Introduction to the Theory of Complex Systems

Stefan Thurner, Rudolf Hanel, and Peter Klimek

Medical University of Vienna, Austria

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2018
Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2018947065
Data available

ISBN 978–0–19–882193–9
DOI: 10.1093/oso/9780198821939.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Preface

This book is for people who are interested in the science of complex adaptive systems and wish to have more than just a casual understanding of it. As with all the sciences, understanding of complex adaptive systems is reached solely in a quantitative, predictive, and ultimately experimentally testable manner. Complex adaptive systems are dynamical systems that are able to change their structure, their interactions, and, consequently, their dynamics as they evolve in time. This is not a book about complicated systems, even though most complex systems are complicated. Indeed, over the last 300 years, scientists have usually dealt with complicated systems that are neither complex nor adaptive.

The theory of complex systems is the theory of generalized time-varying interactions between elements that are characterized by states. Interactions typically take place on networks that connect those elements. The interactions involved may cause the states of the elements themselves to alter over time. The essence of a complex system is that the interaction networks may change and rearrange as a consequence of changes in the states of the elements. Thus, complex systems are systems whose states change as a result of interactions and whose interactions change concurrently as a result of states. Due to this chicken–egg-type problem, complex systems show an extremely rich spectrum of behaviour: they are adaptive and co-evolutionary; they show path-dependence, emergence, power laws; they have rich phase diagrams; they produce and destroy diversity; they are inherently prone to collapse; they are resilient, and so on. The theory of complex systems tries to understand these properties based on its building blocks and on the interactions between those building blocks that take place on networks. It combines mathematical and physical principles with concepts borrowed from biology and the social sciences; it uses new computational techniques and, with the advent of comprehensive large-scale data sets, is becoming experimentally testable. The goal of the theory of complex systems is to understand the dynamical *systemic* outcomes of interconnected systems, and its ultimate goal is to eventually control and design systemic properties of systems such as the economy, the financial system, social processes, cities, the climate, and ecology. The theory of complex systems builds partly on previous attempts to understand systems that interact in non-trivial ways, such as game theory, cybernetics, or systems theory. However, in its current state, the science of complex systems goes well beyond these earlier developments, in so many ways, in fact, that it can be regarded as an independent scientific branch, which—due to its quantitative, predictive, and testable nature—is a natural science.

Even though it is fair to say that the theory of complex systems is not yet complete, in recent years, it has become quite clear just what the theory is going to look like. Its elements and structure are emerging. The current state of the theory of complex

systems is comparable perhaps to the state of quantum mechanics in the 1920s, before the famous Copenhagen meetings and Werner Heisenberg's book. At that time, quantum mechanics was a collection of experimental and theoretical bits and pieces, which had not yet been seen within a fully comprehensive framework. Nevertheless, it was clear that, one day soon, such a framework would exist. The present situation can be compared to an archaeological project, where a mosaic floor has been discovered and is being excavated. While the mosaic is only partly visible and the full picture is still missing, several facts are becoming clear: the mosaic exists; it shows identifiable elements (for instance, people and animals engaged in recognizable activities); there are large patches missing or still invisible, but experts can already tell that the mosaic represents a scene from, say, Homer's *Odyssey*. Similarly, for dynamical complex adaptive systems, it is clear that a theory exists that, eventually, can be fully developed. There are those who say that complex systems will never be understood or that, by their very nature, they are incomprehensible. This book will demonstrate that such statements are incorrect. The elements of a theory of complex systems are becoming clear: dynamical multilayer networks, scaling, statistical mechanics of algorithmic dynamics, evolution and co-evolution, and information theory. The essence of this book is to focus on these components, clarify their meaning in the context of complex systems, and enable the reader with a mathematical skill set to apply them to concrete problems in the world of complex systems.

The book is written in mathematical language because this is the only way to express facts in a quantitative and predictive manner and to make statements that are unambiguous. We aim for consistency. The book should be comprehensible so that no-one with an understanding of basic calculus, linear algebra, and statistics need refer to other works. The book is particularly designed for graduate students in physics or mathematics. We try to avoid ambiguous statements while, at the same time, being as general as possible. The hope is that this work will serve as a textbook and as a starting point for journeys into new and unexplored territory.

Many complex systems are often sensitive to details in their internal setup, to initial and to boundary conditions. Concepts that proved to be extremely robust and effective in non-complex systems, such as the central limit theorem, classical statistical mechanics, or information theory, lose their predictive power when confronted with complex systems. Extreme care is thus needed in any attempt to apply these otherwise distinguished concepts to complex systems: doing so could end in confusion and nonsensical results. In several concrete examples, we will demonstrate the importance of understanding what these methods mean in the context of complex systems and whether they can or cannot be applied. We will discuss how some of these classical concepts can be generalized to become useful for understanding complex systems.

The book is also a statement about our belief that the exact sciences may be entering a phase of transition from a traditional analytical description of nature, as used with tremendous success since Galileo and Newton, towards an *algorithmic* description. Whereas the analytical description of nature is, conceptually, based largely on differential equations and analytical equations of motion, the algorithmic view takes into account evolutionary and co-evolutionary aspects of dynamics. It provides a framework for

systems that can endogenously change their internal interaction networks, rules of functioning, dynamics, and even environment, as they evolve in time. Algorithmic dynamics, which is characteristic of complex dynamical systems, may be a key to the quantitative and predictive understanding of many natural and man-made systems. In contrast to physical systems, which typically evolve analytically, algorithmic dynamics describe certainly how living, social, environmental, and economic systems unfold. This algorithmic view is not new but has been advocated by authors like Joseph A. Schumpeter, Stuart Kauffman, and Brian Arthur. However, it has not, to date, been picked up by mainstream science, and it has never been presented in the context of the theory of complex systems.

This book is based on a two-semester course, that has been held at the Medical University of Vienna since 2011. We are grateful to our students and to Kathryn Platzer and Anita Wanjek for helping us with the manuscript.

ST Vienna January 2018

Contents

1

Introduction to Complex Systems

1.1 Physics, biology, or social science?

The science of complex systems is not an offspring of physics, biology, or the social sciences, but a unique mix of all three. Before we discuss what the science of complex systems is or is not, we focus on the sciences from which it has emerged. By recalling what physics, biology, and the social sciences are, we will develop an intuitive feel for complex systems and how this science differs from other disciplines. This chapter thus aims to show that the science of complex systems combines physics, biology, and the social sciences in a unique blend that is a new discipline in its own right. The chapter will also clarify the structure of the book.

1.2 Components from physics

Physics makes quantitative statements about natural phenomena. Quantitative statements can be formulated less ambiguously than qualitative descriptions, which are based on words. Statements can be expressed in the form of predictions in the sense that the trajectory of a particle or the outcome of a process can be anticipated. If an experiment can be designed to test this prediction unambiguously, we say that the statement is experimentally testable. Quantitative statements are validated or falsified using quantitative measurements and experiments.

Physics is the experimental, quantitative, and predictive science of matter and its interactions.

Pictorially, physics progresses by putting specific questions to nature in the form of experiments; surprisingly, if the questions are well posed, they result in concrete answers that are robust and repeatable for an arbitrary number of times by anyone who can do the same experiment. This method of generating knowledge about nature, by using experiments to ask questions of it, is unique in the history of humankind and is called *the scientific method*. The scientific method has been at the core of all technological progress since the time of the Enlightenment.

Introduction to the Theory of Complex Systems. Stefan Thurner, Rudolf Hanel, and Peter Klimek, Oxford University Press (2018). © Stefan Thurner, Rudolf Hanel, and Peter Klimek.
DOI: 10.1093/oso/9780198821939.001.0001

Physics deals with matter at various scales and levels of granularity, ranging from macroscopic matter like galaxies, stars, planets, stones, and projectiles, to the scale of molecules, atoms, hadrons, quarks, and gauge bosons. There are four fundamental forces at the core of all interactions between all forms of matter: gravity, electromagnetism, and two types of nuclear force: the weak force and the strong force. According to quantum field theory, all interactions in the physical world are mediated by the exchange of gauge bosons. The graviton, the boson for gravity, has not yet been confirmed experimentally.

1.2.1 The nature of the fundamental forces

The four fundamental forces are very different in nature and strength. They are characterized by a number of properties that are crucial for understanding how and why it was possible to develop physics without computers. These properties are set out here.

Usually, the four fundamental forces are homogeneous and isotropic in space (and time). Forces that are homogeneous act in the same way everywhere in space; forces that are isotropic are the same, regardless of the direction in which they act. These two properties drastically simplify the mathematical treatment of interactions in physics. In particular, forces can be written as derivatives of potentials, two-body problems can effectively be treated as one-body problems, and the so-called mean field approach can be used for many-body systems. The mean field approach is the assumption that a particle reacts to the single field generated by the many particles around it. Often, such systems can be fully understood and solved even without computers. There are important exceptions, however; one being that the strong force acts as if interactions were limited to a 'string', where flux-tubes are formed between interacting quarks, similar to type II superconductivity.

The physical forces differ greatly in strength. Compared to the strong force, the electromagnetic force is about a thousand times weaker, the weak force is about 10^{16} times weaker, and the gravitational force is only 10^{-41} of the strength of the strong force [404]. When any physical phenomenon is being dealt with, usually only a single force has to be considered. All the others are small enough to be safely neglected. Effectively, the superposition of four forces does not matter; for any phenomenon, only one force

Matter	**Relevant interaction types**	**Characteristic length scale**
Macroscopic matter	gravity, electromagnetism	all ranges
Molecules	electromagnetism	all ranges
Atoms	electromagnetism, weak force	$\sim 10^{-18}$ m
Hadrons and leptons	electromagnetism, weak and strong force	$10^{-18} - 10^{-15}$ m
Quarks and gauge bosons	electromagnetism, weak and strong force	$10^{-18} - 10^{-15}$ m

matters. We will see that this is drastically different in complex systems, where a multitude of different interaction types of similar size often have to be taken into account.

Typically, physics does not specify which particles interact with each other, as they interact in identical ways. The interaction strength depends only on the relevant interaction type, the form of the potential, and the relative distance between particles. In complex systems, interactions are often *specific*. Not all elements, only certain pairs or groups of elements, interact with each other. Networks are used to keep track of which elements interact with others in a complex system.

1.2.2 What does predictive mean?

Physics is an experimental and a predictive science. Let us assume that you perform an experiment repeatedly; for example, you drop a stone and record its trajectory over time. The predictive or theoretical task is to predict this trajectory based on an understanding of the phenomenon. Since Newton's time, understanding a phenomenon in physics has often meant being able to describe it with differential equations. A phenomenon is understood dynamically if its essence can be captured in a differential equation. Typically, the following three-step process is then followed:

1. Find the differential equations to encode your understanding of a dynamical system. In the example of our stone-dropping experiment, we would perhaps apply Newton's equation,

$$m\frac{d^2x}{dt^2} = F(x),$$

   where t is time, $x(t)$ is the trajectory, m is mass of the stone, and F is force on the stone. In our case, we would hope to identify the force with gravity, meaning that $F = gm$.
2. Once the equation is specified, try to solve it. The equation can be solved using elementary calculus, and we get, $x(t) = x_0 + v_0 t + \frac{1}{2}gt^2$. To make a testable prediction we have to fix the boundary or initial conditions; in our case we have to specify what the initial position x_0 and initial velocity v_0 are in our experiment. Once this is done, we have a prediction for the trajectory of the stone, $x(t)$.
3. Compare the result with your experiments. Does the stone really follow this predicted path $x(t)$? If it does, you might claim that you have understood something on a quantitative, predictive, and experimental basis. If the stone (repeatedly) follows another trajectory, you have to try harder to find a better prediction.

Fixing initial or boundary conditions means simply taking the system out of its context, separating it from the rest of the universe. There are no factors, other than the boundary conditions, that influence the motion of the system from the outside. That

such a separation of systems from their context is indeed possible is one reason why physics has been so successful, even before computing devices became available. For many complex systems, it is impossible to separate the dynamics from the context in a clear way. This means that many outside influences that are not out of control will simultaneously determine its dynamics.

In principle, the same thinking used to describe physical phenomena holds for arbitrarily complicated systems. Assume that a vector $X(t)$ represents the state of a system at a given time (e.g., all positions and momenta of its elements), we then get a set of equations of motion in the form,

$$\frac{d^2X(t)}{dt^2} = G(X(t)),$$

where G is a high-dimensional function. Predictive means that, in principle, these equations can be solved. Pierre-Simon Laplace was following this principle when he introduced a hypothetical daemon familiar with the Newtonian equations of motion and all the initial conditions of all the elements of a large system (the universe) and thus able to solve all equations. This daemon could then predict everything. The problem, however, is that such a daemon is hard to find. In fact, these equations can be difficult, even impossible, to solve. For three bodies that exert a gravitational force on each other, the famous three-body problem (e.g. Sun, Earth, Moon), there is no general analytical solution provided by algebraic and transcendental functions. This was first demonstrated by Henri Poincaré and paved the way for what is today called chaos theory. In fact, the strict Newton–Laplace program of a predictable world in terms of unambiguously computable trajectories is completely useless for most systems composed of many bodies. Are these large systems not then predictable? What about systems with an extremely large number of elements, such as gases, which contain of the order of $\mathcal{O}(10^{23})$ molecules?

Imagine that we perform the following experiment over and over again: we heat and cool water. We gain the insight that if we cool water to 0°C and below, it will freeze, that if we heat it to 100°C it will start to boil and, under standard conditions, ultimately evaporate. These phase transitions will happen with certainty. In that sense, they are predictable. We cannot predict from the equations of motion which molecule will be the first to leave the liquid. Given appropriate instrumentation, we can perhaps measure the velocity of a few single gas molecules at a point in time, but certainly not all 10^{23}. What can be measured is the probability distribution that a gas molecule is observed with a specific velocity v,

$$p(v) \sim v^2 \exp\left(-\frac{mv^2}{2kT}\right),$$

where T is temperature, and k is Boltzmann's constant. Given this probability distribution, it is possible to derive a number of properties of gases that perfectly describe the *macroscopic* behaviour of gases and make them predictable on a macroscopic

(or systemic) level. For non-interacting particles, these predictions can be extremely precise. The predictions immediately start to degenerate as soon as there are strong interactions between the particles or if the number of particles is not large enough. Note that the term prediction now has a much weaker meaning than in the Newton–Laplace program. The meaning has shifted from being a description based on the exact knowledge of each component of a system to one based on a probabilistic knowledge of the system. Even though one can still make extremely precise predictions about multiparticle systems in a probabilistic framework, the concept of determinism is now diluted. The framework for predictions on a macroscopic level about systems composed of many particles on a probabilistic basis is called statistical mechanics.

1.2.3 Statistical mechanics—predictability on stochastic grounds

The aim of statistical mechanics is to understand the macroscopic properties of a system on the basis of a statistical description of its microscopic components. The idea behind it is to link the microscopic world of components with the macroscopic properties of the aggregate system. An essential concept that makes this link possible is Boltzmann–Gibbs entropy.

A system is often prepared in a macrostate, which means that aggregate properties like the temperature or pressure of a gas are known. There are typically many possible microstates that are associated with that macrostate. A microstate is a possible microscopic configuration of a system. For example, a particular microstate is one for which all positions and velocities of gas molecules in a container are known. There are usually many microstates that can lead to one and the same macrostate; for example, the temperature and pressure in the container. In statistical mechanics, the main task is to compute the probabilities for the many microstates that lead to that single macrostate. In physics, the macroscopic description is often relatively simple. Macroscopic properties are often strongly determined by the phase in which the system is. Physical systems often have very few phases—solid, gaseous, or liquid.

Following the Newton–Laplace framework, traditional physics works with extreme precision for a few particles and for many non-interacting particles, where the statistical mechanics of Boltzmann–Gibbs applies. In other words, the class of systems that can be understood with traditional physics is not that big. Most systems are composed of many strongly interacting particles. Often, the interactions are of multiple types, are non-linear, and vary over time. Very often, such systems are complex systems.

1.2.4 The evolution of the concept of predictability in physics

The concept of prediction and predictability has changed in significant ways over the past three centuries. Prediction in the eighteenth century was quite different from the concept of prediction in the twenty-first. The concept of determinism has undergone at least three transitions [299].

In the *classical mechanics* of the eighteenth and nineteenth centuries, prediction meant the exact prediction of trajectories. Equations of motion would make exact statements about the future evolution of simple dynamical systems. The extension to more than two bodies has been causing problems since the very beginning of Newtonian physics; see, for example, the famous conflict between Isaac Newton and John Flamsteed on the predictability of the orbit of the Moon. By about 1900, when interest in understanding many-body systems arose, the problem became apparent. The theory of Ludwig Boltzmann, referred to nowadays as statistical mechanics, was effectively based on the then speculative existence of atoms and molecules, and it drastically changed the classical concept of predictability.

In *statistical mechanics*, based on the assumption that atoms and molecules follow Newtonian trajectories, the law of large numbers allows stochastic predictions to be made about the macroscopic behaviour of gases. Statistical mechanics is a theory of the macroscopic or collective behaviour of non-interacting particles. The concepts of predictability and determinism were subject to further change in the 1920s with the emergence of quantum mechanics and non-linear dynamics.

In *quantum mechanics*, the concept of determinism disappears altogether due to the fundamental simultaneous unpredictability of the position and momentum of the (sub-)atomic components of a system. However, quantum mechanics still allows us to make extremely high-quality predictions on a collective basis. Collective phenomena remain predictable to a large extent on a macro- or systemic level.

In *non-linear systems*, it became clear that even in systems for which the equations of motion can be solved in principle, the sensitivity to initial conditions can be so enormous that the concept of predictability must, for all practical purposes, be abandoned. A further crisis in terms of predictability arose in the 1990s, when interest in more general forms of interactions began to appear.

In *complex systems*, the situation is even more difficult than in quantum mechanics, where there is uncertainty about the components, but not about its interactions. For many complex systems, not only can components be unpredictable, but the interactions between components can also become specific, time-dependent, non-linear, and unpredictable. However, there is still hope that probabilistic predictions about the dynamics and the collective properties of complex systems are possible. Progress in the science of complex systems will, however, be impossible without a detailed understanding of the dynamics of how elements specifically interact with each other. This is, of course, only possible with massive computational effort and comprehensive data.

1.2.5 Physics is analytic, complex systems are algorithmic

Physics largely follows an *analytical* paradigm. Knowledge of phenomena is expressed in analytical equations that allow us to make predictions. This is possible because interactions are homogeneous, isotropic, and of a single type. Interactions in physics typically do not change over time. They are usually given and fixed. The task is to work out specific solutions regarding the evolution of the system for a given set of initial and boundary conditions.

This is radically different for complex systems, where interactions themselves can change over time as a consequence of the dynamics of the system. In that sense, complex systems change their internal interaction structure as they evolve. Systems that change their internal structure dynamically can be viewed as *machines* that change their internal structure as they operate. However, a description of the operation of a machine using analytical equations would not be efficient. Indeed, to describe a steam engine by seeking the corresponding equations of motion for all its parts would be highly inefficient. Machines are best described as *algorithms*—a list of rules regarding how the dynamics of the system updates its states and future interactions, which then lead to new constraints on the dynamics at the next time step. First, pressure builds up here, then a valve opens there, vapour pushes this piston, then this valve closes and opens another one, driving the piston back, and so on.

Algorithmic descriptions describe not only the evolution of the states of the components of a system, but also the evolution of its internal states (interactions) that will determine the next update of the states at the next time step. Many complex systems work in this way: states of components and the interactions between them are simultaneously updated, which can lead to the tremendous mathematical difficulties that make complex systems so complicated to understand. These difficulties in their various forms will be addressed time and again in this book. Whenever it is possible to ignore the changes in the interactions in a dynamical system, analytic descriptions become meaningful.

Physics is generally analytic: complex systems are algorithmic. Quantitative predictions that can be tested experimentally can be made with analytic or algorithmic descriptions.

1.2.6 What are complex systems from a physics point of view?

From a physics point of view, one could try to characterize complex systems by the following extensions to physics.

- Complex systems are composed of many elements, components, or particles. These elements are typically described by their state, velocity, position, age, spin, colour, wealth, mass, shape, and so on. Elements may have stochastic components.
- Elements are not limited to physical forms of matter; anything that can interact and be described by states can be seen as generalized matter.
- Interactions between elements may be specific. Who interacts with whom, when, and in what form is described by interaction networks.
- Interactions are not limited to the four fundamental forces, but can be of a complicated type. Generalized interactions are not limited to the exchange of gauge bosons, but can be mediated through exchange of messages, objects, gifts, information, even bullets, and so on.

continued

- Complex systems may involve superpositions of interactions of similar strengths.
- Complex systems are often chaotic in the sense that they depend strongly on the initial conditions and details of the system. Update equations that algorithmically describe the dynamics are often non-linear.
- Complex systems are often driven systems. Some systems obey conservation laws, some do not.
- Complex systems can exhibit a rich phase structure and have a huge variety of macrostates that often cannot be inferred from the properties of the elements. This is sometimes referred to as *emergence*. Simple forms of emergence are, of course, already present in physics. The spectrum of the hydrogen atom or the liquid phase of water are emergent properties of the atoms involved and their interactions.

With these extensions, we can derive a physics-based definition for what the theory of complex systems is.

The theory of complex systems is the experimental, quantitative, and predictive science of generalized matter interacting through generalized interactions.

Generalized interactions are described by the interaction type and who interacts with whom at what time and at what strength. If there are more than two interacting elements involved, interactions can be conveniently described by time-dependent networks,

$$M_{ij}^{\alpha}(t),$$

where i and j label the element in the system, and α denotes the interaction type. $M_{ij}^{\alpha}(t)$ are matrix elements of a structure with three indices. The value $M_{ij}^{\alpha}(t)$ indicates the strength of the interaction of type α between element i and j at time t. $M_{ij}^{\alpha}(t)=0$ means no interaction of that type. Interactions in complex systems remain based on the concept of exchange; however, they are not limited to the exchange of gauge bosons. In complex systems, interactions can happen through communication, where messages are exchanged, through trade where goods and services are exchanged, through friendships, where bottles of wine are exchanged, and through hostility, where insults and bullets are exchanged.

Because of more specific and time-varying interactions and the increased variety of types of interaction, the variety of macroscopic states and systemic properties increases drastically in complex systems. This diversity increase of macrostates and phenomena emerges from the properties both of the system's components and its interactions. The phenomenon of collective properties arising that are, a priori, unexpected from the elements alone is sometimes called *emergence*. This is mainly a consequence of the presence of generalized interactions. Systems with time-varying generalized interactions can exhibit an extremely rich phase structure, and may be adaptive. Phases may co-exist in particular complex systems. The plurality of macrostates in a system leads to new types

of questions that can be addressed, such as: What is the number of macrostates? What are their co-occurrence rates? What are the typical sequences of occurrence? What are the life-times of macrostates? What are the probabilities of transition between macrostates? As yet, there are no general answers to these questions, and they remain a challenge for the theory of complex systems. For many complex systems, the framework of physics is incomplete. Some of the missing concepts are those of non-equilibrium, evolution, and co-evolution. These concepts will be illustrated in the sections that follow.

1.2.7 A note on chemistry—the science of equilibria

In chemistry, interactions between atoms and molecules are specific in the sense that not every molecule binds to (interacts with) any other molecule. So why is chemistry usually not considered to be a candidate for a theory of complex systems? To a large extent, chemistry is based on the law of mass action. Many particles interact in ways that lead to equilibrium states. For example, consider two substances A and B that undergo a reaction to form substances S and T,

$$\alpha A + \beta B \rightleftharpoons \sigma S + \tau T,$$

where $\alpha, \beta, \sigma, \tau$ are the stoichiometric constants, and k_+ and k_- are the forward and backward reaction rates, respectively. The forward reaction happens at a rate that is proportional to $k_+\{A\}^\alpha\{B\}^\beta$, the backward reaction is proportional to $k_-\{S\}^\sigma\{T\}^\tau$. The brackets indicate the active (reacting) masses of the substances. Equilibrium is attained if the ratio of the reaction rates equals a constant K,

$$K = \frac{k_+}{k_-} = \frac{\{S\}^\sigma\{T\}^\tau}{\{A\}^\alpha\{B\}^\beta}.$$

Note that the solution to this equation gives the stationary concentrations of the various substances. Technically, these equations are fixed point equations. In contrast to chemical reactions and statistical mechanics, many complex systems are characterized by being out-of-equilibrium. Complex systems are often so-called driven systems, where the system is (exogenously) driven away from its equilibrium states. If there is no equilibrium, there is no way of using fixed-point-type equations to solve the problems. The mathematical difficulties in dealing with out-of-equilibrium or non-equilibrium systems are tremendous and beyond analytical reach. One way that offers a handle on understanding driven out-of-equilibrium systems is the concept of self-organized criticality, which allows essential elements of the statistics of complex systems to be understood; in particular, the omnipresence of power laws.

Many complex systems are driven systems and are out-of-equilibrium.

By comparing the nature of complex systems and basic equilibrium chemistry, we learn that the mere presence of specific interactions does not automatically lead us to complex systems. However, cyclical catalytic chemical reactions [22, 113, 204], are classic prototypes of complex systems.

1.3 Components from the life sciences

We now present several key features of complex systems that have been adopted from biology. In particular, we discuss the concepts of evolution, adaptation, self-organization, and, again, networks.

The life sciences describe the experimental science of living matter. What is living matter? A reasonable minimal answer has been attempted by the following three statements [222],

- Living matter must be self-replicating.
- It must run through at least one Carnot cycle.
- It must be localized.

Life without self-replication is not sustainable. It is, of course, conceivable that non-self-replicating organisms can be created that live for a time and then vanish and have to be recreated. However, this is not how we experience life on the planet, which is basically a single, continuous, living germ line that originated and has existed ever since. A Carnot cycle is a thermodynamic cyclical process that converts thermal energy into work, or vice versa. Starting from an initial state, after the cycle is completed, the system returns to the same initial state. The notion that living matter must perform at least one Carnot cycle is motivated by the fact that all living organisms use energy gradients (usually thermal) to perform work of some kind. For example, this work could be used for moving or copying DNA molecules. This view also pays tribute to the fact that all living objects are out-of-equilibrium and constantly driven by energy gradients. If, after performing work, a system were not able to reach its previous states, it would be hard to call it a living system. Both self-replication and Carnot cycles require some sort of localization. On this planet, this localization typically happens at the level of cells.

Living matter uses energy and performs work on short timescales without significantly transforming itself. It is constantly driven by energy gradients and is out-of-equilibrium. Self-replication and Carnot cycles require localization.

1.3.1 Chemistry of small systems

Living matter, as we know it on this planet, is a self-sustained sequence of genetic activity over time. By genetic activity we mean that genes can be turned 'on' and 'off'. If a gene is

on, it triggers the production of molecular material, such as ribonucleic acid (RNA) that can later be translated into proteins. A gene is typically turned on by a cluster of proteins that bind to each other to form a so-called 'complex'. If such a cluster binds to a specific location on the DNA, this could cause a copying process to be activated at this position; the gene is then active or 'on'; see Figure 1.1.

Genetic activity is based on chemical reactions that take place locally, usually within cells or their nuclei. However, these chemical reactions are special in the sense that only a few molecules are involved [340]. In traditional chemistry, reactions usually involve billions of atoms or molecules. What happens within a cell is chemistry with a *few* molecules. This immediately leads to a number of problems:

- It can no longer be assumed that molecules meet by chance to react.
- With only a few molecules present that might never meet to react, the concept of equilibrium becomes useless.
- Without equilibrium, there is no law of mass action.

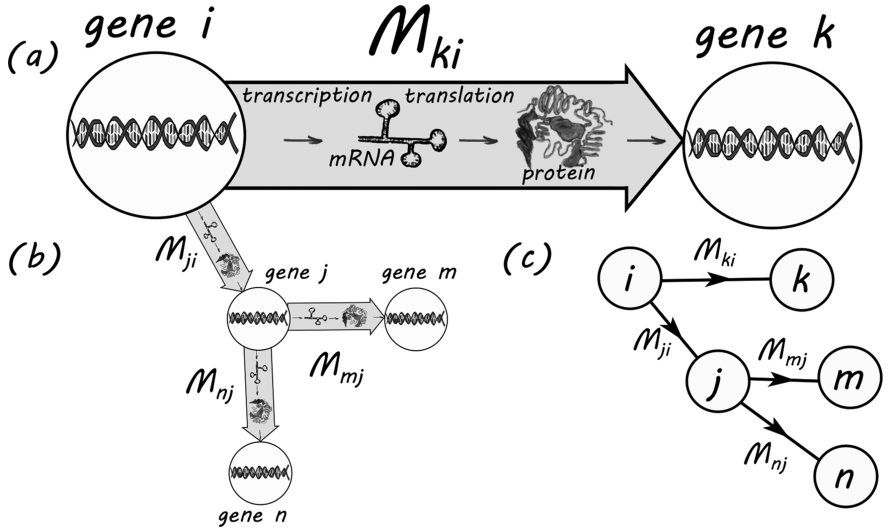

Figure 1.1 *Schematic view of genetic activity and what a link M_{ik} means in a genetic regulatory network. (a) Gene i activates gene j if something like the following process takes place: the activity of gene i means that a specific sub-sequence of the deoxyribonucleic acid (DNA) (gene) is copied into a complementary structure, an mRNA molecule. This mRNA molecule from gene i, might get 'translated' (copied again) into a protein of type i. This protein can bind with other proteins to form a cluster of proteins, a 'complex'. (b) Some of these complexes can bind to other regions of the DNA (say, the region that is associated with gene k) and thereby cause the activation of gene j. Effectively, gene i causes gene j to become active. (c) The whole process, where the activity of gene i triggers the activity of gene j, can be represented as a directed link A_{ij} in a genetic regulatory network. Complexes can also deactivate genes. If gene j is already active, another complex might deactivate it.*

If there is no law of mass action, how can chemistry be done? Classical equilibrium chemistry is inadequate for dealing with molecular mechanisms in living matter. In cells, molecules are often actively transported from the site of production (typically, the nucleus, for organisms that have one) to where they are needed in the cell. This means that diffusion of molecules no longer follows the classical diffusion equation. Instead, molecular transport is often describable by an anomalous diffusion equation of the form,

$$\frac{d}{dt}p(x,t) = D\frac{d^{2+\nu}}{dx^{2+\nu}}p(x,t)^{\mu},$$

where $p(x,t)$ is the probability of finding a molecule at position x at time t, D is the diffusion constant, and μ and ν are exponents that make the diffusion equation non-linear.

