our exact computations show the existence of five distinct regions in the $(c, 0 < \mu \leq 2)$ strip where these quantities display qualitatively different behaviors. We also present numerical simulation results that verify our analytical predictions.

10.1 Introduction

The statistical properties of record-breaking events in stochastic processes have been a popular subject of research in recent years. The theory of records has found many interesting applications. Record events are very important in sports [1, 2] and climatology [3–6], but have also been found relevant in biology [7], in the theory of spin-glasses [8, 9] and in models of growing networks [10]. Also in finance, record-breaking events, e.g., when the price of a stock breaks its previous records, can lead to increased financial activities [11, 12]. In all of these fields researchers have recently made progress in understanding and modeling the statistics of records by comparing the records in observational data with various kinds of stochastic processes. In this context it has become increasingly important to improve our understanding of the record statistics of elementary stochastic processes. In this paper we focus on one such elementary stochastic process namely a random walk in presence of a constant bias. We show that even for such a simple process, its record statistics is considerably nontrivial and rich.

In general, one is interested in the record events of a discrete-time series of random variables (RV's) $x_0, x_1, ..., x_n$. An (upper) record is an entry x_k, which exceeds all previous entries: $x_k > \max(x_0, x_1, ..., x_{k-1})$. Until the end of the last century record statistics was fully understood only in the case when the entries of the time series are independent and identically distributed (i.i.d.) RV's (see for instance [13–15]). For i.i.d. RV's from a continuous distribution $p(x)$ the probability r_n of a record in the n-th time step is given by [13]

$$r_n := \text{Prob}\left[x_n = x_k > \max(x_0, x_1, ..., x_{n-1})\right] = \frac{1}{n+1}\,, \tag{10.1}$$

which is universal, i.e., independent of the parent distribution $p(x)$. This universality follows simply from the isotropy in ordering, i.e., any one of the $(n+1)$ entries is equally probable to be a record. Let R_n denote the total number of records up to step n. The mean record number is then simply $\langle R_n \rangle = \sum_{m=0}^{n} r_m$, which grows asymptotically as $\sim \ln n$ for large n.

Due to the numerous applications of the theory of records it became interesting to consider more general models. There has been a lot of interest in the record statistics of RV's which are uncorrelated but not identical anymore. For instance Ballerini et al. considered uncorrelated RV's with a linear drift [16]. More recently Franke et al. studied the same problem as well and found numerous new results [17–19] by also considering the correlations between individual record events. This model was then successfully applied to the statistics of temperature records in the context of global warming [5]. In 2006 Krug studied the statistics of records of uncorrelated RV's with a time-increasing standard deviation, a model with important biological implications [20].

Another important issue is the study of record statistics for *correlated* random variables. For *weak* correlations, with a finite correlation time, one would expect that the record statistics for a large sequence to be asymptotically similar to the uncorrelated case. This is no longer true when there are *strong* correlations between the entries. Perhaps, one of the simplest and most natural time series with strong correlations between its entries corresponds to the positions of a one dimensional random walk [21]. Despite the striking importance and abundance of random walk in various areas of research, the record statistics of a single, discrete-time random walk with a symmetric jump distribution was not computed and understood until only a few years ago. In 2008, Majumdar and Ziff [22] computed exactly the record statistics of a one dimensional symmetric random walk model and showed that the record rate of such a process is completely universal for any continuous and symmetric jump distribution, thanks to the so called Sparre Andersen theorem [23]. They considered a time series of RV's x_m given by:

$$x_m = x_{m-1} + \eta_m, \tag{10.2}$$

where η_m's are i.i.d. RV's drawn from a symmetric and continuous jump distribution $f(\eta)$ (it includes even Lévy flights where $f(\eta) \sim 1/|\eta|^{\mu+1}$ with $0 < \mu < 2$). Then, the record rate

r_n for such a process is given by the universal formula [22]

$$r_n = \binom{2n}{n} 2^{-2n} \xrightarrow{n\to\infty} \frac{1}{\sqrt{\pi n}} \,, \tag{10.3}$$

independently of the jump distribution $f(\eta)$. They also computed exactly the mean record number $\langle R_n \rangle$ and even its full distribution [22]. In addition, there exists nice connection between the record statistics and the extreme value statistics for the one dimensional symmetric jump processes and many universal results can be subsequently derived using the Sparre Andersen theorem (see [24] for a review).

Following Ref. [22], there has been considerable interests in generalising them to more general set of strongly correlated stochastic processes. For instance, Sabhapandit discussed symmetric random walks with a random, possibly heavy tailed, waiting time between the individual jumps (the so called Continuous Time Random Walk model) [25]. Recently the present authors considered the record statistics of an ensemble of N independent and symmetric random walks [12]. There, in contrast to the case of a single random walker, the record statistics of N Lévy flights with a heavy-tailed jump distribution was found to be different from the one of N Gaussian random walkers with a jump distribution that has a finite second moment.

Another important generalization is to consider a single one dimensional random walker but with asymmetric jump distribution, for instance, in presence of a constant bias c. First steps towards this generalization were taken by Le Doussal and Wiese in 2009 [26] who derived the exact record statistics for a biased random walker with a Cauchy jump distribution (a special case of Lévy flight with Lévy index $\mu = 1$). More recently in 2011, Wergen et al. showed that a biased random walk is useful to model record-breaking events in daily stock prices [11]. They were able to obtain results in some special limits of a biased random walker with a Gaussian jump distribution. Apart from these two special cases, namely the Cauchy and the Gaussian jump distribution, there are no other analytical results available, to our knowledge, for other jump distributions for a biased random walker. Recently, the record statistics for a biased random walker was also studied numerically in order to quantify the contamination spread in a porous medium via the particle tracking simulations [27].

In this article we present a complete analysis of the record statistics for a biased random walker with arbitrary jump distributions. As we will see, the record statistics depends crucially on the persistence probability $Q(n)$ [see Eq. (10.17) below], the probability that the biased walker stays to the left of its initial starting position up to n steps. While persistence probability for various stochastic processes have been extensively studied in the recent past [28], it seems that for this simple biased jump process, it has not been systematically studied in the literature to the best of our knowledge. Here we provide exact results for the persistence probability $Q(n)$ for a biased random walk arbitrary jump distributions [see Eq. (10.67)], which subsequently leads to the exact record statistics for the same process.

The rest of the paper is organized as follows. Since the paper is long with many detailed results, we provide in section 10.1 a short review on the record statistics for random walks both with and without bias, followed by a summary of the main results of this paper. Readers not interested in the details of the calculations can skip the rest of the paper. In section 10.1, we will show how to use the renewal property of the random walk and a generalized version of the Sparre Andersen theorem [23] to compute the persistence of random walks in presence of both positive and negative drift. The results for the persistence are interesting on their own and will be discussed in detail in section 10.4, but they will also allow us to compute the record statistics. In particular we will show that, in the presence of drift, the complete universality found for the record statistics in the unbiased case [22] breaks down and there are five different types of asymptotic behaviors which emerge depending on the two parameters of the model, namely the drift c and the index $0 < \mu \leq 2$ characterizing the tail of the jump distribution. This record statistics will be discussed in detail in section 10.5. Later, in section 10.6, we will also discuss the extreme value statistics of the ages of the longest (section 10.6.1) and the shortest lasting records (section 10.6.2) in each of the regimes. We will show that the asymptotic behavior of these quantities is also systematically different in the five regimes. Finally in section 10.7, we will conclude with some open problems.

10.2 Record statistics for random walks: A short review and a summary of new results

In this section, we provide a short review on the record statistics of a one dimensional random walk model, with and without external drift. This will also serve to set up our notations for the rest of the paper. At the end of this section, we summarize the main new results obtained in this work.

Let us first start with the driftless case. Consider a sequence of random variables $\{x_0 = 0, x_1, x_2, \ldots, x_n\}$ where x_m represents the position of a discrete-time *unbiased* random walker at step m. The walker starts at the origin and its position evolves via the Markov rule $x_m = x_{m-1} + \eta_m$, where η_m represents the stochastic jump at the m-th step. The jump variables η_m's are i.i.d. random variables, each drawn from the common probability distribution function (pdf) $f(\eta)$, normalized to unity. The pdf $f(\eta)$ is continuous and symmetric with zero mean. Let $\hat{f}(k) = \int_{-\infty}^{\infty} f(\eta)\, e^{ik\eta}\, d\eta$ denote the Fourier transform of the jump distribution. We will henceforth focus on jump distributions $f(\eta)$ whose Fourier transform has the following small k behavior

$$\hat{f}(k) = 1 - (l_\mu\, |k|)^\mu + \ldots \tag{10.4}$$

where $0 < \mu \leq 2$ and l_μ represents a typical length scale associated with the jump. The exponent $0 < \mu \leq 2$ dictates the large $|\eta|$ tail of $f(\eta)$. For jump densities with a finite second moment $\sigma^2 = \int_{-\infty}^{\infty} \eta^2\, f(\eta)\, d\eta$, such as Gaussian, exponential, uniform etc, one evidently has $\mu = 2$ and $l_2 = \sigma/\sqrt{2}$. In contrast, $0 < \mu < 2$ corresponds to jump densities with fat tails $f(\eta) \sim |\eta|^{-1-\mu}$ as $|\eta| \to \infty$. A typical example is $\hat{f}(k) = \exp[-|k|^\mu]$ where $\mu = 2$, which corresponds to the Gaussian jump distribution, while $0 < \mu < 2$ corresponds to Lévy flights (for reviews on these jump processes see [29, 30]).

A quantity that will play a crucial role later is $P_n(x)$ which denotes the probability density of the position of the symmetric random walk at step n. Using the Markov rule in Eq. (10.2), it is easy to see that $P_n(x)$ satisfies the recursion relation

$$P_n(x) = \int_{-\infty}^{\infty} P_{n-1}(x')\, f(x - x')\, dx' \;, \tag{10.5}$$

starting from $P_0(x) = \delta(x)$. This recurrence relation can be trivially solved by taking Fourier transform and using the convolution structure. Inverting the Fourier transform, one gets

$$P_n(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \left[\hat{f}(k)\right]^n e^{-i\,k\,x} \,. \tag{10.6}$$

In the limit of large n, the small k behavior of $\hat{f}(k)$ dominates the integral on the right hand side (rhs) of Eq. (10.6). Substituting the small k behavior from Eq. (10.4), one easily finds that for $0 < \mu < 2$, typically $x \sim l_\mu n^{1/\mu}$ and $P_n(x)$ approaches the scaling form [29]

$$P_n(x) \to \frac{1}{l_\mu\, n^{1/\mu}}\, \mathcal{L}_\mu \left(\frac{x}{l_\mu\, n^{1/\mu}}\right) \;, \quad \text{where} \quad \mathcal{L}_\mu(y) = \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, e^{-|k|^\mu}\, e^{-i\,k\,y} \,. \tag{10.7}$$

For $0 < \mu < 2$, the scaling function $\mathcal{L}_\mu(y)$ decays as a power law for large $|y|$ [29]

$$\mathcal{L}_\mu(y) \xrightarrow[y\to\infty]{} \frac{A_\mu}{|y|^{\mu+1}} \,, \quad \text{where } A_\mu = \frac{1}{\pi}\, \sin(\mu\pi/2)\,\Gamma(1+\mu). \tag{10.8}$$

In particular, for $\mu = 1$, the scaling function $\mathcal{L}_1(y)$ is precisely the Cauchy density itself

$$\mathcal{L}_1(y) = \frac{1}{\pi}\,\frac{1}{1+y^2}\,. \tag{10.9}$$

In contrast, for $\mu = 2$, the central limit theorem holds, $x \sim \sigma\, n^{1/2}$, and $P_n(x)$ approaches a Gaussian scaling form

$$P_n(x) \to \frac{1}{\sigma\, n^{1/2}}\mathcal{L}_2\left(\frac{x}{\sigma\, n^{1/2}}\right) \,, \quad \text{where} \quad \mathcal{L}_2(y) = \frac{1}{\sqrt{2\pi}}\, \exp(-y^2/2)\,. \tag{10.10}$$

From the sequence of symmetric random variables representing the position of a discrete-time *unbiased* random walker, we next construct a new sequence of random variables $\{y_0 = 0, y_1, y_2, \ldots, y_n\}$ where

$$y_m = x_m + c\,m \quad \text{implying} \quad y_m = y_{m-1} + c + \eta_m \;, \tag{10.11}$$

where η_m's are symmetric i.i.d. jump variables each drawn from the pdf $f(\eta)$. Clearly, y_m then represents the position of a discrete-time random walker at step m in presence of a constant bias c.

In this paper, we are interested in the record statistics of this biased sequence $\{y_0 = 0, y_1, y_2, \ldots, y_n\}$. A record happens at step m if $y_m > \max(y_0 = 0, y_1, y_2, \ldots, y_{m-1})$, i.e., if the position of the biased walker y_m at step m is bigger than all previous positions, with the convention that the initial position $y_0 = 0$ is counted as a record. Let R_n denote the number of records up to step n. Clearly, R_n is a random variable and we denote its distribution by

$$P(R,n) = \text{Proba.}\,[R_n = R]\,. \tag{10.12}$$

We would like to compute the asymptotic properties of this record number distribution $P(R,n)$ for large n, for arbitrary drift c and for arbitrary symmetric and continuous jump density $f(\eta)$ whose Fourier transform $\hat{f}(k)$ has the small k behavior as in Eq. (10.4) with the index $0 < \mu \leq 2$.

In absence of a drift, i.e., for $c = 0$, the distribution $P(R,n)$ was computed exactly in Ref. [22] for all R and n, using a renewal property of the record process. Amazingly, the distribution was found to be completely universal, i.e., independent of the jump distribution $f(\eta)$ (as long as it is symmetric and continuous) for all R and n [22]. In particular, for large n, it was shown that $P(R,n)$ has a scaling form [22]

$$P(R,n) \approx \frac{1}{\sqrt{n}}\, g_0\left(\frac{R}{\sqrt{n}}\right) \,, \tag{10.13}$$

where the universal scaling function

$$g_0(x) = \frac{1}{\sqrt{\pi}}\, \exp(-x^2/4) \,, \text{ for } x \geq 0 \tag{10.14}$$

is a half-Gaussian. Consequently, the mean and the variance of the number of records grows asymptotically as [22]

$$\langle R_n \rangle \approx \frac{2}{\sqrt{\pi}}\, n^{1/2}, \quad \langle R_n^2 \rangle - \langle R_n \rangle^2 \approx 2\left(1 - \frac{2}{\pi}\right) n\,. \tag{10.15}$$

The renewal property of the record process derived originally for the unbiased random walker in Ref. [22] was then generalized to the case with a nonzero drift c in Ref. [26]. In

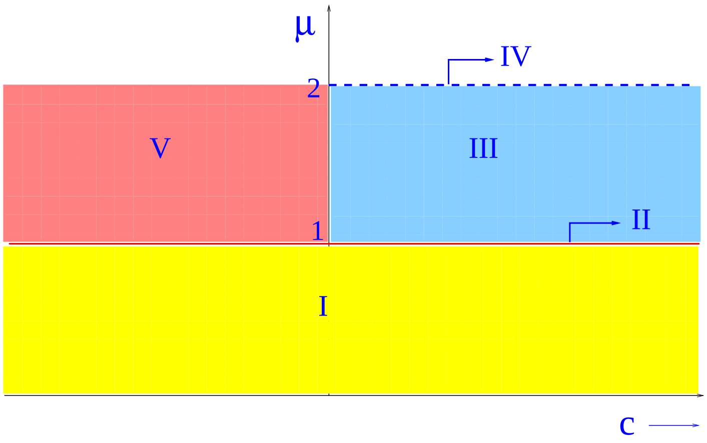

Figure 10.1: Phase diagram in the $(c, 0 < \mu \leq 2)$ strip depicting 5 regimes: (I) $0 < \mu < 1$ and c arbitrary (II) the line $\mu = 1$ and c arbitrary (III) $1 < \mu < 2$ and $c > 0$ (IV) the semi-infinite line $\mu = 2$ and $c > 0$ and (V) $1 < \mu \leq 2$ and $c < 0$. The persistence $Q(n)$, the record number distribution $P(R, n)$ and the mean ages of the longest and the shortest lasting record exhibit different asymptotic behaviors in these 5 regimes (see text).

particular, the authors of Ref. [26] studied in detail the special case of the Cauchy jump distribution $f_{\mathrm{Cauchy}}(\eta) = 1/[\pi(1+\eta^2)]$ [which belongs to the $\mu = 1$ family of jump densities in Eq. (10.4)] and found that the mean number of records $\langle R_n \rangle$ grows algebraically with n for large n with an exponent that depends continuously on c [26]

$$\langle R_n \rangle \approx \frac{1}{\Gamma(1+\theta(c))}\, n^{\theta(c)}, \quad \text{where} \quad \theta(c) = \frac{1}{2} + \frac{1}{\pi} \arctan(c)\,. \tag{10.16}$$

In addition, the asymptotic distribution $P(R, n)$ for large n was found [26] to have a scaling distribution, $P(R, n) \sim n^{-\theta(c)}\, g_c\left(R\, n^{-\theta(c)}\right)$ with a nontrivial scaling function $g_c(x)$ which reduces, for $c = 0$, to the half-Gaussian in Eq. (10.14).

For jump densities with a finite second moment σ^2 and in presence of a nonzero positive drift $c > 0$, the mean number of records $\langle R_n \rangle$ was analysed in Ref. [11] and was found to grow linearly with n for large n, $\langle R_n \rangle \approx a_2(c)\, n$ where the prefactor $a_2(c)$ was computed approximately for the Gaussian jump distribution. However, an exact expression of the prefactor for arbitrary jump densities with a finite σ^2 is missing. In addition, these results were then applied [11] to analyse the record statistics of stock prices from the Standard and Poors 500. The distribution of the record number $P(R, n)$ for large n has not been studied for jump densities with a finite second moment.

In this paper, we present detailed exact results for the asymptotic record number distribution $P(R, n)$ for large n, for arbitrary drift c (both positive and negative) and for arbitrary symmetric and continuous jump densities $f(\eta)$ with Fourier transform $\hat{f}(k)$ having a small k behavior as in Eq. (10.4) parametrized by the exponent $0 < \mu \leq 2$. We find a variety of rather rich behaviors for $P(R, n)$ depending on the value of c and the exponent μ. On the strip $(c, 0 < \mu \leq 2)$ (see Fig. 10.1), we find five distinct regimes: (I) when $0 < \mu < 1$ with c arbitrary (II) when $\mu = 1$ and c arbitrary (III) when $1 < \mu < 2$ and $c > 0$ (IV) when $\mu = 2$ and $c > 0$ and (V) when $1 < \mu \leq 2$ and $c < 0$. In these five regimes the record statistics behave differently, resulting in different asymptotic forms for the record number distribution $P(R, n)$. The line $\mu = 1$ (regime II above) is a critical line on which the record statistics exhibits marginal behavior. These five regimes are summarized in the phase diagram in the $(c, 0 < \mu \leq 2)$ strip in Fig. 10.1.

As we will see later, a quantity that plays a crucial role in the study of record statistics is the persistence $Q(n)$ which denotes the probability that the process y_m in Eq. (10.11) stays below its initial value y_0 up to step n, i.e.,

$$Q(n) = \text{Proba. } [y_i < y_0, \text{for all } i = 1, 2, \dots, n] \ . \tag{10.17}$$

Due to the translational invariance of the process, $Q(n)$ does not depend on y_0. The persistence probability has been studied quite extensively in recent years in a variety of theoretical and experimental systems [28]. We will see that even for the simple stochastic process y_m representing the position of a discrete-time random walker in presence of a drift, the persistence $Q(n)$ has a rather rich asymptotic behavior depending on the parameters μ and c. Hence, even though here our main interest is in the record statistics, we include the results for the persistence $Q(n)$ as a byproduct.

We also analyse the statistics of waiting times between individual record events. In particular we are interested in the expected ages of the longest and the shortest lasting records. The age of the longest lasting record is defined as:

$$l_{\max,n} = \max\left(l_1, l_2, ..., l_R\right), \tag{10.18}$$

where l_i is the length of the time interval between the the i-th and the $(i+1)$-th record. Similarly one defines the age of the shortest lasting record as

$$l_{\min,n} = \min\left(l_1, l_2, ..., l_R\right). \tag{10.19}$$

In [22], the mean values of $l_{\max,n}$ and $l_{\min,n}$ were computed exactly for the symmetric random walk with arbitrary jump distribution. It was found that [22] for large n

$$\langle l_{\rm max,n} \rangle \sim C_0 \, n \ , \tag{10.20}$$

where $C_0 \approx 0.626508...$ is a universal constant independent of the jump distribution. Interestingly, the same constant C_0 also appears in other related problems [31, 32]. In contrast, the shortest record exhibits different behavior for large n [22]

$$\langle l_{\rm min,n} \rangle \sim \sqrt{n/\pi} \ . \tag{10.21}$$

In this paper we generalize these results to the case of a biased random walk and as in the case of record number distribution, we find five different asymptotic behaviors depending on c and μ.

Summary of the new results: Let us then summarize the main new results in this paper for the asymptotic behavior of the persistence $Q(n)$, the record number distribution $P(R, n)$ and the extremal ages of records in the 5 regimes in the (c, μ) strip mentioned above.

Regime I ($0 < \mu < 1$ and c arbitrary): In this regime, we find that the persistence $Q(n)$ decays algebraically for large n

$$Q(n) \approx \frac{B_I}{\sqrt{n}} \ , \tag{10.22}$$

where the prefactor B_I depends on the details of the jump distribution $f(\eta)$ and the drift c and can be computed explicitly [see Eq. (10.78)]. The mean record number up to n steps grows asymptotically for large n as

$$\langle R_n \rangle \approx A_{\rm I} \sqrt{n} \, . \tag{10.23}$$

While the growth exponent 1/2 is universal, i.e. independent of c and the precise form of the jump distribution $f(\eta)$, the prefactor $A_{\rm I}$ depends on c and on the details of the density $f(\eta)$. In addition, the two prefactors $A_{\rm I}$ and $B_{\rm I}$ are related simply via $B_I = 2/(\pi A_{\rm I})$. We find the following exact expression for the prefactor $A_{\rm I}$

$$A_{\rm I} = \frac{2}{\sqrt{\pi}} \exp\left[\frac{1}{\pi} \int\limits_0^\infty \frac{dk}{k} \arctan\left(\frac{\hat{f}(k)\sin(kc)}{1 - \hat{f}(k)\cos(kc)} \right) \right] \ . \tag{10.24}$$

In the scaling limit when $n \to \infty$ and $R \to \infty$, but with the ratio $R/\sqrt{n}$ fixed, we find that the distribution $P(R,n)$ approaches the scaling form

$$P(R,n) \approx \frac{2}{A_\mathrm{I}\sqrt{\pi\, n}}\, g_0\left(\frac{2\,R}{A_\mathrm{I}\sqrt{\pi\, n}}\right) , \quad \text{where} \quad g_0(x) = \frac{1}{\sqrt{\pi}} \exp(-x^2/4)\,. \tag{10.25}$$

Averaging over R evidently reproduces the result in Eq. (10.23). Thus, the record number, rescaled by the nonuniversal scale factor $R \to R/A_\mathrm{I}$, approaches asymptotically the same universal half-Gaussian scaling distribution as in the driftless case $c = 0$ in Eq. (10.14).

The statistics of the longest lasting record is completely unaffected by the drift c in this regime. For the mean value $\langle l_{\max,n}\rangle$ we find that

$$\langle l_{\max,n}\rangle \sim C_\mathrm{I}\, n, \tag{10.26}$$

where the same constant $C_I = C_0 \approx 0.626508...$ was also found in the unbiased case [see Eq. (10.20)]. The age of the shortest lasting record is given by

$$\langle l_{\min,n}\rangle \sim D_\mathrm{I}\, \sqrt{n}, \tag{10.27}$$

with a prefactor $D_\mathrm{I} = B_\mathrm{I}$. Therefore, in contrast to $\langle l_{\max,n}\rangle$, $\langle l_{\min,n}\rangle$ slightly differs from the unbiased case and has a prefactor that depends non-trivially on c.

Regime II (the line $\mu = 1$ and c arbitrary): On this line, we find that the persistence $Q(n)$ decays algebraically for large n but with an exponent that depends continuously on c

$$Q(n) \approx \frac{B_\mathrm{II}}{n^\theta}\,, \tag{10.28}$$

where the exponent $0 \le \theta(c) \le 1$ is given in Eq. (10.16). In this sense the behavior is marginal. The prefactor B_II can be computed exactly [see Eq. (10.85)]. The mean record number also grows marginally for large n

$$\langle R_n\rangle \approx \frac{A_\mathrm{II}}{\Gamma[1+\theta(c)]}\, n^{\theta(c)}\,, \tag{10.29}$$

where the prefactor $A_\mathrm{II} = 1/\left[\Gamma[1-\theta(c)]\, B_\mathrm{II}\right]$. The record number distribution exhibits an asymptotic scaling form

$$P(R,n) \approx \frac{1}{A_\mathrm{II}\, n^{\theta(c)}}\, g_c\left(\frac{R}{A_\mathrm{II}\, n^{\theta(c)}}\right)\,, \tag{10.30}$$

where one can obtain a formal exact expression (10.110) and explicit tails of the scaling function $g_c(x)$ which also exhibits marginal behavior, i.e., depends continuously on c.

Like in regime II we find that the mean age of the longest lasting record grows linearly in n, but this time with a non-trivial c dependent prefactor. We find that

$$\langle l_{\max,n}\rangle \sim C_\mathrm{II}\, n\,, \tag{10.31}$$

where C_II is given in Eq. (10.148). The mean age of the shortest lasting record is more strongly affected by the drift. Here we find that $\langle l_{\min,n}\rangle$ grows algebraically with n with an exponent which depends continuously on c:

$$\langle l_{\min,n}\rangle \sim D_\mathrm{II}\, n^{1-\theta(c)}, \tag{10.32}$$

with $D_\mathrm{II} = B_\mathrm{II}$ as in Eq. (10.28) and $\theta\,(c)$ as defined in Eq. (10.16).

Regime III ($1 < \mu < 2$ and c arbitrary): In this regime, the persistence $Q(n)$ decays for large n as

$$Q(n) \approx \frac{B_\mathrm{III}}{n^\mu}\,, \tag{10.33}$$

where the prefactor $B_{\rm III}$ depends on the details of the jump distribution and can be computed [see Eq. (10.90)]. The mean number of records grows linearly with increasing n

$$\langle R_n \rangle \approx a_\mu(c)\, n \;, \tag{10.34}$$

where the prefactor $a_\mu(c)$ can be computed explicitly [see Eq. (10.116)]. The record number distribution $P(R,n)$, for large n, behaves as

$$P(R,n) \approx \frac{1}{a_\mu(c) n^{1/\mu}}\, V_\mu\left(\frac{R - a_\mu(c) n}{a_\mu(c)\, n^{1/\mu}}\right) \;, \tag{10.35}$$

where the scaling function $V_\mu(u)$ can be computed exactly and it has a non-Gaussian form with highly asymmetric tails

$$\begin{aligned} V_\mu(u) &\approx A_\mu\, |u|^{-\mu-1} \quad \text{as} \quad u \to -\infty & (10.36) \\ &\approx c_1\, u^{(2-\mu)/2(\mu-1)} \exp\left[-c_2\, u^{\mu/(\mu-1)}\right] \quad \text{as} \quad u \to \infty \;, & (10.37) \end{aligned}$$

where A_μ is the same constant as in Eq. (10.8) and the constants c_1 and c_2 are given explicitly by

$$\begin{aligned} c_1 &= \left[2\pi(\mu-1)(\mu B_\mu)^{1/(\mu-1)}\right]^{-1/2} \;, & (10.38) \\ c_2 &= (1-1/\mu)\,(\mu B_\mu)^{-1/(\mu-1)} \;, & (10.39) \end{aligned}$$

where

$$B_\mu = -\frac{1}{2\cos(\mu\pi/2)} > 0 \quad \text{for} \quad 1 < \mu < 2 \,. \tag{10.40}$$

Thus, in this regime, while the mean record number grows linearly with n, the fluctuations around the mean are anomalous $\sim n^{1/\mu}$ and described by a non-Gaussian distribution.

Also the extremal ages of records have an interesting behavior in this regime. In particular we find that the average age of the longest lasting record grows like

$$\langle l_{\max,n} \rangle \sim C_{\rm III}\, n^{\frac{1}{\mu}} , \tag{10.41}$$

where the constant $C_{\rm III}$ can be computed explicitly [see Eq. (10.148)]. On the other hand and in contrast to the results of regime I and II, the mean age of the shortest lasting record converges to a finite value:

$$\langle l_{\min,n} \rangle \sim D_{\rm III} = 1 - a_\mu(c) \;, \tag{10.42}$$

which thus depends continuously on c. The strongly different n dependence of $\langle l_{\max,n} \rangle$ and $\langle l_{\min,n} \rangle$ in the regime I and in the regime III is a consequence of the fact that while in regime I the asymptotic behavior is dominated by the fluctuations, in regime III the effect of the drift is stronger in the large n limit.

Regime IV (the semi-infinite line $\mu = 2$ and $c > 0$): On this semi-infinite line the variance σ^2 of the jump pdf is finite. This leads to an exponential tail of the persistence $Q(n)$ for large n. More precisely we show that

$$Q(n) \approx \frac{B_{\rm IV}}{n^{3/2}} \exp[-(c^2/2\sigma^2)\, n] \;, \tag{10.43}$$

where the nonuniversal prefactor $B_{\rm IV}$ can be computed exactly [see Eq. (10.96)]. We also show that the mean and the variance of the record number both grow linearly for large n

$$\langle R_n \rangle \approx a_2(c)\, n \quad \text{and} \quad \langle R_n^2 \rangle - \langle R_n \rangle^2 \approx b_2(c)\, n \;, \tag{10.44}$$

where the amplitudes $a_2(c)$ and $b_2(c)$ are nonuniversal and depend on the details of the jump distribution $f(\eta)$. We provide exact expressions for these amplitudes respectively in

Eqs. (10.126) and (10.129) as well as in 10.7. The distribution of the record number $P(R,n)$ approaches a Gaussian form asymptotically for large n

$$P(R,n) \approx \frac{1}{\sqrt{2\,\pi\,b_2(c)\,n}} \exp\left[-\frac{1}{2b_2(c)n}(R-a_2(c)\,n)^2\right]. \tag{10.45}$$

Thus, in this regime, the mean record number grows linearly with n with normal Gaussian fluctuations $\sim n^{1/2}$ around the mean.

It is interesting to see that the asymptotic behavior of $\langle l_{\max,n}\rangle$ in regime IV is qualitatively different from regime III. Here we find that $\langle l_{\max,n}\rangle$ grows only logarithmically with n for $n \to \infty$:

$$\langle l_{\max,n}\rangle \sim C_{\rm IV} \ln n\ , \tag{10.46}$$

with an n independent constant $C_{\rm IV} = \frac{2\sigma^2}{c^2}$. Like in regime III, the average age of the shortest lasting record approaches a (different) constant value depending on c:

$$\langle l_{\min,n}\rangle \sim D_{\rm IV} = 1 - a_2(c)\ , \tag{10.47}$$

which depends continuously on c.

Regime V ($1 < \mu \leq 2$ and $c < 0$): In this case, the walker predominantly moves towards the negative axis due to the drift. Consequently, the events where the walker crosses the origin from the negative to the positive side become extremely rare. As a result, with a finite probability the walker stays forever on the negative side. Thus, the persistence $Q(n)$ approaches a constant for large n

$$Q(n) \to \alpha_\mu(c)\,. \tag{10.48}$$

Similarly, the occurrence of the records (with positive record values) are also rare. Subsequently, we find that the mean record number also approaches a constant for large n

$$\langle R_n\rangle \to \frac{1}{\alpha_\mu(c)}\ , \tag{10.49}$$

where the constant $\alpha_\mu(c)$ is given by

$$\begin{aligned} \alpha_\mu(c) &= a_\mu(|c|) \quad \text{for} \quad 1 < \mu < 2\,, && (10.50)\\ &= a_2(|c|) \quad \text{for} \quad \mu = 2\ , && (10.51) \end{aligned}$$

where $a_\mu(c)$ and $a_2(c)$ are precisely the amplitude of the linear growth of the mean record number respectively in regime III and IV [respectively in Eqs. (10.34) and (10.44)]. An explicit expression for $\alpha_\mu(c)$ is given in Eq. (10.101). The record number distribution $P(R,n)$ also approaches a steady state, i.e., n-independent distribution as $n \to \infty$. This distribution has a purely geometric form

$$P(R, n \to \infty) = \alpha_\mu(c)\,[1-\alpha_\mu(c)]^{R-1}\,. \tag{10.52}$$

Physically this result is easy to understand because for $c < 0$ and $\mu > 1$, the walker typically moves away from the origin on the negative side with very rare and occasional excursions to the positive side caused by rare large jumps. As a result, the occurrence of a record is like a Poisson process which eventually leads to a geometric distribution as in Eq. (10.52).

In this regime the statistics of the longest and the shortest lasting records are particularly simple. Since the record number is finite, the longest lasting record will grow linearly in n:

$$\langle l_{\max,n}\rangle \sim C_{\rm V}\, n\ , C_{\rm V} = 1\ . \tag{10.53}$$

For the shortest lasting record we find a similar behavior:

$$\langle l_{\min,n}\rangle \sim \alpha_\mu(c)\, n, \tag{10.54}$$

with the same c dependent constant $\alpha_\mu(c)$ as in Eq. (10.48). Here, the main contributions to these quantities come from trajectories that never cross the origin and stay negative for all n.

The five regimes in the $(c, 0 < \mu \leq 2)$ strip are depicted in Fig. 10.1. As mentioned above, the line $\mu = 1$ is a special 'critical' line with marginal exponents that depend continuously on the drift c. It is not difficult to understand physically why $\mu = 1$ plays a special role. Indeed, writing $y_n = x_n + c\,n$ where x_n represents a symmetric random walk, we see that the two terms x_n and $c\,n$ compete with each other for large n. Since $x_n \sim n^{1/\mu}$ for $0 < \mu \leq 2$ [see Eq. (10.7)], it is clear that for $0 < \mu < 1$, the term x_n dominates over the drift and the presence of a nonzero drift only leads to subleading asymptotic effect. In contrast, for $\mu > 1$, the drift term starts to play an important role in governing the asymptotic record statistics. In the region $1 < \mu < 2$ and $c > 0$ (regime III), while the mean record number increases linearly with n due to the dominance of the drift term, the typical fluctuation around the mean is still dominated by the $x_n \sim n^{1/\mu}$ term [see Eq. (10.35)]. However when $\mu = 2$ and $c > 0$ (regime IV), the drift term completely dominates over the x_n term leading to Gaussian fluctuations around the mean. This competition between x_n and $c\,n$ thus leads to (i) a 'phase transition' in the asymptotic behavior of record statistics of y_n at the critical value $\mu = 1$ and (ii) an anomalous region with non-Gaussian fluctuations around the mean in the region $1 < \mu < 2$ and $c > 0$.

10.3 Record number distribution via renewal property and the generalized Sparre Andersen theorem

The idea of using the renewal property of random walks to compute the distribution of record number was first used in Ref. [22] for symmetric random walks and was subsequently generalized to biased random walks [26]. We briefly summarize below the main idea.

Consider the random sequence $\{y_0, y_1, y_2, \ldots,\}$ representing the successive positions of a discrete-time biased random walker evolving via Eq. (10.11), starting from an arbitrary initial position y_0. Consider first the persistence $Q(n)$ defined in Eq. (10.17). Let us also define

$$F(n) = \text{Proba.}\ [y_1 < y_0,\, y_2 < y_0,\, \ldots,\, y_{n-1} < y_0,\, y_n > y_0] \tag{10.55}$$

which denotes the probability that the walker crosses its initial position y_0 from *below* for the first time at step n. Clearly

$$F(n) = Q(n-1) - Q(n)\,. \tag{10.56}$$

It is also useful to define the generating functions

$$\tilde{Q}(z) = \sum_{n=0}^{\infty} Q(n)\, z^n\ , \quad \tilde{F}(z) = \sum_{n=1}^{\infty} F(n)\, z^n\,. \tag{10.57}$$

Using the relation in Eq. (10.56) it follows that

$$\tilde{F}(z) = 1 - (1-z)\tilde{Q}(z)\,. \tag{10.58}$$

Consider now any realization of the sequence $\{y_0 = 0, y_1, y_2, \ldots, y_n\}$ up to n steps and let R_n be the number of records in this realization. Let $\vec{l} = \{l_1, l_2, \ldots, l_R\}$ denote the time intervals between successive records in this sequence (see Fig. 10.2). Clearly l_i denotes the age of the i-th record, i.e., the time up to which the i-th record survives. The last record, i.e. the R-th record, stays a record till step n. Let $P(\vec{l}, R|n)$ denote the joint distribution of the ages and the number of records up to step n. Using the two probabilities $Q(n)$ and $F(n)$ defined earlier and the fact that the successive intervals between records are statistically independent due to the Markov nature of the process, it follows immediately that

$$P(\vec{l}, R|n) = F(l_1)F(l_2)\ldots F(l_R)\, Q(l_R)\, \delta_{\sum_{i=1}^{R} l_i, N}\ , \tag{10.59}$$

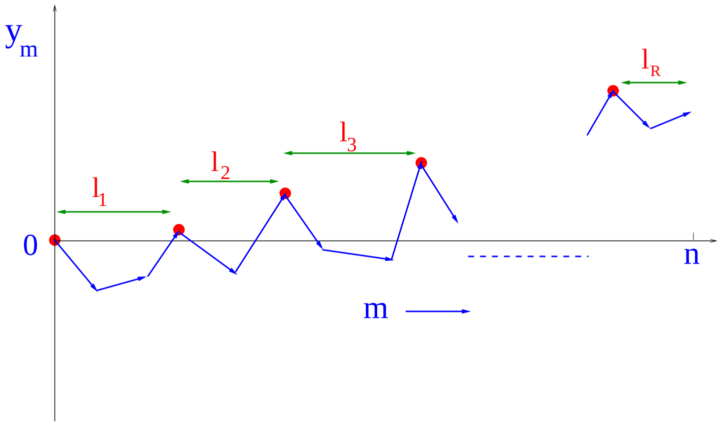

Figure 10.2: A typical realization of the biased random walk sequence $\{y_0 = 0, y_1, y_2, \ldots, y_n\}$ of n steps with R records. Each record is represented by a filled circle. The set $\{l_1, l_2, \ldots, l_{R-1}\}$ represents the time intervals between the successive records and l_R is the age of the last record which is still a record till step n.

where the Kronecker delta enforces the global constraint that the sum of the time intervals equals n. The fact that the last record, i.e. the R-th record, is still surviving as a record at step n indicates that the distribution $Q(l_R)$ of l_R is different from the preceding ones. It is easy to check that $P(\vec{l}, R|n)$ is normalized to unity when summed over $\vec{l}$ and R. The record number distribution $P(R,n) = \sum_{\vec{l}} P(\vec{l}, R|n)$ is just the marginal of the joint distribution when one sums over the interval lengths. Due to the presence of the delta function, this sum is most easily carried out by considering the generating function with respect to n. Multiplying Eq. (10.59) by z^n and summing over $\vec{l}$ and n, one arrives at the fundamental relation

$$\sum_{n=0}^{\infty} P(R,n)\, z^n = \left[\tilde{F}(z)\right]^{R-1} \tilde{Q}(z) = \left[1 - (1-z)\tilde{Q}(z)\right]^{R-1} \tilde{Q}(z) \;, \tag{10.60}$$

where we used the relation in Eq. (10.58). Note that, by definition, $R \le (n+1)$, i.e. $P(R,n) = 0$ if $n < R-1$. Hence, the sum in Eq. (10.60) actually runs from $n = R-1$ to ∞.

Thus the basic object is the generating function $\tilde{Q}(z)$. Once this is determined, one can, at least in principle, compute other quantities such as the statistics of records or their ages using the fundamental renewal equation (10.60). Fortunately, there exists a beautiful combinatorial identity first derived by Sparre Andersen [23] that allows one to compute $\tilde{Q}(z)$

$$\tilde{Q}(z) = \sum_{n=0}^{\infty} Q(n)\, z^n = \exp\left[\sum_{n=1}^{\infty} \frac{z^n}{n}\, p(n)\right] \;, \tag{10.61}$$

where $p(n) = \text{Proba.}\,[y_n < 0]$. Using the relation $y_n = x_n + cn$ where x_n represents the symmetric random walk at step n in Eq. (10.2) one gets, $p(n) = \text{Proba.}\,[x_n < -cn]$. Then, using the pdf $P_n(x)$ of the symmetric walk x_n at step n in Eq. (10.6), one gets

$$p(n) = \text{Proba.}\;[x_n < -cn] = \int_{-\infty}^{-cn} P_n(x)\, dx = \int_{cn}^{\infty} P_n(x)\, dx \;, \tag{10.62}$$

where, in obtaining the last equality we used the symmetry $P_n(x) = P_n(-x)$. Substituting this expression of $p(n)$ in Eq. (10.61) gives

$$\tilde{Q}(z) = \sum_{n=0}^{\infty} Q(n)\, z^n = \exp\left[\sum_{n=1}^{\infty} \frac{z^n}{n} \int\limits_{cn}^{\infty} P_n(x)\, dx\right] . \qquad (10.63)$$

Eq. (10.63), with $P_n(x)$ given by Eq. (10.6), determines $\tilde{Q}(z)$ in terms of the Fourier transform $\hat{f}(k)$ of the jump distribution $f(\eta)$. Subsequently Eq. (10.60) then determines, in principle, the record number distribution $P(R,n)$. In the driftless case $c = 0$, great simplification occurs, since by symmetry $\int_0^\infty P_n(x)dx = 1/2$. This gives, from Eq. (10.63), $\tilde{Q}(z) = 1/\sqrt{1-z}$. This is completely universal as all the dependence on the jump distribution $f(\eta)$ drops out. Subsequently, Eq. (10.60) provides, for $c = 0$, the universal result for the record number distribution [22]

$$\sum_{n=0}^{\infty} P(R,n)\, z^n = \frac{\left(1-\sqrt{1-z}\right)^{R-1}}{\sqrt{1-z}} , \qquad (10.64)$$

which, when inverted, yields [22] for large n the scaling behavior in Eq. (10.13) with the scaling function given by the half-Gaussian form in Eq. (10.14).

However, in presence of a nonzero bias c, extraction of the precise large n behavior of $P(R,n)$ from the set of equations (10.60), (10.63) and (10.6) is more complicated. For the special case of the Cauchy distribution, this was performed in Ref. [26] which showed nontrivial behavior. The rest of this paper is devoted precisely to this technical task of extracting the large n behavior of $P(R,n)$ for a general jump distribution $f(\eta)$ and we will see that a variety of rather rich asymptotic behavior emerges depending on the value of the drift c and the exponent μ characterizing the small k behavior of $\hat{f}(k)$ in Eq. (10.4).

Before finishing this section, let us remark that from Eq. (10.60) one can also compute the generating functions of the moments of the number of records. For example, multiplying Eq. (10.60) by R, summing over R and using the identity $\sum_{n=0}^{\infty} n x^{n-1} = 1/(1-x)^2$ we get for the first moment

$$\sum_{n=0}^{\infty} \langle R_n \rangle\, z^n = \frac{1}{(1-z)^2 \tilde{Q}(z)}. \qquad (10.65)$$

Similarly, multiplying Eq. (10.60) by R^2 and summing over R one gets for the second moment

$$\sum_{n=0}^{\infty} \langle R_n^2 \rangle\, z^n = \frac{2-(1-z)\tilde{Q}(z)}{(1-z)^3\, \tilde{Q}^2(z)}. \qquad (10.66)$$

We will use these two results later in Section IVB.

10.4 Asymptotic behavior of persistence $Q(n)$ for large n

The persistence $Q(n)$, i.e. the probability that the process y_n stays below its initial value y_0 up to step n and its generating function $\tilde{Q}(z)$ is the key ingredient to determine the record number distribution $P(R,n)$ via Eq. (10.60). Apart from its key role as an input for the record statistics, the persistence $Q(n)$ for this process is, by itself, an interesting quantity to study. We will see in this section that even for the simple stochastic process y_n, representing the position of a discrete-time random walker in presence of a drift, the persistence $Q(n)$ has a rather rich asymptotic behavior depending on the parameters μ and c. Before getting into the details of the derivation, it is useful to summarize these asymptotic results. We find that for large n, the persistence $Q(n)$ has the following asymptotic tails depending on

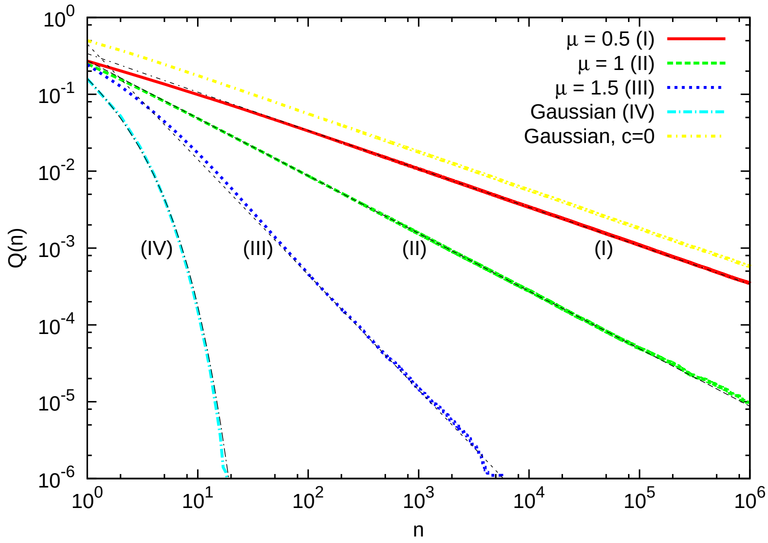

Figure 10.3: Numerical simulations of the persistence $Q\,(n)$, i.e. the probability that a random walker with a bias c stays *below* its initial position up to step n. We considered 4 different Lévy-stable jump distributions characterized respectively by the Lévy index $\mu = 0.5$, 1, 1.5 and $\mu = 2$ (in the last case it is just Gaussian jump distribution). In all cases, we had a constant positive bias $c = 1$ and the data were obtained by averaging over 10^7 samples. For comparison, we also present the result for the unbiased case ($c = 0$) with a Gaussian jump distribution (the top curve). The thin dashed lines give our analytical predictions from Eq. (10.67) with fitted prefactors $B_{\rm I}$, $B_{\rm II}$, $B_{\rm III}$ and $B_{\rm IV}$. For the $\mu = 1$ case we used $\theta\,(c = 1) \approx 0.7498...$.

$\mu > 0$ and c

$$\begin{array}{lll} Q(n) & \sim & B_{\rm I}\, n^{-1/2} \quad \text{for} \quad 0 < \mu < 1 \text{ and } c \text{ arbitrary} \quad \text{(regime I)}\,, \\ & \sim & B_{\rm II}\, n^{-\theta(c)} \quad \text{for} \quad \mu = 1 \text{ and } c \text{ arbitrary} \quad \text{(regime II)}\,, \\ & \sim & B_{\rm III}\, n^{-\mu} \quad \text{for} \quad 1 < \mu < 2 \text{ and } c > 0 \quad \text{(regime III)}\,, \\ & \sim & B_{\rm IV}\, n^{-3/2} \exp[-(c^2/2\sigma^2)\, n] \quad \text{for} \quad \mu = 2 \text{ and } c > 0 \quad \text{(regime IV)}\,, \\ & \sim & \alpha_\mu(c) \quad \text{for} \quad 1 < \mu \leq 2 \text{ and } c < 0 \quad \text{(regime V)}\,, \end{array} \tag{10.67}$$

where the prefactors $B_{\rm I}$, $B_{\rm II}$, $B_{\rm III}$, $B_{\rm IV}$ can be computed explicitly. In regime V, $\alpha_\mu(c)$ is a constant independent of n that can also be computed explicitly [see Eq. (10.95) and 10.7 for $\alpha_2(\mu)$]. The exponent $\theta(c)$ depends continuously on c and is given in Eq. (10.16) [see also Eq. (10.80)]. In Fig. 10.3 these results are confirmed numerically for the regimes I-IV.

To derive these asymptotic behaviors of $Q(n)$ for large n, we start with the key result in Eq. (10.63). Using Cauchy's inversion formula in the complex z plane one can write

$$Q(n) = \int_{C_0} \frac{dz}{2\pi i}\, \frac{1}{z^{n+1}}\, \tilde{Q}(z) \quad \text{with} \quad \tilde{Q}(z) = \exp\left[\sum_{n=1}^{\infty} \frac{z^n}{n} \int_{cn}^{\infty} P_n(x)\, dx \right] \,, \tag{10.68}$$

where the contour C_0 encircles the origin 0 and is free of any singularity of $\tilde{Q}(z)$ (see Fig. 10.4). Let z^* denote the singularity of $\tilde{Q}(z)$ on the real axis closest to the origin. Then, one can deform the contour C_0 to C_1 (see Fig. 10.4) such that the vertical part of C_1 is located just left of z^* and the circular part has radius r_1. By taking the $r_1 \to \infty$ limit, it follows from Eq. (10.68) that for large n, the contribution from the circular part vanishes exponentially. Thus for large n, the leading contribution comes from the vertical part of C_1,

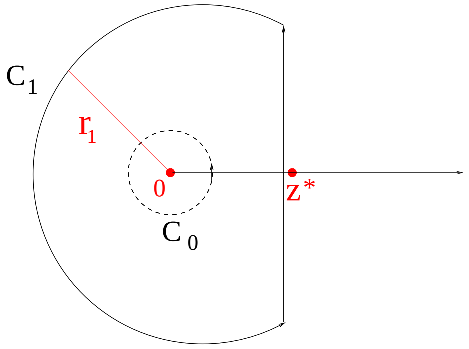

Figure 10.4: The contour C_0 in the complex z plane can be deformed to the contour C_1. In the large n limit, the dominant contribution to the Cauchy integral in Eq. (10.68) comes from the vertical part of C_1.

i.e the imaginary axis located just left of z^*. Next we make a change of variable $z = e^{-sn}$ and define

$$\tilde{q}(s) = \tilde{Q}(z = e^{-s}) = \sum_{n=0}^{\infty} Q(n)\, e^{-sn} = \exp\left[W_{c,\mu}(s)\right] \ , \tag{10.69}$$

$$\text{where} \quad W_{c,\mu}(s) = \sum_{n=1}^{\infty} \frac{e^{-sn}}{n} \int_{cn}^{\infty} P_n(x)\, dx \, . \tag{10.70}$$

Using this expression in the integrand in Eq. (10.68) and retaining only the contribution from the vertical part of the contour C_1 for large n, we get

$$Q(n) \approx \int_{s^*-i\infty}^{s^*+i\infty} \frac{ds}{2\pi i}\, e^{s\,n} \, \exp\left[W_{c,\mu}(s)\right] \ , \tag{10.71}$$

where $W_{c,\mu}(s)$ is given in Eq. (10.70) and $s^* = -\ln(z^*)$ is the singularity of $\tilde{q}(s) = \exp[W_{c,\mu}(s)]$ on the real axis closest to $s = 0$. Identifying the integral on the rhs of Eq. (10.71) as a standard Bromwich integral in the complex s plane, we see that for large n, $Q(n)$ is essentially given by the inverse Laplace transform of the function $\tilde{q}(s) = \exp[W_{c,\mu}(s)]$. To make further progress, we need to first identify the position of the singularity s^* of $W_{c,\mu}(s)$ and then analyse the dominant contribution in the Bromwich integral coming from the neighborhood of s^* for large n. We see below that the singular behavior of $W_{c,\mu}(s)$ as a function of s depends on the parameters c and $\mu > 0$ and there are essentially 5 regimes in the $(c, 0 < \mu \leq 2)$ strip as shown in Fig. 10.1. Below we discuss these regimes separately.

10.4.1 Regime I: $0 < \mu < 1$ and c arbitrary

To analyse the leading singularity of $W_{c,\mu}(s)$ as a function of s in this regime, it is first convenient to use the normalization condition $\int_{-\infty}^{\infty} P_n(x)dx = 1$ and the symmetry $P_n(x) = P_n(-x)$ to rewrite

$$\int_{cn}^{\infty} P_n(x)\, dx = \frac{1}{2} - \int_{0}^{cn} P_n(x)\, dx \, . \tag{10.72}$$

Substituting this in Eq. (10.70) gives

$$W_{c,\mu}(s) = -\frac{1}{2}\ln\left(1-e^{-s}\right) - \sum_{n=1}^{\infty}\frac{e^{-sn}}{n}\int_0^{cn} P_n(x)\,dx\,. \tag{10.73}$$

Now, as $s \to 0$, the sum in Eq. (10.73) converges to a constant for $0 < \mu < 1$

$$S_0 = \sum_{n=1}^{\infty}\frac{1}{n}\int_0^{cn} P_n(x)\,dx\,. \tag{10.74}$$

To see this, let us see how the integral $\int_0^{cn} P_n(x)\,dx$ behaves for large n. For large n, we can use the scaling form for $P_n(x)$ in Eq. (10.7). One finds that $\int_0^{cn} P_n(x)dx \to \int_0^{cn^{(1-1/\mu)}} \mathcal{L}_\mu(y)\,dy$ as $n \to \infty$. For $0 < \mu < 1$, clearly this integral decreases leading to the convergence of the series in Eq. (10.74). Thus, the leading singularity of $W_{c,\mu}(s)$ occurs near $s = s^* = 0$ where it behaves as

$$W_{c,\mu}(s) \approx -\frac{1}{2}\ln(s) - S_0\ . \tag{10.75}$$

Substituting this result in Eq. (10.70) gives

$$\tilde{q}(s) = \sum_{n=0}^{\infty} Q(n)\,e^{-sn} \xrightarrow[s\to 0]{} \frac{e^{-S_0}}{\sqrt{s}}\ . \tag{10.76}$$

We now substitute this singular behavior of the integrand in Eq. (10.71) after setting $s^* = 0$ and perform the standard Bromwich integral (one can use the fact that the inverse Laplace transform $LT^{-1}_{s\to n}[s^{-1/2}] = 1/\sqrt{\pi n}$)

$$Q(n) \xrightarrow[n\to\infty]{} \frac{B_{\rm I}}{\sqrt{n}}\ , \tag{10.77}$$

where the prefactor $B_{\rm I}$ is given by

$$B_{\rm I} = \frac{e^{-S_0}}{\sqrt{\pi}} = \frac{1}{\sqrt{\pi}}\exp\left[-\sum_{n=1}^{\infty}\frac{1}{n}\int_0^{cn} P_n(x)\,dx\right]\ . \tag{10.78}$$

10.4.2 Regime II: $\mu = 1$ and c arbitrary

The case $\mu = 1$ is rather special and marginal as we demonstrate now. Consider the sum $W_{c,1}(s)$ in Eq. (10.70). In this case, it follows from Eq. (10.7) that $P_n(x) \to (1/n)\mathcal{L}_1(x/n)$ as $n \to \infty$, where $\mathcal{L}_1(y) = 1/[\pi(1+y^2)]$ for all y and hence is integrable. Thus the integral $\int_{cn}^{\infty} P_n(x)dx$ converges to a constant for large n

$$\int_{cn}^{\infty} P_n(x)dx \xrightarrow[n\to\infty]{} \int_c^{\infty}\mathcal{L}_1(y)\,dy = 1-\theta(c), \tag{10.79}$$

where

$$\theta(c) = \int_{-\infty}^{c}\mathcal{L}_1(y)dy = \frac{1}{2} + \frac{1}{\pi}\arctan(c)\,. \tag{10.80}$$

Hence, the n-th term of the sum in $W_{c,1}(s)$ behaves, for large n, as $T_n \to (1-\theta(c))\,e^{-sn}/n$. Consequently, the sum $W_{c,1}(s) = \sum_{n\geq 1} T_n$ has a singularity at $s = s^* = 0$. The leading asymptotic behavior near this singularity reads

$$W_{c,1}(s) \xrightarrow[s\to 0]{} -(1-\theta(c))\ln(s) - \gamma_0\ , \tag{10.81}$$

where γ_0 is a constant that depends on the details of $P_n(x)$, in particular on the difference between $P_n(x)$ and its large n asymptotic form $(1/n)\mathcal{L}_1(x/n)$ for finite n

$$\gamma_0 = \sum_{n=1}^{\infty} \left[1 - \theta(c) - \int_{cn}^{\infty} P_n(x)\, dx \right] . \tag{10.82}$$

Using this result on the right hand side (rhs) of Eq. (10.70) gives

$$\tilde{q}(s) \xrightarrow[s\to 0]{} \frac{e^{-\gamma_0}}{s^{1-\theta(c)}} . \tag{10.83}$$

Substituting this result in the Bromwich contour in Eq. (10.71) (after setting $s^* = 0$) and performing the Bromwich integral gives

$$Q(n) \xrightarrow[n\to\infty]{} \frac{B_{\rm II}}{n^{\theta(c)}} , \tag{10.84}$$

where

$$B_{\rm II} = \frac{e^{-\gamma_0}}{\Gamma[1-\theta(c)]} \quad \text{and} \quad \theta(c) = \frac{1}{2} + \frac{1}{\pi} \arctan(c) , \tag{10.85}$$

and γ_0 in given in Eq. (10.82).

Thus, for $\mu = 1$, the persistence $Q(n)$ decays algebraically for large n but with an exponent $\theta(c)$ that depends continuously on the drift c. In this sense the line $\mu = 1$ is *marginally* critical. The exponent $\theta(c)$ in Eq. (10.85) increases continuously with c from $\theta(c \to -\infty) = 0$ to $\theta(c \to \infty) = 1$.

Let us end this subsection with the following remark on the special case of pure Cauchy jump distribution, $f_{\rm Cauchy}(\eta) = 1/[\pi(1+\eta^2)]$. As mentioned before, the record statistics for this case was studied in detail in Ref. [26]. For the Cauchy case, it is well known that $P_n(x) = (1/n) f_{\rm Cauchy}(x/n) = (1/n)\mathcal{L}_1(x/n)$ for *all* n. As a result, it follows from Eq. (10.82) that the constant $\gamma_0 = 0$ in this case. However, in the general $\mu = 1$ case (not necessarily the Cauchy case), γ_0 is generically nonzero. Thus, while the persistence exponent $\theta(c) = 1/2 + \frac{1}{\pi}\arctan(c)$ is universal for all jump densities belonging to the $\mu = 1$ case, the amplitude $B_{\rm II}$ is *nonuniversal* and depends on the details of the jump density.

10.4.3 Regime III: $1 < \mu < 2$ and $c > 0$

To analyse the singular behavior of the sum $W_{c,\mu}(s)$ in Eq. (10.70) in this regime, we consider the n-th term of the sum $T_n = (e^{-sn}/n)\int_{cn}^{\infty} P_n(x)dx$ and substitute, for large n, the scaling behavior of $P_n(x)$ in Eq. (10.7). This gives $T_n \approx (e^{-sn}/n)\int_{cn^{(1-1/\mu)}}^{\infty} \mathcal{L}_\mu(y)dy$. For $1 < \mu < 2$, the lower limit of the integral in T_n becomes large as $n \to \infty$ and we can use the tail behavior in Eq. (10.8) to estimate, $T_n \approx (A_\mu/\mu c^\mu) e^{-sn}/n^\mu$ for large n. Hence the sum, $W_{c,\mu}(s) = \sum_{n=1}^{\infty} T_n$ clearly converges to a constant $W_{c,\mu}(0)$ as $s \to 0$. For small s, one can replace the sum by an integral and estimate exactly the first singular correction to this constant. This gives

$$W_{c,\mu}(s) \xrightarrow[s\to 0]{} W_{c,\mu}(0) - B_\mu\, s^{\mu-1} , \tag{10.86}$$

where the constant $B_\mu = A_\mu \Gamma(2-\mu)/[\mu(\mu-1)c^\mu]$. Using the exact expression of A_μ from Eq. (10.8) and simplify, one finds $B_\mu = -1/[2\cos(\mu\pi/2)] > 0$ as in Eq. (10.40). Note also that from the definition in (10.70)

$$\tilde{q}(0) = \exp[W_{c,\mu}(0)] = \exp\left[\sum_{n=1}^{\infty} \frac{1}{n} \int_{cn}^{\infty} P_n(x)\, dx \right] . \tag{10.87}$$

Substituting the small s behavior from Eq. (10.86) in Eq. (10.70) gives

$$\tilde{q}(s) \xrightarrow[s\to 0]{} \tilde{q}(0)\left[1 - B_\mu\, s^{\mu-1} + \ldots\right] . \tag{10.88}$$

Substituting this singular behavior of $\tilde{q}(s) = \exp[W_{c,\mu}(s)]$ in the Bromwich integral in Eq. (10.71) (upon setting $s^* = 0$) and performing the integral by standard method provides the following large n power law tail for $Q(n)$

$$Q(n) \xrightarrow[n\to\infty]{} \frac{B_{\mathrm{III}}}{n^\mu} , \tag{10.89}$$

where the prefactor B_{III} is given by

$$B_{\mathrm{III}} = \frac{(\mu-1)B_\mu}{\Gamma(2-\mu)c^\mu}\,\tilde{q}(0) = -\frac{(\mu-1)}{2\cos(\mu\pi/2)\,\Gamma(2-\mu)c^\mu}\exp\left[\sum_{n=1}^{\infty}\frac{1}{n}\int_{cn}^{\infty} P_n(x)\,dx\right] . \tag{10.90}$$

10.4.4 Regime IV: $\mu = 2$ and $c > 0$

In this regime, the leading singularity s^* of $W_{c,\mu}(s)$ occurs not at $s = 0$, but at $s = s^* = -s_1$ where $s_1 = c^2/2\sigma^2$. To see this, let us again consider the n-th term of the sum $W_{c,\mu}(s)$, i.e. $T_n = (e^{-sn}/n)\int_{cn}^{\infty} P_n(x)dx$. For large n, $P_n(x)$ now has the Gaussian scaling form in Eq. (10.10) due to the central limit theorem. Substituting this Gaussian form and carrying out the integration one gets,

$$T_n \to \frac{e^{-sn}}{2n}\,\mathrm{erfc}\left(\frac{c}{\sigma\sqrt{2}}\sqrt{n}\right), \quad \text{where} \quad \mathrm{erfc}(y) = \frac{2}{\sqrt{\pi}}\int_y^{\infty} e^{-x^2}\,dx\,. \tag{10.91}$$

Using the asymptotic behavior $\mathrm{erfc}(y) \approx e^{-y^2}/y\sqrt{\pi}$ for large y, one finds that

$$T_n \xrightarrow[n\to\infty]{} \frac{\sigma}{c\sqrt{2\pi}}\,\frac{e^{-(s+s_1)n}}{n^{3/2}}, \quad \text{where} \quad s_1 = \frac{c^2}{2\sigma^2}\,. \tag{10.92}$$

Consequently, the sum $W_{c,\mu}(s) = \sum_{n\geq 1} T_n$ actually, while perfectly analytic near $s = 0$, has a singularity near $s = s^* = -s_1$. Close to this singular value, by taking the limit $s + c^2/2\sigma^2 \to 0$ whereby replacing the sum by an integral over n, one finds the following leading singular behavior of $W_{c,\mu}(s)$ near $s = -s_1$

$$W_{c,\mu}(s) \xrightarrow[s\to -s_1]{} W_{c,\mu}(-s_1) - \sqrt{2}\,\frac{\sigma}{c}\,\sqrt{s+s_1}\,, \tag{10.93}$$

where $W_{c,\mu}(-s_1)$ is just a constant. Substituting this leading singular behavior on the rhs of Eq. (10.70) gives

$$\tilde{q}(s) \xrightarrow[s\to 0]{} e^{W_{c,\mu}(-s_1)}\left[1 - \sqrt{2}\,\frac{\sigma}{c}\,\sqrt{s+s_1} + \ldots\right] . \tag{10.94}$$

We set $s^* = -s_1$ in the Bromwich contour in Eq. (10.71), substitute the singular behavior of $\tilde{q}(s)$ in Eq. (10.94) and perform the Bromwich integral to get

$$Q(n) \xrightarrow[n\to\infty]{} \frac{B_{\mathrm{IV}}}{n^{3/2}}\,e^{-s_1 n} \quad \text{where} \quad s_1 = \frac{c^2}{2\sigma^2} \tag{10.95}$$

and the prefactor

$$B_{\mathrm{IV}} = \frac{\sigma e^{W_{c,\mu}(-s_1)}}{c\sqrt{2\pi}} = \frac{\sigma}{c\sqrt{2\pi}}\exp\left[\frac{e^{s_1 n}}{n}\int_{cn}^{\infty} P_n(x)\,dx\right] . \tag{10.96}$$

Thus, contrary to regimes I, II and III, here the persistence $Q(n)$ has a leading exponential tail (modulated by a power law $n^{-3/2}$).

