Peter Schuster

Stochasticity in Chemistry and Biology

When small population sizes matter and environments fluctuate

July 21, 2013

Springer

Preface

Statistics and stochastic processes are often neglected mathematical disciplines in the education of chemists and biologists, although modern experimental techniques allow for investigations on small sample sizes down to single molecules and most measured data are sufficiently accurate to allow for direct detection of fluctuations. The progress in the development of new techniques and the improvement in the resolution of conventional experiments has been enormous within the last fifty years. Indeed, molecular spectroscopy provided hitherto unimaginable insights into processes down to the hundred attosecond range and current theory in physics, chemistry, and the life sciences cannot be successful without a deeper understanding of randomness and its causes. Sampling of data and reproduction of processes are doomed to produce artifacts in interpretation unless the observer has a solid background in the mathematics of limited reproducibility. As a matter of fact stochastic processes are much closer to observations than deterministic descriptions in modern science and everyday life. Exceptions are the motions of planets and moons as encapsulated in celestial mechanics, which stood at the beginnings of science and modeling by means of differential equations. Fluctuations are so small that the cannot be detected even in highest precision measurements: Sunrise, sunset, and solar eclipses are predictable with practically no scatter. Processes in the life sciences are often entirely different. A famous and characteristic historical example are Mendel's laws of inheritance: Regularities are detectable only in sufficiently large samples of individual observations, and the influence of stochasticity is ubiquitous. Processes in chemistry are between the extremes: The deterministic approach in conventional chemical reaction kinetics has neither suffered a loss in applicability nor did the results become less reliable in the light of modern experiments. What has increased rather dramatically is the accessible resolution in detectable amounts, space, and time. Deeper insights into mechanisms provided new access to molecular information for theory and practice.

Biology is currently in a state of transition: The molecular connection to chemistry revolutionized the sources of biological data and is setting the stage

for a new theoretical biology. Historically biology was based on observation and theory in biology was engaged only in interpretations of the observed regularities. The development of biochemistry at the end of the nineteenth and the first half of twentieth century introduced quantitative thinking in terms of chemical kinetics into some biological subdisciplines. Biochemistry attributed also a new dimension to experiments in biology in the form of *in vitro* studies on isolated and purified biomolecules. A second import of mathematics into biology came in the form of population genetics, which was created in the nineteen twenties as a new theoretical discipline uniting Darwin's natural selection and Mendelian genetics more than twenty years before evolutionary biologists completed the so-called *synthetic theory* performing the same goal. Beginning in the second half of the twentieth century molecular biology started to build a comprehensive bridge from chemistry to biology and enormous progress in experimental techniques created a previously unknown situation in biology insofar as new procedures were required for data handling, analysis, and interpretation since the volume of information is drastically exceeding the capacities of human mind. Biological cells and whole organisms are now accessible to complete description at the molecular level and the overwhelming amount of information thought to be required for a deeper understanding of biological objects is simply a consequence of the complexity of biology and the lack of a universal theoretical biology.

The current flood of results from molecular genetics and genomics to systems biology and synthetic biology requires apart from computer science techniques primarily suitable statistical methods and tools for verification and evaluation of data. Analysis, interpretation, and understanding of experimental results, however, is impossible without proper modeling tools. These tools were so far mainly based on differential equations but it has been realized within the last few years that an extension of the available repertoire by methods derived from stochastic processes is inevitable. Moreover, the enormous complexity of the genetic and metabolic networks in the cell calls for radically new methods of modeling that resemble the mesoscopic level of description in solid state physics. In mesoscopic models the overwhelming and for many purposes dispensable wealth of detailed molecular information is cast into a partially probabilistic description in the spirit of *dissipative particle dynamics*, and such a description cannot be successful without a solid background in stochastic methods. The field of stochastic processes has not been bypassed by the digital revolution. Numerical calculation and computer simulation play a decisive role in present day stochastic modeling in physics, chemistry and biology. Speed of computation and digital storage capacities are growing exponentially since the nineteen sixties with an approximate doubling time of eighteen month, a fact that is commonly addressed as Moore's law [221]. It is not so well known, however, that the spectacular exponential growth in computer power has been overshadowed by the progress in numerical mathematics that led to an enormous increase in the efficiency of algorithms. To give just one example, which was reported

by Martin Grötschel from the Konrad Zuse-Zentrum in Berlin [133, p. 71]: The solution of a benchmark production planning model by linear programming would have taken 82 years CPU time in 1988, using the computers and the linear programming algorithms of the day. In 2003 – fifteen years later – the same model could be solved in one minute and this means an improvement by a factor of about 43 million. Out of this, a factor of roughly 1 000 resulted from the increase in processor speed whereas a factor of 43 000 was due to improvement in the algorithms, and many other examples of similar progress in the design of algorithms can be given. Understanding, analyzing, and designing high-performance numerical methods, however, requires a firm background in mathematics. The availability of cheap computing power has also changed the attitude towards exact results in terms of complicated functions: It does not take so much more computer time to compute a sophisticated hypergeometric function than to calculate an ordinary trigonometric function for an arbitrary argument, and operations on confusing expressions are enormously facilitated by symbolic computation. In this way the present day computational facilities have also large impact on the analytical work.

In the past biologists had often quite mixed feelings for mathematics and reservations against the use of theory. The recent developments in molecular biology, computation, and applied mathematics, however, seem to initiate a change in biological thinking since there is practically no chance to shape modern biology without mathematics, computer science and theory as the biologist Sydney Brenner, an early pioneer of molecular life sciences, points out [26]: *"... it is clear that the prime intellectual task of the future lies in constructing an appropriate theoretical framework for biology. Unfortunately, theoretical biology has a bad name because of its past. ... Even though alternatives have been suggested, such as computational biology, biological systems theory and integrative biology, I have decided to forget and forgive the past and call it theoretical biology."* He and others are calling for a *theoretical biology new* that allows for handling the enormous complexity. Manfred Eigen stated very clearly what can be expected from theory [49, p. xii]: *"Theory cannot remove complexity but it can show what kind of 'regular' behavior can be expected and what experiments have to be done to get a grasp on the irregularities."* Theoretical biology will have to find the appropriate way to combine randomness and deterministic behavior in modeling and it is not very risky to guess that it will need a strong anchor in mathematics in order to be successful.

In this monograph an attempt is made to collect the necessary mathematical background material for understanding stochastic processes. In the sense of Albert Einstein's version of Occam's razor [28, pp. 384-385; p. 475], *"... Everything should be made as simple as possible, but not simpler. ..."*, dispensable deep dwelling in higher mathematics has been been avoided. Some sections that are not required if one is primarily interested in applications are marked for skipping by readers who are willing to accept the basic results without explanations. On the other hand the derivations of analytical solu-

tions for the selected examples are given in full length because the reader who is interested to apply the theory of stochastic processes in practice should be brought in the position to derive solutions on his own. An attempt was made to use a largely uniform notation throughout the book that is summarized in a separate table an the end. A glossary is added to define the most important notions used in the text. We refrained from preparing a separate section with exercises, instead case studies, which may serve as good examples for calculations by the reader, are indicated in the book. Sources from literature were among others the text books [34, 93, 97, 132, 191]. For a brief and concise introduction we recommend [144]. Standard textbooks in mathematics used for the courses were: [22, 204, 249].

This book is derived from the manuscript of a course in stochastic chemical kinetics for graduate students of chemistry and biology held in the years 1999, 2006, 2011, and 2013. Comments by the students of all four courses were very helpful in the preparation of this text and are gratefully acknowledged. Several colleagues gave important advice and critically read the manuscript, among them Christian Höner zu Siederissen, Paul E. Phillipson, and Karl Sigmund. Many thanks to all of them.

Wien, *Peter Schuster*
March 2013

Contents

Chapter 1
Probability

Who considers too much will achieve little.
Wer gar zu viel bedenkt, wird wenig leisten.
Friedrich Schiller, Wilhelm Tell, III.

Abstract . Thinking in terms of probability originated historically from analyzing the chances of success in gambling and its mathematical foundations were laid down together with the development of statistics in the seventeenth century. Since the beginning of the twentieth century statistics is an indispensable tool for bridging the gap between molecular motions and macroscopic observations. The classical notion of probability is based on counting and dealing with finite numbers of observations, the extrapolation to limiting values for hypothetical infinite numbers of observations is the basis of the frequentists' interpretation, and more recently a *subjective* approach derived from the early works of Bayes became useful in modeling and analyzing complex biological systems. The Bayesian interpretation of probability accounts explicitly for incomplete and improvable knowledge of the experimenter. In the twentieth century set theory became the ultimate basis of mathematics and in this sense it became also the fundament of current probability theory that is based on Kolmogorov's axiomatization in 1933. The modern approach allows for handling and comparing countable, countable infinite and the most important class of uncountable sets, which are underlying continuous variables. Borel fields being uncountable subsets of sample spaces allow for defining probabilities for certain uncountable sets like, for example, the real numbers. The notion of random variables is central to the analysis of probabilities and applications to problem solving. Random variables are characterized conventionally in form of their distributions in discrete and countable or continuous and uncountable probability spaces.

Classical probability theory, in essence, can handle all cases that are modeled by discrete quantities. It is based on counting and accordingly runs into problems when it is applied to uncountable sets. Uncountable sets, however, occur with continuous variables and are indispensable therefore for modeling processes in space as well as for handling large particle numbers, which are described in terms of concentrations in chemical kinetics. Current probability theory is based on set theory and can handle variables on discrete – and

countable – as well as continuous – and uncountable – sets. After a general introduction we present historical probability theory by means of examples, different notions of probability are compared, and then we provide a short account of probabilities, which are axiomatically derived from set theoretical operations. Separate sections are dealing with countable and uncountable sample spaces. Random variables are characterized in terms of probability distributions and their properties will be introduced and analyzed insofar as they will be required in the applications to stochastic processes.

1.1 Precision limits and fluctuations

An scientist reproduces an experiment. What is he expecting to observe? If he were a physicist of the early nineteenth century he would expect the same results within the precision limits of the apparatus he is using for the measurement. Uncertainty in observations was considered to be merely a consequence of technical imperfection. Celestial mechanics comes close to this ideal and many of us, for example, could witness the enormous accuracy of astronomical predictions in the precise dating of the eclipse of the sun in Europe on August 11, 1999. Terrestrial reality, however, tells that there are limits to reproducibility that have nothing to do with lack of experimental perfection. Uncontrollable variations in initial and environmental conditions on one hand and large intrinsic diversity of the individuals in a population on the other hand are daily problems in biology. Limitations of correct predictions are commonplace in complex systems: We witness them every day by watching the failures of various forecasts from the weather to the stock market. Another not less important source of randomness comes from irregular thermal motions of atoms and molecules that are commonly characterized as thermal fluctuations. The importance of fluctuations in the description of ensembles depends on the population size: They are – apart from exceptions – of moderate importance in chemical reaction kinetics but highly relevant for the evolution of populations in biology.

Conventional chemical kinetics is handling ensembles of molecules with large numbers of particles, $N \approx 10^{20}$ and more. Under the majority of common conditions, for example near or at chemical equilibria and stable stationary states, random fluctuations in particle numbers are proportional to $\sqrt{N}$. Dealing with substance amounts of about 10^{-4} moles – being tantamount to $N = 10^{20}$ particles – natural fluctuations involve typically $\sqrt{N} = 10^{10}$ particles and thus are in the range of $\pm 10^{-10} N$. Under these conditions the detection of fluctuations would require a precision in the order of $1 : 10^{10}$, which is (almost always) impossible to achieve.[1] Accordingly, the chemist uses concentrations rather than particle numbers, $c = N/(N_L \times V)$ wherein $N_L = 6.23 \times 10^{23}$ and V are Avogadro's number[2] and the volume in dm^3, respectively. Conventional chemical kinetics considers concentrations as continuous variables and applies deterministic methods, in essence differential

[1] Most techniques of analytical chemistry meet serious difficulties when accuracies in concentrations of 10^{-6} or higher are required.

[2] The amount of a chemical compound **A** is commonly measured as the number of molecules, N_A, in the reaction volume V or in solution as concentrations c_X being the numbers of moles in one liter of solution, $c_A = (N_A/N_\mathrm{L})/V$ where $N_\mathrm{L} = 6.023 \times 10^{23}$ is Avogadro's or Loschmidt's number. As a matter of fact there is a difference between the two numbers that is often ignored in the literature: Avogadro's number, $N_\mathrm{L} = 6.02214179 \times 10^{23}\ \mathrm{mole}^{-1}$ refers to one mole substance whereas Loschmidt's constant $n_0 = 2.6867774 \times 10^{25}\ \mathrm{m}^{-3}$ counts the number of particles in one liter gas under normal conditions. The conversion factor between both constants is the molar volume of an ideal gas that amounts to $22.414\ \mathrm{dm}^3 \cdot \mathrm{mole}^{-1}$.

equations, for modeling and analysis of reactions. Thereby, it is implicitly assumed that particle numbers are sufficiently large that the limit of infinite particle numbers neglecting fluctuations is correct. This scenario is commonly not fulfilled in biology where particle numbers are much smaller than in chemistry.

Nonlinearities in chemical kinetics may amplify fluctuations through autocatalysis and then the random component becomes much more important than the $\sqrt{N}$-law suggests. This is the case, for example, with oscillating concentrations or deterministic chaos. Some processes in physics, chemistry, and biology have no deterministic component at all, the most famous of it is Brownian motion,*Brownian motion* which can be understood as a visualized form of diffusion. In biology other forms of entirely random processes are encountered where fluctuations are the only or the major driving force of change. An important example is random drift of population in the space of genotypes in absence of fitness differences or fixation of mutants in evolution where each new molecular species starts out from a single variant.

In 1827 the British botanist Robert Brown detected and analyzed irregular motions of particles in aqueous suspensions that turned out to be independent of the nature of the suspended materials – pollen grains, fine particles of glass or minerals [27]. Although Brown himself had already demonstrated that the motion is not caused by some (mysterious) biological effect, its origin remained kind of a riddle until Albert Einstein [58], and independently Marian von Smoluchowski [298], published satisfactory explanations in 1905 and 1906,[3] which revealed two main points:

(i) The motion is caused by highly frequent collisions between the pollen grain and steadily moving molecules in the liquid in which it is suspended, and
(ii) the motion of the molecules in the liquid is so complicated and irregular that its effect on the pollen grain can only be described probabilistically in terms of frequent, statistically independent impacts.

In order to model Brownian motion Einstein considered the number of particles per volume as a function of space and time, $f(x,t) = N(x,t)/V$,[4] and derived the equation

$$\frac{\partial f}{\partial t} \;=\; D\frac{\partial^2 f}{\partial x^2} \quad \text{with the solution} \quad f(x,t) \;=\; \frac{\varrho}{\sqrt{4\pi D}}\,\frac{\exp\Bigl(-x^2/(4Dt)\Bigr)}{\sqrt{t}}\;,$$

where $\varrho = N/V = \int f(x,t)\,\mathrm{d}x$ is the total number of particles per unit volume and D is a parameter called the *diffusion coefficient*. Einstein showed

[3] The first mathematical model of Brownian motion has been conceived already in 1880 by Thorvald Thiele [175, 276]. Later in 1900 a process using random fluctuations of the Brownian motion type was used by Louis Bachelier [10] in order to describe the stock exchange market at the bourse in Paris.

[4] For the sake of simplicity we consider only motion in one spatial direction, x.

that his equation for $f(x,t)$ is identical with the already known differential equation of diffusion [78], which had been derived fifty years earlier by the German physiologist Adolf Fick. Einstein's original treatment is based on small discrete time steps $\Delta t = \tau$ and thus contains a – well justifiable – approximation that can be avoided by application of the current theory of stochastic processes (section 3.2.3.2). Nevertheless Einstein's publication [58] represents the first analysis based on a probabilistic concept that is by all means comparable to the current theories and Einstein's paper is correctly considered as the beginning of stochastic modeling. Later Einstein wrote four more papers on diffusion with different derivations of the diffusion equation [59]. It is worth mentioning that three years after the publication of Einstein's first paper Paul Langevin presented an alternative mathematical treatment of random motion [173] that we shall discuss at length in the form of the Langevin equation in section 3.4. Since the days of Brown's discovery the interest in Brownian motion has never ceased and publications on recent theoretical and experimental advances document this fact nicely, two interesting examples are [178, 257].

The diffusion parameter D is linked to the mean square displacement that the particle experiences in the x-direction during time t – or its square root λ_x – as Einstein computed from the solution of the diffusion equation:

$$D = \frac{\left\langle \Delta x^2 \right\rangle}{2t} \quad \text{and} \quad \lambda_x \;=\; \sqrt{\bar{x}^2} \;=\; \sqrt{2Dt}\;.$$

Extension to three dimensional space is straightforward and results only in a different numerical factor: $D = \left\langle \Delta x^2 \right\rangle / (6t)$. Both quantities, the diffusion parameter D and the mean displacement λ_x are measurable and Einstein concluded correctly that a comparison of both quantities should allow for an experimental determination of Avogadro's number [239].

Brownian motion was indeed the first completely random process that became accessible to a description within the standards of classical physics. Previously, thermal motion had been identified as the irregular driving force causing collisions of molecules in gases by James Clerk Maxwell and Ludwig Boltzmann but the physicists in the second half of the nineteenth century were not interested in any details of molecular motion unless they were required in order to describe systems in the thermodynamic limit. The desired measurable macroscopic functions were derived the by means of the global averaging techniques of statistical mechanics. Thermal motion as an uncontrollable source of random natural fluctuations has been supplemented by quantum mechanical uncertainty as another limitation of achievable precision in the first half of the twentieth century. Here we shall focus on the mathematical handling of processes that are irregular and often sensitive to small changes, and we shall not be concerned so much with the origin of these irregularities.

Computer assisted analysis of complex dynamical systems was initiated in essence by Edward Lorenz [186] who detected through numerical integration of differential equations what is nowadays called *deterministic chaos*. Complex dynamics in physics and chemistry has been known already much earlier as the works of the French mathematician Henri Pioncaré and the German chemist Wilhelm Ostwald demonstrate. New in the second half of the twentieth century were not the ideas but the tools to study complex dynamics. Quite unexpectedly, easy access to previously unknown computer power and the development of highly efficient algorithms made numerical computation to an indispensable source of scientific information that by now became almost equivalent to theory and experiment. Computer simulations have shown that a large class of dynamical systems modeled by nonlinear differential equations show irregular – that means nonperiodic – variation for certain ranges of parameter values. In these *chaotic regimes* solutions curves were found to be extremely sensitive to small changes in the initial and boundary conditions. Solution curves, which are almost identical at the beginning deviate exponentially from each other and are completely different after sufficiently long time. Thereby they give rise to a kind of deterministic uncertainty. Limitations in the control of initial conditions are inevitable, because any achievable experimental precision is finite, and their consequences are upper bounds for the time spans for which the dynamics of the system can be predicted with sufficient accuracy. It is not accidental that Lorenz detected chaotic dynamics first in the equations for atmospheric motions, which are indeed so complex that forecast is confined to short and medium time spans. Limited predictability of complex dynamics is of a highly important practical nature: Although the differential equations used to describe and analyze chaos are still deterministic, initial conditions of a precision that can never be achieved in reality would be required for correct longtime predictions. Sensitivity to small changes makes a stochastic treatment indispensable.

1.2 The history of thinking in terms of probability

The concept of probability originated from the desire to analyze gambling by rigorous mathematical methods. An early study that has largely remained unnoticed but contained already the basic ideas of probability was done in the sixteenth century by the Italian mathematician Gerolamo Cardano and the beginning of classical probability theory is commonly associated with the story of French mathematician Blaise Pascal and the professional gambler, the Chevalier de Méré, which took place in France 100 years after Cardano and which is found in almost every introduction to probability theory.

In a letter of July 29, 1654, which was addressed to the French mathematician Pierre de Fermat, Blaise Pascal, reports the careful observation of the professional gambler Chevalier de Méré who recognized that obtaining at least one *six* with one die in 4 throws is successful in more than 50% whereas obtaining at least two times the "six" with two dice in 24 throws has less than 50% chance to win. He considered this finding as a paradox because he calculated naïvely and erroneously that the chances should be the same:

$$\text{4 throws with one die yields} \quad 4 \times \frac{1}{6} = \frac{2}{3} ,$$

$$\text{24 throws with two dice yields} \quad 24 \times \frac{1}{36} = \frac{2}{3} .$$

Blaise Pascal became interested in the problem and calculated correctly the probability as we do it now in classical probability theory by counting of events:

$$\text{probability} \;=\; \text{Prob} \;=\; \frac{\text{number of favorable events}}{\text{total number of events}} . \qquad (1.1)$$

Probability according to equation (1.1) is always a positive quantity between zero and one, $0 \leq \text{Prob} \leq 1$. The sum of the probabilities that a given event has either occurred or did not occur thus is always one. Sometimes, as in Pascal's example, it is easier to calculate the probability of the unfavorable case, q, and to obtain the desired probability as $p = 1 - q$. In the one die example the probability not to throw a *six* is 5/6, in the two dice example we have 35/36 as the probability of failure. In case of independent events probabilities are multiplied[5] and we finally obtain for 4 and 24 trials, respectively:

$$q(1) = \left(\frac{5}{6}\right)^4 \quad \text{and} \quad p(1) = 1 - \left(\frac{5}{6}\right)^4 = 0.5177 ,$$

$$q(2) = \left(\frac{35}{36}\right)^{24} \quad \text{and} \quad p(2) = 1 - \left(\frac{35}{36}\right)^{24} = 0.4914 .$$

[5] We shall come back to a precise definition of independent events later when we introduce current probability theory in section 1.

It is remarkable that the gambler could observe this rather small difference in the probability of success – apparently, he must have tried the game very often indeed!

The second example presented here is the *birthday problem*.[6] It can be used to demonstrate the common human weakness in estimating probabilities:

> *"Let your friends guess – without calculating – how many persons you need in a group such that there is a fifty percent chance that at least two of them celebrate their birthday on the same day. You will be surprised by the oddness of some of the answers!"*

With our knowledge on the gambling problem this probability is easy to calculate. First we compute the negative event: all persons celebrate their birthdays on different days in the year – 365 days, no leap-year – and find for n people in the group,[7]

$$q = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365-(n-1)}{365} \quad \text{and} \quad p = 1 - q \; .$$

The function $p(n)$ is shown in figure 1.1. For the above mentioned 50% chance we need only 27 persons, with 41 people we have already more than 90% chance that two celebrate birthday one the same day; 57 yield more than 99% and 70 persons exceed 99,9%. An implicit assumption in this calculation has been that births are uniformly distributed over the year or, in other words, the probability that somebody has the birthday on some day does not depend on the particular day of the year. In mathematical statistics such an assumption is called a *null hypothesis* (see [85] and section 2.5.2).

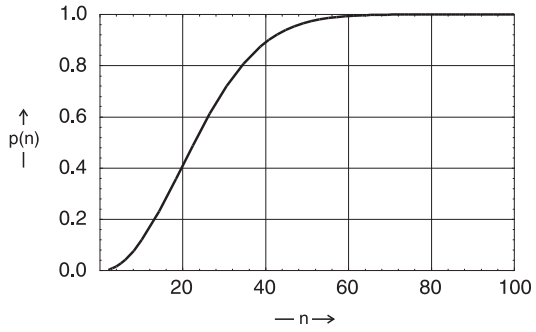

Fig. 1.1 The birthday problem. The curve shows the probability $p(n)$ that two persons in a group of n people celebrate birthday on the same day of the year.

[6] The birthday problem has been invented in 1939 by Richard von Mises [297] and is fascinating mathematicians ever since. It has been discussed and extended in many papers (for example [2, 39, 130, 228] and found its way into textbooks on probability [76, pp. 31-33].

[7] The expressions is obtained by the argument that the first person can choose his birthday freely. The second person must not choose the same day and so he has 364 possible choices. For the third remain 363 choices and the nth person, ultimately, has $365 - (n - 1)$ possibilities.

Table 1.1 Advantage of the second player in Penney's game. Two players choose two triples of digits one after the other, player 2 after player 1. Coin flipping is played until the two triples appear. The player whose triple came first has won. An optimally gambling player 2 (column 2) has the advantage shown in column 3. Code: 1 = *head* and 0 = *tail*. The optimal strategy for player 2 is encoded by grey and boldface (see text).

Player's choice		Outcome
Player 1	Player 2	Odds in favor of player 2
111	**0**11	7 to 1
110	**0**11	3 to 1
101	**1**10	2 to 1
100	**1**10	2 to 1
011	**0**01	2 to 1
010	**0**01	2 to 1
001	**1**00	3 to 1
000	**1**00	7 to 1

The third example deals again with events that occur with counterintuitive probabilities: the coin toss game *Penney Ante* invented by Walter Penney [238]. Before a sufficiently long sequence of *heads* and *tails* is determined through flipping coins, each of two players chooses a sequence of n consecutive flips – commonly $n = 3$ is applied and this leaves the choice of the eight triples shown in table 1.1. The second player has the advantage of knowing the choice of the first player. Then the sequence of coins flips is recorded until both of the chosen triples have appeared in the sequence. The player whose sequence appeared first has won. The advantage of the second player is commonly largely underestimated when guessed without explicit calculation. A simple argument illustrates the disadvantage of player 1: Assume he had chosen '111'. If the second player chooses a triple starting with '0' the only chances for player 1 to win are expressed by the sequences beginning '111... and they have a probability of p=1/8 leading to the odds 7 to 1 for player 2. Eventually, we mention the optimal strategy for player 2: Take the first two digits of the three-bit sequence that player 1 had chosen and precede it with the opposite of the symbol in the middle of the triple (In table 1.1 the shifted pair is shown in grey, the switched symbol in bold).

Laws in classical physics are considered as deterministic in the sense that a single measurement is expected to yield a precise result, deviations from which are interpreted as lack in precision of the used machinery. Random scatter when it is observed is thought to be caused by variation in not sufficiently well controlled experimental conditions. Apart from deterministic laws other regularities are observed in nature, which become evident only when sample

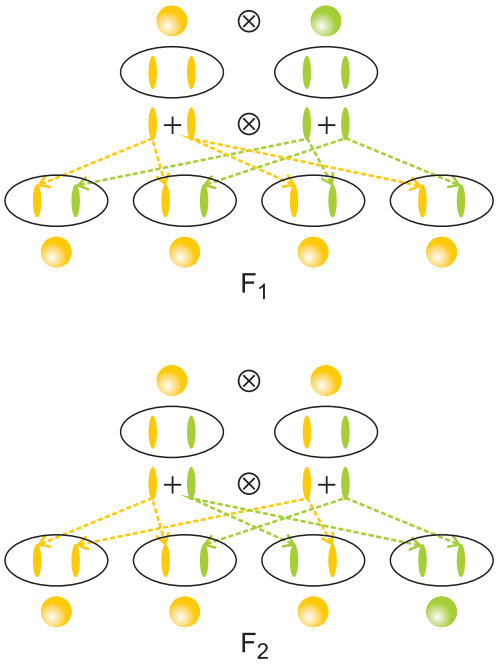

Fig. 1.2 Mendel'l laws of inheritance. The sketch illustrates Mendel's laws of inheritance: (i) the law of segregation and (ii) the law of independent assortment. Every (diploid) organism carried two copies of each gene, which are separated during the process of reproduction. Every offspring receives one randomly chosen copy of the gene from each parent. Encircled are the genotypes formed from two alleles, yellow or green, and above or below the genotypes are the phenotypes expressed as the colors of seeds of the garden pea (*pisum sativum*). The upper part of the figure shows the first generation (F_1) of progeny of two homozygous parents – parents who carry two identical alleles. All genotypes are heterozygous and carry one copy of each allele. The yellow allele is dominant and hence the phenotype expresses yellow color. Crossing two F_1 individuals (lower part of the figure) leads to two homozygous and two heterozygous offspring. Dominance causes the two heterozygous genotypes and one homozygote to develop the dominant phenotype and accordingly the observable ratio of the two phenotypes in the F_2 generation is 3:1 on the average as observed by Gregor Mendel in his statistics of fertilization experiments (see table 1.2).

sizes are made sufficiently large through repetition of experiments. It is appropriate to call such regularities *statistical laws*. Statistics in biology of plant inheritance has been pioneered by the Augustinian monk Gregor Mendel who discovered regularities in the progeny of the *garden pea* in controlled fertilization experiments [210] (figure 1.2). As a fourth and final example we consider some of Mendel's data in order to illustrate a statistical law. In table 1.2 the results of two typical experiments distinguishing roundish or wrinkled seeds

of yellow or green color are listed. The ratios observed with single experiments plants large scatter. In the mean values for ten plants some averaging has occurred but still the deviations from the ideal values are recognizable. Mendel carefully investigated several hundred plants and then the statistical law of inheritance demanding a ratio of 3:1 became evident [210].[8] Ronald Fisher in a somewhat polemic publication [84] reanalyzed Mendel's experiments, questioned Mendel's statistics, and accused him of intentionally manipulating his data because the results are too close to the ideal ratio. Fisher's publication initiated a long lasting debate during which many scientists spoke up in favor of Mendel [226, 227] but there were also critical voices saying that most likely Mendel has unconsciously or consciously eliminated extreme outliers [53]. In 2008 a recent book declared *the end of the Mendel-Fisher controversy* [89]. In section 2.5.2 we shall discuss statistical laws and Mendel's experiments in the light of present day mathematical statistics by applying the χ^2 test.

Probability theory in its classical form is more than three hundred years old. Not accidentally the concept arose in thinking about gambling, which was considered as a domain of chance in contrast to rigorous science. It took indeed rather long time before the concept of probability entered scientific

Table 1.2 Statistics of Gregor Mendel's experiments with the garden pea (*pisum sativum*)**.** The results of two typical experiments with ten plants are shown. In total Mendel analyzed 7324 seeds from 253 hybrid plants in the second trial year, 5474 were round or roundish and 1850 angular wrinkled yielding a ratio 2.96:1. The color was recorded for 8023 seeds from 258 plants out of which 6022 were yellow and 2001 were green with a ratio of 3.01:1.

	Form of seed			Color of seeds		
plants	round	wrinkled	ratio	yellow	green	ratio
1	45	12	3.75	25	11	2.27
2	27	8	3.38	32	7	4.57
3	24	7	3.43	14	5	2.80
4	19	10	1.90	70	27	2.59
5	32	11	2.91	24	13	1.85
6	26	6	4.33	20	6	3.33
7	88	24	3.67	32	13	2.46
8	22	10	2.20	44	9	4.89
9	28	6	4.67	50	14	3.57
10	25	7	3.57	44	18	2.44
total	336	101	**3.33**	355	123	**2.89**

[8] According to modern genetics this ratio as well as other inter ratios are idealized values that are found only for completely independent genes [111], which lie either on different chromosomes or sufficiently far apart on the same chromosome.

thought in the nineteenth century. The main obstacle for the acceptance of probabilities in physics was the strong belief in determinism that has not been overcome before the advent of quantum theory. Probabilistic concepts in physics of the nineteenth century were still based on deterministic thinking, although the details of individual events were considered to be too numerous to be accessible to calculation at the microscopic level. It is worth mentioning that thinking in terms of probabilities entered biology earlier, already in the second half of the nineteenth century through the reported works on the genetics of inheritance by Gregor Mendel and the considerations about pedigrees by Francis Galton (see section 5.2.1). The reason for this difference appears to lie in the very nature of biology: Small sample sizes are typical, most of the regularities are probabilistic and become observable only through the application of probability theory. Ironically, Mendel's investigations and papers did not attract a broad scientific audience before their *rediscovery* at the beginning of the twentieth century. The scientific community in the second half of the nineteenth century was simply not yet prepared for the acceptance of quantitative and moreover probabilistic concepts in biology.

Classical probability theory is dealing successfully with a number of concepts like conditional probabilities, probability distributions, moments and others, which shall be presented in the next section making use of set theoretic concepts that can provide much deeper insight into the structure of probability theory. In addition, the more elaborate notion of probability derived from set theory is absolutely necessary for extrapolation to infinitely large and uncountable sample sizes. From now on we shall use only the set theoretic concept, because it can be introduced straightforwardly for countable sets and discrete variables and, in addition, it can be extended to probability measures for continuous variables where numbers of sample points are not only infinite but also uncountable. In this way real numbers, $\mathbf{x} \in \mathbb{R}^n$, become accessible to probability measures.

1.3 Probability and interpretations

Before a introduction to the currently most popular theory of probability is presented we make a digression into some major philosophical interpretations of probability: (i) the classical interpretations that we adopted in chapter 1.2, (ii) the frequency-based interpretation that will be in the background of the rest of the book, and (iii) the Bayesian or *subjective* interpretation.

The *classical interpretation of probability* goes back to the concepts and works of the Swiss mathematician Jakob Bernoulli and the French mathematician and physicist Pierre-Simon Laplace, who first presented a clear definition of probability [174, pp. 6-7]:

> *"The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all cases possible."*

Clearly, this definition is tantamount to equation (1.1) and the explicitly stated assumption of equal probabilities is now called *principle of indifference*. This classical definition of probability has been questioned during the nineteenth century among others by the two English logicians and philosophers George Boole [23] and John Venn [291], who among others initiated a paradigm shift from the classical view to the modern frequency interpretations of probabilities.

The modern interpretations of the concept of probabilities fall essentially into two categories that can be characterized as *physical probabilities* and *evidential probabilities* [115]. Physical probabilities are often called *objective* or frequency-based probabilities and their proponents are often addressed as *frequentists*. Influential proponents of the frequency-based probability theory were, besides the pioneer John Venn, the Polish American mathematician Jerzy Neyman, the English statistician Egon Pearson, the English statistician and theoretical biologist Ronald Fisher, the Austro-Hungarian American mathematician and scientist Richard von Mises and the German American philosopher of science Hans Reichenbach. The physical probabilities are derived from some real process like radioactive decay, chemical reaction, turning a roulette wheel, or rolling dice. In all such systems the notion of probability makes sense only when it refers to some well defined experiment with a random component. Frequentism comes in two versions: (i) *finite frequentism* and (ii) *hypothetical frequentism*. Finite frequentism replaces the notion of *"total number of events"* in equation (1.1) by *"actually recorded number of events"* and is thus congenial to philosophers with empiricist scruples. Philosophers have a number of problems with finite frequentism, we mention for example

the small sample problems: One can never speak about the probability of a single experiment and there are cases of unrepeated and unrepeatable experiments. A coin that is tossed exactly once yields a relative frequency of heads of zero or one, no matter what its bias really is. Another famous example is the spontaneous radioactive decay of an atom where the probabilities of decaying follow a continuous exponential law but according to finite frequentism it decays with probability one at its actual decay time. The evolution of the universe or the origin of life can serve as cases of unrepeatable experiments, but people like to speak about the probability that the development has been such or such (Personally, I think it would do no harm to replace *probability* by *plausibility* in such estimates concerned with unrepeatable single events).

Hypothetical frequentism complements the empiricism of finite frequentism by the admission of infinite sequences of trials. Let N be the total number of repetitions of an experiment and n_A the number of trials when the event A has been observed, then the relative frequency of recording the event A is an *approximation of the probability* for the occurrence of A:

$$\text{Prob}\,(A) \;=\; P(A) \;\approx\; \frac{n_A}{N}\;.$$

This equation is essentially the same as (1.1) but the claim of the hypothetical frequentist interpretation is that there exists a *true frequency* or *true probability* to which the relative frequency converges when we repeat the experiment an infinite number of times[9]

$$P(A) \;=\; \lim_{N\to\infty}\frac{n_A}{N} \;=\; \frac{|A|}{|\Omega|} \quad \text{with} \quad A\in\Omega. \tag{1.2}$$

The probability of an event A relative to a sample space Ω is then defined as the limiting frequency of A in Ω. As N goes to infinity $|\Omega|$ becomes infinitely large and depending on whether $|A|$ is finite or infinite $P(A)$ is either zero or may adopt a nonzero limiting frequency. It is based on two a priori assumptions that have the character of axioms:

(i) *Convergence:* For any event A exists a limiting relative frequency, the probability $P(A)$ that fulfils $0 \le P(A) \le 1$.
(ii) *Randomness:* The limiting relative frequency of each event in a collective Ω is the same in any *typical* infinite subsequence of Ω.

A typical sequence is *sufficiently random*[10] in order to avoid results biased by predetermined order. As a negative example of an acceptable sequence we consider *"head, head, head, head, ..."* recorded by tossing a coin. If it was obtained with a fair coin – not with a coin with two heads – $|A|$ is 1 and

[9] The absolute value symbol, '$|\cdot|$', means size of n cardinality of or number of elements in a set (section 1.4).

[10] Sequences are sufficiently random when they are obtained through recordings of random events. *Random sequences* are approximated by the sequential outputs of *random number generators*.

$P(A) = 1/|\Omega| = 0$ and we may say this particular events is of measure zero and the sequence is not typical. The sequence *"head, tail, head, tail, ..."* is not typical as well despite the fact that it yields the same result as a fair coin. We should be aware that the extension to infinite series of experiments leaves the realm of empiricism and caused philosophers to reject the claim that the interpretation of probabilities by hypothetical frequentism is more *objective* than others.

Nevertheless, frequentist probability theory is not in conflict with the mathematical axiomatization of probability theory and it provides straightforward guidance in applications to real-world problem. The pragmatic view that stands at the beginning of the dominant concept in current probability theory has been phrased nicely by William Feller, the Croatian-American mathematician and author of the classic introduction to probability theory in two volumes [76, 77, Vol.I, pp. 4-5]:

> *"The success of the modern mathematical theory of probability is bought at a price: the theory is limited to one particular aspect of 'chance'. ... we are not concerned with modes of inductive reasoning but with something that might be called physical or statistical probability."*

He also expresses clearly his attitude towards pedantic scruples of philosophic purists:

> *"..., in analyzing the coin tossing game we are not concerned with the accidental circumstances of an actual experiment, the object of our theory is sequences or arrangements of symbols such as 'head, head, tail, head, ...'. There is no place in our system for speculations concerning the probability that the sun will rise tomorrow. Before speaking of it we should have to agree on an idealized model which would presumably run along the lines 'out of infinitely many worlds one is selected at random ...'. Little imagination is required to construct such a model, but it appears both uninteresting and meaningless."*

We shall adopt the frequentist interpretation throughout this monograph but mention here briefly a few other interpretations of probability in order to show that it is not the only reasonable probability theory.

The *propensity interpretation* of probability was proposed by the American philosopher Charles Peirce in 1910 [237] and reinvented by Karl Popper [242, pp. 65-70] (see also [243]) more than forty years later [115, 215]. *Propensity* is a tendency to do or to achieve something, and in relation to probability propensity means that it makes sense to talk about the probabilities of single events. As an example we mention the probability – propensity – of a radioactive atom to decay within the next one thousand years, and thereby we make a conclusion from the behavior of an ensemble to a single member of the ensemble. For a fair coin we might say that it has a probability of $\frac{1}{2}$ to score "head" when tossed, and precisely expressed we should say that the coin has the propensity to yield a sequence of outcomes, in which the limiting

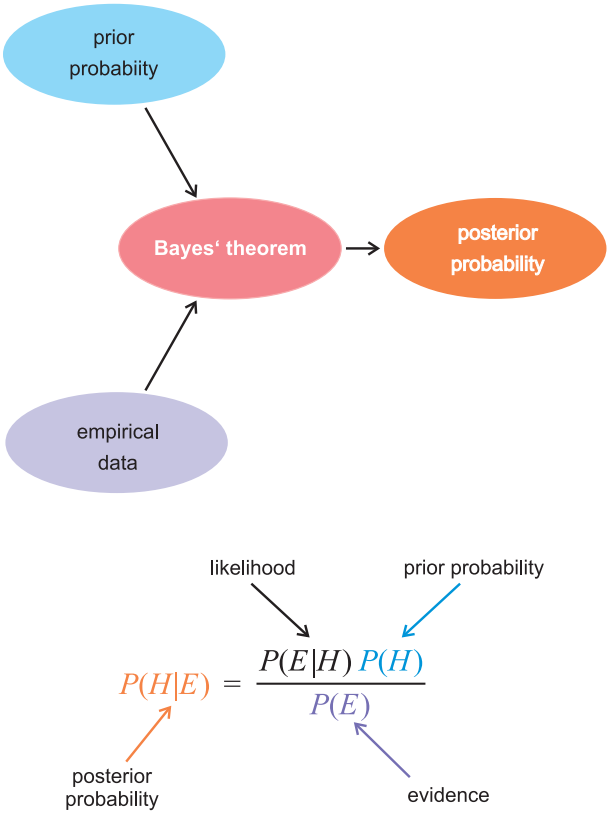

Fig. 1.3 A sketch of the Bayesian method. Prior information of probabilities is confronted with empirical data and converted into a new distribution of probabilities by means of Bayes' theorem according to the formula shown above [52, 268].

frequency of scoring "heads" is $\frac{1}{2}$. The single case propensity is accompanied by, but distinguished from, the *log-run* propensity [107]:

> *"A long-run propensity theory is one in which propensities are associated with repeatable conditions, and are regarded as propensities to produce in a long series of repetitions of these conditions frequencies, which are approximately equal to the probabilities.*

Long-run in these theories is still distinct from infinitely long run in order to avoid basic philosophical problems. As it looks, the use of *propensities* rather than *frequencies* constitutes a language that is somewhat more careful and hence more acceptable in philosophy than the frequentist interpretation.

Finally, we sketch the most popular example of a theory-based on *evidential probabilities*: Bayesian statistics, named after the eighteenth century

English mathematician and Presbyterian minister Thomas Bayes. In contrast to the frequentists' view probabilities are *subjective* and exist only in the human mind. From a practitioner's point of view one major advantage of the Bayesian approach is the direct insight into the process of improving the knowledge on the object of investigation. In order to understand Bayes' theorem we need the notion of conditional probabilities (for a precise definition see section 1.6.3): For a conditional probability the reference ensemble is not the entire sample space Ω but some event, say B. Then, we have

$$P(A|B) \;=\; \frac{P(A \text{ and } B)}{P(B)} \;=\; \frac{P(AB)}{P(B)}\,, \tag{1.3}$$

where "A and B" indicates the joint probability of both events A and B.[11] The conditional probability $P(A|B)$ is obtained as the probability of the simultaneous occurrence of events A and B divided by the probability of the occurrence of B alone. If the event B is the entire sample space, $B \equiv \Omega$ we obtain:

$$P(A|\Omega) \;=\; \frac{P(A \text{ and } \Omega)}{P(\Omega)} \;=\; \frac{P(A\Omega)}{P(\Omega)} \;=\; \frac{P(A)}{1} \;=\; P(A)\,,$$

the conditional probability is equal to the unconditioned probability. Conditional probabilities can be inverted straightforwardly in the sense that we ask about the probability of B under the condition that event A has occurred:

$$P(B|A) \;=\; \frac{P(A \text{ and } B)}{P(A)} \;=\; \frac{P(AB)}{P(A)} \;\text{ since }\; P(AB) \;=\; P(BA)\,, \tag{1.3'}$$

which implies $P(A|B) \neq P(B|A)$ unless $P(A) = P(B)$. In other words the conditional probability can be readily inverted, and as expected $P(A|B)$ and $P(B|A)$ are on equal footing in probability theory. Calculation of $P(AB)$ from both equations, (1.3) and (1.3'), and setting the expressions equal yields

$$P(A|B)\,P(B) \;=\; P(AB) \;=\; P(B|A)\,P(A) \;\Longrightarrow\; (P(B|A) \;=\; P(A|B)\,\frac{P(B)}{P(A)}\,,$$

which properly interpreted represents Bayes' theorem.

Bayes'theorem provides a straightforward interpretation of conditional probabilities and their inversion in terms of models or hypothesis (H) and data (E). The conditional probability $P(E|H)$ corresponds to the conventional procedure in science: Given a set of hypothesis cast into a model H the task is to calculate the probabilities of the different outcomes E. In physics and chemistry where we are dealing with well established theories and models this is, in essence, the common situation. Biology, economics, social sciences

[11] From the next section 1.4 on we shall use the set theoretic symbol intersection, "$\cap$", instead of "and"; AB is an abbreviated notation for "A and B".

and other disciplines, however, are often confronted with situations where no confirmed models exist and then we want to test and improve the probability of a model. We need to invert the conditional probability since we are interested in testing the model in the light of the data available or, in other words, the conditional probability $P(H|E)$ becomes important: What is the probability that a hypothesis H is justified given a set of measured data (evidence E)? The Bayesian approach casts equations (1.3) and (1.3') into Bayes' theorem,

$$P(H|E) \;=\; P(E|H)\,\frac{P(H)}{P(E)} \;=\; \frac{P(E|H)}{P(E)}\cdot P(H)\;, \tag{1.4}$$

and provides a hint on how to proceed – at least in principle (figure 1.3). An *prior probability* in form of a hypothesis $P(H)$ is converted into evidence according to the likelihood principle $P(E|H)$. The basis of the prior understood as all *a priori* knowledge comes form many sources: theory, previous experiments, gut feeling, etc. New empirical information is incorporated in the *inverse probability* computation from data to model, $P(H|E)$, yielding thereby the improved *posterior probability*. The advantage of the Bayesian approach is that a change of opinion in the light of new data is *"part of the game"*. In general, parameters are input quantities of frequentist statistics and if unknown assumed to be available through consecutive repetition of experiments, whereas they are understood as random variables in the Bayesian approach. The direct application of the Bayesian theorem in practice involves quite elaborate computations that were not possible in real world examples before the advent of electronic computers. An example of the Bayesian approach and the calculations involved thereby is presented in section 2.5.4.

Bayesian statistics has become popular in disciplines where model building is a major issue. Examples from biology are among others bioinformatics, molecular genetics, modeling of ecosystems, and forensics. Bayesian statistics is described in a large number of monographs, for example, in references [42, 95, 147, 176].

1.4 Sets and sample spaces

Conventional probability theory is based on several axioms that are rooted in set theory, which will be introduced and illustrated in this section. The development of set theory in the eighteen seventieth was initiated by Georg Cantor and Richard Dedekind and provided the possibility to build among many other things the concept of probability on a firm basis that allows for an extension to certain families of uncountable samples as they occur, for example, with continuous variables. Present day probability theory thus can be understood as a convenient extension of the classical concept by means of set and measure theory. We begin by repeating a few indispensable notions and operations of set theory.

Sets are collections of objects with two restrictions: (i) Each object belongs to one set cannot be a member of two or more sets, and (ii) a member of a set must not appear twice or more often. In other words, objects are assigned to sets unambiguously. In the application to probability theory we shall denote the *elementary objects* by the small Greek letter *omega*, ω – if necessary with various sub- and superscripts – and call them *sample points* or *individual results*. The collection of all objects ω under consideration, the *sample space*, is denoted by Ω with $\omega \in \Omega$. *Events*, A, are subsets of sample points that fulfil some condition[12]

$$A \;=\; \big\{\boldsymbol{\omega}, \omega_k \in \Omega :\; f(\boldsymbol{\omega}) = c\big\} \tag{1.5}$$

with $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots)$ being the set of individual results which fulfil the condition $f(\boldsymbol{\omega}) = c$.

Next we repeat the basic logical operations with sets. Any partial collection of points $\omega_k \in \Omega$ is a *subset* of Ω. We shall be dealing with fixed Ω and, for simplicity, often call these subsets of Ω just sets. There are two extreme cases, the entire sample space Ω and the *empty set*, $\emptyset$. The number of points in a set S is called its size or *cardinality* written as $|S|$, and thus $|S|$ is a nonnegative integer or infinity. In particular, the size of the empty set is $|\emptyset| = 0$. The unambiguous assignment of points to sets can be expressed by[13]

$$\omega \in S \quad \text{exclusive or} \quad \omega \notin S \;.$$

Consider two sets A and B. If every point of A belongs to B, then A is contained in B. A is a subset of B and B is a superset of A:

---

[12] The meaning of *condition* will become clearer later on. For the moment it is sufficient to understand a condition as a restriction cast in a function $f(\boldsymbol{\omega})$, which implies that not all subsets of sample points belong to A. Such a condition, for example, is a score '6' in rolling two dice, which comprises the five sample points: $A = \{\text{'}1+5\text{'},\text{'}2+4\text{'},\text{'}3+3\text{'},\text{'}4+2\text{'},\text{'}5+1\text{'}\}$.

[13] In order to be unambiguously clear we shall write *or* for *and/or* and *exclusive or* for *or* in the strict sense.

$$A \subset B \quad \text{and} \quad B \supset A \,.$$

Two sets are identical if the contain exactly the same points and then we write $A = B$. In other words, $A = B$ iff[14] $A \subset B$ and $B \subset A$.

Some basic operations with sets are illustrated in figure 1.4. We briefly repeat them here:

Complement. The complement of the set A is denoted by A^c and consists of all points not belonging to A:[15]

$$A^c \;=\; \{\omega | \omega \notin A\} \,. \tag{1.6}$$

There are three evident relations which can be verified easily: $(A^c)^c = A$, $\Omega^{\,c} = \emptyset$, and $\emptyset^{\,c} = \Omega$.

Union. The union of the two sets A and B, $A \cup B$, is the set of points, which belong to at least one of the two sets:

$$A \cup B \;=\; \{\omega | \omega \in A \text{ or } \omega \in B\} \,. \tag{1.7}$$

Intersection. The intersection of the two sets A and B, $A \cap B$, is the set of points, which belong to both sets:[16]

$$A \cap B \;=\; AB \;=\; \{\omega | \omega \in A \text{ and } \omega \in B\} \,. \tag{1.8}$$

Unions and intersections can be executed in sequence and are also defined for more than two sets, or even for a countably infinite number of sets:

$$\bigcup_{n=1,\ldots} A_n \;=\; A_1 \cup A_2 \cup \cdots \;=\; \{\omega | \omega \in A_n \text{ for at least one value of } n\} \,,$$

$$\bigcap_{n=1,\ldots} A_n \;=\; A_1 \cap A_2 \cap \cdots \;=\; \{\omega | \omega \in A_n \text{ for all values of } n\} \,.$$

The proof of these relations is straightforward, because the commutative and the associative laws are fulfilled by both operations, intersection and union:

$$A \cup B \;=\; B \cup A \,,\; A \cap B \;=\; B \cap A \,;$$
$$(A \cup B) \cup C \;=\; A \cup (B \cup C) \,,\; (A \cap B) \cap C \;=\; A \cap (B \cap C) \,.$$

Difference. The set theoretic difference, $A \setminus B$, is the set of points, which belong to A but not to B:

[14] The word *'iff'* stands for *if and only if*.

[15] Since we are considering only fixed sample sets Ω these points are uniquely defined.

[16] For short $A \cap B$ is often written simply as AB.

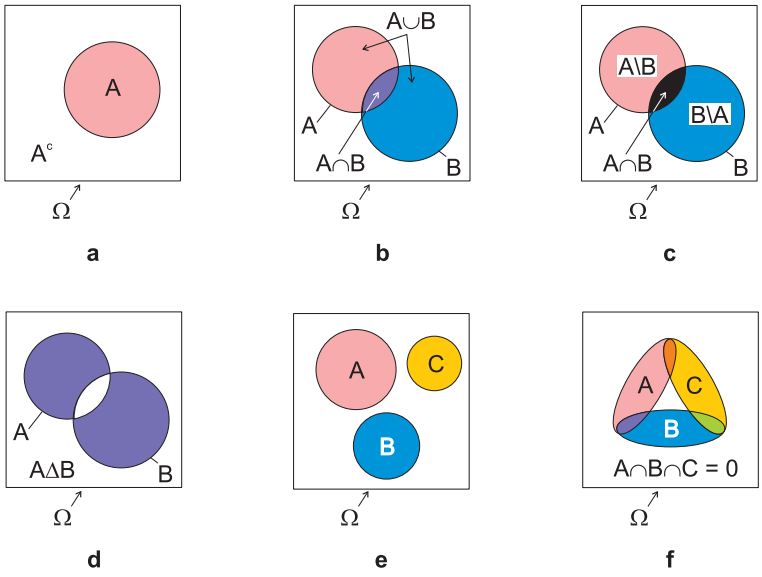

Fig. 1.4 Some definitions and examples from set theory. Part **a** shows the complement A^c of a set A in the sample space Ω. In part **b** we explain the two basic operations union and intersection, $A \cup B$ and $A \cap B$, respectively. Parts **c** and **d** show the set-theoretic difference, $A \setminus B$ and $B \setminus A$, and the symmetric difference, $A \triangle B$. In parts **e** and **f** we demonstrate that a vanishing intersection of three sets does not imply pairwise disjoint sets. The illustrations are made by means of *Venn diagrams* [112, 113, 289, 290].

$$A \setminus B \;=\; A \cap B^c \;=\; \{\omega | \omega \in A \text{ and } \omega \notin B\} \; . \tag{1.9}$$

In case $A \supset B$ we write $A - B$ for $A \setminus B$ and have $A \setminus B = A - (A \cap B)$ as well as $A^c = \Omega - A$.

Symmetric difference. The symmetric difference $A \Delta B$ is the set of points which belongs exactly to one of the two sets A and B. It is used in advanced theory of sets and is symmetric as it fulfils the commutative law, $A \Delta B = B \Delta A$:

$$A \Delta B \;=\; (A \cap B^c) \cup (A^c \cap B) \;=\; (A \setminus B) \cup (B \setminus A) \; . \tag{1.10}$$

Disjoint sets. Disjoint sets A and B have no points in common and hence their intersection, $A \cap B$, is empty. They fulfill the following relations:

$$A \cap B \;=\; \emptyset \; , \quad A \subset B^c \text{ and } B \subset A^c \; . \tag{1.11}$$

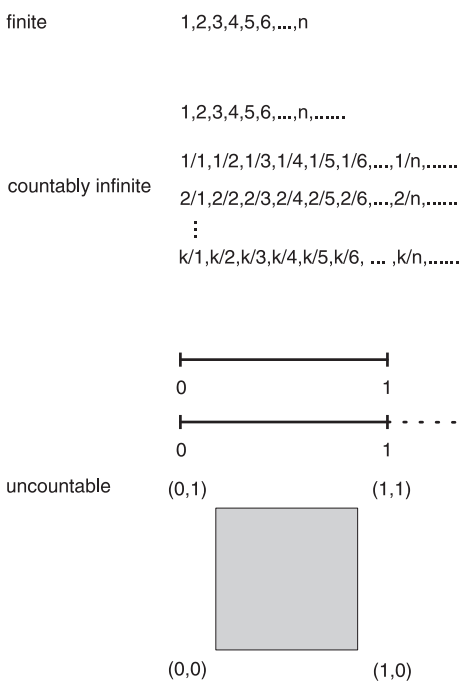

Fig. 1.5 Sizes of sample sets and countability. Finite, countably infinite, and uncountable sets are distinguished. We show examples of every class. A set is countably infinite when its elements can be assigned uniquely to the natural numbers ($\mathbb{N}_{>0} = 1,2,3,\ldots,n,\ldots$).

Several sets are disjoint only if they are pairwise disjoint. For three sets, A, B and C, this requires $A \cap B = \emptyset$, $B \cap C = \emptyset$, and $C \cap A = \emptyset$. When two sets are disjoint the addition symbol is (sometimes) used for the union, $A + B$ for $A \cup B$. Clearly we have always the valid *decomposition*: $\Omega = A + A^c$.

Sample spaces may contain finite or infinite numbers of sample points. As shown in figure 1.5 it is important to distinguish further between different classes of infinity:[17] *countable* and *uncountable* numbers of points. The set of *rational numbers* $\mathbb{Q}$, for example, is a countably infinite since the numbers can be labeled and assigned uniquely to the positive integers $1 < 2 < 3 < \cdots < n < \cdots$ also called *natural numbers* $\mathbb{N}_{>0}$. The set of *real numbers* $\mathbb{R}$, cannot be ordered in such a way and hence it is uncountable (For old and current notations of number systems see the appendix "notations").

[17] Georg Cantors attributed to countably infinite sets the cardinality $\aleph_0$ and characterized uncountable sets by /-*the sizes $\aleph_1$, $\aleph_2$, etc.

1.5 Probability measure on countable sample spaces

For countable sets it is straightforward and almost trivial to measure the size of the set by counting the numbers of sample points they contain. The ratio

$$P(A) \;=\; \frac{|A|}{|\Omega|} \tag{1.12}$$

then gives the probability for the occurrence of event A. For another event, for example B, holds $P(B) = |B|/|\Omega|$. A calculation of the the sum of the two probabilities, $P(A) + P(B)$, requires some care, since we know that only an inequality holds (see previous section 1.4, in particular figure 1.4):

$$|A| \;+\; |B| \;\geq\; |A \cup B| \;.$$

The excess of $|A| + |B|$ over the size of the union $|A \cup B|$ is precisely the size of the intersection $|A \cap B|$ and thus we find

$$|A| \;+\; |B| \;=\; |A \cup B| \;+\; |A \cap B|$$

or by division through the size of sample space Ω we obtain

$$\begin{aligned} P(A) \,+\, P(B) \;&=\; P(A \cup B) \,+\, P(A \cap B) \;\text{ or} \\ P(A \cup B) \;&=\; P(A) \,+\, P(B) \,-\, P(A \cap B) \;. \end{aligned} \tag{1.13}$$

Only in case the intersection is empty, $A \cap B = \emptyset$, the two sets are disjoint and their probabilities are additive, $|A \cup B| = |A| + |B|$, and hence

$$P(A + B) \;=\; P(A) \,+\, P(B) \;\text{ iff }\; A \cap B = \emptyset \;. \tag{1.14}$$

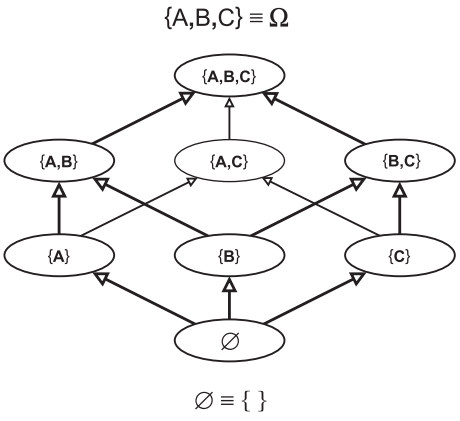

Fig. 1.6 The powerset. The powerset $\Pi(\Omega)$ is a set containing all subsets of Ω including the empty set $\emptyset$ and Ω itself. The figure sketches the powerset for a sample space of three events A, B, and C.

It is important to memorize this condition for later use, because it represents an implicitly made assumption for computing probabilities.

Now are now in the position to define a probability measure by means of basic axioms of probability theory and we present the three axioms as they were first formulated by Andrey Kolmogorov [167]:

A *probability measure* on the sample space Ω is a function of subsets of Ω, $P : S \to P(S)$, which is defined by the three axioms:

(i) For every set $A \subset \Omega$, the value of the probability measure is a nonnegative number, $P(A) \geq 0$ for all A,
(ii) the probability measure of the entire sample set – as a subset – is equal to one, $P(\Omega) = 1$, and
(iii) for any two disjoint subsets A and B, the value of the probability measure for the union, $A \cup B = A + B$, is equal to the sum of its values for A and B,

$$P(A \cup B) \;=\; P(A + B) \;=\; P(A) + P(B) \text{ provided } P(A \cap B) \;=\; \emptyset \; .$$

Condition (iii) implies that for any countable – eventually infinite – collection of disjoint or non-overlapping sets, A_i $(i = 1, 2, 3, \ldots)$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$, the relation called σ-*additivity*

$$P\left(\bigcup_i A_i\right) \;=\; \sum_i P(A_i) \quad \text{or} \quad P\left(\sum_{k=1}^{\infty} A_k\right) \;=\; \sum_{k=1}^{\infty} P(A_k) \tag{1.15}$$

holds.

In other words, the probabilities associated with disjoint sets are additive. Clearly we have also $P(A^c) = 1 - P(A)$, $P(A) = 1 - P(A^c) \leq 1$, and $P(\emptyset) = 0$. For any two sets $A \subset B$ we find $P(A) \leq P(B)$ and $P(B - A) = P(B) - P(A)$, and for any two arbitrary sets A and B we can write the union as a sum of two disjoint sets

$$\begin{aligned} A \cup B \;&=\; A \,+\, A^c \cap B \text{ and} \\ P(A \cup B) \;&=\; P(A) \,+\, P(A^c \cap B) \; . \end{aligned}$$

Since $B \subset A^c \cap B$ we obtain $P(A \cup B) \leq P(A) + P(B)$.

The set of all subsets of Ω is called the *powerset* $\Pi(\Omega)$ (figure 1.6). It contains the empty set $\emptyset$, the entire sample space Ω and the subsets of Ω, and this includes the results of all set theoretic operations that were listed in the previous section 1.4. The relation between the sample point ω, an event A, the sample space Ω and the powerset $\Pi(\Omega)$ is illustrated by means of an example taken from as Penney's game (section 1.2), the repeated coin toss, which we shall analyze as Bernoulli process in section 3.2.1. Flipping a coin has two outcomes: '**0**' for *head* and '**1**' *tail* and one particular coin

toss experiment might give the sequence $(\mathbf{0},\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{0},\ldots,\mathbf{1},\mathbf{0},\mathbf{0})$. Thus the sample points ω for flipping the coin n-times are binary n-tuples or strings, $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ with $\omega_i \in \Sigma = \{0,1\}$.[18] The sample space Ω then is the space of all binary strings of length n commonly denoted by Σ^n and it has the cardinality $|\Sigma^n| = 2^n$. The extension to the set of all strings of any finite length is straightforward,

$$\Sigma^* \;=\; \bigcup_{i\in\mathbb{N}} \Sigma^i \;=\; \{\varepsilon\} \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \ldots \;, \tag{1.16}$$

and this set is called *Kleene star* after the American mathematician Stephen Kleene. Herein $\Sigma^0 = \{\varepsilon\}$ with ε denoting the unique string over Σ^0 called the *empty string*, $\Sigma^1 = \{\mathbf{0},\mathbf{1}\}$, $\Sigma^2 = \{\mathbf{00},\mathbf{01},\mathbf{10},\mathbf{11}\}$, etc. The importance of Kleene star is the *closure* property under concatenation of the sets Σ^i[19]

$$\Sigma^m \Sigma^n \;=\; \Sigma^{m+n} \;=\; \{wv | w \in \Sigma^m \,\text{and}\, v \in \Sigma^n\} \;\text{ with }\; m,n > 0 \;. \tag{1.17}$$

Concatenation of strings is the operation

$$w \,=\, (\mathbf{0001})\,,\; v \,=\, (\mathbf{101}) \;\Longrightarrow\; wv \,=\, (\mathbf{0001101})\;,$$

which can be extended to concatenation of sets in the sense of equation 1.17:

$$\begin{aligned}\Sigma^1\Sigma^2 \;&=\; \{\mathbf{0},\mathbf{1}\}\{\mathbf{00},\mathbf{01},\mathbf{10},\mathbf{11}\} \;= \\ &=\; \{\mathbf{000},\mathbf{001},\mathbf{010},\mathbf{011},\mathbf{100},\mathbf{101},\mathbf{110},\mathbf{111}\} \;=\; \Sigma^3\end{aligned}$$

The set Kleene star Σ^* is the smallest superset of Σ, which contains the empty string ε and which is closed under the string concatenation operation. Although all individual strings in Σ^* have finite length, the set Σ^* itself, however, is countably infinite. We end this brief excursion into strings and string operations by considering infinite numbers of repeats directly in the sense of Σ^n the space of strings of lengths n, $\omega = (\omega_1, \omega_2, \ldots) = (\omega_i)_{i\in\mathbb{N}}$ with $\omega_i \in \{0,1\}$ in the limit $\lim n \to \infty$, as they are used in the theory of computing. Then $\Omega = \{0,1\}^{\mathbb{N}}$ is the sample space of all infinitely long binary strings, whose countability as can be easily verified: Every binary string represents the binary encoding N_k of a natural number including '0', $N_k \in \mathbb{N}$, and hence Ω is countable as the natural numbers are.

A subset of Ω will be called an *event* A when a probability measure derived from axioms (i), (ii), and (iii) has been assigned. Often, one is not interested

[18] There is a trivial but important distinction between strings and sets: In a string the position of an element matters, whereas in a set it does not. The following three sets are identical: $\{1,2,3\} = \{3,1,2\} = \{1,2,2,3\}$. In order to avoid ambiguities strings are written in (normal) parentheses and sets in curly brackets.

[19] Closure under a given operation is an important property of a set that we shall need later on. The natural numbers $\mathbb{N}$, for example, are closed under addition and the integers $\mathbb{Z}$ are closed under addition and substraction.

in the full detail of a probabilistic result and events can be easily adapted to lumping together sample points. We ask, for example, for the probability A that n coin flips yield at least s-times *tail* or a score $k \geq s$:

$$A \;=\; \left\{\omega = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega : \; \sum\nolimits_{i=1}^{n} \omega_i = k \geq s\right\} \;,$$

where the sample space is $\Omega = \{0,1\}^n$. The task is now to find a system of events $\mathcal{F}$ that allows for a consistent assignment of a probability $P(A)$ to all possible events A. For countable sample spaces Ω the powerset $\Pi(\Omega)$ represents such a system $\mathcal{F}$: We characterize $P(A)$ as a probability measure on $\big(\Omega, \Pi(\Omega)\big)$, and the further handling of probabilities following the procedure outlined below is straightforward. In case of uncountable sample spaces Ω the powerset $\Pi(\Omega)$ will turn out to be too large and a more sophisticated procedure is required (section 1.6.3).

So far we have constructed, compared, and analyzed sets but have not yet introduced weights or numbers for application to real world situations. In order to construct a probability measure that can be adapted to calculations on countable sample space, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$, we have to assign a weight ϱ_n to every sample point ω_n that fulfils the conditions

$$\forall\; n: \quad \varrho_n \;\geq\; 0 \;\; \text{and} \;\; \sum_n \varrho_n = 1 \;. \tag{1.18}$$

Then for $P\left(\{\omega_n\}\right) = \varrho_n \;\forall\; n$ the following two equations

$$\begin{aligned} P(A) \;&=\; \sum_{\omega \in A} \varrho(\omega) \;\; \text{for} \;\; A \in \Pi(\Omega) \quad \text{and} \\ \varrho(\omega) \;&=\; P\left(\{\omega\}\right) \;\; \text{for} \;\; \omega \in \Omega \end{aligned} \tag{1.19}$$

represent a bijective relation between the probability measure P on $\big(\Omega, \Pi(\Omega)\big)$ and the sequences $\varrho = \big(\varrho(\omega)\big)_{\omega \in \Omega}$ in [0,1] with $\sum_{\omega \in \Omega} \varrho(\omega) = 1$. Such a sequence is called a discrete probability density.

The function $\varrho(\omega_n) = \varrho_n$ has to be prescribed by some null hypothesis, estimated or determined empirically, because it is the result of factors lying outside mathematics or probability theory. The uniform distribution is commonly adopted as null hypothesis in gambling as well as for many other purposes: The discrete *uniform distribution*, $\mathcal{U}_\Omega$, assumes that all elementary results $\omega \in \Omega$ appear with equal probability and hence $\varrho(\omega) = 1/|\Omega|$.[20] What is meant here by 'elementary' will become clear in the discussion of applica-

[20] The assignment of equal probabilities $\frac{1}{n}$ to n mutually exclusive and collectively exhaustive events, which are indistinguishable except for their tags, is known as *principle of insufficient reason* or *principle of indifference* as it was called by the British economist John Maynard Keynes [159, chap.IV, pp.44-70]. In Bayesian probability theory the *a priori* assignment of equal probabilities is characterized as the simplest *non-informative prior* (see section 1.3).

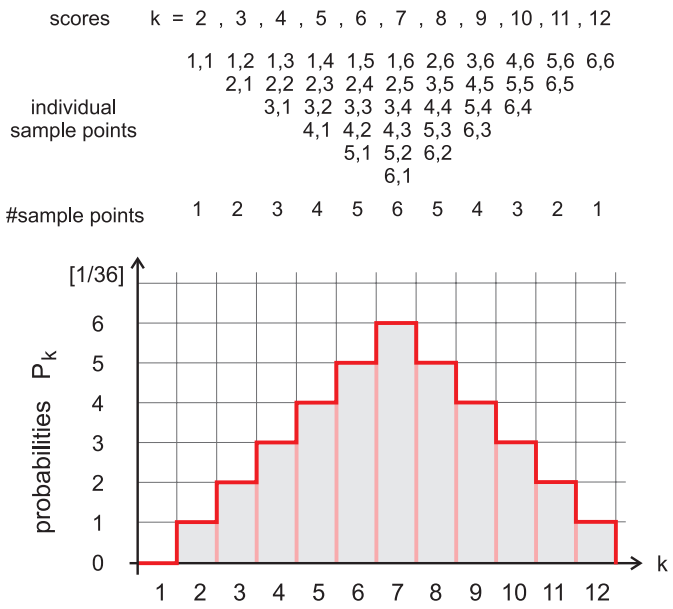

Fig. 1.7 Probabilities of throwing two dice. The probability of obtaining two to twelve counts through throwing two perfect or fair dice are based on the equal probability assumption for obtaining the individual faces of a single die. The probability $P(N)$ raises linearly from two to seven and then decreases linearly between seven and twelve ($P(N)$ is a discretized tent map) and the additivity condition requires $\sum_{k=2}^{12} P(N=k) = 1$. Understood as a probability distribution function of a random variable $\mathcal{Z}$ the shown plot represents the probability mass function (pmf) $f_{\mathcal{Z}}(x)$. It is important to note that the pmf is not a step function but a collection of isolated values at the points $x = k$ with $k \in \{2, 3, \ldots, 12\}$ (see figure 1.11).

tions. Throwing more than one die at a time, for example, can be reduced to throwing one die more often.

In science, particularly in physics, chemistry or biology, the correct assignment of probabilities has to meet the conditions of the experimental setup. An example from *scientific gambling* will make this point clear: The fact whether a die is fair and shows all its six faces with equal probability, whether it is imperfect, or whether it has been manipulated and shows, for example, the 'six' more frequently then the other faces is a matter of physics and not mathematics. Empirical information replaces the principle of indifference – for example, a calibration curve of the faces is determined by doing and recording a few thousand die rolling experiments – and assumptions of the null hypothesis of a uniform distribution become obsolete.

Although the application of a probability measure in the discrete case is rather straightforward, we illustrate by means of a simple example. With the assumption of uniform distribution $\mathcal{U}_\Omega$ we can measure the size of sets by

counting sample points as illustrated best by considering the throws of dice. For one die the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$ and for the *fair* die we make the assumption

$$P(\{k\}) \;=\; \frac{1}{6};\; k = 1, 2, 3, 4, 5, 6\;.$$

that all six outcomes corresponding to different faces of the die are equally likely. Based on the assumption of $\mathcal{U}_\Omega$ we obtain the probabilities for the outcome of two simultaneously rolled fair dice (figure 1.7). There are $6^2 = 36$ possible outcomes with scores in the range $2, 3, \ldots, 12$. The *probability mass function* (pmf) or discrete probability density is a discretized tent function in this case with the most likely outcome being a count of seven points because it has the largest multiplicity, $\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$ (For a generalization to rolling n dice simultaneously see section 1.9.1 and figure 1.20).

1.6 Discrete random variables and distributions

Conventional deterministic variables are not suitable for descriptions of processes with limited reproducibility. In probability theory and statistics we shall make use of *random* or *stochastic variables*, $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$, which were invented especially for dealing with random scatter and fluctuations. Even if an experiment is repeated under precisely the same conditions the random variable will commonly take on a different value. The probabilistic nature of random variables is illustrated well by an expression, which is particularly useful for the definition of probability distribution functions:[21]

$$P_k \;=\; \mathrm{Prob}\left(\mathcal{Z} \,=\, k\right) \;\text{ with }\; k \in \mathbb{N}\;. \tag{1.20}$$

A deterministic variable, $z(t)$, is defined by a function that returns a unique value for a given argument $z(t) = x_t$.[22] In case of the random variable, $\mathcal{Z}(t)$, the single value of the conventional variable has to be replaced by a series of probabilities $P_k(t)$. This series could be visualized, for example, by means of an L_1 normalized probability vector with the probabilities P_k as components: $\mathbf{P} = \left(P_0, P_1, \ldots\right)$ with $\|\mathbf{P}\|_1 = \sum_k P_k = 1$.[23] In probability theory the characterization of a random variable is made by a probability distribution function rather than by a vector, because these functions can

[21] Whenever possible we shall use "k, l, m, n" for discrete counts, $k \in \mathbb{N}$, and "t, x, y, z" for continuous variables, $x \in \mathbb{R}^1$ (see appendix 'Notation').

[22] We use here t as independent variable of the function but do not necessarily imply that t is time.

[23] The notation of vectors and matrices as used in this text is described in the appendix 'Notation'.

be applied with minor modifications to the discrete and the continuous case. Two probability functions are particularly important and in general use (see section 1.6.2): the probability mass function (pmf; see figures 1.7 and 1.11)

$$f_{\mathcal{Z}}(x) \;=\; \begin{cases} \mathrm{Prob}\,(\mathcal{Z} \,=\, k) \;=\; P_k & \forall \;\; x = k \,\in\, \mathbb{N}\;, \\ 0 & \text{anywhere else}\,. \end{cases}$$

or by the *cumulative distribution function* (cdf; see figure 1.10)

$$F_{\mathcal{Z}}(x) \;=\; \mathrm{Prob}\,(\mathcal{Z} \,\leq\, k) \;=\; \sum_{i \leq k} P_i\;.$$

Two properties of the cumulative distribution function (cdf) follow directly from the property of probabilities:

$$\lim_{k \to -\infty} \; F_{\mathcal{X}}(k) \;=\; 0 \quad \text{and} \quad \lim_{k \to +\infty} \; F_{\mathcal{X}}(k) \;=\; 1\;.$$

The limit at low k-values is chosen in analogy to definitions used later on: Taking $-\infty$ instead of zero as lower limit makes no difference, because $f_{\mathcal{X}}(-|k|) = P_{-|k|} = 0$ ($k \in \mathbb{N}$) or negative particle numbers have zero probability. Simple examples of probability functions are shown in figures 1.7 and 1.10.

The probability mass function (pmf) $f_{\mathcal{Z}}(x)$ is not a function in the usual sense, because it has the value zero almost everywhere except at the points $x = k \in \mathbb{N}$ and in this aspect it is closely related to the Dirac delta function (section 1.6.2). All measurable quantities, for example expectation values $E(\mathcal{Z})$ or variances $\mathrm{var}(\mathcal{Z}) = \sigma^2(\mathcal{Z})$, can be computed from either of the two probability functions.

1.6.1 Random variables and continuity

For a precise definition of random variables on countable sample spaces a *probability triple* $(\Omega, \Pi(\Omega), P)$ is required: Ω contains the sample points or individual results, the powerset $\Pi(\Omega)$ provides the events A as subsets, and P eventually represents a probability measure that has been defined in equation (1.19). Based on such a probability triple we can now define a *random variable* as a numerically valued function $\mathcal{Z}$ of ω on the domain of the entire sample space Ω,

$$\omega \in \Omega \;:\;\; \omega \;\to\; \mathcal{Z}(\omega)\;. \tag{1.21}$$

Random variables, $\mathcal{X}(\omega)$ and $\mathcal{Y}(\omega)$, can be manipulated by operations to yield other random variables, such as

$$\mathcal{X}(\omega) + \mathcal{Y}(\omega)\,,\; \mathcal{X}(\omega) - \mathcal{Y}(\omega)\,,\; \mathcal{X}(\omega)\mathcal{Y}(\omega)\,,\; \mathcal{X}(\omega)/\mathcal{Y}(\omega)\;\; [\mathcal{Y}(\omega) \neq 0]\,,$$

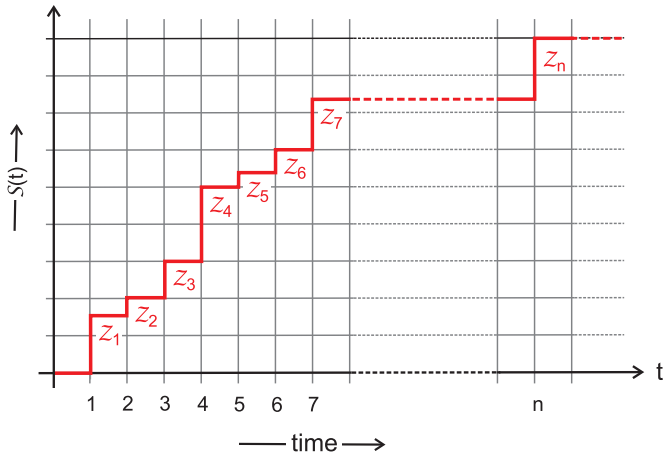

Fig. 1.8 Ordered partial sum of random variables. The sum $\mathcal{S}_n = \sum_{k=1}^{n} \mathcal{Z}_k$ represents the cumulative outcome of a series of events described by a class of random variables, $\mathcal{Z}_k$. The series can be extended to $+\infty$ and such cases will be encountered, for example, with probability distributions. An ordering criterion has still to be specified, it could be *time* t, for example, and then we are dealing with a stochastic process, here a jump process, or a *spatial coordinate* x, y or z.

and, in particular, also any linear combination of random variables such as $\alpha\mathcal{X}(\omega) + \beta\mathcal{Y}(\omega)$ is a random variable too. Just as a function of a function is still a function, a function of a random variable is a random variable,

$$\omega \in \Omega : \quad \omega \to \varphi\left(\mathcal{X}(\omega), \mathcal{Y}(\omega)\right) = \varphi(\mathcal{X}, \mathcal{Y}) \ .$$

Particularly important cases are the partial sums of n variables:

$$S_n(\omega) = \mathcal{Z}_1(\omega) + \ldots + \mathcal{Z}_n(\omega) = \sum_{k=1}^{n} \mathcal{Z}_k(\omega) \ . \tag{1.22}$$

Such a partial sum S_n could be, for example, the cumulative outcome of n successive throws of a die.[24] Consider, for example, an ordered series of events where the current cumulative outcome is given by the sum $S_n = \sum_{k=1}^{n} \mathcal{Z}_k$ as shown in figure 1.8. In principle, the series can be extended to infinity covering thereby entire sample space and then the conservation relation of probabilities, $S_n = \sum_{k=1}^{\infty} \mathcal{Z}_k = 1$, has to be fulfilled. No ordering criterium has been introduced so far. Most frequently and in particular for

[24] The use of *partial* in this context expresses the fact that the sum need not cover the entire sample space at least not for the moment. Series of rolling dice, for example, could be continued in the future.

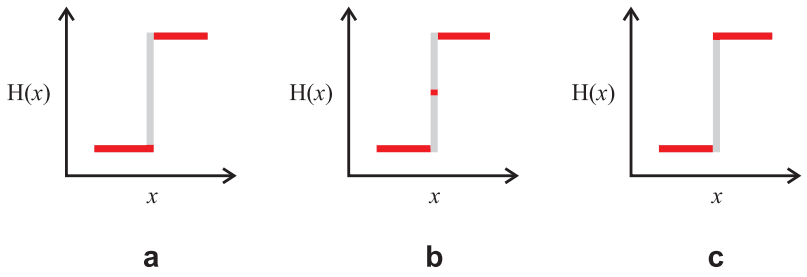

Fig. 1.9 Continuity in probability theory and step processes. Three possible choices of partial continuity at the steps of step functions are shown: (i) left-hand continuity (**a**), (ii) no continuity (**b**), and (iii) right-hand continuity (**c**). The step function in (**a**) is left-hand semi-differentiable, the step function in (**c**) is right-hand semi-differentiable, and the step function in (**b**) is neither right-hand nor left-hand semi-differentiable. Choice (ii) allows for making use of the inherent symmetry of the Heaviside function. Choice (iii) is the standard assumption in probability theory and stochastic processes. It is also known as *càdlàg*-property (section 3.2.1.3).

stochastic processes events will be ordered according the time of occurrence (see chapter 3).

Figure 1.8 represents the plot of a discrete random variable, $\mathcal{S}(t)$, on a continuous axis, time t, which has the from of a step function. In order to avoid ambiguities a convention concerning continuity at the steps is needed. A precise definition, however, is hidden in the equations (1.21) and (1.22). Three definitions for the value of the function at the discontinuity are possible. In the case of the Heaviside step function they are (figure 1.9):

$$H(x) \;=\; \begin{cases} 0\,, & \text{if } x < 0\,, \\ 0,\, \frac{1}{2},\, 1 & \text{if } x = 0\,, \\ 1\,, & \text{if } x > 0\,. \end{cases} \tag{1.23}$$

The value '0' at $x = 0$ implies left hand continuity for $H(x)$ and in terms of a probability distribution would correspond to a definition $P(\mathcal{Z} < x)$, the value $\frac{1}{2}$ implies that $H(x)$ is neither right-hand nor left-hand semi-differentiable at $x = 0$ but this choice is useful in many applications that are based on the inherent symmetry of the Heaviside function, for example the relation $H(x) = \big(1 + \mathrm{sgn}(x)\big)/2$ where $\mathrm{sgn}(x)$ is the sign or signum function:

$$\mathrm{sgn}(x) \begin{cases} -1 & \text{if } x < 0\,, \\ 0 & \text{if } x = 0\,, \\ 1 & \text{if } x > 0\,. \end{cases}$$

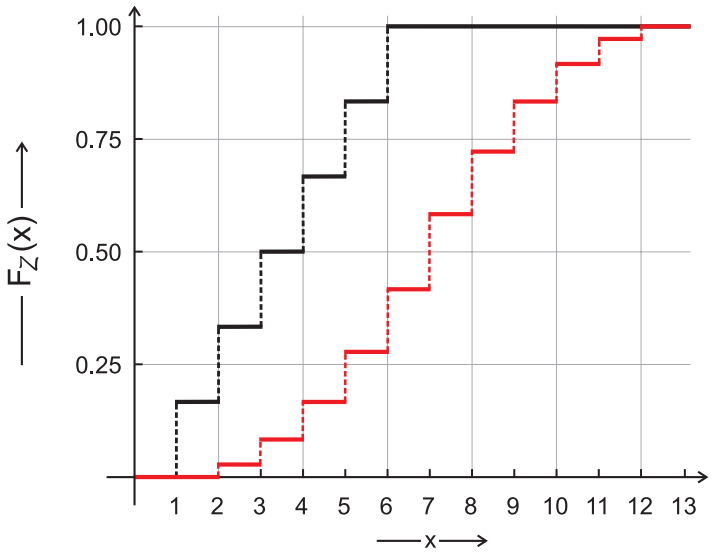

Fig. 1.10 The cumulative distribution function of rolling fair dice. The cumulative probability distribution function (cdf) is a mapping from the sample space Ω onto the unit interval $[0,1]$ of $\mathbb{R}$. It corresponds to the ordered partial sum with the ordering parameter being the score given by the stochastic variable. The example shown deals with throwing fair dice: The distribution for one die (black) consists of six steps of equal height at the scores $1,2,\ldots,6$. The second curve (red) is the probability of throwing two dice yielding the scores $2,3,\ldots,12$. The weights for the individual scores are 1/36,1/18,1/12,1/9,5/36,1/6,5/36,1/9,1/12,1/18,1/36. The two limits of a cdf are $\lim_{x\to-\infty} F_{\mathcal{Z}}(x) = 0$ and $\lim_{x\to+\infty} F_{\mathcal{Z}}(x) = 1$.

The functions in probability theory make use of the third definition determined by $P(\mathcal{Z} \leq x)$ or $H(0) = 1$ in case of the Heaviside function, and this definition leads to right-hand continuity. In other words the step-functions in probability theory are semi-differentiable to the right. Right-hand continuity is an important definition in the conventional handling of stochastic processes, for example in the case of semimartigales (section 3.2.1). Often, the property of right-hand continuity with left limits is denoted as *càdlàg*, which is an acronym from French for "*continue à droite, limites à gauche*". Step functions cannot be integrated by conventional Riemannian integration method, but they are accessible by Stieltjes integration as will be outlined later on in a section on generalizations of the Riemann integral (subsection 1.8.2).

1.6.2 Mass function and cumulative distribution

Two functions of random variables, the probability mass function (pmf) and the cumulative probability distribution (cdf) have been already mentioned and were shown in figures 1.7 and 1.8, respectively. Both functions are equivalent in the sense that essentially all observable properties can be calculated from either one of them. Here, we discuss them again by means of the simple example, rolling two dice ($2\mathcal{D}$), and we present general expressions for them and their interconversions.

As a simple and illustrative example of a probability distributions is the mass function presenting the scores of rolling two dice (figure 1.7) as events. The pmf for two dice is a tent function

$$f_{2\mathcal{D}}(k) \;=\; \begin{cases} \frac{1}{s^2}\,(k-1) & \text{for } \; k \,=\, 1,2,\ldots,s\;, \\ \frac{1}{s^2}\,(2s+1-k) & \text{for } \; k \,=\, s+1,s+2,\ldots,2s \end{cases} .$$

Here k is the score and s the number of faces of the die, which is six in case of the commonly used dice. The cumulative probability distribution function (cdf) is an example of for an ordered sum of random variables. The scores of rolling one die or two dice simultaneously are the events. The cumulative probability distribution is given by the sum of scores (figure 1.10):

$$F_{2\mathcal{D}}(k) \;=\; \sum_{i=2}^{k} f_{2\mathcal{D}}(i)\,;\; k = 2,3,\ldots,2s\;.$$

A generalization to rolling n dice will be presented in chapter 2.5 in the discussion of the *central limit theorem*.

Making use of our knowledge on probability space the probability mass function (pmf) can be formulated as a mapping from sample space into the real numbers and gives the probability that a discrete random variable $\mathcal{Z}$ attains exactly some value x. We assume that $\mathcal{Z}$ is a discrete random variable on the sample space Ω, $\mathcal{Z} : \Omega \to \mathbb{R}$, and then we define the probability mass function as a mapping onto the unit interval, $f_{\mathcal{Z}} : \mathbb{R} \to [0,1]$, by

$$f_{\mathcal{Z}}(x) \;=\; P(\mathcal{Z} = x) \;=\; P\big(\{s \in \Omega :\, \mathcal{Z}(s) = x\}\big)\;. \tag{1.24}$$

Sometimes it is useful to be able to treat a discrete probability distribution as if it were continuous. The function $f_{\mathcal{Z}}(x)$ is defined therefore for all real numbers, $x \in \mathbb{R}$ including those outside the sample set. Then we have: $f_{\mathcal{Z}}(x) = 0\,\forall\, x \notin \mathcal{Z}(\Omega)$. For rolling one die the pmf consists of six isolated peaks, $f_{\mathcal{Z}}(x) = 1/6$ at $x = 1,2,\ldots,6$ and has the value $f_{\mathcal{Z}}(x) = 0$ everywhere else ($x \neq 1,2,\ldots,6$). Figure 1.11 shows the probability mass function of rolling dice, where the probability mass function corresponds to the discretized tent map shown already in figure 1.7. A simple but straightforward representation of the probability mass function makes use of the Dirac

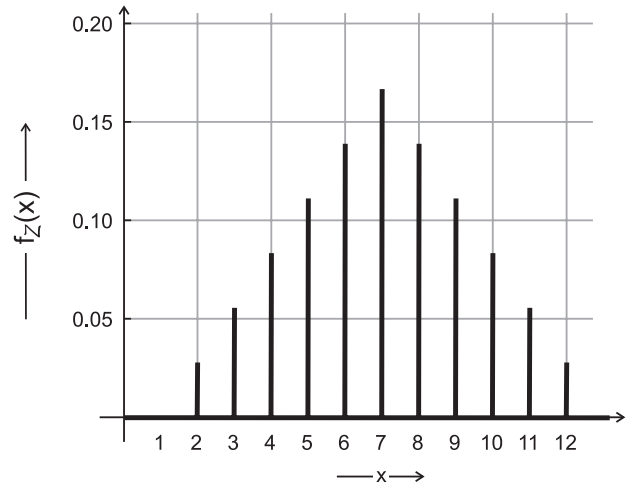

Fig. 1.11 Probability mass function of fair dice. The probability mass function (pmf), $f_{\mathcal{Z}}(x)$, is shown for rolling two dice simultaneously. The scores x are plotted on the abscissa axis. The pmf is zero everywhere on the x-axis except at a set of points, $x = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$, of measure zero where it adopts the values 1/36,1/18,1/12,1/9,5/36,1/6,5/36,1/9,1/12,1/18,1/36. The maximal probability value is obtained for the score $x = 7$ (see also equation (1.24') and figure 1.7).

delta-function.[25] The nonzero score values are assumed to lie exactly at the positions x_k with $k = 1, 2, \ldots$ and $p_k = P(\mathcal{Z} = x_k)$:

$$f_{\mathcal{Z}}(x) \;=\; \sum_{k=1}^{\infty} P(\mathcal{Z} = x_k)\, \delta(x - x_k) \;=\; \sum_{k=1}^{\infty} p_k\, \delta(x - x_k)\;. \qquad (1.24')$$

In this form the probability density function is suitable for the calculation of probabilities by integration.

The step function for the characterization of a discrete probability distribution is the cumulative distribution function (cdf). In essence, it contains the same information as the probability mass function. As a mapping from sample space into the real numbers on the unit interval, $(P(\mathcal{Z} \leq x; \Omega) \Rightarrow (F_{\mathcal{Z}}(x); \mathbb{R} : 0 \leq F_{\mathcal{Z}}(x) \leq 1)$ it is defined by

$$F_{\mathcal{Z}}(x) \;=\; P(\mathcal{Z} \leq x) \;\text{ with }\; \lim_{x \to -\infty} F_{\mathcal{Z}}(x) = 0 \text{ and } \lim_{x \to +\infty} F_{\mathcal{Z}}(x) = 1\,. \quad (1.25)$$

[25] The delta-function is no proper function but a generalized function or *distribution*. It was introduced by Paul Dirac in quantum mechanics. For more details see, for example, [255, pp.585-590] and [251, pp.38-42].

Two examples for throwing one die or two dice are shown in figure 1.10. The distribution function is defined for the entire x-axis, $x \in \mathbb{R}$, but cannot be integrated by conventional Riemann integration. The cumulative distribution function and the partial sums of random variables, however, are continuous and differentiable on the right-hand side of the steps and therefore they are Riemann-Stieltjes or Lebesgue integrable (see sections 1.8.2 and 1.8.3). Since the integral of the Dirac delta-function is the Heaviside function we may write

$$F_{\mathcal{Z}}(x) \;=\; \int_{-\infty}^{x} f_{\mathcal{Z}}(s)\,\mathrm{d}s \;=\; \sum_{k\le x} p_k \;. \tag{1.25'}$$

For more details concerning limitations of integrability see section 1.8.

Finally, we generalize sets, which are defined by the range of a random variable on the closed interval $[a,b]$,[26]

$$\{a \le \mathcal{Z} \le b\} \;=\; \{\omega |\, a \le \mathcal{Z}(\omega) \le b\} \;,$$

and define their probabilities by $P(a \le \mathcal{Z} \le b)$. More generally, the set A of sample points can be defined by the open interval $]a,b[$, the half-open intervals $[a,b[$ and $]a,b]$, the infinite intervals, $]-\infty,b[$ and $]a,+\infty[$, as well as the set of real numbers, $\mathbb{R} =]-\infty,+\infty[$. If A is reduced to the single point x, it is called the *singleton* $\{x\}$:

$$P(\mathcal{Z} = x) \;=\; P(\mathcal{Z} \in \{x\}) \;.$$

For countable, finite or countably infinite, sample spaces Ω the exact range of $\mathcal{Z}$ is just the set of the real numbers w_i below:

$$W_{\mathcal{Z}} \;=\; \bigcup_{\omega\in\Omega} \{\mathcal{Z}(\omega)\} \;=\; \{w_1, w_2, \ldots, w_n, \ldots\} \;.$$

Now we introduce probabilities

$$p_n \;=\; P(\mathcal{Z} = w_n)\,,\; w_n \in W_{\mathcal{Z}} \;,$$

and apparently we have $P(\mathcal{Z} = x) = 0$ if $x \notin W_{\mathcal{Z}}$. An illustrative example was the probability mass function $f_{\mathcal{Z}}(x)$ defined by equation (1.24').

Knowledge of all p_n-values is tantamount to having full information on all probabilities derivable for the random variable $\mathcal{Z}$:

$$P(a \le \mathcal{Z} \le b) \;=\; \sum_{a\le w_n\le b} p_n \quad \text{or, in general,} \quad P(\mathcal{Z} \in A) \;=\; \sum_{w_n\in A} p_n \;. \tag{1.26}$$

[26] The notation we are applying here uses square brackets, '[··]', for closed intervals, open square brackets, ']··[', for open intervals, ']··]' and '[··[' for left-hand or right-hand half-open intervals, respectively. An alternative less common notation uses parentheses instead of open square brackets, e.g., '(··)' instead of ']··['.

An especially important case that has been discussed already in the previous subsection 1.6.2 is obtained when A is the infinite interval $]-\infty, x]$. The function $x \to F_{\mathcal{Z}}(x)$, defined on $\mathbb{R}$ and in particular on the unit interval $[0,1]$, $0 \le F_{\mathcal{Z}}(x) \le 1$, is the cumulative distribution function of $\mathcal{Z}$:

$$F_{\mathcal{Z}}(x) \;=\; P(\mathcal{Z} \le x) \;=\; \sum_{w_n \le x} p_n \;. \tag{1.25"}$$

It fulfils several easy to verify properties:

$$\begin{aligned} F_{\mathcal{Z}}(a) \,-\, F_{\mathcal{Z}}(b) \;&=\; P(\mathcal{Z} \le b) \,-\, P(\mathcal{Z} \le a) \;=\; P(a < \mathcal{Z} \le b) \;, \\ P(\mathcal{Z} = x) \;&=\; \lim_{\epsilon \to 0} \big(F_{\mathcal{Z}}(x+\epsilon) \,-\, F_{\mathcal{Z}}(x-\epsilon)\big) \;, \text{ and} \\ P(a < \mathcal{Z} < b) \;&=\; \lim_{\epsilon \to 0} \big(F_{\mathcal{Z}}(b-\epsilon) \,-\, F_{\mathcal{Z}}(a+\epsilon)\big) \;. \end{aligned}$$

An important special case is an integer valued positive random variable $\mathcal{Z}$ corresponding to a countably infinite sample space which is the set of non-negative integers: $\Omega = \mathbb{N}^0 = \{0, 1, 2, \ldots, n, \ldots\}$ with

$$p_n \;=\; P(\mathcal{Z} = n)\,,\; n \in \mathbb{N}^0 \;\text{ and }\; F_{\mathcal{Z}}(x) \;=\; \sum_{0 \le n \le x} p_n \;. \tag{1.27}$$

Integer valued random variables will be used, for example, for modeling particle numbers or other discrete quantities in stochastic processes.

1.6.3 Conditional probabilities and independence

So far probabilities of events A were defined relative to the entire sample space Ω, $P(A) = |A|/|\Omega| = \sum_{\omega \in A} P(\omega) \;/\; \sum_{\omega \in \Omega} P(\omega)$. We are now interested in the probability of event A relative to a subset of sample space Ω, for example the set S. This means that we attempt to calculate the proportional weight of the part of the subset A in S, which is expressed by the intersection $A \cap S$ relative to the set S, and obtain

$$\sum_{\omega \in A \cap S} P(\omega) \;\Big/\; \sum_{\omega \in S} P(\omega) \;.$$

In other words, we switch from Ω to S as the new universe and the set to be weighted are the sample points belonging to A and to S. It is illustrative to call the event S a *hypothesis* which restricts the sample space Ω for the definition of conditional probabilities.

The *conditional probability* measures the probability of A *relative to* S:

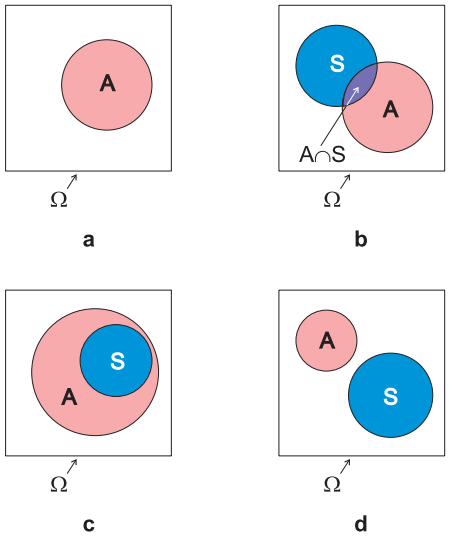

Fig. 1.12 Conditional probabilities. Conditional probabilities measure the intersection of the sets for two events, $A \cap S$ relative to the set S: $P(A|S) = |AS|/|S|$. In essence this is the same kind of weighting that defines the probabilities in sample space: $P(A) = |A|/|\Omega|$ (Part **a** shows $A \subset \Omega$ and **b** shows $A \cap S \subset S$). The two extremes are: $A \cap S = S$ and $P(A|S) = 1$ (**c**) and $A \cap S = 0$ and $P(A|S) = 0$ (**d**).

$$P(A|S) \;=\; \frac{P(A \cap S)}{P(S)} \;=\; \frac{P(AS)}{P(S)} \tag{1.28}$$

provided $P(S) \neq 0$. The conditional probability $P(A|S)$ is undefined for hypothesis of zero probability, $S = \emptyset$. Apparently, the conditional probability vanishes when the intersection is empty: $P(A|S) = 0$ if $A \cap S = AS = \emptyset$,[27] and $P(AS) = 0$. In case S is a true subset of A, $AS = S$ we have $P(A|S) = 1$ (see figure 1.12).

The definition of the conditional probability implies that all general theorems on probabilities hold by the token for conditional probabilities and, for example, we derive from equation (1.13):

$$P(A \cup B|S) \;=\; P(A|S) \,+\, P(B|S) \,-\, P(AB|S)\;. \tag{1.13'}$$

Additivity of conditional probability requires an empty intersection, $AB = \emptyset$.

Equation (1.28) is particularly useful when written in slightly different form:

$$P(AS) \;=\; P(A|S) \cdot P(S)\;, \tag{1.28'}$$

[27] From here on we shall use the short notation for the intersection, $AS \equiv A \cap S$.

which is known as the *theorem of compound probabilities* and which can be easily generalized to more events. For three events we derive [76, chap.V]

$$P(ABC) \;=\; P(A|BC)\cdot P(B|C)\cdot P(C)$$

by applying (1.28') twice – first by setting $BC \equiv S$ and then by setting $BC \equiv AS$, and for n arbitrary events $A_i; i = 1,\ldots,n$ we obtain

$$P(A_1A_2\ldots A_n) \;=\; P(A_1|A_2A_3\ldots A_n)\cdot P(A_2|A_3\ldots A_n)\;\ldots\; P(A_{n-1}|A_n)\cdot P(A_n)$$

provided $P(A_2A_3\ldots A_n) > 0$. If the intersection of event sets $A_2\ldots A_n$ does not vanish, all conditional probabilities are well defined since

$$P(A_n) \geq P(A_{n-1}A_n) \geq \ldots \geq P(A_2A_3\ldots A_n) > 0\ .$$

Next we derive an equation that we shall need in chapter 3 for modeling of stochastic processes. We assume that the sample space Ω is partitioned into n disjoint sets, $\Omega = \sum_n S_n$, then we have for any set A

$$A \;=\; AS_1 \cup AS_2 \cup \ldots \cup S_n$$

and from equation (1.28') we get

$$P(A) \;=\; \sum_n P(A|S_n)\cdot P(S_n)\ . \tag{1.29}$$

From this relation it is straightforward to derive the conditional probability

$$P(S_j|A) \;=\; \frac{P(S_j)P(A|S_j)}{\sum_n P(S_n)P(A|S_n)}$$

provided $P(A) > 0$.

Two or more random variables,[28] for example $\mathcal{X}$ and $\mathcal{Y}$, can be described by a *random vector* $\vec{\mathcal{V}} = (\mathcal{X},\mathcal{Y})$, which is expressed by the *joint probability*

$$P(\mathcal{X} = x_i, \mathcal{Y} = y_j) \;=\; p(x_i, y_j)\ . \tag{1.30}$$

The random vector $\vec{\mathcal{V}}$ is fully determined by the *joint probability mass function*

$$\begin{aligned} f_{\vec{\mathcal{V}}}(x,y) \;&=\; P(\mathcal{X}=x, \mathcal{Y}=y) \;=\; P(\mathcal{X}=x \;\vee\; \mathcal{Y}=y) \;= \\ &=\; P(\mathcal{Y}=y\,|\,\mathcal{X}=x)\cdot P(\mathcal{X}=x) \;= \\ &=\; P(\mathcal{X}=x\,|\,\mathcal{Y}=y)\cdot P(\mathcal{Y}=y)\ . \end{aligned} \tag{1.31}$$

[28] For simplicity we restrict ourselves to the two variable case here. The extension to any finite number of variables is straightforward.

This density constitutes the probabilistic basis of the random vector $\vec{V}$. It is straightforward to define a *cumulative probability distribution* in analogy to the single variable case

$$F_{\vec{\mathcal{V}}}(x,y) \;=\; P(\mathcal{X} \le x, \mathcal{Y} \le y)\;. \tag{1.32}$$

In principle either of the two probability functions contain the complete information on both variables but depending on the specific situation either the pmf or the cdf may be more efficient.

Often no detailed information is required on one particular random variable. Then, by summation over one variable of the vector $\vec{\mathcal{V}}$ we obtain the probabilities for the corresponding *marginal distribution*,

$$\begin{aligned} P(\mathcal{X} = x_i) \;&=\; \sum_{y_j} p(x_i, y_j) \;=\; p(x_i, *) \;\text{ and} \\ P(\mathcal{Y} = y_j) \;&=\; \sum_{x_i} p(x_i, y_j) \;=\; p(*, y_j)\;, \end{aligned} \tag{1.33}$$

of $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Independence of events can be easily formulated in terms of conditional probabilities. The conditional probability can also be interpreted that the information on whether or not an event S has occurred changes the probability of A. Independence, however, implies that an influence of S on A does not exist and hence $P(A|S) = P(A)$ defines *stochastic independence*. Making use of equation (1.28') we define

$$P(AS) \;=\; P(A) \cdot P(S)\;, \tag{1.34}$$

and realize an important symmetry of stochastic independence: A is independent of S implies S is independent of A, and we may account for this symmetry by defining independence by stating that A and S are independent if equation (1.34) holds. We remark that the definition (1.34) is acceptable also for $P(S) = 0$ a case in which $P(A|S)$ is undefined [76, p. 125].

The case of more than two events needs some care and we take three events A, B, C as an example. So far we were dealing with pairwise independence and accordingly we have

$$P(AB) = P(A) \cdot P(B)\,,\; P(BC) = P(B) \cdot P(C)\,,\; P(CA) = P(C) \cdot P(A)\,. \tag{1.35a}$$

Pairwise independence, however, does not necessarily imply that

$$P(ABC) \;=\; P(A) \cdot P(B) \cdot P(C) \tag{1.35b}$$

holds. In addition, examples were constructed where the last equation is fulfilled but nevertheless the sets are not pairwise independent [96]. Although cases of pairwise independence but lacking mutual independence of three events are not common they can be found in general: Case **f** in figure 1.4 al-

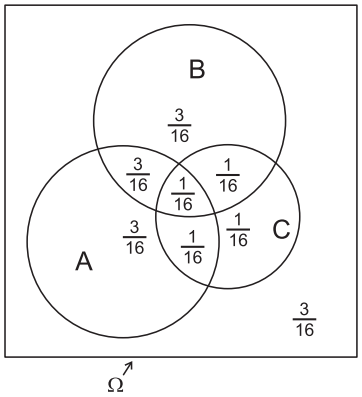

Fig. 1.13 Testing for stochastic independence of three events. The study case show is a example for independence of three events and fulfils equations (1.35a) and (1.35b). It corresponds to example **a** in table 1.3.

Table 1.3 Testing for stochastic independence of three events. We show three examples: Case **a** fulfils equations (1.35a) and (1.35b), and represents a case of mutual independence, case **b** fulfils only equation (1.35a) and not equation (1.35b), and is as example of pairwise independent but not mutually independent events, and case **c** is an especially constructed example for fulfilment of equation (1.35b) by three sets that are pairwise independent. Deviations from equations (1.35a) and (1.35b) are shown in boldface numbers.

	Probabilities P						
	Singles			Pairs			Tripel
	A	B	C	AB	BC	CA	ABC
a	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$
b	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\mathbf{\frac{1}{10}}$
c	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{2}$	$\mathbf{\frac{1}{10}}$	$\mathbf{\frac{6}{25}}$	$\mathbf{\frac{7}{50}}$	$\frac{1}{25}$

lows for straightforward construction of examples with pairwise independence but $P(ABC) = 0$.

Eventually, we present on final example, which is attributed to Sergei Bernstein [76, p. 127]: The six permutations of the three letters a, b and c together with the three triples (aaa), (bbb), (ccc) constitute the sample space and a

probability $P = \frac{1}{9}$ is attributed to each sample point. Now we define three events A_1, A_2 and A_3 according to the appearance of the letter a at the first, second or third place:

$$A_1 \;=\; \{aaa, abc, acb\}\,,\; A_2 \;=\; \{aaa, bac, cab\}\,,\; A_3 \;=\; \{aaa, bca, cba\}\;.$$

Every event has a probability $P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$ and the three events are pairwise independent because

$$P(A_1A_2) \;=\; P(A_2A_3) \;=\; P(A_3A_1) \;=\; \frac{1}{9}\;,$$

but they are not mutually independent because $P(A_1A_2A_3) = \frac{1}{9}$ instead of $\frac{1}{27}$ as required by equation (1.35b). In this case it is easy to detect the cause of the mutual dependence: The occurrence of two events implies the occurrence of the third and therefore we have $P(A_1A_2) = P(A_2A_3) = P(A_3A_1) = P(A_1A_2A_3)$.

Generalization to n events is straightforward [76, p. 128]: The events $A_1\,,\, A_2\,,\, \ldots\,,\, A_n$ are mutually independent if the multiplication rules apply for all combinations $1 \le i < j < k < \ldots \le n$ and hence we have $2^n - n - 1$ conditions,

$$\begin{aligned} P(A_iA_j) \;&=\; P(A_i)\cdot P(A_j) \\ P(A_iA_jA_k) \;&=\; P(A_i)\cdot P(A_j)\cdot P(A_k) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ P(A_1A_2\ldots A_n) \;&=\; P(A_1)\cdot P(A_2)\cdot \ldots \cdot P(A_n)\,, \end{aligned} \tag{1.36}$$

which have to be satisfied.[29]

Independence of random variables will be a highly relevant problem in the forthcoming chapters. Countably-valued random variables $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are defined to be *independent* if and only if for any combination $x_1, \ldots, x_n$ of real numbers the joint probabilities can be factorized:

$$P(\mathcal{X}_1 = x_1, \ldots, \mathcal{X}_n = x_n) \;=\; P(\mathcal{X}_1 = x_1)\cdot \ldots \cdot P(\mathcal{X}_n = x_n)\;. \tag{1.37}$$

An extension of equation (1.37) replaces the single values x_i by arbitrary sets S_i

$$P(\mathcal{X}_1 \in S_1, \ldots, \mathcal{X}_n \in S_n) \;=\; P(\mathcal{X}_1 \in S_1)\cdot \ldots \cdot P(\mathcal{X}_n \in S_n)\;.$$

In order to proof this extension we sum over all points belonging to the sets $S_1, \ldots, S_n$:

[29] The number of conditions consists of $\binom{n}{2}$ equations in the first line, $\binom{n}{3}$ equations in the second line, and so on, down to $\binom{n}{n} = 1$ in the last line. The summation yields $\sum_{i=2}^{n}\binom{n}{i} = (1+1)^n - \binom{n}{1} - \binom{n}{0} = 2^n - n - 1$.

$$\sum_{x_1 \in S_1} \cdots \sum_{x_n \in S_n} P(\mathcal{X}_1 = x_1, \ldots, \mathcal{X}_n = x_n) \;=$$

$$= \sum_{x_1 \in S_1} \cdots \sum_{x_n \in S_n} P(\mathcal{X}_1 \in S_1) \cdot \ldots \cdot P(\mathcal{X}_n \in S_n) \;=$$

$$= \left(\sum_{x_1 \in S_1} P(\mathcal{X}_1 \in S_1) \right) \cdot \ldots \cdot \left(\sum_{x_n \in S_n} P(\mathcal{X}_n \in S_n) \right) \;,$$

which is equal to the right hand side of the equation to be proven. □

Since the factorization is fulfilled for arbitrary sets $S_1, \ldots S_n$ it holds also for all subsets of $(\mathcal{X}_1 \ldots \mathcal{X}_n)$ and accordingly the events

$$\{\mathcal{X}_1 \in S_1\}, \ldots, \{\mathcal{X}_n \in S_n\}$$

are also independent. It can also be verified that for arbitrary real-valued functions $\varphi_1, \ldots, \varphi_n$ on $]-\infty, +\infty[$ the random variables $\varphi_1(\mathcal{X}_1), \ldots, \varphi_n(\mathcal{X}_n)$ are independent too.

Independence can also be extended in straightforward manner to the joint distribution function of the random vector $\vec{\mathcal{V}} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$

$$F_{\vec{\mathcal{V}}}(x_1, \ldots, x_n) \;=\; F_{\mathcal{X}_1}(x_1) \cdot \ldots \cdot F_{\mathcal{X}_n}(x_n) \;,$$

where the $F_{\mathcal{X}_j}$'s are the marginal distributions of the $\mathcal{X}_j$'s , $1 \leq j \leq n$. Thus, the marginal distributions determine the joint distribution in case of independence of the random variables.

1.7 Probability measure on uncountable sample spaces[30]

So far we were dealing with countable, finite or infinite, sample spaces. A new situation arises when the sample space Ω is uncountable (see, e.g., figure 1.5) and this is very common, for example, for continuous variables defined on non-zero, open or closed segments of the real line, $]a,b[$, $]a,b]$, $[a,b[$, or $[a,b]$ for $a<b$, respectively. The most straightforward way to illustrate a measure is to assign length, area, volume, or generalized volume to a set. Sometimes mass of a homogeneous object is easier to visualize than volume. In order to illustrate the problem we may ask a very natural question: Does every arbitrary proper subset of the real line, $-\infty < x < +\infty$, have a length? It seems trivial to assign length 1 to the interval $[0,1]$ and length $b-a$ to the interval $[a,b]$ with $a \leq b$. Now we assign mass to sets in the sense of bars of uniform density. For example, we attribute a bar of length 1 that has mass 1 to $[0,1]$, and accordingly, a bar of mass $b-a$ to $[a,b]$, two bars corresponding to the set $[0,1]\cup[3,5]$ together have mass 3, etc. The question now is: What is the mass of the set of the rational numbers $\mathbb{Q}$ given the mass of the interval $[0,1]$ is one? Since the rational numbers are *dense* in the real numbers,[31] any nonnegative value for the mass of the rational numbers may appear acceptable. The real numbers, however, are uncountable and so are the irrational numbers, $\mathbb{R}\backslash\mathbb{Q}$. Assigning mass $b-a$ to the interval $[a,b]$ leaves no weight for the rational numbers and indeed the rational numbers $\mathbb{Q}$ have measure zero like any other set of countably many objects, more precisely Lebesgue measure zero, $\lambda(\mathbb{Q})=0$ as we shall see in the forthcoming sections the Lebesgue measure indeed assigns precisely the values given above to intervals on the real axis: $\lambda([0,1])=1$ or $\lambda([a,b])=b-a$. The real line $\mathbb{R}$ allows for the definition of a Borel measure, which assigns also $\mu([a,b])=b-a$ for every interval $[a,b]$. It is defined on the σ-algebra of the Borel sets $\mathcal{B}(\mathbb{R})$ and this is the smallest σ-algebra that contains the open intervals of $\mathbb{R}$. In practice, however, the Borel measure is not the most useful measure defined on the σ-algebra of Borel sets, because the Lebesgue-measure on Borel sets is a *complete measure* in contrast to the Borel measure. A complete measure refers to a complete measure space in which every subset of every null set is measurable with measure zero. Indeed, the Lebesgue measure λ is an extension of the Borel measure μ in the sense that every Borel-measurable set E is also a Lebesgue-measurable set, and the two measures coincide on Borel sets: $\lambda(E)=\mu(E)$.

Before we develop a measure for uncountable sample spaces we recall the three indispensable properties of probability measures $\mu : \mathcal{F} \to [0,\infty[$ with $\mathcal{F}$

[30] This section can be skipped by readers who are willing to except the fact that all uncountable sample spaces needed in the forthcoming discussions are measurable notwithstanding the existence of non-measurable sets.

[31] A subset D of real numbers is said to be *dense* in $\mathbb{R}$ if every arbitrarily small interval $]a,b[$ with $a<b$ contains at least one element of D. Accordingly, the set of rational numbers $\mathbb{Q}$ as well as the set of irrational numbers $\mathbb{R}\backslash\mathbb{Q}$ are dense in $\mathbb{R}$.

being a measurable collection of events A: (i) nonnegativity, $\mu(A) \geq 0 \; \forall \; A \in \mathcal{F}$,
(ii) normalization, $P(\Omega) = 1$, and (iii) additivity, $\mu(A) + \mu(B) = \mu(A \cup B)$ provided $P(A \cap B) = \emptyset$. Problems concerning measurability arise from the impossibility to assign a probability to every subset of Ω. The task is to develop measures for uncountable sets that are derived from collections of subsets, whose cardinality is $\aleph_0$, infinite but countable. To do this in full generality is highly demanding and it requires advanced mathematical techniques, in particular sufficient knowledge of measure theory. For the probability concept we are using here, however, the simplest bridge from countability to uncountability is sufficient and we need only derive a measure for sets of a certain family of sets called *Borel sets*, $\mathcal{B} \subset \Omega$. For this goal the introduction of σ-additivity (1.15) and Lebesgue measure $\lambda(A)$ is sufficient, and as said, σ-additivity comes close to mass in the above given example. Still unanswered remains the question whether unmeasurable sets do exist at all.

1.7.1 Existence of non-measurable sets

In the case of a countable sample space the powerset $\Pi(\Omega)$ is the set of all subsets of the sample space Ω and contains the results of all set theoretic operations of section 1.4. Although is seems straightforward to proceed in the same way for uncountable sample spaces Ω, it turns out, however, that the powerset $\Pi(\Omega)$ is too large, because it contains uncountably many subsets. A general proof of this conjecture is difficult but Giuseppe Vitali [294, 295] provided a proof by means of contradiction that mappings $P\colon \Pi(\Omega) \to [0,1]$ exist, which fulfils all three indispensable properties for probabilities, do not exists for the infinitely repeated coin flip, $\Omega = \{0,1\}^{\mathbb{N}}$ [97, p. 9,10]:

(N) normalization: $P(\Omega) = 1$,
(A) σ-additivity: for pairwise disjoint events $A_1, A_2, \ldots \subset \Omega$ holds

$$P\left(\bigcup_{i\geq 1} A_i\right) \;=\; \sum_{i\geq 1} P\left(A_i\right) \;, \quad \text{and}$$

(I) invariance: For all $A \subset \Omega$ and $k \geq 1$ holds $P(\hat{T}_k A) = P(A)$, where $\hat{T}_k$ is an operator that inverts the outcome of the k-th toss,

$$T_k \;:\; \omega = (\omega_1, \ldots, \omega_{k-1}, \omega_k, \omega_{k+1} \ldots) \to (\omega_1, \ldots, \omega_{k-1}, 1-\omega_k, \omega_{k+1} \ldots),$$

and $T_k A = \{T_k(\omega) : \omega \in A\}$ is the image of A under the operation T_k.

The first two conditions are the criteria for probability measures and the invariance condition (I) is specific for coin flipping and encapsulates the properties derived from the uniform distribution, $\mathcal{U}_\Omega$.

In order to proof the conjecture of incompatibility with all three conditions we define an equivalence relation '$\sim$' in Ω: $\omega \sim \omega'$ iff $\omega_k = \omega'_k$ for all sufficiently long sequences ($n \geq k$). According to the axiom of choice[32] guarantees the existence of a set $A \subset \Omega$, which contains exactly one element of each equivalence class.

We define $\mathcal{S} = \{S \subset \mathbb{N} : |S| < \infty\}$ the set containing all finite subsets of $\mathbb{N}$. Since $\mathcal{S}$ is the union of a countable number of finite sets, $\{S \subset \mathbb{N} : \max S = m\}$ with $m \in \mathbb{N}$, $\mathcal{S}$ is countable too. For $S = \{k_1, \ldots, k_n\} \in \mathcal{S}$ we define $T_S = \prod_{k_i \in S} T_{k_i} = T_{k_1} \circ \ldots \circ T_{k_n}$ the simultaneous flip of the digits in S. Then we have:

(i) $\Omega = \bigcup_{S \in \mathcal{S}} T_S A$ since for each sequence $\omega \in \Omega$ there exists an $\omega' \in A$ with $\omega \sim \omega'$, and accordingly an $S \in \mathcal{S}$ such that $\omega = T_S\, \omega' \in T_S\, A$,

(ii) the sets $(T_S A)_{S \in \mathcal{S}}$ are pairwise disjoint: If $T_S A \cup T_{S'} A \neq \emptyset$ were true for $S, S' \in \mathcal{S}$ then there existed an $\omega, \omega' \in A$ with $T_S\, \omega = T_{S'}\, \omega'$ and accordingly $\omega \sim T_S\, \omega = T_S\, \omega \sim \omega'$. By definition of A we had $\omega = \omega'$ and hence $S = S'$.

Applying the properties (N), (A), and (I) of the probability P we find

$$1 \;=\; P(\Omega) \;=\; \sum_{S \in \mathcal{S}} P(T_S A) \;=\; \sum_{S \in \mathcal{S}} P(A)\ . \tag{1.38}$$

Equation (1.38) cannot be fulfilled for infinitely large series of coin tosses, since all values $P(A)$ or $P(T_S A)$ are the same and infinite summation by σ-additivity (A) is tantamount to an infinite sum of the same number, which yields either 0 or ∞ but never 1 as required to fulfil (N). It is straightforward to show that the set of all binary strings with countably infinite length, $B = \{0,1\}^{\mathbb{N}}$, is bijective to the unit interval $[0,1]$. A more or less explicit bijection $f : B \leftrightarrow [0,1]$ can be obtained by defining an auxiliary function

$$g(x) \;:=\; \sum_{k=1}^{\infty} \frac{x_k}{2^k}\ .$$

which interprets a binary string $x = (x_1, x_2, \ldots) \in B$ as an infinite binary fraction

$$\frac{x_1}{2} + \frac{x_2}{4} + \ldots\ .$$

The function $g(x)$ maps B only *almost* bijectively onto $[0,1]$, because the dyadic rationals in $]0,1[$ have two preimages each, for example $g(1,0,0,0,\ldots) = g(0,1,1,1,\ldots) = \frac{1}{2}$. In order to fix this problem we reorder the rationals:

$$\big(q_n\big)_{n \geq 1} \;=\; \left(\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \frac{1}{16}, \ldots\right)\ ,$$

[32] Axiom of choice: Suppose that $A_\theta : \theta \in \Theta$ is a decomposition of Ω into nonempty sets. The axiom of choice exists at least one set C, which contains exactly one point from each A_θ: $C \cap A_\theta$ is a singleton for each θ in Θ (see [20, p. 572] and [50]).

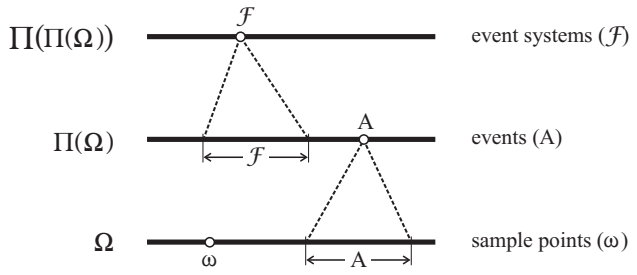

Fig. 1.14 Conceptual levels of sets in probability theory. The lowest level is the sample space Ω, it contains the sample points or individual results ω as elements, and events A are subsets of Ω: $\omega \in \Omega$ and $A \subset \Omega$. The next higher level is the powerset $\Pi(\Omega)$. Events A are its elements and event systems $\mathcal{F}$ constitute its subsets: $A \in \Pi(\Omega)$ and $\mathcal{F} \subset \Pi(\Omega)$. The highest level finally is the power powerset $\Pi\big(\Pi(\Omega)\big)$ that houses event systems $\mathcal{F}$ as elements: $\mathcal{F} \in \Pi\big(\Pi(\Omega)\big)$ (Drawn after [97, p 11]).

and find for the bijection

$$f(x) \;:=\; \begin{cases} q_{2n-1} & \text{if } g(x) = q_n\,, \text{ and } x_k = 1 \text{ for almost all } k\,, \\ q_{2n} & \text{if } g(x) = q_n\,, \text{ and } x_k = 0 \text{ for almost all } k\,, \\ g(x) & \text{otherwise}\,. \end{cases} \tag{1.39}$$

Hence Vitali's theorem applies as well to the unit interval $[0, 1]$, where we are also dealing with an uncountable number of non-measurable sets. For other more detailed proofs of Vitali's theroem see, e.g., [20, p. 47].

Accordingly, the proof of Vitali's theorem demonstrates the existence of non-measurable subsets of the real numbers called Vitali sets – precisely subsets of the real numbers that are not Lebesgue measurable (see next subsection 1.7.2). The problem to be solved is a reduction of the powerset to an event system $\mathcal{F}$ such that the subsets causing the lack of countability are left aside (figure 1.14).

1.7.2 Borel σ-algebra and Lebesgue measure

Before we define minimal requirements for an event system $\mathcal{F}$, we consider the three levels of sets in set theory that are relevant for our construction (figure 1.14). The objects on the lowest level are the sample points corresponding to individual results, $\omega \in \Omega$. The next higher level is the powerset $\Pi(\Omega)$ housing the events $A \in \Pi(\Omega)$. The elements of the powerset are subsets of the sample space, $A \subset \Omega$. To illustrate the role of event systems $\mathcal{F}$ we

need a higher level, the powerset of the powerset, $\Pi\big(\Pi(\Omega)\big)$: Event systems $\mathcal{F}$ are elements of the power powerset, $\mathcal{F} \in \Pi\big(\Pi(\Omega)\big)$ and subsets of the powerset, $\mathcal{F} \subset \Pi(\Omega)$.[33]

The minimal requirements for an *event system* $\mathcal{F}$ are summarized in the following definition of a *σ-algebra* on Ω with $\Omega \neq \emptyset$ and $\mathcal{F} \subset \Pi(\Omega)$:

(1) $\Omega \in \mathcal{F}$,
(2) $A \in \mathcal{F} \implies A^c := \Omega \backslash A \in \mathcal{F}$, and
(3) $A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_{i \geq 1} A_i \in \mathcal{F}$.

Condition (2), demanding the existence of a complement A^c for every subset $A \in \mathcal{F}$, defines the logical negation as expressed by the difference between the entire sample space and the event A, and condition (3) represents the logical *or* operation. The pair $(\Omega, \mathcal{F})$ is called an *event space* or a *measurable space*. From the three properties (1) to (3) follow other properties: The intersection, for example, is the complement of the union of the complements $A \cap B = (A^c \cup B^c)^c \in \mathcal{F}$, and the argument is easily extended to the intersection of countable many subsets of $\mathcal{F}$ that belongs to $\mathcal{F}$ as well. Thus, a σ-algebra is closed under the operations '^{c}', '$\cup$' and '$\cap$'.[34] Trivial examples of σ-algebras are $\{\emptyset, \Omega\}$, $\{\emptyset, A, A^c, \Omega\}$ or the family of all subsets. The Borel σ-algebra on Ω is the smallest σ-algebra, which contains all open sets or equivalently, all closed sets.

A construction principle for σ-algebras starts out from some event system $\mathcal{G} \subset \Pi(\Omega)$ (for $\Omega \neq \emptyset$) that is sufficiently small and otherwise arbitrary. Then, there exists exactly one smallest σ-algebra $\mathcal{F} = \sigma(\mathcal{G})$ in Ω with $\mathcal{F} \supset \mathcal{G}$, and we call $\mathcal{F}$ the σ-algebra induced by $\mathcal{G}$. In other words, $\mathcal{G}$ is the generator of $\mathcal{F}$. Here are three important examples:

(i) the powerset with Ω being countable where $\mathcal{G} = \big\{\{\omega\} : \omega \in \Omega\big\}$ is the system of the subsets of Ω containing a single element, the σ-algebra $\sigma(\mathcal{G}) = \Pi(\Omega)$, each $A \in \Pi(\Omega)$ is countable, and $A = \bigcup_{\omega \in A}\{\omega\} \in \sigma(\mathcal{G})$ (countable sample spaces as discussed in section 1.5),

(ii) the Borel σ-algebra $\mathcal{B}$ containing all open or all closed sets in one dimension (the uncountable sample space of real numbers $\Omega = \mathbb{R}$, see below), and

(iii) the product σ-algebra for sample spaces Ω that are Cartesian products of sets $\mathcal{E}_k$, $\Omega = \prod_{k \in \mathcal{I}} \mathcal{E}_k$ where $\mathcal{I}$ is a set of indices with $\mathcal{I} \neq \emptyset$. We assume $\mathcal{B}_k$ is a Borel σ-algebra on $\mathcal{E}_k$ with $\mathcal{X}_k : \Omega \to \mathcal{E}_k$ being the projection onto the k-th coordinate and the generator
$$\mathcal{G} = \{\mathcal{X}_k^{-1} A_k : k \in \mathcal{I}, A_k \in \mathcal{B}_k\}$$
is the system of all sets in Ω, which are determined by an event on a single coordinate. Then, $\bigotimes_{k \in I} \mathcal{B}_k := \sigma(\mathcal{G})$ is called the product σ-algebra

[33] Recalling the situation in the case of countability we were choosing the entire power set $\Pi(\Omega)$ as reference instead of a smaller event system $\mathcal{F}$.

[34] A family of sets is called closed under an operation if the operation can be applied a countable number of times without producing a set that lies outside the family.

of the sets $\mathcal{B}_k$ on Ω. In the important case of equivalent Cartesian coordinates, $\mathcal{E}_k = \mathcal{E}$ and $\mathcal{B}_k = \mathcal{B}$ for all $k \in \mathcal{I}$, the short-hand notion $\mathcal{B}^{\bigotimes \mathcal{I}}$ is common. The Borel σ-algebra on $\mathbb{R}^n$ is represented by the n-dimensional product σ-algebra of the Borel σ-algebra $\mathcal{B} = \mathcal{B}^1$ on $\mathbb{R}$ (for $n = 1$ one commonly writes $\mathcal{B}$ instead of $\mathcal{B}^1$), or $\mathcal{B}^n = \mathcal{B}^{\bigotimes n}$ (Cartesian product sample spaces, $\Omega = \mathbb{R}^n$).

All three examples are required for the understanding of probability measures: (i) The powerset provides the frame for discrete sample spaces, (ii) the Borel σ-algebra to be discussed below sets the stage for one-dimensional continuous sample spaces, and (iii) the product σ-algebra represents the natural extension from one dimension to the n-dimensional Cartesian space.

For the construction of the Borel σ-algebra[35] we define a generator representing the set of all compact cuboids in n-dimensional Cartesian space, $\Omega = \mathbb{R}^n$, which have rational corners,

$$\mathcal{G} \;=\; \left\{ \prod_{k=1}^{n} [a_k, b_k] : a_k < b_k; a_k, b_k \in \mathbb{Q} \right\} \tag{1.40}$$

where $\mathbb{Q}$ is the set of all rational numbers. The σ-algebra induced by this generator is denoted as the Borel σ-algebra, $\mathcal{B}^n := \sigma(\mathcal{G})$ on $\mathbb{R}^n$ and each $A \in \mathcal{B}^n$ is a Borel set.

Five properties of the Borel σ-algebra are useful for application and for imagination of its enormous size.

(i) Each open set ']..[' $A \subset \mathbb{R}^n$ is Borelian. Every $\omega \in A$ has a neighborhood $Q \in \mathcal{G}$ with $Q \subset A$ and therefore we have $A = \bigcup_{Q \in \mathcal{G},\, Q \subset A} Q$ representing a union of countably many sets in $\mathcal{B}^n$, which follows from condition (3) of σ-algebras.

(ii) Each closed set '[..]' $A \subset \mathbb{R}^n$ is Borelian since A^c is open and Borelian according to item (i).

(iii) The σ-algebra $\mathcal{B}^n$ cannot be described in a constructive way, because is consists of much more than the union of cuboids and their complements. In order to create $\mathcal{B}^n$ the operation of adding complements and countable unions has to be repeated as often as there are countable ordinal numbers (and this leads to uncountable many times [19, pp.24, 29]). It is sufficient to memorize for practical purposes that $\mathcal{B}^n$ covers almost all sets in $\mathbb{R}^n$ – but not all of them.

(iv) The Borel σ-algebra $\mathcal{B}$ on $\mathbb{R}$ is generated not only by the system of compact sets (1.40) but also by the system of closed left-hand open infinite intervals:

$$\tilde{\mathcal{G}} \;=\; \{]-\infty, c]; c \in \mathbb{R}\} \;. \tag{1.40'}$$

Condition (2) requires $\tilde{\mathcal{G}} \subset \mathcal{B}$ and – because of minimality of $\sigma(\tilde{\mathcal{G}})$ – $\sigma(\tilde{\mathcal{G}}) \subset \mathcal{B}$ too. Alternatively, $\sigma(\tilde{\mathcal{G}})$ contains all left-open intervals since

[35] Sometimes a Borel σ-algebra is also called a Borel field.

$]a,b] =]-\infty,b] \setminus]-\infty,a]$ and also all compact or closed intervals since $[a,b] = \bigcap_{n\geq 1}]a-\frac{1}{n},b]$ and accordingly also the σ-algebra $\mathcal{B}$ generated from these intervals (1.40). In full analogy $\mathcal{B}$ is generated from all open left-unbounded, from all closed and open right-unbounded intervals.

(v) The event system $\mathcal{B}^n_\Omega = \{A \cap \Omega : A \in \mathcal{B}^n\}$ on $\Omega \subset \mathbb{R}^n$, $\Omega \neq \emptyset$ represents a σ-algebra on ω, which is denoted as the Borel σ-algebra on Ω.

All intervals discussed in items (i) to (iv) are Lebesgue measurable while other sets are not.

The Lebesgue measure is the conventional mean of assigning lengths, areas, and volumes to subsets of three-dimensional Euclidean space and in formal Cartesian spaces to objects with higher dimensional volumes. Sets to which generalized volumes[36] can be assigned are called Lebesgue measurable and the measure or the volume of such a set A is denoted by $\lambda(A)$. The Lebesgue measure on $\mathbb{R}^n$ has the following properties:

(1) If A is a Lebesgue measurable set, then $\lambda(A) \geq 0$.
(2) If A is a Cartesian product of intervals, $I_1 \otimes I_2 \otimes \ldots \otimes I_n$, then A is Lebesgue measurable and $\lambda(A) = |I_1| \cdot |I_2| \cdot \ldots \cdot |I_n|$.
(3) If A is Lebesgue measurable, its complement A^c is so too.
(4) If A is a disjoint union of countably many disjoint Lebesgue measurable sets, $A = \bigcup_k A_k$, then A is Lebesgue measurable and $\lambda(A) = \sum_k \lambda(A_k)$.
(5) If A and B are Lebesgue measurable and $A \subset B$, then $\lambda(A) \leq \lambda(B)$ holds.
(6) Countable unions and countable intersections of Lebesgue measurable sets are Lebesgue measurable.[37]
(7) If A is an open or closed subset or Borel set of $\mathbb{R}^n$, then A is Lebesgue measurable.
(8) The Lebesgue measure is strictly positive on non-empty open sets, and so its support is the entire $\mathbb{R}^n$.
(9) If A is a Lebesgue measurable set with $\lambda(A) = 0$, called a null set, then every subset of A is also a null set, and every subset of A is measurable.
(10) If A is Lebesgue measurable and $\mathbf{r}$ is an element of $\mathbb{R}^n$, then the translation of A by $\mathbf{r}$ that is defined by $A + \mathbf{r} = \{\mathbf{a} + \mathbf{r} : \mathbf{a} \in A\}$ is also Lebesgue measurable and has the same measure as A.
(11) If A is Lebesgue measurable and $\delta > 0$, then the dilation of A by δ defined by $\delta A = \{\delta \mathbf{r} : \mathbf{r} \in A\}$ is also Lebesgue measurable and has measure $\delta^n \lambda(A)$.

---

[36] We generalize volume here to arbitrary dimension n: The *generalized volume* for $n = 1$ is a length, for $n = 2$ an area, for $n = 3$ a (conventional) volume and for arbitrary dimension n a cuboid in n-dimensional space.