Chemical binding often depends on the three-dimensional structure of the molecules involved. This structure can depend on the 'state' of the molecules. For example, a molecule can be in a normal or a phosphorylated state. Phosphorylation happens through the addition of a phosphoryl group (PO_3^{2-}) to a molecule, which may change its entire structure. This means that for a particular state of a molecule it binds to others, but does not bind if it is in the other state. A further complication in the chemistry of a few particles arises with the reaction rates. By definition, the term reaction rate only makes sense for sufficiently large systems. The speed of reactions depends crucially on the statistical mechanics of the underlying small system and fluctuation theorems may now become important [122].

1.3.2 Biological interactions happen on networks—almost exclusively

Genetic regulation governs the temporal sequence of the abundance of proteins, nucleic material, and metabolites within any living organism. To a large extent, genetic regulation can be viewed as a discrete interaction: a gene is active or inactive; a protein binds to another or it does not; a molecule is phosphorylated or not. Discrete interactions are well-described by networks. In the context of the life sciences, three well-known networks are the metabolic network, the protein–protein binding network, and the Boolean gene-regulatory network. The metabolic network[1] is the set of linked chemical reactions occurring within a cell that determine the cell's physiological and biochemical properties. The metabolic network is often represented in networks of chemical reactions, where nodes represent substances and directed links (arrows) correspond to reactions or catalytic influences. The protein–protein networks represent empirical findings about protein–protein interactions (binding) in network representations [102]. Nodes are proteins, and links specify the interaction type between them. Different interaction types include stable, transient, and homo- or hetero-oligomer interactions.

[1] For an example of what metabolic networks look like, see http://biochemical-pathways.com/#/map/1

1.3.3 Evolution

> 'Nothing in biology makes sense except in the light of evolution'. Theodosius Dobzhansky

Evolution is a natural phenomenon. It is a process that increases and destroys diversity, and it looks like both a 'creative' and a 'destructive' process. Evolution appears in biological, technological, economical, financial, historical, and other contexts. In that sense, evolutionary dynamics is universal. Evolutionary systems follow characteristic dynamical and statistical patterns, regardless of the context. These patterns are surprisingly robust and, as a natural phenomenon, they deserve a quantitative and predictive scientific explanation.

What is evolution? Genetic material and the process of replication involve several stochastic components that may lead to variations in the offspring. Replication and variation are two of the three main ingredients of evolutionary processes. What evolution means in a biological context is captured by the classic Darwinian narrative. Consider a population of some kind that is able to produce offspring. This offspring has some random variations (e.g. mutations). Individuals with the optimal variations with respect to a given environment have a selection advantage (i.e. higher fitness). Fitness manifests itself by higher reproductive success. Individuals with optimal variations will have more offspring and will thus pass their particular variations on to a new generation. In this way 'optimal' variations are selected over time. This is certainly a convincing description of what is going on; however, in this form it may not be useful for predictive science. How can we predict the fitness of individuals in future generations, given that life in future environments will look very different from what it is today? Except over very short time periods, this is a truly challenging task that is far from understood. There is a good prospect, however, of the *statistics* of evolutionary systems being understood. The Darwinian scenario fails to explain essential features about evolutionary systems, such as the existence of boom and crash phases, where the diversity of systems radically changes within short periods of time. An example is the massive diversification (explosion) of species and genera about 500 million years ago in the Cambrian era. It will almost certainly never be possible to predict what species will live on Earth, even 500,000 years from now, but it may be perfectly possible to understand the statistics of these events and the factors that determine the statistics. In particular, statistical statements about expected diversity, diversification rates, robustness, resilience, and adaptability are coming within reach. In Chapter 5 we will discuss approaches to formulating evolutionary dynamics in ways that make them accessible both combinatorially and statistically.

The concept of evolution is not limited to biology. In the economy, the equivalent of biological evolution is innovation, where new goods and services are constantly being produced by combination of existing goods and services. Some new goods will be selected in markets, while the majority of novelties will not be viable and will vanish. The industrial revolution can be seen as one result of evolutionary dynamics, leading, as it did, to an ongoing explosion of diversification of goods, services, and innovations.

Another example of evolutionary dynamics outside biology is the sequence of invention and discovery of chemical compounds. The history of humankind itself is an example of evolutionary dynamics. Evolutionary dynamics can take place simultaneously at various scales. In biological settings, it works at the level of cells, organisms, and populations; in economic settings, it can work at product, firm, corporation, and country level. A famous application of evolutionary dynamics in computer science are so-called genetic algorithms [194]. These algorithms mimic natural selection by iteratively producing copies of computer code with slight variations. Those copies that perform best for a given problem (usually an optimization task) are iteratively selected.

1.3.3.1 Evolution is not physics

To illustrate that evolution is not a process that can be described with traditional physics, we define an evolutionary process as a three-step process:

1. A new thing comes into existence within a given *environment*.
2. The new thing has the chance to interact with its environment. The result of this interaction is that it gets 'selected' (survives) or is destroyed.
3. If the new thing gets selected in the environment, it becomes part of this environment (boundary) and thus transforms the old environment into a new one. New and arriving things in the future will experience the new environment. In that sense, evolution is an algorithmic process that co-evolves its boundaries.

If we try to interpret this three-step process in terms of physics, we immediately see that even if we were able to write down the dynamics of the system in the form of equations of motion, we would not be able to fix the system's boundary conditions. Obviously, the environment plays the role of the boundary conditions within which the interactions happen. The boundary conditions evolve as a consequence of the dynamics of the system and change at every instant. The dynamics of the boundary conditions is dynamically coupled with the equations of motion. Consequently, as the boundary conditions cannot be fixed, this set of equations cannot, in general, be solved and the Newtonian method breaks down. A system of dynamical equations that are coupled dynamically to their boundary conditions is a mathematical monster. That is why an algorithmic process like evolution is hard to solve using analytical approaches.[2]

The second problem associated with evolutionary dynamics, from a physics point of view, is that the phasespace is not well-defined. As new elements may arrive at any point in time, it is impossible to prestate what the phasespace of such systems will be. Obviously, this poses problems in terms of producing statistics with these systems. The situation could be compared to trying to produce statistics by rolling a dice, the face of which changes from one throw to the next.

[2] Such systems can be treated analytically where the characteristic timescales of the processes involved are different. In our example, this would be the case if the dynamics of the interactions of the 'new thing' with the environment happens on a fast timescale, while changes in the environment happen slowly.

Evolutionary dynamics is radically different from physics for two main reasons:

- In evolutionary systems, boundary conditions cannot be fixed.
- In evolutionary systems, the phasespace is not well defined—it changes over time. New elements may emerge that change the environment and therefore also the dynamics for all the existing elements of the system.

Evolutionary aspects are essential for many complex systems and cannot be ignored. A great challenge in the theory of complex systems is to develop a consistent framework that is nevertheless able to deal with evolutionary processes in quantitative and predictive terms. We will see how a number of recently developed mathematical methods can be used to address and deal with these two fundamental problems. In particular, in Chapter 5, we will discuss combinatorial evolution models. These models are a good example of how algorithmic approaches lead to quantitative and testable predictions.

1.3.3.2 The concept of the adjacent possible

A helpful steppingstone in addressing the problem of dynamically changing phasespaces is the concept of the *adjacent possible*, proposed by Stuart Kauffman [222]. The adjacent possible is the set of all possible states of the world that could potentially exist in the next time step, given the present state of the world. It drastically reduces the size of phasespace from all possible states to a set of possibilities that are conditional on the present. Obviously, not everything can be produced within the next time step. There are many states that are impossible to imagine, as the components required to make them do not yet exist. The adjacent possible is the subset of all possible worlds that are *reachable* within the next time step and depends strongly on the present state of the world. In this view, evolution is a process that continuously 'fills' the adjacent possible. The concrete realization of the adjacent possible at one time step determines the adjacent possible at the next time step.

Thus, in the context of biological evolution or technological innovation, the adjacent possible is a huge set, in which the present state of the world determines the potential realization of a vast number of possibilities in the next time step. Typically, the future states are not known. In contrast, in physics, a given state often determines the next state with high precision. This means that the adjacent possible is a very small set. For example, the adjacent possible of a falling stone is given by the next position (point) on its parabolic trajectory. In contrast, the adjacent possible of an ecosystem consists of all organisms that can be born within the next time step, with all possible mutations and variations that can possibly happen—a large set of possibilities indeed. The concept of the adjacent possible introduces path-dependence in the stochastic dynamics of phasespace. We will discuss the statistics of path-dependent evolutionary processes in Chapters 5 and 6.

1.3.3.3 Summary evolutionary processes

We now summarize what we have learned about evolutionary processes that are relevant to the treatment of complex systems.

- For evolutionary systems, boundary conditions cannot usually be fixed. This means that it is impossible to take the system apart and separate it from its context without massively altering and perhaps even destroying it. The concept of reductionism is inadequate for describing evolutionary processes.
- Evolutionary complex systems change their boundary conditions as they unfold in time. They co-evolve with their boundary conditions. Frequently, situations are difficult or impossible to solve analytically.
- For complex systems, the adjacent possible is a large set of possibilities. For physics, it is typically a very small set.
- The adjacent possible itself evolves.
- In many physical systems, the realization of the adjacent possible does not influence the next adjacent possible; in evolutionary systems, it does.

1.3.4 Adaptive and robust—the concept of the edge of chaos

Many complex systems are robust and adaptive at the same time. The ability to adapt to changing environments and to be robust against changing environments seem to be mutually exclusive. However, most living systems are clearly adaptive and robust at the same time. As an explanation for how these seemingly contradictory features could co-exist, the following view of the *edge of chaos* was proposed [245]. Every dynamical system has a maximal Lyapunov exponent, which measures how fast two initially infinitesimally close trajectories diverge over time. The exponential rate of divergence is the Lyapunov exponent, λ,

$$|\delta X(t)| \sim e^{\lambda t}|\delta X(0)|,$$

where $|\delta X(t)|$ is the distance between the trajectories at time t and $|\delta X(0)|$ is the initial separation. If λ is positive, the system is called *chaotic* or strongly mixing. If λ is negative, the system approaches an attractor, meaning that two initially infinitesimally separated trajectories converge. This attractor can be a trivial point (fixed point), a line (limit cycle), or an entire fractal object. If the system approaches a non-trivial attractor, the system is periodic. An interesting case arises when exponent λ is exactly zero. The system is then called quasi-periodic or at the 'edge of chaos'. There are many low-dimensional examples where systems exhibit all three possibilities—they can be chaotic, periodic, or at the edge of chaos, depending on their control parameters. The simplest of these is the logistic map.

The intuitive understanding of how a system can be adaptive and robust at the same time if it operates at the edge of chaos, is given by the following. If λ is close to zero, it takes only tiny changes in the system to move it from a stable and periodic mode (λ slightly negative) to the chaotic phase (λ positive). In the periodic mode, the system is stable and robust; it returns to the attractor when perturbed. When it transits into the chaotic phase, say, through a strong perturbation in the environment, it will sample

large regions of phasespace very quickly. By sampling large volumes of phasespace, the system has the chance to find new 'solutions' that are compatible with the perturbed environment and can then again settle into a periodic phase. The system adapted by sampling other optima. This situation is similar to a simulated annealing procedure in a computer code.

1.3.4.1 How does nature find the edge of chaos?

For many non-linear dynamical systems, the set of points at which the Lyapunov exponents are exactly zero is very limited and often of measure zero; see Figure 1.2a. If living systems operate at the edge of chaos, how did evolution find and select these points? How can a mechanism of evolution detect something that is of measure zero? One possible explanation is that the regions where the Lyapunov exponent is zero do not measure zero, but are extended areas in parameter space. Indeed, simple evolutionary models do show inflated regions of the edge of chaos [355]; see Figure 1.2b. We discuss situations where this possibility seems to be realized in Chapter 5. Another explanation is self-organized criticality. This is a mechanism that allows systems to endogenously organize themselves to operate at a critical point that separates the chaotic from the regular regime [24, 225]. A critical point in physics is a point in parameter space where a system is at a phase transition point. These points are reached at conditions (temperature, pressure, slope in a sand pile, etc.) where characteristic length-scales (e.g. correlation length) become divergent. Think, for example, of the critical temperature at which a

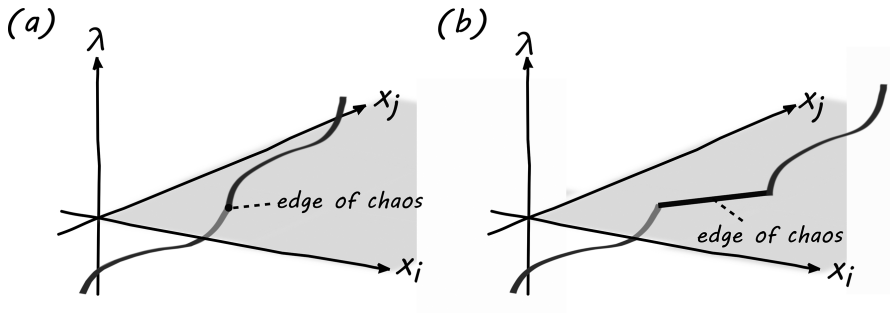

Figure 1.2 *Schematic view of the edge of chaos. (a) Shows the largest Lyapunov exponent in the parameter space (indicated by x_i and x_j) of a dynamical system. The largest exponent dominates the dynamics of the system. For every point in parameter space, the Lyapunov exponent is either positive (chaotic), negative (periodic), or exactly zero—which means at the 'edge of chaos'. The edge of chaos offers a pictorial understanding of how systems can be adaptive and robust at the same time. Systems at the edge of chaos can be thought of as being usually just inside the regular regime. They can transit to the chaotic region very quickly. Once in the chaotic regime, vast volumes of phasespace can be explored which possibly contain regions that are 'better' for the system. Once such a region is found, the system can transit back to the periodic region. Note that the regions of parameter space where the Lyapunov exponent is zero is tiny. Some evolutionary complex systems show that this region can be drastically inflated. This will be discussed in Chapter 5. (b) Shows a situation for an inflated edge of chaos.*

magnetization transition occurs, the Curie temperature T_c. Above that temperature there is no magnetization. Approaching the Curie temperature shows that, at the critical point, certain quantities, such as the magnetic susceptibility χ start to diverge as a power law, $\chi = (T - T_c)^{-\gamma}$, where γ is a critical exponent. In many physical systems, these transition points are unique and parameters have to be fine-tuned to these critical points to find the power law behaviour. Self-organized critical systems manage to find these critical points endogenously, without any fine-tuning. Self-organized criticality seems to be realized in very different systems, including sand piles, precipitation, heartbeats, avalanches, forest fires, earthquakes, financial markets, combinatorial evolution, and so on. It often occurs in slowly driven systems, where driven means that they are driven away from equilibrium.

Self-organized critical systems are dynamical, out-of-equilibrium systems that have a critical point as an attractor. These systems are characterized by (approximate) scale invariance or 'scaling'. Scale invariance means the absence of characteristic scales, and it often manifests itself in the form of power laws in the associated probability distribution functions. Self-organized criticality is one of the classic ways of understanding the origin of power laws, which are omnipresent in complex systems. Other ways of understanding power laws include criticality, multiplicative processes with constraints, preferential dynamics, entropy methods, and sample space reducing processes. We will discuss the mechanisms for understanding the origin of power laws in Chapter 3.

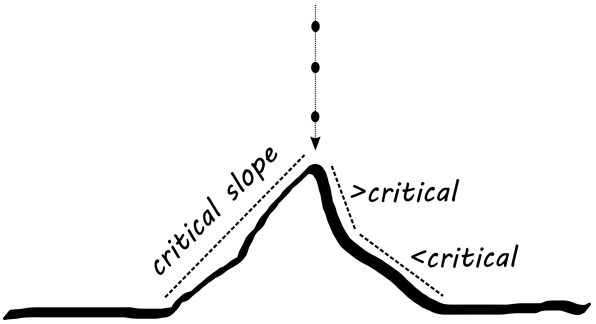

Figure 1.3 *Sand pile models are models for self-organized criticality. These systems self-organize towards a critical state, which is characterized by the fact that the system develops dynamics that are felt across the entire system; in other words it develops diverging correlation lengths. In sand piles, the system organizes towards a critical slope, above which avalanches of sand will occur to reduce the slope, and below which sand is deposited close to the sites where the sand falls down, which makes the slope steeper. The size distribution of avalanches is a power law, meaning that it covers a large spectrum of sizes. The frequency distribution of the occurrence of avalanches is also a power law. Besides being a playground for understanding self-organized criticality, sand pile models have practical applications in the science of earthquakes and collapse.*

1.3.4.2 Intuition behind self-organized critical systems—sand pile models

Imagine sand is dropped on to a table as shown in Figure 1.3. A pile gradually builds up. If you consider the slopes of the pile, you will observe that they are not constant but that they vary in terms of their slope angles. If the slope becomes too steep, avalanches take place and the slope becomes flatter again. If the slope is flat, sand becomes deposited and the slope becomes steeper. In other words, the pile *self-organizes* itself towards a critical slope. The system is robust and adaptive.

1.3.5 Components taken from the life sciences

Let us now put together the components that we need for a description of complex systems adapted from the life sciences.

- Interactions between elements happen specifically. They take place on networks. Often, interactions are discrete; they either happen or not. Often, systems operate algorithmically.
- Complex systems are out-of-equilibrium, which makes them hard to deal with analytically. If they are self-organized critical, the statistics behind them can be understood.
- Many complex systems follow evolutionary dynamics. As they progress, they change their own environment, their context, or their boundary conditions.
- Many complex systems are adaptive and robust at the same time. They operate at the 'edge of chaos'. Self-organized criticality is a mechanism that regulates systems towards their edge of chaos.
- Most evolutionary complex systems are path-dependent and have memory. They are therefore non-ergodic and non-Markovian. The adjacent possible is a way of conceptualizing the evolution of the 'reachable' phasespace.

1.4 Components from the social sciences

Social science is the science of social interactions and their implications for society.

Usually, social science is neither very quantitative or predictive, nor does it produce experimentally testable predictions. It is largely qualitative and descriptive. This is because, until recently, there was a tremendous shortage of data that are both time-resolved (longitudinal) and multidimensional. The situation is changing fast with the new tendency of homo sapiens to leave electronic fingerprints in practically all dimensions of life. The centuries-old data problem of the social sciences is rapidly disappearing. Another fundamental problem in the social sciences is the lack of reproducibility or

repeatability. On many occasions, an event takes place once in history and no repeats are possible.

As in biology, social processes are hard to understand mathematically because they are evolutionary, path-dependent, out-of-equilibrium, and context-dependent. They are high-dimensional and involve interactions on multiple levels. The methodological tools used by traditional social scientists have too many shortcomings to address these issues appropriately. Unfortunately, the tools of social science practised today rarely extend much beyond linear regression models, basic statistics, and game theory. In some branches of social science, such as economics, there is a tradition of ignoring the scientific principle in the sense that data, even if they are available, are not taken seriously. There are influential 'theories' that are in plain contrast to experimental evidence, including the capital asset pricing model, the efficient market hypothesis, or the Markowitz portfolio theory. These concepts earned their inventors Nobel Prizes. What has also not occurred to date in the social sciences are the massive joint and coordinated efforts between scientists that have been taking place in physics (CERN), biology (the genome project), and climate science. However, two important components have been developed in the social sciences and play a crucial role in the theory of complex systems:

- *Multilayer interaction networks.* In social systems, interactions happen simultaneously at more or less the same strength scale on a multitude of superimposed interaction networks. Social scientists, in particular sociologists, have recognized the importance of social networks[3] since the 1970s [156, 397].
- *Game theory.* Another contribution from the social sciences, game theory is a concept that allows us to determine the outcome of rational interactions between agents trying to optimize their payoff or utility [392]. Each agent is aware that the other agent is rational and that he/she also knows that the other agent is rational. Before computers arrived on the scene, game theory was one of the very few methods of dealing with complex systems. Game theory can easily be transferred to dynamical situations, and it was believed for a long time that iterative game-theoretic interactions were a way of understanding the origin of cooperation in societies. This view is now severely challenged by the discovery of so-called zero-determinant games [315]. Game theory was first developed and used in economics but has penetrated other fields of social and behavioural sciences.

1.4.1 Social systems continuously restructuring networks

Social systems can be thought of as time-varying multilayer (multiplex) networks. Nodes are individuals or institutions, and links are interactions of different types. Interactions

[3] Interestingly, until very recently, networks were not recognized as relevant by the 'queen' of the social sciences, economics, even though networks clearly dominate practically every domain of the economy. Mainstream economics has successfully ignored networks—production networks, distribution networks, trading and consumption networks, ownership networks, information networks, and financial networks—for more than a century in favour of a rather unrealistic equilibrium view of the economy.

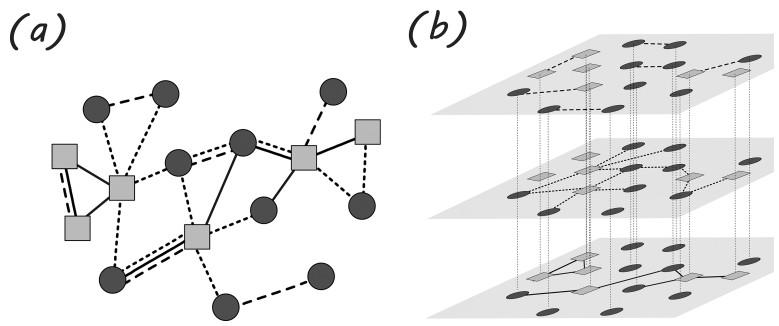

Figure 1.4 *Two schematic representations of the same multilayer network. Nodes are characterized by a two-dimensional state vector. The first component is given by colours (light- and dark-grey) the second by shapes (circles, squares). Nodes interact through three types of interaction that are represented by (full, broken, and dotted) lines. (a) Shows the projection of the multilayer network to a single layer, whereas in (b) each type of link is shown in a different layer. The system is complex if states simultaneously change as a function of the interaction network and if interactions change as a function of the states; see Equation 1.1. The multilayer network could represent a network of banks at a given moment, where shapes represent the wealth of the bank and the links could represent financial assets that connect banks with each other. A full line could mean a credit relation; a broken line could represent derivatives trading, and dotted lines indicate if one bank owns shares in another bank. Depending on that network, the banks will make a profit or loss in the next time step.*

change over time. The types of link can be friendship, family ties, processes of good exchange, payments, trust, communication, enmity, and so on. Every type of link is represented by a separate network layer; see, for example, Figure 1.4. Individuals interact through a superposition of these different interaction types (multilayer network), which happen simultaneously and are often of the same order of magnitude in 'strength'. Often, networks at one level interact with networks at other levels. Networks that characterize social systems show a rich spectrum of growth patterns and a high level of plasticity. This plasticity of networks arises from restructuring processes through link creation, relinking, and link removal. Understanding and describing the underlying restructuring dynamics can be very challenging; however, there are a few typical and recurring dynamical patterns that allow for scientific progress. We will discuss network dynamics in Chapter 4.

Individuals are represented by states, such as their wealth, gender, education level, political opinion, age, and so on. Some of these states can dynamically change over time. States typically have an influence on the linking dynamics of their corresponding node. If that is the case, a tight connection exists between network structure and node states. The joint dynamics of network restructuring and changes of states by individuals is a classic example of co-evolution.

1.5 What are Complex Systems?

We present a one-sentence summary of what complex systems are, which covers most of the features discussed in the previous sections:

Complex systems are co-evolving multilayer networks.

This statement summarizes the following ten facts about complex systems and provides an intuitive picture of the essence of complex systems.

1. Complex systems are composed of many elements. These are labelled with latin indices i.
2. These elements interact with each other through one or more interaction types, labelled with greek indices α. Interactions are often specific between elements. To keep track of which elements interact, we use networks. Interactions are represented as links in the interaction networks. The interacting elements are the nodes in these networks. Every interaction type can be seen as one network layer in a multilayer network; see Figure 1.4. A multilayer network is a collection of networks linking the same set of nodes. If these networks evolve independently, multilayer networks are superpositions of networks. However, there are often interactions between interaction layers.
3. Interactions are not static but change over time. We use the following notation to keep track of interactions in the system. The strength of an interaction of type α between two elements i and j at time t is denoted by,

   $$M_{ij}^{\alpha}(t) \quad \text{interaction strength.}$$

   Interactions can be physical, chemical, social, or symbolic. Most interactions are mediated through some sort of exchange process between nodes. In that sense, interaction strength is often related to the quantity of objects exchanged (gauge bosons for physical interactions, electrons for chemical interactions, financial assets for economical interactions, bottles of wine for positive social interactions, and bullets for aggressive ones, etc.). Interactions can be deterministic or stochastic.
4. Elements are characterized by states. States can be scalar; if an element has various independent states, it will be described by a state vector or a state tensor. States are not static but evolve with time. We denote the state vectors by,

   $$\sigma_i(t) \quad \text{state vector.}$$

   States can be the velocity of a planet, spin of an atom, state of phosphorylation of proteins, capitalization of a bank, or the political preference of a person. State changes can be deterministic or stochastic. They can be the result of an endogenous dynamics or of external driving.
5. Complex systems are characterized by the fact that states and interactions are often not independent but evolve together by mutually influencing each other; states and interactions *co-evolve*. The way in which states and interactions are coupled can be deterministic or stochastic.

6. The dynamics of co-evolving multilayer networks is usually highly non-linear.
7. Complex systems are context-dependent. Multilayer networks provide that context and thus offer the possibility of a self-consistent description of complex systems. To be more precise, for any dynamic process happening on a given network layer, the other layers represent the 'context' in the sense that they provide the only other ways in which elements in the initial layer can be influenced. Multilayer networks sometimes allow complex systems to be interpreted as 'closed systems'. Of course, they can be externally driven and usually are dissipative and non-Hamiltonian. In that sense, complex systems are hard to describe analytically.
8. Complex systems are *algorithmic*. Their algorithmic nature is a direct consequence of the discrete interactions between interaction networks and states.
9. Complex systems are path-dependent and consequently often non-ergodic. Given that the network dynamics is sufficiently slow, the networks in the various layers can be seen as a 'memory' that stores and records the recent dynamical past of the system.
10. Complex systems often have memory. Information about the past can be stored in nodes if they have a memory, or in the network structure of the various layers.

In this book, we assume that a co-evolving multilayer network structure is the fundamental dynamical backbone of complex systems. This assumption not only provides a simple conceptual framework, it also allows us to explain several of the essential properties of complex systems. These include:

- The emergence and origin of power laws.
- Self-organized criticality.
- Collapse and boom phases in evolutionary dynamics.
- The nature of phase transitions.
- The origin of edge of chaos.
- Statistics of path-dependent processes.

A snapshot of a co-evolving multilayer network is shown in Figure 1.4. Nodes are represented by a state vector with two components, colour (light- and dark-grey) and shape (circles and boxes). Nodes interact through three types of interaction (full, broken, and dotted lines). The system is a complex system if states change as a function (deterministic or stochastic) of the interaction network and, simultaneously, interactions (the networks) change as a function of the states. See also Equation 1.1. For example, the multilayer network shown could represent a network of banks on a given day, where the shape of the nodes represents the wealth of the bank (circle indicates rich, box indicates poor) and the colour represents the gender of the CEO of the bank (light-grey is female, dark-grey is male). The links represent financial contracts (assets) between banks; a full line could mean a credit relation, a broken line could represent derivatives trading, and dotted lines could indicate that one bank owns shares in another. The set of links

associated with a bank can be seen as its portfolio. Clearly, the wealth state of a bank will influence the network structure. If a bank is poor, it is not allowed to issue new credits. At the same time, the network structure of assets (the portfolio) has a huge effect on the future wealth of the banks. On the other hand, the gender of the CEO may have very little effect on the interbank network structure, and the networks will certainly not have any effect on the gender of the CEO. While the wealth-asset network is a complex system, the gender-network system is not.