10.4.5 Regime V: $1 < \mu \leq 2$ and $c < 0$

In this regime $c = -|c| < 0$ and $\mu > 1$. It is convenient, using the normalization condition $\int_{-\infty}^{\infty} P_n(x)dx = 1$, to first reexpress the sum $W_{c,\mu}(s)$ in Eq. (10.70) as

$$W_{c,\mu}(s) = \sum_{n=1}^{\infty} \frac{e^{-sn}}{n} \int\limits_{-|c|n}^{\infty} P_n(x)dx = \sum_{n=1}^{\infty} \frac{e^{-sn}}{n} \left[1 - \int\limits_{|c|n}^{\infty} P_n(x)\,dx\right] . \tag{10.97}$$

Performing the sum, and using the definition of $W_{c,\mu}(s)$ in Eq. (10.70) one gets

$$W_{c,\mu}(s) = -\ln\left(1 - e^{-s}\right) - W_{|c|,\mu}(s) . \tag{10.98}$$

For $\mu > 1$, $W_{|c|,\mu}(0)$ is a constant as was demonstrated in the previous two subsections. Hence, one gets from Eq. (10.98), the leading singular behavior for small s

$$W_{c,\mu}(s) \xrightarrow[s\to 0]{} -\ln(s) - W_{|c|,\mu}(0) \tag{10.99}$$

which yields, via Eq. (10.70)

$$\tilde{q}(s) \xrightarrow[s\to 0]{} \frac{\exp[-W_{|c|,\mu}(0)]}{s} . \tag{10.100}$$

Thus, in this regime, the leading singularity of $\tilde{q}(s)$ occurs at $s = s^* = 0$. Setting $s^* = 0$ and the result (10.100) in the Bromwich integral in Eq. (10.71) gives

$$Q(n) \xrightarrow[n\to\infty]{} \alpha_\mu(c) = \exp[-W_{|c|,\mu}(0)] = \exp\left[-\sum_{n=1}^{\infty} \frac{1}{n} \int\limits_{|c|n}^{\infty} P_n(x)\,dx\right] . \tag{10.101}$$

The fact that the persistence $Q(n)$ approaches to a constant for large n in this regime can be understood physically because for $c < 0$ and $\mu > 1$, a finite fraction of trajectories escape to $-\infty$ as $n \to \infty$.

10.5 Asymptotic record number distribution $P(R,n)$ for large n

In this section, we analyse the asymptotic large n properties of the mean record number $\langle R_n\rangle$ and its full distribution $P(R,n)$ for arbitrary c by analysing the set of equations (10.6), (10.60), (10.63) and (10.65) with arbitrary jump distribution $f(\eta)$. Consider first the mean record number. As in Section IV, we invert Eq. (10.65) by using the Cauchy inversion formula, deform the contour (as in Fig. 10.4), keep only the vertical part of the contour C_1 for large n and finally make the substitution $z = e^{-s}$ to obtain the following Bromwich formula

$$\langle R_n\rangle \approx \int\limits_{s^*-i\infty}^{s^*+i\infty} \frac{ds}{2\pi i}\, e^{s\,n}\, \frac{1}{(1-e^{-s})^2 \tilde{q}(s)} , \tag{10.102}$$

where $\tilde{q}(s)$ is given in Eq. (10.70) and its small s properties have already been analysed in section IV in different regimes in the $(c, 0 < \mu \leq 2)$ strip. As in section IV, s^* denotes the singularity of $\tilde{q}(s)$ on the real line in the complex plane that is closest to the origin at $s = 0$.

Similarly, the record number distribution is obtained by inverting Eq. (10.60) in the same way

$$P(R,n) \approx \int\limits_{s^*-i\infty}^{s^*+i\infty} \frac{ds}{2\pi i}\, e^{s\,n}\, \tilde{q}(s)\, \left[1 - (1-e^{-s})\tilde{q}(s)\right]^{R-1} . \tag{10.103}$$

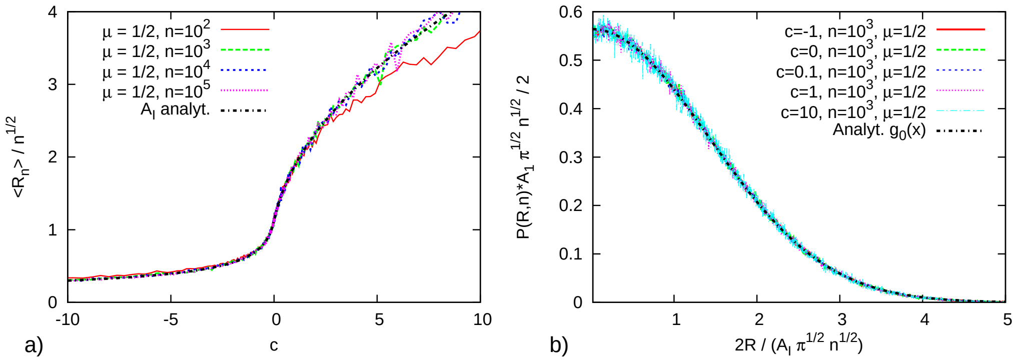

Figure 10.5: a): Rescaled mean record number $\langle R_n\rangle/\sqrt{n}$ for a Lévy-stable distribution with Lévy index $\mu = 1/2$ and different series length $n = 10^2, 10^3, 10^4$ and 10^5. For each n the average was performed over 10^3 samples. For $n \gg 1$ the results collapse and agree with the predicted analytical behavior for $A_{\rm I}\,(c)$ in Eq. (10.104). **b):** Rescaled distribution $A_{\rm I}\sqrt{\pi n}\,P(R,n)/2$ as a function of $2R/A_{\rm I}\sqrt{\pi n}$ of the record number R_n after n steps for a random walk with a Lévy-stable jump distribution of Lévy index $\mu = 1/2$, $n = 10^3$ and different values of the drift $c = -1, 0, 0.1, 1$ and 10. We also plotted the asymptotic analytical result $g_0(x)$ given in Eq. (10.25). All curves collapse nicely. In regime I, the record number has a half-gaussian distribution.

In this section, we use the already derived results for $\tilde{q}(s)$ in Section IV and analyse the asymptotic behavior of $\langle R_n\rangle$ and $P(R,n)$ respectively in Eqs. (10.102) and (10.103) in different regimes of the $(c, 0 < \mu \leq 2)$ strip and on the critical line $\mu = 1$.

10.5.1 Regime I: $0 < \mu < 1$ and c arbitrary

Let us first consider the asymptotic behavior of the mean number of records $\langle R_n\rangle$ for large n in this regime. Consider the Bromwich integral in Eq. (10.102). For large n, this integral can be shown to be dominated by the small s region of the integrand. Taking $s \to 0$ limit in the integrand, substituting the result (10.76) on the rhs of Eq. (10.102), and performing the Bromwich integral we get the leading asymptotic behavior for large n

$$\langle R_n\rangle \approx A_{\rm I}\sqrt{n}\;, \quad \text{where} \quad A_{\rm I} = \frac{2}{\sqrt{\pi}}\,e^{S_0} = \frac{2}{\sqrt{\pi}}\,\exp\left[\sum_{n=1}^{\infty}\frac{1}{n}\int_0^{cn} P_n(x)\,dx\right]\,. \tag{10.104}$$

Comparing this to the amplitude of persistence in Eq. (10.78) we see that the two prefactors are related simply via $B_{\rm I} = 2/(\pi A_{\rm I})$. The prefactor $A_{\rm I}$ can further be expressed explicitly in terms of the Fourier transform of the jump distribution $\hat{f}(k)$ as in Eq. (10.24). This is shown in Appendix A where we also compute the asymptotic behavior of $A_{\rm I}$ for large $|c|$ [see Eq. (10.181)]. In Fig. 10.5 a) we compare this result for $\langle R_n\rangle$ to numerical simulations. The numerical results for $n \gg 1$, $\langle R_n\rangle/\sqrt{n}$ agree nicely with our analytical values for $A_{\rm I}\,(c)$.

Next we turn to $P(R,n)$ in the limit of large n. To extract the scaling behavior of $P(R,n)$ from Eq. (10.103), we substitute on the rhs the small s behavior of $\tilde{q}(s)$ from Eq. (10.76) and use the notation $e^{-S_0} = (2/\sqrt{\pi})\,A_{\rm I}$. The appropriate scaling limit is clearly $R \to \infty$, $s \to 0$ but keeping the product $\sqrt{s}\,R$ fixed. Taking this limit in Eq. (10.103) gives,

$$P(R,n) \approx \int_{-i\infty}^{+i\infty}\frac{ds}{2\pi i}\,e^{s\,n}\,\frac{2}{A_{\rm I}\sqrt{\pi\,s}}\,\exp\left[-\frac{2}{A_{\rm I}\sqrt{\pi}}\,\sqrt{s}\,R\right]\,. \tag{10.105}$$

One can simply evaluate the Bromwich integral by using the identity,

$$LT^{-1}_{s\to n}[e^{-bR\sqrt{s}}/\sqrt{s}] = e^{-b^2R^2/4n}/\sqrt{\pi n}. \tag{10.106}$$

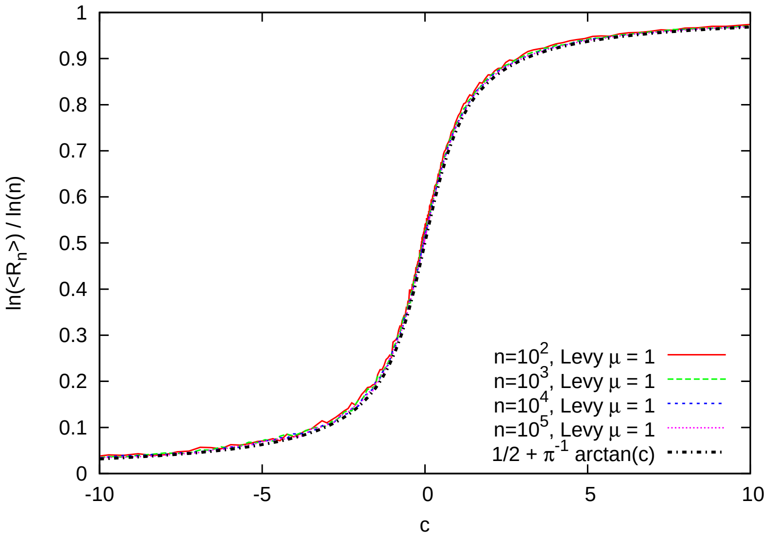

Figure 10.6: $\ln\langle R_n\rangle/\ln n$ as a function of the drift c for the Cauchy distribution with Lévy index $\mu = 1$ and for different values of $n = 10^2, 10^3, 10^4$ and 10^5. For each n and c, the average was performed over 10^3 samples. The results from the numerical simulations collapse and agree with the predicted analytical behavior of $\ln\langle R_n\rangle/\ln n = \theta(c)$ and $\theta(c) = \frac{1}{2} + \frac{1}{\pi}\arctan(c)$ as in Eq. (10.85).

This leads to the asymptotic result announced in Eq. (10.25) in the scaling limit $n \to \infty$, $R \to \infty$ with the ratio $R/\sqrt{n}$ fixed. In Fig. 10.5 b) we computed numerically the rescaled distribution $A_{\rm I}\sqrt{\pi n}\,P(R,n)/2$ as a function of $2R/A_{\rm I}\sqrt{\pi n}$ and compared it with $g_0(x)$ Eq. (10.25). The figure confirms that in regime I, the record number has a half-Gaussian distribution with a width that depends non-trivially on the drift c and the Lévy-index μ.

In summary, for $0 < \mu < 1$, the drift is not strong enough to change the $\sqrt{n}$ growth of the mean record number. The presence of drift just modifies the prefactor of the $\sqrt{n}$ growth. Similarly, the distribution of the record number in Eq. (10.25) in presence of a drift, when appropriately scaled, remains unchanged from the universal half-Gaussian form in the driftless case.

10.5.2 Regime II: $\mu = 1$ and c arbitrary

As mentioned in the introduction, on the critical line $\mu = 1$, the record statistics was investigated in detail in Ref. [26] for the special case of Cauchy jump distribution $f_{\rm Cauchy}(\eta) = 1/[\pi(1+\eta^2)]$. For a general jump distribution with $\mu = 1$ (not necessarily of the Cauchy form), the record statistics has a very similar mathematical structure that can be derived from the general framework developed in this paper.

Let us first consider the growth of the mean record number $\langle R_n\rangle$ in Eq. (10.102). Substituting the small s behavior of $\tilde{q}(s)$ from Eq. (10.83) and performing the Bromwich integral upon setting $s^* = 0$ we get for large n

$$\langle R_n\rangle \approx \frac{A_{\rm II}}{\Gamma(1+\theta(c))}\, n^{\theta(c)} \quad \text{where} \quad A_{\rm II} = e^{\gamma_0}\,. \tag{10.107}$$

Note that γ_0 is a distribution dependent constant while the exponent $\theta(c) = \int_{-\infty}^{c} \mathcal{L}_1(y)dy = 1/2 + \frac{1}{\pi}\arctan(c)$ is universal. In Fig. (10.6) this exponent is plotted and compared with numerical simulations of random walks with a Cauchy jump distribution ($\mu = 1$).

Turning now to the distribution $P(R, n)$ in Eq. (10.103), as before, we substitute the small s expansion of $\tilde{q}(s)$ from Eq. (10.83). It turns out that the appropriate scaling limit for $P(R,n)$ is $n \to \infty$, $R \to \infty$ but keeping the ratio $R/n^{\theta(c)}$ fixed. To see this, we first set $s^* = 0$, set R large but fixed, and keep the leading terms for small s to get

$$P(R,n) \approx e^{-\gamma_0} \int_{-i\infty}^{+i\infty} \frac{ds}{2\pi i}\, e^{s\,n}\, \frac{1}{s^{1-\theta(c)}}\, \exp\left[-e^{-\gamma_0}\, s^{\theta(c)}\, R\right] . \tag{10.108}$$

Rescaling $s\,n \to s$ and keeping the scaled variable $R/n^{\theta(c)}$ fixed gives the asymptotic scaling distribution

$$P(R,n) \approx \frac{1}{A_{\rm II}\, n^{\theta(c)}}\, g_c\left(\frac{R}{A_{\rm II}\, n^{\theta(c)}}\right) , \tag{10.109}$$

where the scaling function $g_c(u)$, which depends continuously on c, is given by the formal Bromwich integral

$$g_c(u) = \int_{-i\infty}^{+i\infty} \frac{ds}{2\pi i}\, s^{\theta(c)-1}\, e^{s-us^{\theta(c)}} \quad \text{with} \quad u \geq 0 , \tag{10.110}$$

where we recall that $0 \leq \theta(c) \leq 1$.

One can easily extract the tail behavior of the scaling function $g_c(u)$ by analysing the integral in Eq. (10.110). For instance, when $u \to 0$, $g_c(u)$ approaches a constant

$$g_c(0) = \int_{-i\infty}^{+i\infty} \frac{ds}{2\pi i}\, s^{\theta(c)-1}\, e^{s} = \frac{1}{\pi}\, \Gamma[\theta(c)]\, \sin[\pi\theta(c)] = \frac{1}{\Gamma[1-\theta(c)]} . \tag{10.111}$$

The integral in Eq. (10.111) can be performed by wrapping the contour around the branch cut on the negative real s axis.

In the opposite limit, when $u \to \infty$, the integral in Eq. (10.110) can be performed using the standard steepest descent method. Skipping details and using the shorthand notation $\theta = \theta(c)$ we get

$$\begin{aligned} g_c(u \to \infty) \quad &\approx \quad \left[2\pi(1-\theta)\, \theta^{(1-2\theta)/(1-\theta)}\right]^{-1/2}\, u^{-(1-2\theta)/2(1-\theta)} \times \\ &\quad \times\, \exp\left[-(1-\theta)\, \theta^{\theta/(1-\theta)}\, u^{1/(1-\theta)}\right] . \end{aligned} \tag{10.112}$$

Thus the distribution has a non-Gaussian tail. The function $g_c(u)$ can be expressed in terms of the one-sided Lévy distribution, which was discussed for instance in Ref. [33]. In some particular case, the Bromwich integral in Eq. (10.110) can be evaluated explicitly. For rational values of $\theta(c)$, $g_c(u)$ can be expressed as a finite sum of hypergeometric functions. A very special case corresponds to $c = -1/\sqrt{3}$ where one has $\theta = 1/3$, such that

$$g_{c=-1/\sqrt{3}}(u) = 3^{2/3} \mathrm{Ai}\left(\frac{u}{3^{1/3}}\right) , \; u \geq 0 . \tag{10.113}$$

where $\mathrm{Ai}(x)$ is the Airy function. Its asymptotic behaviors are then given by

$$g_{c=-1/\sqrt{3}}(u) \sim 1/\Gamma(2/3) , \; u \to 0 \tag{10.114}$$

$$g_{c=-1/\sqrt{3}}(u) \sim \frac{3^{3/4}}{2\sqrt{\pi}} u^{-1/4} \exp\left(-\frac{2}{3\sqrt{3}} u^{3/2}\right) , \tag{10.115}$$

which agree with the general analysis presented above (10.111, 10.112). In Fig. 10.7 we show a plot of the rescaled probability $A_{\rm II}\, n^{\theta(c)}\, P(R,n)$ as a function of $R/A_{\rm II} n^{\theta(c)}$ computed numerically for $c = -1/\sqrt{3}$, which agrees reasonably well with our exact analytical result in Eq. (10.113).

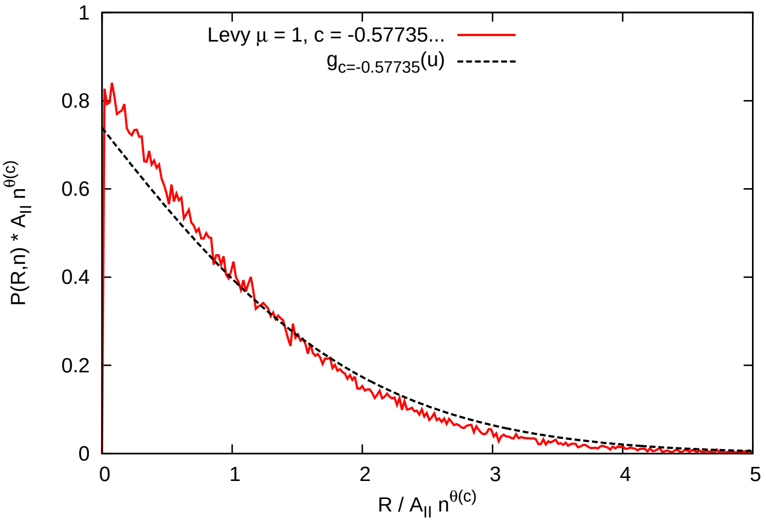

Figure 10.7: Rescaled plot of $A_{\mathrm{II}}\, n^{\theta(c)}\, P(R,n)$ as a function of $R/A_{\mathrm{II}} n^{\theta(c)}$ for $\mu = 1$ and $c = -1/\sqrt{3} = -0.57735$, and $\theta(c) = 1/3$ (regime II). These data have been obtained for $n = 10^5$ and averaged over 10^5 samples. The dotted line corresponds to our exact result in Eq. (10.113).

10.5.3 Regime III: $1 < \mu < 2$ and $c > 0$

We first compute the asymptotic growth of the mean number of records in this regime. Substituting the leading singular behavior of $\tilde{q}(s)$ from Eq. (10.88) on the rhs of Eq. (10.102) and performing the Bromwich integral gives

$$\langle R_n \rangle \approx a_\mu(c)\, n \quad \text{where} \quad a_\mu(c) = \frac{1}{\tilde{q}(0)} = \exp\left[-\sum_{n=1}^{\infty} \frac{1}{n} \int_{cn}^{\infty} P_n(x)\, dx \right] . \tag{10.116}$$

Note that we used above the expression of $\tilde{q}(0)$ in Eq. (10.87). We have checked numerically this linear growth and in Fig. 10.10 the bottom curve shows a plot of $\langle R_n \rangle / n$ as a function of c, although we have not tried to evaluate $a_\mu(c)$ numerically.

We next consider the distribution $P(R,n)$ in Eq. (10.103). We substitute the small s behavior of $\tilde{q}(s)$ from Eq. (10.88) on the rhs of Eq. (10.103), set $s^* = 0$, R large and keep only leading small s terms to get

$$P(R,n) \approx \tilde{q}(0) \int_{-i\infty}^{+i\infty} \frac{ds}{2\pi i} \exp\left[-s\left(\tilde{q}(0)R - n\right) + B_\mu \tilde{q}(0) R s^\mu \right] . \tag{10.117}$$

Next we set

$$R = a_\mu(c)\, n + a_\mu(c)\, n^{1/\mu}\, u \ , \tag{10.118}$$

where $a_\mu(c) = 1/\tilde{q}(0)$ and take the limit $R \to \infty$, $n \to \infty$ but keeping the scaled variable u above fixed. We substitute Eq. (10.118) on the rhs of Eq. (10.117). Keeping only the two leading terms for large n and fixed u gives

$$P(R,n) \approx \tilde{q}(0) \int_{-i\infty}^{+i\infty} \frac{ds}{2\pi i} \exp\left[-s n^{1/\mu} u + B_\mu\, n\, s^\mu \right] . \tag{10.119}$$

Note that for fixed u, both terms inside the exponential are of the same order. In fact, the scaling in Eq. (10.118) is chosen so as to make the two leading terms precisely of the same

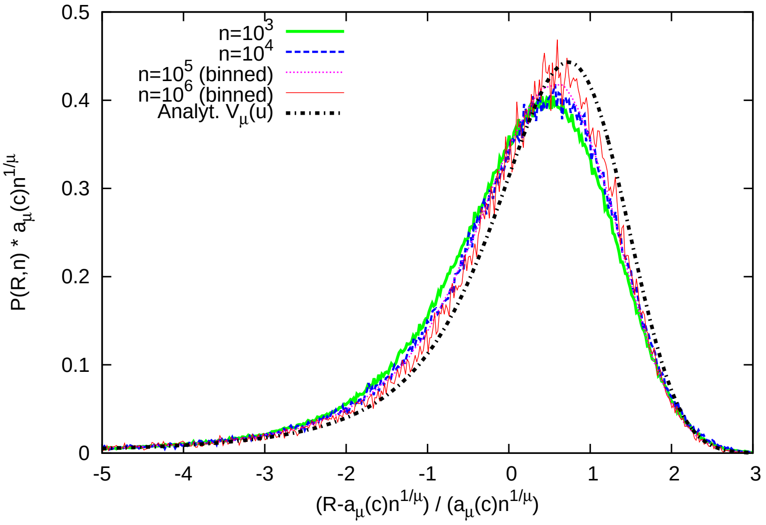

Figure 10.8: Rescaled distribution $a_\mu(c)\, n^{1/\mu} P(R,n)$ of the record number R_n after n steps for a random walk with a Lévy-stable jump distribution of Lévy index $\mu = 1.5$. The data are plotted as a function of the shifted and scaled variable $u = (R - a_\mu(c)\, n)/(a_\mu(c)\, n^{1/\mu})$. The different curves correspond to different values of $n = 10^3, 10^4, 10^5$ and 10^6 and for a drift $c = 1$. They were obtained by averaging over 10^6 samples. For $n = 10^5$ and $n = 10^6$ the numerical results were binned for technical reasons. We also plotted our analytical results for the scaling function $V_\mu\,(u)$ given by Eq. (10.122). While for smaller values of n, there is still a significant difference between the simulations and our analytical result, it converges to the behavior in Eq. (10.122) when n increases.

order for large n. Rescaling s by $n^{1/\mu}$, i.e., $s\, n^{1/\mu} \to s$ and using $a_\mu(c) = 1/\tilde{q}(0)$ reduces Eq. (10.119) to a nicer scaling form announced in Eq. (10.35)

$$P(R,n) \approx \frac{1}{a_\mu(c) n^{1/\mu}}\, V_\mu(u), \quad \text{where} \quad u = \frac{R - a_\mu(c)\, n}{a_\mu(c)\, n^{1/\mu}}\;, \tag{10.120}$$

and the scaling function $V_\mu(u)$ is formally given by the Bromwich integral

$$V_\mu(u) = \int_{-i\infty}^{i\infty} \frac{ds}{2\pi i}\, e^{-u\, s + B_\mu\, s^\mu}\;, \tag{10.121}$$

where the constant $B_\mu > 0$ is given in Eq. (10.40).

Interestingly, the same scaling function $V_\mu(u)$ also appeared in Ref. [34] in the context of the partition function of the zero range process on a ring. The asymptotic tails of the function $V_\mu(u)$ were analysed in great detail in [34] (see Eqs. (78)-(83) and Fig. 5 in Ref. [34] and note that in [34], the index μ was denoted by $\gamma - 1$). We do not repeat the computations here, but just quote the results. It was found that $V_\mu(u)$ has highly asymmetric tails. For $u \to -\infty$, it decays as a power law, $V_\mu(u) \to K_\mu |u|^{-\mu-1}$ where the prefactor $K_\mu = B_\mu \Gamma(1+\mu) \sin[\pi(\mu+1)]/\pi$. Using our expression $B_\mu = -1/(2\cos(\mu\pi/2))$ from Eq. (10.40), it is easy to show that $K_\mu = A_\mu$ where the constant A_μ is defined in Eq. (10.8). This gives Eq. (10.36). In contrast, when $u \to \infty$, $V_\mu(u)$ has a faster than Gaussian tail as described precisely in Eq. (10.37). To plot this scaling function, a convenient real space representation can be used from Ref. [34]. Replacing $\gamma - 1$ by μ in Eq. (84) of Ref. [34] and using $B_\mu = -1/2\cos(\mu\pi/2)$, we obtain

$$V_\mu(u) = \frac{1}{\pi} \int_0^\infty dy\, e^{-y^\mu/2} \cos\left[\frac{1}{2}\tan(\mu\pi/2)\, y^\mu + y\, u\right]. \tag{10.122}$$

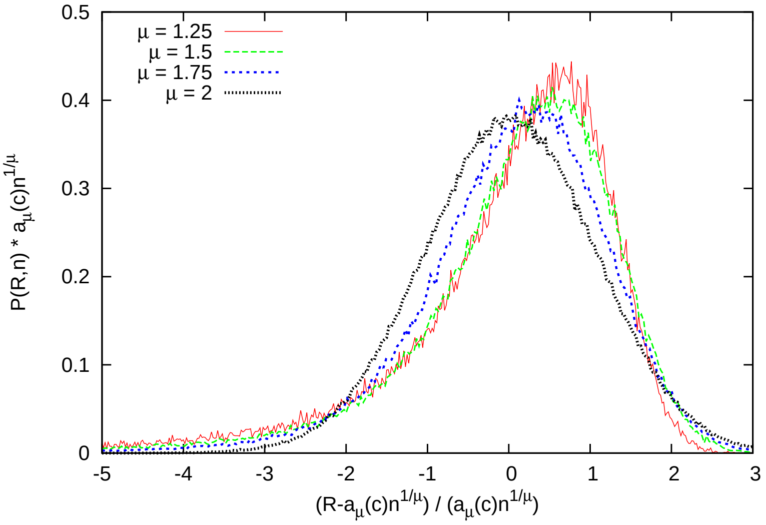

Figure 10.9: Rescaled distribution $a_\mu(c)\,n^{1/\mu}P(R,n)$ of the record number R_n after $n = 10^4$ steps for a random walk with a Lévy-stable jump distribution with different Lévy indices $\mu = 1.25, 1.5, 1.75$ and $\mu = 2$. The data are plotted as a function of the shifted and scaled variable $u = (R - a_\mu(c)\,n)/(a_\mu(c)\,n^{1/\mu})$. For all these data, the value of the drift is $c = 1$ and they have been obtained by averaging over 10^6 samples. The figure shows that for $\mu \to 2$ this rescaled distribution approaches the Gaussian form given in Eq. (10.135).

We compared this result for a Lévy index of $\mu = 1.5$ to numerical simulations in Fig. 10.8. Even though the convergence of the numerically obtained distributions is slow, it is clear that the asymptotic distribution $V_\mu\,(u)$ is approached for $n \to \infty$. In Fig. 10.9 we plotted numerical simulations of the rescaled record number distribution for different values of μ. One finds both numerically and by taking the limit in Eq. (10.135) that, for $\mu \to 2$, this rescaled distribution approaches a Gaussian form (see regime IV).

To summarize, in this regime the mean record number increases linearly with increasing n, but the typical fluctuations around the mean are anomalously large of $\mathcal{O}(n^{1/\mu})$ (superdiffusive) as described in Eq. (10.118). In addition, the probability distribution of these typical fluctuations around the mean are described by a highly non-Gaussian form described precisely in Eq. (10.120).

10.5.4 Regime IV: $\mu = 2$ and $c > 0$

In this regime, as explained in section IV.C, $\tilde{q}(s) = \exp[W_{c,\mu}(s)]$ in Eq. (10.70) is analytic at $s = 0$. This can be seen by expanding the sum $W_{c,\mu}(s)$ in Eq. (10.70) in a Taylor series in s

$$W_{c,\mu}(s) = \sum_{m=0}^{\infty} d_m\, s^m, \quad \text{where} \quad d_m = \frac{(-1)^m}{m!} \sum_{n=1}^{\infty} n^{m-1} \int\limits_{cn}^{\infty} P_n(x)\, dx\,. \tag{10.123}$$

The coefficient d_m, for each m, is finite as the sum over n is convergent since the integral $\int_{cn}^{\infty} P_n(x)\,dx$ decreases with n faster than exponentially for large n (see section IV.C), as long as $\mu = 2$ and $c > 0$. Consequently, for small s, $\tilde{q}(s)$ also has a Taylor series expansion

$$\tilde{q}(s) = \tilde{q}(0) + \tilde{q}'(0)\,s + \frac{1}{2}\tilde{q}''(0)s^2 + \ldots \tag{10.124}$$

Let us start with the asymptotic behavior of the mean record number $\langle R_n \rangle$ in Eq. (10.102). Once again, the dominant contribution to the integral in Eq. (10.102) for large n comes

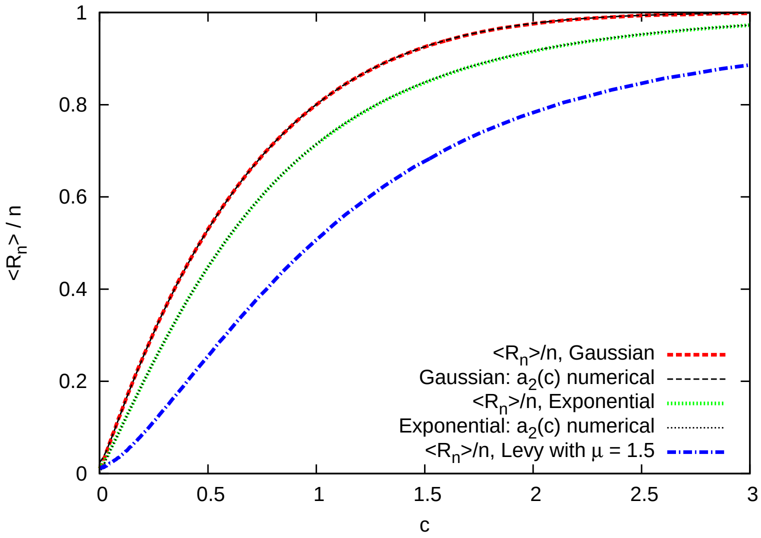

Figure 10.10: Numerical simulations of $\langle R_n \rangle / n$ for random walks with a Gaussian (with variance $\sigma = 1$), an exponential [with parameter $b = 1$, see its definition below Eq. (10.127)], both regime IV, and a Lévy-stable jump distribution with $\mu = 1.5$, in regime III, with positive drift $c > 0$. For each distribution we show data for $n = 10^4$ which were obtained by averaging over 10^4 samples. For the Gaussian and the exponential distribution we also plotted a numerical evaluation of our exact formula for $a_2\,(c)$ using Eq. (10.127) for the Gaussian case and Eq. (10.199) for the exponential case. Those curves agree perfectly with the numerical simulations.

from the small s region. Taking the $s \to 0$ limit in the integrand and using the small s expansion in Eq. (10.124), keeping only the leading terms and performing the Bromwich integral term by term one gets for large n

$$\begin{aligned} \langle R_n \rangle &\approx \int_{s^*-i\infty}^{s^*+i\infty} \frac{ds}{2\pi i}\, e^{s\,n}\, \frac{1}{\tilde{q}(0)s^2} \left[1 + (1 - \frac{\tilde{q}'(0)}{\tilde{q}(0)})\, s + O(s^2) \right] \\ &\approx a_2(c) n + \kappa_2(c) + \mathcal{O}(1/n) \end{aligned} \tag{10.125}$$

where

$$a_2(c) = \frac{1}{\tilde{q}(0)} = \exp\left[- \sum_{n=1}^{\infty} \frac{1}{n} \int_{cn}^{\infty} P_n(x)\, dx \right] \tag{10.126}$$

and $\kappa_2(c) = \left[1 - \tilde{q}'(0)/\tilde{q}(0)\right]/\tilde{q}(0)$.

For example, for a Gaussian jump distribution $f(\eta) = (2\pi\sigma^2)^{-1/2} e^{-\eta^2/2\sigma^2}$, we have $P_n(x) = (2\pi n \sigma^2)^{-1/2}\, e^{-x^2/2\sigma^2 n}$ and hence $a_2(c)$ in Eq. (10.126) is given by the explicit formula

$$a_2(c) = \exp\left[- \sum_{n=1}^{\infty} \frac{1}{2n} \mathrm{erfc}\left(\frac{c\sqrt{n}}{\sigma\,\sqrt{2}} \right) \right] . \tag{10.127}$$

For instance, for $c = 1$ and $\sigma = 1$, one gets $a_2(c = 1) = 0.800543\ldots$. Another example is the exponential jump distribution $f(\eta) = (2\,b)^{-1} \exp(-|x|/b)$. In this case, one can also compute (see the Appendix B) the constant $a_2(c) = \lambda$ where λ is given by the solution of the transcendental equation $\exp(-\lambda\, c/b) = 1 - \lambda^2$. For example, for $c = 1$, $b = 1$, one gets $\lambda = 0.714556\ldots$. For these two examples, we have confirmed the leading asymptotic result for the mean record number in Eq. (10.125) with the exactly computed prefactors $a_2(c)$ (as discussed above) in our numerical simulations (see Fig. 10.10).

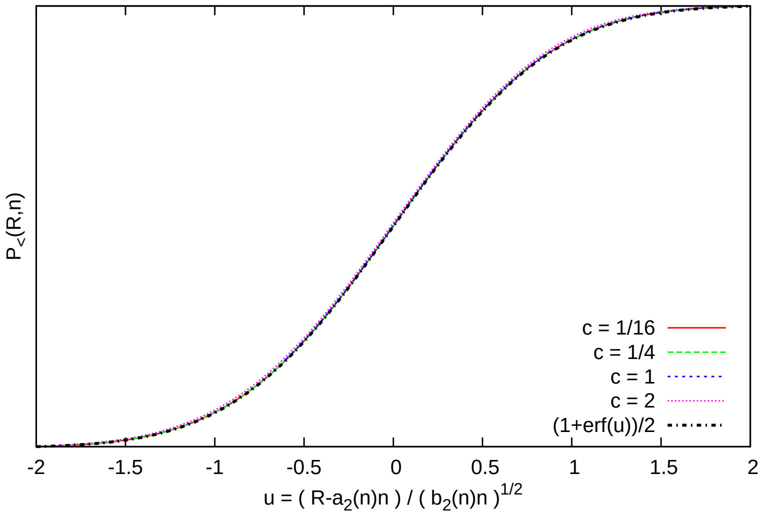

Figure 10.11: Plot of the cumulative distribution of record numbers $P_<(R, n) = \text{Proba.}\,[R_n \le R]$ as a function of the shifted and scaled variable $u = (R - a_2(c)\,n)/(\sqrt{b_2(c)\,n})$ for a random walk with Gaussian jump distribution (with $\sigma = 1$) of $n = 10^4$ steps. The different curves correspond to different values of positive drift $c = 1/16, 1/4, 1$ and 2. For each c the data were obtained by averaging over 10^6 samples. We compared the numerical results to the cumulative distribution of $V_2\,(\mu)$, which we obtained analytically (Eq. (10.135)). All curves collapse nicely, confirming that the asymptotic record number of a biased Gaussian random walk with a positive drift has the Gaussian distribution given by Eq. (10.133).

In a similar way, one can also analyse Eq. (10.66) for the large n behavior of the second moment $\langle R_n^2 \rangle$. Skipping details, we get the following leading large n behavior

$$\langle R_n^2 \rangle \approx a_2^2(c)\,n^2 + \rho_2(c)\,n + \mathcal{O}(1)\ , \quad \text{where} \quad \rho_2(c) = \frac{1}{\tilde{q}^2(0)}\left[3 - \tilde{q}(0) - 4\frac{\tilde{q}'(0)}{\tilde{q}(0)}\right]\ . \tag{10.128}$$

Consequently, the variance of the record number grows for large n as

$$\langle R_n^2 \rangle - \langle R_n \rangle^2 \approx b_2(c)\,n \quad \text{where} \quad b_2(c) = \frac{1}{\tilde{q}^2(0)}\left[1 - \tilde{q}(0) - 2\frac{\tilde{q}'(0)}{\tilde{q}(0)}\right]\ . \tag{10.129}$$

Thus, in this regime, the mean record number grows linearly with n for large n while the size of typical fluctuations around this mean grows as $\sim \sqrt{n}$.

How are these typical fluctuations around the mean distributed? To answer this, we need to analyse $P(R, n)$ in Eq. (10.103) in the scaling limit where both n and R are large, but the ratio $(R - a_2(c)n)/\sqrt{n}$ is fixed. To proceed, we set $s^* = 0$ and substitute the small s expansion of $\tilde{q}(s)$ in Eq. (10.124) on the rhs of Eq. (10.103), take R large but fixed to get

$$P(R, n) \approx \tilde{q}(0) \int_{-i\infty}^{+i\infty} \frac{ds}{2\pi i}\, \exp\left[-s\,(\tilde{q}(0)R - n) + (1/2)\,b_2(c)\,\tilde{q}^3(0)\,Rs^2\right] \tag{10.130}$$

where $b_2(c)$ is given in Eq. (10.129). Next we set

$$R = a_2(c)\,n + \sqrt{b_2(c)}\,\sqrt{n}\,u\ , \tag{10.131}$$

where $a_2(c) = 1/\tilde{q}(0)$ is given in Eq. (10.126) and take the scaling limit where $R \to \infty$, $n \to \infty$ but keeping the scaled variable u above fixed. Substituting R from Eq. (10.131) into Eq. (10.130) and keeping only the two leading terms for large n gives

$$P(R,n) \approx \tilde{q}(0) \int\limits_{-i\infty}^{+i\infty} \frac{ds}{2\pi i} \exp\left[-\sqrt{b_2(c)}\,\tilde{q}(0)\,\sqrt{n}\,s\,u + (1/2)\,b_2(c)\,\tilde{q}^2(0) n\, s^2\right] . \tag{10.132}$$

Note that for fixed u, both terms inside the exponential are of the same order. Indeed, as in the section VB, the scaling in Eq. (10.131) is chosen so as to make the two leading terms precisely of the same order for large n. Rescaling $\sqrt{b_2(c)}\tilde{q}(0)\sqrt{n}\,s \to s$ simplifies to

$$P(R,n) \approx \frac{1}{\sqrt{b_2(c) n}}\, V_2(u) \quad \text{where} \quad u = \frac{R - a_2(c)\,n}{\sqrt{b_2(c)\,n}} , \tag{10.133}$$

and the scaling function $V_2(u)$ is given by the Bromwich integral

$$V_2(u) = \int\limits_{-i\infty}^{i\infty} \frac{ds}{2\pi i}\, e^{-u\,s + s^2/2} , \tag{10.134}$$

which can be exactly computed (since it is a Gaussian integral) to give

$$V_2(u) = \frac{1}{\sqrt{2\pi}} \exp[-u^2/2] . \tag{10.135}$$

This then proves that $P(R,n)$ is asymptotically Gaussian as announced in Eq. (10.45). Fig. 10.11 confirms this result numerically. We plotted the cumulative distribution of record numbers $P_<(R,n) = \text{Proba.}\,[R_n \le R]$ as a function of the shifted and scaled variable $u = (R - a_2(c)\,n)/(\sqrt{b_2(c)\,n})$ after $n = 10^4$ steps for different values of positive drift c and compared them to a Gaussian cdf (cumulative distribution function). All numerical results collapsed perfectly on the analytical curve.

10.5.5 Regime V: $1 < \mu \le 2$ and $c < 0$

In this regime, we set $s^* = 0$ in Eq. (10.103) and substitute on its rhs the small s expansion of $\tilde{q}(s)$ from Eq. (10.100). Keeping only leading order behavior for small s gives, for large n,

$$P(R,n) \approx \alpha_\mu(c)[1-\alpha_\mu(c)]^{R-1} \int\limits_{s^*-i\infty}^{s^*+i\infty} \frac{ds}{2\pi i}\, e^{s\,n}\, \frac{1}{s} , \tag{10.136}$$

where the constant $\alpha_\mu(c) = \exp[-W_{|c|,\mu}(0)] = \exp\left[-\sum_{n=1}^{\infty} \frac{1}{n} \int_{|c|n}^{\infty} P_n(x)\,dx\right]$ as given in Eq. (10.101).

Using the fact that $LT^{-1}_{s\to n}[1/s] = 1$ gives the large n (but R fixed) behavior of $P(R,n)$

$$P(R,n) \xrightarrow[n\to\infty]{} \alpha_\mu(c)\,[1-\alpha_\mu(c)]^{R-1} . \tag{10.137}$$

Thus, the distribution becomes independent of n for large n and has a simple geometric form with mean $\langle R_n \rangle \to 1/\alpha_\mu(c)$. Comparing the expression of $\alpha_\mu(c)$ as given in Eq. (10.101) and those of $a_\mu(c)$ in Eq. (10.116) and $a_2(c)$ in Eq. (10.126) for $c > 0$, one immediately finds that $\alpha_\mu(c) = a_\mu(|c|)$ for $1 < \mu < 2$ while $\alpha_2(c) = a_2(|c|)$, the results mentioned respectively in Eqs. (10.50) and (10.51).

In Fig. 10.12 we compared Eq. (10.137) to numerical simulations of negatively biased Gaussian random walks with different values of c. For large n the rescaled distribution of $u = R\,a_2(|c|)$ approaches the geometric distribution e^{-u}.

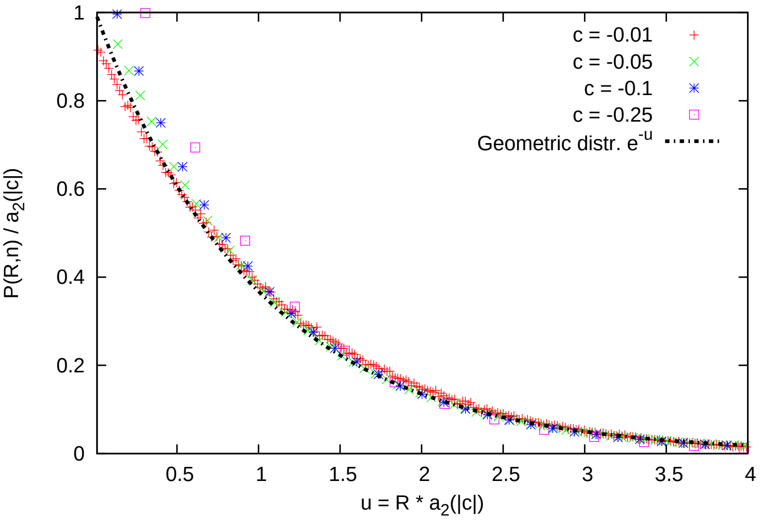

Figure 10.12: Rescaled distribution $a_2(|c|)P(R,n)$ of the record number R_n after $n = 10^4$ steps for a random walk with a Gaussian jump distribution, of variance $\sigma = 1$, with different negative values of the drift $c = -0.01, c = -0.05, -0.1$ and -0.25. The data are plotted as a function of the rescaled variable $u = R\,a_2(|c|)$. For each value of c the data were obtained by averaging over 10^4 samples. We compared the numerical results with a simple geometric distribution. The good agreement confirms our analytical findings given by Eq. (10.137).

10.6 Extreme statistics of the age of a record

From the previous study of the mean number of records $\langle R_n \rangle$, one deduces that the typical age (see Fig. 10.2)) of a record is given by $l_{\rm typ} \sim n/\langle R_n \rangle$. However, following Ref. [22] for the unbiased case, it turns out that the extreme ages of records do not share the typical behavior. In this section, we probe such atypical extremal statistics by considering the longest and shortest lasting records characterized by their respective ages (durations) $l_{\max,n}$ and $l_{\min,n}$. We focus on their mean values $\langle l_{\max,n} \rangle$, $\langle l_{\min,n} \rangle$ and find rather different asymptotic behaviors in the five regimes in the $(c, 0 < \mu \leq 2)$ strip mentioned before (Fig. 10.1).

10.6.1 Age of the longest lasting record $l_{\max,n}$

We first consider the longest lasting record whose age $l_{\max,n}$ is given by (see Fig. 10.2)

$$l_{\max} = \max(l_1, l_2, \cdots, l_R) \ . \tag{10.138}$$

The cumulative distribution $\mathcal{F}_n(m) = \text{Proba.}\,(l_{\max,n} \leq m)$ was studied in Ref. [22], where an explicit formula for its generating function (GF) was obtained:

$$\sum_{n=0}^{\infty} \mathcal{F}_n(m) z^n = \frac{\sum_{l=1}^{m} Q(l) z^l}{1 - \sum_{l=1}^{m} F(l) z^l} \ , \tag{10.139}$$

where $F(l) = Q(l-1) - Q(l)$, from which one deduces the generating function of the mean $\langle l_{\max,n} \rangle = \sum_{m=1}^{\infty} [1 - \mathcal{F}_n(m)]$

$$\sum_{n=0}^{\infty} z^n \langle l_{\max,n} \rangle = \sum_{m=1}^{\infty} \frac{1}{1-z} - \frac{\sum_{l=1}^{m} Q(l) z^l}{1 - \sum_{l=1}^{m} F(l) z^l} \tag{10.140}$$

$$= \frac{1}{1-z} \sum_{m=1}^{\infty} \frac{\sum_{l=m}^{\infty} F(l) z^l + (1-z) \sum_{l=m}^{\infty} Q(l) z^l}{(1-z)\tilde{Q}(z) + \sum_{l=m}^{\infty} F(l) z^l} \ , \tag{10.141}$$

where we have used that $\tilde{F}(z) = 1 - (1-z)\tilde{Q}(z)$ (10.58).

In the absence of drift, $c = 0$, it was shown in Ref. [22] that $\langle l_{\max,n} \rangle$ behaves, for large n, linearly with n with a non trivial coefficient, independently of the jump distribution $f(\eta)$

$$\begin{aligned} \langle l_{\max,n} \rangle \sim C_0\, n \ , C_0 &= \int\limits_0^\infty dy \frac{1}{1 + y^{1/2} e^y \int_0^y dx\, x^{-1/2} e^{-x}} \\ &= 0.626508... \end{aligned} \tag{10.142}$$

Interestingly, this constant C_0 appears also in the study of the longest excursion of Brownian motion [31, 32]. Note that to obtain the large n behavior of $\langle l_{\max,n} \rangle$ from Eq. (10.140) one has to analyse the above formula (10.140) in the limit $z \to 1$. We will see that in this limit the above sum over m is dominated by the large values of m, which thus depends crucially on the large m behavior of the persistence probability $Q(m)$. Consequently $\langle l_{\max,n} \rangle$ behaves quite differently in the five regimes in the $(c, 0 < \mu \leq 2)$ strip in Fig. 10.1 and are summarized as follows:

$$\begin{aligned} \langle l_{\max,n} \rangle &\sim n \quad \text{for} \quad 0 < \mu < 1 \text{ and } c \text{ arbitrary} \quad \text{(regime I)} \ , \\ &\sim n \quad \text{for} \quad \mu = 1 \text{ and } c \text{ .arbitrary} \quad \text{(regime II)} \ , \\ &\sim n^{\frac{1}{\mu}} \quad \text{for} \quad 1 < \mu < 2 \text{ and } c > 0 \quad \text{(regime III)} \ , \\ &\sim \ln n \quad \text{for} \quad \mu = 2 \text{ and } c > 0 \quad \text{(regime IV)} \ , \\ &\sim n \quad \text{for} \quad 1 < \mu \leq 2 \text{ and } c < 0 \quad \text{(regime V)} \ . \end{aligned} \tag{10.143}$$

In the following we will discuss the behavior of $\langle l_{\max,n} \rangle$ separately for the five regimes.

10.6.1.1 Regime I: $0 < \mu < 1$, c arbitrary:

In this regime, we remind that $Q(m)$ behaves, for large m, as

$$Q(m) \sim \frac{B_{\rm I}}{\sqrt{m}} \ , \ F(m) \sim \frac{B_{\rm I}}{2 m^{3/2}} \ , \tag{10.144}$$

where $B_{\rm I}$ is given in Eq. (10.78). Setting $z = e^{-s}$ we are interested in the limit $s \to 0$ in the formula in Eq. (10.140) where one can replace $F(m)$ and $Q(m)$ by their asymptotic behaviors

$$\sum_{n=0}^{\infty} \langle l_{\max,n} \rangle e^{-sn} \sim \frac{1}{s} \sum_{m=1}^{\infty} \frac{\frac{1}{2} \sum_{l=m}^{\infty} l^{-3/2} e^{-sl} + s \sum_{l=m}^{\infty} l^{-1/2} e^{-sl}}{\sqrt{\pi} s^{1/2} + \frac{1}{2} \sum_{l=m}^{\infty} l^{-3/2} e^{-sl}} \ , \tag{10.145}$$

where we have used $\tilde{q}(s) \sim \sqrt{\pi} B_{\rm I}/\sqrt{s}$ when $s \to 0$ (10.76, 10.78). In the limit $s \to 0$, the discrete sums over l and m can be replaced by integrals and one finds that the right hand side in Eq. (10.145) behaves like $1/s^2$ when $s \to 0$ with a prefactor which we can compute to obtain the large n behavior of $\langle l_{\max,n} \rangle$ as

$$\langle l_{\rm max,n} \rangle \sim C_{\rm I}\, n \ , \ C_{\rm I} = \int\limits_0^\infty dy \frac{y^{-1/2} e^{-y}}{\sqrt{\pi} + \frac{1}{2} \int_y^\infty dx\, x^{-3/2} e^{-x}} = C_0 \ , \tag{10.146}$$

where C_0 is given above (10.142) and where the last equality is simply obtained by performing an integration by part in the integral over x in the denominator. In Fig. 10.13, we have plotted the results of our numerical estimate of $\langle l_{\max,n} \rangle$ (obtained by averaging over 10^4 different realizations of random walks) for $\mu = 0.5$ and two different values of $c = \pm 1.0$.

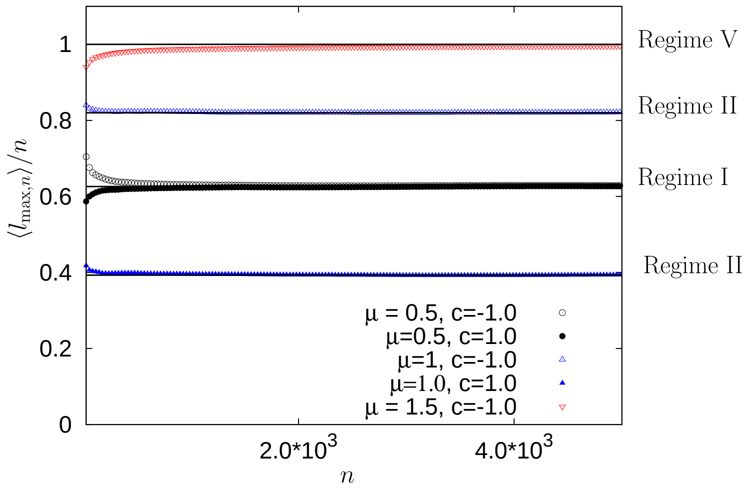

Figure 10.13: Plot of $\langle l_{\max,n}\rangle/n$ in the different regimes I, II and V: the points are the results of our numerical simulations. For regime II ($\mu = 1$), we present two curves, one with a positive drift ($c = 1$) (the second curve from top) and one with a negative drift ($c = -1$) (the bottom curve). These data indicate that in all these cases $\langle l_{\max,n}\rangle \propto n$, for large n, with an amplitude which agree quite well with our analytical results, which are represented in solid line for each of these cases and corresponds to the formula given in Eq. (10.146, 10.148, 10.159).

This plot shows that $\langle l_{\max,n}\rangle/n$ saturates rather quickly to the constant C_0, independently of c, in agreement with Eq. (10.146).

Thus in this regime the large n behavior of $\langle l_{\max,n}\rangle$ is unaffected by the presence of the drift c. This result could have been anticipated as $l_{\max,n}$ can be considered as the longest excursion between two consecutive zeros of a renewal process with a persistence exponent 1/2. This quantity was studied in Ref. [32] and its average was computed, yielding the large n behavior obtained in Eq. (10.146).

10.6.1.2 Regime II: $\mu = 1$ and c arbitrary:

In this regime, we recall that the persistence probability $Q(m)$ behaves algebraically for large m with an exponent $\theta(c)$ which depends continuously on c

$$Q(m) \sim \frac{B_{\rm II}}{m^{\theta(c)}} \;, \quad \theta(c) = \frac{1}{2} + \frac{1}{\pi}\arctan(c) \;, \tag{10.147}$$

where the amplitude $B_{\rm II}$ is given in Eq. (10.85). Here again we can use the result obtained in Ref. [32] for the longest excursion between consecutive zeros of a renewal process with a persistence exponent $\theta(c)$ to obtain

$$\langle l_{\max,n}\rangle \sim C_{\rm II}\, n \;, \quad C_{\rm II} = \int_0^\infty dy \frac{1}{1 + y^{\theta(c)} e^y \int_0^y dx\, x^{-\theta(c)} e^{-x}} \;, \tag{10.148}$$

which depends continuously on c and is independent of the non-universal amplitude $B_{\rm II}$ (10.147). In Fig. 10.14 we show a comparison of $C_{\rm II}$ obtained numerically (the squares symbols) and from our exact formula (solid line), which shows a very good agreement between both.

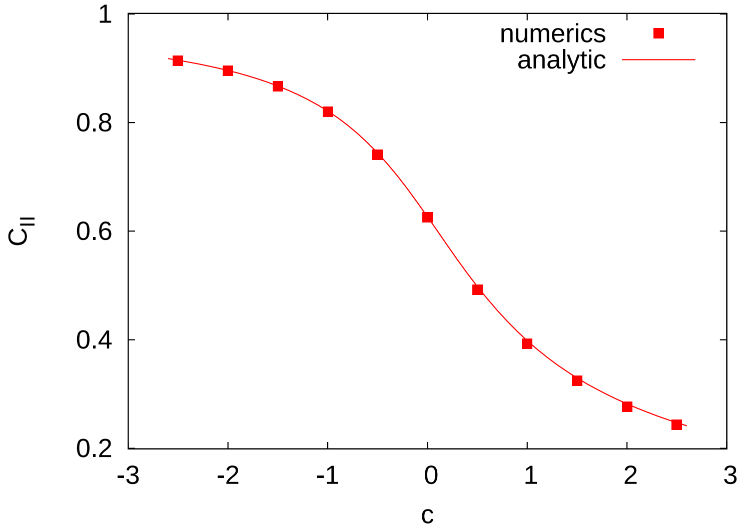

Figure 10.14: Plot of $C_{\rm II}$ as a function of c. The red squares correspond to numerical data while the solid line corresponds to our analytical result in Eq. (10.148) together with Eq. (10.147).

10.6.1.3 Regime III: $1 < \mu < 2$ and $c > 0$:

In this regime the persistence probability $Q(m)$ behaves for large m as

$$Q(m) \sim \frac{B_{\rm III}}{m^{\mu}} \; , \tag{10.149}$$

where the amplitude $B_{\rm III}$ is given in Eq. (10.90). Using again the results obtained in Ref. [32] one obtains that

$$\langle l_{\max,n} \rangle \sim C_{\rm III} \, n^{1/\mu} \; , \tag{10.150}$$

where, however, the amplitude $C_{\rm III}$ was not given in Ref. [32]. A careful analysis of the above formula (10.140) allows to obtain the amplitude $C_{\rm III}$ as

$$C_{\rm III} = \frac{1}{c}\Gamma(1-1/\mu)\left[\frac{1}{\pi}\sin\left(\frac{\mu\pi}{2}\right)\Gamma(\mu)\right]^{1/\mu} \; , \tag{10.151}$$

which diverges as $C_{\rm III} \sim (\pi(\mu-1))^{-1}$ when $\mu \to 1$ and vanishes as $C_{\rm III} \sim \sqrt{\pi(2-\mu)/2}$ when $\mu \to 2$. In Fig. 10.15 we show a plot of our numerical data for $\langle l_{\max,n} \rangle$ (averaged again over 10^4 different realizations) for different values of $\mu = 1.4, 1.5, 1.7, 1.9$ and for a fixed value of the drift $c = 5.0$. The solid lines indicate the corresponding exact asymptotic behaviors in Eq. (10.150, 10.151): the agreement between both is quite good although the convergence to the asymptotic behavior gets slower as μ decreases to 1.

10.6.1.4 Regime IV: $\mu = 2$ and $c > 0$:

In this case the persistence $Q(m)$ behaves quite differently as it vanishes exponentially for large m as

$$Q(n) \sim \frac{B_{\rm IV}}{n^{3/2}} \, e^{-s_1 n} \quad \text{where} \quad s_1 = \frac{c^2}{2\sigma^2} \; , \tag{10.152}$$

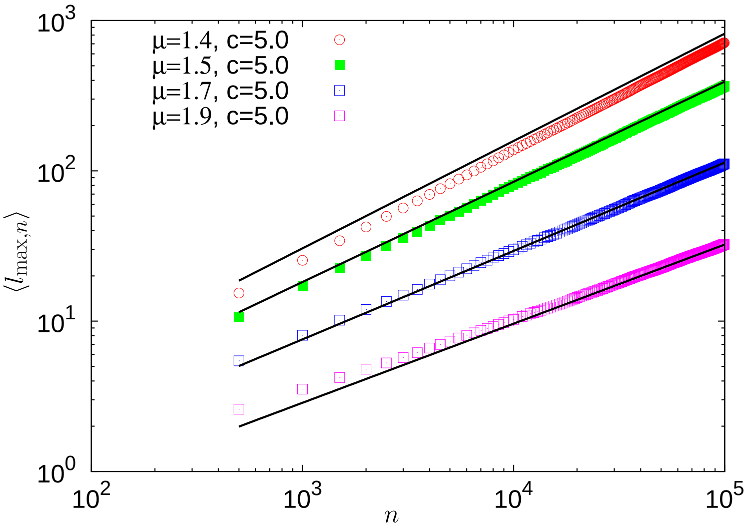

Figure 10.15: Plot, in a log-log scale, of $\langle l_{\rm max,n} \rangle$ as a function of n in regime III: the different curves correspond to different values of $\mu = 1.4, 1.5, 1.7, 1.9$ with a fixed value of $c = 5.0$. The solid line are the exact results given in Eqs (10.150, 10.151), without any fitting parameter.

where the amplitude $B_{\rm IV}$ is given in Eq. (10.96). This case was not analyzed in Ref. [32]. From Eq. (10.140) one has in this case

$$\begin{aligned} \sum_{n=0}^{\infty} \langle l_{\max,n} \rangle e^{-sn} &\sim \frac{1}{s} \sum_{m=1}^{\infty} \frac{\sum_{l=m}^{\infty} F(l)}{s\tilde{q}(0) + \sum_{l=m}^{\infty} F(l)} \\ &= \frac{1}{s} \sum_{m=1}^{\infty} \frac{Q(m)}{s\tilde{q}(0) + Q(m)} \; . \end{aligned} \tag{10.153}$$

Therefore in the limit when $s \to 0$ one can estimate the leading behavior of the sum over m as

$$\sum_{n=0}^{\infty} \langle l_{\max,n} \rangle e^{-sn} \sim \frac{m^*}{s} \; , \tag{10.154}$$

where m^* is such that

$$Q(m^*) \sim s\tilde{q}(0) \; . \tag{10.155}$$

From the asymptotic behavior above (10.152) one finds that $m^* \sim -\frac{\sigma^2}{2c^2} \log s$ so that finally

$$\langle l_{\max,n} \rangle \sim C_{\rm IV} \log n \; , \; C_{\rm IV} = \frac{2\sigma^2}{c^2} \; , \tag{10.156}$$

which is in sharp contrast with the algebraic growth obtained above in Eq. (10.150) for $1 < \mu < 2$ and $c > 0$. In Fig. 10.16 we show a plot of $\langle l_{\max,n} \rangle$ as a function of $\log n$: the straight line suggests indeed a logarithmic growth, in agreement with our analytic result (10.156). However, a more precise comparison with this exact asymptotic result, as shown in the inset of Fig. 10.16, suggests that the leading corrections are proportional to $\log\log n$, and hence quite strong.

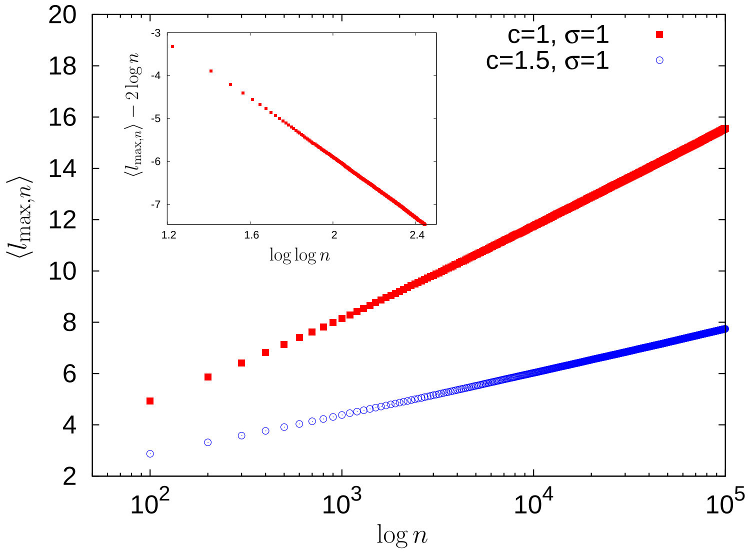

Figure 10.16: Plot of $\langle l_{\max,n}\rangle$ as a function of $\log n$ in the regime IV: here $\mu = 2$ and the two curves correspond to $c = 1$ and $c = 1.5$ ($\sigma = 1$ in both cases). The two curves suggest a logarithmic growth, as expected from Eq. (10.156). **Inset:** Plot of $\langle l_{\max,n}\rangle - 2\log n$ where $2\log n$ is the exact asymptotic result from Eq. (10.156) and $2\sigma^2/c^2 = 2$. This plot suggests rather strong corrections $\propto \log\log n$ to the leading logarithmic growth of $\langle l_{\max,n}\rangle$.

10.6.1.5 Regime V: $1 < \mu \leq 2$ and $c < 0$:

In this case the persistence probability $Q(m)$ tends asymptotically to a constant (10.101):

$$\begin{aligned} Q(m) \xrightarrow[m\to\infty]{} \alpha_\mu(c) &= \exp[-W_{|c|,\mu}(0)] \\ &= \exp\left[-\sum_{n=1}^{\infty}\frac{1}{n}\int_{|c|n}^{\infty} P_n(x)\,dx\right] . \end{aligned} \tag{10.157}$$

In addition from (10.86) one has that $Q(m) - \alpha_\mu(c) \propto n^{\mu-1}$ so that $F(m) \propto m^{-\mu}$ for large m. Therefore, the terms entering into the sum in Eq. (10.140) are given, to leading order when $1 - z = e^{-s} \to 0$ and large m (which are terms which give the leading contribution to this sum over m)

$$\frac{\sum_{l=m}^{\infty} F(l)z^l + (1-z)\sum_{l=m}^{\infty} Q(l)z^l}{(1-z)\tilde{Q}(z) + \sum_{l=m}^{\infty} F(l)z^l} \sim \frac{\alpha_\mu(c)}{\tilde{q}(0)} e^{-sm} = e^{-sm} . \tag{10.158}$$

Therefore this yields

$$\langle l_{\max,n}\rangle \sim C_{\rm V}\, n \; , \; C_{\rm V} = 1 \; . \tag{10.159}$$

This result, which is corroborated by our numerical simulations (see Fig. 10.13), can be physically understood as in this regime where $c < 0$ and $\mu > 1$ the number of records is finite and these records typically occur during the first steps of the random walks, where the walker might stay positive for a short while before it escapes to negative values when $n \to \infty$, and no record happens any more.