[37] This is not a consequence of items (3) and (4): A family of sets, which is closed under complements and countable disjoint unions, need not be closed under countable non-disjoint unions. Consider, for example, the set

$$\{\emptyset, \{1,2\}, \{1,3\}, \{2,4\}, \{3,4\}, \{1,2,3,4\}\}.$$

(12) In generalization of items (10) and (11), if L is a linear transformation and A is a measurable subset of $\mathbb{R}^n$, then $T(A)$ is also measurable and has the measure $|\det(T)|\,\lambda(A)$.

All twelve items listed above can be succinctly summarized in one lemma:

The Lebesgue measurable sets form a σ-algebra on $\mathbb{R}^n$ containing all products of intervals, and λ is the unique complete translation-invariant measure on that σ-algebra with

$$\lambda\big([0,1]\otimes[0,1]\otimes\ldots\otimes[0,1]\big)=1.$$

We conclude this section on Borel σ-algebra and Lebesgue measure by mentioning a few characteristic and illustrative examples:

- Any closed interval $[a,b]$ of real numbers is Lebesgue measurable, and its Lebesgue measure is the length $b-a$. The open interval $]a,b[$ has the same measure, since the difference between the two sets consists of the two endpoint a and b only and has measure zero.
- Any Cartesian product of intervals $[a,b]$ and $[c,d]$ is Lebesgue measurable and its Lebesgue measure is $(b-a)\cdot(d-c)$ the area of the corresponding rectangle.
- The Lebesgue measure of the set of rational numbers in an interval of the line is zero, although this set is dense in the interval.
- The Cantor set[38] is an example of an uncountable set that has Lebesgue measure zero.
- Vitali sets are examples of sets that are not measurable with respect to the Lebesgue measure.

In the forthcoming sections we make use of the fact that the continuous sets on the real axes become countable and Lebesgue measurable if rational numbers are chosen as beginnings and end points of intervals. Hence, we can work with real numbers with almost no restriction for practical purposes.

[38] The Cantor set is generated from the interval $[0,1]$ through consecutively taking out the open middle third: $[0,1]\to[0,\frac{1}{3}]\cup[\frac{2}{3},1]\to[0,\frac{1}{9}]\cup[\frac{2}{9},\frac{1}{3}]\cup[\frac{2}{3},\frac{7}{9}]\cup[\frac{8}{9},1]\to\ldots$. An explicit formula for the set is: $C=[0,1]\setminus\bigcup_{m=1}^{\infty}\bigcup_{k=0}^{(3^{m-1}-1)}\,\big]\frac{3k+1}{3^m},\frac{3k+2}{3^m}\big[$.

1.8 Limits and integrals

Before we can discuss now continuous random variables and their distributions, we need to mention a few technicalities concerning the definition of limits and the methods of integration. Limits of sequences are required for problems convergence of and for approximations to random variables. Taking limits of stochastic variables often needs some care and problems might arise because there is ambiguity in the definition of limits and therefore precise definitions are required. We introduced already functions like the probability mass function (pmf) and the cumulative probability distribution function (cdf) of discrete random variables that contain peaks and steps, which cannot be subjected to conventional Riemannian integration.

1.8.1 Limits of series of random variables

A sequence of random variables, $\mathcal{X}_n$, is defined on a probability space Ω and it is assumed to have the limit

$$\mathcal{X} \;=\; \lim_{n\to\infty} \mathcal{X}_n \;. \tag{1.41}$$

The probability space Ω, we assume now, has elements ω which have a probability density $p\,(\omega)$. Four different definitions of the limit are common in probability theory [93, pp.40,41].

Almost certain limit: The series $\mathcal{X}_n$ converges *almost certainly* to $\mathcal{X}$ if for all ω except a set of probability zero

$$\mathcal{X}(\omega) \;=\; \lim_{n\to\infty} \mathcal{X}_n(\omega) \;. \tag{1.42}$$

is fulfilled and each realization of $\mathcal{X}_n$ converges to $\mathcal{X}$.

Limit in the mean: The limit in the mean or the mean square limit of a series requires that the mean square deviation of $\mathcal{X}_n(\omega)$ from $\mathcal{X}(\omega)$ vanishes in the limit and the condition is

$$\lim_{n\to\infty} \int_\Omega \mathrm{d}\omega\, p\,(\omega)\Big(\mathcal{X}_n(\omega) - \mathcal{X}(\omega)\Big)^2 \;\equiv\; \lim_{n\to\infty} \big\langle (\mathcal{X}_n - \mathcal{X})^2 \big\rangle \;=\; 0 \;. \tag{1.43}$$

The mean square limit is the standard limit in Hilbert space theory and it is commonly used in quantum mechanics.

Stochastic limit: The limit in probability also called the stochastic limit fulfils the condition: $\mathcal{X}$ is the stochastic limit if for any $\varepsilon > 0$ the relation

$$\lim_{n\to\infty} P(|\mathcal{X}_n \;-\; \mathcal{X}| > \varepsilon) \;=\; 0 \;. \tag{1.44}$$

Limit in distribution: Probability theory uses also a weaker form of convergence than the previous three limits, the limit in distribution, which requires that for any continuous and bounded function $f(x)$ the relation

$$\lim_{n\to\infty} \langle f(\mathcal{X}_n)\rangle \;\xrightarrow{d}\; \langle f(\mathcal{X})\rangle \tag{1.45}$$

holds, where the symbol "$\xrightarrow{d}$" stand for *convergence in distribution*. This limit, for example, is particularly useful for characteristic functions (section 2.2.3), $\phi(s) = \int_{-\infty}^{+\infty} \exp(\dot{\imath}xs) f(x)\,\mathrm{d}x$: If two characteristic functions approach each other, the probability density of $\mathcal{X}_n$ converges to that of $\mathcal{X}$.

Finally we mention stringent conditions on convergence of functions that will be also important for probability distributions. We distinguish *pointwise convergence* and *uniform convergence*. for functions. A series of functions $f_0(x), f_1(x), f_2(x), \ldots$ is defined on some interval $I \in \mathbb{R}$. The series converges pointwise to the function $f(x)$ if the limit

$$\lim_{n\to\infty} f_n(x) \;=\; f(x) \;\;\forall\;\; x \in I \;, \tag{1.46}$$

is fulfilled for every point x. It is easily verified that a series of functions can be written as a sum of functions whose convergence is to be tested:

$$\begin{aligned} f(x) \;&=\; \lim_{n\to\infty} f_n(x) \;=\; \lim_{n\to\infty} \sum_{i=1}^{n} g_i(x) \;, \\ g_i(x) \;&=\; \varphi_{i-1}(x) - \varphi_i(x) \;\text{ and hence }\; f_n(x) \;=\; \varphi_0(x) - \varphi_n(x) \;, \end{aligned} \tag{1.47}$$

because $\sum_{i=1}^{n} g_i(x)$ expressed in the functions φ_i is a telescopic sum. An example of a series of curves with $\varphi_n(x) = (1 + nx^2)^{-1}$ and accordingly $f_n(x) = nx^2 \,/\, (1+nx^2)$ showing pointwise convergence is shown in figure 1.15. It is easily verified that the limit takes on the form:

$$f(x) \;=\; \lim_{n\to\infty} \frac{nx^2}{1 + nx^2} \;=\; \begin{cases} 1 & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases} .$$

All functions $f_n(x)$ are continuous on the interval $]-\infty, \infty[$ but the limit $f(x)$ is discontinuous at $x = 0$. An interesting historical detail is mentioned here: In 1821 the famous mathematician Augustin Louis Cauchy gave the wrong answer to the question whether or not infinite sums of continuous functions are necessarily continuous and his obvious error had been corrected only thirty years later. It is easy to visualize that pointwise convergence is compatible with discontinuities in the convergence limit: At two neighboring points the convergent series may have very different limits. There are many examples of series of functions, which have a discontinuous infinite limit, two

further cases that we shall need later on are $f_n(x) = x^n$ with $I = [0,1] \in \mathbb{R}$ and $f_n(x) = \cos(\pi x)^{2n}$ on $I =]-\infty, \infty[\in \mathbb{R}$.

Uniform convergence is the stronger condition, which guarantees among other things that the limit of a series of continuous functions is continuous. It can be defined in terms of equation (1.47): The sum $f_n(x) \sum_{i=1}^{n} g_i(x)$ with $\lim_{n\to\infty} f_n(x) = f(x)$ and $x \in I$ is uniformly convergent in the interval $x \in I$ for every given positive error bound ϵ if there exists a value $\nu \in \mathbb{N}$ such that for any $\nu \geq n$ the relation $|f(x) - f_\nu(x)| < \epsilon$ is fulfilled for all $x \in I$. In compact form the convergence condition may be expressed by

$$\lim_{n\to\infty} \sup\{|f_n(x) - f(x)|\} = 0 \quad \forall \quad x \in I \; . \tag{1.48}$$

A simple but illustrative example is given by the power series on the unit interval, $f(x) = \lim_{n\to\infty} x^n$ with $x \in [0,1]$ which converges pointwise to the discontinuous function $f(x) = 1$ for $x = 1$ and 0 otherwise. A slight modification to $f(x) = \lim_{n\to\infty} x^n/n$ leads to a uniformly converging series, because $f(x) = 0$ is now valid for the entire domain $[0,1]$ (including the point $x = 1$).

1.8.2 Stieltjes integration

Here we provide a short repetition of some generalizations of the conventional Riemann integral, which are important in probability theory. The sketch presented in figure 1.16 compares the Riemann and the Lebesgue approach to integration. Stieltjes integration is a generalization of Riemann or Lebesgue integration, which allows for the calculation of integrals over step functions as they occur, for example, in the context of properties derived from cumulative probability distributions. The *Stieltjes integral* is commonly written in the form

$$\int_a^b g(x)\, \mathrm{d}h(x) \; . \tag{1.49}$$

Herein $g(x)$ is the integrand, $h(x)$ is the integrator, and the conventional Riemann integral is retained for $h(x) = x$. The integrator can be visualized best as a weighting function for the integrand. In case $g(x)$ and $h(x)$ are continuous and continuously differentiable the Stieltjes integral can be resolved by partial integration:

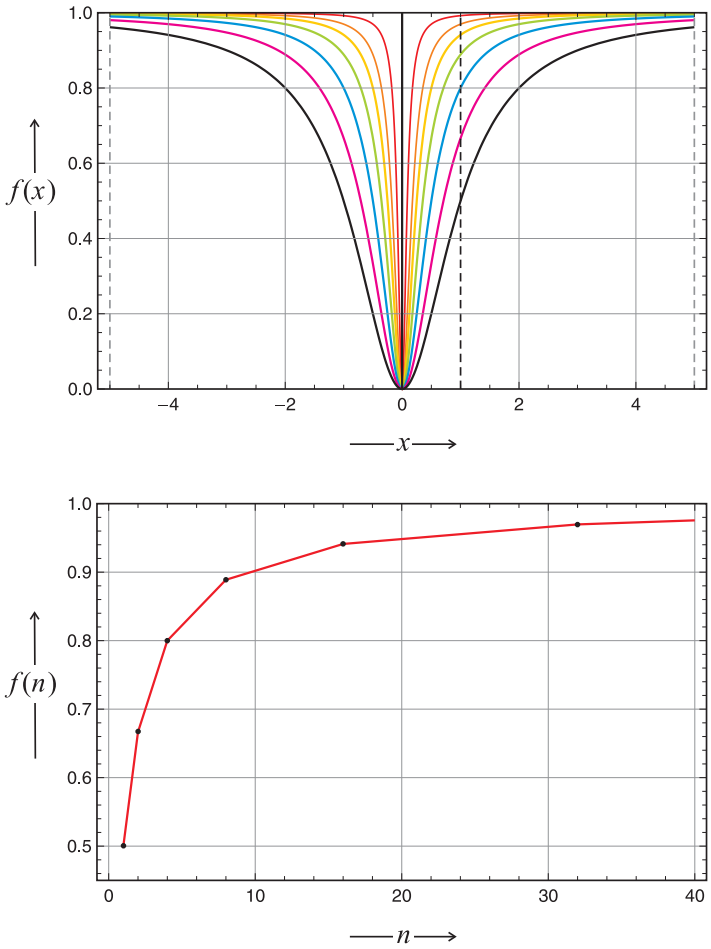

Fig. 1.15 Pointwise convergence. The upper part shows the convergence of the series of functions $f_n(x) = nx^2/(1+nx^2)$ to the limit $\lim_{n\to\infty} f_n(x) = f(x)$ on the real axis $I =]-\infty, \infty[$. The lower plot illustrates the convergence as a function of n at the point $x = 1$. Color code of the upper plot: n=1, black; n=2, violet; n=4, blue; n=8, chartreuse; n=16, yellow; n=32, orange; and n=128, red.

$$\begin{aligned}\int_a^b g(x)\,\mathrm{d}h(x) &= \int_a^b g(x)\,\frac{\mathrm{d}h(x)}{\mathrm{dx}}\,\mathrm{d}x \;= \\ &= \Big(g(x)h(x)\Big)\Big|_{x=a}^{b} - \int_a^b \frac{\mathrm{d}g(x)}{\mathrm{dx}}\,h(x)\,\mathrm{d}x \;= \\ &= g(b)h(b) - g(a)h(a) - \int_a^b \frac{\mathrm{d}g(x)}{\mathrm{dx}}\,h(x)\,\mathrm{d}x\;.\end{aligned}$$

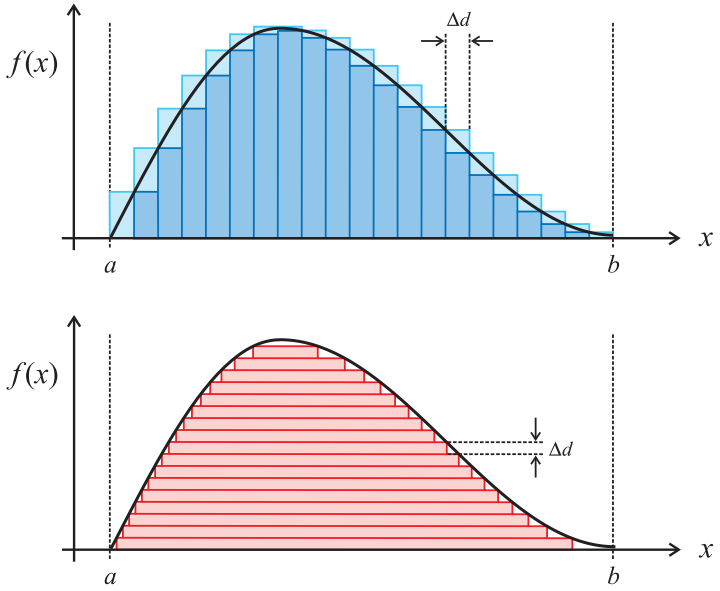

Fig. 1.16 Comparison of Riemann and Lebesgue integrals. In the conventional Riemannian-Darboux integration† the integrand is embedded between an upper sum (light blue) and a lower sum (dark blue) of rectangles. The integral exists iff the upper sum and the lower sum converge to the integrand in the limit $\Delta d \to 0$. The Lebesgue integral can be visualized as an approach to calculating the area enclosed by the x-axis and the integrand through partitioning it into horizontal stripes (red) and considering the limit $\Delta d \to 0$. The definite integral $\int_a^b f(x)\,\mathrm{d}x$ is confining the integrand to a closed interval: $[a, b]$ or $a \leq x \leq b$.

† The concept of representing the integral by the convergence of two sums is due to the French mathematician Gaston Darboux. A function is Darboux integrable iff it is Riemann integrable, and the values of the Riemann and the Darboux integral are equal in case they exist.

The integrator $h(x)$, however, may also be a step function $F(x)$. For $g(x)$ being continuous and $F(x)$ making jumps at the points $x_1, \ldots, x_n \in]a, b[$ with the heights $\Delta F_1, \ldots, \Delta F_n \in \mathbb{R}$, and $\sum_{i=1}^{n} \Delta F_n \leq 1$, respectively, the Stieltjes integral is of the form

$$\int_a^b g(x)\,\mathrm{d}F(x) \;=\; \sum_{i=1}^{n} g(x_i)\,\Delta F_i \;, \tag{1.50}$$

where the limitation of $\sum_i \Delta F_i$ refers to the normalization of probabilities. With $g(x) = 1$, $b = x$ and in the limit $\lim_{a \to -\infty}$ the integral becomes identical with the (discrete) cumulative probability distribution function (cdf).

Riemann-Stieltjes integration is used in probability theory for the computation of functions of random variables, for example, for the computation of moments of probability densities (section 2.1). If $F(x)$ is the cumulative probability distribution of a random variable $\mathcal{X}$ for the discrete case, the expected value (see section 2.1) for any function $g(\mathcal{X})$ is obtained from

$$E\big(g(\mathcal{X})\big) \;=\; \int_{-\infty}^{+\infty} g(x)\,\mathrm{d}F(x) = \;\sum_i g(x_i)\,\Delta F_i \;.$$

If the random variable $\mathcal{X}$ has a probability density $f(x) = \mathrm{d}F(x)/\mathrm{d}x$ with respect to the Lebesgue measure, continuous integration can be used

$$E\big(g(\mathcal{X})\big) \;=\; \int_{-\infty}^{+\infty} g(x)\,f(x)\,\mathrm{d}x \;.$$

Important special cases are the moments: $E(\mathcal{X}^n) = \int_{-\infty}^{+\infty} x^n\,\mathrm{d}F(x)$.

1.8.3 Lebesgue integration

Lebesgue integration differs from the conventional integration in two aspects: The basis are set theory and measure theory and the integrand is partitioned in horizontal segments whereas Riemannian integration makes use of vertical slices. An important difference for nonnegative functions – like probability functions – between the two integration methods can be visualized in three dimensional space: The volume below a surface given by the function $f(x, y)$ is measured by summation of the volumes of cuboids with squares of edge length Δd, whereas the Lebesgue integral is summing the volumes of layers with thickness Δd between constant level sets. Every continuous bounded function on a compact finite interval, $f \in C[a, b]$, is Riemann integrable and also Lebesgue integrable, and the Riemann and the Lebesgue integrals coincide. The Lebesgue integral is a generalization of the Riemann integral in the sense that certain functions may be Lebesgue integrable in cases where the Riemann integral does not exist. The opposite situations might occur with improper Riemann integrals:[39] Partial sums with alternate signs may converge for the improper Riemann integral whereas Lebesgue integration leads to divergence as shown in case of the alternate harmonic series. The

[39] An improper integral is the limit of a definite integral in a series in which the endpoint of the interval of integration approaches either a finite number b at which the integrand diverges or $\pm\infty$:

$$\int_a^b f(x)\,\mathrm{d}x \;=\; \lim_{\varepsilon\to+0}\int_a^{b-\varepsilon} f(x)\,\mathrm{d}x \;\;\text{or}\;\; \lim_{b\to\infty}\int_a^b f(x)\,\mathrm{d}x \;\;\text{and}\;\; \lim_{a\to-\infty}\int_a^b f(x)\,\mathrm{d}x \;.$$

Lebesgue integral can be generalized by the Stieltjes integration technique very much in the same way as the Riemann integral does.

Lebesgue theory of integration assumes the existence of a probability space defined by the triple $(\Omega, \mathcal{F}, \mu)$, which represents the sample space Ω, a σ-algebra $\mathcal{F}$ of subsets $A \in \Omega$, and a probability measure $\mu \geq 0$ satisfying $\mu(\Omega) = 1$, respectively. The construction of the Lebesgue integral is similar to the construction of the Riemann integral: The shrinking rectangles (or cuboids in higher dimensions) of Riemannian integration is replaced by horizontal stripes of shrinking height that can be represented by simple functions. Lebesgue integrals on A over nonnegative functions,

$$\int_{\Omega} f \,\mathrm{d}\mu \quad \text{with} \quad f : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}_{\geq 0}, \mathcal{B}, \lambda) \;, \tag{1.51}$$

are defined for measurable functions f, which fulfill

$$f^{-1}\big([a,b]\big) \in \Omega \quad \text{for all} \quad a < b \;. \tag{1.52}$$

This condition is equivalent to the requirement that the pre-image of any Borel subset $[a, b]$ of $\mathbb{R}$ is an element of the event system $\mathcal{B}$. The set of measurable functions is closed under algebraic operation and also closed under certain pointwise sequential limits like

$$\sup_{k\in\mathbb{N}} \; f_k \,, \quad \liminf_{k\in\mathbb{N}} \; f_k \quad \text{or} \quad \limsup_{k\in\mathbb{N}} \; f_k \;,$$

which are measurable if the sequence of functions $(f_k)_{k\in\mathbb{N}}$ contains only measurable functions.

The construction of an integral $\int_\Omega f \,\mathrm{d}\mu = \int_\Omega f(x)\,\mu(\mathrm{d}x)$ is done in steps and we begin with the introduction of an *indicator function*:

$$\mathbf{1}_A(x) \;=\; \begin{cases} 1 & \text{iff } x \in A \\ 0 & \text{otherwise} \end{cases} \;, \tag{1.53}$$

which provides a possibility to define the integral over $A \in \mathcal{B}^n$ by

$$\int_A f(x)\,dx \;:=\; \int \mathbf{1}_A(x)\, f(x)\,dx \;.$$

The indicator function $\mathbf{1}_A$ assigns a volume to Lebesgue measurable sets A by setting $f \equiv 1$

$$\int \mathbf{1}_A \,d\mu \;=\; \mu(A) \;,$$

which is the Lebesgue measure $\mu(A) = \lambda(A)$ for a mapping $\lambda : \mathcal{B} \to \mathbb{R}$.

Next we define *simple functions*, which are understood as finite linear combinations of indicator functions $g = \sum_j \alpha_j \,\mathbf{1}_{A_j}$ and they are measurable if the coefficients α_j are real numbers and the sets A_j are measurable subsets

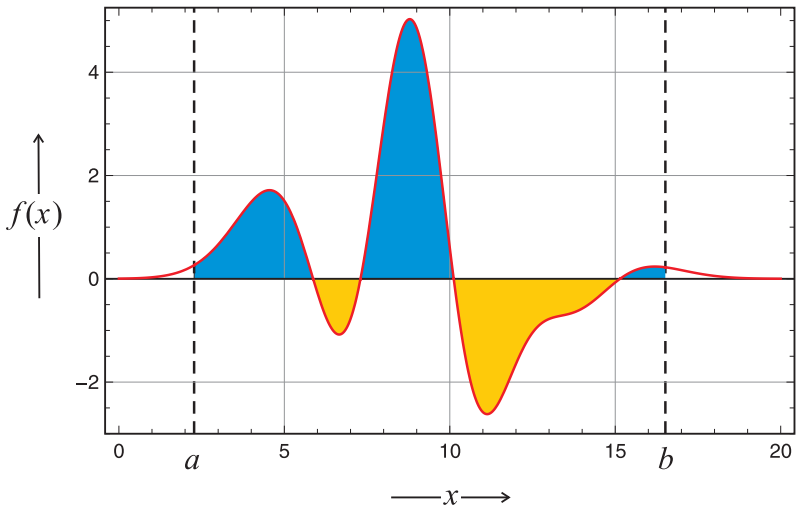

Fig. 1.17 Lebesgue integration of general functions. Lebesgue integration of general functions, i.e. functions with positive and negative stretches, is performed in steps: (i) The integral $I = \int_a^b f \mathrm{d}\mu$ is split into two parts, $I_+ = \int_a^b f_+ \,\mathrm{d}\mu$ (blue) and $I_- = \int_a^b f_- \,\mathrm{d}\mu$ (yellow) function, (ii) the positive part $f_+(x) := \max\{0, f(x)\}$ is Lebesgue integrated like a nonnegative function yielding $I_+ = \int_a^b f_+ \,\mathrm{d}\mu$ and the negative part $f_-(x) := \max\{0, -f(x)\}$ is first mirrored at the x-axis and then Lebesgue integrated like a nonnegative function yielding $I_- = \int_a^b f_- \,\mathrm{d}\mu$, and (iii) the value of the integral is obtained as $I = I_+ - I_-$.

of Ω. For nonnegative coefficients α_j the linearity property of the integral leads to a measure for nonnegative simple functions:

$$\int \left(\sum_j \alpha_j \, \mathbf{1}_{A_j} \right) \mathrm{d}\mu \;=\; \sum_j \alpha_j \int \mathbf{1}_{A_j} \,\mathrm{d}\mu \;=\; \sum_j \alpha_j \, \mu(A_j) \;.$$

Often a simple function can be written in several ways as a linear combination of indicator functions but the value of the integral will necessarily be the same. Sometimes care is needed for the construction of a real-valued simple function $g = \sum_j \alpha_j \mathbf{1}_{A_j}$ in order to avoid undefined expressions of the kind $\infty - \infty$. Choosing $\alpha_i = 0$ implies that $\alpha_i \, \mu(A_i) = 0$ because $0 \cdot \infty = 0$ by convention in measure theory.

An arbitrary nonnegative function $g : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}_{\geq 0}, \mathcal{B}, \lambda)$ is measurable iff there exists a sequence of simple functions $(g_k)_{k\in\mathbb{N}}$ that converges *pointwise*[40] and growing monotonously to g: $g = \lim_{k\to\infty} g_k$. The Lebesgue

[40] Pointwise convergence of a sequence of functions $\{f_n\}$, $\lim_{n\to\infty} f_n = f$ *pointwise* is fulfilled iff $\lim_{n\to\infty} f_n(x) = f(x)$ for every x in the domain (see figure 1.15 and section 1.8.1).

integral of a nonnegative and measurable function g is defined by

$$\int_\Omega g \, \mathrm{d}\mu \;=\; \lim_{k\to\infty} \int_\Omega g_k \, \mathrm{d}\mu \tag{1.54}$$

with g_k being simple functions that converge pointwise and monotonously towards g. The limit is independent of the particular choice of the functions g_k. Such a sequence of simple functions is easily visualized, for example, by the bands below the function $g(x)$ in figure 1.16: The band widths Δd decrease and converge to zero as the index increases, $k \to \infty$.

The extension to general functions with positive and negative value domains is straightforward. As shown in figure 1.17 the function to be integrated, $f(x) : [a, b] \to \mathbb{R}$, is split into two regions that many consist of disjoint domains:

$$\begin{aligned} f_+(x) &:= \max\{0, f(x)\} \\ f_-(x) &:= \max\{0, -f(x)\} \;, \end{aligned}$$

which are considered separately. The function is Lebesgue integrable on the entire domain $[a, b]$ iff both $f_+(x)$ and $f_-(x)$ are Lebesgue integrable and then we have

$$\int_a^b f(x) \, \mathrm{d}x \;=\; \int_a^b f_+(x) \, \mathrm{d}x \;-\; \int_a^b f_-(x) \, \mathrm{d}x \;, \tag{1.55}$$

and this yields precisely the same result as obtained for the Riemann integral. Lebesgue integration readily yields the value for the integral of the absolute value of the function

$$\int_a^b |f(x)| \, \mathrm{d}x \;=\; \int_a^b f_+(x) \, \mathrm{d}x \;+\; \int_a^b f_-(x) \, \mathrm{d}x \;. \tag{1.56}$$

Whenever the Riemann integral exists it is identical with the Lebesgue integral and for practical purposes the calculation by the conventional technique of Riemannian integration is to be preferred since much more experience is available.

Finally, we consider cases where Riemann and Lebesgue integration yield different results. For $\Omega = \mathbb{R}$ and the Lebesgue measure λ holds that functions, which are Riemann integrable on a compact and finite interval $[a, b]$, are Lebesgue integrable too and the values of both integrals are the same, but the inverse is not true: Not every Lebesgue integrable function is Riemann integrable. As an example we consider the Dirichlet step function, $D(x)$, which is the characteristic function of the rational numbers and assumes the

value 1 for rational x and the value 0 for irrational x:[41]

$$D(x) \;=\; \begin{cases} 1\,, \text{ if } x \in \mathbb{Q}\,, \\ 0\,, \text{ otherwise}\,, \end{cases} \qquad \text{or} \quad D(x) \;=\; \lim_{k\to\infty}\Big(\lim_{n\to\infty} \cos^{2n}(k!\,\pi\,x)\Big)\;.$$

$D(x)$ has no Riemann but a Lebesgue integral. The proof is straightforward: $D(x)$ is lacking Riemann integrability for every arbitrarily small interval: Each partitioning S of the integration domain $[a,b]$ into intervals $[x_{k-1},x_k]$ leads to parts that contain necessarily at least one rational and one irrational number. Hence the lower Darboux sum,

$$\Sigma_{\text{low}}(S) = \sum_{k=1}^{n}(x_k - x_{k-1}) \cdot \inf_{x_{k-1}<x<x_k} D(x) = 0\,,$$

vanishes because the infimum is always zero, and the upper Darboux sum,

$$\Sigma_{\text{high}}(S) = \sum_{k=1}^{n}(x_k - x_{k-1}) \cdot \sup_{x_{k-1}<x<x_k} D(x) = b - a\,,$$

is the length of the integration interval, $b-a = \sum_k (x_k - x_{k-1})$, because the supremum is always one and the summation runs over all partial intervals. Since Riemann integrability requires

$$\sup_S\, \Sigma_{\text{low}}(S) \;=\; \int_a^b f(x)dx \;=\; \inf_S\, \Sigma_{\text{high}}(S)$$

$D(x)$ cannot be Riemann integrated.
$D(x)$, on the other hand, has a Lebesgue integral for every interval: $D(x)$ is a nonnegative simple function and therefore we can write the Lebesgue integral over an interval S through sorting into irrational and rational numbers:

$$\int_S D\,\mathrm{d}\lambda \;=\; 0\cdot\lambda(S\cap\mathbb{R}\backslash\mathbb{Q}) \,+\, 1\cdot\lambda(S\cap\mathbb{Q})\;,$$

with λ being the Lebesgue measure. The evaluation of the integral is straightforward. The first term vanishes since multiplication by zero yields zero no matter how large $\lambda(S\cap\mathbb{R}\backslash\mathbb{Q})$ is – we recall that $0\cdot\infty$ is zero by the convention of measure theory – and the second term is also zero as $\lambda(S\cap\mathbb{Q})$ is zero since the set of rational numbers, $\mathbb{Q}$, is countable. Hence we have $\int_S D\,\mathrm{d}\lambda = 0$. □

Another difference between Riemann and Lebesgue integration, however, can occur when the integration is extended to infinity in the improper Riemann integral. Then, the positive and negative contributions may cancel locally

[41] It is worth noticing that the highly irregular, nowhere continuous Dirichlet function $D(x)$ can be formulated as the (double) pointwise convergence limit of a trigonometric function.

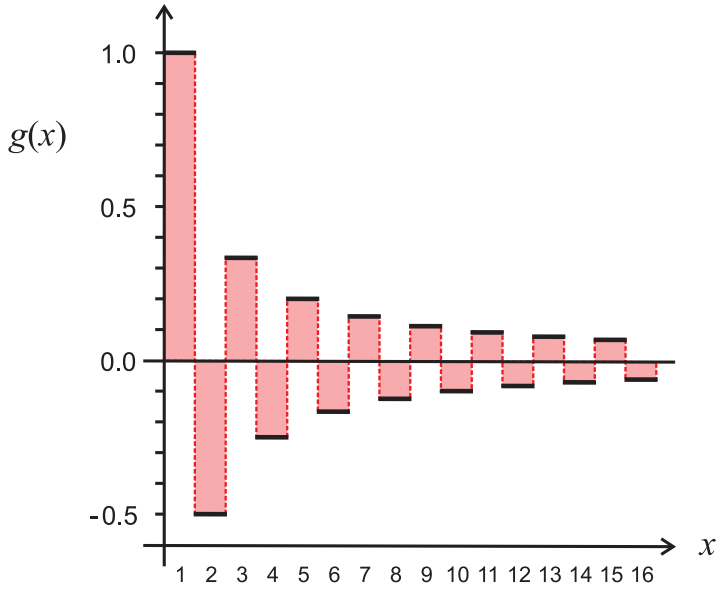

Fig. 1.18 The alternating harmonic series. The alternating harmonic step function, $h(x) = n_k = (-1)^{k+1}/k$ with $(k-1) \leq x < k$ and $n_k \in \mathbb{N}$, has an improper Riemann integral since $\sum_{k=1}^{\infty} n_k = \ln 2$. It is not Lebesgue integrable because the series $\sum_{k=1}^{\infty} |n_k|$ diverges.

in the Riemann summation, whereas divergence may occur in both $f_+(x)$ and in $f_(x)$ since all positive parts and all negative parts are added first in the Lebesgue integral. An example is the improper Riemann integral, $\int_0^{\infty} \cos x \, \mathrm{d}x$, which has a limit inferior, $\liminf_{n\to\infty} x_n = -1$, and a limit superior, $\limsup_{n\to\infty} x_n = +1$, whereas the corresponding Lebesgue integral does not exist.

A typical example of a function that has an improper Riemann integral but is not Lebesgue integrable is the step function with alternatingly positive and negative stretches of size $\frac{1}{n}$, $(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots)$ (see figure 1.18):
The function $h(x) = (-1)^{k+1}/k$ with $(k-1) \leq x < k$ and $k \in \mathbb{N}$ on Riemann integration yields a series of contributions of alternating sign that has a finite infinite sum

$$\int_0^{\infty} h(x) \, \mathrm{d}x \;=\; 1 - \frac{1}{2} + \frac{1}{3} - \ldots \;=\; \ln 2 \;,$$

whereas Lebesgue integrability of h requires $\int_{\mathbb{R}_{\geq 0}} |h| \, \mathrm{d}\lambda < \infty$ and this is not fulfilled since both f_+ and f_- diverge as the harmonic series, $\sum_{k=1}^{\infty} k^{-1}$, does. The proof is straightforward if one uses Leonhard Euler's result that the series of reciprocal prime number diverges:

$$\sum_{p \text{ prime}} \frac{1}{p} \;=\; \frac{1}{2}+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\frac{1}{11}+\frac{1}{13}+\ldots \;=\; \infty \;,$$

$$\sum_{o \text{ odd}} \frac{1}{o} \;=\; 1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\frac{1}{9}+\frac{1}{11}+\frac{1}{13}+\ldots \;>\; \sum_{p \text{ prime}} \frac{1}{p} \;,$$

$$1+\sum_{e \text{ even}} \frac{1}{e} \;=\; 1+\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\frac{1}{10}+\frac{1}{12}+\ldots \;>\; \sum_{o \text{ odd}} \frac{1}{o} \;.$$

Since $\infty - 1 = \infty$ both partial sums $\sum_{o \text{ odd}} \frac{1}{o}$ and $\sum_{e \text{ even}} \frac{1}{e}$ and diverge. □

The first case discussed here – no Riemann integral but Lebesgue integrability – is the more important issue since it provides a proof that the set of rational numbers, $\mathbb{Q}$ is of Lebesgue measure zero.

Finally, we introduce the Lebesgue-Stieltjes integral in a way that allows for summarizing the most important results of this section. For each right-hand continuous and monotonously increasing function $F : \mathbb{R} \to \mathbb{R}$ exists a uniquely determined Lebesgue-Stieltjes measure λ_F that fulfils

$$\lambda_F\big((a,b]\big) \;=\; F(b) \,-\, F(a) \quad \text{for all} \quad (a,b] \subset \mathbb{R}$$

Such functions $F : \mathbb{R} \to \mathbb{R}$ – being righthand continuous and monotonously increasing – are therefore called *measure generating*. The Lebesgue integral of a λ_F integrable function f is called Lebesgue-Stieltjes integral

$$\int_A f \,\mathrm{d}\lambda_F \quad \text{with} \quad A \in \mathcal{B} \tag{1.57}$$

being Borel measurable. Let F be the identity function on $\mathbb{R}$,[42]

$$F = \mathrm{id} : \mathbb{R} \to \mathbb{R},\; \mathrm{id}(x) = x \;,$$

then the corresponding Lebesgue-Stieltjes measure is the Lebesgue measure itself: $\lambda_F = \lambda_{\mathrm{id}} = \lambda$. For proper Riemann integrable functions f we have stated that the Lebesgue integral is identical with the Riemann integral:

$$\int\limits_{[a,b]} f \,\mathrm{d}\lambda \;=\; \int_a^b f(x)\,\mathrm{d}x \;.$$

The interval $[a,b] = a \le x \le b$ is partitioned into a sequence

$$\sigma_n = (a = x_0^{(n)}, x_1^{(n)}, \ldots, x_r^{(n)} = b)$$

[42] The identity function $\mathrm{id}(x) := x$ maps a domain, for example $[a,b]$, point by point onto itself.

where the superscript $'^{(n)}$' indicates a Riemann sum with $|\sigma_n| \to 0$ and the Riemann integral on the righthand side is replaced by the limit of the Riemann summation:

$$\int\limits_{[a,b]} f \, \mathrm{d}\lambda \;=\; \lim_{n\to\infty} \sum_{k=1}^{r} f(x_{k-1}^{(n)}) \left(x_k^{(n)} - x_{k-1}^{(n)} \right) \;=$$

$$= \lim_{n\to\infty} \sum_{k=1}^{r} f(x_{k-1}^{(n)}) \left(\mathrm{id}(x_k^{(n)}) - \mathrm{id}(x_{k-1}^{(n)}) \right) \; .$$

The Lebesgue measure λ has been introduced above as the special case $F = \mathrm{id}$ and therefore the Stieltjes-Lebesgue integral is obtained by replacing λ by λ_F and 'id' by F

$$\int\limits_{[a,b]} f \, \mathrm{d}\lambda_F \;=\; \lim_{n\to\infty} \sum_{k=1}^{r} f(x_{k-1}^{(n)}) \left(F(x_k^{(n)}) - F(x_{k-1}^{(n)}) \right) \; .$$

The details of the derivation are found in [31, 208].

In summary, we define a Stieltjes-Lebesgue integral or F-integral by $F, g : \mathbb{R} \to \mathbb{R}$, where the two functions F and f are partitioned on the interval $[a, b]$ by the sequence $\sigma = (a = x_0, x_1, \ldots, x_r = b)$:

$$\sum_{\sigma} f \, \mathrm{d}F \;:=\; \sum_{k=1}^{r} f(x_{k-1}) \big(F(x_k) - F(x_{k-1}) \big) \; .$$

The function f is F-integrable on [a,b] if

$$\int\limits_{a}^{b} f \, \mathrm{d}F \;=\; \lim_{|\sigma|\to 0} \sum_{\sigma} f \, \mathrm{d}F \tag{1.58}$$

exists in $\mathbb{R}$ and then $\int_a^b f \, \mathrm{d}F$ is called the Stieltjes-Lebesgue integral or F-integral of f. In the theory of stochastic processes the Stieltjes-Lebesgue integral is required for the formulation of the Itō integral, which is used in Itō calculus applied to the integration of stochastic differential equations (SDEs; section 3.4) [140, 141].

1.9 Continuous random variables and distributions

Random variables on uncountable sets are completely characterized by a *probability triple* $(\Omega, \mathcal{F}, P)$. The *triple* is essentially the same as in the case of discrete variables (section 1.6.1) except that the power set $\Pi(\Omega)$ has been replaced by the event system $\mathcal{F} \subset \Pi(\Omega)$. We recall that the powerset $\Pi(\Omega)$ is too large for defining probabilities since it contains uncountably many subsets or events A (figure 1.14). The sets in $\mathcal{F}$ are the Borel σ-algebras, they are measurable, and they *alone* have probabilities. Accordingly, we are now in the position to handle also probabilities on uncountable sets:

$$\{\omega | \mathcal{X}(\omega) \leq x\} \in \mathcal{F} \text{ and } P(\mathcal{X} \leq x) = \frac{|\{\mathcal{X}(\omega) \leq x\}|}{|\Omega|} \tag{1.59a}$$

$$\{a < \mathcal{X} \leq b\} = \{\mathcal{X} \leq b\} - \{\mathcal{X} \leq a\} \in \mathcal{F} \text{ with } a < b \tag{1.59b}$$

$$P(a < \mathcal{X} \leq b) = \frac{|\{a < \mathcal{X} \leq b\}|}{|\Omega|} = F_{\mathcal{X}}(b) - F_{\mathcal{X}}(a) \; . \tag{1.59c}$$

Equation (1.59a) contains the definition of a real-valued function $\mathcal{X}$ that is called a random variable iff it fulfils $P(\mathcal{X} \leq x)$ for any real number x, equation (1.59b) is valid since $\mathcal{F}$ is closed under difference, and finally equation (1.59c) provides the basis for defining and handling probabilities on uncountable sets. The three equations (1.59) together constitute the basis of the probability concept on uncountable sample spaces that will be applied throughout this book.

1.9.1 Densities and distributions

Random variables on uncountable sets Ω are commonly characterized by *probability density functions* (pdf). The probability density function – or density for short – is the continuous analogue to the probability mass function (pmf). A density is a function f on $\mathbb{R} =]-\infty, +\infty[$, $u \to f(u)$, which satisfies the two conditions:[43]

$$\begin{aligned} &\text{(i)} \quad \forall u: \quad f(u) \geq 0 \; , \quad \text{and} \\ &\text{(ii)} \quad \int_{-\infty}^{+\infty} f(u) \, \mathrm{d}u = 1 \; . \end{aligned} \tag{1.60}$$

[43] From here on we shall omit the random variable as subscript and simply write $f(x)$ or $F(x)$ unless a nontrivial specification is required.

Now we can define a class of random variables[44] on general sample spaces: $\mathcal{X}$ is a function on $\Omega : \omega \to \mathcal{X}(\omega)$ whose probabilities are prescribed by means of a density function $f(u)$. For any interval $[a, b]$ the probability is given by

$$P(a \leq \mathcal{X} \leq b) \;=\; \int_a^b f(u)\,\mathrm{d}u \;. \tag{1.61}$$

If A is the union of not necessarily disjoint intervals – some of which may be even infinite – the probability can be derived in general from the density

$$P(\mathcal{X} \in A) \;=\; \int_A f(u)\,\mathrm{d}u \;,$$

in particular, A can be split in disjoint intervals, $A = \bigcup_{j=1}^{k} [a_j, b_j]$ and then the integral can be rewritten as

$$\int_A f(u)\,\mathrm{d}u \;=\; \sum_{j=1}^{k} \int_{a_j}^{b_j} f(u)\,\mathrm{d}u \;.$$

For the interval $A =]-\infty, x]$ we define the *cumulative probability distribution function* (cdf) $F(x)$ of the continuous random variable $\mathcal{X}$

$$F(x) \;=\; P(\mathcal{X} \leq x) \;=\; \int_{-\infty}^{x} f(u)\,\mathrm{d}u \;.$$

If f is continuous then it is the derivative of F as follows from the fundamental theorem of calculus

$$F'(x) \;=\; \frac{\mathrm{d}F(x)}{\mathrm{d}x} \;=\; f(x).$$

If the density f is not continuous everywhere, the relation is still true for every x at which f is continuous.

If the random variable $\mathcal{X}$ has a density, then we find by setting $a = b = x$

$$P(\mathcal{X} = x) \;=\; \int_x^x f(u)\,\mathrm{d}u \;=\; 0$$

reflecting the trivial geometric result that every line segment has zero area. It seems somewhat paradoxical that $\mathcal{X}(\omega)$ must be some number for every ω whereas any given number has probability zero. The paradox is resolved by looking at countable and uncountable sets in more depth as we did in sections 1.5 and 1.6.3.

As an illustrative example for continuous probability functions we present here the *normal distribution*, which is of primary importance in probability theory for two reasons: (i) It is mathematically simple and well behaved, and

[44] Random variables having a density are often called *continuous* in order to distinguish them from *discrete* random variables defined on countable sample spaces.

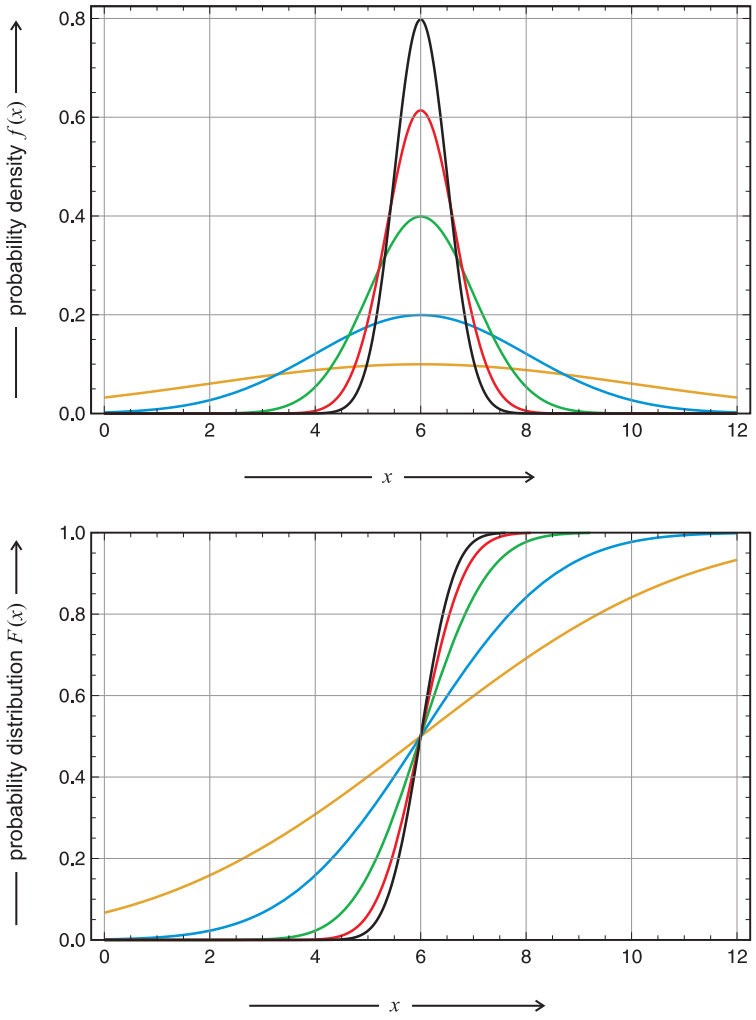

Fig. 1.19 Normal density and distribution. In the plots the normal distribution, $\mathcal{N}(\mu, \sigma)$, is shown in from of the probability density $f(x) = \exp\left(-(x-\mu)^2/(2\sigma^2)\right) / (\sqrt{2\pi}\,\sigma)$ and the probability distribution $F(x) = \left(1 + \mathrm{erf}\left((x-\mu)/\sqrt{2\sigma^2}\right) / 2\right)$ where 'erf' represents the error function. Choice of parameters: $\mu = 6$ and $\sigma = 0.5$ (black), 0.65 (red), 1 (green), 2 (blue) and 4 (yellow).

(ii) all distributions converge to the normal distribution in the limit of large sample numbers as expressed by the *central limit theorem* (subsection 2.3.6). The density of the normal distribution is a Gaussian function named after the German mathematician Carl Friedrich Gauß and is also called symmetric bell curve.

$$\mathcal{N}(x;\mu,\sigma^2): \quad f(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\;\exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right)\;, \tag{1.62}$$

$$F(x) \;=\; \frac{1}{2}\left(1+\mathrm{erf}\Big(\frac{x-\mu}{\sqrt{2\sigma^2}}\Big)\right)\;. \tag{1.63}$$

Herein 'erf' is the error function.[45] This function and its complement, 'erfc', are defined by

$$\mathrm{erf}(x) \;=\; \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2}\mathrm{d}z \quad \text{and} \quad \mathrm{erfc}(x) \;=\; \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-z^2}\mathrm{d}z$$

The two parameters of the normal distribution, μ and σ, are the expectation value and the mean deviation of a normally distributed random variable.

Although the central limit theorem will be discussed separately in section 2.3.6, we present here an example for the convergence of a probability distribution towards the normal distribution we are already familiar with: the rolling dice problem extended to n dice. A collection of n dice is thrown simultaneously and the total score of all dice together is recorded (figure 1.20). The probability of a total score of k points obtained through rolling n dice with s faces can be calculated by means of combinatorics:

$$f_{s,n}(k) \;=\; \frac{1}{s^n}\;\sum_{i=0}^{\lfloor\frac{k-n}{s}\rfloor}(-1)^i\binom{n}{i}\binom{k-s\,i-1}{n-1} \tag{1.64}$$

The results for small values of n and ordinary dice ($s = 6$) are illustrated in Fig. 1.20. The convergence to a continuous probability density is nicely illustrated. For $n = 7$ the deviation from a the Gaussian curve of the normal distribution is hardly recognizable.

Sometimes it is useful to discretize a density function in order to yield a set of elementary probabilities. The x-axis is divided up into m pieces (figure 1.21), not necessarily equal and not necessarily small, and we denote the piece of the integral on the interval $\Delta_k = x_{k+1} - x_k$, i.e. between the values $u(x_k)$ and $u(x_{k+1})$ of the variable u, by

[45] We remark that $\mathrm{erf}(x)$ and $\mathrm{erfc}(x)$ are not normalized in the same way as the normal density: $\mathrm{erf}(x) + \mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_0^\infty \exp(-t^2)\,\mathrm{dt} = 1$, but $\int_0^\infty f(x)\mathrm{d}x = \frac{1}{2}\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \frac{1}{2}$.

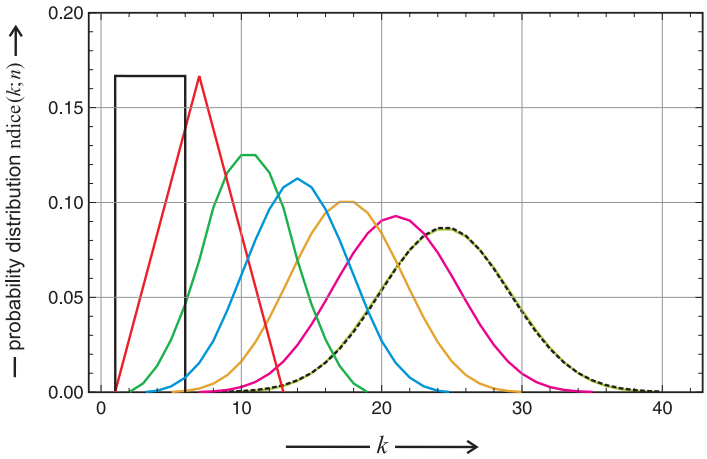

Fig. 1.20 Convergence of the probability mass function for rolling n dice to the normal density. The series of probability mass functions for rolling n dice, $f_{n\mathcal{D}}(k)$, begins with a pulse function $f_{1\mathcal{D}}(k) = 1/6$ for $i = 1, \ldots, 6$ $(n = 1)$, next comes a tent function $(n = 2)$, and then follows a gradual approach towards the normal distribution, $(n = 3, 4, \ldots)$. For $n = 7$ we show the fitted normal distribution (broken black curve) coinciding almost perfectly with $f_{7\mathcal{D}}(k)$. Choice of parameters: $s = 6$ and $n = 1$ (black), 2 (red), 3 (green), 4 (blue), 5 (yellow), 6 (magenta), and 7 (chartreuse).

$$p_k \;=\; \int_{x_k}^{x_{k+1}} f(u)\,\mathrm{d}u\,,\; 0 \le k \le m-1\;, \tag{1.65}$$

where the p_k-values fulfil.

$$\forall\; k:\;\; p_k \ge 0 \;\;\text{and}\;\; \sum_{k=0}^{m-1} p_k \;=\; 1\;.$$

If we choose $x_0 = -\infty$ and $x_m = +\infty$ we are dealing with a partition that is not finite but countable, provided we label the intervals suitably, for example $\ldots, p_{-2}, p_{-1}, p_0, p_1, p_2, \ldots$. Now we consider a random variable $\mathcal{Y}$ such that

$$P(\mathcal{Y} = x_k) \;=\; p_k\;, \tag{1.65'}$$

where we may replace x_k by any value of x in the subinterval $[x_k, x_{k+1}]$. The random variable $\mathcal{Y}$ can be interpreted as the discrete analogue of the random variable $\mathcal{X}$.

Making the intervals Δ_k smaller increases the accuracy of the approximation through discretization and this procedure has a lot in common with Riemann integration.

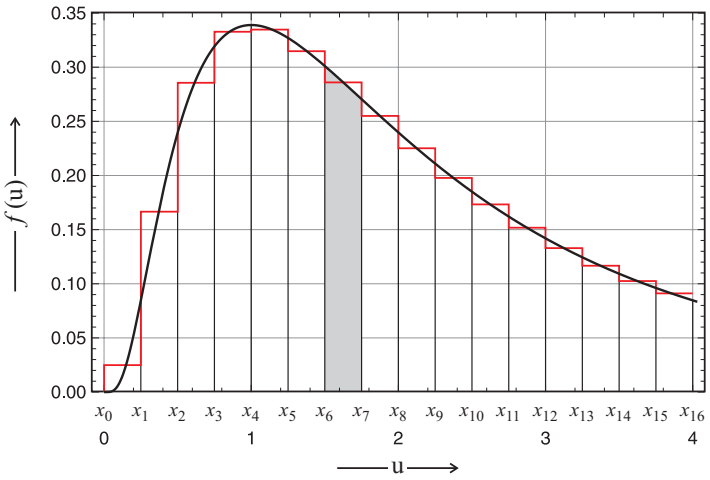

Fig. 1.21 Discretization of a probability density. A segment $[x_0, x_m]$ on the u-axis is divided up into m not necessarily equal intervals and elementary probabilities are obtained by integration. The curve shown here is the density of the lognormal distribution $\ln \mathcal{N}(\nu, \sigma^2)$:

$$f(u) \;=\; \frac{1}{u\,\sqrt{2\pi\,\sigma^2}}\, e^{-(\ln u - \nu)^2/(2\sigma^2)}.$$

The red step function represents the discretized density. The hatched area is the probability $p_6 = \int_{x_6}^{x_7} f(u)\,du$ with the parameters $\nu = \ln 2$ and $\sigma = \sqrt{\ln 2}$.

1.9.2 Continuous variables and independence

In the joint distribution function of the random vector $\vec{\mathcal{V}} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ independence is tantamount to factorizability:

$$F(x_1, \ldots, x_n) \;=\; F_1(x_1) \cdot \ldots \cdot F_n(x_n)\ ,$$

where the F_j's are the marginal distributions of the random variables, the $\mathcal{X}_j$'s $(1 \le j \le n)$. As in the discrete case the marginal distributions are sufficient to calculate joint distributions of independent random variables.

For the continuous case we can formulate the definition of independence for sets $S_1, \ldots, S_n$ forming a Borel family. In particular, when there is a joint density function $f(u_1, \ldots, u_n)$, we have

$$
\begin{aligned}
P(\mathcal{X}_1 \in S_1, \ldots, \mathcal{X}_n \in S_n) &= \int_{S_1} \cdots \int_{S_n} f(u_1, \ldots, u_n)\, du_1 \ldots du_n = \\
&= \int_{S_1} \cdots \int_{S_n} f_1(u_1) \ldots f_n(u_n)\, du_1 \ldots du_n = \\
&= \left(\int_{S_1} f_1(u_1)\, du_1 \right) \cdot \ldots \cdot \left(\int_{S_n} f_n(u_n)\, du_n \right) ,
\end{aligned}
$$

where $f_1, \ldots, f_n$ are the marginal densities, for example

$$
f_1(u_1) = \int_{S_2} \cdots \int_{S_n} f(u_1, \ldots, u_n)\, du_2 \ldots du_n , \tag{1.66}
$$

and eventually we find for the density case:

$$
f(u_1, \ldots, u_n) = f_1(u_1) \ldots f_n(u_n) . \tag{1.67}
$$

As we have seen here, stochastic independence is the basis for factorization of joint probabilities, distributions, densities, and other functions.

1.9.3 Probabilities of discrete and continuous variables

A comparison of the formalisms of probability theory on countable and uncountable sample spaces closes this chapter. For this goal we repeat and compare in table 1.4 the basic features of discrete and continuous probability distributions as they have been discussed in section 1.6.2 and 1.9.1, respectively.

Discrete probability distribution are defined on countable sample spaces and their random variables are discrete sets of events $\omega \in \Omega$, for example sample points on an closed interval $[a, b]$:

$$
\{a \leq \mathcal{X} \leq b\} = \{\omega | a \leq \mathcal{X} \leq b\} .
$$

If the sample space Ω is finite or countable infinite the exact range of $\mathcal{X}$ is a set of real numbers w_i

$$
W_{\mathcal{X}} = \{w_1, w_2, \ldots, w_n, \ldots\} \text{ with } w_k \in \Omega \, \forall \, k .
$$

Introducing probabilities for individual events, $p_n = P(\mathcal{X} = w_n)$; $w_n \in W_{\mathcal{X}}$ and $P(\mathcal{X} = x) = 0$ if $x \notin W_{\mathcal{X}}$, yields

$$
P(\mathcal{X} \in A) = \sum_{w_n \in A} p_n \text{ with } A \in \Omega
$$

or, in particular,

Table 1.4 Comparison of the formalism of probability theory on countable and uncountable sample spaces.

	Expression	Countable case	Uncountable case
Domain	$\mathbb{R}^1$	$w_n,\ n = 1, 2, \ldots$	$-\infty < u < +\infty$
Probability	$P(\mathcal{X} \in A);\ A \in \Omega$	p_n	$dF(u) = f(u)\,du$
Interval	$P(x \leq \mathcal{X} \leq b)$	$\sum_{a \leq w_n \leq b} p_n$	$\int_a^b f(u)\,du$
PDF	$f(x) = P(\mathcal{X} = x)$	$\begin{cases} p_n & \text{if } x \in W_{\mathcal{X}} \\ 0 & \text{if } x \notin W_{\mathcal{X}} \end{cases}$	$f(u)\,du$
CDF	$F(x) = P(\mathcal{X} \leq x)$	$\sum_{w_n \leq x} p_n$	$\int_{-\infty}^{x} f(u)\,du$
Expectation	$E(\mathcal{X})$	$\sum_n p_n\, w_n$	$\int_{-\infty}^{\infty} u\, f(u)\,du$
	with	$\sum_n p_n\, \lvert w_n \rvert < \infty$	$\int_{-\infty}^{\infty} \lvert u \rvert\, f(u)\,du < \infty$

$$P(a \leq \mathcal{X} \leq b) \;=\; \sum_{a \leq w_n \leq b} p_n \;. \tag{1.26}$$

Two probability functions are in common use, the probability mass function (pmf)

$$f_{\mathcal{X}}(x) \;=\; P(\mathcal{X} = x) \begin{cases} p_n & \text{if } x = w_n \in W_{\mathcal{X}} \;, \\ 0 & \text{if } x \neq W_{\mathcal{X}} \;, \end{cases}$$

and the cumulative distribution function (cdf)

$$F_{\mathcal{X}}(x) \;=\; \text{Prob}\,(\mathcal{X} \;\leq\; x) \;=\; \sum_{w_n \leq x} p_n \;,$$

with two properties following form the property of probabilities:

$$\lim_{x \to -\infty} \; F_{\mathcal{X}}(x) \;=\; 0 \quad \text{and} \quad \lim_{x \to +\infty} \; F_{\mathcal{X}}(x) \;=\; 1 \;.$$

Continuous probability distributions are defined on uncountable, Borel measurable sample spaces and their random variables $\mathcal{X}$ have densities. A *probability density function* (pdf) is a mapping

$$f : \mathbb{R}^1 \implies \mathbb{R}^1_{\geq 0} \;,$$

which satisfies the two conditions:

$$\begin{aligned} &\text{(i)} \qquad f(u) \geq 0 \;\forall\; u \in \mathbb{R}^1 \;, \quad \text{and} \\ &\text{(ii)} \qquad \int_{-\infty}^{+\infty} f(u)\, du \;=\; 1 \;. \end{aligned} \tag{1.68}$$

Random variables $\mathcal{X}$ are functions on Ω: $\omega \Longrightarrow \mathcal{X}(\omega)$ whose probabilities are derived from density functions $f(u)$:

$$P(a \leq \mathcal{X} \leq b) \;=\; \int_a^b f(u)\, du \;. \tag{1.61}$$

As in the discrete case the probability functions come in two forms: (i) the probability density function (pdf) defined above,

$$dF(u) \;=\; f(u)\, du \;,$$

and (ii) the cumulative distribution function (cdf)

$$F(x) \;=\; P(\mathcal{X} \leq x) \;= \int_{-\infty}^{x} f(u)\, \mathrm{d}u \;\; \text{with} \;\; \frac{\mathrm{d}F(x)}{\mathrm{dx}} \;=\; f(x)$$

provided the function $f(x)$ is continuous.

Conventional thinking in terms of probabilities has been extended in two important ways in the last two sections: (i) Handling of uncountable sets allowed for definition of and calculation with probabilities when comparison by counting is not possible and (ii) Lebesgue-Stieltjes integration provided an extension of calculus to the step functions encountered with discrete probabilities.

Chapter 2
Distributions, moments, and statistics

Make things as simple as possible but not simpler.
Albert Einstein 1950.

Abstract . The moments of probability distributions represent the link between theory and observations since they are readily accessible to measurement. Generating functions looking rather abstract became important as highly versatile concepts and tools for solving specific problems. The probability distributions, which are most important in application are reviewed. Then the central limit theorem being the basis of the law of large numbers is presented and the chapter is closed by discussing real world samples that cover only a (small) part of sample space.

In this chapter we make an attempt to bring probability theory closer to applications. Probability distributions and densities are used to calculate measurable quantities like expectation values, variances, and higher moments. The moments provide partial information on the probability distributions since the full information would require a series expansion up to infinite order.

2.1 Expectation values and higher moments

Random variables are accessible to analysis via their probability distributions. In addition, straightforward information can be derived also from ensembles defined on the entire sample space Ω. Complete coverage, of course, is an ideal reference that can never be achieved in real situations. Samples collected in observations or experiments are commonly much smaller than an exhaustive collection. We begin here with a discussion of the theoretical reference and introduce mathematical statistics afterwards. Most important are the first two moments having a straightforward interpretation: The expectation value $E(\mathcal{X})$ is the mean or average value of a distribution and the variance $\mathrm{var}(\mathcal{X})$ or $\sigma^2(\mathcal{X})$ measures the width of distributions.

2.1.1 First and second moments

The most important example of an ensemble property is the *expectation value* or *mean value*, $E(\mathcal{X})$ or $\langle\mathcal{X}\rangle$. We begin with a countable sample space Ω:

$$E(\mathcal{X}) \;=\; \sum_{\omega\in\Omega} \mathcal{X}(\omega)P(\omega) \;=\; \sum_n v_n\, p_n \;. \tag{2.1}$$

In the special case of a random variable $\mathcal{X}$ on $\mathbb{N}$ we have $v_n = n$ and find

$$E(\mathcal{X}) \;=\; \sum_{n=0}^{\infty} n\, p_n \;.$$

The expectation value (2.1) exists when the series converges in absolute values, $\sum_{\omega\in\Omega} |\mathcal{X}(\omega)|\, P(\omega) < \infty$. Whenever the random variable $\mathcal{X}$ is bounded, which means that there exists a number m such that $|\mathcal{X}(\omega)| \le m$ for all $\omega \in \Omega$, then it is summable and in fact

$$E(|\mathcal{X}|) \;=\; \sum_{\omega} |\mathcal{X}(\omega)|\, P(\omega) \;\le\; m \sum_{\omega} P(\omega) \;=\; m \;.$$

It is straightforward to show that the sum of two random variables, $\mathcal{X} + \mathcal{Y}$ is summable iff $\mathcal{X}$ and $\mathcal{Y}$ are summable:

$$E(\mathcal{X} + \mathcal{Y}) \;=\; E(\mathcal{X}) \,+\, E(\mathcal{Y}) \;.$$

The relation can be extended to an arbitrary countable number of random variables:

$$E\left(\sum_{k=1}^{n} \mathcal{X}_k\right) \;=\; \sum_{k=1}^{n} E(\mathcal{X}_k) \;.$$

In addition, the expectation values fulfill the following relations $E(a) = a$, $E(a\mathcal{X}) = a \cdot E(\mathcal{X})$ which can be combined in

$$E\left(\sum_{k=1}^{n} a_k\, \mathcal{X}_k\right) \;=\; \sum_{k=1}^{n} a_k \cdot E(\mathcal{X}_k) \;. \tag{2.2}$$

Accordingly, $E(\cdot)$ is a *linear operator*.

For a random variable $\mathcal{X}$ on an arbitrary sample space Ω the expectation value may be written as an abstract integral on Ω or as an integral over $\mathbb{R}$ provided the density $f(u)$ exists:

$$E(\mathcal{X}) \;=\; \int_{\Omega} \mathcal{X}(\omega)\, d\omega \;=\; \int_{-\infty}^{+\infty} u\, f(u)\, \mathrm{d}u \;. \tag{2.3}$$

It is worth to reconsider the discretization of a continuous density (figure 1.21) in this context: The discrete expression for the expectation value is based upon $p_n = P(\mathcal{Y} = x_n)$ as outlined in equations (1.65) and (1.65'),

$$E(\mathcal{Y}) \;=\; \sum_n x_n\, p_n \;\approx\; E(\mathcal{X}) \;=\; \int_{-\infty}^{+\infty} u\, F(u)\, \mathrm{d}u \;,$$

and approximates the exact value similarly as the Darboux sum does in case of a Riemann integral.

For two or more variables, for example $\vec{\mathcal{V}} = (\mathcal{X}, \mathcal{Y})$ described by a joint density $f(u, v)$, we have

$$E(\mathcal{X}) \;=\; \int_{-\infty}^{+\infty} u\, f(u, *)\, \mathrm{d}u \;\text{ and }\; E(\mathcal{Y}) \;=\; \int_{-\infty}^{+\infty} v\, f(*, v)\, \mathrm{d}v \;,$$

where $f(u, *) = \int_{-\infty}^{+\infty} f(u, v)\, \mathrm{d}v$ and $f(*, v) = \int_{-\infty}^{+\infty} f(u, v)\, \mathrm{d}u$ are the marginal densities.

The expectation value of the sum of the variables, $\mathcal{X} + \mathcal{Y}$, can be evaluated by iterated integration:

$$\begin{aligned}
E(\mathcal{X} + \mathcal{Y}) \;&=\; \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (u + v)\, f(u, v)\, \mathrm{d}u\, \mathrm{d}v \;= \\
&=\; \int_{-\infty}^{+\infty} u\, \mathrm{d}u \left(\int_{-\infty}^{+\infty} f(u, v)\, \mathrm{d}v \right) + \int_{-\infty}^{+\infty} v\, \mathrm{d}v \left(\int_{-\infty}^{+\infty} f(u, v)\, \mathrm{d}u \right) \;= \\
&=\; \int_{-\infty}^{+\infty} u\, f(u, *)\, \mathrm{d}u + \int_{-\infty}^{+\infty} v\, f(*, v)\, \mathrm{d}v \;= \\
&=\; E(\mathcal{X}) + E(\mathcal{Y}) \;,
\end{aligned}$$

which establishes the previously derived expression.

The *multiplication theorem* of probability theory requires that the two variables $\mathcal{X}$ and $\mathcal{Y}$ are independent and summable and this implies for the discrete and the continuous case,[1]

$$E(\mathcal{X} \cdot \mathcal{Y}) \;=\; E(\mathcal{X}) \cdot E(\mathcal{Y}) \;\text{ and} \tag{2.4a}$$

$$\begin{aligned}
E(\mathcal{X} \cdot \mathcal{Y}) \;&=\; \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} u\, v\, f(u, v)\, \mathrm{d}u\, \mathrm{d}v \;= \\
&=\; \int_{-\infty}^{+\infty} u\, f(u, *)\, \mathrm{d}u \int_{-\infty}^{+\infty} v\, f(*, v)\, \mathrm{d}v \;= \\
&=\; E(\mathcal{X}) \cdot E(\mathcal{Y}) \;,
\end{aligned} \tag{2.4b}$$

[1] A proof is found in [34, pp.164-166].

respectively. The multiplication theorem is easily extended to any finite number of independent and summable random variables:

$$E(\mathcal{X}_1, \ldots, \mathcal{X}_n) \;=\; E(\mathcal{X}_1) \cdot \ldots \cdot E(\mathcal{X}_n) \;. \tag{2.4c}$$

Let us now consider the expectation values of special functions of random variables, in particular their powers $\mathcal{X}^n$, which give rise to the *raw moments* of the probability distribution, $\hat{\mu}_r$. For a random variable $\mathcal{X}$ we distinguish the r-th moments $E(\mathcal{X}^r)$ and the so-called *centered moments*[2] $\mu_r = E(\tilde{\mathcal{X}}^r)$ referring to the random variable

$$\tilde{\mathcal{X}} \;=\; \mathcal{X} \,-\, E(\mathcal{X}) \;.$$

Clearly, the first raw moment is the expectation value and the first centered moment vanishes, $E(\tilde{\mathcal{X}}) = \mu_1 = 0$. Often the expectation value is denoted by $\mu = \hat{\mu}_1 = E(\mathcal{X}) = \langle \mathcal{X} \rangle$, notations that we shall use too for the sake of convenience but it is important not to confuse μ and μ_1.

In general, a moment is defined about some point a by means of the random variable

$$\mathcal{X}^{(a)} \;=\; \mathcal{X} \,-\, a \;.$$

For $a = 0$ we obtain the raw moments

$$\hat{\mu}_r \;=\; \alpha_r \;=\; E(\mathcal{X}^r) \tag{2.5}$$

whereas $a = E(\mathcal{X})$ yields the centered moments.

The general expressions for the raw r-th moments and centered moments as derived from the density $f(u)$ are

$$E(\mathcal{X}^r) \;=\; \hat{\mu}_r(\mathcal{X}) \;=\; \int_{-\infty}^{+\infty} u^r \, f(u) \, \mathrm{d}u \quad \text{and} \tag{2.6a}$$

$$E(\tilde{\mathcal{X}}^r) \;=\; \mu_r(\mathcal{X}) \;=\; \int_{-\infty}^{+\infty} (u - \mu)^r \, f(u) \, \mathrm{d}u \;. \tag{2.6b}$$

The second centered moment is called the *variance*, $\mathrm{var}(\mathcal{X})$ or $\sigma^2(\mathcal{X})$, and its positive square root is the *standard deviation* $\sigma(\mathcal{X})$. The variance is always a non-negative quantity as can be easily shown. Further we can derive:

[2] Since the moments centered around the expectation value will be used more frequently than the raw moments, we denote them by μ and the raw moments by $\hat{\mu}$. The r-th moment of a distribution is also called the moment of order r.

$$\begin{aligned}\sigma^2(\mathcal{X}) \;&=\; E(\tilde{\mathcal{X}}^2) \;=\; E\left(\Big(\mathcal{X}-E(\mathcal{X})\Big)^2\right) \;= \\ &=\; E\Big(\mathcal{X}^2 \,-\, 2\,\mathcal{X}\,E(\mathcal{X}) \,+\, E(\mathcal{X})^2\Big) \;= \\ &=\; E(\mathcal{X}^2) \,-\, 2\,E(\mathcal{X})\,E(\mathcal{X}) \,+\, E(\mathcal{X})^2 \;= \\ &=\; E(\mathcal{X}^2) \,-\, E(\mathcal{X})^2 \;.\end{aligned} \tag{2.7}$$

If $E(\mathcal{X}^2)$ is finite, than $E(|\mathcal{X}|)$ is finite too and fulfils the inequality

$$E(|\mathcal{X}|)^2 \;\leq\; E(\mathcal{X}^2) \;,$$

and since $E(\mathcal{X}) \leq E(|\mathcal{X}|)$ the variance is necessarily a non-negative quantity, $\sigma^2(\mathcal{X}) \geq 0$.

If $\mathcal{X}$ and $\mathcal{Y}$ are independent and have finite variances, then we obtain

$$\sigma^2(\mathcal{X} \,+\, \mathcal{Y}) \;=\; \sigma^2(\mathcal{X}) \,+\, \sigma^2(\mathcal{Y}) \;,$$

as follows readily by simple calculation:

$$\begin{aligned}E\big((\tilde{\mathcal{X}}+\tilde{\mathcal{Y}})^2\big) \;&=\; E\big(\tilde{\mathcal{X}}^2 \,+\, 2\,\tilde{\mathcal{X}}\,\tilde{\mathcal{Y}} \,+\, \tilde{\mathcal{Y}}^2\big) \;= \\ &=\; E\big(\tilde{\mathcal{X}}^2\big) \,+\, 2\,E(\tilde{\mathcal{X}})\,E(\tilde{\mathcal{Y}}) \,+\, E\big(\tilde{\mathcal{Y}}^2\big) \;=\; E\big(\tilde{\mathcal{X}}^2\big) \,+\, E\big(\tilde{\mathcal{Y}}^2\big) \;.\end{aligned}$$

Here we use the fact of vanishing first centered moments: $E(\tilde{\mathcal{X}}) = E(\tilde{\mathcal{Y}}) = 0$.

For two general – non necessarily independent – random variables $\mathcal{X}$ and $\mathcal{Y}$, the Cauchy-Schwarz inequality holds for the mixed expectation value:

$$E(\mathcal{X}\mathcal{Y})^2 \;\leq\; E(\mathcal{X}^2)\,E(\mathcal{Y}^2) \;. \tag{2.8}$$

If both random variables have finite variances, the *covariance* is defined by

$$\begin{aligned}\mathrm{cov}(\mathcal{X},\mathcal{Y}) \;&=\; \sigma^2(\mathcal{X},\mathcal{Y}) \;=\; E\Big(\big(\mathcal{X}-E(\mathcal{X})\big)\big(\mathcal{Y}-E(\mathcal{Y})\big)\Big) \;= \\ &=\; E\Big(\mathcal{X}\mathcal{Y} \,-\, \mathcal{X}\,E(\mathcal{Y}) \,-\, E(\mathcal{X})\,\mathcal{Y} \,+\, E(\mathcal{X})\,E(\mathcal{Y})\Big) \;= \\ &=\; E(\mathcal{X}\mathcal{Y}) \,-\, E(\mathcal{X})\,E(\mathcal{Y}) \;.\end{aligned} \tag{2.9}$$

The covariance $\mathrm{cov}(\mathcal{X},\mathcal{Y})$ and the *coefficient of correlation* $\rho(\mathcal{X},\mathcal{Y})$,

$$\mathrm{cov}(\mathcal{X},\mathcal{Y}) \;=\; E(\mathcal{X}\mathcal{Y}) \,-\, E(\mathcal{X})\,E(\mathcal{Y}) \;\text{ and }\; \rho(\mathcal{X},\mathcal{Y}) \;=\; \frac{\mathrm{cov}(\mathcal{X},\mathcal{Y})}{\sigma(\mathcal{X})\,\sigma(\mathcal{Y})} \;, \tag{2.9'}$$

are a measure of correlation between the two variables. As a consequence of the Cauchy-Schwarz inequality we have $-1 \leq \rho(\mathcal{X},\mathcal{Y}) \leq 1$. If covariance and correlation coefficient are equal to zero, the two random variables $\mathcal{X}$ and $\mathcal{Y}$

are *uncorrelated*. Independence implies lack of correlation but the latter is in general the weaker property (section 2.3.4).

In addition to the expectation value two more quantities are used to characterize the center of probability distributions (figure 2.1): (i) The *median* $\bar{\mu}$ is the value at which the number of points of a distribution at lower values of matches exactly the number of points at higher values as expressed in terms of two inequalities,

$$\begin{aligned} P(\mathcal{X} \leq \bar{\mu}) \;\geq\; \frac{1}{2} \;\text{ and }\; P(\mathcal{X} \geq \bar{\mu}) \;\geq\; \frac{1}{2} \;\text{ or} \\ \int_{-\infty}^{\bar{\mu}} dF(x) \;\geq\; \frac{1}{2} \;\text{ and }\; \int_{\bar{\mu}}^{+\infty} dF(x) \;\geq\; \frac{1}{2}\,, \end{aligned} \tag{2.10}$$

where Lebesgue-Stieltjes integration is applied or in case of an absolutely continuous distribution the condition simplifies to

$$P(\mathcal{X} \leq \bar{\mu}) \;=\; P(\mathcal{X} \geq \bar{\mu}) \;=\; \int_{-\infty}^{\bar{\mu}} f(x)\,\mathrm{d}x \;=\; \frac{1}{2}\,, \tag{2.10'}$$

and (ii) the *mode* $\tilde{\mu}$ of a distribution is the most frequent value – the value that is most likely to obtain through sampling – and it is obtained as the maximum of the probability mass function for discrete distribution or as the maximum of the probability density in the continuous case. An illustrative example for the discrete case is the probability mass function of the scores for throwing to dice (The mode in figure 1.11 is $\tilde{\mu} = 7$). A probability distribution may have more than one mode. Bimodal distributions occur occasionally and then the two modes provide much more information on the expected outcomes than mean or median (see also subsection 2.4.8).

Median and mean are related by an inequality, which says that the difference between both is bounded by one standard deviation [192, 211]:

$$\begin{aligned} |\mu - \overline{\mu}| \;&=\; |E(\mathcal{X} - \overline{\mu}| \;\leq\; E(|\mathcal{X} - \overline{\mu}|) \;\leq \\ &\leq\; E(|\mathcal{X} - \mu|) \;\leq\; \sqrt{E\Big((\mathcal{X} - \mu)^2\Big)} \;=\; \sigma\,. \end{aligned} \tag{2.11}$$

The absolute difference between mean and median can't be larger than one standard deviation of the distribution.

For many purposes a generalization of the median from two to n equally sized data sets is useful. The *quantiles* are points taken at regular intervals from the cumulative distribution function $F(x)$ of a random variable $\mathcal{X}$. Ordered data are divided into n essentially equal-sized subsets and accordingly, $(n-1)$ points on the x-axis separate the subsets. Then, the k-th n-quantile is defined by $P(\mathcal{X} < x) \leq \frac{k}{n} = p$ (figure 2.2) or equivalently

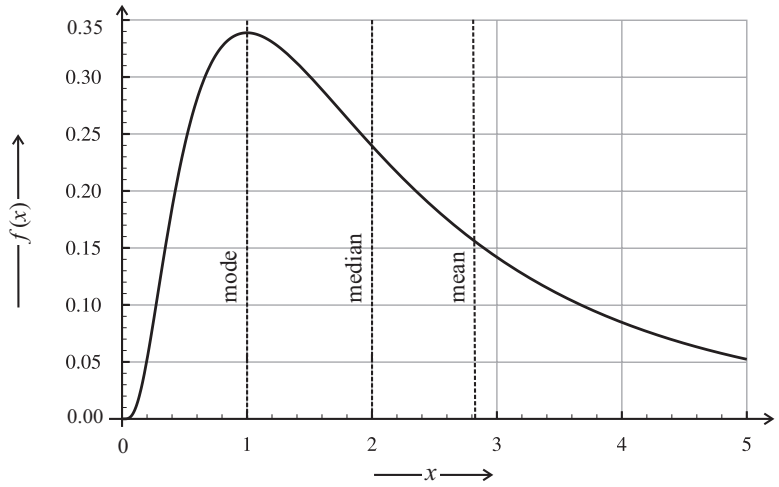

Fig. 2.1 Probability densities and moments. As an example of an asymmetric distribution with highly different values for mode, median, and mean, the lognormal density

$$f(x) = \tfrac{1}{\sqrt{2\pi}\,\sigma\,x} \exp\bigl(-(\ln x - \nu)^2/(2\sigma^2)\bigr)$$

is shown. Parameters values: $\sigma = \sqrt{\ln 2}$, $\nu = \ln 2$ yielding $\tilde{\mu} = \exp(\nu - \sigma^2/2) = 1$ for the mode, $\bar{\mu} = \exp(\nu) = 2$ for the median and $\mu = \exp(\nu + \sigma^2/2) = 2\sqrt{2}$ for the mean, respectively. The sequence mode < median < mean is characteristic for distributions with positive skewness whereas the opposite sequence mean < median < mode is found in cases of negative skewness (see also figure 2.3).

$$F^{-1}(p) := \inf\{x \in \mathrm{R} : F(x) \geq p\} \quad \text{and} \quad p = \int_{-\infty}^{x} \mathrm{d}F(u) \; . \tag{2.12}$$

In case the random variable has a probability density the integral simplifies to $p = \int_{-\infty}^{x} f(u)\mathrm{d}u$. The median is simply the value of x for $p = \frac{1}{2}$. For partitioning into four parts we haver the first or lower quartile at $p = \frac{1}{4}$, the second quartile or median at $p = \frac{1}{2}$, and the third or upper quartile at $p = \frac{3}{4}$. The lower quartile contains 25 % of the data, the median 50 %, and the upper quartile eventually 75 %.

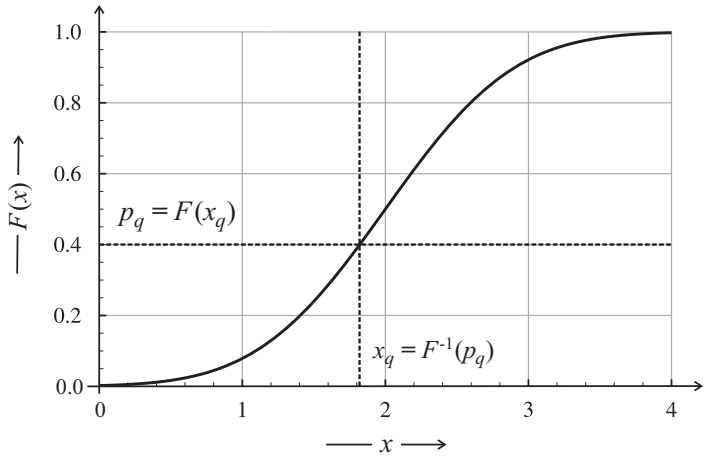

Fig. 2.2 Definition and determination of quantiles. A quantile q with $p_q = k/n$ defines a value x_q at which the (cumulative) probability distribution reaches the value $F(x_q) = p_q$ corresponding to $P(\mathcal{X} < x) \leq p$. The figure shows how the position of the quantile $p_q = k/n$ is used to determine its value $x_q(p)$. In particular we use here the normal distribution $\mathcal{N}(x)$ as function $F(x)$ and the computation yields $F(x_q) = \frac{1}{2}\Big(1 + \mathrm{erf}\big(\frac{x_q - \nu}{\sqrt{2\sigma^2}}\big)\Big) = p_q$. Parameter choice: $\nu = 2$, $\sigma^2 = \frac{1}{2}$, and for the quantile $(n = 5, k = 2)$, yielding $p_q = 2/5$ and $x_q = 1.8209$.

2.1.2 Higher moments

Two other quantities related to higher moments are frequently used for a more detailed characterization of probability distributions:[3] (i) The *skewness*

$$\gamma_1 \;=\; \frac{\mu_3}{\mu_2^{3/2}} \;=\; \frac{\mu_3}{\sigma^3} \;=\; \frac{E\Big(\big(\mathcal{X} - E(\mathcal{X})\big)^3\Big)}{\Big(E\Big(\big(\mathcal{X} - E(\mathcal{X})\big)^2\Big)\Big)^{3/2}} \tag{2.13}$$

and (ii) *kurtosis*, which is either defined as the fourth standardized moment β_2 or in terms of cumulants given as *excess kurtosis*, γ_2,

[3] In contrast to expectation value, variance and standard deviation, skewness and kurtosis are not uniquely defined and it is necessary therefore to check carefully the author's definitions when reading text from literature.

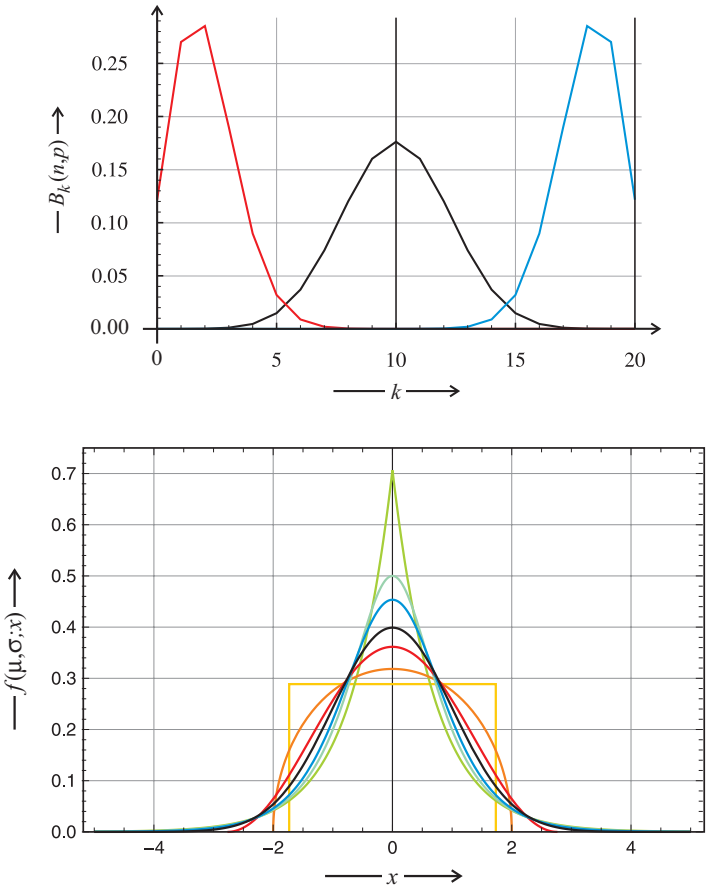

Fig. 2.3 Skewness and kurtosis. The upper part of the figures illustrates the sign of skewness with asymmetric density functions. The examples are taken form the binomial distribution $B_k(n,p)$: $\gamma_1 = (1-2p)/\sqrt{np(1-p)}$ with $p = 0.1$ (red), 0.5 (black; symmetric), and 0.9 (blue) with the values $\gamma_1 = 0.596$, 0, -0.596.
Densities with different kurtosis are compared in the lower part of the figure: The Laplace distribution (chartreuse), the hyperbolic secant distribution (green), and the logistic distribution (blue) are leptokurtic with excess kurtosis values 3, 2, and 1.2, respectively. The normal distribution is the reference curve with excess kurtosis 0 (black). The raised cosine distribution (red), the Wigner semicircle distribution (orange), and the uniform distribution (yellow) are platykurtic with excess kurtosis values of -0.593762, -1, and -1.2 respectively. All densities are calibrated such that $\mu = 0$ and $\sigma^2 = 1$ (The picture is recalculated and redrawn from `http://en.wikipedia.org/wiki/Kurtosis`, March 30,2011).

$$\beta_2 \;=\; \frac{\mu_4}{\mu_2^2} \;=\; \frac{\mu_4}{\sigma^4} \;=\; \frac{E\Big(\big(\mathcal{X}-E(\mathcal{X})\big)^4\Big)}{\Big(E\Big(\big(\mathcal{X}-E(\mathcal{X})\big)^2\Big)\Big)^2} \quad \text{and}$$
$$\gamma_2 \;=\; \frac{\kappa_4}{\kappa_2^2} \;=\; \frac{\mu_4}{\sigma^4} \;-\; 3 \;=\; \beta_2 \;-\; 3\,. \tag{2.14}$$

Skewness is a measure of the asymmetry of the probability density: curves that are symmetric about the mean have zero skew, *negative skew* implies a longer left tail of the distribution caused by more low values, and *positive skew* is characteristic for a distribution with a longer right tail. Positive skew is quite common with empirical data (see, for example the log-normal distribution in section 2.4.1).

Kurtosis characterizes the degree of *peakedness* of a distribution. High kurtosis implies a sharper peak and flat tails, low kurtosis in contrary characterizes flat or round peaks and thin tails. Distributions are called *leptokurtic* if they have a positive excess kurtosis and therefore are sharper peak and a thicker tail than the *normal distribution* (section 2.3.3), which is taken as a reference with zero kurtosis. Distributions are characterized as *platykurtic* if they have a negative excess kurtosis, a broader peak and thinner tails (see figure 2.3; the distributions compared there with respect to kurtosis are standardized to $\mu = 0$ and $\sigma^2 = 1$):

(i) Laplace distribution: $f(x) = \frac{1}{2b}\,\exp\Big(-\frac{|x-\mu|}{b}\Big)$, $b = \frac{1}{\sqrt{2}}$,
(ii) hyperbolic secant distribution: $f(x) = \frac{1}{2}\,\mathrm{sech}\,\big(\frac{\pi}{2}\,x\big)$,
(iii) logistic distribution: $f(x) = \frac{e^{-(x-\mu)/s}}{s\,(1+e^{-(x-\mu)/s})^2}$, $s = \sqrt{3}/\pi$,
(iv) normal distribution: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-(x-\mu)^2/(2\sigma^2)}$,
(v) raised cosine distribution: $f(x) = \frac{1}{2s}\Big(1+\cos(\pi(x-\mu)/s)\Big)$, $s = \frac{1}{\sqrt{\frac{1}{3}-\frac{2}{\pi^2}}}$,
(vi) Wigner's semicircle: $f(x) = \frac{2}{\pi r^2}\sqrt{r^2-x^2}$, $r = 2$, and
(vii) uniform distribution: $f(x) = \frac{1}{b-a}$, $b-a = 2\sqrt{3}$.

These seven functions span the whole range of maxima from a sharp peak to a completely flat plateau with the normal distribution chosen as reference function (figure 2.3).

One property of skewness and kurtosis being caused by definition is important to note: The expectation value, the standard deviation, and the variance are quantities with dimensions, whereas skewness and kurtosis are defined as dimensionless numbers.

The cumulants κ_n are the coefficients of a series expansion of the logarithm of the *characteristic function* (2.28), which in turn is the Fourier transform of the probability density function, $f(x)$, or the logarithm of the *moment generating function* (2.27)(see section 2.2):

$$
\begin{aligned}
h(s) \;&=\; \ln\,\phi(s) \;=\; \sum_{n=1}^{\infty} \kappa_n \,\frac{(\dot{u}\,s)^n}{n!} \;\text{ with} \\
\phi(s) \;&=\; \int_{-\infty}^{+\infty} \exp(\dot{u}\,s\,x)\,f(x)\,\mathrm{d}x \;.
\end{aligned}
\tag{2.15}
$$

The first five cumulants κ_n $(n = 1, \ldots, 5)$ expressed in terms of the expectation value μ and the central moments μ_n $(\mu_1 = 0)$ are

$$
\begin{aligned}
\kappa_1 \;&=\; \mu \\
\kappa_2 \;&=\; \mu_2 \\
\kappa_3 \;&=\; \mu_3 \\
\kappa_4 \;&=\; \mu_4 \,-\, 3\,\mu_2^2 \\
\kappa_5 \;&=\; \mu_5 \,-\, 10\,\mu_2\mu_3 \;.
\end{aligned}
\tag{2.16}
$$

We shall come back to the use of cumulants κ_n in sections 2.3 and 2.4 where we shall compare frequently used individual probability densities and in section 2.5 when we apply k-statistics in order to compute empirical moments from incomplete data sets.

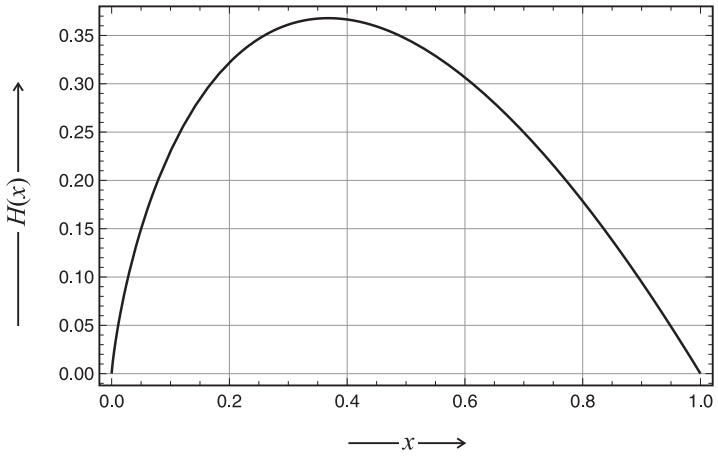

Fig. 2.4 The functional relation of information entropy. The plot shows the function $H = -\,x\,\ln x$ in the range $0 \le x \le 1$. For $x = 0$ the convention of probability theory, $-\,0\ln 0 = 0 \cdot \infty = 0$, is applied.

2.1.3 Information entropy

Information theory has been developed during World War Two as theory of communication of secret messages. No wonder that the theory has been developed at Bell Labs and the person who was the leading figures in this area was an American cryptographer, electronic engineer and computer scientist, Claude Elwood Shannon [261, 262]. One of the central issues of information theory is *self-information* or the *content of information* that can be encoded, for example, in a sequence of given length. Commonly one thinks about binary sequences and therefore the information is measured in *binary digits* or *bits*.[4]

$$I(\omega) \;=\; \mathrm{lb}\left(\frac{1}{P(\omega)}\right) \;=\; -\,\mathrm{lb}\,P(\omega) \tag{2.17}$$

The rationale behind this expression is the definition of a measure of information that is positive and additive for independent events. From equation (1.34) follows:

$$P(AB) \;=\; P(A)\cdot P(B) \;\;\Longrightarrow\;\; I(A\cap B) \;=\; I(AB) \;=\; I(A) \,+\, I(B)\;,$$

and this relation is fulfilled by the logarithm. Since $P(\omega) \leq 1$ by definition, the negative logarithm is a positive quantity. Equation (2.17) yields zero information for an event taking place with certainty, $P(\omega) = 1$. The outcome of the fair coin toss with $P(\omega) = \frac{1}{2}$ provides 1 bit information, and rolling two 'six' with two dice in one throw has a probability $P(\omega) = \frac{1}{36}$ and yields 5.17 bits (For a modern treatise of information theory and entropy see [110]).

Finite sample space. In order to measure the information content of a probability distribution Claude Shannon introduced the information entropy, which is simply the expectation value of the information content and which is represented by a function that resembles the expression for the thermodynamic entropy in statistical mechanics. We consider first the discrete case of a probability mass function $p_k = P(\mathcal{X} = x_k)\,,\; k \in \mathbb{N}_{>0}, k \leq n$:

$$H(p) \;=\; -\sum_{k=1}^{n} p_k \,\log\, p_k \;\;\text{ with }\;\; p_k \geq 0\,, \sum_{k=1}^{n} p_k = 1\;. \tag{2.18}$$

Thus, the entropy can be visualized as the expectation value of the negative logarithm of the probabilities

[4] The logarithm is taken to the base 2 and it is commonly called *binary logarithm* or *logarithmus dualis*: $\log_2 \equiv \mathrm{lb} \equiv \mathrm{ld}$. In informatics the conventional unit of information is the *byte*: 1 byte (B) = 8 bits being tantamount to the coding capacity of an eight digit binary sequence. Although there is little chance of confusion, one should be aware that in the International System of Units 'B' is the abbreviation for the acoustical unit 'bel', which is the unit for measuring the signal strength of sound.

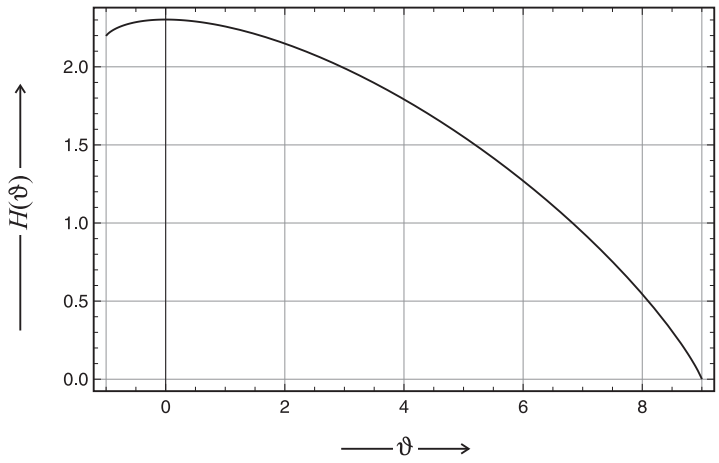

Fig. 2.5 Maximum information entropy. The discrete probability distribution with maximal information entropy in the uniform distribution $\mathcal{U}_p$. The entropy of the probability distribution $p_1 = \frac{1-\vartheta}{n}$ and $p_j = (1 - \frac{\vartheta}{n-1})/n \,\forall\, j = 2, 3, \ldots, n$ with $n = 10$ is plotted against the parameter ϑ. All probabilities p_k are defined and the entropy $H(\vartheta)$ is real and non-negative on the interval $-1 \leq \vartheta \leq 9$ and passes a maximum at $\vartheta = 0$.

$$H(p) \;=\; E(-\log\, p_k) \;=\; E\left(\log\Big(\frac{1}{p_k}\Big)\right) \;,$$

where the term $\log(1/p_k)$ can be viewed as the number of bits to be assigned to to the point x_k provided the binary logarithm is used ($\log \equiv \mathrm{lb}$).

The functional relationship, $H = -x \log x$, on the interval $0 \leq x \leq 1$ underlying the information entropy is a concave function (figure 2.4). It is easily shown that the entropy of a discrete probability distribution is always non-negative. A verification of this conjecture can be given, for example, by considering the two extreme cases: (i) there almost certainly only one outcome, $p_1 = P(\mathcal{X} = x_1) = 1$ and $p_j = P(\mathcal{X} = x_j) = 0 \;\forall\; j \in \mathbb{N}_{>0}, j \neq 1$, and the information entropy $H = 0$ in this completely determined case, and (ii) all events have the same probability, we are dealing with the uniform distribution, $p_k = P(\mathcal{X} = x_k) = \frac{1}{n}$, or a case of the principle of indifference, the entropy is positive, and takes on its maximum value, $H(p) = \log n$. The entropies of all other discrete distributions lie in between:

$$0 \;\leq\; H(p) \;\leq\; \log n \quad \text{or} \quad H(p) \;\leq\; \log\, n \;, \tag{2.19}$$

and the value of the entropy is a measure of the lack of information on the distribution. Case (i) is deterministic and we have the full information on the outcome *a priori*, whereas case (ii) provides maximal uncertainty because all outcomes have the same probability. A rigorous proof that the uniform

Table 2.1 Probability distributions with maximum information entropy. The table compares three probability distributions with maximum entropy: (i) the discrete uniform distribution on the support $\Omega = \{1 \leq k \leq n, k \in \mathbb{N}\}$, (ii) the exponential distribution on $\Omega = \mathbb{R}_{\geq 0}$, and the normal distribution on $\Omega = \mathbb{R}$.

Distribution	Space Ω	Density	Mean	Var	Entropy
uniform	$\mathbb{N}_{>0}$	$\frac{1}{n} \; \forall \; k = 1, \ldots, n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\log\, n$
exponential	$\mathbb{R}_{\geq 0}$	$\frac{1}{\mu}\, e^{-x/\mu}$	μ	μ^2	$1 + \log\, \mu$
normal	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu^2)}{2\sigma^2}}$	μ	σ^2	$\Big(1 + \log(2\pi\sigma^2)\Big)/2$

distribution has maximum information entropy among all discrete distributions is found in the literature [36, 40]. We dispense from reproducing the proof here but we illustrate by means of figure 2.5: The starting point is the uniform distribution of n events with a probability of $p = \frac{1}{n}$ for each one, and then we attribute a different probability to a single event: $p_1 = \frac{1-\vartheta}{n}$ and $p_j = (1 - \frac{\vartheta}{n-1})/n$ $(j = 2, 3, \ldots, n)$. The entropy of the distribution is considered as a function of ϑ and indeed the maximum occurs at $\vartheta = 0$.

Infinite measurable sample space. The information entropy of a continuous probability density $p(x)$ with $x \in \mathbb{R}$ is calculated by means of integration

$$H(p) \;=\; -\int_{-\infty}^{+\infty} p(x)\, \log\, p(x)\, \mathrm{d}x \;\; \text{with} \;\; p_k \geq 0\,, \int_{-\infty}^{+\infty} p(x)\, \mathrm{d}x = 1\,, \;\; (2.18')$$

and as in the discrete case we can write the entropy as an expectation value of $\log(1/p)$:

$$H(p) \;=\; E\Big(-\log\, p(x)\Big) \;=\; E\left(\log\Big(\frac{1}{p(x)}\Big)\right) \;.$$

We consider two specific examples that are distributions with maximum entropy: the exponential distribution (section 2.4.4) on $\Omega = \mathbb{R}_{\geq 0}$ with the density

$$f_{\exp}(x) \;=\; \frac{1}{\mu}\, e^{-\frac{x}{\mu}} \;,$$

the mean μ and the variance $\mathrm{var} = \mu^2$, and the normal distribution (section 2.3.3) on $\Omega = \mathbb{R}$ with the density

$$f_{\mathcal{N}}(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \;,$$

the mean μ and the variance $\mathrm{var} = \sigma^2$

In the discrete case we made a seemingly unconstrained search for the distribution of maximum entropy, although the discrete uniform distribution contained the number of sample points n as input restriction and indeed, n appears as parameter in the analytical expression for the entropy (table 2.1). Now, in the continuous case the constraints become more evident since we shall use fixed mean (μ) or fixed variance (σ^2) as the basis of comparison in the search for distributions with maximum entropy.

The entropy of the exponential density on the sample space $\Omega = \mathbb{R}_{\geq 0}$ with mean μ and variance $\mathrm{var} = \mu^2$ is calculated to be

$$H(f_{\exp}) \;=\; -\int_0^\infty \frac{1}{\mu}\, e^{-x/\mu} \left(-\log\,\mu \,-\, \frac{x}{\mu} \right) \mathrm{d}x \;=\; 1 \,+\, \log\,\mu\;. \qquad (2.20)$$

In contrast to the discrete case the entropy of the exponential probability density can become negative for small μ-values as can be easily visualized by considering the shape of the density: Since $\lim_{x\to 0} f_{\exp}(x) = 1/\mu$, an appreciable fraction of the density function adopts values $f_{\exp}(x) > 1$ for sufficiently small μ and then $-p\,\log p < 0$ is negative. Among all continuous probability distributions with mean $\mu > 0$ on the support $\mathbb{R}_{\geq 0} = [0, \infty[$ the exponential distribution has the maximum entropy. Proofs for this conjecture are available in the literature [36, 40, 231].

For the normal density we obtain from equation (2.18'):

$$\begin{aligned} H(f_{\mathcal{N}}) \;&=\; -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2 \right) \mathrm{d}x \\ &=\; \frac{1}{2}\Big(1 + \log(2\pi\sigma^2)\Big)\;. \end{aligned} \qquad (2.21)$$

It is not unexpected that the information entropy of the normal distribution is independent of the mean μ, which causes nothing but a shift of the whole distribution along the x-axis: all Gaussian densities with the same variance σ^2 have the same entropy. Again we see that the entropy of the normal probability density can become negative for sufficiently small values of σ^2. The normal distribution is distinguished among all continuous distributions on $\Omega = \mathbb{R}$ with given variance σ^2, since it the normal is the distribution with maximum entropy. Several proofs for this theorem were developed, we refer again to the literature [36, 40, 231]. The three distributions with maximum entropy are compared in table 2.1.

Principle of maximum entropy. The information entropy can be interpreted as the required amount of information we would need in order to fully describe the system. Equations (2.18) and (2.18') are the basis of a search for probability distribution with maximum entropy under certain constraints, for example constant mean μ or constant variance σ^2. The maximum entropy principle has been introduced by the American physicist Edwin Thompson Jaynes as a method of statistical inference [145, 146]: He suggests to use those probability distributions, which satisfy the prescribed constraints and have

the largest entropy. The rationale for this choice is to use a probability distribution that reflects our knowledge and does not contain any unwarranted information. The predictions made on the basis of a probability distribution with maximum entropy should be least surprising. If we chose a distribution with smaller entropy, this distribution would contain more information than justified by our *a priori* understanding of the problem. It is useful to illustrate a typical strategy [36]: *"..., the principle of maximum entropy guides us to the best probability distribution that reflects our current knowledge and it tells us what to do if experimental data do not agree with predictions coming from our chosen distribution: Understand why the phenomenon being studied behaves in an unexpected way, find a previously unseen constraint, and maximize the entropy over the distributions that satisfy all constraints we are now aware of, including the new one."* We realize a different way of thinking about probability that becomes even more evident in Bayesian statistics, which is sketched in sections 1.3 and 2.5.4.

The choice of the word *entropy* for the expected information content of a distribution is not accidental. Ludwig Boltzmann's statistical formula[5]

$$S \;=\; k_{\mathrm{B}}\,\ln W \quad \text{with} \quad W \;=\; \frac{N!}{N_1!N_2!\cdots N_m!}\;, \tag{2.22}$$

with W being the so-called thermodynamic probability and k_{B} Boltzmann's constant: $k_{\mathrm{B}} = 1.38065 \times 10^{-23}\,\mathrm{Joule{\cdot}Kelvin}^{-1}$ and $N = \sum_{j=1}^{m} N_j$ being the total number of particles being distributed over m states with the frequencies $p_k = N_k/N$ and $\sum_{j=1}^{m} p_j = 1$. The number of particles N is commonly very large and Stirling's formula named after the Scottish mathematician James Stirling applies: $n! \approx n\ln n$, and this leads to:

$$S \;=\; k_{\mathrm{B}}\left(N\,\ln N - \sum_{i=1}^{m} N_i\,\ln N_i\right) \;=\; -\,k_{\mathrm{B}}\,N\left(-\,\ln N + \sum_{i=1}^{m}\frac{N_i}{N}\,\ln N_i\right) \;=$$

$$=\; -\,k_{\mathrm{B}}\,N\sum_{i=1}^{m} p_i\,\ln p_i\;.$$

For a single particle we obtain an entropy

$$s \;=\; \frac{S}{N} \;=\; -\,k_{\mathrm{B}}\sum_{i=1}^{m} p_i\,\ln p_i\;, \tag{2.22'}$$

which is identical with Shannon's formula (2.18) except the factor containing the universal constant k_{B}.

[5] A few remarks are important: Equation (2.22) in Max Planck's expression for the entropy in statistical mechanics, although it has been carved in Boltzmann's tomb stone, and W is called a probability despite the fact that it is not normalized, $W \geq 0$.

Eventually we point at some important differences between thermodynamic entropy and information entropy that should be kept in mind when discussing analogies between them. The thermodynamic principle of maximum entropy is a physical law known as the second law of thermodynamics: The entropy of an isolated system[6] is nondecreasing in general and increasing if processes are taking place, and hence approaches a maximum. The principle of maximum entropy in statistics is a rule for appropriate design of distribution functions and has the rank of a guideline and not that of a natural law. Thermodynamic entropy is an *extensive property* and this means that it increases with the size of the system. Information entropy, on the other hand, is an *intensive property* and insensitive to size. An illustrative example of this difference is due to the Russian biophysicist Mikhail Vladimirovich Volkenshtein [296]: Considering the process of flipping a coin in reality and calculating all contributions to the process shows that the information entropy is a minute contribution to the thermodynamic entropy only. The total thermodynamic entropy change as a result of the coin flipping process is dominated by far by the metabolic contributions of the flipping individual, as there are muscle contraction, joint rotations, and by the heat production on the surface where the coin lands, etc. Imagine the thermodynamic entropy production if you flip a coin two meters high – the gain in information remains still one bit!

[6] A isolated system exchanges neither matter nor energy with its environment (For isolated, closed, and open systems see also section 4.3).

2.2 Generating functions

In this section we introduce auxiliary functions, which allow for the derivation of compact representations of probability distributions and which provide convenient tools for handling functions of probabilities. The generating functions commonly contain one or more auxiliary variables – here denoted by s, which are lacking direct physical meaning but enable straightforward calculation of properties of random variables at certain values of$s$. In particular we shall make use of probability generating functions $g(s)$, moment generating functions $M(s)$ and characteristic functions $\phi(s)$. The characteristic function $\phi(s)$ exists for all distributions but we shall encounter cases where no probability nd moment generating functions exist (see, for example, the Cauchy-Lorentz distribution in subsection 2.4.6). In addition to the three generating functions mentioned here other functions are in use as well. An example is the *cumulant generating function* that is lacking a uniform definition. It is either the logarithm of the moment generating function or the logarithm of the characteristic function – we shall mention both.

2.2.1 Probability generating functions

Let $\mathcal{X}$ be a random variable taking only non-negative integer values with a probability distribution given by

$$P(\mathcal{X} = j) \;=\; a_j \,;\; j = 0, 1, 2, \ldots \;. \tag{2.23}$$

A auxiliary variable s is introduced and the *probability generating function* is expressed by an infinite power series

$$g(s) \;=\; a_0 \,+\, a_1\, s \,+\, a_2\, s^2 \,+\, \ldots \;=\; \sum_{j=0}^{\infty} a_j\, s^j \;. \tag{2.24}$$

As we shall show later, the full information on the probability distribution is encapsulated in the coefficients a_j ($j \geq 0$). In most cases s is a real valued variable, although it can be of advantage to consider also complex s. Recalling $\sum_j a_j = 1$ from (2.23) we verify easily that the power series (2.24) converges for $|s| \leq 1$:

$$|g(s)| \;\leq\; \sum_{j=0}^{\infty} |a_j| \cdot |s|^j \;\leq\; \sum_{j=0}^{\infty} a_j \;=\; 1 \;, \quad \text{for } |s| \leq 1 \;.$$

For $|s| < 1$ we can differentiate[7] the series term by term in order to calculate the derivatives of the generating function $g(s)$

$$\frac{\mathrm{d}g}{\mathrm{d}s} \;=\; g'(s) \;=\; a_1 \,+\, 2\,a_2 s \,+\, 3\,a_3 s^2 \,+\, \ldots \;=\; \sum_{n=1}^{\infty} n\,a_n s^{n-1} \;,$$

$$\frac{\mathrm{d}^2 g}{\mathrm{d}s^2} \;=\; g''(s) \;=\; 2\,a_2 \,+\, 6\,a_3 s \,+\, \ldots \;=\; \sum_{n=2}^{\infty} n(n-1)\,a_n s^{n-2} \;,$$

and, in general, we have

$$\frac{\mathrm{d}^j g}{\mathrm{d}s^j} \;=\; g^{(j)}(s) \;=\; \sum_{n=j}^{\infty} n(n-1)\ldots(n-j+1)\,a_n s^{n-j} \;=$$

$$= \sum_{n=j}^{\infty} (n)_j \; a_n \, s^{n-j} \;=\; \sum_{n=j}^{\infty} \binom{n}{j} j\,!\; a_n \, s^{n-j} \;,$$

where $(x)_n$ stands for the falling Pochhammer symbol.[8] Setting $s = 0$, all terms vanish except the constant term

$$\frac{\mathrm{d}^j g}{\mathrm{d}s^j}\,\Big|_{s=0} = \; g^{(j)}(0) \;=\; j\,!\,a_j \quad \text{or} \quad a_j \;=\; \frac{1}{j\,!}\,g^{(j)}(0) \;.$$

In this way all a_j's may be obtained by consecutive differentiation from the generating function and alternatively the generating function can be determined from the known probability distribution.

Putting $s = 1$ in $g'(s)$ and $g''(s)$ we can compute the first and second moments of the distribution of $\mathcal{X}$:

$$\begin{aligned}
g'(1) \;&=\; \sum_{n=0}^{\infty} n\,a_n \;=\; E(\mathcal{X}) \;, \\
g''(1) \;&=\; \sum_{n=0}^{\infty} n^2\,a_n \;-\; \sum_{n=0}^{\infty} n\,a_n \;=\; E(\mathcal{X}^2) \;-\; E(\mathcal{X}) \qquad (2.25) \\
E(\mathcal{X}) \;&=\; g'(1)\,, \quad \text{and} \\
E(\mathcal{X}^2) \;&=\; g'(1) + g''(1) \quad \text{and} \quad \sigma^2(\mathcal{X}) \;=\; g'(1) + g''(1) - g'(1)^2 \;.
\end{aligned}$$

[7] Since we need the derivatives very often in this section, we make advantage of short notations: $\mathrm{d}g(s)/\mathrm{d}s = g'(s)$, $\mathrm{d}^2 g(s)/\mathrm{d}s^2 = g''(s)$, and $\mathrm{d}^j g(s)/\mathrm{d}s^j = g^{(j)}(s)$ and for simplicity also $(\mathrm{d}g/\mathrm{d}s)|_{s=k} = g'(k)$ and $(\mathrm{d}^2 g/\mathrm{d}s^2)|_{s=k} = g''(k)$ $(k \in \mathbb{N})$.

[8] The Pochhammer symbol $(x)_n = x(x-1)\ldots(x-n+1)$ is used here as, for example, in combinatorics for the *falling factorial*, the *rising factorial* is written as $x^{(n)} = x(x+1)\ldots(x+n-1)$. We remark that in the theory of special function, in particular for the hypergeometric functions, $(x)_n$ is used for the rising factorial.

We summarize: The probability distribution of a non-negative integer values random variable can be converted into a generating function without loosing information. The generating function is uniquely determined by the distribution and *vice versa*.

2.2.2 Moment generating functions

Basis of the moment generating function is the series expansion of the exponential of the random variable $\mathcal{X}$

$$e^{\mathcal{X}s} \;=\; 1 \,+\, \mathcal{X}\,s \,+\, \frac{\mathcal{X}^2}{2!}\,s^2 \,+\, \frac{\mathcal{X}^3}{3!}\,s^3 \,\ldots\;.$$

The *moment generating function* allows for direct computation of the moments of a probability distribution as defined in equation (2.23) since we have:

$$M_{\mathcal{X}}(s) \;=\; E(e^{\mathcal{X}\,s}) \;=\; 1 + \hat{\mu}_1\,s + \frac{\hat{\mu}_2}{2!}\,s^2 + \frac{\hat{\mu}_3}{3!}\,s^3 \,\ldots \;=\; 1 + \sum_{n=1}^{\infty} \hat{\mu}_n \frac{s^n}{n!}\;. \quad (2.26)$$

wherein $\hat{\mu}_i$ is the i-th raw moment. The moments are obtained by differentiating $M_{\mathcal{X}}(s)$ i times with respect to s and then setting $s = 0$

$$E(\mathcal{X}^n) \;=\; \hat{\mu}_n \;=\; M_{\mathcal{X}}^{(n)} \;=\; \frac{\mathrm{d}^n M_{\mathcal{X}}}{\mathrm{d}s^n}\,\Big|_{s=0}\;.$$

A probability distribution thus has (at least) as many moments as many times the moment generating function can be continuously differentiated (see also characteristic function in subsection 2.2.3). If two distributions have the same moment generating functions they are identical at all points:

$$M_{\mathcal{X}}(s) \;=\; M_{\mathcal{Y}}(s) \iff F_{\mathcal{X}}(x) \;=\; F_{\mathcal{Y}}(x)\;.$$

This statement, however, does not imply that two distributions are identical when they have the same moments, because in some cases the moments exist but the moment generating function does not, since the limit $\lim_{n\to\infty} \sum_{k=0}^{n} \frac{\hat{\mu}_k\, s^k}{k!}$ diverges as, for example, in case of the logarithmic normal distribution.

The real *cumulant generating function* is the formal logarithm of the moment generating function that can be expanded in a power series

$$
\begin{aligned}
k(s) \;&=\; \ln\Big(E\big(e^{\mathcal{X}s}\big)\Big) \;=\; -\sum_{n=1}^{\infty}\frac{1}{n}\Big(1-E\big(e^{\mathcal{X}s}\big)\Big)^n \;= \\
&=\; -\sum_{n=1}^{\infty}\frac{1}{n}\left(-\sum_{m=1}^{\infty}\hat{\mu}_m\frac{s^m}{m!}\right)^n \;= \qquad (2.27)\\
&=\; \hat{\mu}_1 s \;+\; \big(\hat{\mu}_2-\hat{\mu}_1^2\big)\frac{s^2}{2!} \;+\; \big(\hat{\mu}_3-3\hat{\mu}_2\hat{\mu}_1+2\hat{\mu}_1^3\big)\frac{s^3}{3!} \;+\; \ldots
\end{aligned}
$$

The cumulants κ_n are obtained from the cumulant generating function through the n-th differentiation of $k(s)$ and calculating the derivative at $s=0$:

$$
\begin{aligned}
\kappa_1 \;&=\; \frac{\partial k(s)}{\partial s}\Big|_{s=0} \;=\; \hat{\mu}_1 \;=\; \mu\,,\\
\kappa_2 \;&=\; \frac{\partial^2 k(s)}{\partial s^2}\Big|_{s=0} \;=\; \hat{\mu}_2 \;-\; \hat{\mu}_1^2 \;=\; \sigma^2\,,\\
\kappa_3 \;&=\; \frac{\partial^3 k(s)}{\partial s^3}\Big|_{s=0} \;=\; \hat{\mu}_3 \;-\; 3\hat{\mu}_2\mu_1 \;+\; 1\hat{\mu}_1^3 \;=\; \mu_3\,, \qquad (2.16')\\
&\;\;\vdots\\
\kappa_n \;&=\; \frac{\partial^n k(s)}{\partial s^n}\Big|_{s=0}\,,\\
&\;\;\vdots
\end{aligned}
$$

As shown in equation (2.16) the first three cumulants coincide with the centered moments μ_1, μ_2, and μ_3. All higher cumulants are polynomials of two or more centered moments.

2.2.3 Characteristic functions

Like the moment generating function the *characteristic function* $\phi(s)$ of a random variable $\mathcal{X}$ completely describes the probability distribution $F(x)$. It is defined by

$$
\phi(s) \;=\; \int_{-\infty}^{+\infty}\exp(\dot{\imath}\,s\,x)\,\mathrm{d}F(x) \;=\; \int_{-\infty}^{+\infty}\exp(\dot{\imath}\,s\,x)\,f(x)\,\mathrm{d}x\;, \qquad (2.28)
$$

where the integral over $\mathrm{d}F(x)$ is of Riemann-Stieltjes type. In case a probability density $f(x)$ exists for the random variable $\mathcal{X}$ the characteristic function is (almost) the Fourier transform of the density.[9] From equation (2.28') fol-

[9] The difference between the Fourier transform $\hat{f}(s)$ and the characteristic function $\phi(s)$ of a function $f(x)$,

lows the useful expression $\phi(s) = E(e^{\dot{\imath}s\mathcal{X}})$ that we shall use, for example, in solving most equations for stochastic processes (chapter 3).

The characteristic function exists for all random variables since it is an integral of a bounded continuous function over a space of finite measure. There is a bijection between distribution functions and characteristic functions:

$$\phi_{\mathcal{X}}(s) \;=\; \phi_{\mathcal{Y}}(s) \;\Longleftrightarrow\; F_{\mathcal{X}}(x) \;=\; F_{\mathcal{Y}}(x)\;.$$

If a random variable $\mathcal{X}$ has moments up to k-th order, then the characteristic function $\phi(x)$ is k times continuously differentiable on the entire real line and vice versa if a characteristic function $\phi(x)$ has a k-th derivative at zero, then the random variable $\mathcal{X}$ has all moments up to k if k is even and up to $k-1$ if k is odd:

$$E(\mathcal{X}^k) \;=\; (-\dot{\imath})^k \frac{\mathrm{d}^k\phi(s)}{\mathrm{d}s^k}\,\bigg|_{s=0} \quad\text{and}\quad \frac{\mathrm{d}^k\phi(s)}{\mathrm{d}s^k}\,\bigg|_{s=0} \;=\; \dot{\imath}^k\,E(\mathcal{X}^k)\,. \qquad (2.29)$$

An interesting example is presented by the Cauchy distribution (subsection 2.4.6) with $\phi(s) = \exp(|s|)$: It is not differentiable at $s = 0$ and the distribution has no moments including the expectation value.

The moment generating function is related to the probability generating function $g(s)$ (subsection 2.2.1) and the characteristic function $\phi(s)$ (subsection 2.2.3) by

$$g\,(e^s) \;=\; E\left(e^{\mathcal{X}s}\right) \;=\; M_{\mathcal{X}}(s) \;\text{ and }\; \phi(s) \;=\; M_{\dot{\imath}\mathcal{X}}(s) \;=\; M_{\mathcal{X}}(\dot{\imath}s)\;.$$

All three generating functions are closely related but it may happen that not all three are existing. As said, characteristic functions exist for all probability distributions.

$$\hat{f}(s) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(x)\exp(+\dot{\imath}\,s\,x)\,\mathrm{d}x \quad\text{and}\quad \phi(s) \;=\; \int_{-\infty}^{+\infty} f(x)\exp(+\dot{\imath}\,s\,x)\,\mathrm{d}x\;,$$

is only a matter of the factor $(\sqrt{2\pi})^{-1}$. The Fourier convention above is the one used in modern physics, for other convention see, e.g., *Mathematica*.

Table 2.2 Comparison of several common probability densities. Abbreviation and notations used in the table are: $\Gamma(r,x) = \int_x^\infty s^{r-1}e^{-s}\mathrm{d}s$ and
$\gamma(r,x) = \int_0^x s^{r-1}e^{-s}\mathrm{d}s$ are the upper and lower incomplete gamma function, respectively; $I_x(a,b) = B(x;a,b)/B(1;a,b)$ is the regularized incomplete beta function with $B(x;a,b) = \int_0^x s^{a-1}(1-s)^{b-1}\mathrm{d}s$. For more details see [62].

Name	Parameters	Support	pmf / pdf	cdf	Mean	Median	Mode	Variance	Skewness	Kurtosis	mgf	cf
Poisson $\pi(\alpha)$	$\alpha > 0 \in \mathbb{R}$	$k \in \mathbb{N}^0$	$\frac{\alpha^k}{k!}e^{-\alpha}$	$\frac{\Gamma(\lfloor k+1\rfloor,\alpha)}{\lfloor k\rfloor!}$	α	$\approx \lfloor \alpha+\frac{1}{3}-\frac{0.02}{\alpha}\rfloor$	$\lceil\alpha\rceil-1$	α	$\frac{1}{\sqrt{\alpha}}$	$\frac{1}{\alpha}$	$\exp\big(\alpha(e^s-1)\big)$	$\exp\big(\alpha(e^{\mathbb{i}s}-1)\big)$
Binomial $B(n,p)$	$n \in \mathbb{N}$ $p \in [0,1]$	$k \in \mathbb{N}^0$ $p \in [0,1]$	$\binom{n}{k}p^k(1-p)^{n-k}$	$I_{1-p} = (n-k, 1+k)$	np	$\lfloor np\rfloor$ or $\lceil np\rceil$	$\lfloor(n+1)p\rfloor$ or $\lfloor(n+1)p\rfloor-1$	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-6p(1-p)}{np(1-p)}$	$(1-p+p^s)^n$	$(1-p+p^{\mathbb{i}s})^n$
Normal $\varphi(\nu,\sigma)$	$\nu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}^+$	$x \in \mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\nu)^2}{2\sigma^2}}$	$\frac{1}{2}\left(1+\mathrm{erf}\left(\frac{x-\nu}{\sqrt{2\sigma^2}}\right)\right)$	ν	ν	ν	σ^2	0	0	$\exp(\nu s+\frac{1}{2}\sigma^2 s^2)$	$\exp(\mathbb{i}\nu s-\frac{1}{2}\sigma^2 s^2)$
chi-square $\chi^2(k)$	$k \in \mathbb{N}$	$x \in [0,\infty[$	$\frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}$	$\frac{\gamma(\frac{k}{2},\frac{x}{2})}{\Gamma(\frac{k}{2})}$	k	$\approx k\left(1-\frac{2}{9k}\right)^3$	$\max\{k-2,0\}$	$2k$	$\sqrt{\frac{8}{k}}$	$\frac{12}{k}$	$(1-2s)^{-\frac{k}{2}}$ for $s<\frac{1}{2}$	$(1-2\mathbb{i}s)^{-\frac{k}{2}}$
Logistic	$a \in \mathbb{R}, b > 0$	$x \in \mathbb{R}$	$\frac{\mathrm{sech}^2\left((x-a)/2b\right)}{4b}$	$\frac{1}{1+\exp\left(-(x-a)/b\right)}$	a	a	a	$\pi^2b^2/3$	0	4.2	$\frac{\pi bs\,e^{as}}{\sin(\pi bs)}$	$\frac{\mathbb{i}\pi bs\,e^{as}}{\sin(\mathbb{i}\pi bs)}$
Laplace	$\nu \in \mathbb{R}$ $b > 0$	$x \in \mathbb{R}$	$\frac{1}{2b}e^{-\frac{\lvert x-\nu\rvert}{b}}$	$\begin{cases}\frac{1}{2}e^{-\frac{\nu-x}{b}}, & x<a\\ 1-\frac{1}{2}e^{-\frac{x-\nu}{b}}, & x\geq a\end{cases}$	ν	ν	ν	$2b^2$	0	3	$\frac{\exp(\nu s)}{1-b^2s^2}$ for $\lvert s\rvert<\frac{1}{b}$	$\frac{\exp(\mathbb{i}\nu s)}{1-b^2s^2}$
Uniform $\mathcal{U}(a,b)$	$a < b$ $a,b \in \mathbb{R}$	$x \in [a,b]$	$\begin{cases}\frac{1}{b-a}, & x\in[a,b]\\ 0 & \text{otherwise}\end{cases}$	$\begin{cases}0, & x<a\\ \frac{x-a}{b-a}, & x\in[a,b]\\ 1, & x\geq b\end{cases}$	$\frac{a+b}{2}$	$\frac{a+b}{2}$	$\tilde{m}\in[a,b]$	$\frac{(b-a)^2}{12}$	0	$-\frac{6}{5}$	$\frac{e^{bs}-e^{as}}{(b-a)s}$	$\frac{e^{\mathbb{i}bs}-e^{\mathbb{i}as}}{\mathbb{i}(b-a)s}$
Cauchy	$x_0 \in \mathbb{R}$ $\gamma \in \mathbb{R}^+$	$x \in \mathbb{R}$	$\frac{1}{\pi\gamma\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)}$	$\frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right)$	–	x_0	x_0	–	–	–	–	$\exp(\mathbb{i}x_0s-\gamma\lvert s\rvert)$

2.3 Most common probability distributions

Before entering a discussion of individual probability distributions we present an overview over the important characteristics of the most frequently used distributions in table 2.2. Poisson, binomial and normal distributions and transformations in the limits between them are discussed in this section. The central limit theorem and the law of large numbers are presented in a separate section following the normal distribution. We have listed also several less common but nevertheless frequently used probability distributions, which are of importance for special purposes. In the forthcoming chapters 3, 4, and 5 dealing with stochastic processes and applications we shall make use of them.

2.3.1 The Poisson distribution

The Poisson distribution, named after the French physicist and mathematician Siméon Denis Poisson, is a discrete probability distribution expressing the probability of occurrence of independent events within a given interval. An popular example is dealing with the arrivals of phone calls within a fixed time interval Δt. The expected number of occurring calls per time interval, α, is the only parameter of the distribution $\mathrm{Pois}(\alpha; k)$, which returns the probability that k calls are received during Δt. In physics and chemistry the

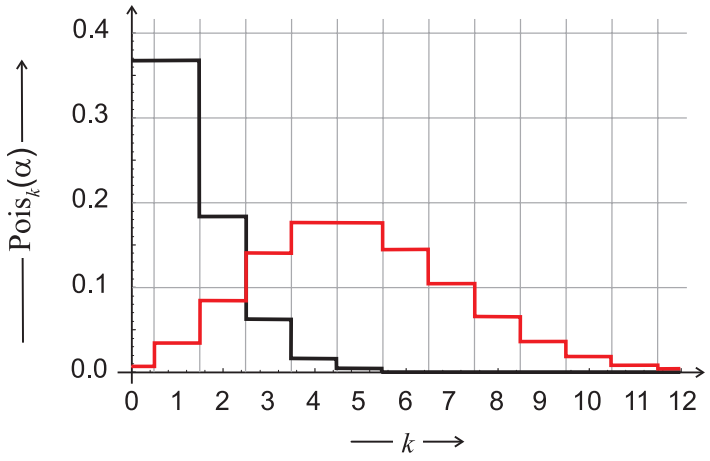

Fig. 2.6 The Poisson probability density. Two examples of Poisson distributions, $\pi_k(\alpha) = \alpha^k e^{-\alpha}/k!$, with $\alpha = 1$ (black) and $\alpha = 5$ (red) are shown. The distribution with the larger α has the mode shifted further to the right and a thicker tail.

Poisson process is the stochastic basis of first order processes, for example radioactive decay or irreversible first order reactions and the Poisson distribution is the probability distribution underlying the time course of particle numbers, $N(t) = \alpha(t)$. The events to be counted need not be on the time axis, the interval can also be defined as a given distance, area, or volume.

Despite its major importance in physics and biology the Poisson distribution with the probability mass function (pmf) $\mathrm{Pois}(\alpha; k)$, is a fairly simple mathematical object. As said it contains a single parameter only, the real valued positive number α:

$$P(\mathcal{X} = k) \;=\; \mathrm{Pois}(\alpha; k) \;=\; \pi_k(\alpha) \;=\; \frac{e^{-\alpha}}{k!}\,\alpha^k \;;\;\; k \in \mathbb{N}^0 \;. \tag{2.30}$$

As an exercise we leave to verify the following properties:[10]

$$\sum_{k=0}^{\infty} \pi_k = 1\,,\; \sum_{k=0}^{\infty} k\,\pi_k = \alpha \;\text{ and }\; \sum_{k=0}^{\infty} k^2\,\pi_k = \alpha + \alpha^2$$

Examples of Poisson distributions with two different parameter values, $\alpha = 1$ and 5, are shown in figure 2.6. The cumulative distribution function (cdf) is obtained by summation

$$P(\mathcal{X} \le k) \;=\; \exp(-\lambda) \sum_{j=0}^{\lfloor k \rfloor} \frac{\lambda^j}{j!} \;=\; \frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !} \;, \tag{2.31}$$

where $\Gamma(x, y)$ is the incomplete Gamma function.

By means of a Taylor expansion we can find the generating function of the Poisson distribution,

$$g(s) \;=\; e^{\alpha(s-1)} \;. \tag{2.32}$$

From the generating function we calculate easily

$$g'(s) \;=\; \alpha\, e^{\alpha(s-1)} \;\text{ and }\; g''(s) \;=\; \alpha^2\, e^{\alpha(s-1)} \;.$$

Expectation value and second moment follow straightforwardly from equation(2.25):

$$E(\mathcal{X}) \;=\; g'(1) \;=\; \alpha \;, \tag{2.32a}$$

$$E(\mathcal{X}^2) \;=\; g'(1) + g''(1) \;=\; \alpha + \alpha^2 \;,\;\text{ and} \tag{2.32b}$$

$$\sigma^2(\mathcal{X}) \;=\; \alpha \;. \tag{2.32c}$$

---

[10] In order to be able to solve the problems some basic infinite series should be recalled:
$e = \sum_{n=0}^{\infty} \frac{1}{n!}$, $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ for $|x| < \infty$, $e = \lim_{n\to\infty}(1 + \frac{1}{n})^n$,
$e^{-\alpha} = \lim_{n\to\infty}(1 - \frac{\alpha}{n})^n$.

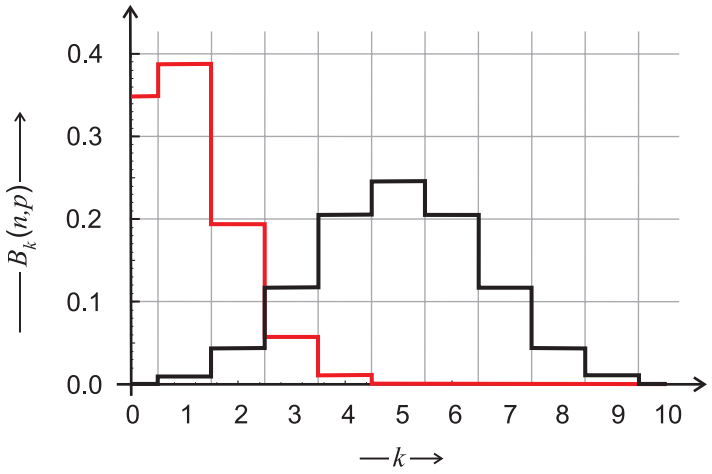

Fig. 2.7 The binomial probability density. Two examples of binomial distributions, $B_k(n,p) = \binom{n}{k} p^k (1-p)^{n-k}$, with $n = 10$, $p = 0.5$ (black) and $p = 0.1$ (red) are shown. The former distribution is symmetric with respect to the expectation value $E(B_k) = n/2$, and accordingly has zero skewness. The latter case is asymmetric with positive skewness (see figure 2.3).

Both, the expectation value and the variance are equal to the parameter α and hence, the standard deviation amounts to $\sigma(\mathcal{X}) = \sqrt{\alpha}$. This remarkable property of the Poisson distribution is not limited to the second moment. The *factorial moments*, $\langle \mathcal{X}^r \rangle_f$, fulfil the equation

$$\langle \mathcal{X}^r \rangle_f \;=\; E\Big(\mathcal{X}(\mathcal{X}-1)\ldots(\mathcal{X}-r+1)\Big) \;=\; \alpha^r \;, \tag{2.32d}$$

which is easily verified by direct calculation.