1.5.1 What is co-evolution?

To provide some insights into what we mean by co-evolution, we formulate it in a slightly more formal way. In general, interactions can change the states of the elements. In physics, gravitational interaction changes the momentum of massive objects; electromagnetic interactions lead to spin flips; chemical interactions may change the binding state of proteins; economic interactions change the portfolios of traders; and social interactions (exchanging gifts) may change sympathy levels.

The interaction partners of a node in a network (or multilayer network) can be seen as the local 'environment' of that node. The environment often determines the future state of the node. In complex systems, interactions can change over time. For example, people establish new friendships or economic relations; countries terminate diplomatic relations. The state of nodes determines (fully or in part) the future state of the link, whether it exists in the future or not, and if it exists, the strength and the direction that it will have.

The essence of co-evolution can be encompassed in the statement:

- The state of the network (topology and weights) determines the future states of the nodes.
- The state of the nodes determines the future state of the links of the network.

More formally, co-evolving multiplex networks can be written as,

$$\frac{d}{dt}\sigma_i(t) \sim F\left(M^{\alpha}_{ij}(t), \sigma_j(t)\right)$$

and

$$\frac{d}{dt}M^{\alpha}_{ij}(t) \sim G\left(M^{\beta}_{ij}(t), \sigma_j(t)\right). \tag{1.1}$$

Here, the derivatives mean 'change within the next time step' and should not be confused with real derivatives. The first equation means that the states of element i change as a 'function', F, that depends on the present states of σ_i and the present multilayer network states, $M^{\alpha}_{ij}(t)$. The function F can be deterministic or stochastic and contains all summations over greek indices and all j. The first equation depicts the analytical nature of physics that has characterized the past 300 years. Once one specifies F and the initial

conditions say $\sigma_i(t=0)$, the solution of the equation provides us with the trajectories of the elements of the system. Note that in physics the interaction matrix $M_{ij}^{\alpha}(t)$ represents the four forces. Usually, it only involves a single interaction type α, is static, and only depends on the distance between i and j. Typically, systems that can be described with the first equation alone are not called complex, however complicated they may be.

The second equation specifies how the interactions evolve over time as a function G that depends on the same inputs, states of elements and interaction networks. G can be deterministic or stochastic. Now interactions evolve in time. In physics this is very rarely the case. The combination of both equations makes the system a co-evolving complex system. Co-evolving systems of this type are, in general, no longer analytically solvable.[4] One cannot solve these systems using the rationale of physics because the environment—or the boundary conditions—specified by M change as the system evolves. From a practical point of view, Equations 1.1 are useless until the functions G and F are specified. Much of the science of complex systems is related to identifying and giving meaning to these functions for a concrete system at hand. This can be done in an analytical or algorithmic way, meaning that F and G can be given by analytical expression or algorithmic 'update rules'. Both can be deterministic or stochastic.

More and more data sets containing full information about a system are becoming available, meaning that all state changes and all interactions between the elements are recorded. It is becoming technically and computationally possible to monitor the cell-phone communication networks on a national scale, to monitor all genetic molecular activities within a cell, or all legal financial transactions on the planet. Data that contain time-resolved information on states and interactions can be used to actually visualize Equations 1.1; all the necessary components are listed in the data at any point in time: the interaction networks M_{ij}^{α}, the states of the elements σ_i, and all the changes $\frac{d}{dt}\sigma_i$ and $\frac{d}{dt}M_{ij}^{\alpha}$ over time. Even though Equations 1.1 might not be analytically solvable, it is becoming possible for more and more situations to 'watch' them. It is often possible to formulate agent-based models of complex systems in the exact form of Equations 1.1, and this allows us to make quantitative and testable predictions.

The structure of Equations 1.1 is, of course, not the most general possible. Immediate generalizations would be to endow the multilayer networks with a second greek index, $M_{ij}^{\alpha\beta}$, which would allow us to capture cross-layer interactions between elements. It is conceivable that elements and interactions are embedded in space and time; indices labelling the elements and interactions could carry such additional information, $i(x,t,\ldots)$ or $\{ij\}^{\alpha\beta}(x,t,\ldots)$. Finally, one could introduce memory to the elements and interactions of the system.

1.5.2 The role of the computer

The science of complex systems is unthinkable without computers. The analytical tools available until the 1980s were good enough to address problems based on differential

[4] Except for simple examples or situations, where the timescale of the dynamics of the states is clearly different from the dynamics of the interaction networks.

equations, for systems in equilibrium, for linear (or sufficiently linearizable) systems, and for stochastic systems with weak interactions. The problems associated with evolutionary processes, non-ergodicity, out-of-equilibrium systems, self-organization, path-dependence, and so on, were practically beyond scientific reach, mainly because of computational limits. The computational power to address these issues has only become available in the past decades.

Often, computer simulations are the only way of studying and developing insights into the dynamics of complex systems. Simulations are often referred to by agent-based models, where elements and their interactions are modelled and simulated dynamically. Agent-based models allow us to study the collective outcomes of systems comprising elements with specific properties and interactions. The algorithmic description of systems in terms of update rules for states and interactions is fully compatible with the way computer simulations are done. In many real-world situations there is only a single history and this cannot be repeated. This occurs in social systems, in the economy, or in biological evolution. Computer simulations allow us to create artificial histories that are statistically equivalent copies. Ensembles of histories can be created that help us understand the systemic properties of the systems that lead to them. Without simulations, predictive statements about systemic properties like robustness, resilience, efficiency, likelihood of collapse, and so on, would never be possible. The possibility of creating artificial histories solves the problem of repeatability.

With the current availability of computer power and high-dimensional and time-resolved data, computational limits or data issues no longer pose fundamental bottlenecks for understanding complex systems. The bottleneck for progress is the theory of complex systems, a mathematically consistent framework in which data can be systematically transformed into useful knowledge.

The computer has fundamentally changed the evolution of science. It has finally opened and paved the way to understanding complex adaptive systems as a natural science on a quantitative, predictive, and testable basis.

1.6 The structure of the book

Complex systems span an immense universe of phenomena, systems, and processes. We believe that most of these can be mapped, in one way or another, into the framework of the stochastic, co-evolving, multilayer networks that we sketched in Equations 1.1. To be able to treat those systems in quantitative and predictive ways, we need tools and concepts for random processes, networks, and evolutionary dynamics. These needs define the structure of the book.

Complex systems involve many different sources of randomness in their components, interactions, processes, time series, structures, and so on. In Chapter 2 we review the basic notions of randomness and statistics that will be needed in various parts of the

book. In Chapter 3 we discuss the notion of scaling and learn why it is helpful. We review the classic routes to understanding the origin of power laws, which are omnipresent in the statistical description of complex systems. One of the central backbones of complex systems are dynamical networks, which tell us how the building blocks interact with each other. Dynamics can happen on networks (diffusion on networks), or the networks dynamically rearrange themselves. The notions and basic concepts of networks, their structures, characteristics, functions, and ultimately their dynamics will be developed in Chapter 4.

Evolutionary dynamics is central to many complex adaptive systems. After reviewing classic ways of understanding the dynamics of evolutionary systems, in Chapter 5 we show how a very general model of evolutionary systems can be related to co-evolving network structures. We will learn that this approach is an example of an algorithmic description of systems. We will further see that evolutionary systems have familiar phase diagrams and can be naturally associated to self-organized critical systems.

Finally, in Chapter 6 we probe how far methods from statistical mechanics, information theory, and statistical inference methods can be used in the context of stochastic complex systems. These systems are typically path-dependent, evolutionary, non-Markovian, and non-ergodic; thus, the methods that we have learned in physics, statistics, and information theory should be inapplicable. The chapter is designed to show that a very careful generalization of the classic concepts of entropy, information production, and statistical inference methods also allows us to use these concepts for (simple) complex systems and for path-dependent processes in particular.

All chapters start in a relatively simple fashion and become more difficult towards the end. We conclude most chapters with extensive examples that should provide an impression of the status of actual research.

1.6.1 What has complexity science contributed to the history of science?

We conclude this chapter with a somewhat incomplete reading list containing several classics and a few more recent contributions from the science of complex systems. Some of them have already changed our world view.

- Network theory [290]
- Genetic regulatory networks [220, 221]
- Boolean networks [224]
- Self-organized criticality [24, 225]
- Genetic algorithms [194]
- Auto-catalytic networks [22, 113, 204, 205]
- Econophysics [66, 124, 260]
- Theory of increasing returns [13]

- Origin and statistics of power laws [93, 137, 289, 345]
- Mosaic vaccines [31]
- Statistical mechanics of complex systems [175, 178, 281, 377]
- Network models in epidemiology [88, 300]
- Complexity economics [94, 189, 247]
- Systemic risk in financial markets [35, 310]
- Allometric scaling in biology [71, 406]
- Science of cities [36, 49]

In this book, we want to take a few steps towards new directions. In particular, we want to clarify the origin of power laws, especially in the context of driven non-equilibrium systems. We aim at deriving a framework for the statistics of driven systems. We try to categorize probabilistic complex systems into equivalence classes that characterize their statistical properties. We present a generalization of statistical mechanics and information theory, so that they finally become useful for complex systems. In particular, we derive an entropy for complex systems and carefully discuss its meaning. Finally, we make an attempt to unify the many classical approaches to evolution and co-evolution into a single mathematical, algorithmic framework. The overarching theme of the book is to contribute to methodology for understanding co-evolutionary dynamics of states and interactions.

2

Probability and Random Processes

2.1 Overview

Phenomena, systems, and processes are rarely deterministic, but contain stochastic, probabilistic, or random components. Even if nature were to follow purely deterministic laws, which at a quantum level it does not, our scientific notion of the world would still have to remain probabilistic to a large extent. We do not have the means to observe and record deterministic trajectories with the infinite precision required to deterministically describe a world full of non-linear interactions.[1] To some extent we can increase experimental precision, storage capacity, and computing power. Indeed, what we can do today would have been unthinkable only a few decades ago. However, the kind of precision needed to conduct science of non-linear systems at a deterministic level is not only temporarily out-of-reach, but is actually outside our world, especially when biological, social, and economical phenomena are being considered. For that fundamental reason, a probabilistic description of most phenomena in this world is necessary. We will use the terms stochastic, probabilistic, and random interchangeably.

Systems can be intrinsically probabilistic, in which case they typically contain one or more 'sources of randomness'. If systems are deterministic, but we lack detailed information about their states and trajectories, randomness is introduced as a consequence of the coarse-grained description. In both situations, if we want to scientifically understand these systems, it is necessary to understand and quantify the randomness involved. Probability theory provides us with the tools for this task. Here, we will not present a proper course in probability theory, which is far beyond the scope of this book, but we will provide a crash course on the most important notions of probability and random processes.[2] In particular, we will try to attach a precise meaning to words like odds, probability, expectation, variance, and so on. We will begin by describing the most elementary stochastic event—the trial—and develop the notion of urn models, which will

[1] Non-linear or chaotic deterministic systems can often effectively be seen as random processes. Uncertainties about their initial conditions grow exponentially as the system evolves. After some time even small uncertainties increase to a point where it becomes impossible to predict any future states or trajectories based on the underlying deterministic equations.

[2] Readers familiar with probability and random processes may skip the chapter.

Introduction to the Theory of Complex Systems. Stefan Thurner, Rudolf Hanel, and Peter Klimek,
Oxford University Press (2018). © Stefan Thurner, Rudolf Hanel, and Peter Klimek.
DOI: 10.1093/oso/9780198821939.001.0001

play an important role throughout the book. We will discuss basic facts about random variables and the elementary operations that can be performed. Such operations include sums and products and how to compare random variables with each other. We will then learn how to compose simple stochastic processes from elementary stochastic events, such as *independent identical distributed* (i.i.d.) trials and how to use *multinomial statistics*. We will discuss random processes as temporal sequences of trials. If the variables in these sequences are identical and independent, we call them Bernoulli processes. If trials depend only on the outcome of the previously observed variable, they are Markov processes, and if the evolution of the processes depends on their history, we call them path-dependent. We will discuss processes that evolve according to update rules that include random components.

We will discuss how the frequency of observed events can be used to estimate probabilities of their occurrence and how frequentist statistics is built upon this notion. We will touch upon the basic logic of Bayesian reasoning, which is important for testing the likelihood of different hypotheses when we are confronted with specific data. We present a number of classical distribution functions, including power laws and other *fat-* or *heavy*-tailed distributions. These are important for situations where *extreme events* occur frequently. A recurring theme in the context of complex systems is that frequent and large outliers are the rule rather than the exception. Understanding the origin of power law distributions, or other fat-tailed distributions is essential for the realistic assessment of risks, systemic risk in particular, which—despite its importance for society—is still widely based on Gaussian statistics. Gaussian statistics drastically underestimates the frequency of extreme events.

In probability theory, we are dealing with 'known unknowns'. This means that we know the stochastic system and what types of event it is possible for us to observe. What we do not know are the *exact* future outcomes (the unknown), which we cannot predict because of the stochastic nature of the system. However, we do know the *range of possible outcomes* (the known). Take the throwing of a dice for instance. We do not know the outcome of the next throw (trial), but we do know all the possible outcomes, namely, $1, 2, \cdots, 6$, and we might know the associated probabilities. The range of possible outcomes is called the *sample space* of the considered random variable. Statistics and probability theory provide us with a set of tools that allows us to deal with these known unknowns. Many complex systems, however, are systems that involve 'unknown unknowns'. This happens, for example, if it is not possible a priori to specify all the possible outcomes that one might observe in a process. This situation is present in the context of evolution, co-evolution, and innovation, where processes 'invent' new possibilities. It is usually impossible to prestate the future sample space of open-ended evolutionary processes. However, for some driven dissipative systems, for which the dynamics of the sample space is known, there exist fascinating new possibilities to understand the statistics of driven systems on the basis of the dynamics of the sample space.

In this chapter, we will not deal with unknown unknowns. This will be partly covered in Chapter 5. We merely touch upon this tricky subject with a few remarks on so-called urn processes and dynamic sample spaces in Section 2.1.1.5. Here, we mainly focus on situations with static and simple examples of dynamically evolving sample spaces.

The latter will occur in the context of reinforcement processes and driven dissipative systems. Systems of this nature will be discussed in more detail in Chapters 3 and 6.

2.1.1 Basic concepts and notions

2.1.1.1 Trials, odds, chances, and probability

In everyday language the notions of *odds*, *chances*, and *probability* are often used synonymously. In probability theory they carry a distinct meaning, which we should be aware of. We start by defining the most elementary process, the *trial*.

Any process that selects or chooses a particular event from a set of available and possible events is called a trial. The set of all possible events is called the *sample space* of that trial.

In a probabilistic trial, the choice is made randomly. As a consequence, its outcome cannot be predicted with certainty from previously observed trials. For example, in a dice game such as Yahtzee, each throw of the dice is a trial that randomly selects one of six possible outcomes. What is needed next is a possibility of characterizing the *quality* of possible outcomes. For that reason we introduce the notion of *favourable* and *unfavourable* events, and for this we need the concept of a bet.

A *bet* divides all *possible events* into *favourable* and *unfavourable events*. We say a bet is won if a favourable event is selected by a trial. If an unfavourable event occurs, the bet is lost.

The words *chances* and *odds* characterize two different but related ways of characterizing the possible outcomes of a bet. They assign a numerical value to a bet, which allows us to express a 'degree of certainty' about the outcome of a trial.

Assume that a trial can select from W distinct but otherwise equivalent possible events and that a bet divides those events into n_f favourable, and $n_u = W - n_f$ unfavourable events. Then we define:

- the *odds* of winning the bet are $o_{\text{for}} = n_f : n_u$
- the *odds* against winning the bet are $o_{\text{against}} = n_u : n_f$
- the *chances* of winning the bet are $q_{\text{for}} = n_f : W$
- the *chances* against winning the bet are $q_{\text{against}} = n_u : W$

The symbol '$x : y$' stands for the ratio of x to y.

For example, the odds of throwing a six with a fair dice with six faces are 1 : 5; the odds against throwing a six are 5 : 1. The chances of throwing a six are 1 : 6, while the

chances of throwing no six are 5 : 6. While chances add up to one, $q_{\text{for}} + q_{\text{against}} = 1$, for odds we have $o_{\text{for}} o_{\text{against}} = 1$. Moreover, the following relations hold,

$$q_{\text{for}} = \frac{o_{\text{for}}}{1 + o_{\text{for}}} \qquad \text{and} \qquad o_{\text{for}} = \frac{q_{\text{for}}}{1 - q_{\text{for}}}. \tag{2.1}$$

Odds have a practical interpretation. They characterize the expected number of times of winning identical independent bets repeatedly in a row (see Exercise 2.1). Sometimes, *chance* is immediately taken for *probability*. However, it is helpful to distinguish the empirical notion of chance from the more abstract notion of a probability measure. The difference is similar to the way one would distinguish the empirical notion of a distance from the abstract notion of a metric. Later, in Section 2.1.1.5, we will use the notion of chance to define *probabilities*.

2.1.1.2 Experiments, random variables, and sample space

To discuss the fundamental concept of a random variable we must first clarify what an experiment is and what a measurement is.

Experiment, measurement, observation. An experiment is a process that retrieves information about another (random) process. It defines which properties will be observed (measured) and how observations are obtained and recorded. An *experiment* associates *measurable* quantities, called *observables*, with a process and assigns a random variable X to each observable. The simplest experiment is a *trial*, which assigns a definite observed value x to the random variable X as the outcome of the trial. By performing a trial we measure (or sample) a random variable. Experiments can consist of many trials on various variables. We use the terms *to measure*, *to observe*, and *to sample* synonymously.

Thus, any property of a process that can be experimentally observed and recorded in trials can become a *variable* X of the process. The experiment determines which of the potentially observable variables of a process will actually be sampled in the experiment. An experiment A may sample and record property X of a process, while a different experiment B may sample and record both X and Y. For example, X could be the face value of a dice that is thrown; the other observable Y of the same dice could be the position at which the dice comes to a rest on the table. We can now specify what we mean by a random variable.

A *random variable* X is a variable whose possible outcomes can be drawn from a range of possible values. The collection of those values is called the *sample space* of the random variable. Random variables X can be sampled in trials with known or unknown odds of observing particular values. The value of the variable is uncertain before measurement and cannot be inferred with certainty from previous

observations of the same system or process. One can only bet on its outcome (compare [354]). Variables of a process that can be predicted with certainty are called *deterministic*.

2.1.1.3 Systems and processes

We will use the notions of *system* and *process* almost synonymously. The term *system* highlights the systemic aspect of a process and its set of observables that can be observed together. The term *process* highlights the temporal aspect of a system and how a system and its variables evolve over time in successive observations. Let us think of random variables that are independent in the same sense that the outcome of one throw of a dice does not influence the outcome of a later trial. It then does not really matter if we think of N dice as a system of N dice that are all thrown at once in a single trial or as a process where we throw one dice N times in a row in N successive trials. Here the ideas of system and process are indistinguishable,[3] for instance, if an experiment consists of a single measurement or observation. For complex processes with interdependent variables, systemic and temporal aspects need to be very carefully distinguished. In any case, it is useful to characterize what we mean by a random process.

A *random process* is a process that can be studied in an experiment and involves the repeated observation of random variables. Mathematically, we can think of a process as a map, $X : n \to X(n)$, that associates a time point n (of a discrete or continuous time line) with a random variable $X(n)$ that is sampled at time n. In simple processes the random variable may consist of a single value, but it can also be multivalued; in other words, a vector of properties is sampled simultaneously. Famous random processes are Bernoulli processes, Poisson processes, and Brownian or Wiener processes. Random processes can be arbitrarily complicated in terms of the number of stochastic observables and their interrelations, such as path-dependent processes, where the chances of observing the next event depend on the history of the process.

2.1.1.4 Urns

Probability theory is a collection of tools that allows us to understand random processes based on a few elementary processes. Complicated processes are often composed of, or can be segmented into, simple elementary processes. For example, elementary processes include the tossing of coins or the throwing of dice. Another useful class of elementary random processes are so-called *urn models*.

Urns are containers filled with balls (or any other type of object) that are indistinguishable except for one property, which may be the colour or a number written on the

[3] That is why for such simple processes, so-called ensemble averages and time averages of observables yield identical results.

ball. Balls are then drawn blindly from that urn; thus a specific colour can only be drawn by chance. When we draw a ball we call that an *urn trial*. Urns allow us to set up and prepare the chances of observing specific outcomes in an experiment. A fair coin will always have the chances of 1 : 2 possibilities. It would be difficult to design a biased coin that gives 'heads' in 7 out of 61 possibilities. All rational chances, such as 7 : 61, can easily be realized using an urn. For example, place 7 white and $54 = 61 - 7$ black balls into an urn. You have prepared the urn in such a way that the chances of drawing white are 7 : 61 and the chances of drawing black are 54 : 61. Of course, you have the same chances if you use multiples of those numbers, for instance 21 white balls and 162 black balls; see also Figure 2.1. What is important is not the absolute number of balls but their ratios q, which we will call *probability density* or *probability distribution*; see Section 2.1.1.5. Real valued probabilities can always be thought of as limits to rational valued probabilities.

Once a ball is drawn in an urn trial, we can decide what to do next. The decision can depend on the outcome of previous urn trials. For example, the ball chosen can be placed back into the urn (drawing with replacement) or not (drawing without replacement). This gives us the opportunity to use urns to model complex probabilistic processes that can even involve networks of interdependent urns 'connected' through *update rules*. Update rules specify how the outcome of one trial with one urn determines how the content of another urn needs to be prepared for the next trial. Update rules are often presented as *algorithms* that tell us how the sample space and the chances of drawing elements evolve over time.

Processes, where a random variable X is sampled in independent (memoryless) trials over a discrete sample space does not change over time, are called *Bernoulli processes*. We use the term 'Bernoulli process' for situations in which trials sample from two or more (finitely many) elements.[4] An example of a Bernoulli process is the algorithm 'draw a ball—record its colour—put it back in the urn—repeat', which is called *drawing with replacement*. It keeps the chances of drawing particular colours from the urn unchanged

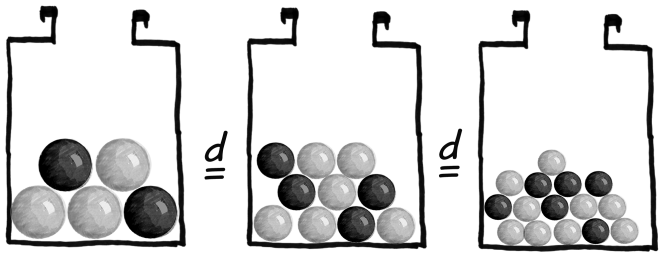

Figure 2.1 *Urn model. The three urns are equivalent in the sense that the chances of drawing a grey ball q_G are 3 : 5. The probability is determined by the ratio given by the chances, $P(G) = 0.6$. Even though the absolute numbers that determine the chances in the three urns are different, the three urns are equivalent in terms of the probability of drawing a black or grey ball.*

[4] In the literature a Bernoulli process is sometimes specifically defined as a two-state process (binary process). Processes with more than two states are called *multivalued* Bernoulli processes or *categorical* processes.

and follows a *multinomial* distribution; see Section 2.5.1.3. Another example of an algorithm is drawing without replacement, which is a simple example of a history-dependent process as, whenever a ball is drawn, the contents of the urn change and consequently, so do the odds for sampling balls in the next trial. This process type follows a *hypergeometric* distribution.

Urns are especially helpful for understanding self-reinforcing complex systems. In this case, we will not just 'draw and replace' a ball but 'draw, replace, and reinforce' by replacing a drawn ball of a certain type (say black) with $\delta > 1$ balls of the same type (black). These urn processes are called Pólya urns, or Pólya urn processes, which show the *rich-get-richer* or the *winner-takes-all* phenomenon. We will discuss these processes in detail in Chapter 6.

It is possible to design arbitrarily complicated algorithms, that is, sets of rules that specify how to prepare and modify the sample space for the next *experiment* based on what has happened in the past. These rules can be deterministic or stochastic. For example, one could think of an algorithm like, 'whenever you draw a red ball, add three balls of a colour not previously used in the urn-process', or 'if you draw green, draw another ball and remove all balls of this colour from the urn'. One may even think of networks of urn processes, where an event sampled in one urn may alter the content of adjacent urns in the network. This flexibility makes urn processes an incredibly useful mathematical vehicle for exploring the properties of many complex path-dependent random processes.

2.1.1.5 Probability and distribution functions

Since Kolmogorov, the mathematical foundations of probability theory have been based on set theory and measure theory. For example, the different colours that can be drawn from an urn can be mapped into an index-set Ω, which is simply the (discrete) sample space of the random variable. An urn containing otherwise identical balls of W different colours will therefore have a sample space of the form $\Omega = \{1, 2, 3, \cdots, W\}$. If we draw colour $i \in \Omega$ in an experiment, we say that we have *observed*, *measured*, or *sampled* i.

The chances of drawing the possible events $i \in \Omega$ in an urn trial depends on the urn's contents before the trial. The urn experiment can be prepared by loading the empty urn with a_i balls for each distinct colour i. We collect those numbers in a vector $a = (a_1, a_2, \cdots, a_W)$. We call a_i the *multiplicity* of possibility (states or events) i. By placing all $|a| = \sum_i a_i$ balls in the urn, the chances of observing i in the next experiment are given by a_i from $|a|$ possibilities. In complex systems, often-severe conceptual and practical problems arise if Ω is not static but evolves in time, or worse, if Ω can not be prestated at all.

We identify the numbers q_i that are defined by $q_i : 1 = a_i : |a|$ with the probability P of observing i, and write,

$$P(\text{to draw } i \text{ from the sample space } \Omega | \text{given } q) = q_i. \tag{2.2}$$

By defining the numbers $q_i = a_i/|a|$, the chances $a_i : |a|$ of observing a particular trial outcome $X = i$, can be identified with the *probability* P of sampling i.

A function $q : i \to q_i$ that maps each possible event $i \in \Omega$ to its *probability* q_i is called the *probability distribution function* of X over Ω. The distribution function is identical to the vector $q = (q_1, \cdots, q_W)$. Note that $|q| = \sum_{i=1}^{W} q_i = 1$ and $q_i \geq 0$ for all i.

The notion of distribution functions is more complicated for continuous sample spaces. For continuous sample spaces such as $\Omega = [0, 1]$, a random variable X over Ω has a probability of $P(y \leq X \leq x)$ of being observed in the interval $[y, x]$. This means that the probability of X of exactly taking the value x is vanishing, $P(x \leq X \leq x) = 0$. However, the limit $\lim_{y \to x} P(y < X \leq x)/|x - y| = \rho(x)$ defines the so-called *probability density function* (pdf) ρ for the random variable X and $P(y \leq X \leq x) = \int_y^x dx \rho(x)$. For further reading see, for instance [318] (p. 25).

Suppose that we perform an urn experiment, where we repeatedly sample with replacement in N subsequent trials. After the experiment we observe that we sampled colour i in k_i out of N trials and $\sum_{i=1}^{W} k_i = N$. We call $k = (k_1, \cdots, k_W)$ the *histogram* of the experiment and $p_i = k_i/N$ the relative frequencies of observing i. Sometimes, $p = (p_1, \cdots, p_W)$ is also called *sample distribution* or *empirical distribution*. Just like probabilities q, the relative frequencies are normalized, $\sum_{i=1}^{W} p_i = 1$. In identical random experiments (with i.i.d. variables) the probabilities q are fixed while the relative frequencies p may vary from experiment to experiment.

A function $k : i \to k_i$ that maps each possible event $i \in \Omega$ to its empirically observed *frequencies* k_i obtained from an experiment sampling a random variable X over Ω for N times, is called the *histogram* of the experiment. The normalized frequencies $p = k/N$ are called the *relative frequency distribution function*.

2.1.2 Probability and information

The notions of *probability* and *information* are tightly related. In the context of probability, information appears in at least four different ways that will be mentioned in various sections of the book. We mention these briefly without any claim of being complete.

Prior information is the knowledge that we have about some observable *prior* to a trial. We can use this information to make a prediction. If we want to predict the outcome of sampling the observable, then the natural choice is to place our bet on the outcome with the highest chances of being sampled. The random variable representing the observable may depend on what we know about the observable before the trial. Consequently, the probabilities also depend on what we know before the trial.[5]

[5] To understand that probabilities can depend on prior knowledge, consider this example: A friend A meets you in a bar. A is accompanied by another person B, that you have never met and know absolutely nothing

Posterior information. Typically, when dealing with a new process, we know little about it; our prior information about the process is limited. We may know the type of process we have before us and thus the parameters that we need to determine in order to fully specify the process. Typically, we do not know the exact value of those parameters. To find out, we gain information about a process after experiments have been performed (a posteriori). This information is then available to us in observation records or histograms. With this empirical data we can infer the parameters of the random process with more certainty than prior to the experiment. Posterior estimates of a priori unknown parameters of a process, based on empirical data, can be made by using *Bayesian inference*, which we will discuss in Section 2.2.5.