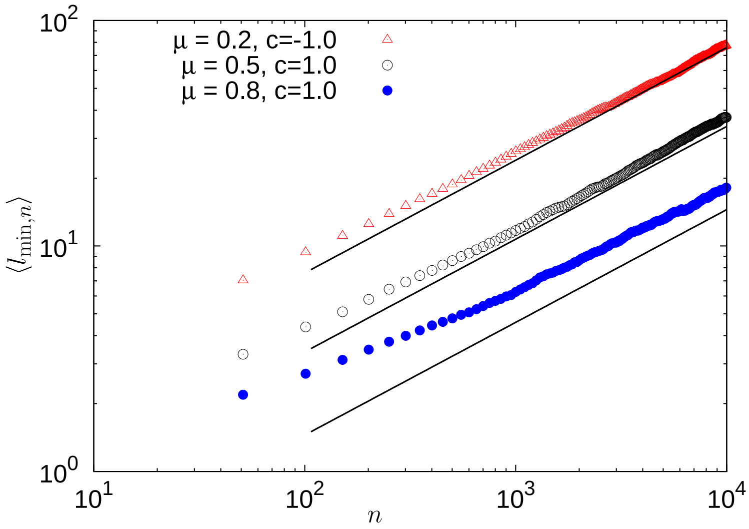

Figure 10.17: Plot, on log-log scale, of $\langle l_{\min,n}\rangle$ as a function of n, for different values of $\mu < 1$ and c (regime I). The points are the results of numerical simulations while solid lines correspond to our exact analytic result given in Eq. (10.165). These data indicate that in this regime $\langle l_{\min,n}\rangle \propto \sqrt{n}$, although the corrections to the exact asymptotic behavior are clearly visible, in particular for $\mu = 0.8$, $c = 1.0$.

10.6.2 Age of shortest lasting record $l_{\min,n}$

We now consider the shortest lasting record whose age $l_{\min,n}$ is given by (see Fig. 10.2)

$$l_{\min,n} = \min(l_1, l_2, \cdots, l_R)\ . \tag{10.160}$$

Note that, given that the final incomplete interval l_R is taken into consideration above, $l_{\min,n}$ can be zero: this happens when a record has been broken at the last step, such that $l_R = 0$.

The cumulative distribution $\mathcal{G}_n(m) = \text{Proba.}\,(l_{\min,n} \geq m)$ was studied in Ref. [22] and an explicit formula was obtained for its generating function:

$$\sum_{n=0}^{\infty} \mathcal{G}_n(m) z^n = \frac{\sum_{l=m}^{\infty} Q(l) z^l}{1 - \sum_{l=m}^{\infty} F(l) z^l}\ , \tag{10.161}$$

from which one gets the generating function of the average value $\langle l_{\rm min,n}\rangle$ as

$$\sum_{n=0}^{\infty} z^n \langle l_{\min,n}\rangle = \sum_{m=1}^{\infty} \frac{\sum_{l=m}^{\infty} Q(l) z^l}{1 - \sum_{l=m}^{\infty} F(l) z^l}\ . \tag{10.162}$$

In the absence of drift, $c = 0$, it was shown in Ref. [22] that

$$\langle l_{\min,n}\rangle \sim D\sqrt{n}\ ,\ D = \frac{1}{\sqrt{\pi}}\ . \tag{10.163}$$

As for $\langle l_{\max,n}\rangle$ we will see that the behavior of $\langle l_{\min,n}\rangle$, in the presence of non zero drift $c \neq 0$, is quite different in the five different regimes discussed above. Again we start by

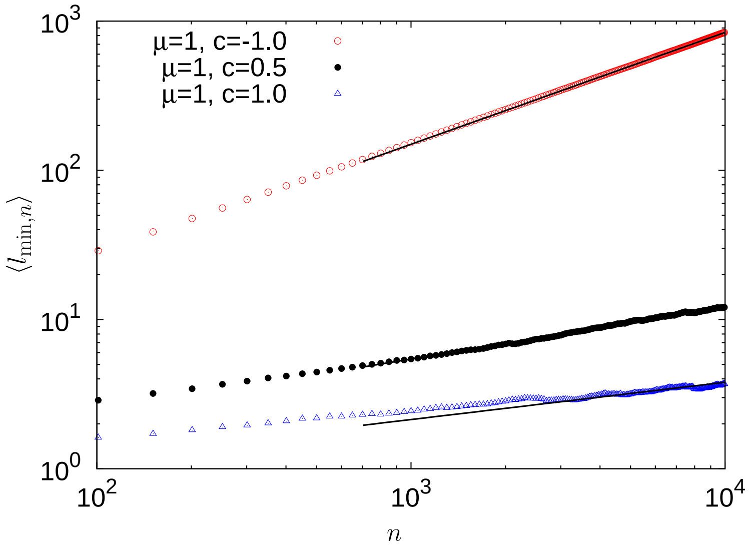

Figure 10.18: Plot, on a log-log scale, of $\langle l_{\min,n}\rangle$ as a function of n for $\mu = 1$ and different values of $c = -1, 0.5$ and $c = 1$. The solid line corresponds to the algebraic growth $n^{1-\theta(c)}$, from Eq. (10.167).

giving a brief summary of our results for $\langle l_{\min,n}\rangle$:

$$\begin{aligned}
\langle l_{\min,n}\rangle \quad &\sim \quad \sqrt{n} \quad \text{for} \quad 0 < \mu < 1 \text{ and } c \text{ arbitrary} \quad (\text{regime I}) \;, \\
&\sim \quad n^{1-\theta(c)} \quad \text{for} \quad \mu = 1 \text{ and } c \text{ arbitrary} \quad (\text{regime II}) \;, \\
&\sim \quad \text{const.} \quad \text{for} \quad 1 < \mu < 2 \text{ and } c > 0 \quad (\text{regime III}) \;, \\
&\sim \quad \text{const.} \quad \text{for} \quad \mu = 2 \text{ and } c > 0 \quad (\text{regime IV}) \;, \\
&\sim \quad n \quad \text{for} \quad 1 < \mu \leq 2 \text{ and } c < 0 \quad (\text{regime V}) \;,
\end{aligned} \tag{10.164}$$

again with $\theta(c)$ as defined in Eq. (10.16). In the following we discuss the behavior of $\langle l_{\min,n}\rangle$ in more detail for each of the five regimes.

10.6.2.1 Regime I: $0 < \mu < 1$ and c arbitrary

In this case the persistence probability decays algebraically as given in Eq. (10.144) and the analysis of $\langle l_{\min,n}\rangle$ can be obtained by noticing that, in the limit $z \to 1$, the denominator in Eq. (10.162) can be simply replaced by 1 while the remaining sums over l (in the numerator) and over m can be replaced by integrals. This yields straightforwardly

$$\langle l_{\min,n}\rangle \sim D_{\mathrm{I}}\sqrt{n} \;, \tag{10.165}$$

$$D_{\mathrm{I}} = B_{\mathrm{I}} = \frac{1}{\sqrt{\pi}} \exp\left[-\frac{1}{\pi}\int_0^\infty \frac{dk}{k} \arctan\left(\frac{\hat{f}(k)\sin(kc)}{1-\hat{f}(k)\cos(kc)}\right)\right], \tag{10.166}$$

where the expression of B_{I} is given in Eq. (10.78). In Fig. 10.17, we show the results of our numerical simulations which are in a rather good agreement with Eq. (10.165), although the corrections to this exact asymptotic behavior are clearly visible, in particular for $\mu = 0.8$, $c = 1.0$. In Fig. 10.18, we show a plot of the numerical computation of $\langle l_{\min,n}\rangle$ for $\mu = 1$ and different values of $c = -1, 0.5$ and $c = 1$: these data are in good agreement with the power law growth in Eq. (10.165), although we have not attempted to estimate numerically the prefactor D_{I}.

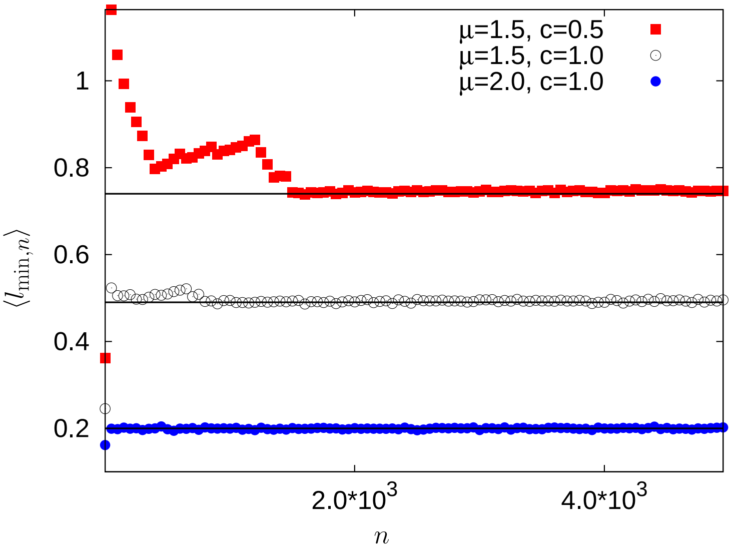

Figure 10.19: Plot of $\langle l_{\min,n} \rangle$ as a function of n for $\mu = 1.5$ and $\mu = 2$ and different values of $c > 0$, therefore corresponding to regime III and IV. The solid line corresponds to the exact result, from Eq. (10.170, 10.171).

10.6.2.2 Regime III: $1 \leq \mu < 2$ and $c > 0$

10.6.2.3 Regime II: $\mu = 1$ and c arbitrary

In this regime where the persistence probability $Q(m)$ decays algebraically as in Eq. (10.84), $\langle l_{\min,n} \rangle$ can be analyzed as in the regime I where in the limit $z \to 1$, the denominator in Eq. (10.162) can be simply replaced by 1 while the remaining sums over l (in the numerator) and over m can be replaced by integrals. This yields straightforwardly:

$$\sum_{n=1}^{\infty} e^{-sn} \langle l_{\min,n} \rangle \sim \frac{B_{\rm II}}{s^{2-\theta(c)}} \int_0^{\infty} dy \int_y^{\infty} dx \, x^{-\theta(c)} e^{-y} = \frac{B_{\rm II}}{s^{2-\theta(c)}} \Gamma[2-\theta(c)] \; , \tag{10.167}$$

which yields

$$\langle l_{\min,n} \rangle \sim D_{\rm II} \, n^{1-\theta(c)} \; , \; D_{\rm II} = B_{\rm II} \; , \tag{10.168}$$

where $B_{\rm II}$ is given in Eq. (10.85) and $\theta(c) = 1/2 + \frac{1}{\pi} \arctan(c)$.

In this case we write the above formula (10.162) as

$$\begin{aligned} \sum_{n=0}^{\infty} z^n \langle l_{\min,n} \rangle &= \frac{1}{1-z} \left(1 - \frac{1}{\tilde{q}(0)} \right) \\ &+ \sum_{m=2}^{\infty} \frac{\sum_{l=m}^{\infty} Q(l) z^l}{1 - \sum_{l=m}^{\infty} F(l) z^l} \; , \end{aligned} \tag{10.169}$$

where we have simply isolated the term $m = 1$ and used $1 - \tilde{F}(0) = (1-z)\tilde{Q}(0)$ (10.58). Now the above sum (10.169), which starts with $m = 2$, is dominated by the large values of m. Because of the algebraic decay of $Q(m) \sim m^{-\mu}$ in this case (10.149) and $\mu > 1$ in this regime one gets that this second term behaves like $(1-z)^{\mu-2}$, which is then subleading, compared to the first term which behaves like $(1-z)^{-1}$. Therefore one gets in this case

$$\langle l_{\min,n} \rangle \sim D_{\rm III} \; , \; D_{\rm III} = 1 - \frac{1}{\tilde{q}(0)} = 1 - \exp\left[- \sum_{n=1}^{\infty} \frac{1}{n} \int_{cn}^{\infty} P_n(x) \, dx \right] \; , \tag{10.170}$$

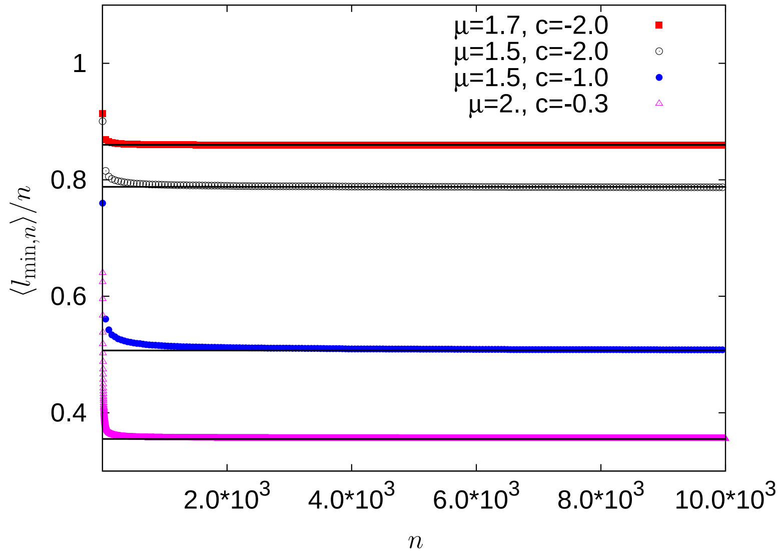

Figure 10.20: Plot of $\langle l_{\min,n}\rangle/n$ as a function of n for different values of $1 < \mu \leq 2$ and different values of $c < 0$, corresponding to regime V.

where we have used the expression for $1/\tilde{q}(0)$ given in Eq. (10.116). In Fig. 10.19 we show a plot of the numerical computation $\langle l_{\min,n}\rangle$ for $\mu = 1.5$ and different values of $c = 0.5$ and $c = 1$, which is in very good agreement with Eq. (10.170). Note that we have extracted the value of $1/\tilde{q}(0)$ which enters into the expression of D_{III} from the linear growth of the mean record number $\langle R_n\rangle$, according to (10.116).

10.6.2.4 Regime IV: $\mu = 2$ and $c > 0$

A similar analysis can be carried out in this case, starting from the same formula (10.169). In this case, in the above sum (10.169), which starts with $m = 2$, one can safely put $z = 1$, because of the behavior of the exponential decay of $Q(m)$ in this case (10.152). Therefore one gets immediately

$$\langle l_{\min,n}\rangle \sim D_{\text{IV}} \ , \ D_{\text{IV}} = 1 - \frac{1}{\tilde{q}(0)} = 1 - \exp\left[-\sum_{n=1}^{\infty}\frac{1}{n}\int_{cn}^{\infty} P_n(x)\,dx\right] \ , \tag{10.171}$$

where we have used the expression for $1/\tilde{q}(0)$ given in Eq. (10.126). In Fig. 10.19 we show a plot of the numerical computation $\langle l_{\min,n}\rangle$ for $\mu = 2$ and $c = 1$, which is good agreement with Eq. (10.171). Note that we have extracted the value of $1/\tilde{q}(0)$ which enters into the expression of D_{IV} from the linear growth of the mean record number $\langle R_n\rangle$, according to Eq. (10.126).

10.6.2.5 Regime V: $1 < \mu \leq 2$ and $c < 0$

In this regime where the persistence goes to a constant $Q(m) \to \alpha_\mu(c)$, for $m \gg 1$, one can simply replace $Q(l)$ by this constant value in the sum of the numerator in Eq. (10.162) while the denominator can be simply approximated by 1 in the limit $1 - z = e^{-s} \to 0$. This yields straightforwardly

$$\langle l_{\min,n}\rangle \sim \alpha_\mu(c)\, n \ . \tag{10.172}$$

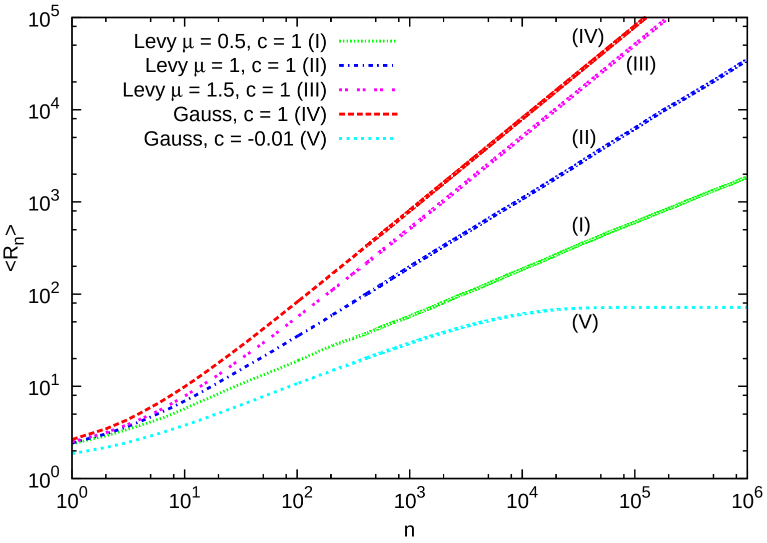

Figure 10.21: The figure shows numerical results for the mean record number $\langle R_n \rangle$ for biased random walks from all five regimes. For regimes I to IV we used a positive bias of $c = 1$, in regime V we simulated a Gaussian random walk (with $\sigma = 1$) with a negative bias of $c = -0.01$. For each jump distribution we averaged over 10^4 samples. In all these cases, as shown in detail in the previous figures, the asymptotic behavior agree very well with our analytical predictions (which are not shown on this figure for clarity).

In Fig. 10.20 we show a plot of $\langle l_{\min,n} \rangle / n$ which we have computed numerically for different values of $\mu = 1.7, 1.5$ and $\mu = 2$ and also for different values of the drift. These results are in very good agreement with our exact asymptotic result in Eq. (10.172), where the value of $\alpha_\mu(c)$ have been extracted from the mean record number $\langle R_n \rangle \sim 1/\alpha_\mu(c)$ (10.49). This result (10.172) can be easily understood by realizing that $l_{\min,n} = n$ if the whole trajectory is on the negative side, which happens with probability $\alpha_\mu(c)$ while $l_{\min,n}$ is of order $\mathcal{O}(1)$ if the walker makes an excursion on the positive side. One also notices that, in this case, $l_{\text{typ}} \sim \langle l_{\min,n} \rangle$.

10.7 Conclusion

In this paper we considered a very simple model of a one dimensional discrete-time random walk in presence of a constant drift c. At each time step the particle jumps by a random distance $c + \eta$ where the noise η is drawn from a continuous and symmetric jump distribution $f(\eta)$, characterized by a Lévy index $0 < \mu \leq 2$. The jump has a finite second moment for $\mu = 2$, while for $0 < \mu < 2$ the second moment diverges. For this discrete-time series consisting of the successive positions of the biased walker, we presented complete analytical studies of the persistence and the record statistics. For the later, we studied the mean and the full distribution of the number of records up to step n and also the statistics of the duration of records, in particular those for the longest and shortest lasting records. As a function of the two parameters c and $0 < \mu \leq 2$, we found that it is necessary to distinguish between five different universal regimes, as shown in the basic phase diagram in Fig. 10.1. In these 5 regimes, the persistence and the record statistics exhibit very different asymptotic behaviors that are summarized in Section 2 and we do not repeat them here. For instance, the growth of the mean record number with n in all five regimes is summarized in the simulation results in Fig. 10.21, in complete agreement with our analytical predictions. The main conclusion is that even though this is a rather simple model, it exhibits very rich and varied universal behaviors for record statistics and persistence depending on the two parameters c and $0 < \mu \leq 2$.

Our results provide a simple yet nontrivial, but fully solvable model for the record statistics, a subject which has gained considerable interest over the last few years. Our results provide one generalization of the previous results for record statistics for symmetric random walks [22]. However, it is important to note that this extension does not yet cover all possible kinds of discrete-time random walks. In principle one could consider more complicated asymmetries of the jump distribution. It might be interesting to consider a jump distribution that has different tail-exponents in the left and in the right tail. Also a generalization of these results to an asymmetric lattice random walk is still missing. In [22] a symmetric lattice random was also considered. It should be possible to compute the record statistics of a lattice random walk that has a higher probability to jump in one direction than in the other.

It might be interesting to see if our results can be applied to financial data, similar to the analysis in [11, 12]. Daily stock data however proved not to be useful for comparison because the asymptotic limit is hardly achieved in the available observational data. An application to stock data with a higher temporal resolution however should be possible and might provide new insights. Such an analysis is definitely an interesting subject for future research. Also the distribution of records in stock prices has not been analysed in detail before and it would be interesting to see if such an analysis for available data can be fitted to our theoretical distributions.

Acknowledgements

SNM and GS acknowledge support by ANR grant 2011-BS04-013-01 WALKMAT and in part by the Indo-French Centre for the Promotion of Advanced Research under Project 4604-3. GW is grateful for the kind hospitality of the Laboratoire de Physique Théorique et Modèles Statistiques during the completion of this work and for the financial support provided by DFG within the Bonn Cologne Graduate School of Physics and Astronomy.

APPENDIX I - The constant A_{I}

The constant A_{I} in Eq. (10.104) can be directly expressed in terms of $\hat{f}(k)$ as announced in Eq. (10.24). To derive this, we use the explicit expression of $P_n(x)$ from Eq. (10.6) in the expression for A_{I} and integrate over x to get

$$A_{\mathrm{I}} = \frac{2}{\sqrt{\pi}} \exp\left[\sum_{n=1}^{\infty} \frac{1}{n} \int_{-\infty}^{\infty} \frac{dk}{2\pi} \, [\tilde{f}(k)]^n \frac{1 - e^{-i\,k\,c\,n}}{ik} \right]. \tag{10.173}$$

Next we use the symmetry $\hat{f}(k) = \hat{f}(-k)$ which leads to

$$A_{\mathrm{I}} = \frac{2}{\sqrt{\pi}} \exp\left[\frac{1}{\pi} \int_{0}^{\infty} \frac{dk}{k} \sum_{n=1}^{\infty} \frac{\sin(kcn)}{n} \, [\tilde{f}(k)]^n \right]. \tag{10.174}$$

The sum on the rhs can be explicitly evaluated using the identity

$$\sum_{n=1}^{\infty} \frac{x^n}{n} \sin(an) = \arctan\left[\frac{x \sin(a)}{1 - x\cos(a)} \right] \tag{10.175}$$

which then leads to the exact expression in Eq. (10.24).

We then analyze the behavior of A_{I} when $|c|$ is large and in the case where $\hat{f}(k) = \exp(-|k|^{\mu})$, with $\mu < 1$. In that case one has $P_n(x) = n^{-1/\mu}\mathcal{L}_{\mu}(x/n^{1/\mu})$ for all n and it is easier to start from the formula given in the text in Eq. (10.104)

$$A_{\rm I} = \frac{2}{\sqrt{\pi}} e^{S_0} \; , \; S_0 \equiv S_0(c) = \sum_{n=1}^{\infty} \frac{1}{n} \int_0^{cn} \mathcal{L}_\mu(x/n^{1^{1/\mu}}) dx/n^{1/\mu} \; . \tag{10.176}$$

Note that, given that $P_n(x) = P_n(-x)$ one has $S_0(c) = S_0(-c)$ and we thus present the analysis for $c > 0$. Performing the change of variable $y = x/n^{1/\mu}$ in the integral above (10.176) we write

$$S_0(c) = \sum_{n=1}^{\infty} \frac{1}{n} \int_0^{cn^{\frac{\mu-1}{\mu}}} \mathcal{L}_\mu(y)\, dy \; , \tag{10.177}$$

and take the derivative with respect to c

$$S_0'(c) = \sum_{n=1}^{\infty} n^{-\frac{1}{\mu}} \mathcal{L}_\mu \left(\frac{c}{n^{\frac{1-\mu}{\mu}}} \right) \; . \tag{10.178}$$

In this expression, one notices that $c/n^{\frac{1-\mu}{\mu}} = (n/c^{\frac{\mu}{1-\mu}})^{\frac{\mu-1}{\mu}}$ so that when $c \to \infty$ the discrete sum over n in Eq. (10.178) can be replaced by an integral (we recall that $\mu < 1$ here), which leads to

$$S_0'(c) \sim \frac{1}{c} \int_0^{\infty} \mathcal{L}_\mu \left(y^{\frac{\mu-1}{\mu}} \right) y^{-1/\mu}\, dy \; . \tag{10.179}$$

Finally, performing the change of variable $z = y^{\frac{\mu-1}{\mu}}$ in Eq. (10.179) yields

$$S_0'(c) \sim \frac{1}{c} \frac{\mu}{1-\mu} \int_0^{\infty} \mathcal{L}_\mu(z) dz = \frac{1}{c} \frac{\mu}{2(1-\mu)} \; , \tag{10.180}$$

so that one gets

$$A_{\rm I} = \frac{2}{\sqrt{\pi}} e^{S_0} \propto c^{\frac{\mu}{2(1-\mu)}} \; , \; c \to \infty \; . \tag{10.181}$$

This power law behavior (10.181) can be understood from the following scaling argument. We are indeed interested in the records statistics of the variables y_n, with $y_n = x_n + cn$ (10.11) where x_n behaves for large n as $x_n = \mathcal{O}(n^{1/\mu})$. Therefore for small n, $n < n^*$ when c is large, y_n is dominated by the drift term and n^* is such that $cn^* \sim {n^*}^{1/\mu}$, which yields

$$n^* \sim c^{\frac{\mu}{1-\mu}} \; . \tag{10.182}$$

On the other hand, for small n, $n < n^*$, y_n is dominated by the (positive) drift and hence is almost deterministic which yields $\langle R_n \rangle \sim n$, for $n < n^*$ while $\langle R_n \rangle \sim A_{\rm I} \sqrt{n}$ for $n > n^*$. By matching these two behaviors for $n = n^*$ one obtains

$$A_{\rm I} \sim \sqrt{n^*} \propto c^{\frac{\mu}{2(1-\mu)}} \; , \tag{10.183}$$

which yields the result obtained above (10.179).

Note finally that, by using $S_0(c) = -S_0(-c)$ one obtains

$$A_{\rm I} \sim (-c)^{\frac{-\mu}{2(1-\mu)}} \; , \; c \to -\infty \; . \tag{10.184}$$

APPENDIX II - Computation of $\alpha_2(c) = a_2(|c|)$, $c < 0$ for exponential jump distribution with $c < 0$

The expression for the amplitude $\alpha_2(c)$ in regime V (with $c < 0$) and for a general jump distribution is given in Eq. (10.101). By comparing with Eq. (10.126) we see that $\alpha_2(c < 0) = a_2(|c|)$ where $a_2(|c|)$ is the prefactor of the leading linear growth of mean record number in regime IV with drift positive $|c|$. For a general jump distribution $f(\eta)$, we then have

$$\alpha_2(c) = \exp\left[-\sum_{n=1}^{\infty} \frac{1}{n} \int_{|c|n}^{\infty} P_n(x)\, dx\right] , \tag{10.185}$$

where we recall that

$$P_n(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \left[\hat{f}(k)\right]^n e^{-i\,k\,x} \tag{10.186}$$

and

$$\hat{f}(k) = \int_{-\infty}^{\infty} f(\eta)\, e^{ik\eta}\, d\eta \tag{10.187}$$

is the Fourier transform of the jump distribution. Thus, in general, computing the prefactor $\alpha_2(c) = a_2(|c|)$ explicitly is difficult for arbitrary $f(\eta)$. It can be done explicitly for Gaussian distribution where

$$P_n(x) = (2\pi n\sigma^2)^{-1/2} \exp[-x^2/2n\sigma^2] \tag{10.188}$$

itself is Gaussian and $\alpha_2(c) = a_2(|c|)$ is then given by the formula in Eq. (10.127). In this appendix, we show that $\alpha_2(c) = a_2(|c|)$ can also be computed explicitly for the symmetric exponential distribution $f(\eta) = (2\,b)^{-1} \exp(-|x|/b)$.

For this exponential jump distribution, the Fourier transform has the Lorentz-ian form, $\hat{f}(k) = 1/[\pi(b^2\,k^2 + 1)]$. One can then substitute this in the expression for $P_n(x)$ and eventually in Eq. (10.185). After a quite convoluted computation involving contour integration in the complex plane, one can find $\alpha_2(c)$ explicitly. However, as we show below, for the exponential case, there is an alternative simpler way to compute $\alpha_2(c)$ directly (without going through the formula in Eq. (10.185).

The first observation is that $\alpha_2(c)$ is just the limiting value of the persistence probability $Q(n)$ (the probability that the walker stays *below* 0 up to n steps starting at 0) when $n \to \infty$ in presence of a negative drift $c < 0$. By symmetry, $Q(n)$ is then also the probability that the walker, starting at the origin, stays *above* the origin up to n steps, but in presence of a positive drift $|c| > 0$. So, the idea is to compute this probability $Q(n)$ directly for the exponential jump distribution and then take the limit $n \to \infty$ to compute $\alpha_2(c) = Q(n \to \infty)$.

To compute $Q(n)$, we first define

$$q_n^+(y) = \text{Proba. that the walker, starting at } y \geq 0 \text{ stays pos. up to step } n\,. \tag{10.189}$$

If we can compute $q_n^+(y)$, then $Q(n)$ is simply obtained by putting the starting position to be 0, i.e., $Q(n) = q_n^+(0)$. To compute $q_n^+(y)$, we can write a backward recurrence relation for $q_n^+(y)$ by considering the jump that happens at the first step from y to $y' \geq 0$

$$q_n^+(y) = \int_0^{\infty} q_{n-1}^+(y') f(y + |c| - y')\, dy' \ , \tag{10.190}$$

$$q_0^+(y) = 1 \text{ for } y \geq 0\,. \tag{10.191}$$

In the limit of large n, we expect that $q_n^+(y)$ approaches to an n independent stationary value, $q_n^+(y) \to q^+(y)$, that just denotes the eventual probability with which the walker escapes to infinity (starting from y) in presence of a positive drift $|c|$. Taking $n \to \infty$ limit on both sides of Eq. (10.190) gives the integral equation for $y \geq 0$

$$q^{+}(y) = \int_{0}^{\infty} q^{+}(y') f(y + |c| - y')\, dy' \,. \tag{10.192}$$

Note that this equation is valid for arbitrary jump distribution $f(\eta)$. However, this half-space Wiener-Hopf type integral equation with asymmetric kernel can not be solved in general. However, for the special case of the exponential distribution, $f(\eta) = \frac{1}{2b}\exp(-|\eta|/b)$, this integral equation (10.192) can be transformed into a differential equation using

$$f''(\eta) = -\frac{1}{b^2}\delta(\eta) + \frac{1}{b^2} f(\eta) \,. \tag{10.193}$$

By differentiating twice Eq. (10.192) with respect to y one then obtains [using Eq. (10.193)]

$$\frac{d^2 q^{+}(y)}{dy^2} = -\frac{1}{b^2} q^{+}(y + |c|) + \frac{1}{b^2} q^{+}(y) \,. \tag{10.194}$$

Note that the solution $q^{+}(y)$ must approach to 1 as $y \to \infty$: $q^{+}(y \to \infty) = 1$. This follows from the fact that if the particle starts at the positive infinity, it escapes to positive infinity with probability 1 in presence of any positive drift $|c| > 0$.

Note that the differential equation (10.194), though linear, is actually *nonlocal* in y due to the first term on the rhs and hence is still not completely trivial to solve. Fortunately, it turns out that it admits a solution of the form

$$q^{+}(y) = 1 - F \exp\left(-\lambda y / b\right) \,, \tag{10.195}$$

where F and λ are two dimensionless constants (independent of y) that are yet to be determined. Note that this ansatz manifestly satisfies the boundary condition $q^{+}(y \to \infty) = 1$. Substituting this ansatz in Eq. (10.194) we see that indeed Eq. (10.195) is a solution provided λ satisfies the equation

$$\exp\left(-\lambda\, |c|/b\right) = 1 - \lambda^2 \,; \text{ with } \lambda > 0 \,. \tag{10.196}$$

The transcendental equation has a unique positive solution which then determines λ uniquely. For example, for $b/c = 1$, we get using Mathematica the root $\lambda = 0.714556\ldots$. But we still need to determine the prefactor F in the ansatz in Eq. (10.195). The amplitude F in Eq. (10.195) is obtained by injecting this solution back into the integral equation (10.194) and performing the integral. Indeed, one finds that Eq. (10.195) is a solution of the integral equation provided

$$F = 1 - \lambda \,. \tag{10.197}$$

This then uniquely determines the solution of the integral equation (10.194)

$$q^{+}(y) = 1 - (1 - \lambda) \exp\left(-\lambda y / b\right) \tag{10.198}$$

where λ is the unique positive solution of the transcendental equation (10.196).

Noting finally that $\alpha_2(c) = Q(n \to \infty) = q^{+}(0)$ gives

$$\alpha_2(c) = a_2(|c|) = q^{+}(0) = \lambda \,, \tag{10.199}$$

where $\lambda > 0$ is the solution of Eq. (10.196). We have checked that we indeed get exactly the same expression by evaluating the original general expression in Eq. (10.185) for the exponential jump distribution, though this was not completely trivial to check (we do not give details of this check here).

Bibliography

[1] D. Gembris, J. G. Taylor, and D. Suter, Nature **417**, 506 (2002).
[2] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).
[3] S. Redner and M. R. Petersen, Phys. Rev. E **74**, 061114 (2006).
[4] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).
[5] G. Wergen and J. Krug, EPL **92**, 30008 (2010).
[6] A. Anderson and A. Kostinski, J. Appl. Meteo. and Climat. **50**, 1859 (2011).
[7] J. Krug and K. Jain, Physica A **358**, 1 (2005).
[8] L. P. Oliveira *et al.*, Phys. Rev. B **71**, 104526 (2005).
[9] P. Sibani, G. F. Rodriguez, and G. G. Kenning, Phys. Rev. B **74**, 224407 (2006).
[10] M. Bauer, C. Godreche, and J. Luck, J. Stat. Phys. **96**, 963 (1999).
[11] G. Wergen, M. Bogner, and J. Krug, Phys. Rev. B **83**, 051109 (2011).
[12] G. Wergen, S. N. Majumdar, and G. Schehr, Phys. Rev. E **86**, 011119 (2012).
[13] F. G. Foster and A. Stuart, J. Roy. Stat. Soc. **16**, 1 (1954).
[14] B. C. Arnold, N. Balakrishnan, and H. N. Nagraja, *Records*, 1st ed. (Wiley-Interscience, 1998).
[15] V. B. Nevzorov, *Records: Mathematical Theory* (American Mathematical Society, 2004).
[16] R. Ballerini and S. Resnick, J. Appl. Probab. **22**, 487 (1985).
[17] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).
[18] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011).
[19] J. Franke, G. Wergen, and J. Krug, Phys. Rev. Lett. **108**, 064101 (2012).
[20] J. Krug, J. Stat. Mech.: Theor. Exp. **07**, 07001 (2007).
[21] G. H. Weiss, *Aspects and applications of the random walk* (North-Holland, 1994).
[22] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).
[23] E. Sparre Andersen, Math. Scand. **20**, 195 (1954).
[24] S. N. Majumdar, Physica A **389**, 4299 (2010).
[25] S. Sabhapandit, EPL **94**, 20003 (2011).
[26] P. Le Doussal and K. J. Wiese, Phys. Rev. E **79**, 051105 (2009).
[27] Y. Edery, A. Kostinski, and B. Berkowitz, Geophys. Res. Lett. **389**, L16403 (2011).
[28] S. N. Majumdar, Curr. Sci. **77**, 370 (1999).
[29] J.-P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990).
[30] R. Metzler and J. Klafter, Phys. Rep. **339**, 1 (2000).
[31] J. Pitman and M. Yor, Ann. Probab. **25**, 855 (1997).
[32] C. Godreche, S. N. Majumdar, and G. Schehr, Phys. Rev. Lett. **102**, 240602 (2009).
[33] G. Schehr and P. Le Doussal, J. Stat. Mech.: Theor. Exp. **P01009** (2010).
[34] M. R. Evans, S. N. Majumdar, and R. K. P. Zia, J. Stat. Phys. **123**, 357 (2006).

Chapter 11

Record statistics for multiple random walks

Gregor Wergen[1], **Satya N. Majumdar**[2] **and Grégory Schehr**[2]

[1]*Institute for Theoretical Physics, University of Cologne*
[2]*Laboratoire de Physique Théorique et Modèles Statistiques, Université Paris Sud 11 and CNRS, Orsay, France*

Published in: Phys. Rev. E 86, 011119 (2012)

Abstract: We study the statistics of the number of records $R_{n,N}$ for N identical and independent symmetric discrete-time random walks of n steps in one dimension, all starting at the origin at step 0. At each time step, each walker jumps by a random length drawn independently from a symmetric and continuous distribution. We consider two cases: (I) when the variance σ^2 of the jump distribution is finite and (II) when σ^2 is divergent as in the case of Lévy flights with index $0 < \mu < 2$. In both cases we find that the mean record number $\langle R_{n,N} \rangle$ grows universally as $\sim \alpha_N \sqrt{n}$ for large n, but with a very different behavior of the amplitude α_N for $N > 1$ in the two cases. We find that for large N, $\alpha_N \approx 2\sqrt{\log N}$ independently of σ^2 in case I. In contrast, in case II, the amplitude approaches to an N-independent constant for large N, $\alpha_N \approx 4/\sqrt{\pi}$, independently of $0 < \mu < 2$. For finite σ^2 we argue, and this is confirmed by our numerical simulations, that the full distribution of $(R_{n,N}/\sqrt{n} - 2\sqrt{\log N})\sqrt{\log N}$ converges to a Gumbel law as $n \to \infty$ and $N \to \infty$. In case II, our numerical simulations indicate that the distribution of $R_{n,N}/\sqrt{n}$ converges, for $n \to \infty$ and $N \to \infty$, to a universal nontrivial distribution, independently of μ. We discuss the applications of our results to the study of the record statistics of 366 daily stock prices from the Standard & Poors 500 index.

11.1 Introduction

A record is an entry in a series of events that exceeds all previous entries. In recent years there has been a surge of interest in the statistics of record-breaking events, both from the theoretical point of view as well as in multiple applications. The occurrence of record-breaking events has been studied for instance in sports [1, 2], in evolution models in biology [3, 4], in the theory of spin-glasses [5, 6] and in models of growing networks[7]. Recently there has been some progress in understanding the phenomenon of global warming via studying the occurrence of record-breaking temperatures [8–11].

More precisely, let us consider a sequence or a discrete-time series of random variables $\{x(0), x(1), x(2), \dots, x(n)\}$ with $n+1$ entries. This sequence may represent for example the daily maximum temperature in a city or the daily maximum price of a stock. A record is said to happen at step m if the m-th member of the sequence is bigger than all previous members, i.e., if $x(m) > x(i)$ for all $i = 0, 1, 2, \dots, (m-1)$. Let R_n denote the number of records in this sequence of $n+1$ entries. Clearly R_n is a random variable whose statistics depends on the joint distribution of $P(x(0), x(1), \dots, x(n))$ of the members of the sequence. When the members of the sequence are independent and identically distributed (i.i.d) random variables each drawn from a distribution $p(x)$, i.e., the joint distribution factorizes, $P(x(0), x(1), \dots, x(n)) = \prod_{i=1}^{n+1} p(x(i))$, the record statistics is well understood from classical theories [12–14]. In particular, when $p(x)$ is a continuous distribution, it is known that the distribution of record number $P(R_n, n)$ is universal for all n, i.e., independent of the parent distribution $p(x)$. The average number of records up to step n, $\langle R_n \rangle = \sum_{m=1}^{n+1} 1/m$ for all n and the universal distribution, for large n, converges to a Gaussian distribution with mean $\approx \ln(n)$ and variance $\approx \ln n$.

While the statistical properties of records for i.i.d random variables (RV's) are thus well understood for many years, numerous questions remain open for more realistic systems with time-dependent or correlated RV's. In principle there are many different ways to generalize the simple i.i.d. RV scenario described above. For instance, one can consider time series of RV's that are independent, but not identically distributed. One example for this case is the so called Linear Drift Model with RV's from probability distributions with identical shape, but with a mean value that increases in time. This model was first proposed in the 1980's [15] and was recently thoroughly analyzed in Refs. [16–18]. In 2007 Krug also considered the case of uncorrelated RV's from distributions with increasing variance [4].

Another possible generalization is the one where RV's are correlated. Perhaps, the simplest and the most natural model of correlated RV's is an n-step one dimensional discrete-time random walk with entries

$$\{x(0) = 0, x(1), x(2), \dots, x(n)\} \tag{11.1}$$

where the position $x(m)$ of the walker at discrete time m evolves via the Markov jump process

$$x(m) = x(m-1) + \eta(m) \ , \tag{11.2}$$

with $x(0) = 0$ and $\eta(m)$ represents the random jump at step m. The noise variables $\eta(m)$'s are assumed to be i.i.d variables, each drawn from a symmetric distribution $f(\eta)$. For instance, it may include Lévy flights where $f(\eta) \sim |\eta|^{-1-\mu}$ for large η with the Lévy index $0 < \mu < 2$ which has a divergent second moment. Even though this model represents a very simple Markov chain, statistical properties of certain observables associated with such a walk may be quite nontrivial to compute, depending on which observable one is studying [19–21]. For instance, in recent years there has been a lot of interest in the extremal properties of such random walks. These include the statistics of the maximal displacement of the walk up to n steps with several applications [21–26] and the order statistics, i.e., the statistics of the ordered maxima [27, 28] as well as the universal distribution of gaps between successive ordered maxima of a random walk [28].

The statistics of the number of record-breaking events in the discrete-time random walk process in Eq. (11.2) has also been studied in a number of recent works with several interesting results [29–33]. In 2008, Majumdar and Ziff computed exactly the full distribution $P(R_n, n)$ of the record number up to n steps and found that when the jump distribution $f(\eta)$ is continuous and symmetric, the record number distribution $P(R_n, n)$ is completely universal for all n, i.e., independent of the details of the jump distribution [29]. In particular, for instance, the Lévy flight with index $0 < \mu < 2$ (thus with a divergent second moment of the jump distribution $f(\eta)$) has the same record number distribution as for a Gaussian walk (with a finite second moment of $f(\eta)$). This is a rather amazing result and the deep reason for this universality is rooted [29] in the so called Sparre Andersen theorem [34]. In particular, for large n, $P(R_n, n) \sim n^{-1/2} G(R_n/\sqrt{n})$ where the scaling function $G(x) = e^{-x^2/4}/\sqrt{\pi}$ is universal [29]. The mean number of records $\langle R_n \rangle \approx \sqrt{4n/\pi}$ for large n [29]. In contrast, this universal result does not hold for symmetric but discontinuous $f(\eta)$. For example, if $f(\eta) = \frac{1}{2}\delta(\eta - 1) + \frac{1}{2}\delta(\eta + 1)$, then x_m represents the position of a random walker at step m on a 1-d lattice with lattice spacing 1. In this case, the mean number of records still grows as $\sqrt{n}$ for large n but with a smaller prefactor, $\langle R_n \rangle \approx \sqrt{2n/\pi}$ [29].

These results were later generalized to several interesting cases, for instance, to the record statistics of one dimensional random walk in presence of an external drift [30, 32] and one dimensional continuous-time random walk with a waiting-time distribution between successive jumps [31]. The record statistics of the distance traveled by a random walker in higher dimensions with and without drift has been studied numerically in the context of contamination spread in porous medium [33]. In [32], it was also found that the record statistics of stock markets is very similar to the ones of biased random walks.

While in Refs. [29–33] the record statistics of a single discrete-time random walker was studied, the purpose of this article is to generalize these results to the case where one has N independent one dimensional discrete-time random walks. In this N-walker process, a record happens at an instant when the maximum position of all the walkers at that instant exceeds all its previous values. We will see that despite the fact that the walkers are independent, the record statistics is rather rich, universal and nontrivial even in this relatively simple model.

Let us first summarize our main results. We derive asymptotic results for the mean of the record number $\langle R_{n,N} \rangle$ up to a time n and also discuss its full distribution. It turns out that for $N > 1$, while the full universality with respect to the jump distribution found for $N = 1$ case is no longer valid, there still remains a vestige of universality of a different sort. In our analysis, it is important to distinguish two cases: case (I) where the jump distribution $f(\eta)$ has a finite variance $\sigma^2 = \int_{-\infty}^{\infty} \eta^2 f(\eta)\, d\eta$ and case (II) where σ^2 is divergent as in the case of Lévy flights with Lévy index $0 < \mu < 2$. In both cases, we find that the mean record number $\langle R_{n,N} \rangle$ grows *universally* as $\sim \alpha_N \sqrt{n}$ for large n. However, the N dependence of the prefactor α_N, in particular for large N, turns out to be rather different in the two cases

$$\alpha_N \xrightarrow[N\to\infty]{} \begin{cases} 2\sqrt{\log N} & \text{in Case I} \quad (\text{independent of}\, \sigma^2) \\ \\ 4/\sqrt{\pi} & \text{in Case II} \quad (\text{independent of}\, \mu) \end{cases} \tag{11.3}$$

In addition, we also study the distribution of the record number $R_{n,N}$. For finite σ^2 we argue and confirm numerically that the distribution of the random variable $(R_{n,N}/\sqrt{n} - 2\sqrt{\log N})\sqrt{\log N}$ converges to the Gumbel law asymptotically for large n and N (see section II for details). In contrast, in case II, we find numerically that the distribution of $R_{n,N}/\sqrt{n}$ converges, for large n and N, to a nontrivial distribution independent of the value of $0 < \mu < 2$ (see section II for details). We were however unable to compute this asymptotic distribution analytically and it remains a challenging open problem. Finally, we discuss the applications of our results to the study of the record statistics of 366 daily stock prices from the Standard & Poors 500 index [35]. We analyze the evolution of the record number in subsets of N stocks that were randomly chosen from this index and compare the results to our analytical findings. While the strong correlations between the individual stocks seem to play an important effect in the record statistics, the dependence of the record number on N still seems to be the same as in the case of the N independent random walkers.

The rest of the paper is organized as follows. In section II, we define the N-walker model precisely and summarize the main results obtained in the paper. In section III, we present the analytical calculation of the mean number of records for multiple random walkers, in both cases where σ^2 is finite (case I) and σ^2 is infinite (case II). Section IV is devoted to an analytic study of the distribution of the record number in the case where σ^2 is finite. In section V we present a thorough numerical study of the record statistics of multiple random walks, and in section VI we discuss the application of our results to the record statistics of stock prices. Finally, we conclude in section VII and present the technical details of some of the analytical computations concerning the computation of the mean number of records and the distribution of the record number for lattice random walks in the three Appendices A, B and C..

11.2 Record Statistics for Multiple Random Walks: The model and the main results

Here we consider the statistics of records of N independent random walkers all starting at the origin 0. The position $x_i(m)$ of the i-th walker at discrete time step m evolves via the Markov evolution rule

$$x_i(m) = x_i(m-1) + \eta_i(m) \ , \tag{11.4}$$

where $x_i(0) = 0$ for all $i = 1, 2, \ldots, N$ and the noise $\eta_i(m)$'s are i.i.d variables (independent from step to step and from walker to walker), each drawn from a symmetric distribution $f(\eta)$. We are interested in the record statistics of the composite process. More precisely, consider at each step m, the maximum position of all N random walkers

$$x_{\max}(m) = \max\left[x_1(m), x_2(m), \ldots, x_N(m)\right]. \tag{11.5}$$

A record is said to happen at step m if this maximum position at step m is bigger than all previous maximum positions, i.e. if $x_{\max}(m) > x_{\max}(k)$ for all $k = 0, 1, \ldots, (m-1)$ (see Fig. 11.1). In other words, we are interested in the record statistics of the stochastic discrete-time series $\{x_{\max}(m)\}$, with the convention that the initial position $x_{\max}(0) = 0$ is counted as a record. Note that even though the position of each walker evolves via the simple independent Markovian rule in Eq. (11.4), the evolution of the maximum process $\{x_{\max}(m)\}$ is highly non-Markovian and hence is nontrivial.

Let $R_{n,N}$ denote the number of records up to step n for this composite N-walker process. Clearly $R_{n,N}$ is a random variable and we are interested in its statistics. For a single walker $N = 1$, we have already mentioned that the probability distribution of the record number $R_{n,1}$ is completely universal, i.e., independent of the jump distribution $f(\eta)$ as long as $f(\eta)$ is symmetric and continuous [29]. In particular, for example, the record number distribution is the same for simple Gaussian walkers as well for Lévy flights with index $0 < \mu < 2$. Here we are interested in the opposite limit when $N \to \infty$.

We find that while the complete universality of the record statistics is no longer true for $N > 1$, a different type of universal behavior emerges in the $N \to \infty$ limit. In this large N limit, there are two universal asymptotic behaviors of the record statistics depending on whether the second moment $\sigma^2 = \int_{-\infty}^{\infty} \eta^2 \, f(\eta) \, d\eta$ of the jump distribution is finite or divergent. For example, for Gaussian, exponential, uniform jump distributions σ^2 is finite. In contrast, for Lévy flights where $f(\eta) \sim |\eta|^{-\mu-1}$ for large η with the Lévy index $0 < \mu < 2$, the second moment σ^2 is divergent. In these two cases, we find the following behaviors for the record statistics.

Case I (σ^2 finite): In this case, we consider jump distributions $f(\eta)$ that are symmetric with a finite second moment $\sigma^2 = \int_{-\infty}^{\infty} \eta^2 \, f(\eta) \, d\eta$. In this case, the Fourier transform of

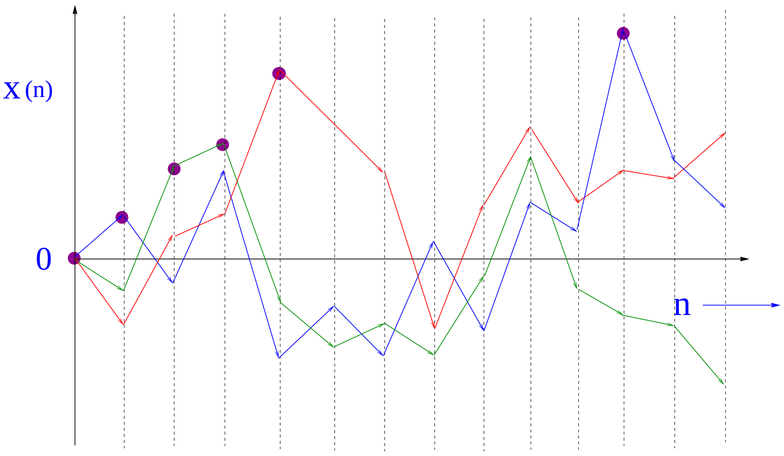

Figure 11.1: Schematic trajectories of $N = 3$ random walkers. Each walker starts at the origin and evolves via the Markov jump process in Eq. (11.4). A record happens at step m if the maximum position at step m $x_{\max}(m) > x_{\max}(k)$ for all $k = 0, 1, 2, \ldots (m-1)$. The record values are shown by filled circles.

the jump distribution $\hat{f}(k) = \int_{-\infty}^{\infty} f(\eta)\, e^{ik\eta}\, d\eta$ behaves, for small k, as

$$\hat{f}(k) \approx 1 - \frac{\sigma^2}{2}\, k^2 + \ldots \tag{11.6}$$

Examples include the Gaussian jump distribution, $f(\eta) = \sqrt{a/\pi}\, e^{-a\,\eta^2}$, exponential jump distribution $f(\eta) = (b/2)\, \exp[-b|\eta|]$, uniform jump distribution over $[-l, l]$ etc. For such jump distributions, we find that for large number of walkers N, the mean number of records grows asymptotically for large n as

$$\langle R_{n,N} \rangle \xrightarrow[N\to\infty]{n\to\infty} 2\sqrt{\ln N}\,\sqrt{n}\,. \tag{11.7}$$

Note that this asymptotic behavior is universal in the sense that it does not depend explicitly on σ as long as σ is finite.

Moreover, we argue that for large N and large n, the scaled random variable $R_{n,N}/\sqrt{n}$ converges, in distribution, to the Gumbel form, i.e,

$$\text{Prob.}\left[\frac{R_{n,N}}{\sqrt{n}} \leq x\right] \xrightarrow[N\to\infty]{n\to\infty} F_1\left[\left(x - 2\sqrt{\ln N}\right)\sqrt{\ln N}\right]\,, \tag{11.8}$$

where

$$F_1(z) = \exp\left[-\exp[-z]\right]. \tag{11.9}$$

Indeed, for large N and large n, the scaled variable $R_{n,N}/\sqrt{n}$ converges, in distribution, to the maximum of N independent random variables

$$\frac{R_{n,N}}{\sqrt{n}} \xrightarrow[N\to\infty]{n\to\infty} M_N \quad \text{where} \quad M_N = \max(y_1, y_2, \ldots, y_N) \tag{11.10}$$

where $y_i \geq 0$'s are i.i.d non-negative random variables each drawn from distribution $p(y) = \frac{1}{\sqrt{\pi}}\, e^{-y^2/4}$ for $y \geq 0$ and $p(y) = 0$ for $y < 0$.

Case II (σ^2 divergent): In this case we consider jump distributions $f(\eta)$ such that the second moment σ^2 is divergent. In this case, the Fourier transform $\hat{f}(k)$ of the noise distribution behaves, for all k, as

$$\hat{f}(k) = 1 - |a\, k|^{\mu} + \ldots \tag{11.11}$$

where $0 < \mu < 2$. Examples include Lévy flights where $f(\eta) \sim |\eta|^{-\mu-1}$ with the Lévy index $0 < \mu < 2$. For the noise distribution in Eq. (11.11), we find, quite amazingly, that in the large N and large n limit, the record statistics is (i) completely universal, i.e., independent of μ and a (ii) more surprisingly and unlike in Case-I, the record statistics also becomes independent of N as $N \to \infty$. For example, we prove that for large N, the mean number of records grows asymptotically with n as

$$\langle R_{n,N} \rangle \xrightarrow[N\to\infty]{n\to\infty} \frac{4}{\sqrt{\pi}}\sqrt{n} \; , \tag{11.12}$$

which is exactly twice that of one walker, i.e., $\langle R_{n,N\to\infty} \rangle = 2\,\langle R_{n,1} \rangle$ for large n. Similarly, we find that the scaled variable $R_{n,N}/\sqrt{n}$, for large n and large N, converges to a universal distribution

$$\text{Prob.}\left[\frac{R_{n,N}}{\sqrt{n}} \leq x\right] \xrightarrow[N\to\infty]{n\to\infty} F_2(x) \; , \tag{11.13}$$

which is independent of the Lévy index μ as well as of the scale a in Eq. (11.11). While we have computed this universal distribution $F_2(x)$ numerically rather accurately, we were not able to compute its analytical form.

11.3 Mean number of records for multiple walkers

Let $R_{n,N}$ be the number of records up to step n for N random walkers, i.e., for the maximum process $x_{\max}(n)$. Let us write

$$R_{m,N} = R_{m-1,N} + \xi_{m,N} \; , \tag{11.14}$$

where $\xi_{m,N}$ is a binary random variable taking values 0 or 1. The variable $\xi_{m,N} = 1$ if a record happens at step m and $\xi_{m,N} = 0$ otherwise. Clearly, the total number of records up to step n is

$$R_{n,N} = \sum_{m=1}^{n} \xi_{m,N} \; . \tag{11.15}$$

So, the mean number of records up to step n is

$$\langle R_{n,N} \rangle = \sum_{m=1}^{n} \langle \xi_{m,N} \rangle = \sum_{m=1}^{n} r_{m,N} \; , \tag{11.16}$$

where $r_{m,N} = \langle \xi_{m,N} \rangle$ is just the record rate, i.e., the probability that a record happens at step m. To compute the mean number of records, we will first evaluate the record rate $r_{m,N}$ and then sum over m.

To compute $r_{m,N}$ at step m, we need to sum the probabilities of all trajectories that lead to a record event at step m. Suppose that a record happens at step m with the record value x (see Fig. 11.2). This corresponds to the event that one of the N walkers (say the dashed trajectory in Fig. 11.2), starting at the origin at step 0, has reached the level x for the first time at step m, while the rest of the $N-1$ walkers, starting at the origin at step 0, have all stayed below the level x till the step m. Also, the walker that actually reaches x at step m can be any of the N walkers. Finally this event can take place at any level $x > 0$ and one needs to integrate over the record value x. Using the independence of N walkers and taking into account the event detailed above, one can then write

$$r_{m,N} = N \int_0^\infty p_m(x)\; \left[q_m(x)\right]^{N-1}\, dx \; , \tag{11.17}$$

where $q_m(x)$ denotes the probability that a single walker, starting at the origin, stays below the level x up to step m and $p_m(x)$ is the probability density that a single walker reaches the level x for the first time at step m, starting at the origin at step 0.

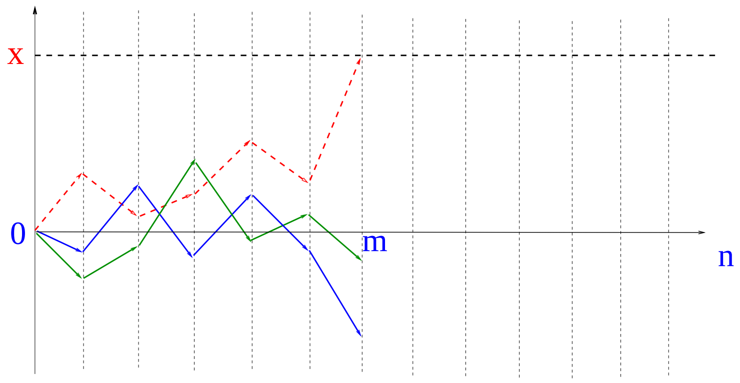

Figure 11.2: A record happens at step m with record value x for $N = 3$ walkers, all starting at the origin. This event corresponds to one walker (the dashed line) reaching the level x for the first time at step m while the other walkers stay below the level x up to step m.

The two quantities $p_m(x)$ and $q_m(x)$ can be reinterpreted in terms of slightly more familiar objects via the following observation. Note that by shifting the origin to the level x and using the time-reversal property of the trajectory of a single random walker, it is easy to see that $p_m(x)$ is just the probability density that a single walker, starting at the origin at step 0, reaches x at step m while staying positive at all intermediate steps. By a similar shift of the origin to level x and using the reflection symmetry of the trajectories around the origin, it is clear that $q_m(x)$ can be interpreted as the probability that a single walker, starting at an initial position $x > 0$ at step 0, stays positive (i.e., does not cross the origin) up to step m. This is then the familiar persistence or the survival probability of a single random walker [21]. In fact, both these quantities $p_m(x)$ and $q_m(x)$ can be regarded as special cases of the more general restricted Green's function in the following sense. Consider a single random walker starting at position x at step 0 and evolving its position via successive uncorrelated jumps as in Eq. (11.2). Let $G_+(y, x, m)$ denote the probability density that the walker reaches $y > 0$ at step m, starting at $x > 0$ at step 0, while staying positive at all intermediate steps. The subscript $+$ denotes that it is indeed the restricted Green's function counting only the trajectories that reaches y at step m without crossing the origin in between. It is then clear from our discussion above that

$$p_m(x) = G_+(x, 0, m) \tag{11.18}$$

$$q_m(x) = \int_0^\infty G_+(y, x, m)\, dy\,. \tag{11.19}$$

In the second line, the survival probability $q_m(x)$ is obtained from the restricted Green's function by integrating over all possible final positions of the walker. Note also, from Eqs. (11.18) and (11.19), that the survival probability starting exactly at the origin is

$$q_m(0) = \int_0^\infty p_m(x)\, dx. \tag{11.20}$$

Hence, if we know the restricted Green's function $G_+(y, x, m)$, we can in principle compute the two required quantities $p_m(x)$ and $q_m(x)$. Using the Markov evolution rule in Eq. (11.2), it is easy to see that the restricted Green's function $G_+(y, x, m)$ satisfies an integral equation in the semi-infinite domain [21]

$$G_+(y,x,m) = \int_0^\infty G_+(y',x,m-1)\, f(y-y')\, dy' \; , \tag{11.21}$$

starting from the initial condition, $G_+(y,x,0) = \delta(y-x)$. Such integral equations over the semi-infinite domain are called Wiener-Hopf equations and are notoriously difficult to solve for arbitrary kernel $f(z)$. Fortunately, for the case when $f(z)$ represents a continuous and symmetric probability density as in our case, one can obtain a closed form solution for the following generating function (rather its Laplace transform) [36]

$$\int_0^\infty dy\, e^{-\lambda\, y} \int_0^\infty dx\, e^{-\lambda_0\, x} \left[\sum_{m=0}^\infty G_+(y,x,m)\, s^m\right] = \tilde{G}(\lambda,\lambda_0,s) = \frac{\phi(s,\lambda)\,\phi(s,\lambda_0)}{\lambda+\lambda_0} \; , \tag{11.22}$$

where

$$\phi(s,\lambda) = \exp\left[-\frac{\lambda}{\pi}\int_0^\infty \frac{\ln[1-s\hat{f}(k)]}{\lambda^2+k^2}\, dk\right] \quad \text{and} \quad \hat{f}(k) = \int_{-\infty}^\infty f(x)\, e^{i\,k\,x}\, dx \; . \tag{11.23}$$

While the formula in Eq. (11.22) is explicit, it is rather cumbersome and one needs further work to extract the asymptotic behavior of $p_m(x)$ and $q_m(x)$ from this general expression. To make progress, one can first make a change of variable on the left hand side (lhs) $\lambda_0 x = z$ and then take the $\lambda_0 \to \infty$ limit. Using $\phi(s,\lambda_0\to\infty) = 1$ and the definition $G_+(y,0,m) = p_m(y)$, and replacing y by x we then obtain the following relation

$$\sum_{m=0}^\infty s^m \int_0^\infty p_m(x)\, e^{-\lambda x}\, dx = \phi(s,\lambda) \tag{11.24}$$

where $\phi(s,\lambda)$ is given in Eq. (11.23). Similarly, putting $\lambda = 0$ on the lhs of Eq. (11.22), using the definition $q_m(x) = \int_0^\infty G_+(y,x,m)\, dy$ and replacing λ_0 by λ, it is easy to see that

$$\sum_{m=0}^\infty s^m \int_0^\infty q_m(x)\, e^{-\lambda x}\, dx = \frac{1}{\lambda\sqrt{1-s}}\, \phi(s,\lambda) \, . \tag{11.25}$$

The formula in Eq. (11.25) is known in the literature as the celebrated Pollaczek-Spitzer formula [37, 38] and has been used in a number of works to derive exact results on the maximum of a random jump process [24, 39–41]. Interestingly, this formula has also been useful to compute the asymptotic behavior of the flux of particles to a spherical trap in three dimensions [25, 42, 43].

Let us also remark that by making a change of variable $\lambda x = y$ on the lhs of Eq. (11.25) and taking $\lambda \to \infty$, one obtains the rather amazing universal result for all m

$$\sum_{m=0}^\infty q_m(0)\, s^m = \frac{1}{\sqrt{1-s}} \Longrightarrow q_m(0) = \binom{2m}{m}\frac{1}{2^{2m}} \; , \tag{11.26}$$

which is known as the Sparre Andersen theorem [34]. In particular, for large m, $q_m(0) \approx 1/\sqrt{\pi m}$. note that for the case of a single walker $N = 1$, it follows from Eq. (11.17) that the record rate at step m is simply given by

$$r_{m,1} = \int_0^\infty p_m(x)\, dx = q_m(0) = \binom{2m}{m}\frac{1}{2^{2m}} \xrightarrow{m\to\infty} \frac{1}{\sqrt{\pi m}} \; , \tag{11.27}$$

where we have used Eq. (11.20) and the Sparre Andersen theorem (11.26). Thus, one obtains the rather surprising universal result for the $N = 1$ case: for all continuous and symmetric jump distributions, the mean number of records up to step n, $\langle R_{n,N} \rangle = \sum_{m=1}^{n} r_{m,N}$ is universal for all n and grows as $\sqrt{4n/\pi}$ for large n [29]. The universality in this case can thus be traced back to Sparre Andersen theorem.

In contrast, for $N > 1$, we need the full functions $p_m(x)$ and $q_m(x)$ to compute the record rate in Eq. (11.17). This is hard to compute explicitly for all m. However, one can make progress in computing the asymptotic behavior of the record rate $r_{m,N}$ for large m and large N, as we show below. In turns out that for large m, the integral in Eq. (11.17) is dominated by the asymptotic scaling behavior of the two functions $p_m(x)$ and $q_m(x)$ for large m and large x. To extract the scaling behavior of $p_m(x)$ and $q_m(x)$, our starting point would be the two equations (11.24) and (11.25). The next step is to use these asymptotic expressions in the main formula in Eq. (11.17) to determine the record rate $r_{m,N}$ at step m for large m and large N. The procedure to extract the asymptotics is somewhat subtle and algebraically cumbersome. To facilitate an easy reading of the paper, we relegate this algebraic procedure in the appendices. Here we just use the main results from these appendices and proceed to derive the results announced in Eqs. (11.7) and (11.12). The asymptotic behavior of $p_m(x)$ and $q_m(x)$ depend on whether $\sigma^2 = \int_{-\infty}^{\infty} \eta^2 \, f(\eta) \, d\eta$ is finite or divergent. This gives rise to the two cases mentioned in Section II.