2.3.2 The binomial distribution

The binomial distribution, $B(n,p)$, characterizes the cumulative result of independent trials with two-valued outcomes, for example, yes-no decisions or successive coin tosses as we discussed in sections 1.2 and 1.5:

$$\mathcal{S}_n \;=\; \sum_{i=1}^{n} \mathcal{X}_i \,, i \in \mathbb{N}_{>0};\; n \in \mathbb{N}_{>0} \;. \tag{1.22'}$$

In general, we assume that head is obtained with probability p and tail with probability $q = 1 - p$. The $\mathcal{X}_i$'s are commonly called Bernoulli random vari-

ables named after the Swiss mathematician Jakob Bernoulli, and the sequence of events is named Bernoulli process after him (section 3.2.1.1). The corresponding random variable is said to have a Bernoulli or binomial distribution::

$$P(\mathcal{S}_n = k) \;=\; B_k(n,p) \;=\; \binom{n}{k} p^k\, q^{n-k}\,, \qquad q = 1-p \text{ and } k \in \mathbb{N};\, k \le n\ . \tag{2.33}$$

Two examples are shown in figure 2.7. The distribution with $p = 0.5$ is symmetric with respect to $k = n/2$.

The generating function for the single trial is $g(s) = q + ps$. Since we have n independent trials the complete generating function is

$$g(s) \;=\; (q \,+\, ps)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} q^{n-k}\, p^k\, s^k\ . \tag{2.34}$$

From the derivatives of the generating function,

$$g'(s) \;=\; n\,p\,(q+ps)^{n-1} \;\text{ and }\; g''(s) \;=\; n(n-1)\,p^2\,(q+ps)^{n-2}\ ,$$

we compute readily expectation value and variance:

$$E(\mathcal{S}_n) \;=\; g'(1) \;=\; n\,p\ , \tag{2.34a}$$

$$E(\mathcal{S}_n^2) \;=\; g'(1) \,+\, g''(1) \;=\; np \,+\, n^2p^2 \,-\, np^2 \;=\; n\,p\,q \,+\, n^2\,p^2\ , \tag{2.34b}$$

$$\sigma^2(\mathcal{S}_n) \;=\; n\,p\,q\ , \;\text{ and} \tag{2.34c}$$

$$\sigma(\mathcal{S}_n) \;=\; \sqrt{npq}\ . \tag{2.34d}$$

For $p = 1/2$, the case of the unbiased coin, we are dealing with the *symmetric binomial distribution* with $E(\mathcal{S}_n) = n/2$, $\sigma^2(\mathcal{S}_n) = n/4$, and $\sigma(\mathcal{S}_n) = \sqrt{n}/2$. We note that the expectation value is proportional to the number of trials, n, and the standard deviation is proportional to its square root, $\sqrt{n}$.

Relation between binomial and Poisson distribution. The binomial distribution $B(n,p)$ can be transformed into a Poisson distribution $\pi(\alpha)$ in the limit $n \to \infty$. In order to show this we start from

$$B_k(n,p) \;=\; \binom{n}{k} p^k\,(1-p)^{n-k}\,,\ 0 \le k \le n\ (k \in \mathbb{N}^0, k \le n)\ .$$

The symmetry parameter p is assumed to vary with n, $p(n) = \alpha/n$ for $n \in \mathbb{N}{>}\,0$, and thus we have

$$B_k\left(n, \frac{\alpha}{n}\right) \;=\; \binom{n}{k} \left(\frac{\alpha}{n}\right)^k \left(1 - \frac{\alpha}{n}\right)^{n-k}\ ,\ (k \in \mathbb{N}^0, k \le n)\ .$$

We let n go to infinity for fixed k and start with $B_0(n,p)$:

$$\lim_{n\to\infty} B_0\left(n,\frac{\alpha}{n}\right) \;=\; \lim_{n\to\infty}\left(1-\frac{\alpha}{n}\right)^n \;=\; e^{-\alpha}\;.$$

Now we compute the ratio of two consecutive terms, B_{k+1}/B_k:

$$\frac{B_{k+1}\left(n,\frac{\alpha}{n}\right)}{B_k\left(n,\frac{\alpha}{n}\right)} \;=\; \frac{n-k}{k+1}\cdot\left(\frac{\alpha}{n}\right)\cdot\left(1-\frac{\alpha}{n}\right)^{-1} \;=\; \frac{\alpha}{k+1}\cdot\left[\left(\frac{n-k}{n}\right)\cdot\left(1-\frac{\alpha}{n}\right)^{-1}\right]\;.$$

Both terms in the square brackets converge to one as $n\to\infty$, and hence we find:

$$\lim_{n\to\infty}\frac{B_{k+1}\left(n,\frac{\alpha}{n}\right)}{B_k\left(n,\frac{\alpha}{n}\right)} \;=\; \frac{\alpha}{k+1}\;.$$

From the two results we compute all terms starting from the limit value of B_0,

$$\begin{aligned}
\lim_{n\to\infty} B_0 &= \exp(-\alpha) \;\text{ and find}\\
\lim_{n\to\infty} B_1 &= \alpha\exp(-\alpha)\;,\\
\lim_{n\to\infty} B_2 &= \alpha^2\exp(-\alpha)/2!\;,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\lim_{n\to\infty} B_k &= \alpha^k\exp(-\alpha)/k!\;.
\end{aligned}$$

Accordingly we have verified Poisson's limit law:

$$\lim_{n\to\infty} B_k\left(n,\frac{\alpha}{n}\right) \;=\; \pi_k(\alpha)\,,\;\; k\in\mathbb{N}\;. \tag{2.35}$$

It is worth keeping in mind that the limit was performed in a peculiar way since the symmetry parameter $p(n)=\alpha/n$ was shrinking with increasing n and as a matter of fact vanished in the limit of $n\to\infty$.

2.3.3 The normal distribution

The normal or Gaussian distribution is of central importance in probability theory because many distributions converge to it in the limit of large numbers since the central limit theorem (CLT) states that under mild conditions the sum of a large number of random variables is approximately normal distributed (section 2.3.6). The normal distribution is a *stable distribution* (section 3.2.4) and this fact is not unrelated to the central limit theorem.

The normal distribution is basic for the estimate of statistical errors and thus we shall discuss it in some detail. Accordingly, the normal distribution

is extremely popular in statistics and quite often 'overapplied'. Many empirical values are not symmetrically distributed but skewed to the right but nevertheless they are often analyzed by means of normal distributions. The log-normal distribution [179] or the Pareto distribution, for example, might do better in such cases. Statistics based on normal distribution is not robust in the presence of outliers where a description by more heavy-tailed distributions like Student's t-distribution is superior. Whether or not the tails have more weight in the distribution can be easily checked by means of the excess kurtosis: Student's distribution has an excess kurtosis of

$$\gamma_2 \;=\; \begin{cases} \frac{6}{\nu-4} & \text{for } \nu > 4\ , \\ \infty & \text{for } 2 < \nu \leq 4\ , \quad \text{and} \\ \text{undefined} & \text{otherwise}\,, \end{cases}$$

which is always positive, whereas the excess kurtosis of the normal distribution is zero.

The normal distribution has also certain advantageous technical features. It is the only absolutely continuous distribution, which has only zero cumulants except the first two corresponding to expectation value and variance, which have the straightforward meaining of the position and the width of the distribution. For given variance the normal distribution has the largest informational entropy of all distributions on $\Omega = \mathbb{R}$ (section 2.1.3). As a matter of fact, the mean μ does not enter the expression for the entropy of the normal distribution (table 2.1),

$$H(\sigma) \;=\; \frac{1}{2}\left(1 \,+\, \log\left(2\pi\sigma^2\right)\right)\ , \tag{2.21'}$$

or in other words, shifting the normal distribution along the x-axis does not change the entropy of the distribution.

The density of the normal distribution[11] is

$$f_{\mathcal{N}}(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{with} \quad \int_{-\infty}^{+\infty} f(x)\,dx \;=\; 1\ , \tag{2.36}$$

and the corresponding random variable $\mathcal{X}$ has the moments $E(\mathcal{X}) = \mu$, $\sigma^2(\mathcal{X}) = \sigma^2$, and $\sigma(\mathcal{X}) = \sigma$. For many purposes it is convenient to use the normal density in centered and normalized form ($\sigma^2 = 1$), which is often called Gaussian bell-shaped curve:

$$f_{\mathcal{N}}(x;0,1) \;=\; \varphi(x) \;=\; \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \quad \text{with} \quad \int_{-\infty}^{+\infty} \varphi(x)\,dx \;=\; 1\ , \tag{2.36'}$$

[11] The notations applied here for the normal distribution are: $\mathcal{N}(\mu,\sigma)$ in general, and $F_{\mathcal{N}}(x;\mu,\sigma)$ for the cumulative distribution or $f_{\mathcal{N}}(x;\mu,\sigma)$ for the density. Commonly, the parameters, (μ,σ) are omitted when no misinterpretation is possible.

In this form we have $E(\tilde{\mathcal{X}}) = 0$, $\sigma^2(\tilde{\mathcal{X}}) = 1$, and $\sigma(\tilde{\mathcal{X}}) = 1$.

Integration of the density yields the distribution function

$$P(\mathcal{X} \leq x) \;=\; F(x) \;=\; \Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du \;. \tag{2.37}$$

The function $F_{\mathcal{N}}(x)$ is not available in analytical form, but it can be easily formulated in terms of the error function, $\mathrm{erf}(x)$. This function as well as its complement, $\mathrm{erfc}(x)$, defined by

$$\mathrm{erf}(x) \;=\; \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\, dt \quad \text{and} \quad \mathrm{erfc}(x) \;=\; \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2}\, dt \;,$$

are available in tables and in standard mathematical packages.[12] Examples of the normal density $f_{\mathcal{N}}(x)$ and the integrated distribution $F_{\mathcal{N}}(x)$ with different values of the standard deviation σ were shown in figure 1.19. The normal distribution is also used in statistics to define confidence intervals: 68.2 % of the data points lie within an interval $\mu \pm \sigma$, 95.4 % within an interval $\mu \pm 2\sigma$, and eventually 99,7 % with an interval $\mu \pm 3\sigma$.

A Poisson density with sufficiently large values of α resembles a normal density (figure 2.6) and it can be shown indeed that the two curves become more and more similar with increasing α:

$$\pi_k(\alpha) \;=\; \frac{\alpha^k}{k!}\, e^{-\alpha} \;\approx\; \frac{1}{\sqrt{2\pi\,\alpha}}\; \exp\!\left(-\frac{(k-\alpha)^2}{2\alpha}\right) \;. \tag{2.38}$$

This fact is an example of the central limit theorem presented and analyzed in section 2.3.6.

The normal density function $f_{\mathcal{N}}(x)$ has, among other remarkable properties, derivatives of all orders. Each derivative can be written as product of $f_{\mathcal{N}}(x)$ by a polynomial, of the order of the derivative, known as Hermite polynomial. The function $f_{\mathcal{N}}(x)$ decreases to zero very rapidly as $|x| \to \infty$. The existence of all derivatives makes the *bell-shaped* Gaussian curve $x \to f(x)$ particularly smooth, and the moment generating function of the normal distribution is especially attractive (see subsection 2.2.2). $M(s)$ can be obtained directly by integration:

[12] We remark that $\mathrm{erf}(x)$ and $\mathrm{erfc}(x)$ are not normalized in the same way as the normal density: $\mathrm{erf}(x) + \mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_0^\infty \exp(-t^2)dt = 1$, but $\int_0^\infty \varphi(x)dx = \frac{1}{2} \int_{-\infty}^{+\infty} \varphi(x)dx = \frac{1}{2}$.

$$\begin{aligned} M(s) &= \int_{-\infty}^{+\infty} e^{x\,s}\, f(x)\,\mathrm{d}x \;=\; \int_{-\infty}^{+\infty} \exp\left(x\,s \,-\, \frac{x^2}{2}\right)\,\mathrm{d}x \;= \\ &= \int_{-\infty}^{+\infty} e^{(\frac{s^2}{2} - \frac{(x-s)^2}{2})}\,\mathrm{d}x \;=\; e^{s^2/2} \int_{-\infty}^{+\infty} f(x-s)\,\mathrm{d}x \;= \\ &= e^{s^2/2}\;. \end{aligned} \tag{2.39}$$

All raw moments of the normal distribution are defined by the integrals

$$\hat{\mu}_n \;=\; \int_{-\infty}^{+\infty} x^n\, f(x)\, dx\;. \tag{2.40}$$

They can be obtained, for example, by successive differentiation of $M(s)$ with respect to s (subsection 2.2.2). In order to obtain the moments more efficiently we expand the first and the last expression in equation (2.39) in a power series of s,

$$\begin{aligned} &\int_{-\infty}^{+\infty} \left(1 \,+\, x\,s \,+\, \frac{(x\,s)^2}{2!} \,+\ldots+\, \frac{(x\,s)^n}{n!} \,+\ldots\right) f(x)\,\mathrm{d}x \;= \\ &= \quad 1 \,+\, \frac{s^2}{2} \,+\, \frac{1}{2!}\left(\frac{s^2}{2}\right)^2 \,+\ldots+\, \frac{1}{n!}\left(s^2 2\right)^n \,+\ldots\;, \end{aligned}$$

or express it in terms of the moments $\hat{\mu}_n$,

$$\sum_{n=0}^{\infty} \frac{\hat{\mu}_n}{n!}\, s^n \;=\; \sum_{n=0}^{\infty} \frac{1}{2^n\, n!}\, s^{2n}\;,$$

from which we compute the moments of $\varphi(x)$ by putting the coefficients of the powers of s equal on both sides of the expansion and find for $n \geq 1$:[13]

$$\hat{\mu}_{2n-1} \;=\; 0 \;\text{ and }\; \hat{\mu}_{2n} \;=\; \frac{(2n)!}{2^n\, n!}\;. \tag{2.41}$$

All odd moments vanish because of symmetry. In case of the fourth moment, kurtosis, a kind of standardization is common, which assigns zero excess kurtosis, $\gamma_2 = 0$ to the normal distribution. In other words, excess kurtosis monitors peak shape with respect to the normal distribution: Positive excess kurtosis implies peaks that are sharper than the normal density, negative excess kurtosis peaks that are broader than the normal density (figure 2.3).

---

[13] The definite integrals are:

$$\int_{-\infty}^{+\infty} x^n \exp(-x^2)\mathrm{d}x = \begin{cases} \sqrt{\pi} & n = 0 \\ 0 & n \geq 1;\ \text{odd} \\ \frac{(n-1)!!}{2^{n/2}}\sqrt{\pi} & n \geq 2;\ \text{even} \end{cases},$$

where $(n-1)!! = 1 \cdot 3 \cdot \ldots \cdot (n-1)$.

As already said all cumulants (2.16) of the normal distribution except $\kappa_1 = \mu$ and $\kappa_2 = sigma^2$ are zero, since the moment generating function of the general normal distribution with mean μ and variance σ^2 is of the form

$$M_{\mathcal{N}}(s) \;=\; \exp\Big(\mu s \,+\, \frac{1}{2}\,\sigma^2\,s^2\Big)\;. \tag{2.42}$$

The expression for the standardized Gaussian distribution is the special case with $\mu = 0$ and $\sigma^2 = 1$. Eventually, we list also the characteristic function of the general normal distribution

$$\phi_{\mathcal{N}}(s) \;=\; \exp\Big(\dot{\imath}\mu s \,-\, \frac{1}{2}\,\sigma^2\,s^2\Big)\;, \tag{2.43}$$

which will be used, for example, in the derivation of the central limit theorem (section 2.3.6).

2.3.4 Multivariate normal distributions

In applications to problems in science it is often necessary to consider probability distributions in multiple dimensions. Then, a random vector, $\vec{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ with the joint probability distribution

$$P(\mathcal{X}_1 = x_1, \ldots, \mathcal{X}_n = x_n) \;=\; p(x_1, \ldots, x_n) \;=\; p(\mathbf{x})\;.$$

replaces the random variable $\mathcal{X}$. This multivariate normal probability density can be written as

$$f(\mathbf{x}) \;=\; \frac{1}{\sqrt{(2\pi)^n\,|\boldsymbol{\Sigma}|}}\,\exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{t}}\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\Big)\;.$$

The vector $\boldsymbol{\mu}$ consists of the (raw) first moments along the different coordinates, $\boldsymbol{\mu}(\mu_1, \ldots, \mu_n)$ and the variance-covariance matrix $\boldsymbol{\Sigma}$ contains the n variances in the diagonal and the covariances are combined as off-diagonal elements

$$\boldsymbol{\Sigma} \;=\; \begin{pmatrix} \sigma^2(\mathcal{X}_1) & \mathrm{cov}(\mathcal{X}_1,\mathcal{X}_2) & \ldots & \mathrm{cov}(\mathcal{X}_1,\mathcal{X}_n) \\ \mathrm{cov}(\mathcal{X}_2,\mathcal{X}_1) & \sigma^2(\mathcal{X}_2) & \ldots & \mathrm{cov}(\mathcal{X}_2,\mathcal{X}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathcal{X}_n,\mathcal{X}_1) & \mathrm{cov}(\mathcal{X}_n,\mathcal{X}_2) & \ldots & \sigma^2(\mathcal{X}_n) \end{pmatrix} \;=\; \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \ldots & \sigma_{nn} \end{pmatrix}$$

which is symmetric, $\mathrm{cov}(\mathcal{X}_i, \mathcal{X}_j) = \mathrm{cov}(\mathcal{X}_j, \mathcal{X}_i) = \sigma_{ij}$, by the definition of covariances.

Mean and variance are given by $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$ and the variance-covariance matrix Σ, the moment generating function expressed in the dummy vector variable

$\mathbf{s} = (s_1, \ldots, s_n)$ is of the form

$$M(\mathbf{s}) \;=\; \exp\left(\boldsymbol{\mu}^{\mathrm{t}}\mathbf{s}\right) \cdot \exp\left(\frac{1}{2}\,\mathbf{s}^{\mathrm{t}}\,\Sigma\mathbf{s}\right) \;,$$

and, finally, the characteristic function is given by

$$\phi(\mathbf{s}) \;=\; \exp\left(\dot{\imath}\,\boldsymbol{\mu}^{\mathrm{t}}\mathbf{s}\right) \cdot \exp\left(-\frac{1}{2}\,\mathbf{s}^{\mathrm{t}}\,\Sigma\mathbf{s}\right)$$

Without showing the details we remark that this particulary simple characteristic function implies that all moments higher than order two can be expressed in terms of first and second moments, in particular expectation values, variances, and covariances. To give an example that we shall require later in subsection 3.4.2, the fourth order moments can be derived from

$$\begin{aligned}
E(\mathcal{X}_i^4) &= 3\,\sigma_{ii}^2 \;, \\
E(\mathcal{X}_i^3\mathcal{X}_j) &= 3\,\sigma_{ii}\sigma_{ij} \;, \\
E(\mathcal{X}_i^2\mathcal{X}_j^2) &= \sigma_{ii}\sigma_{jj} + 2\,\sigma_{ij}^2 \;, \\
E(\mathcal{X}_i^2\mathcal{X}_j\mathcal{X}_k) &= \sigma_{ii}\sigma_{jk} + 2\,\sigma_{ij}\sigma_{ik} \quad \text{and} \\
E(\mathcal{X}_i\mathcal{X}_j\mathcal{X}_k\mathcal{X}_l) &= \sigma_{ij}\sigma_{kl} + \sigma_{li}\sigma_{jk} + \sigma_{ik}\sigma_{jl} \;,
\end{aligned} \tag{2.44}$$

with $i, j, k, l \in \{1, 2, 3, 4\}$.

The entropy of the multivariate normal distribution is readily calculated and appears as a straightforward extension of equation (2.21) to higher dimensions:

$$\begin{aligned}
H(f) &= -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(\mathbf{x})\,\ln f(\mathbf{x})\,\mathrm{d}\mathbf{x} \;= \\
&= \frac{1}{2}\Big(n + \ln\big((2\pi)^n|\boldsymbol{\Sigma}|\big)\Big) \;,
\end{aligned} \tag{2.45}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of the variance-covariance matrix.

The multivariate normal distribution presents and excellent example for discussing the difference between uncorrelatedness and independence. Two random variables are independent if

$$f_{\mathcal{X}\mathcal{Y}}(x, y) \;=\; f_{\mathcal{X}}(x) \cdot f_{\mathcal{Y}}(y) \;\;\forall\, x, y \;,$$

whereas uncorrelatedness of two random variables requires

$$\begin{aligned}
\sigma_{\mathcal{X}\mathcal{Y}} \;=\; \mathrm{cov}(\mathcal{X}, \mathcal{Y}) \;=\; 0 \;&=\; E(\mathcal{X}\mathcal{Y}) \;-\; E(\mathcal{X})E(\mathcal{Y}) \quad \text{or} \\
E(\mathcal{X}\mathcal{Y}) \;&=\; E(\mathcal{X})E(\mathcal{Y}) \;.
\end{aligned}$$

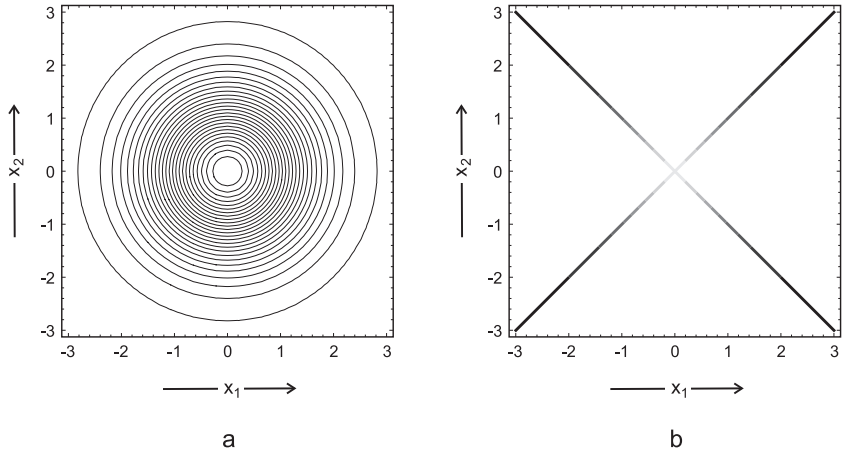

Fig. 2.8 Uncorrelated but not independent normal distributions. The figure compares two different joint densities, which have identical marginal densities. The contour plot on the l.h.s. (**a**) shows the joint distribution $f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)}$, the contour lines are circles equidistant in f and plotted for $f = 0.03, 0.09, \ldots, 0.153$. The marginal distributions of this joint distribution are standard normal distributions in x_1 or x_2. The density in **b** is derived from one random variable $\mathcal{X}_1$ with standard normal density $f(x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2}$ and a second random variable that is modulated by a perfect coin flip: $\mathcal{X}_2 = \mathcal{X}_1 \cdot \mathcal{W}$ with $\mathcal{W} = \pm 1$. The two variables $\mathcal{X}_1$ and $\mathcal{X}_2$ are uncorrelated but not independent.

The covariance between two independent random variables vanishes:

$$\begin{aligned} E(\mathcal{X}\mathcal{Y}) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy\, f_{\mathcal{X},\mathcal{Y}}(x,y)\,\mathrm{d}x\mathrm{d}y \;= \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy\, f_{\mathcal{X}}(x) f_{\mathcal{Y}}(y)\,\mathrm{d}x\mathrm{d}y \;= \\ &= \int_{-\infty}^{+\infty} x\, f_{\mathcal{X}}(x)\mathrm{d}x \int_{-\infty}^{+\infty} y\, f_{\mathcal{Y}}(y)\mathrm{d}y \;= E(\mathcal{X})E(\mathcal{Y})\;. \qquad \square \end{aligned}$$

We remark that the proof made nowhere use of the fact that the variables are normally distributed and the statement *independent variables are uncorrelated* holds in full generality. The inverse, however, is not true as has been shown by means of specific examples [209]: Uncorrelated random variables $\mathcal{X}_1$ and $\mathcal{X}_2$, which both have the same (marginal) normal distribution, need not be independent. The construction of such a contradicting example starts from a two dimensional random vector $\vec{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2)^{\mathrm{t}}$, which obeys a bivariate normal distribution with mean $\boldsymbol{\mu} = (0,0)^{\mathrm{t}}$ and variance $\sigma_1^2 = \sigma_2^2 = 1$ and

covariance $\mathrm{cov}(\mathcal{X}_1, \mathcal{X}_2) = 0$

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1, x_2)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}\right) =$$
$$= \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2} = f(x_1) \cdot f(x_2)\,,$$

and the two random variables are independent. Next we introduce a modification in one of the two random variables: $\mathcal{X}_1$ remains unchanged and has the density $f(x_1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x_1^2)$, whereas the second random variable is modulated with an ideal coin flip, $\mathcal{W}$ with the density $f(w) = \frac{1}{2}\big(\delta(w+1) + \delta(w-1)\big)$. In other words, we have $\mathcal{X}_2 = \mathcal{W} \cdot \mathcal{X}_1 = \pm\mathcal{X}_1$ with equal weights for both signs, and accordingly the density function is $f(x_2) = \frac{1}{2}f(x_1) + \frac{1}{2}f(-x_1) = f(x_1)$, since the normal distribution with zero mean $E(\mathcal{X}_1) = 0$ is symmetric, $f(x_1) = f(-x_1)$. Equality of the two distribution functions with the same normal distribution can also be derived directly:

$$\begin{aligned} P(\mathcal{X}_2 \le x) &= E\big(P(\mathcal{X}_2 \le x|\mathcal{W})\big) = \\ &= P(\mathcal{X}_1 \le x)P(\mathcal{W}=1) + P(-\mathcal{X}_1 \le x)P(\mathcal{W}=-1) = \\ &= \Phi(x)\frac{1}{2} + \Phi(x)\frac{1}{2} = \Phi(x) = P(\mathcal{X}_1 \le x)\,. \end{aligned}$$

The covariance of $\mathcal{X}_1$ and $\mathcal{X}_2$ is readily calculated,

$$\begin{aligned} \mathrm{cov}(\mathcal{X}_1\mathcal{X}_2) &= E(\mathcal{X}_1\mathcal{X}_2) - E(\mathcal{X}_1)\cdot E(\mathcal{X}_2) = E(\mathcal{X}_1\mathcal{X}_2) - 0 = \\ &= E\Big(E(\mathcal{X}_1\mathcal{X}_2)|\mathcal{W}\Big) = E(\mathcal{X}_1^2)P(\mathcal{W}=1) + E(-\mathcal{X}_1^2)P(\mathcal{W}=-1) = \\ &= 1\,\frac{1}{2} + (-1)\,\frac{1}{2} = 0\,, \end{aligned}$$

and hence $\mathcal{X}_1$ and $\mathcal{X}_2$ are uncorrelated. The two random variables, however, are not independent because

$$\begin{aligned} p(x_1, x_2) &= P(\mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2) = \\ &= \frac{1}{2}P(\mathcal{X}_1 = x_1, \mathcal{X}_2 = x_1) + \frac{1}{2}P(\mathcal{X}_1 = x_1, \mathcal{X}_2 = -x_1)\Big) = \\ &= \frac{1}{2}p(x_1) + \frac{1}{2}p(x_1) = p(x_1)\,, \\ f(x_1, x_2) &= f(x_1) \neq f(x_1)\cdot f(x_2)\,, \end{aligned}$$

since $f(x_1) = f(x_2)$. Lack of independence follows also simply from $|\mathcal{X}_1| = |\mathcal{X}_2|$. The example is illustrated in figure 2.8: The fact that marginal distributions are identical does not imply that the joint distribution is also the

same! The statement about independence, however, can be made stronger and then it turns out to be true: *"If random variables have a multivariate normal distribution and are pairwise uncorrelated, then the random variables are always independent."* [209].

The marginal distributions of a multivariate normal distribution are obtained straightforwardly by simply dropping the marginalized variables. If $\vec{\mathcal{X}} = (\mathcal{X}_i, \mathcal{X}_j, \mathcal{X}_k)$ is a multivariate, normally distributed variable with the mean vector $\boldsymbol{\mu} = (\mu_i, \mu_j, \mu_k)$ and variance-covariance matrix $\boldsymbol{\Sigma}$, then after elimination of $\mathcal{X}_j$ the marginal joint distribution of the vector $\tilde{\mathcal{X}} = (\mathcal{X}_i, \mathcal{X}_k)$ is multivariate normal with the mean vector $\tilde{\boldsymbol{\mu}} = (\mu_i, \mu_k)$ and the variance-covariance matrix

$$\tilde{\boldsymbol{\Sigma}} \;=\; \begin{pmatrix} \Sigma_{ii} & \Sigma_{ik} \\ \Sigma_{ki} & \Sigma_{kk} \end{pmatrix} \;=\; \begin{pmatrix} \sigma^2(\mathcal{X}_i) & \mathrm{cov}(\mathcal{X}_i, \mathcal{X}_k) \\ \mathrm{cov}(\mathcal{X}_k, \mathcal{X}_i) & \sigma^2(\mathcal{X}_k) \end{pmatrix} \;.$$

It is worth noticing that non-normal bivariate distributions have been constructed, which have normal marginal distributions [170].

2.3.5 From binomial to normal distributions

The expression *normal distribution* actually originated from the fact that many distributions can be transformed in a natural way for large numbers n to yield the distribution $F_{\mathcal{N}}(x)$. Here we shall derive it from the binomial distribution

$$B_k(n,p) \;=\; \binom{n}{k} p^k \,(1-p)^{n-k} \;,\; 0 \le k \le n \;,$$

through extrapolation to large values of n at constant p.[14] The transformation from the binomial distribution to the normal distribution is properly done in two steps (see also [34, pp.210-217]): (i) standardization and (ii) taking the limit $n \to \infty$.

At first we make the binomial distribution comparable to the standard normal density, $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$, by shifting the maximum towards $x = 0$ and adjusting the width (figures 2.9 and 2.10). For $0 < p < 1$ and $q = 1 - p$ the discrete variable k is replaced by a new variable ξ:[15]

$$\xi \;=\; \frac{k \,-\, np}{\sqrt{npq}} \;;\;\; 0 \le k \le n \;.$$

[14] This is different from an extrapolation performed in a previous section2.3.2 because the limit $\lim_{n\to\infty} B_k(n, \alpha/n) = \pi_k(\alpha)$ leading to the Poisson distribution was performed for vanishing $p = \alpha/n$.

[15] The new variable ξ depends on k and n, but for short we dispense from subscripts.

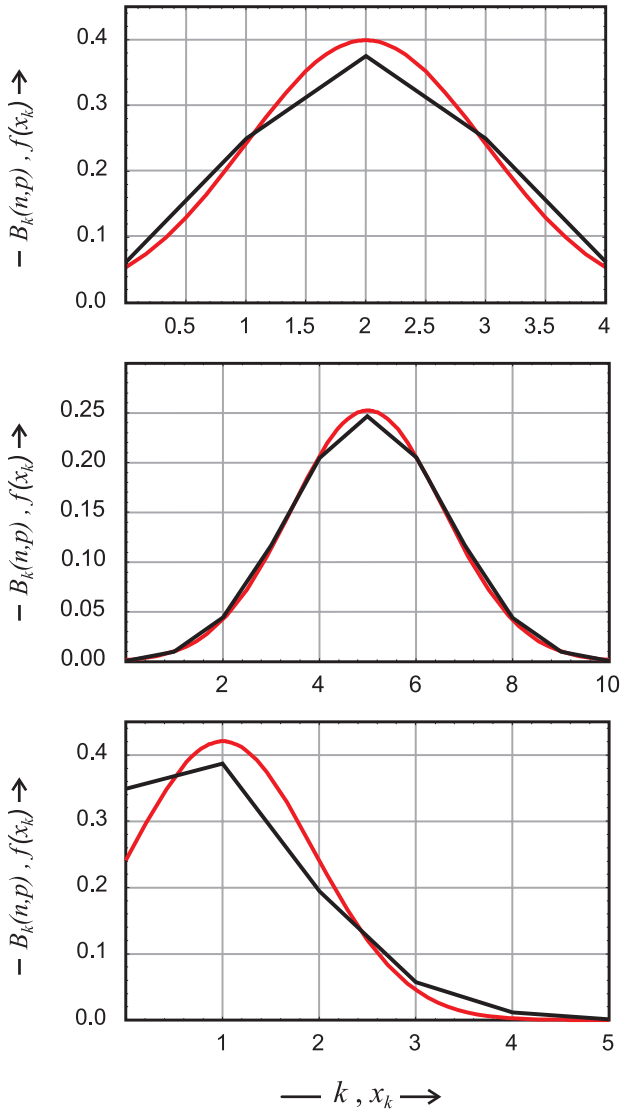

Fig. 2.9 A fit of the normal distribution to the binomial distribution. The curves represent normal densities (red), which were fit to the points of the binomial distribution (black). The three examples. Parameter choice for the binomial distribution: $(n = 4, p = 0.5)$, $(n = 10, p = 0.5)$, and $(n = 10, p = 0.1)$, for the upper, middle, and lower plot, respectively.

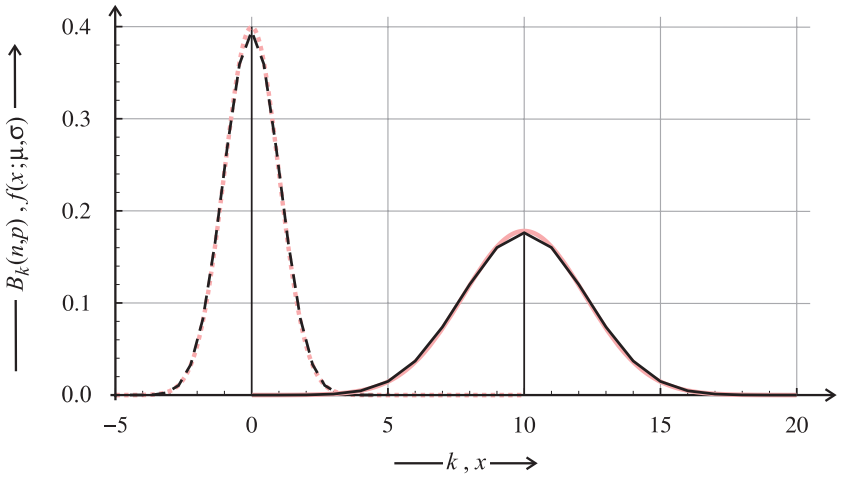

Fig. 2.10 Standardization of the binomial distribution. The figure shows a symmetric binomial distribution $B(20, \frac{1}{2})$, which is centered around $\mu = 10$ (black, full line). The transformation yields a binomial distribution centered around the origin with unit variance: $\sigma = \sigma^2 = 1$ (black, broken line). The pink continuous curve is a normal density $f_{\mathcal{N}}(x) = \exp\big(-(x-\mu)^2/(2\sigma^2)\big)/\sqrt{2\pi\sigma^2}$ with the parameters $\mu = 10$ and $\sigma^2 = np(1-p) = 5$, and the broken pink line is a standardized normal density $\varphi(x)$ ($\mu = 0, \sigma^2 = 1$), respectively.

Instead of the variables $\mathcal{X}_k$ and $\mathcal{S}_k$ in equation (1.22') new random variables, $\mathcal{X}_k^*$ and $\mathcal{S}_n^* = \sum_{k=1}^n \mathcal{X}_k^*$ are introduced, which are centered around $x = 0$ and adjusted to the width of a standard Gaussian, $\varphi(x)$, by making use of the expectation value, $E(\mathcal{S}_n) = np$, and the standard deviation, $\sigma(\mathcal{S}_n) = \sqrt{npq}$, of the binomial distribution.

The theorem of de Moivre and Laplace states now that for k in a neighborhood of $k = np - |\xi| \leq c$ with c being an arbitrary fixed positive constant – the approximation

$$\binom{n}{k} p^k q^{n-k} \;\approx\; \frac{1}{\sqrt{2\pi npq}}\, e^{-\xi^2/2}\,;\; p+q=1,\, p>0,\, ,q>0 \qquad (2.46')$$

becomes exact in the sense that the ration of the l.h.s. to the r.h.s. converges to one as $n \to \infty$ [76, section VII.3]. The convergence is uniform with respect to k in the range specified above. In order to proof the convergence we transform the l.h.s. by making use of Stirling's formula, $n! \approx n^n e^{-n}\sqrt{2\pi n}$ as $n \to \infty$:

$$\binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} \approx \sqrt{\frac{n}{2\pi k(n-k)}} \left(\Big(\frac{k}{np}\Big)^{-k} \Big(\frac{n-k}{nq}\Big)^{-(n-k)} \right) .$$

Next we introduce the variable ξ and transform to the exponential function

$$\binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} \left(\left(1+\xi\sqrt{\frac{q}{np}}\right)^{-k} \left(1-\xi\sqrt{\frac{p}{nq}}\right)^{-(n-k)} \right) =$$

$$= \frac{1}{\sqrt{2\pi npq}} \; e^{\ln\left(\left(1+\xi\sqrt{\frac{q}{np}}\right)^{-k} \left(1-\xi\sqrt{\frac{p}{nq}}\right)^{-(n-k)} \right)} ,$$

and expansion of the logarithm yields

$$\ln\left(\left(1+\xi\sqrt{\frac{q}{np}}\right)^{-k} \left(1-\xi\sqrt{\frac{p}{nq}}\right)^{-(n-k)} \right) =$$

$$= -k \ln\left(1+\xi\sqrt{\frac{q}{np}}\right) - (n-k) \ln\left(1-\xi\sqrt{\frac{p}{nq}}\right) .$$

Making use of the expansion $\ln(1\pm\gamma) \approx \pm\gamma - \gamma^2/2 \pm \gamma^3/3 - \ldots$, and truncation after the second term, and inserting $k = np + \xi\sqrt{npq}$ and $n-k = nq - \xi\sqrt{npq}$ we find

$$\ln\left(\left(1+\xi\sqrt{\frac{q}{np}}\right)^{-k} \left(1-\xi\sqrt{\frac{p}{nq}}\right)^{-(n-k)} \right) =$$

$$= -(np+\xi\sqrt{npq}\,) \left(\xi\sqrt{\frac{q}{np}} - \xi^2\frac{q}{np} + \ldots \right) -$$

$$- (nq-\xi\sqrt{npq}\,) \left(-\xi\sqrt{\frac{p}{nq}} - \xi^2\frac{p}{nq} + \ldots \right) .$$

Evaluation of the expressions eventually yields

$$\ln\left(\left(1+\xi\sqrt{\frac{q}{np}}\right)^{-k} \left(1-\xi\sqrt{\frac{p}{nq}}\right)^{-(n-k)} \right) \approx -\frac{\xi^2}{2}$$

and thereby we have proved the conjecture (2.46'). □

A comparison of figures 2.9 and 2.10 shows that the convergence of the binomial distribution to the normal distribution is particularly effective in the symmetric case, $p = q = 0.5$. The difference is substantially larger for $p = 0.1$. A value of $n = 20$ is sufficient to make the difference hardly recognizable with the unaided eye. Figure 2.10 shows also the effect of standardization on the binomial distribution.

In the context of the central limit theorem (section 2.3.6) it is useful to formulate the theorem of de Moivre and Laplace in a slightly different way: The distribution of the standardized random variable $\mathcal{S}_n^*$ with a binomial distribution converges in the limit of large numbers n to the normal distribution $\varphi(x)$ on any finite constant interval $]\,a, b]$ with $a < b$:

$$\lim_{n\to\infty} P\left(\left(\frac{\mathcal{S}_n - np}{\sqrt{npq}}\right) \in]a,b]\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,\mathrm{d}x \ . \tag{2.46}$$

In the proof [34, p. 215-217] the definite integral $\int_a^b \varphi(x)\,\mathrm{d}x$ is partitioned into n small segments like in Riemannian integration: The segments still reflect the discrete distribution. In the limit $n \to \infty$ the partition becomes finer and eventually converges to the continuous function described by the integral. In the sense in section 1.8.1 we are dealing with convergence to a limit in distribution.

2.3.6 Central limit theorem

In addition to the transformation of the binomial distribution into the normal distribution analyzed in the previous section 2.3.5 we have already encountered two cases where probability distributions approached the normal distribution in the limit of large numbers n: (i) the distribution of scores for rolling n dice simultaneously (section 1.9.1) and (ii) the Poisson distribution (section 2.3.3). It is obvious to conjecture therefore that more general regularities concerning the role of the normal distribution in the limit of large n should exist. The Russian mathematician Aleksandr Lyapunov pioneered the formulation and derivation of such a generalization that got the name central limit theorem (CLT) [189, 190]. Research on CLT has been continued and practically completed by extensive studies during the entire twentieth century [3, 259].

The *central limit theorem* comes in various stronger and weaker forms. We mention here three of them:

(i) The so-called *classical central limit theorem* is commonly associated with the names of the Finnish mathematician Jarl Waldemar Lindeberg [181] and the French mathematician Paul Pierre Lévy [177], and is the most common version used in practice. In essence, the Lindeberg-Lévy central limit theorem provides the generalization of the de Moivre-Laplace theorem (2.46) that has been used in the previous section 2.3.5 to show the transition from the binomial to the normal distribution in the limit $n \to \infty$. This generalization proceeds from Bernoulli variables to *independent and identically distributed* (iid) random variables $\mathcal{X}_i$. The distribution is arbitrary, need not be specified and the only requirements are finite expectation value and variances: $E(\mathcal{X}_i) = \mu < \infty$ and $\mathrm{var}(\mathcal{X}_i) = \sigma^2 < \infty$. Again we consider the sum of n random variables, $\mathcal{S}_n = \sum_{i=1}^{n} \mathcal{X}_i$, standardize to yield $\mathcal{X}_i^*$ and $\mathcal{S}_n^*$, and instead of equation (2.46) we obtain

$$\lim_{n\to\infty} P\left(\frac{\mathcal{S}_n - n\,\mu}{\sqrt{n}\,\sigma} \in]a,b]\right) \;=\; \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,\mathrm{d}x. \tag{2.47}$$

For every segment $a < b$ the arbitrary initial distribution converges to the normal distribution in the limit $n \to \infty$. Although this is already an enormous extension of the validity in the limit of the normal distribution, the results can be made more general.

(ii) Lyapunov's earlier version of the central limit theorem [189, 190] requires only independent not necessarily identically distributes variables $\mathcal{X}_i$ with finite expectation values, μ_i, and variances, σ_i^2 provided a criterium called *Lyapunov condition* is fulfilled by the sum of variances $s_n^2 = \sum_{i=1}^{n} \sigma_1^2$,

$$\lim_{n\to\infty} \frac{1}{s^{2+\delta}} \sum_{i=1}^{n} E\left(|\mathcal{X}_i - \mu_i|^{2+\delta}\right) \;=\; 0\;, \tag{2.48}$$

then the sum $\sum_{i=1}^{n}(\mathcal{X}_i - \mu_i)/s_n$ converges in distribution in the limit $n \to \infty$ to the standard normal variable:

$$\frac{1}{s_n}\sum_{i=1}^{n}(\mathcal{X}_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1) \ . \tag{2.49}$$

Whether or not a given sequence of random variables fulfils the Lyapunov condition is commonly checked in practice by setting $\delta = 1$.

(iii) Lindeberg showed in 1922 [182] that a weaker condition than Lyapunov's condition is sufficient to guarantee the convergence in distribution to the standard normal distribution:

$$\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{i=1}^{n} E\Big((X_i - \mu_i)^2 \cdot \mathbf{1}_{|\mathcal{X}_i - \mu_i| > \epsilon s_n}\Big) \ = \ 0 \ , \tag{2.50}$$

where $\mathbf{1}_{|\mathcal{X}_i - \mu_i| > \epsilon s_n}$ is the indicator function (1.53) identifying the sample space

$$\{|\mathcal{X}_i - \mu_i| > \epsilon\, s_n\} \ := \ \{\omega \in \Omega : \ |\mathcal{X}_i(\omega) - \mu_i| > \epsilon\, s_n\} \ .$$

If a sequence of random variables satisfies Lyapunov's condition it satisfies also Lindeberg's condition but the converse does not hold in general. Lindeberg's condition is sufficient but not necessary in general, and the condition for necessity is

$$\max_{i=1,\ldots,n}\frac{\sigma_i^2}{s_n^2} \ \to \ 0 \ \text{ as } \ n \to \infty \ ,$$

or, in other words, the Lindeberg condition is fulfilled if and only if the central limit theorem holds.

The three versions of the central limit theorem are related to each other: Lindeberg's condition (iii) is the most general form and hence both the classical CLT (i) and the Lyapunov CLT (ii) can be derived as special cases from (iii). It is worth noticing, however, that (i) does not follow necessarily from (ii), because (i) requires a finite second moment whereas the condition for (ii) are finite moments of order $(2 + \delta)$.

In summary the central limit theorem for a sequence of independent random variables $\mathcal{S}_n = \sum_{i=1}^{n}\mathcal{X}_i$ with finite means, $E(\mathcal{X}_i) = \mu_i < \infty$, and variances, $\mathrm{var}(\mathcal{X}_i) = \sigma_i^2 < \infty$ states that the sum $\mathcal{S}_n$ converges in distribution to a standardized normal random variable $\mathcal{N}(0,1)$ without any further restriction on the distributions.

The literature on the central limit theorem is enormous and several proofs with many variants have been derived (see, for example, [33]). We shall present here only a short proof of the CLT in the from of equation (2.47) that demonstrates the usefulness of characteristic functions (section 2.2.3; [34, pp. 222-224]). We assume a sequence $\mathcal{S}_n = \sum_{i=1}^{n}\mathcal{X}_i$ of independent and identically distributed random variables $\mathcal{X}_i$ with finite means $E(\mathcal{X}_i) = \mu$ and variances $\mathrm{var}(\mathcal{X}_i) = \sigma^2$, where the nature of the distribution needs not to

be specified except the existence of finite mean and variance. The first step towards a proof of the central limit theorem is the transformation of variables shifting the maximum to the origin and adjusting the width of the distribution to $\mathrm{var}(\mathcal{S}) = 1$:

$$\mathcal{X}_j^* \;=\; \frac{\mathcal{X}_j - E(\mathcal{X}_j)}{\sigma(\mathcal{X}_j)} \;\text{ and }\; \mathcal{S}_n^* \;=\; \frac{\mathcal{S}_n - E(\mathcal{S}_n)}{\sigma(\mathcal{S}_n)} \;=\; \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathcal{X}_j^* \;. \qquad (2.51)$$

The values for the first and second moments of the sequence are: $E(\mathcal{S}_n) = n\mu$ and $\sigma(\mathcal{S}_n) = \sqrt{n}\,\sigma$. If $\mathcal{V}$ is the finite interval $]a, b]$ then the central limit theorem states that $F(\mathcal{V}) = F(b) - F(a)$ for any distribution function F and we can write the central limit theorem in compact form

$$\lim_{n\to\infty} F_n(\mathcal{V}) \;=\; F_{\mathcal{N}}(\mathcal{V}) \;, \qquad (2.52)$$

and in particular, for any interval $]a, b]$ with $a < b$ the limit

$$\lim_{n\to\infty} P\left(\frac{\mathcal{S}_n - n\,\mu}{\sqrt{n}\,\sigma} \in]a, b]\right) \;=\; \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,\mathrm{d}x. \qquad (2.47')$$

is fulfilled. The proof of the central limit theorem makes use of the characteristic function for the unit normal distribution with $\mu = 0$ and $\sigma^2 = 1$:

$$\phi_{\mathcal{N}}(s) = \;\exp(i\mu\, s \;-\; \frac{1}{2}\,\sigma^2\, s^2) \;=\; e^{-s^2/2} \;=\; \varphi(s) \;. \qquad (2.43')$$

We assume that for every s the characteristic function for $\mathcal{S}_n$ converges to the characteristic function $\phi(s)$,

$$\lim_{n\to\infty} \phi_n(s) \;=\; \varphi(s) \;=\; e^{-s^2/2} \;.$$

Since $\phi_n(s)$ are the characteristic functions associated with an arbitrary distribution function $F_n(x)$ follows for every x

$$\lim_{n\to\infty} F_n(x) \;=\; F_{\mathcal{N}}(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\mathrm{d}u \;., \qquad (2.53)$$

The de Moivre-Laplace theorem (section 2.3.5), for example, follows as a special case.

Characteristic functions $\phi(s)$ of random variables $\mathcal{X}$ with mean zero, $\mu = 0$, and variance one, $\sigma^2 = 1$, have the Taylor expansion

$$\phi(s) \;=\; 1 \;-\; \frac{s^2}{2}\Big(1 + \varepsilon(s)\Big) \;\text{ with }\; \lim_{s\to 0} \varepsilon(s) = 0$$

at $s = 0$ and truncation after the second term. In order to prove this equation we start from the full Taylor expansion up to the second term:

$$\phi(s) \;=\; \phi(0) \,+\, \phi'(0)\, s \,+\, \frac{\phi''(0)}{2}\, s^2 \left(1 + \varepsilon(s)\right) .$$

From $\phi(s) = E(e^{\dot{\imath} s\, \mathcal{X}})$ follows by differentiation

$$\phi'(s) \;=\; E\big(\dot{\imath}\mathcal{X}\, e^{\dot{\imath} s\, \mathcal{X}}\big) \;\text{ and }\; \phi''(s) \;=\; E\big(-\mathcal{X}^2\, e^{\dot{\imath} s\, \mathcal{X}}\big)$$

and hence $\phi'(0) = E(\dot{\imath}\mathcal{X}) = 0$ and $\phi''(0) = E(-\mathcal{X}^2) = -1$ yielding the equation given above.

Next we consider the characteristic function of $\mathcal{S}_n^*$:

$$E\Big(\exp(\dot{\imath} s\, \mathcal{S}_n^*)\Big) \;=\; E\bigg(\exp\Big(\dot{\imath} s\, \big(\textstyle\sum_{j=1}^{n} \mathcal{X}_j^*\big)/\sqrt{n}\Big)\bigg)$$

Since all random variables have the same distribution, the right hand side of the equation can be factorized and yields

$$E\Big(e^{\dot{\imath} s\, \left(\sum_{j=1}^{n} \mathcal{X}_j^*\right)/\sqrt{n}}\Big) \;=\; E\Big(e^{\dot{\imath} s\, \mathcal{X}_j^*/\sqrt{n}}\Big)^n \;=\; \phi\Big(\frac{s}{\sqrt{n}}\Big)^n ,$$

where $\phi(s)$ is the characteristic function of the random variable $\mathcal{X}_j$. Insertion into the expression for the Taylor series yields now

$$\phi\Big(\frac{s}{\sqrt{n}}\Big) \;=\; 1 \,-\, \frac{s^2}{2n}\bigg(1 + \varepsilon\Big(\frac{s}{\sqrt{n}}\Big)\bigg) .$$

Herein the number n is approaching infinity whereas s is fixed:

$$\lim_{n\to\infty} E\Big(e^{\dot{\imath} s\, \mathcal{S}_n^*}\Big) \;=\; \lim_{n\to\infty}\bigg(1 \,-\, \frac{s^2}{2n}\bigg(1 + \varepsilon\Big(\frac{s}{\sqrt{n}}\Big)\bigg)\bigg)^n \;=\; e^{-s^2/2} . \qquad (2.54)$$

For taking the limit in the last step of the derivation we recall the summation of infinite series,

$$\lim_{n\to\infty}\Big(1 - \frac{\alpha_n}{n}\Big)^n \;=\; e^{-\alpha} \;\text{ for }\; \lim_{n\to\infty} \alpha_n \,=\, \alpha\ , \qquad (2.55)$$

and remark that this is a stronger result than the convergence of the conventional exponential series, $\lim_{n\to\infty}(1 - \alpha/n)^n = e^{-\alpha}$. Thus, we have shown that the characteristic function of the normalized sum of random variables, $\mathcal{S}_n^*$, converges to the characteristic function of the standard normal distribution and therefore by equation (2.53) the distribution $F_n(x)$ converges to the unit normal distribution $F_{\mathcal{N}}(x)$ and the validity of (2.52) follows straightforwardly. □

Contrasting the rigorous mathematical derivation, simple practical applications used in *large sample theory* of statistics turn the central limit theorem encapsulated in equation (2.54) into a rough approximation

$$P(\sigma\sqrt{n}\,x_1 \,<\, \mathcal{S}_n - n\,\mu \,<\, \sigma\sqrt{n}\,x_2) \;\approx\; F_{\mathcal{N}}(x_2) \,-\, F_{\mathcal{N}}(x_1) \tag{2.56}$$

or for the spread around the sample mean μ by setting $x_1 = -x_2$

$$P(|\mathcal{S}_n - n\,\mu| < \sigma\sqrt{n}\,x) \;\approx\; 2F_{\mathcal{N}}(x) \,-\, 1\;. \tag{2.56'}$$

For practical purposes equation (2.56) has been used in *pre-computer time* together with extensive tabulations of the functions $F_{\mathcal{N}}(x)$ and $F_{\mathcal{N}}^{-1}(x)$, which are still found in statistics textbooks.

2.3.7 Law of large numbers

The *law of large numbers* is derived as a straightforward consequence of the central limit theorem (2.47) [34, pp.227-233]. For any fixed but arbitrary constant $c > 0$ we have

$$\lim_{n\to\infty} P\left(\left|\frac{\mathcal{S}_n}{n} \,-\, \mu\right| < c\right) \;=\; 1\;. \tag{2.57}$$

The constant c is fixed and therefore we can define a positive constant ℓ that fulfils $\ell < c\sqrt{n}/\sigma$ and for which

$$\left\{\left|\frac{\mathcal{S}_n \,-\, n\mu}{\sqrt{n}\sigma}\right| < \ell\right\} \;\text{ implies }\; \left\{\left|\frac{\mathcal{S}_n \,-\, n\mu}{n}\right| < c\right\}\;,$$

and hence,

$$P\left(\left|\frac{\mathcal{S}_n \,-\, n\mu}{\sqrt{n}\sigma}\right| < \ell\right) \;\leq P\left(\left|\frac{\mathcal{S}_n \,-\, n\mu}{n}\right| < c\right)\;,$$

provided n is sufficiently large. Now we choose a symmetric interval $a = -\ell$ and $b = +\ell$ for the integral and the l.h.s. of the inequality according to (2.47) converges to $\int_{-\ell}^{+\ell} \exp(-x^2/2)\mathrm{d}x/\sqrt{2\pi}$ in the limit $n \to \infty$. For any $\delta > 0$ we can choose ℓ so large that the value of the integral exceeds $1 - \delta$ and we get

$$P\left(\left|\frac{\mathcal{S}_n}{n} \,-\, \mu\right| < c\right) \;=\; 1 \,-\, \delta \tag{2.58}$$

for sufficiently large values of n and this proves that the law of large numbers (2.57) is a corollary of (2.47). □

Related to and a consequence of equation (2.57) is Chebyshev's inequality for random variables $\mathcal{X}$ that have a finite second moment, which is named after the Russian mathematician Pafnuty Lvovich Chebyshev :

$$P(|\mathcal{X}| \geq c) \;\leq\; \frac{E(\mathcal{X}^2)}{c^2} \tag{2.59}$$

and which is true for any constant $c > 0$. We dispense here from a proof that is found in [34, pp. 228-233].

By means of Chebyshev's inequality the law of large numbers (2.57) can extended to a sequence of independent random variables $\mathcal{X}_j$ with different expectation values and variances, $E(\mathcal{X}_j) = \mu^{(j)}$ and $\sigma^2(\mathcal{X}_j) = \sigma_j^2$, with the restriction that there exists a constant $\Sigma^2 < \infty$ such that $\sigma_j^2 \leq \Sigma^2$ is fulfilled for all $\mathcal{X}_j$. Then we have for each $c > 0$:

$$\lim_{n\to\infty} P\left(\left|\frac{\mathcal{X}_1 + \ldots + \mathcal{X}_n}{n} - \frac{\mu^{(1)} + \ldots + \mu^{(n)}}{n}\right| < c\right) = 1 \,. \tag{2.60}$$

The main message of the law of large numbers is that for a sufficiently large numbers of independent events the statistical errors in the sum will vanish and the mean converges to the exact expectation value. Hence, the law of large numbers provides the basis for the assumption of convergence in mathematical statistics (section 2.5).

2.3.8 Law of the iterated logarithm

The *law of the iterated logarithm* consists of two asymptotic regularities of the sums of random variables, which are related to the central limit theorem and the law of large numbers, and in a way complete the predictions of both. The name of the law points at the appearance of the function 'log log' in the forthcoming expressions – it does not refer to the notion of iterated logarithm in computer science[16] – and the derivation is attributed to the two Russian scholars of mathematics Aleksandr Khinchin and Andrey Kolmogorov [160, 166] and the proof for more general case used here was provided later [73, 122]. The law of the iterated logarithm provides upper and lower bounds for the values of sums of random variables and in this ways confines the size of fluctuations.

For a sum of n independent and identically distributed (iid) random variables with expectation value $E(\mathcal{X}_i) = \mu$ and finite variance $\mathrm{var}(\mathcal{X}) = \sigma^2 < \infty$,

$$\mathcal{S}_n = \mathcal{X}_1 + \mathcal{X}_1 + \ldots + \mathcal{X}_n \,,$$

[16] In computer science the iterated logarithm of n is commonly written $\log^* n$ and represents the number of times the logarithmic function must be iteratively applied before the result is is less than or equal to one:

$$\log^* \doteq \begin{cases} 0 & \text{if } n \leq 1 \,, \\ 1 + \log^*(\log n) & \text{if } n > 1 \,. \end{cases}$$

The iterated logarithm is well defined for base 'e', for base '2' and in general for any base greater than $e^{1/e} = 1.444667\ldots$.

Fig. 2.11 Illustration of the law of the iterated logarithm..

the following two limits are fulfilled with probability one:

$$\limsup_{n\to\infty} \frac{\mathcal{S}_n - n\,\mu}{\sqrt{2\,n\,\ln(\ln\,n)}} = +\,|\sigma| \quad \text{and} \tag{2.61a}$$

$$\liminf_{n\to\infty} \frac{\mathcal{S}_n - n\,\mu}{\sqrt{2\,n\,\ln(\ln\,n)}} = -\,|\sigma|\,. \tag{2.61b}$$

The two theorems 2.61 are equivalent and this follows directly from the symmetry of the standardized normal distribution, $\mathcal{N}(0,1)$. We dispense here from the presentation of a proof for the law of the iterated logarithm that can be found, for example, in a monograph by Henry McKean [201] or in a publication by William Feller [73]. For the purpose of illustration we compare with the already mentioned heuristic $\sqrt{n}$-law (see section 1.1), which is based on the properties of the standardized binomial distribution $B(n,p)$ with $p = \frac{1}{2}$ (section 2.3.2): The variance is to $\sigma^2 = np(1-p) = n/4$ and accordingly we have $2\sigma/n = 1/\sqrt{n}$ and accordingly most values of $\mathcal{S}_n - n\mu$ lie in the interval $-|\sigma| \le \mathcal{S}_n \le +|\sigma|$. The corresponding result from the law of the iterated logarithm is

$$-\sqrt{\frac{2\,\ln\,(\ln\,n)}{n}} \;\le\; \mathcal{S}_n \;\le\; +\sqrt{\frac{2\,\ln\,(\ln\,n)}{n}}$$

with probability one. One particular case of iterated Bernoulli trials – tosses of a fair coin, is shown in figure 2.11, where the envelope of the sum $\mathcal{S}_n$ of the cumulative score of n trials, $\pm\sqrt{2\,\ln(\ln n)/n}$ is compared with the results of the square root n law, $\mu \pm \sigma = \pm\sqrt{1/n}$.

The special importance of the results of the law of the iterated logarithm for the Wiener process will be discussed later (section 3.2.3.2).

In essence, we may summarize the results of this section in three statements, which are part of *large sample theory*: For independent and identically distributed (iid) random variables $\mathcal{X}_i$ with $\mathcal{S}_n = \sum_{i=1}^{n} \mathcal{X}_i$ with $E(\mathcal{X}_i) = E(\mathcal{X}) = \mu$ and finite variance $\mathrm{var}(\mathcal{X}_i) = \sigma < \infty$ we have the three large sample results:

(i) the *law of large numbers*: $\mathcal{S}_n \to n\,E(\mathcal{X}) = n\,\mu\,,$

(ii) the *law of the iterated logarithm*: $\begin{cases} \limsup \frac{(S_n - n\mu)}{\sqrt{2n\ln(\ln n)}} & \to\; +|\sigma| \\ \liminf \frac{(S_n - n\mu)}{\sqrt{2n\ln(\ln n)}} & \to\; -|\sigma| \end{cases}$, and

(iii) the *central limit theorem*: $\frac{1}{\sqrt{n}}\big(\mathcal{S}_n - n\,E(\mathcal{X})\big) \to \mathcal{N}(0,1)\,.$

The theorem (i) defines the limit of the expectation value, theorem (ii) determines the size of fluctuations and theorem (iii), eventually, refers to the

limiting distribution function, which turns out to be the normal distribution. All three theorems can be extended in their range of validity to independent random variables with arbitrary distributions provided mean and variance are finite.

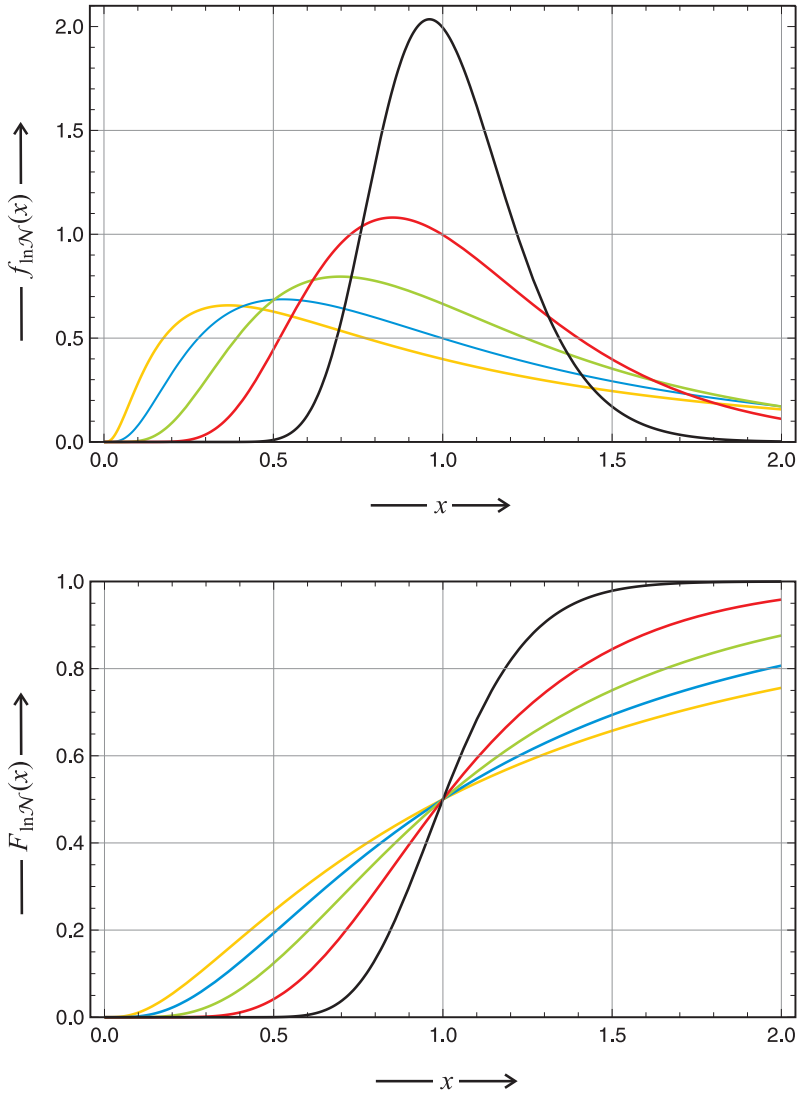

Fig. 2.12 The log-normal distribution. The log-normal distribution, $\ln\mathcal{N}(\mu,\sigma)$, is defined on the positive real axis, $x \in]0,\infty[$ and has the probability density (pdf)

$$f_{\ln\mathcal{N}}(x) = \exp\Big(-(\ln x-\mu)^2/(2\sigma^2)\Big) \;\Big/\; (x\sqrt{2\pi\sigma^2})$$

and the cumulative distribution function (cdf)

$$F_{\ln\mathcal{N}}(x) = \Big(1+\operatorname{erf}\Big((\ln x-\mu)/\sqrt{2\sigma^2}\Big)\Big) \;\Big/\; 2\ .$$

The two parameters are confined by the relations $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Parameter choice and color code: $\mu = 0$, σ =0.2 (black), 0.4 (red), 0.6 (green), 0.8 (blue), and 1.0 (yellow).

2.4 Further probability distributions

In the previous section 2.3 we presented the three most relevant probability distributions: (i) the Poisson distribution because it describes the distribution of occurrence of independent events, (ii) the binomial distribution dealing with independent trials with two outcomes, and (iii) the normal distribution being the limiting distribution of large numbers of individual events irrespectively of the statistics of single events. In this section we shall discuss seven more or less arbitrarily selected distributions, which play an important role in science and/or in statistics. The presentation here is inevitably rather brief and for reading of a detailed treatise we refer to [148, 149].

2.4.1 The log-normal distribution

The *log-normal* distribution in a continuous probability distribution of a random variable $\mathcal{Y}$ with a normally distributed logarithm. In other words, if $\mathcal{X} = \log \mathcal{Y}$ is normally distributed then $\mathcal{Y} = \exp(\mathcal{X})$ has a log-normal distribution. Accordingly $\mathcal{Y}$ can take on only positive real values. Historically, this distribution had several other names the most popular of them being *Galton's distribution* named after the pioneer of statistics in England, Francis Galton or *McAlister's distribution* after the statistician Donald McAlister [148, chap. 14, pp. 207-258].

The log-normal distribution meets the need for modeling empirical data that show frequently observed deviation from the conventional normal distribution: (i) meaningful data are non-negative, (ii) positive skew implying that there are more values above then below the maximum of the probability density function (pdf), and (iii) more obvious meaning of the geometric rather than the arithmetic mean [90, 199]. Despite its obvious usefulness and applicability to problems in science, economics, and sociology the log-normal distribution is not popular among non-statisticians [179].

The log-normal distribution contains two parameters, $\ln \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{>0}$, and is defined on the domain $x \in]0, \infty[$. The density function and the cumulative distribution (cdf) are given by (figure 2.12):

$$\begin{aligned} \text{pdf}: \quad f_{\ln \mathcal{N}}(x) &= \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(\ln x - \mu)^2}{2\sigma^2}\Big) \\ \text{cdf}: \quad F_{\ln \mathcal{N}}(x) &= \frac{1}{2}\left(1 + \operatorname{erf}\Big(\frac{\ln x - \mu}{\sqrt{2\sigma^2}}\Big)\right) . \end{aligned} \tag{2.62}$$

By definition the logarithm of the variable $\mathcal{X}$ is normally distributed, and this implies

$$\mathcal{X} = e^{\mu + \sigma \mathcal{Z}} ,$$

where $\mathcal{N}$ is a standard normal variable. The moments of the log-normal distribution are readily calculated[17]

$$\begin{aligned}
&\text{mean :} && e^{\mu+\sigma^2/2}\ ,\\
&\text{median :} && e^{\mu}\ ,\\
&\text{mode :} && e^{\mu-\sigma^2}\ ,\\
&\text{variance :} && (e^{\sigma^2}-1)\,e^{2\mu+\sigma^2}\ ,\\
&\text{skewness :} && (e^{\sigma^2}+2)\,\sqrt{e^{\sigma^2}-1}\,,\ \text{and}\\
&\text{kurtosis :} && e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-6\ .
\end{aligned} \tag{2.63}$$

The skewness γ_1 is always positive and so is the excess kurtosis since $\sigma^2 = 0$ yields $\gamma_2 = 0$, and $\sigma^2 > 0$ implies $\gamma_2 > 0$.

The entropy of the log-normal distribution is

$$H(f_{\ln\mathcal{N}}) \;=\; \frac{1}{2}\Big(1 \,+\, \ln(2\pi\sigma^2) \,+\, 2\,\mu\Big)\ . \tag{2.64}$$

Like the normal distribution has the maximum entropy of all distribution defined on the real axis, $x \in \mathbb{R}$, the log-normal distribution is the maximum entropy probability distribution for a random variable $\mathcal{X}$ for which mean and variance of $\ln\mathcal{X}$ is fixed.

Finally, we mention that the log-normal distribution can be well approximated by a distribution [273]

$$F(x;\,\mu\sigma) \;=\left(\left(\frac{e^{\mu}}{x}\right)^{\pi/(\sigma\sqrt{3})} \,+\, 1\right)^{-1}$$

that has integrals that can be expressed in terms of elementary functions.

2.4.2 The χ^2-distribution

The χ^2-distribution also written as chi-squared distribution is one of the most frequently used distribution in inferential statistics for hypothesis testing and construction of confidence intervals. In particular, the χ^2 distributions is applied in the common χ^2-test for the quality of the fit of an empirically determined distribution to a theoretical one (section 2.5.2). Many other statistical tests are based on the χ^2-distribution as well.

[17] Here and in the following listings for other distributions t't'kurtosis" stands for excess kurtosis $\gamma_2 = \beta_2 - 3 = \mu_4 \,/\, \sigma^4$.

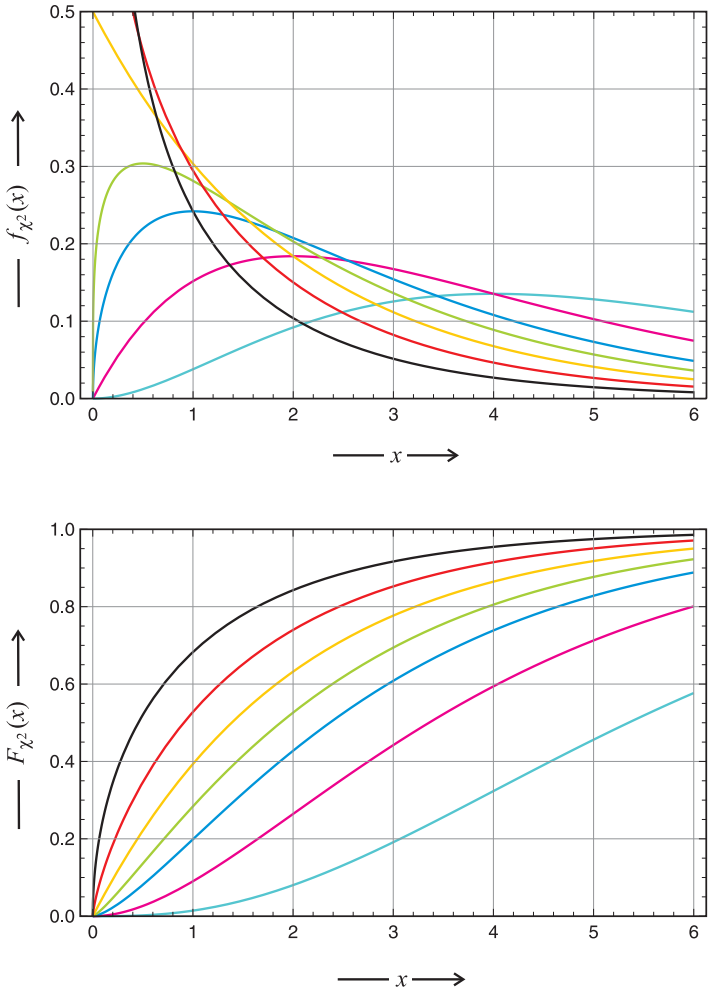

Fig. 2.13 The χ^2 distribution. The chi-squared distribution, χ^2_k , $k \in \mathbb{N}$, is defined on the positive real axis, $x \in [0, \infty[$, with the parameter k called the number of the degrees of freedom, has the probability density (pdf)

$$f_{\chi^2_k}(x) = x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \Big/ \left(2^{\frac{k}{2}} \Gamma\left(\tfrac{k}{2}\right)\right)$$

and the cumulative distribution function (cdf)

$$F_{\chi^2_k}(x) = \gamma\left(\tfrac{k}{2}, \tfrac{x}{2}\right) \Big/ \Gamma\left(\tfrac{k}{2}\right) .$$

Parameter choice and color code: k =1 (black), 1.5 (red), 2 (yellow), 2.5 (green), 3 (blue), 4 (magenta) and 6 (cyan). Although k, the number of degrees of freedom, is commonly restricted to integer values, we show here also the curves for two intermediate values (k=1.5, 2.5).

The chi-squared distribution, χ^2_k,[18] is the distribution of a random variable $\mathcal{Q}$, which is given by the sum of the squares of k independent, standard normal variables with distribution $\mathcal{N}(0,1)$

$$\mathcal{Q} \;=\; \sum_{i=1}^{k} \mathcal{X}_i^2 \;, \tag{2.65}$$

where the only parameter of the distribution, k, is called the number of the *degrees of freedom* being tantamount to the number of independent variables $\mathcal{X}_i$. $\mathcal{Q}$ is defined on the positive real axis (including zero), $x \in [0,\infty[$ and has the following density function and cumulative distribution (figure 2.13):

$$\begin{aligned} \text{pdf}: \quad f_{\chi^2}(x) \;&=\; \frac{x^{\frac{k}{2}-1}\, e^{-\frac{x}{2}}}{2^{\frac{k}{2}}} \;, x \in \mathbb{R}_{\geq 0} \;\; \text{and} \\ \text{cdf}: \quad F_{\chi^2}(x) \;&=\; \frac{\gamma\left(\frac{k}{2},\frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \;=\; P\left(\frac{k}{2},\frac{x}{2}\right) \;. \end{aligned} \tag{2.66}$$

where $\gamma(k,z)$ is the lower incomplete Gamma function and $P(k,z)$ is the regularized Gamma function. The special case with $k = 2$ has the particularly simple form: $F_{\chi^2}(x;2) = 1 - e^{-\frac{x}{2}}$.

The conventional χ^2-distribution is sometimes denoted as *central* χ^2-distribution in order to distinguish it from the *noncentral* χ^2-distribution, which is derived from k independent and normally distributed variables with means μ_i and variances σ_i^2. The random variable

$$\mathcal{Q} \;=\; \sum_{i=1}^{k} \left(\frac{\mathcal{X}_i}{\sigma_i}\right)^2$$

is distributed according to the noncentral χ^2-distribution $\chi^2_k(\lambda)$ with two parameters, k and λ, where $\lambda = \sum_{i=1}^{k} (\mu_i/\sigma_i)^2$ is the *noncentrality* parameter.

The moments of the central χ^2_k-distribution are readily calculated

[18] The chi-squared distribution is sometimes written $\chi^2(k)$ we prefer the subscript since the number of degrees of freedom, the parameter k, specifies the distribution. Often the random variables $\mathcal{X}_i$ fulfil a conservation relation and then the number of independent variables is reduced to $k-1$, and we have χ^2_{k-1} (section 2.5.2).

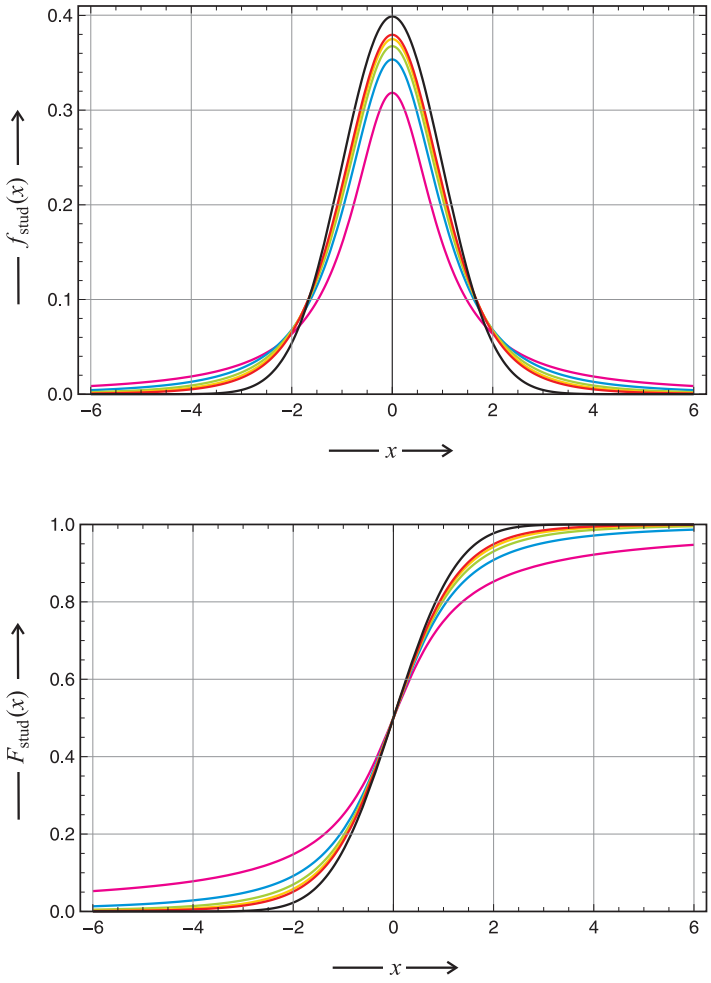

Fig. 2.14 Student's t-distribution. Student's distribution is defined on the real axis, $x \in]-\infty, +\infty[$, with the parameter $r \in \mathbb{N}_{>0}$ called the number of degrees of freedom, has the probability density (pdf)

$$f_{\text{stud}}(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\,\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{\frac{r+1}{2}}$$

and the cumulative distribution function (cdf)

$$F_{\text{stud}}(x) = \frac{1}{2} + x\Gamma\left(\frac{r+1}{2}\right) \cdot \frac{{}_2F_1\left(\frac{1}{2}, \frac{r+1}{2}, \frac{3}{2}, -\frac{x^2}{r}\right)}{\sqrt{\pi r}\,\Gamma\left(\frac{r}{2}\right)} \ .$$

The first curve (magenta, $r = 1$) represents the density of the Cauchy-Lorentz distribution (figure 2.17). Parameter choice and color code: r =1 (magenta), 2 (blue), 3 (green), 4 (yellow), 5 (red) and $+\infty$ (black). The black curve representing the limit $r \to \infty$ of Student's distribution is the standard normal distribution.

$$\begin{aligned}
&\text{mean :} && k\ , \\
&\text{median :} && \approx\ k\,\Big(1-\frac{2}{9k}\Big)^3\ , \\
&\text{mode :} && \max\{k-2\,,\,0\}\ , \\
&\text{variance :} && 2k\ , \\
&\text{skewness :} && \sqrt{8/k}\,,\ \text{and} \\
&\text{kurtosis :} && 12/k\ .
\end{aligned} \tag{2.67}$$

The skewness γ_1 is always positive and so is the excess kurtosis γ_2. The raw moments $\hat{\mu}_n = E(\mathcal{Q}^n)$ and the cumulants of the χ^2_k-distribution have particularly simple expressions:

$$E(\mathcal{Q}^n) \;=\; \hat{\mu}_n \;=\; k(k+2)(k+4)\cdots(k+2n-2) \;=\; 2^n\frac{\Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \quad \text{and} \tag{2.68}$$

$$\kappa_n \;=\; 2^{n-1}\,(n-1)!\,k\ . \tag{2.69}$$

The entropy of the χ^2_k-distribution is readily calculated by integration:

$$H(f_{\chi^2}) \;=\; \frac{k}{2} \,+\, \ln\left(2\,\Gamma\big(\frac{k}{2}\big)\right) \,+\, \left(1-\frac{k}{2}\right)\cdot\psi\left(\frac{k}{2}\right)\ , \tag{2.70}$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\ln\Gamma(x)$ is the digamma function.

The χ^2_k-distribution has a simple characteristic function

$$\phi_{\chi^2}(s) \;=\; (1\,-\,2i\!\!\!^{-}\,s)^{-k/2}\ . \tag{2.71}$$

The moment generating function is defined only for $s < \frac{1}{2}$:

$$M_{\chi^2}(s) \;=\; (1\,-\,2\,s)^{-k/2} \quad \text{for} \quad s \,<\, \frac{1}{2}\ . \tag{2.72}$$

Because of its central importance for tests of significance numerical tables of the χ^2-distribution are found in almost every textbook of mathematical statistics.

2.4.3 Student's t-distribution

Student's t-distribution has a remarkable history. It has been discovered by the famous English statistician William Sealy Gosset who published his works under the pen name *Student* [234]. Gosset was working at the brewery of Arthur Guinness in Dublin, Ireland, where it was forbidden to publish any

paper regardless of the contained information, because Guinness was afraid that trade secrets and other confidential information could be disclosed. Almost all of Gosset's paper including the one describing the t-distribution were published under the pseudonym "Student" [272]. Gosset's work has been known to and was supported by Karl Pearson but it was Ronald Fisher who appreciated the importance of Gosset's work on small samples [81].

Student's t-distribution is a family of continuous, normal probability distributions that applies to situations when the sample size is small, the variance is unknown and one wants to derive a reliable estimate of the mean. Student's distribution plays a role in a number of commonly used tests in analyzing statistical data an example being Student's test accessing the significance of differences between two sample means – for example to find out whether or not a difference in mean body height between basketball players and soccer players is significant – or the construction of confidence intervals for the difference between population means. In a way Student's t-distribution is required for *higher order statistics* in the sense of a statistics of statistics, for example, to estimate, how likely it is to find the true mean within a given range around the finite sample mean (section 2.5). In other words, n samples are taken from a population with a normal distribution having fixed but unknown mean and variance, the sample mean and the sample variance are computed from these n points and the t-distribution is the distribution of the location of the true mean relative to the sample mean, calibrated by the sample standard deviation.

To make the meaning of Student's t-distribution precise we assumes n independent random variables $\mathcal{X}_i$, $i = 1, \ldots, n$ drawn from the same population which is normally distributed with mean value $E(\mathcal{X}_i) = \mu$ and variance $\mathrm{var}(\mathcal{X}_i) = \sigma^2$. Then the sample mean and the unbiased sample variance are the random variables

$$\overline{\mathcal{X}}_n \;= \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i \;\text{ and }\; \mathcal{S}_n^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n} (\mathcal{X}_i - \overline{\mathcal{X}}_n)^2 \;.$$

As follows from Cochran's theorem [35] the random variable $\mathcal{V} = (n-1)\mathcal{S}_n^2/\sigma^2$ follows a χ^2-distribution with $r = n-1$ degrees of freedom. The deviation of the sample mean from the population mean is properly expressed by the variable

$$\mathcal{Z} \;=\; (\overline{\mathcal{X}}_n - \mu)\,\frac{\sqrt{n}}{\sigma} \;, \tag{2.73}$$

which is the basis for the calculation of z-scores.[19] The variable $\mathcal{Z}$ is normally distributed with mean zero and variance one as follows from the fact that the sample mean $\overline{\mathcal{X}}_n$ obeys a normal distribution with mean μ and variance σ^2/n. In addition, the two random variables $\mathcal{Z}$ and $\mathcal{V}$ are independent, and the *pivotal quantity*[20]

$$\mathcal{T} \;:=\; \frac{\mathcal{Z}}{\sqrt{\mathcal{V}/(n-1)}} \;=\; \overline{\mathcal{X}}_n - \mu)\,\frac{\sqrt{n}}{\mathcal{S}_n} \tag{2.74}$$

follows a Student's t-distribution, which depends on the degrees of freedom $r = n-1$ but neither on μ nor on σ.

Student's distribution is a one parameter distribution with r being the number of sample points or the so-called degree of freedom. It is symmetric and bell-shaped like the normal distribution but the tails are heavier in the sense that more values fall further away from the mean. Student's distribution is defined on the real axis, $x \in]-\infty,+\infty[$ and has the following density function and cumulative distribution (figure 2.14):

$$\begin{aligned}
\text{pdf}: \quad f_{\text{stud}}(x) \;&=\; \frac{\Gamma\Big(\frac{r+1}{2}\Big)}{\sqrt{\pi r}\;\Gamma\big(\frac{r}{2}\big)}\,\left(1+\frac{x^2}{r}\right)^{-\frac{r+1}{2}}\;,\; x\in\mathbb{R} \;\text{ and} \\
\text{cdf}: \quad F_{\text{stud}}(x) \;&=\; \frac{1}{2} + x\,\Gamma\left(\frac{r+1}{2}\right)\cdot\frac{{}_2F_1\left(\frac{1}{2},\frac{r+1}{2},\frac{3}{2},-\frac{x^2}{r}\right)}{\sqrt{\pi r}\;\Gamma\big(\frac{r}{2}\big)}\;.
\end{aligned} \tag{2.75}$$

where ${}_2F_1$ is the hypergeometric function. The t-distribution has simple expressions for several special cases:

(i) $r=1$, Cauchy-distribution: $f(x) = \frac{1}{\pi(1+x^2)}$, $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$,

(ii) $r=2$: $f(x) = \frac{1}{(2+x^2)^{\frac{3}{2}}}$, $F(x) = \frac{1}{2}\left(1+\frac{x}{\sqrt{2+x^2}}\right)$,

(iii) $r=3$: $f(x) = \frac{6\sqrt{3}}{\pi(3+x^2)^2}$, $F(x) = \frac{1}{2} + \frac{\sqrt{3}x}{\pi(3+x^2)} + \frac{1}{\pi}\arctan\left(\frac{x}{\sqrt{3}}\right)$,

(iv) $r=\infty$, normal distribution: $f(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{x^2}{2}}$, $F(x) = \Phi(x)$.

Formally the t-distribution represents an interpolation between the Cauchy-Lorentz distribution (section 2.4.6) and the normal distribution both standardized to mean zero and variance one. In this sense it has a lower maximum and heavier tails than the normal distribution and a higher maximum and less heavy tails than the Cauchy-Lorentz distribution.

---

[19] In mathematical statistics (section 2.5) the quality of measured data is often characterized by scores. The z-score of a sample corresponds to the random variable $\mathcal{Z}$ (2.73) and it is measured in standard deviations from the population mean as unites. In bone densitometry this evaluation relative to the *peak bone mass* it is often replaced by a so-called T-value that measures the bone density of an individual relative to the mean within a subpopulation consisting of his age group.

[20] A pivotal quantity or a pivot is a function of measurable and unmeasurable parameters whose probability distribution does not depend on the unknown parameters.

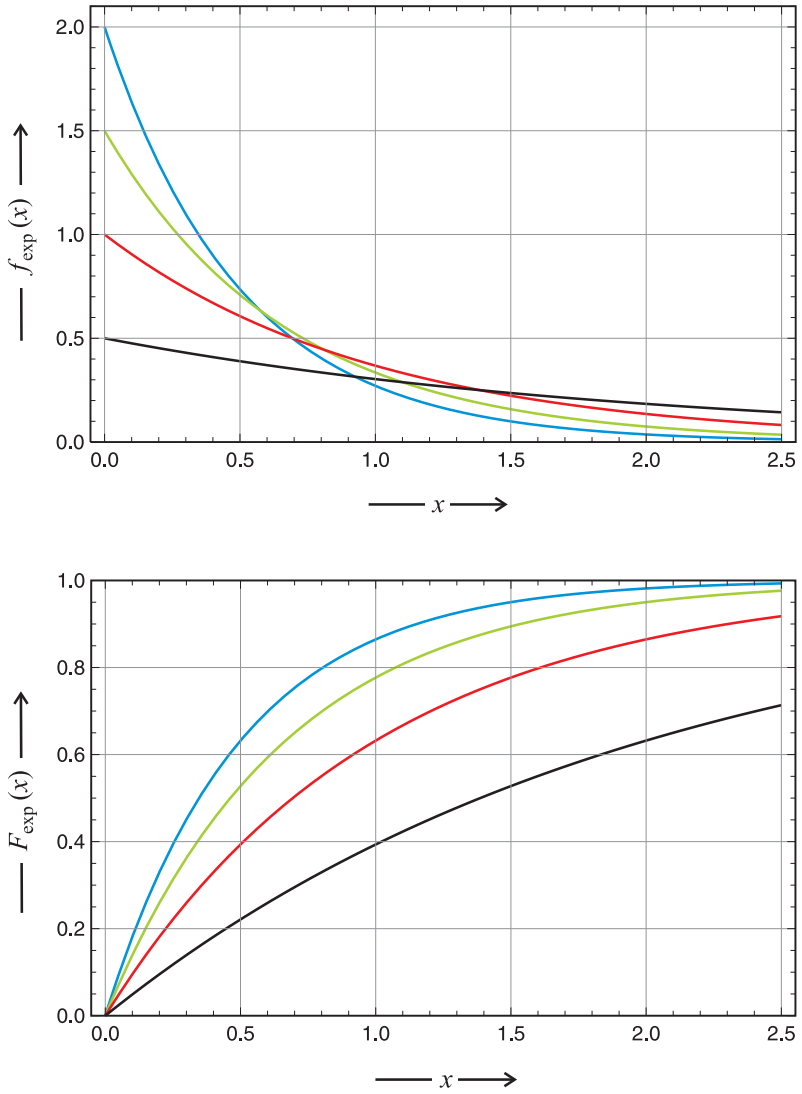

Fig. 2.15 The exponential distribution. The exponential distribution is defined on the real axis including zero, $x \in [0, +\infty[$, with the parameter $\lambda \in \mathbb{R}_{>0}$ called the rate parameter, and has the probability density (pdf)

$$f_{\exp}(x) = \lambda \exp(-\lambda x)$$

and the cumulative distribution function (cdf)

$$F_{\exp}(x) = 1 - \exp(-\lambda x) \ .$$

Parameter choice and color code: λ =0.5 (black), 2 (red), 3 (green), and 4 (blue).

The moments of Student's distribution are readily calculated

$$\begin{aligned}
&\text{mean}: && 0,\ \text{for } r>1,\ \text{otherwise undefined}\,,\\
&\text{median}: && 0\,,\\
&\text{mode}: && 0\,,\\
&\text{variance}: && \begin{cases}\infty & \text{for}\ \ 1<r\le 2\,,\\ \frac{r}{r-2} & \text{for}\ \ r>2\,,\\ \text{undefined} & \text{otherwise}\,,\end{cases}\\
&\text{skewness}: && 0,\ \text{for } r>3,\ \text{otherwise undefined}\,,\ \text{and}\\
&\text{kurtosis}: && \begin{cases}\infty & \text{for}\ \ 2<r\le 4\,,\\ \frac{6}{r-4} & \text{for}\ \ r>4\,,\\ \text{undefined} & \text{otherwise}\,.\end{cases}
\end{aligned} \tag{2.76}$$

If the variance of Student's distribution is defined it is larger one the variance of the standard normal distribution. In the limit of infinite degrees of freedom Student's distribution converges to the standard normal distribution and so does the variance, inevitably: $\lim_{r\to\infty}\frac{r}{r-2}=1$. Student's distribution is symmetric and hance the skewness γ_1 is either zero or undefined, and the excess kurtosis γ_2 is undefined or positive and converges to zero in the limit $r\to\infty$.

The raw moments $\hat{\mu}_n = E(\mathcal{T}^n)$ of the t-distribution have fairly simple expressions:

$$E(\mathcal{T}^k) \;=\; \begin{cases} 0 & k\ \text{odd}\,,\ 0<k<r\,,\\ \frac{1}{\sqrt{\pi}\,\Gamma\left(\frac{r}{2}\right)}\, r^{\frac{k}{2}}\,\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{r-k}{2}\right) & k\ \text{even}\,,\ 0<k<r\,,\\ \text{undefined} & k\ \text{odd}\,,\ 0<r\le k\,,\\ \infty & k\ \text{even}\,,\ 0<r\le k\,. \end{cases} \tag{2.77}$$

The entropy of Student's t-distribution is readily calculated by integration:

$$H(f_{\text{stud}}) \;=\; \frac{k+1}{2}\,\left(\psi\Big(\frac{1+r}{2}\Big)-\psi\Big(\frac{r}{2}\Big)\right)\,+\,\ln\left(\sqrt{r}\,B\Big(\frac{r}{2},\frac{1}{2}\Big)\right)\ , \tag{2.78}$$

where $\psi(x)=\frac{\mathrm{d}}{\mathrm{d}x}\ln\Gamma(x)$ is the digamma function and
$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{dt}$ is the beta function.

Student's-distribution has the characteristic function

$$\phi_{\text{stud}}(s) \;=\; \frac{K_{r/2}(\sqrt{r}\,|s|)\cdot(\sqrt{r}\,|s|)^{r/2}}{2^{\frac{r}{2}-1}\cdot\Gamma\left(\frac{r}{2}\right)} \quad \text{for}\ \ r>0\ . \tag{2.79}$$

where $K_\alpha(x)$ is a modified Bessel function.

2.4.4 The exponential and the geometric distribution

The *exponential distribution* is a family of continuous probability distributions, which describe the distribution of the time intervals between events in a Poisson process (section 3.2.3.5), which is a process where the number of events in any time interval has a Poisson distribution.[21] The Poisson process is a process where events occur steadily, independently of each other and at a constant average rate $\lambda \in \mathbb{R}_{>0}$, which is the only parameter of the exponential distribution (and the Poisson process).

The exponential distribution has widespread applications in science and sociology. It describes the time to decay of radioactive atoms, other irreversible first order processes in chemistry and biology, the waiting times in all kinds of queues from independently acting customers, the time to failure of components with constant failure rates, and many other events.

The exponential distribution is defined on the positive real axis, $x \in [0, \infty[$, with a positive rate parameter $\lambda \in]0, \infty[$. The density function and cumulative distribution are of the form (figure 2.15):

$$\begin{aligned} \text{pdf}: \quad & f_{\exp}(x) \;=\; \lambda \exp(-\lambda x),\ x \in \mathbb{R}_{>0} \ \text{ and} \\ \text{cdf}: \quad & F_{\exp}(x) \;=\; 1 - \exp(-\lambda x),\ x \in \mathbb{R}_{>0}\ . \end{aligned} \tag{2.80}$$

The moments of exponential distribution are readily calculated

$$\begin{aligned} &\text{mean}: && \lambda^{-1} \;=\; \mu\ , \\ &\text{median}: && \lambda^{-1} \ln 2\ , \\ &\text{mode}: && 0\ , \\ &\text{variance}: && \lambda^{-2}\ , \\ &\text{skewness}: && 2\ , \ \text{and} \\ &\text{kurtosis}: && 6\ . \end{aligned} \tag{2.81}$$

A commonly used alternative parametrization uses a *survival parameter* $\beta = \mu = \lambda^{-1}$ instead of the rate parameter, and survival is often measured in terms of *half-life*, which is the expectation value of the time when one half of the events have taken place – for example 50 % of the atoms have decayed

[21] It is important to distinguish the exponential distribution and the class of *exponential families of distributions*, which comprises many other distributions like the normal distribution, the Poisson distribution, the binomial distribution and many others [62].

– and represents just another name for the median: $\overline{\mu} = \beta \ln 2 = \ln 2/\lambda$. The exponential distribution provides an easy to verify test case for the median-mean inequality:

$$|E(\mathcal{X}) - \overline{\mu}| \;=\; \frac{1 - \ln 2}{\lambda} \;<\; \frac{1}{\lambda} \;=\; \sigma \;.$$

The raw moments of the exponential distribution are given simply by

$$E(\mathcal{X}^n) \;=\; \hat{\mu}_n \;=\; \frac{n!}{\lambda^n} \;. \tag{2.82}$$

Among all probability distribution with the support $[0, \infty[$ and mean μ the exponential distribution with $\lambda = 1/\mu$ has the largest entropy (section 2.1.3):

$$H(f_{\exp}) \;=\; 1 \,-\, \log \lambda \;=\; 1 \,+\, \log \mu \;. \tag{2.20'}$$

The moment generation function of the exponential distribution is

$$M_{\exp}(s) \;=\; \left(1 \,-\, \frac{s}{\lambda}\right)^{-1} \;, \tag{2.83}$$

and the characteristic function is

$$\phi_{\exp}(s) \;=\; \left(1 \,-\, \frac{\mathrm{i}\, s}{\lambda}\right)^{-1} \;. \tag{2.84}$$

Finally, we mention a property of the exponential distribution that makes it unique among all continuous probability distributions: It is *memoryless*. Memorylessness can be encapsulated in an example called "hitchhiker's dilemma": Waiting for hours on a lonely road does not increase the probability of arrival of the next car. Cast into probabilities this means for a random variable $\mathcal{T}$:[22]

$$P(\mathcal{T} > s + t \,|\, T > s) \;=\; P(T > t) \;\; \forall \;\; s, t \,\geq\, 0 \;. \tag{2.85}$$

In other words, the probability of arrival does not change no matter how many events have happened.

The discrete analogue to the exponential distribution is the geometric distribution. Considered is a sequence of independent Bernoulli trials with p being the probability of success and the only parameter of the distribution: $0 < p \leq 1$. The random variable $\mathcal{X} \in \mathbb{N}$ is the number of trials before the first success.

The probability mass function and the cumulative distribution function of the geometric distribution are:

[22] We remark that memoryless is not tantamount to independence. Independence would require $P(\mathcal{T} > s + t \,|\, T > s) \;=\; P(T > s + t)$.

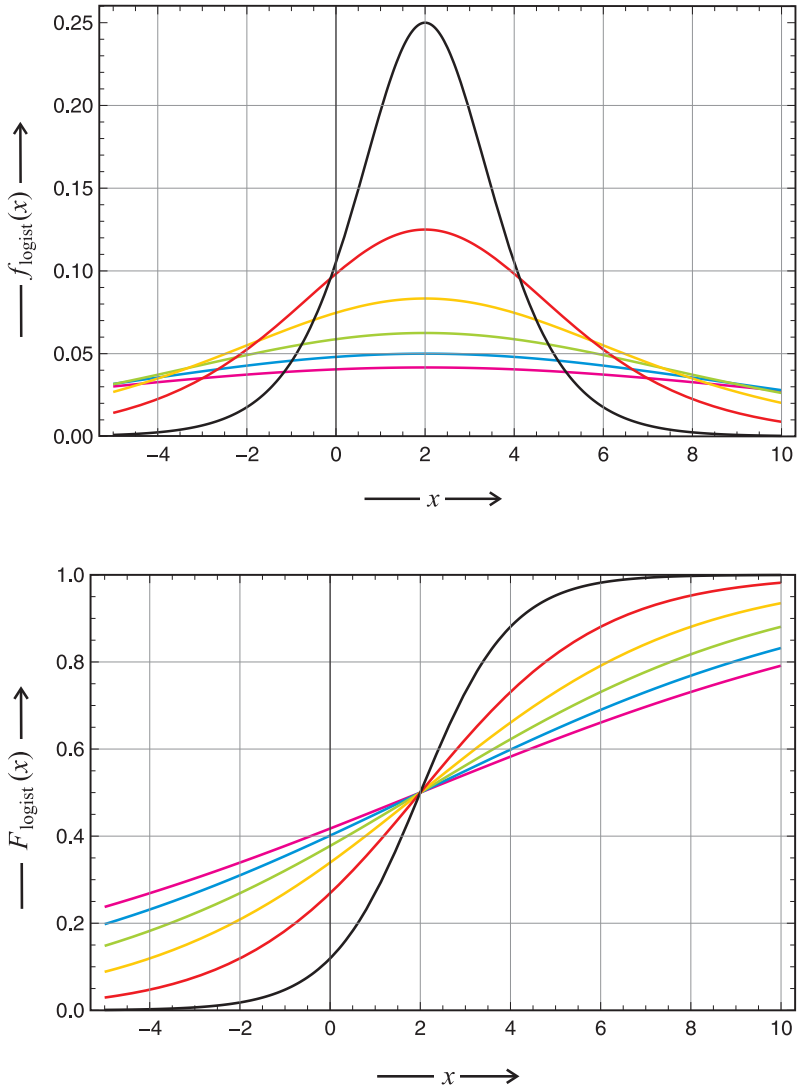

Fig. 2.16 The logistic distribution. The logistic distribution is defined on the real axis, $x \in]-\infty, +\infty[$, with two parameters, the location $\mu \in \mathbb{R}$ and the scale $b \in \mathbb{R}_{<0}$, has the probability density (pdf)

$$f_{\text{logist}}(x) = \frac{e^{-(x-\mu)b}}{b\,(1+e^{-(x-\mu)/b})^2}$$

and the cumulative distribution function (cdf)

$$F_{\text{logist}}(x) = \frac{1}{1+e^{-[x-\mu)/b}} \ .$$

Parameter choice and color code: $\mu = 2$, $b = 1$ (black), 2 (red), 3 (yellow), 4 (green), 5 (blue) and 6 (magenta).

$$\begin{aligned} \text{pmf}: \quad & f^{\text{geom}}_{k;p} \;=\; p \cdot (1-p)^k,\ k \in \mathbb{N} \ \text{ and} \\ \text{cdf}: \quad & F^{\text{geom}}_{k;p} \;=\; 1 \,-\, (1-p)^{k+1}),\ k \in \mathbb{N}\ . \end{aligned} \tag{2.86}$$

The moments of geometric distribution are readily calculated

$$\begin{aligned} &\text{mean}: && \frac{1-p}{p}\ , \\ &\text{median}: && \lambda^{-1}\ \ln 2\ , \\ &\text{mode}: && 0\ , \\ &\text{variance}: && \frac{1-p}{p^2}\ , \\ &\text{skewness}: && \frac{2-p}{\sqrt{1-p}}\ , \text{ and} \\ &\text{kurtosis}: && 6\ +\ \frac{p^2}{1-p}\ . \end{aligned} \tag{2.87}$$

Like the exponential distribution the geometric distribution is lacking memory in the sense of equation (2.85). The information entropy has the form

$$H(f^{\text{geom}}_{k;p}) \;=\; -\frac{1}{p}\Big((1-p)\ \log(1-p) + p\ \log p\Big)\ . \tag{2.88}$$

Finally, we present the moment generating function and the characteristic function of the geometric distribution:

$$M_{\text{geom}}(s) \;=\; \frac{p}{1-(1-p)\exp(s)} \quad \text{and} \tag{2.89}$$

$$\phi_{\text{geom}}(s) \;=\; \frac{p}{1-(1-p)\exp(\mathrm{i}\, s)}\ , \tag{2.90}$$

respectively.

2.4.5 The logistic distribution

The logistic distribution is commonly used as a model for growth with limited resources. It is applied, for example, in economics to model the market penetration of a new product, in biology for population growth in an ecosystem, in agriculture for the expansion of agricultural production or to weight gain in animal fattening. It is a continuous probability distribution with two parameters, the position of the mean μ and the scale b. The cumulative distribution function of the logistic distribution is the *logistic function*.

The logistic distribution is defined on the real axis, $x \in]+\infty, \infty[$, with two parameters, the position of the mean $\mu \in \mathbb{R}$ and the scale $b \in \mathbb{R}_{>0}$. The density function and cumulative distribution are of the form (figure 2.16):

$$\begin{aligned} \text{pdf}: \quad & f_{\text{logist}}(x) \;=\; \lambda \, \exp(-\lambda x), \;\; x \in \mathbb{R}_{>0} \;\; \text{and} \\ \text{cdf}: \quad & F_{\text{logist}}(x) \;=\; 1 \,-\, \exp(-\lambda x), \;\; x \in \mathbb{R}_{>0} \; . \end{aligned} \tag{2.91}$$

The moments of logistic distribution are readily calculated

$$\begin{aligned} &\text{mean}: && \mu \; , \\ &\text{median}: && \mu \; , \\ &\text{mode}: && \mu \; , \\ &\text{variance}: && \frac{\pi^2 \, b^2}{3} \; , \\ &\text{skewness}: && 0 \; , \; \text{and} \\ &\text{kurtosis}: && \frac{6}{5} \; . \end{aligned} \tag{2.92}$$

A frequently used alternative parametrization uses the variance as parameter, $\sigma = \pi b/\sqrt{3}$ or $b = \sqrt{3}\,\sigma/\pi$. The density and the cumulative distribution can be expressed also in terms of hyperbolic functions

$$f_{\text{logist}}(x) \;=\; \frac{1}{4b} \operatorname{sech}^2\left(\frac{x-\mu}{2b}\right) \quad \text{and} \quad F_{\text{logist}}(x) \;=\; \frac{1}{2} \,+\, \frac{1}{2} \tanh\left(\frac{x-\mu}{2b}\right) \; .$$

The logistic distribution resembles the normal distribution and like Student's distribution the logistic distribution has heavier tails and a lower maximum than the normal distribution. The entropy takes on the simple form

$$H(f_{\text{logist}}) \;=\; \log b \,+\, 2 \; . \tag{2.93}$$

The moment generating of the logistic distribution is

$$M_{\text{logist}}(s) \;=\; \exp(\mu s)\, B(1-bs, 1+bs \; , \tag{2.94}$$

for $|bs| < 1$ and $B(x, y)$ being the Beta function. The characteristic function is

$$\phi_{\text{logist}}(s) \;=\; \frac{\pi b s \exp(\imath\!\!\imath \, \mu s)}{\sinh(\pi b s)} \; . \tag{2.95}$$

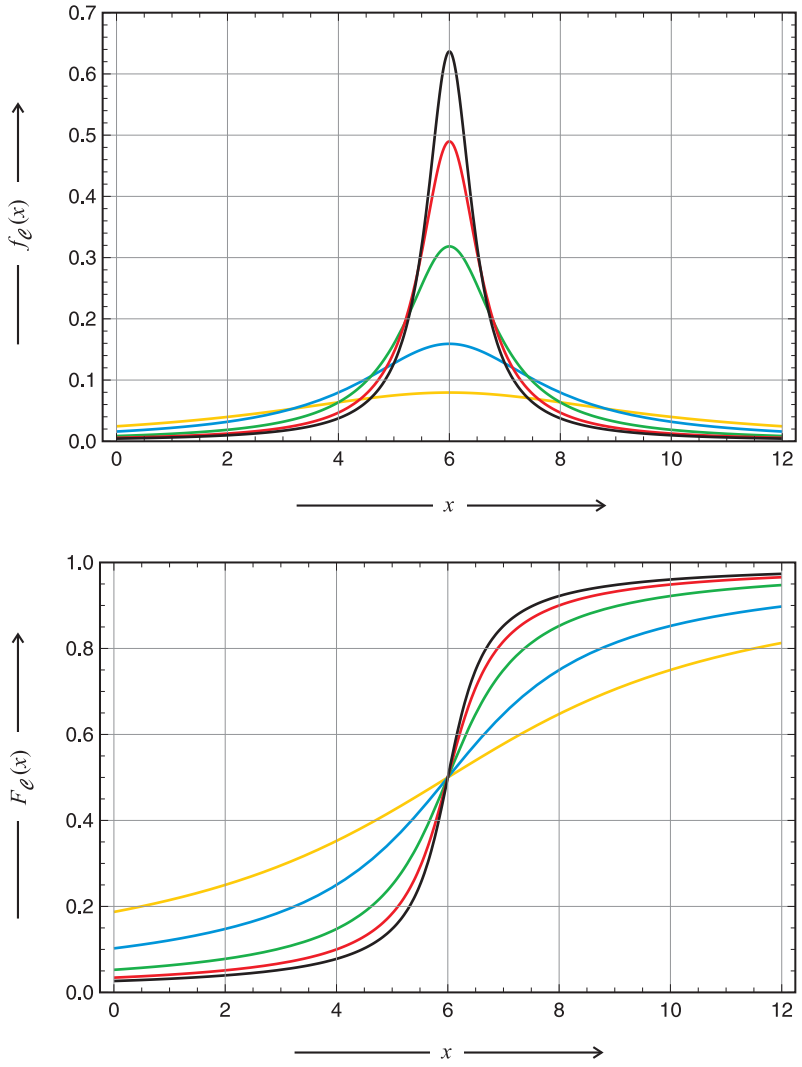

Fig. 2.17 Cauchy-Lorentz density and distribution. In the two plots the Cauchy-Lorentz distribution, $\mathcal{C}(\delta, \gamma)$, is shown in from of the probability density

$$f_{\mathcal{C}}(x) = \gamma \,/\, \Big(\pi\big((x-\delta)^2 + \gamma^2\big)\Big)$$

and the probability distribution

$$F_{\mathcal{C}}(x) = \tfrac{1}{2} \,+\, \arctan\Big((x-\delta)/\gamma\Big) \,/\, \pi\,.$$

Choice of parameters: $\delta = 6$ and $\gamma = 0.5$ (black), 0.65 (red), 1 (green), 2 (blue) and 4 (yellow).

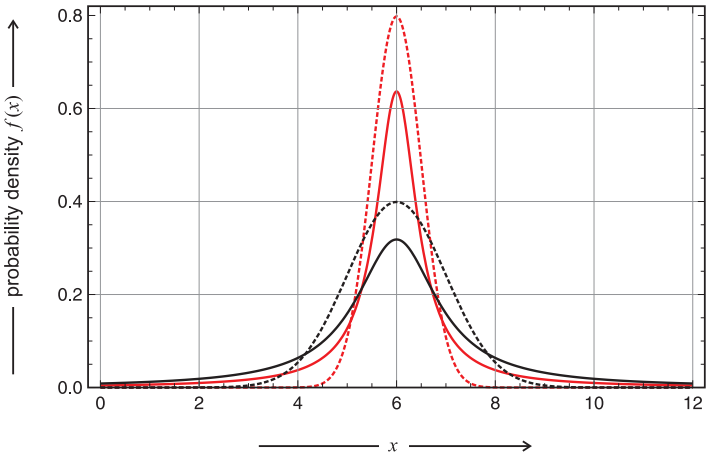

Fig. 2.18 Comparison of Cauchy-Lorentz and normal density. The plots compare the Cauchy-Lorentz density, $\mathcal{C}(\delta, \gamma)$ (full lines), and the normal density $\mathcal{N}(\mu, \sigma^2)$ (broken lines). In the flanking regions the normal density decays to zero much faster than the Cauchy-Lorentz density, and this is the cause of the abnormal behavior of the latter. Choice of parameters: $\delta = \mu = 6$ and $\gamma = \sigma^2 = 0.5$ (black), 1 (red).

2.4.6 The Cauchy-Lorentz distribution

The Cauchy-Lorentz distribution $\mathcal{C}(\gamma, \delta)$ is a continuous distribution with two parameters, the position δ and the scale γ. It is named after the French mathematician Augustin Louis Cauchy and the Dutch physicist Hendrik Antoon Lorentz and is important in mathematics and in particular in physics where it occurs as the solution to the differential equation for forced resonance. In spectroscopy the Lorentz curve is used for the description of spectral lines that are homogeneously broadened. The Cauchy distribution is a typical heavy-tailed distribution in the sense that larger values of the random variable are more likely to occur in the right tail than in the exponential distribution. Heavy-tailed distributions may also have heavy left tails, or both tails may be heavy as in the Cauchy distribution. As we shall see in section 3.2.4 the Cauchy distribution is a stable distribution and can be partitioned into a sum of Cauchy distributions.

The Cauchy probability density function and the cumulative probability distribution are of the form (figure 2.17)

$$\begin{aligned} \text{pdf}: \quad f_{\mathcal{C}}(x) &= \frac{1}{\pi\,\gamma} \cdot \frac{1}{1+\left(\frac{x-\delta}{\gamma}\right)^2} = \\ &= \frac{1}{\pi} \cdot \frac{\gamma}{(x-\delta)^2+\gamma^2} \; x \in \mathbb{R} \;\text{ and} \qquad (2.96) \\ \text{cdf}: \quad F_{\mathcal{C}}(x) &= \frac{1}{2} + \frac{1}{\pi}\,\arctan\left(\frac{x-\delta}{\gamma}\right) . \end{aligned}$$

The two parameters define the position of the peak, δ, and the width of the distribution, γ (figure 2.17). The peak height or amplitude is $1/(\pi\gamma)$. The function $F_{\mathcal{C}}(x)$ can be inverted

$$F_{\mathcal{C}}^{-1}(p) \;=\; \delta \,+\, \gamma\,\tan\Big(\pi\big(p-\frac{1}{2}\big)\Big) \qquad (2.96')$$

and we obtain for the quartiles and the median the values: $(\vartheta-\gamma, \vartheta, \vartheta+\gamma)$. As with the normal distribution we define a standard Cauchy distribution $\mathcal{C}(\delta,\gamma)$ with $\delta=0$ and $\gamma=1$, which is identical with Student's t-distribution with one degree of freedom, $r=1$ (section 2.4.3). Another remarkable relation concerns the ratio between two independent normally distributed random variables that fulfils a standard Cauchy distribution: $\mathcal{N}_1(0,1)/\mathcal{N}_2(0,1) = \mathcal{C}(0,1)$.

Compared to the normal distribution the Cauchy distribution has heavier tails and accordingly a lower maximum (figure 2.18. In this case we cannot use the (excess) kurtosis as an indicator because all moments of the Cauchy distribution are undefined, but we can compute and compare the heights of the standard densities: $f_{\mathcal{C}}(x=\delta) = \frac{1}{\pi}\cdot\frac{1}{\gamma}$ and $f_{\mathcal{N}}(x=\mu) = \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sigma}$, which yields $f_{\mathcal{C}}(\delta) = \frac{1}{\pi}$ and $f_{\mathcal{N}}(\mu) = \frac{1}{\sqrt{2\pi}}$ for $\gamma=\sigma=1$ with $\frac{1}{\pi} < \frac{1}{\sqrt{2\pi}}$. □The Cauchy distribution has, nevertheless, a defined median and mode, which both coincide with the position of the maximum of the density function, $x=\delta$.

The entropy of the Cauchy density is: $H(f_{\mathcal{C}(\delta,\gamma)}) = \log\gamma + \log 4$. It cannot be compared with the entropy of the normal distribution in the sense of the maximum entropy principle (section 2.1.3), because this principle refers to distributions with variance σ^2 whereas the variance of the Cauchy distribution is undefined.

The Cauchy distribution has no moment generating function but a characteristic function:

$$\phi_{\mathcal{C}}(s) \;=\; \exp(\dot{\imath}\,\delta\,s \,-\, \gamma\,|s|) \; . \qquad (2.97)$$

A consequence of the lack of defined moments is that the central limit theorem cannot be applied to a sequence of Cauchy variables. It is can be shown by means of the characteristic function that the mean of a sequence of independent and identically distributed random variables with standard Cauchy distribution, $\mathcal{S} = \sum_{i=1}^{n} \mathcal{X}_i/n$ has the sam standard Cauchy distribution and is not normally distributed as the central limit theorem predicts.

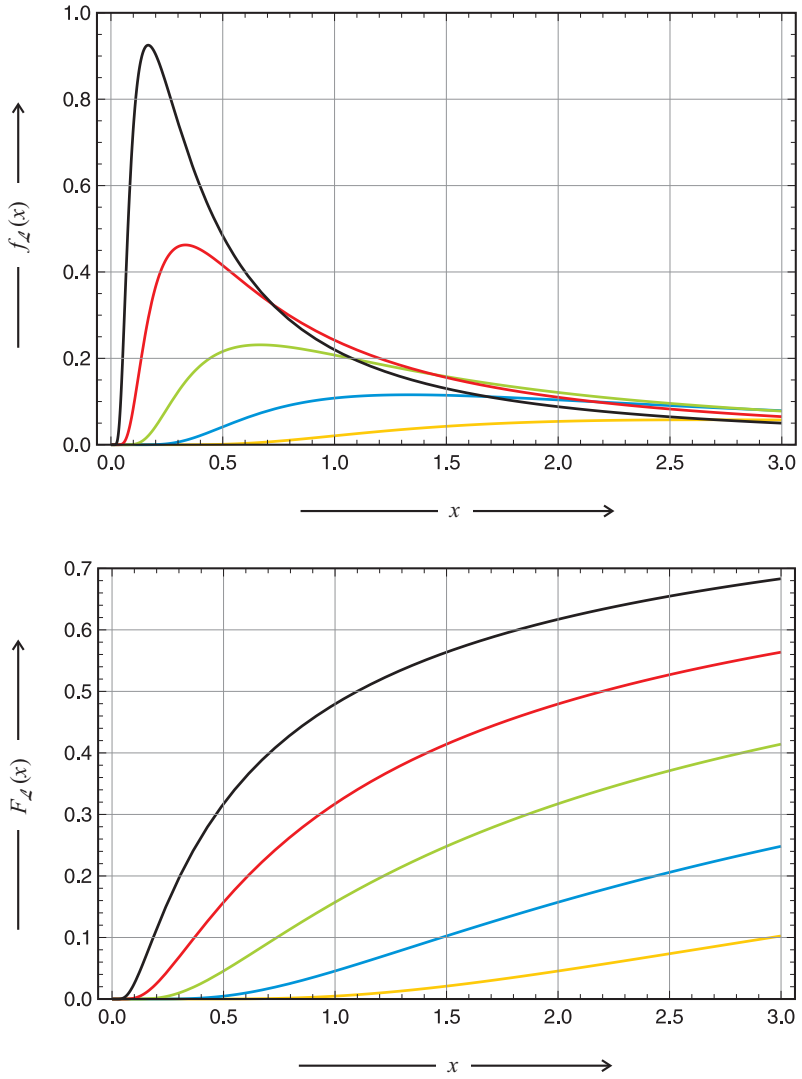

Fig. 2.19 Lévy density and distribution. In the two plots the Lévy distribution, $\mathcal{L}(\delta, \gamma)$, is shown in from of the probability density

$$f_{\mathcal{L}}(x) = \sqrt{\frac{\gamma}{2\pi}} \exp\left(-\frac{\gamma}{2(x-\delta)}\right) \,/\, (x-\delta)^{3/2}$$

and the probability distribution

$$F_{\mathcal{L}}(x) = \operatorname{erfc}\left(\sqrt{\frac{\gamma}{2(x-\delta)}}\right).$$

Choice of parameters: $\delta = 0$ and $c = 0.5$ (black), 1 (red), 2 (green), 4 (blue) and 8 (yellow).

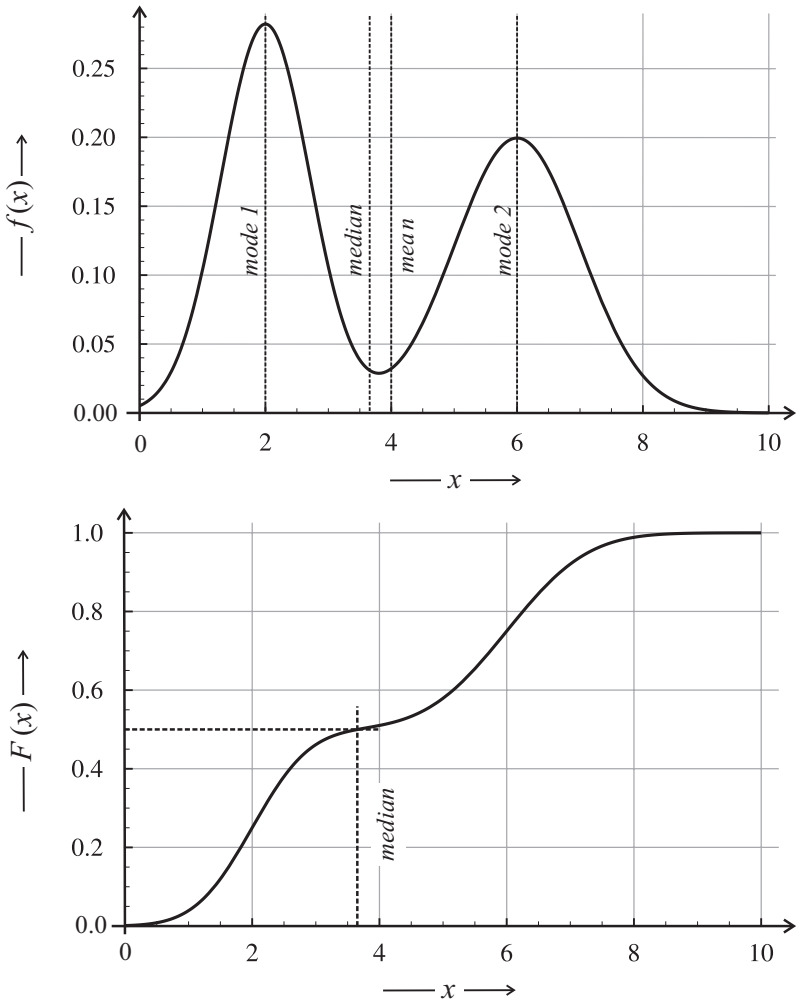

Fig. 2.20 A bimodal probability density. The figure illustrates a bimodal distribution modeled as a superposition of two normal distributions (2.100) with $\alpha = 1/2$ and different values for mean and variance ($\nu_1 = 2, \sigma_1^2 = 1/2$) and ($\nu_2 = 6, \sigma_2^2 = 1$): $f(x) = (\sqrt{2}e^{-(x-2)^2} + e^{-(x-6)^2/2}) \;/\; (2\sqrt{2\pi})$. The upper part shows the probability density corresponding to the two modes $\tilde{\mu}_1 = \nu_1 = 2$ and $\tilde{\mu}_2 = \nu_2 = 6$. Median $\bar{\mu} = 3.65685$ and mean $\mu = 4$ are situated near the density minimum between the two maxima. The lower part presents the cumulative probability distribution, $F(x) = \frac{1}{4}\Big(2 + \mathrm{erf}\big(x-2\big) + \mathrm{erf}\Big(\frac{x-6}{\sqrt{2}}\Big)\Big)$, as well as the construction of the median. The variances in this example are: $\hat{\mu}_2 = 20.75$ and $\mu_2 = 4.75$.

2.4.7 The Lévy distribution

The Lévy distribution $\mathcal{L}(\gamma, \delta)$ is a continuous one-sided probability distribution, which is defined for values of the variable x that are larger or equal a shift parameter δ: $x \in [\delta, \infty[$. It is a special case of the *inverse gamma distribution* and belongs together with the normal and the Cauchy distribution the class of analytically accessible *stable distributions*.

The Lévy probability density function and the cumulative probability distribution are of the form (figure 2.19)

$$\begin{aligned} \text{pdf}: \quad f_{\mathcal{L}}(x) &= \sqrt{\frac{\gamma}{2\pi}} \cdot \frac{1}{(x-\delta)^{3/2}} \exp\left(-\frac{\gamma}{2(x-\delta)}\right), \; x \in [\delta, \infty[, \quad \text{and} \\ \text{cdf}: \quad F_{\mathcal{L}}(x) &= \operatorname{erfc}\left(\sqrt{\frac{\gamma}{2(x-\delta)}}\right). \end{aligned} \tag{2.98}$$

The two parameters $\delta \in \mathbb{R}$ and $\gamma \in \mathbb{R}_{>0}$ are the location of $f_{\mathcal{L}}(x) = 0$ and the scale parameter. Mean and variance of the Lévy distribution are infinite, skewness and kurtosis undetermined. For $\delta = 0$ the mode of the distribution appears at $\tilde{\mu} = \gamma/3$ and the median takes on the value $\bar{\mu} = \gamma/\big(2(\operatorname{erfc}^{-1}(1/2))^2\big)$.

The entropy of the Lévy distribution is

$$H\big(f_{\mathcal{L}}(x)\big) = \frac{1 + 3\,\gamma + \ln(16\pi\gamma^2)}{2} \quad \text{with } \gamma \text{ being Euler's constant,}$$

and the characteristic function

$$\phi_{\mathcal{L}}(s) = \exp(\dot{u}\,\delta\,s - \sqrt{-2\dot{u}\,\gamma\,s}), \tag{2.99}$$

is the only defined generating function.

2.4.8 Bimodal distributions

As the name of the bimodal distribution indicates that the density function $f(x)$ has two maxima. It arises commonly as a mixture of two unimodal distribution in the sense that the bimodally distributed random variable $\mathcal{X}$ is defined as

$$\text{Prob}\,(\mathcal{X}) \;=\; \begin{cases} P(\mathcal{X} \,=\, \mathcal{Y}_1) \;=\; \alpha \;\text{ and} \\ P(\mathcal{X} \,=\, \mathcal{Y}_2) \;=\; (1-\alpha)\,. \end{cases}$$

Bimodal distributions commonly arise from statistics of populations that are split into two subpopulations with sufficiently different properties. The sizes of weaver ants give rise to a bimodal distributions because of the existence of two classes of workers [301]. In case the differences are too small as in case of the combined distribution of body heights for men and women monomodality is observed [253].

As an illustrative model we choose the superposition of two normal distributions with different means and variances (figure 2.20). The probability density for $\alpha = 1/2$ is then of the form:

$$f(x) \;=\; \frac{1}{2\sqrt{2\pi}}\Big(e^{\frac{-(x-\nu_1)^2}{2\sigma_1^2}}\Big/\sqrt{\sigma_1^2} \,+\, e^{\frac{-(x-\nu_2)^2}{2\sigma_2^2}}\Big/\sqrt{\sigma_2^2}\Big)\,. \tag{2.100}$$

The cumulative distribution function is readily obtained by integration. As in the case of the normal distribution the result is not analytical but formulated in terms of the error function, which is available only numerically through integration:

$$F(x) \;=\; \frac{1}{4}\left(2 \,+\, \text{erf}\left(\frac{x-\nu_1}{\sqrt{2\sigma_1^2}}\right) \,+\, \text{erf}\left(\frac{x-\nu_2}{\sqrt{2\sigma_2^2}}\right)\right)\,. \tag{2.101}$$

In the numerical example shown in figure 2.20 the distribution function shows two distinct steps corresponding to the maxima of the density $f(x)$.

As an exercise first an second moments of the bimodal distribution can be readily computed analytically. The results are:

$$\hat{\mu}_1 = \mu = \frac{1}{2}\,(\nu_1 + \nu_2)\,,\;\; \mu_1 = 0 \;\text{ and}$$

$$\hat{\mu}_2 = \frac{1}{2}\,(\nu_1^2 + \nu_2^2) + \frac{1}{2}\,(\sigma_1^2 + \sigma_2^2)\,,\;\; \mu_2 = \frac{1}{4}\,(\nu_1 - \nu_2)^2 + \frac{1}{2}\,(\sigma_1^2 + \sigma_2^2)\,.$$

The centered second moment illustrates the contributions to the variance of the bimodal density. It is composed of the sum of the variances of the subpopulations and the square of the difference between the two means, $(\nu_1 - \nu_2)^2$.

2.5 Mathematical statistics

Mathematical statistics provides the bridge between probability theory and the analysis of real data, which will always represent incomplete since finite samples. Nevertheless, it turned out very appropriate to use infinite samples as a reference (section 1.3). Large sample theory and in particular the law of large numbers (section 2.3.6) deal with the asymptotic behavior of series of samples with increasing size. Although mathematical statistics is a discipline in its own right and would require a separate course, we mention here briefly only three basic concepts, which is of general importance for every scientist.[23]

First we shall be concerned with approximations to moments derived from finite samples. In practice, we can collect data for all sample points of the sample space Ω only in very exceptional cases. Otherwise exhaustive measurements are impossible and we have to rely on limited samples as they are obtained in physics through experiments or in sociology through opinion polls. As an example for the evaluation of the justification of assumptions we introduce Pearson's chi-squared test and finally we illustrate statistical inference by means of an example applying Bayes' theorem.

2.5.1 Sample moments

As we did before for complete sample spaces, we evaluate functions Z from incomplete random samples $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ and obtain as output random variables $\mathcal{Z} = Z(\mathcal{X}_1, \ldots, \mathcal{X}_n)$. Similarly we compute sample expectation values, also called sample means, sample variances, sample standard deviations and other quantities as estimators from limited sets of data, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. They are calculated in the same way as if the sample set would cover the entire sample space. In particular we compute the *sample mean*

$$m \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.102}$$

and the moments around the sample mean. For the *sample variance* we obtain

$$m_2 \;=\; = \; \frac{1}{n} \sum_{i=1}^{n} x_i^2 \;-\; \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \;, \tag{2.103}$$

[23] For the reader who is interested in more details on mathematical statistics we recommend the classical textbook by the Polish mathematician Marek Fisz [87] and the comprehensive treatise by Stuart and Ord [270, 271], which is a new edition of Kendall's classic on statistics. A text that is useful as an not too elaborate introduction is found in [131], the monograph [38] is particularly addressed to experimentalists practicing statistics, and a great variety of other and equally well suitable texts are, of course, available in the rich literature on mathematical statistics.

and for the third and fourth moments after some calculations

$$m_3 = \frac{1}{n}\sum_{i=1}^{n} x_i^3 - \frac{3}{n^2}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} x_j^2\right) + \frac{2}{n^3}\left(\sum_{i=1}^{n} x_i\right)^3 \tag{2.104a}$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{n} x_i^4 - \frac{4}{n^2}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} x_j^3\right) + \\ + \frac{6}{n^3}\left(\sum_{i=1}^{n} x_i\right)^2\left(\sum_{j=1}^{n} x_j^2\right) - \frac{3}{n^4}\left(\sum_{i=1}^{n} x_i\right)^4 . \tag{2.104b}$$

These naïve estimators, m_i $(i = 2, 3, 4, \ldots)$, contain a bias because the exact expectation value μ around which the moments are centered is not known and has to be approximated by the sample mean m. For the variance we illustrate the systematic deviation by calculating a correction factor known as Bessel's correction but more properly attributed to Carl Friedrich Gauss [156, part 2, p.161]. In order to obtain an expectation value for the sample moments we repeat drawing of samples with n elements and denote their expectation values by $< m_i >$.[24] In particular we have

$$\begin{aligned} m_2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i^2 + \sum_{i,j=1,\, i\neq j}^{n} x_i\, x_j\right) = \\ &= \frac{n-1}{n^2}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\sum_{i,j=1,\, i\neq j}^{n} x_i\, x_j \ . \end{aligned}$$

The expectation value is now of the form

$$< m_2 > = \frac{n-1}{n}\left\langle \frac{1}{n}\sum_{i=1}^{n} x_i^2 \right\rangle - \frac{1}{n^2}\left\langle \sum_{i,j=1,\, i\neq j}^{n} x_i\, x_j \right\rangle ,$$

and by using $< x_i x_j >=< x_i >< x_j >=< x_i >^2$ we find

[24] It is important to note that $< m_i >$ is the expectation value of an average over a finite sample, where the expectation value refers to the entire sample space. In particular, we find

$$< m >= \left\langle \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i \right\rangle = \mu = \alpha_1 \ ,$$

where μ is the first (raw) moment. For the higher moments the situation is more complicated and requires some care (see text).

$$< m_2 > \;=\; \frac{n-1}{n}\left\langle \frac{1}{n}\sum_{i=1}^{n} x_i^2 \right\rangle - \frac{n(n-1)}{n^2}\left\langle \sum_{i=1}^{n} x_i \right\rangle^2 \;=$$

$$= \frac{n-1}{n}\,\alpha_2 - \frac{n(n-1)}{n^2}\,\mu^2 \;=\; \frac{n-1}{n}\,(\alpha_2 - \mu^2)\;,$$

where $\alpha_2 = \hat{\mu}_2$ is the second raw moment or second moment about zero. Using the identity $\alpha_2 = \mu_2 + \mu^2$ we find eventually

$$< m_2 > = \frac{n-1}{n}\,\mu_2 \;\text{ and }\; \mathrm{var}(\mathbf{x}) \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2\;. \tag{2.105}$$

Further useful measures of correlation between pairs of random variables can be derived straightforwardly: (i) the unbiased *sample covariance*

$$\mathcal{M}_{\mathcal{XY}} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)\,(y_i - m)\;, \tag{2.106}$$

and (ii) the *sample correlation coefficient*

$$\mathcal{R}_{\mathcal{XY}} \;=\; \frac{\sum_{i=1}^{n}(x_i - m)\,(y_i - m)}{\sqrt{\left(\sum_{i=1}^{n}(x_i - m)^2\right)\left(\sum_{i=1}^{n}(y_i - m)^2\right)}}\;. \tag{2.107}$$

For practical purposes Bessel's correction is often unimportant when the data sets are sufficiently large but the recognition of the principle is important in particular for statistical properties more involved than variances. Sometimes a problem is encountered in cases where the second moment of a distribution, μ_2, does not exist, which means it diverges. Then, computing variances from incomplete data sets is also unstable and one may choose the *mean absolute deviation*,

$$\mathcal{D}(\mathcal{X}) \;=\; \frac{1}{n}\sum_{i=1}^{n}|\mathcal{X}_i - m|\;, \tag{2.108}$$

as a measure for the width of the distribution [244, pp.455-459], because it is commonly more robust than variance or standard deviation.