Information-theoretic information. Frequently, we wish to communicate important information to friends on our cell phone, such as, 'I'm calling you with my cellphone'. In this context the concept of information becomes exactly quantifiable in terms of the code length that is minimally required to transmit the information through the information channel that consists of cell phones and transmission towers. The information content of messages is typically measured by the number of letters (typically from a binary alphabet consisting of 0 and 1) required to communicate a message. We will deal with this code-based notion of information in more detail in Chapter 6.

Algorithmic information. There is another notion of information based on the minimal length of codes. Here, there is no sender involved wanting to send a message to a receiver; however, there is a computer running computer codes that algorithmically predict the behaviour of a system or a process. The minimum number of bits required to write a code that, when run on a computer, emulates a specific random process, is called the *algorithmic complexity* or the *Kolmogorov–Chaitin–Solomonoff complexity* of that process [77, 235, 343]. The Kolmogorov–Chaitin–Solomonoff complexity is not identical to the information-theoretic information rates. However, the two notions are closely related.[6] In this book, we will not discuss the issues

about. The conversation touches upon the topic of tattoos and A proposes the following wager to you: if you can guess correctly whether or not B has a tattoo, then the next drinks are on A, otherwise you are the one who pays. Suppose B does not have a tattoo. While B knows exactly that she has no tattoo, *you* do not have that knowledge. Thus, the question, tattoo or no tattoo, can be decided with certainty by B (as long as she is not blind and suffers from amnesia) while for you both possibilities exist. The random experiment therefore is not 'tattoo-or-no-tattoo', which would obviously lead to ambiguous results, depending on who you ask. A reasonable random experiment is '*you predicting B having a tattoo or not*' while the process '*B predicting B having a tattoo or not*' is another (deterministic) experiment. If we write $P(B \text{ has a tattoo}|q)$ then the probabilities q stands for the information available to the predictor (the person, machine, or algorithm that predicts). This information conditions the probability. q refers to the chances of the person making the prediction, and the probabilities $P(B \text{ has a tattoo}|q_{\text{you}})$ and $P(B \text{ has a tattoo}|q_B)$ are distinct probabilities and will have distinct values.

[6] In a nutshell, if we know that Kolmogorov complexity of a system (symbol string) A is given by the length of the code of the shortest program B that prints A, then it is plausible that the program B can be interpreted as a string that results from a loss-free compression algorithm applied to string A. Here, the information-theoretic aspect enters into play. A can again be obtained from B by re-expanding the compressed sequence. However, as B is the program printing A, B needs to be a little longer than the compressed version of A itself, as we have to add the code for expansion.

involved in measures of complexity. The topic has been discussed extensively, and there are a number of excellent classical works to which we refer the interested reader [146, 147, 254].

2.1.2.1 Why are random processes important?

Perhaps the most important practical reason for studying probability theory is that the uncertainties associated with individual trials sometimes translate into quite predictable phenomena on an aggregate scale. It may be possible to link elementary stochastic processes (micro description) to aggregate systemic properties of systems composed of such elementary processes (macro description). In particular, the *frequencies* of events observed in populations of trials (distribution functions) often follow regularities or 'laws' that tell us how these frequencies are related to other frequencies or parameters that characterize a process. Sometimes such laws allow us to obtain predictability of stochastic systems at a coarse-grained macroscopic level of description. If we can understand distribution functions at the macro level on the basis of the underlying stochastic dynamics, there is frequently much to be gained. For example, if we understand how to reduce the frequency of occurrence of a particular disease at the population level by randomly vaccinating at the individual level [300]. These laws clarify why we observe specific distribution functions and why certain processes produce them. They allow us to understand how distribution functions change over time and how to decompose processes into deterministic and simple random components. Sometimes even physical laws can be derived from laws that govern random variables. *Statistical physics* is perhaps the most famous example. Here, the *law of large numbers* is a probabilistic law of this kind, governing how velocities of gas molecules are distributed in a container. This then allows us to deduce the relations between pressure, temperature, and volume of gases. Since Boltzmann's times [59, 60] the ideal gas law has become a probabilistic law based on *average* numbers of molecules bouncing into the walls of a container. In that sense, thermometers are physical tools for measuring statistical properties of gas molecules at the macro level. One of today's greatest challenges is to discover and develop the analogue of 'thermometers' for complex non-equilibrium systems, such as ecosystems or societies. In this book, we encounter probability theory in the context of network theory, data analytics and time series analysis, information theory, risk assessment, prediction and forecasts, epidemiology, classification problems, parameter estimation, and hypothesis testing. Before we dive into the heart of the chapter, we summarize the main points so far.

- Probability theory deals with *possibilities* and the *uncertainty* of their realization. One can *bet* on the realization of different possibilities.
- The prospects of winning a bet are characterized in terms of *odds* or *chances* for or against the realization of an event. The notion of *probability* is derived from, and closely related to, the notion of chances.

- Random processes are characterized as independent or interdependent *trials*.
- The collection of all the possible outcomes of a trial is called *sample space*.
- In an experiment a processes is characterized by observable *variables*, which are *deterministic* if their value can be predicted with *certainty*, and *stochastic* otherwise.
- The function that maps the elements of the sample space to their probability of occurrence is called *probability distribution function*.
- *Urn processes* allow us to conceptualize and model random processes by controlling the evolution of the sample space of the urn (contents of the urn). They can be seen as elementary 'programmable' random generators, which can be used to implement algorithms that model complicated random processes.
- Distinct ways of drawing from an urn correspond to distinct random processes that are often associated with distinct distribution functions. For example, *drawing with replacement* and *drawing without replacement* lead to *multinomial* and the *hypergeometric* distribution, respectively.
- Probability theory provides a means of dealing with insufficient information and uncertainty under given conditions. It allows us to predict the occurrence frequencies of (typically independent) events in large samples, large systems, or long time series.
- In the remainder of the book, will use q for denoting chances (probabilities of elementary events), p for relative frequencies, k for histograms, W for the number of states $i = 1, \cdots, W$, and $n = 1, \cdots, N$ and N (or $t = 1, \cdots, T$ and T) for the number of elements (or time steps).

2.2 Probability

So far we have discussed basic notions of probability theory derived from the concept of odds and chances in experiments performed with a random variable X. We write $\Omega(X)$ to denote the sample space of variable X. We have also identified the chances q_i in urn models of drawing colour i with the probability $P(X = i|q) = q_i$. The distribution function q determines the conditions under which the random variable X is drawn. We also have identified the empirical distribution functions $p_i = k_i/N$ as the normalized histogram k_i, that is, the relative frequency of occurrence of i. In the following, we will see how probability can be characterized as a property of sample spaces.

2.2.1 Basic probability measures and the Kolmogorov axioms

Perhaps the most influential formalization of *probabilities* is Kolmogorov's set theoretic approach [236], which has been tremendously successful. Some people have even noted:

'One might come to believe that probability theory began with Kolmogorov' [208]. Kolmogorov's axioms can be formulated as follows:

Kolmogorov axioms. Any function P fulfilling the following three axioms is called a *probability measure* over the sample space Ω.

KA1 (positivity) for any subset $A \subset \Omega$, $P(A)$ is a real number and $P(A) \geq 0$.

KA2 (unitarity) $P(\Omega) = 1$.

KA3 (σ-additivity) for any countable sequence of mutually disjoint subsets $(A_1, A_2, \cdots)$ of Ω, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Axiom KA3 is formulated in this way in order to deal with sample spaces that contain a continuum of infinitely many elements, such as the interval of real numbers between zero and one, $[0, 1]$. However, if sample spaces are discrete and finite $\Omega = \{1, 2, \cdots, W\}$, axiom KA3 simply states that for any subset $A \subset \Omega$,

$$P(A) = \sum_{i \in A} P(i). \tag{2.3}$$

Suppose the chances of drawing different colours $i \in \Omega$ in an urn experiment are given by $q_i = a_i : |a|$. What are the chances of drawing a particular colour i *or* another particular colour j? The chances of drawing i or j are given by $a_i + a_j$ out of $|a|$ possibilities and $P(i \text{ or } j|q) = q_i + q_j$. As a result, the probability of drawing any colour contained in the urn $P(\text{draw any } i \in \Omega|q) = 1$, which exactly fulfills KA2. In other words, we identify certainty with a probability value 1.

We sometimes write $P(X = i)$ instead of $P(i)$ or $P(X \in A)$ instead of $P(A)$, to explicitly indicate that it is the random variable X that assumes the value i, or one of the values in A. Sometimes distinct probability measures P need to be distinguished from distinct random variables $X(n)$, $n = 1, 2, \cdots$, which are considered in the same context. This can be done by writing $P(X(n) \in A)$, or alternatively, with $q(n)$ being the distribution function of $X(n)$, by writing $P(A|q(n))$. Moreover, if we are interested in particular ranges of values of A, we can define those by using statements of the form: expression$(X) =$ true. For example, $P(x_{\min} \leq X \leq x_{\max})$ denotes the probability that we observe X has a value larger than $x_{\min}$ and smaller than $x_{\max}$.

For a random number X, the function,

$$Q(x) = P(X \leq x), \tag{2.4}$$

is called the cumulative distribution function (cdf) of X. The probability distribution function (pdf) $q(x)$ can be recovered using differences $q(i) = Q(i) - Q(i-1)$ (for discrete variables) or derivatives (for continuous variables), $q(x) = \frac{d}{dx} Q(x)$. For continuous variables $q(x)$ is also called the *probability density function*. Discrete variables $q(i)$ are often denoted by q_i and are called the *weights* of i.

2.2.2 Histograms and relative frequencies

While chances (or odds) tell us something about the probability of a random variable X taking a particular value, repeated observations of a random process, $n \to X(n)$, $n = 1, 2, \ldots, N$, provide us with collections of definite values $x(N) = (x_1, x_2, \cdots, x_N)$. These are called the sample. This is one particular realization of the process $X = (X(1), X(2) \cdots, X(N))$. Each random variable $X(n)$ belongs to a sample space $\Omega(X(n))$, and each sampled value x_n is drawn from $\Omega(X(n))$. We call $x(N)$ a *realization*, a *sequence of observations*, or a *sample* of the random process. Probability theory provides us with *statistical* tools to analyse samples $x(N)$. The simplest way of characterizing samples $x(N)$ is in terms of their *histograms* and *relative frequency distributions*.

Given a sequence of observations $x(N) = (x_1, x_2, \cdots, x_N)$ with $x_n \in \Omega = \{1, 2, \cdots, W\}$,

- the *histogram* $k = (k_1, \cdots, k_W)$ counts the number of events $i \in \Omega$ recorded in $x(N)$. $k_i(x)$ denotes the number of times i appears in the sequence $x(N)$;
- the function $p = (p_1, \cdots, p_W)$ that maps each possible event $i \in \Omega$ to the normalized histogram $p_i = k_i/N$ is the *empirical distribution function*. p are also called *relative frequencies*.

Relative frequency distributions are non-negative ($p_i \geq 0$) and normalized $\left(\sum_{i \in \Omega} p_i = 1\right)$. Note that one should strictly distinguish between probabilities $P(i) = q_i$ and relative frequencies p_i, even though both fulfill Kolmogorov's probability axioms for discrete sample spaces. *Probability* is a property of the random variable; *relative frequency* is a property of the experiment. Note that the entire histogram is also a random variable. It is unknown prior to the sampling experiment. Note that questions such as 'how probable is it to find the relative frequencies p after N observations of a random process?' can be answered for several classes of random processes. In other words, one can compute how empirical distribution functions are distributed. We will discuss this in Section (2.5.1.3).

2.2.3 Mean, variance and higher moments

The notions of *mean* (sometimes also called *expectation value*, *average*, or *first moment*) and *variance*, refer either to measures corresponding to a random variable X or to samples $x(N) = (x_1, \cdots, x_N)$. To distinguish the notions, the latter are sometimes called *sample mean* and *sample variance*; see Figure 2.2. We can define expectation values and sample means of functions f that map the sample space Ω into the real numbers.

The expectation value of a function f, $\langle f(X) \rangle$, with respect to random variable X over Ω is given by,

$$\langle f(X) \rangle = \sum_{i \in \Omega} P(X = i) f(i). \tag{2.5}$$

continued

Note that in this context $i \in \Omega$ need not have a numerical value. One sometimes finds the notation $E(f)$ instead of $\langle f \rangle$ to denote the expectation value.[a]

[a] For random variables X over continuous sample spaces Ω, the expectation value is computed as an integral over the probability density distribution ρ of X. The expectation value becomes $\langle X \rangle = \int_{x \in \Omega} dx \rho(x) x$.

Similarly, one defines the sample mean of a function:

The sample mean of a function f is $\langle f(x) \rangle$ with respect to the sample $x = (x_1, \cdots, X_N)$ over Ω is given by,

$$\langle f(x) \rangle = \frac{1}{N} \sum_{n=1}^{N} f(x_n). \tag{2.6}$$

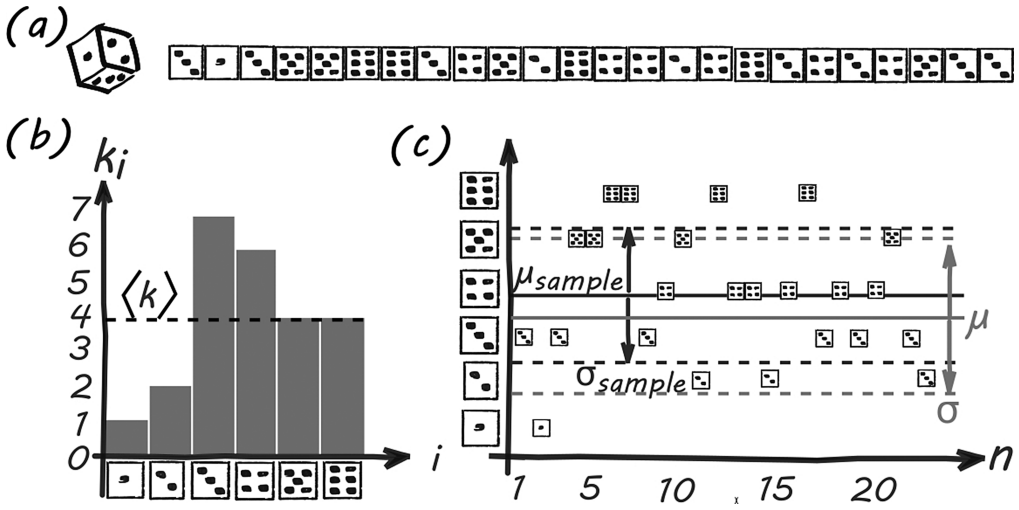

Figure 2.2 *Illustration of some basic notation. A fair dice with $W = 6$ faces has a sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, where the chances of throwing a particular face are 1 : 6. The probability of throwing a face i is therefore $q_i = 1/6$. (a) We throw the dice $N = 50$ times and record the outcomes of those experiments in the sample $x(N) = (x_1, \cdots, x_{50})$. The histogram $k = (k_1, \cdots, k_6)$, that is, the relative frequency distribution $p_i = k_i/N$ of sequence $x(N)$ is shown in (b). (c) The sample mean, μ_{Sample} and sample variance, σ_{Sample} (shown in grey) are different from the expectation value of the random process X that generates the sequences $x(N)$, $\langle X \rangle = \mu = 3.5$ and its variance $\sigma^2(X) = 2.9167$ (black).*

If the elements in sample space Ω are numbers, so-called *moments* can be computed. The expectation value and the variance can then be defined as moments.

Moments are the expectation values of the functions $f(x) = x^m$, $m = 0, 1, 2, \cdots$. The *m'th moment* of a random variable X is defined as

$$\langle X^m \rangle = \sum_{i \in \Omega} P(X = i) i^m. \tag{2.7}$$

- The expectation value $\mu(X) = \langle X \rangle$ is the first moment of X.
- The variance of X is defined by $\sigma^2(X) = \langle (X - \langle X \rangle)^2 \rangle$.

Analogously, for samples the *m'th sample moment* of a sample $x(N)$ is given by,

$$\langle x^m \rangle = \frac{1}{N} \sum_{n=1}^{N} x_n^m. \tag{2.8}$$

- The sample mean $\mu(x) = \langle x \rangle$ is the first sample moment of x.
- The sample variance of x is defined by $\sigma^2(x) = \langle (x - \langle x \rangle)^2 \rangle$.

Sometimes in the literature the sample variance gets 'corrected',[7] by a factor $N/(N-1)$. The variance is a measure that informs us how strongly the random variable X (or the samples $x(N)$) is concentrated around its mean.

There are measures called *cumulants* that are closely related to moments, but might in some cases be more informative than moments. In particular, cumulants allow us to detect a fundamental property of random processes that distinguishes simple (memoryless) from more complex (path-dependent) random processes. One practical use of cumulants is that simply by evaluating the third cumulant of a process (compare Exercise 2.8) one can infer if a process is a Markov process. If it is, the third cumulant vanishes asymptotically. If the third cumulant does not not vanish, the process is path-dependent [143]. Cumulants are the coefficients $\kappa_n(X)$ in the expansion of the so-called cumulant-generating function,

$$K(t) = \log \langle e^{tX} \rangle = \sum_{n=0^\infty} \kappa_n(X) \frac{t^n}{n!}. \tag{2.9}$$

Note that $\kappa_1(X) = \mu(X)$ and $\kappa_2(X) = \sigma^2(X)$.

[7] The corrected sample variance can be obtained as a Bayesian estimate of the true variance of the variable X that generates the sample.

2.2.4 More than one random variable

In path-dependent processes (e.g., Pólya processes), subsequent random variables $X(n)$, $n = 1, 2, \cdots$, will not be identical. How do we find out if random variables are identical? We need a formal way of stating that one random variable is equivalent to another. We also need formal tools to detect if observing one random variable changes the probability distribution of another random variable.

2.2.4.1 Comparing random variables

If we have two real numbers x and y, by writing $x = y$ we mean that x and y have the same value. This notion of equality is too strong for random variables X and Y, as they have no definite value but take values with a certain probability. If we throw two identical dice, then only in one out of six throws will both show the same face value. Nevertheless, we would call the two dice equivalent.

If X and Y are two random variables over the same sample space Ω then

(i) we call X and Y identical (in distribution) if X and Y have identical distribution functions. In this case, we write $X \stackrel{d}{=} Y$;

(ii) we call X and Y of the same *type*, if there are two real numbers a and b such that,

$$X \stackrel{d}{=} aY + b. \tag{2.10}$$

The definition follows [292].

Note that for two random variables to be of the same type, the observable events have to be numbers; they cannot be colours. In that case, the elements of the sample space $x \in \Omega(X)$ can be mapped one-to-one into elements $y \in \Omega(Y)$ through an *affine transformation*, $ax + b \in \Omega(Y)$. In other words, we can multiply and add some constants a and b to the elements of the sample spaces $\Omega(X)$ and $\Omega(Y)$.

2.2.4.2 Joint and conditional probabilities

So far, we have discussed the meaning of a probability $P(X \in A)$ of observing a single random variable X in the range (or set) A. However, a process may incorporate more than one random variable that we can observe at the same time. Suppose that we observe a random variable X_1 over sample space $\Omega(X_1)$ and a variable X_2 over $\Omega(X_2)$, then we can treat (X_1, X_2) as a joint random variable over $\Omega(X_1, X_2) = \Omega(X_1) \times \Omega(X_2)$ where,

$$\Omega(X_1) \times \Omega(X_2) = \{(x_1, x_2) | x_1 \in \Omega(X_1) \text{ and } | \ x_2 \in \Omega(X_2)\}, \tag{2.11}$$

is called the *Cartesian product* of the two sample spaces. The Cartesian product of two sets is simply the set of all possible ordered pairs (x_1, x_2) that we can generate from the two sets. We can now define the joint probability.

We call a measure $P(X_1 \in A_1, X_2 \in A_2) = P((X_1, X_2) \in A_1 \times A_2)$ the *joint probability* of the random variables X_1 and X_2. Joint probabilities have to respect the following consistency conditions,

$$\begin{aligned} P(X_1 \in A_1) &= P(X_1 \in A_1, X_2 \in \Omega_2) \\ P(X_2 \in A_2) &= P(X_1 \in \Omega_1, X_2 \in A_2). \end{aligned} \tag{2.12}$$

These conditions state that if one of the two variables can take any possible value, then the joint probability reduces to a probability that measures the chances of finding the other variable within the specified range. These probability measures $P(X_1 \in A_1)$ and $P(X_2 \in A_2)$ are called *marginal probabilities*. The process of obtaining the marginal probabilities is called *marginalization*.

If two random variables X_1 and X_2 can be observed such that the chances of observing one variable do not depend on the realization of the other variable, then the two random variables are called *statistically independent*.

Two random variables are called *statistically independent* if,

$$P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1) P(X_2 \in A_2), \tag{2.13}$$

for all $A_1 \subset \Omega_1$ and $A_2 \subset \Omega_2$. Here we use the symbol $\subset$ for 'subset of'.

An illustration of the concept of statistical independence is shown in Figures 2.3 and 2.4. Imagine an urn containing two black (B) and three grey (G) balls. The chances of drawing B are two out of five and of drawing G are three out of five. If we draw with replacement, that is, we draw a ball, record the colour, and then put it back into the urn, our experiment will not change the chances of drawing B or G in the next experiment. For processes with replacement, subsequent experiments, $X(n)$ and $X(n+1)$, $n = 1, 2, 3, \cdots$, are identical in distribution and statistically independent. However, if we do not replace the (say, B) ball after drawing it, there are only one B and three G balls left in the urn for the next trial. The chances of drawing B are now one out of four. In drawing without replacement, subsequent variables, $X(n)$ and $X(n+1)$, are no longer identical in distribution and no longer statistically independent. The distribution of the second variable clearly depends on the outcome of the first variable. The same is true for Pólya urn processes, where additional balls of the same colour are added after every draw. Again, subsequent experiments are not independent.

If $X(n)$, $n = 1, 2, \cdots, N$ are statistically independent variables and identical in distribution those random variables are called *independent identically distributed*. In short, $X(n)$ are i.i.d. variables.

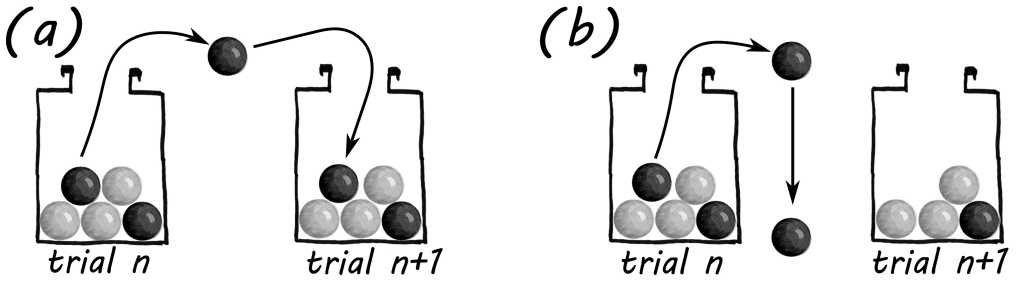

Figure 2.3 *Illustration of statistical independence and dependence in urn processes: (a)* with replacement *and (b)* without replacement.

We are now in a position to ask how probable it is to observe $X_1 \in A_1$, given that we observed $X_2 \in A_2$. This results in the definition of the conditional probability.

The probability of observing $X_1 \in A_1$, given that we know that $X_2 \in A_2$, is called the *conditional probability*. It is defined as,

$$P(X_1 \in A_1 | X_2 \in A_2) = \frac{P(X_1 \in A_1, X_2 \in A_2)}{P(X_2 \in A_2)}. \tag{2.14}$$

As a consequence, if X_1 and X_2 are statistically independent random variables, the conditional probability reduces to the marginal probability, $P(X_1 \in A_1) = P(X_1 \in A_1 | X_2 \in A_2)$. This is generally true for Bernoulli processes, which we will discuss in Section 2.5.1.1.

If X_1 and X_2 are two independent random variables (imagine two distinct urns), what is the probability of jointly drawing $i \in \Omega(X_1)$ and $j \in \Omega(X_2)$? If the respective chances are $q_i = a_i : |a|$ and $q'_j = a'_j : |a'|$, the chances of observing the pair (i,j) are given by $a_i a'_j$ out of $|a||a'|$ possibilities. Thus, $P(X_1 = i \text{ and } X_2 = j | q \times q') = q_i q'_j$, or $P(X_1 = i \text{ and } X_2 = j | q \times q') = P(X_1 = i|q) P(X_2 = j|q')$.

Why does 'X_1 *and* X_2' mean to *multiply* two probabilities? Suppose we can draw from two events 0 and 1 in two identical independent random experiments with chances q_0 and q_1? Suppose $q_0 = 2/5$ and $q_1 = 3/5$. We can represent these chances by putting five balls into an urn, two black and three grey; see Figure 2.4. Let us replace the balls after each trial, so that the chances of picking balls remain the same. To pick black twice in two trials, we have two out of five possibilities for the first ball *and* then two out of five possibilities again. This means for $N = 2$ trials we have to choose from $4 = 2 \times 2$ out of $25 = 5 \times 5$ possibilities to pick a black ball twice, that is, $P(\text{black and black}) = q_0^2$. Similarly, we can construct all other probabilities to draw any combination of black and grey balls.

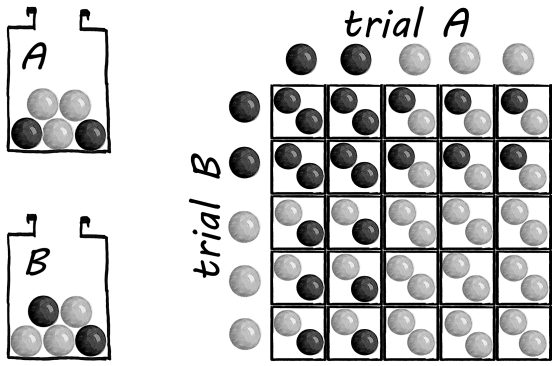

Figure 2.4 *Urn experiment with replacement. The reason why the probability of drawing* i and j *independently is given by the product of* $P(i \text{ and } j) = P(i)P(j)$ *can be visualized by drawing a table of all possible outcomes of the two independent experiments. To draw two black balls there are four out of twenty-five possibilities; for drawing two grey balls there are nine possibilities. For drawing one black and one grey ball there are* $12 = 6 + 6$ *possibilities.*

The notion of *statistical independence* in Equation (2.13) is built on the product property of joint observations. The product captures the insight that two random experiments are statistically independent if the outcome of one experiment does not influence the chances for the outcomes of another variable, and vice versa.

2.2.5 A note on Bayesian reasoning

Bayesian reasoning provides us with a simple and effective tool that can be used to estimate the relative likelihood of two or more alternative hypotheses, or for estimating uncertain parameters that condition random processes. Computations in Bayesian arguments quickly turn out to be lengthy; however, the basic idea underlying all Bayesian reasoning is easily explained.

Consider two random variables X and Y with their respective sample spaces $\Omega(X)$ and $\Omega(Y)$ with elements $i \in \Omega(X)$ and $f \in \Omega(Y)$. Given their joint probabilities $P(X = i, Y = f)$ and the marginal probabilities $P(X = i)$ and $P(Y = f)$, defined in Equation (2.14), the conditional probabilities are given by $P(X = i|Y = f) = P(X = i, Y = f)/P(Y = f)$, and $P(Y = f|X = i) = P(X = i, Y = f)/P(X = i)$.

From this observation Bayes' rule follows,

$$P(X = i|Y = f) = P(Y = f|X = i)\frac{P(X = i)}{P(Y = f)}. \tag{2.15}$$

It allows us to flip the arguments in the conditional probabilities.

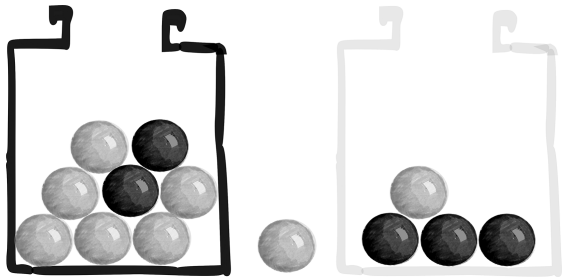

Figure 2.5 *Illustration of a typical Bayesian problem. There are two urns, black and white, each filled with black and grey balls. The chances for drawing each type of ball from both urns are known. Imagine that a blindfolded person draws a ball from one of the urns but does not know from which. The ball drawn is grey. What is the probability that the urn it was drawn from was the black one? In other words, we are testing the hypothesis that the 'urn was black' given the sample 'the ball is grey'. The prior information is that the urn was either black or white.*

What must appear as a mathematical triviality becomes a powerful tool for probabilistic reasoning if we read Bayes' rule in the following way.

If we know the process (or a model of it) that maps uncertain initial conditions $i = X$ (such as the parameters of the model) to the final observations f (data produced by the process), that is, if we know $P(Y=f|X=i)$, then we can infer the probability distribution of an unknown initial condition $X = i$, conditional on observed data $Y = f$. In other words, it allows us to ask questions like: given specific data what are the best parameters of a model to explain it? Or what is the best model (if we have more than one) that describes the data? Bayes' principle further allows us to deal with cause and effect or other forms of statistical interdependencies of variables simultaneously. If new data become available, Bayes' rule provides a scheme to iteratively update our level of certainty by mapping prior probabilities (before new data have been considered) to posterior distribution functions (after new data have been considered).