Case I(σ^2 finite): In this case, we show in Appendix A that in the scaling limit $x \to \infty$, $m \to \infty$ but keeping the ration $x/\sqrt{m}$ fixed, $p_m(x)$ and $q_m(x)$ approach the following scaling behavior

$$p_m(x) \quad \to \quad \frac{1}{\sqrt{2\sigma^2 \, m}} \, g_1\left(\frac{x}{\sqrt{2\,\sigma^2\, m}}\right) \, , \quad \text{where} \quad g_1(z) = \frac{2}{\sqrt{\pi}} \, z \, e^{-z^2} \, , \tag{11.28}$$

$$q_m(x) \quad \to \quad h_1\left(\frac{x}{\sqrt{2\,\sigma^2\, m}}\right) \, , \quad \text{where} \quad h_1(z) = \text{erf}(z) \, , \tag{11.29}$$

where $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} \, du$. Note that $dh_1(z)/dz = g_1(z)/z$.

Case II (σ^2 divergent): For the case when the Fourier transform of the jump distribution $\hat{f}(k)$ has the small k behavior as in Eq. (11.11), we show in Appendix B that in the scaling limit when $x \to \infty$, $m \to \infty$, but keeping the ratio $x/m^{1/\mu}$ fixed,

$$p_m(x) \quad \to \quad \frac{1}{m^{1/2+1/\mu}} \, g_2\left(\frac{x}{m^{1/\mu}}\right) \tag{11.30}$$

$$q_m(x) \quad \to \quad h_2\left(\frac{x}{m^{1/\mu}}\right) . \tag{11.31}$$

While it is hard to obtain explicitly the full scaling functions $g_2(z)$ and $h_2(z)$ for all z, one can compute the large z asymptotic behavior and obtain

$$g_2(z) \underset{z\to\infty}{\sim} \frac{A_\mu}{z^{1+\mu}} \, , \tag{11.32}$$

$$h_2(z) \underset{z\to\infty}{\sim} 1 - \frac{B_\mu}{z^\mu} \tag{11.33}$$

where the two amplitudes are

$$A_\mu \quad = \quad \frac{2\mu}{\sqrt{\pi}} \, \beta_\mu \tag{11.34}$$

$$B_\mu \quad = \quad \beta_\mu \tag{11.35}$$

with the constant β_μ having different expressions for $0 < \mu < 1$ and $1 \leq \mu < 2$

$$\beta_\mu = \frac{a^\mu}{\pi\Gamma(1-\mu)} \int_0^\infty \frac{u^\mu}{1+u^2}\, du \quad \text{for} \quad 0 < \mu < 1 \tag{11.36}$$

$$\beta_\mu = \frac{2a^\mu}{\pi\Gamma(2-\mu)} \int_0^\infty \frac{u^\mu}{(1+u^2)^2}\, du \quad \text{for} \quad 1 \leq \mu < 2\,. \tag{11.37}$$

The expressions above (11.36, 11.37) can be written in a unified way for any $0 < \mu < 2$ as

$$\beta_\mu = \frac{a^\mu}{2\Gamma(1-\mu)\cos(\frac{\mu\pi}{2})} = \frac{a^\mu \Gamma(\mu)\sin\left(\frac{\mu\pi}{2}\right)}{\pi} \;, \tag{11.38}$$

where, in the last equality, we have used $\Gamma(1-\mu)\Gamma(\mu) = \dfrac{\pi}{\sin \mu\pi}$. We recall that here we are considering discrete time random walks (11.2). In the continuous time random walk framework, with an exponential waiting time between jumps, the quantity $g_2(z)$ was studied in Ref. [44]. By performing an asymptotic analysis similar to the one presented in Appendix B, the authors showed that $g_2(z)$ behaves, for large z, like in Eq. (11.32) with the same exponent albeit with a different amplitude. On the other hand, the exact asymptotic result (11.32), together with Eq. (11.38) can also be used to study the normalized pdf $\tilde{p}_m(x)$ of the position after m steps, with the condition that the walker stays positive at all intermediate steps, which was recently studied in Ref. [41]. It reads

$$\tilde{p}_m(x) = \frac{p_m(x)}{\int_0^\infty p_m(x)dx} \to \frac{1}{m^{1/\mu}} \tilde{g}_2\left(\frac{x}{m^{1\mu}}\right) \;,\; \tilde{g}_2(z) = \sqrt{\pi} g_2(z) \;, \tag{11.39}$$

where we have used the Sparre-Andersen theorem $\int_0^\infty p_m(x)\,dx = q_m(0) \sim 1/\sqrt{\pi m}$ for large m. From Eq. (11.32), one obtains the large z behavior of $\tilde{g}_2(z)$ as

$$\tilde{g}_2(z) \underset{z\to\infty}{\sim} \frac{\tilde{A}_\mu}{z^{1+\mu}} \;,\; \tilde{A}_\mu = \frac{2a^\mu \sin\left(\frac{\mu\pi}{2}\right)\Gamma(\mu+1)}{\pi} \;, \tag{11.40}$$

where we have used $\mu\Gamma(\mu) = \Gamma(\mu+1)$. On the other hand, if one considers the probability density function $P_m(x)$ of the position of a free Lévy random walk after m steps, with a jump distribution as in Eq. (11.11) after m steps, it assumes the scaling form, valid for large m, $P_m(x) \sim m^{-1/\mu} p(x/m^{1/\mu})$ where the asymptotic behavior is given by

$$p(z) \underset{z\to\infty}{\sim} \frac{C_\mu}{z^{1+\mu}} \;,\; C_\mu = \frac{a^\mu \sin\left(\frac{\mu\pi}{2}\right)\Gamma(\mu+1)}{\pi} \;. \tag{11.41}$$

Therefore the above result (11.40) establishes that $\tilde{A}_\mu = 2C_\mu$: this result was recently obtained analytically in perturbation theory for μ close to 2, $2-\mu \ll 1$, and conjectured to hold for any μ, on the basis of thorough numerical simulations [41]. Here this result is established exactly for any $\mu \in (0,2)$. While the large z behavior of $g_2(z)$ is the most relevant one for our study, we mention, for completeness, that its small z behavior was also studied in Ref. [44, 45], yielding $g_2(z) \sim z^{\mu/2}$. Finally we remark that the asymptotic behavior of $h_2(z)$ for large z has been computed in great detail recently in Ref. [40], only the first two leading terms are presented in Eq. (11.33) here.

We are now ready to use these asymptotic behavior of $p_m(x)$ and $q_m(x)$ in Eq. (11.17) to deduce the large m behavior of the record rate. Noting that for large m, the integral is dominated by the scaling regime, we substitute in Eq. (11.17) the scaling forms of $p_m(x)$ and $q_m(x)$ found in Eqs. (11.28), (11.29), (11.30) and (11.31). We then get, for large m,

$$r_{m,N} \approx \frac{N}{\sqrt{m}} \int_0^\infty g(z)\,[h(z)]^{N-1}\,dz \;, \tag{11.42}$$

where $g(z) = g_{1,2}(z)$ and $h(z) = h_{1,2}(z)$ depending on the two cases. So, we notice that in all cases the record rate decreases as $m^{-1/2}$ for large m, albeit with different N-dependent prefactors in the two cases. Hence, the mean number of records $\langle R_{n,N}\rangle$ up to step n grows, for large n, as

$$\langle R_{n,N}\rangle \approx \alpha_N \sqrt{n}\,, \quad \text{where} \quad \alpha_N = 2N \int_0^\infty g(z)\,[h(z)]^{N-1}\,dz\,. \tag{11.43}$$

Next we estimate the constant α_N for large N. We first note that α_N in Eq. (11.43) can be expressed as

$$\alpha_N = 2\int_0^\infty \frac{g(z)}{h'(z)}\,\frac{d}{dz}\{[h(z)]^N\}\,dz\,, \tag{11.44}$$

where $h'(z) = dh/dz$. Noticing that $h(z)$ is an increasing function of z approaching 1 as $z \to \infty$, the dominant contribution to the integral for large N comes from the large z regime. Hence, we need to estimate how the ratio $g(z)/h'(z)$ behaves for large z. Let us consider the two cases separately.

Case I (σ^2 finite): In this case, we have explicit expressions for $g_1(z)$ and $h_1(z)$ respectively in Eqs. (11.28) and (11.29). Hence we get

$$\begin{aligned} \alpha_N &= 2\int_0^\infty dz\, z\,\frac{d}{dz}[\mathrm{erf}(z)]^N && (11.45)\\ &= \int_0^\infty dy\, y\,\frac{d}{dy}[\mathrm{erf}(y/2)]^N. && (11.46)\end{aligned}$$

The rhs of Eq. (11.46) has a nice interpretation. Consider N i.i.d positive random variables $\{y_1, y_2, \dots, y_N\}$, each drawn from the distribution: $p(y) = \frac{1}{\sqrt{\pi}}\, e^{-y^2/4}$ for $y \geq 0$ and $p(y) = 0$ for $y < 0$. Let M_N denote their maximum. Then the cdf of the maximum is given by

$$\mathrm{Prob}[M_N \leq y] = \left[\int_0^y p(y')\,dy'\right]^N = [\mathrm{erf}(y/2]^N\,. \tag{11.47}$$

The probability density of the maximum is then given by: $\frac{d}{dy}[\mathrm{erf}(y/2]^N$. Hence, the rhs of Eq. (11.46) is just the average value $\langle M_N\rangle$ of the maximum. This gives us an identity for all N

$$\alpha_N = \langle M_N\rangle\,. \tag{11.48}$$

From the standard extreme value analysis of i.i.d variables [46], it is easy to show that to leading order for large N, $\langle M_N\rangle \approx 2\sqrt{\ln N}$ which then gives, via Eq. (11.43), the leading asymptotic behavior of the mean record number

$$\langle R_{n,N}\rangle \xrightarrow[N\to\infty]{n\to\infty} 2\sqrt{\ln N}\,\sqrt{n}\,. \tag{11.49}$$

Case II (σ^2 divergent): To evaluate α_N in Eq. (11.44), we note that when σ^2 is divergent, unlike in Case I, we do not have the full explicit form of the scaling functions $g_2(z)$ and $h_2(z)$. Hence evaluation of α_N for all N is difficult. However, we can make progress for large N. As mentioned before, for large N, the dominant contribution to the integral in Eq. (11.44) comes from large z. For large z, using the asymptotic expressions in Eqs. (11.32) and (11.33), we get

$$\frac{g_2(z)}{h_2'(z)} \xrightarrow{z\to\infty} \frac{A_\mu}{\mu\, B_\mu} = \frac{2}{\sqrt{\pi}}\,, \tag{11.50}$$

where we have used Eqs. (11.34) and (11.35) for the expressions of A_μ and B_μ. We next substitute this asymptotic constant for the ratio $g_2(z)/h'_2(z)$ in the integral on the rhs of Eq. (11.44). The integral can then be performed trivially and we get, for large N,

$$\alpha_N \xrightarrow{N\to\infty} \frac{4}{\sqrt{\pi}}. \tag{11.51}$$

From Eq. (11.43) we then get for the mean record number

$$\langle R_{n,N}\rangle \xrightarrow[N\to\infty]{n\to\infty} \frac{4}{\sqrt{\pi}}\sqrt{n}\,. \tag{11.52}$$

In contrast to case I in Eq. (11.49), here the mean record number becomes independent of N for large N.

11.4 The distribution of the number of records for finite σ^2

In the previous section, we performed a very precise study of the mean number of records $\langle R_{n,N}\rangle$ up to step n, in both cases where σ^2 is finite and divergent. In the present section, we investigate the full probability distribution function (pdf) of the record number $R_{n,N}$. However, we have been able to make analytical progress for the record number distribution only in case I where σ^2 is finite to which we restrict ourselves below.

The clue that leads to an analytical computation of the record number distribution is actually already contained in the exact expression of the mean record number in Eqs. (11.43) and (11.46). This result suggests that there perhaps is a relation between the record number $R_{n,N}$ and the stochastic variable $Y_{n,N}$ defined as

$$Y_{n,N} = \max_{0\le m\le n} x_{\max}(m) = \max_{0\le m\le n}\max_{0\le i\le N}[x_i(m)]. \tag{11.53}$$

Note that $Y_{n,N}$ simply denotes the maximum position of all the walkers *up to* step n. In this section, we will see that for case I where σ^2 is finite, indeed there is a close relation between the two random variables $R_{n,N}$ and $Y_{n,N}$.

To uncover this relation, it is actually instructive to consider first the case of N independent lattice random walks defined by Eq. (11.4) where the noise $\eta_i(m)$'s are i.i.d. random variables with a distribution $f(\eta) = \frac{1}{2}\delta(\eta-1)+\frac{1}{2}\delta(\eta+1)$. Consider now the time evolution of the two random processes $R_{n,N}$ and $Y_{n,N}$. At the next time step $(n+1)$, if a new site on the positive axis is visited by any of the walkers for the first time, the process $Y_{n,N}$ increases by 1, otherwise its value remains unchanged. Whenever this event happens, i.e., a new site on the positive side is visited for the first time, one also has a record event, i.e., the process $R_{n,N}$ also increases by 1. Otherwise $R_{n,N}$ remains unchanged. Thus, the two random processes $Y_{n,N}$ and $R_{n,N}$ are completely locked with each other at all steps: whenever one of them increases by unity at a given step the other does the same simultaneously and when one of them does not change, the other also remains unchanged. In other words, for every realization, we have, $Y_{n+1,N} - Y_{n,N} = R_{n+1,N} - R_{n,N}$. Now, initially all walkers start at the origin indicating $Y_{0,N} = 0$ while $R_{0,N} = 1$ since the initial point is counted as a record by convention. This allows us to write the following identity for all n and N

$$R_{n,N} = Y_{n,N} + 1 = \max_{0\le m\le n}\max_{0\le i\le N}[x_i(m)] + 1\ . \tag{11.54}$$

We can now take advantage of this identity to compute the probability $P(R_{n,N} = M, n)$ as the distribution of $Y_{n,N}$, i.e., the maximum of N independent lattice walkers up to n steps can be computed using the standard method of images. One obtains, after some manipulations left in Appendix 11.7, for $0 \leq M \leq N+1$

$$\begin{aligned} P(R_{n,N} = M, n) &= \frac{1}{2^{n\,N}} \left(\sum_{k=0}^{\lfloor \frac{n+M}{2} \rfloor} \left[\binom{n}{k} - \binom{n}{k-M} \right] \right)^N \\ &- \frac{1}{2^{n\,N}} \left(\sum_{k=0}^{\lfloor \frac{n+M-1}{2} \rfloor} \left[\binom{n}{k} - \binom{n}{k-M+1} \right] \right)^N , \end{aligned} \tag{11.55}$$

where $\lfloor x \rfloor$ is the largest integer not greater than x. For instance, for $N = 1$ one gets from Eq. (11.55)

$$P(R_{n,N} = M, n) = \frac{1}{2^n} \binom{n}{\lceil \frac{n+M-1}{2} \rceil} , \tag{11.56}$$

where $\lceil x \rceil$ is the smallest integer not less than x. We have checked that this formula for $N = 1$ (11.56) yields back the result for the first moment $\langle R_{n,1} \rangle$ as obtained in Ref. [29]. From the above formula (11.55) one can also compute $\langle R_{n,N} \rangle$, for instance with Mathematica, although obtaining a simple closed form formula for it for $N > 1$ seems rather difficult. In Table 11.1 we have reported the first few values of $\langle R_{n,N} \rangle$ for $N = 1$ to $N = 4$.

	n=0	n=1	n=2	n=3	n=4
$N = 1$	1	$\frac{3}{2} = 1.5$	$\frac{7}{4} = 1.75$	2	$\frac{35}{16}$ $= 2.187...$
$N = 2$	1	$\frac{7}{4} = 1.75$	$\frac{35}{16}$ $= 2.187...$	$\frac{81}{32}$ $= 2.531...$	$\frac{723}{256}$ $= 2.824...$
$N = 3$	1	$\frac{15}{8} = 1.875$	$\frac{157}{64}$ $= 2.453...$	$\frac{731}{256}$ $= 2.855...$	$\frac{13145}{4096}$ $= 3.209...$
$N = 4$	1	$\frac{31}{16}$ $= 1.937...$	$\frac{671}{256}$ $= 2.621...$	$\frac{6303}{2048}$ $= 3.077...$	$\frac{227343}{65536}$ $= 3.468...$

Table 11.1: First values of $\langle R_{n,N} \rangle$ obtained from Eq. (11.55).

Using the identity (11.54), one can also obtain the large n behavior of $R_{n,N}$. Indeed, in this limit, each rescaled ordinary random walk $x_i(\tau n)/\sqrt{n}$ converges, when $n \to \infty$, to a Brownian motion $B_{D=\frac{1}{2},i}(\tau)$ with a diffusion coefficient $D = 1/2$, on the unit time interval, $\tau \in [0, 1]$. Therefore from the above identity (11.54) one gets, in the limit $n \to \infty$

$$\frac{R_{n,N}}{\sqrt{n}} \to \tilde{M}_N = \max_{1 \leq i \leq N} \max_{0 \leq \tau \leq 1} \left[B_{D=\frac{1}{2},i}(\tau) \right] . \tag{11.57}$$

Now, the distribution of $\max_{0 \leq \tau \leq 1} B_{D=\frac{1}{2},i}(\tau)$, i.e., the maximum of a single Brownian motion (with diffusion constant $D = 1/2$) over a unit interval is well known (see e.g., in [21])

$$\text{Proba.}\,[\tilde{M}_1 \leq m] = \sqrt{\frac{2}{\pi}} \int_0^m \exp\left(-\frac{x^2}{2}\right) \, dx = \text{erf}\left(\frac{m}{\sqrt{2}}\right) , \tag{11.58}$$

Eq. (11.57) demonstrates that in this case, $R_{n,N}/\sqrt{n}$ is distributed like the maximum of a collection N i.i.d positive random variables $\{z_1, z_2, \ldots, z_N\}$, each drawn from the distribution: $p(z) = \sqrt{\frac{2}{\pi}}\, e^{-z^2/2}$ for $z \geq 0$ and $p(z) = 0$ for $z < 0$. From Eq. (11.58) one obtains also that $\langle R_{1,n} \rangle \approx \sqrt{2n/\pi}$, for $n \gg 1$, as obtained in Ref. [29], using a different method. More generally for any N one has

$$\lim_{n\to\infty} \text{Proba.} \left[\frac{R_{n,N}}{\sqrt{n}} \leq m \right] = \left(\text{Proba.} \, [\tilde{M_1} \leq m] \right)^N = \left[\text{erf} \left(\frac{m}{\sqrt{2}} \right) \right]^N \, . \tag{11.59}$$

Having discussed the lattice random walk, let us now return to the case where the jump distribution is continuous in space but with a finite σ^2. In this case, we do not have an identity between $R_{n,N}$ and $Y_{n,N}$ analogous to Eq. (11.54) for lattice random walks. Nevertheless, we conjecture below and later provide numerical evidence in section VA, that for large n, the scaled random variable $R_{n,N}/\sqrt{n}$ converges, in distribution, to the scaled variable $Y_{n,N}/\sqrt{n}$ up to a prefactor $\sigma/\sqrt{2}$, i.e.,

$$\lim_{n\to\infty} \frac{R_{n,N}}{\sqrt{n}} \equiv \lim_{n\to\infty} \frac{\sqrt{2}}{\sigma} \frac{Y_{n,N}}{\sqrt{n}} \tag{11.60}$$

where $\equiv$ indicates that the random variables on both sides of Eq. (11.60) have the same probability distribution. On the other hand, using central limit theorem, it is easy to see that the position of each rescaled walker $\frac{\sqrt{2}}{\sigma} \frac{x_i(\tau n)}{\sqrt{n}}$ converges in distribution, as $n \to \infty$, to a continuous-time Brownian motion $B_{D=1,i}(\tau)$ with a diffusion coefficient $D = 1$, over the unit interval $\tau \in [0, 1]$. Thus the conjecture in Eq. (11.60) is equivalent to

$$\lim_{n\to\infty} \frac{R_{n,N}}{\sqrt{n}} \equiv M_N = \max_{1\leq i\leq N} \max_{0\leq\tau\leq 1} \left[B_{D=1}(\tau) \right] \, . \tag{11.61}$$

The argument leading to this conjecture in Eq. (11.61) can be framed as follows. Consider first the case for $N = 1$. In this case, the full distribution of $R_{n,1}$ was computed in Ref. [29] for all n and in particular, for large n, it was shown that [29] that

$$\lim_{n\to\infty} \frac{R_{n,1}}{\sqrt{n}} \equiv M_1 = \max_{0\leq\tau\leq 1} \left[B_{D=1}(\tau) \right] \, , \tag{11.62}$$

where $B_{D=1}(\tau)$ is a Brownian motion with diffusion coefficient $D = 1$ on the unit time interval starting from $B_{D=1}(0)$. This result (11.62), for continuous jump distribution, is very similar to the one obtained for a lattice random walk in Eq. (11.57) for $N = 1$, except that the diffusion coefficient of the Brownian motion involved in the discrete case is $D = 1/2$ while it is $D = 1$, independently of σ^2, for continuous jump distributions. In particular, from Eq. (11.62) one obtains

$$\text{Proba.} \left(\frac{R_{n,1}}{\sqrt{n}} \leq x \right) \xrightarrow[n\to\infty]{} \text{erf} \left(\frac{x}{2} \right) \, . \tag{11.63}$$

Hence, at least for $N = 1$, we know that for large n, $R_{n,1}/\sqrt{n}$ for the continuous jump distribution in Eq. (11.62) behaves in a statistically similar way as that for lattice walks, the only difference is that the effective diffusion constant of the underlying Brownian motion is $D = 1$ in the continuous case (Eq. (11.61)), while it is $D = 1/2$ for the lattice case (Eq. (11.57)). Based on this exact relation for $N = 1$, it is then natural to presume that this asymptotic equality in law between record numbers for continuous jump process and lattice walks holds even for $N > 1$, thus leading to the conjecture in Eq. (11.61). As a first check of the validity of this conjecture, we note that the result for the first moment of the record number in Eqs. (11.43) and (11.48) is fully consistent with the conjecture in Eq. (11.61).

As announced in the introduction (11.10), the conjecture in Eq. (11.61) is equivalent to say that $R_{n,N}/\sqrt{n}$ converges, in distribution, to the maximum of N independent random variables $M_N = \max(y_1, y_2, \ldots, y_N)$, where $y_i \geq 0$'s are i.i.d non-negative random variables each drawn from distribution $p(y) = \frac{1}{\sqrt{\pi}}\, e^{-y^2/4}$ for $y \geq 0$ and $p(y) = 0$ for $y < 0$. In particular, the cdf of $R_{n,N}/\sqrt{n}$ is given by

$$\text{Prob.}\left(\frac{R_{n,N}}{\sqrt{n}} \leq x\right) \xrightarrow[n\to\infty]{} \left[\text{erf}\left(\frac{x}{2}\right)\right]^N \ . \tag{11.64}$$

In section VB we will demonstrate that this conjecture is well supported by our numerical simulations. From Eq. (11.61), one can then use standard results for the extreme statistics of independent random variables [46] to obtain that for large N and large n, the scaled random variable $R_{n,N}/\sqrt{n}$ (properly shifted and scaled) converges, in distribution, to the Gumbel distribution as announced in Eq. (11.9).

Although we have not found a rigorous proof of the above result (11.61), the fact that both formulae for random walk with discrete (11.57) and continuous (11.61) jump distribution differ essentially by a factor of $\sqrt{2}$ is reminiscent of a similar difference, by the same factor of $\sqrt{2}$, for the survival probability $q_n(0)$ corresponding to both random walks (starting from the origin). This quantity $q_n(0)$ plays indeed a crucial role in the computation of the record statistics of a random walk [29]. For the lattice random walk, one has indeed $q_n(0) \sim \sqrt{2/\pi}\, n^{-1/2}$ while for the continuous random walk one has, from Sparre-Andersen's theorem, $q_n(0) \sim \sqrt{1/\pi}\, n^{-1/2}$, independently of σ^2. This fact certainly deserves further investigations.

In the case of divergent σ^2, the two random variables $R_{n,N}$ and $Y_{n,N}$ do not seem to be related in any simple way. This is already evident from the result for the mean record number $\langle R_{n,N}\rangle$ in Eq. (11.52) for $0 \leq \mu < 2$. As $n \to \infty$, $\langle R_{n,N}\rangle \approx 4/\sqrt{\pi}\,\sqrt{n}$ where the prefactor does not depend on N for large N. In contrast, using standard extreme value statistics [46], it is easy to show that the mean value $\langle Y_{n,N}\rangle \sim (N\,n)^{1/\mu}$ for large n and N with $0 < \mu < 2$. This rather different asymptotic behavior of the mean thus already rules out any relationship between $R_{n,N}$ and $Y_{n,N}$ for case II. So, for this case, our result for the distribution of $R_{n,N}$ is only restricted to numerical simulations that are presented in the next section.

11.5 Numerical simulations

In this section we present the results of our numerical simulations of N independent random walks both with a finite σ^2 (case I) and with a divergent σ^2 (case II) and compare them with our analytical results. In the first subsection 11.5.1 we study the statistics of the record numbers (both its mean value and its full distribution). Since, at least in case I, the mean record number is strongly related to the expected maximum of the process we will then analyze the evolution of the largest of the N random walkers. This will be done in section 11.5.1. We find that the statistics of the maximum significantly differs between the cases I and II. Finally, in section 11.5.3 we will consider the correlations between individual record events in the two different cases and show that, at least asymptotically, these correlations are not different from the case of only one single random walker.

11.5.1 Statistics of the record numbers

Case I (σ^2 finite). In Fig. (11.3), we show our numerical results for $\langle R_{n,N}\rangle$ for σ^2 finite, which were obtained by a direct simulation of the jump process in Eq. (11.2) with $n = 10^4$ steps, with a Gaussian distribution of the jump variables $\eta_i(m)$'s (mean 0 and $\sigma^2 = 1$). These results have been obtained by averaging over 10^3 different realizations of the random walks. These data, on Fig. (11.3) are indexed by the label 'Gaussian'. Our numerical data show a very nice agreement with our analytical result obtained in Eq. (11.43) yielding $\langle R_{n,N}\rangle/\sqrt{n} = \alpha_N$. The large N behavior of α_N can be easily obtained by a saddle point

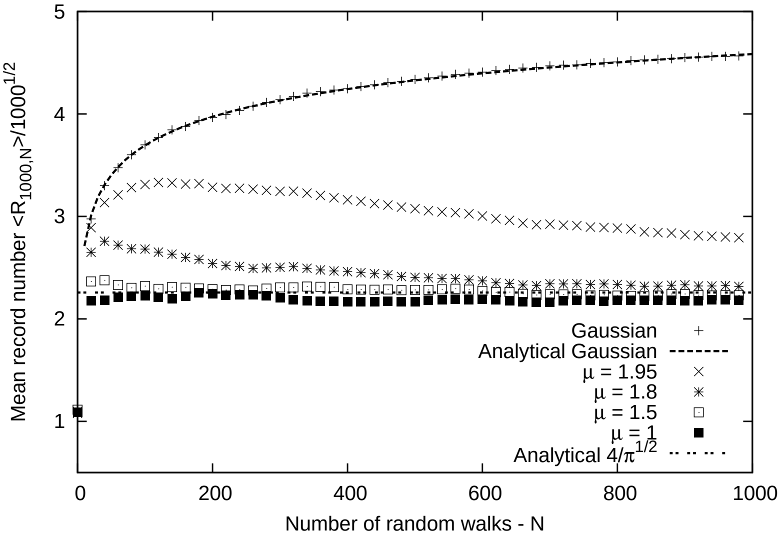

Figure 11.3: Rescaled mean record number $\langle R_{n,N}\rangle/\sqrt{n}$ for a fixed $n = 1000$ plotted against the number of random walkers N. We performed simulations with jump distributions of the type $f(\eta) \sim |\eta|^{-\mu-1}$ and different $\mu = 1, 1.5, 1.8$ and 1.95 and for the Gaussian jump distribution with zero mean and $\sigma^2 = 1$. The Gaussian case is compared to our analytical finding for the finite σ^2 case (Eq. (11.49)), which is given by the dashed line. The thick black line gives the analytical result for the infinite σ^2 regime (Eq. (11.52)). With increasing N all $\langle R_{n,N}\rangle/\sqrt{n}$ with $\mu < 2$ approach this line of value $4/\sqrt{\pi}$.

analysis, yielding:

$$\frac{\langle R_{n,N}\rangle}{\sqrt{n}} = 2\sqrt{\log N} - \frac{\log(\log N)}{2\sqrt{\log N}} + \mathcal{O}[(\log N)^{-1/2}]\ . \tag{11.65}$$

It turns out that for $N \sim 1000$, the sub-leading corrections (11.65) are still sizeable.

We have also computed numerically the distribution of the (scaled) record numbers $R_{n,N}/\sqrt{n}$ and compared it to our conjecture in Eq. (11.61). The results of this comparison, for different values of $N = 2, 4$ are shown in Fig. 11.4 ,where the data were obtained by averaging over 5×10^4 realizations of independent random walks of $n = 10^4$ steps. The data, for two different continuous jump distributions (exponential and uniform) show a very nice agreement with our analytical prediction in Eq. (11.64), which we argue to be an exact result. As mentioned above, one expects that this distribution will converge, for $N \to \infty$ to a Gumbel distribution (11.9), albeit with strong finite N effects.

Case II (σ^2 divergent). In Fig. (11.3), we show our numerical results for $\langle R_{n,N}\rangle$ for σ^2 divergent, obtained by a direct simulation of the random walk (11.2) where the distribution of $\eta_i(m)$'s has a power law tail $f(\eta) \sim |\eta|^{-1-\mu}$ with $\mu < 2$, and different values of μ. The data presented there were also obtained by averaging over 10^3 different realizations of random walks, with 10^4 steps. These data show that, in this case, $\langle R_{n,N}\rangle/\sqrt{n}$ approaches a constant value for fixed (and large) n and $N \to \infty$, which is fully consistent with the value of $4/\sqrt{\pi}$ obtained analytically in Eq. (11.52). The simulations also show how the speed of this convergence is modified when $\mu < 2$ is varied. While for small $\mu \ll 2$, $\langle R_{n,N}\rangle/\sqrt{n}$ approaches the universal limit value of $4/\sqrt{\pi}$ very quickly, we find a slower convergence for $2 - \mu \ll 1$.

The numerical computation of the distribution of the (scaled) record number $R_{n,N}/\sqrt{n}$ is of special interest because an analytical study of it, beyond the first moment, is still lacking. The two plots on Fig. (11.5) show our numerical results for this distribution, which were obtained by averaging over 10^4 independent random walks of length $n = 10^4$. The left panel in Fig. (11.5) shows the rescaled distribution of $R_{n,N}/\sqrt{n}$ at a fixed time

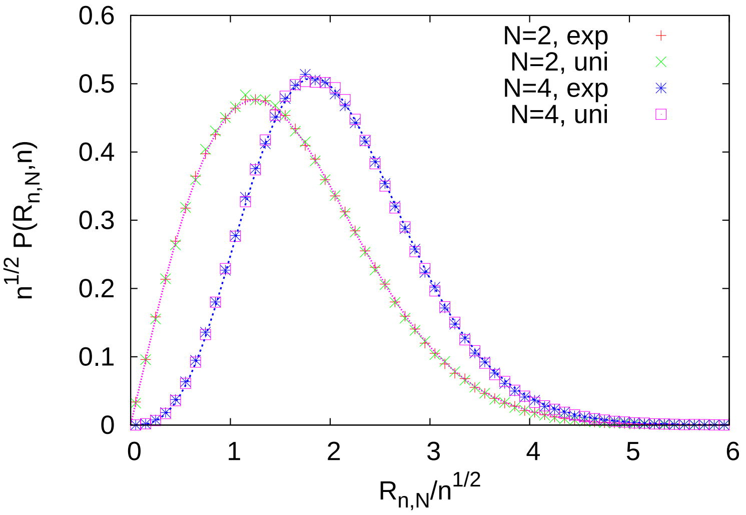

Figure 11.4: Scaled probability distribution function $\sqrt{n}P(R_{n,N}, n)$ as a function of $R_{n,N}/\sqrt{n}$ for $N = 2$ and $N = 4$ independent of random walks of length $n = 10^4$. For each value of N, we show the result for the case where the jumps are distributed exponentially ('exp') and uniformly between $-1/2$ and $1/2$ ('uni'). The dotted lines correspond to the result in Eq. (11.64) which we conjectured to be the exact one. There are no fitting parameters.

step $n = 10^3$ and $\mu = 1$. Apparently all curves for $N = 10, 10^2, 10^3$ and 10^4 collapse on one line. In the inset of the left figure we kept $n = 10^3$ and $N = 10^3$ fixed and varied μ: one can see that all the cdf's collapse. These results suggest that

$$\text{Prob.}\left[\frac{R_{n,N}}{\sqrt{n}} \leq x\right] \xrightarrow[N\to\infty]{n\to\infty} F_2(x) \;, \tag{11.66}$$

where the limiting distribution function $F_2(x)$ is independent of $\mu < 2$. We tried to guess the analytic form of $F_2(x)$ by comparing with several known continuous distributions that are defined for positive real numbers. We are certain that $F_2(x)$ is not a Gaussian distribution. By far the best results were obtained by fitting with a Weibull distribution

$$F_2(x) = 1 - \exp\left(-(\lambda x)^k\right) \;, \tag{11.67}$$

with two free real parameters $\lambda > 0, k > 0$. Fitting with the least-squares method gives values of $\lambda \approx 0.8944 \pm 0.0003$ and $k = 2.558 \pm 0.003$. The right panel in Fig. (11.5) compares this fit with the distribution obtained from a simulation of $N = 10^4$ random walks of length $n = 10^4$. While the agreement, both for the cdf and the probability density function (pdf) is quite good, there are still some deviations between the two, particularly for small values close to zero. We are not sure, if this difference is a finite N effect or if the real limit distribution of $R_{n,N}/\sqrt{n}$ for $N \to \infty$ effectively differs from a Weibull distribution.

11.5.2 Temporal evolution of two stochastic processes: the record number $R_{n,N}$ and the global maximum $Y_{n,N}$ up to step n

In the case of jump distributions with a finite second moment σ^2 (case I), we have shown that the mean $\langle Y_{n,N}\rangle$ of the maximum of all walkers up to step n and the mean record number $\langle R_{n,N}\rangle$ are proportional to each other in the limit $n \to \infty$, both grow with $\sqrt{n \ln N}$. In contrast, for Lévy walks with index $\mu < 2$ (case II) the relation between these two observables does not hold any more and the mean record number grows much slower than

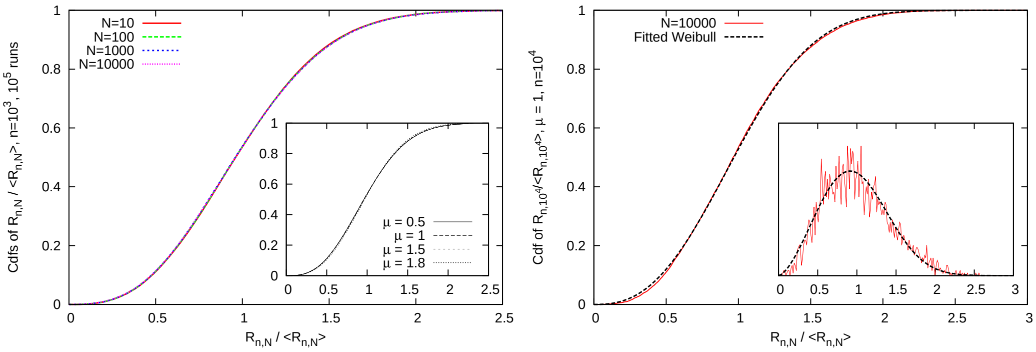

Figure 11.5: Left: Cumulative distribution function (cdf) of the scaled variable $R_{n,N}/\sqrt{n}$ for the Lévy index $\mu = 1$, which approaches the universal distribution $F_2(x)$. The figure gives results for a fixed $n = 10^3$ and different N, for each N we performed 10^5 simulations. The inset gives simulations of the cdf for fixed $N = 10^3$ while Lévy index is varied. **Right:** The cdf for $\mu = 1$ and $N = 10^4$. We have fitted the data with a Weibull distribution as in Eq. (11.67), where the fitting parameters were $\lambda \approx 0.8944 \pm 0.0003$ and $k = 2.558 \pm 0.003$. The inset gives the corresponding curves for the pdf's.

the maximum (see the discussion at the end of section IV). To illustrate the similarities and differences in the growth rate of $R_{n,N}$ and $Y_{n,N}$ in the two cases (I and II), we compare their respective time evolution for 4 different samples in Fig. (11.6). On the left panel, we consider the Gaussian jump distribution with zero mean and $\sigma^2 = 1$ and we see that the process $R_{n,N}$ and $(\sqrt{2}/\sigma)\, Y_{n,N}$ become identical very quickly. Moreover, the two processes evolve almost in a deterministic fashion with growing n and hardly fluctuate from one sample to another. In contrast, on the right panel where we plot the two processes for $\mu = 1$, their behavior change drastically. First of all, the two processes $R_{n,N}$ and $Y_{n,N}$ do not seem to have relation to each other. While $R_{n,N}$ again evolves almost deterministically and in a self-averaging way, the trajectories of the process $Y_{n,N}$ differ strongly from one sample to another and $Y_{n,N}$ is clearly non self-averaging. In particular, the process $Y_{n,N}$ can, like in a single Lévy flight, perform very large jumps exceeding its previous value by several orders of magnitude.

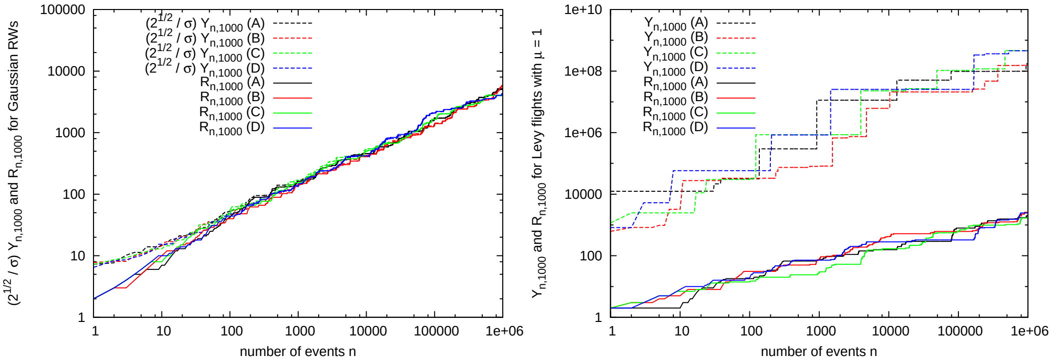

Figure 11.6: Left: Time evolution of the record number $R_{n,N}$ and the rescaled maximum value $(\sqrt{2}/\sigma)\, Y_{n,N}$ reached up to the n-th step for four different realizations of $N = 1000$ random walks (labeled A,B,C,D) with Gaussian jump distribution (zero mean and $\sigma = 1$) (case I). Here, the results look rather deterministic and for $n \to \infty$, we find $(\sqrt{2}/\sigma)\, Y_{n,N} \approx R_{n,N}$ for every realization. **Right:** $R_{n,N}$ and $Y_{n,N}$ for four different realizations of $N = 1000$ Lévy flights (labeled A,B,C,D) with $\mu = 1$. The behavior of $Y_{n,N}$ for the Lévy flight is completely different from $R_{n,N}$: while $R_{n,N}$ is self-averaging, $Y_{n,N}$ fluctuates widely from one sample to another and is not self-averaging.

11.5.3 Correlations between record events

An important feature of the record statistics of a single random walk ($N = 1$) is its renewal property, which leads to the fact, that each time after a record event, the record statistics is, in some sense, *reseted.* After a record event the process evolves as a new process with the record value as its new origin. Therefore it is very simple to give the pairwise correlations between record events. In fact, from the above argument, we have

$$\text{Prob}\,[\text{rec. at } n-k \text{ and } n] = \text{Prob}\,[\text{rec. at } n-k] \times \text{Prob}\,[\text{rec. at } k] = r_{n-k,1} r_{k,1}\ . \quad (11.68)$$

Using the results from [29] this gives the following (exact) result for Prob [rec. at $n-k$ and n]:

$$\text{Prob}\,[\text{rec. at } n-k \text{ and } n] = \binom{2\,(n-k)}{n-k}\binom{2k}{k} 2^{-2n}\ . \quad (11.69)$$

In the special case of $k = 1$ we find Prob [rec. at $n-1$ and n] $= \frac{1}{2} r_{n-1}$. With this we find that the conditional probability of a second record directly following a record that just occurred is always given by

$$\text{Prob}\,[\text{rec. at } n | \text{rec. at } n-1] = \frac{1}{2}\ . \quad (11.70)$$

In our efforts to understand and compute the distribution of the record number $R_{n,N}$ for Lévy flights (case II), we considered the correlations between successive record events also for $N \gg 1$ random walks. If the correlations between successive record events in the large N limit would vanish, one could assume that the asymptotic distribution of $R_{n,N}$ approaches a Gaussian. However, we found that this is not the case. Fig. (11.7) gives the behavior of Prob [rec. at $n-k$ and n] for the $N = 1$ case, as well as for $N = 10^3$ for random walks of the two cases I and II with Lévy indices $\mu = 2$ and $\mu = 1$. In all three cases Prob [rec. at $n-1$ and n] approaches $\frac{1}{2}$Prob [rec. at $n-1$] $= \frac{1}{2} r_{n,N}$ for large n, proving that for $n \to \infty$ the probability for a second record after an occurred one is just 1/2. The inset of Fig. (11.7) also shows this behavior. Here, while for $N = 1$ the conditional probability for a second record is always 1/2, this value is only approached for larger n in the case of $N \gg 1$. For small n the conditional probability is larger.

11.6 Comparison to stock prices

The oldest application of the random walk model, which was already proposed by Le Bachelier [47] in 1900, is the one to stock data [48, 49]. In his model the stock prices perform a so-called geometric random walk and trends in the stocks are modeled by a linear drift in the logarithms of the stock prices. In [32] the record statistics of stocks in the Standard and Poors 500 [35] (S&P 500) index were compared to the records in a random walk with a drift. The authors could show that on average, on a time interval of $n = 100$ trading days, the statistics of upper records in individually detrended stocks are in good agreement with the same statistics of a random walk with a symmetric jump distribution. The lower records however showed a significant deviation from this model.

Here, we want to extend this analysis to the record statistics of N stocks. The question is, to what degree, the record statistics of N normalized and randomly chosen stocks from the S&P 500 can be compared to the record statistics of N independent random walks. As in [32], the observational data we used consisted of 366 stocks that remained in the S&P 500 index for the entire time-span from January 1990 to March 2009. Overall, we had data for 5000 consecutive trading days for each stock at our disposal. In [32] we found that it is useful to analyze this data over smaller intervals, on which one can then detrend the measurements. We decided to split up the 5000 trading days into 20 consecutive intervals of each 250 trading days, which is roughly one calendar year. In each of these

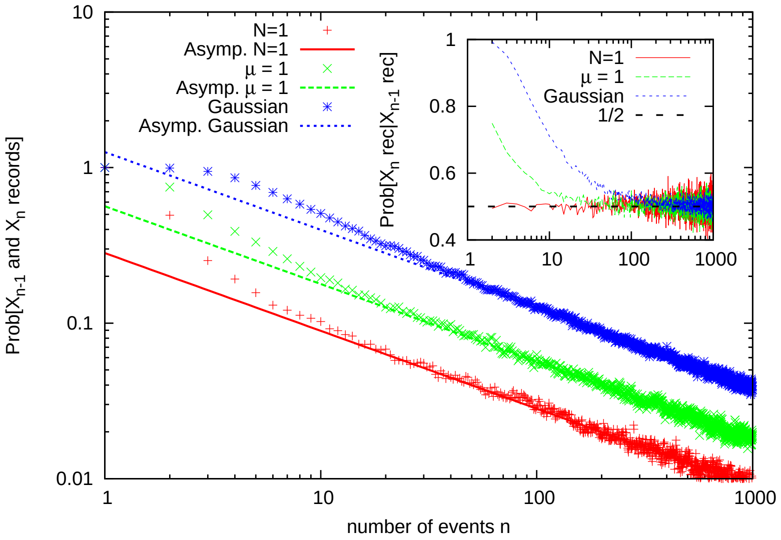

Figure 11.7: Probability for two successive records at times $n-1$ and n for a single random walker as well as N random walks with jump distributions of tail-exponents $\mu = 2$ (case 1) and $\mu = 1$ (case 2). In all three cases this probability asymptotically approaches a value of 1/2 times the probability for a record in the nth step. Therefore, for large n, the correlations between the record events in the $N \gg 1$ regime are the same as in the $N = 1$ case. This is also shown by the inset, where we plotted the probability for a record in the nth step conditioned on a record in the previous one, which approaches 1/2 in all three cases.

intervals we considered the logarithms $X_i = \ln S_i/S_0$ of the stock prices S_i, where S_0 is the first trading day. The random walk model then suggest that these logarithms X_i perform a biased random walk that starts at the origin ($X_0 = 0$). Since our analytical theory presented in this paper only works for symmetric random walks we had to detrend the stocks. We subtracted an index-averaged linear trend obtained by linear regression from the X_i's in order to obtain symmetric random walkers. Finally, in order to make the stocks comparable to our model of N random walkers of the same jump distribution, we had to normalize the X_i's by dividing through the standard deviations of the respective individual jump distributions. After this detrending and normalization we can assume that the jump distributions in the individual time series have at least the same mean and the same variance, which should then be given by 0 and 1.

For a fixed N, we randomly selected subsets of size N from the 366 detrended and normalized stocks for each of the 20 intervals of length $n = 250$ and computed the evolution of the record number $R_{n,N}$ in these subsets. To get reliable statistics we average $R_{n,N}$ over 10^4 different subsets with a fixed N and also averaged over the 20 consecutive intervals. The resulting $\langle R_{n,N}\rangle$'s for the upper and the lower record number and N between 1 and 100 are given in Fig. (11.8). We find that both the curve for the upper and the curve for the lower mean record number are not in agreement with our theoretical prediction for the case of N Gaussian random walks given by $\langle R_{n,N}\rangle = 2\sqrt{n \ln N}$. However, Fig. (11.8) shows, that the $\langle R_{n,N}\rangle$ for the stocks increase with N. We also considered subsets of size $N > 100$ and found that for N closer to the maximal value of 366, $\langle R_{n,N}\rangle$ gets almost constant in N.

While the increase of $\langle R_{n,N}\rangle$ for smaller N indicates that the statistics behave like N independent Gaussian random walks, the fact that it saturates for large N could indicate that they behave like a Lévy flight with Lévy index $\mu < 2$. We know however, that the tail exponent of the daily returns $\ln S_i/S_{i-1}$ in the stock data is much larger than $\mu = 2$ and that they definitely do not perform a Lévy flight [50–52]. Much more likely is that the correlations between the individual stocks play an important role. In addition, we observed

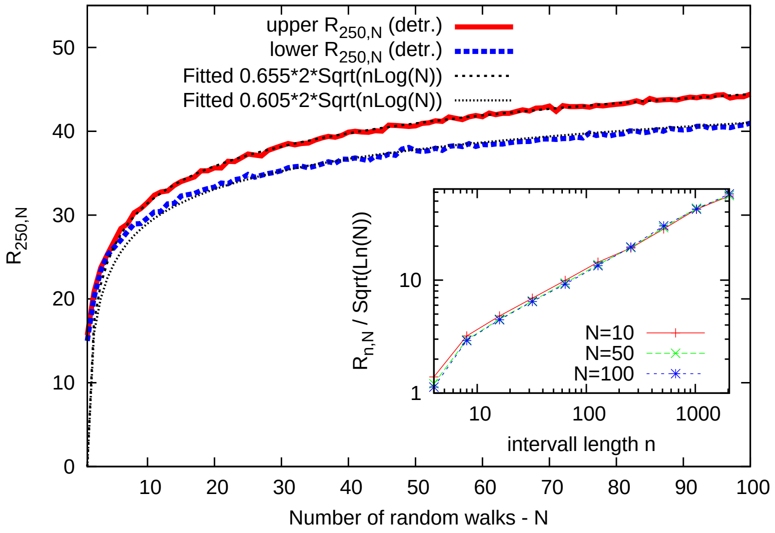

Figure 11.8: The averaged upper and lower record number after $n = 250$ trading days in the S&P 500 stock data. The 5000 trading days in [35] were subdivided in 20 intervals of 250 days and then linearly detrended in these intervals using the average linear trend of the index. Then we chose N stocks randomly out of the total number of 366 stocks and analyzed the evolution of the record number in this set. This random picking was repeated 10^4 times and the results were averaged to obtain the figure. The dashed lines give our analytical prediction for N Gaussian random walks multiplied by fitted prefactors. The inset gives the behavior of the $\langle R_{n,N}\rangle/\sqrt{\ln N}$ for different N plotted against the interval length n, confirming the proportionality $\langle R_{n,N}\rangle \propto \sqrt{\ln N}$.

that at least for $N < 100$, $\langle R_{n,N}\rangle$ grows proportional to $2\sqrt{n \ln N}$. One way to interpret this finding is the following: When we assume that in N stocks only a smaller number of N^γ (with $\gamma \in \mathbb{R}^+$ and $\gamma < 1$) is effectively independent and that only these N^γ stocks contribute to the record statistics, the mean record number should be given by

$$\langle R_n, N\rangle = \langle R_n, N^\gamma\rangle^{(\text{Gaussian})} = 2\sqrt{\gamma\; n \ln N}\;, \tag{11.71}$$

and saturate if the value of N^γ_{max} is achieved, where N_{max} is the total number of stocks. In Fig. (11.8) we fitted curves of the form $\sqrt{\gamma_\pm}\; 2\sqrt{n \ln N}$ with $\sqrt{\gamma_+} \approx 0.655$ and $\sqrt{\gamma_-} \approx 0.605$ to the development of the upper and lower $\langle R_{n,N}\rangle$. The good agreement with the fitted curves and the data confirms our assumption. Apparently, the record statistics of N detrended and normalized stocks is the same as the one of N^γ independent Gaussian random walks. This finding is also confirmed by the inset in Fig. (11.8). There we plotted $\langle R_{n,N}\rangle/\sqrt{\ln N}$ for different interval length n and some different subset sizes N. The fact that all the lines collapse tells us that $\langle R_{n,N}\rangle/\sqrt{\ln N}$ is independent of N and therefore

$$\langle R_{n,N}\rangle \propto \sqrt{n \ln N}. \tag{11.72}$$

11.7 Conclusion

In conclusion we have presented a thorough analysis of the record statistics of N independent random walkers with continuous and symmetric jump distributions. For $N > 1$, we have found two distinct cases: the case where the variance of the jump distribution σ^2 is finite and the case where σ^2 does not exist (case II) as in the case of Lévy random walkers with index $0 < \mu < 2$. In the first case we have found that the mean record number behaves like $\langle R_{n,N}\rangle \approx 2\sqrt{\log N}\sqrt{n}$ for $n, N \gg 1$ while in case II, $\langle R_{n,N}\rangle \approx \sqrt{4/\pi}\sqrt{n}$ for $n, N \gg 1$.

We have then argued that, in the first case, the full distribution of the scaled number of records $R_{n,N}/\sqrt{n}$ is given by the distribution of the maximum of N independent Brownian motions with diffusion coefficient $D = 1$. This statement was suggested by an exact result for lattice random walks and it was corroborated (i) by our exact calculation of the first moment $\langle R_{n,N}\rangle/\sqrt{n}$ valid for any value of N and (ii) by our numerical simulations. Of course it would be very interesting to obtain a proof of this result. From this connection with extreme value statistics, one thus expects that the distribution of $R_{n,N}/\sqrt{n}$ converges, for $N \to \infty$, to a Gumbel form (11.9). This connection between record statistics and extreme value statistics could also be useful to compute the record statistics in other models discussed in the introduction, for instance, in the Linear Drift Model [10, 16, 32]. In the case of Lévy random walkers, we have shown numerically that the full distribution of $\langle R_{n,N}\rangle/\sqrt{n}$ converges, when $N \to \infty$, to a limiting distribution $F_2(x)$ which is independent of μ.

The exact computation of this universal distribution remains a challenging problem. Other interesting questions concern the extension of these results to include a linear drift [30] or to the case of constrained Lévy walks, like Lévy bridges which were recently studied in the context of real space condensation phenomena [53]. Finally the applications of our results to the record statistics of stock prices from the Standard & Poors 500 index suggest that, among a set of N stocks, only a smaller number, which scales like N^γ, with $0 < \gamma < 1$, are effectively independent. The record statistics of these N^γ stocks is then very similar to the statistics of N^γ independent random walkers. This idea might be useful for future investigations of the fluctuations of such ensemble of stock prices.

Acknowledgements

We acknowledge support by ANR grant 2011-BS04-013-01 WALKMAT and the Department of Business Administration and Finance at the University of Cologne for providing access to the stock data from the S&P 500. GW is grateful for the kind hospitality of the Laboratoire de Physique Théorique et Modèles Statistiques during the completion of this work and for the financial support provided by DFG within the Bonn Cologne Graduate School of Physics and Astronomy.

APPENDIX I - Scaling behavior of $p_m(x)$ and $q_m(x)$ for finite σ^2

We start from Eqs. (11.24) and (11.25). When σ^2 is finite, by central limit theorem, the typical position of a walker after m steps scales as $m^{1/2}$ for large m. Hence the natural scaling variable is $z = x/m^{1/2}$. Consider first Eq. (11.24) satisfied by $p_m(x)$. To extract the leading scaling function in the scaling limit $x \to \infty$, $m \to \infty$ but keeping $z = x/m^{1/2}$ fixed, we need to investigate $\phi(s, \lambda)$, given explicitly in Eq. (11.23), in the limit when $\lambda \to 0$, $s \to 1$ but keeping the ratio $\lambda/\sqrt{1-s}$ fixed. To extract the behavior of $\phi(s, \lambda)$ in this scaling limit, it is advantageous to work with an alternative expression of $\phi(s, \lambda)$ derived in Ref. [25] for finite σ^2

$$\phi(s,\lambda) = \frac{1}{[\sqrt{1-s} + \sigma\lambda\sqrt{s/2}]} \exp\left[-\frac{\lambda}{\pi}\int_0^\infty \frac{dk}{\lambda^2+k^2} \ln\left[\frac{1-s\hat{f}(k)}{1-s+s\sigma^2k^2/2}\right]\right]. \qquad (11.73)$$

This expression is more suitable for extracting the scaling limit. In the limit $\lambda \to 0$ and $s \to 1$, the expression inside the exponential in Eq. (11.73) tends to 0 and hence, to leading order, we have

$$\phi(s,\lambda) \approx \frac{1}{[\sqrt{1-s} + \sigma\lambda\sqrt{s/2}]}\,. \qquad (11.74)$$

Inverting the Laplace transform with respect to λ (it has a simple pole at $\lambda = -\frac{1}{\sigma}\sqrt{2(1-s)}$) one gets from Eq. (11.24)

$$\sum_{m=0}^{\infty} s^m p_m(x) \approx \frac{\sqrt{2}}{\sigma}\, e^{-\sqrt{2(1-s)}\, x/\sigma}\,. \tag{11.75}$$

Setting $s = 1 - p$ with $p \to 0$ in the scaling limit, the sum on the lhs of Eq. (11.75) can be approximated, to leading order, by a continuous integral:

$$\sum_{m=0}^{\infty} s^m p_m(x) \approx \int_0^{\infty} p_m(x)\, e^{-pm} dm \tag{11.76}$$

and we have

$$\int_0^{\infty} p_m(x)\, e^{-p\,m}\, dm \approx \frac{\sqrt{2}}{\sigma}\, e^{-\sqrt{2p}\, x/\sigma}\,. \tag{11.77}$$

Next we need to invert the Laplace transform with respect to p. We use the explicit inversion formula, $\mathcal{L}^{-1}_{p\to m}[e^{-b\sqrt{p}}] = \frac{b}{2\sqrt{\pi}\, m^{3/2}} \exp[-b^2/4m]$. Applying this to Eq. (11.77) gives, to leading order, in the scaling limit

$$p_m(x) \approx \frac{x}{\sigma^2\sqrt{\pi}\, m^{3/2}} \exp\left[-\frac{x^2}{2\sigma^2 m}\right]\,, \tag{11.78}$$

which can be reorganized in the scaling form

$$p_m(x) \to \frac{1}{\sqrt{2\sigma^2\, m}}\, g_1\left(\frac{x}{\sqrt{2\,\sigma^2\, n}}\right)\,, \quad \text{where} \quad g_1(z) = \frac{2}{\sqrt{\pi}}\, z\, e^{-z^2}\,. \tag{11.79}$$

Next we consider $q_m(x)$ given in Eq. (11.25). Following exactly the same procedure as in the case of $p_m(x)$ we find, in the scaling limit,

$$\int_0^{\infty} q_m(x)\, e^{-pm}\, dm \approx \frac{1}{p}\left[1 - e^{-\sqrt{2p}\, x/\sigma}\right]. \tag{11.80}$$

Inverting the Laplace transform with respect to p upon using the explicit inversion formula, $\mathcal{L}^{-1}_{p\to m}[e^{-b\sqrt{p}}/p] = \mathrm{erfc}(b/\sqrt{4m})$, we get, to leading order in the scaling limit

$$q_m(x) \approx 1 - \mathrm{erfc}\left(\frac{x}{\sqrt{2\sigma^2\, m}}\right) = \mathrm{erf}\left(\frac{x}{\sqrt{2\sigma^2\, m}}\right)\,, \tag{11.81}$$

which proves the result in Eq. (11.29).

APPENDIX II - Scaling behavior of $p_m(x)$ and $q_m(x)$ for divergent σ^2

We consider jump distribution $f(\eta)$ such that its Fourier transform behaves, for small k, as $\hat{f}(k) \approx 1 - |ak|^\mu$ with $0 < \mu < 2$. In this case, the position of the walker after m steps, grows as $m^{1/\mu}$ for large m [24]. Hence the natural scaling limit is $x \to \infty$, $m \to \infty$ with the ratio $x/m^{1/\mu}$ fixed. For $p_m(x)$, we expect a scaling form $p_m(x) \approx m^{-1/2-1/\mu} g_2(x/m^{1/\mu})$. The power of m outside the scaling function is chosen to ensure that $\int_0^\infty p_m(x)dx \sim m^{-1/2}$. This is needed since we know from Eq. (11.20) and the Sparre Andersen theorem in Eq. (11.26) that $\int_0^\infty p_m(x)dx = q_m(0) \sim 1/\sqrt{\pi m}$ for large m. Similarly, for $q_m(x)$, we expect a scaling form $q_m(x) \approx h_2(x/m^{1/\mu})$ in the scaling limit.

To extract the leading scaling functions $g_2(z)$ and $h_2(z)$ respectively from Eqs. (11.24) and (11.25), we need to investigate the function $\phi(s,\lambda)$ in Eq. (11.23) in the corresponding scaling limit $\lambda \to 0$, $s \to 0$ but keeping the ratio $\lambda/(1-s)^{1/\mu}$ fixed. Fortunately, this was already done in Ref. [24] in a different context. Setting $s = 1-p$ with $p \to 0$, the leading behavior of $\phi(s,\lambda)$ in the scaling limit is given by (see Eqs. (43)-(47) of Ref. [24])

$$\phi(s,\lambda) \approx \frac{1}{\sqrt{p}} \exp\left[-\frac{1}{\pi}\int_0^\infty \frac{du}{1+u^2} \ln\left[1+\frac{1}{p}(a\,\lambda\,u)^\mu\right]\right]. \tag{11.82}$$

Let us first consider the function $p_m(x)$ in Eq. (11.24). We substitute the anticipated scaling form $p_m(x) = m^{-1/2-1/\mu} g_2(x/m^{1/mu})$ on the lhs of Eq. (11.24). As before, setting $p = 1-s$, we can replace, in the scaling limit, the sum over m by a continuous integral over m

$$\sum_{m=0}^\infty s^m \int_0^\infty p_m(x)\, e^{-\lambda x}\, dx \approx \int_0^\infty\int_0^\infty dx\, dm e^{-\lambda x - p\, m} m^{-1/2-1/\mu}\, g_2(xm^{-1/\mu}). \tag{11.83}$$

We then make a change of variable $x\, m^{-1/\mu} = z$ and $p\, m = y$ to get

$$\sum_{m=0}^\infty s^m \int_0^\infty p_m(x)\, e^{-\lambda x}\, dx \approx \frac{1}{\sqrt{p}} \int_0^\infty\int_0^\infty dz\, dy\, g_2(z)\, y^{-1/2}\, e^{-(\lambda\, p^{-1/\mu})\, y^{1/\mu}\, z - y} \tag{11.84}$$

We next substitute Eq. (11.84) on the lhs of Eq. (11.24) and Eq. (11.82) on the rhs of Eq. (11.24). Writing the scaled variable as $\lambda\, p^{-1/\mu} = w$ and comparing lhs with the rhs, we see that the $1/\sqrt{p}$ cancels from both sides leaving us with

$$\begin{aligned} \int_0^\infty dz\, g_2(z) \int_0^\infty dy\, y^{-1/2}\, e^{-y}\, e^{-w\, y^{1/\mu}\, z} &= \exp\left[-\frac{1}{\pi}\int_0^\infty \frac{du}{1+u^2} \ln\left[1 + a^\mu w^\mu u^\mu\right]\right] \\ &\equiv J_\mu(w)\,. \end{aligned} \tag{11.85}$$

Similarly, by substituting the anticipated scaling form $q_m(x) = h_2(x/m^{1/\mu})$ on the lhs of Eq. (11.25) and doing exactly the same series of manipulations, we get

$$\int_0^\infty dz\, h_2(z) \int_0^\infty dy\, y^{1/\mu}\, e^{-y}\, e^{-w\, y^{1/\mu}\, z} = \frac{1}{w}\, J_\mu(w) \tag{11.86}$$

where $J_\mu(w)$ is defined in Eq. (11.85).

For later purposes, it is further convenient to define a pair of Laplace transforms

$$\tilde{g}_2(\rho) = \int_0^\infty g_2(z)\, e^{-\rho\, z}\, dz \tag{11.87}$$

$$\tilde{h}_2(\rho) = \int_0^\infty h_2(z)\, e^{-\rho\, z}\, dz \tag{11.88}$$

in terms of which Eqs. (11.85) and (11.86) read

$$\int_0^\infty dy\, y^{-1/2}\, e^{-y}\, \tilde{g}_2(w\, y^{1/\mu}) = J_\mu(w) \tag{11.89}$$

$$\int_0^\infty dy\, y^{1/\mu}\, e^{-y}\, \tilde{h}_2(w\, y^{1/\mu}) = \frac{1}{w}\, J_\mu(w) \tag{11.90}$$

The equations (11.85) and (11.86) determine, in principle, the two scalings functions $g_2(z)$ and $h_2(z)$ for all z. In practice, it is hard to invert these two equations to obtain $g_2(z)$ and $h_2(z)$ for all z. However, it is possible to extract the large z asymptotics of these two functions by analyzing the leading singular behavior of $J_\mu(w)$ in Eq. (11.85) as $w \to 0$. Clearly, it follows from the definition of $J_\mu(w)$ in Eq. (11.85) that $J_\mu(0) = 1$. We are, however, interested in the leading singular correction term in $J_\mu(w)$ as $w \to 0$ which, as it turns out, depends on whether $0 < \mu < 1$, $1 < \mu < 2$ or $\mu = 1$. Below, we consider the three cases separately.

The case $0 < \mu < 1$ We consider $J_\mu(w)$ in Eq. (11.85) and compute the derivative $J'_\mu(w)$ as $w \to 0$. Simple computation shows that

$$J'_\mu(w) \xrightarrow{w\to 0} -\mu\, b_\mu w^{\mu-1}; \quad \text{where} \quad b_\mu = \frac{a^\mu}{\pi} \int\limits_0^\infty \frac{u^\mu\, du}{1+u^2}. \tag{11.91}$$

Note that the integral defining b_μ is convergent as $u \to \infty$ for $0 < \mu < 1$. Integrating over w and using $J_\mu(0) = 1$ we get the leading correction term as $w \to 0$

$$J_\mu(w) \approx 1 - b_\mu w^\mu + \dots \tag{11.92}$$

where b_μ is given in Eq. (11.91).