Ronald Fisher conceived k-statistics in order to derive estimators for the moments of finite samples [82]. The cumulants of a probability distribution are derived as expectation values, $< k_i >= \kappa_i$, of finite set cumulants calculated in the same way as the complete sample set analogues [157, pp.99-100]. The first four terms of k-statistics for n sample points are

$$\begin{aligned} k_1 &= m \ , \\ k_2 &= \frac{n}{n-1}\, m_2 \ , \\ k_3 &= \frac{n^2}{(n-1)(n-2)}\, m_3 \ \text{ and} \\ k_4 &= \frac{n^2 \left((n+1)\, m_4 \ - \ 3\,(n-1)\, {m_2}^2 \right)}{(n-1)(n-2)(n-3)} \ , \end{aligned} \tag{2.109}$$

which can be derived by inversion of the well known relationships

$$\begin{aligned} < m > &= \mu \ , \\ < m_2 > &= \frac{n-1}{n}\, \mu_2 \ , \\ < m_3 > &= \frac{(n-1)(n-2)}{n^2}\, \mu_3 \ , \\ < m_2^2 > &= \frac{(n-1) \Big((n-1)\, \mu_4 \ + \ (n^2 - 2n + 3)\, {\mu_2}^2 \Big)}{n^3} \ , \ \text{ and} \\ < m_4 > &= \frac{(n-1) \Big((n^2 - 3n + 3)\, \mu_4 \ + \ 3\,(2n-3)\, {\mu_2}^2 \Big)}{n^3} \ . \end{aligned} \tag{2.110}$$

The usefulness of these relations becomes evident in various applications.

The statistician computes moments and other functions from his empirical data sets, which is almost always non-exhaustive, for example $\{x_1, \ldots, x_n\}$ or $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ by means of the equations (2.102) and (2.105) to (2.107). The underlying assumption, of course, is that the values of the empirical functions converge to the corresponding exact moments as the random sample increases and the theoretical basis for this assumption is provided by the law of large numbers.

2.5.2 Pearson's chi-squared test

The main issue of mathematical statistics, however, is not so much to compute approximations to the moments but and has always been and still is the development of independent tests that allow for the derivation of information on the appropriateness of models and the quality of data. Predictions on the reliability of the computed values are made by means of a great variety of tools. We dispense from details, which are extensively treated in the literature [88, 270, 271]. Karl Pearson conceived a test in 1900 [236], which became popular under the name chi-squared test. This test has also been used by Ronald Fisher when he analyzed Gregor Mendel's data on genetics of the garden pea *pisum sativum* and we shall use the data given in table 1.2 to illustrate the application of the chi-squared test.

The formula of Pearson's test is made plausible by means of a simple example [132, pp. 407-414]: A random variable $\mathcal{Y}_1$ is binomially distributed according to $B_k(n, p_1)$ with expectation value $E(\mathcal{Y}_1) = n\,p_1$ and variance $\sigma_1^2 = n\,p_1(1-p_1)$ (section 2.3.2) and then, following the central limit theorem the random variable

$$\mathcal{Z} \;=\; \frac{\mathcal{Y}_1 \;-\; np_1}{\sqrt{np_1(1-p_1)}}$$

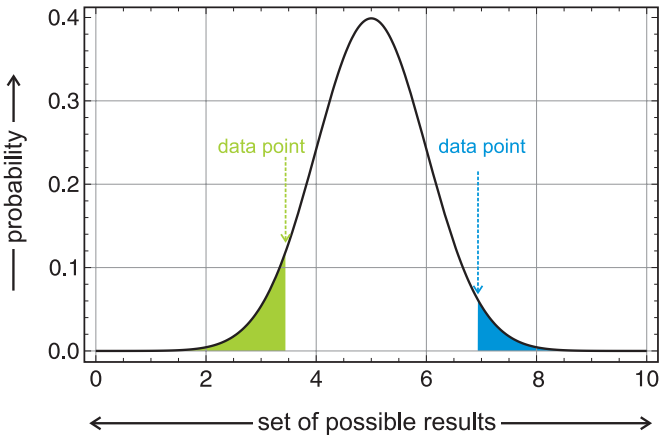

Fig. 2.21 The p-value in significance test of null hypothesis. The figure shows the definition of the p-value. The bell-shaped curve is the probability density function (PDF) of possible results. Two specific data points are shown one at values above the most frequent outcome at $x = 5$ near $x = 7$ (blue) and the other one at $x \approx 3.5$ (green). The p-value – not to be mistaken for a score – is the cumulative probability of more extreme cases, i.e., results that are further away of the most frequent outcome than the data point and obtained as the integral under the PDF. Depending of the position of the observed result this integral has to be taken to higher (blue) or lower (green) values of x, respectively.

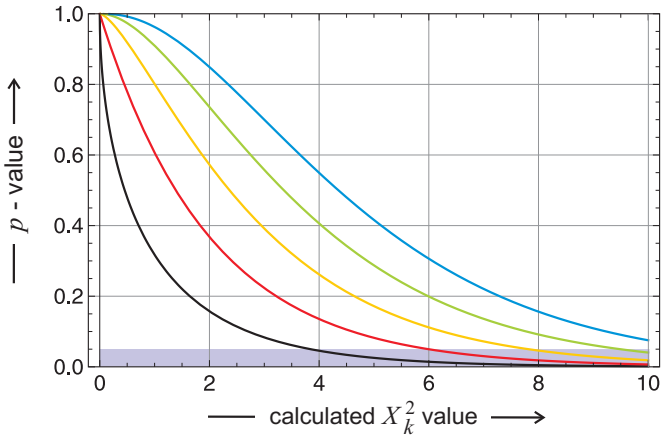

Fig. 2.22 Calculation of the p-value in significance test of null hypothesis. The figure shows the p-values from equation (2.116) as a function of the calculated values of X_k^2 for the k-values 1 (black), 2 (red), 3 (yellow), 4 (green), and 5 (blue). The highlighted area at the bottom of the figure shows the range where the null hypothesis is rejected.

has a standardized binomial distribution, which approximates $\mathcal{N}(0,1)$ for sufficiently large n (section 2.3.5). A second random variable is $\mathcal{Y}_2 = n - \mathcal{Y}_1$ with expectation value $E(\mathcal{Y}_2) = n\,p_2$ and variance $\sigma_2^2 = \sigma_1^2 = n\,p_2(1-p_2) = n\,p_1(1-p_1)$, since $p_2 = (1-p_1)$. The sum $\mathcal{Z}^2 = \mathcal{Y}_1^2 + \mathcal{Y}_2^2$ is approximately χ^2 distributed:

$$\mathcal{Z}^2 \;=\; \frac{(\mathcal{Y}_1 - np_1)^2}{np_1(1-p_1)} \;=\; \frac{(\mathcal{Y}_1 - np_1)^2}{np_1} + \frac{(\mathcal{Y}_2 - np_2)^2}{np_2} \quad \text{since}$$

$$(\mathcal{Y}_1 - np_1)^2 \;=\; \Big(n \,-\, \mathcal{Y}_1 \,-\, n(1-p_1)\Big)^2 \;=\; (\mathcal{Y}_2 - np_2)^2 \;.$$

We can now rewrite the expression by introducing the expectation values

$$\mathcal{Q}_1 \;=\; \sum_{i=1}^{2} \frac{\Big(\mathcal{Y}_i - E(Y_i)\Big)^2}{E(\mathcal{Y}_i)} \;,$$

and indicating the number of independent random variables as a subscript. Provided all products $n\,p_i$ are sufficiently large – a conservative estimate would be $np_i \geq 5\,\forall\, i$ – the quantity $\mathcal{Q}_1$ has an approximate chi-squared distribution with one degree of freedom: χ_1^2.

The generalization to an experiment with k mutually exclusive and exhaustive outcomes $A_1, A_2, \ldots, A_k$ of the variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_k$, is straightforward. All variables $\mathcal{X}_i$ are assumed to have finite mean μ_i and finite variance σ_i^2 such that central limit theorem applies and the distribution for large n

converges to the normal distribution $\mathcal{N}(0,1)$. We define the probability to obtain the result A_i by $P(A_i) = p_i$ and by conservation of probabilities we have $\sum_{i=1}^{k} p_i = 1$. One variable is thus lacking independence and we choose it to be $\mathcal{X}_k$:

$$\mathcal{X}_k \;=\; n \,-\, \sum_{i=1}^{k-1} \mathcal{X}_i \;, \tag{2.111}$$

then the joint distribution of of $k-1$ variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{k-1}$ has the joint probability mass function (pmf)

$$f(x_1, x_2, \ldots, x_{k-1}) \;=\; P(\mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2, \ldots, \mathcal{X}_{k-1} = x_{k-1}) \;.$$

Next we consider n independent trials, x_1 times yielding A_1, x_2 times yielding $A_2, \ldots$,x_k times yielding A_k, where the particular outcome has a probability

$$P(\mathcal{X}_1 = x_1, \mathcal{X}_1 = x_1, \ldots, \mathcal{X}_{k-1} = x_{k-1}) \;=\; p_1^{x_1} \cdot p_2^{x_2} \cdot \ldots \cdot p_k^{x_k} \quad \text{with}$$

the frequency factor or statistical weight

$$g(x_1, x_2, \ldots, x_k) \;=\; \binom{n}{x_1, x_2, \ldots, x_k} \;=\; \frac{n!}{x_1!, x_2!, \ldots, x_k!} \;,$$

and eventually we find for the pmf

$$\begin{aligned} f(x_1, x_2, \ldots, x_{k-1}) \;&=\; g(x_1, x_2, \ldots, x_k) \cdot P(\mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2, \ldots, \mathcal{X}_{k-1} = x_{k-1}) \;= \\ &=\; \frac{n!}{x_1!, x_2!, \ldots, x_k!}\, p_1^{x_1} \cdot p_2^{x_2} \cdot \ldots \cdot p_k^{x_k} \;, \end{aligned} \tag{2.112}$$

with the two restrictions $y_k = n - \sum_{i=1}^{k-1} x_i$ and $p_k = 1 - \sum_{i=1}^{k-1} p_i$. Pearson's construction follows the lines we have shown for the binomial with $k = 2$ and yields under consideration of equation 2.111:

$$\mathcal{Q}_{k-1} \;=\; X_{k-1}^2 \;=\sum_{i=1}^{k} \frac{\Big(\mathcal{X}_i \,-\, E(\mathcal{X}_i)\Big)^2}{E(\mathcal{X}_i)} \;. \tag{2.113}$$

The sum of squares $\mathcal{Q}_{k-1}$ in (2.113) is called Pearson's *cumulative test statistic*. It has an approximate chi-squared distribution with k-1 degrees of freedom, χ^2_{k-1},[25] and again if n is sufficiently large to fulfil $n\,p_1 \geq 5\;\forall\, i$ the distributions are close enough for most practical purposes.

In order to be able to test hypotheses we divide our sample space into k cells and record observations falling into individual cells. In essence, these cells $\mathcal{C}_i$ are tantamount to the outcomes A_i but we can define them to be completely general, for example collecting all instances that falls in a certain

[25] We indicate the expected converge in the sense of the central limit theorem by choosing the symbol X^2_{k-1} for the finite n expression.

range. At the end of the registration period the number of observations is n and the number of those that were falling into the cell $\mathcal{C}_i$ is ν_i with $\sum_{i=1}^{k} \nu_i = n$. Equation (2.113) is now applied to test a (null) hypothesis H_0 against empirically recorded values for all outcomes,

$$H_0 : \quad E_i^{(0)}(\mathcal{X}_i) \; = \; \varepsilon_{i0} \, , \; i = 1, \ldots, k \; . \tag{2.114}$$

In other words, the null hypothesis predicts the distribution score values falling into the cells $\mathcal{C}_i$ to be ε_{i0} ; $i = 1, \ldots, k$ and this in the sense of expectation values $E_i^{(0)}$. If the null hypothesis were, for example, the uniform distribution we had $\varepsilon_{i0} = n/k \, \forall \, i = 1, \ldots, k$. The cumulative test statistic X^2 converges to the χ^2 distribution in the limit $n \to \infty$ – just as the mean value of a stochastic variable, $\overline{\mathcal{Z}} = \sum_{i=1}^{n} z_i$ converges to the expectation value $\lim_{n\to\infty} \overline{\mathcal{Z}} = E\{\mathcal{Z}\}$. This implies that X^2 is never exactly equal to χ^2 and the approximation that will always become better when the sample size is increased. Usually a lower limit is defined for the number of entries in the cells to be considered, values between 5 and 10 are common.

If the null hypotheses H_0 were true, ν_i and ε_{i0} should be approximately equal. Thus we expect the deviation expressed by

$$X_d^2 \; = \; \sum_{i=1}^{k} \frac{(\nu_i \, - \, \varepsilon_{i0})^2}{\varepsilon_{i0}} \; \approx \; \chi_d^2 \tag{2.115}$$

should be small if H_0 is acceptable. Otherwise, we shall reject H_0 if the deviation is too large: $X_d^2 \geq \chi_d^2(\alpha)$, where α is the predefined level of significance for the test. Two quantities are still undefined (i) the degree of freedom d and (ii) the significance level α.

Next the number of degrees of freedom d of the theoretical distribution to which the data are fitted has to be determined. The number of cells, k, represents the maximal number of degrees of freedom, which is reduced by one because of the conservation relation discussed above: $d = k - 1$. The dimension d is reduced further when parameters are needed in fitting the distribution of the null hypothesis. If the number of such parameters is s we get $d = k - 1 - s$. Choosing the uniform distribution $\mathcal{U}$ that is parameter free we find $d = n - 1$.

The significance of the null hypothesis for a given set of data is commonly tested by means of the so-called p-value: For $p < \alpha$ the null hypothesis is rejected. Precisely, the p-value is the probability of obtaining a test statistic that is at least as extreme as the actually observed one under the assumption that the null hypothesis is true. We call a probability $P(A)$ *more extreme* than $P(B)$ if A is less likely to occur under the null hypothesis as B. As shown in figure 2.21 this probability is calculated as the integral below the probability density function from the calculated X_d^2-value to $+\infty$. For the χ_d^2 distribution we have

$$p \;=\; \int_{X_d^2}^{+\infty} \chi_d^2(x)\,\mathrm{d}x \;=\; 1 \;-\; \int_0^{X_d^2} \chi_d^2(x)\,\mathrm{d}x \;= 1 \;-\; F(X^2;d)\;, \qquad (2.116)$$

which involves the cumulative distribution function $F(x;d)$ defined in equation (2.66). Commonly, the null hypothesis is rejected when p is smaller than the significance level: $p < \alpha$ with $0.02 \leq \alpha \leq 0.05$. If the condition $p < \alpha$ is fulfilled one says the null hypothesis is statistically significantly rejected. In other words, the null hypothesis is statistically significant or statistically confirmed in the range $\alpha \leq p \leq 1$.

A simple example is used for the purpose of illustration: Two random samples of N animals were drawn from a population, ν_1 were males and ν_2 were females with $\nu_1 + \nu_2 = n$. The first sample $\nu_1 = 170, \nu_2 = 152$,

$$n = 322,\ \nu_1 = 170,\ \nu_2 = 152 : X_1^2 = \frac{(170-161)^2 + (152-161)^2}{322} = 0.503\,,$$
$$p = 1 - F(0.503;1) = 0.478\,,$$

clearly supports the null hypothesis that that males and females are equally frequent since $p > \alpha \approx 0.05$. The second sample $\nu_1 = 207, \nu_2 = 260$,

$$n = 467,\ \nu_1 = 207,\ \nu_2 = 260 : X_1^2 = \frac{(207-233.5)^2 + (260-233.5)^2}{233.5} = 6.015\,,$$
$$p = 1 - F(6.015;1) = 0.0142\,,$$

leads to a p-value, which is below the critical limit of significance and the rejection of the null hypothesis, the numbers of males and females are equal, is statistically significant or there is very likely another reason than random fluctuation responsible for the difference.

As a second example we test Gregor Mendel's experimental data on the garden pea, *pisum sativum*, given in table 1.2. Here the null hypothesis to be tested is the ratio between phenotypic features developed by genotypes. We consider two features: (i) the shape, roundish and wrinkled, and (ii) the color of seeds, yellow and green, which are determined by two independent loci and two alleles each, A and a or B and b, respectively. The two alleles form four diploid genotypes, AA, Aa, and aA, aa, or, BB, Bb, and bB, bb. Since the alleles a and b are *recessive* only the the genotypes aa or bb develop the second phenotype, wrinkled and green, respectively, and based on the null hypothesis of a uniform distribution of genotypes we expect a 3:1 ratio of phenotypes. In table 2.3 we apply Pearson's chi-square hypothesis to the null hypothesis of 3:1 ratios for the phenotypes roundish and wrinkled or yellow and green. As examples we have chosen the total sample of Mendel's experiments as well as three plants ('1', '5', and '8') in table 1.2) being typical ('1') or showing extreme ratios ('5' having the best and the worst value for shape and color, respectively, and '8' showing the largest ratio, 4.89). All

Table 2.3 Pearson χ^2-test of Gregor Mendel's experiments with the garden pea (*pisum sativum*)**.** The total results as well as the data for three selected plants are analyzed by means of Karl Pearson's chi-square statistics. Two characteristic features of the seeds are reported: the shape, roundish or angular wrinkled, and the color, yellow or green. The phenotypes of the two dominant alleles are: A = *round* and B = *yellow*. The recessive phenotypes are a = *wrinkled* and b = *green*The data are taken from table 1.2.

Property	Sample space	Number of seeds		χ^2-statistics	
		A/B	a/b	X_1^2	p
shape (A,a)	total	5 474	1 850	0.2629	0.6081
color (B,b)	total	6 022	2 001	0.0150	0.9025
shape (A,a)	plant 1	45	12	0.4737	0.4913
color (B,b)	plant 1	25	11	0.5926	0.4414
shape (A,a)	plant 5	32	11	0.00775	0.9298
color (B,b)	plant 5	24	13	2.0405	0.1532
shape (A,a)	plant 8	22	10	0.6667	0.4142
color (B,b)	plant 8	44	9	1.8176	0.1776

p-values in this table are well above the critical limit and without further discussion required confirm the 3:1 ratio.[26]

The test of independence is relevant for situations when an observation registers two outcomes and the null hypothesis is that these outcomes are statistically independent. Each observation is allocated to one cell of a two-dimensional array of cells called a *contingency table* (see next section 2.5.3). In the general case there are m rows and n columns in a table. Then, the theoretical frequency for a cell under the null hypothesis of independence is

$$\varepsilon_{ij} \;=\; \frac{\sum_{k=1}^{n} \nu_{ik} \sum_{k=1}^{m} \nu_{kj}}{N} \;, \tag{2.117}$$

where N is the (grand) total sample size or the sum of all cells in the table. The value of the X^2 test-statistic is

$$X^2 \;=\; \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(\nu_{ij} - \varepsilon_{ij})^2}{\varepsilon_{ij}} \;. \tag{2.118}$$

Fitting the model of independence reduces the number of degrees of freedom by $\pi = m + n - 1$. Originally the number of degrees of freedom is equal to the

[26] We should remember that the claim of Ronald Fisher and others had been that Mendel's data are too good to be true.

number of cells, $m \cdot n$, and after reduction by π we have $d = (m-1) \cdot (n-1)$ degrees of freedom for comparison with the χ^2 distribution. The p-value is again obtained by insertion into the cumulative distribution function (cdf), $p = 1 - F(X^2; d)$, and a value of p less than a predefined critical value, commonly $p < \alpha = 0.05$, is considered as justification for rejection of the null hypothesis or in other words the row variable does not appear to be independent of the column variable.

2.5.3 Fisher's exact test

The second example out of many statistical significance test developed in mathematical statistics we mention here is *Fisher's exact test* for the analysis of contingency tables. In contrast to the χ^2-test Fisher's test is valid for all sample sizes and not only for sufficiently large samples. We begin by defining a contingency table, which in general is a $m \times n$ matrix M where all possible outcomes of one variable x enter the columns in one row and distribution of outcomes of the second variable y is contained in the columns for a given row. The most common case – and the one that is most easily analyzed – is 2×2, two variables with two values each. The the contingency table has the form

	x_1	x_2	total
y_1	a	b	$a+b$
y_2	c	d	$c+d$
total	$a+c$	$b+d$	N

where every variable, x and y, has two outcomes and $N = a+b+c+d$ is the grand total. Fisher's contribution was to prove that the probability to obtain the set of values (x_1, x_2, y_1, y_2) is given by the hypergeometric distribution

$$\begin{aligned} \text{probability mass function} \quad & f_{\mu,\nu}(k) = \frac{\binom{\mu}{k}\binom{N-\mu}{\nu-k}}{\binom{N}{\nu}} \ , \\ \text{cumulative density function} \quad & F_{\mu,\nu}(k) = \sum_{i=0}^{k} \frac{\binom{\mu}{k}\binom{N-\mu}{\nu-k}}{\binom{N}{\nu}} \ , \end{aligned} \tag{2.119}$$

where $N \in \mathbb{N} = \{1, 2, \ldots\}$, $\mu \in \{0, 1, \ldots, N\}$, $\nu \in \{1, 2, \ldots, N\}$, and the support $k \in \{\max(0, \nu+\mu-N), \ldots, \min(\mu, \nu)\}$. Translating the contingency table into the notation of probability functions we have: $a \equiv k$, $b \equiv \mu - k$, $c \equiv \nu - k$, and $d \equiv N + k - (\mu + \nu)$ and hence Fisher's result for the p-value of the general 2×2 contingency table is

$$p \;=\; \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}} \;=\; \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,N!}\;, \tag{2.120}$$

where the expression on the rhs shows beautifully the equivalence between rows and columns.

We present the right- or left-handedness of human males or females as an example for the illustration of Fisher's test: A sample consisting of 52 males and 48 females yields 9 left-handed males and 4 left-handed females. Is the difference statistically significant and allows for the conclusion that left-handedness is more common among males than females? The contingency table in this case reads

	x_m	x_f	total
y_r	43	44	87
y_l	9	4	13
total	52	48	100

The calculation yields $p \approx 0.10$ which is above the critical value $0.02 \leq \alpha \leq 0.05$ and $p > \alpha$ confirms the rejection of the assumption that men are more likely to be left-handed for these data.

2.5.4 Bayesian inference

In this section we present a simple but analytically tractable example as an illustration for the application of Bayesian statistics [52], which has been adapted from the original work of Reverend Thomas Bayes posthumous publication in 1763 [245]. More detailed applications of the Bayesian approach can be found in a number of excellent monographs, for example [95].

The example is called table game and is played by two persons, Alice (A) and Bob (B) as well as a third person (C) acting as game master and being *neutral* and a random number generator simulating a uniform distribution of pseudorandom numbers in the range $0 \leq \mathcal{R} < 1$. The pseudorandom number generator is operated by the game master and cannot be seen by the two players. In essence, A and B are completely passive, have no information on the game except the basic setup and know the scores, which are $a(t)$ for A and $b(t)$ for B. The person who reaches a predefined score value, z, first has won. This simple game starts through drawing a pseudorandom number, $\mathcal{R} = r_0$, by the game master. Consecutive drawings yielding numbers r_i assign points to A iff $0 \leq r_i < r_0$ is fulfilled and to B iff $r_0 \leq r_i < 1$ holds. The game is continued until one person, A or B, reaches the score z.

The problem is to compute fair odds of winning for A and B when the game is terminated premature, and r_0 is unknown. Let us assume that the scores at the time of termination were: $a(t) = a$ and $b(t) = b$ with $a < z$

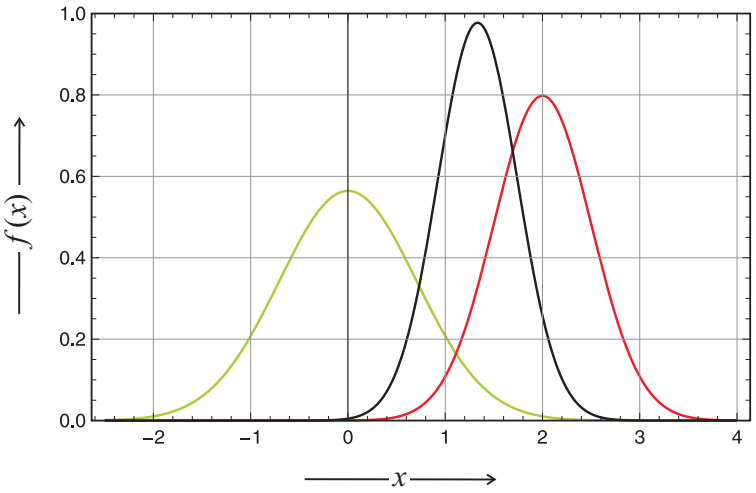

Fig. 2.23 The Bayesian method of inference. The figure sketches the Bayesian method by means of normal density functions. The sample data are given in form of the likelihood function ($P(\mathcal{Y}|\mathcal{X}) = \mathcal{N}(2, \frac{1}{2})$, red) and additional external information on the parameters enters the analysis as prior distribution ($P(\mathcal{X}) = \mathcal{N}(0, 1/\sqrt{2})$, green). The resulting posterior distribution $P(\mathcal{X}|\mathcal{Y}) = P(\mathcal{Y}|\mathcal{X}) \cdot P(\mathcal{X})/P(\mathcal{Y})$ (black) is here again a normal distribution with mean $\bar{\mu} = (\mu_1\sigma_2^2 + \mu_2\sigma_1^2)/(\sigma_1^2 + \sigma_2^2)$ and variance $\bar{\sigma}^2 = (\sigma_1^2\sigma_2^2)/(\sigma_1^2 + \sigma_2^2)$. It is straightforward to show that the mean $\bar{\mu}$ lies between μ_1 and μ_2 and variance has become smaller $\bar{\sigma} \leq \min(\sigma_1, \sigma_2)$ (see text).

and $b < z$, and to make the calculations easy we assume that A is only one point away from winning, $a = z - 1$ and $b < z - 1$. If r_0 were known the answer would be trivial. In the conventional approach we would make an assumption about the parameter r_0. In the lack of knowledge we could make the null hypothesis $r_0 = \hat{r}_0 = \frac{1}{2}$, and find simply

$$P_0(B) \;=\; \text{Prob}\,(B \text{ is winning}) \;=\; (1 - \hat{r}_0)^{z-b} \;=\; \left(\frac{1}{2}\right)^{z-b} \;,$$

$$P_0(A) \;=\; \text{Prob}\,(A \text{ is winning}) \;=\; 1 \,-\, (1 - \hat{r}_0)^{z-b} \;=\; 1 \,-\, \left(\frac{1}{2}\right)^{z-b} \;,$$

because the only way for B to win is to make $z - b$ scores in a row. Thus fair odds for A to win would be $(2^{z-b} - 1) : 1$. An alternative approach is to make the *maximum likelihood* estimate on the unknown parameter $r_0 = \tilde{r}_0 = a/(a+b)$ and again we calculate the probabilities and find by the same token

$$P_{\mathrm{ml}}(B) \;=\; \mathrm{Prob}\,(B \text{ is winning}) \;=\; (1-\tilde{r}_0)^{z-b} \;=\; \left(\frac{b}{a+b}\right)^{z-b} ,$$

$$P_{\mathrm{ml}}(A) \;=\; \mathrm{Prob}\,(A \text{ is winning}) \;=\; 1 \,-\, (1-\tilde{r}_0)^{z-b} \;=\; 1 \,-\, \left(\frac{b}{a+b}\right)^{z-b} ,$$

and for the odds in favor of A: $\left(\frac{a+b}{b}\right)^{z-b} - 1$.

The Bayesian solution considers $r_0 = p$ as a unknown but variable parameter about which no estimate is made. Instead the uncertainty is modeled rigorously by integrating over all possible values: $0 \le p \le 1$. The expected probability for B to win is then

$$E\Big(P(B)\Big) \;=\; \int_0^1 (1-p)^{z-b}\, P(p\,|a,b)\,\mathrm{d}p \;,$$

where $(1-p)^{z-b}$ is the probability for winning B and $P(p\,|a,b)$ is the probability of a certain value of p provided the data a and b were obtained at the termination of the game. The probability $P(p\,|a,b)$ formally written as $P(\mathrm{model}|\mathrm{data})$ is the inversion of the common problem $P(\mathrm{data}|\mathrm{model})$ – given a certain model what is the probability to find a certain set of data – and a so-called *inverse probability* problem. The solution of the problem is provided by Bayes' theorem, which is an almost trivial truism for two random variables $\mathcal{X}$ and $\mathcal{Y}$:

$$P(\mathcal{X}|\mathcal{Y}) \;=\; \frac{P(\mathcal{Y}|\mathcal{X})\cdot P(\mathcal{X})}{P(\mathcal{Y})} \;=\; \frac{P(\mathcal{Y}|\mathcal{X})\cdot P(\mathcal{X})}{\sum_{\mathcal{Z}} P(\mathcal{Y}|\mathcal{Z})\cdot P(\mathcal{Z})} \;, \qquad (1.4')$$

where the sum over the random variable $\mathcal{Z}$ covers entire sample space. Equation (1.4') yields in our example

$$P(p\,|a,b) \;=\; \frac{P(a,b\,|p)\cdot P(p)}{\int_0^1 P(a,b\,|\varrho)\cdot P(\varrho)\,\mathrm{d}\varrho} \;.$$

The interpretation of the equation is straightforward: The probability of a particular choice of p given the data (a,b) called the *posterior probability* (figure 1.3) is proportional to the probability to obtain the observed data if p were true – the *likelihood* of p – multiplied by the *prior probability* of this particular value of p relative to all other values of p. The integral in the denominator takes care of the normalization of the probability – the summation is replaced by an integral, because p is a continuous variable, and $0 \le p \le 1$ is the entire domain of p.

The likelihood term is calculated readily from the binomial distribution

$$P(a,b\,|p\,) \;=\; \binom{a+b}{b} p^a\,(1-p)^b \;,$$

but the probability prior requires more care. By definition $P(p)$ is the probability of p before the data have been recorded. How can we estimate p before we have seen any data? We are thus referred to the situation how r_0 is determined, and we know it has been picked from the uniform distribution and hence, $P(p)$ is a constant that appears in the numerator and in the denominator and thus cancels in the equation for Bayes' theorem (1.4'). After some algebraic computation we eventually obtain for winning of B:

$$E\Big(P(B)\Big) \;=\; \frac{\int_0^1 p^{\,a}\,(1-p)^z\,\mathrm{d}p}{\int_0^1 p^{\,a}\,(1-p)^b\,\mathrm{d}p}\;.$$

Integration is straightforward, because the integrals are known as Euler integrals of the first kind, which have the Beta-function as solution

$$B(x,y) \;=\; \int_0^1 z^{x-1}\,(1-z)^{y-1}\,\mathrm{d}z \;=\; \frac{(x-1)!\,(y-1)!}{(x+y-1)!} \;=\; \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\,. \qquad (2.121)$$

Finally, we obtain the following expression for the chance of winning of B

$$E\big(P(B)\big) \;=\; \frac{z!\,(a+b+1)!}{b!\,(a+z+1)!}\;,$$

and the Bayesian estimation for fair odds yields

$$\left(\frac{b!\,(a+z+1)!}{z!\,(a+b+1)!} - 1\right) \;:\; 1\;.$$

A specific numerical example is given in [52]: $a = 5$, $b = 3$, and $z = 6$. The null hypothesis of equal probabilities of winning for A and B, $\hat{r}_0 = 0.5$ yields an advantage of 7:1 for A, the maximum likelihood approach with $\tilde{r}_0 = a/(a+b) = 5/8$ yields ≈18:1, and the Bayesian estimate yields 10:1. The large differences should not be surprising since the sample size is very small. The correct answer of the table game with the values for a, b, and z is indeed 10 as can be easily verified by numerical computation with a small computer program.

Finally, we show how the Bayesian approach operates on probability distributions (a simple but straightforward description is found in [268]). According to equation (1.4') the posterior probability $P(\mathcal{X}|\mathcal{Y})$ is obtained through multiplication of the prior probability $P(X)$ by the data likelihood function $P(\mathcal{Y}|\mathcal{X})$ and normalization. We illustrate the relation between the probability function by means of two normal distributions and their product (figure 2.23). For the prior probability and the data function we assume

$$P(\mathcal{X}) \;=\; f_1(x) \;=\; \frac{1}{\sqrt{2\pi\sigma_1^2}}\, e^{-(x-\mu_1)^2/(2\sigma_1^2)} \;\text{ and}$$

$$P(\mathcal{X}|\mathcal{Y}) \;=\; f_2(x) \;=\; \frac{1}{\sqrt{2\pi\sigma_2^2}}\, e^{-(x-\mu_2)^2/(2\sigma_2^2)}\;,$$

and obtain for the product with the normalization factor $\mathcal{N} = \mathcal{N}(\mu_1, \mu_2, \sigma_1, \sigma_2)$

$$P(\mathcal{Y}|\mathcal{X}) \;=\; \mathcal{N}\, f_1(x)\, f_2(x) \;=\; \mathcal{N}\, g\; e^{-(x-\bar{\mu})^2/(2\bar{\sigma}^2)} \;\text{ with}$$

$$\bar{\mu} \;=\; \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\,,\quad \bar{\sigma}^2 \;=\; \frac{\sigma_1^2\,\sigma_2^2}{\sigma_1^2\,+\,\sigma_2^2}\,,\quad g \;=\; \frac{1}{2\pi\,\sigma_1\,\sigma_2}\, e^{-\frac{1}{2}\frac{(\mu_2-\mu_1)^2}{\sigma_1^2+\sigma_2^2}}\;,\;\text{ and}$$

$$\mathcal{N}\, g \;=\; \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sqrt{2\pi}\;\sigma_1\,\sigma_2} \;=\; \frac{1}{\sqrt{2\pi\,\bar{\sigma}^2}}\,,$$

as required for normalization of the Gaussian curve.

Two properties of the posterior probability are easily tested by means of our example: (i) the averaged mean, $\bar{\mu}$, lies always between μ_1 and μ_2 and (ii) the product distribution is sharper than the two factor distributions

$$\frac{\sigma_1^2\,\sigma_2^2}{\sigma_1^2+\sigma_2^2} \;\leq\; \min\{\sigma_1^2, \sigma_2^2\}\;,$$

with the equals sign requiring either $\sigma_1 = 0$ or $\sigma_2 = 0$. The improvement of the Bayesian analysis thus reduces the difference in the mean values between expectation and model, and the distribution becomes smaller in the sense of reducing uncertainty.

Whereas the Bayesian approach does not seem to provide a lot more information in situations where the models are confirmed by many other independent applications like, for example, in the majority of problems in physics and chemistry, the highly complex situations in modern biology, economics, or social sciences require highly simplified and flexible models and there is an ample field for application of Bayesian statistics.

Chapter 3
Stochastic processes

With four parameters I can fit an elephant and with five I can make him wiggle his trunk.
Enrico Fermi citing John von Neumann, 1953 [51].

Abstract Different classes of stochastic processes are defined and their properties are listed. The Chapman-Kolmogorov equation is introduced, derived in differential form, and used as basis for the classification of the major stochastic processes: drift, diffusion, and jump processes, which in pure form are described by Liouville equations, stochastic diffusion equations, and master equations, respectively. The most popular and frequently used equation is the Fokker-Planck equation that describes the evolution of a probability density through drift and diffusion. Stochastic differential equations (SDEs) model processes at the level of random variables by solving an ordinary differential equation upon which a diffusion or Wiener process is superimposed. Ensembles of individual trajectories of SDEs are equivalent to densities described by Fokker-Planck equations. Master equations are dealing with jump processes only and represent the appropriate tool for modeling processes described by discrete variables. For technical reasons they are difficult to handle unless population sizes are relatively small.

Stochastic processes introduce time into probability theory and represent the most prominent possibility to combine dynamical phenomena and randomness as a result of incomplete information. In physics and chemistry the dominant source of randomness is thermal motion but in biology the overwhelming complexity of systems is commonly prohibitive for a complete description, and then lack of information results also from the simplifications on the model. In essence, there are two ways to visualize stochasticity in processes: (i) calculation or recording of stochastic variables as functions of time called trajectories and sampling of trajectories from repetitions yielding bundles of curves that represent the stochastic process, and (ii) modeling of the time dependence of entire probability densities. For an illustrative example comparing superposition of trajectories and migration of the probability density we refer to the Ornstein-Uhlenbeck process shown in figures 3.8 and 3.9. The expectation value of a random variable as a function of time, $E\big(\mathcal{X}(t)\big)$, often coincides with the deterministic solution of the corresponding differ-

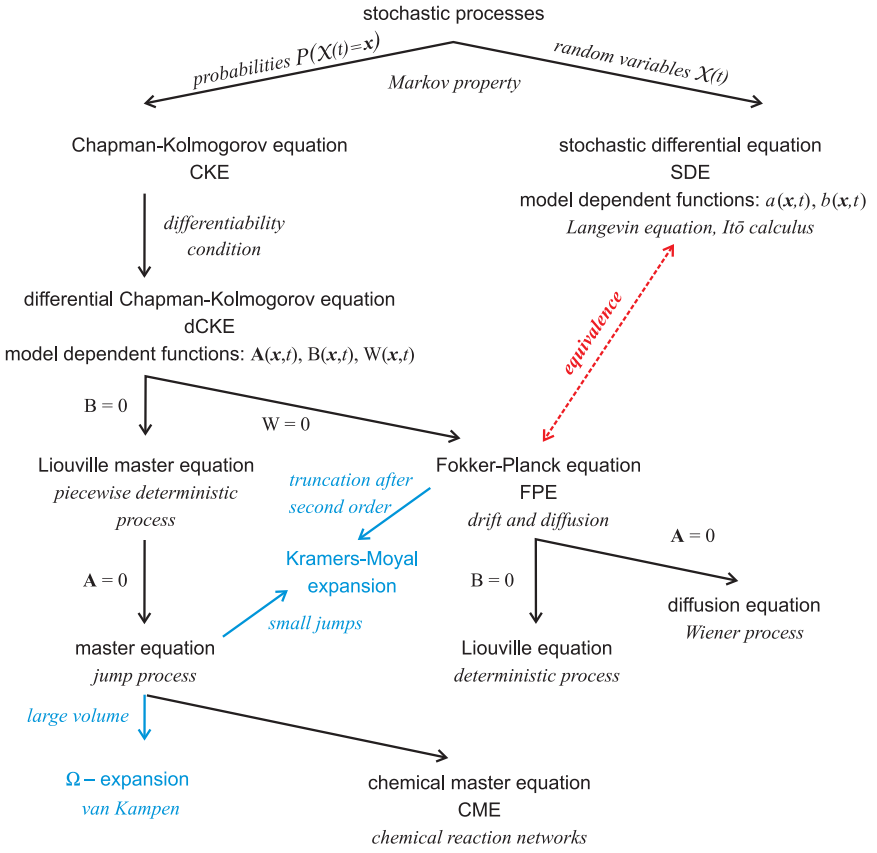

Fig. 3.1 Models of stochastic processes. The sketch presents a *family tree* of stochastic models [282]. Almost all stochastic models used in science are based on the Markov property of processes, which – in a nutshell – states that full information on the system at present is sufficient for predicting or modeling the future (section 3.2.1.3). Models fall into two major classes depending on the objects they are dealing with: (i) random variables $\mathcal{X}(t)$ or (ii) probability densities $P(\mathcal{X}(t) = x)$. In the center of stochastic modeling stands the Chapman-Kolmogorov equation (CKE) that introduces the Markov property into time series of probability densities. In differential form CKE contains three model dependent functions $\mathbf{A}(\mathbf{x}, t)$, $\mathrm{B}(\mathbf{x}, t)$, and $\mathrm{W}(\mathbf{x}, t)$, which determine the nature of the stochastic process. Different combinations of these functions yield the most important equations for stochastic modeling: the Fokker-Planck equation with $\mathbf{A} \neq 0$ and $\mathrm{B} \neq 0$, the stochastic diffusion equation with $\mathrm{B} \neq 0$, and the master equation with $\mathrm{W} \neq 0$. For stochastic processes without jumps the stochastic differential equation to the evolution of a probability density, $P(\mathcal{X}(t) = x)$, described by a Fokker-Planck equation (red arrow). Common approximations by means of expansions are shown in blue.

ential equation. In absence of bifurcations the typical solutions of ordinary (ODEs) or partial differential equations (PDEs) consists of single trajectories whereas the solutions of stochastic processes correspond to bundles of trajectories, which differ in the sequence of random events and which surround the deterministic solution. Commonly, a sharp reference state is chosen as initial condition and the bundle of trajectories diverges either in the future or in the past depending on whether the process is studied in the *forward* or in the *backward* direction. The stochastic equations in the forward and backward direction are different and in a way the typical symmetry of differential equations with respect to time reversal is no more existent because of the diffusion term [4, 60, 263]. In the forward direction the time dependent variance, $\sigma^2\big(\mathcal{X}(t)\big)$ allows for the distinction of two types of processes: (i) The variance increases with time and grows without limits, a behavior that is typical for spatial diffusion and some biologically important processes involving populations in abstract spaces, and (ii) the variance approaches a finite long time limit, which corresponds to thermodynamic equilibrium or to a stationary state and fulfil an approximate $\sqrt{N}$-law.

Figure 3.1 presents a listing of the most frequently used general stochastic models,[1] which are introduced in this chapter, and shows how they are interrelated [282, 283]. The two classes of equations of central importance are (i) the differential form of the Chapman-Kolmogorov equation (dCKE; section 3.2.2) for the evolution of probability densities and (ii) the stochastic differential equation (SDE; section 3.4) describing stochastic trajectories. The Fokker-Planck equation and the master equation are derived from the differential Chapman-Kolmogorov equation through restriction to continuous processes or jump processes, respectively. The *chemical master equation* is a master equation adapted for modeling chemical reaction networks where the jumps are integer changes in the particle numbers of chemical species (section 4.2.1). In this chapter we shall present an introduction to and a general formalism for modeling stochastic processes that is essentially based on two textbooks [41, 93, 287] and use the notation introduced by Crispin Gardiner [92]. A few examples of stochastic processes of general importance will be discussed here for the purpose of illustration of formalisms. Applications are presented in the forthcoming two chapters 4 and 5. Analysis of stochastic processes by mathematics is complemented by numerical simulations [106], which became more and more important over the years for two reasons: (i) the accessibility of cheap and extensive computing power, and (ii) the need for stochastic treatments of complex kinetics in chemistry and biology. Numerical simulation methods will also be presented and discussed in the two forthcoming chapters 4 and 5.

[1] By *general* we mean here methods that are widely applicable and not tailored specifically for one case or a few examples only.

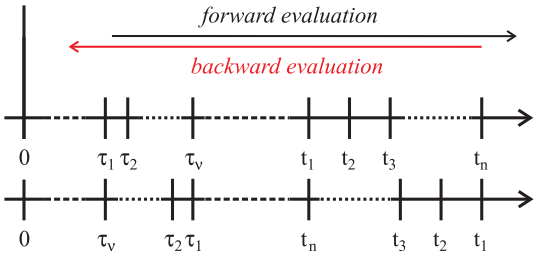

Fig. 3.2 Time order in modeling stochastic processes. Time is progressing from left to right and the most recent event is given by the rightmost recording. Conventional numbering of instances in physics starts at some time $t_0 = 0$ and ends at time t_n. It might be useful to have two series of events, one beginning at τ_1 and ending at τ_ν, and a second and later one ranging from t_1 to t_n (upper time axis). In the theory of stochastic processes an opposite ordering of times is often preferred and τ_1 and t_1, respectively, are the latest events of the series (lower time axis). The Chapman-Kolmogorov equation describing stochastic processes comes in two forms: (i) the forward equation predicting the future from past and present and (ii) the backward equation that extrapolates back in time from present to past.

3.1 Trajectories and processes

Systems evolving probabilistically in time are described and modeled as *stochastic processes*. More precisely, we postulate the existence of a time dependent random variable $\mathcal{X}(t)$ or random vector

$$\vec{\mathcal{X}}(t) = \big(\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\big).^2$$

and we distinguish the simpler discrete case,

$$P_n(t) \;=\; P\Big(\mathcal{X}(t) = x_n\Big) \quad \text{with} \quad n \in \mathbb{N}\,, \tag{3.1}$$

from the continuous or probability density case,

$$dF(x,t) \;=\; f(x,t)\,dx \;=\; P\Big(x \le \mathcal{X}(t) \le x + dx\Big) \quad \text{with} \quad x \in \mathbb{R}\,. \tag{3.2}$$

In both cases an experiment, a *sample path* or *trajectory*, is understood as a recording of the particular values of $\mathcal{X}$ at certain times:

$$\mathcal{T} \;=\; \Big((x_1,t_1),(x_2,t_2),(x_3,t_3),\cdots,(x_k,t_k),(x_{k+1},t_{k+1}),\cdots\Big)\,. \tag{3.3}$$

[2] At first we need not specify whether $\mathcal{X}(t)$ is a simple random variable or a random vector. Later on, when a distinction between problems of different dimensionality becomes necessary, we shall make clear, in which sense $\mathcal{X}(t)$ is used $\big($variable in one dimension or vector $\vec{\mathcal{X}}(t)\big)$.

Although it is not essential for the application of probability theory, but for the sake of clearness we shall always assume that the recorded values are time ordered, here with the earliest or oldest values on the rightmost position and the most recent values at the latest entry on the left-hand side (figure 3.2):

$$t_1 \geq t_2 \geq t_3 \geq \cdots \geq t_k \geq t_{k+1} \geq \cdots \ .$$

A trajectory thus is a sequence of time ordered doubles (x, t). It is worth noticing that the conventional time axis in drawings of processes in physics goes in opposite direction, from left to right:

$$\mathcal{T} \ = \ \Big((x_n, t_n), (x_{n-1}, t_{n-1}), (x_{n-2}, t_{n-2}), \cdots, (x_k, t_k), (x_{k-1}, t_{k-1}), \cdots\Big) \ ,$$

with the ordering

$$t_n \geq t_{n-1} \geq t_{n-2} \geq \cdots \geq t_k \geq t_{k-1} \geq \cdots \ .$$

In order to avoid confusion we shall always state explicitly when we are not using the convention shown in (3.3).[3]

So far we have only used the vague notion of *scores* and not yet specified, which quantities the random variables $(\mathcal{A}, \mathcal{B}, \ldots, \mathcal{W}) \in \Omega$ describe and what their realizations in some measurable space, $(a, b, \ldots, w) \in \mathbb{R}$, are. Stochastic processes in chemistry and biology are commonly modeling the time development of ensembles or populations. In spatially homogeneous chemical reaction systems the variables are discrete particle numbers or continuous concentrations, $\mathcal{A}(t)$ or $a(t)$. Spatial heterogeneity can be accounted for by diffusion resulting in reaction-diffusion systems, where the solutions are visualized best as migration of evolving probability densities in time and space, which is conventional physical space in three dimensions. Then, the variables are functions in space and time, $\mathcal{A}(\mathbf{r}, t)$ or $a(\mathbf{r}, t)$, with $\mathbf{r} = (x, y, z) \in \mathbb{R}^3$ being a vector in space. In biology the variables can be numbers of individuals in populations and then they depend exclusively on time or as in chemistry they can depend on three-dimensional space when migration processes are considered. Sometimes it is of advantage to consider stochastic processes in formal spaces like the genotype or sequence space, which is a discrete space where the points represent individual genotypes and the distance of two genotypes counts the minimal number of mutations required to bridge the interval between them. Neutral evolution (section 5.3.2) can then be visualized as a diffusion process and Darwinian selection as a hill climbing process in genotype space. Biology has to face also another problem that gives rise to lack of precision: When modeled at the molecular level, which is the appropriate reference state, biological processes become exceedingly complex and the complexity is prohibitive for modeling in full detail at the current situation. Simplifica-

[3] The different writing of sequences mentioned here should not be confused with forward and backward processes to be discussed later on (section 3.3).

Table 3.1 Notation used in modeling stochastic processes. Four different approaches to model stochastic processes by probability densities are compared: (i) discrete values of the random variable $\mathcal{X}$ and discrete time, (ii) discrete values and continuous time, (iii) continuous values and discrete time, and eventually (iv) continuous values and continuous time.

Values	Time: discrete	Time: continuous
discrete	$P_{n,k} = P(\mathcal{X}_k = x_n)\,;\, k,n \in \mathbb{N}$	$P_n(t) = \big(\mathcal{X}(t) = x_n\big)\,;\, n \in \mathbb{N}, t \in \mathbb{R}$
continuous	$p_k(x)\,\mathrm{d}x = P(x \le \mathcal{X}_k \le x + \mathrm{d}x) =$ $= f_k(x)\,\mathrm{d}x = \mathrm{d}F_k(x)$ $k \in \mathbb{N}\,,\ x \in \mathbb{R}$	$p\,(x,t)\,\mathrm{d}x = P(x \le \mathcal{X}_k \le x + \mathrm{d}x) =$ $= f(x,t)\,\mathrm{d}x = \mathrm{d}F(x,t)$ $x,t \in \mathbb{R}$

tion and drastic reduction of variables is indispensable and introduces model inherent errors and randomness into the description.

3.2 Modeling stochastic processes

The use of conventional differential equations for modeling dynamical systems implies determinism in the sense that full information at a single instant t_0 allows for computation of future and past. Stochasticity provides also the possibility for modeling a rich variety of different behavior with respect to influence from the past on the future. In this section we shall present different types of stochastic processes with characteristic properties with respect to *memory effects*. Markov processes are of particular importance in science because they have no memory in the sense that probabilistically the future can predicted form the presence and no knowledge on previous events is required.

A stochastic process, as we shall assume, is determined by a set of joint probability densities the existence and analytical form of which is presupposed.[4] The probability density encapsulates the physical nature of the process and contains all parameters and data on external conditions and hence we can assume that they determine the system completely:

[4] The joint density p is defined in the same way as in equations (1.30) and subsection 1.9.2 but with a slightly different notation. In describing stochastic processes we are always dealing with doubles (x,t), and therefore we separate individual doubles by a semicolon: $\cdots\,; x_k, t_k; x_{k+1}, t_{k+1}; \cdots$.

$$p(x_1, t_1; x_2, t_2; x_3, t_3; \cdots ; x_n, t_n; \cdots) \ . \tag{3.4}$$

By the phrase '*the determination is complete*' we mean that no additional information is needed to describe the progress in terms of a time ordered series (3.3) and we shall call such a process a *separable stochastic process*. Although more general processes are conceivable, they play little role in current physics, chemistry, and biology and therefore we shall not consider them here.

Calculation of probabilities from (3.4) by means of marginal densities (1.33) and (1.66) is straightforward. For the discrete case the result is obvious:

$$P(\mathcal{X} = x_1) \ = \ p(x_1, *) \ = \sum_{x_k \neq x_1} p(x_1, t_1; x_2, t_2; x_3, t_3; \cdots ; x_n, t_n; \cdots) \ .$$

The probability to record the value x_1 for the random variable $\mathcal{X}$ at time t_1 is obtained through summation over all previous values $x_2, x_3, \ldots$. In the continuous case we obtain analogously

$$P(\mathcal{X}_1 = x_1 \in [a, b]) \ = \ \int_a^b \mathrm{d}x_1 \iiint_{-\infty}^{\infty} \mathrm{d}x_2 \mathrm{d}x_3 \cdots \mathrm{d}x_n \cdots \\ p(x_1, t_1; x_2, t_2; x_3, t_3; \cdots ; x_n, t_n; \cdots) \ .$$

Time ordering allows us to formulate predictions of future values from the known past in terms of conditional probabilities:

$$p(x_1, t_1; x_2, t_2; \cdots | x_k, t_k; x_{k+1}, t_{k+1}, \cdots) \ = \\ = \ \frac{p(x_1, t_1; x_2, t_2; \cdots ; x_k, t_k; x_{k+1}, t_{k+1}, \cdots)}{p(x_k, t_k; x_{k+1}, t_{k+1}, \cdots)} \ ,$$

with $t_1 \geq t_2 \geq \cdots \geq t_k \geq t_{k+1} \geq \cdots$. In other words, we may compute

$$\{(x_1, t_1), (x_2, t_2), \cdots\} \text{ from known } \{(x_k, t_k), (x_{k+1}, t_{k+1}), \cdots\} \ .$$

With respect to the temporal progress of the process we shall distinguish discrete and continuous time: A trajectory in discrete time is just a time ordered sequence of random variables, $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ where time is implicitly included in the index of the variable in the sense that $\mathcal{X}_1$ is recorded at time t_1, $\mathcal{X}_2$ at time t_2, and so on. For the probability distribution we require to indices, n for the values the random variable can adopt and k for time: $P_{n,k} = P(\mathcal{X}_k = x_n)$ with $n, k \in \mathbb{N}$ (table 3.1). The introduction of continuous time is straightforward, since we need only replace $k \in \mathbb{N}$ by $t \in \mathbb{R}$. The random variable is still discrete and the probability mass function becomes a function of time, $P_{n,k} \Rightarrow P_n(t)$. The transition to a continuous sample space for the random variable is done in precisely the same way as in the case of probability mass functions described in section 1.9.

Before we derive a general concept that allows for flexible modeling and stochastic descriptions, which are applicable to chemical kinetics and biological modeling, we introduce a few common classes of stochastic processes with characteristic properties and different behavior behavior with respect to past, present and future.

3.2.1 Memory in stochastic processes

Three simple stochastic processes with characteristic memory effects will be discussed here: (i) the fully factorizable process with probability densities that are independent of other events with the special case of the Bernoulli process where the probability densities are also independent of time, (ii) the martingale where the (sharp) initial value of the stochastic variable is equal to the conditional mean value of the variable in the future, and (iii) the Markov process where the future is completely determined by the presence.

3.2.1.1 Independence and Bernoulli processes

The simplest class of stochastic processes is characterized by *complete independence* of events,

$$p(x_1, t_1; x_2, t_2; x_3, t_3; \cdots) \;=\; \prod_i p(x_i, t_i) \;, \tag{3.5}$$

which implies that the current value $\mathcal{X}(t)$ is completely independent of its values in the past. A special case is the sequence of Bernoulli trials (see in previous chapters, in particular in sections 1.5 and subsection 2.3.2) where the probability densities are also independent of time: $p(x_i, t_i) = p(x_i)$, and then we have

$$p(x_1, t_1; x_2, t_2; x_3, t_3; \cdots) \;=\; \prod_i p(x_i) \;. \tag{3.5'}$$

Further simplification occurs, of course, when all trials are based on the same probability distribution – for example, if the same coin is tossed in Bernoulli trials – and then the product is replaced by $p(x)^n$.

3.2.1.2 Martingales

The notion of *martingale* has been introduced by the French mathematician Paul Pierre Lévy and the development of the theory of martingales is due to the American mathematician Joseph Leo Doob. Appropriately, we distinguish discrete time and continuous time processes. A discrete-time martingale is a

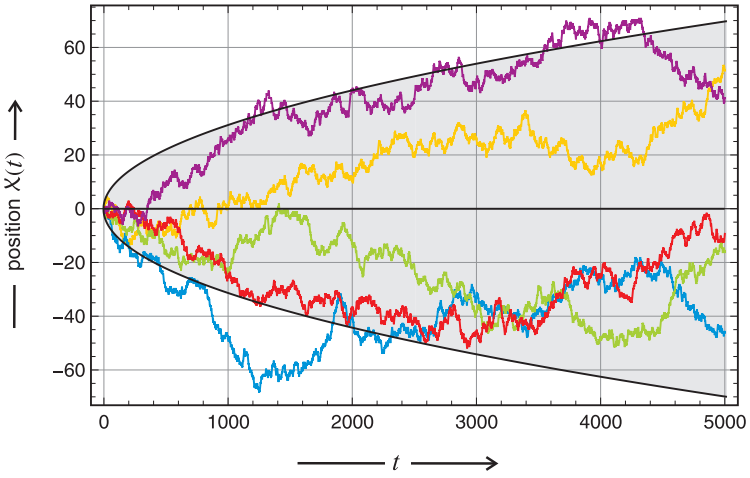

Fig. 3.3 The discrete time one-dimensional random walk. The one-dimensional random is shown as an example of a martingale. Five trajectories were calculated with different seeds for the random number generator. The expectation value $E\big(\mathcal{X}(t)\big) = x_0 = 0$ is constant and the variance grows linearly with time $\sigma^2\big(\mathcal{X}(t)\big) = n = \tau^{-1}t = 2\vartheta t$. The three black lines in the figure correspond to $E(t)$ and $E(t) \pm \sigma(t)$, and the grey area represent the confidence interval of 68,2 %. Choice of parameters: $\tau^{-1} = 1$ $(= 2\vartheta)$; $l = 1$; random number generator: *Mersenne Twister*, seeds: 491 (yellow), 919 (blue), 023 (green), 877 (red), 127 (violet).

sequence of random variables, $\mathcal{X}_1, \mathcal{X}_2, \ldots$, which satisfy the conditions

$$E(\mathcal{X}_{n+1}|\mathcal{X}_1, \ldots, \mathcal{X}_n) \;=\; \mathcal{X}_n \quad \text{and} \quad E(|\mathcal{X}_n|) < \infty \;. \tag{3.6}$$

Given all past values $\mathcal{X}_1, \ldots, \mathcal{X}_n$ the conditional expectation value for the next observation $E(\mathcal{X}_{n+1})$ is equal to the last recorded value $\mathcal{X}_n$.

A continuous time martingale refers to a random variable $\mathcal{X}(t)$ with the expectation value $E\big(\mathcal{X}(t)\big)$. We define first the conditional expectation value of the random variable for $\mathcal{X}(t_0) = x_0$ and $E(|\mathcal{X}(t)|) < \infty$:

$$E\big(\mathcal{X}(t)|(x_0, t_0)\big) \;:=\; \int \mathrm{d}x\, p(x, t|x_0, t_0) \;.$$

In a martingale the conditional mean is simply given by

$$E\big(\mathcal{X}(t)|(x_0, t_0)\big) \;= x_0 \;. \tag{3.7}$$

The mean value at time t is identical to the initial value of the process. The martingale property is rather strong and we shall use it for several specific situations.

As an example of a martingale we show the unlimited symmetric random walk in one dimension (figure 3.3): Equally sized steps of length l to the right and to the left are taken with equal probability. In the discrete time random walk the waiting time between two steps is τ and appropriately we measure time in multiples of the waiting time: $t = k \cdot \tau$. The corresponding probability to be at location $x = n \cdot l$ at time is simply expressed by

$$P(n, k+1 \,|\, n_0, k_0) \;=\; \frac{1}{2}\Big(P(n+1, k \,|\, n_0, k_0) + P(n-1, k \,|\, n_0, k_0)\Big) \;. \quad (3.8)$$

Equation (3.8) can be readily solved by means of the characteristic function $\phi(s,k) = \left\langle e^{\dot\imath n s} \right\rangle = \sum_n P(n, k \,|\, n_0, k_0)\, e^{\dot\imath n s}$, and yields

$$\phi(s,k) \;=\; \cosh^k(\dot\imath s) \;\text{ and }\; P(n, k \,|\, 0, 0) \;=\; \left(\frac{1}{2}\right)^k \frac{k!}{\left(\frac{k-n}{2}\right)!\,\left(\frac{k+n}{2}\right)!} \,, \quad (3.9)$$

where the inial conditions were simplified without loosing generality: $n_0 = 0$ and $k_0 = 0$. Calculation of first and second moments is straightforward and can be achieved best by using the derivatives of the characteristic function as shown in equation (2.29):

$$\begin{aligned} \frac{\partial \phi(s,k)}{\partial s} \;&=\; \dot\imath\, n \, \cosh^{n-1}(\dot\imath s) \cdot \sinh(\dot\imath s) \;\text{ and} \\ \frac{\partial^2 \phi(s,k)}{\partial s^2} \;&=\; -n \left(\cosh^n(\dot\imath s) + (n-1)\, \cosh^{n-2}(\dot\imath s) \cdot \sinh^2(\dot\imath s)\right) \end{aligned}$$

Insertion of $s = 0$ yields $(\partial\phi/\partial s)|_{s=0} = 0$ and $(\partial^2\phi/\partial s^2)|_{s=0} = -n$ and by equation (2.29) we obtain with $n(0) = n_0$ and $k(0) = k_0$ for the moments

$$E\big(\mathcal{X}(t)\big) \;=\; x_0 \;=\; n_0 \cdot l \quad\text{and}\quad \sigma^2\big(\mathcal{X}(t)\big) \;=\; t - t_0 \;=\; (k - x k_0)\, \tau \;. \quad (3.10)$$

The unlimited symmetric (discrete) random walk in one dimension is a martingale and the standard deviation $\sigma\big(\mathcal{X}(t)\big)$ increases with $\sqrt{t}$ as predicted in the path-breaking works of Albert Einstein [58] and Marian von Smoluchowski [298].

The somewhat relaxed notion of a *semimartingale* is of importance because it covers the majority of processes that are accessible to modeling by *stochastic differential equations*. A semimartingale is composed of a *local martingale* and an *adapted càdlàg-process*[5] with bounded variation

$$\mathcal{X}(t) \;=\; \mathcal{M}(t) + \mathcal{A}(t)$$

[5] The property *càdlàg* is an acronym from French for "*continue à droite, limites à gauche*". It is a common property of step functions in probability theory (section 1.6.1).

A local martingale is a stochastic process that satisfies locally the martingale property (3.7) but its expectation value $\langle \mathcal{M}(t) \rangle$ may be distorted at long times by large values of low probability. Hence, every martingale is a local martingale and every bounded local martingale is a martingale. In particular, every driftless diffusion process is a local martingale but need not be a martingale.

An adapted process $\mathcal{A}(t)$ is *nonanticipating* in the sense that it *cannot see into the future*. An informal interpretation [303, section II.25] would say: A stochastic process $\mathcal{X}(t)$ is adapted if and only if for every realization and for every time t, $\mathcal{X}(t)$ is known at time t and not before. The notion 'nonanticipating' is irrelevant for deterministic processes but matters for processes containing fluctuating elements, because the independence of random or irregular increments makes it impossible to look into the future. The concept of adapted processes is essential, for the Itō stochastic integral, which requires that the integrand is an adapted process (section 3.4.2).

Two generalizations of martingales are in common use: (i) A discrete time *submartingale* is a sequence $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \ldots$, of random variables that satisfy

$$E(\mathcal{X}_{n+1} | \mathcal{X}_1, \ldots, \mathcal{X}_n) \ \geq \ \mathcal{X}_n \ , \tag{3.11}$$

and for the continuous time analogue we have the condition

$$E\big(\mathcal{X}(t) | \{\mathcal{X}(\tau) : \tau \leq s\}\big) \ \geq \mathcal{X}(s) \ \forall \ s \ \leq \ t \ . \tag{3.12}$$

(ii) The relations for *supermartingales* are in complete analogy to the submartingales, only the '$\geq$' relations have to be replaced by '$\leq$':

$$E(\mathcal{X}_{n+1} | \mathcal{X}_1, \ldots, \mathcal{X}_n) \ \leq \ \mathcal{X}_n \ , \tag{3.13}$$

$$E\big(\mathcal{X}(t) | \{\mathcal{X}(\tau) : \tau \leq s\}\big) \ \leq \mathcal{X}(s) \ \forall \ s \ \leq \ t \ . \tag{3.14}$$

A straightforward consequence of the property of martingales is: If a sequence or a function of random variables is a simultaneously submartingale and a supermartingale it is a martingale.

3.2.1.3 Markov processes

Another simple concept assumes that knowledge of the present only is sufficient to predict the future. It is realized in *Markov processes* named after the Russian mathematician Andrey Markov[6] and can formulated easily in terms

[6] The Russian mathematician Andrey Markov (1856-1922) is one of the founders of Russian probability theory and pioneered the concept of memory free processes, which is named after him. He expressed more precisely the assumptions that were made by Albert Einstein [58] and Marian von Smoluchowski [298] in their derivation of the diffusion process.

of conditional probabilities:

$$p(x_1,t_1;x_2,t_2;\cdots|x_k,t_k;x_{k+1},t_{k+1},\cdots) \;=\; p(x_1,t_1;x_2,t_2;\cdots|x_k,t_k)\;. \tag{3.15}$$

In essence, the Markov condition expresses independence of the history of the process prior to time t_k, or in other words and said more sloppily: *"A Markov process has no memory and the future is completely determined by the presence"*. In particular, we have

$$p(x_1,t_1;x_2,t_2;x_k,t_k) \;=\; p(x_1,t_1|x_2,t_2)\,p(x_2,t_2|x_k,t_k)\;.$$

As we have seen in section 1.6.3 any arbitrary joint probability can be simply expressed as products of conditional probabilities:

$$\begin{aligned} &p(x_1,t_1;x_2,t_2;x_3,t_3;\cdots;x_n,t_n) \;= \\ &=\; p(x_1,t_1|x_2,t_2)\,p(x_2,t_2|x_3,t_3)\cdots p(x_{n-1},t_{n-1}|x_n,t_n)\,p(x_n,t_n) \end{aligned} \tag{3.15'}$$

under the assumption of time ordering $t_1 \geq t_2 \geq t_3 \geq \ldots \geq t_{n-1} \geq t_n$.

Stationarity is an important property of Markov processes, and we shall make use of it in the search for thermodynamic equilibria and stationary states. Several definitions of stationarity are possible and we shall adopt here a rather strict one [93, pp. 56-65]: A stochastic process is called *stationary* if $\mathcal{X}(t)$ and $\mathcal{X}(t+\Delta t)$ obey the same statistics for every Δt. Accordingly, joint probability densities are invariant to time translation:

$$p(x_1,t_1;x_2,t_2;\ldots;x_n,t_n) \;=\; p(x_1,t_1+\Delta t;x_2,t_2+\Delta t;\ldots;x_n,t_n+\Delta t)\;. \tag{3.16}$$

In other words, the probabilities are only functions of the differences in time, $t_k - t_j$, and this leads to time independent stationary *one-time probabilities*

$$p(x,t) \;\Longrightarrow\; \bar{p}(x)\;, \tag{3.17}$$

and two-times joint or conditional probabilities of the form

$$\begin{aligned} p(x_1,t_1;x_2,t_2) \;&\Longrightarrow\; \bar{p}(x_1,t_1-t_2;x_2,0) \;\text{ and} \\ p(x_1,t_1|\,x_2,t_2) \;&\Longrightarrow\; \bar{p}(x_1,t_1-t_2|\,x_2,0)\;. \end{aligned} \tag{3.18}$$

Since all joint probabilities of a Markov process can be written as products of two-time conditional probabilities and a one-time probability (3.15'), the necessary and sufficient condition for stationarity is cast into the requirement to be able to write all one- and two-time probabilities as shown in equations (3.17) and (3.18). A Markov process that becomes stationary in the limit $t \to \infty$ or $t_0 \to -\infty$ is called a *homogeneous Markov process*.

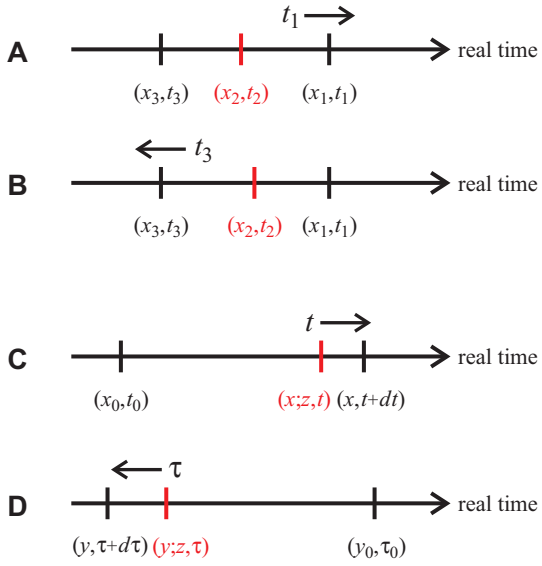

Fig. 3.4 Notation of time dependent variables. In the following sections we shall require several time dependent variables and adopt the following notations: In case of the Chapman-Kolmogorov equation we require three variables at different times denoted by x_1, x_2, and x_3. The variable x_2 is associated with intermediate time t_2 (red) and disappears through integration. In the forward equation (x_3, t_3) are fixed initial conditions and (x_1, t_1) is moving (**A**). For backward integration the opposite relation is assumed: (x_1, t_1) being fixed and (x_3, t_3) moving (**B**). In both cases *real time* progresses from the left to the right. The lower part of the figure shows notations used for forward and backward differential Chapman-Kolmogorov equations: In the forward equation (**C**) $x(t)$ is the variable, the initial conditions are denoted by (x_0, t_0) and (z, t) is an intermediate double (red). In the backward equation the time order is reversed (**D**): $y(\tau)$ is the variable and (y_0, τ_0) are the final conditions.

3.2.1.4 Continuity in stochastic processes

The condition of continuity in Markov processes requires a more detailed discussion. For this goal we consider a process that progresses from location z at time t to location x at time $t + \Delta t$ denoted as $(z, t) \to (x, t + \Delta t)$.[7] Then the process is continuous if and only if in the limit $\lim \Delta t \to 0$ the probability of x to be finitely different from z goes to zero faster than Δt as expressed by the equation

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|>\varepsilon} dx\, p\,(x, t + \Delta t | z, t) \;=\; 0\;, \tag{3.19}$$

[7] For the time dependent variables we use the notation listed in figure 3.4.

and this convergence is uniform in z, t, and Δt. In other words, the difference in probability as a function of $|x - z\,|$ approaches zero sufficiently fast and therefore no jumps occur in the random variable $\mathcal{X}(t)$.

Two illustrative examples for the analysis of continuity are chosen and sketches in figure 3.5: (i) the Einstein-Smoluchowski solution of Brownian motion, which is a continuous version of the random walk in one dimension shown in figure 3.3,[8] which leads to a normally distributed probability,

$$p(x, t + \Delta t|z, t) \;=\; \frac{1}{\sqrt{4\pi D \Delta t}} \, \exp\left(-\frac{(x-z)^2}{4D\Delta t}\right) \,, \tag{3.20}$$

and (ii) the so-called Cauchy process following the Cauchy-Lorentz distribution,

$$p(x, t + \Delta t|z, t) \;=\; \frac{\Delta t}{\pi} \, \frac{1}{(x-z)^2 + \Delta t^2} \,. \tag{3.21}$$

In case of the Wiener process we exchange the limit and the integral, introduce $\vartheta = (\Delta t)^{-1}$, perform the limit $\vartheta \to \infty$, and have

$$\begin{aligned}
&\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|>\varepsilon} \mathrm{d}x \, \frac{1}{\sqrt{4\pi D}} \, \frac{1}{\sqrt{\Delta t}} \, \exp\left(-\frac{(x-z)^2}{4D\Delta t}\right) \;= \\
&= \int_{|x-z|>\varepsilon} \mathrm{d}x \, \lim_{\Delta t \to 0} \frac{1}{\Delta t} \, \frac{1}{\sqrt{4\pi D}} \, \frac{1}{\sqrt{\Delta t}} \, \exp\left(-\frac{(x-z)^2}{4D\Delta t}\right) \;= \\
&= \int_{|x-z|>\varepsilon} \mathrm{d}x \, \lim_{\vartheta \to \infty} \frac{1}{\sqrt{4\pi D}} \, \frac{\vartheta^{3/2}}{\exp\left(\frac{(x-z)^2}{4D}\,\vartheta\right)} \,, \quad \text{where} \\
&\lim_{\vartheta \to \infty} \frac{\vartheta^{3/2}}{1 + \frac{(x-z)^2}{4D} \cdot \vartheta + \frac{1}{2!}\left(\frac{(x-z)^2}{4D}\right)^2 \cdot \vartheta^2 + \frac{1}{3!}\left(\frac{(x-z)^2}{4D}\right)^3 \cdot \vartheta^3 + \ldots} = 0 \,.
\end{aligned}$$

Since the power expansion of the exponential in the denominator increases faster than every finite power of ϑ, the ratio vanishes in the limit $\vartheta \to \infty$, the value of the integral is zero, and the Wiener process is continuous everywhere. Although it is continuous, the curve of Brownian motion [27] is extremely irregular since it is nowhere differentiable (figure 3.5).

In the second example, the Cauchy process, we exchange limit and integral as in case of the Wiener process, and perform the limit $\Delta t \to 0$:

[8] Later on we shall discuss the continuous version of this stochastic process in more detail and call it a *Wiener process*.

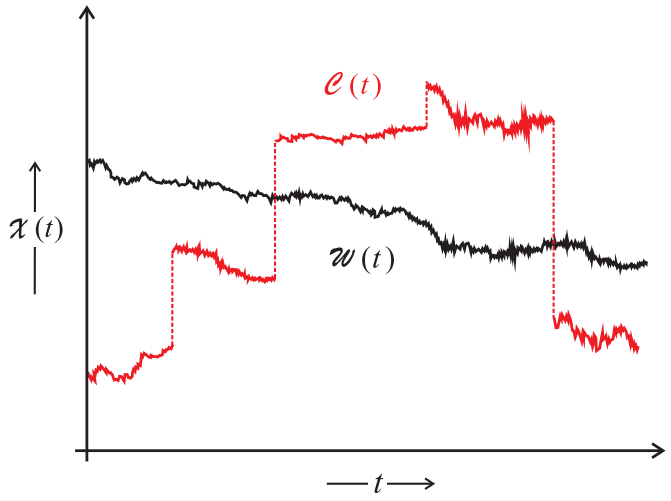

Fig. 3.5 Continuity in Markov processes. Continuity is illustrated by means of two stochastic processes of the random variable $\mathcal{X}(t)$, the Wiener process $\mathcal{W}(t)$ (3.20; black) and the Cauchy process $\mathcal{C}(t)$ (3.21; grey). The Wiener process describes Brownian motion and is continuous but almost nowhere differentiable. The even more irregular Cauchy process is wildly discontinuous.

$$\begin{aligned}
&\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{|x-z|>\varepsilon}\mathrm{d}x\,\frac{\Delta t}{\pi}\,\frac{1}{(x-z)^2+\Delta t^2} \;=\\
&=\int_{|x-z|>\varepsilon}\mathrm{d}x\,\lim_{\Delta t\to 0}\frac{1}{\Delta t}\,\frac{\Delta t}{\pi}\,\frac{1}{(x-z)^2+\Delta t^2} \;=\\
&=\int_{|x-z|>\varepsilon}\mathrm{d}x\,\lim_{\Delta t\to 0}\frac{1}{\pi}\,\frac{1}{(x-z)^2+\Delta t^2} \;=\; \int_{|x-z|>\varepsilon}\frac{1}{\pi(x-z)^2}\,\mathrm{d}x \;\neq\; 0\;.
\end{aligned}$$

The value of the last integral, $I=\int_{|x-z|>\varepsilon}^{\infty} dx/(x-z)^2 = 1/\big(\pi(x-z)\big)$, is of the order $I\approx 1/\varepsilon$ and accordingly finite. Consequently, the curve for the Cauchy-process is not only irregular but also discontinuous (figure 3.5).

Both processes, as required for consistency, fulfill the relation

$$\lim_{\Delta t\to 0} p\,(x,t+\Delta t|\,z,t) \;=\; \delta(x-z)\;,$$

where $\delta(\cdot)$ is the so-called delta-function (see section 1.6.2).

We are now in the position to give a concise mathematical definition for continuity in Markov processes [93, p. 46], which will be used to derive a comprehensive and convenient equation for stochastic processes. For general validity we use vector notation for the locations:

A Markov process has – with probability one – sample paths that are continuous functions of time t, if for any $\varepsilon > 0$ the limit

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|\boldsymbol{x}-\boldsymbol{z}|<\varepsilon} \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x}, t+\Delta t|\, \boldsymbol{z}, t) \;=\; 0 \;. \tag{3.22}$$

is approached uniformly in $\boldsymbol{z}$, t, and Δt.

In essence, equation (3.22) expresses the fact that probabilistically the difference between $\boldsymbol{x}$ and $\boldsymbol{z}$ converges to zero faster than Δt does.

3.2.2 Chapman-Kolmogorov equations

At the basis of general modeling of stochastic processes stands a straightforward consideration concerning the propagation of probability distributions in time: How to calculate the probability to come from $\mathcal{N}_3 = n_3$ at time $t = t_3$ to $\mathcal{N}_1 = n_1$ at time $t = t_1$. We assume an intermediate state ($\mathcal{N}_2 = n_2$ at $t = t_2$) and an implicit order in time: $t_1 \geq t_2 \geq t_3$. The value of the variable $\mathcal{N}_2$ need not be unique or, in other words, there may be different paths or trajectories leading from (n_3, t_3) to (n_1, t_1). If the individual values of the random variables are replaced by probabilities, $\mathcal{N} = n \implies P(\mathcal{N} = n, t) = P(n, t)$, an equation is obtained that encapsulates the full diversity of sources of randomness with the exception of quantum uncertainty [94]. The only restriction in the generally used form of this equation is the Markov property of the stochastic process. The equation is named *Chapman-Kolmogorov equation* after the British geophysicist and mathematician Sydney Chapman and the Russian mathematician Andrey Kolmogorov and for the rest of this section we shall be concerned with it.

3.2.2.1 Discrete and continuous Chapman-Kolmogorov equations

The relation between the three random variables $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ can be illustrated by application of set theoretical considerations. If all mutually exclusive events of one kind are included in the summation the corresponding variable B is eliminated:

$$\sum_B P(A \cap B \cap C) \;=\; P(A \cap C) \;.$$

First we assume to be dealing with a discrete state space and accordingly the random variables $\mathcal{N} \in \mathbb{N}$ are defined on the integers. Then we can simply make use of state space covering and find for the marginal probability

$$P(n_1,t_1) \;=\; \sum_{n_2} P(n_1,t_1;n_2,t_2) \;=\; \sum_{n_2} P(n_1,t_1|\,n_2,t_2)\,P(n_2,t_2)\;.$$

Now we introduce a third event (n_3,t_3) and describe the process by the equations for conditional probabilities

$$\begin{aligned} P(n_1,t_1|\,n_3,t_3) \;&=\; \sum_{n_2} P(n_1,t_1;n_2,t_2|\,n_3,t_3) \;= \\ &=\; \sum_{n_2} P(n_1,t_1|\,n_2,t_2;n_3,t_3)\,P(n_2,t_2|\,n_3,t_3)\;. \end{aligned}$$

Both equations are of general validity for all stochastic processes, and the series could be extended further to four, five events and so on. Adopting the Markov assumption and introducing the time order $t_1 \geq t_2 \geq t_3$ provides the basis for dropping the dependence on (n_3,t_3) in the doubly conditioned probability and leads to

$$P(n_1,t_1|n_3,t_3) \;=\; \sum_{n_2} P(n_1,t_1|n_2,t_2)\,P(n_2,t_2|n_3,t_3)\;. \tag{3.23}$$

This is the *Chapman-Kolmogorov equation* in its simplest general form. Equation (3.23) can be interpreted as a matrix multiplication where the size of the matrices depends on the event space of n_2 – it could even be countably infinite.

The extension from the discrete case to probability densities is straightforward. By the same token we find for the continuous case

$$p(x_1,t_1) \;=\; \int dx_2\,p(x_1,t_1;x_2,t_2) \;=\; \int dx_2\,p(x_1,t_1|x_2,t_2)\,p(x_2,t_2)\;,$$

and the extension to three events leads to

$$\begin{aligned} p(x_1,t_1|x_3,t_3) \;&=\; \int dx_2\,p(x_1,t_1;x_2,t_2|x_3,t_3) \;= \\ &=\; \int dx_2\,p(x_1,t_1|x_2,t_2;x_3,t_3)\,p(x_2,t_2|x_3,t_3)\;. \end{aligned}$$

For $t_1 \geq t_2 \geq t_3$ and making again use of the Markov assumption we obtain the continuous version of the Chapman-Kolmogorov equation:

$$p(x_1,t_1|x_3,t_3) \;=\; \int dx_2 p(x_1,t_1|x_2,t_2)\,p(x_2,t_2|x_3,t_3)\;. \tag{3.24}$$

Equation (3.24) is of very general nature. The only relevant approximation is the assumption of a Markov process, which is empirically full justified in physics, chemistry and biology. General validity is commonly accompanied by a variety of different solutions and the Chapman-Kolmogorov equation is

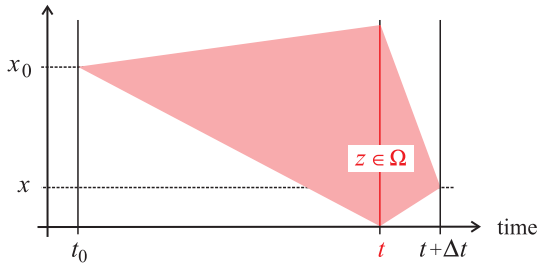

Fig. 3.6 Time order in the differential Chapman-Kolmogorov equation (dCKE) **.** The one-dimensional sketch shows the notation used in the derivation of the forward dCKE. The variable $\mathbf{z}$ is integrated over the entire sample space Ω in order to sum up all trajectories leading from $(\mathbf{x}_0, t_0)$ via $(\mathbf{z}, t)$ to $(\mathbf{x}, t + \Delta t)$.

no exception in this aspect. The generality of (3.24) in the description of a stochastic process becomes evident when the evolution in time is continued $t_1 \geq t_2 \geq t_3 \geq t_4 \geq t_5 \ldots$, where summations over all intermediate states are performed. Sometimes it is useful – and we shall adopt this notation here – to indicate an initial state by the doublet (x_0, t_0). As said before, all expressions are valid also for vectors $\mathbf{x}$ in space.

3.2.2.2 Differential Chapman-Kolmogorov forward equation

Since we aim at a description of processes the Chapman-Kolmogorov equations in discrete and continuous form as expressed in equations (3.23) and (3.24), respectively, provide a general definition of Markov processes but they are not really useful to describe the temporal evolution. Much better suited for describing stochastic processes as well as analyzing nature and properties of solutions or performing actual calculations is an equation in differential form. In a way the differential formulation of basic stochastic processes can be compared to the invention of calculus by Gottfried Wilhelm Leibniz and Isaac Newton, which provides the ultimate basis for all modeling by means of differential equations. Analytical solution or numerical integration of such a *differential Chapman-Kolmogorov equation* (dCKE) is then expected to provide the desired description of the process. A differential form of the Chapman-Kolmogorov equation has been derived by Crispin Gardiner [93, pp. 48-51][9] . We shall follow here, in essence, a simpler approach given more recently by Mukhtar Ullah and Olaf Wolkenhauer[282, 283].

[9] The derivation is contained already in the first edition of Gardiner's *Handbook of stochastic methods* [92] and it has been Crispin Gardiner who coined the name *differential Chapman-Kolmogorov equation*.

The Chapman-Kolmogorov equation is considered for an interval $t \to t+\Delta t$ defined for a sample space Ω and the initial conditions $(\mathbf{x}_0, t_0)$:

$$p\left(\mathbf{x}, t+\Delta t|\, \mathbf{x}_0, t_0\right) \;=\; \int_{\Omega} \mathrm{d}\mathbf{z}\; p\left(\mathbf{z}, t+\Delta t|\, \mathbf{x}_0, t_0\right) p\left(\mathbf{z}, t|\, \mathbf{x}_0, t_0\right). \qquad (3.24\,')$$

The probability of a transition $(\mathbf{x}_0, t_0) \to (\mathbf{x}, t+\Delta t)$ is obtained by summation of all probabilities to occur via an intermediate, $(\mathbf{x}_0, t_0) \to (\mathbf{z}, t) \to (\mathbf{x}, t+\Delta t)$ as illustrated in figure 3.6. In order to simplify derivation and notation we shall assume fixed initial conditions $(\mathbf{x}_0, t_0)$ for conditioning probability and transition:

$$p\left(\mathbf{x}, t\right) \;=\; p\left(\mathbf{x}, t|\, \mathbf{x}_0, t_0\right). \qquad (3.25)$$

We introduce the time derivative by tacitly assuming that the probability $p\left(\mathbf{x}, t\right)$ is differentiable with respect to time:

$$\frac{\partial}{\partial t} p\left(\mathbf{x}, t\right) \;=\; \lim_{\Delta t \to 0} \frac{1}{\Delta t}\Big(p\left(\mathbf{x}, t+\Delta t\right) \;-\; p(\mathbf{x}, t)\Big) \qquad (3.26)$$

Introducing the CKE in form (3.24') and multiplying $p\left(\mathbf{x}, t\right)$ formally by one in the form of the normalization condition of probabilities,[10]

$$1 \;=\; \int_{\Omega} \mathrm{d}\mathbf{z}\, p\left(\mathbf{z}, t+\Delta t|\, \mathbf{x}, t\right),$$

we can rewrite equation (3.26) as

$$\begin{aligned} \frac{\partial}{\partial t} p\left(\mathbf{x}, t\right) \;=\; \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{\Omega} \mathrm{d}\mathbf{z}\Big(& p\left(\mathbf{x}, t+\Delta t|\, \mathbf{z}, t\right) p\left(\mathbf{z}, t\right) - \\ & - \; p\left(\mathbf{z}, t+\Delta t|\, \mathbf{x}, t\right) p\left(\mathbf{x}, t\right)\Big). \end{aligned} \qquad (3.27)$$

For the purpose of integration the sample space Ω is divided up into to parts with respect to an arbitrarily small parameter $\epsilon > 0$. Using the notion of continuity (section 3.2.1.4) the region D_1 defined by $\|\mathbf{x}-\mathbf{z}\| < \epsilon$ represents a continuous process.[11] Part two of sample space, D_2 with $\|\mathbf{x}-\mathbf{z}\| \geq \epsilon$, corresponds to a jump process, and for the derivative taken on the entire sample space Ω we get:

[10] It is important to note that the trick in the derivation is that the time order is reversed in this integral.

[11] The notation $\|\cdot\|$ refers to a suitable vector norm – in the one-dimensional case we would just use the absolute value $|x-z|$.

$$\begin{aligned}
\frac{\partial}{\partial t}\,p\,(\mathbf{x},t) \;&=\; D_1 \,+\, D_2\,, \quad \text{with} \\
D_1 \;&=\; \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{\|\mathbf{x}-\mathbf{z}\|<\epsilon} \mathrm{d}\mathbf{z}\Big(p\,(\mathbf{x},t+\Delta t|\,\mathbf{z},t)\,p\,(\mathbf{z},t) - \\
&\qquad\qquad -\; p\,(\mathbf{z},t+\Delta t|\,\mathbf{x},t)\,p\,(\mathbf{x},t)\Big)\,, \quad \text{and} \qquad (3.28)\\
D_2 \;&=\; \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{\|\mathbf{x}-\mathbf{z}\|\geq\epsilon} \mathrm{d}\mathbf{z}\Big(p\,(\mathbf{x},t+\Delta t|\,\mathbf{z},t)\,p\,(\mathbf{z},t) - \\
&\qquad\qquad -\; p\,(\mathbf{z},t+\Delta t|\,\mathbf{x},t)\,p\,(\mathbf{x},t)\Big)\,.
\end{aligned}$$

In the first region with $\|\mathbf{x}-\mathbf{z}\| < \epsilon$ the integrand is expanded in a Taylor series with $\mathbf{r} = \mathbf{x} - \mathbf{z}$

$$\begin{aligned}
D_1 \;&=\; \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{\|\mathbf{r}\|<\epsilon} \mathrm{d}\mathbf{r}\Big(p\,(\mathbf{x},t+\Delta t|\,\mathbf{x}-\mathbf{r},t)\,p\,(\mathbf{x}-\mathbf{r},t) - \\
&\qquad\qquad -\; p\,(\mathbf{x}-\mathbf{r},t+\Delta t|\,\mathbf{x},t)\,p\,(\mathbf{x},t)\Big)\,.
\end{aligned}$$

In this Taylor expansion all terms higher than second order are vanishing for consistence [93, 282]. Provided the differentiability conditions are fulfilled we obtain in the limit $\epsilon \to 0$:

$$D_1 \;=\; -\sum_i \frac{\partial}{\partial x_i}\Big(A_i(\mathbf{x},t)\,p\,(\mathbf{x},t)\Big) \,+\, \frac{1}{2}\sum_i\sum_j \frac{\partial^2}{\partial x_i \partial x_j}\Big(B_{ij}(\mathbf{x}.t)\,p\,(\mathbf{x},t)\Big)\,. \qquad (3.29)$$

which defines a Fokker-Planck equation. In the limit $\epsilon \to 0$ the continuous part of the process becomes equivalent to an equation for the differential increments of the random vector $\vec{\mathcal{X}}(t)$ describing a single trajectory:

$$\vec{\mathcal{X}}(t+\,\mathrm{dt}) \;=\; \vec{\mathcal{X}}(t) \,+\, \mathbf{A}\big(\vec{X}(t),t\big)\;\mathrm{dt} \,+\, \Big(\mathrm{B}\big(\vec{\mathcal{X}}(t),t\big)\;\mathrm{dt}\Big)^{\frac{1}{2}}\,. \qquad (3.30)$$

Equation (3.30) is a stochastic differential equation or Langevin equation (see section 3.4.1).
The second part of the integration over sample space Ω involves the probability rate for jumps:

$$\begin{aligned}
D_2 \;&=\; \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{\|\mathbf{x}-\mathbf{z}\|\geq\epsilon} \mathrm{d}\mathbf{z}\Big(p\,(\mathbf{x},t+\Delta t|\,\mathbf{z},t)\,p\,(\mathbf{z},t) - \\
&\qquad\qquad -\; p\,(z,t+\Delta t|\,\mathbf{x},t)\,p\,(\mathbf{x},t)\Big)\,.
\end{aligned}$$

The condition for a jump process is $\|\mathbf{x}-\mathbf{z}\| \geq \epsilon$ (section 3.2.1.4) and accordingly we have

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(p\,(\mathbf{x},t+\Delta t|\,\mathbf{z},t)\,p\,(\mathbf{z},t)\right) \;=\; W(\mathbf{x}|\,\mathbf{z},t)\,p\,(\mathbf{z},t)\,, \qquad (3.31)$$

where $W(\mathbf{x}|\,\mathbf{z},t)$ is the transition rate for the jump $\mathbf{z} \to \mathbf{x}$. By the same token we define a transition rate for the jump in the reverse direction $\mathbf{x} \to \mathbf{z}$. As $\epsilon \to 0$ the integration is extended over the whole same space Ω and eventually we obtain

$$\lim_{\epsilon \to 0} D_2 \;=\; \int_{\Omega} \mathrm{d}\mathbf{z}\Big(W(\mathbf{x}|\,\mathbf{z},t)\,p\,(\mathbf{z},t) \;-\; W(\mathbf{z}|\,\mathbf{z},t)\,p\,(\mathbf{x},t)\Big)\;, \tag{3.32}$$

which completes the somewhat simplified derivation of the differential Chapman-Kolmogorov equation.

The evolution of the system is now expressed in terms of functions $\mathbf{A}(\mathbf{x},t)$, which correspond to the functional relations in conventional differential equations, a diffusion matrix $\mathrm{B}(\mathbf{x},t)$, and transition matrix for discontinuous jumps $\mathrm{W}(\mathbf{x}|\mathbf{z},t)$:

$$\frac{\partial p\,(\mathbf{x},t)}{\partial t} \;=\; -\sum_i \frac{\partial}{\partial x_i}\Big(A_i(\mathbf{x},t)\,p\,(\mathbf{x},t)\Big) + \tag{3.33a}$$

$$+\;\frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\Big(B_{ij}(\mathbf{x},t)\,p\,(\mathbf{x},t)\Big) + \tag{3.33b}$$

$$+\int \mathrm{d}\mathbf{z}\;\Big(\mathrm{W}(\mathbf{x}\,|\mathbf{z},t)\,p\,(\mathbf{z},t) - \mathrm{W}(\mathbf{z}\,|\mathbf{x},t)\,p\,(\mathbf{x},t)\Big). \tag{3.33c}$$

Equation (3.33) is called a *forward equation* in the sense of figure 3.15. In the derivation surface terms at the boundary of the domain of $\mathbf{x}$ have been neglected [93, p. 50]. This assumption is not critical for most cases considered here. It is always correct for infinite domains because the probabilities vanish: $\lim_{\mathbf{x}\to\pm\infty} p(\mathbf{x},t) = 0$.

From a mathematical purist's point of view it is not clear from the derivation that solutions of the differential Chapman-Kolmogorov equation (3.33) exist, are unique and are solutions to the Chapman-Kolmogorov equation (3.24) as well. It is true, however, that the set of conditional probabilities obeying equation (3.33) does generate a Markov process in the sense that the joint probabilities produced satisfy all probability axioms. It has been shown, however, that a non-negative solution to the differential Chapman-Kolmogorov equations exists and satisfies the Chapman-Kolmogorov equation under certain conditions (see [100, Vol.II]):

(1) $\mathbf{A}(\mathbf{x},t) = \{A_i(\mathbf{x},t);\, i = 1,\ldots\}$ and $\mathrm{B}(\mathbf{x},t) = \{B_{ij}(\mathbf{x},t);\, i,j = 1,\ldots\}$ are specific vectors and positive semidefinite matrices[12] of functions, respectively,
(2) $\mathrm{W}(\mathbf{x}|\mathbf{z},t)$ and are $\mathrm{W}(\mathbf{z}|\mathbf{x},t)$ non-negative quantities,
(3) the initial condition has to satisfy $p(\mathbf{x},t|\mathbf{x}_0,t_0) = \delta(\mathbf{x}_0 - \mathbf{x})$ which follows from the definition of a conditional probability density, and

[12] A positive definite matrix has exclusively positive eigenvalues, $\lambda_k > 0$ whereas a positive semidefinite matrix has non-negative eigenvalues, $\lambda_k \geq 0$.

(4) appropriate boundary conditions have to be fulfilled.

The boundary conditions are very hard to specify for the full equation but can be discussed precisely for special cases, for example in the case of the Fokker-Planck equation [250].

The nature of the different stochastic processes associated with the three terms in equation (3.33), $\mathbf{A}(\mathbf{x},t)$, $\mathrm{B}(\mathbf{x},t)$, $\mathrm{W}(\mathbf{x}\,|\,\mathbf{z},t)$ and $\mathrm{W}(\mathbf{z}\,|\,\mathbf{x},t)$, is visualized by setting some parameters equal to zero and analyzing the remaining equation. We shall discuss here four cases that are modeled by different equations (for relations between them see figure 3.1).

(i) $\mathrm{B}=0$, $\mathrm{W}=0$, deterministic drift process: Liouville equation,
(ii) $\mathbf{A}=0$, $\mathrm{W}=0$, drift free diffusion process or Wiener process,
(iii) $\mathrm{W}=0$, drift and diffusion process: Fokker-Planck equation, and
(iv) $\mathbf{A}=0$, $\mathrm{B}=0$, pure jump process: master equation.

The first term in differential Chapman-Kolmogorov equation, equation (3.33a) is the probabilistic version of a differential equation describing deterministic motion, which is known as *Liouville equation* named after the French mathematician Joseph Liouville. It is a fundamental equation of statistical mechanics and will be discussed in some detail subsection 3.2.3.1. With respect to the theory of stochastic processes (3.33a) encapsulates the drift of a probability distribution.

The second term in equation (3.33) describes spreading of probability densities by diffusion and is called a stochastic *diffusion equation*. In pure form it is represented by the Wiener process, which got the name from the American mathematician Norbert Wiener and which can be understood as the continuous time and continuous space limit of the one-dimensional random walk (see figure 3.3). The Wiener process is fundamental for understanding stochasticity in continuous space and time and will be discussed in subsection 3.2.3.2.

Combining equations (3.33a) and (3.33b) yields the Fokker-Planck equation, which we repeat here because of its general importance:

$$\frac{\partial p\,(\mathbf{x},t)}{\partial t} \;=\; -\sum_i \frac{\partial}{\partial x_i}\Big(A_i(\mathbf{x},t)\,p\,(\mathbf{x},t)\Big) \;+\; \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\Big(B_{ij}(\mathbf{x},t)\,p\,(\mathbf{x},t)\Big)\,. \quad (3.34)$$

The equation is named after two physicists, the Dutchman Adriaan Daniël Fokker and the German Max Planck. Fokker-Planck equations are frequently used in physics to model and analyze processes with fluctuations [250].

If only the third term of the differential Chapman-Kolmogorov equation, (3.33c), has nonzero elements, the variables $\mathbf{x}$ and $\mathbf{z}$ change only in steps and the corresponding differential equation is called a *master equation*. Master equations are the most important tools for describing processes in discrete spaces, $\mathcal{X}(t) \in \mathbb{N}$. We shall discuss specific examples in sections 3.2.3.5 and 3.2.3.6 and treat them in a whole section (section 3.2.5. In particular, master equations are indispensable for modeling chemical reactions or biological processes with small particle numbers. Specific applications in chemistry and biology will be presented in two separate chapters 4 and 5.

It is important to stress that the mathematical expressions of the three contributions to the general stochastic process represent a pure formalism that can be applied equally well to problems in physics, chemistry, biology, sociology, economics or other disciplines. Specific empirical knowledge enters the model in form of the parameters: the drift vector **A**, the diffusion matrix B, and the jump transition matrix W. By means of examples we shall show how physical laws are encapsulated in the regularities between parameters.

3.2.3 Examples of stochastic processes

In this section we present examples of stochastic processes with characteristic properties that will be used in the forthcoming applications: (i) the Liouville process, (ii) the Wiener process, (iii) the Ornstein-Uhlenbeck process, (iv) the Poisson process, and (v) the random walk in one dimension.

3.2.3.1 Liouville equation

The Liouville equation is the straightforward link between deterministic motion and stochastic processes. As shown in figure 3.1 all elements of the jump transition matrix W and the diffusion matrix B are zero and what remains is a differential equation falling into the class of *Liouville equations* from classical mechanics. A Liouville equation is used commonly for the description of the deterministic motion of particles in *phase space*.[13] Following [93, p. 54] we show that deterministic trajectories are identical to solutions of the differential Chapman-Kolmogorov equation with D = 0 and W = 0 and then relate the result to Liouville's theorem in classical mechanics [184, 185].

The probability density $p\,(\mathbf{x},t)$ with sharp initial conditions $p\,(\mathbf{x},t_0) = \delta(\mathbf{x}-\mathbf{x}_0)$.[14] From the dCKE we obtain

$$\frac{\partial p\,(\mathbf{x},t)}{\partial t} \;=\; -\sum_i \frac{\partial}{\partial x_i}\Big(A_i(\mathbf{x},t)\,p(\mathbf{x},t)\Big)\;, \tag{3.35}$$

and the goal is to show equivalence to the differential equation

$$\frac{\mathrm{d}\boldsymbol{\xi}(t)}{\mathrm{dt}} \;=\; \mathbf{A}\big(\boldsymbol{\xi}(t),t\big) \;\;\text{with}\;\; \boldsymbol{\xi}(t_0) \,=\, \mathbf{x}_0 \tag{3.36}$$

[13] Phase space is an abstract space, which is particularly useful for the visualization of particle motion. The six independent coordinates of particle S_k are the position coordinates $\mathbf{q}_k = (q_{k1}, q_{k2}, q_{k3})$ and the (linear) momentum coordinates $\mathbf{p}_k = (p_{k1}, p_{k2}, p_{k3})$. In Cartesian coordinates they read: $\mathbf{q}_k = (x_k, y_k, z_k)$ and $\mathbf{p}_k = m_k \cdot \mathbf{v}_k$ with $\mathbf{v} = (v_x, v_y, v_z)$ being the velocity vector.

[14] For simplicity we write $p\,(\mathbf{x},t)$ instead of the conditional probability $p\,(\mathbf{x},t|\,\mathbf{x}_0,t_0)$ as long as the initial conditions $(\mathbf{x}_0,t_0)$ refer to the sharp density $p\,(\mathbf{x},t_0) = \delta(\mathbf{x}-\mathbf{x}_0)$.

in form of the common solution

$$p\,(\mathbf{x},t) \;=\; \delta\big(\mathbf{x}-\boldsymbol{\xi}(t)\big)\;. \tag{3.37}$$

The proof is done by direct substitution

$$\begin{aligned}\sum_i \frac{\partial}{\partial x_i}\Big(A_i(\mathbf{x},t)\,\delta\big(\mathbf{x}-\boldsymbol{\xi}(t)\big)\Big) \;&=\; \sum_i \frac{\partial}{\partial x_i}\Big(A_i\big(\boldsymbol{\xi}(t),t\big)\;\delta\big(\mathbf{x}-\boldsymbol{\xi}(t)\big)\Big) \;=\\ &=\; \sum_i\Big(A_i\big(\boldsymbol{\xi}(t)\big)\,\frac{\partial}{\partial x_i}\;\delta\big(\mathbf{x}-\boldsymbol{\xi}(t)\big)\Big)\,,\\ \text{and}\quad \frac{\partial}{\partial t}\;\delta\big(\mathbf{x}-\boldsymbol{\xi}(t)\big) \;&=\; -\sum_i\Big(\frac{\mathrm{d}\xi_i(t)}{\mathrm{dt}}\cdot\frac{\partial}{\partial x_i}\;\delta\big(\mathbf{x}-\boldsymbol{\xi}(t)\big)\Big)\,.\end{aligned}$$

Making use of equation (3.36) we see that the sums in the expressions on the last two lines are equal. □

Deterministic motion as described be equation (3.36) is a special case of a Markov process in which the distribution $p\,(\mathbf{x},t)$ degenerates to a Dirac delta-function. We may relax the initial conditions $p\,(\mathbf{x},t_0) = \delta(\mathbf{x}-\mathbf{x}_0) \to p\,(\mathbf{x},t_0) \;=\; p\,(\mathbf{x}_0)$ and then the result is a distribution migrating through space with unchanged shape instead of a delta function travelling on a single trajectory (see equation (3.39') below).

The following part on Liouville's equation[15] illustrates how empirical science – here Newtonian mechanics – enters a formal stochastic equation. In Hamiltonian mechanics [117, 118] dynamical systems may be represented by a *density function* or *classical density matrix* $\varrho\,(\mathbf{q},\mathbf{p})$ in phase space. The density function allows for the calculation of system properties. Commonly it is normalized such that the expected total number of particles is the integral over phase space:

$$N \;=\; \int\cdots\int \varrho\,(\mathbf{q},\mathbf{p})\,(\mathrm{d}q)^n(\mathrm{d}p)^n\;.$$

The evolution of the system is described by a time dependent density that is commonly denoted as $\varrho\,\big(\mathbf{q}(t),\mathbf{p}(t),t\big)$ with $\varrho\,\big(\mathbf{q}_0,\mathbf{p}_0,t_0\big)$ being the initial conditions. For a particle S_k the generalized spatial coordinated q_{ki} are related to conjugate momenta p_{ki} by Newton's equations of motion

$$\frac{\mathrm{d}p_{ki}}{\mathrm{dt}} \;=\; f_{ki}(\mathbf{q}) \quad\text{and}\quad \frac{\mathrm{d}q_{ki}}{\mathrm{dt}} \;=\; \frac{1}{m_k}\,p_{ki};\quad i=1,2,3\;,$$

where f_{ki} is the component of the force acting on particle S_k in the direction of q_{ki} and m_k the particle mass, respectively. Liouville's theorem based on Hamiltonian mechanics of an n particle system makes a statement on the

[15] The name Liouville equation has been created by Josiah Willard Gibbs [98].

evolution of the density ϱ

$$\frac{\mathrm{d}\varrho\,(\mathbf{q},\mathbf{p},t)}{\mathrm{dt}} \;=\; \frac{\partial\varrho}{\partial t} + \sum_{k=1}^{n}\sum_{i=1}^{3}\left(\frac{\partial\varrho}{\partial q_{ki}}\,\frac{\mathrm{d}q_{ki}}{\mathrm{dt}} + \frac{\partial\varrho}{\partial p_{ki}}\,\frac{\mathrm{d}p_{ki}}{\mathrm{dt}}\right) \;=\; 0 \qquad (3.38)$$

Insertion of the individual time derivatives yields:

$$\frac{\partial\varrho\,(\mathbf{q},\mathbf{p},t)}{\partial t} \;=\; -\sum_{k=1}^{n}\sum_{i=1}^{n}\left(\frac{1}{m_i}p_{ki}\,\frac{\partial}{\partial q_{ki}}\varrho\,(\mathbf{q},\mathbf{p},t) \;+\; f_{ki}\,\frac{\partial}{\partial p_{ki}}\varrho\,(\mathbf{q},\mathbf{p},t)\right). \qquad (3.39)$$

Equation (3.39) is already of the form of a differential Chapman-Kolmogorov equation (3.35) with B = 0 and W = 0 as follows from

$$\begin{aligned}\varrho\,(\mathbf{q},\mathbf{p},t) \;&\equiv\; p\,(\mathbf{x},t) \;\text{ with}\\ \mathbf{x} \;&\equiv\; (q_{11},\ldots,q_{n3},p_{11},\ldots,p_{n3}) \;\text{ and}\\ \mathbf{A} \;&\equiv\; \left(\frac{1}{m_1}p_{11},\ldots,\frac{1}{m_n}p_{n3},f_{11},\ldots,f_{n3}\right)\end{aligned}$$

where the $6n$ coordinates represent the $3n$ coordinates determining the positions and the $3n$ coordinates for the linear momenta of n particles.

Finally, we indicate how the case of an extended probability density is handled in equation (3.35). The density function is the expectation value of the probability distribution,

$$\varrho\big(\mathbf{q}(t),\mathbf{p}(t),t\big) \;=\; E\Big(\varrho\big(\mathbf{q}(t),\mathbf{p}(t),t\big)\Big)\,, \qquad (3.40)$$

and it fulfils the Chapman-Kolmogorov equation:

$$\frac{\partial\varrho\,(\mathbf{q},\mathbf{p},t)}{\partial t} \;=\; -\sum_{i=1}^{3n}\left(\frac{1}{m_i}p_i\,\frac{\partial}{\partial q_i}\varrho\,(\mathbf{q},\mathbf{p},t) \;+\; f_i\,\frac{\partial}{\partial p_i}\varrho\,(\mathbf{q},\mathbf{p},t)\right). \qquad (3.39')$$

The Liouville equation states the conservation of density in phase space or in other words the distribution function $\varrho\,(\mathbf{q},\mathbf{p},t)$ is constant along any trajectory in phase space.

3.2.3.2 Wiener process and diffusion equation

The Wiener process named after the American mathematician and logician Norbert Wiener is fundamental in many aspects. It is often used synonymous to Brownian motion or *white noise* and describes among other things diffusion due to random fluctuations caused by thermal motion. The fluctuation driven random variable is denoted by $\mathcal{W}(t)$ and is characterized by the cumulative probability distribution,

$$P\left(\mathcal{W}(t) \leq w\right) = \int_{-\infty}^{w} p\left(v,t\right) \mathrm{d}v \ .$$

From the point of view of stochastic processes the probability density of the Wiener process is the solution of the differential Chapman-Kolmogorov equation in one variable with a diffusion term $\mathrm{B} = D = 1$, zero drift $\mathbf{A} = 0$ and no jumps $\mathrm{W} = 0$:

$$\frac{\partial p\left(w,t\right)}{\partial t} \ = \ \frac{1}{2} \, \frac{\partial^2}{\partial w^2} \, p\left(w,t\right) \ \text{ with } \ p\left(w,t_0\right) = \delta(w - w_0) \ . \tag{3.41}$$

Again a sharp initial condition (w_0, t_0) is assumed and we write for short $p\left(w,t\right) = p\left(w,t|w_0,t_0\right)$.

In physics and chemistry equation (3.41) – apart from the factors $1/2$ and D, respectively – occurs in connection with particle numbers or concentrations as functions of space and time: $c(x,t)$ in the one-dimensional case, which fulfils

$$\frac{\partial \, c\left(x,t\right)}{\partial t} \ = \ D \, \frac{\partial^2}{\partial x^2} \, c\left(x,t\right) \ \text{ with } \ c\left(x,t_0\right) = c_0(x) \tag{3.42}$$

as initial condition. Equation (3.42) is called *diffusion equation*,[16] because $c(x,t)$ describes the spreading of concentrations in homogeneous media driven by thermal molecular motion (for a detailed mathematical description of diffusion see, for example, [44]). The parameter D is called the diffusion coefficient and here it is assumed to be a constant. The diffusion equation has been derived first by Adolf Fick in 1855 [239]. Replacing the concentration by the temperature distribution in an one-dimensional object $c(x,t) \Leftrightarrow u(x,t)$ and the diffusion constant by the thermal diffusivity, $D \Leftrightarrow \alpha$, the diffusion equation (3.42) becomes the heat equation, which describes the distribution of heat in a given region over time.

Solutions of equation (3.41) can be derived readily by means of the characteristic function

$$\phi(s,t) \ = \ \int_{-\infty}^{+\infty} \mathrm{d}w \, p\left(w,t\right) \, \exp(\mathrm{i}s \, w) \ .$$

First we derive a differential equation for the characteristic function by applying integration by parts twice.[17] The first and second integration steps yield

---

[16] We distinguish the two formally identical equations (3.41) and (3.42), because the interpretation is different: The first equation (3.41) describes the evolution of a probability distribution with the conservation relation $\int \mathrm{d}w \, p\left(w,t\right) = 1$, whereas the second equation (3.42) deals with a concentration profile, which fulfils $\int \mathrm{d}x \, c\left(x,t\right) = c_{\mathrm{tot}}$ corresponding to mass conservation. In case of the heat equation the conserved quantity is total heat.

[17] Integration by parts is a standard integration method in calculus. It is encapsulated in the formula

$$\dot{\imath}s\,\phi(s,t) \;=\; p\,(w,t)\,e^{\dot{\imath}sw}\,\Big|_{-\infty}^{\infty} \;-\; \int_{-\infty}^{+\infty} \mathrm{d}w\,\frac{\partial p\,(w,t)}{\partial w}\,\exp(\dot{\imath}s\,w) \quad \text{and}$$

$$-\,s^2\,\phi(s,t) \;=\; \frac{\partial p\,(w,t)}{\partial w}\,e^{\dot{\imath}sw}\,\Big|_{-\infty}^{\infty} \;-\; \int_{-\infty}^{+\infty} \mathrm{d}w\,\frac{\partial^2 p\,(w,t)}{\partial w^2}\,\exp(\dot{\imath}s\,w)\;.$$

The function $p\,(w,t)$ is a probability density and accordingly has to vanish in the limits $w \to \pm\infty$. The same is true for the first derivatives, $\partial p\,(w,t)/\partial w$. Differentiation of $\phi(s,t)$ in equation (2.28) with respect to t and using equation (3.41) we obtain

$$\frac{\partial \phi(s,t)}{\partial t} \;=\; -\frac{1}{2}\,s^2\,\phi(s,t) \tag{3.43}$$

Next we compute the characteristic function by integration:

$$\phi(s,t) \;=\; \phi(s,t_0)\cdot\exp\Big(-\frac{1}{2}\,s^2\,(t-t_0)\Big)\;. \tag{3.44}$$

With the initial condition $\phi(s,t_0) = \exp(\dot{\imath}s\,w_0)$ we complete the characteristic function

$$\phi(s,t) \;=\; \exp\Big(\dot{\imath}s\,w_0 - \frac{1}{2}\,s^2\,(t-t_0)\Big) \tag{3.45}$$

and eventually obtain the probability density through inverse Fourier transformation

$$p\,(w,t|w_0,t_0) \;=\; \frac{1}{\sqrt{2\pi\,(t-t_0)}}\,\exp\left(-\frac{(w-w_0)^2}{2\,(t-t_0)}\right)\;. \tag{3.46}$$

Hence, the density function of the Wiener process is a normal distribution with expectation value and variance,

$$E\Big(\mathcal{W}(t)\Big) \;=\; w_0 \quad \text{and} \quad \sigma(t)^2 \;=\; E\Big(\big(\mathcal{W}(t)-w_0\big)^2\Big) \;=\; t-t_0\;, \tag{3.47}$$

respectively. The standard deviation, $\sigma(t) = \sqrt{t-t_0}$, is proportional to the square root of the time elapsed since the start of the process, $t-t_0$, and fulfils the famous $\sqrt{t}$-law. Starting the Wiener process at time $t_0 = 0$ at the origin $w_0 = 0$ yields $E\big(\mathcal{W}(t)\big) = 0$ and $\sigma\big(\mathcal{W}(t)\big)^2 = t$. An initially sharp distribution spreads in time as illustrated in figure 3.7, and this is precisely what is experimentally observed in diffusion. The infinite time limit of (3.46) is a

---

$$\int_a^b u(x)\,v'(x)\,\mathrm{d}x \;=\; u(x)\,v(x)\,\Big|_a^b \;-\; \int_a^b u'(x)\,v(x)\,\mathrm{d}x\;.$$

Characteristic functions are especially well suited for partial integration, because exponential functions, $v(x) = \exp(\dot{\imath}sx)$, can be easily integrated and probability densities $u(x) = p\,(x,t)$ as well as their first derivatives $u(x) = \partial p\,(x,t)/\partial x$ vanish in the limits $x \to \pm\infty$.

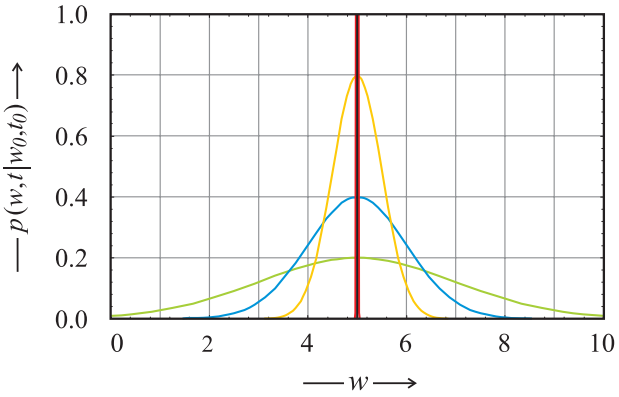

Fig. 3.7 Probability density of the Wiener process. In the figure we show the conditional probability density of the Wiener process, which is identical with the normal distribution (figure 1.19),

$$p(w,t|w_0,t_0) = \exp\Big(-(w-w_0)^2/\big(2(t-t_0)\big)\Big)/\sqrt{2\pi(t-t_0)}.$$

The values used are $w_0 = 5$ and $t - t_0$ =0 (black), 0.01 (red), 0.5 (yellow), 1.0 (blue), and 2.0 (green). The initially sharp distribution, $p(w,t|w_0,t_0) = \delta(w-w_0)$ spreads with increasing time until it becomes completely flat in the limit $t \to \infty$.

uniform distribution $\mathcal{U}(w) = 0$ on the whole real axis and hence $p\,(w,t|w_0,t_0)$ vanishes in the limit $t \to \infty$,.

Although the expectation value $E\big(\mathcal{W}(t)\big) = w_0$ is well defined and independent of time in the sense of a martingale, the mean square $E\big(\mathcal{W}(t)^2\big)$ becomes infinite as $t \to \infty$. This implies that the individual trajectories, $\mathcal{W}(t)$, are extremely variable and diverge after short time (see, for example, the five trajectories of the forward equation in figure 3.3). We shall encounter such a situation with finite mean but diverging variance also in biology in the case of multiplication as a pure birth and death process (chapter 5): The mean although well defined looses its value in practice when the standard deviation becomes larger than the expectation value.

An important generalization of Wiener processes is the Gaussian process $\mathcal{X}_t$ with $t \in \mathcal{T} = (t_1,\ldots,t_n)$, for which any finite linear combination of samples has a joint normal distribution. The Gaussian property can be defined in terms of normal distributions: $(\mathcal{X}, t \in \mathcal{T})$ is Gaussian if and only if for every finite index set $t_1,\ldots,t_n$ there exist real numbers μ_k and σ_{kl} with $\sigma_{kk} > 0$ such that

$$E\left(\exp\Big(\dot{u}\sum_{i=1}^{n} t_i\mathcal{X}_{t_i}\Big)\right) \;=\; \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}t_it_j + \dot{u}\sum_{i=1}^{n}\mu_it_i\Big)\,, \qquad (3.48)$$

where μ_k $(k = 1, \ldots, n)$ are the mean value of the variables $\mathcal{X}_k$ and σ_{kl} $(k, l = 1, \ldots, n)$ are the elements of the covariance matrix Σ. Whereas the Wiener process is nonstationary since the variance grows with $\sqrt{t}$, the Ornstein-Uhlenbeck process (section 3.2.3.4) is an example for a stationary Gaussian process.

Continuity of sample paths of the Wiener process has been discussed already in subsection 3.2.2. Here we present proofs for two more features of the Wiener process: (i) individual trajectories, although being continuous, are nowhere differentiable and (ii) the increments of the Wiener Process are independent of each other. The nondifferentiability of the trajectories of the Wiener process has a consequence for the physical interpretation as Brownian motion: The moving particle has no defined velocity. Independence of increments is indispensable for the integration of stochastic differential equations (section 3.4).

In order to show nondifferentiability we consider the convergence behavior of the difference quotient

$$\lim_{h \to 0} \left| \frac{\mathcal{W}(t+h) - \mathcal{W}(t)}{h} \right| ,$$

where the random variable $\mathcal{W}$ has the conditional probability (3.46). Ludwig Arnold [7, p.48] illustrates the nondifferentiability in a heuristic way: The difference quotient $\big(\mathcal{W}(t+h) - \mathcal{W}(t)\big)/h$ follows the normal distribution $\mathcal{N}(0, 1/|h|)$, which diverges as $h \downarrow 0$ – the limit of a normal distribution with exploding variance is undefined – and hence for every bounded measurable set S we have

$$P\Big(\big(\mathcal{W}(t+h) - \mathcal{W}(t)\big)/h \in \mathrm{S}\Big) \to 0 \quad \text{as} \quad h \downarrow 0 .$$

Accordingly, the difference quotient cannot converge with nonzero probability to a random variable with finite value. The information on the convergence can be made more precise by using the law of the iterated logarithm: We obtain for almost every sample function and arbitrary ϵ in the interval $0 < \epsilon < 1$ as $h \downarrow 0$

$$\frac{\mathcal{W}(t+h) - \mathcal{W}(t)}{h} \geq (1-\epsilon) \sqrt{\frac{2 \ln(\ln(1/h))}{h}} \quad \text{infinitely often}$$

and simultaneously

$$\frac{\mathcal{W}(t+h) - \mathcal{W}(t)}{h} \leq (-1+\epsilon) \sqrt{\frac{2 \ln(\ln(1/h))}{h}} \quad \text{infinitely often} .$$

Since expressions on the r.h.s. approach $\pm\infty$ as $h \downarrow 0$, the difference quotient $\big(\mathcal{W}(t+h)-\mathcal{W}(t)\big)/h$ has with probability one, for every fixed t, the extended real line $[-\infty,+\infty]$ as its limit set of cluster points.

Because of the general importance of the Wiener process it is essential to present a proof for the statistical independence of nonoverlapping increments of $\mathcal{W}(t)$ [93, pp. 67,68]. We are dealing with a Markov process and hence we can write the joint probability as a product of conditional probabilities (3.15'), where $t_n - t_{n-1}, \ldots, t_1 - t_0$ are subintervals of the time span $t_n \geq t \geq t_0$

$$p\,(w_n, t_n; w_{n-1}, t_{n-1}; \ldots; w_0, t_0) \;=\; \prod_{i=0}^{n-1} p\,(w_{i+1}, t_{i+1}|\, w_i, t_i)\, p(w_0, t_0)\;.$$

Next we introduce new variables that are consistent with this partition: $\big(\Delta w_i \equiv \mathcal{W}(t_i) - \mathcal{W}(t_{i-1}), \Delta t_i \equiv t_i - t_{i-1}\big)\,\forall\, i = 1, \ldots, n$. Since $\mathcal{W}(t)$ is a Gaussian process the probability density of any partition is normally distributed and we express the conditional probabilities in terms of (3.46):

$$p\,(\Delta w_n, \Delta t_n; \Delta w_{n-1}, \Delta t_{n-1}; \ldots; w_0, t_0) \;=\; \prod_{i=i}^{n} \frac{\exp\left(-\frac{\Delta w_i^2}{2\Delta t_i}\right)}{\sqrt{2\pi\Delta t_i}}\, p\,(w_0, t_0)\;.$$

The joint probability distribution is factorized into distributions from individual intervals and provided the intervals don't overlap the increments Δw_i are stochastically independent random variables in the sense of section 1.6.3, and they are independent of the initial condition $\mathcal{W}(t_0)$. The independence relation is readily cast in precise form

$$\mathcal{W}(t) \;-\; \mathcal{W}(s) \;\text{ is independent of }\; \{\mathcal{W}(\tau)\}_{\tau \leq s} \;\text{ for any }\; 0 \leq s \leq t\;, \quad (3.49)$$

which will be used in the forthcoming sections on stochastic differential equations (section 3.4).

Applying equation (3.47) to the probability distribution within a partition we find for an the interval $\Delta t_k = t_k - t_{k-1}$:

$$E\big(\mathcal{W}(t_k) - \mathcal{W}(t_{k-1})\big) \;=\; E(\Delta w_k) \;=\; w_{k-1} \;\text{ and }\; \sigma^2(\Delta w_k) \;=\; t_k \;-\; t_{k-1}\,,$$

It is now straightforward to calculate the autocorrelation function, which is defined by

$$\begin{aligned} \langle \mathcal{W}(t)\mathcal{W}(s)|(w_0,t_0)\rangle \;&=\; E\big(\mathcal{W}(t)\mathcal{W}(s)|(w_0,t_0)\big) \;= \\ &=\; \iint \mathrm{d}w_t \mathrm{d}w_s\, w_t w_s\, p\,(w_t, t; w_s, s|w_0, t_0)\;. \end{aligned} \quad (3.50)$$

Substraction and addition of $\mathcal{W}(s)^2$ inside the expectation value yields

$$E\big(\mathcal{W}(t)\mathcal{W}(s)|(w_0,t_0)\big) \;=\; E\Big(\big(\mathcal{W}(t) - \mathcal{W}(s)\big)\mathcal{W}(s)\Big) \;+\; E\big(\mathcal{W}(s)^2\big)\;,$$

where the first term vanishes because of independence of the increments and the second term follows from (3.47):

$$E\big(\mathcal{W}(t)\mathcal{W}(s)|(w_0,t_0)\big) \;=\; \min\{t-t_0, s-t_0\} \;+\; w_0^2 \;, \tag{3.51}$$

and simplifies to $E\big(\mathcal{W}(t)\mathcal{W}(s)\big) = \min\{t,s\}$ for $w_0 = 0$ and $t_0 = 0$. This expectation value reproduces also the diagonal element, the variance σ, since for $s = t$ we find $E\big(\mathcal{W}(t)^2\big) = t$. In addition, several other useful relations can be derived from the autocorrelation relation. We summarize:

$$\begin{aligned} E\big(\mathcal{W}(t)-\mathcal{W}(s)\big) \;&=\; 0\,,\; E\big(\mathcal{W}(t)^2\big) \;=\; t\,,\; E\big(\mathcal{W}(t)\mathcal{W}(s)\big) \;=\; \min\{t,s\}\,, \\ E\Big(\big(\mathcal{W}(t)-\mathcal{W}(s)\big)^2\Big) \;&=\; E\big(\mathcal{W}(t)^2\big) \;-\; 2E\big(\mathcal{W}(t)\mathcal{W}(s)\big) \;+\; E\big(\mathcal{W}(t)^2\big) \;= \\ &=\; t \;-\; 2\min\{t,s\} \;+\; s \;=\; |t-s|\;, \end{aligned}$$

and remark that these results are not independent of the *càdlàg* convention for stochastic processes.

The Wiener process has the property of self-similarity: Assume that $\mathcal{W}_1(t)$ is a Wiener process. Then, for every $c > 0$,

$$\mathcal{W}_2(t) \;=\; \mathcal{W}(ct) \;=\; \sqrt{c}\,\mathcal{W}_1(t)$$

is also a Wiener process. Accordingly, we can change the scale at will and the process remains a Wiener process. The power of the scaling factor is called the Hurst factor H (see sections 3.2.3.8 and 3.2.4.3), and accordingly the Wiener process has $H = 1/2$. In one and two dimensions the Wiener process is *recurrent* implying that every trajectory will return to the origin. In three and higher dimensions this is not the case and the process is called *transient*. The three dimensional trajectory revisits the origin in 34 % of the cases only, and this value decreases further in higher dimensions. Joking one can say a drunken sailor finds his way back home for sure, but a drunken pilot only in one out of three trials.

The Wiener process is readily extended to higher dimension. For the multivariate Wiener process, defined as

$$\vec{\mathcal{W}}(t) \;=\; \Big(\mathcal{W}_1(t),\ldots,\mathcal{W}_n(t)\Big) \tag{3.52}$$

satisfying the Fokker-Planck equation

$$\frac{\partial p(\mathbf{w},t|\mathbf{w}_0,t_0)}{\partial t} \;=\; \frac{1}{2}\sum_i \frac{\partial^2}{\partial w_i^2}\, p(\mathbf{w},t|\mathbf{w}_0,t_0)\;. \tag{3.53}$$

The solution is a multivariate normal density

$$p(\mathbf{w},t|\mathbf{w}_0,t_0) \;=\; \frac{1}{\sqrt{2\pi\,(t-t_0)}}\,\exp\left(-\frac{(\mathbf{w}-\mathbf{w}_0)^2}{2\,(t-t_0)}\right)\;. \tag{3.54}$$

with mean $E\big(\vec{\mathcal{W}}(t)\big) = \mathbf{w}_0$ and variance-covariance matrix

$$\big(\boldsymbol{\Sigma}\big)_{ij} \;=\; E\Big(\big(\mathcal{W}_i(t) - w_{0i}\big)\big(\mathcal{W}_j(t) - w_{0j}\big)\Big) \;=\; (t - t_0)\,\delta_{ij}\ ,$$

where all off-diagonal elements – the covariances – are zero. Hence, Wiener processes along different Cartesian coordinates are independent.

The Wiener process $W = \big(\mathcal{W}(t), t \geq 0\big)$ is characterized by ten important features and definitions:

(i) initial condition $\mathcal{W}(t_0) = \mathcal{W}(0) \equiv 0$,
(ii) trajectories are continuous functions of $t \in [0, \infty[$,
(iii) expectation value $E\big(\mathcal{W}(t)\big) \equiv 0$,
(iv) correlation function $E\big(\mathcal{W}(t)\mathcal{W}(s)\big) = \min\{t, s\}$,
(v) Gaussian property implies that for any $(t_1, \ldots, t_n)$ the random vector $\big(\mathcal{W}(t_1), \ldots, \mathcal{W}(t_n)\big)$ is a Gaussian process, ,
(vi) moments $E\big(\mathcal{W}(t)^2\big) = t$, $E\big(\mathcal{W}(t) - \mathcal{W}(s)\big) = 0$, and $E\Big(\big(\mathcal{W}(t) - \mathcal{W}(s)\big)^2\Big) = |t - s|$,
(vii) increments of the Wiener process on non-overlapping intervals are independent, for $(s_1, t_1) \cap (s_2, t_2) = \emptyset$ the random variables $\mathcal{W}(t_2) - \mathcal{W}(s_2)$ and $\mathcal{W}(t_1) - \mathcal{W}(s_1)$ are independent,
(viii) nondifferentiability of trajectories $\mathcal{W}(t)$,
(ix) self-similarity of the Wiener process $\mathcal{W}_2(t) = \mathcal{W}(\gamma t) = \sqrt{\gamma}\mathcal{W}_1(t)$, and
(x) martingale property, for $\mathcal{W}_0^s = \mathcal{W}(u)\,\forall\, 0 \leq u \leq s\}$ we have $E\big(\mathcal{W}(t)|\mathcal{W}_0^s\big) = \mathcal{W}(s)$ and $E\Big(\big(\mathcal{W}(t) - \mathcal{W}(s)\big)^2 \;\big|\; \mathcal{W}_0^s\Big) = t - s$.

Out of these ten properties three will be most important for the goals we will pursue here:(i) continuity of sample paths, (ii) nondifferentiability of sample paths, and (iii) independence of increments.

3.2.3.3 Autocorrelation functions and spectra

Analysis of experimentally recorded or computer created trajectories is often largely facilitated by the usage of additional tools complementing moments and probability distributions since they can, in principle, be derived from a single recording. These tools are autocorrelation functions and spectra of random variables, which provide direct insight into the dynamics of the process, since they are dealing with relations between sample points collected at different times (for an extensive treatment of *time series analysis* see, for example, [302]).

The *autocorrelation function* of the random variable $\mathcal{X}(t)$ is a measure of the influence the value of $\mathcal{X}$ recorded at time θ, $x(\theta)$ has on the measurement of the same variable at time $\theta + \tau$

$$\begin{aligned} G(\tau) &= \langle \mathcal{X}(\theta)\mathcal{X}(\theta+\tau)\rangle = E\big(\mathcal{X}(\theta)\mathcal{X}(\theta+\tau)\big) = \\ &= \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathrm{d}\theta\, x(\theta)\, x(\theta+\tau)\ . \end{aligned} \tag{3.55}$$

The autocorrelation function is the time average of the product of two values recorded at different times. It is of high relevance in the analysis of experimental data because technical devices called *autocorrelators* have been built [232], which sample data and can record directly the autocorrelation function of a process under investigation.

Another relevant quantity is the *spectrum* or the spectral density of the quantity $x(t)$. In order to derive the spectrum, we construct a new variable $y(\omega)$ by means of the transformation $y(\omega) = \int_0^t \mathrm{d}\theta\, e^{\mathrm{i}\omega\theta}\, x(\theta)$. The spectrum is then obtained from y by performing the limit $t \to \infty$:

$$S(\omega) = \lim_{t\to\infty} \frac{1}{2\pi t}\,|y(\omega)|^2 = \lim_{t\to\infty} \frac{1}{2\pi t}\left|\int_0^t \mathrm{d}\theta\, e^{\mathrm{i}\omega\theta}\, x(\theta)\right|^2 . \tag{3.56}$$

The autocorrelation function and the spectrum are closely connected. By some calculations one finds

$$S(\omega) = \lim_{t\to\infty} \left(\frac{1}{\pi}\int_0^t \cos(\omega\tau)\,\mathrm{d}\tau\ \frac{1}{t}\int_0^{t-\tau} x(\theta)\, x(\theta+\tau)\,\mathrm{d}\theta\right) .$$

Under certain assumptions, which insure the validity of the interchanges of order, we may take the limit $t \to \infty$ and find

$$S(\omega) = \frac{1}{\pi}\int_0^\infty \cos(\omega\tau)\, G(\tau)\,\mathrm{d}\tau\ .$$

This result relates the Fourier transform of the autocorrelation function to the spectrum and can be cast in an even prettier form by using

$$G(-\tau) = \lim_{t\to\infty} \frac{1}{t}\int_{-\tau}^{t-\tau} \mathrm{d}\theta\, x(\theta)\, x(\theta+\tau) = G(\tau)$$

to yield the *Wiener-Khinchin theorem* named after Norbert Wiener and the Russian mathematician Aleksandr Khinchin

$$S(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\mathrm{i}\omega\tau}\, G(\tau)\,\mathrm{d}\tau \ \text{ and } \ G(\tau) = \int_{-\infty}^{+\infty} e^{\mathrm{i}\omega\tau} S(\omega)\,\mathrm{d}\omega\ . \tag{3.57}$$

Spectrum and autocorrelation function are related to each other by the Fourier transformation and its inversion.

Equation (3.57) allows for a straightforward proof that the Wiener process $\vec{\mathcal{W}}(t) = W(t)$ gives rise to *white noise* (subsection 3.2.3.2). Let $\mathbf{w}$ be a zero-mean random vector with the identity matrix as (auto)covariance or autocorrelation matrix:

$$E(\mathbf{w}) \;=\; \boldsymbol{\mu} \;=\; \mathbf{0} \;\text{ and }\; \operatorname{cov}(\mathcal{W},\mathcal{W}) \;=\; E(\mathbf{w}\mathbf{w}') \;=\; \sigma^2\,\mathbb{I}\;,$$

then the Wiener process $W(t)$ fulfils the relations,

$$\begin{aligned} \mu_W(t) \;&=\; E\big(\mathcal{W}(t)\big) \;=\; 0 \;\text{ and} \\ G_W(\tau) \;&=\; E\big(\mathcal{W}(t)\mathcal{W}(t+\tau)\big) \;=\; \delta(\tau)\;, \end{aligned}$$

defining it as a zero-mean process with infinite power at zero time shift. For the spectral density of the Wiener process we obtain:

$$S_W(\omega) \;=\; \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\hbar\omega\tau}\,\delta(\tau)\,\mathrm{d}\tau \;=\; \frac{1}{2\pi}\;. \tag{3.58}$$

The spectral density of the Wiener process is a constant and hence all frequencies in the noise are represented with equal weight. All colors are mixed with equal weight in light yields white light and this property of visible light gave the name for white noise, in case of *colored noise* the noise frequencies do not fulfil the uniform distribution. Pink or flicker noise, for example, has a spectrum close to $S(\omega) \propto \omega^{-1}$ and red or Brownian noise fulfils $S(\omega) \propto \omega^{-2}$.