The following example is a typical application of Bayesian inference. Assume that you have two urns, one black, the other white; both are filled with black and grey balls. You know the chances of drawing black (B) and grey (G) balls for both urns, $P(\text{ball} = \text{G}|\text{urn} = \text{black})$ and $P(\text{ball} = \text{G}|\text{urn} = \text{white})$. Imagine you close your eyes, and somebody hands you one of the two urns to draw from. You do not see which urn it is. After you open your eyes, you see that you have drawn a grey ball; see Figure 2.5. The question now is: what is the probability that the urn you drew from was black? In other words, we want to know the value of $P(\text{urn} = \text{black}|\text{ball} = G)$. The answer to that question is given by Bayes' rule in Equation (2.15),

$$\begin{aligned} P(\text{urn} = \text{black}|\text{ball} = G) &= \frac{P(\text{ball}=G|\text{urn}=\text{black})P(\text{urn}=\text{black})}{P(\text{ball}=G)} \\ &= \frac{P(\text{ball}=G|\text{urn}=\text{black})P(\text{urn}=\text{black})}{P(\text{ball}=G|\text{urn}=\text{black})P(\text{urn}=\text{black})+P(\text{ball}=G|\text{urn}=\text{white})P(\text{urn}=\text{white})} \\ &= \left(1+\frac{P(\text{ball}=G|\text{urn}=\text{white})}{P(\text{ball}=G|\text{urn}=\text{black})}\,\frac{P(\text{urn}=\text{white})}{P(\text{urn}=\text{black})}\right)^{-1}. \end{aligned} \tag{2.16}$$

If we know that the conditional probabilities have the values $P(\text{ball} = G|\text{urn} = \text{black}) = \frac{6}{8}$ and $P(\text{ball} = G|\text{urn} = \text{white}) = \frac{1}{4}$, then we still face the problem of not knowing the value of the prior distributions $P(\text{urn} = \text{black})$. With no additional information available, our best guess is to assume a 50 : 50 chance; $P(\text{urn} = \text{black}) = P(\text{urn} = \text{white}) = \frac{1}{2}$. This is the so-called maximum ignorance assumption for the prior distributions. From this, it follows that $P(\text{urn} = \text{black}|\text{ball} = G) = 1/(1 + (1/4)/(6/8)) = 1/(1 + (1/3)) = 3/4$; the probability we have drawn from the black urn is 0.75; see Exercise 2.2.

Questions of this type arise in countless situations. For example, in medicine, you can rephrase this example by identifying the sampled black ball with a patient with tumor X, of subtype Y, and genome Z. If you are treating this patient, you can ask for the probability that the box was white, which in the medical example could mean the probability that medication B works for that patient. In other words: to which class (box) of patients does this patient belong? To the class where medication B works or to the other one where it fails?

Bayesian statistics is perfectly applicable to these types of question, and yields testable predictions. In other words, Bayesian inference is made for problems where you are given a sample $x(N) = (x_1, \cdots, x_N)$ that records the outcome of N identical and independent experiments. However, in the experiment we do not know the distribution function q_i of drawing particular values, $i \in \Omega$. Using Bayes' rule one can estimate the distribution function q (for instance, by predicting the most likely q_i) from data x and perhaps additional available information I.

Let us consider another example. Suppose you are drawing from an urn with replacement. You are given the information that the urn only contains $W = 2$ colours, grey (G) and black (B). You also know that there are only $M = 5$ balls in the urn. This information might be collected in the vector $I = (W, M)$. However, you do not know how many balls are B and how many are G. Let n_G be the number of G and $n_B = N - n_G$ the number of B balls. As you know there are two colours, $n_B = N - n_G$ has to be one of the numbers $n_B \in \{1, 2, 3, 4\}$. Suppose you have drawn $N = 20$ times from the urn and obtained the sample $x(N) = (x_1, \cdots, x_{20})$, which contains $k_G = 13$ times L and $k_B = 7$ times B. What is the sample $x(N)$ telling us about the probability of finding n_B and n_G balls in the urn? In other words, we would like to know the probability[8] $P(q|x, I)$ of observing the distribution function q given the samples $x(N)$ and the information I. One may also interpret this as an example of *Bayesian hypothesis testing*. n_B determines q uniquely given the information I, and there are four possible hypotheses, $H_b = \{n_L = b\}$, $b = 1, \cdots, 4$, to choose from. Moreover $q_B(b) = b/5$ and $q_G(b) = 1 - b/5$. We therefore want to know $P(H_b|x, I)$, the probability of the hypotheses H_b being true. We can use Bayes' rule and the definition of conditional probabilities 2.14 to obtain,

$$\begin{aligned} P(H_b|x, I) &= P(H_b, x, I)/P(x, I) \\ &= P(x|H_b, I)P(H_b, I)/P(x, I) \\ &= P(x|H_b, I)P(H_b|I)/P(x|I). \end{aligned} \tag{2.17}$$

[8] We now drop the (N) in the notation of $x(N)$ for a moment.

Estimating $P(H_b|x,I)$ now becomes a matter of estimating $P(x|q(b),I) = P(x|H_b,I)$, $P(H_b|I)$ and $P(x|I)$. Be careful with the notation of conditional (|) and joint (,) distribution.

1. *The prior probabilities*: Given the information I, the best (maximal ignorance) estimate of $P(H_b|I) = 1/4$, for any of the four possible values $b \in \{1,2,3,4\}$. Note that this choice depends on the fact that we have no previous information about the urn, as this is our first sequence of N experiments.[9] $P(H_b|I)$ are called the *prior probabilities* of H_b, as they represent the probability of the predictor guessing H_b correctly before the experiment.
2. *The random process*: $P(x|q,I)$ is the probability of drawing a particular sample x containing k_B times B and k_G times G in $N = k_B + k_G$ independent trials. This probability is known to us from basic probabilistic and combinatorial facts,

$$P(x|q,I) = q_B^{k_B} q_G^{k_G}. \tag{2.18}$$

3. *Normalization*: We next compute $P(x|I)$. This can be done by marginalizing $P(x,H_b|I)$ with respect to H_b. This means summing $P(x,H_b|I)$ over all hypotheses b. Using the properties of conditional probabilities, we find $P(x,H_b|I) = P(x|H_b,I)P(H_b|I)$. Using $P(H_b|I) = 1/4$, it follows that,

$$P(x|I) = \sum_{b=1}^{4} P(x,H_b|I) = \sum_{b=1}^{4} P(x|H_b,I)P(H_b|I) = \frac{1}{4}\sum_{b=1}^{4}\left(\frac{b}{5}\right)^{k_B}\left(1-\frac{b}{5}\right)^{k_G}. \tag{2.19}$$

4. *Posterior probabilities*: Finally, we insert the results (1–3) into Equation (2.16) and find,

$$P(H_{n_B}|x,I) = \frac{\left(\frac{n_B}{5}\right)^7\left(1-\frac{n_B}{5}\right)^{13}}{\sum_{b=1}^{4}\left(\frac{b}{5}\right)^7\left(1-\frac{b}{5}\right)^{13}}. \tag{2.20}$$

The Bayesian estimates for the probabilities that n_B takes the particular value $1,2,3$, or 4, are given by the probabilities,

$$\begin{aligned} P(H_1|x,I) &= 0.2321 \\ P(H_2|x,I) &= 0.7059 \\ P(H_3|x,I) &= 0.0620 \\ P(H_4|x,I) &= 0.0001, \end{aligned}$$

[9] In a subsequent sequence of experiments, one might use different values for $P(H_b|I)$ drawing upon information obtained from previous experiments. The priors can be updated as one learns more about a system by performing more experiments.

and we can predict the most likely hypothesis H_2 with the probability of 0.7059. We also know that with probability 0.2941 this prediction will be wrong. The probabilities $P(H_b|x,I)$ are called the *posterior probabilities* of H_b, as they describe the chances of the predictor correctly guessing H_b *after* having used the information contained in the sample $x(N)$.

Again, we can fix a significance level (for instance, $p_s = 0.1$) and then perform as many experiments as necessary to accept or reject a hypothesis. In our example, we might say at this point that a probability of 0.2941 of failing in our prediction is too high, as it is larger than the significance level $p_s = 0.1$. We might decide to perform another sequence of $N = 20$ experiments and sample the data $x'(N) = (x'_1, \cdots, x'_{20})$. In this second round of experiments, we can use the information from the first sequence $x(N)$ by identifying the prior probabilities for the new data with the posterior probabilities $P(H_b|I') = P(H_b|x,I)$ that were obtained previously. In other words, in the second experiment we use the information $I' = (x', I)$. Suppose we throw a sequence $x'(20)$ with $k'_B = 9$ and $k'_G = 11$. Following the same reasoning as before, we now get the equation,

$$P(H_{n_B}|x',I') = \frac{\left(\frac{n_B}{9}\right)^5 \left(1 - \frac{n_B}{5}\right)^{11} P(H_{n_B}|I')}{\sum_{b=1}^{4} \left(\frac{b}{5}\right)^9 \left(1 - \frac{b}{5}\right)^{11} P(H_b|I')}, \tag{2.21}$$

which updates the chances for predicting hypothesis H_b to,

$$\begin{aligned} P(H_1|x',I') &= 0.0144 \\ P(H_2|x',I') &= 0.9486 \\ P(H_3|x',I') &= 0.0370 \\ P(H_4|x',I') &= 2.2 \cdot 10^{-7}. \end{aligned}$$

We now predict H_2 with probability 0.9486. This prediction will still fail with a probability of 0.0514. If we are happy with a significance level of $p_s = 0.1$ we stop here and predict H_2 at the significance level of 0.1. If, however, we had chosen a significance level of 0.01, we may need to perform additional experiments to reach this level of confidence. If none of the possible hypotheses reaches the specified confidence level, even after many additional experiments are performed, then we have to reject all possible hypotheses and conclude that the information I was probably incorrect and that the data generation model we used to infer the probabilities was therefore inadequate.

Bayesian reasoning makes use of the fact that the chances of making correct predictions depend on information available prior to an experiment. As experimental data accumulate, posterior probabilities from previous experiments can be used as prior probabilities for new experiments. In Bayesian reasoning, the prior information (and prior distribution functions) is updated as we gain more information about the system.

2.2.6 Bayesian and frequentist thinking

Sometimes a distinction is made between the 'frequentist' and the 'Bayesian' notion of probabilities. These are also referred to as 'objective' and 'subjective' probabilities. This can be very misleading. In the frequentist notion the probabilities of events are considered as 'real' and the data we obtain are just samples. In the Bayesian view the data are 'real' (and not probabilistic) and the probabilities of events (or of the model of the data-generating process) are uncertain. In the frequentist world, one therefore tries to infer probabilities of events from repeatable independent experiments, where with more data, one is assured of estimating the probabilities with increasing accuracy. The frequentist uses the *p-value* to reject a null hypothesis.

In the Bayesian world, one often considers reasonable data-generating models (or a parametric family of models) and then infers that the probability distribution for those models has produced the data. In this task the Bayesian tries to use any available information to infer the data-generation process and its parameters. Bayesians talk not about the probabilities of actual events but about the probabilities of predicting them correctly. If the available information about a particular phenomenon changes, the probabilities of possible models describing the phenomenon can be updated. This allows the Bayesian to think probabilistically about events that may never have occurred before—or cannot be observed in repeatable independent trials—while the strict frequentist has no notion of probabilities of this type. The mathematical theory of probabilities, however, does not depend on whether one feels inclined towards a Bayesian or a frequentist point of view. Bayes' rule Equation (2.15), is a necessary tool in the pocket of every twenty-first-century scientist.

2.2.6.1 Hypothesis testing in the frequentist approach

For completeness we recall the basic philosophy behind testing a statistical hypothesis in the frequentist view; see also Section 4.7. It follows several steps. First, the *null hypothesis* H_0 is formulated, specifying precisely what is being tested. It can be as follows. The null hypothesis H_0 is: the samples from measurement A are from the same distribution as other samples from measurement B. The basic idea is now to try to reject the null hypothesis for a given significance level α. If we can, we have shown that, for the specified significance level α, the samples from measurement A are not from the same distribution as the samples of B. To be able to reject the null hypothesis, we need the *test statistic*, which is a convenient test function $d(x)$ that measures a 'distance' between sample x_A and sample x_B. The idea is that for the test statistic the correct distribution function is known, and a significance level can be specified. If the test statistic is larger than α, the null hypothesis must be rejected. Typically, null hypotheses are set up to be rejected. If they are rejected at a given significance level, this signals that there might be a statistically significant difference between A and B.

We briefly summarize this section on basic notions in probability theory.

- Probabilities can be axiomatically defined as *normalized measures* on the sample space. This is accomplished by the three Kolmogorov axioms of probability theory.

- Probabilities are abstract measures that measure our chances of observing specific events before sampling them. Histograms and relative frequencies (normalized histograms) are *empirical* distribution functions that we observe when we are sampling a process.
- Probability distribution functions can have a mean, a variance, and higher moments. Empirical distribution functions have a sample mean (average) and a sample variance.
- Probability distributions of more than one variable are called *joint probabilities*.
- A *conditional probability* is the ratio between a joint probability and one of its marginal probabilities, the condition.
- Two random variables are *statistically independent* if the distribution function of one variable does not depend on the outcome of the other.
- Frequentist hypothesis testing uses a null hypothesis and a test statistic to test if the null hypothesis for the given data can be rejected at a specified confidence level. If not, the hypothesis is not rejectable at the confidence level.
- Bayesian hypothesis testing allows one to compare the quality of one data-generating model with that of another, regardless of whether the models are correct. New data can be used to update prior and posterior probabilities iteratively. If enough data are available and the data-generating models are reasonable, the process can be iterated until a prespecified significance level is reached.

2.3 The law of large numbers—adding random numbers

How can we do algebra with random numbers? If we add random numbers, we obtain a new random number. What is the distribution function of a sum of random numbers? In this context we will encounter the mathematical notion of limits, which we summarize in the law of large numbers. This law is a statement about the convergence of sums of random numbers, given that we add infinitely many of those. We will then review the convolution product and, with its help, discuss the central limit theorem, which is at the core of classical statistics. This theorem tells us that if we add many i.i.d. random numbers originating from sources with a finite variance, the distribution function will be a Gaussian, (or normal) distribution function. We will discuss the notion of α-stable processes and learn that the Gauss, Lévy, and Cauchy distributions are the only possibilities for α-stable distribution functions.

The concept of *limits* is essential for dealing with stochastic processes. Intuitively, a limit means that elements of a sequence of 'objects' become increasingly 'similar' to a particular 'limit-object', until the elements become virtually indistinguishable from the limit object—the sequence *converges* to its *limit*. The similarity can be defined in several ways, it has to be clarified beforehand which *similarity measure* is to be used.

What is a limit of a sequence of random variables? Assume the random variables $X(n)$, $n = 1, 2, 3 \cdots$, over the sample space $\Omega(X(n)) = \{1, \cdots, W\}$, and a random variable

Y over the same $\Omega(X(n))$. What do we mean by a limit $\lim_{n\to\infty} X(n) \to Y$ of random variables? For two real numbers one can easily visualize what it means to converge.

One possibility for giving meaning to a limit for random numbers is to consider limits of distribution functions. If the probability $q_i(n) = P(X(n) = i)$ and $\bar{q}_i = P(Y = i)$, then the limit $X(n) \to Y$ can be understood in terms of the limit $\lim_{t\to\infty} q_i(n) \to \bar{q}_i$ for all $i \in \Omega$. This is the so-called *point-wise limit* of distribution functions.

The sequence $q_i(n)$, $n = 1, 2, 3, \cdots$, converges point-wise to the limit distribution $\bar{q}_i$, if for any real number $\epsilon > 0$ there is a number n_0, such that for any $n' > n_0$, it is true that $|\bar{q}_i(n) - q_i(n')| < \epsilon$.

Note that the term point-wise convergence tells us nothing about the speed of convergence[10] or whether the speed of convergence is the same for all i. Depending on the context, different notions of convergence are sometimes used, usually for the sake of simplicity of mathematical proofs. For the most part, however, the exact notion of convergence bears little or no relevance for practical questions. One of such practical questions addresses the conditions under which sampling processes converge to an average. The Italian mathematician Gerolamo Cardano noted in the sixteenth century that sample averages improve as more measurements are taken. If we consider the random variables $X(n)$, $n = 1, 2, 3, \cdots, N$, where each of the variables $X(n)$ has an average $\langle X(n)\rangle = \mu$, does its average converge to μ? If we define the random variable,

$$Y(N) = \frac{1}{N}(X(1) + X(2) + \cdots + X(N)), \tag{2.22}$$

and $y_N = (x_1 + x_2 + \cdots + x_N)/N$, where the x_n are samples of the variables $X(n)$, and y_n is a sample of $Y(N)$, is it then reasonable to assume that $\lim_{N\to\infty} y_N = \mu$? The *law of large numbers* provides the answer to this question. There are various versions of this law, the best known ones being the weak and the strong law, both of which we will mention. However, the difference will not matter for our purposes.

The *weak law of large numbers* states that if we have independent and identically distributed (i.i.d.) random variables $X(n)$ with a mean $\langle X(n)\rangle = \mu$, for any number $\epsilon > 0$ (however small) it is true that,

$$\lim_{N\to\infty} P(|y_N - \mu| > \epsilon) = 0. \tag{2.23}$$

A mathematician would say that $Y(t)$ converges to μ *in probability*.

[10] To learn more about the speed of convergence we refer to the Berry–Esseen theorem [48, 120]. For methods that can be used to derive bounds on the speed of convergence, see, e.g., Stein's method [353].

The *strong law of large numbers*[a] states that under the same conditions,

$$P(\lim_{N\to\infty} y_N = \mu) = 1. \tag{2.24}$$

A mathematician would say that $Y(t)$ converges to μ *almost surely*.

[a] The strong law indeed makes the stronger claim. It states that if one performs random sampling experiments many times with variables $Y(n), n = 1, 2, \cdots$, then the fraction of sampled sequences y_n, where $\lim_{n\to\infty} y_n \to \mu$ is not fulfilled vanishes. The weak law only implies that, no matter how small we choose the error ϵ to be, the fraction of sampled sequences that will fail to converge to μ by a margin of ϵ, will go to zero.

The law of large numbers does not tell us anything about the expected error, when estimating the average μ using finite sample averages. The theorems only tell us that if you have taken sufficiently many samples, then it is usually safe to estimate the mean μ of the i.i.d. processes from its sample average. They tell us nothing about what 'sufficiently many' means.

2.3.1 The central limit theorem

The law of large numbers does require the random variables $X(n)$ to be i.i.d., but it does not require a finite variance. If we have both finite mean and finite variance, we can state the single best-known limit theorem, the *central limit theorem*. This is at the heart of Gaussian statistics and explains the ubiquity of the so-called *Gaussian* or *normal distribution function*. Before we turn to the theorem, let us recollect some facts about Gaussian distribution functions and the notion of the *convolution product*.

2.3.1.1 Gaussian distribution function

The *standard normal distribution function* has the form,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right). \tag{2.25}$$

All *normal* distribution functions can be obtained by stretching and dilating the standard normal distribution,

$$p_{\text{normal}}(x|\sigma,\mu) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), \tag{2.26}$$

where μ is the mean and σ the standard deviation of the distribution.

A random variable X over the real numbers, with mean μ and standard deviation σ, is called a *Gaussian random variable*, if the probability of finding $a < X < b$ is given by,

continued

$$P(a < X < b|\sigma,\mu) = \int_a^b dx\, p_{\text{normal}}(x|\mu,\sigma). \tag{2.27}$$

The variable X is then said to be in $\mathcal{N}(\mu,\sigma^2)$.

We can immediately compute the moments of Gaussian random variables from Equation (2.25). As the standard normal distribution is symmetric, all odd moments with $n = 1,3,5,\cdots$, vanish, meaning that $\int_{-\infty}^{\infty}\phi(x)x^n = 0$. The even moments, with $n = 0,2,4,\cdots$, can be computed by taking derivatives repeatedly in the following way,

$$\begin{aligned}\int_{-\infty}^{\infty} dx\,\phi(x)x^{2m} &= \left(-2\,\frac{d}{d\alpha}\bigg|_{\alpha=1}\right)^m \int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi}}\,e^{-\frac{\alpha}{2}x^2} = \left(-2\,\frac{d}{d\alpha}\bigg|_{\alpha=1}\right)^m \int_{-\infty}^{\infty}\frac{dy}{\sqrt{2\pi\alpha}}\,e^{-\frac{1}{2}y^2}\\ &= \left(-2\,\frac{d}{d\alpha}\bigg|_{\alpha=1}\right)^m \alpha^{-\frac{1}{2}} = 1\cdot 3\cdot 5\cdots(2m-1) = \frac{(2m)!}{m!2^m}.\end{aligned} \tag{2.28}$$

Here the expression $(d/d\alpha)^m$ means: apply the derivative m times.

2.3.1.2 The convolution product

The convolution product has many applications. One of them is to describe the distribution of sums of independent random variables.

The *convolution product* of two functions is defined as,

$$f * g(x) = \int_{-\infty}^{\infty} dx' f(x')g(x-x'). \tag{2.29}$$

It has similar algebraic properties to the usual product of real numbers, $f * g = g * f$, $(f * g) * h = f * (g * h)$, and it is linear $f * (g + h) = f * g + f * h$. Here the sum of two functions is defined as $(g + h)(x) = g(x) + h(x)$.

To see how the convolution product yields the sum of two random numbers, assume that X and Y are two independent random variables over the real numbers with distribution functions $f(x)$ and $g(y)$, respectively. What is the distribution function of the random variable $Z = X + Y$? To find the probability that the sum of two random numbers is exactly z, $P(x + y = z)$, we must consider all possible realizations x and y that fulfill this condition. We do this in the following way,

$$P(x+y=z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x)g(y)\delta(z-x-y), \tag{2.30}$$

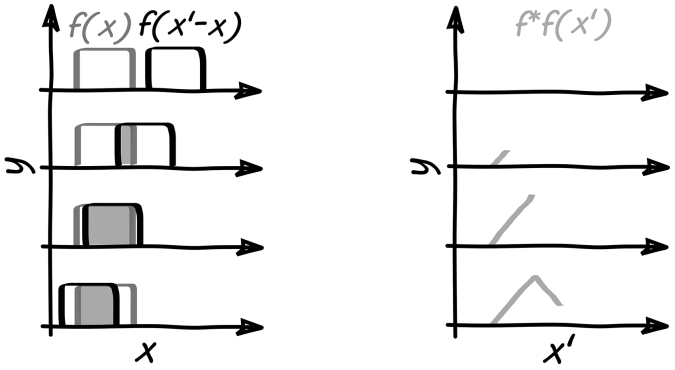

Figure 2.6 *The schematic diagram shows how the convolution product of a function can be understood as the moving overlap of two functions. Initially, we have two box functions f and start shifting the right box to the left. The area by which the boxes overlap (shaded area) is the value of the corresponding convolution product shown in the right panel. The convolution product of two box functions is a triangle. Think of the box functions as the probabilities of the outcomes of a fair dice, one to six. The convolution product is the distribution function of the sum of the face values of two dice that are thrown simultaneously. It is a triangle that spans from two to twelve. Repeated iterations of the convolution product on box functions produce functions that approach a normal distribution very quickly; see Figure 2.7.*

where $\delta(x)$ is the Dirac delta function; see Section 8.1.2. The Dirac delta ensures that $x+y=z$. We can now integrate over y, using the property of the Dirac delta that $\int_{-\infty}^{\infty} dx\, f(x)\delta(x-y) = f(y)$,

$$P(z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x)g(y)\delta(z-x-y) = \int_{-\infty}^{\infty} dx f(x)g(z-x) = f * g(z), \qquad (2.31)$$

where in the last step we used the definition of the convolution product Equation (2.29). We see that the distribution function of $Z = X + Y$ is given by the convolution product of the distribution functions of X and Y. For a graphical illustration of the convolution product, see Figure 2.6. As a result, the probability density function of the sum $X(1) + X(2) + \cdots + X(N)$, of N independent random variables $X(n)$ with the same probability density functions $f(n)$, $n = 1, \cdots, N$, is given by the convolution product $f(1) * f(2) * \cdots * f(N)$. The essence of the central limit theorem is that for a large N, this iterated convolution product converges to the Gaussian distribution. This is true, no matter what f looks like, as long as f has finite mean and variance. We demonstrate this for a pathological-looking distribution function f, shown in Figure 2.7a. As it is convoluted four times with itself, it already looks 'quite' Gaussian; Figure 2.7d.

We can now state the central limit theorem more formally. As for the law of large numbers, there are various ways of stating the theorem, and we use the Lévy–Lindeberg version [253].

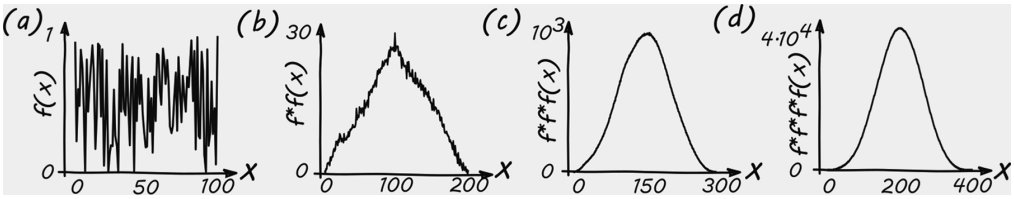

Figure 2.7 *Starting from a pathologically spiky distribution function f (a) we demonstrate how quickly the repeated convolution of f with itself converges to a Gaussian function (b and c). After four iterations, (d) the convolution product already looks Gaussian. The Gaussian distribution is the 'attractor' for an iterated convolution of distribution functions with existing mean and variance.*

Central limit theorem. Assume that $X(n)$, $n = 1, 2, 3, \cdots$, are i.i.d. random variables with an existing mean $\langle X(n) \rangle = \mu$ and an existing variance $\langle X(n)^2 \rangle = \sigma^2$. Then, in the limit of large N, the variable $Y(N) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (X(n) - \mu)$ converges to a Gaussian random variable Z, with mean 0 and variance σ^2. A mean and variance exist if they have finite values and do not diverge for large N.

The general proof is not trivial and we do not reproduce it here. However, to convince ourselves that Gaussian distributions arise in the limit of sums of i.i.d. random variables, it is sufficient to note that the sum of two Gaussian random variables is again a Gaussian random variable. This implies that Gaussian distributions are so-called *attractors* or limit distributions of the convolution product. To see this, let X and Y be two independent Gaussian random variables, both with mean zero and respective standard deviations, σ_1 and σ_2. Using Gaussians for f and g in Equation (2.31), the probability density function of $X + Y$ is found to be,

$$P(x + y = z) = \frac{1}{\sqrt{2\pi}\bar{\sigma}} e^{-\frac{z^2}{2\bar{\sigma}^2}}, \tag{2.32}$$

where $\bar{\sigma}$ is given by $\bar{\sigma} = \sigma_1 \sigma_2 \sqrt{1/\sigma_1^2 + 1/\sigma_2^2}$. The sum of two Gaussian random variables is again a Gaussian variable. This is explicitly shown in Exercise (2.10) at the end of the chapter. See also Equation (2.39). This proves that the Gaussian distributions is a fixed point under the convolution product but it does not prove that the Gaussian distributions are also stable attractors for the iterated convolution product. The central limit theorem (in its various versions) proves exactly this, for example, [253]. However, we can easily convince ourselves of this fact computationally; see Figure 2.7.

2.3.1.3 Log-normal distribution function

The central limit theorem tells us that the expected distribution function for sums of many random numbers is the Gaussian. What about the product of random numbers, $Y(N) = \prod_{i=1}^{N} X(i)$, if the $X(i)$ are i.i.d.? If we take the logarithm of Y and define the variables $G = \log Y$ and $Z = \log X$ we get,

$$G(N) = \log Y(N) = \sum_{i=1}^{N} \log X(i) = \sum_{i=1}^{N} Z(i). \tag{2.33}$$

Recalling the central limit theorem, note that if $Z(i)$ has a finite mean $\langle Z(i) \rangle = \mu$ and variance $\sigma^2(Z) < \infty$, then $Z(i)$ fulfills the conditions for the central limit theorem, and the properly centred and scaled G is a Gaussian random number, $V = G/\sqrt{N} - \sqrt{N}\mu$, with mean 0 and variance σ^2, given, of course, that N is large enough.