Substituting this small w behavior of $J_\mu(w)$ on the rhs of Eq. (11.89), it follows that to match the powers of w on both sides, the Laplace transform $\tilde{g}_2(\rho)$ must have the following small ρ behavior

$$\tilde{g}_2(\rho) \underset{\rho\to 0}{\sim} \frac{1}{\sqrt{\pi}} - \frac{2}{\sqrt{\pi}}\, b_\mu\, \rho^\mu\,. \tag{11.93}$$

Using the classical Tauberian theorem (for a simple derivation see the appendix A.2 of Ref. [54]), one immediately gets the following large z behavior of $g_2(z)$

$$g_2(z) \underset{z\to\infty}{\sim} \frac{A_\mu}{z^{1+\mu}} \tag{11.94}$$

with the amplitude

$$A_\mu = \frac{2\mu}{\sqrt{\pi}}\, \frac{b_\mu}{\Gamma(1-\mu)} = \frac{2\mu}{\sqrt{\pi}}\, \beta_\mu \tag{11.95}$$

where

$$\beta_\mu = \frac{b_\mu}{\Gamma(1-\mu)} = \frac{a^\mu}{\pi\Gamma(1-\mu)} \int\limits_0^\infty \frac{u^\mu}{1+u^2}\, du\,. \tag{11.96}$$

Similarly, substituting the small w behavior of $J_\mu(w)$ on the rhs of Eq. (11.90) and matching powers of w on both sides, we get

$$\tilde{h}_2(\rho) \underset{\rho\to 0}{\sim} \frac{1}{\rho} - b_\mu\, \rho^{\mu-1}\,. \tag{11.97}$$

Once again, using the Tauberian theorem of inversion, we get

$$h_2(z) \underset{z\to\infty}{\sim} 1 - \frac{B_\mu}{z^\mu} \tag{11.98}$$

with the amplitude

$$B_\mu = \frac{b_\mu}{\Gamma(1-\mu)} = \beta_\mu\,, \tag{11.99}$$

where β_μ is given in Eq. (11.96).

Finally, note that the ratio

$$\frac{A_\mu}{\mu\, B_\mu} = \frac{2}{\sqrt{\pi}} \tag{11.100}$$

is universal in the sense that it is independent of $\mu \in (0, 1)$ as well as on the scale factor a.

The case $1 < \mu < 2$ Unlike in the previous case, one finds that the first derivative of $J_\mu(w)$ at $w = 0$ is finite when $\mu \in [0, 2]$ and is given by

$$\alpha_\mu = J'_\mu(0) = -\frac{a\,\mu}{\pi} \int_0^\infty \frac{z^{\mu-2}\,dz}{1+z^\mu}. \tag{11.101}$$

Note that for $1 < \mu < 2$, the integral in Eq. (11.101) is convergent as $z \to 0$. Thus, as $w \to 0$, $J_\mu(w) \to 1 - \alpha_\mu\, w$. To obtain the leading non-analytic singular term, we need to compute the next term. By taking two derivatives with respect to w near $w = 0$ and then re-integrating back, we find the following leading singular behavior of $J_\mu(w)$ near $w = 0$

$$J_\mu(w) \approx 1 - \alpha_\mu\, w + c_\mu\, w^\mu + \ldots \quad \text{where} \quad c_\mu = \frac{2a^\mu}{\pi(\mu-1)} \int_0^\infty \frac{u^\mu\, du}{(1+u^2)^2}. \tag{11.102}$$

Substituting this small w behavior of $J_\mu(w)$ on the rhs of Eq. (11.89) and matching powers of w on both sides we get

$$\tilde{g}_2(\rho) \underset{\rho\to 0}{\sim} \frac{1}{\sqrt{\pi}} - \frac{\alpha_\mu}{\Gamma(1/2+1/\mu)}\,\rho + \frac{2}{\sqrt{\pi}}\, c_\mu\, \rho^\mu \tag{11.103}$$

where α_μ and c_μ are given respectively in Eqs. (11.101) and (11.102). Again, inverting via the Tauberian theorem (see Ref. [54]), we get

$$g_2(z) \underset{z\to\infty}{\sim} \frac{A_\mu}{z^{1+\mu}}\ , \tag{11.104}$$

with the amplitude

$$A_\mu = \frac{2}{\sqrt{\pi}}\, \frac{\mu(\mu-1)c_\mu}{\Gamma(2-\mu)} = \frac{2\mu}{\sqrt{\pi}}\,\beta_\mu \quad \text{where} \quad \beta_\mu = \frac{2a^\mu}{\pi\Gamma(2-\mu)} \int_0^\infty \frac{u^\mu}{(1+u^2)^2}\, du\,. \tag{11.105}$$

Exactly in a similar way, we substitute the small w behavior of $J_\mu(w)$ on the rhs of Eq. (11.90), match powers of w on both sides and find that

$$\tilde{h}_2(\rho) \underset{\rho\to 0}{\sim} \frac{1}{\rho} - \frac{\alpha_\mu}{\Gamma(1+2/\mu)} + c_\mu\, \rho^{\mu-1} \tag{11.106}$$

where α_μ and c_μ are defined in Eqs. (11.101) and (11.102). Inverting via the Tauberian theorem gives the desired result

$$h_2(z) \underset{z\to\infty}{\sim} 1 - \frac{B_\mu}{z^\mu} \tag{11.107}$$

with the amplitude

$$B_\mu = \frac{(\mu-1)c_\mu}{\Gamma(2-\mu)} = \beta_\mu \tag{11.108}$$

where β_μ is given in Eq. (11.105).

In this case, also we note that the ratio

$$\frac{A_\mu}{\mu\, B_\mu} = \frac{2}{\sqrt{\pi}} \tag{11.109}$$

is universal and does not depend explicitly on $1 < \mu < 2$ and a.

The case $\mu = 1$ In this case, from Eq. (11.85)

$$J_1(w) = \exp[-I_1(w)] \quad \text{where} \quad I_1(w) = \frac{1}{\pi}\int_0^\infty \frac{du}{1+u^2}\,\ln(1+a\,w\,u)\ . \tag{11.110}$$

Let us first derive the leading singular behavior of $I_1(w)$ as $w \to 0$. Making a change of variable $x = a\,w\,u$ in the integral we get

$$I_1(w) = \frac{aw}{\pi}\int_0^\infty \frac{dx}{x^2+a^2w^2}\,\ln(1+x)\ . \tag{11.111}$$

Next, we divide the integration range $[0,\infty)$ into two parts $[0,1]$ and $[1,\infty)$ and write $I_1(w) = Z_1(w)+Z_2(w)$. The second part $Z_2(w)$, i.e., the integral over $[1,\infty]$ is a completely analytic function as $w \to 0$. Thus the leading singular behavior of $I_1(w)$ as $w \to 0$ is contained only in the first part

$$Z_1(w) = \frac{aw}{\pi}\int_0^1 \frac{dx}{x^2+a^2w^2}\,\ln(1+x)\ . \tag{11.112}$$

In this integral, we can now safely expand $\ln(1+x) = x - x^2/ + x^3/3 + \ldots$ and perform the integral term by term. The leading singularity comes from the first term of this expansion

$$Z_1(w) \approx \frac{aw}{\pi}\int_0^1 \frac{x}{x^2+a^2w^2}\,dx = \frac{aw}{\pi}\ln\left[\frac{\sqrt{1+a^2w^2}}{aw}\right] \underset{w\to 0}{\sim} -\frac{a}{\pi}\,w\,\ln w \tag{11.113}$$

which indicates, from Eq. (11.110), that

$$J_1(w) \underset{w\to 0}{\sim} 1 + \frac{a}{\pi}\,w\,\ln w\ . \tag{11.114}$$

Substituting this small w behavior of $J_1(w)$ on the rhs of Eq. (11.89) and matching the leading behavior of w on both sides indicates that

$$\tilde{g}_2(\rho) \underset{\rho\to 0}{\sim} \frac{1}{\sqrt{\pi}} + \frac{2}{\sqrt{\pi}}\,\frac{a}{\pi}\,\rho\,\ln\rho\,. \tag{11.115}$$

This indicates, using Tauberian inversion theorem (see Ref. [54]),

$$g_2(z) \underset{z\to\infty}{\sim} \frac{A_1}{z^2} \quad \text{where} \quad A_1 = \frac{2}{\sqrt{\pi}}\,\frac{a}{\pi}. \tag{11.116}$$

Similarly, substituting the small w behavior of $J_\mu(w)$ on the rhs of Eq. (11.90) and matching leading behavior of w on both sides we get

$$\tilde{h}_2(\rho) \underset{\rho\to 0}{\sim} \frac{1}{\rho} + \frac{a}{\pi}\,\ln\rho \tag{11.117}$$

which, when inverted, provides the following large z behavior

$$h_2(z) \underset{z\to\infty}{\sim} 1 - \frac{B_1}{z} \quad \text{where} \quad B_1 = \frac{a}{\pi}\ . \tag{11.118}$$

Finally, we notice that even for this marginal $\mu = 1$ case, the ratio $A_1/B_1 = 2/\sqrt{\pi}$ has the same value as in the other two cases, namely for $0 < \mu < 1$ and $1 < \mu < 2$.

Let us remark that if one puts $\mu = 1$ in the expression of β_μ in Eq. (11.37), we get $\beta_1 = a/\pi$. Correspondingly $A_1 = 2a/\pi^{3/2}$ from Eq. (11.34) and $B_1 = a/\pi$ from Eq. (11.35), we find that they are consistent respectively with A_1 in (11.116) and and B_1 in (11.118). In other words, the final asymptotic results for $g_2(z)$ and $h_2(z)$ in the marginal case $\mu = 1$ are included in the range $\mu \in [1,2]$, even though the details for $\mu = 1$ are quite different, as it has logarithmic singularities.

APPENDIX III - Distribution of the maximum of a lattice random walk

In this appendix we consider N lattice random walks (RW) starting at $x_i(0) = 0$, for $i = 1, 2, \cdots, N$ and evolving as

$$x_i(m) = x_i(m-1) + \eta_i(m) \; , \tag{11.119}$$

where the noise $\eta_i(m)$'s are i.i.d. random variables with a distribution $f(\eta) = \frac{1}{2}\delta(\eta - 1) + \frac{1}{2}\delta(\eta + 1)$. The aim is to show the result in Eq. (11.55), taking advantage of the relation (11.54).

We first consider a single random walk, $N = 1$, and denote by $W(j, n)$ the number of lattice RW starting at $x_1(0) = 0$ and ending in j after n steps. One has

$$W(j,n) = \left\{ \begin{array}{l} \binom{n}{k} \; , \; 2k = n + j \; , \; n + j \text{ even} \; , \\ 0 \; , \; n + j \text{ odd} \; . \end{array} \right. \tag{11.120}$$

To compute the cumulative distribution function (cdf) of the maximal displacement of N walkers we need to compute the number of walks, for a single walker $N = 1$, which stay strictly below a given value M. We thus denote, for $N = 1$, by $W_M(j, n)$ the number of walks which stays strictly below an integer M and end up in j after n steps. To do this we use the reflection principle, e. g. the method of images: $W_M(j, n)$ can be obtained by subtracting to $W(j, n)$ the number of free walks which start in $x(0) = 2M$ and end in j after n steps. This yields:

$$W_M(j,n) = \left\{ \begin{array}{l} \binom{n}{k} - \binom{n}{k-m} \; , \; 2k = n + j \; , \; n + j \text{ even} \; , \\ 0 \; , \; n + j \text{ odd} \; . \end{array} \right. \tag{11.121}$$

The total number of walks $W_M(n)$ which start at $x_1(0) = 0$ and stay strictly below M after n steps are obtained by summing $W(j, n)$ in Eq. (11.121) over the endpoint $j < M$. This yields

$$W_M(n) = \sum_{k=0}^{\lfloor \frac{n+M}{2} \rfloor} \left[\binom{n}{k} - \binom{n}{k-M} \right] \; , \tag{11.122}$$

where $\lfloor x \rfloor$ is the largest integer not greater than x. Therefore one has

$$\begin{aligned} \text{Proba.} \left[\max_{0 \le m \le n} x_1(m) < M \right] &= \frac{W_M(n)}{2^n} \\ &= \frac{1}{2^n} \sum_{k=0}^{\lfloor \frac{n+M}{2} \rfloor} \left[\binom{n}{k} - \binom{n}{k-M} \right] \; . \end{aligned}$$

We can now write the cdf of the maximal displacement of N independent walkers as

$$\begin{aligned} \text{Proba.} \left[\max_{0 \le m \le n} x_{\max}(m) < M \right] &= \left(\frac{W_M(n)}{2^n} \right)^N \\ &= \left(\frac{1}{2^n} \sum_{k=0}^{\lfloor \frac{n+M}{2} \rfloor} \left[\binom{n}{k} - \binom{n}{k-M} \right] \right)^N \; , \end{aligned}$$

where $x_{\max}(m) = \max_{1 \le i \le N} x_i(m)$, from which one gets

$$\text{Proba.} \left[\max_{0 \le m \le n} x_{\max}(m) = M \right] = \frac{1}{2^{nN}} \left([W_{M+1}(n)]^N - [W_M(n)]^N \right) \; .$$

Finally, using the identity (11.54), one obtains the result given in the text in Eq. (11.55).

Bibliography

[1] D. Gembris, J. G. Taylor, and D. Suter, Nature **417**, 506 (2002).

[2] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).

[3] J. Krug and K. Jain, Physica A **358**, 1 (2005).

[4] J. Krug, J. Stat. Mech.: Theor. Exp. **07**, 07001 (2007).

[5] L. P. Oliveira *et al.*, Phys. Rev. B **71**, 104526 (2005).

[6] P. Sibani, G. F. Rodriguez, and G. G. Kenning, Phys. Rev. B **74**, 224407 (2006).

[7] M. Bauer, C. Godreche, and J. Luck, J. Stat. Phys. **96**, 963 (1999).

[8] S. Redner and M. R. Petersen, Phys. Rev. E **74**, 061114 (2006).

[9] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).

[10] G. Wergen and J. Krug, EPL **92**, 30008 (2010).

[11] A. Anderson and A. Kostinski, J. Appl. Meteo. and Climat. **50**, 1859 (2011).

[12] F. G. Foster and A. Stuart, J. Roy. Stat. Soc. **16**, 1 (1954).

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

[14] V. B. Nevzorov, *Records: Mathematical Theory* (American Mathematical Society, 2004).

[15] R. Ballerini and S. Resnick, J. Appl. Probab. **22**, 487 (1985).

[16] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).

[17] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011).

[18] J. Franke, G. Wergen, and J. Krug, Phys. Rev. Lett. **108**, 064101 (2012).

[19] G. H. Weiss, *Aspects and applications of the random walk* (North-Holland, 1994).

[20] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, 2001).

[21] S. N. Majumdar, Physica A **389**, 4299 (2010).

[22] E. G. C. Jr. and P. W. Shor, Algorithmica **9**, 253 (1993).

[23] E. G. Coffman *et al.*, Probab. Eng. Inform. Sc. **12**, 373 (1998).

[24] A. Comtet and S. N. Majumdar, J. Stat. Mech.: Theor. Exp. **P06013** (2005).

[25] S. N. Majumdar, A. Comtet, and R. M. Ziff, J. Stat. Phys. **122**, 833 (2006).

[26] J. Franke and S. N. Majumdar, J. Stat. Mech.: Theor. Exp. **P05024** (2012).

[27] N. R. Moloney, K. Ozogany, and Z. Racz, Phys. Rev. E **84**, 061101 (2011).

[28] S. N. Majumdar and G. Schehr, Phys. Rev. Lett. **108**, 040601 (2012).

[29] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).

[30] P. Le Doussal and K. J. Wiese, Phys. Rev. E **79**, 051105 (2009).

[31] S. Sabhapandit, EPL **94**, 20003 (2011).

[32] G. Wergen, M. Bogner, and J. Krug, Phys. Rev. B **83**, 051109 (2011).

[33] Y. Edery, A. Kostinski, and B. Berkowitz, Geophys. Res. Lett. **389**, L16403 (2011).

[34] E. Sparre Andersen, Math. Scand. **20**, 195 (1954).

[35] Thomson Reuters, “Thomson Datastream Advance 4.0 SP4,” (2003).

[36] V. V. Ivanov, Astron. Astrophys. **286**, 328 (1994).

[37] F. Pollaczek, Cr. Acad. Sci. I-Math. , 2334 (1952).

[38] F. Spitzer, Duke Math. J. **24**, 327 (1957).

[39] D. A. Darling, T. Am. Math. Soc. **83**, 164 (1956).

[40] M. Kwasnicki, J. Malecki, and M. Ryznar, arXiv:1103.0935 (2011).

[41] R. Garcia Garcia, A. Rosso, and G. Schehr, Phys. Rev. E **86**, 011101 (2012).

[42] R. M. Ziff, S. N. Majumdar, and A. Comtet, J. Phys.: Condens. Mat. **19**, 065102 (2007).

[43] R. M. Ziff, S. N. Majumdar, and A. Comtet, J. Chem. Phys. **130**, 204104 (2009).

[44] G. Zumofen and J. Klafter, Phys. Rev. E **51**, 2805 (1995).

[45] A. Zoia, A. Rosso, and S. N. Majumdar, Phys. Rev. Lett. **102**, 120602 (2009).

[46] E. J. Gumbel, *Statistics of Extremes* (Dover, 1958).

[47] L. Bachelier, Ann. Sci. Ecole Norm. S. **17**, 21 (1900).

[48] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000).

[49] J. Voit, *The Statistical Mechanics of Financial Markets* (Springer Berlin, 2001).

[50] X. Gabaix *et al.*, Nature **423**, 267 (2003).

[51] V. Plerou *et al.*, Phys. Rev. E **60**, 6519 (1999).

[52] G. Wergen, “Unpublished,” (2012).

[53] G. Schehr and S. N. Majumdar, J. Stat. Mech.: Theor. Exp. **P08005** (2010).

[54] M. R. Evans, S. N. Majumdar, and R. K. P. Zia, J. Stat. Phys. **123**, 357 (2006).

Part IV

Records in finance

12 Modeling record-breaking stock prices **205**
12.1 Introduction. 206
12.2 Records in i.i.d. RV's & symmetric random walks 207
12.3 Records in biased Gaussian random walks 209
12.4 Records in autoregressive processes 211
12.5 Analysis of single stocks 212
12.5.1 Data introduction 212
12.5.2 Jump distribution of the S&P 500 214
12.5.3 Record statistics of the increments 215
12.5.4 Record statistics of the stocks. 218
12.5.4.1 Correlations between stock records 219
12.5.4.2 Full distribution of the record number 220
12.5.5 First-passage times of the stocks. 222
12.6 Record statistics of N stocks 223
12.7 Weekly distributions of records 225
12.8 Summary and conclusions. 226

Chapter 12

Modeling record-breaking stock prices

Gregor Wergen

Institute for Theoretical Physics, University of Cologne

in preparation

Abstract: We study the statistics of record-breaking events in daily stock prices of 366 stocks from the Standard and Poors 500 stock market. Both the record events in the daily stock prices themselves and the records in the daily returns are discussed. In both cases we try to describe the record statistics of the stock data with simple theoretical models. The daily returns are compared to i.i.d. RV's and the stock prices are modeled using a biased random walk, for which the record statistics are known. These models agree partly with the behavior of the stock data, but we also identify several interesting deviations. Most importantly, the number of records in the stocks appears to be systematically decreased in comparison with the random walk model. Considering the slightly more complicated autoregressive AR(1) process, we can predict the record statistics of the daily stock prices more accurately. To better understand our findings, we discuss the survival and first-passage times of stock prices on certain intervals and analyze the correlations between the individual record events. After recapitulating some recent results for the record statistics of ensembles of N stocks, we also present some new observations for the weekly distributions of record events.

12.1 Introduction

Not only because of the recent financial crises, the study of extremes in stock markets is of great importance for scientists and economists [1–4]. Traders are eagerly interested in the statistics of extreme events in stock prizes at the worlds stock exchanges. In the context of extreme and dramatic developments in finance, people also talk a lot about record-breaking events. A record is an entry in a series of events that exceeds all previous values. In the context of stock markets, a record stock prize is often considered to be an important and remarkable event that attracts more attention and media coverage than others.

In recent years, the theory of records has found many applications in various areas of science. Most extensively studied was the statistics of record temperatures and their connection with global warming [5–10]. In 2010, Wergen and Krug [7] presented a simple analytical model that predicts the effect of climatic change on the occurrence of daily and monthly temperature records to a good accuracy (see also [9, 10]). Furthermore, the statistics of records found applications in evolutionary biology [11], physics [12–14], hydrology [15] and of course also in sports [16, 17]. Additionally, a lot of progress was made from the mathematical point of view. Often motivated by the multitude of applications, the theory of records from time-dependent and correlated random variables was developed further. The interesting problem of the record statistics of independent random variables with a linear drift is well understood by now [18, 19]. Also the ramifications involved with discrete distributions and ties due to rounding were studied [20, 21]. But most important for this work was the full characterization of the universal record statistics of symmetric random walks by Majumdar and Ziff [22] in 2008.

Recently, we started analyzing the statistics of record-breaking stock prices. Preliminary results were already published in 2011 in the context of a study of the record statistics of biased random walks [23] (see also [24]) and, in 2012, in an analysis of ensembles of independent random walkers [25]. The purpose of this article is to present a more thorough and comprehensive discussion of record-breaking events in stock prices and returns.

As in [23], we study record events in daily stock data from the Standard and Poors 500 (S&P 500) stock index [26]. Our data set contains 5000 consecutive trading days from 366 stocks that stayed in the S&P 500 for the entire time-span from January 1, 1990 to March 31, 2009. We are interested both in the record events in the time series of the stocks themselves and in record-breaking daily returns. For a series of stock prices $S_0, S_1, ..., S_n$, we have a record at the nth day if

$$S_n > \max\{S_0, S_1, ..., S_{n-1}\}. \tag{12.1}$$

Analogously, we have a record breaking return $\Delta_n := S_n - S_{n-1}$, if

$$\Delta_n > \max\{\Delta_0, \Delta_1 ..., \Delta_{n-1}\}, \tag{12.2}$$

where we set $\Delta_0 = 0$. When we consider the time series of stock prices S_i or returns Δ_i, the most important quantity for us, is the probability that a certain entry in such a series is a record. This probability P_n for a stock price S_n is defined as follows:

$$P_n := \text{Prob}\left[S_n > \max\{S_0, S_1 ..., S_{n-1}\}\right]. \tag{12.3}$$

For the returns Δ_i we define the probability p_n in the same manner:

$$p_n := \text{Prob}\left[\Delta_n > \max\{\Delta_0, \Delta_1, ..., \Delta_{n-1}\}\right]. \tag{12.4}$$

In the following, we will also refer to these quantities as the *record rates*. Of similar importance are the closely related record numbers R_n and r_n, the numbers of records that occur in a time series up to step n. One obtains the important mean record numbers $\langle R_n \rangle$ and $\langle r_n \rangle$ of the stocks and the returns by summing over the record rates:

$$\langle R_n \rangle := \sum_{k=0}^{n} P_k \qquad \text{and} \qquad \langle r_n \rangle := \sum_{k=0}^{n} p_k. \tag{12.5}$$

This article discusses the record rates and record numbers of the stock prizes and returns in the S&P 500 and compares them with several simple stochastic models such as simple i.i.d. RV's or biased random walks. The aim of this work is to better understand the occurrence of record-breaking events in the stock markets and to find useful and accurate models that reproduce and predict them correctly.

Since this work summarizes multiple observations and results, we will now give a short outline of the rest of this article: We start by briefly recapitulating some important classical results from the theory of records in time series of independent and identically distributed (i.i.d.) random variables (RV's) in section 12.2. There, we also present the findings for the symmetric random walk derived by Majumdar and Ziff [22]. In section 12.3, we discuss the record statistics of biased random walks with a Gaussian jump distribution following the results derived in Wergen et al. [23] and Majumdar et al. [27]. Subsequently, in section 12.4, we introduce the more complicated, so-called autoregressive AR(1) process, which might be able to describe the statistics of record-breaking stock prices more accurately. In recent years, this model and its continuous analog, the Ornstein-Uhlenbeck process, was used to model stock data by several research groups [28, 29]. We analyze and discuss its record statistics numerically.

Section 12.5 is about the record statistics of individual stocks from the S&P 500 index. After introducing the data and analyzing some of the basic statistical properties of the time series of stock prices, we first have a look at the record statistics of the daily returns. We compute the record rate and the mean record number of the daily returns and discuss the correlations between individual return records. Then, the record statistics of the stock prices themselves are analyzed. In this context, we also consider the full distribution of the record number of stock prices and compare this distribution with theoretical predictions. We will briefly discuss the first-passage statistics and survival probabilities of the stocks, since they are closely related to the statistics of records in random walks.

In section 12.6, we consider ensembles of stocks and compare their record statistics to the one of multiple independent random walks. The results in this section have mostly been already published in [25], here, we present them also to make this work more self-contained.

Some new findings for the weekly distribution of the record rate and the record number of the daily stock prices are presented in section 12.7. These observations are partly in contradiction with the assumptions and findings discussed before and show that simple, random walk type models can only predict certain features of the stock behavior, while others are not captured.

Finally, in section 12.8, we will briefly summarize and evaluate our findings and discuss possibilities for future research.

12.2 Records in i.i.d. RV's & symmetric random walks

The classical theory of records from i.i.d. RV's has been developed several decades ago. A detailed introduction can for instance be found in the books of Arnold et al. [30] or Nevzorov [31]. For a series of i.i.d. RV's $\xi_1, ...\xi_n$ sampled from a single probability density $f(\xi)$, we can easily give the probability for a certain entry in this series to be a record. Using the so-called *stick-shuffling* argument one finds that, on average, the nth entry is a new record in $1/n$ cases. Therefore, the record rate of i.i.d. RV's is given by $p_n^{(\mathrm{iid})} = \frac{1}{n}$. With this, it is also straightforward to compute the mean record number $\langle r_n \rangle$. By summing over the

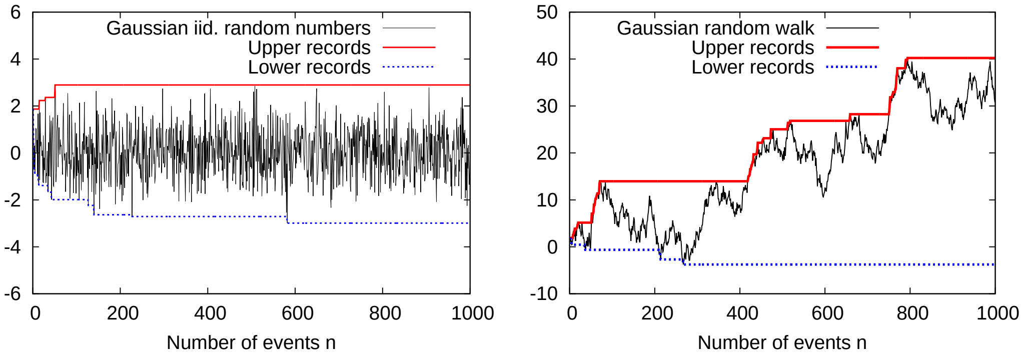

Figure 12.1: Left figure: 1000 numerically generated random numbers from a Gaussian standard normal distribution (standard deviation $\sigma = 1$). The red line gives the progression of the upper record, the blue line the progression of the lower record. **Right figure:** A numerical realization of a random walk with 1000 steps. The jump distribution is again a Gaussian standard normal distribution with mean value zero and standard deviation unity. Again, the red and blue lines give the progressions of the upper and lower record values.

record rate we find that, for $n \gg 1$, $\langle r_n \rangle$ grows logarithmically in n:

$$\langle r_n \rangle = \sum_{k=1}^{n} p_k = \sum_{k=1}^{n} \frac{1}{k} \approx \ln(n) + \gamma. \tag{12.6}$$

Here, $\gamma \approx 0.577215...$ is the Euler-Mascheroni constant [30]. Interestingly, these results are completely independent from the choice of the underlying distribution $f(\xi)$.

In 2008, Majumdar and Ziff [22] found that the record statistics of a discrete-time random walk with a symmetric jump distribution has similar universal properties. They considered random walks with entries $X_0, X_1, ..., X_n$ given by

$$X_i = X_{i-1} + \xi_i, \tag{12.7}$$

with i.i.d. RV's ξ_i sampled from a symmetric and continuous jump distribution $f(\xi)$. As an initial value they set $X_0 = 0$. The different characteristics of the record processes of i.i.d. RV's and random walks are illustrated in Fig. 12.1. Majumdar and Ziff computed the full distribution of the record number R_n of such a process.

They obtained the probability of having R_n records in a random walk of n steps by subdividing the process into a series of R_n first-passage and survival problems (cf. [32]). Those were then solved using a celebrated theorem by Sparre Andersen [33, 34]. Majumdar and Ziff showed that the probability P_n for a record in the nth event of a symmetric random walk is the same as the survival probability Q_n, i.e. the probability that a random walk starting from the origin stays above the origin without crossing it for the next n steps:

$$Q_n := \text{Prob}\left[X_1, X_2, ..., X_n > 0\right]. \tag{12.8}$$

According to Sparre Andersen [22, 33, 34], this probability and therefore also the record rate of the symmetric random walk is universal for symmetric random walks with a continuous jump distribution and given by

$$Q_n = P_n = \binom{2n}{n} 2^{-2n}. \tag{12.9}$$

Using Stirling's formula [35], one finds that, for $n \gg 1$, the record rate P_n decays as

$$P_n \approx \frac{1}{\sqrt{\pi n}}. \tag{12.10}$$

This allows to compute the mean record number of the symmetric random walk:

$$\langle R_n \rangle = n \binom{2n}{n} 2^{-2n+1} \approx \sqrt{\frac{4n}{\pi}}. \tag{12.11}$$

Here, the mean record number grows with $\sqrt{n}$, much faster than in the case of i.i.d. RV's. The most important and surprising feature of these results is that, for arbitrary n, they are again completely independent from the choice of the symmetric and continuous jump distribution $f(\xi)$. They also hold for the so-called Lévy flights, i.e. random walks with a jump distribution that do not have a finite second moment.

12.3 Records in biased Gaussian random walks

More important for an application to financial data are random walks with a bias. In the contexts of stocks, such a bias represents an inherent growth in the system, a long term interest-rate or economic growth. The entries $X_0, X_1, ..., X_n$ of a discrete-time random walk with bias c can be described as

$$X_i = X_{i-1} + \xi_i + c, \tag{12.12}$$

where the ξ_i's are again sampled from a symmetric and continuous distribution $f(\xi)$. Unfortunately, the full universality, which was found for the symmetric random walk is not conserved here. In the biased case, the record statistics depends on the shape of the jump distribution.

In a recent publication of Majumdar et al. [27], the asymptotic record statistics of such biased random walks were studied thoroughly. The authors used a generalized version of Sparre Andersen's theorem to compute the survival probabilities of biased random walks for different classes of jump distributions. With their findings, they could derive asymptotically exact results for the distribution of the record number, as well as the extreme value statistics of the ages of the shortest and longest lasting records.

The most relevant parameter for the asymptotic survival and record statistics of a biased random walk is the so-called Lévy-index of the jump distribution $f(\xi)$. Majumdar et al. considered jump distributions, whose Fourier transforms $\tilde{f}(k) := \int_{-\infty}^{\infty} f(\xi) e^{-ik\xi} \mathrm{d}\xi$ have the following small k behavior:

$$\tilde{f}(k) \approx 1 - (\alpha_\mu |k|)^\mu . \tag{12.13}$$

Here, μ is called the Lévy-index and α_μ is a parameter, that represents a characteristic length scale of the jump distribution. The Lévy index describes the tail-behavior of the jump distribution ($|\xi| \gg 1$). For distributions with a finite second moment, like a Gaussian, an exponential or a uniform distribution, one always finds a Lévy-index of $\mu = 2$. For $n \to \infty$, these jump distributions lead to a random walk that behaves like classical Brownian motion with a mean square displacement $\langle \sum_{k=1}^{n} X_k^2 \rangle \propto n$ (cf. [36]).

More complicated is the regime of $0 < \mu < 2$, where the jump density has no finite variance. In this case, one finds that, for large $|\xi| \to \infty$, the real-space representation of the jump distribution is of the form $f(\xi) \propto |\xi|^{-1-\mu}$ and decays with a power law broader than $1/|\xi|^3$. These jump distributions are called *heavy-tailed*, random walks with heavy-tailed jumps are called Lévy flights.

Majumdar et al. found that, for extremely heavy-tailed distributions with $\mu < 1$, the drift does not change the asymptotic behavior of the mean record number $\langle R_n \rangle$. Here, we have $\langle R_n \rangle \propto \sqrt{n}$ independent of c. For the marginal case of $\mu = 1$, which includes the well known Cauchy distribution (c.f. [35]), they found that the mean record number depends non-trivially on c. In this regime one can show that $\langle R_n \rangle \propto n^{\Theta(c)}$ with $\Theta(c) = 1/2 + 1/\pi \arctan(c)$.

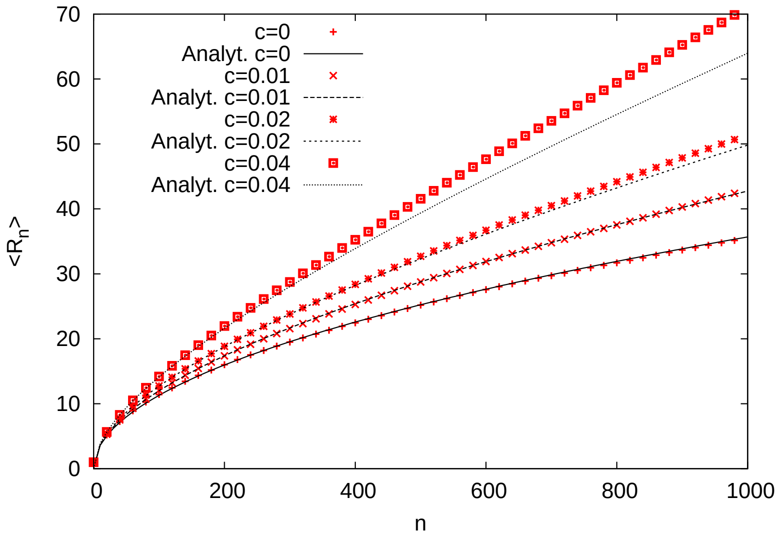

Figure 12.2: Numerical simulations of the mean record number $\langle R_n \rangle$ of Gaussian random walks with different drift rates $c = 0, 0.01, 0.02$ and 0.04. For each c we averaged over 10^4 realizations. The lines are our analytical approximations for $\langle R_n \rangle$ in the regime of $c\sqrt{n} \ll \sigma$. For small c the analytical result is in good agreement with the simulations.

More interesting for us are the distributions, which are less heavy-tailed than the Cauchy distribution, with $\mu > 1$. In this regime, for $c > 0$, the mean record number grows linearly in n and one finds $\langle R_n \rangle \approx \alpha_\mu (c) n$ with $\alpha_\mu (c)$ depending on the exact shape of $f(\xi)$. This asymptotic behavior for the mean record number was also found for the Brownian case with $\mu = 2$ and $c > 0$. However, the asymptotic distributions of the record number and the statistics of the extremal ages of the longest and shortest lasting record are systematically different between the $1 < \mu < 2$-regime and the the case of $\mu = 2$ with a positive c.

In the context of our interest in stock prices, the Brownian regime with $\mu = 2$ will prove to be particularly important. For the special case of a Gaussian jump distribution with probability density

$$f(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\xi^2}{2\sigma^2}}, \tag{12.14}$$

Majumdar et al. derived an explicit expression for the prefactor $\alpha_2(c)$. Here, for $n \to \infty$, they found that

$$\langle R_n \rangle \approx n \exp\left[-\sum_{k=1}^{\infty} \frac{1}{2k} \mathrm{erfc}\left(\frac{c\sqrt{k}}{\sigma\sqrt{2}} \right) \right]. \tag{12.15}$$

Another regime, which has proven to be useful for the analysis of stock prices, has been studied by Wergen et al. [23] in 2011. They considered the case of a small drift c and a finite n with $c\sqrt{n} \ll \sigma$ for random walks with a Gaussian jump distribution. Here, the generalized Sparre Andersen theorem for the survival probability is helpful, too. For $c\sqrt{n} \ll \sigma$, the survival probability Q_n of a Gaussian random walker is given by

$$Q_n \approx \binom{2n}{n} 2^{-2n} - \frac{c}{\sqrt{2}\sigma} \approx \frac{1}{\sqrt{\pi n}} - \frac{c}{\sqrt{2}\sigma}. \tag{12.16}$$

With this one finds that, for $c \ll \sigma/\sqrt{n}$,

$$\langle R_n \rangle \approx \frac{2\sqrt{n}}{\sqrt{\pi}} + \frac{c}{\sigma} \frac{\sqrt{2}}{\pi} \left(n \arctan\left(\sqrt{n}\right) - \sqrt{n} \right) \tag{12.17}$$

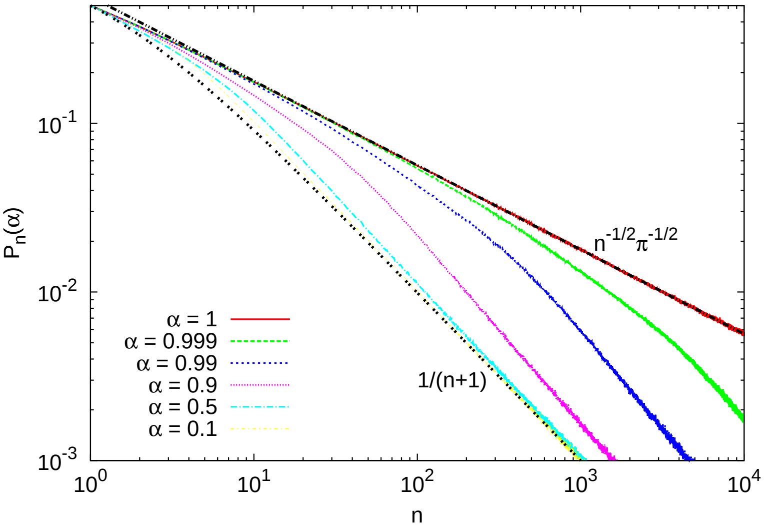

Figure 12.3: Numerical simulations of the record rate P_n for the AR1 process with different values of α and $n = 10^4$ steps. We averaged over 10^6 realizations for each value of α and simulated $\alpha = 1$ (the random walk case), $\alpha = 0.999$,$\alpha = 0.99$,$\alpha = 0.9$,$\alpha = 0.5$ and $\alpha = 0.1$. We also plotted our analytical results for the random walk and i.i.d. RV's (black dotted lines). For large α the process behaves like a random walk for a long time. For smaller α he converges to the i.i.d. behavior quite fast. Eventually we expect $P_n \propto 1/n$ for all $\alpha < 1$ in the $n \to \infty$ limit.

and

$$P_n \approx \frac{1}{\sqrt{\pi n}} + \frac{c}{\sigma}\frac{\sqrt{2}}{\pi}\text{arctan}\left(\sqrt{n}\right). \tag{12.18}$$

In Fig. 12.2, we compare Eq. 12.17 with numerical simulations of a biased Gaussian random walk. In [23], it is shown that these results also apply for other jump distributions with an existing second moment.

In the subsequent section 12.5 we will discuss these findings for biased random walks, especially the ones for the Gaussian process given in Eqs. 12.15 and 12.17, in the context of records in the evolution of stock prices. Prior to this, we will now introduce a more complicated type of process, which might be even more accurate in the modeling of stock records, and analyze its record statistics numerically.

12.4 Records in autoregressive processes

The fact that the record statistics of i.i.d. RV's as well as of symmetric random walks is well understood by now, and also our previous work on financial data has motivated us to consider another stochastic process, which, in some sense, intermediates between uncorrelated RV's and random walks. The natural way to interpolate between these two processes is a process of entries $X_0, X_1, ..., X_n$ with $X_0 = 0$ and

$$X_i = \alpha X_{i-1} + \xi_i, \tag{12.19}$$

and α being a parameter between zero and one. The ξ_i's are again i.i.d. RV's sampled from a continuous jump distribution $f(\xi)$. In the special case of $\alpha = 0$, this process is just a series of i.i.d. RV's with $X_1 = \xi_1, X_2 = \xi_2, ...$. For $\alpha = 1$ we recover the symmetric random walk (cf. Eq. 12.7). An important feature of such a process is that it is not invariant under translations and the distribution of the jumps $X_i - X_{i-1}$ depends on the value of X_{i-1}.

Therefore the distance of the walker from the origin is relevant. Such a model is also known as an AR(1) process.

Unfortunately, the methods introduced in the previous publications concerning the record statistics of random walks do not allow to compute the record statistics of the AR(1) process. Here, Sparre Andersen's theorem loses its validity and, by now, it was not possible to derive the record rate and the mean record number using an alternative analytical approach.

In Fig. 12.3 we show numerical results for the record rate of the AR(1) process for different values of α. Apparently, for small values of α, this process behaves almost like a series of i.i.d. RV's. If α is close to one the record statistics resembles the one of a symmetric random walk for a long time, but eventually the record rate decreases and approaches the $1/n$ behavior as well. These simulations allow us to conjecture the asymptotic behavior of the record rate of the AR(1) process. For $n \to \infty$, we assume that

$$P_n^{(\alpha)} \approx P_n^{(\mathrm{iid})} = \frac{1}{n} \tag{12.20}$$

In the opposite regime of finite n and $\alpha \approx 1$, we expect that the record statistics of the process differs only slightly from a random walk. The mean record number will grow roughly proportional to $\sqrt{n}$ as in the case of the symmetric random walk. As indicated by several studies of the AR(1) process [37–40] and our own numerical simulations, the survival probability of this process decays exponentially with n (and not with $\sqrt{n}$ as in the case of the symmetric random walk). Building upon the work of Novikov [40], we can make an educated guess for the record rate $P_n^{(\alpha)}$ and the mean record number $\langle R_n^{(\alpha)} \rangle$ for $(1-\alpha) \ll 1$:

$$\langle R_n^{(\alpha)} \rangle \approx \langle R_n^{(1)} \rangle e^{-\alpha B_n} \qquad \text{and} \qquad P_n^{(\alpha)} \approx P_n^{(1)} e^{-\alpha B_n} \tag{12.21}$$

with a positive parameter B_n depending on n. Interestingly, a very good agreement with numerical results was found for $B_n = \langle R_n^{(1)} \rangle$. Despite these observations, we will use numerical results for our analysis of the stock data. It would be an interesting and challenging goal for future research to compute the record statistics of the AR(1) process analytically.

12.5 Analysis of single stocks

12.5.1 Data introduction

As in [23], we consider a data set of daily prices of stocks contained in the Standard and Poors 500 (S&P 500) index [26]. The S&P 500 is an important stock market index and includes mostly U.S. companies, which are selected by a committee. With a market capitalization of more than 10^4 billion USD, the index is supposed to represent all relevant branches of the U.S. industry.

While the index is computed as a weighted average of the stock prices depending on how many shares of the stocks are publicly tradable, we will consider only the prices of the contained stocks themselves and not the score of the index. To further simplify our studies, we examine a set of daily closing prizes of 366 stocks that remained within the S&P 500 index for the entire time span from January 1990 to March 2009. This allowes us to analyze 366 time series with an identical length of $n = 5000$ trading days.

Probably the oldest model of stock prices is the so-called Geometric Random Walk Model (GRM) [41, 42], which was introduced already more than 100 years ago by Le Bachelier [43]. In this model the logarithms $s_n := \ln(S_n/S_0)$ of the stock prices perform a random walk with a constant bias:

$$s_n = s_{n-1} + \xi_n + c. \tag{12.22}$$

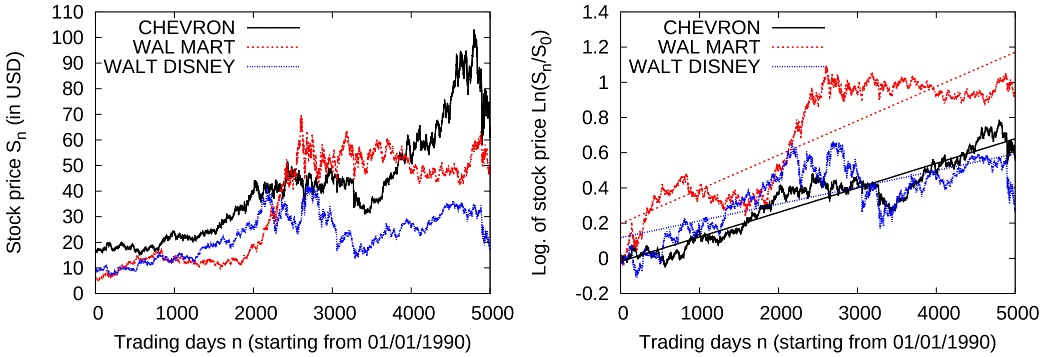

Figure 12.4: **Left:** The daily stock prices S_n (in Dollar) of three stocks (Chevron, Wal Mart and Disney) from the S&P 500 index. **Right:** The logarithms $s_n = \ln(S_n/S_0)$ for the same three stocks along with linear regressions for these logarithmic stock prices.

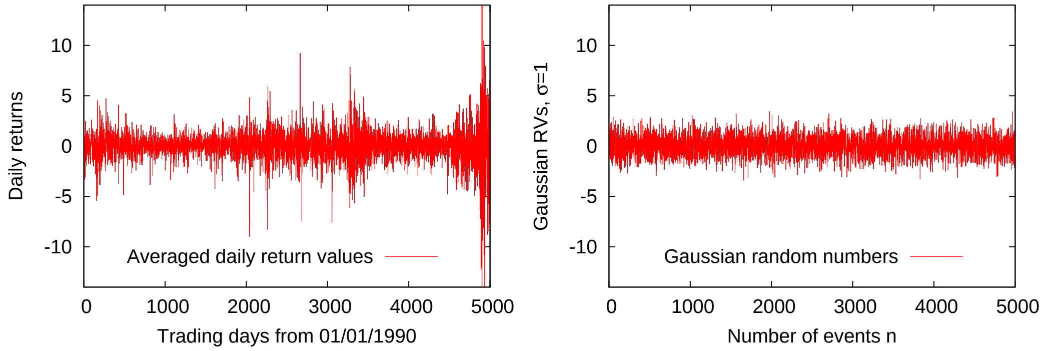

Figure 12.5: **Left:** The averaged and normalized daily returns $\delta_n = \ln S_n/S_{n-1}$ of the stocks in the S&P 500 index. **Right:** 5000 Gaussian random numbers with standard deviation unity.

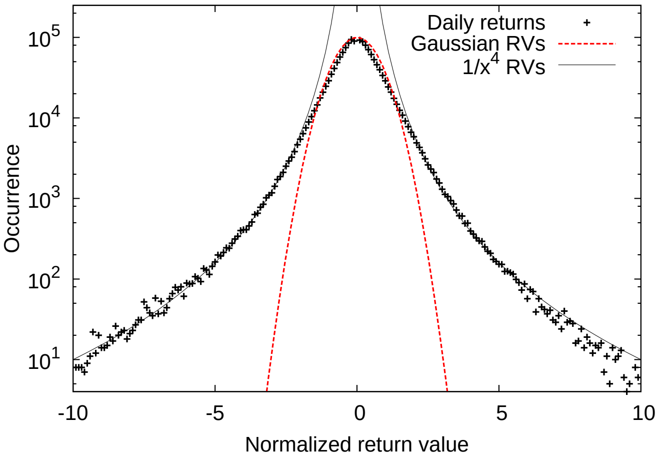

Figure 12.6: The distribution of the normalized and detrended logarithmic daily returns of the stocks in the S&P 500. We considered the daily returns $\tilde{\delta}_i = (\delta_i - c)/\sigma$ of the logarithmically detrended stocks, where the drift c and the standard deviation σ were obtained from the stock data. These normalized returns have standard deviation unity. The plot was obtained by binning the values into bins of size 0.01. For comparison, we also plotted a manually fitted Gaussian (red dashed line) with standard deviation $\sigma = 1$ and a symmetric Pareto distribution with $f(x) = 1/x^4$ (black line). Apparently, the daily returns of the detrended data are not Gaussian and are instead well described by the $1/x^4$-power-law.

This bias c is supposed to represent some kind of inherent growth like a long-term interest rate or an exponentially growing amount of money in the market. Fig. 12.4 illustrates this model with three randomly selected stocks from the S&P 500 index. It is important to notice that the occurrence of records in the stocks is not affected by the logarithm. Due to the monotony of the logarithm, we have

$$S_n > \max\{S_0, S_1, ..., S_{n-1}\} \quad \Leftrightarrow \quad s_n > \max\{0, s_1, ..., s_{n-1}\} \tag{12.23}$$

and therefore a record breaking stock prices S_n is also a record in the series of the logarithms $s_n = \ln(S_n/S_0)$. On the other hand, a record of the daily returns $\Delta_n = S_n - S_{n-1}$ is not necessarily a record of the logarithmic returns $\delta_n := s_n - s_{n-1}$. However, it is easy to show that if a logarithmic return δ_n is a record, this return is also a new record in the series of relative daily changes of the stock price S_n/S_{n-1}:

$$\frac{S_n}{S_{n-1}} > \max\{0, \frac{S_1}{S_0}, ..., \frac{S_{n-1}}{S_{n-2}}\} \quad \Leftrightarrow \quad \delta_n > \max\{0, \delta_1, ..., \delta_{n-1}\}. \tag{12.24}$$

Since these relative return records are usually more interesting in a growing stock market, we will consider them in the following. In the context of the GRM, the logarithmic daily returns δ_n should be i.i.d. RV's sampled from a symmetric distribution plus a constant linear drift c ($\delta_n \equiv \xi_n + c$).

For some applications, we will also detrend the logarithmic stocks, i.e. a linear trend is subtracted from the logarithms $s_n = \ln S_n/S_0$ to obtain stock prices more comparable to symmetric random walks. Of course, in this case, the record statistics of the stocks are altered by removing the trend. A linear regression analysis of the individual stocks for the entire time-span of 5000 trading days gives an average normalized drift of $c/\sigma \approx 0.025$, where σ is the standard deviation of the distribution of the returns δ_n.

12.5.2 Jump distribution of the S&P 500

Today, it is well known that this simple geometric random walk model does not represent a complete and accurate model of the stock markets. In fact, it is particularly useless in times of high market activity and during crashes since it assumes that the daily changes in a stock prize are random and uncorrelated. Actual stocks are of course correlated with other stocks and, in addition, usually non-Markovian [41, 42].

In Fig. 12.5, we compare the 5000 averaged and normalized logarithmic daily returns δ_i/σ (where σ is the standard deviation of the return distribution) of the S&P 500 with 5000 computer-generated Gaussian random numbers with standard deviation unity. While the pattern of the Gaussian RV's is homogeneous, the amplitudes of the stock returns fluctuate over time. Despite these findings, many statistical properties of stock prices can be modeled and understood using the GRM and, as it was already shown in previous studies, it is also, to some degree, useful to model the record statistics of stocks in the S&P 500.

Since, in the context of the record statistics of biased random walks, the shape of the jump distribution is of importance, we measured the probability density of the logarithmic daily returns δ_n. In Fig. 12.6, the probability density of the these returns, after linearly detrending the logarithms of the stock prizes, is plotted. Apparently, while the daily returns close to zero are approximately normally distributed, their tails are highly non-Gaussian and decay like a power-law with $1/x^4$. In spite of that, this return distribution still has a finite second moment. The corresponding Lévy index to the discovered power-law is $\mu = 3$ and is much larger than the critical value of $\mu = 2$ (see section 12.3). A random walk with such a jump distribution is not a Lévy flight, we will henceforward compare the record statistics of the stocks with *regular* random walks with a jump distribution that has a finite variance.

Knowing that the return distributions of stock prizes on shorter time-scales, such as minutes or seconds, are much broader than the distributions of daily returns [41, 42], the results of Majumdar et al. [27] for Lévy-indices $\mu < 2$ might still be useful in future studies of the record statistics of stock prices with a higher temporal resolution.

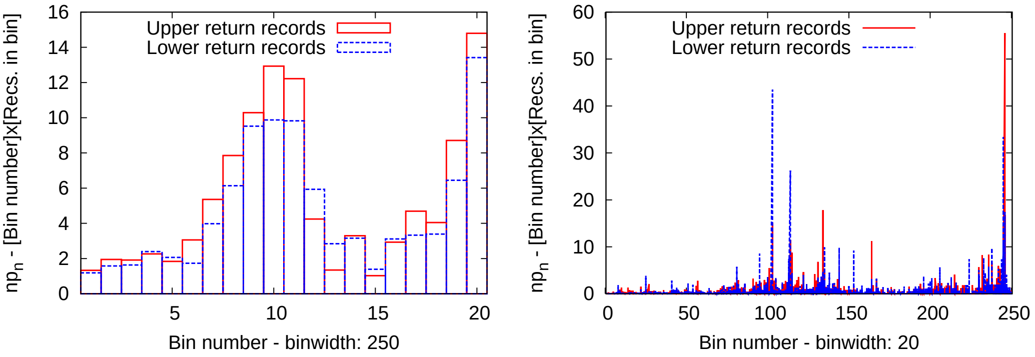

Figure 12.7: Left: Binned normalized record rate of the daily returns $\delta_n = \ln S_n/S_{n-1}$ in the S&P 500. We computed the normalized record rate np_n in bins of 250 trading days length. For each bin, we counted the number of upper and lower records of δ_n and multiplied it with the number of the bin. As discussed in the main text, the upper and lower record rates are highly correlated. **Right:** The same analysis but for 250 bins of 20 trading days length. Note the extremely high number of new upper and lower return records during the financial crisis in 2009.

12.5.3 Record statistics of the increments

We begin our study of the record statistics of stocks in the S&P 500 with an analysis of record-breaking daily returns. On the basis of the simple Geometric Random Walk Model (GRM), the returns $\delta_n = \ln S_N/S_{n-1}$ should have the same record statistics as i.i.d. RV's. Therefore, the record process of these increments gives an opportunity to check if the returns of the stock prices and i.i.d. RV's are comparable. If we find that the returns have the same record statistics as i.i.d. RV's this would indicate that they are also more or less uncorrelated and have no systematic time-dependence.

We analyzed the records in the logarithmic daily returns δ_n and computed the mean record number $\langle r_n \rangle$ in these time series. At the end of the entire period of 5000 trading days, we found an average number of 11.09 upper and 11.12 lower records. For i.i.d. RV's we would have expected only $\langle r_n^{(\mathrm{iid})} \rangle \approx 9.094 \pm 0.389$ records, so, at least on such a long time-scale, the record statistics seems to differ from the one of i.i.d. RV's. However, the actual error margins might be much larger since we assumed 366 independent stocks to compute them. It turns out that, for this long time-span of 5000 trading days, the record statistics of returns is dominated by fluctuations and a significant amount of the records that contribute to the mean record number are set on a very small number of trading days.

In Fig. 12.7, we illustrate how the record rate of the returns varies over time. The figure shows a normalized version of the return record rate. We computed the rate p_n of new return records of δ_n for each trading day and multiplied this rate with the number of the trading day n. For i.i.d. RV's one expects $np_n^{(\mathrm{iid})} = 1$ for an arbitrary value of n and, therefore, we call np_n the normalized record rate. In Fig. 12.7, the distribution of np_n for the returns δ_n of the 366 stocks is binned in bins of 250 (left figure) and 20 (right figure) trading days length. The left part of Fig. 12.7 shows roughly the annual distribution of the normalized return record rate np_n. Both the upper and the lower record rate of δ_n were very high in the years between 1997 and 2002, as well as in 2007 and 2008. In some years the record rate was more than 10 times as high as expected on the basis of i.i.d. returns. Interestingly, the plot shows that the upper and lower record rates are highly correlated. These rates never differ by more than 20%, a result that is not found for i.i.d. RV's, where upper and lower records are uncorrelated.

In the right part of Fig. 12.7, we plotted np_n for the daily logarithmic returns δ_n with a smaller bin length of 20 trading days, which is roughly one calendar month. This figure shows how extremely the record rate of δ_n fluctuates over time. A significant amount of the return records was set in only a few trading days.

	Upper Recs.	Lower Recs.		Upper Recs.	Lower Recs
2040	0	**156**	4752	0	15
2041	**54**	4	4880	0	13
2260	0	**85**	4883	22	0
2266	37	0	4884	20	0
2293	28	0	4890	0	**50**
2663	49	0	4891	14	0
2684	0	27	4898	0	16
2873	0	25	4900	**83**	0
3055	25	0	4902	0	26
3277	0	15	4911	25	0

Table 12.1: The 20 trading days with the highest normalized (upper or lower) return record rate np_n of the daily returns δ_n together with the corresponding number of upper and lower records. The top 20 days are ordered chronologically. Days with 40 upper or lower return records were additionally highlighted. Note that 9 of these 20 trading days with the highest relative daily changes in the stock prices S_n occurred during only 31 trading days (in the fall of 2008). Another 7 occurred in only 3 years following trading day $n = 2040$ (1998-2000).

In Tab. 12.1, we list the 20 trading days with the highest normalized (positive or negative) record rate np_n along with the total number of (upper and lower) return records of δ_n for these trading days. The table shows how strongly the record statistics of the returns is affected by a few periods of very high market activity. 9 of these 20 *record*-days fall into a short time-period of only 31 trading days in the fall of 2008, in which the recent financial crisis had its climax.

The table also explains the high correlations between upper and lower records, which was found for the binned data of the normalized record rate np_n. While there are no trading days with both, a very high number of upper and a very high number of lower records, days with many lower records are often quickly succeeded by days with a large number of upper records and vice versa. In fact, of all upper return records in the entire data set, 9.77% are followed by a lower record in the next trading day. Similarly 9.33% of the lower records are immediately followed by an upper record. If one considers the five days following a record, these values double. 18.00% (18, 54%) of all upper (lower) records entail a lower (an upper) record within the next five trading days. For i.i.d. RV's, it is easy to show that the corresponding rates should be below 0.5% for the interval length of $n = 5000$ trading days.

Tab. 12.1 and Fig. 12.7 indicate that the record statistics of the returns δ_n might differ between times of relatively calm market activity and those of financial crisis. To reduce the effect of these events and to get more reliable statistics, we also analyzed the data set in shorter intervals.

We split the 5000 trading days into 20 consecutive intervals of each 250 trading days. We computed the mean record numbers of the δ_n's separately in each of these intervals and averaged over the results. Fig. 12.8 shows the progressions of the averaged upper and lower mean record number for these intervals and compares them with the i.i.d.-curve given by $\langle r_n^{(\text{iid})} \rangle = \ln n + \gamma$. Apparently, the mean upper record number agrees quite well with the i.i.d. result. The number of lower records is slightly increased.

We also subdivided the data in shorter intervals of each 100 trading days. For these intervals we have enough individual time series to plot the record rate p_n of the daily returns δ_n. Fig. 12.9 compares the upper and lower record rates in these intervals with the i.i.d. prediction of $p_n^{(\text{iid})} = 1/n$. Here, the curves for the stock returns are in good agreement with the i.i.d. behavior. For this interval length, also the behavior of the mean record number $\langle r_n \rangle$ (inset in Fig. 12.9) agrees almost perfectly with the i.i.d. result.

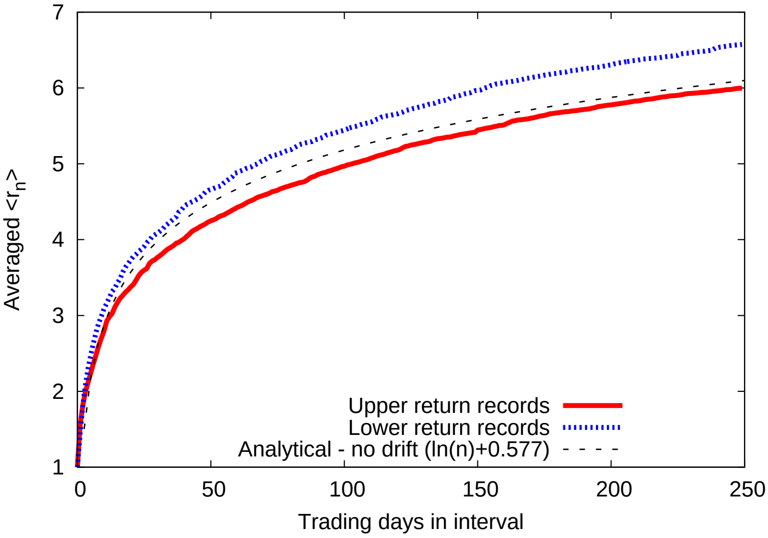

Figure 12.8: Averaged mean upper (red) and lower (blue) record number of the daily returns $\delta_n = \ln S_n/S_{n-1}$ in the S&P 500. The data set was split up into 20 consecutive intervals of each 250 trading days. We computed the evolution of the mean record number of the 366 stocks separately in each of the intervals before we averaged over all intervals. The black dashed line gives the theoretical result for the mean record number of i.i.d. RV's with $\langle r_n \rangle \approx \ln n + 0.577\ldots$.

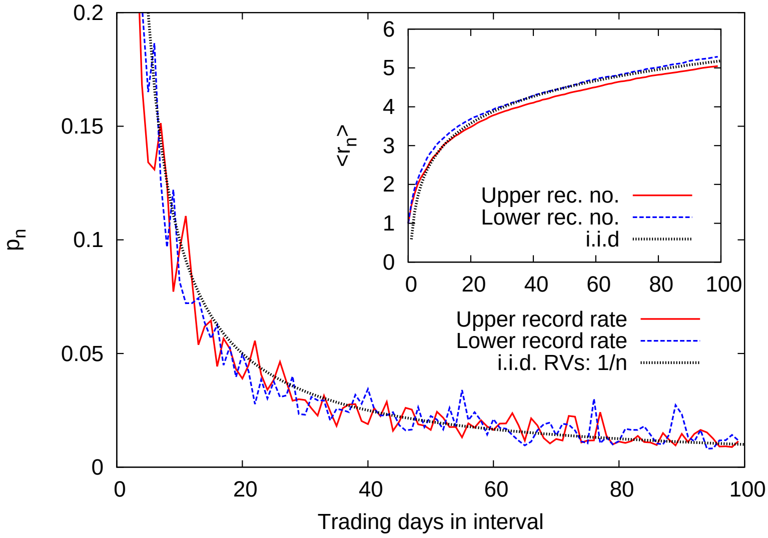

Figure 12.9: Averaged upper (red) and lower (blue) record rate of the daily returns $\ln S_n/S_{n-1}$ in the S&P 500. Here, the data was split up into 50 consecutive intervals of each 100 trading days. Again, we first computed the record rates for all 366 stocks in the individual intervals before we averaged over all intervals. The black dashed line gives the theoretical i.i.d. result with $P_n^{(\mathrm{iid})} = 1/n$. The inset shows the same analysis for the averaged mean record number.

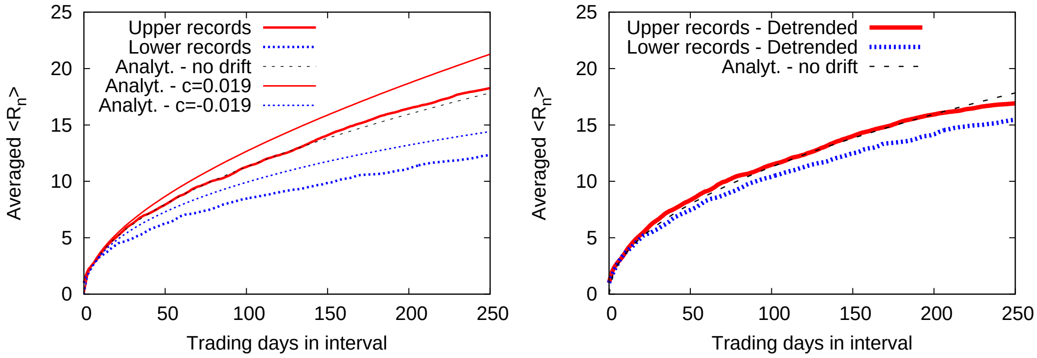

Figure 12.10: Left: Averaged upper (red) and lower (blue) mean record number of the daily stock prices S_n in the S&P 500. The data set was split up into 20 consecutive intervals of each 250 trading days. The records in the stock prices are compared to our analytical result for the biased random walk with a Gaussian jump distribution with standard deviation $\sigma = 1$ and a drift $c = 0.019$ (see Eq. 12.17). The drift was obtained by a linear regression analysis of the logarithmic stock prices $s_n = \ln S_n/S_{n-1}$. Both the mean record numbers and the values for the drift were first computed separately for each interval and then averaged over all intervals. **Right:** The same analysis but for detrended daily stock prices. We subtracted a linear trend from all individual stocks in the individual intervals before we computed and averaged the record numbers. The detrended data is compared to the analytical prediction for the unbiased random walk (black dashed line).