The time average of a signal as expressed by an autocorrelation function is complemented by the *ensemble average*, $\langle\cdot\rangle$, or expressed by the expectation value of the corresponding random variable, $E(\cdot)$, which implies an (infinite) number of repeats of the same measurement. In case the assumption of *ergodic behavior* is true, the time average is equal to the ensemble average. Thus we find for a fluctuating quantity $\mathcal{X}(t)$ in the ergodic limit

$$E\big(\mathcal{X}(t),\mathcal{X}(t+\tau)\big) \;=\; \langle x(t)x(t+\tau)\rangle \;=\; G(\tau)\;.$$

It is straightforward to consider dual quantities which are related by Fourier transformation and get:

$$x(t) \;=\; \frac{1}{2\pi}\int \mathrm{d}\omega\, c(\omega)\, e^{\hbar\omega t} \;\text{ and }\; c(\omega) \;=\; \int \mathrm{d}t\, x(t)\, e^{-\hbar\omega t}\;.$$

We use this relation to derive several important results. Measurements refer to real quantities $x(t)$ and this implies: $c(\omega) = c^*(-\omega)$. From the condition of stationarity, $\langle x(t)x(t')\rangle = f(t-t')$ and does not depend on t otherwise follows

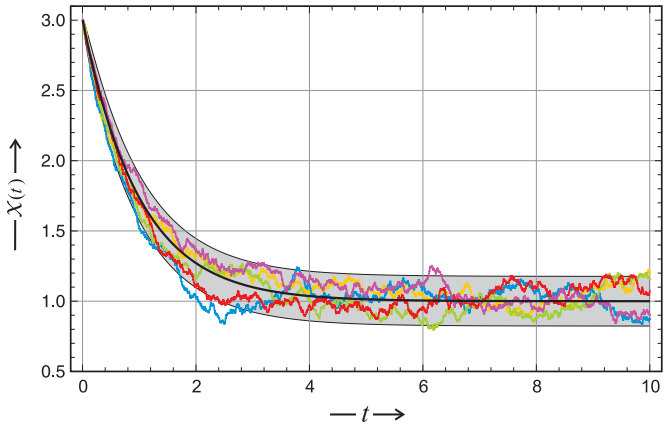

Fig. 3.8 The Ornstein-Uhlenbeck process. Individual trajectories of the process are simulated by $\mathcal{X}_{i+1} = \mathcal{X}_i\, e^{-k\,\vartheta} + \mu(1-e^{-k\,\vartheta}) + \sigma\sqrt{\frac{1-e^{-2k\,\vartheta}}{2k}}\,(\mathcal{R}_{0,1}-0.5)$, where $\mathcal{R}_{0,1}$ is a random number drawn by a random number generator from the uniform distribution on the interval $[0,1]$. The figures shows several trajectories differing only in the choice of seeds for *Mersenne Twister* as random number generator. The black lines represent the expectation value $E\big(\mathcal{X}(t)\big)$ and the curves $E\big(\mathcal{X}(t)\big) \pm \sigma\big(\mathcal{X}(t)\big)$. The area highlighted in grey is the confidence interval $E \pm \sigma$. Choice of parameters: $\mathcal{X}(0) = 3$, $\mu = 1$, $k = 1$, $\sigma = 0.25$, $\vartheta = 0.002$ or total time $t_f = 10$. Seeds: 491 (yellow), 919 (blue), 023 (green), 877 (red), and 733 (violet). For the simulation of the Ornstein-Uhlenbeck model see [105, 284].

$$\begin{aligned}\langle c(\omega)c^*(\omega')\rangle &= \frac{1}{(2\pi)^2}\iint \mathrm{d}t\,\mathrm{d}t'\,e^{-i\omega t+i\omega t'}\,\langle x(t)x(t')\rangle \;=\\ &= \frac{\delta(\omega-\omega')}{2\pi}\int \mathrm{d}\tau\,e^{i\omega\tau}\,G(\tau) \;=\; \delta(\omega-\omega')\,S(\omega)\;.\end{aligned}$$

The last expression relates not only the mean square $\big\langle |c(\omega)|^2 \big\rangle$ with the spectrum of the random variable, it shows also that stationarity alone implies that $c(\omega)$ and $c^*(\omega')$ are uncorrelated.

3.2.3.4 Ornstein-Uhlenbeck process and Fokker-Planck equation

The Ornstein-Uhlenbeck process is named after two Dutch physicists Leonard Ornstein and George Uhlenbeck [281] and represents presumably the simplest stochastic process that approaches a stationary state with a defined

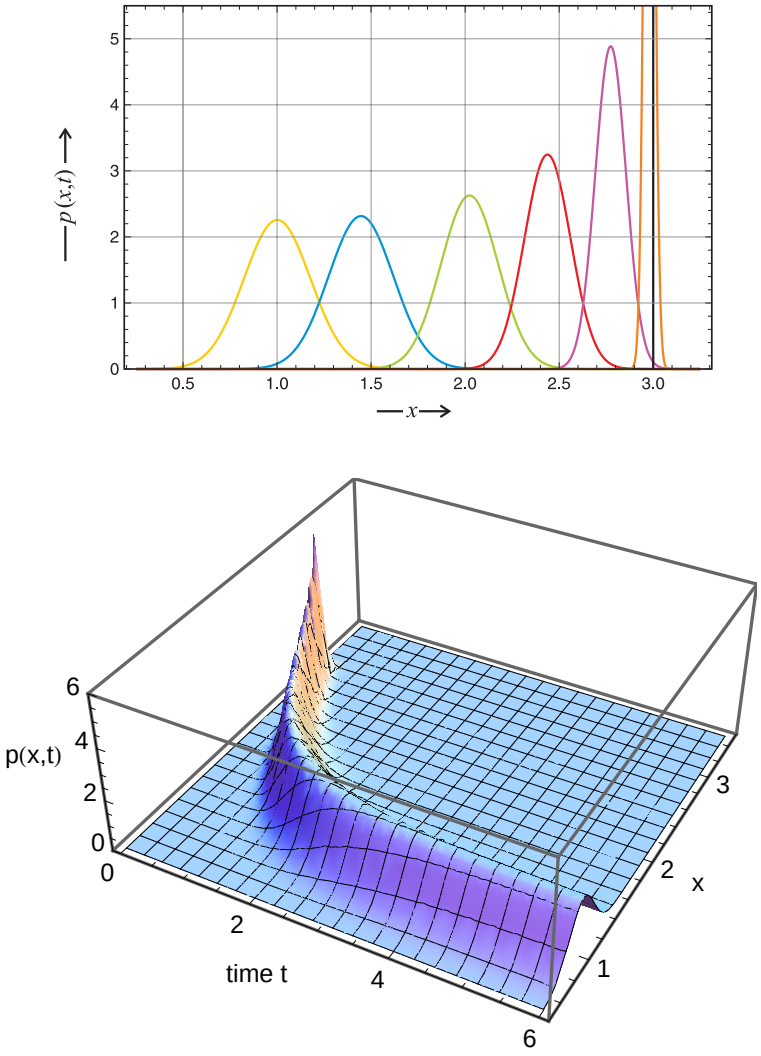

Fig. 3.9 The probability density of the Ornstein-Uhlenbeck process. Starting from the initial condition $p\,(x, t_0) = \delta(x - x_0)$ (black) the probability density (3.60) broadens and migrates until it reaches the stationary distribution (yellow). The lower plot presents an illustration in 3D. Choice of parameters: $x_0 = 3$, $\mu = 1$, $k = 1$, and $\sigma = 0.25$. Times: $t = 0$ (black), 0.12 (orange), 0.33 (violet), 0.67 (green), 1.5 (blue), and 8 (yellow).

variance.[18] It is a stationary Gaussian process and can be understood as the continuous-time analogue of the discrete first-order autoregressive $\big(AR(1)\big)$ process [116, 302]. The Ornstein-Uhlenbeck process found wide-spread applications, for example in economics for modeling irregular behavior of financial markets [288]. In physics it is among other applications a model for the velocity of a Brownian particle under the influence of friction. In essence, the Ornstein-Uhlenbeck process describes exponential relaxation to a stationary state or to an equilibrium superimposed by a Wiener process. Figure 3.8 presents several trajectories of the Ornstein-Uhlenbeck process, which show nicely the drift and the diffusion component of the individual runs.

The Fokker-Planck equation of the Ornstein-Uhlenbeck process for the probability density $p\,(x,t)$ of the random variable $\mathcal{X}(t)$ with the initial condition $p\,(x,t_0) = \delta(x - x_0)$ is of the form

$$\frac{\partial p\,(x,t)}{\partial t} \;=\; k\,\frac{\partial}{\partial x}\Big((x-\mu)\,p\,(x,t)\Big) \;+\; \frac{\sigma^2}{2}\,\frac{\partial^2 p\,(x,t)}{\partial x^2}\;, \tag{3.59}$$

with k is the rate parameter of the exponential decay, μ the expectation value of the random variable in the long-time or stationary limit, $\mu = \lim_{t\to\infty} E\big(\mathcal{X}(t)\big)$, and $\sigma^2/(2k)$ being the stationary variance. For the initial condition $p\,(x,0) = \delta(x - x_0)$ the probability density can be obtained by standard techniques

$$p\,(x,t) \;=\; \sqrt{\frac{k}{\pi\sigma^2\,(1-e^{-2k\,t})}}\;\exp\left(-\frac{k}{\sigma^2}\,\frac{(x-\mu-(x_0-\mu)e^{-kt})^2}{1-e^{-2kt}}\right)\;. \tag{3.60}$$

This expression can be easily checked by performing the two limits $t \to 0$ and $t \to \infty$. The first limit has to yield the initial conditions and it is indeed recalling a common definition of the Dirac delta-function.

$$\delta_\alpha(x) \;=\; \lim_{\alpha\to 0}\,\frac{1}{\alpha\sqrt{\pi}}\;e^{-x^2/\alpha^2}\;, \tag{3.61}$$

Inserting $\alpha^2 = \sigma^2(1-e^{-2kt})/k$ leads to

$$\lim_{t\to 0} p\,(x,t) \;=\; \delta(x-x_0)\;.$$

The long time limit of the probability density is calculated straightforwardly:

$$\lim_{t\to\infty} p\,(x,t) \;=\; \sqrt{\frac{k}{\pi\sigma^2}}\,e^{-k\,(x-\mu)^2/\sigma^2}\;,$$

which is a normal density with expectation value μ and variance $\sigma^2/(2k)$. □

[18] The variance of the Wiener process diverges in the limit, $\lim_{t\to\infty} \mathrm{var}\big(\mathcal{W}(t)\big) = \infty$. The same is true for the Poisson process and the random walk, which are discussed in the next two sections.

The evolution of probability density $p\,(x,t)$ from the δ-function at $t = 0$ to the stationary density $\lim_{t\to\infty} p\,(x,t)$ is shown in Fig. 3.9.

The Ornstein-Uhlenbeck process can be modeled efficiently also by the stochastic differential equation (SDE, see section 3.4.6.2):

$$\mathrm{d}x(t) \;=\; k\left(\mu - x(t)\right)\,\mathrm{dt} \;+\; \sigma\, dW(t)\;. \tag{3.62}$$

The individual trajectories shown in figure 3.8 [105, 284] were simulated by means of the following equation

$$\mathcal{X}_{i+1} \;=\; \mathcal{X}_i\, e^{-k\,\vartheta} \;+\; \mu(1 - e^{-k\,\vartheta}) \;+\; \sigma\sqrt{\frac{1 - e^{-2k\,\vartheta}}{2k}}\,(\mathcal{R}_{0,1} - 0.5)\;,$$

where $\vartheta = \Delta t/n_{st}$ is the number of steps per time interval Δt. The probability density can be derived, for example, from a sufficiently large ensemble of simulated trajectories. Expectation value and variance of the random variable $\mathcal{X}(t)$ can be calculated directly from the solution of the SDE (3.62) as shown in section 3.4.6.2.

3.2.3.5 Poisson process

The three processes discussed so far in this section were all dealing with continuous variables and their probability distributions. We continue by presenting two examples of processes dealing with discrete variables and pure jump processes according to equation (3.33c), which are modeled by master equations: the Poisson process and the discrete, one-dimensional random walk (see also section 3.2.5). To be stressed once more, master equations and related techniques to model and analyze stochasticity at low particle numbers are of particular importance in present day chemistry and biology.

The master equation (3.33c), rewritten for the discrete case by replacing the integral by a summation, is of the form[19]

$$\frac{\partial P(n,t)}{\partial t} \;=\; \int \mathrm{d}x\,\Big(\mathrm{W}(n\,|x,t)\,p\,(x,t) - \mathrm{W}(x\,|n,t)\,p\,(n,t)\Big) \;= \tag{3.33c'}$$

$$\;=\; \sum_{x=0}^{\infty}\Big(\mathrm{W}(n\,|x,t)\,P_x(t) - \mathrm{W}(x\,|n,t)\,P_n(t)\Big) \;=\; \frac{\mathrm{d}P_n(t)}{\mathrm{dt}}. \tag{3.33c'}$$

where we are assuming sharp initial conditions (n_0, t_0) or $P_n(t_0) = \delta_{n,n_0}$.[20] The matrix $W(m\,|n,t)$ is called the transition matrix that contains the prob-

[19] Riemann-Stieltjes integration converts the integral into a sum and since we are dealing with discrete events exclusively we use an index on the probability, $P_n(t)$,rather than apply an additional variable, $P(n,t)$.

[20] By δ_{ij} we denote the Kronecker delta named after the German mathematician Leopold Kronecker and means

abilities attributed to jump of variables, and from both equations follows that the diagonal elements, $W(n\,|n,t)$, cancel. The domain of the random variable is implicitly included in the domain of integration or summation, respectively.

The Poisson process is commonly applied to model certain classes of independent cumulative random events. These may be, for example, electrons arriving at an anode, customers entering a shop, telephone calls arriving at a switch board or e-mails being registered at an account. Aside from independence the requirement is an unstructured time profile of events or, in other words, the probability of occurrence of events i a constant. The cumulative number of these events is denoted by the random variable $\mathcal{N}(t) \in \mathbb{N}$. In other words $\mathcal{N}(t)$ is counting the number of arrivals and hence can only increase. The probability of arrival is assumed to be α per unit time, or $\alpha \cdot \Delta t$ is the expected number of events recorded in a time interval of length Δt. The Poisson process can also be interpreted as a *one-sided* random walk in the sense that the walker takes a step, for example to the right, with a probability α within a unit time interval. The transition frequencies are of the form

$$W(m|\,n,t) \;=\; \begin{cases} \alpha & \text{if } \; m \,=\, n+1\,, \\ 0 & \text{otherwise} \end{cases} \quad , \tag{3.63}$$

where the probability that two or more arrivals occur within the differential time interval dt is of measure zero. According to (3.33c) the master equation takes on the form

$$\frac{\mathrm{d}P_n(t)}{\mathrm{dt}} \;=\; \alpha \Big(P_{n-1}(t) \;-\; P_n(t) \Big) \tag{3.64}$$

with the initial condition $P_n(t_0) \,=\, \delta_{n,n_0}$. In other words, the number of arrivals recorded before $t = t_0$ is n_0. The interpretation of (3.64) is straightforward: the increase in the probability to have n recorded events between time t and $t +$ dt is proportional to the difference in probabilities between $n-1$ and n recorded events, because the elementary single arrival processes, $(n-1 \to n)$ and $(n \to n+1)$, increase or decrease the probability of n events, respectively.

The method of probability generating functions (section 2.2.1) is now applied for deriving solutions of the master equation (3.64). The probability generating function for the Poisson process is

$$g(s,t) \;=\; \sum_{n=0}^{\infty} P_n(t)\, s^n \;,\;\; |s| \le 1 \;\; \text{with} \;\; g(s,t_0) \,=\, s^{n_0} \;. \tag{2.24'))}$$

$$\delta_{ij} \begin{cases} 1 & \text{if } \; i = j \\ 0 & \text{if } \; i \neq j \end{cases}.$$

It represents the discrete analogue of the Dirac delta-function.

The time derivative of the generation function is obtained by insertion of equation (3.64)

$$\frac{\partial g(s,t)}{\partial t} \;=\; \sum_{n=0}^{\infty} \frac{\partial P_n(t)}{\partial t}\, s^n \;=\; \alpha \sum_{n=0}^{\infty} \Big(P_{n-1}(t) - P_n(t) \Big)\, s^n \;,$$

the first sums is readily evaluated

$$\sum_{n=0}^{\infty} \frac{\partial P_{n-1}(t)}{\partial t}\, s^n \;=\; s \sum_{n=0}^{\infty} \frac{\partial P_{n-1}(t)}{\partial t}\, s^{n-1} \;=\; s\, g(s,t)$$

and the second sum is identical to the definition of the generating function. This yields the equation for the generating function

$$\frac{\partial g(s,t)}{\partial t} \;=\; \alpha\,(s-1)\, g(s,t) \;. \tag{3.65}$$

Since the equation does not contain a derivative with respect to the dummy variable s we are dealing with an ODE and the solution by conventional calculus is straightforward:

$$\int_{\ln g(s,t_0)}^{\ln g(s,t)} \mathrm{d} \ln g(s,t) \;=\; \int_{t_0}^{t} \alpha(s-1)\, \mathrm{dt} \;,$$

which yields

$$g(s,t) \;=\; s^{n_0}\, e^{\alpha(s-1)(t-t_0)} \quad \text{or} \quad g(s,t) \;=\; e^{\alpha(s-1)t} \text{ for } (n_0 = 0, t_0 = 0) \tag{3.66}$$

with $g(s,0) = s^{n_0}$. The assumption $(n_0 = 0, t_0 = 0)$ is meaningful, because it implies that counting arrivals starts at time $t = 0$, and the expressions become especially simple: $g(0,t) = \exp(-\alpha t)$ and $g(s,0) = 1$. The individual probabilities $P_n(t)$ are obtained through expansion of the exponential function and equating the coefficients for the powers of s:

$$\begin{aligned} \exp\big(\alpha(s-1)t\big) \;&=\; \exp\big(\alpha\, st\big)\, e^{-\alpha t} \quad \text{and} \\ \exp\big(\alpha\, st\big) \;&=\; 1 \,+\, s\,\frac{\alpha t}{1!} \,+\, s^2\,\frac{(\alpha t)^2}{2!} \,+\, s^3\,\frac{(\alpha t)^3}{3!} \,+\, \ldots \;, \end{aligned}$$

and eventually we are obtaining the solution

$$P_n(t) \;=\; e^{-\alpha t}\,\frac{(\alpha t)^n}{n!} \;=\; e^{-\lambda}\,\frac{\lambda^n}{n!} \;, \tag{3.67}$$

which is the well-known Poisson distribution (2.30) with the expectation value $E\big(\mathcal{N}(t)\big) = \alpha t = \lambda$ and variance $\sigma^2\big(\mathcal{N}(t)\big) = \alpha t = \lambda$. Since the standard devi-

ation is $\sigma\big(\mathcal{N}(t)\big) = \sqrt{\alpha t}$ and accordingly the Poisson process fulfils perfectly the $\sqrt{N}$ relation for fluctuations.

It is easily verified that expectation value and variance can be directly obtained from the generating function through differentiation (2.25):

$$\begin{aligned} E\big(\mathcal{N}(t)\big) &= \frac{\partial g(s,t)}{\partial s}\Big|_{s=1} = \alpha\, t\ , \\ \sigma^2\big(\mathcal{N}(t)\big) &= \frac{\partial g(s,t)}{\partial s}\Big|_{s=1} + \frac{\partial^2 g(s,t)}{\partial s^2}\Big|_{s=1} - \left(\frac{\partial g(s,t)}{\partial s}\Big|_{s=1}\right)^2 = \alpha\, t\ , \end{aligned} \tag{3.68}$$

We remark that equation (3.64) can be solved also by using the characteristic function (section 2.2.3), which will be applied for the purpose of illustration in solving the master equation of the one-dimensional random walk (section 3.2.3.6).

The Poisson process can be viewed from a slightly different perspective by considering the arrival times of individual independent events as random variables $\mathcal{T}_1, \mathcal{T}_2, \cdots$. We shall assume that they are positive and follow an exponential density $\varrho(a,t) = a \cdot e^{-a\cdot t}$ with $a > 0$ and $\int_0^\infty \varrho(a,t)\, dt = 1$, and thus for each index j we have

$$P(\mathcal{T}_j \le t) = 1 - e^{-a\,t} \ \text{ and thus } \ P(\mathcal{T}_j > t) = e^{-a\,t},\ t \ge 0\ .$$

Independence of the individual events implies the validity of

$$P(\mathcal{T}_1 > t_1, \ldots, \mathcal{T}_n > t_n) = P(\mathcal{T}_1 > t_1)\ \ldots\ P(\mathcal{T}_n > t_n) = e^{-a(t_1+\ldots+t_n)}\ ,$$

which determines the joint probability distribution of the arrival times $\mathcal{T}_j$'s. The expectation value of the *inter-arrival* times, or times between consecutive arrivals, is simply given by $E(\mathcal{T}_j) = a^{-1}$. Clearly, the smaller a is, the longer will be the mean inter-arrival time, and thus a can be addressed as the intensity of flow. In comparison to the previous derivation we have $a \equiv \alpha$. For $\mathcal{S}_0 = 0$ and $n \ge 1$ we define by the cumulative random variable

$$\mathcal{S}_n = \mathcal{T}_1 + \ldots\, \mathcal{T}_n = \sum_{j=1}^{n} \mathcal{T}_j$$

the waiting time until the nth arrival. The event $\mathcal{I} = (\mathcal{S}_n \le t)$ implies that the nth arrival has occurred before time t. The connection between the arrival times and the cumulative number of arrivals, $\mathcal{N}(t)$, is easily performed and illustrates the usefulness of the dual point of view:

$$P(\mathcal{I}) = P(\mathcal{S}_n \le t) = P(\mathcal{N}(t) \ge n)\ .$$

More precisely, $\mathcal{N}(t)$ is determined by the whole sequence $(\mathcal{T}_j, j \ge 1)$, and depends on the elements ω of the sample space through the individual arrival times $\mathcal{T}_j$. In fact, we can compute the number of arrivals exactly by

$$\{\mathcal{N}(t) = n\} \;=\; \{\mathcal{S}_n \le t\} \;-\; \{\mathcal{S}_{n+1} \le t\} \;=\; \{\mathcal{S}_n \le t \le \mathcal{S}_{n+1}\}\;.$$

We may interpret this equation directly: there are exactly n arrivals in $[0,t]$ if and only if the arrival n occurs before t *and* the arrival $(n+1)$ occurs after t. For each value of t the probability distribution of the random variable $\mathcal{N}(t)$ is given by

$$P\big(\mathcal{N}(t) = n\big) \;=\; P\{\mathcal{S}_n \le t\} \;-\; P\{\mathcal{S}_{n+1} \le t\}\,,\; n \in \mathbb{N}_0\;,$$

where we used already the initial condition $\mathcal{S}_0 = 0$. As we have shown before this distribution of $\mathcal{N}(t)$ is the Poisson distribution $\pi(at) = \pi(\alpha t) = \pi(\lambda)$.

3.2.3.6 Continuous time random walk in one dimension

The random walk in one dimension is a classical and famous problem of probability theory, which we have used already to illustrate the properties of a martingale in section 3.2.1.2, where we made the assumption of discrete space and time: A walker moves along a line and takes steps to the left or to the right with equal probability and length l, and regularly after a constant waiting time τ. The location of the walker is thus $n \cdot l$ with n being an integer, $n \in \mathbb{Z}$. Here we keep the step size discrete but time is assumed to be continuous – *continuous time random walk* (CTRW), a probability that the walker takes a step is defined, and then the random walk can be readily modeled by a master equation. In the next section 3.2.3.8 we shall consider a random walk with probability distributions for the moves in space and time, step sizes and waiting times.

For the master equation we require transition probabilities per unit time, which are simply defined to be a constant, ϑ, for single steps and zero otherwise:

$$W(m|\,n,t) \;=\; \begin{cases} \vartheta & \text{if}\;\; m \,=\, n+1\,, \\ \vartheta & \text{if}\;\; m \,=\, n-1\,, \\ 0 & \text{otherwise} \end{cases}\;. \tag{3.69}$$

Hence, the master equation describing the evolution of the probability for the walker to be in location $n \cdot l$ at time t is

$$\frac{\mathrm{d}P_n(t)}{\mathrm{dt}} \;=\; \vartheta\left(P_{n+1}(t) \;+\; P_{n-1}(t) \;-\; 2\,P_n(t)\right)\;, \tag{3.70}$$

provided he started at location $n_0 \cdot l$ at time t_0: $P_n(t_0) = \delta_{n,n_0}$.

The master equation (3.70) can be solved by means of the time dependent characteristic function (see equations (2.28) and (2.28')):

$$\phi(s,t) \;=\; E(e^{\mathrm{i}s\,n(t)}) \;=\; \sum_n P_n(t)\,\exp(\mathrm{i}s\,n)\;. \tag{3.71}$$

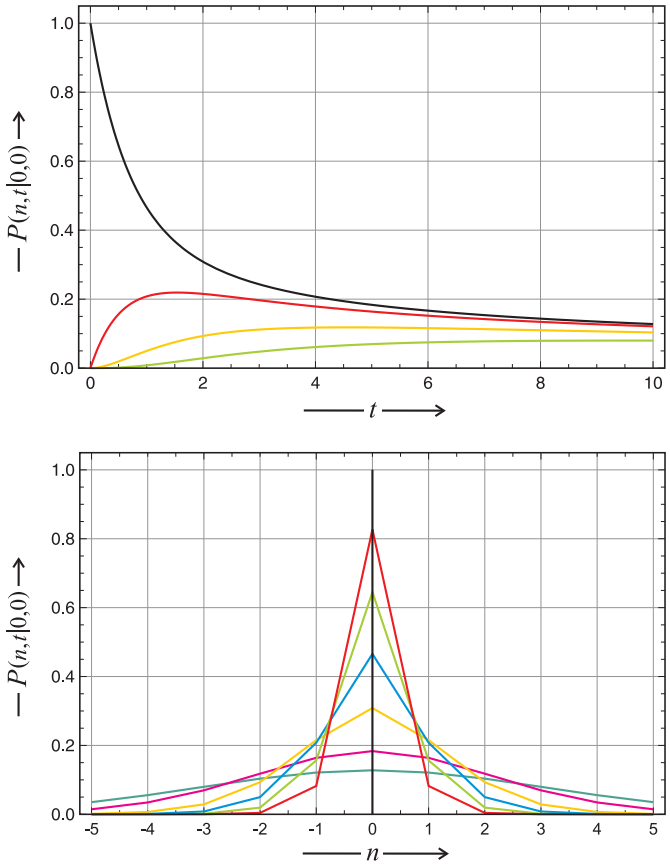

Fig. 3.10 Probability distribution of the random walk. The figure presents the conditional probabilities $P_n(t)$ of a random walker to be in location $n \in \mathbb{Z}$ at time t for the initial condition to be at $n = 0$ at time $t = t_0 = 0$. The upper part shows the dependence on t for given values of n: $n = 0$ (black), $n = 1$ (red), $n = 2$ (yellow), and $n = 3$ (green). The lower plot shows the probability distribution as a function of n at a given time t_k. Parameter choice: $\vartheta = 0.5$; $t_k = 0$ (black), 0.2 (red), 0.5 (green), 1 (blue), 2 (yellow), 5 (magenta), and 10 (cyan).

Combining (3.70) and (3.71) yields

$$\frac{\partial \phi(s,t)}{\partial t} \;=\; \vartheta \left(e^{\dot{\imath} u s} + e^{-\dot{\imath} u s} - 2 \right) \phi(s,t) \;=\; 2\,\vartheta \left(\cosh(\dot{\imath}\, u\, s) - 1 \right) \phi(s,t) \;.$$

Accordingly, the solution for the initial condition $n_0 = 0$ at $t_0 = 0$ is

$$\begin{aligned}\phi(s,t) &= \phi(s,0)\,\exp\Big(2\vartheta\,t\,\big(\cosh(\dot{\imath}\,s)-1\big)\Big) = \\ &= \exp\Big(2\vartheta\,t\,\big(\cosh(\dot{\imath}\,s)-1\big)\Big) = e^{-2\vartheta t}\exp\Big(2\vartheta\,t\big(\cosh(\dot{\imath}\,s)-1\big)\Big).\end{aligned} \tag{3.72}$$

Comparison of the coefficients for individual powers of s through insertion of

$$\cosh(\dot{\imath}s)-1 = \frac{(\dot{\imath}s)^2}{2!}+\frac{(\dot{\imath}s)^4}{4!}+\frac{(\dot{\imath}s)^6}{6!}+\ldots = -\frac{s^2}{2!}+\frac{s^4}{4!}-\frac{s^6}{6!}+\ldots$$

yields the individual probabilities:

$$P_n(t) = I_n(2\vartheta t)\,e^{-2\vartheta t}\,,\ n\in\mathbb{Z}\,. \tag{3.73}$$

where the pre-exponential term is written in terms of modified Bessel functions $I_k(\theta)$ with $\theta = 2\vartheta t$ (for details see [6, p. 208 ff.]), which are defined by

$$\begin{aligned}I_k(\theta) &= \sum_{j=0}^{\infty}\frac{(\theta/2)^{2j+k}}{j!(j+k)!} = \sum_{j=0}^{\infty}\frac{(\theta/2)^{2j+k}}{j!\,\Gamma(j+k+1)} = \\ &= \sum_{j=0}^{\infty}\frac{(\vartheta t)^{2j+k}}{j!(j+k)!} = \sum_{j=0}^{\infty}\frac{(\vartheta t)^{2j+k}}{j!\,\Gamma(j+k+1)}\,.\end{aligned} \tag{3.74}$$

The probability that the walker is found in his initial location, $n_0\,l$, for example, is given by

$$P_0(t) = I_0(2\vartheta\,t)\,e^{-2\vartheta\,t} = \big(1+(\vartheta\,t)^2+\frac{(\vartheta\,t)^4}{4}+\frac{(\vartheta\,t)^6}{36}+\ldots\big)\,e^{-2\vartheta\,t}$$

Illustrative numerical examples are shown in figure 3.10. It is straightforward to calculate first and second moments from the characteristic function $\phi(s,t)$ by means of equation (2.29) and the result is:

$$E\big(\mathcal{N}(t)\big) = n_0 \quad\text{and}\quad \sigma^2\big(\mathcal{N}(t)\big) = 2\vartheta\,(t-t_0)\,. \tag{3.75}$$

The expectation value is constant and coincides with the starting point of the random walk and the variance increases linearly with time.

The density function $P_n(t)$ allows for straightforward calculation of practically all interesting quantities. For example, we might like to know the probability that the walker reaches a given point at distance $n\cdot l$ from the origin within a predefined time span, which is simply obtained by $P_n(t)$ with $P_n(t_0)=\delta_{n,0}$ (figure 3.10). The probability distribution is symmetric because of the symmetric initial condition $P_n(t_0)=\delta_{n,0}$ and hence $P_n(t)=P_{-n}(t)$. For long times the probability density $P(n,t)$ becomes flatter and flatter and eventually converges to the uniform distribution over the spatial domain. In case $n\in\mathrm{Z}$ all probabilities vanish: $\lim_{t\to\infty}P_n(t)=0$ for all n.

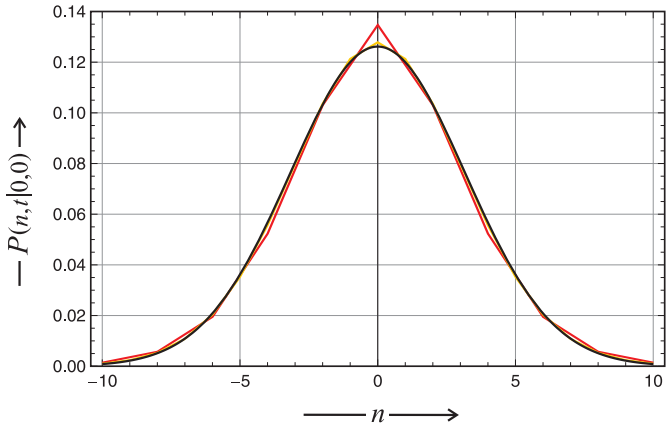

Fig. 3.11 Transition from random walk to diffusion. The figure presents the conditional probabilities $P(n,t|0,0)$ during convergence from a discrete space random walk to diffusion. The black curve is the normal distribution (2.36) resulting from the solution of the stochastic diffusion equation (3.41') with $D = 2\lim_{l\to 0}(l^2\vartheta) = 2$. The yellow curve is the random walk approximation with $l = 1$ and $\vartheta = 1$, the red curve was calculated with $l = 2$ and $\vartheta = 0.25$. Smaller step width of the random walk, $l \leq 0.5$, led to curves that are indistinguishable from the normal distribution. In order to obtain comparable curves, the probability distributions were scaled by a factor $\sigma = l^{-1}$. Choice of other parameters: $t = 5$.

3.2.3.7 From random walks to diffusion

In order to derive the stochastic diffusion equation (3.41) we start from a discrete time random walk of a single particle on an infinite one-dimensional lattice where the lattice sites are denoted by $i = \ldots, -1, 0, 1, \ldots$ or $i \in \mathbb{Z}$. Because of its general importance we present two derivations, (i) from the discrete time and space random walk model presented and solved in section 3.2.1.2, and (ii) from the continuous time discrete space random walk (CTRW) discussed in the previous section 3.2.3.6.

The particle is assumed to be at position i at time t and within a discrete time interval Δt it is obliged to jump to one of the neighboring sites, $i+1$ or $i-1$. This time elapsed between two jumps is called the *waiting time*. Spatial isotropy demands that the probabilities to jump to the right or to the left are the same and equal to one half. The probability to be at site 'i' at time $t + \Delta t$ is given by[21]

[21] It is worth to point at the difference between equations (3.70) and (3.8): The term containing $-P_i(t)$ is missing in the latter, because moving is obligatory in the discrete time model.

$$P_i(t+\Delta t) \;=\; \frac{1}{2}\,P_{i-1}(t) \;+\; \frac{1}{2}\,P_{i+1}(t)\;. \tag{3.8'}$$

Next we make a Taylor expression in time and truncate after the linear term in Δt assuming t is a continuous variable:

$$P_i(t+\Delta t) \;=\; P_i(t)\,\Delta t\,\frac{\mathrm{d}P_i(t)}{\mathrm{dt}} \;+\; \mathcal{O}\big((\Delta t)^2\big)\;.$$

Now we convert the site number into a continuous spatial variable, $i \Rightarrow x$ and $P_i(t) \Rightarrow p(x,t)$ and find

$$P_{i\pm1} \;=\; p(x,t) \pm \Delta x\,\frac{\partial p(x,t)}{\partial x} \;+\; \frac{(\Delta x)^2}{2}\,\frac{\partial^2 p(x,t)}{\partial x^2} \;+\; \mathcal{O}\big((\Delta x)^3\big)\;.$$

Here we truncate after the quadratic term in Δx because the terms with the first derivatives cancels, and obtain by insertion into equation (3.8') and omitting residuals

$$\Delta t\,\frac{\partial p(x,t)}{\partial t} \;=\; \frac{(\Delta x)^2}{2}\,\frac{\partial^2 p(x,t)}{\partial x^2}\;.$$

The next and final task is carrying out the limits to infinitesimal differences in time and space:

$$\lim_{\Delta t\to 0,\Delta x\to 0}\frac{(\Delta x)^2}{\Delta t} \;=\; D\;, \tag{3.76}$$

where D is called the diffusion coefficient. According to (3.76) the dimension of D is $[\mathrm{length}^2/\mathrm{time} = cm^2 \times sec^{-1}]$. Eventually we obtain the stochastic version of the diffusion equation

$$\frac{\partial p\,(x,t)}{\partial t} \;=\; \frac{D}{2}\,\frac{\partial^2 p(x,t)}{\partial x^2}\;, \tag{3.41'}$$

which is fundamental in physics and chemistry for the description of passive transport by thermal motion (see also equation (3.42) in section 3.2.3.2).

It is also straightforward to consider the continuous time random walk in the limit of continuous space. This is achieved by setting the distance traveled to $x = n \cdot l$ and performing the limit $l \to 0$. For that purpose we start from the characteristic function of the distribution in x,

$$\phi(s,t) \;=\; E\Big(e^{\mathrm{i}sx(t)}\Big) \;=\; \Phi(ls,t) \;=\; \exp\Big(2\vartheta\,t\,\big(\cosh(\mathrm{i}\,ls)-1\big)\Big)\;,$$

make use of the series expansion of the function cosh,

$$\cosh y \;=\; \sum\nolimits_{k=0}^{\infty}\frac{y^{2k}}{(2k)!} \;=\; 1 \;+\; \frac{y^2}{2!} \;+\; \frac{y^4}{4!} \;+\; \frac{y^6}{6!} \;+\; \ldots\;,$$

and take the limit of infinitesimally small steps, $\lim l \to 0$,

$$\lim_{l\to 0} \exp\Big(2\vartheta\, t\, \big(\cosh(\hat{u}\, ls) - 1\big)\, t\Big) \;=\; \lim_{l\to 0} \exp\Big(\vartheta\, t\, (-l^2 s^2 \,+\, \ldots)\Big) \;=$$
$$=\; \lim_{l\to 0} \exp(-s^2\, l^2\, \vartheta\, t) \;=\; \exp(-s^2\, D\, t/2)\;,$$

where we used the definition $D = 2\lim_{l\to 0}(l^2\vartheta)$ for the diffusion coefficient D (figure 3.11).[22] Since this is the characteristic function of the normal distribution we obtain for the probability density (2.36):

$$p\,(x,t) \;=\; \frac{1}{\sqrt{2\pi D t}}\, \exp\big(-x^2/(2Dt)\big) \tag{2.36}$$

for the sharp initial condition $\lim_{t\to 0} p(x,t) = p(x,0) = \delta(x)$. We could also have proceeded directly from equation (3.70) and expanded the right-hand side as a function of x up to second order in l, which yields again the stochastic diffusion equation

$$\frac{\partial p\,(x,t)}{\partial t} \;=\; \frac{D}{2}\, \frac{\partial^2 p\,(x,t)}{\partial x^2}\;, \tag{3.42}$$

where D stands for $2\lim_{l\to 0}(l^2\vartheta)$ as before.

The stochastic diffusion equation can be Fourier transformed in order to yield an equation for the Fourier transformed probability density $\hat{p}\,(|k|,t)$ with $|k|$ being the wave number with dimension $[l^{-1} = cm^{-1}]$:

$$\frac{\partial \hat{p}(|k|,t)}{\partial t} \;=\; -\frac{D}{2}\, |k|^2\, \hat{p}\,(|k|,t)\;. \tag{3.77}$$

The solution of (3.77) after normalization is of the form

$$\hat{p}\;(|k|,t) \;=\; \sqrt{\frac{Dt}{2\pi}}\, \exp\big(-\frac{D}{2}\, |k|^2\, t\big) \tag{3.78}$$

and represents a relaxation equation of the mode with a fixed value $|k|$ for the wave number.

3.2.3.8 Universality class of the continuous time random walk

In order to facilitate the comparison of normal diffusion and anomalous diffusion discussed in the next section 3.2.4 we present the one-dimensional continuous time random walk (CTRW) from a slightly different perspective [25, 213]. The random variable $\mathcal{X}(t)$ is the sum of all step increments:

$$\mathcal{X}(t) \;=\; \sum_{j=1}^{n} \xi_j \quad \text{with} \quad t \;=\; \sum_{j=1}^{n} \tau_j\;.$$

[22] The most straightforward way to perform the limit is to introduce a *scaling assumption* using a variable σ such that $l = l_0\sigma$ and $\vartheta = \vartheta_0/\sigma^2$. Then we have $l^2\vartheta = l_0^2\vartheta_0 = D$ and taking the limit $\sigma \to 0$ is trivial.

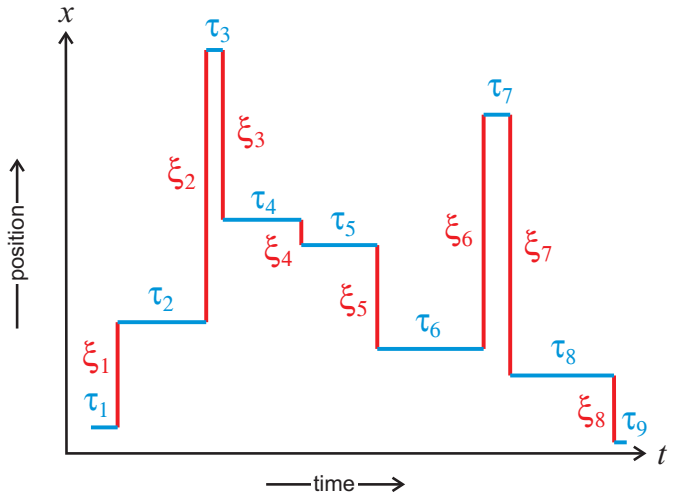

Fig. 3.12 Continuous time random walk model. Both, the jump lengths, ξ_k, and the waiting times, τ_k, are variable. The sketch indicates a case of high variability.

The model is built upon the concept that both, the *jump lengths* and the length of the time between two jumps denoted as *waiting time* are variable (figure 3.12),and have the joint density function

$$\begin{aligned} \psi(\xi,\tau) &= f(\xi)\,w(\tau) \quad \text{with} \\ w(\tau) &= \int_{-\infty}^{+\infty} \mathrm{d}\xi\,\psi(\xi,t) \quad \text{and} \quad f(\xi) = \int_{0}^{\infty} \mathrm{d}\tau\,\psi(\xi,\tau) \end{aligned} \tag{3.79}$$

being the two marginal distributions. This equation is based on the assumption that waiting times and jump lengths are independent random variables. In case they were coupled we had $\psi(\xi,\tau) = p(\xi|\tau)w(\tau) = p(\tau|\xi)f(\xi)$. Coupling could imply, for example, that it is impossible to jump a certain distance within a time span shorter than a minimum time required.

For the classification of random walks two moments of the distributions $w(\tau)$ and $f(\xi)$ are important:
(i) the *characteristic waiting time* $\overline{\tau} = \tau_w = \int_0^\infty \mathrm{d}\tau\,\tau\,w(\tau)$, and
(ii) the *variance of the jump length* $\overline{\xi^2} = 2\sigma^2 = \int_{-\infty}^{+\infty} \mathrm{d}\xi\,\xi^2\,f(\xi)$.[23]
In case of Brownian motion or normal diffusion both quantities, $\overline{\tau}$, and $\overline{\xi^2}$, are finite since the probability densities are Poissonian and Gaussian, respec-

[23] As in several previous examples we assume that the random walk is symmetric and started at the origin. Then the expectation value of the location of the particle stays at the origin and we have $\overline{\xi} = 0$ and $\overline{\xi}^{\,2} = 0$, and $\mathrm{var}(\xi) = \overline{\xi^2}$.

tively:

$$w(\tau) \;=\; \frac{1}{\tau_w}\;\exp\!\left(-\frac{\tau}{\tau_w}\right) \quad\text{and}\quad f(\xi) \;=\; \frac{1}{\sqrt{4\pi\sigma^2}}\;\exp\!\left(-\frac{\xi^2}{4\sigma^2}\right)\;.$$

The Laplace transform of $w(\tau)$ and the Fourier transform of $f(\xi)$ are of the asymptotic form

$$\hat{w}(u) \;=\; \int_0^\infty \mathrm{d}\tau\, w(\tau)\, e^{-u\tau} \;=\; 1 \;-\; \tau_w\, u \;+\; \mathcal{O}(u^2) \quad\text{and} \tag{3.80}$$

$$\hat{f}(k) \;=\; \int_{-\infty}^{+\infty} \mathrm{d}\xi\, f(\xi)\, e^{-2\pi i\, k\xi} \;=\; 1 \;-\; \sigma^2\, k^2 \;+\; \mathcal{O}(k^4)\;. \tag{3.81}$$

In both cases the transformed probability distributions are given in expressions that allow for direct readout of the universality exponents, which are $\alpha = 2$ for the spatial density $\hat{f}(k)$ and $\vartheta = 1$ for the temporal density.

As a matter of fact any pair of probability density functions with finite τ_w and σ^2 leads to the same asymptotic result and this is a beautiful manifestation of the central limit theorem (section 2.3.6): In the inner part of the transformed densities all representatives of the universality class of CTRWs with finite mean waiting time and positional variance fulfil equations (3.80) and (3.81) and the individuality of the densities comes into play only within the higher order terms $\mathcal{O}(\tau^2)$ and $\mathcal{O}(k^4)$.

Finally, we mention a feature that will be brought up again and generalized in the next section 3.2.4: the Wiener process or Brownian motion are self similar. A stochastic process is self-similar with Hurst index H, named after the British hydrologist Harold Edwin Hurst, if the two processes

$$(\mathcal{Y}(a\,t), t \geq 0) \quad\text{and}\quad (a^H \mathcal{Y}(t), t \geq 0)$$

with the same initial condition $\mathcal{Y}(0) = 0$ have the same finite-dimensional distribution for all $a \geq 0$. Expressed in popular language if you look on a self-similar process with a magnifying glass it looks the same as without the magnifier no matter how large the magnification factor is.

3.2.4 Lévy processes

Lévy processes were defined precisely in mathematical terms and analyzed in detail by the famous French mathematician Paul Lévy. Many stochastic processes from physics fall into this class, and Lévy processes are of particular importance in financial mathematics [5]. Examples of Lévy processes are Brownian motion (section 3.2.3.2), the Poisson process (section 3.2.3.5), the Cauchy process (section 3.2.1.4), and many others. In physics, Lévy pro-

cesses are used for example in the mathematical theory of anomalous diffusion [25, 213], in other forms of fractional kinetics, and Lévy flights were found to occur in foraging strategies of animals. We are interested here in Lévy processes, because they allow for a general analytic treatment combining all three classes of processes in the dCKE, drift, diffusion, and jump, and they can handle probability densities with heavy tails (section 2.4.6).

A Lévy process $\mathcal{X} = (\mathcal{X}(t), t \geq 0)$ is a stochastic process that satisfies the following four properties:

(i) the random variable $\mathcal{X}(t)$ has independent increments as expressed by the property that the variables $\mathcal{Z}_k = \mathcal{X}(t_k) - \mathcal{X}(t_{k-1})$ with $k = 1, 2, \ldots$ are statistically independent,

(ii) the increments $\mathcal{Z}_k$ of the random variable $\mathcal{X}(t)$ are stationary in the sense that the probability distributions of the increments $\mathcal{Z}_k$ depend only on the length of the time interval $t_k - t_{k-1}$ but do not depend explicitly on time t, and increments on equal time intervals are identically distributed,

(iii) the process starts at the origin, $\mathcal{X}_0 = 0$, with probability one, and

(iv) the trajectory of the random variable $\mathcal{X}(t)$ is at least piecewise *stochastically continuous* in the sense that it fulfils the relation

$$\lim_{t \to \tau} P(|\mathcal{X}(t) - \mathcal{X}(\tau)| > a) = 0$$

for all $a > 0$ and for all $\tau \geq 0$.

The conditions (i), (ii), and (iii) are fulfilled by a dCKE with the parameters

$$A(x,t) \to a\,, \;\; B(x,t) \to \frac{1}{2}\,\sigma^2\,, \text{ and } W(z|x,t) \to w(z-x)\,, \tag{3.82}$$

and for the initial condition (x_0, t_0) the dCKE has the form

$$\begin{aligned} \frac{\partial p(x,t)}{\partial t} \;&=\; -\; a\,\frac{\partial p(x,t)}{\partial x} \;+\; \frac{1}{2}\,\sigma^2\,\frac{\partial^2 p(x,t)}{\partial x^2} + \\ &\quad + \!\!\!\!\!\!\int\!\!\!\!\!\!-\, \mathrm{d}z\, w(z)\Big(p(x-z,t) - p(x,t)\Big)\,. \end{aligned} \tag{3.83}$$

Lévy processes are thus fully characterized by the Lévy-Khinchin triplet (a, σ^2, w) that is named after Paul Lévy and the Russian mathematician Aleksandr Khinchin. As follows from condition (ii) a Lévy process is a homogeneous Markov process.

As seen from equation (3.82) the choice of parameters for Lévy processes replaces the functions $A(x,t)$ and $B(x,t)$ by the constants a and $\frac{1}{2}\sigma^2$, and time is eliminated from the jump probability $W(z|x,t)$. In this sense the differential Chapman-Kolmogorov equation becomes the analogue of a linear equation and indeed the corresponding Liouville equation becomes exceedingly simple. For a deeper understanding of Lévy processes the analogy although superficial often becomes useful.

3.2.4.1 Characteristic function of Lévy processes

The characteristic function of a Lévy process starting at $t_0 = 0$ from $p(x,0) = \delta(x_0)$ is defined as stated in section 2.2.3

$$\phi(s,t) \;=\; \int_{-\infty}^{+\infty} \mathrm{d}x\, e^{\dot{\imath}\, sx}\, p(x,t)\ .$$

Insertion in equation (3.83) yields the differential equation [93, pp. 248-250]

$$\frac{\partial \phi(s,t)}{\partial t} \;=\; \left(\dot{\imath}\, a\, s \;-\; \frac{1}{2}\,\sigma^2 s^2 \;+ \fint_{-\infty}^{+\infty} \mathrm{d}u\, \left(e^{\dot{\imath}\, su} - 1\right) w(u)\right) \phi(s,t)\ ,$$

which is readily solved to yield the expression

$$\begin{aligned} \phi(s,t) \;&=\; \int_{-\infty}^{+\infty} \mathrm{d}x\, e^{\dot{\imath} sx}\, p(x,t|0,0) \;= \\ &=\; \exp\!\left(\left(\dot{\imath}\, a\, s \;-\; \frac{1}{2}\,\sigma^2 s^2 \;+ \fint_{-\infty}^{+\infty} \mathrm{d}u\, (e^{\dot{\imath} su} - 1)\, w(u)\right) t\right)\ . \end{aligned} \tag{3.84}$$

A principle value integral is needed, because of the possibility of a singularity $\lim_{u\to 0} w(u) = \infty$, which we define as

$$w(u) \;\approx\; |u|^{-\alpha-1} \;\text{ as }\; |u| \to 0\,,\ \text{povided } \alpha < 2\ .$$

For $\alpha \leq 1$ the process has finite intensity and there is no need for a principal value integral since $e^{\dot{\imath} su} - 1 \approx \dot{\imath} su$ near $u = 0$. However, for $1 < \alpha < 2$ the intensity is infinite and allowing for asymmetry we write the function $w(u)$ near $u = 0$ in the following way

$$w(u) \;=\; \begin{cases} \vartheta_-\, |u|^{-\alpha-1} & \text{if } u < 0 \text{ and} \\ \vartheta_+\, u^{-\alpha-1} & \text{if } u > 0\ . \end{cases}$$

Now the principal value integral is defined as

$$\begin{aligned} &\fint \mathrm{d}u\, w(u)\left(e^{\dot{\imath} su} - 1\right) \;\equiv \\ &\equiv\; \lim_{\epsilon\to 0}\left(\int_{-\infty}^{-\delta(\epsilon)} \mathrm{d}u\, w(u)\big(e^{\dot{\imath} su} - 1\big) \;+\; \int_{\epsilon}^{+\infty} \mathrm{d}u\, w(u)\big(e^{\dot{\imath} su} - 1\big)\right)\ . \end{aligned} \tag{3.85}$$

and the function $\delta(\epsilon)$ is given by

$$\delta(\epsilon)^{-\alpha+1} \;=\; \frac{\vartheta_+}{\vartheta_-}\,\epsilon^{-\alpha+1} \;+\; \kappa \tag{3.86}$$

which excludes the case $\alpha = 1$ that will be treated separately. We remark that this procedure leads to a cancellation of the divergence at the upper limit of the fist integral with that in the lower limit of the second integral for any value of κ provided $\alpha < 2$.

The precise definition and the evaluation of the principle value integral impede the analytical work on Lévy processes and can be circumvent by the Lévy-Khinchin formula that separates the principal value integral from the rest of the expression:

$$\begin{aligned} &⨍_{-1}^{1} \mathrm{d}u\, \dot{\imath}\, s u\, w(u) \;\equiv \\ &\lim_{\epsilon \to 0} \left(\int_{-1}^{-\delta(\epsilon)} \mathrm{d}u\, \dot{\imath}\, s u\, w(u) + \int_{\epsilon}^{1} \mathrm{d}u\, \dot{\imath}\, s u\, w(u) \right) \equiv \dot{\imath}\, a_S\, s \;. \end{aligned} \tag{3.87a}$$

In this way a_S can be incorporated into the drift constant and evaluated for any special case together with the arbitrary constant κ:

$$A \;=\; a + a_S \;. \tag{3.87b}$$

The final expression of the Lévy-Khinchin formula is then of the form

$$\begin{aligned} &\phi(s,t) \;= \\ &= \exp\!\left(\left(\dot{\imath}\, A\, s - \frac{1}{2}\, \sigma^2 s^2 + \int_{-\infty}^{+\infty} \mathrm{d}u \left(e^{\dot{\imath} s u} - 1 - \dot{\imath}\, s\, u\; \mathbf{1}_{]-1,1[}(u) \right) w(u) \right) t \right), \end{aligned} \tag{3.87c}$$

where the indicator function is used to exclude the range of the principal value integral

$$\mathbf{1}_{]-1,1[}(u) \;=\; \begin{cases} 1 & \text{if } |u| < 1 \text{ and} \\ 0 & \text{if } |u| \geq 1 \;. \end{cases} \tag{3.87d}$$

Although the characteristic function can be written down and solved for any Lévy process the probability density need not be expressible in analytic functions. The only three known exceptions for stable Lévy processes are the normal distribution (section 2.3.3), the Cauchy distribution (section 2.4.6), and the Lévy distribution (section 2.4.7). Next we need to define two basic properties of distribution functions: (i) *infinite divisibility* and (ii) *stability*.

3.2.4.2 Infinite divisibility and stability

The property of infinite divisibility is defined for a probability density $p\,(x)$ and demands that the random variable $\mathcal{X}$ with the density $p\,(x)$ can be partitioned into any arbitrary number n of independent random variables with $n \in \mathbb{N}_{>0}$ such that the sum $\mathcal{S}_n = \mathcal{X}_1 + \mathcal{X}_2 + \ldots + \mathcal{X}_n$ has the probability density $p\,(x)$. In general the probability distributions of the individual parts

$\mathcal{X}_k$ will be different and different from the density $p(x)$. Lévy processes are homogeneous Markov processes and they are infinitely divisible therefore.

A Lévy process, $(\mathcal{X}_t, t \geq 0)$ is called stable if every random variable $\mathcal{X}_t$ has a stable distribution [225]. The random variables of a stable distribution fulfil the equation

$$a\,\mathcal{X}_1 + b\,\mathcal{X}_2 \stackrel{d}{=} c\,\mathcal{X} + d\,, \tag{3.88}$$

wherein a and b are positive constants, c is some positive number dependent on a, b and the summation properties of $\mathcal{X}$, $d \in \mathbb{R}$, and the symbol 'd' above the equals sign means equality in distribution. *Stability* or *stability in the broad sense* is to be distinguished from *strict stability* or *stability in the narrow sense* in which case the equality 3.88 holds with $d = 0$ for all choices of a and b. A random variable is *symmetric stable* if it is stable and symmetrically distributed around zero, $\mathcal{X} \stackrel{d}{=} -\mathcal{X}$.

We demonstrate stability of a distribution by means of the normal distribution, and use the central limit theorem (CLT) for this purpose:

$$\begin{aligned} \mathcal{S}_n &= \sum_{i=1}^{n} \mathcal{X}_i \quad \text{with} \quad E(\mathcal{X}_i) = \mu\,,\ \mathrm{var}(\mathcal{X}_i) = \sigma^2 \ \forall\ i = 1, \ldots, n \\ E(\mathcal{S}_n) &= n\,\mu \quad \text{and} \quad \mathrm{var}(\mathcal{S}_n) = (n\,\sigma)^2 \ \forall\ i = 1, \ldots, n\,. \end{aligned} \tag{3.89}$$

From the two equations (3.88) and (3.89) follow the conditions for the constants a, b, c, and d:

$$\begin{aligned} \mu(a\mathcal{X}) = a\,\mu(\mathcal{X})\,,\ \mu(b\mathcal{X}) = b\,\mu(\mathcal{X})\,,\ \mu(c\mathcal{X}+d) = c\,\mu(\mathcal{X}) + d \ \Rightarrow \\ \Rightarrow\ d = (a+b-c)\,\mu \\ \mathrm{var}(a\mathcal{X}) = (a\sigma)^2\,,\ \mathrm{var}(b\mathcal{X}) = (b\sigma)^2\,,\ \mathrm{var}(c\mathcal{X}+d) = (c\sigma)^2 \ \Rightarrow \\ \Rightarrow\ c^2 = a^2 + b^2\,. \end{aligned}$$

The two conditions $d = (a+b-c)\,\mu$ and $c = \sqrt{a^2+b^2}$ with $d \neq 0$ are readily fulfilled for pairs of arbitrary positive constants $a, b \in \mathbb{N}_{>0}$ and accordingly, the normal distribution $\mathcal{N}(\mu, \sigma)$ is stable. Strict stability, on the other hand, requires $d = 0$ and this can be fulfilled by zero-centered normal distributions $\mathcal{N}(0, \sigma)$ only.

An other definition of stability [225] is presented here, because it introduces four parameters that are required to fully characterize a stable distribution:

(i) *characteristic exponent*: $\alpha \in]\,0, 2\,]$,
(ii) *skewness parameter*: $\beta \in [-1, 1\,]$,
(iii) *scale parameter*: $\gamma \geq 0$, and
(iv) *location parameter*: $\delta \in \mathbb{R}$.

The characteristic exponent α is also called *index of stability* and will turn out as the parameter determining asymptotic behaviour in the sense of the

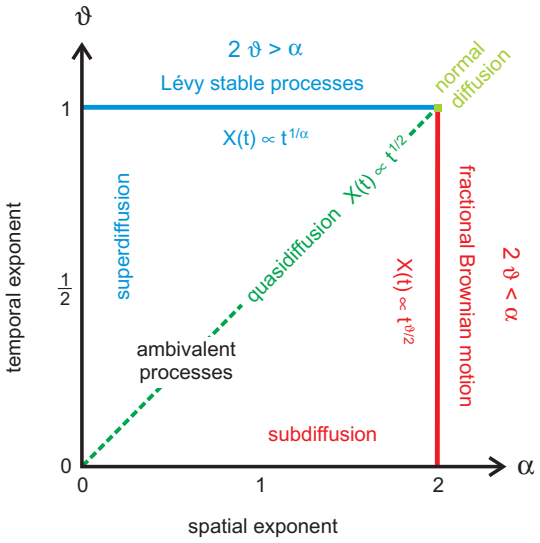

Fig. 3.13 Normal and anomalous diffusion. The figure sketches continuous time random walks (CTRW) as of the universality exponents of space, $0 < \alpha \leq 2$, and time, $0 < \vartheta \leq 1$. Lévy flights, normal diffusion, and fractional Brownian motion are limiting cases with the asymptotic behavior $(0 < \alpha < 2, \vartheta = 1)$, $(\alpha = 2, \vartheta = 1)$, and $(\alpha = 2, 0 < \vartheta < 1)$, respectively, of the general class of *ambivalent processes*.

spatial universality exponent. The parameters α and β together determine the shape of the distribution and are called *shape parameters* therefore. A scale parameter of $\gamma = 0$ is only meaningful as the limiting case of a degenerate distribution, which is concentrated at δ. The three stable distributions with analytical densities are:

(i) the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with $\alpha = 2$, $\beta = 0$, $\gamma = \frac{\sigma}{\sqrt{2}}$, $\delta = \mu$,
(ii) the Cauchy distribution $\mathcal{C}(\delta, \gamma)$ with $\alpha = 1$, $\beta = 0$, γ, δ and
(iii) the Lévy distribution $\mathcal{L}(\delta, \gamma)$ with $\alpha = \frac{1}{2}$, $\beta = 1$, γ, δ.

Easy access to extensive computing power, however, makes it possible to work highly efficiently with non-analytical distributions too and, after all, the characteristic functions (3.87c) are always available analytically.

3.2.4.3 Universality and self-similarity

Self-similarity and shapes of objects fitting fractal dimension – non-integer – dimensions are the topics of Benoît Mandelbrot's seminal book [193]. Self-similarity of stochastic processes has been mentioned already at the end of section 3.2.3.8 in the context of continuous time random walks. Here we

shall generalize the processes and discuss the properties of processes with the universality exponents $0 < \alpha \leq 2$ in space and $0 < \vartheta \leq 1$ in time.

The continuous time random walk considered here is assumed to Lévy distributed step lengths and waiting times. The derivation is analogous to section 3.2.3.8 and starts from the joint distribution $\psi(\xi, \tau) = f(\xi)w(\tau)$ where independence according to (3.79) is assumed. The spatial density $f(\xi)$ is given by a Lévy stable distribution that is defined in terms of its characteristic function, which we write here in form of the long distance (x) or short frequency (k) limit

$$\hat{f}(k; a, \alpha) \;=\; E\big(\exp(\dot{\imath}kx)\big) \;=\; \exp(-|ak|^\alpha) \;=\; 1 - |ak|^\alpha + \mathcal{O}(|k|^{2\alpha}) \,. \quad (3.90)$$

The condition to obtain an acceptable probability density – being nonnegative everywhere and normalizable – by inverse Fourier transform defines the domain for the universality exponent: $0 < \alpha \leq 2$. Thus we generalize now the previous account, which was exclusively dealing with $\alpha = 2$. In order to give a second illustrative example[24] we consider the Cauchy distribution

$$f(x) \;=\; \frac{a}{\pi(a^2 + x^2)} \quad \text{and} \quad \hat{f}(k) \;=\; \exp(-|ak|) \;=\; 1 - |ak| + \mathcal{O}(|k|^2) \,,$$

and the universality exponent is $\alpha = 1$.

The length of the CTRW is expressed by the *width* of the density $f(\frac{x}{n}; a, \alpha)$. Stability of the distribution of the random variable requires that a linear combination of independent copies of the variable has the same distribution as the copy:

$$f_n\Big(\sum_i x_i; a, \alpha\Big) \;=\; f(x_1; a, \alpha) \circ f(x_2; a, \alpha) \circ \ldots \circ f(x_n; a, \alpha) \,,$$

where '$\circ$' stands for convolution. Transformation in Fourier space yields

$$\hat{f}_n(k) \;=\; \prod_{i=1}^{n} \hat{f}(k_i; a, \alpha) \;=\; \exp\big(-|a\, n^{\frac{1}{\alpha}}\, k|^\alpha\big)$$

Backtransformation into space and time yields a generalization of the center limit theorem:

$$f_n\Big(\sum_i x_i; a, \alpha\Big) \;=\; f_n\bigg(\frac{x}{n^{\frac{1}{\alpha}}}; a, \alpha\bigg) \,. \quad (3.91)$$

In the context of a random walk the width of the distribution is related to the length of the walk. Equation (3.91) provides the answer for the exponential scaling of the walk lengths: $x(n) = n^{\frac{1}{\alpha}}$. In normal diffusion the length grows with $\sqrt{n}$, for Lévy stable distributions with $\alpha < 2$ the walks

[24] The case $\alpha = 2$ dealing with the normal distribution is extensively treated in section 3.2.3.8.

become longer because of heavier tails compared to the normal distribution. The corresponding trajectories are called Lévy flights and will be discussed at the final paragraph of this section. In polymer theory the length of the walk corresponds to the end-to-end distance of the polymer chain for which probability densities are available [267].

For the density of the waiting times we proceed similarly only the Fourier transform is replaced by a Laplace transform because time τ is limited to the nonnegative part of the axis, $\tau \geq 0$ and obtain the same result as for the random walk:

$$\hat{w}(u;\tau_w,\vartheta) \;=\; \int_0^\infty \mathrm{d}\tau\, w(\tau) e^{-u\tau} \;=\; \frac{1}{1+\tau_w u} \;=\; 1-\tau_w u+\mathcal{O}(u^2)\;,$$

and the universality exponent in $\vartheta = 1$.

The transformed joint distribution function can be obtained from the Montroll-Weiss equation named after Elliot Montroll and George Weiss [220]:

$$\hat{\psi}(k,u) \;=\; \frac{1\,-\,\hat{w}(u)}{u}\,\frac{1}{1\,-\,\hat{w}(u)\,\hat{f}(k)} \;\approx\; \frac{\theta\, u^{\vartheta-1}}{\theta u^{\vartheta}\,+\,\lambda|k|^{\alpha}}\;. \tag{3.92}$$

The expression can be easily checked by Laplace and Fourier transform of the density of normal diffusion with $\alpha = 2$, $\vartheta = 1$, and $\lambda = \sigma^2/2$:

$$\psi(x,t) \;=\; \frac{1}{\sqrt{4D\pi\, t}}\,\exp\Big(-\frac{x^2}{4D\,t}\Big) \;\longrightarrow\; \hat{\psi}(k,u) \;\approx\; \frac{1}{u\,+\,D\,k^2}\;,$$

where $D = \lambda\theta = \sigma^2/2\theta$ is the diffusion constant.

The next step is to perform inverse Laplace and inverse Fourier transform on the expression of the right hand side of equation (3.92) in order to yield the joint space-time distribution

$$\begin{aligned}\psi(x,t) \;&\approx\; \int_0^\infty \mathrm{d}u \int_{-\infty}^{+\infty} \mathrm{d}k\; e^{-\mathrm{i}|k|x+ut}\,\frac{\theta\, u^{\vartheta-1}}{\theta u^{\vartheta}\,+\,\lambda|k|^{\alpha}} \;=\\ &=\; \int_{-\infty}^{+\infty} \mathrm{d}k\; e^{-\mathrm{i}|k|x}\, E_{\vartheta}(-|k|^{\alpha}\, t^{\vartheta})\;.\end{aligned} \tag{3.93}$$

Herein we made use of the Mittag-Leffler function $E_\vartheta(-|k|^\alpha\, t^\vartheta)$, which is named after Magnus Gösta Mittag-Leffter, occurs in inverse Laplace transforms of functions of the Laplace transform parameter $p^\alpha(a+bp^\beta)$ [198], and has the form of an infinite series [216]:

$$E_\alpha(z) \;=\; \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(1+\alpha k)}\;,\; \alpha\in\mathbb{C}, \Re(\alpha)>0, z\in\mathbb{C}\;,$$

which leads to quite involved expressions except in some simple cases, for example $E_1(z) = \exp(z)$ or $E_0(z) = 1/(1-z)$ [123]. The evaluation of the inverse Fourier transform (3.93) is even more complicated but we shall need to consider only the form of the leading terms: The function of the form $\hat{\psi}(t^{\vartheta}|k|^{\alpha})$in the integrand becomes a function $\psi(\frac{x^{\alpha}}{t^{\vartheta}})$ after the inverse Fourier transform. If we express distance as a function of time we obtain eventually: $\frac{x^{\alpha}}{t^{\vartheta}} = c \rightarrow x(t) \propto t^{\frac{\vartheta}{\alpha}}$. The expression covers normal diffusion with $\alpha = 2$ and $\vartheta = 1$ leading to the relation $x(t) \propto \sqrt{t}$ and fractional diffusion with $\alpha = 2$ and $\vartheta < 1$ resulting in $x(t) \propto t^{\vartheta/2}$.

In figure 3.13 we summarize the results of this section. All continuous time random walks are characterized by two universality exponents, $0 < \alpha \leq 2$ and $0 < \vartheta \leq 1$, for scaling behavior in space and time. Normal diffusion is the limiting case with $\alpha = 2$ and $\vartheta = 1$. The probability densities of time steps or waiting times and jump length, the Poisson distribution and the normal distribution, respectively, have both finite expectation values and variances. Lévy stable distributions with $\alpha < 2$ have heavy tails and the variance of the jump length diverges. Heavy tails makes larger jump increments more probable and the processes are characterized by longer walk lengths, $x(n) \propto n^{1/\alpha}$. Alternatively the variance of the step size is kept finite in *anomalous diffusion* but the jumps are delayed and the waiting times diverge. The inner part of the square is filled by so-called *ambivalent processes* were the distributions of waiting times have diverging expectation values and no finite variances of the jump sizes (for details see [25, 213]).

Lévy processes derived from jump distributions with diverging variances and $0 < \alpha < 2$ are called *Lévy flights* by Benoît Mandelbrot [193]. He distinguishes them from *Rayleigh flights* with $\alpha = 2$, which are based on ordinary Brownian motion. In the special class of *Cauchy flights*, $\alpha = 1$, the step sizes are drawn from a Cauchy distribution. Normal random walks create paths on a sufficiently large space-time scale that look like Brownian motion. Dimension one and two are fully covered in the limit $t \rightarrow \infty$. In the plane the visited zones have a comparable densities in the segments covered by the trajectory. The paths of Lévy flights have a very different appearance: Small more or less densely covered patches are separated by long jumps.

Prey foraging strategies of marine predators, for example those of sharks, were found to come close to Lévy flights. An optimal strategy consists in the combination of local searches by Brownian motion like movements and long jumps into distant regions where the next local search can start. The whole trajectory of such a combined search resembles the path of a Lévy flight [137, 293].

3.2.5 Master equations

Master equations have been applied for modelling two processes on discrete spaces, $\mathcal{X}(t) \in \mathbb{N}$: occurrence of independent events (section 3.2.3.5) and random walks (section 3.2.3.6). Because of their general importance in particular in chemical kinetics and population dynamics in biology we shall present here a more detailed discussion of properties and versions of master equations.

3.2.5.1 General master equations

The master equations we are considering here describe continuous time processes. Then, the starting point is the dCKE for pure jump processes (3.33c') with the integral converted into a sum by Riemann-Stieltjes integration (section 3.2.3.5)

$$\frac{\mathrm{d}P_n(t)}{\mathrm{dt}} \;=\; \sum_{m=0}^{\infty} \Big(W(n\,|\,m,t)\,P_m(t) - W(m\,|\,n,t)\,P_n(t) \Big)\,;\; n,m \in \mathbb{N}\;, \quad (3.94)$$

where we have implicitly assumed sharp initial conditions: $P_n(t_0) = \delta_{n,n_0}$. The transition probabilities $W(n\,|\,m,t)$ form a (eventually infinite) transition matrix $\mathrm{W}(t)$ with one special property: The diagonal elements $W(n\,|\,n,t)$ cancel in the master equation and hence can be defined at will without changing the dynamics of the process. Two assumptions are common: (i) W is a stochastic matrix (normalization),

$$\sum_m W(n\,|\,m,t) \;=\; 1 \;\text{ and }\; W(n\,|\,n,t) \;=\; 1 - \sum_{m,m\neq n} W(n\,|\,m,t)\;,$$

or (ii) the diagonal elements vanish, $W(n\,|n,t) = 0$ (annihilation). A Markov process in general, and a master equation is called *time homogeneous* if the transition matrix W does not depend on time and in most cases we shall be dealing with a finite sample or *state space*: $m,n \in \{0,1,\ldots,N\}$ – this is tantamount to saying we are always dealing with a finite numbers of molecules in chemistry or to stating that population sizes in biology are finite.

In the derivation of the dCKE – and also of the master equation – the limit of infinitesimal time steps, $\lim \Delta t \to 0$, excludes the simultaneous occurrence of two or more jumps, but the general master equation allows for jumps of all sizes $\Delta n = n - m$ and this might seem quite unrealistic in most of the realistic systems. In the next section we shall introduce a powerful simplification in form of *death-and-birth* processes that restricts the size of jumps and thus truncates the sum of the terms in the master equation.

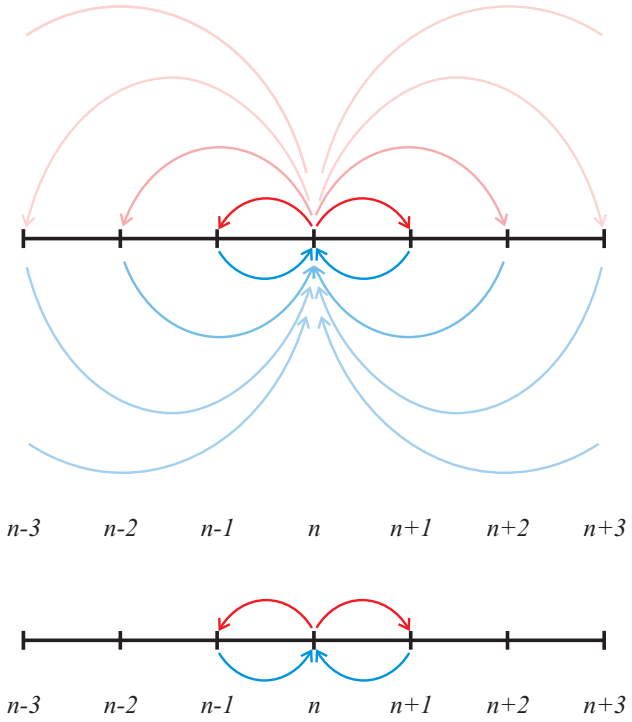

Fig. 3.14 Sketch of the transition probabilities in master equations. In the general master equation steps of any size are admitted (upper drawing) whereas in birth-and-death processes all jumps have the same size. The simplest and most common case is dealing with the condition that the particles *are born* and *die* one at a time (lower drawing).

3.2.5.2 Birth-and-death master equations

The concept of birth-and-death processes has been created in biology (section 5.2.2) and is based on the assumption that only a finite number of individuals are produced – born – or destroyed – die – in a single event. Therefore the jump size is a matter of physics, chemistry or biology and has to be known from empirical observations. In chemical kinetics the jump size is determined by the stoichiometry of the process and in population biology it is the litter size for birth[25] and commonly one for death. The simplest and the only case, we shall discuss here, occurs when births and deaths are confined single individuals. Then,the processes are commonly called one step

[25] The litter size is defined as the number of offspring produced by animal at one birth.

birth-and-death processes.[26] In figure 3.14 the transitions in a general jump process and a birth-and-death process are illustrated. Restriction to single events is tantamount to the choice of a sufficiently small time interval of recording, Δt, such that the simultaneous occurrence of two events has a probability of measure zero (see also section 4.7). This small time step is often called the *blind interval*, because no information on things happening within Δt is available.

Now we can rewrite the transition probabilities in the form

$$W(n|m,t) \;=\; W_{nm} \;=\; w_m^- \,\delta_{n,m+1} \,+\, w_m^+ \,\delta_{n,m-1}\,, \quad \text{or}$$

$$W_{nm} \;=\; \begin{cases} w_m^+ & \text{if } \; m = n-1\,, \\ w_m^- & \text{if } \; m = n+1\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{3.95}$$

as we are dealing with only two allowed processes per event with the transition probabilities

$$w_n^+ \quad \text{for} \;\; n \,\to\, n+1 \quad \text{and} \tag{3.96}$$

$$w_n^- \quad \text{for} \;\; n \,\to\, n-1\,, \;\text{respectively.} \tag{3.97}$$

In section 3.2.3.5 we discussed the Poisson process which can be understood as a birth-and-death process with zero death rate, or birth process, on $n \in \mathbb{N}_{\geq 0}$. The one-dimensional random walk (section 3.2.3.6) is a birth-and-death process with equal birth and death probabilities when the spatial coordinate is changed to a population variable and negative particle numbers are avoided. Modeling of chemical reactions by birth-and-death processes turns out to be a very useful approach for reaction mechanisms, which can be described by changes in a single variable.

The stochastic process can now be described by a birth-and-death master equation

$$\frac{\mathrm{d}P_n(t)}{\mathrm{dt}} \;=\; w_{n-1}^+ \,P_{n-1}(t) \,+\, w_{n+1}^- \,P_{n+1}(t) \,-\, (w_n^+ + w_n^-)\,P_n(t)\;. \tag{3.98}$$

There is no general technique that allows to find the time-dependent solutions of equation (3.98). Special cases, however, are important in chemistry and biology and therefore we shall present several examples later on. In section 5.2.2 we shall give also a detailed overview of the exactly solvable single step birth-and-death processes [108]. Nevertheless, it is possible to analyze the stationary case in full generality.

[26] In addition, one commonly distinguishes between birth-and-death processes in one variable and in many variables [93]. We shall restrict the analysis here to the simpler single variable case here.

Provided a stationary solution of equation (3.98), $\lim_{t\to\infty} P_n(t) = \bar{P}_n$, exists, we can compute it in straightforward manner. We define a probability current $\varphi(n)$ for the n-th step in the series:

$$\begin{array}{lccccccccccc} \text{Particle number } 0 & \rightleftharpoons & 1 & \rightleftharpoons & \ldots & \rightleftharpoons & n-1 & \rightleftharpoons & n & \rightleftharpoons & n+1 & \ldots \\ \text{Reaction step} & & 1 & & 2\ \ldots\ n-1 & & & & n & & n+1 & \ldots\,, \end{array}$$

which is of the form

$$\varphi_n \;=\; w_n^- \, \bar{P}_n \;-\; w_{n-1}^+ \, \bar{P}_{n-1} \;. \tag{3.99}$$

Now, the conditions for the stationary solution are given by

$$\frac{\mathrm{d}P_n(t)}{\mathrm{dt}} = \;0\; = \;\varphi_{n+1} \;-\; \varphi_n \;, \tag{3.100}$$

Restriction to positive particle numbers, $n \in \mathbb{N}_{\geq 0}$, implies $w_0^- = 0$ and $P_n(t) = 0$ for $n < 0$, which in turn leads to $\varphi_0 = 0$.

Now we add the vanishing flow terms according to equation (3.100) and obtain from the telescopic sum:

$$0 \;=\; \sum_{j=0}^{n-1} \varphi_{j+1} \;-\; \varphi_z \;=\; \varphi_n \;-\; \varphi_0 \;.$$

Thus we find $\varphi_n = 0$ for arbitrary n which leads to

$$\bar{P}_n \;=\; \frac{w_{n-1}^+}{w_n^-} \, \bar{P}_{n-1} \quad \text{and finally} \quad \bar{P}_n \;=\; \bar{P}_0 \prod_{z=1}^{n} \frac{w_{j-1}^+}{w_j^-} \;. \tag{3.101}$$

The vanishing flux condition $\varphi_n = 0$ for every reaction step at equilibrium is known in chemical kinetics as the principle of detailed balance, which has been formulated first by the American mathematical physicist Richard Tolman [278] (see also, for example, [93, pp.142-158]).

3.3 Forward and backward equations

Time inversion in a conventional differential equation changes the direction in which trajectories are passed and this changes the phase portrait of the dynamical system: ω-limits become α-limits and vice versa, stable equilibrium points and limit cycles become unstable and so on, but the trajectories – without the arrow of time – remain unchanged. This has the consequence that integrating forward in time yields precisely the same results as integrating backward from the endpoint of the forward trajectory. The same is true, of course, for a Liouville equation but it does not hold for a Wiener process or a Langevin equation: Spreading of individual trajectories occurs likewise in the forward and in the backward direction. Time reversal of diffusion processes has been studied extensively in the nineteen eightieth [4, 60, 124] and it was shown that under mild conditions the time reversed process is a diffusion process as well. This fact is sketched in figure 3.15 where we observe trajectories diverging in the backward direction. In other words, the commonly chosen reference conditions are such that a forward process has the sharp initial conditions at the beginning of the ordinary time scale – t_0 for t progressing into the future – whereas a backward process has sharp final conditions at the end – τ_0 for a virtual or *computational time* τ progressing backwards into the past. Accordingly, the Chapman-Kolmogorov equation can be interpreted in two different ways giving rise to forward and backward equations that are equivalent to each other and basic difference concerns the set of variables, which is held fixed. In case on the forward equation we hold $(\mathbf{x}_0, t_0)$ fixed, and consequently solutions exist for $t \geq t_0$, so that $p(\mathbf{x}, t_0|\mathbf{x}_0, t_0) = \delta(\mathbf{x} - \mathbf{x}_0)$ is an *initial condition* for the forward equation. The backward equation has solutions for $\tau \leq \tau_0$ and hence it expresses development in τ. Accordingly, $p(\mathbf{y}_0, \tau_0|\mathbf{y}, \tau_0) = \delta(\mathbf{y}_0 - \mathbf{y})$ is an appropriate *final* condition (rather than an initial condition).[27]

Naïvely we could expect to find symmetry between forward and backward computation there is, however, one fundamental difference between calculations progressing in opposite directions, which will become evident when we consider backward equations in detail: In addition to the two different computational time scales for forward and backward equations – t and τ, respectively, in figure 3.15 – we have the real or physical time of the process, which has the same direction as t, unless we use some scaling factor it is even identical to t and we shall only distinguish the two time scales if necessary. The basic difference breaking the symmetry between forward and the backward equation thus concerns the arrow of time: The forward calculations progress in the direction of real time whereas in the backward equations the arrow of computational time is opposite to real time. The difference can also be expressed by saying the forward equations make prediction of the future

[27] In order to avoid confusion we shall reserve the variable $y(\tau)$ and $y(0) = y_0$ for backward computation.

and the backward equations reconstruct the past. In the eyes of mathematicians the backward equation is (somewhat) better defined than its forward analogue (see [74] and [77, pp. 321 ff.]).

3.3.1 Backward Chapman-Kolmogorov equations

The Chapman-Kolmogorov equations (3.23 and 3.24) are interpreted in two different ways giving rise to the two formulations known as forward and backward equation. In the forward equation the double (x_3, t_3) is considered to be fixed and (x_1, t_1) expresses the variable in the sense of $x_1(t)$, where the time t_1 is progressing in the direction of positive real time (see figure 3.4). The backward equation, in contrary, is exploring the past of a given situation: Here, the double (x_1, t_1) is fixed and (x_3, t_3) is propagating backwards in time. The fact that real time proceeds in the forward direction has the consequence of somewhat different forms of forward and backward equations. Both Chapman-Kolmogorov differential expressions, the forward and the backward equation, are useful in their own rights. The forward equation gives directly the values of measurable quantities as functions of the *observed* or *real time*. Accordingly, it is preferentially used in describing actual processes and modeling experimental systems, and it is suited for predictions of probabilities in the future. The backward equation finds applications in the computation of the evolution towards given events, for example *first passage times* or *exit problems*, which are dealing with the search for the probability that a particle leaves a region at a certain time.

Since the difference in the derivation of forward and backward equations is essential for the interpretation of the results, we repeat here this derivation for the backward case, which is similar to but not identical with the procedure for the forward equation. The starting point again is the conditional probability of a Markov process from a recording $(\mathbf{y}, \tau)$ in the past to the final condition $(\mathbf{y}_0, \tau_0)$ at present: $p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{y}, \tau) = \delta(\mathbf{y}_0 - \mathbf{y})$ for all values of τ. As the term *backward* indicates we shall, however, assume that the computational time τ progresses from τ_0 into the past ($\tau = -t$ and $\frac{\partial}{\partial \tau} = -\frac{\partial}{\partial t}$; see figure 3.15).

In essence, we proceed in the same way as in section 3.2.2.2 and write down the infinitesimal limit of the difference equation:

$$
\begin{aligned}
&\frac{\partial p\,(\mathbf{y}_0, \tau_0 | \mathbf{y}, \tau)}{\partial \tau} \;= \\
&= \lim_{\Delta\tau \to 0} \frac{1}{\Delta\tau} \Big(p\,(\mathbf{y}_0, \tau_0 | \mathbf{y}, \tau + \Delta\tau) \;-\; p\,(\mathbf{y}_0, \tau_0 | \mathbf{y}, \tau) \Big) \;= \\
&= \lim_{\Delta\tau \to 0} \frac{1}{\Delta\tau} \int_{\Omega} \mathrm{d}\mathbf{z}\, p\,(\mathbf{z}, \tau + \Delta\tau | \mathbf{y}, \tau) \Big(p\,(\mathbf{y}_0, \tau_0 | \mathbf{z}, \tau + \Delta\tau) \;-\; p\,(\mathbf{y}_0, \tau_0 | \mathbf{y}, \tau + \Delta\tau) \Big)
\end{aligned}
$$

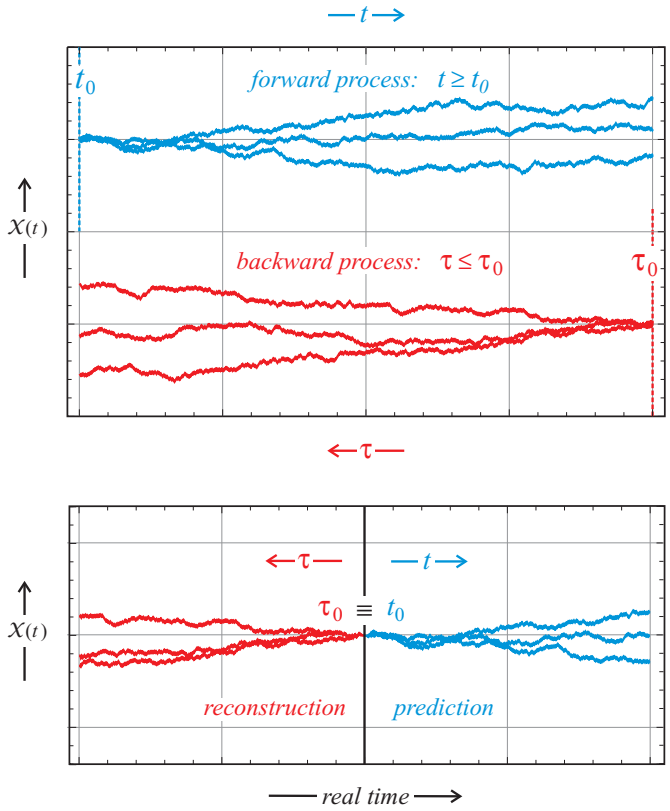

Fig. 3.15 Illustration of forward and backward equations. The forward differential Chapman-Kolmogorov equation is used in calculations of the future development of ensembles or populations. The trajectories (blue) start from an initial condition $(\mathbf{x}_0, t_0)$ commonly corresponding to the sharp distribution $p(\mathbf{x}, t_0) = \delta(\mathbf{x} - \mathbf{x}_0)$, and the probability density unfolds with time, $t \geq t_0$. The backward equation is commonly applied to the calculation of first passage times or the solution of exit problems. In order to minimize the risk of confusion we choose in backward equations the notation $\mathbf{y}$ and τ for the variable and the time, respectively, and we have the apparent correspondence $(\mathbf{y}(\tau), \tau) \Leftrightarrow (\mathbf{x}(t), t)$. In backward equations the latest time the corresponding value of the variable at this time, $(\mathbf{y}_0, \tau_0)$, are held constantτ_0 and a sharp initial condition – better called *final condition* in this case – is applied $p(\mathbf{y}, t_0 | \mathbf{y}, t) = \delta(\mathbf{y} - \mathbf{y}_0)$ and the time dependence of the probability density corresponds to samples unfolding into the past, $\tau \leq \tau_0$ (trajectories in red). In the lower part of the figure and alternative interpretation is given: The forward and the backward process start at the same time into different time directions, computation of the forward process makes predictions of the future whereas the backward process is calculated for the reconstruction of the past.

where we have applied the same two operations as used for the derivation of equation (3.27): (i) resolution of unity,

$$1 \;=\; \int_{\Omega} \mathrm{d}\mathbf{z}\, p\,(\mathbf{z}, \tau + \Delta\tau |\, \mathbf{y}, t)\;,$$

and (ii) insertion of the Chapman-Kolmogorov equation in the second term with $\mathbf{z}$ being the intermediate variable. In the second line of the equation we fixed the error of time. Although the time difference $\Delta\tau$ is vanishing in the limit, the two terms are ordered by the minus sign, a *memory* on the order remains at $\Delta\tau = 0$, and it determines the final result of the derivation. All further steps in the derivation are similar as in the forward case:(i) separation of the domain of integration into two parts with the integrals I_1 and I_2 with $\|\mathbf{z} - \mathbf{y}\| < \epsilon$ and $\|\mathbf{z} - \mathbf{y}\| \geq \epsilon$, respectively, (ii) expansion of I_1 into a Taylor series, (iii) neglect of higher order residual terms, and (iv) integration by parts,n and eventually we obtain:

$$\frac{\partial p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{y}, \tau)}{\partial \tau} \;=\; + \sum_i A_i(\mathbf{y}, \tau)\, \frac{\partial p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{y}, \tau)}{\partial y_i} \;+ \tag{3.102a}$$

$$+\; \frac{1}{2} \sum_{i,j} B_{ij}(\mathbf{y}, \tau)\, \frac{\partial^2 p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{y}, \tau)}{\partial y_i \partial y_j} \;+ \tag{3.102b}$$

$$+\; \int \mathrm{d}\mathbf{z}\; W(\mathbf{z} |\, \mathbf{y}, \tau) \left(p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{y}, \tau) \;-\; p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{z}, \tau) \right). \tag{3.102c}$$

This equation is called the *backward differential Chapman-Kolmogorov equation* in contrast to the previously derived forward equation (3.33). The appropriate finial condition (figures 3.4 and 3.15) is

$$p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{y}, \tau) \;=\; \delta(\mathbf{y}_0 - \mathbf{y}) \quad \text{for all} \quad \tau\;,$$

which expresses the fact that the probability density for finding the particle at location $\mathbf{y}$ at time t if it is at $\mathbf{y}_0$ at the same time is $\delta(\mathbf{y}_0 - \mathbf{y})$, or in other words the (classical and non-quantum-mechanical) particle can be simultaneously at $\mathbf{y}$ and $\mathbf{y}_0$ if and only if $\mathbf{y} = \mathbf{y}_0$.

The Liouville equation (section 3.2.3.1) is a partial differential equation whose physically relevant solutions coincide with the solution of an ordinary differential equation, and therefore the trajectories are invariant under time reversal – only the direction of the process is reversed: going backwards in time changes the sign of the components of $\mathbf{A}$ and the particle travels in opposite direction along the same trajectory, which is fixed by the initial or final condition $(\mathbf{x}_0, t_0)$ or $(\mathbf{y}_0, \tau_0)$.

The diffusion process described by equation (3.102b) spreads in opposite direction as a consequence of the inverse arrow of time. The mathematics of time reversal in diffusion has been studied extensively in the nineteen

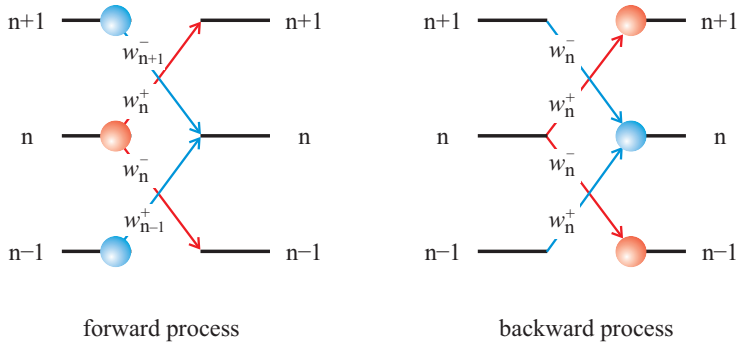

Fig. 3.16 Jumps in the single event master equations. The sketch on the left hand side shows the four single steps in the forward birth-and-death master equations, which are determined by the four transition probabilities w_n^+, w_{n-1}^+, w_{n+1}^-, and w_n^-. Transitions leading to a gain in probability P_n are indicated in blue, those reducing P_n are shown in red. On the right hand side we show the situation in the backward master equation: Only two transition probabilities, w_n^+ and w_n^- enter the equations, and the probabilities determining the amount of gain or loss in P_n are given at the final jump destinations rather than the beginnings.

eighties [4, 60, 124, 263] and rigorous mathematical proofs were derived, which confirmed that inversion of time leads to indeed to a diffusion process in the direction of past time in the sense of the backward processes sketched in figure 3.15. Starting from a sharp final condition the trajectories diverge in the direction of $\tau = -t$.

The third term (3.102c) describes the jump processes and will be handled in the following section 3.3.2 on backward master equations. As in case of the forward equation the limit to vanishing $\Delta\tau$ is encapsulated in the transition probabilities:

$$\lim_{\Delta\tau\to 0} \frac{1}{\Delta\tau}\, p\,(\mathbf{z}, \tau + \Delta\tau|\mathbf{y}, \tau) \;=\; W(\mathbf{z}|\mathbf{y}, \tau)\;. \tag{3.103}$$

Some care is needed in applications to problem solution, because the transition probabilities depend on the definition of the time axis.

3.3.2 Backward master equations

The backward master equation follows directly from the third term in the backward dCKE (3.102c). Since the difference in the derivation of forward and backward equations is essential for the interpretation of the results, we repeat here the derivation of a backward equation by means of the master equation

as an example. The starting point again is the conditional probability of a Markov process from a recording $(\mathbf{y}, \tau)$ in the past to the final condition $(\mathbf{y}_0, \tau_0)$ at present: $p\,(\mathbf{y}_0, \tau_0 |\, \mathbf{y}, \tau) = \delta(\mathbf{y}_0 - \mathbf{y})$ for all values of τ. As the term *backward* indicates we shall, however, assume that the computational time τ progresses from τ_0 into the past.

Because of its general insight into the forward-backward asymmetry we present here a brief derivation of the backward master equation. The jump term (3.102c) is subjected to Riemann-Stieltjes integration,

$$\begin{aligned}\frac{\partial p\,(y_0, \tau_0 |\, y, \tau)}{\partial \tau} \;&=\; \int_\Omega \mathrm{d}z\; W(z |\, y, \tau) \Big(p\,(y_0, \tau_0 |\, y, \tau) \;-\; p\,(y_0, \tau_0 |\, z, \tau) \Big) \;= \\ &=\; \sum_{z=0}^{\infty} W(z |\, y, \tau) \Big(p\,(y_0, \tau_0 |\, y, \tau) \;-\; p\,(y_0, \tau_0 |\, z, \tau) \Big)\;,\end{aligned}$$

and we introduce the notation for discrete particle numbers, $y \Leftrightarrow n \in \mathbb{N}_{\geq 0}$, $z \Leftrightarrow m \in \mathbb{N}_{\geq 0}$, and $y_0 \Leftrightarrow n_0 \in \mathbb{N}_{\geq 0}$:

$$\frac{\partial P(n_0, \tau_0 | n, \tau)}{\partial \tau} \;=\; \sum_{m=0}^{\infty} W(m|n, \tau) \Big(P(n_0, \tau_0 | n, \tau) \;-\; P(n_0, \tau_0 | m, \tau) \Big)\,. \tag{3.104}$$

As previously we assume now time independent transition rates and restrict transitions to single births and deaths:

$$W(m|n, t) \;=\; W_{mn} \;=\; w_n^+\, \delta_{n+1,n} \;+\; w_n^-\, \delta_{n-1,n}\,, \quad \text{or}$$

$$W_{mn} \;=\; \begin{cases} w_n^+ & \text{if} \;\; m = n+1\;, \\ w_n^- & \text{if} \;\; m = n-1\;, \\ 0 & \text{otherwise}\;, \end{cases} \tag{3.95'}$$

Then, then master equation is of the form

$$\begin{aligned}\frac{\partial P(n_0, \tau_0 | n, \tau)}{\partial \tau} \;&=\; w_n^+ \Big(P(n_0, \tau_0 |\, n, \tau) \;-\; P(n_0, \tau_0 |\, n+1, \tau) \Big) \;+ \\ &\quad +\;\; w_n^- \Big(P(n_0, \tau_0 |\, n, \tau) \;-\; P(n_0, \tau_0 |\, n-1, \tau) \Big) \;= \\ &=\; -\, w_n^+\, P(n_0, \tau_0 |\, n+1, \tau) \;-\; w_n^-\, P(n_0, \tau_0 |\, n-1, \tau) \;+ \\ &\quad +\, (w_n^+ \;+\; w_n^-)\, P(n_0, \tau_0 |\, n, \tau)\;.\end{aligned} \tag{3.105}$$

The different conditions for the jumps in the forward and the backward single step master equations are compared in figure 3.16. The interpretation of the forward jumps is straightforward the transition rates $w_k^{\pm}$ $(k = n-1, n, n+1)$ are multiplied by the probabilities to be in the state before the jump at the instant of hopping. The different directions of real time and computational

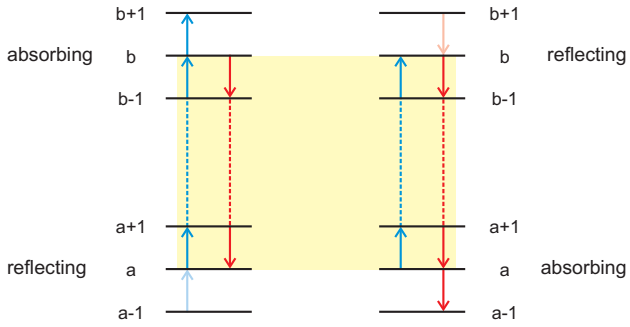

Fig. 3.17 Boundaries in single-step birth-and-death master equations. The figure on the l.h.s. sketches an interval, $a \leq n \leq b$ (indicated by yellow background), with a reflecting boundary at $n = a$ and an absorbing boundary at $n = b$ whereas the interval on the r.h.s. has the absorbing boundary at $n = a$ and the reflecting boundary at $n = b$. The step-up transition probabilities w_n^+ are shown in blue, the step-down transition probabilities w_n^- in red, a reflecting boundary has a zero outgoing probability, w_a^- or w_b^+, and the incoming probabilities, w_{a-1}^+ or w_{b+1}^-, are zero at an absorbing boundary. The incoming transition probabilities at the reflection boundaries are shown in light colors and play no role in the stochastic process because the probabilities of the corresponding virtual states are zero by definition: $P_{a-1}(t) = P_{b+1}(t) = 0$.

time in the backward process change the situation: The probabilities involved in jumps of the backward process are the probabilities after the jump.

3.3.3 Mean first passage times

A *first passage time* is a random variable $\mathcal{T}$ that measures the instant when a particle passes a predefined location or state the first time. Its expectation value $E(\mathcal{T})$ is called *mean first passage time*. We need to stress *first*, because in the majority of processes we are discussing here the variables may take on certain values finitely or in infinite time processes even infinitely often. In order to facilitate precise definition we introduce boundaries, which determine the behavior of the particle at the boundary of the accessible domain. For master equations two classes of boundaries are commonly defined: (i) *absorbing boundaries* and (ii) *reflecting boundaries*. When a particle reaches an absorbing boundary it disappears whereas from a reflecting boundary it automatically returns to the range of allowed values. Both boundaries are readily incorporated into master equations.

3.3.3.1 Boundaries in birth-and-death master equations

The implementation of boundary conditions for single-step birth-and-death processes is straightforward. The process is, for example, assumed to be confined to the interval $a \leq n \leq b$, $n \in \mathbb{Z}$, and we only need to choose the appropriate transition probabilities that forbid the exit from the interval in case of a reflecting boundary or the return to the interval for an absorbing boundary (figure 3.17). Confining the process to $[a, b]$ we need two boundaries,[28] a lower boundary at $n = a$ and an upper one at $n = b$. Because of symmetry it is sufficient to consider only the lower boundary.

The boundary at $n = a$ is absorbing when the particle after it left the domain $[a, b]$ cannot return to it in forthcoming jumps, which is easily achieved by setting $w^+_{a-1} = 0$. A reflecting boundary results from the assumption $w^-_a = 0$: the particle cannot leave the domain. By symmetry we have $w^-_{b+1} = 0$ and $w^+_b = 0$ for the absorbing and the reflecting upper boundary.

In the forward single-step birth and death master equation the flux across the boundaries is only relevant for the equations of the states at the boundaries, $n = a$ and $n = b$. According to equation (3.98) with the initial condition $P_n(t_0) = \delta_{n,n_0}$ the differential change in the probability density at the lower boundary is

$$\frac{\mathrm{d}P_a(t)}{\mathrm{dt}} \;=\; w^-_{a+1}\,P_{a+1}(t) \;-\; w^+_a\,P_a(t) \;+\; w^+_{a-1}\,P_{a-1}(t) \;-\; w^-_a\,P_a(t)\;.$$

The two rightmost terms are effected by the boundary condition. In the case of reflection at the boundary the condition is that nothing flows out of the domain and this is fulfilled by $w^-_a = 0$ (figure 3.17), if the reflecting boundary is combined with no influx either w^+_{a-1} or $P_{a-1}(t)$ (or both) must be zero. In general the assumption $P_{a-1}(t) = 0$ is reasonable because it is not very meaningful to assume a finite probability density outside the domain $[a, b]$. Nevertheless, an influx can be modeled readily under the assumption of a virtual state $n = a - 1$. An alternative assumption is the equivalent to noflux or Neumann boundary conditions in partial differential equations: The flux at the boundary has to vanish and this implies

$$w^+_{a-1}\,P_{a-1}(t) \;=\; w^-_a\,P_a(t)\;. \tag{3.106}$$

Absorption at the lower boundary also allows for an alternative to setting $w^+_{a-1} = 0$: Introducing a virtual state $n = a - 1$ and demanding $P_{a-1}(t) = 0$ yields the same effect. It is straightforward to show that the assumption of a virtual state $n = b + 1$ and the two conditions, $w^+_b P_b(t) = w^-_{b+1} P_{b+1}(t) = 0$

[28] Boundaries are also called *barriers* in the literature and both notions are used as synonyms. We shall use here exclusively the word 'boundary'. The expression barrier will be reserved for obstacles of motion inside the domain of the random variable.

and $P_{b+1} = 0$, do the same job for a reflecting or an absorbing upper barrier, respectively.

Alternative conditions can be found also for the backward master equation (3.105) on the interval $[a, b]$. At the lower boundary $n = a$ we find:

$$\begin{aligned}\frac{\mathrm{d}P(n_0, t_0|a, t)}{\mathrm{dt}} &= w_a^+ \, P(n_0, t_0|a+1, t) \; - \; w_a^+ \, P(n_0, t_0|a, t) + \\ &+ w_a^- \, P(n_0, t_0|a-1, t) \; - \; w_a^- \, P(n_0, t_0|a, t)\end{aligned}$$

for $n_0 \in [a, b]$. Again only the last two terms – the second line – are affected by the boundary conditions, and setting

$$P(n_0, t_0|a-1, t) \; = \; P(n_0, t_0|a, t) \tag{3.107}$$

is equivalent to putting $w_a^-(t) = 0$ in order to introduce a reflecting lower boundary through equating the second line to zero. The introduction of an absorbing lower boundary is a bit more tricky, since the transition rate w_{a-1}^+ does not appear in the backward master equation. Clearly, the condition $P(n_0, t_0|n, t) = 0$ with $n_0 \in [a, b]$ and $n < a$ will have the same effect as $w_{a-1}^+ = 0$. In single-step birth-and-death processes only the term with the largest value of n will be relevant for the process confined to the domain $[a, b]$ and hence $P(n_0, t_0|a-1, t) = 0$ is sufficient. At the upper boundary the corresponding two equations having the same effect as $w_b^+ = 0$ and $w_{b+1}^- = 0$ are: $P(n_0, t_0|b+1, t) = P(n_0, t_0|b, t)$ and $P(n_0, t_0|b+1, t) = 0$ for the reflecting and the absorbing boundary, respectively.

In this context it should be mentioned that in case of the chemical master equation equation we shall encounter *natural boundaries* where reaction kinetics itself takes care of reflecting or absorbing boundaries. If we are dealing with a reversible chemical reaction approaching a thermodynamic equilibrium in a system with a total number of N molecules the states $n_{\mathbf{K}} = 0$ and $n_{\mathbf{K}} = N$ are reflecting for each molecular species **K**, whereas in an irreversible reaction the state $n_{\mathbf{K}} = 0$ is absorbing when the reactant **K** is at shortfall. Similarly, in absence of migration the state of extinction $n_{\mathbf{S}} = 0$ is an absorbing boundary for species **S**.

3.3.3.2 First passage time in birth-and-death master equations

The calculation of a mean first passage time is illustrated by means of a simple example: The escape of a particle from a domain $[a, b]$ with a reflecting boundary at $n = a$ and an absorbing boundary at $n = b$ [191, pp. 90-92]. We make use of the backward master equation (3.105) and according to last section 3.3.3.1 we adopt the following conditions for the boundaries

$$P(n_0, t_0|a-1, t) \; = \; P(n_0, t_0|a, t) \;\; \text{and} \;\; P(n_0, t_0|b+1, t) \; = \; 0 \; .$$

The probability that the particle is still in the interval $[a, b]$ is calculated by summation over all states in the accessible domain:

$$I_n(t) \;=\; \sum_{m=a}^{b} P(m,t|n,0)\,,\; m \in \mathbb{Z}\,. \tag{3.108}$$

Insertion of the individual terms from the backward master equation (3.105) yields for the time derivative:

$$\begin{aligned} -\,\frac{\mathrm{d}I_n(t)}{\mathrm{dt}} \;&=\; \sum_{m=a}^{b} \frac{\mathrm{d}P(m,t|n,0)}{\mathrm{dt}} \;= \\ &=\; w_n^+\Big(I_n(t) \,-\, I_{n+1}(t)\Big) \,+\, w_n^-\Big(I_n(t) \,-\, I_{n-1}(t)\Big) \end{aligned} \tag{3.109}$$

with the conditions $I_{a-1}(t) = I_a(t)$ for the reflecting boundary at $n = a$ and $I_{b+1}(t) = 0$ for the absorbing boundary at $n = b$. The minus sign expresses the decrease in probability to be still within the interval $[a, b]$ in real time and is a consequence of the two time scales in backward processes, $\mathrm{dt} = -\mathrm{d}\tau$.

The probability of leaving the interval $[a, b]$ – the probability of absorption – within an infinitesimal interval of time $[t, t + \mathrm{dt}]$ is calculated to be

$$I_n(t) \,-\, I_n(t + \Delta t) \;=\; -\,\frac{\partial I_n}{\partial t}\;\mathrm{dt}\;,$$

and we can now obtain the mean first passage time for the escape from state n, $\langle \mathcal{T}_n \rangle$ by integration

$$\langle \mathcal{T}_n \rangle \;=\; -\int_0^\infty t\;\frac{\partial I_n}{\partial t}\;\mathrm{dt} \;=\; \int_0^\infty I_n\;\mathrm{dt}\;, \tag{3.110}$$

where the last expression results from integration by parts. Integration of equation (3.109) yields

$$\int_0^\infty -\,\frac{\partial I_n(t)}{\partial t}\,\mathrm{dt} \;=\; 1$$

for the l.h.s. since absorption of the particle or escape from the domain is certain. Integration of the r.h.s. yields mean passage times, and finally we obtain

$$1 \;=\; w_n^+\Big(\langle \mathcal{T}_n \rangle - \langle \mathcal{T}_{n+1} \rangle\Big) \,+\, w_n^-\Big(\langle \mathcal{T}_n \rangle - \langle \mathcal{T}_{n-1} \rangle\Big) \tag{3.111}$$

the equation for the calculation of $\langle \mathcal{T}_n \rangle$. The boundary conditions are: $\langle \mathcal{T}_{a-1} \rangle = \langle \mathcal{T}_a \rangle$ and $\langle \mathcal{T}_{b+1} \rangle = 0$.

The solution of equation (3.111) for $\langle \mathcal{T}_n \rangle$ is facilitated by the introduction of new variables S_n and auxiliary functions φ_n:

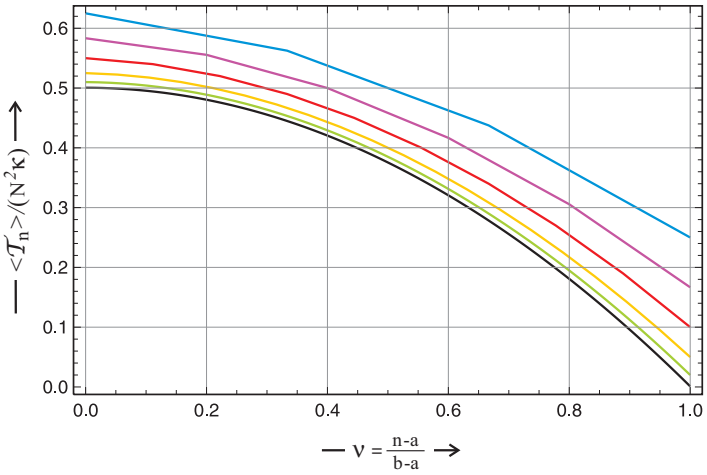

Fig. 3.18 Mean first passage times of a single-step birth-and-death process. Mean first passage times are computed from equation(3.113). In order to be able to compare the results for different sizes of the interval $[a, b]$ the interval is are normalized: $a = 0$ and $b = 1$, or $\nu = (n - a)/(b - a)$. Computed mean first passage times are scaled by a factor $(N^2\kappa)^{-1}$ with $N = b - a + 1$. The values for N chosen in the computations and the color code are: 4 (blue), 6 (violet), 10 (red), 20 (yellow), 50 (green), and 1000 (black).

$$S_n \;=\; \frac{\langle \mathcal{T}_{n+1}\rangle \;-\; \langle \mathcal{T}_n\rangle}{\varphi_n}\,,\; n \in [a,b] \;\text{ and}$$

$$\varphi_n \;=\; \prod_{m=a+1}^{n} \frac{w_m^-}{w_m^+}\,,\; n \in [a+1,b] \;\text{ with }\; \varphi_a \;=\; 1\;,$$

and in the new variables equation (3.111) takes on the form

$$-1 \;=\; w_n^+\,\phi_n\left(S_n \;-\; S_{n-1}\right)\;,$$

which allows for deriving a solution for the new variables

$$S_k \;=\; -\sum_{m=a}^{k} \frac{1}{w_m^+\,\varphi_m}\;.$$

From $\varphi_k S_k = \langle \mathcal{T}_{k+1}\rangle \;-\; \langle \mathcal{T}_k\rangle$ we obtain by means of the telescopic sum from $k = n$ to $k = b$

$$\sum_{k=n}^{b} \langle \mathcal{T}_{k+1} \rangle - \langle \mathcal{T}_k \rangle = \langle \mathcal{T}_{n+1} \rangle - \langle \mathcal{T}_n \rangle + \langle \mathcal{T}_{n+2} \rangle - \langle \mathcal{T}_{n+1} \rangle + \ldots + \langle \mathcal{T}_{b+1} \rangle - \langle \mathcal{T}_b \rangle =$$

$$= - \langle \mathcal{T}_n \rangle \ ,$$

because of the boundary condition $\langle \mathcal{T}_{b+1} \rangle = 0$, and we obtain the desired result

$$\langle \mathcal{T}_n \rangle \ = \ \sum_{k=n}^{b} \varphi_k \sum_{m=a}^{k} \frac{1}{w_m^+ \, \varphi_m} \ = \ \sum_{k=n}^{b} \frac{1}{w_k^+ \, \bar{P}_k} \sum_{m=a}^{k} \bar{P}_m \ , \tag{3.112}$$

where we have used the stationary probabilities $\bar{P}$ (3.101) instead of the functions φ to calculate the mean passage times.

For the purpose of illustration we choose a example that yields simple analytical expressions for the mean first passage times. The simplification is made with the transition probabilities:

$$w_k^+ \ = \ w_k^- \ = \ \kappa \ \ \forall \ \ k = 1, \ldots, N \ \ \text{and} \ \ w_0^+ = \kappa \, , \ w_0^- = 0 \, , \ w_{b+1}^- = 0 \ .$$

The number of states n in the interval $[a, b]$ with $a, b, n \in \mathbb{Z}$ is $N = b - a + 1$. Since $\bar{P}_n = \bar{P}_0$ follow from (3.101) we obtain by means of the normalization condition $\sum_n \bar{P}_n = 1$ the same probability $\bar{P}_n = 1/N \ \forall \, n$. Insertion in equation(3.112) yields the expression

$$\langle \mathcal{T}_n \rangle \ = \ \frac{1}{2\kappa} \, (b + n - 2a + 2)(b - n + 1) \ , \tag{3.113}$$

which has the leading term $-n^2$ in n. Numerical results are given in figure 3.18 and indeed the curves approach a negative quadratic function for large N.

Mean first passage times find widespread applications in chemistry and biology. Important study cases are the escape from potential traps, for example classical motion in the double well potential, fixation of alleles in population, and extinction times.

3.4 Stochastic differential equations

The Chapman-Kolmogorov equation had been conceived in order to be able to model the propagation of probabilities in sample space. An alternative modeling approach starts out from a deterministic description by means of a difference or differential equation and superimposes a fluctuating random element. The idea of introducing stochasticity into deterministic modeling of processes goes back to the beginning of the twentieth century: In 1900 the French mathematician Louis Bachelier conceived and analyzed in his thesis the stochastic difference equation

$$\mathcal{X}(t_{n+1}) \;=\; \mathcal{X}(t_n) \,+\, \mu\,\Delta t \,+\, \sigma\sqrt{\Delta t}\; W_{n+1}$$

in order to model the fluctuating prices in the Paris stock exchange. Herein $\mu(\mathcal{X}_t, t)$ is a function related to the foreseeable development, $\sigma(\mathcal{X}_t, t)$ describes the amplitude of the random fluctuations and the W_n's are independent normal variables with mean zero and variance one in the sense of Brownian increments [10]. Remarkable is the fact that Bachelier's thesis preceded Einstein's and von Smoluchowski's famous works by five and six years, respectively.

The concept of stochastic differential equations is commonly attributed to the French mathematician Paul Langevin who proposed an equation named after him that allows for the introduction of random fluctuations into conventional differential equations [173]. The idea was to find a sufficiently simple approach to model successfully Brownian motion. In its original form the Langevin equation was written as

$$m\,\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{dt}^2} \;=\; -\,\gamma\,\frac{\mathrm{d}\mathbf{r}}{\mathrm{dt}} \,+\, \boldsymbol{\xi}(t) \quad \text{or} \quad \frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{dt}} \;=\; -\frac{\gamma}{m}\,\mathbf{p}(t) \,+\, \boldsymbol{\xi}(t)\;. \tag{3.114}$$

It describes the motion of a Brownian particle of mass m where $\mathbf{r}(t)$ and $\mathrm{d}\mathbf{r}/\,\mathrm{dt} = \mathbf{v}(t)$ are location and velocity of the particle, respectively. The term on the l.h.s. is the Newtonian gain in linear momentum $\mathbf{p}$ due to the force, $\mathrm{d}\mathbf{p}/\,\mathrm{dt}$, the first term on the r.h.s. is the loss of momentum due to friction, and the second term, $\boldsymbol{\xi}(t)$, represents the irregularly fluctuating *Brownian random force*. The Langevin equation can be written in terms of the momentum $\mathbf{p}$ and then it takes on the more familiar form. The parameter $\gamma = 6\pi\,\eta\,r$ is the friction coefficient according to Stokes law with η being the viscosity coefficient of the medium and r the size of the particle. The analogy of (3.114) to Newton's equation of motion is evident: The deterministic force, $f(x) = -(\partial V/\partial x)$ with $V(x)$ being the potential energy, is replaced by $\xi(t)$.

In figure 3.1 stochastic differential equations were shown as an alternative to the Chapman-Kolmogorov equation in modeling Markov processes. As said, the basic difference between the Chapman-Kolmogorov and the Langevin approach is the object whose time dependence is investigated: The Langevin equation 3.114 considers a single instant of a particle moving in

physical 3D-space that is exposed to thermal motion and the integration yields a single stochastic trajectory. The Chapman-Kolmogorov equation of continuous motion leads to a Fokker-Planck equation 3.34, which describes the migration of a probability density in the same 3D-space where the trajectory is defined. *Equivalence* of both approaches expresses the fact that sampling of trajectories of a Langevin equation converges in distribution to the (time dependent) Fokker-Planck probability density in the limit of (infinitely) large samples. The equivalence of the Langevin and the Chapman-Kolmogorov approach is discussed in more detail in section 3.4.5. In case an analytical solution to the stochastic differential equation is available, the solution can be used to calculate moments of the probability distribution of $\mathcal{X}(t)$ and their time-dependence (section 3.4.6), especially mean and variance, which in practice are often sufficient for the description of a process.

In the literature one can find an enormous variety of detailed treatises of stochastic differential equations. We mention here the monograph [7] and two books that are available on the internet: [206, 229]. The forthcoming presentation of stochastic differential equations follows in essence the line of thought chosen by Crispin Gardiner [93, pp.77-96].

3.4.1 Mathematics of stochastic differential equations

Generalization of equation (3.114) from Brownian motion to an arbitrary stochastic process yields

$$\frac{\mathrm{d}x}{\mathrm{dt}} \;=\; a(x,t) \;+\; b(x,t)\,\xi(t)\;, \tag{3.115}$$

where x is the variable under consideration and $\xi(t)$ is an irregularly fluctuating term often called *noise*. The two functions $a(x,t)$ and $b(x,t)$ are defined by the process to be investigated and the letters are chosen in order to point at the analogy to Fokker-Planck equations (3.34). If the fluctuating term is independent of x, one speaks of *additive noise*.

From the mathematical point of view we require statistical independence for $\xi(t_1)$ and $\xi(t_2)$ if and only if $t_1 \neq t_2$. Furthermore we assume $\langle \xi(t) \rangle = 0$ without loosing generality since any drift term can be absorbed in $a(x,t)$, and encapsulate all requirements in an irregularity condition

$$\langle \xi(t_1)\,\xi(t_2) \rangle \;=\; \delta(t_1 - t_2)\;. \tag{3.116}$$

The Dirac δ-function diverges as $|t_1 - t_2| \to 0$ this has the consequence that $\langle \xi(t)\,\xi(t) \rangle$ and the variance $\sigma^2\big(\xi(t)\big) \;=\; \langle \xi(t)\,\xi(t) \rangle - \langle \xi(t) \rangle^2$ are infinite for $t_1 = t_2 = t$.

In order to be able to work with the differential equation (3.115) we require existence of an integral of the form

$$\omega(t) \;=\; \int_0^t \xi(\tau)\,\mathrm{d}\tau \;.$$

If $\omega(t)$ is a continuous function of time it has the Markov property, which can be proven by partitioning the integral

$$\begin{aligned}\omega(t_2) \;&=\; \int_0^{t_1} \xi(\tau_1)\,\mathrm{d}\tau_1 \;+\; \int_{t_1}^{t_2} \xi(\tau_2)\,\mathrm{d}\tau_2 \;=\\ &=\; \lim_{\varepsilon\to 0}\left(\int_0^{t_1-\varepsilon} \xi(\tau_1)\,\mathrm{d}\tau_1\right) \;+\; \int_{t_1}^{t_2} \xi(\tau_2)\,\mathrm{d}\tau_2\end{aligned}$$

and hence for every $\varepsilon > 0$ the $\xi(\tau_1)$ in the first integral is independent of the $\xi(\tau_2)$ in the second integral. By continuity $\omega(t_1)$ and $\omega(t_2) - \omega(t_1)$ are statistically independent in the limit $\varepsilon \to 0$, and further $\omega(t_2) - \omega(t_1)$ is independent of all $\omega(\hat{t})$ with $\hat{t} < t_1$. In other words, $\omega(t_2)$ is completely determined in probabilistic terms by the value $\omega(t_1)$ and no information on any past values is required: $\omega(t)$ is Markovian. □

Recalling now the differential Chapman-Kolmogorov equation (3.33) and because of the continuity of $\omega(t)$, we postulate the existence a Fokker-Planck equation that describes $\omega(t)$. Computation of the drift and diffusion term is straightforward [93, pp. 78,79] and yields $A(t) = 0$ and $B(t) = 1$. This is the Fokker-Planck equation of the Wiener process (3.41) and we identify

$$\int_0^t \xi(\tau)\,\mathrm{d}\tau \;=\; \omega(t) \;=\; W(t) \;.$$

Considering rigorously the consequences of equation (3.41) we have the paradoxical situation that the integral of $\xi(t)$ is $W(t)$, which is continuous but *nowhere* differentiable and therefore neither the Langevin equation (3.114) nor the stochastic differential equation (3.115) exist in strict mathematical terms. Only the integral equation,

$$x(t) \;-\; x(0) \;=\; \int_0^t a\big(x(\tau),\tau\big)\,\mathrm{d}\tau \;+\; \int_0^t b\big(x(\tau),\tau\big)\,\xi(\tau)\,\mathrm{d}\tau \;, \qquad (3.117)$$

can be accessed by consistent interpretation. The relation to the Wiener process becomes more visible by writing

$$\xi(t)\;\mathrm{dt} \;=\; \mathrm{d}W(t) \;\equiv\; W(t+\mathrm{dt}) \;-\; W(t) \;,$$

which eventually yields:

$$x(t) \;-\; x(0) \;=\; \int_0^t a\big(x(\tau),\tau\big)\,\mathrm{d}\tau \;+\; \int_0^t b\big(x(\tau),\tau\big)\,\mathrm{d}W(\tau) \;. \qquad (3.117')$$

The second integral is a stochastic Stieltjes integral the evaluation of which will be discussed in the next section 3.4.2. In differential form we find now for the correctly formulated stochastic differential equation

$$\mathrm{d}x \;=\; a\big(x(t),t\big)\;\mathrm{dt} \;+\; b\big(x(t),t\big)\,\mathrm{d}W(t)\;. \qquad (3.117")$$

There are several different ways to make the Langevin equation (3.115) compatible with standard mathematics. We followed here the approach of Crispin Gardiner, assumed continuity of $\omega(t)$ and got the answer that $\xi(t)$ follows the normal distribution. The inverse sequel of arguments starting out from assumption of the Gaussian nature of the probability density $\xi(t)$ is equally justifiable and it is definitely a matter of taste, which assumption is preferred.

3.4.2 Stochastic integration

A *stochastic integral* requires additional definitions compared to ordinary Riemann integration. We shall explain this rather unexpected fact and give some practical recipes for integration (for more details see [246]). Let $G(t)$ be an arbitrary function of time and $W(t)$ the Wiener process, then the stochastic integral $I(t,t_0)$ is defined as a Riemann-Stieltjes integral (1.49) of the form

$$I(t,t_0) \;=\; \int_{t_0}^{t} G(\tau)\,\mathrm{d}W(\tau)\;. \qquad (3.118)$$

The integral is partitioned into n subintervals, which are separated by the points t_i: $t_0 \le t_1 \le t_2 \le \cdots \le t_{n-1} \le t$ (figure 3.19). Intermediate points τ_i are defined within each of the subintervals $t_{i-1} \le \tau_i \le t_i$ and they will be used for the evaluation of the function $G(\tau_i)$ and the value of the integral depends on the position chosen for τ_i within the subintervals.

The stochastic integral $\int_0^t G(\tau)\,\mathrm{d}W(\tau)$ is defined as the limit of the partial sums

$$S_n \;=\; \sum_{i=1}^{n} G(\tau_i)\,\Big(W(t_i) - W(t_{i-1})\Big)$$

and it is not difficult to recognize that the integral is different for different choices of the intermediate point τ_i. As a particular important example we consider the case $G(t) = W(t)$:

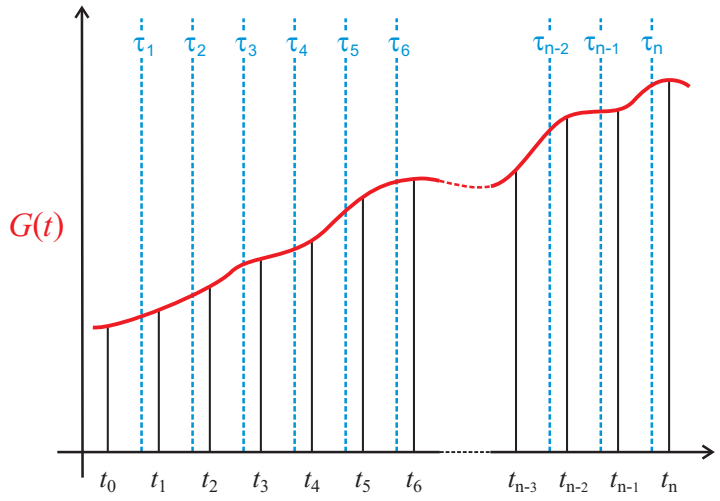

Fig. 3.19 Stochastic integral. The time interval $[t_0, t]$ is partitioned into n segments and an intermediate point τ_i is defined in each segment: $t_{i-1} \le \tau_i \le t_i$.

$$\begin{aligned}
\langle S_n \rangle &= \left\langle \sum_{i=1}^{n} W(\tau_i) \Big(W(t_i) - W(t_{i-1}) \Big) \right\rangle = \\
&= \sum_{i=1}^{n} \langle W(\tau_i) W(t_i) \rangle - \sum_{i=1}^{n} \langle W(\tau_i) W(t_{i-1}) \rangle = \\
&= \sum_{i=1}^{n} \Big(\min(\tau_i, t_i) - \min(\tau_i, t_{i-1}) \Big) = \sum_{i=1}^{n} (\tau_i - t_{i-1}) \; .
\end{aligned}$$

Next we choose the same intermediate position τ for all subintervals 'i'

$$\tau_i = \alpha\, t_i + (1-\alpha)\, t_{i-1} \quad \text{with} \quad 0 \le \alpha \le 1 \tag{3.119}$$

and obtain for the telescopic sum[29]

$$\langle S_n \rangle = \sum_{i=1}^{n} (t_i - t_{i-1})\, \alpha = (t - t_0)\, \alpha \; .$$

Accordingly, the mean value of the integral may adopt any value between zero and $(t - t_0)$ depending on the choice of the position of the intermediate points as expressed by the parameter α. Out of different possible choices two are popular leading to the Itō and the Stratonovich stochastic integral.

[29] In a telescopic sum all terms except the first and the last summand cancel.

3.4.2.1 Itō stochastic integral

The most frequently used definition of the stochastic integral is due to the Japanese mathematician Kiyoshi Itō [140, 141]. Semimartingales (section 3.2.1.2), in particular local martingales, are the most common stochastic processes that allow for straightforward application of Itō's formulation of stochastic calculus.

The choice $\alpha = 0$ or $\tau_i = t_{i-1}$ defines the Itō stochastic integral of a function $G(t)$:

$$\int_{t_0}^{t} G(\tau)\,\mathrm{d}W(\tau) \;=\; \lim_{n\to\infty} \sum_{i=1}^{n} G(t_{i-1})\Big(W(t_i) \;-\; W(t_{i-1})\Big)\;, \qquad (3.120)$$

where the limit is taken as the mean square limit (1.43). As an example we compute the previously discussed integral $\int_{t_0}^{t} W(\tau)\,\mathrm{d}W(\tau)$ and find for the sum S_n:

$$\begin{aligned} S_n \;&=\; \sum_{i=1}^{n} W(t_{i-1})\,\big(W(t_i) \;-\; W(t_{i-1})\big) \;\equiv\; \sum_{i=1}^{n} W(t_{i-1})\,\Delta W(t_i) \;= \\ &=\; \frac{1}{2}\sum_{i=1}^{n}\Big(\big(W(t_{i-1}) + \Delta W(t_i)\big)^2 \;-\; W(t_{i-1})^2 \;-\; \Delta W(t_i)^2\Big) \;= \\ &=\; \frac{1}{2}\Big(W(t)^2 \;-\; W(t_0)^2\Big) \;-\; \frac{1}{2}\sum_{i=1}^{n} \Delta W(t_i^2)\;, \end{aligned}$$

where the second line results from: $2\sum ab = \sum(a+b)^2 - \sum a^2 - \sum b^2$. It is now necessary to calculate the mean square limit of the second term in the last line of the equation. For a finite sum we have the expectation values

$$\left\langle \sum_{i=1}^{n} \Delta W(t_i)^2 \right\rangle \;=\; \sum_i \Big\langle \big(W(t_i) - W(t_{i-1})\big)^2 \Big\rangle \;=\; \sum_i (t_i - t_{i-1}) \;=\; t - t_0\;, \quad (3.121)$$

where the second equality results from the Gaussian nature of the probability density (3.46):[30]

$$\big\langle \big(W(t_i) - W(t_j)\big)^2 \big\rangle = \big\langle W(t_i)^2 \big\rangle - \big\langle W(t_j)^2 \big\rangle = \sigma^2\big(W(t_i)\big) - \sigma^2\big(W(t_j)\big) = t_i - t_j\;.$$

Next we calculate the expectation of the mean square deviation in (3.121):

[30] For the derivation of this relation we used the fact that the stochastic variables of the Wiener process at different times are uncorrelated, $\langle W(t_i)W(t_j)\rangle = 0$ and the variance is $\sigma^2\big(W(t_i)\big) = \langle W(t_i)^2\rangle - \langle W(t_i)\rangle^2 = \langle W(t_i)^2\rangle - \mu^2$.

$$\Big\langle \Big(\sum_{i=1}^{n} \big(W(t_i) - W(t_{i-1})\big)^2 \; - \; (t - t_0) \Big)^2 \Big\rangle \; =$$

$$= \Big\langle \sum_i \big(W(t_i) - W(t_{i-1})\big)^4 \; + \; 2 \sum_{i<j} \big(W(t_i) - W(t_{i-1})\big)^2 \big(W(t_j) - W(t_{j-1})\big)^2 \; -$$

$$- \; 2(t - t_0) \sum_i \big(W(t_i) - W(t_{i-1})\big)^2 \; + \; (t - t_0)^2 \Big\rangle \; .$$

We start the evaluation of the individual terms in the second line: According to (2.44) the fourth moment of a Gaussian variable can be expressed in terms of the variance

$$\Big\langle \big(W(t_i) - W(t_{i-1})\big)^4 \Big\rangle \; = \; 3 \Big\langle \big(W(t_i) - W(t_{i-1})\big)^2 \Big\rangle^2 \; = \; 3\,(t_i - t_{i-1})^2$$

Making use again of the independence of Gaussian variables we find

$$\Big\langle \big(W(t_i) - W(t_{i-1})\big)^2 \big(W(t_j) - W(t_{j-1})\big)^2 \Big\rangle \; = \; (t_i - t_{i-1})(t_j - t_{j-1}) \; .$$

Insertion into the expectation value eventually yields:

$$\Big\langle \Big(\sum_{i=1}^{n} \big(W(t_i) - W(t_{i-1})\big)^2 \; - \; (t - t_0) \Big)^2 \Big\rangle \; =$$

$$= \; 2 \sum_i (t_i - t_{i-1})^2 \; + \; \Big(\sum_i (t_i - t_{i-1}) - (t - t_0) \Big) \Big(\sum_j (t_j - t_{j-1}) - (t - t_0) \Big) \; =$$

$$= \; 2 \sum_i (t_i - t_{i-1})^2 \; \to \; 0 \;\; \text{as} \;\; n \to \infty \, .$$

Accordingly, $\lim_{n\to\infty} \sum_i \big(W(t_i) - W(t_{i-1})\big)^2 \; = \; t - t_0$ in the mean square limit. □

Eventually, we obtain for the Itō stochastic integral of the Wiener process:

$$\int_{t_0}^{t} W(\tau)\, \mathrm{d}W(\tau) \; = \; \frac{1}{2} \Big(W(t)^2 - W(t_0)^2 - (t - t_0) \Big) \; . \tag{3.122}$$

We remark that the Itō integral differs from the conventional Riemann-Stieltjes integral where the term $t - t_0$ is absent. An illustrative explanation for this unusual behavior of the limit of the sum S_n is the fact that the quantity $|W(t + \Delta t) - W(t)|$ is almost always of the order $\sqrt{t}$ and hence – unlike in ordinary integration – the terms of second order in $\Delta W(t)$ do not vanish on taking the limit.

It is also worth noticing that the expectation value of the integral (3.122) vanishes,

$$\Big\langle \int_{t_0}^{t} W(\tau)\, \mathrm{d}W(\tau) \Big\rangle \; = \; \frac{1}{2} \Big(\big\langle W(t)^2 \big\rangle - \big\langle W(t_0)^2 \big\rangle - (t - t_0) \Big) \; = \; 0 \; , \tag{3.123}$$

since the intermediate terms $\langle W(t_{i-1})\Delta W(t_i)\rangle$ vanish because $\Delta W(t_i)$ and $W(t_{i-1})$ are statistically independent.

3.4.2.2 Nonanticipating functions

The concept of a nonanticipating or adapted process has been mentioned in section 3.2.1.2: A stochastic process $\mathcal{X}(t)$ is adapted if and only if for every trajectory and for every time t, $\mathcal{X}(t)$ is known at time t and not before, in other words, a nonanticipating or adapted process 'does not look into the future', in other words, a function $G(t)$ is nonanticipating or adapted to the process $\mathrm{d}W(t)$ if the value of $G(t)$ at time t depends only on the random increments $\mathrm{d}W(\tau)$ for $t \leq \tau$. Here we shall require this property in order to be able to solve certain classes of Itō stochastic integrals, which can be expressed as functions or functionals[31] of the Wiener process $W(t)$ by means of a stochastic differential or integral equation of the form

$$x(t) - x(t_0) = \int_{t_0}^{t} a\big(x(\tau),\tau\big)\,\mathrm{d}\tau + \int_{t_0}^{t} b\big(x(\tau),\tau\big)\,\mathrm{d}W(\tau)\ . \qquad (3.117')$$

A function $G(t)$ is *nonanticipating* with respect to t if $G(t)$ is probabilistically independent of $\big(W(s)-W(t)\big)$ for all s and t with $s > t$. In other words, $G(t)$ is independent of the behavior of the Wiener process in the future $s > t$. This is a natural and physically reasonable requirement for a solution of equation (3.117") because it boils down to the condition that $x(t)$ involves $W(\tau)$ only for $\tau \leq t$. Examples of important nonanticipating functions are

(i) $W(t)$,
(ii) $\int_{t_0}^{t} F\big(W(\tau)\big)\,\mathrm{d}\tau$,
(iii) $\int_{t_0}^{t} F\big(W(\tau)\big)\,\mathrm{d}W(\tau)$,
(iv) $\int_{t_0}^{t} G(\tau)\,\mathrm{d}\tau$, when $G(t)$ itself is nonanticipating, and
(v) $\int_{t_0}^{t} G(\tau)\,\mathrm{d}W(\tau)$, when $G(t)$ itself is nonanticipating.