If a random variable V is normally distributed with variance $\mu(V)$ and $\sigma^2(V)$ then the variable $U = \exp(V)$ is called log-normally distributed with $\mu(V)$ and $\sigma^2(V)$. Note that here $\mu(V)$ and $\sigma^2(V)$ denote the mean and variance of the underlying Gaussian variable V. The mean $\mu(U)$ and $\sigma^2(U)$ of the log-normal variable U can be computed,

$$\mu(U) = e^{\mu(V)+\frac{1}{2}\sigma^2(V)} \quad \text{and} \quad \sigma^2(U) = e^{2\mu(V)+\sigma^2(V)} \left(e^{\sigma^2(V)} - 1 \right). \tag{2.34}$$

Now, as $V = \lim_{N\to\infty}(G(N)/\sqrt{N} - \sqrt{N}\mu)$ is a normally distributed variable with $\mu(V) = 0$ and $\sigma(V) = \sigma$ one concludes that,

$$U(N) = Y(N)^{\frac{1}{\sqrt{N}}} e^{-\mu\sqrt{N}} \tag{2.35}$$

converges towards a log-normally distributed variable U, with $\mu(U) = e^{\frac{1}{2}\sigma^2}$ and $\sigma^2(U) = e^{\sigma^2}\left(e^{\sigma^2} - 1\right)$. The log-normal variable U follows the probability density distribution,

$$\begin{aligned} p_{\log-\text{normal}}(x|\mu,\sigma) &= \tfrac{1}{dx} P(x < U < x + dx) \\ &= \tfrac{1}{dx} P(\log x < \log(U) < \log(x + dx)) \\ &= \tfrac{1}{dx} P(\log x < V < \log(x + dx)) \\ &= \tfrac{1}{dx} P(\log x < V < \log x + \tfrac{dx}{x}) \\ &= \tfrac{1}{x\ d\log x} P(\log x < V < \log x + d\log x) \\ &= p_{\text{normal}}(\log x|\mu,\sigma)\tfrac{1}{x}, \end{aligned} \tag{2.36}$$

where we used $d\log x = dx/x$ and Equation 2.4. Keep in mind that μ and σ are the mean and standard deviation of $V = \log U$ and not of the variable U. For multiplicative random variables (and the geometric mean of random variables) the log-normal distribution plays the same role as the normal distribution plays for additive variables (and the arithmetic mean) of random variables; see Figure 2.8.

2.3.2 Generalized limit theorems and α-stable processes

The central limit theorem is certainly the best-known probabilistic limit theorem. However, it is not the only one. In fact, there is an entire family of limit theorems that

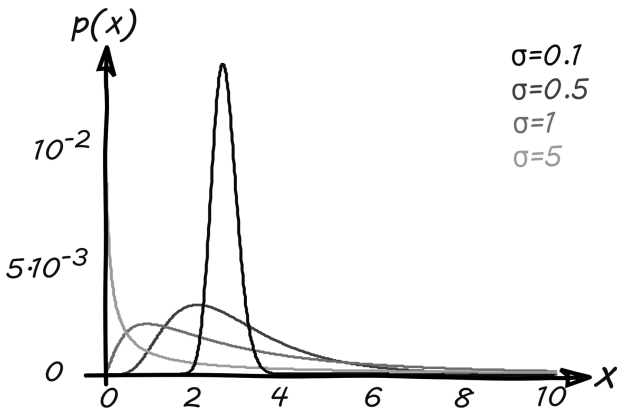

Figure 2.8 *The log-normal distribution for $\mu = 1$ and various values of σ. It is sometimes hard to decide if a given empirical distribution derived from data is a log-normal distribution or a power law.*

are tightly associated with so-called *α-stable processes*. Gaussian random processes are members of this larger family and the central limit theorem is a special case within it. The basic idea behind more general limit theorems is the concept of stable processes. In a nutshell, a process is stable if the sum of two random variables from a given distribution leads to a random variable that is of the same *type*. More precisely:

A random variable X is called *stable* if for the i.i.d. variables $X(1)$, $X(2)$, and $X(3)$ ($X(i) \stackrel{d}{=} X$ for $i = 1,2,3$) there are four constants a, b, c, and d such that,

$$aX(1) + bX(2) \stackrel{d}{=} cX(3) + d. \tag{2.37}$$

Before we discuss the general case, let us focus on the Gaussian example. We have seen before in Equation (2.32) that Equation (2.37) is true for Gaussian random variables; the sum of two Gaussian random variables is again a Gaussian. To show that Gaussian random variables are stable, we can start by assuming $X(1)$, $X(2)$, and $X(3)$ to be i.i.d. Gaussian random variables with unit variance and use the definition of the cumulative distribution function,

$$P(X(i) < x) = \Phi(x) = \int_{-\infty}^{x(i)} dx' \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}, \tag{2.38}$$

for $i = 1,2,3$, and where $x(1) = a$, $x(2) = b$ and $x(3) = c$. The computation proceeds as in Equation (2.31), just the constraint we want to satisfy is now $aX(1) + bX(2) \stackrel{d}{=} cX(3)$, for three constants a, b, and c that satisfy $c^2 = a^2 + b^2$. Remembering that the Dirac

delta function is the derivative of the Heaviside step function (see Section 8.1.1) we compute,

$$\begin{aligned}
&\tfrac{d}{dz}P(aX(1)+bX(2)<cz)\\
&\quad= c\int_{-\infty}^{\infty}dx\tfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\int_{-\infty}^{\infty}dy\tfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\delta(ax+by-cz)\\
&\quad= c\int_{-\infty}^{\infty}dx'\tfrac{1}{2\pi ab}e^{-\frac{x'^2}{2a^2}-\frac{(cz-x')^2}{2b^2}}\\
&\quad= c\int_{-\infty}^{\infty}dx'\tfrac{1}{2\pi ab}e^{-\frac{x'^2}{2}\left(\frac{1}{a^2}+\frac{1}{b^2}\right)+\frac{czx'}{b^2}-\frac{z^2c^2}{2b^2}}\\
&\quad= \underbrace{\frac{c}{\sqrt{2\pi(a^2+b^2)}}e^{-\frac{z^2c^2}{2(a^2+b^2)}}}_{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}}\underbrace{\int_{-\infty}^{\infty}dx'\sqrt{\frac{a^2+b^2}{2\pi a^2b^2}}e^{-\frac{1}{2}\left(\frac{x'\sqrt{a^2+b^2}}{a^2b^2}-\frac{za}{b\sqrt{a^2+b^2}}\right)^2}}_{\int_{-\infty}^{\infty}dx''\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x''^2}=1}\\
&\quad= \tfrac{d}{dz}P(X(3)<z),
\end{aligned}\tag{2.39}$$

where in the last line we used $c^2=a^2+b^2$. This shows that if $X(1)$ and $X(2)$ are standard Gaussian distributed, then so is $X(3)$. The Gaussian is not the only example, however. We can now define a stable process. If a random variable X is stable, then there exist constants c_n and d_n and an i.i.d. sequence $X(n)\stackrel{d}{=}X$, $n=0,1,2,\cdots$, such that for any $n>0$,

$$\sum_{n=1}^{N}X(n)\stackrel{d}{=}c_NX+d_N.\tag{2.40}$$

Stable random variables can be parametrized with four parameters.

- α is the *characteristic exponent*. Parameter α characterizes how the c_n scale with n, that is, $c_n\sim n^{1/\alpha}$ for some $0<\alpha\leq 2$.
- β is the *skewness* and quantifies how asymmetric a distribution function is.[11]
- γ is the *scale*. If we rewrite the stability condition $aX(1)+bX(2)=\gamma X(3)+\delta$, the scale parameter γ may still be meaningful, even if the variance of a process does not exist. γ characterizes how 'stretched' $aX(1)+bX(2)$ is with respect to $X(3)$.
- δ is the *position* that quantifies how much $aX(1)+bX(2)$ is shifted with respect to $X(3)$.

[11] If the mean and variance of a random variable exist, the skewness of the distribution is given by the third moment of the standardized random variable $\beta=\langle((X-\mu)/\sigma)^3\rangle$. Another way of defining skewness is $\beta=\kappa_3(X)/\kappa_2(X)^{3/2}$, where κ_2 and κ_3 are the second and third cumulants of X, respectively.

We can now characterize α-stable processes in terms of limits of random variables,

A random variable X is said to be in the *domain of attraction* of the random variable Y, if, for an i.i.d. sequence with variables $X(n) \stackrel{d}{=} X$, there exist constants $a_n > 0$ and b_n, such that,

$$\lim_{N\to\infty} a_N \sum_{n=1}^{N} X(n) - b_N \stackrel{d}{=} Y. \tag{2.41}$$

If there is at least one random variable X with characteristic exponent $0 < \alpha \leq 2$ that is in the domain of attraction of Y, then Y is called α-stable [292].

2.3.2.1 Distribution functions of α-stable processes

To this day, only three stable random processes are known for which the distribution functions can be written in closed form, meaning that they can be written as a formula. Other stable distribution functions can only be handled numerically.

2.3.2.2 Gaussian distribution function

We discussed Gaussian or normal distribution functions at the beginning of this section. The normal distribution is obtained by stretching and shifting the standard normal distribution $\phi(x)$ of Equation (2.25),

$$p_{\text{normal}}(x|\mu,\sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right). \tag{2.42}$$

Normal distributions are an attractor of the convolution product. They arise in countless situations, from the distribution of velocities in gases in thermal equilibrium to the distribution of errors in repeated measurements of a quantity. Measuring the weight of a tomato with your kitchen scales thousands of times will provide you with a Gaussian distribution of weights. Try it.

2.3.2.3 Cauchy distribution function

This continuous distribution is also called Cauchy–Lorentz or Breit–Wigner distribution. It is given by,

$$p_{\text{Cauchy}}(x|\gamma,\mu) = \frac{1}{c\pi}\frac{c^2}{(x-\mu)^2 + c^2}, \tag{2.43}$$

where c and μ are real-valued parameters. Its asymptotic power law is characterized by an exponent of two, $p_{\text{Cauchy}}(x|\gamma,\mu) \sim x^{-2}$. Note that for the Cauchy distribution neither the mean, variance, nor any other higher moment can be defined. In physics the distribution typically appears in resonance phenomena.

2.3.2.4 Lévy distribution function

This continuous probability distribution is named after Paul P. Lévy and is given by,

$$p_{\text{Levy}}(x|\gamma,\mu) = \sqrt{\frac{c}{2\pi}}\frac{e^{-\frac{c}{2(x-\mu)}}}{(x-\mu)^{3/2}}. \tag{2.44}$$

c and μ are real-valued parameters, and $x > \mu$ is required. Lévy distributions behave asymptotically as $p_{\text{Levy}}(x|\gamma,\delta) \sim x^{-\frac{3}{2}}$. The variance and higher moments of the Lévy distribution do not exist; they are infinite. It has applications in the area of random walks, in particular, in 'first-hit problems', where one wants to know how much time a particle in Brownian motion (random walk) needs to hit a particular target. The Lévy distribution is a special case of the inverse gamma distribution, which is important in the context of Bayesian statistics. Lévy distributions have physical applications, for instance, in the context of spectroscopy or geomagnetic reversal time statistics.

The Cauchy and Lévy distributions are shown in Figure 2.9. Both decay very slowly when compared to the normal distribution. The slow decay is often referred to as a *fat-tailed* distribution, which often decay as an asymptotic power law (for large x) . This implies that *extreme events* (large values of the random variable), are much more likely for fat-tailed processes than they are for processes with Gaussian or exponentially decaying distribution functions. Cauchy and Lévy distributions indeed are asymptotic power laws, as seen in the loglog plot in the inset of Figure 2.9b.

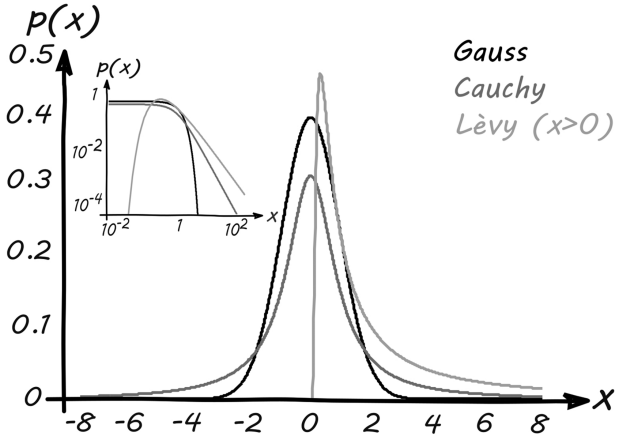

Figure 2.9 *The three families of α-stable distribution functions that can be written as a formula are the Gaussian (black), the Cauchy (dark-grey), and the Lévy (light-grey) distributions. The inset shows the distributions in a loglog plot. For large x they become increasingly linear, indicating an asymptotic power law. Power laws decay* much *more slowly than exponential or Gaussian distributions. They are therefore called* fat-tailed.

A fat-tailed distribution function $p(x)$ decays slowly in comparison with exponential or Gaussian distribution functions. Typically it decays as a power law for large x. The terms fat-tailed, heavy-tailed, long-tailed, scale-free, complex, or power law distribution are often used interchangeably.

2.3.2.5 Lévy flights

The term Lévy flight was coined by Benoît Mandelbrot to refer to random walks with fat-tailed increment distribution functions. *Lévy flights* are random walks $X_{t+1} = X_t + \Delta X_t$, where the step sizes ΔX_t are i.i.d. random variables that have fat-tailed distribution functions.[12] Lévy processes are a continuous generalization of Lévy flights, and are defined as a continuous stochastic process. The corresponding distribution function of Lévy processes can be described by a generalized Fokker–Planck equation of the form,

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}v(x,t)p(x,t) + \frac{\partial^\alpha}{\partial x^\alpha}D(x,t)p(x,t), \tag{2.45}$$

where v represents a flow term, D is a diffusion constant, and $\frac{\partial^\alpha}{\partial x^\alpha}$ is a so-called fractional derivative, which has a clear meaning in the Fourier transformed version of Equation (2.45). In Figure 2.10 we show several examples of two-dimensional Lévy flights with power law distributed increments $\Delta X = (\Delta X^x, \Delta X^y)$, $p(\Delta X) \sim (\Delta X)^{-\alpha}$. Various values of the exponent α are shown. Smaller values of α imply heavier tails in the increment distribution. As a consequence, extreme increments become more likely and the random walk becomes increasingly 'jumpy' for smaller α. We will say more about random walks in Sections 2.5.1.5, 6.4.3.1, and 6.5.3.

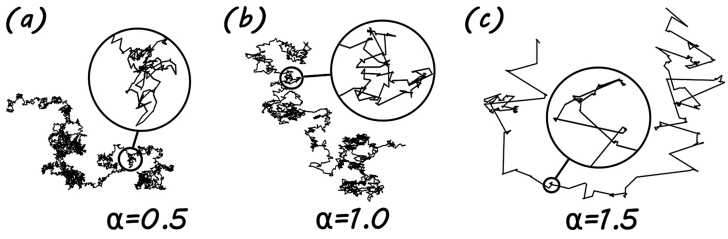

Figure 2.10 *Three examples of Lévy flights with increments are drawn from a power law distribution with three different values of $\alpha = 0.5,\ 1,\ 1.5$. The smaller α becomes, the more 'jumpy' the flights become.*

[12] It is a common misconception that the increments in a Lévy flight X must be from a Lévy distribution. They can be from any fat-tailed distribution.

We briefly summarize this section, which has dealt with the law of large numbers and the fundamental limit theorems of probability theory.

- There are different ways of defining the equivalence of two random variables. Equivalence in distribution means that both variables have the same distribution function.
- The *law of large numbers* tells us that for large numbers of independent trials the sample mean converges to the mean of the random variable used in those trials.
- For a random variable with existing mean and variance the *central limit theorem* states that for a large number of independent trials the sum of the random variable converges to the *normal distribution*.
- Products of many random numbers under appropriate conditions are distributed according to the *log-normal* distribution.
- α-stable distributions are distributions that are 'type-invariant' under summation. This means that sums of i.i.d. random variables are random variables of the same type (with the same distribution function). Stable distributions are characterized by four parameters. Their definition as limits of sums of i.i.d. random variables leads to a generalization of the central limit theorem.
- The only stable limit distributions with known explicit formulas are the Gaussian, Cauchy, and Lévy distributions.
- *Fat-tailed* distributions concentrate much statistical weight in the tails of the distribution. Large values occur relatively often. Fat-tailed distributions appear everywhere in statistical data of complex systems.
- Lévy flights are random walks with fat-tailed increment statistics, $X_{t+1} = X_t + \Delta X_t$. This means that the increment random variable ΔX_t is drawn from a fat-tailed distribution (not necessarily the Lévy distribution).

2.4 Fat-tailed distribution functions

Gaussian distribution functions arise naturally from the central limit theorem whenever random numbers are added. This explains the ubiquity of Gaussian distributions in countless data. One understands the Cauchy and Lévy distributions, which are fat-tailed distributions, on the basis of a slightly more general version of the central limit theorem. There are, however, many more types of fat-tailed distribution functions that appear in the context of complex systems. Is there an equivalent to the central limit theorem for fat-tailed distributions other than Cauchy and Lévy? Not to our current knowledge. However, there are about five classic mechanisms that generate power laws. We discuss these mechanisms in detail in Section 3.3. They include critical and self-organized (critical) systems, multiplicative processes with constraints, preferential processes, and sample space reducing processes. In the following, we review the most

important fat-tailed distributions that appear in the context of complex systems. We present the following distribution functions as *probability density functions* $p(x)$ where the probability of finding a realization of the random variable X in the interval $[x, x+dx]$ is $p(x)dx = P(x < X < x+dx)$. We do so because these distribution density functions, when confronted with data, will frequently be useful for fitting the empirical relative frequency distributions.

2.4.1 Distribution functions that show power law tails

In this Sections, we present well-known distribution functions with power law tails. For historical reasons, multiple names are sometimes attached to one specific form of power law. Some of them are strongly related, some are even exactly equivalent.

2.4.1.1 Pareto distribution

The Pareto distribution, named after Vilfredo Pareto, is a continuous, pure power law with exponent $\alpha + 1$, given by,

$$p_{\text{Pareto}}(x|x_0,\alpha) = \begin{cases} \frac{ax_0^\alpha}{x^{\alpha+1}} & \text{if} \quad x \geq x_0 \\ 0 & \text{if} \quad x < x_0. \end{cases} \tag{2.46}$$

The mean and variance exist if $\alpha > 1$ and $\alpha > 2$, respectively. In that case, the mean is $\langle X \rangle = \frac{\alpha x_0}{\alpha - 1}$ if $\alpha > 1$, and the variance $\sigma^2(X) = \left(\frac{\alpha x_0}{\alpha-1}\right)^2 \frac{\alpha}{\alpha-2}$ for $\alpha > 2$. For every higher moment m to exist, $\alpha > m$ must hold. The Pareto distribution was initially used to describe wealth distributions in economies. Pareto distributions with various values of α are shown in Figure 2.11.

2.4.1.2 Zipf distribution

The Zipf distribution is a discrete distribution function defined as,

$$p_{\text{Zipf}}(x_i|\alpha) = \frac{1}{Z}\frac{1}{x_i^{\alpha+1}}, \tag{2.47}$$

where Z is a normalization constant. If $x_i = i$, $i = 1, 2, \cdots$ then $Z = \zeta(\alpha+1)$, where $\zeta(n) = \sum_{i=1}^{\infty} i^{-n}$ is the zeta function. The Zipf distribution is a discrete version of the Pareto distribution. In the literature, the Zipf probability density distribution is often associated with a value of $\alpha = 0$, for which it cannot be normalized if the number of states i is infinite. However, in many data-related applications this problem does not arise, as all data are finite. In this case, the normalization factor is no longer the zeta function, as the sum $\sum_{i=1}^{N} i^{-n}$ is now taken up to a finite N. Often, Zipf distributions occur in *rank distributions*. Compare, for instance, Section 3.3 or [289]. If a *rank distribution* displays an exponent of -1, it is often called *Zipf's law*. This law was first noted for population sizes in cities by Felix Auerbach and later for word frequencies in written texts by George K. Zipf [418].

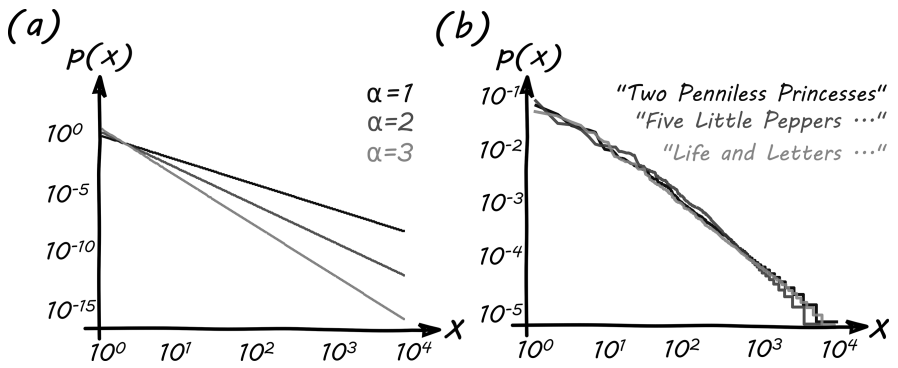

Figure 2.11 *(a) Pareto distributions are pure power laws $p \sim x^{-\alpha}$ defined for $0 < x_0 < x$. For the distribution to be normalizable $\alpha > 1$ is necessary. We show the Pareto distribution in loglog scale for $\alpha = 1, 2, 3$. (b) Empirically observed Zipf's law. The rank-frequency distribution of word occurrences from three novels are shown,* Two Penniless Princesses *by Charlotte M. Yonge (black),* Five Little Peppers and How They Grew *by Margaret Sidney (grey), and* Life and Letters of Lord Macaulay *by Sir George O. Trevelyan (light-grey). The Zipf distribution is usually associated with a scaling exponent of -2. In the rank-frequency distribution shown in (b), this becomes an exponent of -1.*

There are several ways of understanding the origin of this distribution. Zipf himself proposed a 'principle of least effort' [419], which did not become popular. Nowadays, more accepted ways of understanding it are *random languages*, where letters, including a white-space symbol, are randomly typed into a typewriter (a machine used in the last century for writing characters similar to those produced by printers' type) [273], *preferential processes* or Yule–Simon processes [339, 415], or through *sample space reducing* processes [366]. These processes are discussed in more detail in Sections 3.3 and 6.6.2. Empirical Zipf rank distributions of word frequencies are shown in Figure 2.11 for three novels.

2.4.1.3 Zipf–Mandelbrot distribution

An extension of the Zipf distribution is the Zipf–Mandelbrot distribution, which is also known by the name of Zipf–Pareto distribution. It is a discrete distribution function given by,

$$p_{\text{Zipf}-\text{Mandelbrot}}(x_i|c,\alpha) = \frac{1}{Z}\frac{1}{(x_i+c)^{\alpha+1}}, \tag{2.48}$$

where c is a constant, Z is a normalization constant, and $x_i > -c$ must hold for all i. For $x_i = i$ where $i = 1, 2, 3, \cdots, N$, one finds that $Z = H(\alpha, s, N)$ is the Hurwitz function defined as $H(\alpha, c, N) = \sum_{i=1}^{N}(i+c)^{-\alpha}$. For finite N, the distribution function is always normalizable.

2.4.1.4 The q-exponential distribution function—Tsallis distribution

The Tsallis distribution is the distribution function obtained by maximization of the so-called Tsallis entropy, which plays an important role in the statistics of complex systems, as we learn in Chapter 6. The Tsallis distribution is given by,

$$p(x|q,\gamma) = \frac{1}{Z}\left[1-\gamma(1-q)x\right]^{\frac{1}{1-q}}, \tag{2.49}$$

where Z is a normalization constant, x may be continuous or discrete $x = 1, 2, 3, \cdots$, q parametrizes the characteristic exponent of the distribution, and γ is a scale parameter. The asymptotic power law exponent is $1/(1-q)$. The expression, $e_q(x) = [1+(1-q)x]^{\frac{1}{1-q}}$, is called the q-exponential. From the q-exponential one obtains the so-called q-Gaussian if one uses a squared argument, $e_q(-x^2)$. For $q=1$, the Tsallis distribution becomes the exponential or Boltzmann distribution. In this limit, the q-Gaussian converges to the normal distribution. The Tsallis distribution is equivalent to the Zipf–Mandelbrot distribution. If x has no upper bound, it is normalizable for values of $q < 2$. For values $q < 1$, the q-exponential needs to be set to zero, $e_q(x) = 0$ for $x > 1/(1-q)$. This is a function of finite support, meaning that it extends over a limited range of x, and is zero outside. In Figure 2.12a we show q-exponential functions for different values of q and γ. Note that for $q < 1$, q-exponentials have a finite support.

2.4.1.5 Student-t distribution

The Student-t distribution is a continuous distribution that arises naturally in situations with small sample sizes or samples with unknown standard deviations. It is given by,

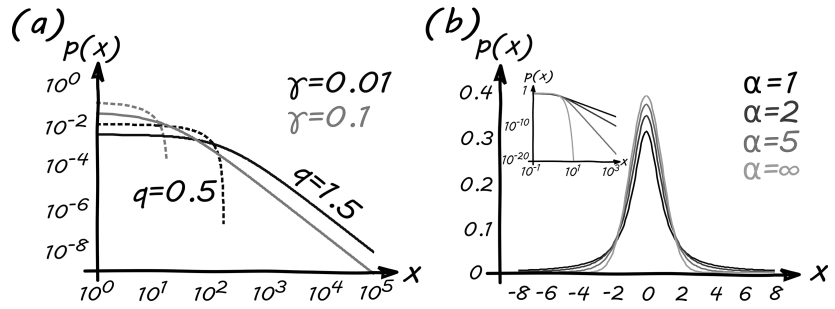

Figure 2.12 *(a) q-exponential distribution functions for two values of the scale parameter γ and two values of q in a loglog plot. For $q = 0.5 < 1$ the distribution has finite support and exists only in a limited region of x. For $q = 1.5 > 1$ the q-exponential distribution is an asymptotic power law with exponent $\alpha = 1/(1-q)$. (b) Student t-distribution for various values of α. As α becomes large, the Gaussian distribution is recovered. In the inset (loglog plot) we see that the Student-t distribution is a power law for large x.*

$$p_{\text{Student}}(x|\alpha) = \frac{\Gamma(\frac{\alpha+1}{2})}{\sqrt{\alpha\pi}\Gamma(\frac{\alpha}{2})}\left(1+\frac{x^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}, \tag{2.50}$$

where Γ is the Gamma function; see Equation (8.12). It was developed at the brewery company, Guinness, by William S. Gosset who used the pseudonym 'Student' for his publications. It was popularized by Ronald A. Fisher [129]. Its characteristic scaling exponent (it is a power law) is $\alpha + 1$. The limit $\alpha \to \infty$ recovers the Gaussian distribution function. In the original context, $\alpha + 1$ is related to the sample size. The Student-t distribution is equivalent to the Tsallis distribution. The Student-t distribution is often used in statistical hypothesis testing because it describes the distribution of the average μ of N i.i.d. random numbers around the true mean μ_0. More precisely, it describes the distribution of the standardized random variable $T = \sqrt{N}(X - \mu_0)/\sigma$. This is where the name t-distribution comes from. The Student-t distribution plays the role of the test statistic for accepting and rejecting the hypothesis that a value μ_0 represents the true mean of the variable X, $\langle X \rangle$; compare Section 2.2.6.1. In Figure 2.12b we show Student-t distributions for several values of α.

2.4.2 Other distribution functions

There are a number of distribution functions that appear frequently in the context of simple and complex systems.

2.4.2.1 Exponential- or Boltzmann distribution

Exponential distribution functions,

$$p(x|\beta) = \frac{1}{Z}e^{-\beta x}, \tag{2.51}$$

where Z is the normalization factor. Exponential distributions emerge whenever independent events i occur exactly n times in a row, given that the probability of sampling i is $q(i)$. The probability of this happening is then $q(i)^n(1 - q(i))$. Another situation where the exponential distribution occurs is the distribution of inter-event times in a Poisson process; see Section 2.5.1.4. Probably the most famous exponential distribution function to appear in physics is the Boltzmann distribution or the Boltzmann factor. It appears as a statistical factor in systems in thermal equilibrium and expresses the fact that energy states E are exponentially distributed,

$$p(E|\beta) = \frac{1}{Z}e^{-\beta E}. \tag{2.52}$$

In a physics context, β is the *inverse temperature*. Exponential distribution functions (see Figure 2.13a) have one property that distinguishes them from all other distribution functions. Changing the off-set of the distribution function (shifting it) does not change

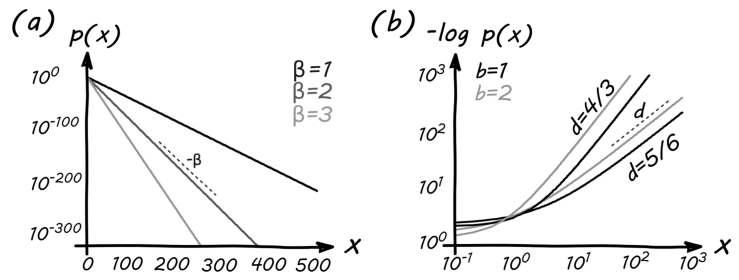

Figure 2.13 *(a) Exponential distribution functions for $\beta = 1, 2, 3$ in a semilog plot (y-axis is logarithmic). (b) Stretched exponential distributions (also called Kohlrausch–Williams–Watts functions) are plotted for different stretching exponents d and a scale parameter b. Note that $-\log(p(x)) = \log(Z) - bx^d$ is a power law. Therefore, if we plot the log of a stretched exponential distribution $-\log(p(x))$ in a loglog plot, it appears as a straight line (asymptotically).*

its overall shape, that is, $p(x|x_{\min}) = Z^{-1}\exp(-\beta(x - x_{\min}))$ remains an exponential. This is not true for any other family of distribution functions [276].