In summary, over longer time-spans of several years the record statistics of the logarithmic daily returns $\delta_n = \ln S_n/S_{n-1}$ of the stocks in the S&P 500 differs significantly from the behavior i.i.d. RV's. Nevertheless, this effect seems to be caused by a few short periods of high market activity, in which large numbers of stocks collectively set new records. On shorter time frames the daily returns are more similar to i.i.d. RV's, here both the record rate p_n and the mean record number $\langle r_n \rangle$ of the δ_n's are modeled accurately by the i.i.d. results. In the context of the GRM, we have therefore reason to believe that, on these shorter intervals, the record statistics of the logarithmic stock prices $s_n = \ln S_n/S_0$ is accurately modeled by a biased random walk.

12.5.4 Record statistics of the stocks

With these findings for the daily returns, we can now discuss the record statistics of the stock prices S_n themselves. In [23], we already presented an analysis of the mean record number of the undetrended stock prices in the S&P 500 for the full time-span of $n = 5000$ trading days and also an analysis of shorter intervals of length $n = 100$. While the Geometric Random Walk Model (GRM) yielded an accurate description for the long interval, we found a significant reduction of the number of lower records in the detrended data for shorter intervals.

In the context of the findings for the returns discussed above, the good agreement for the full time-span could be a coincidence. However, especially because of Figs. 12.8 and 12.9, the record statistics of the stocks on the shorter intervals should resemble the GRM. The asymmetry between upper and lower records on the interval of 100 trading days that was discovered by Wergen et al. [23] is not explained by this model.

A similar asymmetry is found when we consider the data set subdivided into intervals of each 250 trading days, which we also considered for the daily returns. Fig. 12.10 shows the averaged mean record number in the undetrended and detrended stock prices. In the left figure, we compare the mean record number of the undetrended data with the analytical predictions from the GRM of a biased random walk. The analytical results were computed for Gaussian jump distribution with standard deviation unity and a drift of $c = 0.019$. We computed the average drift c from the data and plotted the curve described by Eq. 12.17.

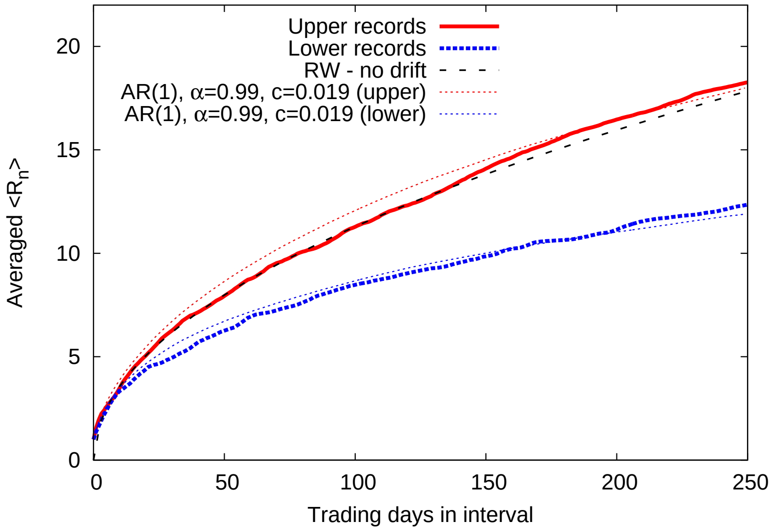

Figure 12.11: Averaged upper (red) and lower (blue) mean record number of the daily stock prices in the S&P 500. The data set was split up into 20 consecutive intervals of each 250 trading days. We compare the observational data with numerical simulations for the AR(1) process with $\alpha = 0.99$ and a bias of $c = 0.019$ (dotted lines).

	Symm. RW.	N=5000	N=1000	N=250	N=100	N=25
Upper rec.	0.5	0.440	0.444	0.448	0.443	0.451
Lower rec.	0.5	0.460	0.459	0.452	0.450	0.444

Table 12.2: Correlations $P_{n+1|n}$ between records in successive trading days in daily stock data from the S&P 500 for different interval length $n = 5000, 1000, 250, 100, 25$ compared with the result of $P_{n+1|n} = 1/2$ for the symmetric random walk.

The GRM results predict the difference between the number of upper and lower records correctly, but the observational curves lie distinctly below the theoretical ones. Apparently, both the upper and the lower record numbers in the stocks are systematically decreased in comparison to the GRM.

The right plot in Fig. 12.10 shows the averaged upper and lower mean record number in the detrended data. The stocks were detrended separately in each interval before we analyzed and averaged the records. The records in the stocks are compared to the analytical result for the symmetric random walk with $\langle R_n \rangle \approx \sqrt{4n/\pi}$. The upper record numbers are in good agreement with the (unbiased) random walk result, but, here, in contrast to our analysis of the daily returns (see Fig. 12.8), the lower record number is decreased.

Since, as demonstrated in Fig. 12.10, the GRM of biased random fails partly in modeling the mean record number of daily stocks, we compared the data for the intervals of 250 trading days to the more sophisticated AR(1) process introduced in section 12.4. Even though the parameter α of the AR(1) process is arbitrarily set to $\alpha = 0.99$, this process seems to describe the mean record numbers of the stocks more accurately. Both the total upper and lower record numbers and the differences between the two are modeled correctly.

12.5.4.1 Correlations between stock records

Motivated by these observations, we also considered the correlations between the record events in the individual stocks. In a simple symmetric random walk the probability $P_{n+1|n}$ of a record in the $(n + 1)$th entry given that the nth entry was already a record is simply

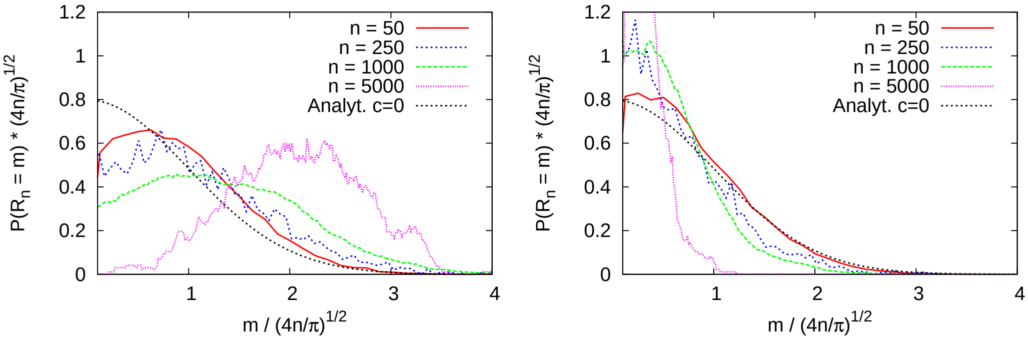

Figure 12.12: **Left:** The rescaled upper record number of the daily stock prices S_n of the S&P 500 index. We computed the upper record number at the end of intervals with different interval lengths of $n = 50, 250, 1000, 5000$ trading days. For the intervals shorter than $n = 5000$ we averaged the distributions over consecutive intervals. All distributions are normalized and rescaled with respect to the mean record number $\langle R_n^{(\mathrm{GRM})}\rangle = \sqrt{4n/\pi}$ of the symmetric random walk. The distribution of the symmetric random walk is also plotted for comparison. **Right:** The same analysis but for the lower record number of the daily stock prices.

$P_{n+1|n} = \mathrm{Prob}\left[\xi_{n+1} > 0\right] = 1/2$. This is just the probability that the walker makes a positive jump. Accordingly, for a small $c \ll \sigma$, The probability for a second upper record in step $n+1$ directly following a record in step n of a biased random walk is given by

$$P_{n+1|n} := \mathrm{Prob}\left[X_{n+1}\ \mathrm{rec.}|X_n\ \mathrm{rec.}\right] = \int\limits_{-\infty}^{c} \mathrm{d}\xi\ f\left(\xi\right) \approx \frac{1}{2} + f\left(0\right) c > \frac{1}{2}. \tag{12.25}$$

Here, we assumed that the symmetric jump distribution $f\left(\xi\right)$ is sufficiently smooth around zero. We compared this prediction with the daily stock prices in the S&P 500 and found that the probability $P_{n+1|n}$ in these series is always significantly smaller than predicted by the GRM. On average, the correlations both for upper and lower records were around $P_{n+1|n} \approx 0.45$. More detailed results for this probability in time series of different interval length N can be found in Tab. 12.2. Apparently, the conditional probabilities $P_{n+1|n}$ for upper and the lower records are much smaller than 0.5 relatively independent of the interval length N.

When we compare these findings with numerical simulations of $P_{n+1|n}$ for the AR(1) process, we find a better agreement. When simulating a $n = 5000$ step symmetric AR(1) process with $\alpha = 0.99$ (as in Fig. 12.11), we find a value of $P_{n+1|n} \approx 0.4583...$. A small bias of the order $c = 0.02$, which was found in the data, had only a weak effect, smaller than 0.01, on this probability. Also from this point of view, the autoregressive AR(1) process models the record statistics of the daily stocks more precisely.

12.5.4.2 Full distribution of the record number

To conclude our study of the record statistics of the stock prices S_n, we also analyzed the full distribution of the record number. Again, we tried to compare the distribution of the record number in the stocks with the predictions from the GRM. As shown in [22], for an unbiased random walk the record number R_n has a half-Gaussian distribution. For $n \to \infty$, we have

$$\mathrm{Prob}\left[R_n = m\right] \approx \frac{1}{\sqrt{n\pi}} \exp\left(-\frac{m^2}{4n}\right), \tag{12.26}$$

for a positive integer $m \in [1, n]$. Therefore, the rescaled record number $R_n/\langle R_n\rangle$ is distributed according to a half Standard-Normal distribution. Interestingly, the most probable

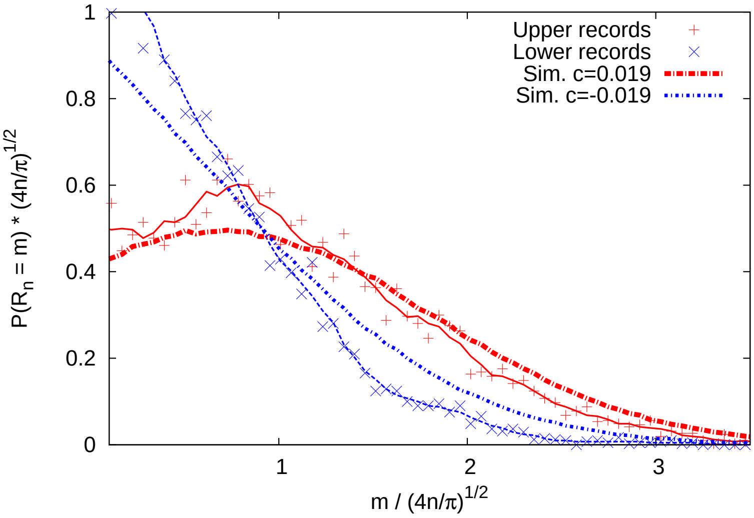

Figure 12.13: The rescaled upper and lower record number of the daily stock prices S_n of the S&P 500 index. Here, we computed the upper and lower record number at the end of 20 intervals with 250 trading days length. Again, these distributions are normalized and rescaled with respect to the mean record number of the symmetric random walk with $\langle R_n^{(GRM)} \rangle = \sqrt{4n/\pi}$. The blue and red dashed lines are results from numerical simulations of biased random walks with a Gaussian jump distribution ($\sigma = 1$) and a drift of $c = 0.019$. The drift was obtained from the S&P 500 data by linear regression in the intervals.

record number in the unbiased case is always $R_n = 1$. In [27], the asymptotic behavior of the distribution $P\left[R_n = m\right]$ was studied also for the biased case. It was shown that for $n \to \infty$ the record number of a Gaussian random walk (Lévy index $\mu = 2$) approaches a Gaussian distribution. Unfortunately, the results presented in this publication do not hold in the regime of $c\sqrt{n} \ll \sigma$, which is most interesting for our analysis of stock data.

To illustrate how the distribution of the record number of the stocks in the S&P 500 evolves in time, we computed this distribution for different interval length n. In Fig. 12.12, we show rescaled distributions of the mean record number divided by $\langle R_n^{(GRM)} \rangle = \sqrt{4n/\pi}$ at the end of intervals with length n. As in our previous considerations, if possible, we averaged over successive intervals.

The left plot in Fig. 12.12 shows distributions of upper records of S_n. Here, for small interval length the distribution of the record number R_n is still very similar to the analytical prediction for the unbiased random walk, but already for $n = 50$ the most probable record number is not one anymore and the maximum of the distribution is shifted. For $n = 5000$ the distribution resembles a full Gaussian and appears to be symmetric around its mean value.

The right plot in Fig. 12.12 shows rescaled distributions of lower records. Here, the distribution after $n = 50$ trading days is again very similar to the half-Gaussian. With increasing n the distributions get narrower and small record numbers become more and more likely.

In Fig. 12.13, we compare the rescaled distribution of the upper and lower record number R_{250} after $n = 250$ trading days with numerical simulations of biased random walks. We simulated a Gaussian GRM with a jump distribution of $\sigma = 1$ and the same linear drift of $c = 0.019$ as before. Both the distributions of the upper and the lower mean record number agree relatively well with the model prediction, but again there are some essential deviations between the model and the data. Both in the case of the upper records and in the case of the lower records, the number of series with a small record number is larger than predicted and in both cases the tails of the distributions decay faster.

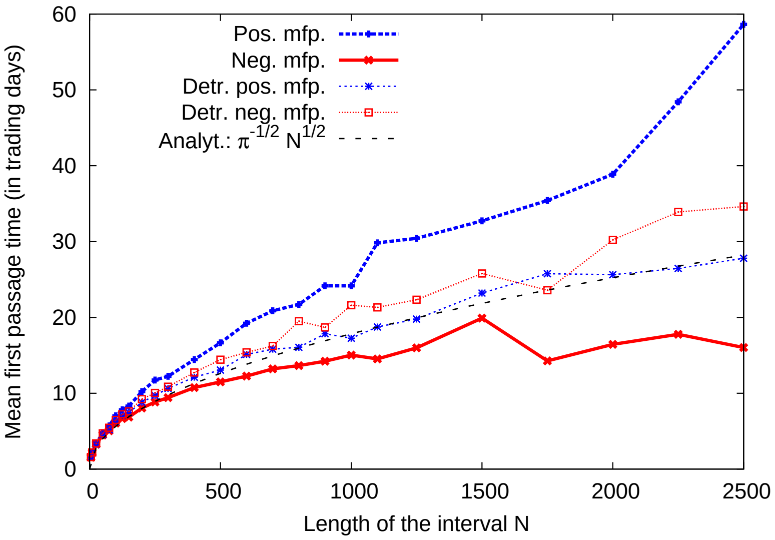

Figure 12.14: Averaged mean first-passage times (fpt's) for the logarithmic stock prices in the S&P 500 index. The plot shows both the mean negative (red) and positive (blue) fpt's of the daily stock prices for different interval length N. Only the first-passage times smaller than the respective interval length N contributed to the individual data points. We computed the fpt's for the undetrended (bold lines) and the lin-log-detrended (dashed lines) stock prices. The analytical result for the symmetric random walk is given by the black dashed line.

12.5.5 First-passage times of the stocks

To improve our understanding of the record statistics of the S&P 500 index, we also considered the mean first-passage times (fpt's) $f_{\pm}(n)$ of the logarithmic stock prices. Since a process with R_n records can be seen as a chain of R_{n-1} first-passage problems (and one survival problem), we assume that we can learn something from considering these quantities.

Due to the finite length of the time series, it is of course impossible to compute the full mean fpt including first-passage events with arbitrarily large $n > 5000$. Because of that, we considered the averaged fpt's on shorter intervals of several different interval lengths N. For each entry S_n in a given time series, we computed the time that it took until the logarithmic stock prize $\ln S_n/S_{n-1}$ of this trading day was first deceeded (negative fpt) or exceeded (positive fpt) by a succeeding logarithmic prize. Stocks that did not cross this initial value within the next N steps were not considered. Then we averaged these fpt's of over all entries and all stocks in an interval, to get the positive and negative mean fpt's for the individual intervals. Eventually, those mean values were averaged over all intervals. We performed this analysis both for the undetrended logarithmic stock prices and the lin-log-detrended stocks, as well as for numerous different choices for the interval length N. The results are summarized in Fig. 12.14.

In this figure, we also plotted the analytical result for the symmetric random walk with $f(N) = \sqrt{N/\pi}$. The results for the undetrended mean fpt's are not surprising. The negative mean fpt is significantly increased in good agreement with the decreased number of lower records in the undetrended data (see also [23]). Of course, if it takes longer until the next negative first-passage event occurs after a lower record, the lower record rate is smaller. Accordingly also the fact that the positive fpt is decreased agrees well with the large number of upper records in the undetrended S&P 500 data.

If we detrend the data, the mean fpt's of the stocks shift much closer to the analytical result for the symmetric random walk. Nevertheless, our analysis indicates a small asymmetry, which might be related to the slightly decreased number of lower records in the detrended data discussed in the previous section. While the detrended negative fpt's behave

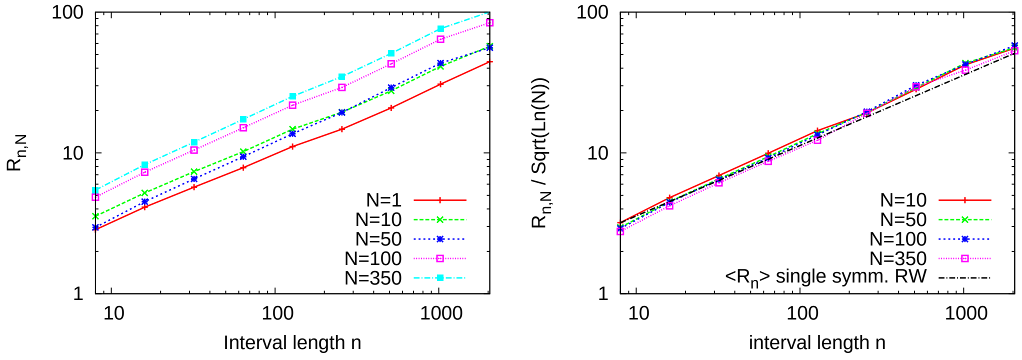

Figure 12.15: **Left:** The mean record number $\langle R_{n,N}\rangle$ of the maximum of N randomly selected detrended and normalized stocks from the S&P 500. The 5000 trading days were subdivided in intervals of n trading days and then linearly detrended in these intervals using the average linear trend of the index. Then, we chose N stocks randomly out of the total number of 366 stocks and analyzed the evolution of the record number in this set. We computed $\langle R_{n,N}\rangle$ at the end of the intervals and averaged over k subsequent intervals with $nk \leq 5000$. We averaged over 10^4 randomly selected ensembles for each $N = 1, 10, 50, 100, 350$. **Right:** $\langle R_{n,N}\rangle/\sqrt{\ln N}$ for the stocks in the S&P 500 and the analytical result for the single random walk plotted against the interval length n. The fact that all lines collapse confirms that the mean record number of the maximum of N stocks behaves like $\langle R_{n,N}\rangle \propto \sqrt{\ln N}$

exactly as predicted by the random walk model, the positive mean fpt's are increased for most interval lengths $N > 500$.

12.6 Record statistics of N stocks

In a recent article of Wergen et al. [21], the record statistics ensembles of N independent and symmetric random walks was studied. In that publication, it was shown that the mean record number $\langle R_{n,N}\rangle$ of the maximum of $N \gg 1$ independent Brownian random walks with a common jump distribution that has a finite variance ($\mu = 2$), is given by

$$\langle R_{n,N}\rangle \approx 2\sqrt{n \ln N}. \tag{12.27}$$

The record rate $P_{n,N}$ of the maximum of these N Brownian walkers behaves like

$$P_{n,N} \approx \sqrt{\frac{\ln N}{n}}. \tag{12.28}$$

Furthermore, one can compute the full distribution of the record number $R_{n,N}$ in this case, which approaches a Gumbel form [27, 35]. Wergen et al. also considered the regime of $\mu < 2$, where, surprisingly, $\langle R_{n,N}\rangle$ becomes completely independent of N behavior for $N \gg 1$.

In [21], the results for the Brownian case were compared to stock data from the S&P 500 index. To make this comparison, one can consider randomly chosen subsets of size N of the 366 stocks and analyze the record statistics of their maximum. To make these N stocks comparable, it is necessary to detrend and rescale the individual time series. In [21], the logarithms $\ln S_n/S_0$ of the stocks were first detrended with respect to the index mean value, before they were normalized in a way so that the standard deviation of the jump distribution was one. This way, one compares N detrended and rescaled stocks with a common variance, similar to N symmetric random walks with a Gaussian Standard Normal jump distribution.

In the left part of Fig. 12.15, the mean record number $\langle R_{n,N}\rangle$ of the maximum of N index-detrended and rescaled stocks at the end of time series with length n is plotted against

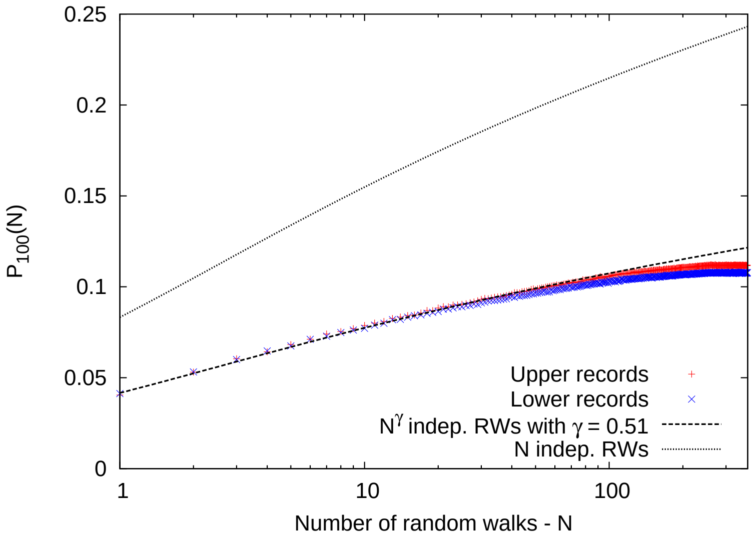

Figure 12.16: The averaged upper and lower record rate $P_{n,N}$ after $n = 100$ trading days in the S&P 500 stock data. The 5000 trading days were subdivided in intervals of 100 days and then linearly detrended in these intervals using the average linear trend of the index. The jump distributions of the detrended stocks were rescaled to a unit standard deviation. We chose N stocks randomly out of the total number of 366 stocks and analyzed the evolution of the record rate in this set. This random picking was repeated 10^4 times and the results were averaged to obtain the figure. The dashed line gives our analytical prediction for N Gaussian random walks multiplied with a manually fitted prefactor of $\gamma = 0.51$. The bold black line is the analytical prediction for N independent random walks (see [21]).

n for different values of the ensemble size N. For each N, we averaged over many different randomly selected subsets. For a given interval length n, we considered $\langle R_{n,N} \rangle$ as many consecutive intervals as we could obtain from the total number of 5000 trading days. The figure shows that $\langle R_{n,N} \rangle$ has the same n dependence as the mean record number $\langle R_{n,1} \rangle$ of a single stock. The curves for the different values of $N > 1$ are parallel to the result for $\langle R_{n,1} \rangle$ and shift upwards with increasing N. Clearly we have $\langle R_{n,N} \rangle \propto \sqrt{n}$ with a different, N-dependent, prefactor.

The right part of Fig. 12.15 shows a plot of the ratio $\langle R_{n,N} \rangle / \sqrt{\ln N}$ against the interval length n. Here, the curves for all considered values of N collapse and we find that $\langle R_{n,N} \rangle / \sqrt{\ln N}$ is independent of N. This confirms that the record statistics of the maximum of N index-detrended and normalized stocks from the S&P 500 has the same dependence on N as the maximum of an ensemble of N independent random walks. In other words, we find that, in both cases

$$\langle R_{n,N} \rangle \propto \sqrt{n \ln N}. \tag{12.29}$$

In Fig. 12.16, we present an analysis of the record rate $P_{n,N}$ of N randomly selected, index-detrended and normalized stocks compared with the analytical prediction from [21] (Eq. 12.28). The record rate was evaluated at the end of intervals with a fixed length of $n = 100$. This figure is consistent with the findings in [21]. Apparently, both the the record rates (and the mean record numbers) of the stocks and of the N independent random walkers are proportional to $\sqrt{\ln N}$, but the prefactors are different. In this case, for $n = 100$, the record rate of the stocks is only 0.51 times the record rate of the random walkers.

In [21], a similar prefactor was found considering the mean record number on intervals of $n = 250$ trading days length. A possible way to explain this prefactor is the following: As it is well known, while the theory in [21] was developed for N entirely independent random walkers, the stocks in the S&P 500 are strongly correlated. Therefore, the fact

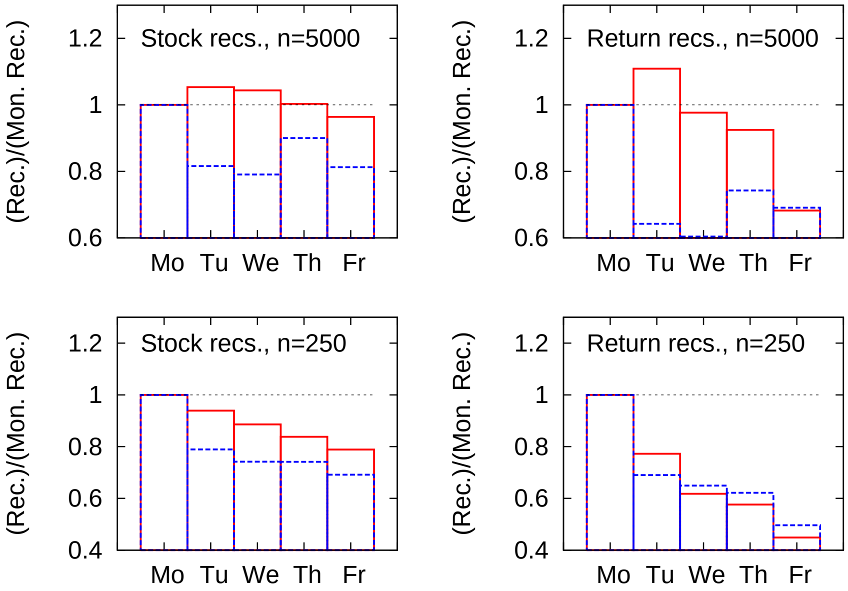

Figure 12.17: Weekly distributions of record-breaking stock prizes and daily returns. We counted the number of records on a given weekday and divided by the number of records on Mondays to make the distributions comparable. **Top left:** Records of the stock prizes $\ln S_n/S_0$ in the time series of $n = 5000$ trading days. **Top right:** Records of the daily returns $\ln S_n/S_{n-1}$ for the time series of $n = 5000$ trading days. **Bottom left:** Records of the stock prizes in 20 intervals of each $n = 250$ trading days. **Bottom Right:** Records of the daily returns in 20 intervals of each $n = 250$ trading days.

that the record number of the maximum of the stock prices has a smaller prefactor, can be explain by assuming that only a smaller number $\tilde{N} < N$ of effectively independent stocks contributes to the record statistics.

Fig. 12.16 shows that the record rate of N randomly selected stocks is the same as the record rate of N^γ independent random walkers (with $\gamma \approx 0.51$). Therefore, we can conjecture that only N^γ stocks are independent in the context of record statistics. This hypothesis is supported by the fact that the record rate of the stocks saturates at a constant value for $N > 100$. Apparently, here, the number of effectively independent stocks saturates at an upper limit of 366^γ.

12.7 Weekly distributions of records

Since not all of the findings for the record statistics of the stock data are in perfect agreement with the predictions of our simple stochastic models, we further investigated the specific patterns of record occurrence in the stock data. One thing that we were interested in is the weekly distribution of record-breaking events. The simple question was: Does the number of upper and lower records in the stocks depend on the weekday?

The weekly distributions of records in stock prices and stock returns for two different interval length are presented in Fig. 12.17. In the upper two figures this distribution was computed for records in the entire time series of 5000 trading days. The upper left plot shows the distribution of the records of the stock prices and the upper right the one of the return records. The number of records that was recorded on a given weekday was divided by the number of records that was set on Mondays to make the plots more comparable. In both cases, records are more likely on Mondays than on Fridays and, with few exceptions, the record rate decreases over the week.

The picture gets clearer if we consider the weekly distribution of record stock prizes and record returns set on intervals with a shorter length of only 250 trading days. This analysis can be found in the two lower plots of Fig. 12.17. Here, the occurrence of records decreases monotonically over the week. For the return records this effect is a lot stronger and the record rate on Mondays is roughly twice as large as on Fridays.

These significant effects on shorter time-scales show that the simple models introduced above are only capable to predict certain properties of the record statistics of the stock prices. Up to a certain degree, they can model the averaged record rate and the mean record number of the stocks on time-scales much longer than one week. More complicated factors, such as the weekly fluctuations of market activity, are not reproduced.

12.8 Summary and conclusions

We presented a detailed analysis of the statistics of record breaking events in time series of daily stock prizes. Both the record process of the daily returns δ_n and the record process of the daily stock prizes S_n behave similar to what is predicted for the Geometric Random Walk Model (GRM). In both cases, the agreement gets better if the length of the considered series length decreases. Indeed, for time series length of $n = 50$ trading days or smaller, the GRM describes the averaged record statistics of the stocks accurately.

We found that the record rate of the daily returns over the full period of 5000 trading days fluctuates strongly and is dominated by a few periods of very high market activity (or crisis). Additionally, the records of the returns are strongly correlated. Upper return records are often followed by lower return records and vice versa.

For time series with an intermediate length of $n = 250$ trading days, the record statistics of the daily returns is already more similar to the i.i.d. behavior, but we find that the rate of lower records is slightly increased by a mechanism that we can not explain. For $n = 250$ there are more lower than upper return records. For shorter time series of $n = 100$ trading days this effect vanishes and the record rate of the returns becomes very similar to the i.i.d. record rate of $p_n^{(\mathrm{iid})} = 1/n$.

Our analysis of the stock prizes showed that the GRM slightly overestimates the mean record numbers $\langle R_n \rangle$ on intervals of intermediate length with $100 < n < 1000$. We demonstrated that an autoregressive AR(1) process with a manually fitted parameter $\alpha = 0.99$ models the behavior of the record numbers more accurately. The AR(1) process is also better in describing the inter-record correlations of the stock prices, which are significantly weaker than in the random walk case. We considered the full distribution of the mean record number of the stock prices, which are, as expected, slightly narrower than the distributions corresponding from the GRM. These findings support the hypothesis that the stock prizes are more accurately described by an autoregressive process, which is basically like a random walker in a quadratic potential that draws the walker back to some mean value. It would be nice to compute the record statistics of such an AR(1) process analytically. Such a result might eventually lead to a better understanding of record-breaking stock prices.

As in [25], we discussed the record statistics of the maximum of multiple stocks that were detrended and normalized. Our analysis showed that the mean record number $\langle R_{n,N} \rangle$ of the maximum of N of these stocks has the same dependence on N as the maximum of N independent Gaussian random walks. Nevertheless, the mean record number of the stocks had a different prefactor that can be explained by introducing an effective number N^γ of independent stocks. We assume that only these N^γ stocks are effectively uncorrelated. It would be interesting to see if this effective number of stocks and the coefficient γ can also be measured by other methods unrelated to record statistics. By now, we are not aware of similar results for stock data, even though such a measure of correlation between the stocks might be useful in the estimation of possible risks for instance for stock portfolios and index funds.

In the penultimate section, we also studied the weekly distribution of the occurrence of record events both in stock prizes and in daily returns. We showed that record events are most likely to occur in the beginning of the week. This is not surprising, since it is known that the market activity varies over the week and usually decreases when the weekend approaches.

In summary, we demonstrated, where the simple Geometric Random Walk Model is useful in describing the record statistics of stock prizes from the S&P 500 stock market and we illustrated where it fails. With the slightly more complicated AR(1) process we already gave an idea how to improve the modeling. However, a model that describes the strong fluctuations of the record rate discussed in this article and effects like the distinct weekday-dependence of the record occurrence is still missing. Such a model would be an interesting goal for future research.

Acknowledgements

The author is grateful to Joachim Krug for many interesting and insightful discussions on the subject of this article and Ivan Georg Szendro Terán for his input towards the completion of this work.

Bibliography

[1] D. Sornette, *Why stock markets crash: critical events in complex financial systems* (Princeton University Press, Oxford, 2003).
[2] E. F. Fama, J. Bus. **38**, 34 (1965).
[3] F. M. Longin, J. Bus. **69**, 383 (1996).
[4] A. Johansen and D. Sornette, Eur. Phys. J. B **1**, 141 (1998).
[5] S. Redner and M. R. Peterson, Phys. Rev. E **74**, 061114 (2006).
[6] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).
[7] G. Wergen and J. Krug, EPL **92**, 30008 (2010).
[8] N. Elguindi, S. A. Rauscher, and F. Giorgi, Climatic Change **114**, 1 (2012).
[9] S. Rahmstorf and D. Coumou, Proc. Natl. Acad. Sci. USA **108**, 17905 (2011).
[10] G. Wergen, A. Hense, and J. Krug, arXiv:1210.5416 (2012).
[11] J. Krug and K. Jain, Physica A **358**, 1 (2005).
[12] L. P. Oliveira *et al.*, Phys. Rev. B **71**, 104526 (2005).
[13] P. Sibani, G. F. Rodriguez, and G. G. Kenning, Phys. Rev. B **74**, 224407 (2006).
[14] P. Sibani, Eur. Phys. J. B **58**, 483 (2007).
[15] R. M. Vogel, A. Zafirakou-Koulouris, and N. C. Matalas, Water Resour. Res. **37**, 1723 (2001).
[16] D. Gembris, J. G. Taylor, and D. Suter, Nature **417**, 506 (2002).
[17] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).
[18] R. Ballerini and S. Resnick, Adv. Appl. Prob. **19**, 801 (1987).
[19] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).
[20] R. Gouet, F. J. Lopez, and G. Sanz, Adv. Appl. Prob. **37**, 781 (2005).
[21] G. Wergen *et al.*, Phys. Rev. Lett. **109**, 164102 (2012).
[22] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).
[23] G. Wergen, M. Bogner, and J. Krug, Phys. Rev. B **83**, 051109 (2011).
[24] M. Bogner, “Unpublished thesis: Rekordstatistik in Finanzdaten,” (2009).
[25] G. Wergen, S. N. Majumdar, and G. Schehr, Phys. Rev. E **86**, 011119 (2012).
[26] Thomson Reuters, “Thomson Datastream Advance 4.0 SP4,” (2003).
[27] S. N. Majumdar, G. Schehr, and G. Wergen, J. Phys. A: Math. Theor. **45**, 355002 (2012).

[28] O. E. Barndorff-Nielsen and N. Shephard, J. Roy. Stat. Soc. B **63**, 167 (2001).

[29] R. Cont and P. Tankov, *Financial modelling with jump processes* (Chapman & Hall, 2003).

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

[31] V. B. Nevzorov, *Records: Mathematical Theory* (American Mathematical Society, 2004).

[32] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, 2001).

[33] E. Sparre Andersen, Math. Scand. **1**, 263 (1953).

[34] E. Sparre Andersen, Math. Scand. **20**, 195 (1954).

[35] M. Abramowitz and I. Stegun, *Handbook of mathematical functions* (Dover Publications Inc., New York, 1970).

[36] G. H. Weiss, *Aspects and applications of the random walk* (North-Holland, 1994).

[37] A. A. Novikov, Th. Probab. Appl. **35**, 269 (1990).

[38] L. Alili and P. Patie, Stoch. Mod. **21**, 967 (2005).

[39] S. Ditlevsen, Stat. Probab. Lett. **77**, 1744 (2007).

[40] A. A. Novikov, Teor. Veroyatnost. i Primenen. **53**, 458 (2008).

[41] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000).

[42] J. Voit, *The Statistical Mechanics of Financial Markets* (Springer Berlin, 2001).

[43] L. Bachelier, Ann. Sci. Ecole Norm. S. **17**, 21 (1900).

Part V

Records statistics - A review

13 Record statistics beyond the standard model - Theory and applications 231

13.1 Why records? 232
13.2 Classical theory of records 233
13.3 Records in uncorrelated and time-dependent RV's 235
  13.3.1 The Linear Drift Model 236
    13.3.1.1 The regime of small cn 238
    13.3.1.2 Correlations in the Linear Drift Model 239
  13.3.2 The Increasing Variance Model 240
13.4 Discreteness, rounding and ties 241
  13.4.1 Rounding effects 243
  13.4.2 δ-records 244
  13.4.3 Geometric records 246
13.5 Records in correlated processes 246
  13.5.1 Records in symmetric, discrete-time random walks 246
  13.5.2 Biased random walks 249
  13.5.3 Multiple random walks 252
  13.5.4 Continuous-time random walks 254
  13.5.5 Records in higher-dimensional processes 255
13.6 Applications. 255
  13.6.1 Climate records 255
  13.6.2 Records in finance 257
  13.6.3 Physics and biology 258
  13.6.4 Athletic records. 259
13.7 Summary and outlook 260

14 Summary & Conclusions 265
14.1 Part I - Records in uncorrelated random variables 265
14.2 Part II - Record-breaking temperatures 267
14.3 Part III - Record statistics of random walks 268
14.4 Part IV - Records in finance 269
14.5 Closing remarks & bigger picture 270

Bibliography for chapters 1, 2 & 14 271

Acknowledgements 277

Chapter 13

Record statistics beyond the standard model - Theory and applications

Gregor Wergen

Institute for Theoretical Physics, University of Cologne

submitted to Journal of Physics A: Mathematical and Theoretical (arXiv:1211.6005)

Abstract: In recent years there has been a surge of interest in the statistics of record-breaking events in stochastic processes. Along with that, many new and interesting applications of the theory of records were discovered and explored. The record statistics of uncorrelated random variables sampled from time-dependent distributions was studied extensively. The findings were applied in various areas to model and explain record-breaking events in observational data. Particularly interesting and fruitful was the study of record-breaking temperatures and their connection with global warming, but also records in sports, biology and some areas in physics were considered in the last years. Similarly, researchers have recently started to understand the record statistics of correlated processes, such as random walks, which can be helpful to model record events in financial time series. This review is an attempt to summarize and evaluate the progress that was made in the field of record statistics throughout the last years.

13.1 Why records?

In our competitive society, we care a lot about performance and we often feel the need to outperform others. Maybe this is why, recently, also researchers have become more and more interested in records. A record is simply an achievement, a result or some other kind of measurement in a given chain of events that exceeds everything that has been encountered previously. Therefore a new record is always something remarkable which attracts attention, regardless of whether or not the occurrence of this record is considered good or bad. Records receive more attention and are remembered longer than other measurements because they show the boundary of what has been possible so far. In this context, the famous book 'Guinness World Records' holds its own record as the best-selling copyrighted book in history [1].

An area, where records are certainly of great interest is, of course, sports. Particularly in athletics and in swimming records, like Olympic- or world-records, are always something special and noteworthy [2, 3]. But also in the context of global warming, records have recently become particularly important and interesting for climatologist. The question, how a changing climate affects the number of record temperatures that we encounter has bothered both the general public and researchers [4–12]. By now it is well established that global warming leads to many new heat-records and to a decreased number of new record-breaking cold temperatures.

Records are important also in countless other areas of science. In physics, they were discussed in the context of the theory of spin-glasses [13–15] and high-temperature superconductors [13, 16], but they also found applications in evolutionary biology [17–20]. Curiously, in 2010, the dynamics of ant movements were studied using results from the theory of records [21]. Thanks to new theoretical result it was recently possible to analyze and model the statistics of records in stock prices [22–24].

These data-oriented studies were accompanied and complemented by a substantial number of new theoretical results. The classical theory of records in time series of independent and identically distributed (i.i.d.) random variables was already developed many decades ago [25–27], but to understand the record statistics of more complicated systems like the worlds climate or evolutionary pathways, new techniques beyond this standard model were needed. In this context, various processes of uncorrelated random numbers sampled from time-dependent distributions were studied. Most importantly the Linear Drift Model (LDM), which was introduced already in the 80's, where random numbers are drawn from a distribution of unvarying shape but with an increasing mean value, was studied extensively [28–33]. Some authors also considered record events from broadening distributions [34, 35].

Also connected with problems in the adaptation of theoretical results from record statistics on observational data, are the so-called discreteness or rounding effects. Even though most of the classical theory is developed for random numbers sampled from continuous distributions, practical measurements are always unprecise and rounded to a certain accuracy. Both, the record statistics of random numbers from discrete distributions [36–39], as well as the consequences of analyzing records in time series of random numbers that were drawn from continuous distributions and then discretizes in a measuring process were discussed in recent years [40].

In 2008, Majumdar and Ziff computed the record statistics of symmetric random walks [41]. Their findings entailed a series of new theoretical results and applications. By now, the complicated record statistics of biased random walks and Lévy flights [22, 23] as well as the one of ensembles of multiple independent random walks [42] is well understood. Records in continuous time random walks [43] and also in the distance of higher dimensional jump processes from their origin [44] were studied, similarly.

The purpose of this work is to summarize and evaluate these recent developments mostly from a theoretical point of view, but also with a short evaluation of recent data-driven studies in the field. The rest of this review is organized as follows: We will start with a brief introduction to the classical theory of records, where we introduce the important

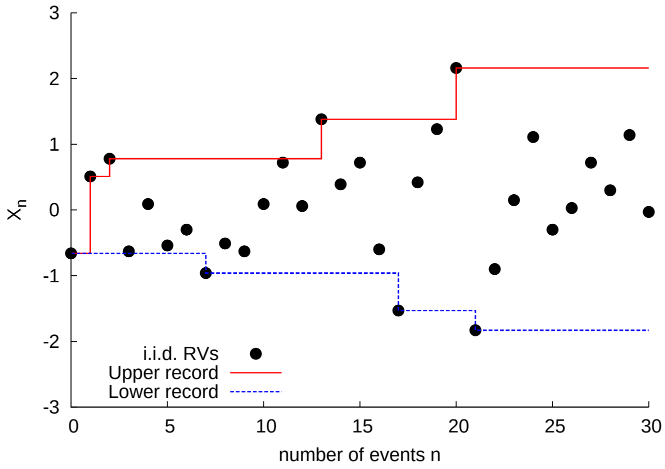

Figure 13.1: Sketch of the record process of i.i.d. RV's. The dots represent a time series $X_0, X_1, X_2, ...$ of RV's drawn from a continuous distribution $f(x)$ (in this case a standard normal distribution). The red (blue dotted) lines illustrate the progressions of the upper (lower) record. Here, we find 5 upper and 4 lower records. In both cases X_0 is the first record.

notation and present some elementary results. Then, in section 13.3, we describe recent developments in the field of record statistics of uncorrelated random variables that are sampled from time-dependent distributions. In this context we consider the important Linear Drift Model of random variables with a linearly increasing mean value, as well as a model of increasing variance.

In the subsequent section 13.4 we discuss various alternative models of records in continuous and discrete random variables. In particular, we will consider the effects of rounding and present generalized concepts like δ- and geometric records, which are record events that are only counted if they exceed a certain barrier above or a certain multiple of the last record.

Then, in section 13.5, we analyze various stochastic processes with correlated entries, starting with the symmetric random walk. After discussing the important results of Majumdar and Ziff on the symmetric discrete-time random walk [41], we will demonstrate how these findings can be generalized to biased random walks, to ensembles of multiple symmetric random walks and to symmetric continuous time random walks.

Various important applications that were mentioned above are presented and discussed in section 13.6. We will start by describing the progress made in the study of temperature records in 13.6.1. As an important application of the random walk model we outline some recent results about the statistics of record-breaking stock prizes in section 13.6.2. Subsequently, we briefly mention some other applications, for instance in physics, biology and in athletics (13.6.3 and 13.6.4). Afterwards, in section 13.7, we give a brief summary in which we assess the current state of research in the field of record statistics and point out a number of interesting open questions and suggestions for future research.

13.2 Classical theory of records

Let us consider a time series $X_0, X_1, ..., X_n$ of random variables (RV's), which can, for instance, be a series of temperatures, stock prices, sports results or some other kind of measurement process. In such a time series an entry X_n is an upper record if it exceeds all previous entries:

$$X_n > \max\{X_0, X_1, ..., X_{n-1}\}. \tag{13.1}$$

Analogously, a lower record is an entry with $X_n < \min\{X_0, X_1, ..., X_{n-1}\}$. In general, one defines the first entry X_0 as the first (upper and lower) record. The record process in the simple case of independent and identically distributed (i.i.d.) RV's is illustrated in Fig. 13.1. Probably the two most studied quantities in the theory of records are the record number R_n and the probability P_n for a record at time n. This probability P_n for an upper record is defined as

$$P_n := \mathrm{Prob}\left[X_n > \max\{X_0, X_1, ..., X_{n-1}\}\right]. \tag{13.2}$$

In the following we will also refer to P_n as the record rate. The record number R_n is simply the number of records that occurred in the time series up to time n. The mean record number $\langle R_n \rangle$, the expected average record number of a stochastic process, can by expressed in terms of the record rate:

$$\langle R_n \rangle = \sum_{k=0}^{n} P_k. \tag{13.3}$$

In the case of i.i.d. RV's sampled from a continuous distribution with probability density function (pdf) $f(x)$, one can easily compute the record rate and the mean record number: With the so-called *stick-shuffling* argument one finds that the probability P_n for a record at time n in a time series of i.i.d. RV's is given by

$$P_n = \frac{1}{n+1}. \tag{13.4}$$

This is just the probability that in a random ordering of $n+1$ RV's (sticks) the last one (X_n) is the largest. With this result, the mean record number $\langle R_n \rangle$ in a series of i.i.d. RV's takes the form

$$\langle R_n \rangle = \sum_{k=0}^{n} P_k = \sum_{k=0}^{n} \frac{1}{k+1} = H_{k+1} \xrightarrow{n\to\infty} \ln n + \gamma, \tag{13.5}$$

where H_k is the kth Harmonic number (cf. [45]) and $\gamma \approx 0.577215...$ the Euler-Mascheroni constant [25, 45].

An important feature of record events in i.i.d. RV's is that they are stochastically independent. The probability for a record in the nth event is independent from records in previous entries [25, 31]. One can shown that the joint probability $P_{n,m}$ of records both at times n and m factorizes:

$$P_{n,m} := \mathrm{Prob}\left[X_n, X_m \text{ both records}\right] = P_n \cdot P_m \tag{13.6}$$

By now, a lot more is known about the record statistics of i.i.d. RV's. A good review can be found in the book by Arnold et al. [25], or Nevzorov [26] (see also [27]). There, quantities like the distributions of record values with a given record number, or the interesting waiting-time statistics between individual record events are discussed in detail. A noteworthy finding is that the mean time $\langle T_{R_n} \rangle$, at which a record with record number R_n occurs is infinite (see also [46, 47]). Similarly the inter-record times $\Delta_{R_n} := T_{R_n} - T_{R_{n-1}}$ have a divergent mean value $\langle \Delta_{R_n} \rangle$.

Furthermore, in the book by Arnold et al. [25], it is shown how to compute the probability density function of a record value with a given record number k. Arnold argues that due to the so-called *lack-of-memory* property of the exponential distribution with $f(x) = e^{-x}$ (for $x > 0$), the pdf $f_k(x)$ of such a record value from this distribution is given by

$$f_n(x) = \Gamma[k]^{-1} x^{-k} e^{-x}, \tag{13.7}$$

where $\Gamma[k] = (k-1)!$ is the Gamma-function [45]. It is easy to show that a record with record number k from an exponential distribution is given by the value of the $(k-1)$th record plus an exponential RV sampled from $f(x)$. Therefore, the pdf of the kth record is just the convolution of $f_{k-1}(x)$ and $f(x)$. By iteration this leads eventually to the Gamma-distribution in Eq. 13.7.

This result can be used to compute the distribution of the kth record value in time series of RV's from arbitrary continuous distributions. For that purpose one has to realize that a RV X_i from any continuous pdf $f(x)$ has the same distribution as

$$F^{-1}\left(1-\exp\left(X_i^{(\exp)}\right)\right), \tag{13.8}$$

where $F^{-1}(x)$ is the inverse cumulative of $f(x)$ and $X_i^{(\exp)}$ an exponentially distributed RV with pdf e^{-x} as before. This is easy to see since $1-\exp\left(X_i^{(\exp)}\right)$ is just a uniform distribution on the interval $[0,1)$, which is the image space of $F(x)$. Using Eq. 13.8 we can infer that

$$\begin{aligned} F(x) = \text{Prob}\left[X_i < x\right] &= \text{Prob}\left[F^{-1}\left(1-e^{-X_i^{(\exp)}}\right) < x\right] \\ &= \text{Prob}\left[X_i^{(\exp)} < -\ln\left(1-F(x)\right)\right]. \end{aligned} \tag{13.9}$$

With this result it is clear how to compute $f_k(x)$ in the general case. We just have to replace the x in Eq. 13.7 by $-\ln(1-F(x))$. This leads to

$$f_k(x) = \Gamma[k]^{-1}\left(-\ln\left(1-F(x)\right)\right)^k f(x). \tag{13.10}$$

An important classical result is that the three universality classes of extreme value statistics (EVS) [48] that describe the limiting distributions of the maximal value of large sets of i.i.d. RV's (for an introduction cf. [49–51]), are also relevant in record statistics. In 1973 Resnick showed that the distribution $f_k(x)$ of a record value with a large record number $R_n \to \infty$ can be rescaled to one of three limiting forms [52]:

- Negative-log-normal distribution — This distribution corresponds to the Weibull class of EVS. Here the negative logarithmic values of the records are normal distributed.
- Normal distributed — Record values in series of RV's from the Gumbel class of EVS approach a normal distribution for $R_n \to \infty$.
- Log-normal distribution — In the Fréchet class the logarithms of the record values are normal distributed.

In the following, we will find that the three universality classes of EVS are also of importance for the record statistics of time-dependent RV's. Many of the results presented in this article will be characterized and discussed in the context of these classes. In time series of correlated RV's however, they lose their importance and one finds different interesting universal characteristics.

13.3 Records in uncorrelated and time-dependent RV's

While the classical theory introduced above deals with identically distributed RV's drawn from a single, stationary pdf $f(x)$, one can also consider the more general scenario of uncorrelated, but non-identically distributed random numbers $X_0, X_1, ..., X_n$ from a time series of probability densities $f_i(x_i)$. In this general case, it is more complicated to compute the record rate P_n and the mean record number $\langle R_n \rangle$. Here, the record rate can be obtained from the integral (cf. [31])

$$P_n = \int \mathrm{d}x_n\, f_n(x_n) \prod_{k=0}^{n-1} F_k(x_n), \tag{13.11}$$

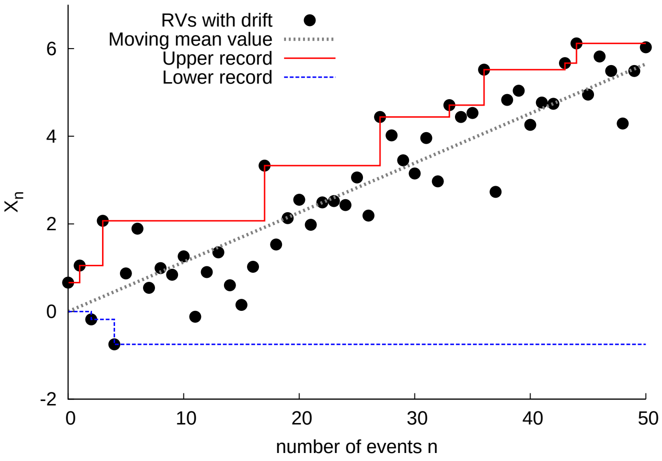

Figure 13.2: Sketch of the record process of uncorrelated RV's with a linear drift ($c = 0.1$). The dots represent a time series $X_0, X_1, X_2, ...$ of RV's drawn from a series continuous distributions with $f_k(x) = f(x - ck)$. The red (blue dotted) lines illustrate the progressions of the upper (lower) records.

where $F_k(x_n) = \int_n^x \mathrm{d}x_k \; f_k(x_k)$ is the cumulative distribution function (cdf) of the pdf $f_k(x_k)$. This expression is easy to understand, since it just integrates the probability that the X_n's entry is equal to x_n, while all previous entries X_k, with $k < n$, are below x_n, over all possible values of x_n.

In the following we present two possible choices for the series of probability densities $f_k(x_k)$ that were studied in the literature and that proved to be useful in the analysis of observational data.

13.3.1 The Linear Drift Model

The Linear Drift Model (LDM) was first introduced by Ballerini and Resnick in the 1980's [28, 29] and later studied by Borovkov [30] and more recently by Franke et al. [31] as well as Wergen et al. [32]. The model describes RV's drawn from a distribution that retains its shape but has a time-dependent mean value. In particular, the RV's are sampled from a series of pdf's $f_k(x_k) = f(x_k - ck)$, where c is a real constant, which is called the drift. The entries of such a time series are of the form

$$X_k = Y_k + ck, \tag{13.12}$$

where $Y_0, Y_1, ..., Y_n$ is a time series of i.i.d. RV's sample from $f(x)$. The record process of a series of RV's from the LDM is illustrated in Fig. 13.2. Here, the general, time-dependent expression for the record rate (Eq. 13.11) takes the following form:

$$\begin{aligned} P_n(c) &= \int \mathrm{d}x \; f(x - cn) \prod_{k=0}^{n-1} F(x - ck) \\ &= \int \mathrm{d}x \; f(x) \prod_{k=1}^{n} F(x + ck) \end{aligned} \tag{13.13}$$

Ballerini and Resnick [28] could prove that this record rate has an asymptotically constant limiting value $P(c) = \lim_{n\to\infty} P_n(c)$ if the pdf $f(x)$ has a finite first moment $\int \mathrm{d}x \; xf(x) < \infty$. To determine the behavior of $P_n(c)$ in more detail, is however a difficult problem, which, in general, can not be solved exactly.

There is an example of a pdf, for which the record rate $P_n(c)$ in the LDM can be calculated for arbitrary n: For the Gumbel distribution with the probability density $f(x) = e^{-x}\exp(e^{-x})$, Franke et al. [31] found that

$$P_n^{\mathrm{Gumbel}}(c) = \frac{1-e^{-c}}{1-e^{-nc}} \tag{13.14}$$

Similarly, it is possible to compute the asymptotic record rate for the exponential distribution with $f(x) = \nu e^{-\nu x}$ (with $x > 0$ and $\nu > 0$). In this case, the record rate $P_n(c)$ assumes the following form:

$$\begin{aligned} P_n(c) &= \int_0^\infty \mathrm{d}x\ \nu e^{-\nu x} \prod_{k=1}^{n} \left(1 - e^{-\nu(x+ck)}\right) \\ &= \int_0^1 \mathrm{d}y\ \frac{(y, e^{-c\nu})_n}{1-y}, \end{aligned} \tag{13.15}$$

where $(a,q)_n$ is the q-Pochhammer symbol with $(a,q)_n := \prod_{k=0}^{n}\left(1 - aq^k\right)$ [45]. With this we can expand the asymptotic record rate $P(c)$ in powers of $e^{-c\nu}$ and find

$$\begin{aligned} P(c) &= \int_0^1 \mathrm{d}y\ \frac{(y, e^{-c\nu})_\infty}{1-y} \\ &\approx 1 - \frac{1}{2}e^{-c\nu} - \frac{1}{2}e^{-2c\nu} - \frac{1}{6}e^{-3c\nu} - \frac{1}{6}e^{-4c\nu} + O\left(e^{-5c\nu}\right). \end{aligned} \tag{13.16}$$

By using computer algebra software, such as Mathematica, it is possible to compute arbitrary higher order terms of this expansion. However, we found that, in comparison with numerical simulations of $P(c)$, the expansion up to the 4th order in Eq. 13.16 is already very accurate and fails only for small $c \to 0$ [53]. For this case, one can compute the record rate $P_n(c)$ by a different approach. Replacing the product in Eq. 13.15 by the exponential of a sum of logarithms leads to

$$\begin{aligned} P(c) &= \int_0^\infty \mathrm{d}x\ \nu e^{-\nu x} \exp\left(-\nu x - \sum_{k=1}^{n} \ln\left(1 - e^{-\nu(x-ck)}\right)\right) \\ &\approx \int_0^\infty \mathrm{d}x\ \exp\left(\frac{e^{-\nu x}}{c\nu}\left(1 - e^{-c\nu n}\right)\right), \end{aligned} \tag{13.17}$$

where, for the second step, we replaced the sum by an integral assuming that $n \gg 1$ and $c\nu \ll 1$. With this we obtain the small c behavior of $P_n(c)$ for the exponential case:

$$P_n(c) \approx c\nu \frac{1 - e^{-\frac{1}{c\nu}\left(1-e^{-c\nu n}\right)}}{1 - e^{-c\nu n}} \xrightarrow[n\to\infty]{} c\nu\left(1 - e^{-\frac{1}{c\nu}}\right), \tag{13.18}$$

which, for small $c\nu \ll 1$, approaches $c\nu$. Apparently, for small c, the record rate of the exponential distribution depends linearly on c. Comparing with numerical simulations, we found that Eq. 13.18 is computes $P_n(c)$ accuratly for $c\nu < \frac{1}{2}$ [53].

In the article by Franke et al. [31], the record rate for a more general set of continuous probability distributions was computed in two different regimes. For small c, Franke et al. derived approximate results for finite values of n in the regime of $nc \ll \sigma$, where σ is usually the standard deviation or some other measure of the width of $f(x)$. The same can be done in the opposite regime of $c \to \infty$. It turns out that the behavior of $P_n(c)$ depends systematically on the three classes of EVS. Franke et al. [31] discussed their findings in the context of these classes.

13.3.1.1 The regime of small cn

In the small c regime, it is possible to expand the record rate $P_n(c)$ into powers of c. Expansion up to the first order yields

$$\begin{aligned} P_n(c) &= \int \mathrm{d}x\, f(x) \prod_{k=1}^{n} F(x+ck) \\ &\approx \int \mathrm{d}x\, f(x) \prod_{k=1}^{n} \left(F(x) + ckf(x)\right) \\ &\approx \int \mathrm{d}x\, f(x)\, F^{n}(x) + \frac{c}{2} n(n+1) \int \mathrm{d}x\, f^{2}(x)\, F^{n-1}(x). \end{aligned} \tag{13.19}$$

The first summand in the last line is the stationary record rate with $P_n(c=0) = 1/n$. With

$$\mathrm{I}_n := \int \mathrm{d}x\, f^{2}(x)\, F^{n-1}(x) \tag{13.20}$$

this leads to

$$P_n(c) \approx \frac{1}{n+1} + \frac{c}{2} n(n+1)\, \mathrm{I}_n. \tag{13.21}$$

This expansion is accurate if the underlying distribution $f(x)$ varies only slowly between x and $x+cn$. For many probability densities this can be translated into $cn \ll \sigma$ (with $\sigma^2 := \int \mathrm{d}x\, x^2 f(x)$).

In [31], I_n was computed for several representative distributions from the three classes of extreme value statistics. Here, we want to derive I_n and $P_n(c)$ for a Generalized Pareto Distribution (GPD). We consider RV's from the cdf

$$F_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-\frac{1}{\xi}}, & \text{for } \xi \neq 0 \\ 1 - e^{-x}, & \text{for } \xi = 0. \end{cases} \tag{13.22}$$

$\xi \in \mathbb{R}$ is the shape parameter of $F(x)$. For $\xi \geq 0$ this distribution has an infinite support and is defined for $x > 0$, for $\xi < 0$ is it defined on the finite interval $x \in \left[0, 1-\xi^{-1}\right]$.

Depending on ξ, the GPD can be in all three classes of EVS. For $\xi < 0$, $F_\xi(x)$ is in the Weibull class, for $\xi = 0$ in the Gumbel class and for $\xi > 0$ in the Fréchet class.

For this distribution, the integral I_n can be evaluated by elementary means and we find that

$$\mathrm{I}_n = \begin{cases} \xi \mathrm{B}(n, 2+\xi), & \text{for } \xi \neq 0, \xi > -2 \\ 1/(n(n+1)), & \text{for } \xi = 0, \end{cases} \tag{13.23}$$

where $\mathrm{B}(x,y) = \Gamma[x]\,\Gamma[y]\,/\Gamma[x+y]$ is the Beta-function [45]. Using Stirling's approximation [45], we find that for $n \gg 1$, the record rate of the GPD with a drift $c \ll n^{\xi-1}$ is given by

$$P_n(c) \approx \frac{1}{n+1} + c \begin{cases} \xi \Gamma[2+\xi]\, n^{-\xi}, & \text{for } \xi \neq 0, \xi > -2 \\ \frac{1}{2}, & \text{for } \xi = 0. \end{cases} \tag{13.24}$$

This result summarizes how a small linear drift affects the record rates depending on the extreme value class of the underlying distribution. Although this is no conclusive proof, we conjecture that the effect of the drift generally increases with n for distributions of the Weibull class. In the Fréchet class the effect decays with n and the drift is asymptotically negligible. The Gumbel class is intermediate between these two cases. Interestingly, since for $\xi > 1$, $n^{\xi-1}$ grows with n, some of the results for the Fréchet class (for $\xi > 1$) are also correct in the asymptotic limit with $n \to \infty$.

To better understand the behavior of $P_n(c)$ in the Gumbel class, Franke et al. [31] considered the Generalized Gaussian Distribution (GGD) with

$$f(x) = 2\Gamma\left[1+\beta^{-1}\right]^{-1} e^{-|x|^{\beta}} \tag{13.25}$$

with $\beta > 0$. They could show that, here, I_n grows logarithmically with n:

$$\mathrm{I}_n \propto \ln(n)^{1-\frac{1}{\beta}}. \tag{13.26}$$

This expression includes the important case of a Gaussian distribution for $\beta = 2$. For the Gaussian probability density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2\sigma^2}$ one obtains

$$P_n(c) \approx \frac{1}{n} + \frac{c}{\sigma}\frac{2\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)}. \tag{13.27}$$

13.3.1.2 Correlations in the Linear Drift Model

An interesting subtlety that was discovered in the study of the LDM is the fact that record events in this process are not stochastically independent as in the i.i.d. case. In particular, it was found by Wergen et al. [32] that the probability $P_{n,m}(c)$ of records in the entries n and m in a series of RV's with a linear drift can differ from the product of the record rates $P_n(c)$ and $P_m(c)$. In [32], the probability $P_{n,n+1}(c)$ of having two consecutive records was studied in detail. They showed that, depending on the choice of the underlying distribution, $P_{n,n+1}(c)$ can be both, smaller and larger than $P_n(c) \cdot P_{n+1}(c)$. Therefore, the probability for a second record in step $n+1$ after a record in step n can be both increased and decreased with respect to the unconditional probability $P_{n+1}(c)$.

Wergen et al. [32] defined the ratio

$$l_{n,n+1}(c) := \frac{P_{n,n+1}(c)}{P_n(c) \cdot P_{n+1}(c)}, \tag{13.28}$$

which is always given by $l_{n,n+1}(c=0) = 1$ in the i.i.d. case. In the regime of small c and $n \gg 1$, we can again use the GPD (Eq. 13.22) to illustrate the asymptotic behavior of this quantity. Expanding $l_{n,n+1}(c)$ up to first order in c with the same method as before, we find that

$$l_{n,n-1} \approx 1 + c\begin{cases}\xi\Gamma[2+\xi]\,n^{1-\xi}, & \text{for } \xi \neq 0, \xi > -2\\ \frac{1}{2}, & \text{for } \xi = 0,\end{cases} \tag{13.29}$$

which is again valid for $c \ll n^{\xi-1}$. Apparently, the exponential distribution with $f(x) = e^{-x}$ ($\xi = 0$) plays an outstanding role. For the distributions of the Weibull class with $\xi < 0$, the inter-record correlations are negative for a positive drift $c > 0$, for the representatives of the Fréchet class ($\xi > 0$), $l_{n,n-1}(c > 0)$ is larger than one and grows with n. Only in the exponential case, $l_{n,n+1}(c)$ assumes an n-independent value slightly above unity.

Again, the intermediate Gumbel regime can be studied more systematically using the GGD with $f(x) \propto e^{-|x|^{\beta}}$ as before. In a lengthy calculation, Wergen et al. [32] showed that, here, for $n \gg 1$, the ratio $l_{n,n+1}(c)$ behaves like

$$l_{n,n+1}(c) \approx 1 - cnA\left(1 - \frac{1}{\beta}\right)\ln(n)^{1-\frac{1}{\beta}}. \tag{13.30}$$

with a positive constant A, which depends on n. Apparently, for $\beta < 1$, stretched exponential distributions, broader than the exponential have positive correlations that grow logarithmically with n. Distributions decaying faster than the exponential ($\beta > 1$) lead to negative correlations.