The items (iii) and (v) depend on the fact that in Itō's version the stochastic integral is defined as the limit of a sequence in which $G(\tau)$ and $W(\tau)$ are involved exclusively for $\tau < t$.

Three reasons for the specific discussion of nonanticipating functions are important:
1. Many results can be derived that are only valid for nonanticipating functions.
2. Nonanticipating function occur naturally in situations, in which *causality* can be expected in the sense that the future cannot affect the presence.
3. The definition of stochastic differential equations requires nonanticipating

[31] A function assigns a value to the argument of the function, $x_0 \to f(x_0)$ whereas a functional relates a function to the value of a function, $f \to f(x_0)$.

functions.
In conventional calculus we never encounter situations in which the future acts back on the presence or even on the past.

Several relations are useful and required in Itō calculus:

$$\mathrm{d}W(t)^2 \;=\; \mathrm{d}t\,, \tag{3.124a}$$

$$\mathrm{d}W(t)^{2+n} \;=\; 0 \quad \text{for} \quad n>0\,, \tag{3.124b}$$

$$\mathrm{d}W(t)\,\mathrm{d}t \;=\; 0\,, \tag{3.124c}$$

$$\int_{t_0}^{t} W(\tau)^n\,\mathrm{d}W(\tau) \;=\; \frac{1}{n+1}\Big(W(t)^{n+1}-W(t_0)^{n+1}\Big) - \frac{n}{2}\int_{t_0}^{t} W(\tau)^{n-1}\,\mathrm{d}\tau\,, \tag{3.124d}$$

$$\mathrm{d}f\big(W(t),t\big) \;=\; \left(\frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial W^2}\right)\mathrm{d}t + \frac{\partial f}{\partial W}\,\mathrm{d}W(t)\;, \tag{3.124e}$$

$$\left\langle \int_{t_0}^{t} G(\tau)\,\mathrm{d}W(\tau)\right\rangle \;=\; 0\,, \quad \text{and} \tag{3.124f}$$

$$\left\langle \int_{t_0}^{t} G(\tau)\,\mathrm{d}W(\tau)\int_{t_0}^{t} H(\tau)\,\mathrm{d}W(\tau)\right\rangle \;=\; \int_{t_0}^{t} \langle G(\tau)H(\tau)\rangle\;\mathrm{d}\tau\;. \tag{3.124g}$$

The expressions are easier to memorize when we assign a dimension $[t^{1/2}]$ to $W(t)$ and discard all terms of order t^{1+n} with $n>0$.

3.4.2.3 Stratonovich stochastic integral

As said already, the value of a stochastic integral depends on the particular choice of the intermediate points, τ_i. The Russian physicist and engineer Ruslan Leontevich Stratonovich [269] and the American mathematician Donald LeRoy Fisk [86] developed simultaneously an alternative approach to Itō's stochastic integration, which is commonly called Stratonovich integration:[32]

$$\oint_{t_0}^{t} G(\tau)\,\mathrm{d}W(\tau)$$

The intermediate points are chosen such that the unconventional term $(t-t_0)$ does not appear any more. The integrand as a function of $W(t)$ is evaluated precisely in the middle, namely at the value $\tau_i=(t_i-t_{i-1})/2$ and it is straight-

[32] In order to distinguish the two versions of stochastic integrals we use the symbol $\int_{t_0}^{t}$ for the Itō integral and $\oint_{t_0}^{t}$ for the Stratonovich integral [144, p. 86]. The distinction from ordinary integrals is automatically provided by the differential $\mathrm{d}W$.

forward to show that the mean square limit converges to the expression for the integral over $W(t)$

$$\begin{aligned} \int\!\!\!\!\!\!\diagup_{t_0}^{t} W(\tau)\,\mathrm{d}W(\tau) \;&=\; \lim_{n\to\infty}\sum_{i=1}^{n}\frac{W(t_i)+W(t_{i-1})}{2}\Big(W(t_i)-W(t_{i-1})\Big) \;= \\ &=\; \frac{1}{2}\Big(W(t)^2 - W(t_0)^2\Big)\;. \end{aligned} \tag{3.125}$$

Stratonovich integration in contrast to the Itō integral obeys the rules of conventional calculus but the Stratonovich integral is not nonanticipating.

We compare here the derivation of the Stratonovich and the Itō integral [144, pp. 85-89], because additional insights into the nature of stochastic processes are gained. The starting point is the general Itō difference equation

$$\Delta x \;=\; F(x,t)\,\Delta t \;+\; G(x,t)\,\Delta W\;, \tag{3.126}$$

choosing $x_k = x(t_k)$, $t_k = k\,\Delta t$, and ΔW_0 as the first random increment we obtain $x_k = x_{k-1} + \Delta x_{k-1}$ with

$$\Delta x_{k-1} \;=\; F(x_{k-1},t_{k-1})\Delta t \;+\; G(x_{k-1},t_{k-1})\Delta W(t_{k-1})$$

for $k = 1,\ldots,n$ with equal intervals as shown in figure 3.19. We choose the starting point $t_0 = 0$ and $x(0) = x_0$, and find the general solution of the difference equation at $t = t_n$:

$$x_n \;=\; x(t_n) \;=\; x_0 \;+\; \sum_{k=0}^{n-1} F(x_k,t_k)\,\Delta t \;+\; \sum_{k=0}^{n-1} G(x_k,t_k)\,\Delta W(t_k)\;. \tag{3.127}$$

Equation (3.127) represents also the explicit formula for the Cauchy-Euler integration (figure 3.20) and is used in numerical SDE integration. We apply it here for definition the Itō integral

$$\int_0^t G(x,t)\,\mathrm{d}W \;\equiv\; \lim_{n\to\infty}\sum_{k=0}^{n-1} G(x_k,t_k)\,\Delta W(t_k)\;, \tag{3.120'}$$

and the Stratonovich integral

$$\int\!\!\!\!\!\!\diagup_0^t G\big(x,t\big)\,\mathrm{d}W \;\equiv\; \lim_{n\to\infty}\sum_{k=0}^{n-1} G\Big(\frac{x_{k+1}+x_k}{2},t_k\Big)\,\Delta W(t_k)\;, \tag{3.128}$$

and use it for a calculation of the relationship between them.

First we expand the function $G(x,t)$ in the Stratonovich analogue of the noise term in equation (3.127)

$$G\Big(\frac{x_{k+1}+x_k}{2},t_k\Big)\,\Delta W(t_k) \;=\; G\Big(x_k+\frac{\Delta x_k}{2},t_k\Big)\,\Delta W(t_k)\;,$$

in a power series around the point (x_n, t_n). For the expansion we simplify the notation by defining $F_n \equiv F(x_n, t_n)$ and $G_n \equiv G(x_n, t_n)$,

$$G\Big(x_k + \frac{\Delta x_k}{2}, t_k\Big) \;=\; G_n \;+\; \frac{\Delta x_n}{2}\,\frac{\partial G_n}{\partial x} \;+\; \Big(\frac{\Delta x_n}{2}\Big)^2\,\frac{1}{2}\,\frac{\partial^2 G_n}{\partial x^2} \;+\; \dots \;,$$

insert $\Delta x_n = F_n \Delta t + G_n \Delta W(t_n)$, and by considering that $\Delta W(t)^2 = \Delta t$ we find by omitting the higher order terms, because they will not contribute since all differences with higher powers, $(\Delta t)^\gamma$ with $\gamma > 1$ and $\big(\Delta W(t)\big)^\alpha$ with $\alpha > 2$ (3.124), vanish in the continuum limits $\Delta t \to \mathrm{dt}$ and $\Delta W \to \mathrm{d}W(t)$

$$G\Big(x_k + \frac{\Delta x_k}{2}, t_k\Big) \;=\; G_n + \left(\frac{F_n}{2}\frac{\partial G_n}{\partial x} + \frac{G_n^2}{4}\frac{\partial^2 G_n}{\partial x^2}\right)\Delta t + \frac{G_n}{2}\frac{\partial G_n}{\partial x}\,\Delta W(t_n)\,.$$

Next we insert this result into the discrete sum for the Stratonovich integral (3.128), omit the term with $\Delta t\,\Delta W$ since $\Delta t\,\Delta W \to \mathrm{dt}\,\mathrm{d}W(t) = 0$, and find

$$\sum_{k=0}^{n-1} G\Big(x_k + \frac{\Delta x_k}{2}, t_k\Big)\,\Delta W(t_k) \;=\; \sum_{k=0}^{n-1} G_k\,\Delta W(t_k) \;+\; \sum_{k=0}^{n-1} \frac{G_k}{2}\,\frac{\partial G_k}{\partial x}\,\Delta t \;.$$

Taking the continuum limit we obtain the desired relation between Itō and Stratonovich integrals

$$\oint_0^t G(x,t)\,\mathrm{d}W(t) \;=\; \int_0^t G(x,t)\,\mathrm{d}W(t) + \frac{1}{2}\int_0^t \frac{\partial G(x,t)}{\partial x}\,G(x,t)\;\mathrm{dt}\;. \quad (3.129)$$

The Stratonovich integral is equal to the Itō integral plus an additional contribution, which can be assimilated into the drift term.

In summary we derived two integration methods for the stochastic differential equation

$$\mathrm{d}x \;=\; F(x,t)\;\mathrm{dt} \;+\; G(x,t)\,\mathrm{d}W(t) \;\;: \quad (3.130)$$

(i) the Itō method yielding

$$x(t) \;=\; x(0) \;+\; \int_0^t F(x,t)\;\mathrm{dt} \;+\; \int_0^t G(x,t)\,\mathrm{d}W(t) \;\;\text{and}$$

(ii) the Stratonovich method resulting in a different solution, which we denote by $z(t)$ for the purpose of distinction

$$\begin{aligned} z(t) \;&=\; z(0) \;+\; \int_0^t F(z,t)\;\mathrm{dt} \;+\; \oint_0^t G(z,t)\,\mathrm{d}W(t) \;= \\ &=\; z(0) \;+\; \int_0^t \left(F(z,t) + \frac{G(z,t)}{2}\,\frac{\partial G(z,t)}{\partial z}\right)\;\mathrm{dt} \;+\; \int_0^t G(z,t)\,\mathrm{d}W(t)\;. \end{aligned}$$

On the other hand we would obtain the same solution $z(t)$ if we applied the Itō calculus to the stochastic differential equation

$$\mathrm{d}z \;=\; \left(F(z,t) + \frac{G(z,t)}{2}\,\frac{\partial G(z,t)}{\partial z} \right)\,\mathrm{dt} \;+\; G(z,t)\,\mathrm{d}W(t)\;. \tag{3.131}$$

Since the Stratonovich calculus is much more involved than the Itō calculus, we can readily see a strategy for obtaining Stratonovich solutions: Use equation (3.131) and derive the solution by means of Itō calculus. It is worth mentioning that a stand-alone Stratonovich integral has no relationship to a stand-alone Itō integral or, in other words, there is no connection between the two classes of integrals for an arbitrary function $G(t)$. When the stochastic differential equation is known to which the two integrals refer, a formula can be derived – as we did here – that relates the Itō integral to the Stratonovich integral.

At the end of this section we are left with the dilemma that the Itō integral is mathematically and technically most satisfactory but the more natural choice would be the Stratonovich integral that enables the usage of conventional calculus. In addition, the noise term $\xi(t)$ in the Stratonovich interpretation can be real noise with finite correlation time whereas the idealized white noise assumed as reference in Itō's formalism gives rise to divergence of variances and correlations. The Stratonovich and not the Itō calculus, for example, is adequate for dealing with multiplicative noise in physical systems.

3.4.3 Integration of stochastic differential equations

A stochastic variable $x(t)$ is consistent with an Itō stochastic differential equation (SDE)

$$\mathrm{d}x(t) \;=\; a\big(x(t),t\big)\,\mathrm{d}t \;+\; b\big(x(t),t\big)\,\mathrm{d}W(t) \tag{3.115'}$$

if for all t and t_0 the integral equation (3.117') is fulfilled. Time is ordered,

$$t_0 \;<\; t_1 \;<\; t_2 \;<\; \cdots \;<\; t_n \;=\; t\;,$$

and the time axis may be assumed to be split into (equal or unequal) increments, $\Delta t_i = t_{i+1} - t_i$. We visualize a particular solution curve of the SDE for the initial condition $x(t_0) = x_0$ by means of a discretized version

$$x_{i+1} \;=\; x_i \;+\; a(x_i,t_i)\,\Delta t_i \;+\; b(x_i,t_i)\,\Delta W(t_i)\;, \tag{3.117"}$$

wherein $x_i = x(t_i)$, $\Delta t_i = t_{i+1} - t_i$, and $\Delta W(t_i) = W(t_{i+1}) - W(t_i)$. Figure 3.20 illustrates the partitioning of the stochastic process into a deterministic drift component, which is the discretized solution curve of the ODE

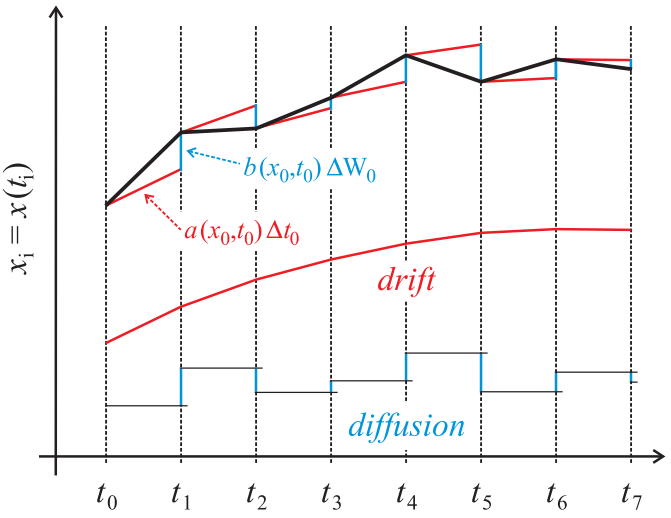

Fig. 3.20 Stochastic integration. The figure illustrates the Cauchy-Euler procedure for the construction of an approximate solution of the stochastic differential equation (3.115'). The stochastic process consists of two different components: (i) the drift term, which is the solution of the ODE in absence of diffusion (red; $b(x_i, t_i) = 0$) and (ii) the diffusion term representing a Wiener process $W(t)$ (blue; $a(x_i, t_i) = 0$). The superposition of the two terms gives the stochastic process (black). The two lower plots show the two components in separation. The increments of the Wiener process $\Delta W(t_i)$ are independent or uncorrelated. An approximation to a particular solution of the stochastic process is constructed by letting the mesh size approach zero, $\lim \Delta t \to 0$.

obtained by setting $b\big(x(t), t\big) = 0$ in equation (3.117"), and a stochastic diffusion component, which is a Wiener process $W(t)$ that is obtained by setting $a\big(x(t), t\big) = 0$ in the SDE. The increment of the Wiener process, $\Delta W(t_i)$, is independent of x_i provided (i) x_0 is independent of all $W(t) - W(t_0)$ for $t > t_0$ and (ii) $a(x, t)$ is a nonanticipating function of t for any fixed x. Condition (i) is tantamount to the requirement that any random initial condition must be nonanticipating.

A particular solution to equation (3.117") is constructed by letting the mesh size go to zero, $\lim n \to \infty$ implying $\Delta t \to 0$. In the construction of an approximate solution x_i is always independent of $\Delta W(t_j)$ for $j \geq i$ as we verify easily that by inspection of (3.117"). Uniqueness of solutions refers to individual trajectories in the sense that a particular solution is uniquely obtained for a given sample function $\mathcal{W}(t)$ of the Wiener Process $W(t)$. The existence of a solution is defined for the whole ensemble of sample functions: A solution of equation (3.117") exists if – with probability one – a particular solution exists for any choice of sample function $\mathcal{W}(t)$ of the Wiener process.

Existence and uniqueness of solutions to Itō stochastic differential equations can be proven for two conditions [7, pp.100-115]: (i) the Lipschitz condition and (ii) the growth condition. Existence and uniqueness of a nonanticipating solution $x(t)$ of an Itō SDE within the time interval $[t_0, t]$ require:

(i) *Lipschitz condition*: there exists a κ such that

$$|a(x,\tau) - a(y,\tau)| \,+\, |b(x,\tau) - b(y,\tau)| \,\leq\, \kappa\,|x-y|$$

for all x and y and $\tau \in [t_0, t]$, and

(ii) *growth condition*: a κ exists such that for all $\tau \in [t_0, t]$

$$|a(x,\tau)|^2 \,+\, |b(x,\tau)|^2 \,\leq\, \kappa^2\,(1+|x|^2)\;.$$

The Lipschitz condition is almost always fulfilled for stochastic differential equations in practice, because in essence it is a smoothness condition. The growth condition, however, may often be violated in abstract model equations, for example, when a solution *explodes* and progresses to infinity at finite time. In other words, the value of x may become infinite at some finite time. We shall encounter such situations in the applied chapter 5. As a matter of fact this is a typical model behavior since no population or spatial variable can approach infinity at finite times in a finite world.

Several other properties known to apply to solutions of ordinary differential equations can be shown without major modifications to apply to SDE's too: Continuity in the dependence on parameters and boundary conditions as well as the Markov property (for proofs we refer to [7]).

3.4.4 Changing variables in Itō calculus

Changing variables is a technical issue but important for applications and boring when one makes errors. Since Itō calculus is different from ordinary calculus, we expect differences also in the rules of substituting variables. In order to see the general effect of substitutions in Itō's stochastic differential equations we consider an arbitrary function, $\mathbf{x}(t) \Rightarrow f\big(\mathbf{x}(t)\big)$, and calculate $\mathrm{d}x(t) \Rightarrow \mathrm{d}f\big(\mathbf{x}(t)\big)$. The major difference compared to ordinary calculus comes from the necessity to extend all expansions up to second order because $\mathrm{d}W(t)^2 = \mathrm{dt}$ and hence $\Delta W(t)^2$ does not approach zero faster than Δt in the limit $\Delta t \to \mathrm{dt}$. We start with the simpler case of a single variable and afterwards introduce the multidimensional situation.

3.4.4.1 Single variable case

Starting out from the SDE $\mathrm{d}x = a(x,t)\,\mathrm{d}t + b(x,t)\mathrm{d}W(t)$ and making use of our previous results on nonanticipating functions we expand $\mathrm{d}f\big(x(t)\big)$ up to second order but retain only the term in $\mathrm{d}W(t)$, because by the Itō rules we have $\mathrm{dt}^2 = 0$ and $\mathrm{d}W(t)\,\mathrm{dt} = 0$ (and write x instead $x(t)$):

$$\begin{aligned} \mathrm{d}f(x) \;&=\; f\big(x+\mathrm{d}x\big) \,-\, f\big(x\big) \;= \\ &=\; \frac{\partial f(x)}{\partial x}\,\mathrm{d}x \,+\, \frac{1}{2}\frac{\partial^2 f(x)}{\partial x^2}\,\mathrm{d}x^2 \,+\, \cdots \;= \\ &=\; \frac{\partial f(x)}{\partial x}\,\Big(a(x,t)\,\mathrm{d}t \,+\, b(x,t)\,\mathrm{d}W(t)\Big) \,+\, \frac{1}{2}\frac{\partial^2 f(x)}{\partial x^2}\,b(x,t)^2\,\mathrm{d}W(t)^2\,, \end{aligned} \tag{3.132}$$

where all terms higher than second order have been neglected. According to Itō calculus (3.124) we introduce $\mathrm{d}W(t)^2 = \mathrm{d}t$ into the last line of this equation and obtain Itō's formula:

$$\begin{aligned} \mathrm{d}f\big(x(t)\big) \;=\; &\Big(a\big(x(t),t\big)\,\frac{\partial f\big(x(t)\big)}{\partial x} \,+\, \frac{1}{2}\,b\big(x(t),t\big)^2\,\frac{\partial^2 f\big(x(t)\big)}{\partial x^2}\Big)\,\mathrm{d}t + \\ &+\, b\big(x(t),t\big)\,\frac{\partial\big(x(t)\big)}{\partial x}\,\mathrm{d}W(t)\;. \end{aligned} \tag{3.133}$$

It is worth noticing that Itō's formula and ordinary calculus lead to different results unless $f(x)$ is linear in $x(t)$ and accordingly $\frac{\partial^2 f(x)}{\partial x^2}$ vanishes.

As an exercise we suggest to calculate the substitution by the function $f(x) = x^2$. The result is

$$\mathrm{d}(x^2) \;=\; \big(2x\,a(x,t) \,+\, b(x,t)^2\big)\,\mathrm{dt} \,+\, 2\,b(x,t)\,\mathrm{d}W(t)\;,$$

which is, for example, useful to calculate the time derivative of the variance: $\mathrm{d}\,\mathrm{var}\big(x(t)\big)/\,\mathrm{dt} = \mathrm{d}\,\big\langle x^2\big\rangle/\,\mathrm{dt} + 2\,\langle x\rangle\;\mathrm{d}\,\langle x\rangle/\,\mathrm{dt}$.

3.4.4.2 Multiple variable case

The application of Itō's formalism to many dimensions, in general, becomes very complicated. The most straightforward simplification is the extension of Itō calculus to the multivariate case by making use of the rule that $\mathrm{d}W(t)$ is an infinitesimal of order $t^{1/2}$. Then we can show that the following relations hold for an n-dimensional Wiener process $\mathbf{W}(t) = \big(W_1(t), W_2(t), \ldots, W_n(t)\big)$:

$$\mathrm{d}W_i(t)\,\mathrm{d}W_j(t) \;=\; \delta_{ij}\;\mathrm{dt}\;, \tag{3.134a}$$

$$\mathrm{d}W_i(t)^{2+N} \;=\; 0\;, \qquad (N>0)\;, \tag{3.134b}$$

$$\mathrm{d}W_i(t)\;\mathrm{dt} \;=\; 0\;, \tag{3.134c}$$

$$\mathrm{dt}^{1+N} \;=\; 0\;, \qquad (N>0)\;. \tag{3.134d}$$

The first relation is a consequence of the independence of increments of Wiener processes along different coordinate axes, $\mathrm{d}W_i(t)$ and $\mathrm{d}W_j(t)$. Making use of the drift vector $\mathbf{A}(\mathbf{x},t)$ and the diffusion matrix $\mathrm{B}(\mathbf{x},t)$ the multidimensional stochastic differential equation

$$\mathrm{d}\mathbf{x} \;=\; \mathbf{A}(\mathbf{x},t)\;\mathrm{dt} \;+\; \mathrm{B}(\mathbf{x},t)\,\mathrm{d}\mathbf{W}(t)\;. \tag{3.135}$$

Following Itō's procedure we obtain for an arbitrary well-behaved function $f\big(\mathbf{x}(t)\big)$ the result

$$\begin{aligned}\mathrm{d}f(\mathbf{x}) \;=\; &\Bigg(\sum_i A_i(x,t)\,\frac{\partial}{\partial x_i}f(\mathbf{x}) + \\ &+\;\frac{1}{2}\sum_{i,j}\big(\mathrm{B}(x,t)\cdot\mathrm{B}'(x,t)\big)_{ij}\;\frac{\partial^2}{\partial x_i\partial x_j}f(\mathbf{x})\Bigg)\,\mathrm{d}t + \\ &+\;\sum_{i,j}B_{ij}\frac{\partial}{\partial x_i}f(\mathbf{x})\,\mathrm{d}W_j(t)\;.\end{aligned} \tag{3.136}$$

Again we observe the additional term introduced through the definition of the Itō integral.

3.4.5 Fokker-Planck equations and SDEs

The expectation value of an arbitrary function $f\big(x(t)\big)$ can be calculated by means of Itō's formula. We begin with a single variable:

$$\begin{aligned}\frac{\big\langle \mathrm{d}f\big(x(t)\big)\big\rangle}{\mathrm{d}t} \;&=\; \left\langle\frac{\mathrm{d}f\big(x(t)\big)}{\mathrm{dt}}\right\rangle \;=\; \frac{\mathrm{d}}{\mathrm{d}t}\big\langle f\big(x(t)\big)\big\rangle = \\ &=\; \left\langle a\big(x(t),t\big)\frac{\partial f\big(x(t)\big)}{\partial x} \;+\; \frac{1}{2}\,b\big(x(t),t\big)\frac{\partial^2 f\big(x(t)\big)}{\partial x^2}\right\rangle\;.\end{aligned}$$

The stochastic variable $\mathcal{X}(t)$ has the conditional probability density $p(x,t|\,x_0,t_0)$ and hence we can compute the expectation value by integration – again we simplify notation $f(x)\equiv f\big(x(t)\big)$ and $p\,(x,t)\equiv p\,(x,t|\,x_0,t_0)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle f(x)\rangle \;=\; \int \mathrm{d}x\, f(x)\,\frac{\partial}{\partial t}p(x,t) \;=$$
$$\;=\; \int \mathrm{d}x\left(a(x,t)\,\frac{\partial f(x)}{\partial x} \;+\; \frac{1}{2}\,b(x,t)^2\,\frac{\partial^2 f(x)}{\partial x^2}\right)p(x,t)$$

The further derivation follows the procedure that is used in the of the differential Chapman-Kolmogorov equation [93, 48-51] – in particular integration by parts and neglect of surface terms – and we obtain

$$\int \mathrm{d}x\, f(x)\,\frac{\partial}{\partial t}p(x,t) \;=\; \int \mathrm{d}x\, f(x)\Big(-\frac{\partial}{\partial x}\Big(A(x,t)\,p(x,t)\Big)+\frac{1}{2}\frac{\partial^2}{\partial x^2}\Big(B(x,t)^2\,p(x,t)\Big)\Big)\,.$$

Since the choice of a function $f(x)$ has been arbitrary we can drop it now and finally obtain a Fokker-Planck equation

$$\begin{aligned}\frac{\partial p(x,t|\,x_0,t_0)}{\partial t} \;&=\; -\;\frac{\partial}{\partial x}\Big(A(x,t)\,p(x,t|\,x_0,t_0)\Big)\;+\\ &\quad+\;\frac{1}{2}\frac{\partial^2}{\partial x^2}\Big(B(x,t)^2\,p(x,t|\,x_0,t_0)\Big)\;.\end{aligned}\tag{3.137}$$

The probability density $p(x,t)$ thus obeys an equation that is completely equivalent to the equation for a diffusion process characterized by a drift coefficient $a(x,t)\equiv A(x,t)$ and a diffusion coefficient $b(x,t)\equiv B(x,t)$ as derived from the Chapman-Kolmogorov equation. Hence, Itō's stochastic differential equation provides indeed a local approximation to a drift and diffusion process in probability space. The extension to the multidimensional case based on Itō's formula is straightforward, and we obtain for the conditional probability density $p(\mathbf{x},t|\mathbf{x}_0,t_0)\equiv p$ the following Fokker-Planck equation:

$$\frac{\partial p}{\partial t} \;=\; -\sum_i \frac{\partial}{\partial x_i}\big(A_i(\mathbf{x},t)\,p\big) \;+\; \frac{1}{2}\sum_{i,j}\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\Big(\big(\mathrm{B}(\mathbf{x},t)\cdot\mathrm{B}'(\mathbf{x},t)\big)_{i,j}\;p\Big)\;.\tag{3.138}$$

Here, we derive one additional property, which is relevant in practice. The stochastic differential equation,

$$\mathrm{d}\mathbf{x} \;=\; \mathbf{A}(\mathbf{x},t)\;\mathrm{dt} \;+\; \mathrm{B}(\mathbf{x},t)\,\mathrm{d}\mathbf{W}(t)\;,\tag{3.135'}$$

is mapped into a Fokker-Planck equation that depends only on the matrix product $\mathrm{B}\cdot\mathrm{B}'$ and accordingly, the same Fokker-Planck equation arises from all matrices B that give rise to the same product $\mathrm{B}\cdot\mathrm{B}'$. Thus, the Fokker-Planck equation is invariant to a replacement $\mathrm{B}\Rightarrow\mathrm{B}\cdot\mathrm{S}$ when S is an orthogonal matrix: $\mathrm{S}\cdot\mathrm{S}'=\mathbb{I}$. If S fulfils the orthogonality relation it may depend on $\mathbf{x}(t)$, but for the stochastic handling it has to be *nonanticipating*.

Eventually we proof the redundancy directly from the SDE and define a transformed Wiener process

$$\mathrm{d}\mathbf{V}(t) \;=\; \mathrm{S}(t)\,\mathrm{d}\mathbf{W}(t)\;.$$

The random vector $\mathbf{V}(t)$ is a normalized linear combination of Gaussian variables $\mathrm{d}W_i(t)$ and $\mathrm{S}(t)$ in nonanticipating, and accordingly, $\mathrm{d}\mathbf{V}(t)$ is itself Gaussian with the same correlation matrix. Averages $\mathrm{d}W_i(t)$ to various powers and taken at different times factorize and the same is true for the $\mathrm{d}V_i(t)$. Accordingly, the infinitesimal elements $\mathrm{d}\mathbf{V}(t)$ are increments of a Wiener process: The orthogonal transformation mixes trajectories without, however, changing the stochastic nature of the process, and equation (3.135) can be rewritten and yields

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{A}(\mathbf{x},t)\ \mathrm{dt} + \mathrm{B}(\mathbf{x},t)\,\mathrm{S}'(t)\cdot\mathrm{S}(t)\,\mathrm{d}\mathbf{W}(t) = \\ &= \mathbf{A}(\mathbf{x},t)\ \mathrm{dt} + \mathrm{B}(\mathbf{x},t)\,\mathrm{S}'(t)\cdot\,\mathrm{d}\mathbf{V}(t) = \\ &= \mathbf{A}(\mathbf{x},t)\ \mathrm{dt} + \mathrm{B}(\mathbf{x},t)\,\mathrm{S}'(t)\cdot\,\mathrm{d}\mathbf{W}(t)\ , \end{aligned}$$

since $\mathbf{V}(t)$ is as good a Wiener process as $\mathbf{W}(t)$ is, and both SDEs give rise to the same Fokker-Planck equation. □

3.4.6 Examples of stochastic differential equations

In order to show how stochastic differential equations can be handled in practice we show how to calculate first the expectation value and the variance of stochastic differential equations and then consider two cases: (i) the Ornstein-Uhlenbeck process that has been discussed as an example of a process that can be handled easily with a Fokker-Planck equation in section 3.2.3.4, and (ii) the general linear stochastic differential equations.

3.4.6.1 Low moments of stochastic differential equations

In many cases it is sufficient to know the expectation value and the variance of the stochastic variable of a process as a function of time. These low moments[33] can be calculated without solving the stochastic equations explicitly. We consider the general SDE

$$\mathrm{d}x = a(x,t)\ \mathrm{dt} + b(x,t)\,\mathrm{d}W(t)$$

and compute the mean value by taking the average and recall that the second term on the r.h.s. vanishes because $\langle \mathrm{d}W(t)\rangle = 0$:

$$\mathrm{d}\,\langle x\rangle = \langle \mathrm{d}x\rangle = \langle a(x,t)\rangle\ \mathrm{dt} \quad \text{or} \quad \frac{\mathrm{d}\langle x\rangle}{\mathrm{dt}} = \langle a(x,t)\rangle\ . \tag{3.139}$$

[33] Expectation value and variance are considered as low moments.

Thus, the calculation of the expectation value boils down to solving an ODE. For a derivation of an expression for the second moment and the variance we have to calculate the differential of the square of the variable. By means of equation (3.133) we find:

$$\mathrm{d}(x^2) \;=\; \big(2x\,a(x,t) \,+\, b(x,t)^2\big)\;\mathrm{dt} \,+\, 2\,b(x,t)\,\mathrm{d}W(t)\;,$$

and forming the average yields

$$\big\langle \mathrm{d}(x^2)\big\rangle \;=\; \mathrm{d}\,\big\langle x^2\big\rangle \;=\big\langle 2x\,a(x,t) \,+\, b(x,t)^2\big\rangle\;\mathrm{dt}\;,$$

where we made use of the relation $\langle \mathrm{d}W(t)\rangle = 0$. Provided we knew the expectation values, a differential equation for the variance would be given by

$$\frac{\mathrm{d}\,\mathrm{var}(x)}{\mathrm{dt}} \;=\; \frac{\mathrm{d}\big\langle x^2\big\rangle}{\mathrm{dt}} - \frac{\mathrm{d}\langle x\rangle^2}{\mathrm{dt}} \;=\; \frac{\mathrm{d}\big\langle x^2\big\rangle}{\mathrm{dt}} - 2\,\langle x\rangle\;\frac{\mathrm{d}\langle x\rangle}{\mathrm{dt}}\;.$$

The continuation of the calculations requires knowledge of the functions $a(x,t)$ and $b(x,t)$.

As an example we consider the simple linear SDE with $a(x,t) = \alpha x$ and $b(x,t) = \beta x$,

$$\mathrm{d}x \;=\; \alpha\,x\,\mathrm{dt} \,+\, \beta\,x\,\mathrm{d}W(t) \;=\; x\,\big(\alpha\;\mathrm{dt} \,+\, \beta\,\mathrm{d}W(t)\big)\;,$$

and find for the expectation value

$$\langle x(t)\rangle \;=\; \langle x(0)\rangle\;e^{\alpha t} \;=\; x_0\,e^{\alpha t} \quad\text{for}\quad p\,(x,0) \,=\, \delta(x-x_0) \tag{3.140a}$$

and for the variance

$$\begin{aligned}\mathrm{var}\big(x(t)\big) \;&=\; \big\langle x(t)^2\big\rangle \,-\, \langle x(t)\rangle^2 \;=\\ &=\; \big\langle x(0)^2\big\rangle\;e^{(2\alpha+\beta^2)t} \,-\, \langle x(0)\rangle^2\;e^{2\alpha t} \;=\\ &=\; x_0^2\;\Big(e^{(2\alpha+\beta^2)t} \,-\, e^{2\alpha t}\Big) \quad\text{for}\quad p\,(x,0) \,=\, \delta(x-x_0)\;.\end{aligned} \tag{3.140b}$$

The expressions are easily generalized to time dependent coefficients $\alpha(t)$ and $\beta(t)$ as we shall see in section 3.4.6.3.

3.4.6.2 The Ornstein-Uhlenbeck process

The general SDE for the Ornstein-Uhlenbeck process has been given in (3.62). Without loosing generality but simplifying the solution we shift the long-time expectation value to the origin, $\mu = 0$:

$$\mathrm{d}x \;=\; -k\,x\,\mathrm{dt} \,+\, \sigma\,\mathrm{d}W(t)\;. \tag{3.62'}$$

The solution of the deterministic equation is simply and exponential decay or relaxation to the long-time value $\lim_{t\to\infty} x(t) = 0$,

$$\mathrm{d}x \;=\; -k\,x\;\mathrm{dt} \quad\text{and}\quad x(t) \;=\; x(0)\,e^{-kt}\;,$$

and we make a substitution that compensates for the exponential decay

$$x(t) \;=\; y(t)\,e^{-kt} \quad\text{and}\quad y(t) \;=\; x(t)\,e^{kt} \quad\text{with}\quad y(0) \;=\; x(0)\;.$$

Now we expand $\mathrm{d}y$ up to second order

$$\mathrm{d}y \;=\; \mathrm{d}x\,e^{kt} + x\,\mathrm{d}(e^{kt}) + (\mathrm{d}x)^2 + \mathrm{d}x\,\mathrm{d}(e^{kt}) + \big(\mathrm{d}(e^{kt})\big)^2 \quad\text{with}\quad \mathrm{d}\big(e^{kt}\big) \;=\; ke^{kt}\,\mathrm{dt}\,.$$

All second order terms vanish because the expansion contains no term with $\mathrm{d}W(t)^2$ and we find by integration,

$$\mathrm{d}y \;=\; \sigma\,e^{kt}\,\mathrm{d}W(t) \quad\text{and}\quad y(t) \;=\; y(0) \;+\; \sigma\int_0^t e^{k\tau}\,\mathrm{d}W(\tau)\;,$$

and resubstitution yields the solution

$$x(t) \;=\; x(0)\,e^{-kt} \;+\; \sigma\int_0^t e^{-k(t-\tau)}\,\mathrm{d}W(\tau)\;. \tag{3.141}$$

The calculation of expectation value and variance is straightforward:

$$\langle x(t)\rangle \;=\; \left\langle x(0)\,e^{-kt} \;+\; \sigma\int_0^t e^{-k(t-\tau)}\,\mathrm{d}W(\tau)\right\rangle \;=\; \langle x(0)\rangle\;e^{-kt}\;, \tag{3.142a}$$

and with

$$\begin{aligned}\big\langle x(t)^2\big\rangle \;&=\; \left\langle\Big(x(0)\,e^{-kt} \;+\; \sigma\int_0^t e^{-k(t-\tau)}\,\mathrm{d}W(\tau)\Big)^2\right\rangle \;=\\ &=\; \big\langle x(0)^2\big\rangle\;e^{-2kt} \;+\; \frac{\sigma^2}{2k}\left(1 - e^{-2kt}\right)\end{aligned}$$

we obtain

$$\mathrm{var}\big(x(t)\big) \;=\; \left(\mathrm{var}\big(x(0)\big) - \frac{\sigma^2}{2k}\right)\,e^{-2kt} \;+\; \frac{\sigma^2}{2k}\;, \tag{3.142b}$$

and with sharp initial conditions, $p(x,0) = \delta(x - x_0)$, we find

$$\mathrm{var}\big(x(t)\big) \;=\; \left(\frac{1}{2k}\left(1 - e^{-2kt}\right)\right)\;. \tag{3.142c}$$

Finally we mention that the analysis of the Ornstein-Uhlenbeck process can be readily extended to many dimensions and time dependent parameters, $k(t)$ and $\sigma(t)$ [93].

3.4.6.3 The linear stochastic differential equations

As last example we consider again the linear SDE but allow time dependent parameters

$$\mathrm{d}x \;=\; \alpha(t)\,x\,\mathrm{d}t \;+\; \beta(t)\,x\,\mathrm{d}W(t) \;=\; x\,\big(\alpha(t)\;\mathrm{d}t \;+\; \beta(t)\,\mathrm{d}W(t)\big)\;.$$

Now we make the substitution $y = \ln x$, expand up to second order

$$\mathrm{d}y \;=\; \frac{\mathrm{d}x}{x} - \frac{\mathrm{d}x^2}{x^2} \;=\; \alpha(t)\;\mathrm{d}t + \beta(t)\,\mathrm{d}W(t) \;-\; \frac{1}{2}\beta(t)^2\,\mathrm{d}t$$

and find the solution by integration and resubstitution

$$x(t) \;=\; x(0)\,\exp\left(\int_0^t \Big(\alpha(\tau) - \frac{1}{2}\beta(\tau)^2\Big)\,\mathrm{d}\tau \;+\; \int_0^t \beta(\tau)\,\mathrm{d}W(\tau)\right)\;. \tag{3.143}$$

We make use of the relation $\langle e^z\rangle = \exp\big(\frac{1}{2}\langle z^2\rangle\big)$, which is fulfilled by all Gaussian variables,[34] and find for the n-th raw moment [93, p. 109]:

$$\begin{aligned}\langle x(t)^n\rangle \;&=\; \langle x(t)^n\rangle\,\left\langle \exp\Big(n\int_0^t \big(\alpha(\tau) - \frac{1}{2}\beta(\tau)^2\big)\,\mathrm{d}\tau \;+\; n\int_0^t \beta(\tau)\,\mathrm{d}W(\tau)\Big)\right\rangle \\ &=\; \langle x(t)^n\rangle\,\exp\Big(n\int_0^t \alpha(\tau)\,\mathrm{d}\tau \;+\; \frac{1}{2}\,n(n-1)\int_0^t \beta(\tau)^2\,\mathrm{d}\tau\Big)\;.\end{aligned} \tag{3.144}$$

All moments can be calculated from this expression and for the low moments we find:

$$\langle x(t)\rangle \;=\; \langle x(0)\rangle\,\exp\left(\int_0^t \alpha(\tau)\,\mathrm{d}\tau\right) \tag{3.145a}$$

$$\begin{aligned}\mathrm{var}\big(x(t)\big) \;&=\; \mathrm{var}\big(x(0)\big)\,\exp\left(2\int_0^t \alpha(\tau)\,\mathrm{d}\tau\right) \;+ \\ &\quad +\; \big\langle x(0)^2\big\rangle\,\exp\left(\int_0^t \beta(\tau)^2\,\mathrm{d}\tau\right)\;.\end{aligned} \tag{3.145b}$$

Analytical solutions have been derived also for the inhomogeneous case, $a(x,t) = \alpha_0 + \alpha_1 x$ and $b(x,t) = \beta_0 + \beta_1 x$ and the raw moments are readily calculated [93, p. 109].

[34] In order to proof the conjecture one makes use of the fact that all cumulants κ_n with $n > 2$ vanish (see section 2.3.3). The reader encouraged to complete the proof.

In this last part we have shown that analytical expressions derived from stochastic differential equations can be used successfully to compute the most important quantities of stochastic processes and in this sense are also equivalent to Fokker-Planck equations in practice.

Chapter 4
Applications in chemistry

There is nothing so practical as a good theory.
Kurt Lewin, 1952.

Abstract In chemistry the master equation is the best suited and most commonly used tool to model stochasticity in chemical reactions. We review the common elementary reaction in mass action kinetics and discuss Michaelis-Menten as kinetics as an example of combining several elementary steps into an over-all reaction. Reaction networks are considered and a formal mathematical theory that allows for the derivation of general properties of networks is presented. After a formal introduction of the chemical master equation we digress into the origin of rate parameters. Then, a selection of simple reactions is presented for which the master equation can be solved exactly. The exact solutions are also used to illustrate the relation between the mathematical approach and the recorded data. A separate chapter is dealing with correlation functions, fluctuation spectroscopy, single molecule data and their stochastic modeling. Deterministic and stochastic parts of solutions can be separated by means of size expansions.Most reaction mechanisms are not accessible to the analytical approach and therefore we present a numerical approach that is exact within the concept of the chemical master equation is presented and applied to some selected examples of chemical reactions.

Conventional chemical reaction kinetics commonly does not require a stochastic approach because the numbers of particles are very large. There are exceptions when the particle numbers of certain species become very small during reactions – oscillations of species may serve as examples. Such cases will be mentioned and discussed in this and in the next chapter but even more important is the requirement of a stochastic approach for direct measurements of fluctuations, which became possible because of the progress in spectroscopy leading to spectacular increases in sensitivities. Single molecule techniques are another not completely unrelated and also rapidly developing field where a stochastic approach is indispensable. On the other hand, if one wants to resolve reaction dynamics at the molecular level the situation is different, because conventional statistical mechanics is blurring the details of interest. Molecules are involved in large numbers of collisions, which

considered individually in the vapor phase could be calculated by means of advanced quantum mechanics – at least in principle, although we have to admit that the situation in solution where molecules are densely packed would be helpless.

Stochastic chemical kinetics is based on the assumption that knowledge on the transformation of molecules in chemical reactions is not accessible in full atomistic detail or if it would, the information would be overwhelming and obscuring the essential features. Thus, it is assumed that chemical reactions have a probabilistic element and can be modeled properly by means of stochastic processes. The random processes are caused by thermal noise as well as by random encounters of molecules in collisions. Fluctuations, therefore, play an important role and they are responsible for the limitations in the reproduction of experiments. This concept is not substantially different from the ideas underlying equilibrium statistical mechanics although statistics applied to thermodynamic equilibrium is on saver grounds than statistics applied to chemical reaction kinetics. On the other hand, the current theory of chemical reaction rates is around for more than fifty years and so far it has not yet been replaced by some better founded and applicable theory.

Particle numbers change necessarily in jumps requiring a discrete stochastic description, for example, by means of a master equation. Other descriptions are branching processes and other special cases of stochastic processes, which we will shall discuss in the next chapter 5, because they are more frequently addressed in biology. Commonly different approaches do not exclude each other as, for example, birth-and-death processes are frequently solved by application of precisely the same techniques as used for master equations. Birth-and-death master equations were already discussed in section 3.2.5.2. Continuous descriptions play a role in the case of the population size expansions, which allow for the separation of a deterministic part of the solution from a diffusion term.

Conventional chemical reaction kinetics is dealing, in essence, with two classes of problems: (i) *forward problems*, which deal with the determination of time dependent concentrations as solutions of kinetic model equations, where the kinetic parameters are assumed to be known (for an introduction in to traditional chemical kinetics see [171], a modern textbook is [136]), and (ii) *inverse problems*, which aim at the determination of parameters from measured data, where the kinetic model is commonly assumed to be known [275]. The first problem boils down to deriving the solution curves or performing qualitatively analysis of a kinetic ODE, or a PDE in case the spatial distribution is nonhomogeneous. The inverse task is often addressed as parameter identification problem. Qualitative analysis allows for a reconstruction of bifurcation patterns of dynamical systems, and there exists an inverse problem too: The determination of the regions in parameter space from where parameter combinations give rise to a certain dynamic behavior. In order to distinguish from simple or *level one* parameter identification the inversion

of the determination of bifurcation diagrams has been as a typical *level two* inverse problem [61].

The chapter starts with an introduction into chemical kinetics (section 4.1) consisting of short reviews of elementary step kinetics (section 4.1.1), Michaelis-Menten kinetics, which is discussed as an example of a reaction mechanism merging single elementary steps into one over-all reaction (section 4.1.2), and a formal theory of reaction networks conceived for the qualitative analysis of multidimensional kinetic differential equations (section 4.1.3). Stochasticity in chemical reactions is introduced in terms of the chemical master equation and we shall ask the question how the parameters are derived, which are required for modeling stochastic chemical kinetics (section 4.2). Then, examples of exactly solvable chemical master equations are presented: (i) the equilibration of particle numbers or concentrations in the flow reactor, (ii) irreversible and reversible monomolecular reactions, and (iii) bimolecular reactions that can be still solved exactly but where the solutions become so complicated that practical work with them has to rely on numerical computation (section 4.3). A separate chapter is dealing with correlation functions, fluctuation spectroscopy, single molecule techniques and their implications for stochastic modeling (section 4.5). The next section deals with the transition from microscopic to macroscopic systems by means of the size expansion technique. Size expansion is particularly useful if the particle numbers are sufficiently large (section 4.6). Most reaction mechanisms involve many reactions steps and commonly analytical solutions are neither available for the conventional deterministic approach nor for stochastic methods. Stochastic methods applied to reaction networks are discussed in section 4.4. The last sections handle a numerical approach to stochastic chemical kinetics in which probability distribution are obtained by sampling a sufficiently large number of numerically calculated individual trajectories (section 4.7).

4.1 A glance on chemical reaction kinetics

Chemical reactions will be modeled as Markov processes and analyzed in form of the corresponding master equations. In a few cases Fokker-Planck equations will be applied. As an appropriate criterium for classification of single elementary steps we shall use the molecularity of reactions[1] and the complexity of the dynamical behavior. With respect to reaction dynamics we shall consider reactions and reaction networks with (i) linear behavior, (ii) nonlin-

[1] The *molecularity* of a reaction is the number of molecules that are involved in the reaction, for example two in a reactive collision between molecules or one in a conformational change. An elementary step is a reaction at the molecular level that cannot be resolved further in mass action kinetics (section 4.1.1). We shall distinguish elementary steps and elementary processes: the latter are more general and need not be referring to the level of molecules.

ear behavior with simple dynamics in the sense of a monotonous approach to thermodynamic equilibrium or towards a unique stationary state, and (iii) complex behavior as exhibited by dynamical systems showing multiple stable stationary states, oscillations or deterministic chaos.

The stochastic approach to chemical reaction kinetics has some tradition, which began in the late fifties from two different initiatives: (i) approximation of the complex vibrational relaxation in small molecules [15, 218, 264] and its application to chemical reactions, and (ii) direct simulation of chemical reactions as stochastic processes [12, 13, 14]. The latter approach can be viewed in the sense of initially mentioned limited information on reaction details and has been taken up and developed further by several groups [47, 139, 161, 202, 205]. The major part of these works has been summarized in an early review [203], which is recommended here for further reading. Anthony Bartholomay's studies are also highly relevant for biological models of evolution, because he investigated reproduction as a linear birth-and-death process. Exact solutions to master or Fokker-Planck equations can be found only for particularly simple special cases. Often approximations are used or the analysis has been restricted to expectation values and variances of the variables. Later on computer assisted approximation techniques and numerical simulation methods were developed, which allow for handling stochastic phenomena in chemical kinetics on a general level [93, 106, 286].

4.1.1 Elementary steps of chemical reactions

Chemical reactions at the level of mass action kinetics are defined by mechanisms, which can be decomposed into *elementary steps*. An elementary step describes the transformation of zero, one or two molecules into products. Common elementary steps written as stoichiometric equations are:[2]

$$\star \longrightarrow \mathbf{A} \tag{4.1a}$$

$$\mathbf{A} \longrightarrow \oslash \tag{4.1b}$$

$$\mathbf{A} \longrightarrow \mathbf{B} \tag{4.1c}$$

$$\mathbf{A} \longrightarrow 2\,\mathbf{B} \tag{4.1d}$$

$$\mathbf{A} \longrightarrow \mathbf{B} + \mathbf{C} \tag{4.1e}$$

$$\mathbf{A} + \mathbf{B} \longrightarrow \mathbf{C} \tag{4.1f}$$

$$\mathbf{A} + \mathbf{B} \longrightarrow 2\,\mathbf{A} \tag{4.1g}$$

$$\mathbf{A} + \mathbf{B} \longrightarrow \mathbf{C} + \mathbf{B} \tag{4.1h}$$

$$\mathbf{A} + \mathbf{B} \longrightarrow \mathbf{C} + \mathbf{D} \tag{4.1i}$$

$$2\,\mathbf{A} \longrightarrow \mathbf{B} \tag{4.1j}$$

$$2\,\mathbf{A} \longrightarrow 2\,\mathbf{B} \tag{4.1k}$$

$$2\,\mathbf{A} \longrightarrow \mathbf{B} + \mathbf{C} \tag{4.1l}$$

The molecularity of a reaction is defined by the number of – different or identical – molecules on the reactant side of the stoichiometric equation and we distinguish *zero-*, *mono-*, *bi-*, or *termolecular*, reactions and so on. The list shown above contains one zero-molecular reaction, (4.1a), four monomolecular reactions, (4.1b) - (4.1e), and seven bimolecular reactions, (4.1f) - (4.1l). Nonreactive events, which occur in open systems, for example in flow reactors, like the creation of a molecules through influx (4.1a) or the annihilation of a molecule through outflux (4.1b) are included in the list. Molecularities of three and higher are not included in the list, because simultaneous encounters of three and more molecules are extremely improbable and therefore, elementary steps involving three or more molecules are not considered in conventional chemical kinetics.[3]

---

[2] Stoichiometry deals with the relative quantities of reactants and products in chemical reactions. Reaction stoichiometry, in particular, determines the molar ratios of the reactants, which are converted into products, and the products that are formed. For example, in the reaction $2\mathbf{H}_2 + \mathbf{O}_2 \rightarrow 2\mathbf{H}_2\mathbf{O}$ the stoichiometric ratios of $\mathbf{H}_2 : \mathbf{O}_2 : \mathbf{H}_2\mathbf{O}$ are $2 : 1 : 2$ (see also the stoichiometric matrix in section 4.7.2).

[3] Exceptions are reactions involving surfaces as third partner, which are important in gas phase kinetics and, for example, biochemical reactions involving macromolecules.

The elementary step shown in equation (4.1g) is an example of an *autocatalytic* elementary process. In practice, autocatalytic reactions commonly involve many elementary steps and are the results of complex reaction mechanisms (see, e.g., the review [252]). In order to study basic features of autocatalysis or chemical self-enhancement, single step autocatalytic reactions rather than autocatalytic multistep reaction networks are used as model systems. One particular termolecular autocatalytic process,

$$2\,\mathbf{A} \;+\; \mathbf{B} \quad \longrightarrow \quad 3\,\mathbf{A}\;, \tag{4.2}$$

became very famous [224] despite its termolecular nature, which makes it unlikely to occur in real systems. The elementary step (4.2) is the essential step in the so-called Brusselator model, it can be straightforwardly addresses by analytical mathematical techniques, and it gives rise to complex dynamical phenomena in space and time which are otherwise rarely observed in chemical reaction systems. Among other features such special phenomena are: (i) multiple stationary states, (ii) chemical hysteresis, (iii) oscillations in concentrations, (iv) deterministic chaos, and (v) spontaneous formation of spatial structures. The last example is known as Turing instability [280] and is frequently used as a model for pattern formation or morphogenesis in biology [207]. The formal kinetics of reproduction involves autocatalysis, and because of its fundamental importance of in biology we shall discuss autocatalysis in section 5.1 within the chapter dealing with applications in biology.

Although chemists were intuitively familiar with mass action throughout the nineteenth century, the precise formulation of a *law of mass action* is due to two Norwegians, the mathematician and chemist Cato Maximilian Guldberg and the chemist Peter Waage [299]. For reaction (4.1f), for example, mass action rate law (κ_{ma}) yields

$$\frac{\mathrm{d}[\mathbf{A}]}{\mathrm{dt}} \;=\; \frac{\mathrm{d}[\mathbf{B}]}{\mathrm{dt}} \;=\; -\,\frac{\mathrm{d}[\mathbf{C}]}{\mathrm{dt}} \;=\; k\;[\mathbf{A}]\cdot[\mathbf{B}]\;,$$

the rate of the reaction is proportional to the particle numbers or concentrations of both reactants, [**A**] and [**B**].

Precisely, the law of mass action states that the rate of any given chemical reaction is proportional to the product of the concentrations or activities of the reactants.[4] In particular, the numbers of identical molecules that are consumed in a reaction step – called the stoichiometric coefficients[5] ν_A and

[4] Several idealized regularities hold only in the limit of vanishing concentrations, $\lim c \to 0$. The idealized laws are retained through replacing concentrations by activities, $a_X = [\mathbf{X}] = f_X c_X$. Unless stated otherwise we shall approximate activities by concentrations here and for the sake of simplicity use lower case letters to indicate the species: $f_X \approx 1$ and $[\mathbf{X}] = x$. The units used for concentrations are [mole×dm^{-3}].

[5] The stoichiometric coefficients of the reactants in the reaction $\mathbf{R}_j$ will be denoted by $\nu_{Aj}, \nu_{Bj}, \ldots$, for the products we shall use $\nu'_{Aj}, \nu'_{Bj}, \ldots$ and the elements of the stoichiometric matrix are $\mathrm{S} = \{s_{ij} = \nu'_{ij} - \nu_{ij}\}$ (see sections 4.1.3.1 and 4.7.2).

ν_B – appear as exponents of concentrations, v is the reaction rate, and k is a reaction rate parameter:

$$\begin{aligned} & \nu_A \mathbf{A} + \nu_B \mathbf{B} \xrightarrow{k} \text{products} \quad \Longrightarrow \\ \Longrightarrow \quad & \text{reaction rate} = v = -\frac{1}{\nu_A}\frac{\mathrm{d}a}{\mathrm{d}t} = -\frac{1}{\nu_B}\frac{\mathrm{d}b}{\mathrm{d}t} = k\, a^{\nu_A}\, b^{\nu_B} . \end{aligned} \tag{4.3}$$

In a reversible reaction,[6] which is an acceptable chemical reaction in both directions and can be understood as a special combination of two elementary steps compensating each other, the reverse reaction is accounted for by a minus sign:[7]

$$\begin{aligned} & \nu_A \mathbf{A} + \nu_B \mathbf{B} \underset{h}{\overset{k}{\rightleftharpoons}} \nu_C \mathbf{C} + \nu_D \mathbf{D} \quad \Longrightarrow \\ \Longrightarrow \quad & v = -\frac{1}{\nu_A}\frac{\mathrm{d}a}{\mathrm{d}t} = -\frac{1}{\nu_B}\frac{\mathrm{d}b}{\mathrm{d}t} = \frac{1}{\nu_C}\frac{\mathrm{d}c}{\mathrm{d}t} = \frac{1}{\nu_D}\frac{\mathrm{d}d}{\mathrm{d}t} = \\ & = k\, a^{\nu_A}\, b^{\nu_B} - h\, c^{\nu_C}\, d^{\nu_D} . \end{aligned} \tag{4.4}$$

The condition of zero net reaction rate yields an expression for the equilibrium parameter commonly denoted as *equilibrium constant* as in the formulation of mass action at equilibrium by Guldberg and Waage:

$$K = \frac{k}{h} = \frac{c^{\nu_C}\, d^{\nu_D}}{a^{\nu_A}\, b^{\nu_B}} . \tag{4.5}$$

Later derivations of mass action are using the chemical potentials of reactants and products as introduced by Josiah Willard Gibbs around nineteen hundred [98] (see also [99, pp. 56, 64, 65]).

Strictly speaking, the notion of elementary steps implies the application of mass action kinetics, and this means that on the level of molecules – not necessarily molecular states – no further resolution is possible. The advances in spectroscopy made it possible to distinguish between different states of molecules – the ground state and various excited states in quantum molecular physics or the minimum free energy structures and suboptimal conformations

[6] The notions *reversible* and *irreversible* for chemical reactions are used differently from thermodynamics: In chemical kinetics a reaction is *irreversible* if the occurrence of the reaction in opposite direction is not observable on realistic time scales and hence can be neglected. Strict chemical irreversibility causes an instability in thermodynamics. All chemical reactions that proceed with nonzero velocity are *irreversible* in the sense of thermodynamics as reversibility requires infinitely slow progress of processes and chemical reactions are no exception.

[7] We shall be dealing with multistep reaction networks of reversible reactions and apply a notation that allows for straightforward identification of reaction steps by choosing k_m and h_m as reaction parameters for reaction m.

in biopolymers – and then the ultimate resolution has to be pushed further down to individual states in order to be able to adequately describe processes.

Elementary step resolution and mass action kinetics often lead to complex reaction networks with a great number of variables, which are hard to analyze and which yield results that are difficult to interpret. It is sometimes useful to reduce the number of variables and to introduce a simpler *higher level kinetics*. The difference between mass action and higher level kinetics and is illustrated by means of an old and well studied example, the Michaelis-Menten reaction kinetics of enzyme catalyzed reactions in biochemistry.

4.1.2 Michaelis-Menten kinetics

Chemical kinetics became relevant for biology already at the end of the nineteenth century since biochemical processes gained a quantitative perspective. In particular, enzyme catalyzed reaction were studied and biochemical kinetics was initiated by the path-breaking work of Leonor Michaelis and Maud Menten [214]. General enzyme catalysis is modeled by three elementary steps, which at first are assumed to be reversible:

$$\mathbf{S}+\mathbf{E} \rightleftharpoons \mathbf{S}\cdot\mathbf{E} \rightleftharpoons \mathbf{E}\cdot\mathbf{P} \rightleftharpoons \mathbf{E}+\mathbf{P}\,, \tag{4.6}$$

(i) binding of the substrate **S** to the enzyme **E**, (ii) conversion of substrate into product, both being bound to the enzyme, and (iii) the release of the product **P** through dissociation of the enzyme-product complex. Then, the full mechanism of the simplest enzyme catalyzed reaction comprises six elementary steps, which in mass action kinetics (κ_{ma}) are of the form

$$\mathbf{S} + \mathbf{E} \xrightarrow{k_1} \mathbf{S}\cdot\mathbf{E}\,, \tag{4.7a}$$

$$\mathbf{S}\cdot\mathbf{E} \xrightarrow{h_1} \mathbf{S} + \mathbf{E}\,, \tag{4.7b}$$

$$\mathbf{S}\cdot\mathbf{E} \xrightarrow{k_2} \mathbf{E}\cdot\mathbf{P}\,, \tag{4.7c}$$

$$\mathbf{E}\cdot\mathbf{P} \xrightarrow{h_2} \mathbf{S}\cdot\mathbf{E}\,, \tag{4.7d}$$

$$\mathbf{E}\cdot\mathbf{P} \xrightarrow{k_3} \mathbf{P} + \mathbf{E}\,, \tag{4.7e}$$

$$\mathbf{P} + \mathbf{E} \xrightarrow{h_3} \mathbf{E}\cdot\mathbf{P}\,. \tag{4.7f}$$

For an efficient enzyme reaction it is essential that the steps (4.7d) and (4.7f) are negligibly slow. The latter reaction (4.7f), in particular at high concentrations of the product **P**, can lead to *product inhibition*. It is useful for the

catalytic efficiency if reaction (4.7b) is slow too. In section single molecule techniques we shall come back to full Michaelis-Menten kinetics when we discuss the reaction in a system containing a single enzyme molecule [247].

In the genuine Michaelis-Menten mechanism step (i) is considered as reversible binding whereas (ii) is thought to be an irreversible chemical reaction. Step (iii) follows the irreversible reaction step (ii) and hence need not be considered explicitly. Michaelis-Menten enzyme kinetics deals with four molecular species, $\mathbf{S}$, $\mathbf{E}$, $\mathbf{S}\cdot\mathbf{E}$, and $\mathbf{P}$ being substrate, enzyme,substrate-enzyme complex, enzyme-product complex and product, respectively – the enzyme-product complex, $\mathbf{E}{\cdot}\mathbf{P}$ is not considered explicitly, and the concentration of the product is interpreted best as total concentration: $p \approx p_0 = [\mathbf{P}] + [\mathbf{E}\cdot\mathbf{P}]$. Again we denote concentrations by small letters, $[\mathbf{S}] = s$, $[\mathbf{E}] = e$, $[\mathbf{P}] = p$, and for the complexes we use $[\mathbf{S}\cdot\mathbf{E}] = c_S = c$ and $[\mathbf{E}\cdot\mathbf{P}] = c_P = c$, respectively. Total concentrations will be denoted by: $e_0 = e + c + c_p$, $s_0 = s + c$ and p_0 as said above. In Michaelis-Menten kinetics (κ_{MM}) the stoichiometric equations kinetic and the equation take on the form:

$$\mathbf{S}+\mathbf{E} \rightleftharpoons \mathbf{S}\cdot\mathbf{E} \to \mathbf{E}+\mathbf{P} \quad\Longrightarrow\quad \text{reaction rate} \;=\; \frac{\mathrm{d}[\mathbf{P}]}{\mathrm{dt}} \;=\; \frac{v_{\max}\cdot s}{K_M + s}\;.$$

The parameters $v_{\max}$ and K_M denote the maximal reaction rate and the Michaelis-Menten constant, respectively. The Michaelis-Menten constant is the free substrate concentration s at the half maximal reaction rate, $v_{\max}/2$. For more than half a century after the pioneering works of Michaelis and Menten, the Michaelis-Menten constant K_M has been the most important quantitative parameter of enzymes, and it has been used, for example, to determine the purity of enzyme preparations. Biochemical kinetics became a discipline in its own right and recently led to the ambitious goal of systems biology consisting in biochemical modeling of all processes on the levels of cells and whole organisms. Beginning in the nineteen sixties new spectroscopic and kinetic techniques were developed that allowed for resolution of reaction kinetics into individual reaction steps.

In order to derive the Michaelis-Menten equation we start from the mechanism given above and assign rate parameters to individual reaction steps

$$\mathbf{S} + \mathbf{E} \underset{h_1}{\overset{k_1}{\rightleftharpoons}} \mathbf{S}\cdot\mathbf{E} \xrightarrow{k_2} \mathbf{E} + \mathbf{P}\,. \tag{4.8}$$

The differential equation for the enzyme substrate complex is of the form

$$\frac{\mathrm{d}c}{\mathrm{dt}} \;=\; k_1\, e\, s \;-\; (h_1 + k_2)\, c$$

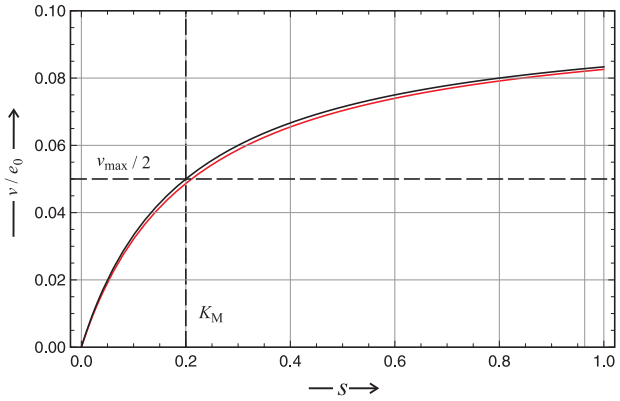

Fig. 4.1 The Michaelis-Menten mechanism for enzyme catalyzed reaction. The reaction rate $v = dp/dt = k_2(s_0 - s - p) = k_2(e_0 - e)$ is determined form a plot of v against s: v reaches a plateau value after an initial nonlinear increase, and this plateau value may be estimated from the maximum $v|_{dv/ds=0}$. The maximal rate is approximated by $v_{\max} = k_2(e_0 - e) \approx e_0$ because all enzyme **E** is converted into complex **S** · **E** at high substrate concentration, $s \gg e_0$. Choice of parameters: $k_1 = 1$, $h_1 = k_2 = 0.1$, $e_0 = 0.01$, and hence $K_M = 0.2$. The black curve $v(s)$ is compared with the plot of v against s_0 (red).

and we obtain for the steady state:[8]

$$\frac{\mathrm{d}c}{\mathrm{dt}} = 0 \quad \Longrightarrow \quad (h_1 + k_2)\,\hat{c} = k_1\,\hat{e}\,\hat{s}$$

Now we define the Michaelis-Menten constant and introduce $e_0 = e + c$ as the total enzyme concentration in order to eliminate the free enzyme concentration variable e:

$$\frac{h_1 + k_2}{k_1} = K_M = \frac{(e_0 - \hat{c})\,\hat{s}}{\hat{c}} \quad \Longrightarrow \quad \hat{c} = \frac{e_0 \cdot \hat{s}}{K_M + \hat{s}}\ .$$

The rate of product formation is obtained through multiplication by the rate constant of the irreversible reaction

$$v = \frac{\mathrm{d}p}{\mathrm{dt}} = k_2\,\frac{e_0 \cdot s}{K_M + s} = \frac{v_{max} \cdot s}{K_M + s} \quad \text{with} \quad v_{\max} = k_2\,e_0\ , \tag{4.9}$$

and the result is the equation reported above. □

[8] To indicate a true equilibrium state we would use the symbol $\bar{\cdot}$, e.g., $\bar{c}$. Since the assumption for the derivation of the Michaelis-Menten equation will be that the enzyme catalyzed reaction is sufficiently slow in order to keep the system in an approximate equilibrium state we are using $\hat{c}$, $\hat{s}$, etc., instead.

Often it is quite demanding to measure the free substrate concentration s and in the initial phase of the reaction or for $s_0 \gg e_0$, [**S**] can be approximated by the total substrate concentration $s \approx s_0$. An exact calculation is possible if the rate of reaction is zero and substrate binding is at equilibrium, $k_2 \ll h_1$:

$$\bar{s} \;=\; \frac{1}{2}\left((s_0 - e_0 - H) + (s_0 - e_0 + H)\sqrt{1 \,+\, \frac{4\,e_0 H}{(s_0 - e_0 + H)^2}}\right) \;, \quad (4.10)$$

with $H = h_1/k_1$ being the dissociation constant of the enzyme-substrate complex, $\mathbf{S}\cdot\mathbf{E}$. Equation (4.10) has a very simple solution under two conditions: (i) substrate **S** in large excess over enzyme **E**, $e_0 \ll s_0$ (and $e_0 \ll H$), and (ii) fast dissociation of the complex $h_1 \gg k_1$ or $\lim H \to \infty$

$$\bar{s} \;\approx\; s_0 \,-\, e_0 \;\approx\; s_0 \;.$$

Without the equilibrium approximation Michaelis-Menten enzyme kinetics is described by two ODEs. The total concentrations of substrate and enzyme are according to stoichiometry

$$s(0) \;=\; s_0 \;=\; s + c + p\,,\; e(0) \;=\; e_0 \;=\; e + c\,,\;\; \text{and}\;\; c \;=\; e_0 - e\,, \quad (4.11)$$

where we have assumed that initially there was not product in the reaction mixture, $p(0) = p_0 = 0$:

$$\begin{aligned} \frac{\mathrm{d}p}{\mathrm{d}t} \;&=\; k_2\,(s_0 \,-\, s \,-\, p) \quad \text{and} \\ \frac{\mathrm{d}s}{\mathrm{d}t} \;&=\; -\,k_1\,s\cdot(e_0 \,-\, s_0 \,+\, s \,+\, p) \,+\, h_1\,(s_0 \,-\, s \,-\, p)\;. \end{aligned} \quad (4.12)$$

Results from computer integration of equation (4.12) are shown in figure 4.1. The Michaelis-Menten constant is obtained straightforwardly from the substrate concentration s at half-maximal reaction rate $v_{\max}/2$. It is also worth noticing how small the differences between s and s_0 are in this particular case.

The most important results of the Michaelis-Menten analysis of enzyme catalyzed reactions are: (i) A small value of the Michaelis-Menten constant K_M means that the enzyme reaches its maximal turnover already at small substrate concentrations, (ii) a large value of K_M implies the opposite – the maximal reaction rate is achieved only at high substrate concentrations, and (iii) the Michaelis-Menten constant K_M is proportional to the sum $h_1 + k_2$ and therefore large K_M does not necessarily imply a high catalytic rate parameter $k_2 = k_{\mathrm{cat}}$, it can also indicate weak binding of the substrate.

4.1.3 Reaction network theory

So far we have considered only single step processes of chemical reactions.[9] Almost all interesting chemical systems, however, consist of networks of reactions that are characterized by a variety of interacting molecular species, and this leads to dynamical systems of more than one variable, often many variables, for which analytical solutions are available very rarely only.

In the second half of last century, when chemists and physicists began to consider kinetic differential equations as dynamical systems and started to apply qualitative analysis, new questions in addition to forward and inverse problems became relevant. The new questions are concerned with general properties of reaction networks, for example, to prove (i) whether or not a network can sustain multiple steady states in the positive orthant of concentration space, (ii) whether or not undamped oscillations resulting from a stable limit cycle are possible or (iii) whether of not a specific reaction network can display deterministic chaos. A general recent technique that can be applied for finding answers to these questions consists in the inversion of qualitative analysis [188, 187]: Inverse bifurcation analysis aims at an exploration of the domains in parameter space that give rise to certain forms of complex dynamics.

A formal deterministic theory of chemical reaction networks has been developed already in the nineteen seventieth by Fritz Horn, Roy Jackson, and Martin Feinberg [69, 134] in order to complement conventional chemical kinetics by tools that allow for the derivation of general results for entire classes of reaction networks. The theoretical approach became really popular only recently when chemical reaction kinetics has been applied in systems biology and it was realized that stochastic modeling of extended chemical reaction networks is required for any deeper understanding of regulation and control of cellular dynamics and cellular metabolism [37, 114]. Before we consider modeling of stochastic chemical reaction networks (SCRNs) in section 4.7 we present a brief introduction to the Feinberg-Horn-Jackson-theory, which allows for straightforward answers to other wise difficult to predict properties of chemical reaction networks, for example, the nonexistence of multiple steady states or the absence of oscillating concentrations in reaction networks. The theory is not aiming at deducing the properties of networks for given sets of rate parameters but derives tools for studying features of whole classes of networks irrespectively of the particular choice of parameters.

[9] Two trivial exceptions were the influx and outflux of a compound **A** in the flow reactor and the reversible reaction **A**⇌**B**. In both cases, however, we were dealing with a single stochastic variable counting the numbers of molecules **A**.

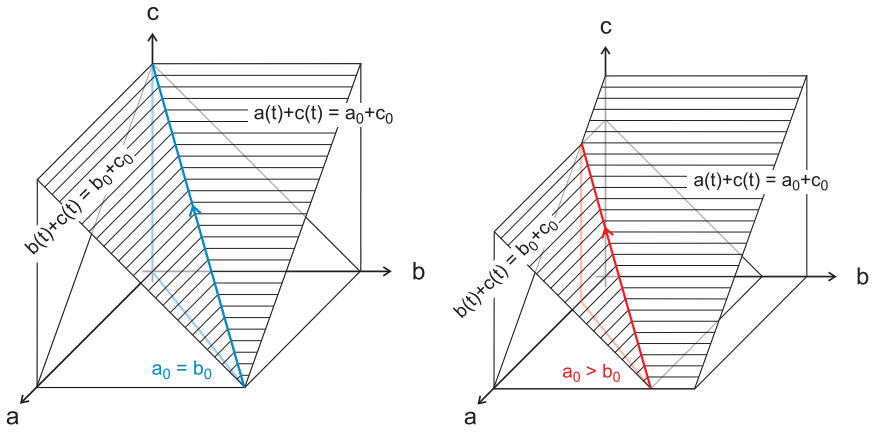

Fig. 4.2 Stoichiometric subspace and compatibility class. The figure on the r.h.s. sketches the stoichiometric subspace, $\mathbf{S} = \text{span}_j\{\mathbf{s}_j\}$, of the irreversible reaction $\mathbf{A}+\mathbf{B} \to \mathbf{C}$. The concentration space $\mathbf{X} = \{a, b, c\} \in \mathbb{R}^3$ is three dimensional, two independent conservation relations, $a(t) = a_0+c_0-c(t)$ and $b(t) = b_0+c_0-c(t)$, introduce linear dependencies and hence the stoichiometric subspace is one-dimensional. The stoichiometric compatibility class is formed by adding a constant vector $\mathbf{c} \in \mathbb{R}^M$, for example the initial conditions $\mathbf{x}_0 = (a_0, b_0, c_0)$ to the stoichiometric subspace: $\mathbf{x}_0 + S$. The two initial conditions applied here are: $\mathbf{x}_0 = (a_0, b_0 = a_0, 0)$ and $\mathbf{x}_0 = (a_0, b_0 < a_0, 0)$.

4.1.3.1 Formal stoichiometry

For the forthcoming discussions it is of advantage to formalize the concept of stoichiometry by means of linear algebra. For this goal we assume a set of M chemical species $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M\}$, which are interconverted by K chemical reactions, $\mathbf{R}_1, \mathbf{R}_2 \ldots, \mathbf{R}_K$. It is useful to define a row vector of species: $\overrightarrow{\mathbf{S}} = (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M)$. Each individual chemical reaction $\mathbf{R}_j$

$$\sum_{i=1}^{M} \nu_{ij}\, \mathbf{S}_i \;\to\; \sum_{i=1}^{M} \nu'_{ij}\, \mathbf{S}_i \tag{4.13}$$

is characterized by two column vectors containing the stoichiometric coefficients $\boldsymbol{\nu}_j = \left(\nu_{1j}, \nu_{2j}, \ldots, \nu_{Mj}\right)^{\mathrm{t}}$ and $\boldsymbol{\nu}'_j = \left(\nu'_{1j}, \nu'_{2j}, \ldots, \nu'_{Mj}\right)^{\mathrm{t}}$ of reactants and products, respectively.[10] Now we can write the stoichiometric equation of reaction $\mathbf{R}_j$ (4.13) in compact form

[10] We introduce here temporarily *primed* stoichiometric coefficients for reaction products.

$$\mathbf{R}_j : \quad \overrightarrow{\mathbf{S}} \cdot \boldsymbol{\nu}_j \;\rightarrow\; \overrightarrow{\mathbf{S}} \cdot \boldsymbol{\nu}'_j \;\text{ and }\; \overrightarrow{\mathbf{S}} \cdot (\boldsymbol{\nu}'_j - \boldsymbol{\nu}_j) \;=\; \overrightarrow{\mathbf{S}} \cdot \mathbf{s}_j \;. \qquad (4.13')$$

The linear combination of species as defined by the stoichiometry of a chemical reaction is called a *reaction complex* (section 4.1.3.2):[11] $\mathbf{C}_j = \overrightarrow{\mathbf{S}} \cdot \boldsymbol{\nu}_j$ or $\mathbf{C}_{j'} = \overrightarrow{\mathbf{S}} \cdot \boldsymbol{\nu}'_j$ being the reactant complex and the product complex of reaction $\mathbf{R}_j$, respectively. The stoichiometric coefficients of all N complexes appearing in a chemical reaction network together form the $M \times N$ *matrix of complexes*

$$\mathrm{C} \;=\; \left(\overrightarrow{\mathbf{S}} \cdot \boldsymbol{\nu}_1 \;\; \overrightarrow{\mathbf{S}} \cdot \boldsymbol{\nu}_2 \;\; \ldots \;\; \overrightarrow{\mathbf{S}} \cdot \boldsymbol{\nu}_N \right) \;.$$

As indicated already in equation (4.13') we combine the stoichiometric vectors belonging to the reactants and the products of the same reaction whereby we count reactant coefficients as being negative in order to provide a measure of the change introduced by the reaction. The stoichiometry of the entire reaction network is properly encapsulated in the $M \times K$ *stoichiometric matrix*:

$$\mathrm{S} \;=\; (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K) \;=\; \{s_{ij}; i = 1, \ldots, M; j = 1, \ldots, K\} \qquad (4.14)$$

The stoichiometric matrix allows for a compact written form of the kinetic differential equations and their solutions

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{dt}} \;=\; \mathrm{S} \cdot \mathbf{v} \;\text{ and }\; \mathbf{x}(t) - \mathbf{x}_0 \;=\; \sum_{j=1}^{K} \left(\int_0^t v_j\big(\mathbf{x}(\tau)\big)\, \mathrm{d}\tau \right) \mathbf{s}_j \;, \qquad (4.15)$$

where $\mathbf{v} = \Big(v_1\big(\mathbf{x}(t)\big), v_2\big(\mathbf{x}(t)\big), \ldots, v_K\big(\mathbf{x}(t)\big) \Big)^{\mathrm{t}}$ is the vector of reaction rates, here mass action rates κ_{ma} according to equation (4.4), the variables are concentrations described by a vector $\mathbf{x}(t) = \big(x_1(t), x_2(t), \ldots, x_M(t)\big) \in \mathbb{R}^M$, and $\mathbf{x}_0 = \big(x_1(0), x_2(0), \ldots, x_M(0)\big)$ are the initial conditions.

A number of restrictions apply to chemical kinetics: (i) concentrations are positive real numbers, $x_j(t) \in \mathbb{R}_{>0}\, \forall\, j = 1, \ldots, M$,[12] (ii) the solutions have to fulfil the stoichiometric relations for all reactions $\mathbf{R}_j$ $(j = 1, \ldots, K)$ and this is encapsulated in the restriction to *stoichiometric compatibility classes*. We define the *stoichiometric subspace* of a reaction system by

$$\mathbf{S} \;=\; \mathrm{span}\{\mathbf{s}_j \,|\, j = 1, \ldots, K\} \subset \mathbb{R}^M \;\text{ and }\; R \;:=\; \dim(\mathbf{S}) \;. \qquad (4.16)$$

---

[11] The notion of reaction complex needs affirmation, since it is different from an association complex like the enzyme-substrate complex in the Michaelis-Menten reaction: A reaction complex is a combination of molecules in the correct stoichiometric ratio as it appears at the reactant side or at the product side of a stoichiometric equation.

[12] In chemistry concentrations of molecular species are commonly required to be positive quantities, whereas extinction corresponding to concentration zero is often an important issue than *positive* has to be replaced by *nonnegative*, $\mathbb{R}_{>0} \rightarrow \mathbb{R}_{\geq 0}$.

The stoichiometric compatibility class contains the stoichiometric subspace shifted by some constant vector, $\mathbf{c}+\mathbf{S}$, and we restrict the variables to positive values of the concentrations is of the form

$$\mathcal{D} \;=\; (\mathbf{c} \,+\, \mathrm{span}\{\mathbf{s}_j \,|\, j = 1, \ldots, K\}) \cap \mathbb{R}^M_{>0} \;=\; (\mathbf{c} \,+\, \mathbf{S}) \cap \mathbb{R}^M_{>0}\,. \qquad (4.17)$$

Figure 4.2 shows a simple example of a one-dimensional compatibility call embedded in a three-dimensional concentration space. Since the linear span is built from all reaction vectors $\mathbf{s}_j$, linear dependencies will occur in most cases. The number of independent vectors in $\mathrm{span}_j(\mathbf{s}_j)$, the dimension or the rank R of the stoichiometric subspace, is the number of independent concentration variables or the number of degrees of freedom in the kinetic reaction system. For small systems, like the examples in section 4.1.3.4, it is useful and illustrative to reduce the degrees of freedom by means of easy to find conservation relations, but for larger system with several hundred variables and more, a stable numerical procedure is commonly to be preferred: The rank R of the stochastic matrix represents the number of degrees of freedom of the kinetic system and is computed straightforwardly by routine software.

4.1.3.2 Chemical reaction networks

The notion of a chemical reaction network stands in the center of the reaction network theory. Each network consists of three commonly finite sets of objects

(i) a set of M molecular species, $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M\}$, which interact through a finite number of chemical reactions,
(ii) a set of N complexes, $\mathcal{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_N\}$, which are linear combinations of species, $\mathbf{C}_j = \sum_{i=1}^{M} \nu_{ij}\mathbf{S}_i$ with $\nu_{ij} \in \mathbb{N}_{>0}$, and
(iii) a set of K molecular reactions, $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_K\}$, with $\mathcal{R} \subset \mathcal{C} \times \mathcal{C}$ in the sense of individual elements being directed combinations of two complexes, $(\mathbf{C}_\mathrm{R}, \mathbf{C}_\mathrm{P}) \in \mathcal{R}$ is written as $\mathbf{C}_\mathrm{R} \to \mathbf{C}_\mathrm{P}$ where R and P stand for *reactants* or *products*, respectively.

Restrictions are imposed on the sets $\mathcal{S}$ and $\mathcal{C}$: Each element of $\mathcal{S}$ has to be found in at least one reaction complex or, in other words, there are no superfluous species. Condition (iii) is supplemented by two exclusions: No complex may react into itself, $\mathbf{C}_\mathrm{R} \neq \mathbf{C}_\mathrm{P}$, and isolated complexes are not allowed, in the sense that every element of $\mathcal{C}$ must be the reactant or the product complex of some reaction. It is worth reminding that a reversible reaction (see e.g. section 4.3.2.2) is represented by two reactions: $\mathbf{C}_\mathrm{R} \to \mathbf{C}_\mathrm{P}$ and $\mathbf{C}_\mathrm{P} \to \mathbf{C}_\mathrm{R}$.

The mentioned restriction can be cast in a somewhat different form that is presented here for making the definitions clearer. Complexes and species are related through

(a) $\mathcal{C} \subset \mathbb{R}^{\mathcal{S}}$ where $\mathbb{R}^{\mathcal{S}}$ stands for a vector space spanned by unit vectors representing individual species. Commonly, the coefficients in the linear combinations of species called complexes are natural numbers, $s_{ij} \in \mathbb{N} > 0$,

(b) $\bigcup_{\text{supp, } \mathbf{C}_j \in \mathcal{C}} \mathbf{C}_j = \mathcal{S}$ the union of the species in all complexes is the species set and no species can exist in $\mathcal{S}$, which does not appear in at least one complex.[13]

Species $\mathbf{S}_i$ and reactions $\mathbf{R}_j$ are directly related by the stoichiometric matrix $\mathrm{S} = \{s_{ij}\}$. The columns of S refer to reactions and the rows to species. We shall make use of S also in section 4.7 for the implementation of a simulation tool for chemical master equations.

The fourth components of a reaction system is the kinetics of the reactions, $\mathcal{K}$. Mass action kinetics (κ_{ma}) has been discussed in section 4.1.1 and Michaelis-Menten kinetics (κ_{MM}) as an example of higher-level kinetics in section 4.1.2. In the majority of the examples discussed here mass action we be applied. We repeat the basic equation (4.3) for reaction $\mathbf{R}_j$:

$$s_{1j}\,\mathbf{S}_1 + s_{2j}\,\mathbf{S}_2 + \ldots \implies v_j = k_j\,[\mathbf{S}_1]^{s_{1j}} \cdot [\mathbf{S}_2]^{s_{2j}} \cdots . \tag{4.18}$$

In mass action kinetics κ_{ma} we need one reaction parameter k_j for every elementary step and hence the number of rate parameters is equal to R, the number of reactions. Eventually, a reaction system consists of the four components $\{\mathcal{S}, \mathcal{C}, \mathcal{R}, \mathcal{K}\}$ and the evolution in time of the reaction system can be encapsulated in an ODE or in a master equation in case of a stochastic description.

4.1.3.3 Reaction graphs

Some general properties of reaction networks can be predicted directly from the reaction graph (figure 4.3), which is a directed graph with *complexes*, $\mathbf{C}_k \in \mathcal{C}\,, (k = 1, \ldots, N)$, being the nodes and three symbols indicating forward ($\rightarrow$), backward ($\leftarrow$) and reversible reaction ($\rightleftharpoons$) for the edges. A reaction graph may have several components called *linkage classes*. Two properties are important for reaction graphs: (i) A complex appears only once as a node of the graph and (ii) different linkage classes to not share complexes.

The network in figure 4.3 has two linkage classes since the two clusters don't share a single complex. The information on the number of complexes and the number of linkage classes is contained in the reaction graph. The same is true for the classification of a network as reversible, weak reversible or not reversible. A (strongly) reversible network contains exclusively reversible reactions in the strict thermodynamic sense. Weak reversibility relaxes the condition of (*strong*) *reversibility*: A network is weakly reversible when for

[13] The notion 'supp' stands for the support of a vector which is the subset of unit vectors for which the vector has nonzero coefficients.

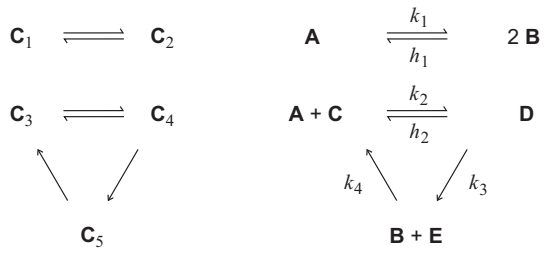

Fig. 4.3 The graph corresponding to the chemical reaction network (4.21a)**.** Each node of the graph (l.h.s.) corresponds to a reaction complex, three different symbols characterize the directed edges: $\rightarrow$, $\leftarrow$, and $\rightleftharpoons$ for *forward*, *backward*, and *reversible reaction*, respectively. This graph consists of $L = 2$ linkage classes. On the r.h.s. we show the Feinberg mechanism, which is an implementation of the reaction graph on the l.h.s. The mechanism differs from the graph by additional information: (i) the molecular realization of the reaction complexes and the rate parameters.

every pair of complexes there exist a directed arc leading from one complex to the other. The network in figure 4.3 fulfils the condition of weak reversibility, it would be (strongly) reversible if it would be complemented by the arrows $\mathbf{C}_3 \rightarrow \mathbf{C}_5$ and $\mathbf{C}_5 \rightarrow \mathbf{C}_4$. For the determination of linkage classes only the existence or absence of arrows between complexes matters. Clearly, the direction of arrows is is required too for the classification of reversibility.

A reaction graph differs from a reaction mechanism in three aspects: The reaction complexes are not defined in terms of chemical compounds and therefore the reaction graph does not consider stoichiometry, it does not specify the algebraic relations of reaction rates in the form of mass action, Michaelis-Menten or other kinetic functions, and it does not contain weighting factors of edges in the sense of rate parameters. The reaction graph represents nothing more than the topology of a reaction network and general properties derived from the graph are valid for a large number of concrete cases irrespective of stoichiometries, kinetic functions, and rate constants.

4.1.3.4 Examples

We illustrate chemical reaction network theory by means of examples.

The irreversible addition reaction: $\mathbf{A} + \mathbf{B} \rightarrow \mathbf{C}$. The first example is the irreversible addition reaction (4.1f):

$$\mathbf{A} + \mathbf{B} \xrightarrow{k} \mathbf{C} \,. \tag{4.19a}$$

For the three sets of the chemical reaction network we have

$$\mathcal{S} \;=\; \{\mathbf{A},\mathbf{B},\mathbf{C}\}\;, \tag{4.19c}$$

$$\mathcal{C} \;=\; \{\mathbf{C}_1 = \mathbf{A}+\mathbf{B}, \mathbf{C}_2 = \mathbf{C}\}\;, \quad \text{and} \tag{4.19d}$$

$$\mathcal{R} \;=\; \{\mathbf{R}_1 = \mathbf{C}_1 \rightarrow \mathbf{C}_2\}\;. \tag{4.19e}$$

The stoichiometric matrix S is of dimension 3×1:

$$\mathrm{S} \;=\; \begin{pmatrix} -1 \\ -1 \\ +1 \end{pmatrix}\;. \tag{4.19f}$$

In deterministic mass action kinetics, κ_{ma}, the variables are the concentrations of the molecular species, $[\mathbf{A}] = a(t)$, $[\mathbf{B}] = b(t)$, and $[\mathbf{C}] = c(t)$. In order to solve the kinetic differential equation we require a rate parameters k and three initial conditions $a(0) = a_0$, $b(0) = b_0$, and $c(0) = c_0$. The three variables are stoichiometrically related by two conservation relations derived from equation (4.19a), which can be used to eliminate two variables, $b(t)$ and $c(t)$ for example, yielding the remaining single degree of freedom as $\frac{\mathrm{d}a}{\mathrm{dt}} = \frac{\mathrm{d}b}{\mathrm{dt}} = -\frac{\mathrm{d}c}{\mathrm{dt}}$ corresponding to $R = 1$ (see figure 4.2):

$$\begin{aligned} a(t) \,+\, c(t) \;&=\; a_0 \,+\, c_0 \;=\; \vartheta_0^{(ac)}\;, \\ b(t) \,+\, c(t) \;&=\; b_0 \,+\, c_0 \;=\; \vartheta_0^{(bc)}\;, \quad \text{and} \\ b(t) \,-\, a(t) \;&=\; b_0 \,-\, a_0 \;=\; \vartheta_0^{(b)}\;. \end{aligned}$$

One out of these three conditions is dependent, since the second line minus the first line yields the third line. Eventually one finds:

$$\frac{\mathrm{d}a}{\mathrm{dt}} \;=\; -\,k\;a\,b \;=\; -\,k\,a\,(\vartheta_0^{(b)} - a)\;. \tag{4.19g}$$

The ODE is solved by standard techniques and we obtain the solutions

$$\begin{aligned} a(t) \;&=\; \frac{a_0\,\vartheta_0^{(b)}\,\exp(-\vartheta_0^{(b)}\,kt)}{\vartheta_0 \,+\, a_0\,\big(1-\exp(-\vartheta_0^{(b)}\,kt)\big)} \quad \text{for} \;\; \vartheta_0^{(b)} > 0\,,\; b_0 > a_0\,, \\ a(t) \;&=\; \frac{a_0\,|\vartheta_0^{(b)}|}{a_0 \,-\, (a_0 - |\vartheta_0^{(b)}|)\,\big(1-\exp(-|\vartheta_0^{(b)}|\,kt)\big)} \\ &\qquad \text{for} \;\; \vartheta_0^{(b)} < 0\,,\; b_0 < a_0\,, \quad \text{and} \\ a(t) \;&=\; \frac{a_0}{1 \,+\, a_0\,kt} \quad \text{for} \;\; \vartheta_0^{(b)} = 0\,,\; b_0 = a_0\;. \end{aligned} \tag{4.19h}$$

by direct integration. The three cases differ in the long-time behavior: $\lim_{t\to\infty} a(t) = 0$ for $\vartheta_0^{(b)} \geq 0$, $b_0 > a_0$ and $\lim_{t\to\infty} a(t) = b_0 - a_0$ for $\vartheta_0^{(b)} < 0$, $b_0 > a_0$.

The reversible bimolecular conversion reaction: $\mathbf{A} + \mathbf{B} \to \mathbf{C} + \mathbf{D}$. The second case simply consists of a reversible bimolecular conversion reaction that is decomposed into two elementary reactions of type (4.1i):

$$\mathbf{A} \ + \ \mathbf{B} \quad \xrightarrow{\ k\ } \quad \mathbf{C} \ + \ \mathbf{D} \quad \text{and} \tag{4.20a}$$

$$\mathbf{C} \ + \ \mathbf{D} \quad \xrightarrow{\ h\ } \quad \mathbf{A} \ + \ \mathbf{B} \ . \tag{4.20b}$$

For the three sets of the chemical reaction network we have

$$\mathcal{S} \ = \ \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\} \ , \tag{4.20c}$$

$$\mathcal{C} \ = \ \{\mathbf{C}_1 = \mathbf{A} + \mathbf{B}, \mathbf{C}_2 = \mathbf{C} + \mathbf{D}\} \ , \quad \text{and} \tag{4.20d}$$

$$\mathcal{R} \ = \ \{\mathbf{R}_1 = \mathbf{C}_1 \to \mathbf{C}_2, \mathbf{R}_2 = \mathbf{C}_2 \to \mathbf{C}_1\} \ . \tag{4.20e}$$

The stoichiometric matrix S is of dimension 4×2:

$$\mathrm{S} \ = \ \begin{pmatrix} -1 & +1 \\ -1 & +1 \\ +1 & -1 \\ +1 & -1 \end{pmatrix} \ . \tag{4.20f}$$

In deterministic mass action kinetics, κ_{ma}, the variables are the concentrations of the molecular species, $[\mathbf{A}] = a(t)$, $[\mathbf{B}] = b(t)$, $[\mathbf{C}] = c(t)$, and $[\mathbf{D}] = d(t)$. In order to solve the kinetic differential equation we require two rate parameters, k and h, and four initial conditions: $a(0) = a_0$, $b(0) = b_0$, $c(0) = c_0$, and $d(0) = d_0$. The four variables are stoichiometrically related by three conservation relations in (4.20a) and (4.20b)

$$\begin{aligned} a(t) \ + \ b(t) \ + \ c(t) \ + \ d(t) \ &= \ a_0 \ + \ b_0 \ + \ c_0 \ + \ d_0 \ , \\ a(t) \ - \ b(t) \ &= \ a_0 \ - \ b_0 \ , \quad \text{and} \\ c(t) \ - \ d(t) \ &= \ c_0 \ - \ d_0 \ , \end{aligned}$$

and only one degree of freedom – corresponding to the rank $R = 1$ of the stoichiometric matrix – remains: $\mathrm{d}a/\,\mathrm{dt} = \mathrm{d}b/\,\mathrm{dt} = -\mathrm{d}c/\,\mathrm{dt} = -\mathrm{d}d/\,\mathrm{dt}$. Accordingly, we can substitute $b(t) = b_0 - a_0 + a(t)$, $c(t) = c_0 + a_0 - a(t)$, and $d(t) = d_0 + a_0 - a(t)$ and the ODE for the last remaining variable $a(t)$ takes on the form:

$$\begin{aligned} \frac{\mathrm{d}a}{\mathrm{dt}} \ &= \ -\,k\,a\,b \ + \ h\,c\,d \ = \ -\,k\,a\,(\vartheta_0^{(b)} + a) \ + \ h\,(\vartheta_0^{(c)} - a)(\vartheta_0^{(d)} - a) \ = \\ &= \ (h-k)\,a^2 \ - \ (k\vartheta_0^{(b)} + h\vartheta_0^{(c)} + h\vartheta_0^{(d)})\,a \ + \ h\,\vartheta_0^{(c)}\vartheta_0^{(d)} \ , \end{aligned} \tag{4.20g}$$

where the initial conditions are contained in the quantities $\vartheta_0^{(b)} = b_0 - a_0$, $\vartheta_0^{(c)} = c_0 + a_0$, and $\vartheta_0^{(d)} = d_0 + a_0$.

Equation (4.20g) can be integrated by standard methods to yield an implicit solution of the form $t = f(a)$ but the expression is so clumsy that we dispense here from listing it. The analytical solution for the irreversible forward reaction are identical with the solutions of the addition reaction (4.19h) treated in the previous example, since the kinetic ODEs of an irreversible reaction do not depend on the concentrations on the product side. Clearly, the expressions are also valid for the irreversible backward reaction by replacing $a \leftrightarrow c$, $b \leftrightarrow d$, and $k \leftrightarrow h$.

The Feinberg mechanism shown in figure 4.3. Our third example is taken directly from Martin Feinberg [69, 71] and deals with six elementary reactions involving five chemical species in the following mechanism:

$$\begin{aligned}
\mathbf{A} &\xrightarrow{k_1} 2\,\mathbf{B}\,, \\
2\,\mathbf{B} &\xrightarrow{h_1} \mathbf{A}\,, \\
\mathbf{A} + \mathbf{C} &\xrightarrow{k_2} \mathbf{D}\,, \\
\mathbf{D} &\xrightarrow{h_2} \mathbf{A} + \mathbf{C}\,, \\
\mathbf{D} &\xrightarrow{k_3} \mathbf{B} + \mathbf{E}, \quad \text{and} \\
\mathbf{B} + \mathbf{E} &\xrightarrow{k_4} \mathbf{A} + \mathbf{C}\,.
\end{aligned} \tag{4.21a}$$

The three sets defining the chemical reaction network are:

$$\mathcal{S} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}\}\,, \tag{4.21c}$$

$$\mathcal{C} = \{\mathbf{C}_1 = \mathbf{A}, \mathbf{C}_2 = 2\mathbf{B}, \mathbf{C}_3 = \mathbf{A} + \mathbf{C}, \mathbf{C}_4 = \mathbf{D}, \mathbf{C}_5 = \mathbf{B} + \mathbf{E}\}\,, \quad \text{and} \tag{4.21d}$$

$$\begin{aligned}
\mathcal{R} = \{&\mathbf{R}_1 = \mathbf{C}_1 \to \mathbf{C}_2, \mathbf{R}_2 = \mathbf{C}_2 \to \mathbf{C}_1, \mathbf{R}_3 = \mathbf{C}_3 \to \mathbf{C}_4, \\
&\mathbf{R}_4 = \mathbf{C}_4 \to \mathbf{C}_3, \mathbf{R}_5 = \mathbf{C}_4 \to \mathbf{C}_5, \mathbf{R}_6 = \mathbf{C}_5 \to \mathbf{C}_3\}\,.
\end{aligned} \tag{4.21e}$$

The stoichiometric matrix S for the mechanism (4.21a) is readily obtained:

$$\mathrm{S} = \begin{pmatrix}
-1 & +1 & -1 & +1 & +1 & 0 \\
+2 & -2 & 0 & 0 & -1 & +1 \\
0 & 0 & -1 & +1 & +1 & 0 \\
0 & 0 & +1 & -1 & 0 & -1 \\
0 & 0 & 0 & 0 & -1 & +1
\end{pmatrix}, \tag{4.21f}$$

it has the dimension 5×6 and its rank is $R = 3$. The reaction graph corresponding to this mechanism is found in figure 4.3. The comparison of both graphs is a nice illustration of one already mentioned property of reaction graphs: The graph visualizes only the interconversions between reaction complexes and contains no information about the molecular realization of the kinetic reaction network, whereas the graphical representation of the reaction network in contains the full information except the specific initial conditions. Analytical solutions for the reaction network (4.21a) are not available but numerical integration for given initial conditions is easily achieved. Some qualitative properties will be derived in the next two sections (4.1.3.5) and (4.1.3.6).

4.1.3.5 Definition of deficiency

First, the basic definitions reaction of chemical reaction network theory are repeated and we point out how the relevant properties are obtained:

(i) a *linkage class* is a subset of complexes that are linked be reactions and the number of linkage classes is denoted by L,
(ii) a reaction network is *weakly reversible* if and only if a directed arc leads from every complex to every complex of the network,
(iii) the *reaction vectors* combine reactants and products in the stoichiometric way, $\vec{\mathbf{R}} = -\mathcal{C}_\mathrm{R} + \mathcal{C}_\mathrm{P}$, and
(iv) the *rank* of a reaction network, R is the largest linearly independent set that can be found among its reaction vectors.

The linkage classes of a reaction network are obtained straightforwardly: Each complex is displayed exactly once in the sketch of the network, the complexes are joined by introducing the reaction arrows into the sketch, and linkage classes comprise all complexes joined together. The network in figure 4.3, for example, has $L = 2$ linkage classes.

Strong and weak reversibility are directly seen in the reaction graph: In a strongly reversible network all reactions $\mathbf{R} \in \mathcal{R}$ are reversible,

$$(\mathbf{C}_j \to \mathbf{C}_k \in \mathcal{R}) \implies (\mathbf{C}_k \to \mathbf{C}_j \in \mathcal{R}) \; \forall \; (\mathbf{C}_j, \mathbf{C}_k) \in \mathcal{C} \; . \tag{4.22}$$

Weak reversibility relaxes the condition for strong reversibility in the sense that it is sufficient to be able to reach every species from every species by a sequence of reactions. The network in figure 4.3 is weakly reversible.

The rank of a chemical reaction network is defined as

$$R \; := \; \mathrm{rank}\{\mathbf{C}_\mathrm{P} - \mathbf{C}_\mathrm{R} \in \mathbb{R}^\mathcal{S} : \mathbf{C}_\mathrm{R} \to \mathbf{C}_\mathrm{P} \in \mathcal{R}\} \; . \tag{4.23}$$

We illustrate by means of a simple example: The six reaction vectors of the network (4.21a),

$$\{2\mathbf{B}-\mathbf{A}, \mathbf{A}-2\mathbf{B}, \mathbf{D}-(\mathbf{A}+\mathbf{C}), (\mathbf{A}+\mathbf{C})-\mathbf{D}, (\mathbf{B}+\mathbf{E})-\mathbf{D}, (\mathbf{A}+\mathbf{C})-(\mathbf{B}+\mathbf{E})\},$$

can be contracted to the linearly independent subset of dimension three

$$\{2\mathbf{B}-\mathbf{A}, (\mathbf{A}+\mathbf{C})-\mathbf{D}, (\mathbf{B}+\mathbf{E})-\mathbf{D}\} \ .$$

Although the network (4.21a) consists of six reactions, only three of them are linearly independent and accordingly it has rank $R = 3$. It is straightforward to see that every reversible reaction consists of two reactions but only one of them can be linearly independent. The determination of the rank R in small systems is properly done by means of the conservation relations but for larger systems a computation of the rank of the stochastic variables is usually much faster.

The most important quantity of reaction network theory is the *deficiency* of a reaction system, which is defined in the following equation:

$$\text{Deficiency} \quad \delta \ := \ N \ - \ L \ - \ R \ , \tag{4.24}$$

with N being the number of complexes, L the number of linkage classes, and R the number of degrees of freedom or the rank of the reaction kinetics.

The deficiency of a chemical reaction network is a nonnegative quantity [70] and it determines essential features of the reaction system like the existence of unique equilibria and stationary states.

4.1.3.6 The deficiency zero theorem

The deficiency zero theorem holds for all chemical reaction networks $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$ of deficiency zero and makes three statements [70]:

(i) If the network is not weakly reversible then the ODEs for the reaction system $\{\mathcal{S}, \mathcal{C}, \mathcal{R}, \mathcal{K}\}$ with any arbitrary kinetics $\mathcal{K}$ cannot admit a positive equilibrium, i.e., a stationary point in $\mathbb{R}^M_+$,
(ii) if the network is not weakly reversible then the ODEs for the reaction system $\{\mathcal{S}, \mathcal{C}, \mathcal{R}, \mathcal{K}\}$ with any arbitrary kinetics $\mathcal{K}$ cannot admit a cyclic trajectory containing a positive composition, i.e., a point in $\mathbb{R}^M_+$, and
(iii) if the network is weakly reversible (or reversible) then, for any mass action kinetics $\kappa \in \mathbb{R}^R_+$, the ODEs for the mass action system $\{\mathcal{S}, \mathcal{C}, \mathcal{R}, \kappa\}$ have the following properties: Within each positive stoichiometric compatibility class there exists exactly one equilibrium, this equilibrium is asymptotically stable, and there cannot exist a nontrivial cyclic trajectory in $\mathbb{R}^M_+$.

The third property is a highly important extension of equilibrium thermodynamics because existence and uniqueness of a stable equilibrium in the interior of the positive orthant of concentration space is extended from strictly reversible to weakly reversible systems, from closed systems to closed and open systems of deficiency zero. It is worth stressing again that the statements hold for arbitrary finite dimensions of the reaction system irrespectively of the particular choice of rate parameters – provided they are nonnegative.

4.1.3.7 The deficiency one theorem

The results of the deficiency zero theorem hold for a much wider class of networks than those with deficiency zero. The extension of the range of validity is encapsulated in the *deficiency one theorem*. For the formulation of the theorem it is important to extend the notion of deficiency to individual linkage classes, which are denoted as $\mathcal{L} = \{\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_L\}$. The number of complexes in linkage class $\mathbf{L}_j$ is denoted by N_j and since a complex can appear only in one linkage class we have $\sum_{j=1}^{L} N_j = N$. The number of independent degrees of freedom of the ODE or the rank of a linkage class $\mathbf{L}_j$ is denoted by R_j,

$$R_j \;:=\; \mathrm{rank}\{\mathbf{C}_\mathrm{P} - \mathbf{C}_\mathrm{R} \in \mathbb{R}^{\mathcal{S}} : \mathbf{C}_\mathrm{R} \to \mathbf{C}_\mathrm{P} \in \mathcal{R} \wedge \mathcal{C}_\mathrm{R} \in \mathbf{L}_j\}$$

and we define:

$$\text{Deficiency of class } \mathbf{L}_j : \quad \delta_j \;=\; N_j \,-\, 1 \,-\, R_j \;. \tag{4.25}$$

The class deficiency δ_j is a nonnegative integer like δ. The ranks of the subsystems need not be additive but they fulfil $\sum_{j=1}^{L} R_j \geq R$ and this yields for the deficiency of the total network

$$\delta \;\geq\; \sum_{j=1}^{L} \delta_j \;=\; N \,-\, L \,-\, \sum_{j=1}^{L} R_j \;. \tag{4.24'}$$

It is illustrative to consider zero deficiency networks because they are precisely those networks that fulfil both of the conditions:

$$\delta_j \;=\; 0 \;\forall\; j = 1, 2, \ldots, L\,, \quad \text{and} \quad \delta \;=\; \textstyle\sum_{j=1}^{L} \delta_j \;=\; 0.$$

Now are in a position to introduce the deficiency one theorem [70].

Let $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$ be a reaction network with L linkage classes, let $\delta = N - L - R$ denote the deficiency of the network, $\delta_j = N_j - 1 - R_j$; $j = 1, \ldots, L$ denote the deficiencies of the individual linkage classes, and assume that the two following conditions are fulfilled:

$$\delta_j \;\leq\; 1 \;\forall\; j = 1, 2, \ldots, L\,, \quad \text{and} \quad \delta \;=\; \textstyle\sum_{j=1}^{L} \delta_j \;(=\; 0).$$

If the network is weakly reversible, in particular if it is strongly reversible, then for any mass action kinetics $\kappa \in \mathbb{R}_{>}^{\mathcal{R}}$ the ODEs for the mass action system $\{\mathcal{S}, \mathcal{C}, \mathcal{R}, \kappa\}$ sustains precisely one equilibrium in each positive stoichiometric compatibility class.

Thus deficiency one theorem is a powerful tool for the recognition of reaction system lacking multiple stationary states. In later works the existence of multiple stationary states came in focus [43, 72] and these studies make a bridge between applications in chemistry and in biology. We shall come back to reaction systems with multiple steady states and complex dynamics in the next chapter 5.

4.2 Stochasticity in chemical reactions

Provided particle numbers are assigned to the variables describing the progress of chemical reactions, the stochastic variable $\mathcal{N}(t)$ with the probability $P_n(t) = P(\mathcal{N}(t) = n)$ can take only nonnegative integer values, $n \in \mathbb{N}^0$. In addition we introduce a few simplifications and some conventions in our notation. We shall use the forward equation unless stated differently and assume an infinitely sharp initial density: $P(n, 0|n_0, 0) = \delta_{n,n_0}$ with $n_0 = n(0)$. Then, we can simplify the full notation by $P(n, t|n_0, 0) \Rightarrow P_n(t)$ with the implicit assumption of the initial condition specified above. Other sharp initial values or for initial extended probability densities will be given explicitly. In addition the notation $P_n(t)$ implies already that t is a continuous variable whereas n is discrete. The expectation value of the stochastic variable $\mathcal{N}(t)$ will be denoted by

$$E\big(\mathcal{N}(t)\big) = \langle n(t)\rangle = \sum_{n=0}^{\infty} n \cdot P_n(t) \ . \tag{4.26}$$

Its stationary value, provided it exists, will be expressed as

$$\bar{n} = \lim_{t\to\infty} \langle n(t)\rangle \ . \tag{4.27}$$

Almost always the stationary expectation value $\bar{n}$ will be identical with the long time value of the corresponding deterministic variable. The running index of integers will be denoted by m.[14]

4.2.1 The chemical master equation

The *chemical master equation* is of the form

$$\frac{\partial P_n(t)}{\partial t} \ = \ \sum_m \Big(W(n|m,t)P_m(t) \ - \ W(m|n,t)P_n(t) \Big) \ . \tag{4.28}$$

We have accounted here for the fact that transition probabilities may be time dependent in certain cases. Most frequently we shall assume, however, that they are not and use $W(n|m)$. The probabilities $W(n|m,t)$ can be understood as the elements of a transition matrix $\mathrm{W} := \{W_{nm}; n, m \in \mathbb{N}^0\}$. Diagonal elements W_{nn} cancel in the master equation (4.28) and hence, in principle, need not be defined. According to their nature as *transition probabilities*, all W_{nm} with $n \neq m$ have to be nonnegative. Two definitions of diagonal elements are nevertheless common (i) normalization

[14] In cases were more than one running index are required we shall use n', m', etc.

$$W_{nn} \;=\; 1 \,-\, \sum_{m\neq n} W_{mn} \;\; \text{with} \;\; \sum_{m} W_{mn} \;=\; 1$$

as used for example in the mutation selection problem [55], or (ii) annihilation, where the definition $\sum_m W_{mn} = 0$ is used, which implies $W_{nn} = -\sum_{m\neq n} W_{mn}$ and then insertion into (4.28) leads to a compact form of the master equation master equation

$$\frac{\partial P_n(t)}{\partial t} \;=\; \sum_{m} W_{nm} P_m(t) \;. \tag{4.28'}$$

Introducing vector notation, $\mathbf{P}(t)' = (P_1(t), \ldots, P_n(t), \ldots)$, we obtain

$$\frac{\partial \mathbf{P}(t)}{\partial t} \;=\; \mathrm{W} \times \mathbf{P}(t) \;. \tag{4.28"}$$

With the initial condition $P_n(0) = \delta_{n,n_0}$ stated above we can solve equation (4.28") in formal terms for each n_0 by applying linear algebra and obtain

$$P(n,t|n_0,0) \;=\; \Big(\exp(\mathrm{W}\,t)\Big)_{n,n_0} \;,$$

where the element (n, n_0) of the matrix $\exp(\mathrm{W}t)$ is the probability to have n particles at time t, $\mathcal{N}(t) = n$, when there were n_0 particles at time $t_0 = 0$. The evaluation of this equation boils down to diagonalize the matrix W which can be done analytically in rather few low-dimensional cases only.

For the forthcoming considerations of stochastic processes it is often convenient to express changes in particle numbers in terms of the so-called *jump moments*

$$\alpha_p(n) \;=\; \sum_{m=0}^{\infty} (m-n)^p \, W(m|n) \,;\; p = 1, 2, \ldots \;. \tag{4.29}$$

The usefulness of the first two jump moments ($p = 1, 2$) is easily demonstrated: We multiply equation (4.28) by n and obtain through summation:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle n\rangle \;&=\; \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\Big(m\,W(n|m)\,P_m(t) \,-\, n\,W(m|n)\,P_n(t)\Big) \;= \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(m-n)\,W(m|n)\,P_n(t) \;=\; \langle \alpha_1(n)\rangle \;.
\end{aligned}$$

Only in case $\alpha_1(n)$ is a linear function of n, formation of moment and expectation value may be interchanged and we have the simple equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n\rangle \;=\; \alpha_1(\langle n\rangle) \;.$$

Otherwise this is only a zeroth order approximation which can be improved through expansion of $\alpha_1(n)$ in $(n-\langle n\rangle)$. Break off after the second derivative yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n\rangle \;=\; \alpha_1(\langle n\rangle) \,+\, \frac{1}{2}\,\sigma_n^2\,\frac{\mathrm{d}^2}{\mathrm{d}n^2}\,\alpha_1(\langle n\rangle)\;. \tag{4.29'}$$

In order to obtain a consistent approximation one may apply a similar approximation to the time development of the variance and finds [286]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\sigma_n^2 \;=\; \alpha_2(\langle n\rangle) \,+\, 2\,\sigma_n^2\,\frac{\mathrm{d}}{\mathrm{d}n}\,\alpha_1(\langle n\rangle)\;. \tag{4.29"}$$

These expressions will be simplified in case of the forthcoming examples. We proceed now by discussing first some important special cases where exact solutions are derivable and then present a general and systematic approximation scheme which allows to solve the master equation for sufficiently large systems [93, 286]. This scheme is based on a power series expansion in some extensive physical parameter Ω, for example the size of the system or the total number of particles. It will turn out that $\Omega^{-1/2}$ is the appropriate quantity for the expansion and thus the approximation is based on the smallness of fluctuations. This implies that we shall encounter the limits of reliability of the technique at small population sizes or in situations of self-enhancing fluctuations, for example at instabilities or phase transitions.

Eventually, we consider the stochastic description of our previous example (4.20). The four random variables are $\mathcal{N}_{\mathsf{A}}(t)$, $\mathcal{N}_{\mathsf{B}}(t)$, $\mathcal{N}_{\mathsf{C}}(t)$, and $\mathcal{N}_{\mathsf{D}}(t)$, but only one variable is independent, and we choose again $\mathcal{N}_{\mathsf{A}}(t)$ with $P_n(t) = P(\mathcal{N}_{\mathsf{A}} = n)$. In order to simplify the initial conditions by assuming that only **A** and **B** are present at time $t = 0$ and they have sharp values: n_0 molecules **A**, $P_n(0) = \delta_{n,n_0}$, and b_0 molecules **B**, $P\big(\mathcal{N}_{\mathsf{B}}(0) = b\big) = \delta_{b,b_0}$, and we have $\mathcal{N}_{\mathsf{B}}(t) = \vartheta_0 + \mathcal{N}_{\mathsf{A}}(t)$ with $\vartheta_0 = b_0 - n_0$, and $\mathcal{N}_{\mathsf{C}}(t) = \mathcal{N}_{\mathsf{D}}(t) = n_0 - \mathcal{N}_{\mathsf{A}}(t)$. Under these conditions the master equation becomes

$$\begin{aligned}\frac{\partial P_n(t)}{\partial t} \;&=\; k\,(n+1)(\vartheta_0+n+1)\,P_{n+1}(t) \,+ h\,(n_0-n+1)^2\,P_{n-1} - \\ &-\; \Big(k\,n\,(\vartheta_0+n+1) \,+\, h\,(n_0-n)^2\Big)\,P_n(t)\;.\end{aligned} \tag{4.30}$$

The master equation for the irreversible reaction (4.20a) has been solved and will be discussed in section 4.3.3.1, the full reversible reaction is rather very hard to solve and we dispense from further analysis because size expansion and numerical simulation are to be preferred for practical purposes.

The chemical master equation has been shown to be based on a rigorous microscopic concept of chemical reactions in the vapor phase within the frame of classical collision theory [104]. The two general requirements that have to be fulfilled are: (i) a homogeneous mixture as it is assumed to exits through well stirring and (ii) thermal equilibrium implying that the velocities of molecules follow a Maxwell-Boltzmann distribution. Daniel Gillespie's approach focusses on chemical reactions rather than molecular species and is well suited to handle reaction networks. In addition the algorithm can be

easily implemented for computer simulation. We shall discuss the Gillespie formalism together with the computer program in section 4.7.

The macroscopic rate equations are readily derived from the master equation through computation of the expectation value:

$$\begin{aligned}
\frac{\partial}{\partial t}\,E\big(n(t)\big) \;&=\; \frac{\partial}{\partial t}\left(\sum_{n=0}^{\infty} n\,P_n(t)\right) \;= \\
&=\quad \sum_{n=0}^{\infty} n\left(w_{n-1}^{+}\,P_{n-1}(t) \;-\; w_n^{+}\,P_n(t)\right) \;+ \\
&\quad +\;\sum_{n=0}^{\infty} n\left(w_{n+1}^{-}\,P_{n+1}(t) \;-\; w_n^{-})\,P_n(t)\right) \;= \\
&= \sum_{n=0}^{\infty}\Big((n+1)\,w_n^{+} \;-\; n\,w_n^{+} \;+\; (n-1)\,w_n^{-} \;-\; n\,w^{-}(n)\Big)\,P_n(t) \;= \\
&= \sum_{n=0}^{\infty} w_n^{+}\,P_n(t) \;-\; \sum_{n=0}^{\infty} w_n^{-}\,P_n(t) \;=\; E\big(w_n^{+}\big) \;-\; E\big(w_n^{-}\big)\;.
\end{aligned}$$

Neglect of fluctuations yields the deterministic rate equation of the birth-and-death process

$$\frac{\mathrm{d}\langle n\rangle}{\mathrm{dt}} \;=\; w_{\langle n\rangle}^{+} \;-\; w_{\langle n\rangle}^{-}\;. \tag{4.31}$$

From the condition of stationary $\bar{n} = \lim_{t\to\infty}\langle n(t)\rangle$ we derive $w_{\bar{n}}^{+} = w_{\bar{n}}^{-}$. Compared to this results we note that the maximum value of the stationary probability density, $\max\{\bar{P}_n,\, n \in \mathbb{N}^0\}$, is defined by $\bar{P}_{n+1} - \bar{P}_n \approx -(\bar{P}_n - \bar{P}_{n-1})$ or $\bar{P}_{n+1} \approx \bar{P}_{n-1}$, which coincide with the deterministic value for large n.

4.2.2 Conventional and probabilistic rate parameters

The formulation of a chemical master equation for a population variable $\mathcal{X}(t)$ requires knowledge of some probabilistic features of chemical reactions. In particular we need expressions for the probabilities $\pi(t,\,\mathrm{dt})$ that a reactant molecule or a combination of reactant molecules for reaction **R** randomly selected at time t will react to yield products within the next infinitesimal time interval $[t, t + \mathrm{dt}[$. Under two assumptions, (i) spatial homogeneity assumed to be achieved by fast mixing, and (ii) thermal equilibrium, virtually all chemical reactions fulfil the condition

$$\pi(t,\,\mathrm{dt}) \;=\; \gamma\;\mathrm{dt}\;, \tag{4.32}$$

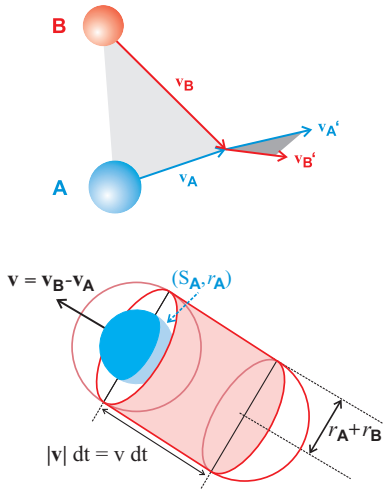

Fig. 4.4 Sketch of a molecular collision in dilute gases. A spherical molecule $\mathbf{S_A}$ with radius $r_\mathbf{A}$ moves with a velocity $\mathbf{v} = \mathbf{v_B} - \mathbf{v_A}$ relative to a spherical molecule $\mathbf{S_B}$ with radius $r_\mathbf{B}$. The upper part of the figure shows the geometry of a typical elastic collision, for which linear angular momentum, $\mathbf{p} = m \cdot \mathbf{v}$, and kinetic energy $E_{\text{kin}} = m \cdot v^2/2$ are conserved: $\mathbf{p_B} + \mathbf{p_A} = \mathbf{p'_A} + \mathbf{p'_B}$ and $m_\mathbf{A} \cdot |\mathbf{v_A}|^2 + m_\mathbf{B} \cdot |\mathbf{v_B}|^2 = m_\mathbf{A} \cdot |\mathbf{v'_A}|^2 + m_\mathbf{A} \cdot |\mathbf{v'_B}|^2$. The lower part of the figure shows the geometry of the collision as seen within the coordinate system of one collision partner. If the two molecules are to collide within the next infinitesimal time interval dt, the center of $\mathbf{S_B}$ has to lie inside a cylinder of radius $r = r_\mathbf{A} + r_\mathbf{B}$ and height $|\mathbf{v}|\,\text{dt} = v\,\text{dt}$. The upper and lower surface of the cylinder are deformed into identically oriented hemispheres of radius r and therefore the volume of the deformed cylinder is identical with that of the non-deformed one.

where the reaction specific *probabilistic rate parameter*[15] γ is independent of t, and then π is simply proportional to dt. The two basic conditions (i) and (ii) are fulfilled likewise for chemical reactions in the vapor phase and in dilute aqueous solutions.

In contrast to most other probabilistic concepts the rate parameters of chemical kinetics, in essence, can be deduced from first principles in quantum mechanics and therefore we make here a brief excursion into the theory of reaction rates which dates back to the beginnings of applications of quantum mechanics to chemistry [68, 172] in order to show how a stochastic treatment may results from an underlying deterministic concept. Molecules or atoms have to come together before they can react and molecular collisions play a key role in the theory of chemical reactions [32] and therefore we begin with a short account on molecular collisions (For an excellent introduction into statistical physics of molecular reactions see [16, pp.803-1018]).

[15] Thus γ is the probabilistic pendant of the deterministic reaction rate parameter k.

4.2.2.1 Molecular collisions

Here, we consider first the rate parameter for a general bimolecular reaction by means of classical collision theory (figure4.4, and then extend briefly to mono- and termolecular reactions. Apart from the quantum mechanical approach the theory of collisions in dilute gases is the best developed microscopic model for chemical reactions and well suited for the rigorous derivation of a probabilistic description of chemical reactions from molecular motion and events in form of a the master equation. First we consider a reaction mixture in the vapor phase for which the Maxwell-Boltzmann theory is valid. This concept is dealing with molecular motions in gases and centers around the assumption that molecules are obeying the laws of Newtonian mechanics and therefore it is also called *classical collision theory* of chemical reactions. Molecules change their motions, their internal states, and their natures in collisions that are classified as elastic, inelastic and reactive, respectively. In an elastic collision the collision partners exchange linear momentum and energy related to it, and as a consequences the directions and the absolute values of the velocities of both collision partners before and after the collision are different (figure 4.4). In an inelastic collision internal energy, rotational and/or vibrational and in exceptional cases also electronic energy is transferred between the reaction partners. Finally, in a reactive collision a chemical reaction takes place between the reaction partners and the molecular species are different before and after the collision.

In order to be able to handle the properties of individual molecules, we must be able to distinguish a molecular species and an individual molecule, e.g. $\mathbf{A}$ and $\mathbf{S_A}$. In the latter case knowledge of the detailed molecular state $\Lambda_\mathbf{A}$ is required, for example, $\mathbf{S}_\mathbf{A}^{\Lambda_\mathbf{A}}$with $\Lambda_\mathbf{A} = (N_\mathbf{A}, \Sigma_\mathbf{A}, n_\mathbf{A}, J_\mathbf{A}; m_\mathbf{A}, \mathbf{r_A}, \mathbf{v_A})$ where $(N_\mathbf{A}, \Sigma_\mathbf{A}, n_\mathbf{A}, J_\mathbf{A})$ stands for a complete set of molecular quantum numbers characterizing electronic and spin state $(N_\mathbf{A}, \Sigma_\mathbf{A})$, vibrational state $(n_\mathbf{A})$, and rotational state $J_\mathbf{A})$ of molecule $\mathbf{S_A}$. The mass of the molecule is $m_\mathbf{A}$, position $(\mathbf{r_A})$ and velocity coordinates $(\mathbf{v_A})$ are commonly measured in a Cartesian (labor) coordinate system: $\mathbf{r_A}(t) = (x_\mathbf{A}, y_\mathbf{A}, z_\mathbf{A})$ and $\mathbf{v_A}(t) = \left(v_x^\mathbf{A}, v_y^\mathbf{A}, v_z^\mathbf{A}\right)$. In the spirit of classical mechanics, apart from spontaneous changes in collisions the position vector is a linear function of time, $\mathbf{r}(t) = \mathbf{r}_0 + \mathbf{v}\cdot t$, and the velocity is constant, $\mathbf{v} = \mathbf{v}_0$, or in other words the molecules travel on a straight line with constant speed between collisions. On this basis we can easily identify the different classes of bimolecular collisions, $\mathbf{A} + \mathbf{B} \rightarrow$, by means of examples where ‘′’ is used to indicate the state after the collision:

(1) Elastic collisions: $S_\mathbf{A} + S_\mathbf{B} \rightarrow S_\mathbf{A} + S_\mathbf{B}$ with $m_\mathbf{A}\mathbf{v_A} + m_\mathbf{B}\mathbf{v_B} = m_\mathbf{A}\mathbf{v'_A} + m_\mathbf{B}\mathbf{v'_B}$ and $\frac{1}{2}(m_\mathbf{A}|\mathbf{v_A}|^2 + m_\mathbf{B}|\mathbf{v_B}|^2) = \frac{1}{2}(m_\mathbf{A}|\mathbf{v'_A}|^2 + m_\mathbf{B}|\mathbf{v'_B}|^2)$ corresponding to conservation of linear momentum and kinetic energy. The set of internal quantum numbers remains unchanged in both molecules.

(2) Inelastic collisions: $S_\mathbf{A}^{\Lambda_\mathbf{A}} + S_\mathbf{B}^{\Lambda_\mathbf{A}} \rightarrow S_\mathbf{A}^{\Lambda'_\mathbf{A}} + S_\mathbf{B}^{\Lambda'_\mathbf{B}}$ where the set of quantum numbers for internal motions has been changed in the collision.

(3) Reactive collisions: $S_{\mathbf{A}} + S_{\mathbf{B}} \to \cdots$ where the two molecules undergo a chemical reaction in which the nature of at least one molecule is changed.

The correct description of translational motion in a macroscopic reaction vessel does not require quantum mechanical treatment and hence elastic collisions are just an exercise in Newtonian mechanics. Internal energy of molecules is converted into translational energy in inelastic collisions, and a quantum mechanical approach is needed for detailed modeling. The same is true for reactive collisions in case one is interested in reactions of molecules in specific states, otherwise the reaction can described by a mean reaction probability that averages over a Boltzmann ensemble (for the theory of molecular collisions see, e.g., [32]).

The two conditions, (i) *perfect mixture* and (ii) *thermal equilibrium*, can now be cast into precise physical meanings. Premise (i), *spatial homogeneity*, requires that the probability of finding the center of an arbitrarily chosen molecule inside a container subregion with a volume ΔV is equal to $\Delta V/V$. The system is spatially homogeneous on macroscopic scales but it allows for random fluctuations from homogeneity. Formally, requirement (i) asserts that the position of a randomly selected molecule is described by a random variable, which is uniformly distributed over the interior of the container. Premise (ii), thermal equilibrium, implies that the velocity of a randomly chosen molecule of mass m will follow a *Maxwell-Boltzmann distribution*. The Maxwell-Boltzmann density

$$f_{MB}(\mathbf{v})\,\mathrm{d}\mathbf{v}^3 \;=\; \left(\frac{m}{2\pi k_{\mathrm{B}}T}\right)^{3/2} e^{-mv^2/(2k_{\mathrm{B}}T)}\,\mathrm{d}\mathbf{v}^3 \;,$$

describes the probability that the velocity of the molecule is found to lie within an infinitesimal region $\mathrm{d}\mathbf{v}^3$ around the velocity $\mathbf{v}$ where the velocity vector is denoted in Cartesian coordinates by $\mathbf{v} = (v_x, v_y, v_z)$, the infinitesimal volume element is given by $\mathrm{d}\mathbf{v}^3 = \mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$, the square of the velocity is $v^2 = v_x^2 + v_y^2 + v_z^2$, and k_{B} is Boltzmann's constant. Formally it states that each Cartesian velocity component of a randomly selected molecule of mass m is represented by a random variable, which is normally distributed with mean 0 and variance $k_{\mathrm{B}}T/m$:

$$f_{\mathrm{M}B}(v_i)\,\mathrm{d}v_i \;=\; \left(\frac{m}{2\pi k_{\mathrm{B}}T}\right)^{1/2} e^{-mv^2/(2k_{\mathrm{B}}T)}\,\mathrm{d}v_i \quad \text{with} \quad i = x, y, z \;. \tag{4.33}$$

Here, premises (i) and (ii) assert that the distribution of molecular velocities is isotropic and only a function of mass m and temperature T. Implicitly, the two conditions guarantee also that the molecular position and velocity components are all statistically independent of each other. For practical purposes, we expect the two premises to be valid for any dilute gas system at constant temperature in which *nonreactive* molecular collisions occur much more frequently than *reactive* molecular collisions.

4.2.2.2 Bimolecular reactions

The occurrence of a bimolecular reaction

$$\mathbf{S_A} + \mathbf{S_B} \longrightarrow \mathbf{S_C} + \dots \tag{4.34}$$

has to be preceded by an encounter of a molecule $\mathbf{S_A}$ with a molecule $\mathbf{S_B}$, and first we shall calculate the probability of such a collision in the reaction volume V. For simplicity molecular species are regarded as spheres with specific masses and radii, for example $m_{\mathbf{A}}$ and $r_{\mathbf{A}}$ for $\mathbf{S_A}$, and $m_{\mathbf{B}}$ and $r_{\mathbf{B}}$ for $\mathbf{S_B}$, respectively. A collision occurs whenever $r_{\mathbf{AB}}$, the center-to-center distance of the two molecules, becomes as small as the sum of the two radii, $(r_{\mathbf{AB}})_{\min} = r_{\mathbf{A}} + r_{\mathbf{B}}$. Next we define the probability that a randomly selected pair of $\mathbf{R}_{\boldsymbol{\mu}}$ reactant molecules – $\boldsymbol{\mu} = (\mathbf{S_A}, \mathbf{S_B})$ – at time t will collide within the next infinitesimal time interval $[t, t + \mathrm{d}t[$ by $\pi^*_{\boldsymbol{\mu}}(t, \mathrm{d}t)$ and calculate it from the Maxwell-Boltzmann distribution of molecular velocities according to the geometry shown in figure 4.4.