2.4.2.2 Stretched exponential distribution function

Stretched exponential distribution functions are exponential functions with a power exponent on the argument x,

$$p_{\mathrm{SE}}(x|b,\beta) = \frac{1}{Z}e^{-(bx)^d} \quad x > 0, \tag{2.53}$$

where b is a scale parameter and d is the exponent characterizing the 'stretching' of the exponential function. For a continuous sample space $x > 0$, the normalization constant Z is $Z = \int_0^\infty dx\, e^{-bx^d} = \frac{1}{bd}\Gamma(1/d)$, where Γ is the Gamma function. Stretched exponential distributions $p(x)$ can be identified as straight lines in plots where $-\log p(x)$ is plotted in a loglog plot, that is, in $\log x$ versus the $\log(-\log p)$. This becomes clear by noting that $\log(-\log p(x) - \log Z) = d\log x + d\log b$. A stretched exponential distribution function in this representation is shown in Figure 2.13b. The stretched exponential function (not the distribution) is used in many physical, chemical, and statistical contexts, for example, in phenomenological descriptions of relaxation processes in disordered systems.

2.4.2.3 Gumbel distribution

The *Gumbel distributions* is named after Emil J. Gumbel, who analysed biases in politically motivated murders. The Gumbel distribution appears naturally in *extreme value theory*. It describes the statistics of the extreme values recorded in finite samples of data that are sampled from exponential or normal distributions. It is also referred to as the *Fisher–Tippet*, the *log-Weibull*, or the *double exponential* distribution. The Gumbel distribution is used for predicting the likelihood of extreme events, for example, in meteorology [161]. It can be obtained by centering and scaling the so-called standard Gumbel distribution

$p_{\text{Gumbel}}(z)$ by inserting standardized variables $z = (x-\delta)/\gamma$, where δ is a position and γ a scale parameter,

$$\begin{aligned} &\text{Standard Gumbel} && p_{\text{Gumbel}}(z) = \tfrac{1}{Z} e^{-(x+\exp(-x))}, && -\infty < x < \infty \\ &\text{Gumbel} && p_{\text{G}}(x|\delta,\gamma) = \tfrac{1}{\gamma} p_{\text{Gumbel}}\left(\tfrac{1}{\gamma}(x-\delta)\right). \end{aligned} \tag{2.54}$$

For fitting a Gumbel distribution to data consisting of N observations, one typically plots the standardized rank $x = r/(N+1)$, where r denotes the rank of the frequency of observations of a given magnitude versus the relative frequency $p(x|\delta,\gamma)$. In Figure 2.14a the Gumbel distribution is shown for position $\delta = 1$ and scale $\gamma = 1,2,3$.

2.4.2.4 Weibull distribution

Weibull distributions describe the statistics of 'ageing' processes. It is named after Waloddi E.H. Weibull [402] and was developed for applications in materials testing. Lifetimes of electronic components, light bulbs, furniture, and so on, either depend on manufacturing errors or break because they wear out. Product failure at short lifetimes (often called *infant mortality*) is typically caused by manufacturing errors (the new light bulb that blows when you turn it on), whereas for longer lifetimes, failure happens mainly due to wear and tear (the light bulb that blows after being switched on thousands of times). The Weibull distribution allows both types of cause and can tune between the importance of infant mortality versus product mortality due to wear and tear. This is done using a parameter k, where $k < 1$ means that the mortality rate decreases over time, $k = 1$ the mortality remains constant,[13] and $k > 1$ means that the mortality rate increases over time. The standard Weibull distribution is,

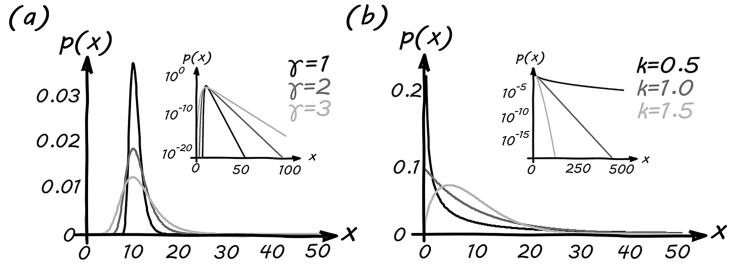

Figure 2.14 *(a) Gumbel distributions are shown for $\delta = 1$ and the values of $\gamma = 1,2,3$. Changing the value of δ would shift the Gumbel distribution along the x-axis (not shown). The inset shows the distribution in a semilog plot. The tail of the distribution decays as an exponential for large x. (b) Weibull distributions for $\gamma = 10$, offset $\delta = -1$, and shape $k = 0.5$ (black), 1 (dark-grey), and 1.1 (light-grey). The shape parameter k is used to control lifetime distribution functions, where the probability of death increases over time: ($k > 1$) remains the same ($k = 1$), or decreases over time ($k < 0$).*

[13] For this situation the exponential distribution is a special case of the Weibull distribution.

$$p_{\text{Weibull}}(z|k) = z^{k-1}e^{-z^k}, \quad z > 0. \tag{2.55}$$

General Weibull distribution functions are obtained using the standardized variable $z = (x-\delta)/\gamma$, where δ determines the position and γ is a scale parameter,

$$p_{\text{Weibull}}(z|k,\delta,\gamma) = \frac{1}{\gamma} p_{\text{Weibull}}\left(\frac{1}{\gamma}(x-\delta)|k\right). \tag{2.56}$$

Parameter k is a stretching parameter of the exponential function. In comparison with the stretched exponential distribution, we see that k corresponds one-to-one to the stretching exponent d of the stretched exponential. In other words, Weibull probability density functions have stretched exponential tails. $k=1$ gives exponential distributions and $k=2$ yields the so-called Rayleigh distribution. A useful thing to remember is that if a random variable X follows the standard exponential distribution, then positive powers $Y \sim X^k$ are Weibull distributed. Examples of Weibull distribution functions are shown in Figure 2.14b for several values of k and γ.

2.4.2.5 Gamma distribution

The Gamma distribution is a combination of a power law and an exponential. It is a two-parameter family of functions that arises in the context of waiting time distributions and Bayesian statistics as solutions of the *maximum entropy principle*; see Section 6.2.3. The exponential distribution and the chi-squared-distribution are special cases. The shape and scale parameters of the distribution are α and β, respectively. Both are positive numbers.

$$p(x|\alpha,\beta) = \frac{1}{Z}x^{\alpha-1}e^{-\beta x} \quad \text{with} \quad Z = \Gamma(\alpha)\beta^{-\alpha}. \tag{2.57}$$

In Figure 2.15a we show the Gamma distribution for various values of α. For Gamma distributed random variables X, the first moment is $\langle X\rangle = \alpha/\beta > 0$ and the variance is $\langle X^2\rangle = \alpha/\beta^2$. The Gamma distribution belongs to the family of the so-called Wishart distributions. The negative binomial distribution is sometimes considered to be a discrete analogue of the Gamma distribution. Famous applications of the Gamma distribution are in the statistics of drought and rainfall patterns [198]. Gamma distribution functions are often used when fitting power laws with an exponential cut-off. These situations often occur as a consequence of limited data size, that is, finite size effects.

2.4.2.6 Gompertz distribution

The Gompertz distribution is a continuous two-parameter distribution function, named after Benjamin Gompertz. Like the Weibull distribution, the Gompertz distribution was developed in the context of lifespan statistics and demographics. More recently, the Gompertz distribution has also found applications in biology and computer science. In the theory of Erdös–Rényi networks (see Section 4.4.2) the Gompertz distribution

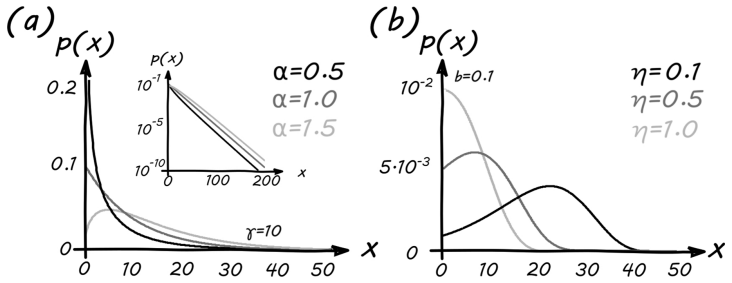

Figure 2.15 *(a) Gamma distributions for shape parameters $\alpha = 0.5, 1.0, 1.5$, and scale parameter $\beta = 1$. They are combinations of power laws and exponential functions. The asymptotic behaviour is dominated by the exponential. (b) Gompertz distributions for various shape parameters $\eta = 0.1, 0.5, 1.0$, and scale parameter $b = 0.1$ are shown. Gompertz distributions appear in lifespan distributions and describe the length distribution of self-avoiding random walks on Erdös–Rényi networks.*

describes the length of *self-avoiding random walks*. The Gompertz distribution function is defined as,

$$p_{\text{Gompertz}}(x|b,\eta) = \frac{1}{Z} e^{bx - \eta e^{bx}} \quad \text{with} \quad Z = \frac{1}{b\eta} e^{-\eta}, \tag{2.58}$$

where $b > 0$ is a scale and $\eta > 0$ a shape parameter. x has to be positive. In Figure 2.15b we show the Gompertz distribution for various parameter values.

2.4.2.7 *Generalized exponential distribution or Lambert-W exponential distribution*

The exponential, the power law, and the stretched exponential distribution functions can all be understood as special cases of the family of *generalized exponential distribution* functions, which we call *Lambert*-W exponential distributions.[14] These are three-parameter functions specified by a scaling parameter $0 < c \leq 1$, a shape parameter d, and a scale parameter r [175]. The generalized exponential distribution function is defined as,

$$p(x|c,d,r) = e^{-\frac{d}{1-c}\left[\mathcal{W}_k\left(B\left(1-\frac{x}{r}\right)^{\frac{1}{d}}\right) - \mathcal{W}_k(B)\right]} \quad \text{with} \quad B = \frac{1-c}{1-(1-c)r} \exp\left(\frac{1-c}{1-(1-c)r}\right), \tag{2.59}$$

where $\mathcal{W}_k$ is kth branch of the Lambert-W function; see Section 8.1.4. The factor B depends on the parameters c, d, and r.[15] Special cases of the generalized

[14] Note that the Weibull and the Gamma distributions are combinations of exponentials and power laws, but they do not contain the stretched exponential.

[15] The parameter r is constrained to $r > \frac{1}{1-c}$, for $d > 0$, $r = \frac{1}{1-c}$, for $d = 0$, and $r < \frac{1}{1-c}$, for $d > 0$. Sometimes, the particular choices $r = \exp(-d/2)/(1-c)$, for $d < 0$, and $r = 1/(1-c+cd)$, for $d > 0$ are useful.

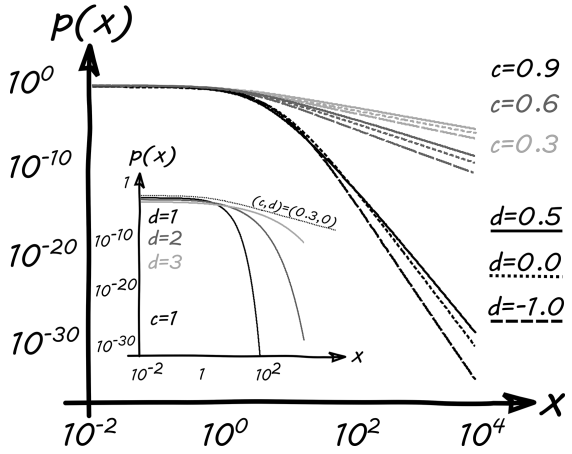

Figure 2.16 *Generalized exponential (Lambert-W exponential) distributions for several choices of scaling parameter c and form parameter d. For $c = 1$ and $d = 1$ the Lambert-W exponential distribution functions yields exponentials. For $c = q$ and $d = 0$ q-exponentials are recovered; see Figure 2.12 and the stretched exponential family appears for $c = 1$ and $d > 0$. Asymptotic power law distributions are shown for $c = 0.9$ (black), $c = 0.6$ (grey), and $c = 0.3$ (light-grey). Different values of d are shown in dotted ($d = 0$), dashed ($d = 0.5$), and dash-dotted ($d = -1$) lines. Stretched exponential distributions for $c = 1$ are shown in the inset for $d = 1$ (black), $d = 2$ (grey), and $d = 3$ (light-grey).*

exponential distribution family are the exponential distributions for $c = 1$ and $d = 1$, the q-exponentials for $c = q$ and $d = 0$, or the stretched exponential family of distributions for $c = 1$ and $d > 0$. This is true for any r. In Figure 2.16 we show the Lambert-W exponential distribution for several values of c and d. Generalized exponential distribution functions appear in empirical distribution functions of history-dependent processes and appear naturally in the context of statistics of strongly correlated systems; compare Sections 6.3, 6.4, and 6.5.

We briefly summarize this section on distribution functions.

- Fat-tailed distribution functions are ubiquitous in complex systems and processes. Various mechanisms are known that produce fat-tailed distribution functions. These include criticality, self-organized criticality, multiplicative processes, preferential attachment, and sample space reducing processes. These will be discussed in Chapter 3.
- We described a number of classical fat-tailed distribution functions. Many of them are related; some are even identical such as the the q-exponential, the Student-t, and the Zipf–Mandelbrot distributions.
- Some important distribution functions that are not fat-tailed such as the exponential, the Gaussian and the stretched exponential distributions, and the Gamma and Weibull distributions have exponential tails.

- Gumbel distributions are important for extreme value theory and record statistics; Weibull distributions allow us to deal with rates that are increasing or decreasing over time.
- Power law, exponential, and stretched exponential distribution functions can be seen as special cases of the Lambert-W exponential or generalized exponential distribution function.
- As a practical rule of thumb, the vast majority of distribution functions are either exponentially decaying or are asymptotical fat-tailed power laws.

2.5 Stochastic processes

Processes are sequences of events that we can observe as trials. Processes evolve over time. A process can be simple, which means that the past does not influence the chances of observing future events. Or the process can be arbitrarily complex, which is typically the case if it has an underlying history-dependent, probabilistic dynamics. The notion of *memory* is essential in the context of stochastic processes. We discuss the Bernoulli process and its multinomial statistics and the Poisson process as examples of simple processes with absolutely no memory. We then proceed to processes where the chances of sampling the next states (events) depend only on the current state of the process. Processes that 'know' where they are, but not where they have been, are called Markov processes. In that sense they have a minimal amount of memory. We finally discuss processes where the chances of future states depend on the history of the process. These are history- or path-dependent processes. Complex systems are typically governed by path-dependent processes.

Mathematically, a random process is a map $t \to X_t$ that associates a random variable X_t with every element of a discrete (or continuous) time line t. In simple processes, all variables X_t draw their possible values from the same discrete (or continuous) sample space; sample space does not evolve. Many complex systems do have co-evolving sample spaces that evolve together with the states of the process over time. The theory of processes described in continuous time and processes with continuous sample spaces requires a set of refined mathematical tools that will not be covered in this book.

We write a process of length N as a set of random variables, $\bar{X}(N) = (X(1), X(2), \cdots, X(N))$. Each random variable $X(i)$ has its own sample space $\Omega(X(i))$. x_i is the concrete outcome of random variable $X(i)$. The probability of a process occurring is defined as the joint probability,

$$P(X(1) = x_1, X(2) = x_2, \cdots, X(N) = x_N). \tag{2.60}$$

Similarly, we denote the conditional probability to find the outcome x_{N+1} at the next time step, given all the previous observations x_i by,

$$P(X(N+1) = x_{N+1} | x_1, x_2, \cdots, x_N). \tag{2.61}$$

2.5.1 Simple stochastic processes

The simplest stochastic processes do not have memory at all: the Bernoulli processes. Or they only have memory about the present: the Markov processes. We begin with processes without memory.

2.5.1.1 Bernoulli processes

Bernoulli processes are sequences of statistically independent trials. An example of a Bernoulli process $t \to X_t$ is a coin that is tossed N times. The outcome at the next time step has nothing to do with the outcomes at previous times. Processes without memory are characterized in terms of their conditional probabilities.

Memoryless processes do not depend on the past, which means that,

$$P(X(N+1) = x_{N+1}) = P(X(N+1) = x_{N+1} | x_1, x_2, \cdots, x_N). \tag{2.62}$$

If the marginal probabilities are all identical, meaning that $P(X(t) = x_t)$ does not depend on time t, the process samples events that are independent and that are from identical distributions at all times, namely, i.i.d. A discrete, memoryless, i.i.d. processes is called a *Bernoulli process*. Sometimes, the term 'Bernoulli process' is used exclusively for a process with a binary sample space, a coin, for instance. We use the term Bernoulli process for i.i.d. processes, no matter how many discrete elements the sample space contains.

As Bernoulli process are statistically independent, the property of Equation (2.62) holds. It can also be shown that the statistical independence of observations in a Bernoulli process follow from Equation (2.62). A *Bernoulli trial* is a single, independent observation of a random variable X_t. Observing a sequence of independent trials of identical random variables X_t constitutes a realization of the *Bernoulli process*. The X_t in the process might represent identical dice or coins.

2.5.1.2 Binomial distribution functions

The simplest Bernoulli processes have *binary* outcomes, {yes, no}, {0, 1}, or {success, failure}. The respective outcomes are distributed according to the *binomial distribution*. Suppose you toss a biased coin N times with probabilities q_0 of getting 'heads', and q_1 of getting 'tails'. We obtain sequences $x(N) = (x_1, \cdots, x_N)$, where x_n is the outcome of the nth trial. We denote the histogram of 'heads' and 'tails' by $k = (k_0, k_1)$, where k_0 (k_1) are the numbers of times 'heads' ('tails') appeared, $k_0 + k_1 = N$. The binomial distribution is the answer to the question: what is the probability of 'heads' occurring exactly n times? We denote this by $p_{\text{Binomial}}(n|N) = P(k_0 = n)$.

How do we obtain this function? As we know that each outcome x_n was obtained by an independent trial, the probability of sampling a sequence $x(N)$ is simply the product of the probability of sampling the outcomes x_n. It follows that the probability of sampling

a sequence with the specific histogram $k = (k_0, k_1)$ is $q_0^{k_0} q_1^{k_1} = q_0^{k_0}(1-q_0)^{N-k_0}$. This probability does not depend on the particular order in which the k_0 'heads' and k_1 'tails' appeared in the sequence. It means that all sequences with the same histogram are equally likely. How many sequences are there with exactly k_0 'heads' out of N observations? We know that this number is given by the binomial factor,

$$\binom{N}{k_0} = \frac{N!}{k_0!(N-k_0)!}. \tag{2.63}$$

We have derived the binomial distribution as the product of the probability of the histogram occurring times the number of possible ways of generating such a histogram in N observations.

The *binomial distribution* is defined as,

$$p_{\text{Binomial}}(n|N) = \binom{N}{n} q_0^n (1-q_0)^{N-n}. \tag{2.64}$$

It is the probability distribution function of a random variable $K = (K_0, K_1)$ with histograms $k_0 = n$ and $k_1 = N - n$, given that there are N samples in a binary i.i.d. process $X = (X_1, \cdots, X_N)$; see Figure 2.17. If $X_n \in \{0, 1\}$; we can then write $K_1 = \sum_{n=1}^{N} X_n$ and $K_0 = N - K_1$.

The binomial distribution is normalized. By using the binomial formula $(a+b)^N = \sum_{n=0}^{N} \binom{N}{n} a^{N-n} b^n$, we see that $1 = (q_0 + q_1)^N = \sum_{n=0}^{N} \binom{N}{n} q_0^{N-n} q_1^n = \sum_{n=0}^{N} p_{\text{Binomial}}(n|N)$. Note that the histogram of N trials is a random variable $K = (K_0, K_1)$. Its possible

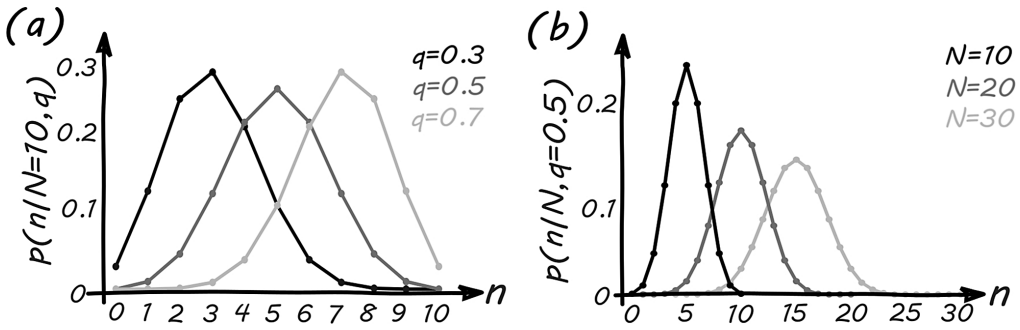

Figure 2.17 *Binomial distribution functions. (a) shows $p_{\text{binomial}}(n|N,q)$ for $N = 10$ identical biased coins with $q = 0.3$ (black), $q = 0.5$ (grey) and $q = 0.7$ (light-grey). In (b) the situation for fair coins $q = 0.5$ and three different values of sequence lengths, $N = 10$ (black), $N = 20$ (dark-grey), and $N = 30$ (light-grey) are shown. The Poisson distribution can be obtained from the binomial distribution in the limit $N \to \infty$.*

outcomes are $k_0 = n$ and $k_1 = N - n$, with $n = 0, \cdots, N$. If $X_n \in \{0, 1\}$, then we can write $K_1 = \sum_{n=1}^{N} X_n$ and $K_0 = N - K_1$. We can also use this to compute the expectation value $\langle K_1 \rangle_N = \sum_{n=0}^{N} p_{\text{Binomial}}(n|N)n$,

$$\begin{aligned}
\langle K_1 \rangle &= \textstyle\sum_{n=1}^{N} \binom{N}{n} (1-q_1)^{N-n} q_1^n n \\
&= \tfrac{d}{d\alpha}|_{\alpha=1} \textstyle\sum_{n=1}^{N} \binom{N}{n} (1-q_1)^{N-n} q_1^n \alpha^n \\
&= \tfrac{d}{d\alpha}|_{\alpha=1} (1 - q_1 + \alpha q_1)^N \\
&= N(1 - q_1 + \alpha q_1)^{N-1} q_1 |_{\alpha=1} \\
&= N q_1 .
\end{aligned} \tag{2.65}$$

Here, we used that $\frac{d}{d\alpha}\alpha^n = n$, if we set $\alpha = 1$. We also get $\langle K_0 \rangle = \langle N - K_1 \rangle = N - N q_1 = (1 - q_1)N = q_0 N$. The second moment is given by,

$$\begin{aligned}
\langle K_1^2 \rangle &= \textstyle\sum_{n=1}^{N} \binom{N}{n} (1-q_1)^{N-n} q_1^n n^2 \\
&= \left[\tfrac{d}{d\alpha} \left(1 + \tfrac{d}{d\alpha} \right) \right] |_{\alpha=1} \textstyle\sum_{n=1}^{N} \binom{N}{n} (1-q_1)^{N-n} q_1^n \alpha^n \\
&= \left[\tfrac{d}{d\alpha} \left(1 + \tfrac{d}{d\alpha} \right) \right] |_{\alpha=1} (1 - q_1 + \alpha q_1)^N \\
&= N q_1 + N(N-1) q_1^2 .
\end{aligned} \tag{2.66}$$

Using that $\sigma^2(K_1) = \langle K_1^2 \rangle - \langle K_1 \rangle^2$ it follows that $\sigma^2(K_1) = N q_1 (1 - q_1) = N q_0 q_1$.

The mean and the variance of the binomial distribution $p_{\text{binomial}}(n|N, q)$ are,

$$\begin{aligned}
\mu &= Nq \\
\sigma^2 &= Nq(1-q).
\end{aligned} \tag{2.67}$$

2.5.1.3 Multinomial processes

Binary Bernoulli processes result in *binomial* distributions. Bernoulli processes that draw from more than two (say, W) values result in *multinomial* distributions of W different events or states. Suppose we have W independent possibilities for the outcome of a Bernoulli trial $i \in \{1, \ldots, W\}$, and the chances of the respective possibilities are $q = (q_1, q_2, \cdots, q_W)$. Assume we perform N trials in a sequence $x(N) = (x_1, \cdots, x_N)$ and record the histogram of its outcomes, $k = (k_1, k_2, \cdots, k_W)$, where obviously $N = k_1 + k_2 + \cdots + k_W$. As before, the essential question that defines the multinomial distribution is: what is the probability of observing a particular histogram k, given that we know the probabilities $q = (q_1, q_2, \cdots, q_W)$ of the underlying Bernoulli process? In complete analogy to the binomial case, one finds the following.

The probability of observing a histogram $k = (k_1, k_2, \cdots, k_W)$ in a Bernoulli process of length N with W states that occur with probabilities $q = (q_1, q_2, \cdots, q_W)$ is given by the *multinomial distribution function*,

$$P(k|q,N) = \binom{N}{k} q_1^{k_1} q_2^{k_2} \cdots q_W^{k_W}, \tag{2.68}$$

where the *multinomial coefficient* is,

$$\binom{N}{k} = \frac{N!}{k_1! k_2! \cdots k_W!}. \tag{2.69}$$

Note that in this notation k is a vector. In the binomial factor k_0 was a number.

2.5.1.4 Poisson processes

Poisson processes are counting processes. They count the number of occurrences of events that happen randomly but at a given rate. One typically thinks of arrivals of events on a time line; see Figure 2.18. The Poisson distribution $p_{\text{Poisson}}(n|\lambda)$ describes the probability of observing n events in a fixed time interval, if the probability density for the events does not vary over time. For example, the number of people entering a specific building at a given rate in ten-minute intervals follows a Poisson distribution. This is no longer true if there is a nearby bus stop and the assumption of a constant arrival rate is wrong. Similarly, the number of drops of water that fall from a dripping faucet at a constant rate in three-minute intervals follows a Poisson distribution. The parameter λ controlling the average rate of events in the time window is called the *intensity* of the Poisson process.

To derive the Poisson distribution, one partitions time into non-overlapping segments of length dt; see Figure 2.19. In every segment a binary Bernoulli trial takes place. The possible outcomes are 'white' or 'grey'. We consider a time interval of length Δt. In

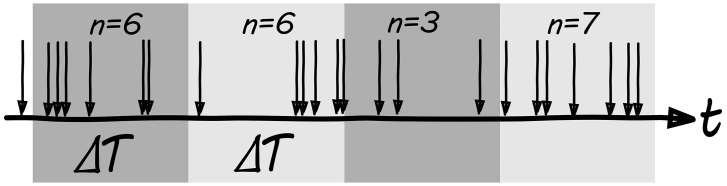

Figure 2.18 *Poisson process. Events happen randomly at a given rate λ. It is not known exactly when they appear, just that in a given time interval ΔT there are, on average, $\lambda \Delta T$ events. The probability of observing n events in a time interval ΔT is described by a Poisson distribution.*

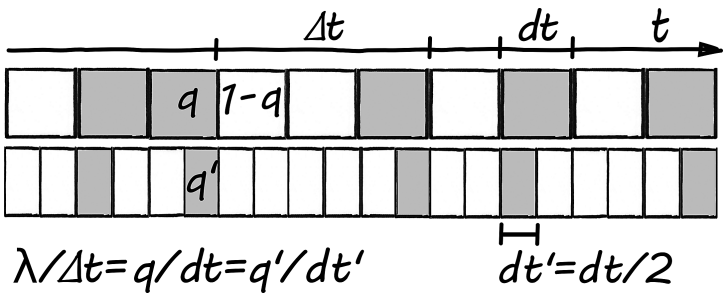

Figure 2.19 *Illustration of how the Poisson distribution can be understood as a continuous version of a binomial distribution. Each square represents one Bernoulli trial, with 'grey' outcomes having a probability q and 'white' having a probability $1-q$. By choosing a finer segmentation of the timeline $dt' = dt/2$, one obtains twice as many Bernoulli trials within the same time interval Δt. Keeping the rate of events, that is, the expected number of 'white' events, $\lambda = q\Delta t/dt$ within a time interval Δt constant, while taking $dt \to 0$, defines the Poisson distribution.*

Δt we find approximately $N \sim \Delta t/dt$ segments of length dt. We can count the number of times 'white' occurred in each segment dt in Δt. Given that at this level of coarse-graining the probability of observing 'white' in a segment dt is given by q, then we will on average observe $\lambda = q\Delta t/dt$ 'white' events in Δt. We then go to a finer segmentation of the time line with segments of length dt'. For instance $dt' = dt/2$. Again we will observe $\lambda = q'\Delta t/dt'$ where q' is the probability of observing 'white' in a segment dt'.