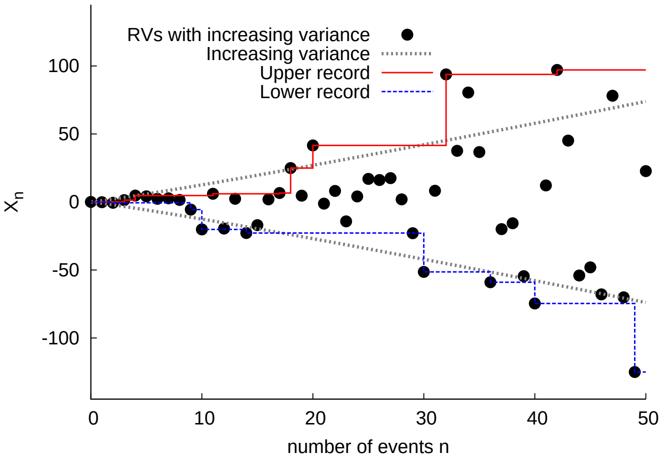

Figure 13.3: Sketch of the record process of uncorrelated (Gaussian) RV's from a broadening distribution. The sticks represent a time series $X_0, X_1, X_2, ...$ of RV's drawn from a series of continuous and symmetric distributions $f_k(x) = f\left(xk^{-\alpha}\right)$ (Here: $\alpha = 1.1$). The red (blue dotted) lines illustrate the progression of the upper (lower) record.

Even though it is not clear how two explain the emergence of these correlations and, in particular, why it is possible that records occur more frequently after a preceding record, these effects turn out to be useful. Since it is a well known problem for experimentalist to decide whether or not a series of measurements is drawn from an underlying distribution with so-called heavy-tails (see for instance [54–56]), Franke et al. [33] proposed a test that uses the findings presented in [32] in this matter.

For this test a set of measurements $X_0, X_1, ..., X_n$ has to be shuffled randomly before adding an artificial linear drift. For a random permutation $\pi_0, \pi_1 ..., \pi_n$ of $0, 1, ..., n$ such a shuffled and artificially drifted set of data is given by

$$X_{\pi_0}, X_{\pi_1} + c, X_{\pi_2} + 2c, ..., X_{\pi_n} + nc. \tag{13.31}$$

Now one can analyze the inter-record correlations in this time series. In particular, one has to compute the ratio $l_{n,n+1}(c)$. As shown in [33], the statistics can be improved significantly by averaging over many different random permutations. If the correlations in the drifted time series are positive ($l_{n,n+1}(c) > 1$) this is a good indicator for measurements from a distribution, which is at least broader than the exponential one. Franke et al. [33] demonstrated that this record-based test allows to detect these heavy-tail properties already in very small data-sets with less than 64 data-points. In this context, the test might be better than standard methods like, for instance, maximum likelihood estimators, which are commonly used for problems of this type. However, a thorough comparison of this new test with the existing ones has not been performed yet.

13.3.2 The Increasing Variance Model

In 1975, Yang [57] introduced a model of growing populations to explain the increased record rate in sports due to a growing number of athletes that tries to break records. Yang showed that any exponentially growing population of athletes leads to an asymptotically constant record rate $\lim_{n\to\infty} P_n > 0$.

Building up on this model, Krug considered random variables $X_0, X_1, ..., X_n$ from a series of probability densities $f_k(x_k)$ with a time-dependent width:

$$f_k(x_k) = \lambda_k f(\lambda_k x). \tag{13.32}$$

In particular, Krug discussed distributions with a power-law time-dependence and

$$\lambda_k = k^{-\alpha} \tag{13.33}$$

Such a process is illustrated in Fig. 13.3 for an α slightly larger than one. Apparently, the distribution broadens for $\alpha > 0$ and gets sharper when $\alpha < 0$. Here, the record rate $P_n(\alpha)$ takes the following form:

$$\begin{aligned} P_n(\alpha) &= \int \mathrm{d}x\, f\left(xn^{-\alpha}\right) \prod_{k=0}^{n-1} k^{-\alpha} F\left(xk^{-\alpha}\right) \\ &= \int \mathrm{d}x\, f(x) \prod_{k=0}^{n-1} F\left(x\left(\frac{k}{n}\right)^{-\alpha}\right) \end{aligned} \tag{13.34}$$

Krug [34] computed the asymptotic behavior of the record rate $P_n(\alpha)$ and the mean record number $\langle R_n(\alpha)\rangle$ for this model in the context of the three universality classes of EVS. The effect of a broadening distribution with $\alpha > 1$ is similar to the effect of a positive drift in the LDM. For distributions of the Fréchet class, the broadening width does not systematically change the large n behavior of the record rate. At the same time, it has the strongest effects in the Weibull class.

Using the findings of Krug [34], we can calculate the asymptotic behavior of the record rate $P_n(\alpha)$ for the GPD $F_\xi(x)$, which was introduced in section 13.3.1. For a large $n \to \infty$ and $\alpha > 1$ we obtain

$$P_n(\alpha) \propto \begin{cases} \alpha^{(1-\xi)^{-1}} n^{-(1-\xi)^{-1}}, & \text{for } \xi < 0 \quad \text{(Weibull class)} \\ \frac{\ln(n)}{n} & \text{for } \xi = 0 \quad \text{(Exp. distribution)} \\ \frac{1}{n} & \text{for } \xi > 0 \quad \text{(Fréchet class)} \end{cases} \tag{13.35}$$

For the mean record number, this leads to

$$\langle R_n(\alpha)\rangle \propto \begin{cases} \alpha^{(1-\xi)^{-1}} n^{-\xi(1-\xi)^{-1}}, & \text{for } \xi < 0 \quad \text{(Weibull class)} \\ (\ln(n))^2 & \text{for } \xi = 0 \quad \text{(Exp. distribution)} \\ \ln(n) & \text{for } \xi > 0 \quad \text{(Fréchet class)} \end{cases} \tag{13.36}$$

For a Gaussian distribution, the asymptotic results only differ in the prefactors from the exponential case.

Krug also studied the correlations between the record events in this model in a numerical manner. In contrast to the LDM, he found only negative correlations between records from RV's with an increasing variance. As in the case of the LDM, it is still controversial how to explain these correlations comprehensibly.

13.4 Discreteness, rounding and ties

The theoretical results in the previous chapter were all derived for RV's from entirely continuous distributions. In the context of experimental measurements and their record statistics, but also for purely mathematical reasons, one can also study models with discrete RV's with respect to records. In this case the statistics of records is more complicated and, in principle, there are several different approaches to this problem.

On the one hand, it is possible to consider the record statistics of RV's from distributions which are inherently discrete. Two prominent examples are

- the discrete uniform distribution with equally likely probabilities for a finite number of RV's: $P[X = k] = 1/N$ (with $N \in \mathbb{N}$ and $k = 1, ..., N$),

- and the geometric distribution with $P[X=k]=(1-p)^{k-1}p$ (with $p\in[0,1]$ and $k\in\mathbb{N}$.)

In the case of a discrete distribution, it is possible that a record value, for instance an entry X_i, gets tied by a succeeding RV $X_j=X_i$. This is impossible for RV's sampled from a continuous pdf. In the case of a tie, one has to decide whether or not one wants to count this tie as a new record. In the literature about record statistics from discrete distributions, records without ties are usually called *strong* records, while records including ties are called *weak* records.

For the discrete uniform distribution, it is very easy to compute the strong record rate P_n, which is given by the following sum:

$$P_n=\sum_{k=1}^{N}\frac{1}{N}\left(\frac{k-1}{N}\right)^{n-1}, \tag{13.37}$$

For $n\to\infty$ this behaves like $P_n\approx N^{-1}\left(1-N^{-1}\right)^n$, which leads, of course, to a finite mean record number $\langle R_n\rangle\propto N$. For the weak record rate p_n the situation is different and one finds that the asymptotic record rate is given by $p_n\approx N^{-1}$, which leads to a divergent weak mean record number of $\langle r_n\rangle\approx n/N$.

The case of the geometric distribution is already much more complicated and was considered by Prodinger in 1996 [37] (see also Vervaat [36]). He derived the asymptotic mean record number $\langle R_n\rangle$ in the strong case for the geometric distribution with $P[X=k]=(1-p)^{k-1}p$. In a rather complicated computation, he showed that for $n\to\infty$:

$$\langle R_n\rangle\approx\frac{p}{\ln\left((1-p)^{-1}\right)}\left(\ln n+\gamma-\sum_{k\neq0}\Gamma\left(\alpha_p\right)n^{\alpha_p}\right)+\frac{p}{2} \tag{13.38}$$

With an imaginary constant $\alpha_p=(2k\pi i)/\ln(1-p)^{-1}$. The occurrence of the oscillatory term in this expression is quite surprising and, to our knowledge, it is difficult to explain this effect intuitively.

Apart from that, it is also interesting to consider discreteness effects in RV's from continuous distributions. While in our considerations in sections 13.2 and 13.3 a record entry X_n was simply a value that was larger than all previous values $X_0,X_1,...,X_{n-1}$, one can also impose different, more complicated, conditions, where records are only counted if they exceed another barrier depending on $X_0,X_1,...,X_{n-1}$. Some important examples that have been studied in the literature are the following:

- Rounded records: An entry X_n is a strong record if the rounded value $\lfloor X_n\rfloor_\Delta$ exceeds the maximum of all previous entries:

$$\lfloor X_n\rfloor_\Delta>\max\{\lfloor X_0\rfloor_\Delta,\lfloor X_1\rfloor_\Delta,...,\lfloor X_n\rfloor_\Delta\}. \tag{13.39}$$

  Here, $\lfloor\cdot\rfloor_\Delta$ means rounding (up or down) to the next integer multiple of $k\cdot\Delta$ with $k\in\mathbb{Z}$. Similarly, we have a weak record if

$$\lfloor X_n\rfloor_\Delta\geq\max\{\lfloor X_0\rfloor_\Delta,\lfloor X_1\rfloor_\Delta,...,\lfloor X_n\rfloor_\Delta\}. \tag{13.40}$$

  This process is illustrated for exponential RV's in Fig. 13.4 (left).

- δ-records: A δ-record is an entry X_n that exceeds all previous entries $X_0,X_1,...,X_{n-1}$ at least by δ:

$$X_n>\max\{X_0+\delta,X_1+\delta,...,X_{n-1}+\delta\}. \tag{13.41}$$

  Note that δ can, in principle, also be negative. Such a record is called a strong record for $\delta>0$ and a weak record if $\delta<0$. This record model is sketched in Fig. 13.4 (right) for $\delta=1$ and RV's from an exponential distribution.

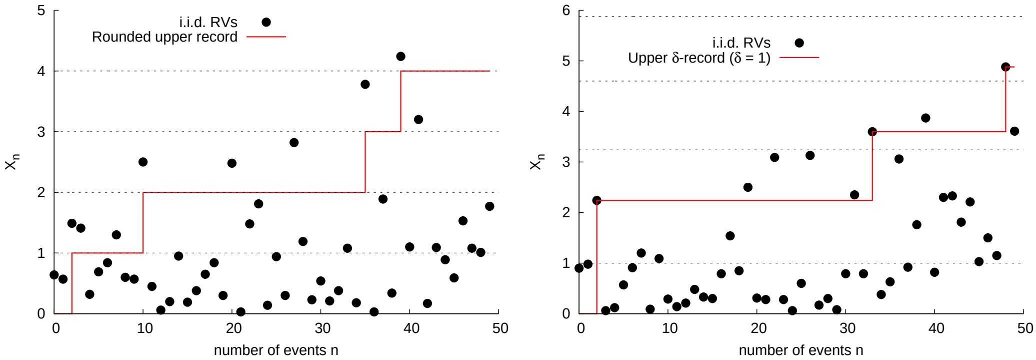

Figure 13.4: **Left:** Sketch of the record process of (exponential) i.i.d. RV's, which are rounded down to the next integer. A continuous RV X_k (dots) is a new record if $\lfloor X_k \rfloor$ exceeds all previous values. The progression of the (upper) record value is given by the red line. **Right:** Sketch of the δ-record process of (exponential) i.i.d. RV's for $\delta = 1$. An entry X_k is counted as a new δ-record, if it is larger than $\max\{X_1, ..., X_{k-1}\} + \delta$. Again, the progression of the (upper) record value is given by the red line.

- Geometric records: For a geometric record, an entry X_n has to exceed a fixed multiple of all previous entries:

$$X_n > \max\{\alpha X_0, \alpha X_1, ..., \alpha X_{n-1}\}, \tag{13.42}$$

  where $\alpha > 0$ is an arbitrary constant. Here, a record is a strong record if $\alpha > 1$ and a weak record for $\alpha < 1$.

In the following, we summarize how the record statistics in these three cases differ from the continuous case. As in the above, the three universality classes of EVS will play an important role. In all three cases, the findings will differ systematically between these classes.

13.4.1 Rounding effects

The record statistics of rounded measurements where first considered by Wergen et al. in 2012 [40]. In a previous study, they analyzed historical temperature measurements from U.S. weather stations [58] that were recorded in whole degrees of Fahrenheit. They found that this discreteness had a significant effect on the record statistics of the temperature data that could, in principle, disguise a possible effect of global warming on the occurrence of record-breaking events [9, 40]. The problem is more general: In all applications, experimental measurements can only be recorded up to a certain accuracy. Usually, one has to deal with RV's, which are sampled from a hypothetical continuous distribution and then discretized by rounding in the measurement process.

In 2012, Wergen et al. [40], studied the strong record rate and the mean record number of i.i.d. RV's for a continuous pdf $f(x)$ that were rounded down to integer multiples of a discretization length Δ. They showed that the strong record rate P_n^{Δ}, in this case, can by computed from the following sum:

$$P_n^{\Delta} = \sum_k \left[F\left((k+1)\Delta\right) - F\left(k\Delta\right)\right] F^{n-1}\left(k\Delta\right). \tag{13.43}$$

This is just the sum over the probabilities for a new record at time n on the individual lattice sites $k\Delta$ (with $k \in \mathbb{Z}$). For $\Delta \to 0$, it is easy to show that P_n^{Δ} approaches the continuous result with $P_n = \int \mathrm{d}x f(x) F^{n-1}(x)$ as in section 13.2.

In the limit of $n \to \infty$, it is possible to analyze the asymptotic behavior of Eq. 13.43 with respect to the universality classes of EVS. For that purpose, we can again use the GPD (Eq.

13.22). Since interesting results can only be expected for a discretization length Δ much smaller than the support of the distribution, we can approach the problem by replacing the sum in Eq. 13.43 by an integral. In this case, however, the bounds of integration have to be chosen carefully. Then, the strong record rate for the GPD is given by

$$P_n^{\Delta} \approx \begin{cases} \int_1^{\frac{1}{\Delta}-1} \mathrm{d}x \left[F\left((k+1)\Delta\right) - F\left(k\Delta\right)\right] F^{n-1}\left(k\Delta\right), & \text{for } \xi < 0 \\ \int_1^{\infty} \mathrm{d}x \left[F\left((k+1)\Delta\right) - F\left(k\Delta\right)\right] F^{n-1}\left(k\Delta\right), & \text{for } \xi \geq 0 \end{cases} \tag{13.44}$$

Here, the upper bound for the Weibull class $\frac{1}{\Delta} - 1$ is simply the finite number of lattice sites in $\left[0, 1-\xi^{-1}\right]$ minus one. In the case of weak records one has to omit the minus one. With Eq. 13.44, we can compute the large n limit of P_n^{Δ} and find that

$$P_n^{\Delta} \xrightarrow[n\to\infty]{} \begin{cases} \frac{1}{n} e^{-n\Delta^{-\frac{1}{\xi}}}, & \text{for } \xi < 0 \quad \text{(Weibull class)} \\ \frac{1}{n\Delta}\left(1 - e^{-\Delta}\right) & \text{for } \xi = 0 \quad \text{(Exp. distribution)} \\ \frac{1}{n} & \text{for } \xi > 0 \quad \text{(Fréchet class)} \end{cases} \tag{13.45}$$

Apparently, the record rate changes systematically in the Weibull class, while, in the Fréchet class, the asymptotic behavior is as in the continuous case. The corresponding results for the weak record rate p_n^{Δ} are the following:

$$p_n^{\Delta} \xrightarrow[n\to\infty]{} \begin{cases} \left(1 + \xi\left(1-\Delta\right)\right)^{-\frac{1}{\xi}}, & \text{for } \xi < 0 \quad \text{(Weibull class)} \\ \frac{1}{n\Delta}\left(e^{\Delta} + e^{-\Delta}\right) & \text{for } \xi = 0 \quad \text{(Exp. distribution)} \\ \frac{1}{n} & \text{for } \xi > 0 \quad \text{(Fréchet class)} \end{cases} \tag{13.46}$$

Here, the asymptotic weak record rate in the Weibull class is constant and equals the probability that a RV falls into the largest lattice site. Both cases show that rounding effects are important for RV's from the Weibull class, while they are negligible in Fréchet class. The behavior in the Gumbel class is, as usual, intermediate between those two. While the (strong and weak) record rates of the exponential distribution are still proportional to $1/n$, Wergen et al. [40] found sublinear corrections to the $1/n$-behavior for the GGD with $f(x) \propto e^{-|x|^{\beta}}$. Here, for $\beta > 1$, the strong and weak record rates decay as

$$P_n^{\Delta} \propto \frac{1}{n} \ln n^{\frac{1}{\beta}-1} \qquad \text{and} \qquad p_n^{\Delta} \propto \frac{1}{n} \ln n^{1-\frac{1}{\beta}}. \tag{13.47}$$

Note that, even though these results were derived for the special case of 'rounding down', they do not change systematically if one considers other kinds of rounding like 'rounding up' or 'rounding to the nearest integer'.

Wergen et al. [40] also considered the interesting regime of very strong discreteness with $\Delta \gg 1$. Here, the occurrence of records becomes predictable on a logarithmic time-scale for certain distributions from the Gumbel class. For a detailed discussion of this phenomenon we refer the reader to [40].

13.4.2 δ-records

The concept of δ-records (or near-records) was discussed by various authors, for instance by Gouet et al. [38, 39, 59] or Balakrishnan et al. [60, 61]. In particular Gouet et al. made important progress on this problem. In [39], they discussed the process of δ-records in detail and proved a limit theorem for the asymptotic distribution of the record number in this case. Instead of describing their rather complex derivations, we will now demonstrate an elementary approach that illustrates the asymptotic behavior of δ-records in time series of RV's from the three classes of EVS in the regime of small $\delta \ll 1$. Our findings are in good agreement with the results of Gouet et al..

In the general case, it is easy to see that the record rate of δ-records can be obtained from the integral

$$P_n^\delta = \int \mathrm{d}x\, f(x)\, F^{n-1}(x-\delta)\,. \tag{13.48}$$

Again, for $\delta = 0$, we obtain the continuous result with $P_n = 1/n$. As in the case of the LDM (see section 13.3) we can now expand this integral for small values of δ and $n \gg 1$. Doing this we find

$$\begin{aligned} P_n^\delta &\approx \int \mathrm{d}x\, f(x) \left(F^{n-1}(x) - \delta\,(n-1)\, f(x)\, F^{n-2}(x)\right) \\ &\approx \frac{1}{n+1} - \delta n \mathrm{I}_n. \end{aligned} \tag{13.49}$$

with the same $\mathrm{I}_n = \int \mathrm{d}x\, f^2(x)\, F^{n-2}(x)$ as in section 13.3. With the results for I_n described in that section, we can now compute the rate of δ-records for RV's from a GPD in the regime of small δ. Here, we find that

$$P_n^\delta \approx \frac{1}{n+1} - \delta \begin{cases} 2\xi\Gamma\,[2+\xi]\, n^{-\xi-1}, & \text{for } \xi \neq 0 \\ \frac{1}{n+1}, & \text{for } \xi = 0 \end{cases} \tag{13.50}$$

For $\xi \geq 0$ (Exponential distribution and Fréchet class) this result is even correct in the limit $n \to \infty$. In the Weibull class with $\xi < 0$ it holds for $\delta \ll n^{\xi}$. The result illustrates nicely how the δ affects the record rate in the $\delta \ll 1$ regime. As in the case of rounding discussed before, the δ is negligible in the Fréchet class and has a strong effect that increases with n in the Weibull class. It is straightforward to show that, in the Weibull class, the record rate will eventually decay exponentially, which leads to a finite asymptotic record number [38, 39].

Again, the case of the Gumbel class is more complicated. For the GGD with $f(x) \propto e^{-|x|^\beta}$ one finds a that, for small $\delta \ll 1$

$$P_n^\delta \approx \frac{1}{n+1} - A_\beta \frac{\delta}{n} \ln(n)^{1-\frac{1}{\beta}}\,. \tag{13.51}$$

With a positive constant A_β, which depends on the tail parameter β. While, for $\beta < 1$, this approximation is valid for arbitrary values of n, it only holds for $\ln(n)^{1-\frac{1}{\beta}} \ll \delta^{-1}$ when β is larger than one. This result indicates that the marginal case of $\beta = 1$ plays an important role.

In fact, it is well known, that the mean spacings $\langle \Delta_k \rangle$ between the subsequent records with record numbers k and $k+1$ from an exponential distribution are equidistant from each other [25]. For all (light-tailed) distributions decaying faster than the exponential, e.g. with $\beta > 1$, these mean spacings are decreasing with increasing k and the record values move closer and closer together. For $\beta < 1$ and (heavy-tailed) distributions broader than the exponential, the spacings increase with k. Only in the regime of $\beta > 1$, the spacings will eventually become smaller than any δ. Eventually, as shown rigorously by Gouet et al. [59], for very large n, this leads to a slow exponential decay of the record rate for all distributions with $\beta > 1$. Using the results of Gouet et al. [59] one can compute the (exact) asymptotic mean record number for the GGD in the large n limit:

$$\langle R_n \rangle \xrightarrow[n\to\infty]{} \begin{cases} \frac{B_\beta}{\delta(\beta-1)} \ln(n)^{1-\frac{1}{\beta}}\, e^{-\delta C_\beta \ln(n)^{\beta+\frac{1}{\beta}-2}}, & \text{for } \beta > 1 \\ \ln(n)\, e^{-\delta}, & \text{for } \beta = 1 \\ \ln(n)\,, & \text{for } \beta < 1 \end{cases} \tag{13.52}$$

with positive constants B_β and C_β depending on β.

13.4.3 Geometric records

The first author who discussed the problem of geometric records was Eliazar in 2005 [62]. A geometric record is a record that is only counted if it exceeds a certain multiple of the previous record. In particular, in order to be a record, X_n has to be larger than $\alpha \cdot \max\{X_0, X_1 ..., X_n\}$ with a positive constant α (not necessarily larger than one). The record rate P_n^α for this problem is given by

$$P_n^\alpha = \int \mathrm{d}x\, f(x)\, F\left(\frac{x}{\alpha}\right)^{n-1}. \tag{13.53}$$

For the exponential distribution with $f(x)$ $(x > 0)$, this integral can be computed exactly. Here we find

$$P_n^\alpha = \int\limits_0^\infty \mathrm{d}x\, e^{-x}\left(1 - e^{-\frac{x}{\alpha}}\right)^{n-1} \xrightarrow[n\to\infty]{} \frac{\alpha\Gamma[\alpha]}{n^\alpha}, \tag{13.54}$$

which reproduces the i.i.d. record rate for $\alpha = 1$. Interestingly, for $\alpha > 1$, this indicates a finite asymptotic mean record number. For the GPD with $\xi \neq 0$, the situation is more complicated and we can not compute the record rate P_n^α directly for other representatives of the distribution. In a recent article, Gouet et al. [59] proved a series of theorems for the record rate of geometric records that can be used to give the asymptotic record rate of the GPD in the geometric case.

Gouet et al. showed that the record rate P_n^α of distributions of the Fréchet class does not differ significantly from the i.i.d. case. In the Weibull class, on the other hand, the asymptotic record rate goes to zero for $\alpha > 1$. In the Gumbel class, the situation is more complicated and, for $\alpha > 1$, the mean record number can be both finite or divergent. Interestingly, in this class, one also finds distributions, were the mean record number goes to infinity with a slower than logarithmic speed. With the results presented in [59], we can infer the record rate P_n^α for the GPD for $n \to \infty$ and $\xi \geq 0$:

$$P_n^\alpha \xrightarrow[n\to\infty]{} \begin{cases} \alpha\Gamma[\alpha]\, n^{-\alpha}, & \text{for } \xi = 0 \quad \text{(Exp. distribution)} \\ \alpha^{-\frac{1}{\xi}} n^{-1}, & \text{for } \xi > 0 \quad \text{(Fréchet class)} \end{cases} \tag{13.55}$$

In the Weibull class, we expect an exponential decay of P_n^α, but, by now, we are not aware of any analytical results for this regime.

13.5 Records in correlated processes

13.5.1 Records in symmetric, discrete-time random walks

An entirely new field of research was established through the work of Majumdar and Ziff [41], who were the first to consider the record statistics of symmetric random walks. In contrast to most of the previous research in the field of record statistics, they considered a correlated process, namely a symmetric, discrete-time random walk (an introduction can be found in [63, 64]), and computed its record rate, its mean record number and also the full distribution of the record number. In the following, we will summarize and discuss their important findings.

A discrete-time random walk (DTRW) $X_0, X_1, ..., X_n$ is a time series with entries of the form

$$X_i = X_{i-1} + \eta_i, \tag{13.56}$$

with i.i.d. increments η_i drawn from a continuous and symmetric distribution $f(\eta)$ (see also Fig. 13.5). Without loss of generality, we can set $X_0 = 0$. Then, by definition, $X_0 = 0$ is also the first record.

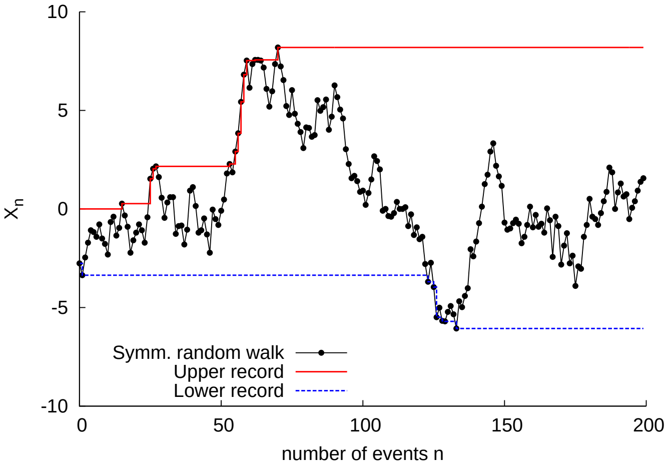

Figure 13.5: Sketch of a symmetric random walk with a Gaussian jump distribution. The red (blue dotted) lines mark the progression of the upper (lower) record of the process.

To compute the record statistics of this process, it is helpful to introduce two generally important quantities, the first-passage probability $\phi(x,n)$ and the survival probability $q(x,n)$ (cf. [65]). The (positive) first-passage probability is the probability that a random walk, starting at 0, crosses $x \geq 0$ in time-step n for the first time:

$$\phi(x,n) := \text{Prob}\left[X_n > x \;\&\; X_0, X_1, ..., X_{n-1} \leq x\right]. \tag{13.57}$$

The related (positive) survival probability $q(x,n)$ is the probability that the random walk remains below x for the first n steps:

$$q(x,n) := \text{Prob}\left[X_0, X_1, ..., X_n \leq x\right]. \tag{13.58}$$

It is easy to see that the first-passage probability can also be obtained by $\phi(x,n) = q(x,n-1) - q(x,n)$.

In the special case of a symmetric DTRW ($f(\eta) = f(-\eta)$) and $x = 0$, these quantities can be computed using an important theorem by Sparre Andersen [66, 67]. He showed that, in this case, the generating function of the survival probability $q(0,n)$, defined as $\tilde{q}(0,z) = \sum_{n=0}^{\infty} q(0,n) z^n$ is given by

$$\tilde{q}(0,z) = \frac{1}{\sqrt{1-z}}. \tag{13.59}$$

Expanding in powers of z this leads to $q(0,n) = \binom{2n}{n} 2^{-2n}$.

This result was the most important requirement for the work of Majumdar and Ziff [41]. They showed that a random walk of length n with R_n records can be described as a chain of $R_n - 1$ first-passage and one survival problem. This is possible because of the so-called *renewal property* of the random walk. After a record at time i, the probability for a record at time $i + j$ is the same as the probability $\phi(0,j)$ that a random walk starting from 0 crosses the origin (from negative to positive) after j steps for the first time. As long as the process stays below the origin set by the record at time i, no further records occur.

Therefore, the probability $P(i_1, ..., i_{R_n}; n)$ for a random walk with records at times $0, i_2, i_3, ..., i_{R_n}$ (with $i_1 = 0$ by definition and $0 < i_2 < i_3 < ... < i_{R_n} \leq n$) can be given by

$$P(i_1, ..., i_{R_n}; n) = \phi(0,i_2) \cdot \phi(0, i_3 - i_2) \cdot ... \cdot \phi\left(0, i_{R_n} - i_{R_{n-1}}\right) \cdot q(0, n - i_{R_n}) \tag{13.60}$$

With this, the distribution $P(R_n|n)$ of the record number R_n can be obtained by summing over all possible sets of inter-record times $0, i_2, i_3, ..., i_{R_n}$ with $0 < i_2 < ... < i_{R_n} \leq n$. The easiest way to compute this sum is via the generating function of $P\left(R_n|n\right)$. Majumdar and Ziff found that $P\left(R_n|n\right)$ obeys

$$\begin{aligned} \sum_{n=R_n-1}^{\infty} P\left(R_n|n\right) z^n &= \left(\tilde{\phi}\left(z\right)\right)^{R_n-1} \tilde{q}\left(z\right) \\ &= \left(1-\left(1-z\right)\tilde{q}\left(z\right)\right)^{R_n-1}\tilde{q}\left(z\right) \end{aligned} \tag{13.61}$$

and, with the survival probability $\tilde{q}\left(z\right) = \sqrt{1-z}^{-1}$ of the symmetric random walk, one finds

$$\sum_{n=R_n-1}^{\infty} P\left(R_n|n\right) z^n = \frac{\left(1-\sqrt{1-z}\right)^{R_n-1}}{\sqrt{1-z}}. \tag{13.62}$$

This result allowed Majumdar and Ziff to extract the exact distribution of the record number R_n:

$$P\left(R_n|n\right) = \binom{2n-R_n+1}{n} 2^{-2n+R_n-1}. \tag{13.63}$$

From this expression, one can easily obtain the mean record number $\langle R_n \rangle$ and the record rate P_n of the symmetric DTRW. For the generating function of $\langle R_n \rangle$, one has to multiply Eq. 13.61 with the record number R_n and sum over all possible values for R_n. This leads to

$$\sum_{n=0}^{\infty} \langle R_n \rangle z^n = \sum_{R_n=0}^{\infty} R_n \left(\tilde{\phi}\left(z\right)\right)^{R_n-1} \tilde{q}\left(z\right) = \frac{1}{\sqrt{1-z}^3}. \tag{13.64}$$

Expanding this result in powers of z we find

$$\langle R_n \rangle = \left(2n+1\right)\binom{2n}{n} 2^{-2n} \qquad \text{and} \qquad P_n = \binom{2n}{n} 2^{-2n}. \tag{13.65}$$

Is is interesting to analyze the asymptotic behavior of these quantities in the limit of $n \to \infty$. Here, the record number approaches a half-Gaussian distribution with

$$P\left(R_n|n\right) \approx \frac{1}{\sqrt{n\pi}} e^{-\frac{R_n^2}{4n}}. \tag{13.66}$$

The mean record number and the record rate converge to

$$\langle R_n \rangle \approx \sqrt{\frac{4n}{\pi}} \qquad \text{and} \qquad P_n \approx \frac{1}{\sqrt{\pi n}}. \tag{13.67}$$

Majumdar and Ziff also considered discrete random walks on a lattice with lattice constant d and a jump distribution $f\left(x\right) = \frac{1}{2}\left(\delta\left(x-d\right)+\delta\left(x+d\right)\right)$. In this case, the asymptotic record statistics is very similar to the continuous case and differs only in a prefactor. For $n \to \infty$, the mean record number and the record rate are reduced by a factor of $1/\sqrt{2}$:

$$\langle R_n \rangle \approx \sqrt{\frac{2n}{\pi}} \qquad \text{and} \qquad P_n \approx \frac{1}{\sqrt{2\pi n}}. \tag{13.68}$$

In the article by Majumdar and Ziff [41] one can also find a discussion of the extreme value statistics of the ages of the longest and shortest lasting records in a symmetric random walk. In particular, they showed that the expected age of the longest lasting record grows proportional to the walk length n and not to $\sqrt{n}$ as the average age of a record and also the age of the shortest lasting record.

13.5.2 Biased random walks

A natural way to generalize the model of a symmetric DTRW considered by Majumdar and Ziff, is to introduce a bias. The entries $X_0, X_1, .., X_n$ of such a biased random walk with a constant drift c are given by

$$X_i = X_{i-1} + \eta_i + c, \tag{13.69}$$

where the η_i's are i.i.d. RV's from a symmetric distribution $f(\eta)$ as in the previous section and again $X_0 = 0$. As in the case of the LDM for uncorrelated time series, the drift causes a non-universal and distribution dependent behavior of the record statistics of the DTRW. The simple, universal version of the Sparre Andersen theorem is not valid in the biased case and the first-passage and survival probabilities of the random walk depend on the choice of the jump distribution $f(\eta)$.

Fortunately, there exists a more general version of Sparre Andersen's theorem that holds also in the biased case. Sparre Andersen [66, 67] showed that the (positive) survival probability with respect to the origin of the biased random walk $q_c(0, n)$ has the generating function

$$\tilde{q}(0, z) = \exp\left(\sum_{n=1}^{\infty} \frac{z^n}{n!} \rho_c(n)\right), \tag{13.70}$$

where $\rho_c(n)$ is the probability that a random walk is negative at time n: $\rho_c(n) := P[X_n < 0]$. In the case of an unbiased random walk with $c = 0$, we have $\rho_0(n) = \frac{1}{2}$, which reduces Eq. 13.70 to the simple symmetric version of the Sparre Andersen theorem introduced in the previous section (Eq. 13.59). In the general case, it is more complicated to compute $\rho_c(n)$ and the probability density of the random walk at time n is needed.

Majumdar et al. [23] found that the asymptotic behavior of $q_c(0, n)$ depends crucially on the tail of the jump distribution $f(\eta)$. Since, the behavior of the distribution $f(\eta)$ for large values of η is dictated by the small k behavior of its Fourier transform $\tilde{f}(k)$, Majumdar et al. considered jump distributions with a Fourier representation of the form

$$\hat{f}(k) \approx 1 + (l_\mu |k|)^\mu \tag{13.71}$$

for small values of k. Here, l_μ is a constant parameter and $\mu \in (0, 2]$ the so-called Lévy-index (also the tail-index) of the jump distribution. While a Lévy-index with $\mu = 2$ corresponds to jump-distributions with a finite second moment $\sigma^2 = \int \mathrm{d}x\ x^2 f(x)$, a value of $\mu < 2$ describes a distribution with heavy-tails, whose variance does not converge. In real-space, the tails of such a distribution decay as $f(x) \propto |x|^{-\mu-1}$ for $x \to \infty$. Distributions with $\mu \leq 1$ have even a divergent mean value $\int \mathrm{d}x\ x f(x)$.

For the simple form of $\tilde{f}(k)$ in Eq. 13.71 one can compute $\rho_c(n)$ by elementary means. Obtaining the generating function $\tilde{q}(0, z)$ in a closed form is, however, far more complicated and requires more sophisticated techniques.

Without going into the details of the computations of Majumdar et al. [23], we will now summarize their results for the survival probability $q_c(0, n)$, the distribution of the record number $P(R_n|n)$, the mean record number $\langle R_n \rangle$ and the record rate P_n of the biased random walk. In fact, they found five different universal regimes depending on the bias c and the Lévy-index μ, in which the asymptotic survival and record statistics have systematically different characteristics. These regimes are the following:

I — The subcritical case ($\mu < 1$): In this regime, the standard deviation of the position of the random walker X_n grows faster than linear and the effects of the drift are therefore negligible in the large n limit. In fact, the survival probability $q_c(0, n)$ is proportional to $1/\sqrt{n}$ as in the unbiased case. The mean record number and the distribution of the record number have the same large n behavior as the symmetric random walk, only their prefactors are different. Also the extremal ages of the shortest and longest lasting records have the same large n asymptotics as in the symmetric case.

II — The marginal case ($\mu = 1$): In this interesting regime, the survival and record statistics have a more complicated dependence on n. In some sense, this is the regime, were the drift and the fluctuations of the process are of the same order. Here, the survival probability $q_c(0,n)$ decays like

$$q_c(0,n) \propto \frac{1}{n^{\Theta(c)}}, \tag{13.72}$$

with a c dependent exponent $\Theta(c) = \frac{1}{2} + \frac{1}{\pi}\arctan(c)$. This non-trivial exponent also appears in the mean record number. Majumdar et al. [23] showed that

$$\langle R_n \rangle \propto n^{\Theta(c)} \qquad \text{and} \qquad P_n \propto n^{\Theta(c)-1}. \tag{13.73}$$

Also the full distribution of the record number $P(R_n|n)$ and the extremal ages of the shortest and longest lasting records depend on this function $\Theta(c)$. In both cases the asymptotic results are more complicated and we refer to [23] for details.

A jump distribution, which falls into this regime, is the Cauchy distribution with $f(x) = \frac{1}{\pi(1+x^2)}$. This special case was already considered by Le Doussal and Wiese [68], prior to the work of Majumdar et al.. They also found the non-trivial exponent $\Theta(c)$ and computed the exact mean record number, as well as its variance, for this case. For a biased random walk with a Cauchy jump distribution the mean record number reads

$$\langle R_n \rangle = \frac{\Gamma[n+2-\Theta(c)]}{\Gamma[n+1]\,\Gamma[2-\Theta(c)]}. \tag{13.74}$$

Interestingly, the function $\Theta(x)$ is also the cumulative distribution $\int^x \mathrm{d}x\, f(x)$ of the Cauchy distribution. The reason for this agreement is, to our knowledge, unclear.

III — The supercritical case with positive drift ($\mu > 1$ & $c > 0$): Here, the survival probability decays faster than in the two previous cases. For $n \to \infty$, $q_c(0,n)$ behaves like

$$q_c(0,n) \propto \frac{1}{n^{\mu}} \tag{13.75}$$

and the mean record number grows linearly with n:

$$\langle R_n \rangle \approx \alpha_\mu(c)\, n \tag{13.76}$$

with a parameter $\alpha_\mu(c)$, which was also computed by Majumdar et al. [23]. The distribution of the record number has an interesting, non-trivial scaling form. In particular, $P(R_n, n)$ is given by

$$P(R_n, n) \xrightarrow[n\to\infty]{} \frac{1}{\alpha_\mu(c)\, n^{\frac{1}{\mu}}} V_\mu\left(\frac{R_n - \alpha_\mu(c)\, n}{\alpha_\mu(c)\, n^{\frac{1}{\mu}}}\right) \tag{13.77}$$

The scaling function $V_\mu(u)$ is of the form

$$V_\mu(u) \approx c_1 u^{\frac{2-\mu}{2(\mu-1)}} e^{-c_2 u^{\frac{\mu}{\mu-1}}} \tag{13.78}$$

with constants c_1 and c_2, which depend only on μ. Here, the age of the longest lasting record grows like the inverse survival probability $\propto n^{-\mu}$ and the age of the shortest approaches a constant value proportional to $1 - \alpha_\mu(c)$.

IV — The Brownian case with positive drift ($\mu = 2$ & $c > 0$): This is the regime of a Brownian random walk with a jump distribution that has a finite variance. Here, a positive drift has a strong effect on the survival probability and the mean record number. In fact, in contrast to regime III, the survival probability decays exponentially. Majumdar et al. [23] showed that

$$q_c(0,n) \propto \frac{1}{n^{\frac{3}{2}}} e^{-\frac{c^2 n}{2\sigma^2}}, \tag{13.79}$$

where σ is the standard deviation of the jump distribution ($\sigma^2 := \int \mathrm{d}x\, x^2 f(x)$). Despite of this systematic difference between the survival probabilities of regime III and IV, the mean record number has the same behavior. As in regime III, we have $\langle R_n \rangle \approx \alpha_2(c)\, n$ (with a distribution dependent prefactor $\alpha_2(c) = \lim_{\mu\to 2} \alpha_\mu(c)$) and an asymptotically constant record rate $P_n = \alpha_2(c) > 0$. Again in contrast to regime III, the distribution of the record number is Gaussian with mean value $\langle R_n \rangle$ and a standard deviation, which grows proportional to $\sqrt{n}$. The age of the longest lasting record grows logarithmically $\propto \ln n$, while the age of the shortest approaches a constant value as in regime III.

V — The supercritical case with negative drift ($\mu > 1$ & $c < 0$): The regime with $\mu > 1$ and negative drift is characterized by an asymptotically constant survival probability as well as a finite record number. Here, the drift eventually dominates the behavior and, beyond a certain time, records will no longer be possible. Majumdar et al. [23] computed the asymptotic survival probability $q_c(0,n) \approx a_\mu(c)$ and showed that the asymptotically finite mean record number is given by the inverse of this value:

$$\langle R_n \rangle \approx \frac{1}{q_c(0,n)} \approx \frac{1}{a_\mu(c)}. \tag{13.80}$$

Here, the parameter $a_\mu(c)$ for $c < 0$ is related to the parameter $\alpha_\mu(c)$ for $c > 0$ from the regimes III and IV. One finds that $a_\mu(c) = \alpha_\mu(|c|)$. The distribution of the record number for $n \to \infty$ has a simple geometric form:

$$P(R_n|n) \approx a_\mu(c)\,(1 - a_\mu(c))^{R_n - 1}. \tag{13.81}$$

Due to the fact that the record number is finite, the ages of the shortest and longest lasting records grow linearly with n in regime V.

In 2011, Wergen et al. [22] also considered the Brownian case (regime IV), but focused on the behavior of the record statistics for finite n in the regime of a small drift c. Wergen et al. showed that, in this case, the survival probability and the record number of any biased random walk with a jump distribution that has a finite variance σ^2 ($\mu = 2$) is very similar to the corresponding quantities of a Gaussian random walk with the same σ^2. Expanding the survival probability $q_c(0,n)$ from Eq. 13.70 up to first order in c, one finds that, for $c\sqrt{n} \ll \sigma$,

$$q_c(0,n) \approx \frac{1}{\sqrt{\pi n}} + \frac{c}{\sqrt{2}\sigma}. \tag{13.82}$$

With the methods described in section 13.5.1 this result can be used to obtain the mean record number and the record rate in the regime of $c\sqrt{n} \ll \sigma$:

$$\begin{aligned} \langle R_n \rangle &\approx \sqrt{\frac{4n}{\pi}} + \frac{c}{\sigma}\frac{\sqrt{2}}{\pi}\left(n \arctan\left(\sqrt{n}\right)\right), & (13.83)\\ P_n &\approx \frac{1}{\sqrt{\pi n}} + \frac{c}{\sigma}\frac{\sqrt{2}}{\pi} \arctan\left(\sqrt{n}\right) & (13.84) \end{aligned}$$

For $n \gg 1$, the arctan-function approaches $\frac{\pi}{2}$, which leads to $\langle R_n \rangle \approx \sqrt{\frac{4n}{\pi}} + \frac{cn}{\sqrt{2}\sigma}$ and $P_n \approx \frac{1}{\sqrt{\pi n}} + \frac{c}{\sqrt{2}\sigma}$.

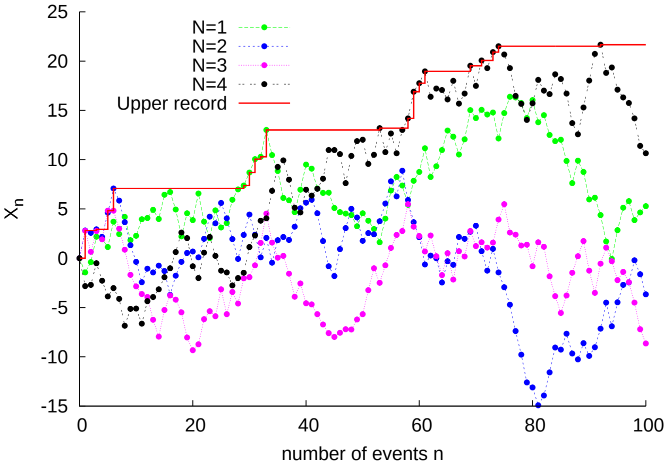

Figure 13.6: Sketch of the record process of the maximum $X_{\max,n}(N)$ of $N=4$ independent (Gaussian) random walks. The progression of the upper record of $X_{\max,n}(N)$ is indicated by the red line.

13.5.3 Multiple random walks

Another way to generalize the fundamental work of Majumdar and Ziff [41] is to consider ensembles of multiple random walks. In 2012, Wergen et al. [40] discussed the record statistics of the maximum $X_{\max,n}(N)$ of N uncorrelated DTRW's with a symmetric jump distribution. For a N random walks with entries

$$X_{i,n} = X_{i,n-1} + \eta_{i,n} \qquad (i = 1, ..., N \text{ and } X_{i,0} = 0) \tag{13.85}$$

and jumps $\eta_{i,n}$ drawn from a single symmetric jump distribution $f(\eta)$. The maximum of N random walks $X_{\max,n}(N)$ is defined as

$$X_{\max,n}(N) = \max\{X_{1,n}, X_{2,n}, ..., X_{N,n}\}. \tag{13.86}$$

Fig. 13.6 illustrates the record process of $X_{\max,n}(N)$ for $N=4$ independent random walks.

Unfortunately, since the maximum of N random walks does not exhibit the same renewal property as the single random walk, it is impossible to compute the record statistics from the survival probability $q(0,n)$ and Sparre Andersen's theorem [66, 67] is not useful here.

Because of that, it is not possible to compute the probability $P(R_n(N), n)$ for $R_n(N)$ records of the maximum $X_{\max,n}(N)$ of the N random walks directly. However, Wergen et al. [40] found a more general way to calculate the record rate $P_n(N)$ that also works in absence of the renewal property.

Wergen et al. [40] argued that the probability that the maximum of N random walks sets a new record with record value x at time n, is given by N times the first-passage probability $\phi(x,n)$ multiplied with the survival probability $q(x,n)$ to the power $N-1$. This is because, $N\phi(x,n)\, q(x,n)^{N-1}$ is just the probability that the value x is first exceeded by one of the N walkers in step n, while the other $N-1$ stay below x. Integration over all possible record values x leads to

$$P_n(N) = N \int_0^\infty \mathrm{d}x\, \phi(x,n)\, q(x,n)^{N-1}. \tag{13.87}$$

Therefore, to compute the record rate, we need the more general survival and first-passage probabilities $q(x,n)$ and $\phi(x,n)$ for an arbitrary $x > 0$.

Wergen et al. [40] computed the asymptotic behavior of these quantities for $n \to \infty$ using a highly non-trivial theorem due to Ivanov [69], which is, in some sense, a very general form of Sparre Andersen theorem. To keep this essay readable, we will not describe the calculations of Wergen et al. and only summarize their results: The large n behavior of $q(x,n)$ and $\phi(x,n)$ depends again on the Lévy-index μ (see section 13.5.2 and Eq. 13.71) and one finds two universal regimes:

I — The Brownian case with finite σ^2 ($\mu = 2$): Here, for $n \to \infty$, the first passage and survival probabilities approach the following forms:

$$\phi(x,n) \approx \frac{x}{\pi\sigma^2 n^{\frac{3}{2}}} e^{-\frac{x^2}{2\sigma^2 n}} \qquad \text{and} \qquad q(x,n) \approx \operatorname{erf}\left(\frac{x}{\sqrt{2\sigma^2 n}}\right). \tag{13.88}$$

With these results one can compute the record rate directly. For a large number of random walks $N \gg 1$ one finds:

$$P_n(N) \xrightarrow[n\to\infty]{N\to\infty} \sqrt{\frac{\ln N}{N}}. \tag{13.89}$$

Apparently, for $n, N \gg 1$, the record rate of N random walks is given by the record rate $P_n(N=1)$ of the single random walk times $\sqrt{\pi \ln N}$. Similarly, the mean record number of $N \gg 1$ random walks approaches $\langle R_n(N)\rangle \approx \sqrt{4n\ln N} \approx \sqrt{\pi \ln N}\langle R_n(1)\rangle$.

II — Lévy flights with divergent σ^2 ($\mu < 2$): In this regime, it is not possible to compute the exact asymptotic behavior of $q(x,n)$ and $\phi(x,n)$. However, Wergen et al. [40] managed to extract the scaling behavior of the product $\phi(x,n)\, q(x,n)^{N-1}$ in Eq. 13.87 and found that

$$P_n(N) \xrightarrow[n\to\infty]{N\to\infty} \frac{2}{\sqrt{\pi n}} \qquad \text{and} \qquad \langle R_n(N)\rangle \xrightarrow[n\to\infty]{N\to\infty} \frac{4\sqrt{n}}{\sqrt{\pi}}. \tag{13.90}$$

In this case, these quantities become completely independent from N and are exactly twice as large as the corresponding results for $N = 1$. The emergence of the prefactor 2 and the complete independence of N are very interesting findings, which are not entirely understood by now.

Wergen et al. [40] also considered the distribution of the record number $P(R_n(N), n)$ in these two cases. As discussed before, because of the lacking renewal property, it is not possible to compute this distribution directly. However, in case I with $\mu = 2$, it is possible to conjecture the asymptotic distribution $P(R_n(N), n)$ from the corresponding distribution of ensembles of random walks with a discrete jump distribution $f(x) = \frac{1}{2}(\delta(x+1) - \delta(x-1))$. For such an ensemble of lattice random walks, the record number is simply given by the maximum $M_n(N)$ of the process. Since the maximum of a single lattice random walk has a finite variance, the maximum value of N random walkers must be distributed according to a Gumbel distribution in the limit of large n and N. In fact, one finds that the mean value of the maximum $\langle M_n(N)\rangle$ of N random walkers converges to $\langle M_n(N)\rangle \approx \sqrt{2n\ln N}$. The mean record number of the N random walkers with a continuous jump distribution only differs by a prefactor of $\sqrt{2}$. Using this analogy one can infer that the record number in the continuous case has the following Gumbel distribution:

$$P(R_n(N)\,|n) \approx e^{-z} e^{-e^{-z}}, \quad \text{with } z = \frac{R_n(N) - \sqrt{4n\ln N}}{\sqrt{n\ln N}}. \tag{13.91}$$

This conjecture was confirmed numerically in [40]. For Lévy flights with a heavy-tailed jump distribution ($\mu < 2$) it was not possible to compute $P(R_n(N)\,|n)$ by similar means. However, performing numerical simulations, Wergen et al. [40] found that this distribution seems to be entirely independent from the Lévy-index $\mu \in [0, 2)$ and the number of walkers $N \gg 1$. Because of this universality, it would be very interesting to find an analytical expression for this hitherto unknown distribution.

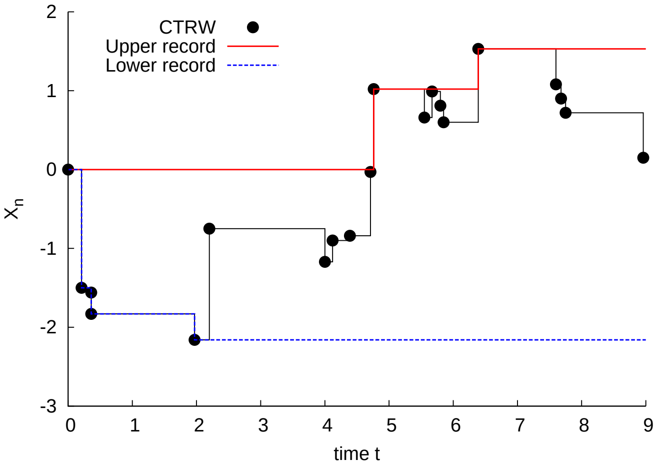

Figure 13.7: Sketch of the record process of a continuous time random walk with random waiting times between the jumps (Here, we sampled the waiting times from an exponential distribution). The progression of the upper (lower) record is indicated by the red (blue dotted) line.

13.5.4 Continuous-time random walks

As another natural generalization of the symmetric DTRW studied by Majumdar and Ziff [41] (section 13.5.1), Sabhapandit [43] studied continuous-time random walks (CTRW's). A CTRW is a process with entries $X(t_0), X(t_1), ..., X(t_n)$ that are recorded at random times $t_0 < t_1 < ... < t_n$ seperated by random waiting-times $\tau_i := t_i - t_{i-1}$ sampled from a (continuous) waiting time distribution $\rho(t)$ (see Fig. 13.7). In this context, a simple discrete-time random walk can be seen as a process with entries $X(0), X(1), ..., X(n)$ at fixed times $t_0 = 0, t_1 = 1, ..., t_n = n$ with a degenerate waiting-time distribution $\rho(t) = \delta(t-1)$.

In the general case, the number of entries in a CTRW is apparently random, this makes it more complicated to consider the statistics of records. The number of records $R(t)$ at an arbitrary, continuous time t is defined as

$$R(t) := \max\{R_n | t_n \leq t\}, \tag{13.92}$$

Sabhapandit discussed the record statistics of CTRW's in the limit of $t \to \infty$. He found, that the asymptotic behavior depends on the tail of the waiting-time distribution $\rho(\tau)$ and therefore introduced the Laplace transform $\tilde{\rho}(s)$ of $\rho(\tau)$:

$$\tilde{\rho}(s) := \int_0^\infty \mathrm{d}\tau\ \rho(\tau)\, e^{-s\tau}. \tag{13.93}$$

The behavior of the waiting-time distribution for large values of t is encapsulated in the small $s \to 0$ behavior of is Laplace transform. In general $\tilde{\rho}(s)$ can be expanded in powers of s and, for small s, one finds $\tilde{\rho}(s) \approx 1 - (\tilde{\tau}s)^\alpha$ with parameters $\tilde{\tau}$ and α depending on the tail of the distribution $\rho(\tau)$. For waiting-time distributions with a finite mean value (in this case $\tilde{\tau}$) one finds $\alpha = 1$, a heavy-tailed waiting-time distribution without a first moment yields an α between 0 and 1.

Sabhapandit showed that the first-passage probability of the CTRW can be obtained from the existing result for the discrete time random walk and computed the asymptotic record statistics for the two cases of $\alpha = 1$ and $0 < \alpha \leq 1$.

In first case, for a finite mean waiting time $\tilde{\tau}$, the asymptotic record statistics of the CTRW is the same as in the time-discrete case. Here, for $t/\tilde{\tau} \to \infty$, the record number

$R(t)$ is distributed according to the half-Gaussian distribution

$$P\left(R\left(t\right)|t\right) \approx \frac{1}{\sqrt{\pi}}\left(\frac{t}{\tilde{\tau}}\right)^{-\frac{1}{2}} \exp\left(-\left(\frac{t}{\tilde{\tau}}\right)^{-1}\frac{R\left(t\right)^{2}}{4}\right), \tag{13.94}$$

which, for $\tilde{\tau} = 1$, is exactly the record number distribution found in the discrete case. Similarly, for $\alpha = 1$, the mean record number and the record rate are given by

$$\langle R\left(t\right)\rangle \approx \frac{2}{\sqrt{\pi}}\left(\frac{t}{\tilde{\tau}}\right)^{\frac{1}{2}} \qquad \text{and} \qquad P\left(t\right) \approx \frac{1}{\sqrt{\pi}}\left(\frac{t}{\tilde{\tau}}\right)^{-\frac{1}{2}}. \tag{13.95}$$

In the case of a divergent mean waiting time with $\alpha < 1$, the record number $R(t)$ approaches a different asymptotic distribution:

$$P\left(R\left(t\right)|t\right) \approx \frac{2}{\alpha}\left(\frac{t}{\tilde{\tau}}\right)^{\frac{1}{2}} R\left(t\right)^{1-\frac{2}{\alpha}} L_{\frac{\alpha}{2}}\left(\left(\frac{t}{\tilde{\tau}}\right)^{-1} R\left(t\right)^{-\frac{2}{\alpha}}\right), \tag{13.96}$$

where $L_{\frac{\alpha}{2}}(z)$ is the pdf of a one-sided Lévy-stable distribution, which, in general, can not be expressed analytically. The Laplace transform of $L_{\frac{\alpha}{2}}(z)$ is given by $\tilde{L}_{\frac{\alpha}{2}}(s) = e^{-s^{-\frac{\alpha}{2}}}$. For $\alpha < 1$, the mean record number and the record rate grow more slowly with n. Sabhapandit [43] found that, for an arbitrary $\alpha \in (0, 1]$,

$$\langle R\left(t\right)\rangle \approx \frac{2}{\alpha\Gamma\left[\frac{\alpha}{2}\right]}\left(\frac{t}{\tilde{\tau}}\right)^{\frac{\alpha}{2}} \qquad \text{and} \qquad P\left(t\right) \approx \frac{1}{\tilde{\tau}\Gamma\left[\frac{\alpha}{2}\right]}\left(\frac{t}{\tilde{\tau}}\right)^{-\frac{\alpha}{2}-1}, \tag{13.97}$$

in good agreement with his findings for $\alpha = 1$ in Eq. 13.95.

13.5.5 Records in higher-dimensional processes

In 2011, Edery et al. [44], considered discrete-time random walks in two and three dimensions and discussed the record statistics of the distance of such a process from the origin. In the case of a one-dimensional random walk this distance at the time n is just given by $|X_n| = \sqrt{X_n^2}$. Already in this case it was not yet possible to compute the exact record rate and the distribution of the record number analytically.

Edery et al. were interested in DTRW's on an orthogonal lattice in two and three dimensions. At each time step, such a random walker jumps from one lattice site to an adjecent site in a random direction. They analyzed the number of records in the series of distances $|\vec{X}_0|, |\vec{X}_1|, ..., |\vec{X}_n|$ from the origin using numerical simulations.

Edery et al. began with a discussion of a symmetric lattice random walk with a symmetric distribution of the jumps. In this case, they could demonstrate that the mean record number of this process has the same scaling behavior as in the case of the discrete-time random walk in one dimension. Without a bias the mean record number grows proportional to $\sqrt{n}$.

In the case of a biased lattice random walk, with a drift in an arbitrary direction, the asymptotic behavior changes. In all three considered dimensions, the asymptotic mean record numbers grows linearly in n.

13.6 Applications

13.6.1 Climate records

The most popular application of the theory of records in the last years was certainly the study of temperature records. The evident and most likely man-made increase of the global mean temperature over the last decades [70] raised the question about the effects of this climatic change on the occurrence and the magnitude of extreme and record-breaking events [4, 71–78]. While, it is intuitively clear to assume that a warming climate also leads to more heat and less cold records, the first systematic application of theoretical results from record statistics was presented by Benestad in 2003 and 2004 [5, 6]. He compared the record process of monthly temperature mean values from Scandinavian weathers stations with i.i.d. RV's and found a small, but significant increase in the number of heat records. Interestingly, he also considered daily precipitation sums, where he could not determine any non-stationary behavior of the record rate [5, 79].

In 2006, Redner and Peterson considered daily temperature measurements from a single weather station in Philadelphia. Even though their data set covered more than 100 years, they had difficulties to quantify the effect of global warming on the measurements from this station. Nevertheless, Redner and Peterson made important progress on the matter. In fact, they proposed a simple model of a Gaussian distribution with a linear drift to describe the record statistics of temperature measurements for individual calendar days. Within this model, a daily (mean, minimum or maximum) temperature T_n in the nth year of an observation period is sampled from a Gaussian distribution with an increasing mean value $\mu_0 + ct$. Here, c is the drift, which is basically the speed of warming. Then, the probability density of the daily temperatures should be of the form

$$f\left(T_n\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(T_n-\mu_0-ct\right)^2}{2\sigma^2}}, \tag{13.98}$$

where σ is the standard deviation that describes the fluctuations of the measurement around the moving mean value. Apparently, this is exactly the Linear Drift Model (LDM) we discussed in section 13.3 for Gaussian RV's.

The work of Redner and Peterson motivated many others to study both, the statistics of record breaking temperatures [8–12, 80] and also the simple LDM [31–33]. In 2009, Meehl et al. [8] analyzed a large number of U.S. weather stations with respect to the occurrence of heat and cold records and found a significant effect of global warming in the ratio of the two of them. In 2010, Wergen and Krug [9] confirmed these findings in an independent study of European station and re-analysis data [58, 81]. The work of Newman et al. [10], Anderson and Kostinski [80], Elguindi et al. [11] as well as Rahmstorf and Coumou [12] lead to similar results.

The main subject of these studies was a comparison between time series of temperature measurements for individual calendar days or months and uncorrelated RV's from the LDM or related, slightly more complicated models. In fact, it is by now well established that a Gaussian LDM, despite its simplicity, can describe the effect of global warming on the occurrence of daily temperature records relatively accurate. From 1960 to 2010, the global mean temperature increased in a roughly linear manner [4, 70]. This effect is also found in measurement from European and U.S. weather stations and re-analysis data. In the same time, the standard deviation of the daily and monthly temperatures around their moving mean value remained more or less constant [4, 9, 70]. Since the magnitude of the warming in recent years is still much smaller than the average standard deviation of daily temperatures, one can compare the temperature measurements with a Gaussian LDM in the regime of small $c \ll \sigma/\sqrt{n}$ (see section 13.3). Here, with the findings of Franke et al. [33], the record rate $P_n\left(c\right)$ in a time series with a linear drift (warming) c and standard deviation σ is given by:

$$P_n\left(c\right) \approx \frac{1}{n} + \frac{c}{\sigma}\frac{2\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)}. \tag{13.99}$$

Apparently, in this approximation, the important degree of freedom is the *normalized drift* $\tilde{c} := \frac{c}{\sigma}$. For a fixed n, a large increase in the record rate can be caused by a large positive drift c, or a small standard deviation σ.

In most of the considered daily data sets, the normalized drift between 1960 and 2010 was of the order $\tilde{c} \approx 0.01\ \mathrm{y}^{-1}$. For this value, Eq. 13.99 predicts an increase in the record rate of about 27% after 30 years and of more than 50% after 50 years of warming. These predictions are in good agreement with the data found in the literature. For instance, Wergen et al. [9] considered daily maximum temperatures measured at 202 European stations [81] between 1975 and 2005 and found an increase of around 40% in the number of heat records along with a normalized drift of $\tilde{c} \approx 0.015\ \mathrm{y}^{-1}$.

For monthly mean temperatures the normalized drift is usually larger, since the standard deviation of these averaged values is much smaller than in the case of daily measurements. Therefore, the rate of monthly upper records can be many times as high as expected in the case of a stationary climate [80, 82]. Of course, for annual mean values this effect is even stronger and, over the last decade, the rate of new global mean temperature records was increased by around 2800% [76].

As shown by Wergen et al. [9] and recently, in more detail, also by Elguindi et al. [11], the normalized drift $\tilde{c}$ has strong regional variations. In Europe, $\tilde{c}$ seems to be generally larger than in the U.S., where $\tilde{c}$ for the daily data is usually much smaller than $0.01\ \mathrm{y}^{-1}$ [9]. Due to the high heat capacity of water, the standard deviation of the daily measurements is much smaller near or over the oceans. Because of that, time series of these measurements can have a very large $\tilde{c}$ and therefore a very strong effect of global warming on the record rate. Stations far away from the sea, on the other hand, can have a very high standard deviation, which reduces the effect of the drift on the record rate. This explains, why the increase of the record rate in Europe was much stronger than in the U.S. [9]: The temperature fluctuations at the U.S. stations, in particular of those in the middle of the continent, are much larger than at the European stations and therefore the normalized drift $\tilde{c}$ of the U.S. stations is much smaller.

In a recent study, Wergen et al. [82] discuss the statistics of the values of record breaking temperatures. While the preceding literature focused only on the occurrence of temperature records, it is also possible to describe the effect of global warming on the values of these records using the LDM. However, here, a Gaussian LDM is appropriate only in the summer months, where the values of heat records are significantly increased by the warming. In the winter, especially in cold, sub-polar regions, the distribution of daily temperatures is highly asymmetric and favors extremely cold temperatures in comparison with the Gaussian case. Therefore, despite of global warming, cold records are still much further away from the temperature mean value than heat records [82]. In other words, despite of global warming, we can still expect extreme, record-breaking cold days in winter.

13.6.2 Records in finance

While, in the study of temperature records, the observational data was compared to uncorrelated random variables, the financial markets yield good example for highly correlated processes. A simple way to model a stock, which is tradable at a stock market, is the Geometric Random Walk model (GRM), which was, in a slightly different form, already proposed by Le Bachelier in 1900 [83]. The GRM describes the logarithms of stocks prizes with a simple biased random walk (as in section 13.5.2). In recent work Wergen et al. [22, 42], as well as Bogner [84], discussed the record statistics of daily stock data from the U.S. Standard and Poors 500 (S&P 500) stock index [85] in the context of this model.