The probability that a randomly selected pair of reactant molecules $\mathbf{R}_{\boldsymbol{\mu}}$, one molecule $\mathbf{S_A}$ and one molecule $\mathbf{S_B}$, has a relative velocity $\hat{\mathbf{v}} = \mathbf{v_B} - \mathbf{v_A}$ lying in an infinitesimal volume element $\mathrm{d}\hat{\mathbf{v}}^3$ around $\hat{\mathbf{v}}$ at time t is denoted by $f\big(\hat{\mathbf{v}}(t), \mathbf{R}_{\boldsymbol{\mu}}\big)$ and can be readily obtained from kinetic theory of gases:

$$f\big(\hat{\mathbf{v}}(t), \mathbf{R}_{\boldsymbol{\mu}}\big) \;=\; \left(\frac{\hat{m}}{2\pi\, k_{\mathrm{B}} T}\right)^{3/2} \exp\big(-\hat{m}\hat{v}^2/(2k_{\mathrm{B}}T)\big)\, \mathrm{d}\hat{\mathbf{v}}^3 \;.$$

Herein $\hat{v} = |\hat{\mathbf{v}}| = |\mathbf{v_B} - \mathbf{v_A}| = \sqrt{\hat{v}_x^2 + \hat{v}_y^2 + \hat{v}_z^2}$ is the absolute value of the relative velocity and $\hat{m} = m_{\mathbf{A}} m_{\mathbf{B}}/(m_{\mathbf{A}} + m_{\mathbf{B}})$ is the reduced mass of the two $\mathbf{R}_{\boldsymbol{\mu}}$ molecules. Two properties of the probabilities $f\big(\hat{\mathbf{v}}(t), \mathbf{R}_{\boldsymbol{\mu}}\big)$ for different velocities $\hat{\mathbf{v}}$ are important:
(i) The elements in the set of all combinations of velocities, $\{\mathcal{E}_{\hat{\mathbf{v}}(t),\mathbf{R}_{\boldsymbol{\mu}}}\}$ are mutually exclusive, and
(ii) they are collectively exhaustive since $\hat{\mathbf{v}}$ is varied over the entire three dimensional velocity space,$-\infty < (\hat{v}_x, \hat{v}_y, \hat{v}_z) < +\infty$.
Now we relate the probability $f\big(\hat{\mathbf{v}}(t), \mathbf{R}_{\boldsymbol{\mu}}\big)$ to a collision event $\mathcal{E}_{\mathrm{col}}$ by calculating the conditional probability $P\big(\mathcal{E}_{\mathrm{col}}(t+\mathrm{d}t)|\mathcal{E}_{\hat{\mathbf{v}}(t),\mathbf{R}_{\boldsymbol{\mu}}}\big)$. In figure 4.4 we sketch the geometry of the collision event between to randomly selected spherical molecules $\mathbf{S_A}$ and $\mathbf{S_B}$ that is assumed to occur with an infinitesimal time interval $\mathrm{d}t$:[16] A randomly selected molecule $\mathbf{S_A}$ moves along the vector $\hat{\mathbf{v}}$ of the relative velocity $\mathbf{v_B} - \mathbf{v_A}$ between $\mathbf{S_A}$ and an also randomly selected molecule $\mathbf{S_B}$. A collision between the molecules will take place in the interval $[t, t + \mathrm{d}t$ if and only if the center of molecule $\mathbf{S_B}$ is inside the spherically distorted cylinder (figure 4.4) at time t. Thus $P\big(\mathcal{E}_{\mathrm{col}}(t + \mathrm{d}t)|\mathcal{E}_{\hat{\mathbf{v}}(t),\mathbf{R}_{\boldsymbol{\mu}}}\big)$ is the probabil-

[16] The absolute time t comes into play because the positions of the molecules, $\mathbf{r_A}$ and $\mathbf{r_B}$, and their velocities, $\mathbf{v_A}$ and $\mathbf{v_B}$, depend on t.

ity that the center of a randomly selected $\mathbf{S_B}$ molecule moving with velocity $\hat{\mathbf{v}}(t)$ relative to the randomly selected $\mathbf{S_A}$ molecule will be situated at time t within a certain subregion of V that has a volume $V_{col} = \hat{v}\,\mathrm{dt} \cdot \pi(r_\mathbf{A} + r_\mathbf{B})^2$, and by scaling with the total volume V we obtain:[17]

$$P\big(\mathcal{E}_{\mathrm{col}}(t+\,\mathrm{dt})|\mathcal{E}_{\hat{\mathbf{v}}(t),\mathbf{R}_{\boldsymbol{\mu}}}\big) \;=\; \frac{\hat{v}(t)\,\mathrm{dt} \cdot \pi(r_\mathbf{A} + r_\mathbf{B})^2}{V}\;. \tag{4.35}$$

By substitution and integration over the entire velocity space we can calculate the desired probability

$$\pi^*_{\boldsymbol{\mu}}(t,\,\mathrm{dt}) \;=\; \iiint\limits_{\mathbf{v}} \left(\frac{\hat{m}}{2\pi\,k_\mathrm{B}T}\right)^{3/2} e^{-\hat{m}\hat{v}^2/(2k_\mathrm{B}T)} \cdot \frac{\hat{v}(t)\,\mathrm{dt} \cdot \pi(r_\mathbf{A} + r_\mathbf{B})^2}{V}\;\mathrm{d}\hat{\mathbf{v}}^3\;.$$

Evaluation of the integral is straightforward and yields

$$\pi^*_{\boldsymbol{\mu}}(t,\,\mathrm{dt}) \;=\; \left(\frac{8\,k_\mathrm{B}T}{\pi V^2}\right)^{1/2} \frac{\pi(r_\mathbf{A} + r_\mathbf{B})^2}{\sqrt{\hat{m}}}\;\mathrm{dt}\;. \tag{4.36}$$

The first factor contains only constants and the macroscopic quantities, volume V and temperature T, whereas the molecular parameters, the radii $r_\mathbf{A}$ and $r_\mathbf{B}$ and the reduced mass $\hat{m}$ appear in the second factor.

A collision is a necessary but not a sufficient condition for a reaction to take place and therefore we introduce a *collision-conditioned reaction probability* $p_{\boldsymbol{\mu}}$ that is the probability that a randomly selected pair of colliding $\mathbf{R}_{\boldsymbol{\mu}}$ reactant molecules will indeed react according to $\mathbf{R}_{\boldsymbol{\mu}}$. By multiplication of independent probabilities we have

$$\pi_{\boldsymbol{\mu}}(t,\,\mathrm{dt}) \;=\; p_{\boldsymbol{\mu}}\,\pi^*_{\boldsymbol{\mu}}(t,\,\mathrm{dt})\;,$$

and with respect to equation (4.32) we find

$$\gamma_{\boldsymbol{\mu}} \;=\; p_{\boldsymbol{\mu}} \left(\frac{8\,k_\mathrm{B}T}{V}\right)^{1/2} \frac{\pi(r_\mathbf{A} + r_\mathbf{B})^2}{\sqrt{\hat{m}}}\;. \tag{4.37}$$

As said before, it is crucial for the forthcoming analysis that $\gamma_{\boldsymbol{\mu}}$ is independent of dt and this will be the case if and only if reaction probability $p_{\boldsymbol{\mu}}$ does not depend on dt. This is highly plausible for the above given definition, and an illustrative check through the detailed examination of bimolecular reactions can be found in [104, pp.413-417].

The results of collision theory for reactive bimolecular encounters can be summarized in a commonly used form for the probabilistic rate parameter and its temperature dependence

[17] Implicitly in the derivation we made use of the infinitesimally small size of dt. Only if the distance $\hat{v}$ dt is vanishingly small, the possibility of collisional interference of a third molecule can be neglected.

$$\gamma_{\boldsymbol{\mu}}(T) \;=\; \zeta\,\rho\,\exp\left(-\frac{\varepsilon_{\mathrm{act}}}{k_{\mathrm{B}}T}\right)\;. \tag{4.38}$$

Equation (4.38) was proposed for the temperature dependence of the deterministic rate parameter already in 1884 by the Swedish physicist and chemist Svante Arrhenius.[18] Herein ζ is the collision frequency as calculated above

$$\zeta \;=\; \sigma_{\mathbf{AB}}\sqrt{\frac{8\,k_{\mathrm{B}}T}{\pi\hat{m}}} \;\text{ with }\; \sigma_{\mathbf{AB}} = (r_{\mathbf{A}} + r_{\mathbf{B}})^2\,\pi.$$

The factor ρ is a denoted as steric factor and $\varepsilon_{\mathrm{act}}$ is called the activation energy of the reaction that is measured here as energy per molecule. Often particle numbers are used instead of concentrations and this implies multiplication by Avogadro's number. Then the activation energy, $E_{\mathrm{act}} = N_{\mathrm{L}}\,\varepsilon_{\mathrm{act}}$, is commonly given in [kJ/mole] and the gas constant $R = N_{\mathrm{L}}\,k_{\mathrm{B}}$ is used instead of Boltzmann's constant. The actual number of collisions in the volume V per time unit is $Z = N_{\mathrm{L}}\,V\,\zeta$. The exponential temperature dependence of the rate parameter on temperature is often fulfilled with astonishingly high accuracy but an interpretation of the steric factor ρ is often unsatisfactory and therefore some chemists prefer to stay away from any rationalization of the steric factor and define it simply as the ratio between the pre-exponential factor and the collision frequency: $\rho = A/\zeta$.

It has to be remarked, however, that the application of classical collision theory to molecular details of chemical reactions can be an illustrative and useful heuristic at best, because the molecular domain falls into the realm of quantum phenomena and any theory that aims at a derivation of reaction probabilities from first principles has to be built upon a quantum mechanical basis (section 4.2.2.3).

4.2.2.3 Bimolecular reaction dynamics

For any detailed understanding of chemical reactions knowledge from quantum mechanics mechanics is indispensable and we refer here to the great variety of text books. Very briefly we sketch the basic idea: In conventional quantum chemistry the fast motion of electrons is separated from slow motion of atomic nuclei and the stationary Schrödinger equation of a molecule or a reaction complex is partitioned into two equations

$$\mathcal{H}_{\mathrm{el}}\,\Psi_{\mathrm{el}}^{(n)} \;=\; E_n(\mathbf{R})\,\Psi_{\mathrm{el}}^{(n)} \;\text{ with }\; \mathcal{H}_{\mathrm{el}} \;=\; \mathcal{T}_{\mathrm{el}} + V(\mathbf{r},\mathbf{R})\;, \tag{4.39a}$$

$$\Big(\mathcal{T}_{\mathrm{nuc}} \;+\; E_n(\mathbf{R})\Big)\,\Xi_{\mathrm{nuc}}^{(k;n)} \;=\; W_{k,n}\,\Xi_{\mathrm{nuc}}^{(k;n)}\;. \tag{4.39b}$$

[18] Svante Arrhenius used a slightly different form, $k = A\cdot\exp(-E_A/RT)$, where A is the so-called pre-exponential factor.

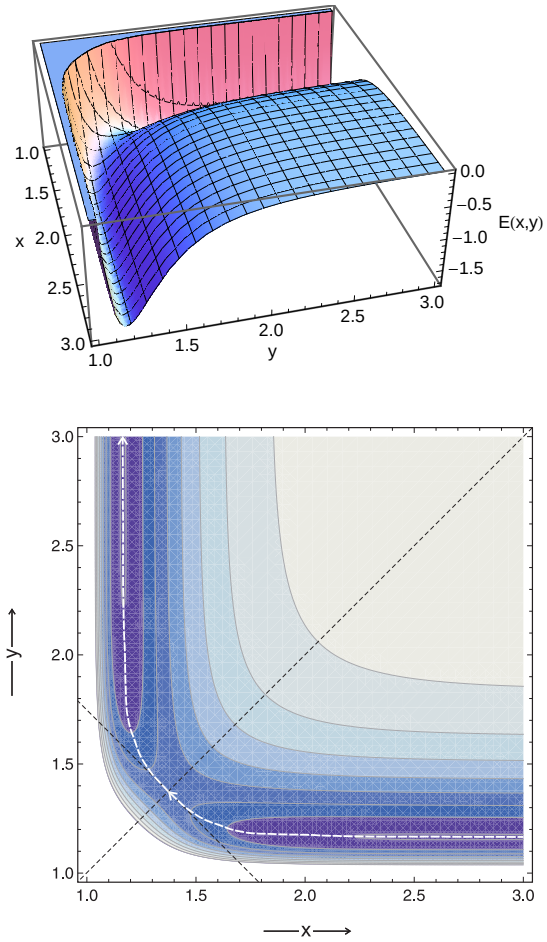

Fig. 4.5 Energy surface of the symmetric bimolecular triatomic exchange reaction $\mathbf{S_A + S_B \rightarrow S_B + S_A}$. The best studied example of such a reaction is the hydrogen isotope exchange reaction $\mathrm{D + HD \rightarrow DH + D}$ for which a highly accurate energy surface is available. The three atoms lie on a straight line. The model surface plotted here is

$$E(x,y) = a/x^{12} - b/x^6 + a/y^{12} - b/y^6 + c/(x+y)^{12}.$$

The upper part of the figure shows a 3D-plot of the energy surface with the reaction path being recognizable as a steep valley. The lower part presents a contour plot of this surface. The dotted white line indicates the reaction path. In the steep horizontal valley at the bottom of the figure the atom is approaching the molecule, then the bond becomes longer and at the saddle point the two bonds are of equal length. Parameters: $a = 10$, $b = 8$, and $c = 1.5 \times 10^5$, leading to a bond length of $r_e = 1.165$ [l.u.] and a bond energy of $\Delta E = -1.6$ [e.u.]. At the saddle point the distance is $x = y = 1.3856$ [l.u.] and the energy amounts to $\Delta E = -1.1303$ [e.u.]. Length and energy are given in arbitrary units, [l.u.] stands for length unit and [e.u.] for energy unit respectively.

Herein the positions of all electrons are subsumed in the vector $\mathbf{r}$, and likewise the nuclei occupy positions denoted by $\mathbf{R}$. Both equations are partial differential equations and they are coupled through the energy (hyper)surface $E_n(\mathbf{R})$ (see figure 4.5). The Hamilton operator $\mathcal{H}_{\mathrm{el}}$ describes the motion of electrons and consists of the kinetic energy operator of electrons $\mathcal{T}_{\mathrm{el}}$ and the electrostatic potential $V(\mathbf{r},\mathbf{R})$ caused by the electric charges of electrons and nuclei, $E_n(\mathbf{R})$ is the n-th eigenvalue of the Schrödinger equation (4.39a), and $\Psi_{\mathrm{el}}^{(n)}$ is the corresponding eigenfunction. The separation of electronic and nuclear motion was introduced into quantum mechanics by Max Born and Robert Oppenheimer in 1927 [24]. Because of the large difference in mass between electrons and nuclei – being at least three orders of magnitude – and the reasonable assumption that linear momenta of electrons and nuclei are roughly the same because the forces acting on them are identical – actio equals reactio – we have

$$M\frac{\mathrm{d}\mathbf{R}}{\mathrm{dt}} \;=\; \mathbf{P} \;\approx\; \mathbf{p} \;=\; m\frac{\mathrm{d}\mathbf{r}}{\mathrm{dt}} \quad \text{with} \quad M \gg m \quad \text{and hence} \quad \frac{\mathrm{d}\mathbf{R}}{\mathrm{dt}} \ll \frac{\mathrm{d}\mathbf{r}}{\mathrm{dt}}\,.$$

Seen from the fast moving electrons nuclei are practically immobile, the total wave function can be factorized, $\Phi(\mathbf{r},\mathbf{R}) = \Psi_{\mathrm{el}}^{(n)}(\mathbf{r}) \cdot \Xi_{\mathrm{nuc}}^{(k;n)}(\mathbf{R})$ or, in other words, the electrons see the nuclei at fixed positions and the nuclei see the electrons in form of a potential coming from a time-averaged mean density. Within the Born-Oppenheimer approximation the connecting piece between the electron density in the quantum state n and nuclear motion but also chemical reactions is the energy (hyper)surface $E_n(\mathbf{R})$. Classical collision theory (section 4.2.2.2) did not account for energetic aspects of reactions and the consideration of an energy surface is an appropriate and important extension. Nuclear motion can be modeled by Newtonian mechanics and the combination of an energy surface of quantum mechanical origin and classical dynamics is often addressed as semiclassical collision theory in contrast to the full quantum mechanical approach based on scattering theory [32].

Roughly a decade after the establishment of quantum mechanics Henry Eyring proposed a theory of chemical reactions [68] that allows to calculate reaction rate parameters. This theory also called *transition state theory* is still use more than 65 years after its invention and provides the alternative to the fully empirical reaction probabilities of collision theory [172]. In order to be activated for the reaction the reaction complex has to be driven up the reaction coordinate (ρ) through energy transfer from other degrees of freedom until the local maximum called transition state is reached (figure 4.6). Then the reaction complex travel down the product valley and looses energy through transfer to other degrees of freedom. The transition state is symbolized by double-dagger (‡) and is treated like a molecular entity except one unstable vibrational mode along the reaction coordinate ρ. Thermodynamics is applied to calculate the reaction rate parameter for the reaction

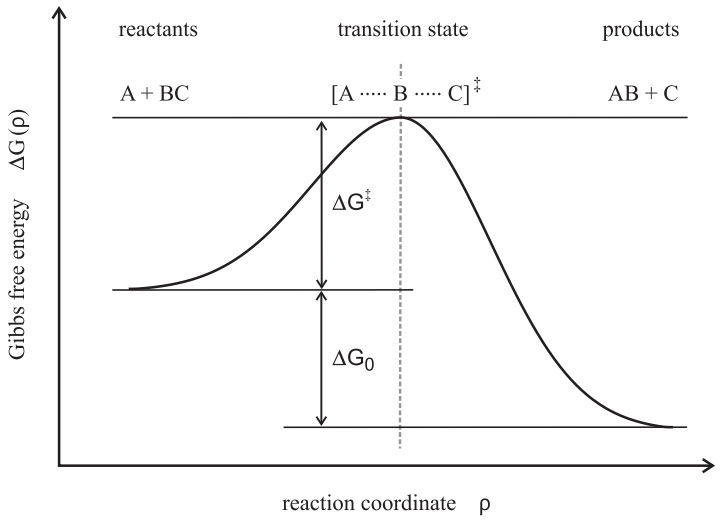

Fig. 4.6 Transition state for the reaction $A+BC \rightarrow AB+C$**.** Reaction dynamics is visualized as a process along a single coordinate called the *reaction coordinate* ρ. The Gibbs free energy of the reaction complex, $\Delta G(\rho)$, is plotted against the reaction coordinate and increases during the the approach of the reactants until it reaches a (local) maximum denoted as transition state. Then the reaction complex looses free energy as it goes down the product valley. The example presented is an exergonic reaction since $\Delta G_0 = \Delta G_{\mathrm{reactants}} - \Delta G_{\mathrm{products}} < 0$
.

$$\mathbf{A} + \mathbf{B} \overset{K^{\ddagger}}{\rightleftharpoons} [\mathbf{AB}]^{\ddagger} \xrightarrow{\quad k^{\ddagger} \quad} \text{products}$$

by making a quasi-equilibrium assumption for the transition state:

$$K^{\ddagger} = \frac{[\mathbf{AB}^{\ddagger}]}{[\mathbf{A}] \cdot [\mathbf{B}]} \,. \tag{4.40}$$

The conventional rate parameter is then obtained from $k = k^{\ddagger} \cdot K^{\ddagger}$ and what remains, is to find an expression for the rate $k^{\ddagger}$ with which the transition state is converted into products. The transition state is considered as a molecular complex with one uncommon degree of freedom consisting of the motion along the reaction coordinate ϱ, which leads to products. All other $3n-7$ or $3n-6$ in case of linear geometries degrees of freedom are handled as in conventional statistical mechanics and the equilibrium constant for complex formation is of the form

$$K^{\ddagger} = \frac{q_{\mathbf{AB}^{\ddagger}}}{q_{\mathbf{A}}\, q_{\mathbf{B}}} \, e^{-\Delta H_0^{\ddagger}/RT} \,,$$

wherein the individual partition functions are denoted by q and the enthalpy difference between the transition state and the reactants is $\Delta H_0^{\ddagger}$.[19] The remaining degree of freedom is responsible for product formation and has the partition function $q^{(\varrho)}_{\mathbf{AB}^{\ddagger}}$. No matter whether this mode is interpreted as a degenerate vibration with a negative harmonic potential or as a translational degree of freedom we find $k^{\ddagger} \cdot q^{(\varrho)}_{\mathbf{AB}^{\ddagger}} = k_{\mathrm{B}}T/h$ with h being Planck's constant and the final result is the same:

$$k \;=\; k^{\ddagger} \cdot K^{\ddagger} \;=\; \kappa \, \frac{k_{\mathrm{B}}T}{h} \, e^{\Delta S_0^{\ddagger}/R} \, e^{-\Delta H_0^{\ddagger}/RT} \;. \tag{4.41}$$

By κ we denote an empirical transmission factor measuring the probability that the vibrating activated complex decomposes into the product valley, and activation entropy and activation are related to the equilibrium constant through:

$$RT \, \ln \, K^{\ddagger} \;=\; -\Delta G_0^{\ddagger} \;=\; -\Delta H_0^{\ddagger} \,+\, T \, \Delta S_0^{\ddagger} \;.$$

Equation (4.41) is Eyring's formula for the value of the reaction rate parameter that corresponds to the rate probability $\gamma_{(\mathbf{S_A}+\mathbf{S_B})}$. The value of the formula is twofold: (i) It shows how reaction rate parameters can be derived from first principles, and (ii) it provides a thermodynamic interpretation of the steric factor ρ by means of an activation entropy $\Delta S_0^{\ddagger}$. Direct calculations of rate constants, however, are highly inaccurate since energy surfaces cannot be obtained with sufficient precision apart from a few special cases like the $H+H_2$ reaction (figure 4.5).

4.2.2.4 Monomolecular reactions

A *monomolecular* or *unimolecular reaction* is of the form $\mathbf{A} \longrightarrow \mathbf{C}$ and describes the spontaneous conversion

$$\mathbf{S_A} \;\longrightarrow\; \mathbf{S_C} \;. \tag{4.42}$$

One molecule $\mathbf{S_A}$ is converted spontaneously into one molecule $\mathbf{S_C}$. The monomolecular reaction was first considered to be particularly simple, because only one type of molecule is involved, but this expectation turned out to be wrong: Most formally monomolecular reactions follow a bimolecular rate law at sufficiently low concentrations and have to be distinguished from true monomolecular conversions. It is worth mentioning also a class of spontaneous dissociation reactions of small cluster ions, for example $(H_3O^+)(H_2O)_n$ or $Cl^-(H_2O)_n$ with n=2-4, where the loss of ligands seems to be initiated by collisions with the wall of the reaction vessel [241].

[19] At constant pressure, for example in solution where the volume change ΔV_0 of a reaction is small, the reaction enthalpy ΔH_0 takes on practically the same values as the reaction energy ΔE_0.

In absence of interaction with an environment the true monomolecular conversion (4.42) is driven by some quantum mechanical mechanism similar as in the case of radioactive decay of a nucleus. Time-dependent perturbation theory in quantum mechanics [212, pp.724-739] shows that almost all weakly perturbed energy-conserving transitions have linear probabilities of occurrence in time intervals δt, when δt is *microscopically large* but *macroscopically small*. Therefore, to a good approximation the probability for a radioactive nucleus to decay within the next infinitesimal time interval dt is of the form $\alpha\,\mathrm{dt}$, were α is some time-independent constant. On the basis of analogy we may expect $\pi_{\boldsymbol{\mu}}(t, \mathrm{dt})$ the probability for a monomolecular conversion to be approximately of the form $\gamma_{\boldsymbol{\mu}}\,\mathrm{dt}$ with $\gamma_{\boldsymbol{\mu}}$ being independent of dt.

The vast majority of apparent monomolecular reactions, however, follow a different mechanisms and involve a reaction partner in the sense of a catalyzed bimolecular conversion

$$\mathbf{S_A} + \mathbf{S_B} \longrightarrow \mathbf{S_C} + \mathbf{S_B} \quad \text{or} \tag{4.34'}$$

$$\mathbf{S_A} + \mathbf{S_A} \longrightarrow \mathbf{S_C} + \mathbf{S_A}\ . \tag{4.34"}$$

In equation (4.34') the conversion $\mathbf{A} \longrightarrow \mathbf{C}$ is initiated by a collision of an $\mathbf{S_A}$ molecule with a $\mathbf{S_B}$ molecule, which acts as a catalyst since it is not consumed by the process.[20] When the collision partner is another $\mathbf{S_A}$ molecule (4.34"), we are dealing with a monomolecular reaction in the conventional sense, which is described straightforwardly as a special class of bimolecular process. The first proposal of mechanism (4.34") for the monomolecular conversion has been made already in 1922 by Frederick Lindemann [183]: The monomolecular conversion follows a two step mechanism of the form

$$\begin{aligned} \mathbf{A} + \mathbf{A} &\underset{h_1}{\overset{k_1}{\rightleftharpoons}} \mathbf{A} + \mathbf{A}^* \quad \text{and} \\ \mathbf{A}^* &\xrightarrow{k_2} \mathbf{C} \end{aligned} \tag{4.43}$$

with $k_2 \ll h_1$. The Lindemann mechanism with a conventional rate parameter k_1 did not fit the experimental data and has been improved by Cyril Hinshelwood [129] by a different interpretation of the activation of molecule $\mathbf{A}$ that was extended to a range of energy values $k_{1(E_0 \to E_1)} \Rightarrow k_{1(E_0 \to E_1 + \delta E)}$. Later on the molecular mechanistic details were improved and the Lindemann-Hinshelwood mechanism has been substantially extended by Oscar Rice, Herman Ramsperger [248], and Louis Kassel [153] through the explicit in-

[20] Formally we are dealing with a reaction that is catalyzed by a molecule of the same kind or another kind and the reaction is are related to the spontaneous conversion by rigorous thermodynamics: Whenever a catalyzed reaction appears in a mechanism the uncatalyzed process has to be considered as well.

troduction of a transition state $\mathbf{A}^{\ddagger}$:

$$\begin{aligned} \mathbf{A} + \mathbf{A} &\underset{h_1}{\overset{k_1}{\rightleftarrows}} \mathbf{A} + \mathbf{A}^* \quad \text{and} \\ \mathbf{A}^* &\xrightarrow{k_2 a} \mathbf{A}^{\ddagger} \xrightarrow{k^{\ddagger}} \mathbf{C} \, . \end{aligned} \tag{4.44}$$

As in transition state theory the rate parameter $k^{\ddagger}$ corresponds to the fast process associated with the reactive mode of the transition state. Since $k^{\ddagger}$ is thought to be larger than any other rate parameter, the rate limiting step of the formation of the product $\mathbf{C}$ is the conversion $\mathbf{A}^* \to \mathbf{A}^{\ddagger}$ and comparing Lindemann and RRK mechanism we have $k_2 \approx k_{2a}$ and $k_{2a} = k^{\ddagger}[\mathbf{A}^{\ddagger}]/[\mathbf{A}^*]$ from the steady state assumption. Eventually, the theory of monomolecular reactions got its present form through a reformulation of the transitions state by Rudolph Marcus and Oscar Rice [196, 194, 195]. The current version of the so-called RRKM theory of monomolecular reactions theory allows for a highly accurate and very detailed description of reactions and it can be readily converted into a stochastic formulation [180].

4.2.2.5 Termolecular and other reactions

Termolecular or trimolecular reactions of the form

$$\mathbf{S_A} + \mathbf{S_B} + \mathbf{S_C} \longrightarrow \mathbf{S_D} + \ldots \tag{4.45}$$

are rare and need not be considered because collisions of three particles do not occur with a probability larger than of measure zero. Exceptions are two classes of reactions: (i) Vapor phase association reactions where a third body is required as collision partner and (ii) the reaction of nitrogen monoxide with oxygen or halogens. A characteristic example of a class (i) reaction is the formation of ozone

$$\mathbf{O} + \mathbf{O}_2 + \mathbf{N}_2 \longrightarrow \mathbf{O}_3 + \mathbf{N}_2 \, ,$$

where the nitrogen molecule removes energy in order to allow for reaching a bound state of ozone [233]. The typical class (ii) reaction is the oxidation of nitrogen oxide with molecular oxygen [230]

$$2\,\mathbf{NO} + \mathbf{O}_2 \longrightarrow 2\,\mathbf{NO}_2 \, .$$

Although nitric oxide oxidation by oxygen is considered as the prototype of a termolecular reaction two competitive two step mechanism involving only bimolecular collisions are discussed:

$$\mathbf{NO} + \mathbf{NO} \rightleftharpoons (\mathbf{NO})_2\,,\ (\mathbf{NO})_2 + \mathbf{O}_2 \rightarrow 2\,\mathbf{NO}_2 \quad \text{or}$$

$$\mathbf{NO} + \mathbf{O}_2 \rightleftharpoons \mathbf{NO}_3\,,\ \mathbf{NO}_3 + \mathbf{NO} \rightarrow 2\,\mathbf{NO}_2\ .$$

A comparison of the data for all three mechanistic variants of the reaction are found in the review [279].

There may also be, however, special situations where approximations of complicated processes by termolecular events is justified. One example is a set of three coupled reactions with four reactant molecules [103, pp. 359-361] where is was shown that $\pi_{\boldsymbol{\mu}}(t, \mathrm{dt})$ is essentially linear in dt.

The last class of reaction to be considered here is no proper chemical reaction but an influx of material into the reactor. It is often denoted as a the zeroth order reaction (4.1a):

$$* \longrightarrow \mathbf{S_A}\ . \tag{4.46}$$

Here, the definition of the influx and the *efficient mixing* or homogeneity condition is essential, because it guarantees that the number of molecules entering the homogeneous system is a constant and does not depend on dt.

4.3 Examples of chemical reactions

In this section we shall present exact solutions of the chemical master equation for examples from three classes of chemical reactions: zeromolecular in form of the flow in a reactor, monomolecular, and bimolecular. Molecularity of a reaction refers to the number of molecules in the reaction complex and in most cases the molecularity is also reflected by the chemical rate law of reaction kinetics in form of the reaction order. In particular, we distinguish fist order and second order kinetics, which is typically observed with monomolecular and bimolecular reactions, respectively.

4.3.1 The flow reactor

The flow reactor is introduced as an experimental device that allows for investigations of systems off thermodynamic equilibrium. The establishment of a stationary state or the *flow equilibrium* in a flow reactor (CFSTR or CSTR: continuous flow stirred tank reactor; figure 4.7) is a suitable case study for the illustration of the search for a solution of a birth-and-death master equation. At the same time the non-reactive flow of a single compound represents the simplest conceivable process in such a reactor. The stock solution contains **A** at the concentration $[\mathbf{A}]_{\text{influx}} = \hat{a} = \bar{a}\,[\text{mole}\cdot\text{l}^{-1}]$. The influx concentration $\hat{a}$ is equal to the stationary concentration $\bar{a}$, because no reaction is assumed to take place in the reactor. The flow is measured by means of the flow rate $r\,[\text{l}\cdot\text{sec}^{-1}]$: This implies an influx of $\bar{a}\cdot r\,[\text{mole}\cdot\text{sec}^{-1}]$ of **A** into the reactor, instantaneous mixing with the content of the reactor, and an outflux of the mixture in the reactor at the same flow rate r.[21] The reactor has a volume of $V\,[\text{l}]$ and thus we have a mean residence time of $\tau_R = V\cdot r^{-1}\,[\text{sec}]$ of a volume element dV in the reactor.

In- and outflux of compound **A** into and from the reactor are modeled by two formal elementary steps or pseudo-reactions

$$\begin{aligned} \star &\longrightarrow \mathbf{A} \\ \mathbf{A} &\longrightarrow \oslash\ . \end{aligned} \tag{4.47}$$

In chemical kinetics the differential equations are almost always formulated in molecular concentrations. For the stochastic treatment, however, we replace concentrations by the numbers of particles, $n = a\cdot V\cdot N_{\text{L}}$ with $n\in\mathbb{N}^0$ and N_{L} being Avogadro's number the number of particles per mole.

The particle number of **A** in the reactor is a stochastic variable with the probability $P_n(t) = P\big(\mathcal{N}(t) = n\big)$. The time derivative of the probability

[21] The assumption of equal influx and outflux rate is required because we are dealing with a flow reactor of constant volume V (CSTR, figure 4.7).

distribution is described by means of the master equation

$$\frac{\partial P_n(t)}{\partial t} \;=\; r\left(\bar{n}\,P_{n-1}(t) + (n{+}1)\,P_{n+1}(t) - (\bar{n}{+}n)\,P_n(t)\right);\; n \in \mathrm{N}^0\;. \quad (4.48)$$

Equation (4.48) describes a birth-and-death process with $w_n^+ = r\bar{n}$ and $w_n^- = rn$. Thus we have a constant birth rate and a death rate which is proportional to n. Solutions of the master equation can be found in text books listing stochastic processes with known solutions, for example [108]. Here we shall derive the solution by means of probability generating functions as introduced in subsection 2.2.1, equation (2.24) in order to illustrate one particularly powerful approach:

$$g(s,t) \;=\; \sum_{n=0}^{\infty} P_n(t)\,s^n\;. \quad (2.24')$$

Sometimes the initial state is included in the notation: $g_{n_0}(s,t)$ implies $P_n(0) = \delta_{n,n_0}$. Partial derivatives with respect to time t and the dummy variable s are readily computed:

$$\frac{\partial g(s,t)}{\partial t} \;=\; \sum_{n=0}^{\infty} \frac{\partial P_n(t)}{\partial t}\cdot s^n \;=$$

$$=\; r\sum_{n=0}^{\infty}\left(\bar{n}\,P_{n-1}(t) \;+\; (n+1)\,P_{n+1}(t) \;-\; (\bar{n}+n)\,P_n(t)\right)s^n \quad \text{and}$$

$$\frac{\partial g(s,t)}{\partial s} \;=\; \sum_{n=0}^{\infty} n\,P_n(t)\,s^{n-1}\;.$$

Proper collection of terms and arrangement of summations – by taking into account: $w_0^- = 0$ – yields

$$\frac{\partial g(s,t)}{\partial t} \;=\; r\bar{n}\sum_{n=0}^{\infty}\left(P_{n-1}(t) \;-\; P_n(t)\right)s^n \;+\; r\sum_{n=0}^{\infty}\left((n+1)\,P_{n+1}(t) \;-\; n\,P_n(t)\right)s^n\,.$$

Evaluation of the four infinite sums

$$\sum\nolimits_{n=0}^{\infty} P_{n-1}(t)\,s^n \;=\; s\sum\nolimits_{n=0}^{\infty} P_{n-1}(t)\,s^{n-1} \;=\; s\,g(s,t)\,,$$

$$\sum\nolimits_{n=0}^{\infty} P_n(t)\,s^n \;=\; g(s,t)\,,$$

$$\sum\nolimits_{n=0}^{\infty} (n+1)\,P_{n+1}(t)\,s^n \;=\; \frac{\partial g(s,t)}{\partial t}\,, \quad \text{and}$$

$$\sum\nolimits_{n=0}^{\infty} n\,P_n(t)\,s^n \;=\; s\sum\nolimits_{n=0}^{\infty} n\,P_n(t)\,s^{n-1} \;=\; s\,\frac{\partial g(s,t)}{\partial t}\;,$$

and regrouping of terms yields a linear partial differential equation of first order

$$\frac{\partial g(s,t)}{\partial t} \;=\; r\Big(\bar{n}(s-1)\,g(s,t) \;-\; (s-1)\,\frac{\partial g(s,t)}{\partial s}\Big)\;. \quad (4.49)$$

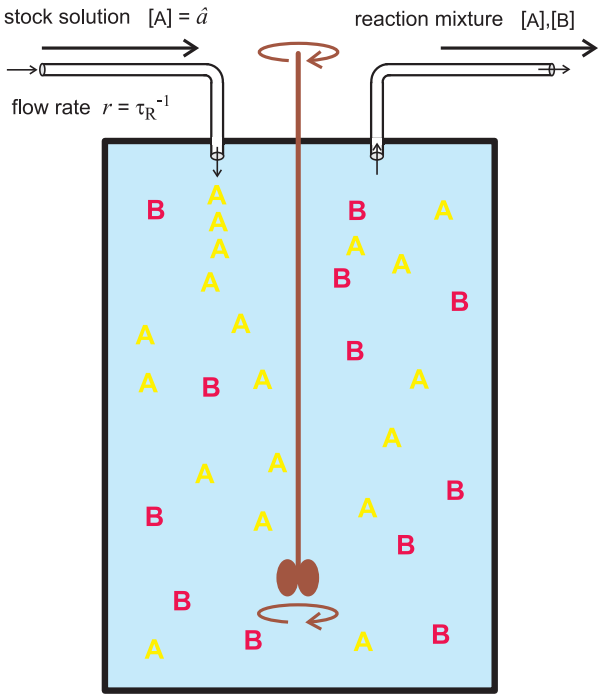

Fig. 4.7 The flow reactor. The reactor shown in the sketch is a device for experimental and theoretical chemical reaction kinetics, which is used to carry out chemical reactions in an open system. The stock solution contains materials, for example **A** at the concentration $[\mathbf{A}]_{\mathrm{influx}} = \hat{a}$, which are usually consumed during the reaction to be studied. The reaction mixture is stirred in order to guarantee a spatially homogeneous reaction medium. Constant volume implies an outflux from the reactor that compensates precisely the influx. The flow rate r is equivalent to the inverse mean residence time of solution in the reactor multiplied by the reactor volume, $\tau_R^{-1} \cdot V = r$. The reactor shown here is commonly called continuously stirred tank reactor (CSTR).

The partial differential equation (PDE) is solved through consecutive substitutions

$$\phi(s,t) \;=\; g(s,t)\,\exp(-\bar{n}\,s) \quad\longrightarrow\quad \frac{\partial\phi(s,t)}{\partial t} \;=\; -r(s-1)\,\frac{\partial\phi(s,t)}{\partial s}\;,$$

$$s-1 \;=\; e^{\rho} \text{ and } \psi(\rho,t) \;=\; \phi(s,t) \quad\longrightarrow\quad \frac{\partial\psi(\rho,t)}{\partial t} \;+\; r\,\frac{\partial\psi(\rho,t)}{\partial\rho} \;=\; 0\;.$$

Computation of the characteristic manifold is equivalent to solving the ordinary differential equation (ODE) $r\,\mathrm{d}t \;=\; -\mathrm{d}\rho$. We find: $rt - \rho = C$ where C is the integration constant. The general solution of the PDE is an arbitrary

function of the combined variable $rt - \rho$:

$$\psi(\rho,t) \;=\; f\big(\exp(-rt+\rho)\big)\cdot e^{-\bar{n}} \quad \text{and} \quad \phi(s,t) \;=\; f\Big(\big(s-1\big)\,e^{-rt}\Big)\cdot e^{-\bar{n}}\;,$$

and the probability generating function

$$g(s,t) \;=\; f\Big(\big(s-1\big)\,e^{-rt}\big)\Big)\cdot\exp\big((s-1)\bar{n}\big)\;.$$

Normalization of probabilities (for $s = 1$) requires $g(1,t) = 1$ and hence $f(0) = 1$. The initial conditions as expressed by the conditional probability $P(n,0|n_0,0) = P_n(0) = \delta_{n,n_0}$ leads to the final expression

$$\begin{aligned} g(s,0) \;&= f(s-1)\,\cdot\exp\big((s-1)\bar{n}\big) \;=\; s^{n_0}\;, \\ f(\zeta) \;&=\; (\zeta+1)^{n_0}\cdot\exp(-\zeta\bar{n}) \;\;\text{with}\;\; \zeta=(s-1)\,e^{-rt}\;, \\ g(s,t) \;&=\; \Big(1+(s-1)\,e^{-rt}\Big)^{n_0}\cdot\exp\big(-\bar{n}(s-1)\,e^{-rt}\big)\cdot\exp\big(\bar{n}(s-1)\big) \;= \\ &= \Big(1+(s-1)\,e^{-rt}\Big)^{n_0}\cdot\exp\big\{-\bar{n}(s-1)\,(1-e^{-rt})\big\}\;. \end{aligned} \tag{4.50}$$

From the generating function we compute with somewhat tedious but straightforward algebra the probability distribution

$$P_n(t) \;=\; \sum_{k=0}^{\min\{n_0,n\}} \binom{n_0}{k}\,\bar{n}^{n-k}\cdot\frac{e^{-krt}\left(1-e^{-rt}\right)^{n_0+n-2k}}{(n-k)!}\cdot e^{-\bar{n}\left(1-e^{-rt}\right)} \tag{4.51}$$

with $n, n_0, \bar{n} \in \mathbb{N}^0$. In the limit $t \to \infty$ we obtain a non-vanishing contribution to the stationary probability only from the first term, $k = 0$, and find

$$\lim_{t\to\infty}\; P_n(t) \;=\; \frac{\bar{n}^n}{n!}\,\exp(-\bar{n})\;.$$

This is a Poissonian distribution with parameter and expectation value $\alpha = \bar{n}$. The Poissonian distribution has also a variance which is numerically identical with the expectation value, $\sigma^2(\mathcal{N}_A) = E(\mathcal{N}_A) = \bar{n}$, and thus the distribution of particle numbers fulfils the $\sqrt{N}$-law at the stationary state.

The time dependent probability distribution allows to compute the expectation value and the variance of the particle number as a function of time

$$\begin{aligned} E\big(\mathcal{N}(t)\big) \;&=\; \bar{n} \;+\; (n_0-\bar{n})\cdot e^{-rt}\;, \\ \sigma^2\big(\mathcal{N}(t)\big) \;&=\; \big(\bar{n} \;+\; n_0\cdot e^{-rt}\big)\cdot\big(1 \;-\; e^{-rt}\big)\;. \end{aligned} \tag{4.52}$$

As expected the expectation value apparently coincides with the solution curve of the deterministic differential equation

$$\frac{\mathrm{d}n}{\mathrm{dt}} \;=\; w_n^+ \;-\; w_n^- \;=\; r\;(\bar{n}-n)\;, \tag{4.53}$$

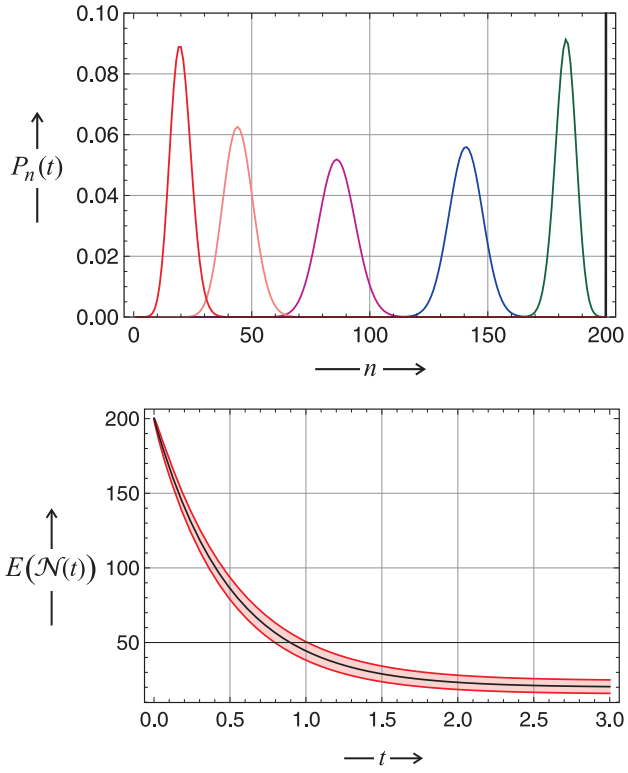

Fig. 4.8 Establishment of the flow equilibrium in the CSTR. The upper part shows the evolution of the probability density, $P_n(t)$, of the number of molecules of a compound **A** which flows through a reactor of the type illustrated in figure 4.7. The initially infinitely sharp density becomes broader with time until the variance reaches its maximum and then sharpens again until it reaches stationarity. The stationary density is a Poissonian distribution with expectation value and variance, $E(\mathcal{N}) = \sigma^2(\mathcal{N}) = \bar{n}$. In the lower part we show the expectation value $E\big(\mathcal{N}(t)\big)$ in the confidence interval $E \pm \sigma$. Parameters used: $\bar{n} = 20$, $n_0 = 200$, and $V = 1$; sampling times (upper part): $\tau = r \cdot t = 0$ (black), 0.05 (green), 0.2 (blue), 0.5 (violet), 1 (pink), and ∞ (red).

which is of the form

$$n(t) \;=\; \bar{n} \,+\, (n_0 - \bar{n}) \cdot e^{-rt} \;. \qquad (4.53')$$

Since we start from sharp initial densities variance and standard deviation are zero at time $t = 0$. The qualitative time dependence of $\sigma^2\{\mathcal{N}_A(t)\}$, however, depends on the sign of $(n_0 - \bar{n})$:

(i) For $n_0 \leq \bar{n}$ the standard deviation increases monotonously until it reaches the value $\sqrt{\bar{n}}$ in the limit $t \to \infty$, and
(ii) for $n_0 > \bar{n}$ the standard deviation increases until it passes through a maximum at

$$t(\sigma_{\max}) \;=\; \frac{1}{r}\left(\ln 2 \,+\, \ln n_0 \,-\, \ln(n_0 - \bar{n})\right)$$

and approaches the long-time value $\sqrt{\bar{n}}$ from above.

In figure 4.8 we show an example for the evolution of the probability density (4.51). In addition, the figure contains a plot of the expectation value $E\big(\mathcal{N}(t)\big)$ inside the band $E - \sigma < E < E + \sigma$. In case of a normally distributed stochastic variable we find 68.3% of all values within this *confidence interval*. In the interval $E-2\sigma < E < E+2\sigma$ we would find even 95.4% of all stochastic trajectories (2.3.3).

4.3.2 Monomolecular chemical reactions

The reversible mono- or monomolecular chemical reaction can be split into two irreversible elementary reactions

$$\mathbf{A} \quad \xrightarrow{k_1} \quad \mathbf{B} \tag{4.54a}$$

$$\mathbf{A} \quad \xleftarrow{h_1} \quad \mathbf{B}\,, \tag{4.54b}$$

wherein the reaction rate parameters, k_1 and h_1, are called *reaction rate constants*. The reaction rate parameters depend on temperature, pressure, and other environmental factors. At equilibrium the rate of the forward reaction (4.54a) is precisely compensated by the rate of the reverse reaction (4.54b), $k_1 \cdot [\mathbf{A}] = h_1 \cdot [\mathbf{B}]$, leading to the condition for the thermodynamic equilibrium:

$$K \;=\; \frac{k_1}{h_1} \;=\; \frac{[\mathbf{B}]}{[\mathbf{A}]}\,. \tag{4.55}$$

The parameter K is called the *equilibrium constant* that depends on temperature, pressure, and other environmental factors like the reaction rate parameters. In an isolated or in a closed system we have a conservation law:

$$\frac{\mathcal{N}_A(t) \,+\, \mathcal{N}_B(t)}{\Omega \cdot N_{\mathrm{L}}} \;=\; [\mathbf{A}] \,+\, [\mathbf{B}] \;=\; c(t) \;=\; c_0 \;=\; \bar{c} \;=\; \text{constant}\,, \tag{4.56}$$

with c being the total concentration and $\bar{c}$ the corresponding equilibrium value, $\lim_{t\to\infty} c(t) = \bar{c}$.

The two irreversible reactions are characterized by vanishing rate parameters, $\lim h_1 \to 0$ or $\lim k_1 \to 0$, respectively. It is worth mentioning that vanishing rate parameters correspond to an instability in the Gibbs free energy at equilibrium, $\Delta G_0 = -RT \ln K$ and are incompatible with rigorous thermodynamics. Nevertheless, the assumption of irreversibility is a good approximation in cases where equilibria are lying almost completely on the side of reactants or products, respectively.

4.3.2.1 Irreversible monomolecular chemical reaction

We start by discussing the simpler irreversible case,

$$\mathbf{A} \quad \xrightarrow{k} \quad \mathbf{B} \,, \tag{4.54a'}$$

which is can be modeled and analyzed in full analogy to the previous case of the flow equilibrium. Although we are dealing with two molecular species, **A** and **B** the process is described by a single stochastic variable, $\mathcal{N}_A(t)$, since we have $\mathcal{N}_B(t) = n_0 - \mathcal{N}_A(t)$ with $n_0 = n(0)$ being the number of **A** molecules initially present because of the conservation relation (4.56). If a sufficiently small time interval is applied, the irreversible monomolecular reaction is modeled by a single step birth-and-death process with $w_n^+ = 0$ and $w_n^- = kn$.[22] The probability density is defined by $P_n(t) = P(\mathcal{N}_A = n)$ and its time dependence obeys

$$\frac{\partial P_n(t)}{\partial t} \;=\; k\,(n+1)\,P_{n+1}(t) \;-\; k\,n\,P_n(t) \;. \tag{4.57}$$

The master equation (4.57) is solved again by means of the probability generating function,

$$g(s,t) \;=\; \sum_{n=0}^{\infty} P_n(t)\,s^n \,;\; |s| \leq 1 \;,$$

which is determined by the PDE

$$\frac{\partial g(s,t)}{\partial t} \;-\; k\,(1-s)\,\frac{\partial g(s,t)}{\partial s} \;=\; 0.$$

The computation of the characteristic manifold of this PDE is tantamount to solving the ODE

$$k\ \mathrm{d}t \;=\; \frac{\mathrm{d}s}{s-1} \quad \Longrightarrow \quad e^{kt} \;=\; s - 1 \;+\; \text{const} \;.$$

[22] We remark that $w_0^- = 0$ and $w_0^+ = 0$ are fulfilled, which are the conditions for a natural absorbing barrier at $n = 0$ (section 5.2.2.3).

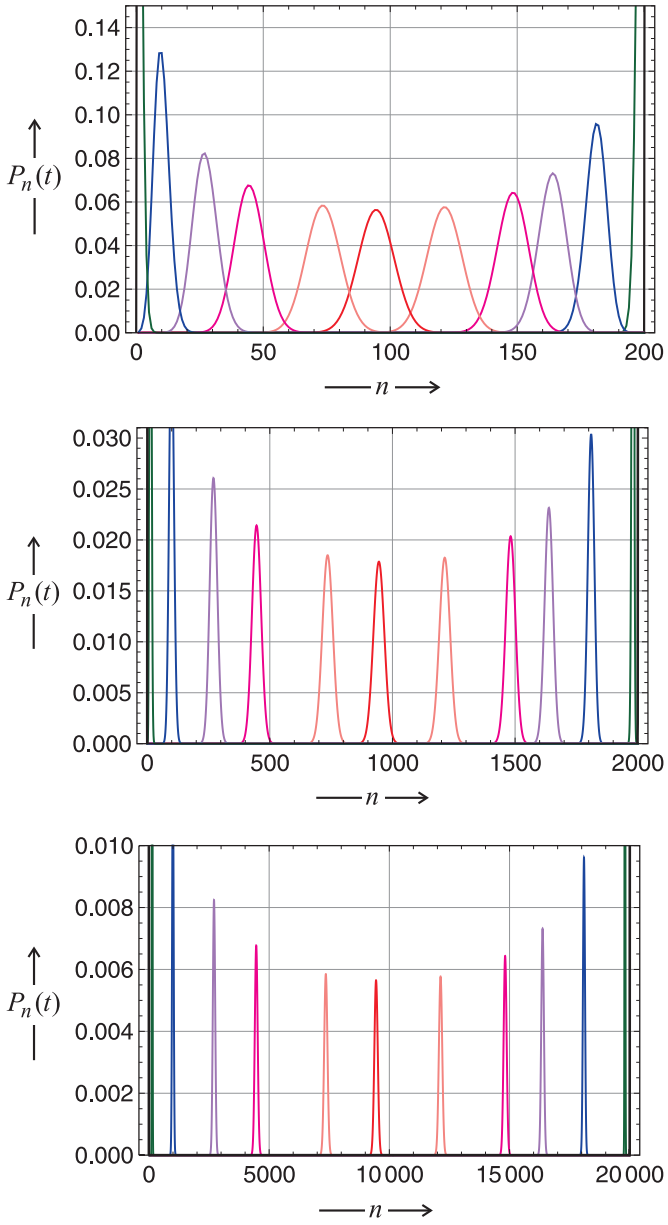

Fig. 4.9 Continued on next page.

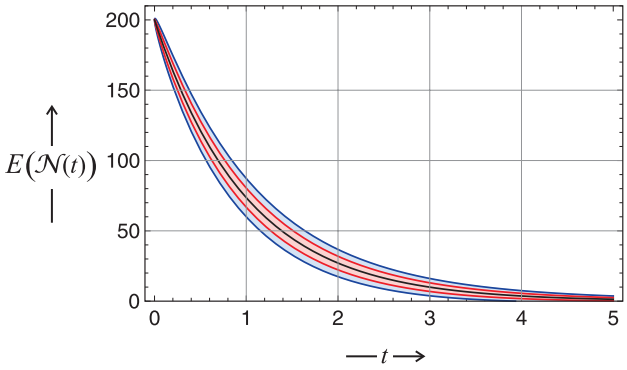

Fig. 4.9 Probability density of an irreversible monomolecular reaction. The three plots on the previous page show the evolution of the probability density, $P_n(t)$, of the number of molecules of a compound **A** which undergo a reaction **A**→**B**. The initially infinitely sharp density $P_n(0) = \delta_{n,n_0}$ becomes broader with time until the variance reaches its maximum at time $t = t_{1/2} = \ln 2/k$ and then sharpens again until it approaches full transformation, $\lim_{t\to\infty} P_n(0) = \delta_{n,0}$. On this page we show the expectation value $E\big(\mathcal{N}_A(t)\big)$ and the confidence intervals $E \pm \sigma$ (68,3%,red) and $\pm 2\sigma$ (95,4%,blue) with $\sigma^2\big(\mathcal{N}_A(t)\big)$ being the variance. Parameters used: $n_0 = 200$, 2000, and 20 000; $k = 1\,[t^{-1}]$; sampling times: 0 (black), 0.01 (green), 0.1 (blue), 0.2 (violet), (0.3) (magenta), 0.5 (pink), 0.75 (red), 1 (pink), 1.5 (magenta), 2 (violet), 3 (blue), and 5 (green).

With $\phi(s,t) = (s-1)\exp(-kt)+\gamma$, $g(s,t) = f(\phi)$, the normalization condition $g(1,t) = 1$, and the boundary condition $g(s,0) = f(\phi)_{t=0} = s^{n_0}$ we find

$$g(s,t) \;=\; \Big(s\cdot e^{-kt} \;+\; 1 \;-\; e^{-kt}\Big)^{n_0} \;. \tag{4.58}$$

This expression is easily expanded in binomial form, which orders with respect to increasing powers of s,

$$g(s,t) = (1-e^{-kt})^{n_0} + \binom{n_0}{1} s e^{-kt}(1-e^{-kt})^{n_0-1} + \binom{n_0}{2} s e^{-2kt}(1-e^{-kt})^{n_0-2} + \\ + \ldots + \binom{n_0}{n_0-1} s^{n_0-1} e^{-(n_0-1)kt}(1-e^{-kt}) + s^{n_0} e^{-n_0 kt} \;.$$

Comparison of coefficients yields the time dependent probability density

$$P_n(t) \;=\; \binom{n_0}{n} \Big(\exp(-kt)\Big)^n \Big(1-\exp(-kt)\Big)^{n_0-n} \;. \tag{4.59}$$

It is straightforward to compute the expectation value of the stochastic variable $\mathcal{N}_A$, which coincides again with the deterministic solution, and its variance

$$\begin{aligned} E\big(\mathcal{N}_A(t)\big) &= n_0\, e^{-kt}\ , \\ \sigma^2\big(\mathcal{N}_A(t)\big) &= n_0\, e^{-kt}\ \big(1 - e^{-kt}\big)\ . \end{aligned} \tag{4.60}$$

The half-life of a population of n_0 particles,

$$t_{1/2}\ :\ E\{\mathcal{N}_A(t)\} = \frac{n_0}{2} = n_0 \cdot e^{-kt_m} \quad \Longrightarrow \quad t_{1/2} = \frac{1}{k}\ln 2\ ,$$

is time of maximum variance or standard deviation, $\mathrm{d}\sigma^2/\,\mathrm{dt} = 0$ or $\mathrm{d}\sigma/\,\mathrm{dt} = 0$, respectively. An example of the time course of the probability density of an irreversible monomolecular reaction is shown in figure 4.9.

4.3.2.2 Reversible monomolecular chemical reaction

The analysis of the irreversible reaction is readily extended to the reversible case (4.54), where we are dealing with a one step birth-and-death process. Again we are dealing with a closed system, the conservation relation $\mathcal{N}_A(t) + \mathcal{N}_B(t) = n_0$ – with n_0 being again the number of molecules of class **A** initially present, $P_n(0) = \delta_{n,n0}$ – holds and the transition probabilities are given by: $w_n^+ = k_2(n_0 - n)$ and $w_n^- = k_1 n$.[23] The master equation is now of the form

$$\begin{aligned} \frac{\partial P_n(t)}{\partial t} &= k_2(n_0 - n + 1)P_{n-1}(t)\ +\ k_1(n+1)P_{n+1}(t) - \\ &\quad -\ \Big(k_1 n + k_2(n_0 - n)\Big)P_n(t)\ . \end{aligned} \tag{4.61}$$

Making use of the probability generating function $g(s,t)$ we derive the PDE

$$\frac{\partial g(s,t)}{\partial t} = \Big(k_1 + (k_2 - k_1)s - k_1 s^2\Big)\frac{\partial g(s,t)}{\partial s}\ +\ n_0\, k_2(s-1)\, g(s,t)\ .$$

The solutions of the PDE are simpler when expressed in terms of parameter combinations, $\kappa = k_1 + k_2$ and $\lambda = k_1/k_2$, and the function $\omega(t) = \lambda \exp(-\kappa t) + 1$:

$$\begin{aligned} g(s,t) &= \left(1\ +\ (s-1)\, e^{-\kappa t}\ -\ \frac{s}{\lambda}\right)^{n_0} \ = \\ &= \left(\frac{\lambda\,(1 - e^{-\kappa t})\ +\ s\,(\lambda e^{-\kappa t} + 1)}{1+\lambda}\right)^{n_0} \ = \\ &= \sum_{n=0}^{n_0}\left(\binom{n_0}{n}\ \left(\lambda e^{-\kappa t} + 1\right)^n\ \Big(\lambda(1 - e^{-\kappa t})\Big)\right)^{n_0 - n}\ \frac{s^n}{(1+\lambda)^{n_0}}\ . \end{aligned}$$

[23] Here we note the existence of barriers at $n = 0$ and $n = n_0$, which are characterized by $w_0^- = 0$, $w_0^+ = k_2 n_0 > 0$ and $w^+ n_0 = 0$, $w_{n_0}^- = k_1 n_0 > 0$, respectively. These equations fulfil the conditions for reflecting barriers (section 5.2.2.3).

The probability density for the reversible reaction is then obtained as

$$P_n(t) \;=\; \binom{n_0}{n} \frac{1}{(1+\lambda)^{n_0}} \left(\lambda e^{-\kappa t}+1\right)^n \left(\lambda(1-e^{-\kappa t})\right)^{n_0-n} . \qquad (4.62)$$

Expectation value and variance of the numbers of molecules are readily computed (with $\omega(t) = \lambda \exp(-\kappa t) + 1$):

$$\begin{aligned} E\big(\mathcal{N}_A(t)\big) &= \frac{n_0}{1+\lambda}\,\omega(t) \;, \\ \sigma^2\big(\mathcal{N}_A(t)\big) &= \frac{n_0\,\omega(t)}{1+\lambda}\left(1-\frac{\omega(t)}{1+\lambda}\right) \;, \end{aligned} \qquad (4.63)$$

and the stationary values are

$$\begin{aligned} \lim_{t\to\infty} E\big(\mathcal{N}_A(t)\big) &= n_0\,\frac{k_2}{k_1+k_2} \;, \\ \lim_{t\to\infty} \sigma^2\big(\mathcal{N}_A(t)\big) &= n_0\,\frac{k_1\,k_2}{(k_1+k_2)^2} \;, \\ \lim_{t\to\infty} \sigma\big(\mathcal{N}_A(t)\big) &= \sqrt{n_0}\,\frac{\sqrt{k_1\,k_2}}{k_1+k_2} \;. \end{aligned} \qquad (4.64)$$

This result shows that the $\sqrt{N}$-law is fulfilled up to a factor that is independent of N: $E/\sigma = \sqrt{n_0}\,k_2/\sqrt{k_1\,k_2}$.

Starting from a sharp distribution, $P_n(0) = \delta_{n,n0}$, the variance increases, may or may not pass through a maximum and eventually reaches the equilibrium value, $\bar{\sigma^2} = k_1 k_2\, n_0/(k_1+k_2)^2$. The time of maximal fluctuations is easily calculated from the condition $\mathrm{d}\sigma^2/\,\mathrm{dt} = 0$ and one obtains

$$t_{\mathrm{var\,max}} \;=\; \frac{1}{k_1+k_2}\,\ln\left(\frac{2\,k_1}{k_1-k_2}\right) \;. \qquad (4.65)$$

Depending on the sign of $(k_1 - k_2)$ the approach towards equilibrium passes a maximum value or not. The maximum is readily detected from the height of the mode of $P_n(t)$ as seen in figure 4.10 where a case with $k_1 > k_2$ is presented.

In order to illustrate fluctuations and their value under equilibrium conditions the Austrian physicist Paul Ehrenfest designed a game called Ehrenfest's urn model [54], which was indeed played in order to verify the $\sqrt{N}$-law. Balls, $2N$ in total, are numbered consecutively, $1, 2, \ldots, 2N$, and distributed arbitrarily over two containers, say **A** and **B**. A lottery machine draws lots, which carry the numbers of the balls. When the number of a ball is drawn, the ball is put from one container into the other. This setup is already sufficient for a simulation of the equilibrium condition. The more balls are in a container, the more likely it is that the number of one of its balls is drawn

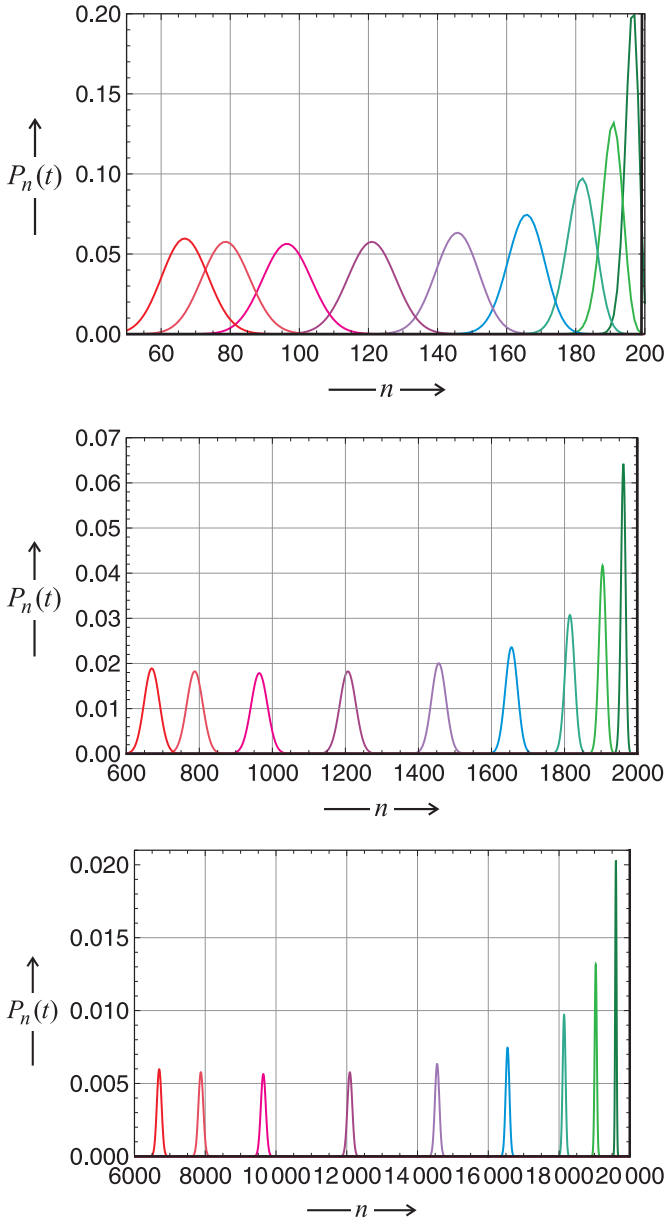

Fig. 4.10 Continued on next page.

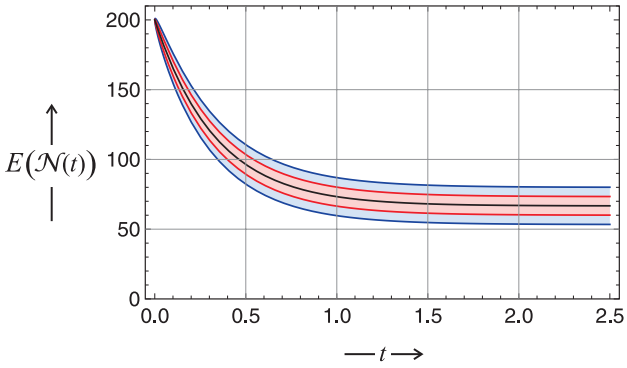

Fig. 4.10 Probability density of a reversible monomolecular reaction The three plots on the previous page show the evolution of the probability density, $P_n(t)$, of the number of molecules of a compound **A** which undergo a reaction **A**⇌**B**. The initially infinitely sharp density $P_n(0) = \delta_{n,n_0}$ becomes broader with time until the variance settles down at the equilibrium value eventually passing a point of maximum variance. On this page we show the expectation value $E\big(\mathcal{N}_A(t)\big)$ and the confidence intervals $E \pm \sigma$ (68,3%,red) and $\pm 2\sigma$ (95,4%,blue) with $\sigma^2\big(\mathcal{N}_A(t)\big)$ being the variance. Parameters used: $n_0 = 200$, 2000, and 20 000; $k_1 = 2\,k_2 = 1\,[t^{-1}]$; sampling times: 0 (black), 0.01 (dark green), 0.025 (green), 0.05 (turquoise), 0.1 (blue), 0.175 (blue violet), 0.3 (purple), 0.5 (magenta), 0.8 (deep pink), 2 (red).

and a transfer occurs into the other container. Just as it occurs with chemical reactions we have self-controlling fluctuations: Whenever a fluctuations becomes large it creates a force for compensation which is proportional to the size of the fluctuation.

4.3.3 Bimolecular chemical reactions

Two classes of bimolecular reactions are accessible to full stochastic analysis:

$$\mathbf{A} + \mathbf{B} \xrightarrow{k} \mathbf{C} \quad \text{and} \tag{4.66a}$$

$$2\,\mathbf{A} \xrightarrow{k} \mathbf{B}\,. \tag{4.66b}$$

Bimolecularity gives rise to nonlinearities in the kinetic differential equations and in the master equations and complicates substantially the analysis of the individual cases. At the same time, these classes of bimolecular equations do not show essential differences in the qualitative behavior compared to the corresponding monomolecular or linear case **A** ⟶ **B** in contrast to autocatalytic

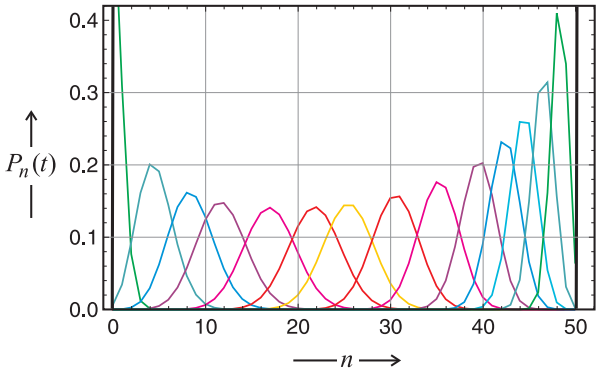

Fig. 4.11 Irreversible bimolecular addition reaction A + B → C. The plot shows the probability distribution $P_n(t) = \text{Prob}\big(\mathcal{N}_C(t) = n\big)$ describing the number of molecules of species **C** as a function of time and calculated by equation (4.72). The initial conditions are chosen to be $\mathcal{N}_A(t) = \delta(a, a_0)$, $\mathcal{N}_B(t) = \delta(b, b_0)$, and $\mathcal{N}_C(t) = \delta(c, 0)$. With increasing time the peak of the distribution moves from left to right. The state $n = \min(a_0, b_0)$ is an absorbing state and hence the long time limit of the system is: $\lim_{t\to\infty} \mathcal{N}_C(t) = \delta\big(n, \min(a_0, b_0)\big)$. Parameters used: $a_0 = 50$, $b_0 = 51$, $k = 0.02\,[t^{-1} \cdot M^{-1}]$; sampling times (upper part): $t = 0$ (black), 0.01 (green), 0.1 (turquoise), 0.2 (blue), 0.3 (violet), 0.5 (magenta), 0.75 (red), 1.0 (yellow), 1.5 (red), 2.25 (magenta), 3.5 (violet), 5.0 (blue), 7.0 (cyan), 11.0 (turquoise), 20.0 (green), and ∞ (black).

processes (section 5.1), which can rise to multiply steady states, and oscillations of concentrations, and deterministic chaos. The following derivations are based upon two publications [205, 139].

4.3.3.1 Addition reaction

In the first example (4.66a) we are dealing with three dependent stochastic variables $\mathcal{N}_A(t)$, $\mathcal{N}_B(t)$, and $\mathcal{N}_C(t)$. Following McQuarrie *et al.* we define the probability $P_n(t) = P\big(\mathcal{N}_A(t) = n\big)$ and apply the standard initial condition $P_n(0) = \delta_{n,n_0}$, $P(\mathcal{N}_B(0) = b) = \delta_{b,b_0}$, and $P(\mathcal{N}_C(0) = c) = \delta_{c,0}$. Accordingly, we have from the laws of stoichiometry $\mathcal{N}_B(t) = b_0 - n_0 + \mathcal{N}_A(t)$ and $\mathcal{N}_C(t) = n_0 - \mathcal{N}_A(t)$. For simplicity we denote $b_0 - n_0 = \Delta_0$. Then the master equation for the chemical reaction is of the form

$$\frac{\partial P_n(t)}{\partial t} \;=\; k\,(n+1)\,(\Delta_0 + n + 1)\,P_{n+1}(t) \;-\; k\,n\,(\Delta_0 + n)\,P_n(t)\;. \quad (4.66a')$$

We remark that the birth and death rates are no longer linear in n. The corresponding PDE for the generating function is readily calculated

$$\frac{\partial g(s,t)}{\partial t} \;=\; k\,(\Delta_0+1)(1-s)\frac{\partial g(s,t)}{\partial s} \;+k\,s(1-s)\,\frac{\partial^2 g(s,t)}{\partial s^2}\;. \qquad (4.67)$$

The derivation of solutions or this PDE is quite demanding. It can be achieved by separation of variables:

$$g(s,t) \;=\; \sum_{m=0}^{\infty} A_m\, Z_m(s)\, T_m(t)\;. \qquad (4.68)$$

We dispense from details and list only the coefficients and functions of the solution:

$$\begin{aligned} A_m \;&=\; (-1)^m\,\frac{(2m+\Delta_0)\Gamma(m+\Delta_0)\Gamma(n_0+1)\Gamma(n_0+\Delta_0+1)}{\Gamma(m+1)\Gamma(\Delta_0+1)\Gamma(n_0-m+1)\Gamma(n_0+\Delta_0+m+1)}\;,\\ Z_m(s) \;&=\; J_m(\Delta_0,\Delta_0+1,s)\;,\quad \text{and}\\ T_m(t) \;&=\; \exp\bigl(-m(m+\Delta_0)\,kt\bigr)\;. \end{aligned}$$

Herein, Γ represents the conventional *gamma function* with the definition $\Gamma(x+1)=x\Gamma(x)$, and $J(p,q,s)$ are the Jacobi polynomials named after the German mathematician Carl Jacobi [1, ch.22, pp.773-802], which are solutions of the differential equation

$$s(1-s)\frac{\mathrm{d}^2 J_n(p,q,s)}{\mathrm{ds}^2} + \Bigl(q-(p+1)s\Bigr)\frac{\mathrm{d}J_n(p,q,s)}{\mathrm{ds}} + n(n+p)\,J_n(p,q,s) \;=\; 0\;.$$

These polynomials fulfil the following conditions:

$$\frac{\mathrm{d}J_n(p,q,s)}{\mathrm{ds}} \;=\; -\,\frac{n(n+p)}{s}\,J_{n-1}(p+2,q+1,s) \quad \text{and}$$

$$\int_0^1 s^{q-1}(1-s)^{p-q}J_n(p,q,s)\,J_\ell(p,q,s)\,ds \;=\; \frac{n!\Bigl(\Gamma(q)\Bigr)^2\Gamma(n+p-q+1)}{(2n+p)\Gamma(n+p)\Gamma(n+q)}\,\delta_{\ell,n}\;.$$

At the relevant value of the dummy variable, $s=1$, we differentiate twice and find:

$$\left(\frac{\partial g(s,t)}{\partial s}\right)_{s=1} = \sum_{m=1}^{n_0}\frac{(2m+\Delta_0)\Gamma(n_0+1)\Gamma(n_0+\Delta_0+1)}{\Gamma(n_0-m+1)\Gamma(n_0+\Delta_0+m+1)}\;T_m(t)\,, \qquad (4.69)$$

$$\left(\frac{\partial^2 g(s,t)}{\partial s^2}\right)_{s=1} =$$

$$= \sum_{m=2}^{n_0}\frac{(m-1)(m+\Delta_0+1)(2m+\Delta_0)\Gamma(n_0+1)\Gamma(n_0+\Delta_0+1)}{\Gamma(n_0-m+1)\Gamma(n_0-\Delta_0+m+1)}\;T_m(t) \qquad (4.70)$$

from which we obtain expectation value and variance according to subsection 2.2.1

$$E\big(\mathcal{N}_A(t)\big) \;=\; \left(\frac{\partial g(s,t)}{\partial s}\right)_{s=1} \quad \text{and}$$

$$\sigma^2\big(\mathcal{N}_A(t)\big) \;=\; \left(\frac{\partial^2 g(s,t)}{\partial s^2}\right)_{s=1} + \left(\frac{\partial g(s,t)}{\partial s}\right)_{s=1} - \left(\left(\frac{\partial g(s,t)}{\partial s}\right)_{s=1}\right)^2 . \quad (2.25')$$

As we see in the current example and we shall see in the next subsubsection, bimolecularity complicates the solution of the chemical master equations substantially and makes it quite sophisticated. We dispense here from the detailed expressions but provide the results for the special case of vast excess of one reaction partner, $|\Delta_0| \gg n_0 > 1$, which is known as *pseudo first order condition* or *concentration buffering*. Then, the sums can be approximated well be the first terms and we find (with $k' = \Delta_0 k$):

$$\left(\frac{\partial g(s,t)}{\partial s}\right)_{s=1} \;\approx\; n_0\,\frac{\Delta_0+2}{n_0+\Delta_0+1}\,e^{-(\Delta_0+1)kt} \;\approx\; n_0\,e^{-k't} \quad \text{and}$$

$$\left(\frac{\partial^2 g(s,t)}{\partial s^2}\right)_{s=1} \;\approx\; n_0\,(n_0-1)\,e^{-2\,k't} \;,$$

and we obtain finally,

$$\begin{aligned} E\big(\mathcal{N}_A(t)\big) \;&=\; n_0\,e^{-k't} \quad \text{and} \\ \sigma^2\big(\mathcal{N}_A(t)\big) \;&=\; n_0\,e^{-k't}\,\left(1-e^{-k't}\right) \;, \end{aligned} \quad (4.71)$$

which is essentially the same result as obtained for the irreversible first order reaction.

For the calculation of the probability density we make use of a slightly different definition of the stochastic variables and use $\mathcal{N}_C(t)$ counting the number of molecules **C** in the system: $P_n(t) = P\big(\mathcal{N}_C(t) = n\big)$. With the initial condition $P_n(0) = \delta(n,0)$ and the upper limit of n, $\lim_{t\to\infty} P_n(t) = c$ with $c = \min\{a_0, b_0\}$ where a_0 and b_0 are the sharply defined numbers of **A** and **B** molecules initially present ($\mathcal{N}_A(0) = a_0$, $\mathcal{N}_B(0) = b_0$), we have

$$\sum_{n=0}^{c} P_n(t) = 1 \;\; \text{and thus} \;\; P_n(t) = 0 \;\forall\; (n \notin [0,c], n \in \mathbb{Z})$$

and the master equation is now of the form

$$\begin{aligned} \frac{\partial P_n(t)}{\partial t} \;&=\; k\,\big(a_0-(n-1)\big)\big(b_0-(n-1)\big)P_{n-1}(t) - \\ &\quad -\; k\,(a_0-n)(b_0-n)P_n(t) \;. \end{aligned} \quad (4.66a")$$

In order to solve the master equation (4.66a") the probability distribution $P_n(t)$ is Laplace transformed in order to obtain a set of pure difference equation from the master equation being a set of differential-difference equation

$$q_n(s) \;=\; \int_0^\infty \exp(-\,s\cdot t)\,P_n(t)\,\mathrm{dt}$$

and with the initial condition $P_n(0) = \delta(n,0)$ we obtain

$$\begin{aligned} -1 \,+\, s\,q_0(s) \;&=\; -\,k\,a_0\,b_0\,q_0(s)\;, \\ s\,q_n(s) \;&=\; k\,\big(a_0-(n-1)\big)\big(b_0-(n-1)\big)\,q_{n-1}(s) - \\ &\quad -\,k\,(a_0-n)(b_0-n)\,q_n(s)\,,\;\; 1\le n\le c\;. \end{aligned}$$

Successive iteration yields the solutions in terms of the functions $q_n(s)$

$$q_n(s) \;=\; \binom{a_0}{n}\binom{b_0}{n}(n!)^2k^n\prod_{j=0}^{n}\frac{1}{s+k(a_0-j)(b_0-j)}\,,\;\; 0\le n\le c$$

and after converting the product into partial fractions and inverse transformation one finds the result

$$\begin{aligned} P_n(t) \;=\; (-1)^n\binom{a_0}{n}\binom{b_0}{n}\sum_{j=0}^{n}(-1)^j\left(1+\frac{n-j}{a_0+b_0-n-j}\right)\times \\ \times\,\binom{n}{j}\binom{a_0+b_0-j}{n}^{-1}e^{-k(a_0-j)(b_0-j)t}\;. \end{aligned} \tag{4.72}$$

An illustrative example is shown in figure 4.11. The difference between the irreversible reactions monomolecular conversion and the bimolecular addition reaction (figure 4.9) is indeed not spectacular.

4.3.3.2 Dimerization reaction

When the dimerization reaction (4.66b) is modeled by means of a master equation [205] we have to take into account that two molecules **A** vanish at a time, and an individual jump involves always $\Delta n = 2$:

$$\frac{\partial P_n(t)}{\partial t} \;=\; \frac{1}{2}\,k\,(n+2)(n+1)\,P_{n+2}(t) \;-\; \frac{1}{2}\,k\,n(n-1)\,P_n(t)\;, \tag{4.66b'}$$

which gives rise to the following PDE for the probability generating function

$$\frac{\partial g(s,t)}{\partial t} \;=\; \frac{k}{2}\,(1-s^2)\,\frac{\partial^2 g(s,t)}{\partial s^2}\;. \tag{4.73}$$

The analysis of this PDE is more involved than it might look at a first glance. Nevertheless, an exact solution similar to (4.68) is available:

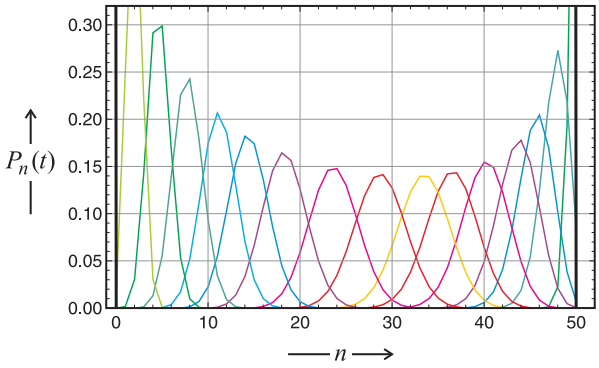

Fig. 4.12 Irreversible dimerization reaction $2\mathbf{A} \rightarrow \mathbf{C}$. The plot shows the probability distribution $P_n(t) = \text{Prob}\left(\mathcal{N}_A(t) = n\right)$ describing the number of molecules of species **C** as a function of time and calculated by equation (4.77). The number of molecules **C** is given by the distribution $P_m(t) = \text{Prob}\left(\mathcal{N}_C(t) = m\right)$. The initial conditions are chosen to be $\mathcal{N}_A(t) = \delta(n, a_0)$, and $\mathcal{N}_C(t) = \delta(m, 0)$ and hence we have $n + 2m = a_0$. With increasing time the peak of the distribution moves from right to left. The state $n = 0$ is an absorbing state and hence the long time limit of the system is: $\lim_{t\to\infty} \mathcal{N}_A(t) = \delta(n, 0)$ $\lim_{t\to\infty} \mathcal{N}_C(t) = \delta(m, a_0/2)$. Parameters used: $a_0 = 100$ and $k = 0.02[t^{-1} \cdot M^{-1}]$; sampling times (upper part): $t = 0$ (black), 0.01 (green), 0.1 (turquoise), 0.2 (blue), 0.3 (violet), 0.5 (magenta), 0.75 (red), 1.0 (yellow), 1.5 (red), 2.25 (magenta), 3.5 (violet), 5.0 (blue), 7.0 (cyan), 11.0 (turquoise), 20.0 (green), 50.0 (chartreuse), and ∞ (black).

$$g(s,t) \;=\; \sum_{m=0}^{\infty} A_m \, C_m^{-\frac{1}{2}}(s) \, T_m(t) \;, \tag{4.74}$$

wherein the parameters and functions are defined by

$$A_m \;=\; \frac{1-2m}{2^m} \cdot \frac{\Gamma(n_0+1)\,\Gamma[(n_0-m+1)/2]}{\Gamma(n_0-m+1)\,\Gamma[(n_0+m+1)/2]} \;,$$

$$C_m^{-\frac{1}{2}}(s) \;:\; (1-s^2)\,\frac{\mathrm{d}^2 C_m^{-\frac{1}{2}}(s)}{\mathrm{ds}^2} \;+\; m(m-1)\,C_m^{-\frac{1}{2}}(s) \;=\; 0 \;,$$

$$T_m(t) \;=\; \exp\{-\frac{1}{2}\,k\,m(m-1)\,t\} \;.$$

The functions $C_m^{-\frac{1}{2}}(s)$ are ultraspherical or Gegenbauer polynomials named after the German mathematician Leopold Gegenbauer [1, ch.22, pp.773-802]. They are solution of the differential equation shown above and belong to the family of hypergeometric functions. It is straightforward to write down expressions for the expectation values and the variance of the stochastic variable $\mathcal{N}_A(t)$ (μ stands for an integer running index, $\mu \in \mathbb{N}$):

$$
\begin{aligned}
E\big(\mathcal{N}_A(t)\big) &= -\sum_{m=2\mu=2}^{2\lfloor\frac{n_0}{2}\rfloor} A_m\, T_m(t) \quad \text{and} \\
\sigma^2\big(\mathcal{N}_A(t)\big) &= -\sum_{m=2\mu=2}^{2\lfloor\frac{n_0}{2}\rfloor} \Big(\frac{1}{2}(m^2-m+2)\, A_m\, T_m(t) \,+\, A_m^2\, T_m^2(t)\Big)\,.
\end{aligned}
\tag{4.75}
$$

In order to obtain concrete results these expressions can be readily evaluated numerically.

There is one interesting detail in the deterministic version of the dimerization reaction. It is conventionally modeled by the differential equation (4.76a), which can be solved readily. The correct ansatz, however, would be (4.76b) for which we have also an exact solution (with [**A**]=$a(t)$ and $a(0) = a_0$):

$$
-\frac{\mathrm{d}a}{\mathrm{d}t} \;=\; k\,a^2 \quad\Longrightarrow\quad a(t) \;=\; \frac{a_0}{1+a_0\,kt} \quad \text{and} \tag{4.76a}
$$

$$
-\frac{\mathrm{d}a}{\mathrm{d}t} \;=\; k\,a(a-1) \quad\Longrightarrow\quad a(t) \;=\; \frac{a_0}{a_0+(1-a_0)e^{-kt}}\,. \tag{4.76b}
$$

The expectation value of the stochastic solution lies always between the solution curves (4.76a) and (4.76b). An illustrative example is shown in figure 4.12.

As the previous subsection 4.3.3.1 we consider also a solution of the master equation by means of a Laplace transformation [139]. Since we are dealing with a step size of two molecules **A** converted into one molecule **C**, the master equation is defined only for odd or only for even numbers of molecules **A**. For an initial number of $2a_0$ molecules and a probability $P_2n(t) = P\big(\mathcal{N}_A(t) = 2n\big)$ we have for the initial conditions $\mathcal{N}_A(0) = 2a_0$, $\mathcal{N}_C(0) = 0$ and the condition that all probabilities outside the interval $[0, 2a_0]$ as well as the odd probabilities P_{2n-1} $(n = 1, \ldots, 2a_0 - 1)$ vanish

$$
\frac{\partial P_{2n}(t)}{\partial t} \;=\; -\frac{1}{2}\,k\,(2n)(2n-1)P_{2n}(t) \,+\, \frac{1}{2}\,k\,(2n+2)(2n+1)P_{2n+2}(t) \tag{4.66b"}
$$

The probability distribution $P_{2n}(t)$ is derived as in the previous subsection by Laplace transformation

$$
q_{2y}(s) \;=\; \int_0^\infty \exp(-\,s\cdot t)\, P_{2y}(t)\;\mathrm{d}t
$$

yielding the set of difference equations

$$
\begin{aligned}
-1 \,+\, s\, q_{2a_0}(s) \;&=\; -\frac{1}{2}\, k\,(2a_0)(2a_0-1)\, q_{2a_0}(s)\,, \\
s\, q_{2n}(s) \;&=\; -\frac{1}{2}\, k\,(2n)(2n-1)\, q_{2n}(s) + \\
&\qquad + \frac{1}{2}\, k\,(2n+2)(2n+1)\, q_{2n+2}(s)\,,\;\; 0 \le y \le a_0 - 1\;,
\end{aligned}
$$

which again can be solved by successive iteration. It is straightforward to calculate first the Laplace transform for 2μ, the number of molecules of species **A** that have reacted to yield **C**: $2\mu = 2(a_0 - m)$ with $m = [\mathbf{C}]$ and $0 \le m \le a_0$:

$$
q_{2(a_0-m)}(s) \;=\; \left(\frac{k}{2}\right)^m \binom{2a_0}{2m} (2m)! \prod_{j=1}^{m} \left(s + \frac{k}{2}(2(a_0-j))\cdot(2(a_0-j)-1)\right)^{-1}\;,
$$

and a somewhat tedious but straightforward exercise in algebra yields the inverse Laplace transform:

$$
\begin{aligned}
P_{2(a_0-m)}(t) \;&=\; (-1)^m \,\frac{a_0!\,(2a_0-1)!!}{(a-m)!\,(2a_0-2m-1)} \times \\
&\qquad \times \sum_{j=0}^{m} (-1)^j \,\frac{(4a_0-4j-1)(4a_0-2m-2j-3)!!}{j!(m-j)!(4a_0-2j-1)!!} \times \\
&\qquad \times\; e^{-k\,(a_0-j)\cdot\left(2(a_0-j)-1\right)\,t}\;.
\end{aligned}
$$

The substitution $i = a_0 - j$ leads to

$$
\begin{aligned}
P_{2(a_0-m)}(t) \;&=\; (-1)^m \,\frac{a_0!\,(2a_0-1)!!}{(a-m)!\,(2a_0-2m-1)} \times \\
&\qquad \times \sum_{i=a_0-m}^{a_0} (-1)^{a_0-i} \,\frac{(4i-1)(2a_0-2m+2i-3)!!}{(a_0-i)!(a_0-i+m)!(2a_0+2i-1)!!} \times \\
&\qquad \times\; e^{-k\,2i\cdot\left(2i-1\right)\,t}\;.
\end{aligned}
$$

Setting now $n = a_0 - m$ in accord with the definition of m we obtain the final result

$$
\begin{aligned}
P_{2n}(t) \;&=\; (-1)^n \,\frac{a_0!(2a_0-1)!!}{n!(2n-1)!!} \times \\
&\qquad \times \sum_{i=1}^{n} (-1)^i \,\frac{(4i-1)(2n+2i-3)!!}{n!(2n-1)!!} \;\times\; e^{-k\,i(2i-1)t}\;.
\end{aligned}
\tag{4.77}
$$

The results are illustrated be means of a numerical example in figure 4.12.

4.4 Stochastic chemical reaction networks

Stochastic chemical reaction networks (SCRNs) therefore are studied mainly by means of computer simulation based on algorithms that converge to solutions, which are exact within the frame of chemical master equations (section 4.7).

4.4.1 Reaction network modeling

Without considering fluctuations reaction networks are commonly modeled by differential equations, ODEs in well mixed homogeneous solution and PDEs in case spatial patterning is of interest.

4.5 Fluctuations and single molecules techniques

The rapid advancements of molecular spectroscopy with respect to signal intensity and temporal within the second half of the twentieth century became the basis for entirely new developments. We menton here two of them as examples.: (i) correlation spectroscopy and (ii) fluorescence spectroscopy at single molecule resolution.

4.6 Scaling and size expansions

Master equations encounter serious limitations with respect to solvability when particle numbers become large whereas Fokker-Planck and stochastic differential equations are much easier to handle and accessible to upscaling. In this section we shall discuss ways to relate master equations to Fokker-Planck equations. In particular, we shall try to solve master equations through approximation methods based on expansions in parameters to be still defined. Straightforward is the expansion of the master equation in a Taylor series with the jump moments as coefficients. Truncation after the second term yields a Fokker-Planck equation. It is important to note that every diffusion process can be approximated by a jump process but the reverse is not true: There are master equations for which no approximation by a Fokker-Planck equation exists. A particularly useful expansion technique based on system sizes has been introduced by Nicholas van Kampen [285, 286]. It can be used to calculate fluctuations without handling full population sizes.

4.6.1 From master to Fokker-Planck equations

The typical example that has been discussed already previously is the random walk (section 3.2.3.6), where the master equation becomes a Fokker-Planck equation in the limit of infinitely small step size. During this transition the jumps must become simultaneously smaller and more probable and this can be taken care by a *scaling assumption*, which is encapsulated by a parameter δ: The average step size is proportional to δ and so is the variance of the step size[24] and thus decreases with δ whereas the jump probabilities increase as δ becomes smaller. In the random walk example we had a step size of $l = l_0\,\delta$ and a probability $\vartheta = \vartheta_0/\delta^2$.

Here we perform first a general transition from the master equation to the Fokker-Planck equation and then illustrate by means of examples. Following [93, pp. 273-274] we rewrite the elements of the transition matrix by introducing a new variable $y = (z - x - A(x)\delta)/\sqrt{\delta}$, where we denote the general drift term by $A(x)$. For the jump probability we write

$$\begin{aligned} W_\delta(z\,|x) \;&=\; \delta^{-3/2}\,\phi(y,x) \;\;\text{with} \\ &\int \mathrm{d}y\;\phi(y,x) \;=\; Q \;\;\text{and}\;\; \int \mathrm{d}y\;y\;\phi(y,x) \;=\; 0\;, \end{aligned} \tag{4.78}$$

where the function $\phi(y,x)$ is given by the concrete example to be studied. Now we define the first three terms for an expansion in jump moments (4.85),

[24] This is automatically fulfilled when the steps follow a Poisson distribution.

$$\begin{aligned}
\alpha_0(x) &\equiv \int \mathrm{d}z\, W_\delta(z|x) = \frac{Q}{\delta} \\
\alpha_1(x) &\equiv \int \mathrm{d}z\,(z-x)\, W_\delta(z|x) = A(x)\,Q \\
\alpha_2(x) &\equiv \int \mathrm{d}z\,(z-x)^2\, W_\delta(z|x) = \int \mathrm{d}y\, y^2\, \phi(y,x)\ ,
\end{aligned} \tag{4.79}$$

and assume that the function $\phi(y,x)$ vanishes sufficiently fast as $y \to \infty$ in order to guarantee that

$$\lim_{\delta\to 0} W_\delta(z|x) = \lim_{y\to\infty} \left(\left(\frac{x}{z-x} \right)^3 \phi(y,x) \right) = 0 \quad \text{for} \quad z \neq x\ .$$

Next we choose some twice differentiable function $f(z)$, carry out a procedure that is very similar to the derivation of the differential Chapman-Kolmogorov equation in section 3.2.2 and find

$$\lim_{\delta\to 0} \left\langle \frac{\partial f(z)}{\partial t} \right\rangle = \left\langle \alpha_1(z)\, \frac{\partial f(z)}{\partial z} + \frac{1}{2}\, \alpha_2(z)\, \frac{\partial^2 f(z)}{\partial z^2} \right\rangle\ .$$

This result has the consequence that in the limit $\delta \to 0$ the master equation

$$\frac{\partial P(x)}{\partial t} = \int \mathrm{d}z \left(W(x|z)\, P(z) - W(z|x)\, P(x) \right) \tag{4.80a}$$

becomes the Fokker-Planck equation

$$\frac{\partial P(x)}{\partial t} = -\frac{\partial}{\partial x}\Big(\alpha_1\, P(x) \Big) + \frac{1}{2}\, \frac{\partial^2}{\partial x^2}\Big(\alpha_2\, P(x) \Big)\ . \tag{4.80b}$$

Accordingly, one can always construct a Fokker-Planck limit for the master equation if the requirements imposed by the three α-functions (4.79) are met. In case these criteria are not fulfilled, there is no approximation possible. The approximation is illustrated now by means of three examples.

Random walk. Based on the notation introduced in subsection 3.2.3.6 we find for $x = n \cdot l$:

$$W(x|z) = \vartheta\,(\delta_{z,x-l} + \delta_{z,x+l}) \implies \alpha_0(x) = 2\vartheta,\ \alpha_1(x) = 0,\ \alpha_2(x) = 2\,l^2\,\vartheta\,.$$

With $\delta = l^2$ and $D = l^2\vartheta$ we obtain the familiar stochastic diffusion equation

$$\frac{\partial P(x,t)}{\partial t} = D\, \frac{\partial^2 P(x,t)}{\partial x^2}\ . \tag{4.81}$$

The final result obtained is exactly the same as in section 3.2.3.6, although we used a much simpler intuitive procedure instead of the transformation (4.78).

Poisson process. With the notation used in section 3.2.3.5 – except α is to be replaced by ϑ – and $x = n \cdot l$ we find:

$$W(x|z) \;=\; \vartheta\, \delta_{z,x+l} \;\Longrightarrow\; \alpha_0(x) = \vartheta,\, \alpha_1(x) = l\,\vartheta,\, \alpha_2(x) = l^2\,\vartheta\,.$$

In this case there is no way to define l and ϑ as functions of δ such that $\alpha_1(x)$ and $\alpha_2(x)$ remain finite in the limit $l \to 0$. Applying, for example, the model assumption made for the one-dimensional random walk we find $l = l_0\,\delta$ and $\vartheta = \vartheta_0/\delta^2$, and hence $\lim_{\delta\to 0} l\vartheta = \infty$. Accordingly, there is no Fokker-Planck limit for the Poisson process.

General approximation of diffusion by birth-and death master equations. We begin with a master equation of the class

$$W_\delta(z|x) \;=\; \left(\frac{A(x)}{2\delta} + \frac{B(x)}{2\delta^2}\right)\,\delta_{z,x+\delta} \;+\; \left(-\frac{A(x)}{2\delta} + \frac{B(x)}{2\delta^2}\right)\,\delta_{z,x-\delta}\;, \quad (4.82)$$

where $W_\delta(z|x)$ is positive for sufficiently small δ. Under the assumption that this is fulfilled for the entire range of interest for x, the process takes place on a range of x that is composed of integer multiples of δ.[25] In the limit $\delta \to 0$ the birth and death master equation is converted into a Fokker-Planck equation with

$$\begin{aligned} &\alpha_0(x) = B(x)\,/\delta^2,\, \alpha_1(x) = A(x),\, \alpha_2(x) = B(x) \;\text{ and} \\ &\lim_{\delta\to 0} W_\delta(z\,|x) = 0 \text{ for } z \neq x\,. \end{aligned} \quad (4.83)$$

Although $\alpha_0(x)$ diverges with $1/\delta^2$ in contrast to (4.79) – where we prescribed the required $1/\delta$ behavior – and the imagination of jumps converging smoothly into a continuous distribution is no longer valid, there exists a limiting Fokker-Planck equation, because the behavior of $\alpha_0(x)$ is irrelevant

$$\frac{\partial P(x,t)}{\partial t} \;=\; -\,\frac{\partial}{\partial x}\Big(A(x)\,P(x,t)\Big) \;+\; \frac{1}{2}\,\frac{\partial^2}{\partial x^2}\Big(B(x)\,P(x,t)\Big)\;. \quad (4.84)$$

Equation (4.83) provides a tool for the simulation of a diffusion process by an approximating birth-and-death process. The method, however, fails for $B(x) = 0$ for all possible ranges of x since then $W_\delta(z,x)$ does not fulfil the criterion of being nonnegative.

[25] We remark that the scaling relations (4.78) and (4.82) not the same but both lead to a Fokker-Planck equation.

4.6.2 Kramers-Moyal expansion

A general expansion of master equations has been proposed by the two physicists Hendrik Anthony Kramers and José Enrique Moyal, which is a kind of Taylor expansion of the integral representation of the master equation

$$\frac{\partial P(x,t)}{\partial t} \;=\; \int \mathrm{d}z \left(W(x|z,t)\, P(z,t) \;-\; W(z|x,t)\, P(x,t) \right) \tag{4.80a}$$

in jump moments (4.85). A comprehensive presentation dealing with different ways to derive the series expansion of the Fokker-Planck equation is found in [250, pp. 63-76]. Starting point is the transition probability from the probability density at time t to the probability density at time $t + \Delta t$:

$$P(x, t + \Delta t) \;=\; \int \mathrm{d}x' W(x, t + \Delta t|x', t)\, P(x', t) \;.$$

In order to derive an expression for the differential ∂P the transition probability $W(x, t + \Delta t|x', t)$ must be known for small Δt at least. In addition, we assume known jump moments

$$\begin{aligned} \alpha_n(x', t, \Delta t) \;&=\; \left\langle \Big(\mathcal{X}(t + \Delta t) - \mathcal{X}(t) \Big)^n \right\rangle \Big|_{\mathcal{X}(t)=x'} = \\ &=\; \int \mathrm{d}x\, (x - x')^n\, W(x, t + \Delta t|x', t) \;. \end{aligned} \tag{4.85}$$

Here $\mathcal{X}(t) = x'$ implies that the random variable $\mathcal{X}(t)$ adopts the sharp value x' at time t. Now we introduce $\Delta x = x - x'$ into the integrand in equation (4.80a) and expand in a Taylor series[26]

$$\begin{aligned} &W(x, t + \Delta t|x', t)\, P(x', t) \;= \\ &\quad=\; W(x + \Delta x - \Delta x, t + \Delta t|x - \Delta x, t)\, P(x - \Delta x, t) \;= \\ &\quad=\; \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n} \Big(W(x + \Delta x, t + \Delta t|x, t)\, P(x - \Delta x, t) \Big) \;. \end{aligned}$$

Insertion into (4.80a) and integration over $\mathrm{d}x' = -\mathrm{d}(\Delta x)$ yields

$$\begin{aligned} P(x, t + \Delta t) \;-\; P(x, t) \;&=\; \frac{\partial P(x,t)}{\partial t}\, \Delta t \;+\; \mathcal{O}(\Delta t^2) \;= \\ &=\; \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n}\, \alpha_n(x, t, \Delta t)\, P(x, t) \;, \end{aligned}$$

[26] The Taylor series is named after the English mathematician Brook Taylor who invented the calculus of finite differences in 1715.

and Taylor expansion of the jump moments and truncation after the linear term yields the desired final result:

$$\frac{\alpha_n(x,t,\Delta t)}{n!} \;=\; \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!}\,\Theta_k^{(n)} \quad \text{with} \quad \Theta_k^{(n)} \;=\; \frac{1}{n!}\,\frac{\partial^k \alpha_n}{\partial \Delta t^k}\;.$$

Since $D_0^{(n)}$ has to vanish because the transition probability has the initial value $W(x,t|x-\Delta x,t)=\delta(\Delta x)$ we find,

$$\frac{\alpha_n(x,t,\Delta t)}{n!} \;=\; \Theta_1^{(n)}\,\Delta t \;+\; \mathcal{O}(\Delta t^2)\;,$$

with the linear term being the only nonzero coefficient, and accordingly we can drop the subscript, $\Theta^{(n)} \equiv \Theta_1^{(n)}$. Eventually, we find for the expansion of the master equation

$$\frac{\partial P(x,t)}{\partial t} \;=\sum_{n=1}^{\infty}(-1)^n \frac{\partial^n}{\partial x^n}\Big(\Theta^{(n)}\,P(x,t)\Big)\;.$$

We remark that the above given derivation corresponds to a forward stochastic process and accordingly there exists also a backward Kramers-Moyal expansion.

Assuming explicit time independence of the transition matrix and the jump moments we obtain the conventional form of the Kramers-Moyal expansion

$$\begin{aligned}\frac{\partial P(x,t)}{\partial t} \;&=\sum_{n=1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial x^n}\Big(\alpha_n(x)\,P(x,t)\Big) \quad \text{with}\\ \alpha_n(x) \;&=\; \int_{-\infty}^{\infty}(z-x)^n\;W(x,z-x)\,\mathrm{d}z\;.\end{aligned} \tag{4.86}$$

In case the Kramers-Moyal expansion is terminated at the second term the result is a Fokker-Planck equation of the form (4.80b):

$$\frac{\partial P(x)}{\partial t} \;=\; -\,\frac{\partial}{\partial x}\Big(\alpha_1(x)\,P(x)\Big) \;+\; \frac{1}{2}\,\frac{\partial^2}{\partial x^2}\Big(\alpha_2(x)\,P(x)\Big)\;. \tag{4.80b}$$

The two jump moments represent the conventional drift and diffusion terms: $\alpha_1(x) \equiv A(x)$ and $\alpha_2(x) \equiv B(x)$. The major difference between the two equations (4.84) and (4.80b) consists in the fact that (4.84) has been derived for one-step birth and death processes whereas (4.80b) is generally valid.

4.6.3 Small noise expansion

For large particle numbers thermal noise fulfilling a $\sqrt{N}$-law my be very small and advantage of this fact can be made in *small noise expansions* of stochastic differential equations and (SDEs) Fokker-Planck. A small noise SDE can be written as:

$$\mathrm{d}x \;=\; a(x)\;\mathrm{dt} \;+\; \varepsilon\, b(x)\;\mathrm{d}W(t)\;, \tag{4.87a}$$

where the solution is assumed to be of the form

$$x_\varepsilon(t) \;=\; x_0(t) \;+\; \varepsilon\, x_1(t) \;+\; \varepsilon^2\, x_2(t) \;+\; \ldots \tag{4.87b}$$

Solutions can be derived term by term and $x_0(t)$, for example, is the solution of the deterministic differential equation, $\mathrm{d}x = a(x)\mathrm{d}x$ with the initial condition $x_0(0) = c_0$.

In the small noise limit suitable Fokker-Planck equation is of the form

$$\frac{\partial P(x,t)}{\partial t} \;=\; -\frac{\partial}{\partial x}\Big(A(x)\,P(x,t)\Big) \;+\; \frac{1}{2}\,\varepsilon^2\,\frac{\partial^2}{\partial x^2}\Big(B(x)\,P(x,t)\Big)\;, \tag{4.88a}$$

where variable and probability density are scaled

$$\xi \;=\; \frac{x - x_0(t)}{\varepsilon} \;\;\text{and}\;\; P_\varepsilon(\xi,t) \;=\; \varepsilon\, P(x,t|c_0,0)\;, \tag{4.88b}$$

and the probability density is assumed to be of the form

$$P_\varepsilon(\xi,t) \;=\; P_\varepsilon^{(0)}(\xi,t) \;+\; \varepsilon\, P_\varepsilon^{(1)}(\xi,t) \;+\; \varepsilon^2\, P_\varepsilon^{(2)}(\xi,t) \;+\; \ldots \tag{4.88c}$$

For both approaches hold two facts: (i) There is no guarantee that the expansion series (4.87b) or (4.88c) converge, and (ii) the explicit calculations involving the series expansions are commonly quite sophisticated [93, pp.169-184].

For the purpose of illustration we consider one special example, the Ornstein-Uhlenbeck process, which is exactly solvable (see section 3.2.3.4). The stochastic differential equations takes on the form

$$\mathrm{d}x \;=\; -\,k\,x\;\mathrm{dt} \;+\; \varepsilon\,\mathrm{d}W(t)\;. \tag{4.89}$$

In the limit $\varepsilon \to 0$ the stochasticity disappears and the resulting ODE remains first order in time and we are dealing with a non-singular limit. The exact solution of (4.89) for the initial condition $x(0) = c_0$ is

$$x_\varepsilon(t) \;=\; c_0\,\exp(-k\,t) \;+\; \varepsilon\int_0^t \exp\Big(-k(t-\tau)\Big)\,\mathrm{d}W(\tau)\;. \tag{4.90}$$

This case is particularly simple since the partitioning according to the series expansion (4.87b) is straightforward

$$x_0(t) \;=\; c_0\,\exp(-k\,t) \;\text{ and }\; x_1(t) \;=\; \int_0^t \exp\Big(-k(t-\tau)\Big)\,\mathrm{d}W(\tau)\;,$$

and $x_0(t)$ is indeed the solution of the ODE obtained by setting $\varepsilon = 0$ in the SDE (4.89).

Now we consider the corresponding Fokker-Planck equation

$$\frac{\partial P(x,t)}{\partial t} \;=\; \frac{\partial}{\partial x}\Big(k\,x\,P(x,t)\Big) \;+\; \frac{1}{2}\,\varepsilon^2\,\frac{\partial^2 P(x,t)}{\partial t^2}\;, \tag{4.91}$$

with the exact solution being a Gaussian with $x_0(t)$ as expectation value

$$\begin{aligned} \langle x(t)\rangle \;&=\; E\big(x(t)\big) \;=\; \alpha(t) \;=\; c_0\,\exp(-k\,t) \;\text{ and} \\ \sigma^2\big(x(t)\big) \;&=\; \varepsilon^2\,\beta(t) \;=\; \varepsilon^2\,\frac{1-\exp(-2k\,t)}{2k}\;, \end{aligned} \tag{4.92}$$

and hence

$$P_\varepsilon(x,t|\,c_0,0) \;=\; \frac{1}{\varepsilon}\,\frac{1}{\sqrt{2\pi\beta(t)}}\,\exp\left(\frac{1}{\varepsilon^2}\,\frac{\big(x-\alpha(t)\big)^2}{2\beta(t)}\right)\;. \tag{4.92'}$$

In the limit $\varepsilon \to 0$ we obtain the expected results for the determinist8ic solution:

$$\lim_{\varepsilon\to 0} P_\varepsilon(x,t|\,c_0,0) \;=\; \delta\big(x-\alpha(t)\big)\;,$$

which is the first order solution of the corresponding SDE and a deterministic trajectory along the path $x(t) = c_0\exp(-kt)$. In the limit $\varepsilon \to 0$ the second order differential equation (4.91) is reduced to a first order equation and this implies a singularity and the requirement to apply singular perturbation theory.

Therefore, the probability density, however, cannot be expanded straightforwardly in a power series in ε, the introduction of a scaled variable is needed before:

$$\xi \;=\; \big(x\,-\,\alpha(t)\big)\,/\,\varepsilon \;\text{ or }\; x \;=\; \alpha(t) \,+\, \varepsilon\,\xi\;.$$

Now we can write down the probability density in ξ in terms of its first and second moments,

$$P_\varepsilon(\xi,t|0,0) \;=\; P_\varepsilon(x,t|c_0,0)\cdot\frac{\mathrm{d}x}{\mathrm{d}\xi} \;=\; \frac{1}{\sqrt{2\pi\beta(t)}}\,\exp\left(-\frac{\xi^2}{2\beta(t)}\right)\;.$$

Scaling has eliminated the singularity as the probability density for ξ does not contain ε: The distribution of the scaled variable ξ is a Gaussian with mean zero and the deviation of x from the deterministic trajectory $\alpha(t)$ is of order ε as ε goes to zero. The coefficient of ε is the random variable ξ. As expected, in the interpretation there is no difference between the Fokker-Planck and the stochastic differential equation.

4.6.4 Size expansion of the master equation

Although quit a few representative examples and model systems can be analyzed by solving the one step birth-and-death master equation exactly (section 4.3), the actual applicability to specific problems of chemical kinetics of this technique is rather limited. In order to apply a chemical master equation to a problem in practice on is commonly dealing with about 10^{12} particles or more. Upscaling discloses one particular problem that is related to size expansion and that becomes virulent in the transition from the master equation to a Fokker-Planck equation. The problem is intimately related to the parameter volume V, which is the best possible estimator of system size in condensed matter. We distinguish two classes of quantities: (i) *intensive quantities* that are independent of system size, and (ii) *extensive quantities* that grow proportional to system size. Examples of intensive properties are temperature, pressure, density, concentrations, and extensive properties are volume, particle numbers, energy, or entropy. In upscaling from say 1000 to 10^{12} particles extensive properties grow by a factor of 10^9 whereas intensive properties remain the same. Some *pairs* of properties – one extensive and one intensive – are of particular importance, for example particle number $\mathcal{N}$ and concentration $c = \mathcal{N}/(V \cdot N_L)$ or mass M and (volume) density $\varrho = M/V$, respectively.

In order to compensate for the lack of generality, approximation methods were developed, which turned out to be particularly illustrative and useful in the limit of sufficiently large particle numbers [286, 287]. The Dutch theoretical physicist Nicholas van Kampen expands the master equation in the inverse square root of some extensive quantity, particle number, mass or volume, which is characteristic of system size and which will be denoted by Ω. In van Kampen's notation,

$$\begin{aligned} a \;\propto\; \Omega \quad &= \text{ extensive variable, and} \\ \alpha \;=\; a/\Omega \;\; &= \text{ intensive variable ,} \end{aligned} \tag{4.93}$$

the limit of interest is a large value of Ω at fixed α, which is tantamount to the transition to a macroscopic system.[27] The transition probabilities are reformulated as

$$W(a|a') \;=\; W(a';\Delta a) \;\text{ with }\; \Delta a = a - a' \;,$$

and scaled according to the assumption

$$W(a|a') \;=\; \Omega\, \psi\Big((\frac{a'}{\Omega}, \Delta a\Big) \;=\; \Omega\, \psi\Big((\alpha', \Delta a\Big) \;.$$

[27] In this section we shall use Greek letters for intensive and Roman letters for extensive variables wherever possible.

The essential trick in the van Kampen expansion is that the size of the jump is expressed in term of an extensive quantity, Δa, whereas the intensive variable α is used for the expression of the dependence on the variable, a'.

The expansion is made now in the new variable z defined by

$$a \;=\; \Omega\,\phi(t) \,+\, \Omega^{1/2}\,z \;\;\text{or}\;\; z \;=\; \Omega^{-1/2}\,a \,-\, \Omega^{1/2}\,\phi(t)\;. \tag{4.94}$$

where the function $\phi(t)$ is still to be determined. The derivative moments $\alpha_n(a)$ are now proportional to the system size Ω and therefore we can scale them accordingly: $\alpha_n(a) = \Omega\,\tilde{\alpha}_n(x)$. In the next step the new variable z is introduced into the Kramers-Moyal expansion (4.86):

$$\begin{aligned}
\frac{\partial P(z,t)}{\partial t} \,-\, \Omega^{1/2}\,\frac{\partial \phi}{\partial t}\,\frac{\partial P(z,t)}{\partial z} \;&=\; \\
&=\; \sum_{n=1}^{\infty} (-1)^n\,\frac{\Omega^{\,1-n/2}}{n!}\,\frac{\partial^n}{\partial z^n}\Big(\tilde{\alpha}_n\big(\phi(t)+\Omega^{-1/2}\,z\big)\,P(z,t)\Big)\;, \\
\frac{\partial P(z,t)}{\partial t} \;&=\; \Omega^{1/2}\cdot\left(\frac{\partial \phi}{\partial t} \,-\, \tilde{\alpha}_1\big(\phi(t)\big)\right)\,\frac{\partial P(z,t)}{\partial z} \,+\, \Omega^0\cdot\Big(\cdots\Big)\ldots\;.
\end{aligned}$$

For general validity of an expansion all terms of a certain order in the expansion parameter must vanish. We make use of this property to define $\phi(t)$ such that the terms of order $\Omega^{1/2}$ are eliminated by demanding

$$\frac{\partial \phi}{\partial t} \;=\; \tilde{\alpha}_1\big(\phi(t)\big)\;. \tag{4.95}$$

This equation is an ODE determining $\phi(t)$ and, of course, it is in full agreement with the deterministic equation for the expectation value of the random variable. Accordingly, $\phi(t)$ is indeed the deterministic part of the solution.

The next step is an expansion of $\tilde{\alpha}_n\big(\phi(t)+\Omega^{-1/2}z\big)$ in $\Omega^{-1/2}$ and reordering of terms yielding

$$\frac{\partial P(z,t)}{\partial t} \;=\; \sum_{m=2}^{\infty}\,\frac{\Omega^{-(m-2)/2}}{m!}\sum_{n=1}^{m}(-1)^n\binom{m}{n}\,\tilde{\alpha}_n^{m-n}\big(\phi(t)\big)\,\frac{\partial^n}{\partial z^n}\Big(z^{m-n}P(z,t)\Big)$$

In taking the limit of large system size Ω all terms vanish except the one with $m = 2$ and we find the result

$$\frac{\partial P(z,t)}{\partial t} \;=\; -\,\tilde{\alpha}_1^{(1)}\big(\phi(t)\big)\,\frac{\partial}{\partial z}\Big(z\,P(z,t)\Big) \,+\, \frac{1}{2}\,\tilde{\alpha}_2\big(\phi(t)\big)\,\frac{\partial^2}{\partial z^2}P(z,t)\;, \tag{4.96}$$

where $\alpha_1^{(1)}$ stands for the linear drift term.

It is straightforward to compare with the result of the Kramers-Moyal expansion (4.86) truncated after two terms:

$$\frac{\partial P(x,t)}{\partial t} \;=\; -\frac{\partial}{\partial x}\Big(\alpha_1(x)P(x,t)\Big) \;+\; \frac{1}{2}\,\frac{\partial^2}{\partial x^2}\Big(\alpha_2(x)P(x,t)\Big)\;.$$

The change of variables $\xi = x/\Omega$ leads to

$$\frac{\partial P(\xi,t)}{\partial t} \;=\; -\frac{\partial}{\partial \xi}\Big(\tilde{\alpha}_1(\xi)P(\xi,t)\Big) \;+\; \frac{1}{2\Omega}\,\frac{\partial^2}{\partial \xi^2}\Big(\tilde{\alpha}_2(\xi)P(\xi,t)\Big)\;.$$

Through application of small noise theory (section 4.6.3) with $\epsilon^2 = \Omega^{-1}$ and using the substitution $\xi = \Omega^{1/2}\big(x-\phi(t)\big)$ one obtains the lowest order Fokker-Planck equation, which is exactly the same as the lowest order approximation in the van Kampen expansion. This result has an important consequence: If we are only interested in the lowest order approximation we may use the Kramers-Moyal equation, which is much easier to derive than the van Kampen equation.

Eventually, we have found a procedure to relate approximately master equations, Fokker-Planck and stochastic differential equations and to close the gap between microscopic stochasticity and macroscopic behavior. It should be stressed, however, that the range of validity of a Fokker-Planck equation derived from a master equation is not independent of the kind of limiting procedure applied. If the transition was made by means of equations (4.78) and (4.79) in the limit $\delta \to 0$, the full nonlinear dependence of $\alpha_1(x)$ and $\alpha_2(x)$ can be seriously analyzed. If, on the other hand, only the small noise approximation is approximately valid than it is appropriate to consider only the linearization of the drift term and individual solutions of this equations are represented by the trajectories of the stochastic equation:

$$\mathrm{d}z \;=\; \tilde{\alpha}_1^{(1)}\big(\phi(t)\big)\,z\;\mathrm{dt} \;+\; \sqrt{\tilde{\alpha}_2\big(\phi(t)\big)}\;\mathrm{d}W(t)\;. \tag{4.97}$$

The choice of the best way of scaling will also depend on the special example to be studied and we close this section by presenting two examples: (i) the flow reactor and (ii) the reversible first order chemical reaction.

Equilibration in the flow reactor. The problem we are considering here is the time dependence of a single chemical substance **A** in a device for performing chemical reactions under controlled conditions as described in section 4.3.1. The concentration of **A** in the solution flowing into the reactor is $\hat{a}$ and it is equal to $\bar{a}$ the concentration of **A** in the reactor after flow equilibrium has been established. The flux in and out of the reactor is controlled by the flow rate r commonly measured in volume/time=[cm^3/sec] and it represents the reciprocal mean residence time of the solution in the reactor: $r = \tau_{\mathrm{R}}^{-1}$. The extensive variables in this case are the numbers of particles of class **A**: $n_{\mathbf{A}} = n = a \cdot V \cdot N_L = a \cdot \Omega$ with $n \in \mathbb{N}^0$. The elements of the transition matrix W are

$$W(n|n') \;=\; r\,\Big(\delta_{n,n'+1}\bar{n} \;+\; \delta_{n,n'-1}n\Big)\;. \tag{4.98}$$

The first term with $n'+1=n$ as the only nonzero contribution describes the increase in the particle number in the reactor through influx,

$$\mathcal{N}(t)=n' \implies \mathcal{N}(t+\mathrm{dt})=n=n'+1\ ,$$

whereas the second term deals with the outflux of a particle **A**,

$$\mathcal{N}(t)=n' \implies \mathcal{N}(t+\mathrm{dt})=n=n'-1\ ,$$

and in both cases we have by simple mass action the probabilities $r\cdot\bar{n}$ and $r\cdot n(t)$, respectively. The reformulation of the elements of the transition matrix leads to

$$W(a';\Delta n) \;=\; \Omega\left(r\,\bar{a}\,\delta_{\Delta n,+1}\;+\;r\,a'\,\delta_{\Delta n,-1}\right)$$

with $\Delta n=n-n'$. Calculation of the first two jump moments yields

$$\alpha_1 \;=\; \sum_{n'=0}^{\infty}(n'-n)\,W(n'|n) \;=\; r(\bar{n}-n) \;=\; \Omega\,r(\bar{a}-a)\ ,$$

$$\alpha_2 \;=\; \sum_{n'=0}^{\infty}(n'-n)^2\,W(n'|n) \;=\; r(\bar{n}+n) \;=\; \Omega\,r(\bar{a}+a)\ ,$$

and the deterministic equation with $\phi(t)=a(t)=n(t)/\Omega$ is of the form

$$\frac{\mathrm{d}a}{\mathrm{dt}} \;=\; r\,(\bar{a}\;-\;a)\;\;\text{and}\;\;a(t) \;=\; \bar{a}\;+\;\big(a(0)-\bar{a}\big)\,e^{-rt}\ .$$

Following the procedure of van Kampen's expansion we define

$$n \;=\; \Omega\,\phi(t)\;+\;\Omega^{1/2}\,z\;\;\text{or}\;\;z \;=\; \Omega^{-1/2}n\;-\;\Omega^{\,1/2}n\,\phi(t) \tag{4.94'}$$

and obtain the Fokker-Planck equation

$$\frac{\partial P(z)}{\partial t} \;=\; r\,\frac{\partial}{\partial z}\Big(z\,P(z)\Big)\;+\;\frac{r}{2}\,\frac{\partial^2}{\partial z^2}\Big(\big(\bar{a}\;+\;a(t)\big)\,P(z)\Big)\ , \tag{4.99}$$

which leads to the expectation value and variance in the scaled variable z:

$$\begin{aligned}\langle z(t)\rangle \;&=\; E\big(z(t)\big) \;=\; z(0)\,e^{-rt}\ ,\\ \sigma^2\big(z(t)\big) \;&=\big(\bar{a}\;+\;a(0)\,e^{-rt}\big)\,(1-e^{-rt})\ .\end{aligned}$$

Transformation into the extensive variable, the particle number n yields

$$\begin{aligned}\langle n(t)\rangle \;&=\; E\big(n(t)\big) \;=\; \bar{n}\;+\;\big(n(0)-\bar{n}\big)\,e^{-rt}\ ,\\ \sigma^2\big(n(t)\big) \;&=\big(\bar{n}\;+\;n(0)\,e^{-rt}\big)\,(1-e^{-rt})\ .\end{aligned} \tag{4.100}$$

The stationary solution of the Fokker-Planck equation is readily calculated

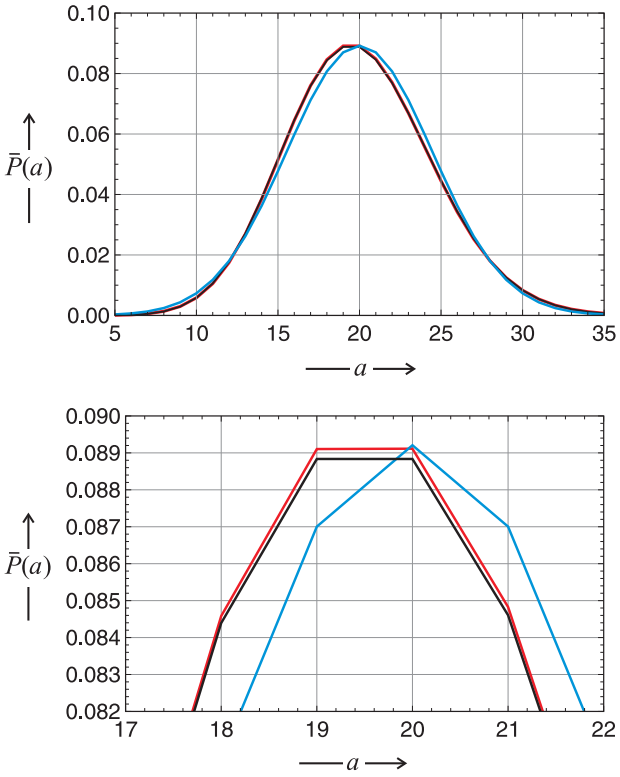

Fig. 4.13 Comparison of expansions of the master equation. The reaction **A**⇋**B** with **B** buffered, $[\mathbf{B}] = b = b_0$, is chosen as example and the exact solution (black) is compared with the results of the Kramers-Moyal expansion (red) and the van Kampen size expansion (blue). Parameter choice: $V = 1$, $k_1 = 2$, $k_2 = 1$, $b = 40$.

$$\bar{P}(z) \;=\; \frac{1}{\sqrt{2\pi\,\bar{a}}}\,\exp\left(-\frac{z^2}{2\bar{a}}\right)$$

and it represents the approximation of the exact stationary Poisson density by means of a Gaussian as mentioned in (2.38):

$$\bar{P}(n) \;=\; \frac{\bar{n}^n}{n!}\,\exp(-\bar{n}) \;\approx\; \frac{1}{\sqrt{2\pi\,\bar{n}}}\,\exp\left(-\frac{(n-\bar{n})^2}{2\bar{n}}\right).$$

The chemical reaction **A**⇋**B**. The transition probabilities for the interval $t' \to t$ of the corresponding single step birth-and-death master equation with $[\mathbf{A}]_t = a(t)$, $[\mathbf{A}]_{t'} = a'$, $[\mathbf{B}] = b_0$, a fixed or buffered concentration, and the reaction rate parameters k_1 and k_2 are:

$$W(a|a') \;=\; \delta_{a,a'+1}\, k_2 b_0 \;+\; \delta_{a,a'-1}\, k_1 a \;.$$

As before we choose the volume of the system times Loschmidt's number, $\Omega = V \cdot N_L$, as size parameter and have: $a = \alpha\Omega$ and $b = \beta\Omega$. This leads to the scaled transition probability,

$$W(\alpha'; \Delta a) \;=\; \Omega\left(k_2\beta\,\delta_{\Delta a,1} \;+\; k_1\alpha'\,\delta_{\Delta a,-1}\right) ,$$

and the first two derivative moments

$$\alpha_1 \;=\; \sum_{(a')}(a'-a)\,W(a'|a) \;=\; k_2 b_0 \;-\; k_1 a \;=\; \Omega(k_2\beta \;-\; k_1\alpha) ,$$

$$\alpha_2 \;=\; \sum_{(a')}(a'-a)^2\,W(a'|a) \;=\; k_2 b_0 \;+\; k_1 a \;=\; \Omega(k_2\beta \;+\; k_1\alpha) .$$

Following the procedure of van Kampen's expansion we define

$$a \;=\; \Omega\,\phi(t) \;+\; \Omega^{1/2}\,z \tag{4.101}$$

and obtain for the deterministic differential equation and its solution:

$$\frac{\mathrm{d}\phi(t)}{\mathrm{dt}} \;=\; k_2\beta \;-\; k_1\,\phi(t) \;\text{ and }\; \phi(t) \;=\; \phi(0)\,e^{-k_1 t} \;+\; \frac{k_2\beta}{k_1}\,(1-e^{-k_1 t}) .$$

The Fokker-Planck equation takes on the form

$$\frac{\partial P(z)}{\partial t} \;=\; k_1\,\frac{\partial}{\partial z}\Big(z\,P(z)\Big) \;+\; \frac{1}{2}\frac{\partial^2}{\partial z^2}\Big(\big(k_2\beta \;+\; k_1\,\phi(t)\big)\,P(z)\Big)$$

The expectation value of z is readily computed to be $\langle z(t)\rangle = z(0)e^{-k_1 t}$. Since the partition of the variable a in equation (4.101) is arbitrary we can assume $z(0) = 0$ – as usual[28] – and find for the variance in z

$$\sigma^2\big(z(t)\big) \;=\; \left(\frac{k_2\beta}{k_1} \;+\; \phi(0)\right)\,(1-e^{-k_1 t})$$

and eventually obtain for the solutions in the macroscopic variable a with $a(0) = \Omega\,\phi(0)$

$$\langle a(t)\rangle \;=\; \Omega\,\phi(t) \;=\; a(0)\,e^{-k_1 t} \;+\; \frac{k_2\,b_0}{k_1}\,(1-e^{-k_1 t}) ,$$

$$\sigma^2\Big(a(t)\Big) \;=\; \Omega\,\sigma^2\big(z(t)\big) \;=\; \left(\frac{k_2\,b_0}{k_1} \;+\; a(0)\right)\,(1-e^{-k_1 t}) .$$

[28] The assumption $z(0) = 0$ implies $z(t) = 0$ and hence the corresponding stochastic variable $\mathcal{Z}(t)$ describes the fluctuations around zero.

Finally, we compare the different stationary state solutions obtained from the van Kampen expansion, $\alpha = k_2 b_0/k_1$,

$$\bar{P}(z) \;=\; \frac{1}{\sqrt{\frac{\pi\alpha}{2}}\left(1+\mathrm{erf}(\sqrt{\frac{\alpha}{2}})\right)}\,\exp\left(-\frac{(z-\alpha)^2}{2\alpha}\right)\;,$$

with those derived from the Kramers-Moyal expansion

$$\bar{P}(a) \;=\; \mathcal{N}\,(k_2\,b_0 \,+\, k_1\,a)^{-1+4k_2\,b_0/k_1}\,e^{-2a}\;,$$

and the exact solution

$$\bar{P}(a) \;=\; \frac{\left(k_2\,b_0/k_1\right)^a\,\exp\!\left(-k_2\,b_0/k_1\right)}{a!} \;=\; \frac{\alpha^a\,e^{-\alpha}}{a!}\;,$$

which is a Poissonian. A comparison of numerical plots is shown in figure 4.13. It is remarkable how well the truncated Kramers-Moyal expansion agrees with the exact probability density. It is easy to understand therefore that it is much more popular than the size expansion, which in addition is also much more sophisticated.

4.6.5 Size expansion of birth-and-death processes

In the previous section (section 4.6.4) we introduced a size expansion for the chemical master equation. Here, we repeat the derivation this technique in the case of a simple birth an death process from biology, the spreading of an epidemic, which is, nevertheless, sufficiently general in order to be transferable to other cases [286, pp.251-258].

Before we discuss the specific example, however, we recall the birth-and-death transition matrix for a single step processes (3.98),

$$W(n|n') = w^+_{n'}\,\delta_{n,n'-1} + w^-_{n'}\,\delta_{n,n'+1}\;,$$

where w^+_n and w^-_n are analytic functions, which are as we shall assume (at least) twice differentiable:

$$\frac{\partial P_n(t)}{\partial t} \;=\; w^+_{n-1}\,P_{n-1}(t) \,+\, w^-_{n+1}\,P_{n+1}(t) \,-\, \left(w^+_n + w^-_n\right)P_n(t)\;.$$

It turns out useful to define a single step difference operator $\widehat{\Theta}$ by

$$\widehat{\Theta}\,f(n) \;=\; f(n+1)\;, \quad \text{and} \quad \widehat{\Theta}^{-1}\,f(n) \;=\; f(n-1)\;. \tag{4.102}$$

Using this operator we can rewrite the master equation in compact form

$$\frac{\partial \mathbf{P}(t)}{\partial t} \;=\; \left\{(\widehat{\Theta}-1)\,w^+_n \,+\, (\widehat{\Theta}^{-1}-1)\,w^-_n\right\}\mathbf{P}(t)\;. \tag{3.98'}$$

The jump moments are now

$$\alpha_p(n) \;=\; (-1)^p\, w_n^+ \;+\; w_n^- \;. \tag{4.29a}$$

We repeat the macroscopic rate equation

$$\frac{\mathrm{d}\langle n\rangle}{\mathrm{dt}} \;=\; -\,w^-_{\langle n\rangle} \;+\; w^+_{\langle n\rangle} \;,$$

find for the coupled equations for expectation value and variance the simpler expressions

$$\frac{\mathrm{d}\langle n\rangle}{\mathrm{dt}} \;=\; w^+_{\langle n\rangle} \;-\; w^-_{\langle n\rangle} \;+\; \frac{1}{2}\left(\frac{\mathrm{d}^2 w^+_{\langle n\rangle}}{\mathrm{dn}^2} - \frac{\mathrm{d}^2 w^-_{\langle n\rangle}}{\mathrm{dn}^2}\right)\sigma_n^2\,, \tag{4.103a}$$

$$\frac{\mathrm{d}\sigma_n^2}{\mathrm{dt}} \;=\; w^+_{\langle n\rangle} \;+\; w^-_{\langle n\rangle} \;+\; 2\left(\frac{\mathrm{d}w^+_{\langle n\rangle}}{\mathrm{dn}} - \frac{\mathrm{d}w^-_{\langle n\rangle}}{\mathrm{dn}}\right)\sigma_n^2\,. \tag{4.103b}$$

With these preliminaries we are in the position to handle the epidemic example by means of the size expansion technique (see [286, pp.251-254] and section 4.6.4).