The average number of events λ observed in Δt does not depend on the coarse-graining dt, so that for a fixed intensity λ one finds that the probability of observing 'white' in a segment of length dt asymptotically behaves like $q = \lambda dt/\Delta t$, as dt goes to zero. For every scale $N = \Delta t/dt$ we can think of the Poisson distribution as a binomial distribution of observing 'white' exactly n times in N binary trials with the probability of 'white' being $q = \lambda/N$. Note that the binomial distribution at scale $dt = \Delta t/N$ is,

$$p(n|\lambda, N) = \binom{N}{n}\left(1 - \frac{\lambda}{N}\right)^{N-n}\left(\frac{\lambda}{N}\right)^{n}. \tag{2.70}$$

The Poisson distribution then emerges as the $N \to \infty$ limit of the binomial distribution,

$$p_{\text{Poisson}}(n|\lambda) = \lim_{N\to\infty}\binom{N}{n}\left(1 - \frac{\lambda}{N}\right)^{N-n}\left(\frac{\lambda}{N}\right)^{n}. \tag{2.71}$$

By using Stirling's formula $N! \sim N^N e^{-N}$, see Section 8.2.1, and the fact that $e^x = \lim_{N\to\infty}(1 + \frac{x}{N})^N$, one finds for large N that $\binom{N}{n}(\lambda/N)^n \to \lambda^n/n!$ and that $(1 - (\lambda/N))^{N-n} \to e^{-\lambda}$. Putting these two results together, one obtains the following:

The Poisson distribution is given by,

$$p_{\text{Poisson}}(n|\lambda) = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}. \tag{2.72}$$

It is the probability of observing exactly n independent events in a time interval Δt if the average number of events in Δt (rate) is given by the intensity of the Poisson distribution λ.

In Figure 2.20 the Poisson distribution is shown for several values of λ. It is not difficult to verify that the Poisson distribution is normalized, $\sum_{k=0}^{\infty} P(k|\lambda) = 1$. The mean and variance of the Poisson distribution also exist. To compute μ and σ^2, note that $\langle Y \rangle$ is the average number of events per unit time and we would therefore expect to find $\langle Y \rangle = \lambda$. This is indeed so,

$$\langle Y \rangle = \sum_{n=0}^{\infty} P(n|\lambda) n = \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda} = \sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!}\mathrm{e}^{-\lambda} = \lambda \sum_{m=0}^{\infty} \frac{\lambda^m}{m!}\mathrm{e}^{-\lambda} = \lambda. \tag{2.73}$$

For the variance, a little algebra along the same lines yields, $\sigma^2(Y) = \langle Y^2 \rangle - \langle Y \rangle^2 = \lambda$.

For the Poisson distribution mean and variance have the same value,

$$\mu = \sigma^2 = \lambda. \tag{2.74}$$

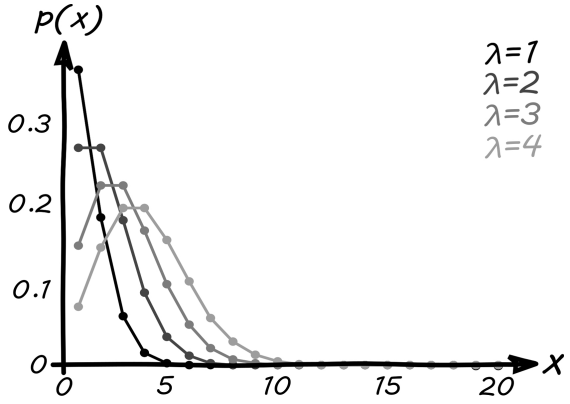

Figure 2.20 *The Poisson distribution for several values of λ, which is a single parameter that controls the distribution. The mean and variance of the Poisson distribution are both equal to $\mu = \sigma^2 = \lambda$. The distribution assigns a probability to the number of events that can be expected to happen in a unit time interval.*

Finally, we can compute the inter-event time distribution p_{IE} of the Poisson process. Looking at Figure (2.19) we could ask how likely it is for two 'white' events to be separated by exactly n 'grey' segments of length dt. Intuitively, it is clear that the average inter-event times $t = ndt$ will be inversely proportional to the intensity λ. Remember that $q = \lambda \Delta t/dt = \lambda/N$, and, as a consequence, we have $n = t/dt = Nt/\Delta t$. Again using $(1 + x/N)^N \sim \exp(x)$ for large N we find,

$$p_{\mathrm{IE}}(t|\lambda, \Delta t) = \lim_{N\to\infty} \frac{1}{Z}(1-q)^n \sim \frac{1}{Z}\left(1 - \frac{\lambda}{N}\right)^{N\frac{t}{\Delta t}} \sim \beta e^{-\beta t}, \tag{2.75}$$

where Z is a normalization constant and $\beta = \lambda/\Delta t$.

The inter-event times of Poisson distributed events follow an exponential distribution,

$$p_{\mathrm{IE}}(t|\lambda, \Delta t) = \frac{\lambda}{\Delta t} e^{-\frac{\lambda}{\Delta t}t}. \tag{2.76}$$

2.5.1.5 Markov processes

Markov processes are processes $t \to X_t$, whose dynamics for the next time step only depend on what was realized in the last time step. In terms of conditional probabilities, Markov processes can be characterized in the following way.

Markov processes are processes for which the probability for the next time step is,

$$P(X_{N+1} = x_{N+1}|x_N) = P(X_{N+1} = x_{N+1}|x_1, x_2, \cdots, x_N). \tag{2.77}$$

This equation states that a Markov process has no memory other than its current position. All older information from times t_n before N with $t_n < t_N$ are irrelevant. We discuss only discrete Markov processes.

In contrast to Bernoulli processes, which have no memory at all, Markov processes remember the last observed event. One way of constructing a Markov process is by 'summing' over Bernoulli trials. How can we see that a Markov process $t \to X_t$ can be written as a sum of a memoryless (independent) 'increment' process $t \to \Delta X_t$, where $t = 0, 1, 2, \cdots$ are elements on a discrete time line? Let us consider an example of a Markov process of this kind; see Figure 2.21. A random walker tosses a coin and walks one step to the left if he tosses 'heads' and one step to the right if he tosses 'tails'. If the walk is associated with the process $t \to X_t$, then the random variable X_t is the position of the walk at time step t, and x_n is the actual value of the position. In the next step $t + 1$, the random variable X_{t+1} can only take values $x_t + 1$ or $x_t - 1$. We can write the random walk,

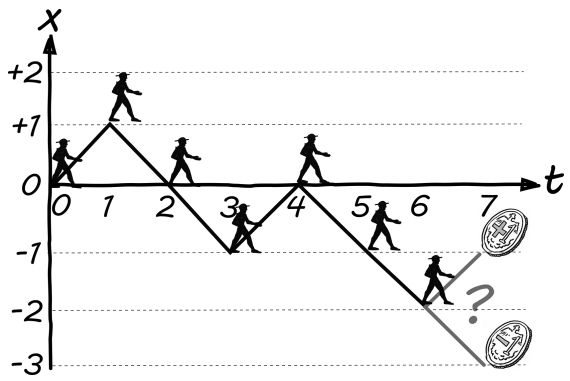

Figure 2.21 *Random walk as an example of a Markov process. A random walker is going left and right depending on the outcome of tossing a coin: one step to the left when tossing 'heads', and one to the right for 'tails'. This random walk is the cumulative sum of the outcomes of binary Bernoulli trials. The next position depends only on the current position. How the walker got there is irrelevant for the next step.*

$$X_{t+1} = x_t + \Delta X_{t+1}, \tag{2.78}$$

where ΔX_n are random variables taking the values $\{+1, -1\}$ with some probabilities $q(n)$ and $1 - q(n)$. We can write $X_t = \sum_{k=0}^{t} \Delta X_k$, where $t \to \Delta X_t$ is the memoryless 'increment' process. If further $q(t) = q$ does not depend on time t, then ΔX is a Bernoulli process. To see that X_n is indeed a Markov process, we need only show that the probability of sampling an event at time n depends only upon events that are not older than the last recorded position x_t at time t. Clearly, the increment process $\Delta X = X_n - X_t = \sum_{k=t+1}^{n} \Delta X_k$ depends only on the outcome of memoryless trials ΔX_k with $k > t$. It follows that information about the path x_k with $k < t$ is irrelevant in terms of predicting the outcome of X_n for $n > t$.

Markov processes are widely used across all sciences. In information theory they are used to model information sources. Information sources can be thought of as random walks on a directed network; whenever a walker passes from one node to the next through a given link, the link emits a specific symbol. Discrete Markov processes with finite sample spaces are equivalent to so-called finite-state machines; see Section 6.2.2, and, in the context of linguistics, they are equivalent to *regular grammars*. Markov processes can be used to describe diffusion processes such as the Brownian motion of particles, which originally refers to the random motion of small particles suspended in liquids, which can be observed in a microscope. The motion is caused by collisions of the particles with water molecules moving at thermal speeds. Much of the area of stochastic calculus, stochastic differential equations, and discrete-time versions of stochastic processes (random walks) have been inspired by this natural phenomenon. Brownian motion was first understood by Albert Einstein on the basis of a random walk model, which led him to a new derivation for the diffusion equation [115]. It allowed him to estimate the size of molecules, which at that time were not widely believed to exist.

Einstein was clever enough to formulate his random walk model of Brownian motion in discrete time. It took more than half a century to clarify the situation mathematically for continuous time—this is now possible, but mathematically rather difficult. The continuous time formulation of stochastic processes, however, does not offer deeper insights into the processes discussed in this book.

An important criterion distinguishes Markovian from non-Markovian processes. For Markovian processes, cumulants of order two and higher vanish, while for history-dependent processes, all cumulants are non-vanishing.[a] If for a random process the third cumulant vanishes, then the process is Markov and all cumulants of order $m > 2$ are zero.

[a] Readers interested in mathematics may wish to look at the *Kramer–Moyal* expansion of stochastic processes and the *Pawula theorem*; see e.g. [143].

The random variables in random walks that describe the increments are usually statistically independent. The main body of the literature on the theory of stochastic processes is built on independent increments. However, random walks can also be built with strongly auto-correlated increments and increments with long-term memory. These will then be non-Markovian walks.

2.5.2 History- or path-dependent processes

Almost all complex systems involve history- or path-dependent processes. In general they are poorly understood. Path-dependent processes are processes $t \to X_t$, where the probability of an outcome at the next time step is a function of the history of prior outcomes.

$$P(X_{N+1} = x_{N+1}|x_N) = f(x_1, x_2, \cdots, x_N). \tag{2.79}$$

It produces sequences $x(N)$. There are 'mild' versions of path-dependent processes that depend not on the exact order of the sequence, but only on the histogram of $x(N)$, and $P(X_{N+1} = x_{N+1}|x_N) = f(k(x(N)))$. Processes of this kind can be partially understood. We will discuss them in Section 6.5.2.3. If path-dependent processes also depend on the order of the sequence, then it is quite difficult to understand their statistics.

In contrast to Markov processes, path-dependent processes are frequently non-ergodic, which means in this context that time averages and ensemble averages do not coincide. What does this mean? Take N dice, throw them all at once, and generate a list of the outcomes $x(N) = (x_1, \ldots, x_N)$. In doing that, we create *ensembles* of realizations. If now we take a single dice and throw it N times in a row and record the outcomes in a sequence, $x'(N) = (x'_1, \ldots, x'_N)$, the averages of X and X' are identical. We say that the ensemble- and sequence pictures coincide. For path-dependent processes this equivalence is no longer true. Using ensemble statistics for path-dependent processes will in general lead to nonsensical results. Think, for example, of processes with symmetry breaking, where

some events early on in the process determine the future of the process. Consider the silly toy process Y_t, where $Y_1 \in \{-1, 1\}$ represents a fair coin-tossing experiment, with the additional rule that after the first event, the process continues deterministically $Y_{t+1} = Y_t$ for all future time steps. The ensemble average over all possible paths[16] is $\langle Y \rangle_{\text{ensemble}} = 0$. The time average, or the sample mean, is either $\langle Y \rangle_{\text{sample}} = -1$ or $\langle Y \rangle_{\text{sample}} = 1$, each occurring with probability $1/2$. This leads us to a fundamentally important observation.

For history- or path-dependent processes, the law of large numbers does not, in general, hold. The intuition that more samples will increase the accuracy of the average breaks down.

There are a few families of history- or path-dependent processes where scientific progress has been made. In particular, these are *reinforcement processes*, *processes with dynamic sample spaces*, and *driven dissipative systems*.

2.5.3 Reinforcement processes

As their name suggests, reinforcement processes reinforce their dynamics as they unfold. A typical reinforcement process is a Pólya urn, see Section 2.1.1, which we will study in detail towards the end of Chapter 6. Pólya urns are urns that contain balls of different colours. Whenever a ball of a given colour is drawn, several new balls of the same colour are added to the urn. It is a self-reinforcing Markov process that shows the rich-get-richer phenomenon or the winner-takes-all dynamics.

More general situations of probabilistic reinforcement processes are found in biology in the context of auto-catalytic cycles of chemical reactions. In auto-catalytic cycles one type of molecule can enhance (catalyse) or suppress the production or degradation of other molecules [113]. Similar cycles exist in production networks in the economy. Neural systems, where the frequent use of a particular neural pathway can modify the synaptic coupling strengths of the pathway, are examples of path-dependent processes that 'learn from experience' [186]. At a macro level, such systems are usually modelled with non-linear differential equations; at a micro level, Boolean networks are often used. It is often possible to picture these processes in terms of network models of interacting urn processes, where an event in one urn can modify the content of others urns.

A common phenomenon observed in reinforcement processes with positive and negative reinforcement mechanisms, is *multistability*, which means that there are multiple attractors in the dynamics. In these systems, an event may cause a path or a trajectory to abandon its basin of attraction and to enter the basin of another attractor. These shifts are often experienced as 'regime shifts'. Such mechanisms are important for many processes that involve complex regulatory systems, such as cells or socio-economic systems. In the

[16] There are only two paths, $(1, 1, \cdots, 1)$ and $(-1, -1, \cdots, -1)$.

context of cells, regime shifts would, for example, be associated with cell differentiation. We will learn more about the statistics of simple reinforcement processes in Chapter 6.

Complex regulatory systems can often be modelled as networks of reinforcing urn experiments, where drawing one event in one urn can increase or decrease the chances of events in dependent urns. Such processes typically show multistable stochastic dynamics. In particular, auto-catalytic systems, which we encounter in Chapter 5, can be modelled in this way.

2.5.4 Driven dissipative systems

Many systems are driven dissipative systems, which usually means that there is an energy gradient that drives the system away from equilibrium. Without that gradient, systems would relax to equilibrium. We often encounter two competing dynamics, one that drives and excites the system, the other being the process of relaxation. The resulting dynamics can be extremely rich and shows fat-tailed statistics in many observable quantities. The reason for this is that the sample spaces of these systems are not static but evolve over time. There has been much progress recently in understanding the probabilistic nature of driven dissipative systems. Think of a pot filled with water on a stove. There are two competing thermal processes. The stove heats the water and the water cools at the interfaces with the surrounding air. Energy flows through the system from the stove through the water to the air. Flows of energy through a system make these systems do things they would never do in equilibrium, such as produce convective currents or turbulence. Systems where one process (a source) 'charges' the system and another (the sink) 'discharges' it is called *driven*. The word dissipative means that energy is flowing in and out of these systems.

With the exception of a few physical phenomena, virtually all processes that take place on this planet, including the entire biosphere and the anthroposphere, are driven non-equilibrium processes. Systems in equilibrium are the exception in nature and exist practically only in laboratories or for systems at relatively short timescales. Any machine, say an electric drill, is a driven dissipative system. An energy flow is needed to keep the machine running. If we interrupt the flow, the machine stops. The energy fed into the system will eventually be converted to work and heat, which dissipate into the environment. The energy source of the earth is the sun, and its energy sink is the approximately 3 Kelvin background radiation of the universe.

Perhaps the simplest example of a driven dissipative process is something that children might do, when home alone and bored. Take a tennis ball, take it upstairs, drop the ball, and watch it bounce downstairs; see Figure 2.22. When the ball reaches the bottom of the stairs, bring it back to the top and drop it again.[17] As the ball bounces down the

[17] One of the authors occasionally found this 'game' to be a fun pastime as a child. Note that if such experiments are carried out in reality, the way a ball bounces downstairs can be quite erratic. A ball does not always keep on landing on one lower stair after another. Sometimes the ball hits an edge of a stair and then bounces straight to the bottom of the stairs, and sometimes the ball gets stuck on a step and never reaches the bottom.

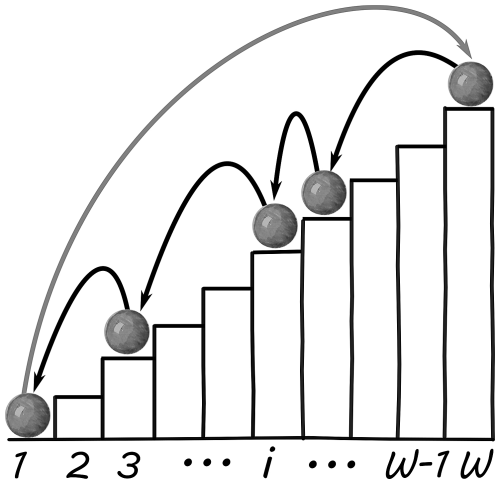

Figure 2.22 *A ball bouncing down a staircase as a simple model of a driven dissipative process. The ball bounces downwards only at random distances. Along the way it samples various steps that are below its current position. Note that the process is a Markov process. The visiting distribution of the stairs is a power law; in this case, it is Zipf's law, from Section 2.4.1.*

stairs and comes to rest at the bottom of the stairs it dissipates energy in the form of heat produced in inelastic collisions with the stairs. What drives the process? The energy that the child uses to take the ball back upstairs—where does that energy go? It dissipates in form of heat into the environment. In the following, we will consider an idealized version of this process. It will allow us to understand the origin of the peculiar statistics of driven dissipative systems; see Section 3.3.5. In particular, we will learn how the dynamics of the sample space is related to the statistics of the system. In the context of complex systems one often associates driven dissipative systems with self-organized critical systems, which are usually more complicated realizations of this example.

Driven dissipative processes often consist of a dynamical sample space that increases and decreases over time. The increase comes from a driving source process that 'loads' the system into an excited state of 'high' energy with many possibilities to dissipate into. Sample space decrease is associated with a 'sink' process, that relaxes the system towards 'low energy' states (sinks) as energy is dissipated. We will discuss sample space reducing processes in Section 6.6.2.

Driven systems are driven by a 'source' and a 'sink' process. The source process heaves a system into a 'high energy' state and is typically sample space expanding. The sink process allows the system to relax towards 'low energy' by dissipating this energy into the environment. The sink process is typically sample space reducing. Source and sink processes will create an 'energy flow' through the system, which at the microscopic level is described with a probabilistic dynamics that violates detailed

continued

balance. This means that statistical currents between two adjacent states of the system do not compensate for each other. In non-physical complex systems the role of energy in the driving process is played by other quantities.

2.5.4.1 Processes with dynamical sample spaces

Many stochastic processes have sample spaces that change over time. For example, they occur in evolutionary processes that involve both *innovations* and *extinctions*, such as in biological evolution or technological innovation. They have dynamical sample spaces that can shrink and expand as they evolve.

Processes with dynamical sample spaces are random processes $t \to X_t$, where the sample spaces $\Omega(X_t)$ of the random variables X_t change with time t. For the example of a ball randomly bouncing down a staircase, let us label the W individual stairs as states i. In the beginning, when the ball is at the top in state W, it can bounce to any other stair below, meaning that $\Omega_W = \{1, 2, \cdots, W-1\}$. If the ball is later at stair $n < W$ its sample space is now smaller, as it can only reach the states $\Omega_n = \{1, 2, \cdots, n-1\}$. At the bottom, the process needs to be restarted. If it is not restarted the process is strictly sample space reducing. Restarting, or charging, or driving usually means to expand the sample space of a stochastic system. At higher energy levels there are simply more ways of dissipating the energy, meaning that more possible states or outcomes can be reached or sampled.

It transpires that pure sample space reducing processes generically produce an exact Zipf law. If sample space reducing (relaxing) and expanding (driving) processes are balanced in appropriate ways, practically every statistical distribution function that we have discussed in this chapter can be recovered for some simple driven dissipative systems. We will show how this is done in Section 3.3.5.

What have we learned about random processes?

- Stochastic processes are sequences of subsequent random trials.
- Bernoulli processes are time-independent and memoryless. They are sequences of i.i.d. Bernoulli trials over a finite and discrete sample space.
- Events in Bernoulli processes are multinomially distributed.
- Poisson distributions are obtained as the $N \to \infty$ limit of binomial distributions.
- Markov processes have a memory of the most recent trial only. The distribution function or the sample space for sampling the next event in a Markov process only depends on the last observation. The next step in a Markovian random walk does not depend on the details of the history of the walk, but only on the last position in the walk.
- Sums of memoryless processes yield Markov processes. In general, increments of Markov processes are random variables $\Delta X(i,t)$, which are independent but not necessarily identically distributed. They can depend on the state i they are in at time t. When at time t the Markov process is in state i it samples the increment $\Delta X(i,t)$ to move to the next state.

- Random walks often have no memory of their increment process. Such random walks are Markov processes. However, random walks can have history-dependent or auto-correlated increment processes. This leads to path-dependent random walks.
- The challenge with history-dependent processes is that ensemble averages and sample means are no longer the same. In general, the law of large numbers does not hold for path-dependent processes; more samples will not necessarily increase the accuracy of measurements of stochastic variables.
- The Pólya urn process is a history-dependent process that is simple enough for its dynamics to be perfectly understood. Pólya urns are frequently used to model (coupled) self-reinforcing processes.
- Driven dissipative systems are at the core of many complex systems. They have a part that drives them and they have a relaxing dynamics. The driving part usually expands sample space volume; the relaxing part reduces it. The statistics of driven dissipative systems can be partially understood if the dynamics of the sample spaces is known.
- Evolutionary processes are processes with dynamical sample spaces, where innovation and extinction events modify the volume of sample space over time.
- In terms of understanding stochastic processes associated with the dynamics of complex systems, there is vast and unexplored terrain in front of us. Progress in the science of complex systems will, to a large extent, depend on the rate of progress in the field of path-dependent processes.

2.6 Summary

We briefly summarize the chapter and what we learned about probability and random processes.

- We learned how to handle and compare random numbers. The iterated addition of random numbers led us to the central limit theorem. The Gaussian distribution arises as an attractor of sums of random numbers. The log-normal distribution arises as an attractor for products of random numbers. We learned that two more distributions arise naturally as attractors for stable processes: the Cauchy and Lévy distributions. Both are asymptotic power laws.
- We discussed those fat-tailed distribution functions that frequently appear in many areas of complex systems. We encountered distribution functions that arise in extreme value theory and distributions that can be seen as generalizations of other distributions. In this context, we found that the Gamma distribution is a combination of exponential and power law, and the generalized exponential function (Lambert-W exponential) is a generalization of the exponential, the

continued

power law, and the stretched exponential function. The generalized exponential function will appear naturally in the context of entropies for complex systems in Chapter 6.

- We learned about stochastic processes. We divided them into classes that have no memory (Bernoulli processes), those that have a one-step memory (Markov process), and those that are history- or path-dependent. We encountered the multinomial distribution and Poisson distribution, and slightly touched upon reinforcement processes. These will be discussed in detail in Chapter 6.
- We concluded the chapter by introducing the notions of driven dissipative systems and systems with dynamical sample spaces. We will develop a theory of driven dissipative systems in detail in Chapters 3 and 6, meaning that a consistent way of computing the statistics and distribution functions for stationary driven systems will be presented. Systems with reducing sample spaces are tightly related to driven systems and allow us to understand the emergence of different distribution functions, including power laws. This will be demonstrated in Chapter 3.
- We mentioned the different philosophies with regard to hypothesis testing behind the Bayesian and the frequentist approaches.

2.7 Problems

Problem 2.1 Suppose an event A can be drawn with probability p_A. How often on average can we expect to draw A in a row? Ask yourself why the probability of drawing A exactly n times is given by $p_A^n(1-p_A)$. Use the relation to compute the average $\langle n\rangle$. Compare the result with Equation (2.1).

Problem 2.2 Show in detail that Equation (2.16) indeed yields 3/4. Use the hints that are given before the equation in the text.

Problem 2.3 Prove that $N!$ counts the number of ways N mutually distinct objects can be placed on to N distinct positions $n = 1, \cdots, N$. Try induction on N, that is, show that 1! counts the number of ways that 1 item can be placed on one position. Then show that if it is true for N that $N!$ counts the number of ways N objects can be placed on N positions, then it is also true for $N+1$.

Problem 2.4 Compute in full detail the limit $N \to \infty$ of the binomial distribution to get the Poisson distribution. Use the outline for the computations given in Section 2.5.1.4.

Problem 2.5 Prove that the Poisson distribution is normalized. You may wish to remind yourself that the Taylor expansion of the exponential function is $e^x = \sum_{n=1}^{\infty} \frac{1}{n!} x^n$.

Problem 2.6 Compute the variance of the Poisson distribution with intensity λ.

Problem 2.7 Show from the definition of the expectation value, Equation (2.7), and $\sigma^2(X) = \langle (X - \langle X\rangle)^2\rangle$ that $\sigma^2(X) = \langle X^2\rangle - \langle X\rangle^2$.

Problem 2.8 Compute the third and fourth cumulant of a random variable X. Use the cumulant-generating function Equation (2.9) to compute the third and fourth cumulant in terms of moments.

Problem 2.9 Compute the variance $\sigma^2(k_1)$ of a binomial distribution after N trials of the underlying Bernoulli process. Use the binomial formula $(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$, the definition of the variance, and the fact that for binomial processes $\langle k_1 \rangle = Nq$.

Problem 2.10 Completing squares can be used to solve the following integral in Equation (2.32). Note that we use the formula for the Gaussian integral, $\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$ in the sixth line.

$$
\begin{aligned}
\tfrac{d}{dy}P(X_1+X_2<y) &= \int_{-\infty}^{\infty} \tfrac{dx_1\,dx_2}{2\pi\sigma_1\sigma_2} e^{-\frac{x_1^2}{2\sigma_1^2}-\frac{x_2^2}{2\sigma_2^2}}\,\delta(z-x_1-x_2) \\
&= \int_{-\infty}^{\infty} \tfrac{dx_1}{2\pi\sigma_1\sigma_2} e^{-\frac{x_1^2}{2\sigma_1^2}-\frac{(x_1-z)^2}{2\sigma_2^2}} \\
&= \int_{-\infty}^{\infty} \tfrac{dx_1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left\{x_1^2\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)-\frac{2x_1 z}{\sigma_2^2}+\frac{z^2}{\sigma_2^2}\right\}} \\
&= \int_{-\infty}^{\infty} \tfrac{dx_1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2\sigma_3^2}\left\{x_1^2-2x_1 z\frac{\sigma_3^2}{\sigma_2^2}+z^2\frac{\sigma_3^2}{\sigma_2^2}\right\}} \\
&= \int_{-\infty}^{\infty} \tfrac{dx_1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2\sigma_3^2}\left\{\left(x_1-z\frac{\sigma_3^2}{\sigma_2^2}\right)^2+z^2\frac{\sigma_3^2}{\sigma_2^2}\left(1-\frac{\sigma_3^2}{\sigma_2^2}\right)\right\}} \\
&= \tfrac{\sigma_3}{\sqrt{2\pi}\sigma_1\sigma_2} e^{-\frac{z^2}{2\sigma_2^2}\left(1-\frac{\sigma_3^2}{\sigma_2^2}\right)} \\
&= \tfrac{1}{\sqrt{2\pi}\sigma_4} e^{-\frac{z^2}{2\sigma_4^2}}.
\end{aligned}
$$

In the third line, we define $1/\sigma_3^2 = 1/\sigma_1^2 + 1/\sigma_2^2$. In the last line, we need to identify σ_4 with two distinct expressions, (i) $\sigma_4 = \sigma_1\sigma_2/\sigma_3$ and (ii) $\sigma_4^2 = \sigma_2^2/(1-\sigma_3^2/\sigma_2^2)$. Show that the identifications (i) and (ii) indeed yield the same result (σ_4), that is, show that

$$
\sigma_1^2\sigma_2^2/\sigma_3^2 = \frac{\sigma_2^2}{1-\left(\frac{\sigma_3}{\sigma_2}\right)^2},
$$

is true.

Problem 2.11 Suppose $X \in \log-\mathcal{N}(\mu,\sigma^2)$ is log-normally distributed, where the μ and σ^2 are the mean and variance of the normal distribution. Compute the relations between μ, σ, and $m = \langle X \rangle$ and $v = \sigma^2(X) = \langle (X-m)^2 \rangle$. Look up the integrals you need to solve in an integral table.

Problem 2.12 Show that the Zipf–Mandelbrot distribution Equation (2.48) can be written exactly as a q-exponential function as in Equation (2.49). Work out the exact relations between the various constants in the corresponding equations.

Problem 2.13 Compute the Lambert-W exponential in Equation (2.59) for (i) $(c,d) = (c,0)$ and (ii) $(c,d) = (1,d)$. Use the fact that for small $x \sim 0$ the Lambert-W function behaves like $W_0(x) \sim x$.