They considered the logarithms of time series of daily stock prizes $S_0, S_1, ..., S_n$. Within the GRM, these logarithmic prizes $\ln S_n$ should behave like a biased random walk with

$$\ln S_i = \ln S_{i-1} + \eta_i + c \tag{13.100}$$

with random jumps (daily returns) η_i from a symmetric (return-) distribution $f(\eta)$ and a bias c. The bias represents a systematic, long-term growth in the system and leads, asymptotically, to an exponential growth of S_n. Since the logarithm is monotonic, a record in a series of stock prizes $S_0, S_1, ..., S_n$ is also a record in the series $\ln S_0, \ln S_1, ..., \ln S_n$ and one can use the results for the record statistics of biased random walks to model the records of the stock prizes.

As it turns out, the GRM is, to a certain degree, useful to predict the record number of daily stock prizes from the S&P 500. Wergen et al. [22] considered series of daily stock prizes of 366 stocks from this index and analyzed the progression of the mean record number $\langle R_n \rangle$ of these stocks.

To compare with the analytical findings for the biased random walk, they determined the drift c and the standard deviation σ of the jump distribution of the daily returns from the stock data. In the relevant regime of $c\sqrt{n} \ll \sigma$, they predicted a mean record number of (see section 13.5.2)

$$\langle R_n \rangle \approx \sqrt{\frac{4n}{\pi}} + \frac{1}{\sqrt{2}}\frac{cn}{\sigma}. \tag{13.101}$$

Again, as in the case of temperature records in a warming climate, the relevant parameter that describes the effects of the bias is the normalized drift $\tilde{c} = \frac{c}{\sigma}$. For the S&P 500 stock data one finds a value of $\tilde{c}$ between $0.015\ d^{-1}$ and $0.025\ d^{-1}$ depending on the length of the observation period.

Wergen et al. [24, 32] found that Eq. 13.101 predicts the qualitative behavior of the mean record number of the stock prizes to some accuracy. Even though it slightly overestimates both the number of upper and the number of lower records, the difference between the two is modeled correctly.

In a similar study, Wergen et al. [40] considered ensembles of multiple stocks from the S&P 500 and compared them with their analytical results for multiple independent random walkers (section 13.5.3). They rescaled and detrended the daily stock data to make them comparable with symmetric random walks with a jump distribution of standard deviation unity. Ensembles of N of these rescaled stocks were then compared with the analytical result for the mean record number of the maximum of N independent random walks (see section 13.5.3):

$$\langle R_{n,N} \rangle \approx 2\sqrt{n \ln N}. \tag{13.102}$$

Interestingly, one finds that the maximum of N detrended and rescaled stocks grows also proportional to $\sqrt{n \ln N}$, but has a different prefactor smaller than the one of the independent random walks. In [40] this was tentatively interpreted with a smaller, effective number of stocks that are stochastically independent in the context of record statistics. In view of the ongoing research on the important role of correlations between stocks in financial markets, it would be interesting to better understand the meaning of the effective number in the future.

13.6.3 Physics and biology

Interestingly, in physics and also in evolutionary biology, one finds several complex dynamical systems that behave like the record process of i.i.d. RV's. In particular, some diffusive processes in random environments, like the random energy landscape by Derrida [86], can be described using record statistics. These systems are usually stable on short time-scales, but run through intermittent events, so-called quakes, which bring them from one stable state to another. As it turns out these quakes can be modeled as record events in time series of i.i.d. RV's.

An important feature of the record process of i.i.d. RV's is that it can be described as a Poisson process in logarithmic time. According to Sibani and Littlewood [87] (see also the reviews by Jensen [13] and Anderson et al. [18]), the distribution of the logarithmic waiting times $\Delta_k := \ln t_k - \ln t_{k-1}$ between the $(k-1)$st and the kth record is given by the exponential distribution with the pdf $\rho(\Delta) = e^{-\Delta}$. With this one can show that the probability $P_k(t)$ of having $k \gg 1$ records up to time $t \gg k$ is given by

$$P_k(t) \approx \frac{1}{t}\frac{(\ln t)^{k-1}}{(k-1)!} \tag{13.103}$$

This kind of log-Poisson statistics is also found in various dynamical systems like, for instance, the Edwards Anderson spin-glass model [13–15]. The time-evolution of the (local) energy $E(t)$ of such a spin-glass, which relaxes towards a lower energy state after an initial quench, is characterized by a series of local energy minima $E_{\mathrm{min}}(k)$ and maxima $E_{\mathrm{max}}(k)$. In order to get from one stable state with energy $E_{\mathrm{min}}(k)$ to the next with energy $E_{\mathrm{min}}(k+1)$, the system needs to overcome an energy barrier with $\Delta E_k = E_{\mathrm{max}}(k) - E_{\mathrm{min}}(k)$. Therefore, the noise driven system needs a fluctuation of the size ΔE_k to relax to the next stable state. Now, as shown by Sibani et al. [13–15], these energy barriers ΔE_k are usually monotonically increasing in k and one finds $\Delta E_k < \Delta E_{k+1}$. Because of that, the fluctuation necessary to overcome ΔE_{k+1} has to be (slightly) larger than the one needed for ΔE_k. In other words, it requires a record-breaking event in the series of fluctuations for the system to relax further. Since these fluctuations are usually assumed to be i.i.d. RV's, one can assume that the process of jumps (quakes) from one stable state to another has the same time-evolution as the record process of i.i.d. RV's, which is describe by Eq. 13.103.

A similar behavior is found in the so-called Restricted Occupancy Model, which was proposed in the context of the theory of type-II superconductors [16, 18]. This model describes the gradual magnetization of a superconducting sample through an external magnetic field. In this context, the number of flux vortices inside a three dimensional model of the type-II superconductor increases step-wise and monotonically in time. The occurrence of the steps (quakes) in the vortex number also exhibits the same log-Poisson statistics as the record process of i.i.d. RV's.

As it turns out, the record process of i.i.d. RV's can be used to describe various dynamical models related to the random energy model. In the context of evolutionary biology, several authors studied the connection between record statistics and adaptation on the fitness landscapes of genotypes. Such a landscape maps the fitness associated with a certain genotype to a (high-dimensional) cubic lattice similar to a lattice of spins. Kaufmann and Levin [17], Sibani et al. [88], as well as Krug and Jain [19] discussed mutations on random fitness landscapes with i.i.d. fitness values and compared the rate of their occurrence with the record rate of i.i.d. RV's. The main idea is that, in order to survive and take over a population, a mutant with random fitness has to be fitter than all previous mutants in an evolutionary process. Therefore he must be a mutant with record-breaking fitness.

A detailed discussion of the applications of record statistics in evolutionary biology can, for instance, be found in [89].

13.6.4 Athletic records

Even though, the occurrence of records in sports, like, for instance, in athletics or in swimming, receives an enormous amount of public attention, only very few have studied the statistical properties of these sport records so far. In the context of the ongoing controversy about the role of legal and illegal doping on the performance of athletes, the theory of records provides a method to distinguish between statistical fluctuations and real improvements. Because of the universal features of the record statistics of i.i.d. RV's, one can analyze the occurrence of records in time series of sports results without detailed knowledge about the underlying distribution from which the results are sampled. In principle, if the number of records, in a series of sports results, is significantly larger (or smaller) than in an i.i.d. series of comparable length, this can not be in agreement with a constant performance level of the athletes. However, when analyzing historical data, it is hard to determine the total number of attempts in a certain event and it is therefore difficult to determine the record rate P_n. Usually only a small number of very good performances is recorded on the leaderboard and one can only analyze the statistics of their values.

By now, the only systematic analysis of athletic records was published by Gembris et al. [2, 3]. They considered the evolution of the record values of several track and field events and compared them to theoretical results for the maximal values of Gaussian i.i.d. RV's.

They estimated the mean value and the standard deviation of the athletic performances for the time series of individual events. Then they compared the record events in the athletic data with series of Gaussian RV's with the same parameters. A comparison of the record values allows to identify events, where the athletes improved significantly over the duration of the time series. It turns out that only in some events the record values actually improve faster than expected on the basis of constant athletic capabilities. In 50% to 80% of all considered track and field events, Gembris et al. [2, 3] could not disprove their null hypothesis of a stationary distribution of athletic performances. Interestingly, in the cases where they could detect a systematic time-dependence, the increase in the performance seemed to be far from linear in time. If the performances of athletes would improve due to better training, nutrition, or just a growing population, one would expect a continuous effect on the record rate. Instead, the progressions of some athletic record values, especially in long distance running and in throwing, are characterized by large jumps, which are probably best explained by instantaneous effects like the introduction, or the prohibition of certain drugs. In fact, while the record values of several long distance running events, like 5000m or 10000m improved drastically after the introduction of blood doping with erythropoietin, almost no records were set in the throwing events since anabolic drugs became detectable in urine samples [2, 3].

Another interesting problem in this context, is the question about an absolute limit to world records in sports, like, for instance, a hypothetical speed that can not be exceeded by humans, which would lead to a lower boundary for the possible outcome of a 100m dash race. Up to now, several different methods were applied to find such a boundary, but the issue is still controversial [90–93]. It might be possible to answer this question using extreme-value or record statistics, since these exhibit different universal properties for distributions with a bounded and an infinite support. A first step towards this goal was done by Einmahl and Magnus in 2008 [94]. They estimated the tail behavior of the distributions of performances in track and field events by comparing them to a Generalized Pareto Distribution. In most considered cases, they could find an absolute limit to the world records. It would interesting to confirm and improve their findings in future studies.

13.7 Summary and outlook

In this review, we tried to summarize numerous interesting and non-trivial results on the statistics of records, which were discovered in the last couple of decades. Especially in the last 10 years, the study of record-breaking events has become a broad and diverse field of research. Additionally, the occurrence and the properties of records were analyzed and discussed in a vast number observational data sets. Researchers have understood that records are often more than just interesting to the observer: One can learn a lot about the properties of a complex dynamical system by considering the record events it generates.

In this context, there a many open problems, which might suit as subjects for future research. The research on the record statistics of uncorrelated random variables with time-dependent distributions has just began and only the very simple cases of a constant linear drift and an increasing standard deviation have been understood to some degree. Of course, one can ask, how the record rate and also the full distribution of the record number is affected by a more complicated time dependence of the underlying distribution. For the Linear Drift Model and the Increasing Variance Model discussed in this review, it would also be interesting to compute the mean values of records that have a certain record number, or occur at a certain time.

Furthermore, the record statistics of correlated random variables are only understood for a few special cases. Especially in the context of possible applications in finance, it would be very interesting to calculate the record rate of more complicate non-Markovian processes. Some interesting candidates for future research are branching random walks, the absolute value of a random walk and the Ornstein-Uhlenbeck process, which is particularly important for the modeling of financial data [95]. In general, it would be desirable to better understand the effect of long-term, or power-law type correlations on the record statistics of stochastic processes.

With respect to the various applications of the theory one easily finds numerous interesting open question. For instance in climatology, it is hardly understood how specific weather conditions affect the occurrence of records. Here, it is also still unclear if one can find a significant effect of climatic change on the record statistics of precipitation events or, for instance, also record-breaking storms. In finance, our understanding of the statistics of record-breaking stock prizes does still not explain some interesting deviations from the classical model of a geometric random walk. It is a challenging problem to find a more accurate description for this record process.

Acknowledgements

The author is grateful to Joachim Krug for many helpful discussions and support on the subject of this review and Ivan Georg Szendro Terán for his input towards the completion of this work.

Bibliography

[1] C. Glenday, ed., *Guinness World Records* (Jim Pattison Group, 2012).
[2] D. Gembris, J. G. Taylor, and D. Suter, Nature **417**, 506 (2002).
[3] D. Gembris, J. G. Taylor, and D. Suter, J. Appl. Stat. **34**, 529 (2007).
[4] P. A. Stott, D. A. Stone, and M. R. Allen, Nature **432**, 610 (2004).
[5] R. E. Benestad, Climate Res. **25**, 3 (2003).
[6] R. E. Benestad, Global Planet. Change **44**, 11 (2004).
[7] S. Redner and M. R. Peterson, Phys. Rev. E **74**, 061114 (2006).
[8] G. A. Meehl *et al.*, Geophys. Res. Lett. **36**, L23701 (2009).
[9] G. Wergen and J. Krug, EPL **92**, 30008 (2010).
[10] W. I. Newman, B. D. Malamud, and D. L. Turcotte, Phys. Rev. E **82**, 066111 (2010).
[11] N. Elguindi, S. A. Rauscher, and F. Giorgi, Climatic Change **114**, 1 (2012).
[12] S. Rahmstorf and D. Coumou, Proc. Natl. Acad. Sci. USA **108**, 17905 (2011).
[13] H. J. Jensen, Adv. Solid State Phys. **45**, 95 (2006).
[14] P. Sibani, G. F. Rodriguez, and G. G. Kenning, Phys. Rev. B **74**, 224407 (2006).
[15] P. Sibani, Eur. Phy. J. B **58**, 483 (2007).
[16] L. P. Oliveira *et al.*, Phys. Rev. B **71**, 104526 (2005).
[17] S. A. Kauffman and S. Levin, J. Theor. Biol. **128**, 11 (1987).
[18] P. Anderson *et al.*, Complexity **10**, 49 (2004).
[19] J. Krug and K. Jain, Physica A **358**, 1 (2005).
[20] J. Franke, PLoS Comp. Biol. **7**, e1002134 (2011).
[21] T. O. Richardson *et al.*, PLoS One **5**, e9621 (2010).
[22] G. Wergen, M. Bogner, and J. Krug, Phys. Rev. E **83**, 051109 (2011).
[23] S. N. Majumdar, G. Schehr, and G. Wergen, J. Phys. A: Math. Theor. **45** (2012).
[24] G. Wergen, “unpublished,” (2012).
[25] B. C. Arnold, N. Balakrishnan, and H. N. Nagraja, *Records*, 1st ed. (Wiley-Interscience, 1998).
[26] V. B. Nevzorov, *Records: Mathematical Theory* (American Mathematical Society, 2004).
[27] N. Glick, Am. Math. Mon. **85**, 2 (1978).
[28] R. Ballerini and S. Resnick, J. Appl. Probab. **22**, 487 (1985).

[29] R. Ballerini and S. Resnick, Adv. Appl. Probab. **19**, 801 (1987).
[30] K. Borovkov, J. Appl. Probab. , 668 (1999).
[31] J. Franke, G. Wergen, and J. Krug, J. Stat. Mech.: Theor. Exp. **P10013** (2010).
[32] G. Wergen, J. Franke, and J. Krug, J. Stat. Phys. **144**, 1206 (2011).
[33] J. Franke, G. Wergen, and J. Krug, Phys. Rev. Lett. **108**, 064101 (2012).
[34] J. Krug, J. Stat. Mech.: Theor. Exp. **07**, 07001 (2007).
[35] I. Eilazar and J. Klafter, Phys. Rev. E **80**, 061117 (2009).
[36] W. Vervaat, Stoch. Proc. Appl. **1**, 317 (1973).
[37] H. Prodinger, Discrete Math. **153**, 253 (1996).
[38] R. Gouet, F. J. Lopez, and G. Sanz, Adv. Appl. Probab. **37**, 781 (2005).
[39] R. Gouet, F. J. Lopez, and G. Sanz, Bernoulli **13**, 754 (2007).
[40] G. Wergen *et al.*, Phys. Rev. Lett. **109**, 164102 (2012).
[41] S. N. Majumdar and R. M. Ziff, Phys. Rev. Lett. **101**, 050601 (2008).
[42] G. Wergen, S. N. Majumdar, and G. Schehr, Phys. Rev. E **86**, 011119 (2012).
[43] S. Sabhapandit, EPL **94**, 20003 (2011).
[44] Y. Edery, A. Kostinski, and B. Borkowitz, Geophys. Res. Lett. **389**, L16403 (2011).
[45] M. Abramowitz and I. Stegun, *Handbook of mathematical functions* (Dover Publications Inc., New York, 1970).
[46] R. Shorrock, J. Appl. Probab. **9**, 219 (1972).
[47] R. Shorrock, J. Appl. Probab. **9**, 316 (1972).
[48] E. J. Gumbel, *Statistics of Extremes* (Dover, 1958).
[49] L. De Haan and A. Ferreira, *Extreme Value Theory* (New York: Springer, 2006).
[50] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools* (Berlin: Springer, 2000).
[51] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics* (Malabar: Krieger, 1987).
[52] S. Resnick, Stoch. Proc. Appl. **1**, 67 (1973).
[53] G. Wergen, "unpublished," (2011).
[54] G. M. Viswanathan *et al.*, Nature **381**, 413 (1996).
[55] A. M. Edwards *et al.*, Nature **449**, 1044 (2007).
[56] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[57] M. C. K. Yang, J. Appl. Probab. , 148 (1975).
[58] C. N. Williams Jr. *et al.*, *Historical Climatology Network Daily Temperature, Precipitation, and Snow Data*, Tech. Rep. (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, 2006).
[59] R. Gouet, F. J. Lopez, and G. Sanz, J. Stat. Mech.: Theor. Exp. **P01005** (2012).
[60] N. Balakrishnan, K. Balasubramanian, and S. Panchapakesan, J. Appl. Statist. Sci. **4**, 123 (1996).
[61] N. Balakrishnan, A. Pakes, and A. Stepanov, Adv. Appl. Probab. **37**, 765 (2005).
[62] I. Eliazar, Physica A **348**, 181 (2005).
[63] G. H. Weiss, *Aspects and applications of the random walk* (North-Holland, 1994).
[64] W. Feller, *An introduction to probability theory and its applications* (Wiley, New York, 1968).
[65] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, 2001).
[66] E. Sparre Andersen, Math. Scand. **1**, 263 (1953).
[67] E. Sparre Andersen, Math. Scand. **20**, 195 (1954).
[68] P. Le Doussal and K. J. Wiese, Phys. Rev. E **79**, 051105 (2009).
[69] V. V. Ivanov, Astron. and Astrophys. **286**, 328 (1996).
[70] Intergovernmental Panel on Climate Change, "Contribution of Working Group 1 to the

Forth Assessment Report of the Intergovernmental Panel on Climate Change," (2007).

[71] L. O. Mearns, R. W. Katz, and S. H. Schneider, J. Appl. Meteorol. **23**, 1601 (1984).

[72] R. A. Kerr, Science **251**, 274 (1991).

[73] D. R. Easterling *et al.*, Science **277**, 384 (1997).

[74] D. V. Hoyt, Climatic Change **3**, 243 (1981).

[75] E. Zorita, T. Stocker, and H. von Storch, Geophys. Res. Lett. **35**, L24706 (2008).

[76] S. Rahmstorf and D. Coumou, Nature Climate Change **20**, 491 (2012).

[77] G. W. Bassett Jr, Climatic Change **21**, 303 (1992).

[78] G. W. Bassett Jr and J. Zhiyong, Climatic Change **25**, 179 (1993).

[79] R. E. Benestad, J. Climate **19**, 630 (2006).

[80] A. Anderson and A. Kostinski, J. Appl. Meteo. and Climat. **50**, 1859 (2011).

[81] Project team ECAD, "European climate assessment and dataset (Tech. Rep., Royal Netherlands Meteorological Institute KNMI)," (2008).

[82] G. Wergen, A. Hense, and J. Krug, arXiv:1210.5416 (2012).

[83] L. Bachelier, Ann. Sci. Ecole Norm. S. **17**, 21 (1900).

[84] M. Bogner, "Rekordstatistik in Finanzdaten, unpublished thesis," (2009).

[85] Thomson Reuters, "Thomson Datastream Advance 4.0 SP4," (2003).

[86] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[87] P. Sibani and P. B. Littlewood, Phys. Rev. Lett. **71** (1993).

[88] P. Sibani, M. Brandt, and P. Alstrom, Int. J. Mod. Phys. B **12**, 361 (1998).

[89] J. Franke, "PhD Thesis: Statistical topography of fitness landscapes," (2012).

[90] M. Deakin, Math. Gaz. **51**, 100 (1967).

[91] R. L. Smith, J. Am. Stat. Assoc. **83**, 331 (1986).

[92] A. M. Nevill and G. Whyte, Med. Sci. Sport. Exer. **37**, 1785 (2005).

[93] G. Lippi *et al.*, Brit. Med. Bull. **87**, 7 (2008).

[94] J. H. J. Einmahl and J. R. Magnus, J. Am. Stat. Assoc. **104**, 1382 (2008).

[95] O. E. Barndorff-Nielsen and N. Shephard, J. Roy. Stat. Soc. B **63**, 167 (2001).

Chapter 14

Summary & Conclusions

> The most exciting phrase to hear in science,
> the one that heralds new discoveries,
> is not 'Eureka!' but 'That's funny...'
>
> *Isaac Asimov*

While the extensive review in chapter 13 was intended to give a general summary of the developments in the field of record statistics, we will now, in a few words, recapitulate the most important findings presented in this thesis.

14.1 Part I - Records in uncorrelated random variables

The main topic of most of the first part was the Linear Drift Model. Together with my collaborators, I studied the record statistics of time series of uncorrelated random variables (RV's) with a constant drift. The entries $X_1, X_2, ..., X_n$ of such a time series are of the form

$$X_k = Y_k + ck, \tag{14.1}$$

with independent and identically distributed (i.i.d.) RV's Y_k sampled from a continuous probability density $f(x)$.

In chapter 3, we discussed the record rate $P_n(c)$, the probability for a record in the nth entry of a time series with a drift c. In the stationary case, for $c = 0$, this record rate, $P_n(0) = 1/n$, is completely independent from the choice of the underlying probability density $f(x)$ that generates the RV's. In the presence of drift however, the situation is more complicated and, with one exception, it was not possible to compute $P_n(c)$ exactly. Instead, approximate results for the regime of a small drift $cn \ll \sigma$[1] and also the asymptotic regime with $c \gg \sigma$ and $n \to \infty$ were derived for various elementary distributions. The first case is important to model the effect of a slowly varying mean temperature on the occurrence of heat and cold records in the context of climatic change. Here, for $cn \ll \sigma$, one can expand $P_n(c)$ in powers of c:

$$P_n(c) \approx \frac{1}{n} + c\frac{n(n-1)}{2}\mathrm{I}_n, \tag{14.2}$$

with $\mathrm{I}_n := \int \mathrm{d}x \; f^2(c)\, F^{n-2}(x)$. It turns out that I_n and therefore the behavior of $P_n(c)$ depends crucially on the extreme value class of the underlying distribution. For RV's sampled from a distribution of the Fréchet class of distributions with power-law tails, the effect of the drift on the record rate is negligible for large n. On the other hand, the drift

---

[1] Where σ is the standard deviation or some other measure of the width of $f(x)$

in the Weibull class of distributions with a bounded support is very important and changes the behavior of $P_n(c)$ systematically. The most interesting case is the intermediate Gumbel class of distributions with an infinite support and tails that decay faster than a power-law. For the important exponential distribution and $cn \ll \sigma$, the drift has an n-independent effect $P_n(c) - P_n(0) \propto c$ on the record rate. For other representatives of the Gumbel class, such as the Gaussian distribution, we found logarithmic corrections to this result. In a similar manner, we also analyzed the ordering probability, that all RV's of a time series occur in ascending order.

The approximations for the record rate, derived in chapter 3, are instructive, but not entirely satisfying. In future research, it would be interesting to compute exact results for the record rate $P_n(c)$ that describe the behavior of the record rate for arbitrary n and c. Additionally, the effect of the linear drift on the distribution of record values has not been discussed in detail.

Probably the most important observation, which was made in our study of the LDM, was that, in the presence of a linear drift, the record events lose their stochastic independence. In particular, the joint probability $P_{n,m}(c)$ of records at times n and m can differ from the product of the individual record rates $P_n(c) \cdot P_m(c)$. In chapter 4, we defined the ratio

$$l_{n,m}(c) := \frac{P_{n,m}(c)}{P_n(c) \cdot P_m(c)}, \tag{14.3}$$

which is always 1 in the i.i.d. case. Surprisingly, for $c \neq 0$, $l_{n,n+1}(c)$ for records in neighboring entries n and $n+1$ can be both smaller and larger than one, i.e. record events can 'repel' and 'attract' each other. As it turns out, for $n \gg 1$, the correlations are negative for distributions that decay faster than exponential, e.g. a Gaussian or a uniform distribution. For distributions at least as broad as the exponential distribution, $l_{n,n+1}(c)$ is larger than one and therefore the probability for a record at time $n+1$ is increased if the nth entry was a record, too. These positive correlations for heavy-tailed RV's were surprising to us and, we do not know yet how to explain them intuitively.

Nevertheless, this effect can be used to design a test that detects heavy-tailed distributions in experimental data. As implied in chapter 5, such a record-based test might be more sensitive than the standard methods as, for instance, maximum-likelihood estimators. To determine whether or not a set of experimental measurements has an underlying heavy-tailed distribution, one has to add an artificial linear drift to a random permutation of the data. If the correlations in the resulting time series are positive ($l_{n,n+1}(c) > 1$), this is a good indicator for RV's from a distribution with a broader than exponential tail. The sensitivity of the test can be increased significantly by averaging over many different random permutations of the original data.

Motivated by observations made in our study of temperature measurements, we considered the effects of rounding on the record statistics of experimental data (chapter 6). When experimental measurements from an underlying continuous distribution are rounded in the measurement process, this can lead to ties. Together with my collaborators, I considered time series of rounded entries $\lfloor X_1 \rfloor_\Delta, \lfloor X_2 \rfloor_\Delta, ..., \lfloor X_n \rfloor_\Delta$, where the $X_i's$ are sampled from a continuous probability density and $\lfloor \cdot \rfloor_\Delta$ gives the result of rounding down to the next integer multiple of the discretization length Δ. In such a series, the nth entry is a new (strong) record if

$$\lfloor X_n \rfloor_\Delta > \max\{\lfloor X_1 \rfloor_\Delta, \lfloor X_2 \rfloor_\Delta, ..., \lfloor X_{n-1} \rfloor_\Delta\}. \tag{14.4}$$

We computed the effect of this rounding on the record rate P_n^Δ for various elementary distributions and discussed our findings in the context of the universality classes of extreme value statistics. Similar to the LDM, these discreteness effects are very strong for RV's from the Weibull class. In this case, the rounding leads to an exponential decay of the record rate P_n^Δ. On the other hand the asymptotic behavior of the record rate in the Fréchet class does not differ significantly from the continuous case and, for $n \to \infty$, one finds $P_n^{\Delta} \approx 1/n$. As before, the most interesting results were obtained in the Gumbel class. For the exponential distribution, the record rate is still proportional to $1/n$, but with a different prefactor, which depends on Δ. For a Gaussian probability density we could show that

$$P_n^{\Delta,(\mathrm{Gaussian})} \approx \frac{\sigma}{n\Delta}\left(\sqrt{\ln\left(n^2/2\pi\right)}\right)^{-1}. \tag{14.5}$$

In the regime of very strong discreteness with $\Delta \gg 1$ and for distributions of the Gumbel class that decay at least as fast as the exponential distribution, the asymptotic record statistics becomes highly regular and record events are almost predictable on a logarithmic time-scale.

14.2 Part II - Record-breaking temperatures

Some of the results derived in part I were applied in our study of record-breaking temperatures in the context of global warming. These studies were subject of part II of this thesis. In Chapter 7, we discussed the occurrence of records in time series of daily temperature measurements for individual calendar days. In these time series, the subsequent measurements are, to a good approximation, uncorrelated from each other. Therefore, we could consider their record statistics in comparison with uncorrelated RV's.

In fact, we showed that the simple model of independent RV's with a linear drift is useful to describe the effect of an increasing global mean temperature on the occurrence of temperature records. We analyzed historical temperature recordings from hundreds of European and U.S. weather stations and also gridded re-analysis data. Over the last decades, the mean temperature in these observational time series increased linearly. Additionally, we found that the distribution of daily temperatures around this moving average value is, to a good approximation, a Gaussian with a constant standard deviation. On this basis, one can predict the effect of a linear warming with speed c on the record rate from the LDM:

$$P_n\left(c\right) = \frac{1}{n} + \frac{c}{\sigma}\frac{2\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)}, \tag{14.6}$$

where σ is the standard deviation of daily temperatures. Apparently, the relevant parameter that determines the increase of the record rate is the normalized drift c/σ. Therefore, a large effect $P_n\left(c\right) - 1/n$ of the warming on the record rate can be caused either by a large drift c or a small standard deviation σ.

We could demonstrate that this simple formula accurately predicts the occurrence of temperature records. The normalized drift in the observational time series was of the order of $0.01\mathrm{y}^{-1}$ leading to a significant increase in the number of daily heat records in the last decades, depending on the choice of the observation period. In Europe, for instance, global warming resulted in a 40% increase in the upper record rate at the end of the time period from 1976 to 2005. In consistence with the model predictions, the number of cold records in the data was decreased in the same extent. In chapter 7, we also showed that the normalized drift and with it the effect of the warming on the record rate, has strong seasonal and spatial variations. In the latter case, it was found that, close to the oceans or on islands, the large heat capacity of water resulted in a very small standard deviation σ. In turn, the small σ leads to a very strong increase of the record rate due to a large normalized drift c/σ. On the other hand, σ is very large for stations far away from the coast, where only a small increase of the record rate was measured.

In a more extensive study, in chapter 8, we analyzed also monthly mean temperatures. These averaged measurements have a smaller standard deviation and, because of that, their record statistics is much more affected by global warming. The record rate of monthly mean temperatures can be many times as high as expected on the basis of a stationary climate.

Furthermore, we considered the values of record-breaking temperatures in European reanalysis data and compared them with analytical predictions from the LDM. We computed the effect of the drift on the mean $\mu_n(c)$ of a record in the nth entry of a Gaussian LDM with a drift c:

$$\mu_n(c) \approx \mu_n(0) + n\frac{c}{\sigma}\frac{2\sqrt{\pi}}{e^2}\ln(4), \qquad (14.7)$$

where $\mu_n(0)$ is the mean value in the i.i.d. case.

In the summer months, the temperature record values are in good agreement with this Gaussian model and the mean values μ_n of the upper records that occur in a given year are significantly increased by global warming. In winter, on the other hand, the lower records are more extreme and the behavior can not be explained with the Gaussian LDM. This effect is caused by a distinct asymmetry of the distribution of daily temperatures in winter. Here, this distribution has a broader lower tail and therefore favors extremely cold temperatures. Because of that, extreme cold records in winter, especially in northern Europe, are not in contradiction with global warming.

In the work described in chapter 8 it became especially clear that the occurrence of record-breaking temperatures might be linked to specific weather patterns as well as parameters like soil moisture or snow coverage. The connection between these factors and record events is not sufficiently understood yet and this would certainly pose a promising objective for future research.

14.3 Part III - Record statistics of random walks

In this part, we discussed two possible generalizations of the symmetric discrete-time random walk considered by Majumdar and Ziff [15]. Chapters 9 and 10 were focused on record-breaking entries in biased discrete-time random walks, chapter 11 on records of the maximum of ensembles of multiple, symmetric random walks.

A discrete-time biased random walk is a process with entries $X_0, X_1, ..., X_n$ of the form

$$X_n = X_{n-1} + \xi_n + c \qquad (14.8)$$

with symmetric i.i.d. RV's ξ_i sampled from a continuous jump distribution $f(\xi)$. In the asymptotic limit ($n \to \infty$), we found five different universal regimes depending on the Lévy index μ of the jump distribution and the sign of the bias c. Subject of chapter 10 was the computation of the asymptotic record statistics of biased random walks in these regimes. The effect of the bias on the record statistics increases with an increasing Lévy index μ. In the subcritical case with $\mu < 1$, the records are not systematically affected by the drift and the mean record number grows proportional to $\sqrt{n}$ as in the unbiased case. In the marginal case of $\mu = 1$, this mean record number depends non-trivially on n with $\langle R_n \rangle \propto n^{\Theta(c)}$ and a function $\Theta(c) = \frac{1}{2} + \frac{1}{\pi}\arctan(c)$, which depends continuously on c. For $\mu > 1$, the sign of the bias c becomes important. While the mean record number grows linearly with n when c is positive, it approaches a constant value for $c < 0$. In the important case of a Gaussian random walk ($\mu = 2$) with a positive bias, the distribution of the record number approaches a Normal distribution. The methods introduced in chapter 10 also allowed us to compute the extremal statistics of the ages $\langle l_{\max,n} \rangle$ and $\langle l_{\min,n} \rangle$ of the longest and shortest lasting records in a biased random walk. The main results of chapter 10 are summarized in the following table:

	$\langle R_n(c) \rangle$	Distr. of R_n	$\langle l_{\max,n} \rangle$	$\langle l_{\min,n} \rangle$
I ($\mu < 1$)	$\propto \sqrt{n}$	half-Gaussian	$\propto n$	$\propto \sqrt{n}$
II ($\mu = 1$)	$\propto n^{\Theta(c)}$	asymm. & non-Gaussian	$\propto n$	$\propto n^{1-\Theta(c)}$
III ($1 < \mu < 2, c > 0$)	$\propto n$	asymm. & non-Gaussian	$\propto n^{1/\mu}$	const.
IV ($\mu = 2, c > 0$)	$\propto n$	Gaussian	$\propto \ln n$	const.
V ($\mu > 1, c < 0$)	const.	geometric	$\propto n$	$\propto n$

Chapter 9, which was published prior to chapter 10, discusses the Brownian case (IV) of a biased discrete-time random walk with a Gaussian jump distribution and a small bias $c\sqrt{n} \ll \sigma$. Here, the relevant quantities, like the survival probability and the mean record number, can be expanded in powers of the bias c. For $c\sqrt{n} \ll \sigma$, the effect of the drift on the mean record number grows linearly in n and one finds that

$$\langle R_n(c) \rangle \approx \sqrt{\frac{4n}{\pi}} + \frac{cn}{\sqrt{2}\sigma}. \tag{14.9}$$

At the end of chapter 9, we demonstrated that this result can be used to describe record-breaking events in time series of daily stock prizes from the Standard and Poors 500 index. For long time series reaching over many years, the records in daily stock prizes are accurately modeled by a biased random walk. On intervals of 100 trading days length, the number of lower records is systematically reduced and can not be explained with the random walk model.

In chapter 11, we studied the record statistics of the maximum of ensembles of multiple random walks with a common jump distribution. For ensembles of $N \gg 1$ random walks with a jump distribution that has a finite variance ($\mu = 2$), the asymptotic distribution of the record number $R_{n,N}$ approaches a Gumbel form. The first moment of this distribution, the mean record number, is given by

$$\langle R_{n,N} \rangle^{(\mu=2)} \approx \sqrt{4n \ln N}. \tag{14.10}$$

For Lévy flights with a heavy-tailed jump distribution ($\mu < 2$), the mean record number gets entirely independent from N and the Lévy index μ and we found that

$$\langle R_{n,N} \rangle^{(\mu<2)} \approx 4\sqrt{\frac{n}{\pi}} = 2\langle R_{n,1} \rangle. \tag{14.11}$$

Numerical simulations showed that the asymptotic distribution of the record number $R_{n,N}$ in this case is also completely independent from N and μ. To obtain an analytical representation of this hitherto unknown universal distribution is definitely an interesting subject for future research. Also, the mean record number of N detrended and normalized stocks from the Standard and Poors 500 index is proportional to $\sqrt{n \ln N}$ but has a different prefactor, which is smaller than in the case of the independent random walks. The emergence of this prefactor can be explained with an effective number of independent stocks, which, due to strong correlations in the market, is much smaller than N.

14.4 Part IV - Records in finance

A more systematic analysis of record-breaking stock prizes was presented in the fourth part. In chapter 12, we compared the occurrence of records in the daily returns to the record statistics of i.i.d. RV's. It was shown that, on long time-scales, record-breaking daily returns are highly correlated and most of the records occur during short periods of high market activity, as for instance the recent financial crisis, which began in 2008. In some sense, upper and lower return records attract each other, trading days with many upper records are often followed by trading days with a large number of lower records and vice versa. On shorter time-scales, these correlations are less pronounced and for observation periods of less than a few hundred trading days length the record statistics of the daily returns approaches the i.i.d. behavior. Interestingly, for some time series, one can identify a significant increase in the number of lower return records that can not be explained within the Geometric Random Walk Model.

As in chapter 9, the statistics of record-breaking stock prizes was compared to the record statistics of biased random walks. We found that, up to a certain degree, this model is useful to describe the occurrence of records in the stock data. However, for interval lengths of a few hundred trading days, both the number of upper and the number of lower records are overestimated by the biased random walk. In this context, we demonstrated that the autoregressive AR(1) process with a fitted regression parameter models the stock data more accurately. The AR(1) process is certainly a nice subject for future research and it would be interesting to compute its record statistics analytically.

The findings for the record statistics are generally in agreement with the survival probabilities for the time series of the daily stock prizes, which were discussed as well. Towards the end of chapter 12, we illustrated that the occurrence of records both in the stocks and in the daily returns decreases over the week. The probability for a record event is highest on Mondays and lowest on Fridays.

14.5 Closing remarks & bigger picture

In retrospect, we - my coworkers and I - considered record processes in several different models of both time-dependent and correlated random variables. We could improve our understanding of the record statistics of random variables with a linear drift and use these findings to describe the occurrence of record-breaking temperatures in a warming climate. Motivated by this application, we also considered the important effects of rounding on records in experimental measurements. In a similar manner, our attempts to model records in financial data encouraged us to consider biased random walks as well as ensembles of uncorrelated random walks. We learned a lot about record-breaking in stock prizes, but our findings are also interesting from a purely theoretical point of view.

In this light, the contributions in this thesis should be considered as progress made towards a better understanding of the statistics of records in time series of time-dependent and correlated random variables as well as their role in climatology and finance. Both from the theoretical and from the applied point of view, many open questions remain and several new problems were posed. Thus it is certain that the study of record-breaking events will continue to be interesting in the future.

The work done in this thesis can also be seen as a contribution to the broader field of extreme value statistics in statistical physics. By now, the study of extremes has become an important branch of theoretical physics and mathematics with many applications in other areas of science. Considerable effort was made to draw a bigger picture of the statistics of extremes in complex systems and stochastic processes, such as diffusion processes [70, 96, 97], stochastic growth [98, 99] or, in general, systems of interacting particles [100–103]. Many of these advances are connected to progress made in the theory of random matrices and, in particular, the extreme value statistics of their eigenvalues [101, 104–106]. In recent years it was discovered that the distribution of extreme values of complex systems, such as the largest eigenvalue of a random matrix, height fluctuations of surface growth models or flux fluctuations of simple exclusion processes exhibit common universal properties related to the so-called Tracy-Widom distribution [105–107].

In the context of these remarkable developments it is important to find and explain universal characteristics in the extreme value behavior of complex and correlated systems. The study of records, as extreme values with respect to the history of a stochastic process, can be a piece in this puzzle. In the future, it would certainly be a major step forward if one could establish a more profound connection between record statistics and the recent progress made in the theory of extremes. A better understanding of the occurrence of record-breaking events in strongly correlated systems, such as, for instance, fractional brownian motion with correlated increments or power-law correlated random variables, might be the next step towards this goal.

Bibliography for chapters 1, 2 & 14

[1] C. Glenday, editor. *Guinness World Records.* Jim Pattison Group, 2012.

[2] N. Glick. Breaking Records and Breaking Boards. *Am. Math. Mon.*, 85(1):2–26, 1978.

[3] V. B. Nevzorov. *Records: Mathematical Theory.* American Mathematical Society, 2004.

[4] B. C. Arnold, N. Balakrishnan, and H. N. Nagraja. *Records.* Wiley-Interscience, 1 edition, 1998.

[5] M. Abramowitz and I. Stegun. *Handbook of mathematical functions.* Dover Publications Inc., New York, 1970.

[6] W. Feller. *An introduction to probability theory and its applications.* Wiley, New York, 1968.

[7] R.W. Shorrock. A limit theorem for inter-record times. *J. Appl. Probab.*, 9(1):219–223, 1972.

[8] R.W. Shorrock. On record values and record times. *J. Appl. Probab.*, 9(2):316–326, 1972.

[9] A. Stepanov. Random intervals based on records. *J. Stat. Plan. Infer.*, 118(1-2):103–113, 2004.

[10] E. J. Gumbel. *Statistics of Extremes.* Dover, 1958.

[11] J. Galambos. *The Asymptotic Theory of Extreme Order Statistics.* Malabar: Krieger, 1987.

[12] L. De Haan and A. Ferreira. *Extreme Value Theory.* New York: Springer, 2006.

[13] S. Resnick. Limit laws for record values. *Stoch. Proc. Appl.*, 1(1):67–82, 1973.

[14] G. H. Weiss. *Aspects and applications of the random walk.* North-Holland, 1994.

[15] S. N. Majumdar and R. M. Ziff. Universal Record Statistics of Random Walks and Levy Flights. *Phys. Rev. Lett.*, 101(5):050601, 2008.

[16] S. Redner. *A Guide to First-Passage Processes.* Cambridge University Press, 2001.

[17] E. Sparre Andersen. On the fluctuations of sums of random variables. *Math. Scand.*, 1:263–285, 1953.

[18] E. Sparre Andersen. On the fluctuations of sums of random variables. II. *Math. Scand.*, 20:195–223, 1954.

[19] IPCC - Intergovernmental Panel on Climate Change, 2007.

[20] Contribution of Working Group 1 to the Forth Assessment Report of the Intergovernmental Panel on Climate Change, 2007.

[21] IPCC Special Report - Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, 2012.

[22] P. A. Stott, D. A. Stone, and M. R. Allen. Human contribution to the European heatwave of 2003. *Nature*, 432:610–614, 2004.

[23] E. Zorita, T.F. Stocker, and H. von Storch. How unusual is the recent series of warm years? *Geophys. Res. Lett.*, 35:L24706, 2008.

[24] S. Rahmstorf and D. Coumou. A decade of weather extremes. *Nature Climate Change*, 2:491–496, 2012.

[25] R. C. Balling Jr., J. A. Skindlov, and D. H. Phillips. The Impact of Increasing Summer Mean Temperatures on Extreme Maximum and Minimum Temperatures in Pheonix, Arizona. *J. Climate*, 3(12):1491–1494, 1990.

[26] L. O. Mearns, R. W. Katz, and S. H. Schneider. Extreme High-Temperature Events: Changes in their Probabilites with Changes in Mean Temperature. *J. Appl. Meteorol.*, 23(12):1601–1613, 1984.

[27] C. H. Reitan and J. M. Moran. The Probability of Record Daily Low Temperatures in Winter. *Mon. Weather Rev.*, 105(11):1442–1446, 1977.

[28] D. V. Hoyt. Weather 'Records' and Climatic Change. *Climatic Change*, 3(3):243–249, 1981.

[29] R. A. Kerr. Global Temperature Hits Record Again. *Science*, 251:274, 1991.

[30] G. W. Bassett Jr. Breaking Recent Global Temperature Records. *Climatic Change*, 21(3):303–315, 1992.

[31] G. W. Bassett Jr. and Zhiyong Lin. Breaking Global Temperature Records After Mt. Pinatubo. *Climatic Change*, 25(2):179–184, 1993.

[32] D. R. Easterling et al. Maximum and Minimum Temperature Trends for the Globe. *Science*, 277:384–387, 1997.

[33] R. E. Benestad. How often can we expect a record event? *Climate Res.*, 25:3–13, 2003.

[34] R. E. Benestad. Record-values, nonstationarity tests and extreme value distributions. *Global Planet. Change*, 44(1-4):11–26, 2004.

[35] R. E. Benestad. Can We Expect More Extreme Precipitation on the Monthly Time Scale? *J. Climate*, 19(4):630–637, 2006.

[36] S. Redner and M. R. Peterson. On the Role of Global Warming on the Statistics of Record-Breaking Temperatures. *Phys. Rev. E*, 74(6):061114, 2006.

[37] G A Meehl et al. Relative increase of record high maximum temperatures compared to record low minimum temperatures in the U.S. *Geophys. Res. Lett.*, 36:L23701, 2009.

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

[39] G. Wergen. Diploma Thesis: More and more temperature records - Is global warming to blame?, 2009.

[40] W. I. Newman, B. D. Malamud, and D. L. Turcotte. Statistical properties of record-breaking temperatures. *Phys. Rev. E*, 82(6):066111, 2010.

[41] N. Elguindi, S. A. Rauscher, and F. Giorgi. Historical and future changes in maximum and minimum temperature records over Europe. *Climatic Change*, 114:1–17, 2012.

[42] S. Rahmstorf and D. Coumou. Increase of extreme events in a warming world. *P. Natl. Acad. Sci. USA*, 108(44):17905–17909, 2011.

[43] G. Wergen, A. Hense, and J. Krug. Record occurrence and record values in daily and monthly temperatures. *(submitted to Climate Dynamics, arXiv:1210.5416)*, 2012.

[44] R. Ballerini and S. Resnick. Records from improving populations. *J. Appl. Probab.*, 22(3):487–502, 1985.

[45] R. Ballerini and S. Resnick. Records in the Presence of a Linear Trend. *Adv. Appl. Probab.*, 19(4):801–828, 1987.

[46] K. Borovkov. On Records and Related Processes for Sequences with Trends. *J. Appl. Probab.*, 36(3):668–681, 1999.

[47] J. Franke, G. Wergen, and J. Krug. Records and sequences of records from random variables with a linear trend. *J. Stat. Mech.: Theor. Exp.*, P10013, 2010.

[48] G. Wergen, J. Franke, and J. Krug. Correlations between record events in sequences of random variables with a linear trend. *J. Stat. Phys.*, 144:1206, 2011.

[49] J. Franke, G. Wergen, and J. Krug. Correlations of record events as a test for heavy-tailed distributions. *Phys. Rev. Lett.*, 108(6):064101, 2012.

[50] J. Krug. Records in a changing world. *J. Stat. Mech.: Theor. Exp.*, P07001, 2007.

[51] I. Eilazar and J. Klafter. Record events in growing populations: Universality, correlation, and aging. *Phys. Rev. E*, 80(6):061117, 2009.

[52] J. Krug and K. Jain. Breaking records in the evolutionary race. *Physica A*, 358(1):1–9, 2005.

[53] S. A. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, 128:11–45, 1987.

[54] P. Anderson, P. Sibani, L. P. Oliveira, and H. J. Jensen. Evolution in complex systems. *Complexity*, 10(1):49–56, 2004.

[55] H. A. Orr. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.*, 6(2):119–127, 2005.

[56] J. Franke, A. Klözer, J. A. G. M. de Visser, and J. Krug. Evolutionary Accessibility of Mutational Pathways. *PLoS Comp. Biol.*, 7(8):e1002134, 2011.

[57] J. Franke. PhD Thesis: Statistical topography of fitness landscapes, 2012.

[58] W. Vervaat. Limit theorems for records from discrete distributions. *Stoch. Proc. Appl.*, 1(4):317–334, 1973.

[59] H. Prodinger. Combinatorics of geometrically distributed random variables: Left-to-right maxima. *Discrete Math.*, 153(1-3):253–270, 1996.

[60] R. Gouet, F. J. Lopez, and G. Sanz. Central limit theorems for the number of records in discrete models. *Adv. Appl. Probab.*, 37(3):331–336, 2005.

[61] E. S. Key. On the Number of Records in an IID Discrete Sequence. *J. Theor. Probab.*, 18:99, 2005.

[62] R. Gouet, F. J. Lopez, and G. Sanz. Asymptotic normality for the counting process of weak records and δ-records in discrete models. *Bernoulli*, 13(3):754–781, 2007.

[63] I. Eliazar. On geometric record times. *Physica A*, 348:181–198, 2005.

[64] R. Gouet, F. J. Lopez, and G. Sanz. On geometric records: rate of appearance and magnitude. *J. Stat. Mech.: Theor. Exp.*, P01005, 2012.

[65] G. Wergen, D. Volovik, S. Redner, and J. Krug. Rounding effects in record statistics. *Phys. Rev. Lett.*, 109(16):164102, 2012.

[66] S. Sabhapandit. Record Statistics of Continuous Time Random Walk. *EPL*, 94:20003, 2011.

[67] G. Wergen, M. Bogner, and J. Krug. Record statistics for biased random walks, with an application to financial data. *Phys. Rev. B*, 83(5):051109, 2011.

[68] Thomson Datastream Advance 4.0 SP4, 2003.

[69] M. Bogner. Unpublished thesis: Rekordstatistik in Finanzdaten, 2009.

[70] P. Le Doussal and K. J. Wiese. Driven particle in a random landscape: disorder correlator, avalanche distribution and extreme value statistics of records. *Phys. Rev. E*, 79(5):051105, 2009.

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

[72] G. Wergen, S. N. Majumdar, and G. Schehr. Record statistics for multiple random walks. *Phys. Rev. E*, 86(1):011119, 2012.

[73] Y. Edery, A. Kostinski, and B. Borkowitz. Record setting during dispersive transport in porous media. *Geophys. Res. Lett.*, 389:L16403, 2011.

[74] L. P. Oliveira et al. Record dynamics and the observed temperature plateau in the magnetic creep-rate of type-II superconductors. *Phys. Rev. B*, 71(10):104526, 2005.

[75] P. Sibani and P. B. Littlewood. Slow Dynamics from Noise Adaptation. *Phys. Rev. Lett.*, 71(10):1482–1485, 1993.

[76] P. Sibani, G. F. Rodriguez, and G. G. Kenning. Intermittent quakes and record dynamics in the thermoremanent magnetization of a spin-glass. *Phys. Rev. B*, 74(22):224407, 2006.

[77] P. Sibani. Linear response in aging glassy systems, intermittency and the Poisson statistics of record fluctuations. *Eur. Phys. J. B*, 58(4):483–491, 2007.

[78] R. M. Vogel, A. Zafirakou-Koulouris, and N. C. Matalas. Frequency of record-breaking floods in the United States . *Water Resour. Res.*, 37(6):1723–1731, 2001.

[79] T. O. Richardson et al. Record Dynamics in Ants. *PLoS One*, 58(3):e9621, 2010.

[80] D. Gembris, J. G. Taylor, and D. Suter. Sports statistics: Trends and random fluctuations in athletics. *Nature*, 417:506, 2002.

[81] D. Gembris, J. G. Taylor, and D. Suter. Evolution of Athletic Records: Statistical Effects versus Real Improvements. *J. Appl. Stat.*, 34(5):529–545, 2007.

[82] J. Pickands. The two dimensional poisson process and extremal processes. *J. Appl. Probab.*, 8(4):745–756, 1971.

[83] J. Bunge and H. N. Nagraja. The distribution of certain record statistics from a random number of observations. *Stoch. Proc. Appl.*, 38(1):167–183, 1991.

[84] H. N. Nagraja. Record statistics from point process models. In J. Galambos, J. Lechner, and E. Simiu, editors, *Extreme Value Theory and Applications*, pages 355–370. Kluwer, 1994.

[85] D. Pfeifer. Characterizations of exponential distributions by independent nonstationary record increments. *J. Appl. Prob.*, 19:127–135, 1982.

[86] C. M. Goldie and S. I. Resnick. Records in a partially ordered set. *Ann. Probab.*, 17:678–689, 1989.

[87] C. M. Goldie and S. I. Resnick. Many multivariate records. *Stoch. Proc. Appl.*, 59(2):185–216, 1995.

[88] D. Sornette. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools.* Berlin: Springer, 2000.

[89] G. M. Viswanathan et al. Lévy flight search patterns of wandering albatrosses. *Nature*, 381:413–415, 1996.

[90] A. M. Edwards et al. Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449:1044–1048, 2007.

[91] S. Redner. How Popular is Your Paper? An Empirical Study of the Citation Distribution. *Eur. Phys. J. B*, 4(2):131–134, 1998.

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

[93] C. N. Williams Jr. et al. Historical Climatology Network Daily Temperature, Precipitation, and Snow Data. Technical report, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, 2006. ORNL/CDIAC-118, NDP-070.

[94] M. R. Haylock et al. A European daily high-resolution gridded dataset of surface temperature and precipitation . *J. Geophys. Res.-Atmos.*, 113:D11110, 2008.

[95] L. Bachelier. Théorie de la Spéculation. *Ann. Sci. Ecole Norm. S.*, 17:21–88, 1900.

[96] S. N. Majumdar and A. Comtet. Exact Maximal Height Distribution of Fluctuating Interfaces. *Phys. Rev. Lett.*, 92(22):225501, 2004.

[97] G. Schehr and P. Le Doussal. Extreme value statistics from the real space renormalization group: Brownian motion, Bessel processes and continuous time random walks. *J. Stat. Mech.: Theor. Exp.*, P01009, 2010.

[98] T. Sasamoto and H. Spohn. One-Dimensional Kardar-Parisi-Zhang Equation: An Exact Solution and its Universality. *Phys. Rev. Lett.*, 104(23):230602, 2010.

[99] T. Kriecherbauer and J. Krug. A pedestrian's view on interacting particle systems, KPZ universality, and random matrices. *J. Stat. Mech.: Theor. Exp.*, 43:403001, 2010.

[100] E. Bertin. Global Fluctuations and Gumbel Statistics. *Phys. Rev. Lett.*, 95(17):170601, 2005.

[101] G. Biroli, J.-P. Bouchaud, and M. Potters. Extreme value problems in random matrix theory and other disordered systems . *J. Stat. Mech.: Theor. Exp.*, P07019, 2007.

[102] Y. V. Fyodorov and J.-P. Bouchaud. Freezing and extreme-value statistics in a random energy model with logarithmically correlated potential . *J. Phys. A-Math. Theor.*, 41(37):372001, 2008.

[103] M. Evans and S. N. Majumdar. Condensation and extreme value statistics . *J. Stat. Mech.*, P05004, 2008.

[104] D. S. Dean and S. N. Majumdar. Extreme value statistics of eigenvalues of Gaussian random matrices. *Phys. Rev. E*, 77(4):041108, 2008.

[105] C. Nadal and S. N. Majumdar. A simple derivation of the Tracy-Widom distribution of the maximal eigenvalue of a Gaussian unitary random matrix . *J. Stat. Mech.: Theor. Exp.*, P04001, 2011.

[106] G. Borot et al. Large deviations of the maximal eigenvalue of random matrices . *J. Stat. Mech.: Theor. Exp.*, P11024, 2011.

[107] C. A. Tracy and H. Widom. Level-spacing distributions and the Airy kernel. *Phys. Lett. B*, 305(1-2):115–118, 1993.

[108] J.-P. Bouchaud and M. Mezard. Universality classes for extreme-value statistics. *J. Phys. A-Math. Theor.*, 30(23):7997–8015.

[109] T. Antal et al. 1/f Noise and Extreme Value Statistics. *Phys. Rev. Lett.*, 87(22):240601, 2001.

[110] J. F. Eichner et al. Extreme value statistics in records with long-term persistence. *Phys. Rev. E*, 73(1):016130, 2006.

Acknowledgements

It was never about the outcome
it was about the process of getting there!

First and foremost I take immense pleasure in thanking my supervisor Joachim Krug. Without Joachim I would never have started to study records. I learned a lot from him during the last five years. He has a unique way of pointing me in the right direction, if directions are in need.

I would like to thank my brilliant coworkers who contributed to this work, most importantly Jasper Franke with whom I collaborated on many different questions. I am grateful to Sidney Redner and Daniel Volovik from the Institute of Polymer Studies in Boston for their hospitality and their input on our joint projects. Similarly, I am also very grateful to Satya N. Majumdar and Grégory Schehr from the Université Paris Sud for welcoming me in Orsay and for the opportunity to work with them. I learned a lot during my visits in Paris. The assistance in climatological questions provided by Andreas Hense from the University of Bonn was very helpful.

The financial support of the Friedrich-Ebert-Stiftung and the Bonn Cologne Graduate School of Astronomy and Physics helped me to study and work independently. Especially the Friedrich-Ebert-Stiftung also enriched my life and my studies in several non-material ways.

Thanks to the members of my group, in particular, Ingo, Su-Chan, Marian, Johannes, Andrea, Ivan, Stefan and Steffen. Without their company my work would have been only half as interesting and pleasant. I am very grateful to all my friends who accompanied my in the last 9 years in Cologne, it was a memorable time! For their assistance in proofreading this thesis at short notice, I am very grateful to Julie, Anna, Rabia and Ivan.

I am deeply grateful to my parents Heinz and Regine Wergen for their tireless support. Thanks also to my sister Johanna and my brothers Clemens and Niklas for supporting me in many ways. Without my family none of this would have been possible. Towards the completion of this work, I had the great fortune to meet Julie Weis, she supported me a lot in the last months. She helped me find the peace and quietness to finish this long-term endeavor.

Anhang gemäß der Prüfungsordnung (deutsch)

Zusammenfassung meiner Beteiligungen zu den Publikationen **281**

Erklärung zur Selbstständigkeit **283**

Lebenslauf **285**

Zusammenfassung meiner Beteiligungen zu den eingebundenen Publikationen

- **Kapitel 3**: Jasper Franke, Gregor Wergen and Joachim Krug,
  Records and sequences of records from random variables with a linear trend,
  J. Stat. Mech.: Theor. Exp. **P10013** (13 October 2010)

  Die analytischen Rechnungen in diesem Artikel habe ich zusammen mit Jasper Franke durchgeführt. Insbesondere sind die Resultate im Regime kleinen Drifts $cn \ll \sigma$ von mir. Weiterhin habe ich zahlreiche numerische Simulationen durchgeführt. Der Artikeltext wurde zusammen mit Jasper Franke und Joachim Krug geschrieben.

- **Kapitel 4**: Gregor Wergen, Jasper Franke and Joachim Krug,
  Correlations Between Record Events in Sequences of Random Variables with a Linear Trend,
  J. Stat. Phys. **144**, 6 (5 August 2011)

  Die analytischen Rechnungen in diesem Artikel wurden von Jasper Franke und mir durchgeführt. Die numerischen Simulationen stammen ebenso von uns beiden. Der Artikeltext wurde zusammen mit Jasper Franke und Joachim Krug geschrieben.

- **Kapitel 5**: Jasper Franke, Gregor Wergen and Joachim Krug,
  Correlations of Record Events as a Test for Heavy-Tailed Distributions,
  Phys. Rev. Lett. **108**, 064101 (7 February 2012)

  Die Idee zu dieser Publikation stammt von Jasper Franke. Ich habe mit ihm zusammen an dem Artikel gearbeitet und insbesondere auch numerische Ergebnisse beigetragen. Der Artikeltext wurde von Jasper Franke mit Unterstützung von Joachim Krug und mir verfasst.

- **Kapitel 6**: Gregor Wergen, Daniel Volovik, Sidney Redner and Joachim Krug,
  Rounding Effects in Record Statistics,
  Phys. Rev. Lett. **109**, 164102 (19. October 2012)

  Die Idee zu diesem Artikel ist bei einem Besuch bei Sid Redner und Daniel Volovik am Boston Center of Polymer Studies im Jahre 2009 entstanden. Ich habe die analytischen Rechnungen sowie die numerischen Simulationen durchgeführt und den Artikeltext mit Unterstützung von Joachim Krug und Sid Redner verfasst.

- **Kapitel 7**: Gregor Wergen and Joachim Krug,
  Record-breaking temperatures reveal a warming climate,
  EPL **92**, 30008 (29 November 2010)

  Die Idee zu dieser Studie stammt von Joachim Krug. Ich habe die analytischen Rechnungen für die Rekordraten im Linearen Drift Modell durchgeführt und die Beobachtungsdaten der Europäischen und Amerikanischen Wetterstationen ausgewertet. Den Artikeltext habe ich mit Unterstützung von Joachim Krug geschrieben.

- **Kapitel 8**: Gregor Wergen, Andreas Hense and Joachim Krug,
  Record occurrence and record values in daily and monthly temperatures,
  submitted to Climate Dynamics (arXiv:1210.5416)

Für diesen, bisher unpublizierten, Artikel habe ich die analytischen Rechnungen am Linearen Drift Modell sowie die numerischen Simulationen durchgeführt. Weiterhin habe ich die historischen Wetterdaten ausgewertet. Bei Fragen aus dem Bereich der Klimatologie bin ich von Andreas Hense unterstützt worden. Andreas Hense und Joachim Krug haben mir auch beim Verfassen des Artikeltextes geholfen.

- **Kapitel 9**: Gregor Wergen, Miro Bogner and Joachim Krug,
  Record statistics for biased random walks, with an application to financial data,
  Phys. Rev. E **83**, 051109 (9 May 2011)

  Die analytischen Berechnungen in diesem Artikel stammen von mir. Ich habe auch die numerischen Simulationen der Random Walks sowie die Auswertung der Aktiendaten durchgeführt. Miro Bogner hat, im Rahmen seiner Staatsexamensarbeit, grundlegende Beobachtungen zur Statistik der Rekorde in den Börsendaten beigetragen. Den Artikeltext habe ich mit Unterstützung von Joachim Krug geschrieben.

- **Kapitel 11**: Gregor Wergen, Satya N. Majumdar and Grégory Schehr,
  Record statistics for multiple random walks,
  Phys. Rev. E **86**, 011119 (18 July 2012)

  Die Publikation ist bei einem mehrwöchigen Besuch bei Satya N. Majumdar und Grégory Schehr an der Université Paris Sud im Jahr 2012 entstanden. Die analytischen Rechnungen haben wir gemeinsam erarbeitet, die überwiegende Zahl der numerischen Simulationen sind von mir durchgeführt worden. Die Untersuchung der Aktiendaten stammt ebenso von mir. Den Artikeltext habe ich zusammen mit Satya N. Majumdar und Grégory Schehr verfasst.

- **Kapitel 10**: Satya N. Majumdar, Grégory Schehr and Gregor Wergen,
  Record statistics and persistence for a random walk with a drift,
  J. Phys. A: Math. Theor. **45**, 355002 (15 August 2012)

  Dieser Artikel ist ebenfalls bei einem Besuch an der Université Paris Sud im Jahr 2012 geschrieben worden. Die analytischen Rechnungen wurden überwiegend von Satya N. Majumdar und Grégory Schehr erarbeitet. Ich habe numerische Simulationen durchgeführt und beim Schreiben der Publikation mitgewirkt.

- **Kapitel 12**: Gregor Wergen,
  Modeling record-breaking stock prices,
  in preparation

  Dieser, bisher unveröffentlichte, Artikel wurde von mir verfasst. Meine Arbeit zu diesem Thema wurde maßgeblich von zahlreichen anregenden Diskussion mit Joachim Krug geprägt.

- **Kapitel 13**: Gregor Wergen,
  Record statistics beyond the standard model - Theory and applications,
  submitted to J. Phys. A: Math. Theor. (arxiv:1211.6005)

  Dieser Review wurde von mir im Rahmen meiner Dissertation verfasst.

Gregor Wergen

Köln, 22. Januar 2013

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit — einschließlich Tabellen, Karten und Abbildungen —, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie — abgesehen von oben angegebenen Teilpublikationen — noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Joachim Krug betreut worden.

Gregor Wergen

Köln, 22. Januar 2013

Lebenslauf - Dipl.-Phys. Gregor Wergen

Persönliche Daten

Name	Gregor Wergen
Geburtdatum	7. Januar 1984
Geburtsort	Köln
Staatsangehörigkeit	deutsch
Adresse	Königswinterstr. 9, 50939 Köln

Schulische Ausbildung

1990 - 1994	GGS Reinshagen in Remscheid
1994 - 2003	Ernst-Moritz-Arndt Gymnasium in Remscheid, Allgemeinene Hochschulreife mit Abschlussnote 1,5

Studium

2003 - 2009	Diplomstudium der Physik an der Universität zu Köln, Titel der Diplomarbeit: 'More and more temperature records - is global warming to blame?', Abschlussnote 'sehr gut'
2009 - 2013	Promotionsstudium im Fachbereich Theoretische Physik an der Universität zu Köln, Promotionsthema: Rekordstatistik mit Anwendungen in Klimatologie und Wirtschaftswissenschaften

Stipendien

2004 - 2009	Grundförderung der Friedrich-Ebert-Stiftung
2008 - 2012	Bonn Cologne Graduate School of Astronomy and Physics
2010 - 2012	Promotionsförderung der Friedrich-Ebert-Stiftung

Qualifikationen und Engagement

2003 - 2006	Erfahrung als Trainer im Leistungssport
seit 2006	Leitung div. Seminare der Friedrich-Ebert-Stiftung
2007 - 2008	Bundesvertreter der Stipendiatenschaft der Friedrich-Ebert-Stiftung
2008 - 2010	Mitglied des Kölner Studierendenparlaments
2008 - 2011	Vorsitzender der Kölner Juso-Hochschulgruppe
seit 2009	Langjährige Tätigkeit als Rhetoriktrainer
seit 2010	Mehrere eingeladene Vorträge über Klimaextreme
	Programmierkenntnisse (u.a. C und C++)
	Englisch in Wort und Schrift, Französisch Grundkenntnisse

Gregor Wergen

Köln, 22. Januar 2013