An epidemic spreads in a population of Ω individuals. We assume that $n(t)$ individuals are already infected. The probability of a new infection is proportional to both, to the number of infected and to the number of uninfected individuals, $w_n^- = \beta\, n(\Omega - n)$. No cure is possible and thus $w_n^+ = 0$. Finally, we have

$$W(n|n') \;=\; \beta\,\delta_{n,n'+1}\,n'\,(\Omega - n')\;,$$

which leads to the master equation

$$\begin{aligned} \frac{\partial P_n(t)}{\partial t} \;&=\; \beta(n-1)(\Omega - n + 1)\,P_{n+1}(t) \;-\; \beta n(\Omega - n)\,P_n(t) \quad \text{or} \\ \frac{\partial \mathbf{P}}{\partial t} \;&=\; \beta\,\left(\widehat{\Theta}^{-1} - 1\right)\,n\,(\Omega - n)\,\mathbf{P}(t)\;. \end{aligned} \tag{4.104}$$

Basic to the expansion is the idea that the density of the stochastic variable $\mathcal{N}$ can be split in a macroscopic part, $\Omega\,\phi(t)$, and fluctuations of the order $\Omega^{1/2}$ around it. As shown in figure 4.14 we assume that $P(n,t)$ is represented by a (relatively) sharp peak located approximately at $\Omega\,\phi(t)$ with a width of order $\Omega^{1/2}$. In other words, we assume that the fluctuations fulfil a $\sqrt{N}$-law, and we make the ansatz

$$n(t) \;=\; \Omega\,\phi(t) \;+\; \Omega^{1/2}\,x(t)\;, \tag{4.105}$$

where x is a new variable describing the fluctuations. The function $\phi(t)$ has to be chosen in accord with the master equation. As said above, $\Omega\,\phi(t)$ is called the *macroscopic part* and $\Omega^{1/2}\,x$ the *fluctuating part* of n. We may refer to

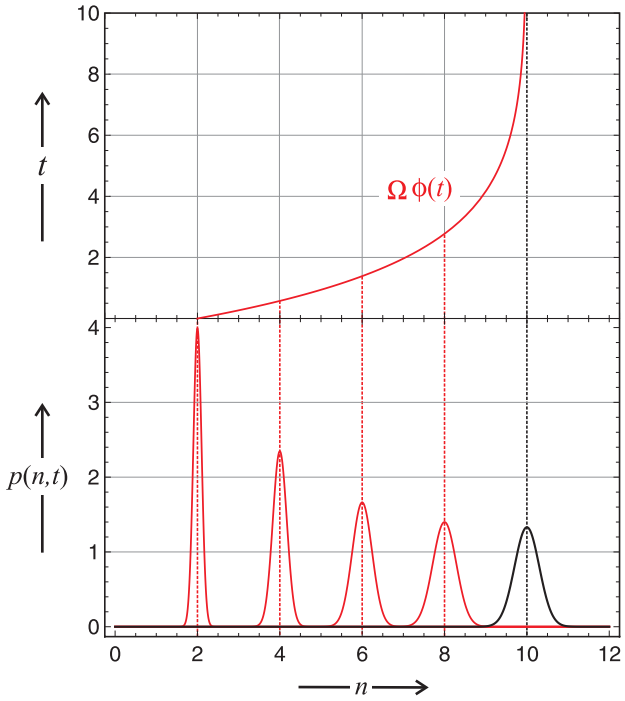

Fig. 4.14 The size expansion of a stochastic variable $\mathcal{N}$. The variable n is partitioned according to into a macroscopic part and the fluctuations around it, $n = \Omega\phi(t) + \Omega^{1/2}x(t)$, wherein Ω is a size parameter, for example the size of the population or the volume of the system. Computations: $\Omega\phi(t) = 5n_0(1 - 0.8e^{-kt})$ with $n_0 = 2$ and $k = 0.5$; $p(n,t) = \Omega^{1/2}x(t) = e^{-(n-\Omega\phi(t))^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$ with $\sigma = 0.1, 0.17, 0.24, 0.285, 0.30$.

the new variables as an Ω *language*. The probability density of n becomes now a probability density $\Pi(x,t)$ of x:

$$\begin{aligned} P(n,t)\,\Delta n &= \Pi(x,t)\,\Delta x\ , \\ \Pi(x,t) &= \Omega^{1/2}\,P\left(\Omega\phi(t) + \Omega^{1/2}x, t\right)\ . \end{aligned} \tag{4.106}$$

Differentiation yields[29]

$$\frac{\partial\Pi}{\partial x} = \Omega^{1/2}\frac{\partial P}{\partial n}\ , \quad \frac{\partial\Pi}{\partial t} = \Omega^{1/2}\left(\Omega\frac{\mathrm{d}\phi}{\mathrm{dt}}\frac{\partial P}{\partial n} + \frac{\partial P}{\partial t}\right)\ ,$$

[29] The somewhat unclear differentiation $\partial P/\partial n$ can be circumvented through direct variation of t by δt and simultaneously of x by $-\Omega^{1/2}\phi(t)\delta t$, which leads to the same final result.

and eventually we obtain

$$\Omega^{1/2}\,\frac{\partial P}{\partial t} \;=\; \frac{\partial \Pi}{\partial t} \;-\; \Omega^{1/2}\,\frac{\mathrm{d}\phi}{\mathrm{dt}}\frac{\partial \Pi}{\partial x}\;. \tag{4.107}$$

Now, the difference operators are also *size expanded* in power series of differential operators

$$\widehat{\Theta} \;=\; 1 \,+\, \Omega^{-1/2}\,\frac{\partial}{\partial x} \,+\, \frac{1}{2}\,\Omega^{-1}\,\frac{\partial^2}{\partial x^2} \,+\, \ldots\;, \tag{4.108a}$$

$$\begin{aligned}
\widehat{\Theta}^{-1} \;&=\; \frac{1}{1 \,+\, \Omega^{-1/2}\,\frac{\partial}{\partial x} \,+\, \frac{1}{2}\,\Omega^{-1}\,\frac{\partial^2}{\partial x^2} \,+\, \ldots} \;= \\
&=\; 1 \,-\, \Omega^{-1/2}\,\frac{\partial}{\partial x} \,-\, \frac{1}{2}\,\Omega^{-1}\,\frac{\partial^2}{\partial x^2} \,+\, \ldots \,+\, \Omega^{-1}\,\frac{\partial^2}{\partial x^2} \,+\ldots \;\approx \\
&\approx\; 1 \,-\, \Omega^{-1/2}\,\frac{\partial}{\partial x} \,+\, \frac{1}{2}\,\Omega^{-1}\,\frac{\partial^2}{\partial x^2}\;,
\end{aligned} \tag{4.108b}$$

$$\widehat{\Theta}^{-1} - 1 \;=\; -\,\Omega^{-1/2}\,\frac{\partial}{\partial x} \,+\, \frac{1}{2}\,\Omega^{-1}\,\frac{\partial^2}{\partial x^2}\;. \tag{4.108c}$$

Insertion of the operator and substitution of the new variables into the master equation (4.104) yields after cancelation of an overall factor $\Omega^{-1/2}$

$$\begin{aligned}
\frac{\partial \Pi}{\partial t} \,-\, \Omega^{1/2}\frac{\mathrm{d}\phi}{\mathrm{dt}}\frac{\partial \Pi}{\partial x} \;=\; \beta\Omega^2 \,&\Big(-\,\Omega^{-1/2}\,\frac{\partial}{\partial x} \,+\, \frac{1}{2}\,\Omega^{-1}\,\frac{\partial^2}{\partial x^2}\Big)\cdot \\
&\cdot\,\Big(\Big(\phi + \Omega^{-1/2}x\Big)\Big(1 - \phi - \Omega^{-1/2}x\Big)\,\Pi(x,t)\Big)\;.
\end{aligned}$$

The right-hand side requires two consecutive differentiations of three factors:

$$\begin{aligned}
&-\,\Omega^{-1/2}\,\frac{\partial}{\partial x}\Big(\Big(\phi + \Omega^{-1/2}x\Big)\cdot\Big(1 - \phi - \Omega^{-1/2}x\Big)\,\Pi(x,t)\Big) \;= \\
=\; &-\,\Big(1 - 2\phi - 2\Omega^{-1/2}x\Big)\,\Pi(x,t) \,-\, \Omega^{1/2}\Big(\phi + \Omega^{-1/2}x\Big)\cdot\Big(1 - \phi - \Omega^{-1/2}x\Big)\,\frac{\partial \Pi}{\partial x}\;, \\
&\frac{1}{2}\,\Omega^{-1}\,\frac{\partial^2}{\partial x^2}\Big\{\Big(\phi + \Omega^{-1/2}x\Big)\cdot\Big(1 - \phi - \Omega^{-1/2}x\Big)\,\Pi(x,t)\Big\} \;= \\
=\; &-\,\Omega^{-3/2}\,\frac{\partial \Pi}{\partial x} \,+\, \frac{1}{2}\Omega^{-1}\Big(\phi + \Omega^{-1/2}x\Big)\cdot\Big(1 - \phi - \Omega^{-1/2}x\Big)\,\frac{\partial^2 \Pi}{\partial x^2}\;.
\end{aligned}$$

For convenience we introduce a new time scale, $\tau = \beta\Omega t$, in order to absorb one factor Ω – and for convenience also the factor β – into the time variable. Collection of terms corresponding to the largest powers in Ω now yields

$$\Omega^{1/2}\ :\quad \frac{\mathrm{d}\phi}{\mathrm{d}\tau}\,\frac{\partial \Pi(x,\tau)}{\partial x} \;=\; \phi\,(1-\phi)\,\frac{\partial \Pi(x,\tau)}{\partial x}\ ,$$

$$\Omega^{0}\quad :\quad \frac{\partial \Pi(x,\tau)}{\partial \tau} \;=\; -\,(1-2\phi)\,\frac{\partial}{\partial x}\Big(x\Pi(x,\tau)\Big) \;+\; \frac{1}{2}\,\phi\,(1-\phi)\,\frac{\partial^2 \Pi(x,\tau)}{\partial x^2}\ .$$

The largest term cancels if

$$\frac{\mathrm{d}\phi}{\mathrm{d}\tau} \;=\; \phi\,(1-\phi)\ , \tag{4.109'}$$

and this yields the differential equation for the macroscopic variable $\phi(t)$, which after transformation back into the original variables leads to the macroscopic rate equation

$$\frac{\mathrm{d}n}{\mathrm{dt}} \;=\; \beta\, n\,(\Omega\, - n)\ . \tag{4.109}$$

Equating the next largest term, the coefficient of Ω^0 to zero results in a linear Fokker-Planck equation with time dependent coefficients $\phi(t)$:

$$\frac{\partial \Pi(x,\tau)}{\partial \tau} \;=\; -\,(1-2\phi)\,\frac{\partial}{\partial x}\Big(x\Pi(x,\tau)\Big) \;+\; \frac{1}{2}\,\phi\,(1-\phi)\,\frac{\partial^2 \Pi(x,\tau)}{\partial x^2}\ . \tag{4.110}$$

Equation (4.110) describes the fluctuations of the random variable $\mathcal{N}(t)$ around the macroscopic part and these fluctuations are of order $\Omega^{1/2}$ as expected and initially assumed.

The strategy for solving the master equation (4.104) is now obvious. At first one determines $\phi(\tau)$ by integrating the differential equation (4.109') with the initial value $\phi(0) = n_0/\Omega$, then one solves the Fokker-Planck equation (4.110) with the initial condition $\Pi(x,0) = \delta(x)$ and finally one obtains the desired probabilities from

$$P(n,t|n_0,0) \;=\; \Omega^{-1/2}\ \Pi\Big(\frac{n-\Omega\phi(\tau)}{\Omega^{1/2}},\,\tau\Big) \tag{4.111}$$

A typical solution is sketched in figure 4.14 and it compares perfectly with the exact solutions for sufficiently large systems (see, for example figures 4.9 and 4.10). Remembering the derivation we remark the terms of relative order $\Omega^{-1/2}$ and smaller have been neglected.

4.7 Numerical simulation of master equations

Almost at the same time when the Feinberg-Horn-Jackson theory was introduced a simulation tool for stochastic chemical reactions was developed by the American physicist and mathematical chemist Daniel Gillespie [101, 102, 104, 106].

History

He conceived and implemented a simple and powerful algorithm for the calculation of single trajectories, and showed later that the chemical master equation and the computation tool can be put on a firm physical and mathematical basis [104]. Meanwhile the Gillespie algorithm became an essential simulation tool in chemistry and biology. Here we present the concept and the implementation of the algorithm and demonstrate the usefulness by means of selected examples.

4.7.1 Basic assumptions

Gillespie's general stochastic model of chemical reaction networks considers a population of M different molecular species, $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M\}$ in a homogeneous medium, which interact through R elementary chemical reactions $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_R\}$ as in the previous deterministic model (section 4.1.3). Again two conditions are assumed to be fulfilled by the system: (i) the container with constant volume V in the sense of a flow reactor (CSTR in figure 4.7) is assumed to be *well mixed* by efficient stirring,[30] and (ii) the system is assumed to be in *thermal equilibrium* at constant temperature T. The primary goals of the simulation are the computation of the time course of the stochastic variables – $\mathcal{X}_k(t)$ being the number of molecules $\mathbf{S}_k$ of species $\mathbf{K}$ at time t – and the description of the evolution of the entire molecular population. The individual computations yield single trajectories, very much in the sense of a single solution of a stochastic differential equation (figure 3.20) or single molecule experiments. Observable results of conventional macroscopic experiments are commonly derived through sampling of trajectories.

For a reaction system involving M species in R reactions the entire population is characterized by an N-dimensional random vector counting numbers of molecules for the various species $\mathbf{S}_k$,

$$\vec{\mathcal{X}}(t) \;=\; \big(\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_M(t)\big)\;. \tag{4.112}$$

The common variables in chemistry are concentrations rather than particle numbers:

[30] As said before the assumption is almost always perfectly fulfilled in the gas phase or and applies also to solution where perfect mixing is still a challenge.

$$\mathbf{x} \;=\; \big(\mathrm{x}_1(t), \mathrm{x}_2(t), \ldots, \mathrm{x}_N(t)\big) \;\; \text{with} \;\; \mathrm{x}_k(t) \;=\; \frac{\mathcal{X}_k(t)}{V \cdot N_L}, \tag{4.113}$$

where the volume V is an appropriate expansion parameter Ω for the system size (section 4.6.4). The chemical master equation and the (probabilistic) rate parameters were derived in section 4.2.2 (see also [104, pp. 407-417]) for reaction channels $\mathbf{R}_{\boldsymbol{\mu}}$ of bimolecular nature

$$\mathbf{S_A} \,+\, \mathbf{S_B} \;\longrightarrow\; \mathbf{S_C} \,+\, \ldots \tag{4.34}$$

like (4.1f,4.1i,4.1j and 4.1k) shown in the list (4.1) by making use of the well-developed collision theory in the vapor phase. The extension to monomolecular reaction channels (section 4.2.2.4) and termolecular reaction channels (section 4.2.2.5) is straightforward. *Zero-molecular* processes like the influx of material into the reactor in the elementary step (4.1a) provide no major problems and reversible reactions, for example (4.54), are handled as two elementary steps, $\mathbf{A} + \mathbf{B} \longrightarrow \mathbf{C} + \mathbf{D}$ and $\mathbf{C} + \mathbf{D} \longrightarrow \mathbf{A} + \mathbf{B}$. As in Feinberg-Horn-Jackson theory (section 4.1.3) we distinguish between *reactant* species – for example $\mathbf{A}$ and $\mathbf{B}$ in equation (4.34) – and *product* species – e.g., $\mathbf{C}$ ... in equation (4.34) – of a reaction $\mathbf{R}_{\boldsymbol{\mu}}$.

4.7.2 Reaction stoichiometry

In section 4.2.2 we succeeded to derive the fundamental fact that for each elementary reaction channel $\mathbf{R}_{\boldsymbol{\mu}}$ with $\boldsymbol{\mu} = 1, \ldots, R$, which is accessible to the molecules of a well-mixed and thermally equilibrated system in the gas phase (or in solution), exists a scalar quantity $\gamma_{\boldsymbol{\mu}}$, which is independent of dt such that [104, p.418]

$$\begin{aligned} \gamma_{\boldsymbol{\mu}}\ \mathrm{dt} \;&= \text{probability that a randomly selected combination of} \\ &\quad\ \mathbf{R}_{\boldsymbol{\mu}} \text{ reactant molecules at time } t \text{ will react accordingly} \\ &\quad\ \text{in the next infinitesimal time interval } [t, t + \mathrm{dt}[\,. \end{aligned} \tag{4.114}$$

The specific probabilistic rate parameter, $\gamma_{\boldsymbol{\mu}}$ is one of three quantities that are required to fully characterize a particular reaction channel $\mathbf{R}_{\boldsymbol{\mu}}$. In addition to $\gamma_{\boldsymbol{\mu}}$ we shall require a function $h_{\boldsymbol{\mu}}(\mathbf{n})$ where the vector $\mathbf{n} = (n_1, \ldots, n_M)^{\mathrm{t}}$ contains the exact numbers of all molecules at time t, $\vec{\mathcal{N}}(t) = \big(\mathcal{N}_1(t), \ldots, \mathcal{N}_M(t)\big)^{\mathrm{t}} = \mathbf{n}(t)$,

$$\begin{aligned} h_{\boldsymbol{\mu}}(\mathbf{n}) \;&\equiv \text{the number of distinct combinations of } \mathbf{R}_{\boldsymbol{\mu}} \text{ reactant} \\ &\quad\ \text{molecules in the system when the numbers of molecules} \\ &\quad\ \mathbf{S}_k \text{ are exactly } n_k \text{ with } k = 1, \ldots, M\ , \end{aligned} \tag{4.115}$$

and an $M \times R$ matrix of integers, $\mathrm{S} = \{s_{k\boldsymbol{\mu}};\, k = 1, \ldots, M,\, \boldsymbol{\mu} = 1, \ldots, R\}$, where

$$s_{k\boldsymbol{\mu}} \equiv \text{the change in the } \mathbf{S}_k \text{ molecular population caused by the occurrence of one } \mathbf{R}_{\boldsymbol{\mu}} \text{ reaction.} \tag{4.116}$$

The functions $h_{\boldsymbol{\mu}}(\mathbf{n})$ and the matrix S are readily deduced by inspection of the algebraic structure of the reaction channels. We illustrate by means of an example:

$$\begin{aligned} \mathbf{R}_1 :\ & \mathbf{S}_1 + \mathbf{S}_2 \longrightarrow \mathbf{S}_3 + \mathbf{S}_4 \ , \\ \mathbf{R}_2 :\ & 2\,\mathbf{S}_1 \longrightarrow \mathbf{S}_1 + \mathbf{S}_5 \ , \quad \text{and} \\ \mathbf{R}_3 :\ & \mathbf{S}_3 \longrightarrow \mathbf{S}_5 \ . \end{aligned} \tag{4.117}$$

The functions $h_{\boldsymbol{\mu}}(\mathbf{n})$ are obtained by simple combinatorics

$$\begin{aligned} h_1(\mathbf{n}) &= n_1\, n_2 \ , \\ h_2(\mathbf{n}) &= n_1\,(n_1 - 1)/2 \ , \quad \text{and} \\ h_3(\mathbf{n}) &= n_3 \ , \end{aligned}$$

and the matrix S is of the form

$$\mathrm{S} = \begin{pmatrix} -1 & -1 & 0 \\ -1 & 0 & 0 \\ +1 & 0 & -1 \\ +1 & 0 & 0 \\ 0 & +1 & +1 \end{pmatrix} \ ,$$

where the rows refer to molecular species, $(\mathbf{S}_1, \mathbf{S}_3, \mathbf{S}_4, \mathbf{S}_5)$, and the columns to individual reactions, $(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3)$. The integers in S reflect the net production of species per elementary reaction: Species $\mathbf{S}_1$ in reaction $\mathbf{R}_2$ has a stoichiometric coefficient -2 on the reactant side sine two molecules are consumed in one elementary step and a coefficient $+1$ on the product side leading to $s_{12} = -2 + 1 = -1$ in the stoichiometric matrix S.

It is worth noticing that the functional form of $h_{\boldsymbol{\mu}}$ is determined exclusive by the reactant side of $\mathbf{R}_{\boldsymbol{\mu}}$. For mass action kinetics there is only one difference between the deterministic and the stochastic expressions: Since the particles are counted exactly in the latter approach we have to use $n(n-1)/2$ instead of $n^2/2$ because $n - 1$ is significantly different from n in small systems. The stoichiometric matrix S refers to the product side of the reaction equations in the sense that products are counted with positive and reactants with negative stoichiometric coefficients and the summation yields the net production of molecular species per one elementary reaction event: $s_{k\boldsymbol{\mu}}$ is the number of molecules $\mathbf{S}_k$ produces by reaction $\mathbf{R}_{\boldsymbol{\mu}}$, these numbers are integers

and negative values indicate the number of molecules, which have disappeared during one reaction. In the forthcoming analysis we shall make use of vectors corresponding to individual reactions $\mathbf{R}_{\boldsymbol{\mu}}$: $\boldsymbol{\nu}_{\boldsymbol{\mu}} = (s_{1\boldsymbol{\mu}}, \ldots, s_{R\boldsymbol{\mu}})^{\mathrm{t}}$.

It is illustrative to consider the relation to conventional deterministic chemical kinetics. If we denote the concentration vector of the molecular species $\mathbf{S}$ by $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{t}}$ and the flux or rate vector by $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_M)^{\mathrm{t}}$ the kinetic equation can be expressed by

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{dt}} \;=\; \mathrm{S} \cdot \boldsymbol{\varphi} \;. \tag{4.118}$$

The individual elements of the flux vector in mass action kinetics are

$$\varphi_{\boldsymbol{\mu}} \;=\; k_{\boldsymbol{\mu}} \prod_{k=1}^{n} x_k^{s_{k\boldsymbol{\mu}}^{(\mathrm{R})}} \quad \text{for} \quad s_{1\boldsymbol{\mu}}^{(\mathrm{R})}\,\mathbf{S}_1 \;+\; s_{2\boldsymbol{\mu}}^{(\mathrm{R})}\,\mathbf{S}_2 \;+\; \ldots \;+\; s_{M\boldsymbol{\mu}}^{(\mathrm{R})}\,\mathbf{S}_N \;\longrightarrow$$

wherein the factors $s_{k\boldsymbol{\mu}}^{(\mathrm{R})}$ are the stoichiometric coefficients on the reactant side of the reaction equations. It is sometimes useful to define analogous factors $s_{k\boldsymbol{\mu}}^{(\mathrm{P})}$ for the product side, both classes of factors can be summarized in matrices $\mathrm{S_R}$ and $\mathrm{S_P}$ and then the stochastic matrix is simply given by the difference $\mathrm{S} = \mathrm{S_P} - \mathrm{S_R}$. We illustrate by means of the model mechanism (4.117) in our example:

$$\mathrm{S_P} \;-\; \mathrm{S_R} \;=\; \begin{pmatrix} 0 & +1 & 0 \\ 0 & 0 & 0 \\ +1 & 0 & 0 \\ +1 & 0 & 0 \\ 0 & +1 & +1 \end{pmatrix} - \begin{pmatrix} +1 & +2 & 0 \\ +1 & 0 & 0 \\ 0 & 0 & +1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & -1 & 0 \\ -1 & 0 & 0 \\ +1 & 0 & -1 \\ +1 & 0 & 0 \\ 0 & +1 & +1 \end{pmatrix} = \mathrm{S}$$

We remark that the entries of $\mathrm{S_R}$ and $\mathrm{S_P}$ are nonnegative integers by definition. The flux $\boldsymbol{\varphi}$ has the same structure as in the stochastic approach, $\gamma_{\boldsymbol{\mu}}$ corresponds to the kinetic rate parameter or rate constant $k_{\boldsymbol{\mu}}$ and the combinatorial function $h_{\boldsymbol{\mu}}$ and the mass action product are identical apart from the simplifications, e.g., $(n-1) \Rightarrow n$, for large particle numbers.

4.7.3 Occurrence of reactions

The probability of occurrence of reaction events within an infinitesimal time interval dt is cast into three theorems:

Theorem 1. If $\vec{\mathcal{X}}(t) = \mathbf{n}$, then the probability that *exactly one* $\mathbf{R}_{\boldsymbol{\mu}}$ will occur in the system within the time interval $[t, t + \mathrm{dt}[$ is equal to

$$\gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})\ \mathrm{dt} \;+\; o(\mathrm{dt})\ ,$$

where $o(\mathrm{dt})$ denotes terms that approach zero with dt faster than dt.

Theorem 2. If $\vec{\mathcal{X}}(t) = \mathbf{n}$, then the probability that *no* reaction will occur within the time interval $[t, t + \mathrm{dt}[$ is equal to

$$1 - \sum_{\boldsymbol{\mu}} \gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})\ \mathrm{dt} + o(\mathrm{dt})\ .$$

Theorem 3. The probability of more than one reaction occurring in the system within the time interval $[t, t + \mathrm{dt}[$ is of order $o(\mathrm{dt})$.

Proofs for all three theorems were derived by Daniel Gillespie and can be found in [104, pp.420,421].

Based on the three theorems an analytical description can be derived for the evolution of the population vector $\vec{\mathcal{X}}(t)$. The initial state of the system at some initial time t_0 is fixed: $\vec{\mathcal{X}}(t_0) = \mathbf{n}_0$. In almost all cases there is no chance to derive an exact solution for the time evolution of the probability function $P(\mathbf{n}, t|\mathbf{n}_0, t_0)$ but a deterministic function for the differential change of the probability for $t \geq t_0$ is readily obtained. We express the probability $P(\mathbf{n}, t+\mathrm{dt}|\mathbf{n}_0, t_0)$ as the sum of the probabilities of several mutually exclusive and collectively exhaustive *routes* from $\vec{\mathcal{X}}(t_0) = \mathbf{n}_0$ to $\vec{\mathcal{X}}(t+\mathrm{dt}) = \mathbf{n}$. These routes are distinguished from one another with respect to the event that happened in the last time interval $[t, t + \mathrm{dt}[$:

$$\begin{aligned} P(\mathbf{n}, t + \mathrm{dt}|\mathbf{n}_0, t_0) \;=\; & P(\mathbf{n}, t|\mathbf{n}_0, t_0) \;\times\; \left(1 - \sum_{\boldsymbol{\mu}=1}^{R} \gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})\,\mathrm{dt} + o(\mathrm{dt})\right) + \\ & + \sum_{\boldsymbol{\mu}=1}^{R} P(\mathbf{n} - \boldsymbol{\nu}_{\boldsymbol{\mu}}, t|\mathbf{n}_0, t_0) \;\times\; \Big(\gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n} - \boldsymbol{\nu}_{\boldsymbol{\mu}})\,\mathrm{dt} + o(\mathrm{dt})\Big) + \\ & + o(\mathrm{dt})\ . \end{aligned} \tag{4.119}$$

The different routes from $\vec{\mathcal{X}}(t_0) = \mathbf{n}_0$ to $\vec{\mathcal{X}}(t+\mathrm{dt}) = \mathbf{n}$ are obvious from the balance equation (4.119):

(i) One route from $\vec{\mathcal{X}}(t_0) = \mathbf{n}_0$ to $\vec{\mathcal{X}}(t + \mathrm{dt}) = \mathbf{n}$ is given by the first term on the right-hand side of the equation: *No reaction* is occurring in the time interval $[t, t + \mathrm{dt}[$ and hence $\vec{\mathcal{X}}(t) = \mathbf{n}$ was fulfilled at time t. The joint probability for route (i) is therefore *the probability to be in* $\vec{\mathcal{X}}(t) = \mathbf{n}$ *conditioned by* $\vec{\mathcal{X}}(t_0) = \mathbf{n}_0$ times *the probability that no reaction has occurred in* $[t, t + \mathrm{dt}[$. In other words, the probability for this route is the probability to go from $\mathbf{n}_0$ at time t_0 to $\mathbf{n}$ at time t and to stay in this state during the next interval dt.

(ii) An alternative route from $\vec{\mathcal{X}}(t_0) = \mathbf{n}_0$ to $\vec{\mathcal{X}}(t + \mathrm{dt}) = \mathbf{n}$ accounted for by one particular term in sum of terms on the right-hand side of the equation: *An* $\mathbf{R}_{\boldsymbol{\mu}}$ *reaction* is occurring in the time interval $[t, t+\mathrm{dt}[$ and hence $\vec{\mathcal{X}}(t) = \mathbf{n} - \boldsymbol{\nu}_{\boldsymbol{\mu}}$ was fulfilled at time t. The joint probability for route (ii) is

therefore *the probability to be in* $\vec{\mathcal{X}}(t) = \mathbf{n} - \boldsymbol{\nu}_{\boldsymbol{\mu}}$ *conditioned by* $\vec{\mathcal{X}}(t_0) = \mathbf{n}_0$ times *the probability that exactly one* $\mathbf{R}_{\boldsymbol{\mu}}$ *reaction has occurred in* $[t, t + \mathrm{dt}[$. In other words, the probability for this route is the probability to go from $\mathbf{n}_0$ at time t_0 to $\mathbf{n} - \boldsymbol{\nu}_{\boldsymbol{\mu}}$ at time t and to undergo an $\mathbf{R}_{\boldsymbol{\mu}}$ during the next interval dt. Obviously, the same consideration is valid for every elementary reaction and we have R terms of this kind.

(iii) A third possibility – neither *no* reaction nor *exactly one* reaction chosen from the set $\{\mathbf{R}_{\boldsymbol{\mu}}; \boldsymbol{\mu} = 1, \ldots, R\}$ – must inevitably invoke *more than one reaction within the time interval* $[t, t + \mathrm{dt}[$. The probability for such events, however, is $o(\mathrm{dt})$ or of measure zero by theorem 3.

All routes (i) and (ii) are mutually exclusive since different events are taking place within the last interval $[t, t + \mathrm{dt}[$.

The last step to derive the *chemical master equation* is straightforward: $P(\mathbf{n}, t|\mathbf{n}_0, t_0)$ is subtracted from both sides in equation (4.119), then both sides are divided by dt, the limit $\mathrm{dt} \downarrow 0$ is taken, all $o(\mathrm{dt})$ terms vanish and finally we obtain

$$\begin{aligned} \frac{\partial}{\partial t} P(\mathbf{n}, t|\mathbf{n}_0, t_0) \;=\; \sum_{\boldsymbol{\mu}=1}^{R} \Big(& \gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n} - \boldsymbol{\nu}_{\boldsymbol{\mu}})\, P(\mathbf{n} - \boldsymbol{\nu}_{\boldsymbol{\mu}}|\mathbf{n}_0, t_0) - \\ & - \gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})\, P(\mathbf{n}, t|\mathbf{n}_0, t_0) \Big) \,. \end{aligned} \tag{4.120}$$

Initial conditions are required to calculate the time evolution of the probability $P(\mathbf{n}, t|\mathbf{n}_0, t_0)$ and we can easily express them in the form

$$P(\mathbf{n}, t_0|\mathbf{n}_0, t_0) \;=\; \begin{cases} 1\,, & \text{if } \mathbf{n} = \mathbf{n}_0\,, \\ 0\,, & \text{if } \mathbf{n} \neq \mathbf{n}_0\,, \end{cases} \tag{4.120'}$$

which is precisely the initial condition used in the derivation of equation (4.119). Any sharp initial probability distribution $P\big(n_k, t_0|n_k^{(0)}, t_0\big) = \delta(n_k - n_k^{(0)})$ is admitted for the molecular particle numbers at t_0. The assumption of extended initial distributions is, of course, also possible but the corresponding master equations become more sophisticated.

4.7.4 The simulation algorithm

The chemical master equation in the form (4.120) as derived from molecular collisions in section 4.2.2 is the basis of Gillespie's stochastic simulation algorithm [106] and it is important to realize how the simulation tool fits into the general theoretical framework of the chemical master equation. The algorithm is not based on the probability function $P(\mathbf{n}, t|\mathbf{n}_0, t_0)$ but on another related probability density $p(\tau, \boldsymbol{\mu}|\mathbf{n}, t)$, which expresses the probability that given $\vec{\mathcal{X}}(t) = \mathbf{n}$ the *next* reaction in the system will occur in the infinitesimal time interval $[t + \tau, t + \tau + \mathrm{d}\tau[$, *and* it will be an $\mathbf{R}_{\boldsymbol{\mu}}$ reaction.

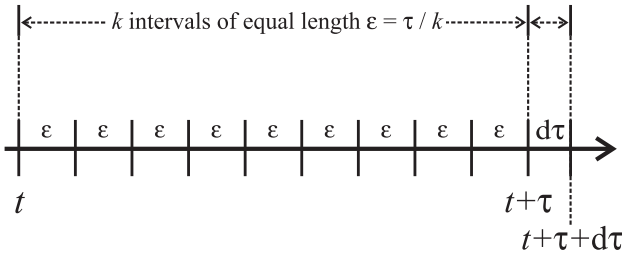

Fig. 4.15 Partitioning of the time interval $[t, t + \tau + \mathrm{d}\tau[$**.** The entire interval is subdivided into $(k+1)$ nonoverlapping subintervals. The first k intervals are of equal size $\varepsilon = \tau/k$ and the $(k+1)$-th interval is of length $\mathrm{d}\tau$.

Within the frame of the theory of random variables, $p(\tau, \boldsymbol{\mu}|\mathbf{n}, t)$ is the joint density function of two random variables: (i) the *time to the next reaction*, τ, and (ii) the *index of the next reaction*, $\boldsymbol{\mu}$. The possible values of the two random variables are given by the domain of the real variable $0 \leq \tau < \infty$ and the integer variable $1 \leq \boldsymbol{\mu} \leq R$. In order to derive an explicit formula for the probability density $p(\tau, \boldsymbol{\mu}|\mathbf{n}, t)$ we introduce the quantity

$$\alpha(\mathbf{n}) \;=\; \sum_{\boldsymbol{\mu}=1}^{R} \gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})$$

and consider the time interval $[t, t + \tau + \mathrm{d}\tau[$ to be partitioned into $k + 1$ subintervals, $k > 1$. The first k of these intervals are chosen to be of equal length $\varepsilon = \tau/k$, and together they cover the interval $[t, t + \tau[$ leaving the interval $[t + \tau, t + \tau + \mathrm{d}\tau[$ as the remaining $(k + 1)$-th part (figure 4.15). With $\vec{\mathcal{X}}(t) = \mathbf{n}$ the probability $p(\tau, \boldsymbol{\mu}|\mathbf{n}, t)$ describes the event *no reaction* occurring in each of the k ε-size subintervals and *exactly one* $\mathbf{R}_{\boldsymbol{\mu}}$ *reaction* in the final infinitesimal $\mathrm{d}\tau$ interval. Making use of theorems 1 and 2 and the multiplication law of probabilities we find

$$p(\tau, \boldsymbol{\mu}|\mathbf{n}, t) \;=\; \Big(1 \,-\, \alpha(\mathbf{n})\,\varepsilon \,+\, o(\varepsilon)\Big)^k \Big(\gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})\,\mathrm{d}\tau \,+\, o(\mathrm{d}\tau)\Big)$$

Dividing both sides by $\mathrm{d}\tau$ and taking the limit $\mathrm{d}\tau \downarrow 0$ yields

$$p(\tau, \boldsymbol{\mu}|\mathbf{n}, t) \;=\; \Big(1 \,-\, \alpha(\mathbf{n})\,\varepsilon \,+\, o(\varepsilon)\Big)^k \gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})$$

This equation is valid for any integer $k > 1$ and hence its validity is also guaranteed for $k \to \infty$. Next we rewrite the first factor on the right-hand side of the equation

$$\begin{aligned}\Big(1 \,-\, \alpha(\mathbf{n})\,\varepsilon \,+\, o(\varepsilon)\Big)^k \;&=\; \left(1 \,-\, \frac{a(\mathbf{n})\,k\varepsilon \,+\, k\,o(\varepsilon)}{k}\right)^k \;= \\ &=\; \left(1 \,-\, \frac{\alpha(\mathbf{n})\,\tau \,+\, \tau\,o(\varepsilon)/\varepsilon}{k}\right)^k ,\end{aligned}$$

and take now the limit $k \to \infty$ whereby we make use of the simultaneously occurring convergence $o(\varepsilon)/\varepsilon \downarrow 0$:

$$\lim_{k\to\infty}\Big(1 \,-\, \alpha(\mathbf{n})\,\varepsilon \,+\, o(\varepsilon)\Big)^k \;=\; \lim_{k\to\infty}\left(1 - \frac{\alpha(\mathbf{n})\,\tau}{k}\right)^k \;=\; e^{-\alpha(\mathbf{n})\,\tau} .$$

By substituting this result into the initial equation for the probability density of the occurrence of a reaction we find

$$\begin{aligned}p(\tau, \boldsymbol{\mu}|\mathbf{n}, t) \;&=\; \alpha(\mathbf{n})\,\frac{\gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})}{\alpha(\mathbf{n})}\, e^{-\alpha(\mathbf{n})\,\tau} \;= \\ &=\; \gamma_{\boldsymbol{\mu}}\, h_{\boldsymbol{\mu}}(\mathbf{n})\; e^{-\sum_{\nu=1}^{R}\gamma_{\boldsymbol{\nu}} h_{\boldsymbol{\nu}}(\mathbf{n})\,\tau}.\end{aligned} \tag{4.121}$$

Equation (4.121) provides the mathematical basis for the stochastic simulation algorithm. Given $\vec{\mathcal{X}}(t) = \mathbf{n}$, the probability density consists of two independent probabilities where the first factor describes the *time to the next reaction* and the second factor the *index of the next reaction*. These factors correspond to two statistically independent random variables η_1 and η_2.

4.7.5 Implementation of the simulation algorithm

Equation (4.121) is implemented now for computer simulation and we inspect the probability densities of the two unit-interval uniform random variables η_1 and η_2 in order to find the conditions to be imposed of a statistically exact sample pair $(\tau, \boldsymbol{\mu})$: η_1 has an exponential density function with the decay constant $\alpha(\mathbf{n})$,

$$\tau \;=\; \frac{1}{\alpha(\mathbf{n})}\;\ln\bigl(1\,/\,\eta_1\bigr)\;, \tag{4.122a}$$

and taking m to be the *smallest* integer which fulfils

$$\boldsymbol{\mu} \;=\; \inf\left\{ m \;\Big|\; \sum_{\mu=1}^{m} c_\mu\, h_\mu(\mathbf{n}) \;>\; \alpha(\mathbf{n})\;\eta_2 \right\}\;. \tag{4.122b}$$

After the values for τ and $\boldsymbol{\mu}$ have been determined the action *advance the state vector* $\vec{\mathcal{X}}(t)$ *of the system* is taking place:

$$\vec{\mathcal{X}}(t) \;=\; \mathbf{n} \quad\longrightarrow\quad \vec{\mathcal{X}}(t+\tau) = \mathbf{n} + \boldsymbol{\nu}_{\boldsymbol{\mu}}\;.$$

Repeated application of the advancement procedure is the essence of the stochastic simulation algorithm. It is important to realize that this advancement procedure is exact as far as η_1 and η_2 are obtained by *fair samplings* from a unit interval uniform random number generator or, in other words, the correctness of the procedure depends on the quality of the random number generator applied. Two further issues are important: (i) The algorithm operates with internal time control that corresponds to real time of the chemical process, and (ii) contrary to the situation in differential equation solvers the discrete time steps are not finite interval approximations of an infinitesimal time step and instead, the population vector $\vec{\mathcal{X}}(t)$ maintains the value $\vec{\mathcal{X}}(t) = \mathbf{n}$ throughout the entire finite time interval $[t, t+\mathrm{d}\tau[$ and then changes abruptly to $\vec{\mathcal{X}}(t+\tau) = \mathbf{n} + \boldsymbol{\nu}_{\boldsymbol{\mu}}$ at the instant $t+\tau$ when the $\mathbf{R}_{\boldsymbol{\mu}}$ reaction occurs. In other words, there is no *blind interval* during which the algorithm is unable to record changes.

4.7.5.1 Structure of the simulation algorithm

The time evolution of the population in described by the vector $\vec{\mathcal{X}}(t) = \mathbf{n}(t)$, which is updated after every individual reaction event. Reactions are chosen from the set $\mathcal{R} = \{\mathbf{R}_{\boldsymbol{\mu}}; \boldsymbol{\mu} = 1, \ldots, R\}$, which is defined by the reaction mechanism under consideration. They are classified according to the criteria listed in table 4.1. The reaction probabilities are contained in a vector $\boldsymbol{\alpha}(\mathbf{n}) = \bigl(c_1 h_1(\mathbf{n}), \ldots, c_R h_R(\mathbf{n})\bigr)^{\mathrm{t}}$, which is also updated after every individ-

Table 4.1 The combinatorial functions $h_{\boldsymbol{\mu}}(\mathbf{n})$ for elementary reactions. Reactions are ordered with respect to reaction order, which in case of mass action is identical to the molecularity of the reaction. Order zero implies that no reactant molecule is involved and the products come from an external source, for example from the influx in a flow reactor. Orders 0, 1, 2, and 3 mean that zero, one, two or three molecules are involved in the elementary step, respectively.

No.	Reaction	Order	$h_{\boldsymbol{\mu}}(\mathbf{n})$
1	$* \longrightarrow$ products	0	1
2	$\mathbf{A} \longrightarrow$ products	1	$n_{\mathbf{A}}$
3	$\mathbf{A}+\mathbf{B} \longrightarrow$ products	2	$n_{\mathbf{A}}n_{\mathbf{B}}$
4	$2\,\mathbf{A} \longrightarrow$ products	2	$n_{\mathbf{A}}(n_{\mathbf{A}}-1)/2$
5	$\mathbf{A}+\mathbf{B}+\mathbf{C} \longrightarrow$ products	3	$n_{\mathbf{A}}n_{\mathbf{B}}n_{\mathbf{C}}$
6	$2\,\mathbf{A}+\mathbf{B} \longrightarrow$ products	3	$n_{\mathbf{A}}(n_{\mathbf{A}}-1)n_{\mathbf{B}}/2$
7	$3\,\mathbf{A} \longrightarrow$ products	3	$n_{\mathbf{A}}(n_{\mathbf{A}}-1)(n_{\mathbf{A}}-2)/6$

ual reaction event. Updating is performed according to the stoichiometric vectors $\boldsymbol{\nu}_{\boldsymbol{\mu}}$ of the individual reactions $\mathbf{R}_{\boldsymbol{\mu}}$, which represent columns of the stoichiometric matrix S. We repeat that the combinatorial functions $h_{\boldsymbol{\mu}}(\mathbf{n})$ are determined exclusively by the reactant side of the reaction equation whereas the stoichiometric vectors $\boldsymbol{\nu}_{\boldsymbol{\mu}}$ represent the net production, (*products*)−(*reactants*).
The algorithm comprises five steps:

(i) *Step 0. Initialization*: The time variable is set to $t = 0$, the initial values of all N variables $\mathcal{X}_1, \ldots, \mathcal{X}_N$ for the species – $\mathcal{X}_k$ for species $\mathbf{S}_k$ – are stored, the values for the R parameters of the reactions $\mathbf{R}_{\boldsymbol{\mu}}$, $c_1, \ldots, c_R$, are stored, and the combinatorial expressions are incorporated as factors for the calculation of the reaction rate vector $\boldsymbol{\alpha}(\mathbf{n})$ according to table 4.1 and the probability density $P(\tau, \boldsymbol{\mu})$. Sampling times, $t_1 < t_2 < \cdots$ and the stopping time t_{stop} are specified, the first sampling time is set to t_1 and stored and the *pseudorandom* number generator is initialized by means of *seeds* or *at random*.

(ii) *Step 1. Monte Carlo step*: A pair of random numbers is created $(\tau, \boldsymbol{\mu})$ by the random number generator according to the joint probability function $P(\tau, \boldsymbol{\mu})$. In essence two explicit methods can be used: the *direct* method and the *first-reaction* method.

(iii) *Step 2. Propagation step*: $(\tau, \boldsymbol{\mu})$ is used to advance the simulation time t and to update the population vector $\mathbf{n}$, $t \to t + \tau$ and $\mathbf{n} \to \mathbf{n} + \boldsymbol{\nu}_{\boldsymbol{\mu}}$, then all changes are incorporated in a recalculation of the reaction rate vector $\mathbf{a}$.

(iv) *Step 3. Time control*: Check whether or not the simulation time has been advanced through the next sampling time t_i, and for $t > t_i$ send current t and current $\mathbf{n}(t)$ to the output storage and advance the sampling time,

$t_i \to t_{i+1}$. Then, if $t > t_{\text{stop}}$ or if no more reactant molecules remain leading to $h_{\boldsymbol{\mu}} = 0 \;\forall\; \boldsymbol{\mu} = 1, \ldots, R$, finalize the calculation by switching to *step 4*, and otherwise continue with *step 1*.

(v) *Step 4. Termination*: Prepare for final output by setting flags for early termination or other unforseen stops and send final time t and final $\mathbf{n}$ to the output storage and terminate the computation.

A caveat is needed for the integration of stiff systems where the values of individual variable can vary by many orders of magnitude and such a situation might caught the calculation in a trap by slowing down time progress.

4.7.5.2 The Monte Carlo step

Pseudorandom numbers are drawn from a random number generator of sufficient quality whereby quality is meant in terms of no or very long recurrence cycles and a the closeness of the distribution of the pseudorandom numbers r to the uniform distribution on the unit interval:

$$0 \le a < b \le 1 \qquad \Longrightarrow \qquad P(a \le \eta \le b) \;=\; b - a \;.$$

With this prerequisite we discuss now two methods which use two output values η of the pseudorandom number generator to generate a random pair $(\tau, \boldsymbol{\mu})$ with the prescribed probability density function $P(\tau, \boldsymbol{\mu})$.

The *direct* method. The two-variable probability density is written as the product of two one-variable density functions:

$$P(\tau, \boldsymbol{\mu}) \;=\; P_1(\tau) \cdot P_2(\boldsymbol{\mu}|\tau) \;.$$

Here, $P_1(\tau)\,\mathrm{d}\tau$ is the probability that the next reaction will occur between times $t + \tau$ and $t + \tau + \mathrm{d}\tau$, irrespective of which reaction it might be, and $P_2(\boldsymbol{\mu}|\tau)$ is the probability that the next reaction will be an $\mathbf{R}_{\boldsymbol{\mu}}$ given that the next reaction occurs at time $t + \tau$.

By the addition theorem of probabilities, $P_1(\tau)\,\mathrm{d}\tau$ is obtained by summation of $P(\tau, \boldsymbol{\mu})\,\mathrm{d}\tau$ over all reactions $\mathbf{R}_{\boldsymbol{\mu}}$:

$$P_1(\tau) \;=\; \sum_{\boldsymbol{\mu}=1}^{R} P(\tau, \boldsymbol{\mu}) \;. \tag{4.123}$$

Combining the last two equations we obtain for $P_2(\boldsymbol{\mu}|\tau)$

$$P_2(\boldsymbol{\mu}|\tau) \;=\; P(\tau, \boldsymbol{\mu}) \;/\; \sum_{\boldsymbol{\nu}}^{R} P(\tau, \boldsymbol{\nu}) \tag{4.124}$$

Equations (4.123) and (4.124) express the two one-variable density functions in terms of the original two-variable density function $P(\tau, \boldsymbol{\mu})$. From equation (4.121) we substitute into $P(\tau, \boldsymbol{\mu}) = p(\tau, \boldsymbol{\mu}|\mathbf{n}, t)$ through simplifying the notation by using

$$\alpha_{\boldsymbol{\mu}} \;\equiv\; \gamma_{\boldsymbol{\mu}} h_{\boldsymbol{\mu}}(\mathbf{n}) \;\text{ and }\; \alpha \;=\; \sum_{\boldsymbol{\mu}=1}^{R} \alpha_{\boldsymbol{\mu}} \;\equiv\; \sum_{\boldsymbol{\mu}=1}^{R} \gamma_{\boldsymbol{\mu}} h_{\boldsymbol{\mu}}(\mathbf{n})$$

and find

$$\begin{aligned} P_1(\tau) \;&=\; \alpha \exp(-\alpha\,\tau)\,, \quad 0 \le \tau < \infty \;\text{ and} \\ P_2(\boldsymbol{\mu}|\tau) \;&=\; P_2(\boldsymbol{\mu}) \;=\; \alpha_{\boldsymbol{\mu}} \,/\, \alpha\,, \quad \boldsymbol{\mu} = 1, \ldots, R\;. \end{aligned} \tag{4.125}$$

As indicated, in this particular case, $P_2(\boldsymbol{\mu}|\tau)$ turns out to be independent of τ. Both one variable density functions are properly normalized over their domains of definition:

$$\int_0^\infty P_1(\tau)\,\mathrm{d}\tau \;=\; \int_0^\infty \alpha\, e^{-\alpha\,\tau}\,\mathrm{d}\tau \;=\; 1 \;\text{ and }\; \sum_{\boldsymbol{\mu}=1}^{R} P_2(\boldsymbol{\mu}) \;=\; \sum_{\boldsymbol{\mu}=1}^{R} \frac{\alpha_{\boldsymbol{\mu}}}{\alpha} \;=\; 1\;.$$

Thus, in the *direct* method a random value τ is created from a random number on the unit interval, η_1, and the distribution $P_1(\tau)$ by taking

$$\tau \;=\; -\,\frac{\ln \eta_1}{\alpha}\;. \tag{4.126}$$

The second task is to generate a random integer $\hat{\boldsymbol{\mu}}$ according to $P_2(\boldsymbol{\mu}|\tau)$ in such a way that the pair $(\tau, \boldsymbol{\mu})$ will be distributed as prescribed by $P(\tau, \boldsymbol{\mu})$. For this goal another random number, η_2, will be drawn from the unit interval and then $\hat{\boldsymbol{\mu}}$ is taken to be the integer that fulfils

$$\sum_{\boldsymbol{\nu}=1}^{\boldsymbol{\mu}-1} \alpha_{\boldsymbol{\nu}} \;<\; \eta_2\,\alpha \;\le\; \sum_{\boldsymbol{\nu}=1}^{\boldsymbol{\mu}} \alpha_{\boldsymbol{\nu}}\;. \tag{4.127}$$

The values α_1, α_2, ... , are cumulatively added in sequence until their sum is observed to be equal or to exceed $\eta_2\alpha$ and then $\hat{\boldsymbol{\mu}}$ is set equal to the index of the last $\alpha_{\boldsymbol{\nu}}$ term that had been added. Rigorous justifications for equations (4.126) and (4.127) are found in [101, pp.431-433]. If a fast and reliable uniform random number generator is available, the *direct* method can be easily programmed and rapidly executed. This it represents a simple, fast, and rigorous procedure for the implementation of the *Monte Carlo* step of the simulation algorithm.

The *first-reaction* method. This alternate method for the implementation of the *Monte Carlo* step of the simulation algorithm is not quite as efficient

as the *direct* method but it is worth presenting here because it adds insight into the stochastic simulation approach. Adopting again the notation $\alpha_{\boldsymbol{\nu}} \equiv \gamma_{\boldsymbol{\nu}} h_{\boldsymbol{\nu}}(\mathbf{n})$ it is straightforward to derive

$$P_{\boldsymbol{\nu}}(\tau)\,\mathrm{d}\tau \;=\; \alpha_{\boldsymbol{\nu}}\,\exp(-\alpha_{\boldsymbol{\nu}}\,\tau)\,\mathrm{d}\tau \tag{4.128}$$

from (4.114) and (4.115). Then, $P_{\boldsymbol{\nu}}(\tau)$ would indeed be the probability at time t for an $\mathbf{R}_{\boldsymbol{\nu}}$ reaction to occur in the time interval $[t+\tau, t+\tau+d\tau[$ were it not for the fact that the number of $\mathbf{R}_{\boldsymbol{\nu}}$ reactant combinations might have been altered between t and $t+\tau$ by the occurrence of other reactions. Taking this into account, a *tentative reaction time* $\tau_{\boldsymbol{\nu}}$ for $\mathbf{R}_{\boldsymbol{\nu}}$ is generated according to the probability density function $P_{\boldsymbol{\nu}}(\tau)$, and in fact, the same can be done for all reactions $\{\mathbf{R}_{\boldsymbol{\mu}}\}$. We draw a random number $\eta_{\boldsymbol{\nu}}$ from the unit interval and compute

$$\tau_{\boldsymbol{\nu}} \;=\; -\frac{\ln \eta_{\boldsymbol{\nu}}}{\alpha_{\boldsymbol{\nu}}}\,,\; \boldsymbol{\nu} = 1, \ldots, R\,. \tag{4.129}$$

From these R *tentative next* reactions the one, which occurs first, is chosen to be the *actual next* reactions:

$$\begin{aligned} \tau \;&=\; \text{smallest } \tau_{\boldsymbol{\nu}} \;\text{ for all }\; \boldsymbol{\nu} = 1, \ldots, R\,, \\ \boldsymbol{\mu} \;&=\; \boldsymbol{\nu} \text{ for which } \tau_{\boldsymbol{\nu}} \text{ is smallest}\,. \end{aligned} \tag{4.130}$$

Daniel Gillespie [101, pp.420-421] provides a straightforward proof that the random $(\tau, \boldsymbol{\mu})$ obtained by the *first reaction* method is in full agreement with the probability density $P(\tau, \boldsymbol{\mu})$ from equation (4.121).

It is tempting to try to extend the *first reaction* methods by letting the *second next* reaction be the one for which $\tau_{\boldsymbol{\nu}}$ has the second smallest value. This, however, is in conflict with correct updating of the vector of particle numbers, $\mathbf{n}$, because the results of the first reaction are not incorporated into the combinatorial terms $h_{\boldsymbol{\mu}}(\mathbf{n})$. Using the second earliest reaction would, for example, allow the second reaction to involve molecules already destroyed in the first reaction but would not allow the second reaction to involve molecules created ion the first reaction.

Thus, the *first reaction* method is just as rigorous as the *direct* method and it is probably easier to implement in a computer code than the direct method. From a computational efficiency point of view, however, the direct method is preferable because for $R \geq 3$ it requires fewer random numbers and hence the first reaction methods is wasteful. This question of economic use of computer time is not unimportant because stochastic simulations in general are taxing the random number generator quite heavily. For $R \geq 3$ and in particular for large R the direct method is probably the method of choice for the Monte Carlo step.

4.7.5.3 The computer code

An early computer code of the simple version of the algorithm described – still in *FORTRAN* – is found in [101]. Meanwhile many attempts were made in order to speed-up computations and allow for simulation of stiff systems (see e.g. [30]. A recent review of the simulation methods also contains a discussion of various improvements of the original code [106].

4.7.6 Examples of simulations

Chapter 5
Applications in biology

Nothing in biology makes sense except in the light of evolution.
Theodosius Dobzhansky, 1972.

Abstract Stochastic phenomena are central to biological modeling: Small numbers of molecules regulate and control genetics, epigenetics, and cellular metabolism, and small numbers of well-adapted individuals drive evolution. Reproduction, the basis of all processes in biology is autocatalysis in the language of chemists and replication of the genetic molecules, DNA and RNA, build the bridge between chemistry and biology. The earliest stochastic models in biology were applying branching processes to find answers to genealogical questions like the fate of family names in pedigrees. Branching processes, birth-and-death processes, and related stochastic models are frequently used in biology and they will be defined, analyzed, and applied to typical problems. Although the master equation is not so dominant in biology as it is chemistry, it is sufficiently important to justify a detailed presentations. Kimura's neutral theory of evolution makes use of a Fokker-Planck equation to describe population dynamics in the absence of fitness differences. Simulations of stochastic reaction networks are a rapidly growing field as illustrated by means of a few examples.

The population aspect is basic to biology, in particular it is highly important in evolution and accordingly we introduce it here again. A population vector

$$\boldsymbol{\Pi}(t) = \big(N_1(t), N_2(t), \ldots, N_n(t)\big) \quad \text{with} \quad N_k \in \mathbb{N}_0\,,\ t \in \mathbb{R}^1_{\geq 0}\ ,$$

counts the numbers of individuals for the different *species*[1] $\mathbf{X}_k$ as a function of time $N_k(t)$. This definition states already that time will be considered as a continuous variable and the use of *counting* implies that the numbers of individuals are discrete. The basic assumptions thus are the same as in the applications of master equations to chemical reaction kinetics (section 4.2.1).

[1] In this chapter we use species in the sense of *molecular species* for any component of a population vector $\boldsymbol{\Pi} = (N_1, N_2, \ldots, N_n)$ and indicate by *biological species* the genuine notion of species in biology.

There is, however, a major difference between the molecular approach based on elementary reactions on one hand side and macroscopic modeling as commonly used in biology on the other hand: The biological objects are no longer single molecules or atoms but modules commonly consisting of a large number of atoms or individual cells or organisms. Elementary step dynamics obeys several conservation relations like conservation of mass or conservation of the numbers of atoms for every chemical element – unless nuclear reactions are admitted, and the laws of thermodynamics provide additional restrictive relations. In the macroscopic models these relations are not violated, of course, but they are hidden in complex networks of interactions, which appear in the model only after averaging on several hierarchical levels. For example, conservation of mass and energy are encapsulated and obscured in the carrying capacity K of the ecosystem as modeled by the logistic equation [292]:

$$\frac{\mathrm{d}N}{\mathrm{dt}} \;=\; h\,N\left(1-\frac{N}{K}\right) \quad \text{and} \quad N(t)\;\frac{N_0\,K}{N_0\;+\;(K-N_0)\,\exp(-ht)}\;, \tag{5.1}$$

with $N = \sum_{i=1}^{n} N_i$ being the population size, h the so-called Malthus or growth parameter,[2] and $N_0 = N(0)$ representing the initial population size at time $t = 0$. Consequently, numbers of individual species may change in biological models, $N_k(t) \to N_k(t + \Delta t) \pm 1$, without a compensation in another variable. A similar situation in chemistry is happening in *buffering* where a large molecular reservoir remains practically unchanged when a single molecule is added or subtracted (see section 4.6.4 figure 4.13, and irreversible addition reaction in section 4.3.3.1). In other biological models, for example in population genetics, the limitation of the population size is part of the specific model, or as done most frequently normalized variables are used. As indicated above the changes ± 1 in the numbers of individuals imply that the time interval considered is sufficiently short that multiple events can be excluded. Exceptions are, of course, processes with m particles reacting in a single event, for example $m\mathbf{X} \to \ldots$, where the changes are $\pm m$. In biology we can interpret the flow reactor (section 4.3.1) as a kind of idealized ecosystem. The analogous processes to influx (4.1a) and outflux (4.1b) in biological systems are migration, immigration and emigration, respectively.

A stochastic process on the population level is – by the same token as in section 3.1 – a recording of time ordered successive events at times T_i:

$$T_0 < T_1 < T_2 < \ldots < T_{k-1} < T_k < T_{k+1} \ldots ,$$

along a continuous time axis t.[3] As an example we consider a birth event or a death event at some time $t = T_r$, which creates or consumes one individual

[2] the Malthus parameter is commonly denoted by r. Since r is defined as the flow rate in the CSTR, we use here h in order to avoid confusion.

[3] The application of discretized time in evolution – mimicking synchronized generations, for example – is straightforward and we shall discuss a specific case in detail (section 5.2.1.2), because we focus here on continuous time birth-and-death processes and master equations.

according to the processes, $\mathbf{X}_j \to 2\mathbf{X}_j$ or $\mathbf{X}_j \to \varnothing$, respectively. Then the population changes according to:

$$\boldsymbol{\Pi} = \begin{cases} \Big(\ldots, N_j(t) = N_j(T_{r-1}), N_k(t) = N_k(T_{r-1}), \ldots\Big) & \text{for } T_{r-1} \le t < T_r \\ \Big(\ldots, N_j(t) = N_j(T_{r-1}) \pm 1, N_k(t) = N_k(T_{r-1}), \ldots\Big) & \text{for } T_r \le t < T_{r+1} \end{cases}.$$

This formulation of a biological birth or death events reflects the previously mentioned convention in probability theory: Right-hand continuity is assumed for steps in stochastic processes (see figure 1.9).

Compared to stochasticity in chemistry stochastic phenomena in biology are not only more important but also much harder to control. The major sources of the problem are small population numbers and the lack of sufficiently simple references systems that are accessible to experimental studies. In biology we are regularly encountering reaction mechanisms that lead to enhancement of fluctuations at non-equilibrium conditions and biology in essence is dealing with processes and stationary states far away from equilibrium whereas in chemistry autocatalysis in non-equilibrium systems became an object of general interest and intensive investigation not before some forty years ago. We start therefore with the classification and analysis of simple autocatalytic processes at the ODE level (section 5.1). In section 5.2 we present an overview of the various stochastic processes that are popular in biology: branching processes (section 5.2.1), solvable birth-and-death processes including boundaries in form of different barriers (section 5.2.2), and special biological models (section 5.2.3). Then, we discuss the usage of master equations in stochastic modeling of biological phenomena (section 5.3). A section on discrete time processes follows, which are important in biology when synchronization occurs with external or internal pace makers, for example seasons (the relation between discrete and continuous time models is discussed by means of a specific example in section 5.2.1.2). Random selection occurs in case of selective neutrality as postulated in the neutral theory of evolution (section 5.3.2) and finally, we shall present examples of numerical analysis and simulation of stochastic processes in biology (section 5.5).

5.1 Autocatalysis and growth

Autocatalysis in its simplest form is found in the bimolecular reaction (4.1g): $\mathbf{A} + \mathbf{X} \rightarrow 2\mathbf{X}$. In the previous chapter 4 we studied already bimolecular reactions, the addition and the dimerization reaction, which allowed for analytical solution and which gave rise to conventional or perfectly *normal* behavior, although the analysis and the solutions were quite sophisticated (section 4.3.3). The nonlinearity in the kinetic equation became manifest in the task to find solutions but did not change effectively the qualitative behavior of the reaction systems, for example the $\sqrt{N}$-law for the fluctuations in the stationary states retained in essence its validity.

The simplest conceivable autocatalytic reaction mechanism consists of two elementary steps, reproduction and extinction, and will be studied as an example for an exactly solvable birth-and-death process in section 5.2.2. In this case the $\sqrt{N}$-law is not valid and fluctuations do not settle down to some value which is proportional to the square root of the size of the system but grow in time without limit as we saw in case of the Wiener process (section 3.2.3.2). Here, we shall set the stage by reviewing the most relevant results from conventional deterministic autocatalysis [240, pp. 9-75] (section 5.1.1 and 5.1.2). A short glance of the relation between autocatalysis and growth (section 5.1.3) ends the section.

5.1.1 Autocatalysis in closed systems

Autocatalysis in its simplest form is described by the single reaction step

$$A + n\,X \underset{h_1}{\overset{k_1}{\rightleftharpoons}} (n+1)\,X\ , \tag{5.2}$$

which for small n is already contained in equation (4.1): $n = 0$ represents the uncatalyzed monomolecular conversion reaction (4.1c) $A \rightleftharpoons X$, and $n = 1$ is the bimolecular reaction of first order autocatalysis (4.1g). Equation (5.2) with $n = 2$ corresponds to a termolecular reaction[4] which is representative for second and higher order autocatalysis. It is a frequently used component of mechanisms exhibiting unconventional nonlinear behavior (see section 5.1.3).

[4] Termolecular and higher reaction steps are commonly neglected in mass action kinetics because they require a highly improbable encounter of three molecules. They are nevertheless frequently used in models and simplified kinetic mechanisms, examples are the Schlögl model [254] and the Brusselator model [224]. The Oregonator model [79, 80, 252] is a five step mechanism showing similar behavior without a termolecular step. Both, Brusselator and Oregonator are simplified models for the Belousov-Zhabotinsky reaction [305].

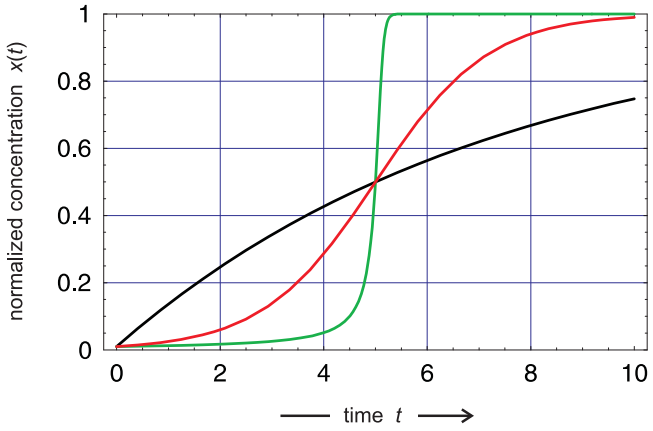

Fig. 5.1 Autocatalysis in a closed system. The concentration of the substance **X** as a function of time, $x(t)$, according to equation (5.5) is compared for the uncatalyzed first order reaction $\mathbf{A} \to \mathbf{X}$ ($n = 0$; black curve), for the first order autocatalytic process $\mathbf{A}+\mathbf{X} \to 2\mathbf{X}$ ($n = 1$; red curve), and for the second order autocatalytic process, $\mathbf{A}+2\mathbf{X} \to 3\mathbf{X}$ ($n = 2$ green curve). The following initial conditions and rate parameters were chosen: $x_0 = 0.01$, $c_0 = a(t) + x(t) = 1$ (normalized concentrations), $h_1 = 0$ (irreversible reaction), and $k_1 = 0.13662$, 0.9190 and 20.519 for the uncatalyzed process, the first order and the second order autocatalytic process, respectively. The rate parameters k_1 are chosen such that all curves pass the point $(x, t) = (0.5, 5)$.

Still higher autocatalytic elementary steps, $n \geq 3$, give rise to qualitative behavior that is very similar to the case $n = 2$.

In the case of mass action kinetics autocatalysis is modeled by the differential equation

$$\frac{dx}{dt} \;=\; -\,\frac{da}{dt} \;=\; k_1\, x^n\, a \;-\; h_1\, x^{n+1}\;. \tag{5.3}$$

The variables are the concentrations of molecular species: $x(t) = [X]$ and $a(t) = [A]$ with the initial concentrations $x(0) = x_0$ and $a(0) = a_0$ and the conservation relation $x(t)+a(t) = c_0$.[5] Equation (5.3) can be solved by means of the integral [109, p.106]:

$$\int \frac{\mathrm{d}x}{x^n(\alpha+\beta x)} \;=\; \sum_{k=1}^{n-1} \frac{(-1)^k\,\beta^{k-1}}{(n-k)\,\alpha^k\,x^{n-k}} \;+\; \frac{(-1)^n\,\beta^{n-1}}{\alpha^n}\,\ln\frac{\alpha+\beta x}{x}\;.$$

with $\alpha = k_1 c_0$, $\beta = -(k_1 + h_1)$, and $n \in \mathbb{N}_{>0}$. For the special case $n = 0$ we have $\int \mathrm{d}x/(\alpha+\beta x) = \ln(\alpha+\beta x)/\beta$.

[5] The conservation law is a result of mass conservation in the closed system considered here.

It is not possible to derive an explicit expression $x(t)$ in general but then the analysis of the implicit equation, $t(x)$ turns out to be quite useful too:

$$t(x) = \sum_{k=1}^{n-1} \frac{(-1)^k\,\beta^{k-1}}{(n-k)\,\alpha^k}\left(\frac{1}{x^{n-k}} - \frac{1}{x_0^{n-k}}\right) + \frac{(-1)^n\,\beta^{n-1}}{\alpha^n}\,\ln\frac{(\alpha+\beta x)\,x_0}{x\,(\alpha+\beta x_0)}\,. \quad (5.4)$$

For numerical calculations of the solution curves it makes practically no difference whether one considers $x(t)$ or $t(x)$.

In figure 5.1 the curves $x(t)$ for first order, $\mathbf{A}+\mathbf{X}\to 2\mathbf{X}$ and second order autocatalysis $\mathbf{A}+2\mathbf{X}\to 3\mathbf{X}$ are compared with the corresponding curve for the uncatalyzed process, $\mathbf{A}\to\mathbf{X}$:

$$n=0:\;\; x(t) = \frac{1}{k_1+h_1}\left(k_1c_0 + (h_1x_0 - k_1a_0)\,e^{-(k_1+h_1)\,t}\right)\,, \quad (5.5a)$$

$$n=1:\;\; x(t) = \frac{k_1\,c_0\,x_0}{(k_1+h_1)x_0\,(1-e^{-k_1c_0\,t}) + k_1c_0\,e^{-k_1c_0\,t}}\,, \quad (5.5b)$$

$$n=2:\;\; t(x) = \frac{1}{k_1c_0}\left(\frac{x-x_0}{x\,x_0} + \frac{k_1+h_1}{k_1\,c_0}\,\ln\frac{\big((k_1+h_1)x_0 - k_1c_0\big)\,x}{\big((k_1+h_1)x - k_1c_0\big)\,x_0}\right)\,. \quad (5.5c)$$

All three curves approach the final state monotonously – this is the state of complete conversion of $\mathbf{A}$ into $\mathbf{X}$, $\lim_{t\to\infty} x(t) = 1$, because we have chosen $h_1 = 0$. Both curves for autocatalysis show self-enhancement at low concentrations of the autocatalyst $\mathbf{X}$, pass through an inflection point, and then approach the final state in form of relaxation kinetics. The difference between first and second order autocatalysis manifests itself in the steepness of the curve, i.e. the value of the tangent at the inflection point, and is remarkably large. In general holds: The higher the coefficient of autocatalysis, the steeper is the curve; already for second order it is close to a step function.

Inspection of equation 5.5 reveals three immediate results:

(i) The autocatalytic reactions require a seeding amount of $\mathbf{X}$, since $x_0 = 0$ has the consequence $x(t) = 0\,\forall\,t$,

(ii) for sufficiently long time the system approaches a stationary state corresponding to chemical equilibrium

$$\lim_{t\to\infty} x(t) = \bar{x} = \frac{k_1}{k_1+h_1}\,c_0 \quad\text{and}\quad \lim_{t\to\infty} a(t) = \bar{a} = \frac{h_1}{k_1+h_1}\,c_0\,,\ \text{and}$$

(iii) The function $x(t)$ increases or decreases monotonously for $t > 0$ depending on whether $x_0 < \bar{x}$ or $x_0 > \bar{x}$ holds.

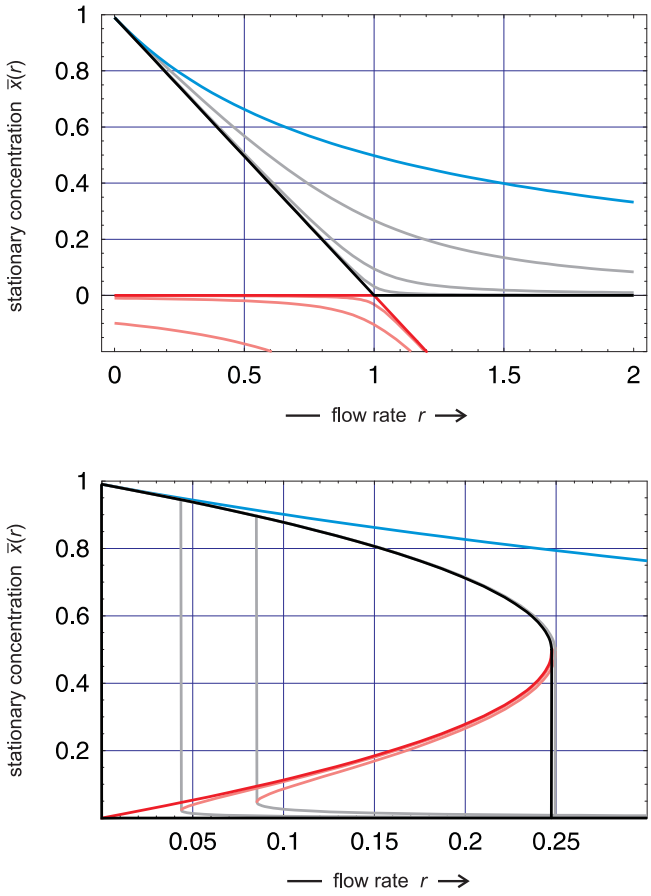

Fig. 5.2 Stationary states of autocatalysis in the flow reactor. The upper plot shows avoided crossing in first order autocatalysis ($n = 1$) when the uncatalyzed reaction is included. Parameter values: $k_1 = 1$, $h_1 = 0.01$, $c_0 = 1$, $\kappa = 0$ (black and red), $\kappa = 0.001$, 0.01, and 0.1 (grey and pink). The uncatalyzed reaction (blue) is shown for comparison. The lower plot refers to second order autocatalysis ($n = 2$) and shows shrinking of the range of bistability as a function of the parameter κ. Parameter values: $k_1 = 1$, $h_1 = 0.01$, $c_0 = 1$, $\kappa = 0$ (black and red), $\kappa = 0.0005$ and 0.002 (grey and pink). Again, the uncatalyzed reaction is shown in blue. The upper stable branch in the bistability range is called equilibrium branch, the lowest branch represent the state of extinction.

5.1.2 Autocatalysis in open systems

The continuously stirred tank reactor (CSTR) is an appropriate open system to study chemical reactions under controlled conditions (figure 4.7). Material consumed during the reaction flows contained in solution into the reactor and the volume increase is compensated by an outflow of reaction mixture. The flow rate is r and represents the reciprocal mean residence time of a volume element in the reactor: $r = \tau_{\mathrm{V}}^{-1}$. Substance **A** flows into the reactor with a concentration c_0 int he stock solution, and all substances being present int the reactor flow out by the same rate r. Both parameters, r and c_0, can be easily varied in experiments. The reaction is initiated by injection of a seeding amount of **X**, x_0. The reaction mechanism is of the form

$$\begin{aligned} * &\xrightarrow{c_0\, r} \mathbf{A}\ , \\ \mathbf{A} + n\,\mathbf{X} &\underset{h_1}{\overset{k_1}{\rightleftharpoons}} (n+1)\,\mathbf{X}\ , \\ \mathbf{A} &\xrightarrow{r} \varnothing\ , \quad \text{and} \\ \mathbf{X} &\xrightarrow{r} \varnothing\ . \end{aligned} \tag{5.6}$$

The stoichiometric factor n again distinguishes different cases, the uncatalyzed reaction with $n = 0$, first order autocatalysis with $n = 1$, and second or higher order autocatalysis with $n \geq 2$. Two kinetic differential equations are required to describe the temporal changes, because the concentrations a and x are now independent:

$$\begin{aligned} \frac{da}{dt} &= -k_1\, a\, x^n + h_1\, x^{n+1} + r\,(c_0 - a) \\ \frac{dx}{dt} &= k_1\, a\, x^n - h_1\, x^{n+1} - r\, x\ . \end{aligned} \tag{5.7}$$

The sum of the concentrations, $c(t) = a(t) + x(t)$, however, converges to the concentration or **A** in the stock solution, c_0, since

$$\frac{dc}{dt} = r\,(c_0 - c)\ .$$

The relaxation time towards the stable steady state $c(t) = \bar{c} = c_0$ is the mean residence time, $\tau_{\mathrm{V}} = r^{-1}$, and accordingly, different orders of autocatalysis, n, have no in influence on the relaxation time.

Steady states analysis, $\mathrm{d}a/\,\mathrm{dt} = 0$ and $\mathrm{d}x/\,\mathrm{dt} = 0$, reveals three different scenarios sharing the limiting cases: At vanishing flow rate r the system approaches thermodynamic equilibrium with $\bar{x} = k_1 c_0/(k_1 + h_1)$, $\bar{a} = h_1 c_0/(k_1 + h_1)$ and $K = k_1/h_1$, and no reaction occurs at sufficiently

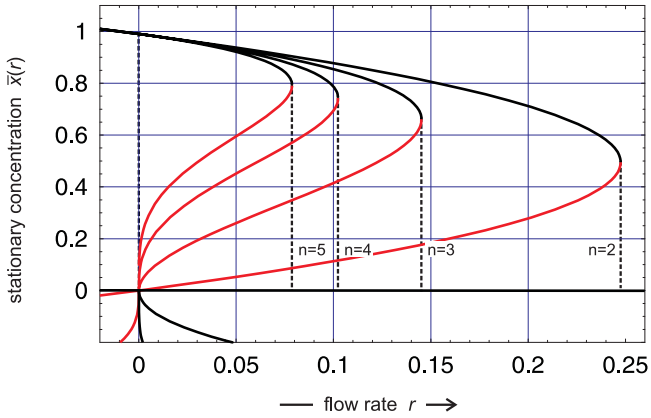

Fig. 5.3 Stationary states of higher order autocatalysis in the flow reactor. The curves show the range of bistability for different orders autocatalysis ($n = 2$, 3, 4, and 5 from right to left) and the parameters $k_1 = 1$, $h_1 = 0.01$, and $c_0 = 1$. The two stable branches, the thermodynamic branch (upper branch) and the state of extinction ($\bar{x} = 0$) are shown in black, the intermediate unstable branch is plotted in red. The vertical dotted lines indicate the critical points of the subcritical bifurcations.

large flow rates, $r > r_{\rm cr}$, when the mean residence time is too short to sustain changes due to the reaction and then we have $\bar{x} = 0$ and $\bar{a} = c_0$ for $\lim r \to \infty$. In the intermediate range, at finite flow rates $0 < r < r_{\rm cr}$, we observe:

(i) The unique steady state for the uncatalyzed process, $n = 0$, $\mathbf{A} \rightleftharpoons \mathbf{X}$ fulfils

$$\bar{x} = \frac{k_1\, c_0}{k_1 + h_1 + r} \quad \text{and} \quad \bar{a} = \frac{(h_1 + r)\, c_0}{k_1 + h_1 + r}$$

and show monotonous change from equilibrium to no reaction.

(ii) In case of first order autocatalysis, $n = 1$, steady state conditions yield two solutions,

$$\bar{x}_1 = \frac{k_1 c_0 - r}{k_1 + h_1}\,,\ \bar{a}_1 = \frac{h_1 c_0 + r}{k_1 + h_1} \quad \text{and} \quad \bar{x}_2 = 0\,,\ \bar{a}_2 = c_0\ . \tag{5.8}$$

The first solution $P_1 = (\bar{x}_1, \bar{a}_1)$ is stable in the range $0 \le r < k_1 c_0$ whereas solution $P_2 = (\bar{x}_2, \bar{a}_2)$ shows stability at high flow rates $r > k_1 c_0$. The change from the active state P_1 to the state of extinction, P_2, occurs abruptly at the transcritical bifurcation point $r = k_1 c_0$ (See the solution for $\kappa = 0$ in figure 5.2).[6]

[6] Bifurcation analysis is a standard topic in the theory of nonlinear systems. Monographs oriented towards practical application are, for example, [142, 143, 260].

(iii) Second and higher order autocatalysis ($n \geq 2$) allow for a common treatment. The steady condition yields[7]

$$r(\bar{x}) \;=\; k_1 c_0\, \bar{x}^{n-1} \;-\; (k_1 + h_1)\, \bar{x}^n \;.$$

Points with a horizontal tangent to $r(\bar{x})$, defined by $dr/d\bar{x} = 0$, in an $(\bar{x}, r)$-plot are points with a vertical tangent to the function $\bar{x}(r)$, which represent subcritical or other bifurcation points (figure 5.2). Such points correspond to maximal or minimal values of r at which branches of $\bar{x}(r)$ end and they can be computed analytically:

$$\bar{x}(r_{\max}) \;=\; \frac{n-1}{n} \cdot \frac{k_1\, c_0}{k_1 + h_1} \text{ for } n \geq 2 \;\text{ and }\; \bar{x}(r_{\min}) \;=\; 0 \text{ for } n \geq 3 \;,$$

with the corresponding flow rates

$$r_{max} \;=\; \left(\frac{n-1}{k_1 + h_1}\right)^{n-1} \cdot \left(\frac{k_1\, c_0}{n}\right)^{n} \;\text{ and }\; r_{\min} \;=\; 0 \;.$$

In figure 5.3 the bifurcation patterns for second and higher order autocatalysis in the flow reactor are compared. All four curves show a range of bistability, $r_{\min} < r < r_{\max}$, with two stable stationary states (black in the figure) that are separated by one unstable state (red in the figure). In case of second order autocatalysis, $n = 2$, the lower limit is built by vanishing flow rate, $r = 0$, for $n = 3, 4$, and 5 the lower limit is given by the minimum of the function $r(\bar{x})$, which coincides with $r = 0$. An increase in the values of n causes the range of bistability to shrink.

The three cases, $n = 0$,1, and $n \geq 2$, provide an illustrative example for the role of a nonlinearity in chemical reactions: The uncatalyzed reaction shows a simple decay to the stationary state with a single negative exponential function. In closed systems all autocatalytic processes have characteristic phases, consisting of a growth phase with a positive exponential at low concentration of the autocatalyst and the (obligatory) relaxation phase with a negative exponential at concentrations sufficiently close to equilibrium (figure 5.1). In the flow reactor the nonlinear systems exhibit characteristic bifurcation patterns (figure 5.2): First order autocatalysis gives rise to a rather smooth transition in the form of a transcritical bifurcation from the equilibrium branch to the state of extinctions, whereas for $n \geq 2$ the transitions are abrupt and as characteristic for a subcritical bifurcation chemical hysteresis is observed.

All cases of autocatalysis in the flow reactor ($n > 0$) discussed so far contradict a fundamental theorem of thermodynamics stating the uniqueness of the equilibrium state. Only a single steady state may occur in the limit $\lim r \to 0$. The incompatibility of the model mechanism (5.6) with basic

[7] Similarly as in the case of the time dependence in the closed system, expressed by equation (5.5c), we make use of the uncommon implicit function $r = f(\bar{x})$ than the direct relation $\bar{x} = f(r)$.

thermodynamics can be corrected by fulfilling the principle: Any catalyzed reaction requires the existence of an uncatalyzed process that approaches the same equilibrium state or, in other words, a catalyst accelerates the forward and the backward reaction by the same factor. Accordingly we have to add the uncatalyzed process to the reaction mechanism (5.6)

$$A \underset{h_1 \cdot \kappa}{\overset{k_1 \cdot \kappa}{\rightleftharpoons}} X \ . \tag{5.6a}$$

The parameter κ represents the ratio of the rate parameters of the uncatalyzed and the catalyzed reaction. In figure 5.2 we show the effect of nonzero values of κ on the bifurcation pattern. In first order autocatalysis the transcritical bifurcation disappears through a phenomenon known in linear algebra as avoided crossing: Two eigenvalues, λ_1 and λ_2 of a 2×2 matrix A plotted as functions of parameter p cross at some critical value: $\lambda_1(p_{\mathrm{cr}}) = \lambda_2(p_{\mathrm{cr}})$ avoid crossing when the variation of a second parameter, q, causes an off-diagonal element of A to change for zero to some non-zero value. Parameter p is represented by the flow rate r and parameter q by κ in the figure. The two steady states are obtained as solutions of a quadratic equation

$$\begin{aligned}\bar{x}_{1,2} \ &= \ \frac{1}{2(k_1+h_1)}\cdot \\ &\cdot \left(k_1c_0 - \kappa(k_1+h_1) - r \pm \sqrt{\big(k_1c_0 - \kappa(k_1+h_1) - r\big) + 4k_1c_0\kappa(k_1+h_1)}\right) \ .\end{aligned}$$

In the limit $\kappa \to 0$ we obtain the solutions (5.8) and in the limit of vanishing flow, $\lim r \to 0$, we find $\bar{x}_1 = k_1c0/(k_1+h_1)$ and $\bar{x}_2 = -\kappa$. As demanded by thermodynamics only one solution, $\bar{x}_1$, the equilibrium state $P_1 = (\bar{x}_1, \bar{a}_1)$ for $r = 0$ occurs within the physically meaningful domain of nonnegative concentrations whereas the second steady state $P_2 = (\bar{x}_2, \bar{a}_2)$ for $r = 0$, has a negative value of the concentration of the autocatalyst.

5.1.3 Unlimited growth

It is worth considering different classes of growth functions $y(t)$ and the behavior of long time solutions of the corresponding ODEs. An intimately related problem concerns population dynamics: What is the long time distribution of genotypes in a normalized population, $\big(x_1(t), x_2(t), \ldots, x_N(t)\big)$ with $\sum_{i=1}^{N} x_i(t) = 1$, provided the initial distribution at time $t = 0$ has been $\big(x_1(0), x_2(0), \ldots, x_N(0)\big)$? Is there a universal long time distribution that is characteristic for certain classes of growth functions?

The results presented below are obtained within the frame of the ODE model, i.e. neglecting stochastic phenomena caused by small particle numbers.

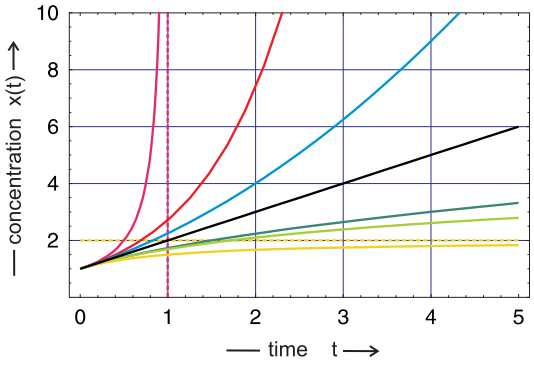

Fig. 5.4 Typical functions describing unlimited growth. All functions are normalized in order to fulfil the conditions $y(0) = 1$ and $dy/dt|_{y=0} = 1$. The individual curves show hyperbolic growth ($y(t) = 1/(1-t)$; magenta; the dotted line shows the position of the instability), exponential growth ($y(t) = \exp(t)$; red), parabolic growth ($y(t) = (1+t/2)^2$; blue), linear growth ($y(t) = 1+t$; black), sublinear growth ($y(t) = \sqrt{1+2t}$; turquoise), logarithmic growth ($y(t) = 1 + \log(1+t)$; green), and sublogarithmic growth ($y(t) = 1 + t/(1+t)$; yellow; the dotted line indicates the maximum value $y_{\max}$: $\lim_{t\to\infty} y(t) = y_{\max}$).

The differential equation describing unlimited growth,

$$\frac{dy}{dt} = f \cdot y^n \tag{5.9}$$

yields two types of general solutions for the initial value $y(0) = y_0$

$$y(t) = \left(y_0^{1-n} + (1-n)ft\right)^{1/(1-n)} \qquad \text{for } n \neq 1 \text{ and} \tag{5.9a}$$

$$y(t) = y_0 \cdot e^{ft} \qquad \text{for } n = 1\,. \tag{5.9b}$$

In order to make the functions comparable we normalize them in order to fulfil $y(0) = 1$ and $dy/dt|_{t=0} = 1$. According to equations (5.9) this yields $y_0 = 1$ and $f = 1$. The different classes of growth functions as shown in figure 5.4 are characterized by the following behavior:

(i) Hyperbolic growth requires $n > 1$; for $n = 2$ it yields the solution curve of the $y(t) = 1/(1-t)$. Characteristic is the existence of an instability in the sense that $y(t)$ approaches infinity at some critical time, $\lim_{t\to t_{\rm cr}} = \infty$ with $t_{\rm cr} = 1$. The selection behavior is illustrated by the Schlögl model: Depending on the initial conditions each of the replicators can be selected. X_m the species with the highest replication parameter, $f_{mm} = \max\{f_{ii}; i = 1, 2, \ldots, N\}$ has the largest basin of attraction. After selection has occurred a new species X_k is extremely unlikely to replace the current species even if its replication parameter

is substantially higher, $f_{kk} > f_{mm}$. We are dealing with *once for ever* selection.

(ii) Exponential growth is observed for $n = 1$ and described by the solution $y(t) = e^t$. It represents the most common growth function in biology. The species with the highest replication parameter X_m, $f_m = \max\{f_i; i = 1, 2, \ldots, N\}$, is always selected on the population level, $\lim_{t\to\infty} x_m = 1$. Injection of a new species X_k with a still higher replication parameter, $f_k > f_m$, leads to selection of the fitter variant X_k.

(iii) Parabolic growth occurs for $0 < n < 1$ and for $n = 1/2$ has the solution curve $y(t) = (1 - t/2)^2$. It is observed, for example, in enzyme free replication of oligonucleotides that form a stable duplex, i.e. a complex of one plus and one minus strand.

(iv) Linear growth follows from $n = 0$ and takes on the form $y(t) = 1 + t$. Linear growth is observed, for example, in replicase catalyzed replication at enzyme saturation.

(v) Sublinear growth occurs for $n < 0$. In particular, for $n = -1$ gives rise to the solution $y(t) = (1 + 2t)^{1/2} = \sqrt{1 + 2t}$.

In addition we mention also two additional forms of weak growth that do not follow from equation (5.9):

(vi) Logarithmic growth that can be expressed by the function $y(t) = y_0 + \ln(1 + ft)$ or $y(1) = 1 + \ln(1 + t)$ after normalization, and

(vii) sublogarithmic growth modeled by the function $y(t) = y_0 + ft/(1 + ft)$ or $y(t) = 1 + t/(1 + t)$ in normalized form.

Hyperbolic growth, parabolic growth, and sublinear growth in figure 5.4 constitute families of solution curves defined by a certain parameter range, for example a range of exponents $n_{\text{low}} < n < n_{\text{high}}$, whereas exponential growth, linear growth and logarithmic growth represent critical curves separating zones of characteristic behavior. Logarithmic growth separates growth functions approaching infinity in the limit $t \to \infty$, $\lim_{t\to\infty} y(t) = \infty$ from those that remain finite, $\lim_{t\to\infty} y(t) = y_\infty < \infty$. Linear growth separates concave from convex growth functions, and exponential growth eventually separates growth functions that reach infinity at finite times from those that don't.

We summarize this section by comparing growth behavior and characteristic dynamics of autocatalysis. Subexponential growth allows for coexistence whereas superexponential growth gives rise to selection that depends on an initial population $\boldsymbol{\Pi}(0) = \boldsymbol{\Pi}_0$. Only the intermediate case of exponential growth results in population independent selection with the Malthus parameter or the fitness of species as selection criterion. It is not accidental therefore that in terms of autocatalysis exponential growth is the result of first order autocatalysis, which in discrete time corresponds to a growth and division process – $\mathbf{X} \to \mathbf{X}^* \to 2\mathbf{X}$ with $\mathbf{X}^*$ being a cell after the internal growth phase during which the genetic material has been duplicated – and which is universal for all cells in biology.

5.2 Stochasticity in biology

In this section we shall discuss several approaches to models of stochastic processes used in biology. These model some specific insights into mechanisms that are missing in applications of the general formalisms like master and Fokker-Planck equations. In particular we shall discuss branching processes (section 5.2.1), birth-and-death processes (section 5.2.2) as well as some specific models in population biology like the Wright-Fisher process and the Moran process (section 5.2.3) before we discuss the general treatment of biological processes by master equations (section 5.3).

5.2.1 Branching processes

According to David Kendall's historical accounts on the centennial of the beginnings of stochastic thinking in population mathematics [154, 155] the name *branching process* was coined only late by Kolmogorov and Dmitriev in their 1947 paper [169]. The interest in stochasticity of the evolving populations, however, is much older. The origin of the problem is the genealogy of human males, which is reflected by the development of family names or *surnames* in the population. Commonly the stock of family names is *eroded* in the sense of steady disappearance of families in particular in small communities. The problem was clearly stated in a book by Alphonse de Candolle [29] and has been brought up by Sir Francis Galton after he had read de Candolle's book. The first rigorous mathematical analysis of a problem by means of a branching process is commonly assigned to Galton and Reverend Henry William Watson [300], the Galton-Watson process named after them has become a standard problem in branching processes. Apparently, Galton and Watson were not aware of earlier work on this topic [128], that had been performed almost thirty years before by Jules Bienaymé and was reported in a publication [18]. Most remarkable Bienaymé discussed already the criticality theorem, which expresses different behavior of the Galton-Watson process for $m < 1$, $m = 1$, and $m > 1$, where m denotes the expected or mean number of sons per father. The three cases were called *subcritical*, *critical*, and *supercritical*, respectively, by Kolmogorov [168]. Watson's original work contained a serious error in the analysis of the supercritical case and this was not detected and reported during more than fifty years before Johan Steffensen published his work on this topic [266]. In the years after 1940 the Galton-Watson model received plenty of attention because of the analogies of genealogies and nuclear chain reactions. In addition, mathematicians became generally more interested in probability theory and stochasticity. The pioneering work related to nuclear chain reactions and criticality of nuclear reactors was done by Stan Ulam at the Los Alamos National Laboratory [125, 63, 64, 65, 66]. Many other applications to biology and physics were found and branching

processes have been studied intensively. By now, it seems, we have a clear picture on the Galton-Watson process and its history [155].

5.2.1.1 The Galton-Watson process

A *Galton-Watson process* [300] deals with the generation of objects from objects of the same kind in the sense of reproduction. These objects can be neutrons, bacteria, or higher organisms, or men as in the family name genealogy problem. The Galton-Watson process is the simplest possible description of consecutive reproduction and falls into the class of branching processes. Recorded are only the population sizes of successive generations, which are considered as random variables: $\mathcal{Z}_0, \mathcal{Z}_1, \mathcal{Z}_2, \ldots$. A question of interest is the extinction of a population in generation n, and this simply means $\mathcal{Z}_n = 0$ from which follows that the random variables are zero in all future generations: $\mathcal{Z}_{n+1} = 0$ if $\mathcal{Z}_n = 0$. Indeed, the extinction or disappearance of aristocratic family names was the problem that Galton wanted to model by means of a stochastic process. In the following presentation and analysis we make use of the two books [9, 120].

In mathematical terms the Galton-Watson process is a Markov chain $(\mathcal{Z}_n;\, n \in \mathbb{N}_0)$ on the nonnegative integers. The transition probabilities are defined in terms of a given probability function $\text{Prob}\{\mathcal{Z}_1 = k\} = p_k;\, k \in \mathbb{N}_0$ with $p_k \leq 0, \sum p_k = 1$ have the

$$P(i,j) \;=\; \text{Prob}\{\mathcal{Z}_{n+1} = j | \mathcal{Z}_n = i\} \;=\; \begin{cases} p_j^{*i} & \text{if} \;\; i \geq 1,\, j \geq 0\,, \\ \delta_{0,j} & \text{if} \;\; i = 0,\, j \geq 0\,, \end{cases} \tag{5.10}$$

wherein δ_{ij} is the *Kronecker delta*[8] and $\{p_k^{*i};\, k \in \mathbb{N}_0\}$ is the i-fold convolution of $\{p_k;\, k \in \mathbb{N}_0\}$, and accordingly the probability mass function $f(k) = p_k$ is the only datum of the process. The use of the convolution of the probability distribution is an elegant mathematical trick for the rigorous analysis of the problem. Convolutions in explicit form are quite difficult to handle as we shall see in the case of the generating function. Nowadays one can use computer assisted symbolic computation but in Galton's times, in the 19th century handling of higher convolutions was quite hopeless.

The process describes an evolving population of particles or individuals and it might be useful although not necessary to define a time axis. The process starts with $\mathcal{Z}_0$ particles at time $T = 0$, each if which produces – independently

[8] The Kronecker delta is named after the German mathematician Leopold Konecker and represents the discrete analogue of Dirac's delta function:

$$\delta_{ij} \;=\; \begin{cases} 0 \;\; \text{if} \;\; i \neq j\,, \\ 1 \;\; \text{if} \;\; i = j\,. \end{cases}$$

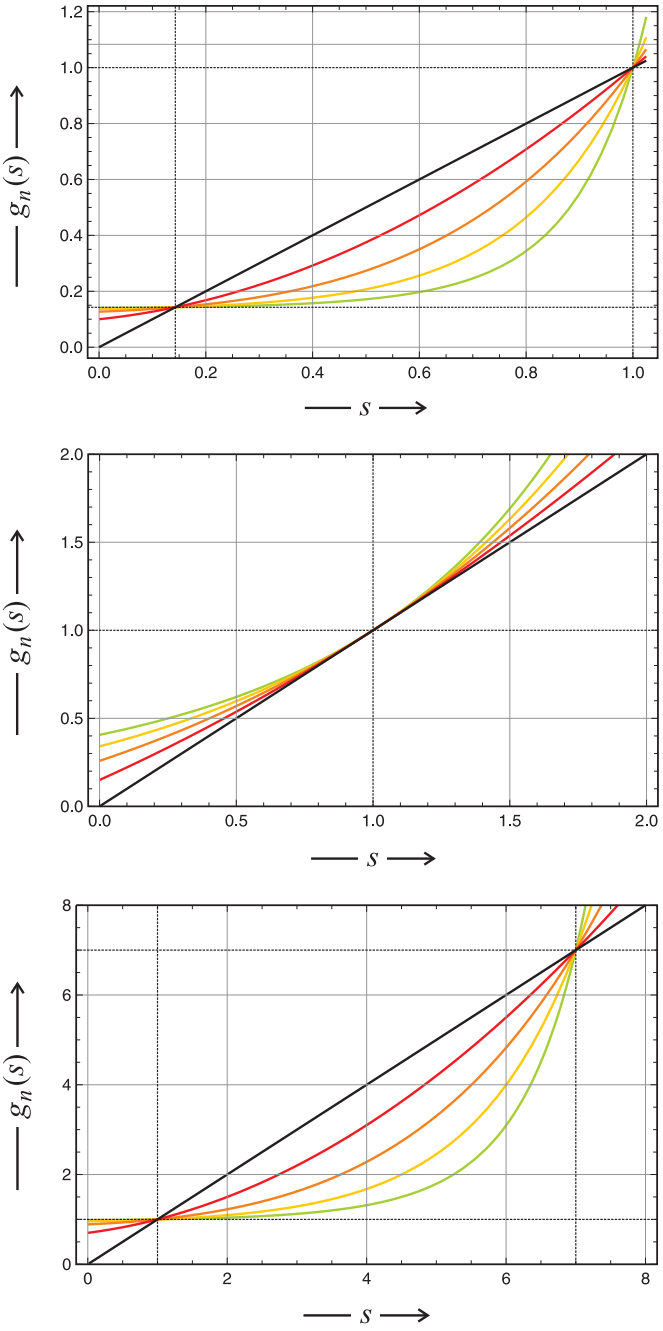

Fig. 5.5 Continued on next page.

Fig. 5.5 Calculation of extinction probabilities for the Galton-Watson process. The individual curves show the iterated generating functions of the Galton-Watson process, $g_0(s) = s$ (black), $g_1(s) = g(s) = p_0+p_1 s+p_2 s^2$ (red), $g_2(s)$ (orange), $g_3(s)$ (yellow), and $g_4(s)$ (green), for different probability densities $\mathbf{p} = (p_0, p_1, p_2)$. Choice of parameters: supercritical case (upper part) $\mathbf{p} = (0.1, 0.2, 0.7)$, $m = 1.6$; critical case (middle part) $\mathbf{p} = (0.15, 0.7, 0.15)$, $m = 1$; subcritical case (lower part) $\mathbf{p} = (0.7, 0.2, 0.1)$, $m = 0.4$.

of the others – a random number of offspring offspring at time $T = 1$ according to the probability density $f(k) = p_k$. The total number of particles in the first generation, $\mathcal{Z}_1$ is the sum of all $\mathcal{Z}_0$ random variables where each was drawn according to the pmf $f(p_k)$. The first generation produces $\mathcal{Z}_2$ particles at time $T = 2$, the second generation gives rise to the third with $\mathcal{Z}_3$ particles at time $T = 3$, and so on. Since discrete times T_n are equivalent to the numbers of generations n we shall refer only to generations ion the following. From (5.10) follows that the future development of the process at any time is independent of the history and this constitutes the Markov property.

The number of offspring produced by a single parent particle in the n-th generation is a random variable $\mathcal{Z}_n^{(1)}$ where the superscript indicates $\mathcal{Z}_0 = 1$. In general we shall write for the branching process $(\mathcal{Z}_n^{(i)};\ n \in \mathbb{N}_0)$ when we want to express that the process started with i particles. Since $i = 1$ is the by far most common case, we write simply $\mathcal{Z}_n$ instead of $\mathcal{Z}_n^{(1)}$. Equation (5.10) tells that $\mathcal{Z}_{n+k} = 0\,\forall\, k \geq 0$ if $\mathcal{Z}_n = 0$. Accordingly, the state $\mathcal{Z} = 0$ is absorbing and reaching $\mathcal{Z} = 0$ is tantamount to becoming extinct.

In order to analyze the process we shall make use of the probability generating function

$$g(s) \;=\; \sum_{k=0}^{\infty} p_k\, s^k\,,\ |s| \leq 1\ , \tag{5.11}$$

where s is complex in general but we shall assume here $s \in \mathbb{R}^1$. In addition, we define the iterates of the generating function:

$$g_0(s) \;=\; s\,,\ g_1(s) \;=\; g(s)\,,\ g_{n+1}(s) \;=\; g\big(g_n(s)\big)\,,\ n = 1, 2, \ldots\ . \tag{5.12}$$

Expressed in terms of transition probabilities the generating function is of the form

$$\sum_{j=0}^{\infty} P(1,j)\, s^j \;=\; g(s) \quad \text{and} \quad \sum_{j=0}^{\infty} P(i,j)\, s^j \;=\; \big(g(s)\big)^i\,,\ i \geq 1\ . \tag{5.13}$$

Denoting the n-step transition probability by $P_n(i,j)$ and using the Chapman-Kolmogorov equation we obtain

$$\sum_{j=0}^{\infty} P_{n+1}(1,j)\, s^j \;=\; \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} P_n(1,k)\, P(k,j)\, s^j \;=$$
$$= \sum_{k=0}^{\infty} P_n(1,k) \sum_{j=0}^{\infty} P(k,j)\, s^j \;=$$
$$= \sum_{k=0}^{\infty} P_n(1,k) \big(g(s)\big)^k \;.$$

Writing $g_{(n)} = \sum_j P_n(1,j)\, s^j$ the last equation has shown that

$$g_{(n+1)}(s) \;=\; g_{(n)}\big(g(s)\big)$$

which yields the fundamental relation

$$g_{(n)}(s) \;=\; g_n(s) \;, \tag{5.14}$$

and by making use of Equ. (5.13) we find

$$\sum_{j=0}^{\infty} P_n(i,j)\, s^j \;=\; \Big(g_n(s)\Big)^i \;. \tag{5.15}$$

Equ. (5.14) expressed as "The generating function of $\mathcal{Z}_n$ is the n-iterate $g_n(s)$", provides a tool for the calculation of the generating function. As stated in Equ. (5.10) the probability distribution of $\mathcal{Z}_n$ is obtained as the n-th convolution or iterate of $g(s)$. The explicit form of an n-th convolution is hard to compute and the true value of (5.14) lies in the calculation of the moments of $\mathcal{Z}_n$ and in the possibility to derive asymptotic laws for large n.

For the purpose of illustration we present the first iterates of the simplest useful generating function

$$g(s) \;=\; p_0 \,+\, p_1\, s \,+\, p_2\, s^2 \;.$$

The first convolution $g_2(s) = g\big(g(s)\big)$ contains ten terms already:

$$\begin{aligned} g_2(s) \;=\;& p_0 + p_0 p_1 + p_0^2 p_2 \,+\, (p_1^2 + 2p_0 p_1 p_2)\, s + \\ & + \,(p_1 p_2 + p_1^2 p_2 + 2p_0 p_2^2)\, s^2 \,+\, 2p_1 p_2^2\, s^3 \,+\, p_2^3\, s^4. \end{aligned}$$

The next convolution, $g_3(s)$, contains already nine constant terms that contribute to the probability of extinction $g_n(0)$, and $g_4(s)$ already 29 terms.

It is straightforward to compute the moments of the probability distributions from the generating function:

$$\frac{\partial g(s)}{\partial s} \;=\; \sum_{k=0}^{\infty} k\, p_k\, s^{k-1} \quad \text{and} \quad \frac{\partial g(s)}{\partial s}\,\Big|_{s=1} = \; E(\mathcal{Z}_1) \;=\; m \;, \tag{5.16a}$$

$$\frac{\partial^2 g(s)}{\partial s^2} \;=\; \sum_{k=0}^{\infty} k(k-1)\, p_k\, s^{k-2} \quad \text{and} \quad \frac{\partial^2 g(s)}{\partial s^2}\,\Big|_{s=1} = \; E(\mathcal{Z}_1^2) \;-\; m \;,$$

$$\text{var}(\mathcal{Z}_1) \;=\; \frac{\partial^2 g(s)}{\partial s^2}\,\Big|_{s=1} + m \;-\; m^2 \;=\; \sigma^2 \;. \tag{5.16b}$$

Next we calculate the moments of the distribution in higher generations and differentiate the last expression in Equ. 5.12 at $|s| = 1$:

$$\begin{aligned} \frac{\partial g_{n+1}(s)}{\partial s}\,\Big|_{s=1} \;&=\; \frac{\partial g(s)}{\partial s}\Big(g_n(s)\,\Big|_{s=1}\Big)\,\frac{\partial g_n(s)}{\partial s}\,\Big|_{s=1} = \\ &=\; \frac{\partial g(s)}{\partial s}\,\Big|_{s=1}\,\frac{\partial g_n(s)}{\partial s}\,\Big|_{s=1} \quad \text{and} \\ E(\mathcal{Z}_{n+1}) \;&=\; E(\mathcal{Z})\,E(\mathcal{Z}_n) \quad \text{or} \quad E(\mathcal{Z}_n) \;=\; m^n \;, \end{aligned} \tag{5.17}$$

by induction. Provided the second derivative of the generating function at $|s| = 1$ is finite, Equ. 5.12 can be differentiated twice:

$$\frac{\partial^2 g_{n+1}(s)}{\partial s^2}\,\Big|_{s=1} = \; \frac{\partial g(s)}{\partial s}\,\Big|_{s=1}\,\frac{\partial^2 g_n(s)}{\partial s^2}\,\Big|_{s=1} + \frac{\partial^2 g(s)}{\partial s^2}\,\Big|_{s=1}\Big(\frac{\partial g_n(s)}{\partial s}\,\Big|_{s=1}\Big)^2 \;,$$

and $\partial^2 g(s) \,/\, \partial s^2\,\big|_{s=1}$ is obtained by repeated application. The final result is:

$$\text{var}(\mathcal{Z}_n) \;=\; E(\mathcal{Z}_n^2) \;-\; E(\mathcal{Z}_n)^2 \;=\; \begin{cases} \frac{\sigma^2\, m^n\,(m^n-1)}{m\,(m-1)} & , \text{ if } m \neq 1 \\ n\,\sigma^2 & , \text{ if } m = 1 \end{cases} \;. \tag{5.18}$$

Thus, we have $E(\mathcal{Z}_n) = m^n$ and provided $\sigma = \text{var}(\mathcal{Z}_1) < \infty$ the variances are given by Equ. (5.18).

Two more assumptions are made to simplify the analysis: (i) Neither the probabilities p_0 and p_1 nor their sum are equal to one, $p_0 < 1$, $p_1 < 1$, and $p_0 + p_1 < 1$, and this implies that $g(s)$ is strictly convex on the unit interval $0 \leq s \leq 1$, and (ii) the expectation value $E(\mathcal{Z}_1) = \sum_{k=0}^{\infty} k\, p_k$ is finite, and from the finiteness of the expectation value follows $\partial g/\partial s|_{s=1}$ is finite too since $|s| \leq 1$.

Eventually we can now consider Galton's extinction problem of family names. The straightforward definition of extinction is given in terms of a random sequence $(\mathcal{Z}_n;\, n = 0, 1, 2, \ldots, \infty)$, which consists of zeros except a finite number of positive integer value at the beginning of the series. The random variable $\mathcal{Z}_n$ is integer valued and hence extinction is tantamount to the event $\mathcal{Z}_n \to 0$. From $P(\mathcal{Z}_{n+1} = 0 | \mathcal{Z}_n = 0) = 1$ follows the equality

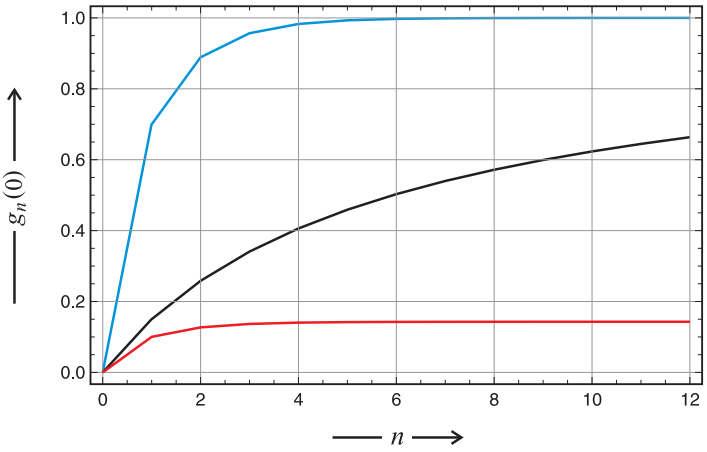

Fig. 5.6 Extinction probabilities in the Galton-Watson process. Shown are the extinction probabilities for the three Galton-Watson processes discussed in figure 5.5. The supercritical process ($\mathbf{p} = (0.1, 0.2, 0.7)$, $m = 1.6$; red) is characterized by a probability of extinction of $q = \lim g_n < 1$ leaving room for a certain probability of survival, whereas both, the critical ($\mathbf{p} = (0.15, 0.7, 0.15)$, $m = 1$; black) and the subcritical process ($\mathbf{p} = (0.7, 0.2, 0.1)$, $m = 0.4$; blue) lead to certain extinction, $q = \lim g_n = 1$. In the critical case we observe much slower convergence than in the super- or subcritical case representing a nice example of *critical slowing down*.

$$\begin{aligned} P(\mathcal{Z}_n \to 0) &= P(\mathcal{Z}_n = 0 \text{ for some } n) = \\ &= P\big((\mathcal{Z}_1 = 0) \cup (\mathcal{Z}_2 = 0) \cup \cdots \cup (\mathcal{Z}_n = 0)\big) = \\ &= \lim_{n\to\infty} P\big((\mathcal{Z}_1 = 0) \cup (\mathcal{Z}_2 = 0) \cup \cdots \cup (\mathcal{Z}_n = 0)\big) = \\ &= \lim P(\mathcal{Z}_n = 0) = \lim g_n(0) \ , \end{aligned} \tag{5.19}$$

and the fact that $g_n(0)$ is a nondecreasing function of n (see also figure 5.6).

We define a probability of extinction, $q = P(\mathcal{Z}_n \to 0) = \lim g_n(0)$ and show that $m = E(\mathcal{Z}_1) \leq 1$ the probability of extinction fulfils $q = 1$, and the family names disappear in finite time. For $m > 1$, however, the extinction probability is the unique solution less than one of the equation

$$s = g(s) \quad \text{for} \quad 0 \leq s < 1 \ . \tag{5.20}$$

It is straightforward to show by induction that $g_n(0) < 1$, $n = 0, 1, \ldots$. From Equ. (5.19) we know

$$0 = g_n(0) \leq g_1(0) \leq g_2(0) \leq \cdots \leq q = \lim g_n(0) \ .$$

Making use of the relations $g_{n+1}(0) = g\big(g_n(0)\big)$ and $\lim g_n(0) = \lim g_{n+1}(0) = q$ we derive $q = g(q)$ for $0 \leq q \leq 1$ – trivially fulfilled for $q = 1$ since $g(1) = 1$:

(i) $m \leq 1$, then $(\partial g(s)/\partial s) < 1$ for $0 \leq s < 1$. Next we use the *law of the mean*[9] express $g(s)$ in terms of $g(1)$ and for $m \leq 1$ we find $g(s) > s$ in the entire range $0 \leq s < 1$. There is only the trivial solution $q = g(q)$ with $q = 1$ and extinction is certain. □

(ii) $m > 1$, then $g(s) < s$ for s sightly less than one because $(\partial g/\partial s)|_{s=1} = m > 1$, whereas for $s = 0$ we have $g(0) > 0$ and hence we have at least one solution $s = g(s)$ in the half-open interval $[0, 1[$. Assume there were two solutions, for example s_1 and s_2 with $0 \leq s_1 < s_2 < 1$ than Rolle's theorem named after the French mathematician Michel Rolle would demand the existence of ξ and η with $s_1 < \xi < s_2 < \eta < 1$ such that $(\partial g(s)/\partial s)|_{s=\xi} = (\partial g(s)/\partial s)|_{s=\eta} = 1$ but this contradicts the fact that that $g(s)$ is strictly convex. In addition $\lim g_n(0)$ cannot be one because $(g_n(0); n = 0, 1, \ldots)$ is a nondecreasing sequence. If $g_n(0)$ were slightly less than one then $g_{n+1}(0) = g\big(g_n(0)\big)$ would be less than $g_n(0)$ and the series were decreasing. Accordingly, $q < 1$ is the unique solution of Equ. 5.20 in $[0, 1[$. □

The answer is simple and straightforward: When a father has on the average one son or less, the family name is doomed to disappear, when he has more than one son there is a finite probability of survival $0 < (1-q) < 1$, which, of course, increases with increasing expectation value m, the average number of sons. Reverend Henry William Watson correctly deduced that the extinction probability is given by a root of Equ. (5.20). He failed, however, to recognize that for $m > 1$ the relevant root is the one with $q < 1$ [91, 300]. It is remarkable that it took almost fifty years for the mathematical community to detect the error that has a drastic consequence for the result.

5.2.1.2 Reproduction and mutation as multitype branching process

The problem of reproduction and mutation has been studied in population genetics in great detail. *In vitro* evolution provided an additional access to population dynamics that can be easily traced down to the molecular level where correct replication and mutation are understood as parallel chemical reactions [17, 55, 56, 57]: Evolution of RNA molecules in cell-free replication assays and reproduction of RNA-viruses is presently understood at the same mechanistic resolution as other chemical reactions. We present here a

[9] The law of the mean expresses the difference in the values of a function $f(x)$ in terms of the derivative at one particular point $x = x_1$ and the difference in the arguments

$$f(b) \; - \; f(a) \; = \; (b-a)\,(\partial f/\partial x)|_{x=x_1} \; , \; a < x < b \; .$$

The law of the mean is fulfilled at least at one point x_1 on the arc between a and b.

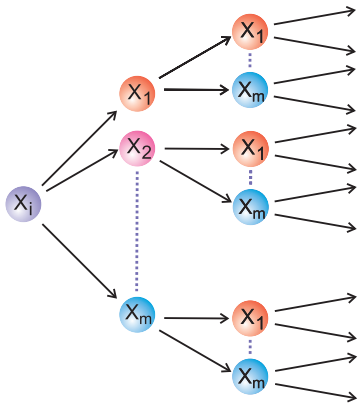

Fig. 5.7 Reproduction as a discrete multitype branching process. An individual $\mathbf{X}_i$ has progeny, $\mathbf{X}_k \in \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_m\}$, which consists of correct copies, $\mathbf{X}_i$, or mutations, $\mathbf{X}_j, j \neq i$. Reproduction is assumed to be homogeneous in time, to occur independently of the other individuals present in the population and in discrete generations. The probabilities for an individual of type $\mathbf{X}_i$ to produce γ_1 offspring of type $\mathbf{X}_1$, γ_2 offspring of type $\mathbf{X}_2$, ... , and γ_m offspring of type $\mathbf{X}_m$ is given by $P_i(\gamma_1^{(i)}, \gamma_2^{(i)}, \ldots, \gamma_i^{(i)}, \ldots \gamma_m^{(i)})$. These probabilities are independent of the generation but, of course, depend on the type of individual.

stochastic treatment of the problem as a branching process and follow the derivation and analysis in [48]. In particular, we shall consider here first the development of the population in discrete time steps corresponding to non-overlapping generations and transform later to continuous time. Eventually we compare the discrete and continuous stochastic models with their deterministic analogues: difference and differential equations.

Discrete time branching process. In the focus of these studies is the evolution of a population of N individuals chosen from m distinct classes or species $\mathbf{X}_k \in \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\}$:

$$\boldsymbol{\Pi}(n) = \vec{\mathcal{Z}}(n) = \big(\mathcal{Z}_1(n), \mathcal{Z}_2(n), \ldots, \mathcal{Z}_m(n)\big) \quad \text{with} \quad \mathcal{Z}_k \in \mathbb{N}\,, \;\; n \in \mathbb{N}\,.$$

The random variable $\mathcal{Z}_k(n)$ $(k = 1, \ldots, m)$ counts the number of individuals $\mathbf{X}_k$ in generation n $(n = 1, 2, \ldots, \infty)$. In order to model discrete time evolution we introduce multitype branching (figure 5.7) and for the purpose of illustration a simple initial conditions by assuming

$$\vec{\mathcal{Z}}(0) \;=\; \mathbf{e}_i \quad \text{with} \quad \mathbf{e}_i \;=\; (0, \ldots, 1, \ldots, 0)\,,$$

being the unit vector pointing in the direction of type $\mathbf{X}_i$. In other words, the initial condition is one individual $\mathbf{X}_i$ at generation $n = 0$. Now we define the

probability to obtain a certain distribution of species through replication and mutation in the first generation by

$$P_i^{(1)}(z_1,\ldots,z_m) \;=\; \mathrm{Prob}\Big(\mathcal{Z}_1(1)=z_1,\ldots,\mathcal{Z}_m(1)=z_m\Big)\;,$$

and analogously in the n-th generation:

$$P_i^{(n)}(z_1,\ldots,z_m) \;=\; \mathrm{Prob}\Big(\mathcal{Z}_1(n)=z_1,\ldots,\mathcal{Z}_m(n)=z_m\Big)\;. \tag{5.21}$$

Next we introduce the generating function $g^{(1)}(\mathbf{s})$ with $\mathbf{s}=(s_1,\ldots,s_m)$ being the vector of auxiliary variables and obtain for the of the first generation $\vec{\mathcal{Z}}(1)$:

$$G^{(1)}(\mathbf{s}) \;=\; g_i(\mathbf{s}) \;=\; \sum_{z_1,\ldots,z_m\geq 0} P^{(1)}(z_1,\ldots,z_m)\, s_1^{z_1}\cdots s_m^{z_m}\;. \tag{5.22}$$

The generalization is straightforward: If $\vec{\mathcal{Z}}(n)=(z_1,\ldots,z_m)$ represents the distribution of individuals at generation n, then $\vec{\mathcal{Z}}(n+1)$ is the sum of $z_1+\ldots+z_m$ random vectors, out of which z_1 have the generating function $f_1(\mathbf{s})$, z_2 the generating function $f_2(\mathbf{s})$, and so on. As typical for convolutions (see section 5.2.1.1) the explicit formula is rather lengthy and provides little additional insight, and we dispense here from showing it.

Instead, we compute a *mean matrix*, which contains the expectation values that $\mathbf{X}_i$ is obtained through replication of $\mathbf{X}_j$:

$$\mathrm{M} \;=\; \{m_{ij} \;=\; E\big(\mathcal{Z}_i(1)|\vec{\mathcal{Z}}(0)=\mathbf{e}_j\big)\;\forall\; i,j=1,\ldots,m\}\;. \tag{5.23}$$

For obvious reason that have a firm background in physics we take it for granted that the first moment exist for all i and j.[10] The matrix element m_{ij} is the mean number of $\mathbf{X}_i$ individuals derived from one type $\mathbf{X}_j$ individual within one generation and this number is readily obtained from the generating function:

$$m_{ij} \;=\; \left(\frac{\partial g_j}{\partial s_i}\right)_{s_1=\ldots=s_m=1} \quad;\quad i,j=1,\ldots,m\;. \tag{5.24}$$

In general, we are dealing with nonnegative first moments $m_{ij}=\geq 0$, and if not stated otherwise we shall assume that the matrix $\mathrm{M}=\{m_{ij}\}$ is positively regular: There exists an $n>0$ such that M^n has strictly positive elements, and M is irreducible, which implies that each type $\mathbf{X}_i$ can be derived from each type $\mathbf{X}_j$ through a finite chain of mutations.[11]

---

[10] In real systems we are always dealing with finite populations in finite time and then expectation values do not diverge (but see, for example, the unrestricted birth-and-death process in section 5.2.2).

[11] Situations may exists where it is for all practical purposes impossible to reach one population from another one through a chain of mutations in any reasonable time span. Then M is not irreducible in reality and we are dealing with two inde-

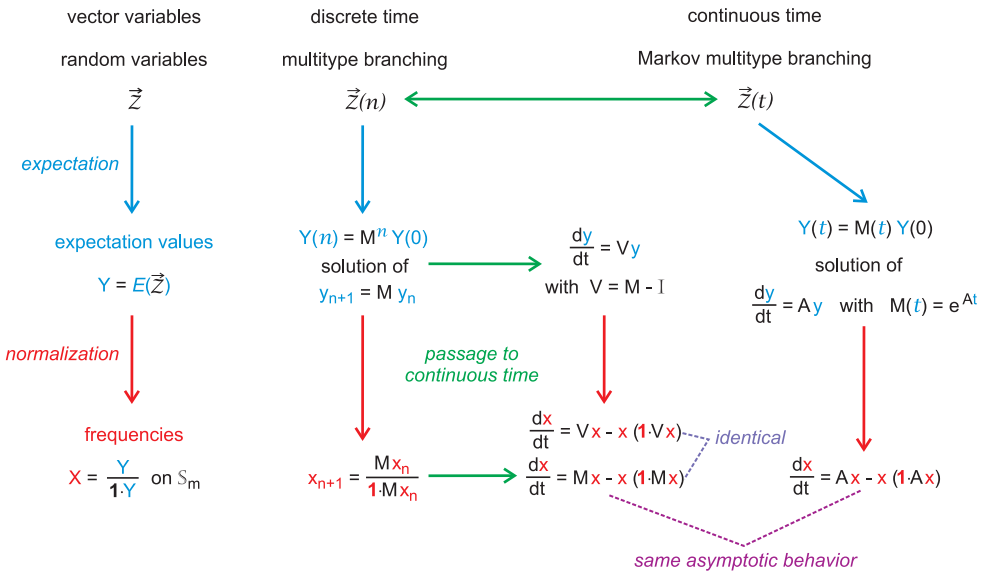

Fig. 5.8 Comparison of mutation-selection dynamics and branching processes. The sketch summarizes the different transformations discussed in text. The distinct classes of transformation are color coded: forming expectation values in blue, normalization in red and transformation between discrete and continuous variables in green (For details see text and [48]).

Perron-Frobenius theorem [258] applies to irreducible matrices M and states that the mean matrix admits a unique simple largest eigenvalue λ, which is dominant in the sense that $|\mu| < \lambda$ is fulfilled for every other eigenvalue μ of M. Since λ is non-degenerate a unique strictly positive right eigenvector, $\mathbf{u} = (u_1, \ldots, u_m)$ with $u_i > 0 \,\forall\, i = 1, \ldots, m$, and a unique strictly positive right eigenvector, $\mathbf{v} = (v_1, \ldots, v_m)$ with $v_i > 0 \,\forall\, i = 1, \ldots, m$ such that

$$\mathrm{M}\,\mathbf{u}^{\mathrm{t}} \;=\; \lambda\,\mathbf{u}^{\mathrm{t}} \quad \text{and} \quad \mathbf{v}\,\mathrm{M} \;=\; \lambda\,\mathbf{v}\;. \tag{5.25}$$

No other eigenvalue μ admits left an right eigenvector whose components are all strictly positive. The left eigenvector is normalized according to an L^1-norm and for the right eigenvector we use a peculiar *scalar product normalization*:

$$\sum_{i=1}^{m} v_i = 1 \;\text{ and }\; (\mathbf{v}, \mathbf{u}) \;=\; \mathbf{v} \cdot \mathbf{u}^{\mathrm{t}} \;=\; 1\;.$$

pendently mutating populations. In particular, when more involved mutation mechanisms comprising point mutations, deletions, and insertions are considered it may be of advantage to deal with disjoint sets of types.

The usage of the L^1-norm rather than the more familiar L^2-norm is a direct consequence of the existence of conservation laws based on addition of particle numbers or concentrations. The somewhat strange normalization has the consequence that the matrix $\mathrm{T} = \mathbf{u}^{\mathrm{t}} \cdot \mathbf{v} = \{t_{ij} = v_i u_j\}$ is idempotent or a *projection operator*:

$$\mathrm{T} \cdot \mathrm{T} \;=\; \mathbf{u}^{\mathrm{t}} \cdot \mathbf{v} \cdot \mathbf{u}^{\mathrm{t}} \cdot \mathbf{v} \;=\; \mathbf{u}^{\mathrm{t}} \cdot 1 \cdot \mathbf{v} \;=\; \mathrm{T}\,,$$

and hence we have in addition

$$\mathrm{T} \cdot \mathrm{M} \;=\; \mathrm{M} \cdot \mathrm{T} \;=\; \lambda\,\mathrm{T} \;\text{ and }\; \lim_{n\to\infty} \lambda^{-n}\,\mathrm{M}^n \;=\; \mathrm{T}\,. \tag{5.26}$$

Despite the fact that λ^n goes to zero, diverges or stays at $\lambda^n = 1$ – a situation of probability measure zero – depending on whether $\lambda < 1$, $\lambda > 1$, or $\lambda = 1$ is fulfilled, respectively.

A population is said to become extinct if $\vec{\mathcal{Z}}(n) = 0$ for some $n > 0$. We denote the probability of extinction for the initial condition $\vec{\mathcal{Z}}(n) = \mathbf{e}_i$ by q_i and define

$$q_i \;=\; \mathrm{Prob}\left(\exists\, n \text{ such that } \vec{\mathcal{Z}}(n) = 0 | \vec{\mathcal{Z}}(0) = \mathbf{e}_i\right)\,. \tag{5.27}$$

The vector $\mathbf{q} = (q_1, \ldots, q_m)$ is given by the smallest nonnegative solution of the equation

$$\mathbf{g}(\mathbf{q}) \;=\; \mathbf{q} \;\text{ or }\; \mathbf{g}(\mathbf{q}) - \mathbf{q} \;=\; 0\,, \tag{5.28}$$

where $\mathbf{g}(\mathbf{s}) = \big(g_1(\mathbf{s}), \ldots, g_m(\mathbf{s})\big)$ with the functions $g_i(\mathbf{s})$ defined by equation (5.22). The conditions for extinction can be expressed in terms of the the dominant eigenvector λ of the mean matrix M:

(i) if $\lambda \leq 1$ then $q_i = 1\,\forall\, i$ and extinction is certain,
(ii) if $\lambda > 1$ then $q_i < 1\,\forall\, i$ and there is a positive probability of survival to infinite time.

In case (ii) it is of interest to compute *asymptotic frequencies* where frequency stands for the normalized random variables

$$\mathcal{X}_i(n) \;=\; \frac{\mathcal{Z}_i(n)}{\sum_{k=1}^m \mathcal{Z}_k(n)} \;\text{ with }\; \mathcal{Z}_i(n) > 0\,\forall\, i\,. \tag{5.29}$$

If $\lambda > 1$, then there exists a random vector $\vec{\mathcal{W}} = (\mathcal{W}_1, \ldots, \mathcal{W}_m)$ and a scalar random variable w such that with probability one we have

$$\lim_{n\to\infty} \lambda^{-n}\,\vec{\mathcal{Z}}(n) \;=\; \vec{\mathcal{W}} \;\text{ and }\; \vec{\mathcal{W}} \;=\; w\,\mathbf{u}\,, \tag{5.30}$$

where $\mathbf{u}$ is the right eigenvector of M given by (5.25). Then follows that

$$\lim_{n\to\infty} \mathcal{X}_i(n) \;=\; \frac{u_i}{\sum_{k=1}^m u_k}\,. \tag{5.31}$$

holds almost always provided the population does not become extinct. Equation (5.31) states that the random variable for the frequency of type $\mathbf{X}_i$, $\mathcal{X}_i(n)$, converges almost certainly to a constant value (provided $w \neq 0$). The asymptotic behavior of the random vector $\vec{\mathcal{X}}(n)$ contrasts sharply the behavior of the total population size, $\mathcal{Z}(n) = \sum_{k=1}^{m} \mathcal{Z}_k$, and that of the population distribution $\vec{\mathcal{Z}}(n)$, which both may undergo large fluctuations accumulating in later generations because of the autocatalytic nature of the replication process. In late generations the system either has become extinct or it has grown to very large population size where in the latter case fluctuations in relative frequencies become small by the law of large numbers.

The behavior of the random variable w can be described completely by means of the results given in [158]: We have either

(i) $w = 0$ with probability one, which is always the case if $\lambda \leq 1$, or
(ii) $E\big(w|\vec{\mathcal{Z}}(0) = \mathbf{e}_i\big) = v_i$,

where v_i is the i-th component of the left eigenvector $\mathbf{v}$ of matrix M. A necessary and sufficient condition for the validity of condition (ii) is

$$E(\mathcal{Z}_j(1)\ \log \mathcal{Z}_j(1)|\vec{\mathcal{Z}}(0) = \mathbf{e}_l)\ <\ \infty\ \text{ for }\ 1 \leq i, j \leq m\ ,$$

which is a condition of finite population size that is always fulfilled in realistic systems.

Continuous time branching process. In case of intermixing generations, in particular in the case of *in vitro* evolution [152] or in absence of generation synchronizing pacemakers, the assumption of discrete generations is not justified, because any initially given synchronization is lost within a reproduction cycles. Continuous time multitype branching Markov processes offer an appropriate description in such cases but one which is technically more complicated. Since the basic results are similar to the discrete case, we will sketch them rather briefly.

For the continuous time model we suppose that an individual of type $\mathbf{X}_l$, independently of other individuals present in the population, persists for an exponentially distributed time with mean α^{-1} (see also section 4.7.3) and then generates a copy by reproduction and mutation according to a distribution whose generating function is $g_i(\mathbf{s})$. As discussed and implemented in case of chemical master equations (section 4.2.1) we assume that in a properly chosen time interval of length Δt – up to probability $o(\Delta t)$ – exactly one the following three alternatives is happening:

(i) no change,
(ii) extinction, or
(iii) survival and production of a copy of type $\mathbf{X}_j$ $(j = 1, \ldots, m)$.

The probabilities for the events (ii) and (iii) are homogeneous in time and up to some $o(\Delta t)$ proportional to Δt. As before we denote by $\mathcal{Z}_i(t)$ the number of individual of type $\mathbf{X}_i$ at time t and by $\vec{\mathcal{Z}}(t)$ the distribution of types. Again we define a mean matrix

$$\mathrm{M} \;=\; \{m_{ij} \;=\; E\big(\mathcal{Z}_i(t)|\vec{\mathcal{Z}}(0)=\mathbf{e}_j\big)\}\;, \tag{5.23'}$$

where we assume again that all first moments are finite for all $t \geq 0$. The mean matrix satisfies the semigroup and the continuity property

$$\mathrm{M}(t+u) \;=\; \mathrm{M}(t)\cdot\mathrm{M}(u) \;\text{ and }\; \lim_{t\to+0} M(t) \;=\; \mathbb{I}\,. \tag{5.32}$$

Conditions (5.32) implies the existence of a matrix A called the infinitesimal generator, which fulfils for all $t > 0$:

$$\mathrm{M}(t) \;=\; e^{\mathrm{A}t} \;\text{ with }\; \mathrm{A} \;=\; \{a_{ij} = \mu_i(b_{ij}-\delta_{ij})\} \;\text{ and }\; b_{ij} \;=\; \left(\frac{\partial g_i}{\partial s_j}\right)_{s_1^=\ldots,s_m=1}\,. \tag{5.33}$$

Again we assume that each type can produce every other type. As in the discrete time case we have $m_{ij}(t) > 0$ for $t > 0$, A is strictly positive, and Perron-Frobenius theorem holds. A has a unique dominant real eigenvalue λ with strictly positive right and left eigenvectors $\mathbf{u}$ and $\mathbf{v}$, respectively. The dominant eigenvalue of $\mathrm{M}(t)$ is $e^{\lambda t}$, again we normalize $\sum_{i=1}^m v_i = \sum_{i=1}^m u_i v_i$, and with $\mathrm{T} = \mathbf{u}^{\mathrm{t}}\cdot\mathbf{v}$ we have

$$\lim_{t\to\infty} e^{-\lambda t}\mathrm{M}(t) \;=\; \mathrm{T}\;, \tag{5.26'}$$

which guarantees the existence of finite solutions in relative particle numbers. As in the discrete case the extinction conditions are determined by λ: If $\mathbf{q} = (q_1,\ldots,q_m)$ denotes the extinction probabilities, then $\mathbf{q}$ is the unique solution of $\mathbf{g}(\mathbf{q})-\mathbf{q} = 0$, where $\mathbf{g}(\mathbf{s}) = (s_1,\ldots,s_m)$ as before, and accordingly (i) if $\lambda \leq 0$ then $q_i = 1\,\forall\, i$ and (ii) if $\lambda > 0$ then $q_i < 1\,\forall\, i$. Again we obtain

$$\lim_{t\to\infty} \mathcal{X}_i(t) \;=\; \frac{u_i}{u_1+\ldots+u_m} \tag{5.31'}$$

whenever the process does no lead to extinction.

The deterministic reference. We shortly repeat here the solutions of the deterministic problem [55, 57, 56], which is described by the differential equation

$$\frac{\mathrm{d}x_i}{\mathrm{dt}} \;=\; \sum_{j=1}^m w_{ij}x_j \;-\; x_i\left(\sum_{r=1}^m\sum_{s=1}^m w_{rs}x_s\right) \tag{5.34}$$

or in vector notation

$$\frac{\mathrm{d}\mathbf{x}^{\mathrm{t}}}{\mathrm{dt}} \;=\; \mathrm{W}\,\mathbf{x}^{\mathrm{t}} \;-\; (\mathbf{1}\cdot\mathrm{W}\,\mathbf{x}^{\mathrm{t}})\mathbf{x}^{\mathrm{t}} \tag{5.34'}$$

with $\mathrm{W} = \{w_{ij}; i,j = 1,\ldots,m\}$, $\mathbf{x} = (x_1\ldots,x_m)$, and $\mathbf{1} = (1,\ldots,1)$ restricted to the unit simplex

$$\mathbb{S}_m \;=\; \{\mathbf{x}\in\mathbb{R}^m :\, x_i \geq 0, \sum\nolimits_{j=1}^m x_j = 1\}\;. \tag{5.35}$$

The matrix W had been characterized as *value matrix* and it is commonly split into a product of a *fitness matrix* F and a *mutation matrix* Q:

$$\mathrm{W} \;=\; \begin{pmatrix} Q_{11}f_1 & Q_{12}f_2 & \dots & Q_{1m}f_m \\ Q_{21}f_1 & Q_{22}f_2 & \dots & Q_{2m}f_m \\ \vdots & \vdots & \ddots & \vdots \\ Q_{m1}f_1 & Q_{m2}f_2 & \dots & Q_{mm}f_m \end{pmatrix} \;=\; \mathrm{Q}\cdot\mathrm{F}\ .$$

The fitness matrix is a diagonal matrix whose elements are the fitness values of the individual species: $\mathrm{F} = \{f_{ij} = f_i \cdot \delta_{ij}\}$. The mutation matrix corresponds to the branching diagram in figure 5.7: $\mathrm{Q} = \{Q_{ij}\}$, where Q_{ij} is the frequency with which species $\mathbf{X}_i$ is obtained through copying $\mathbf{X}_j$. Since every copying event results either in a correct copy or a mutant we have $\sum_{i=1}^{m} Q_{ij} = 1$ and Q is a *stochastic matrix*. Some model assumptions, for example the *uniform error rate model* [274], lead to symmetric Q-matrices, which are then *bistochastic matrices*.[12] It is worth considering the second term on the right hand side of equation (5.34) in the explicit formulation

$$\mathbf{1}\cdot\mathrm{W}\,\mathbf{x}^{\mathrm{t}} \;=\; \sum_{r=1}^{m}\sum_{s=1}^{m} w_{rs}x_s \;=\; \sum_{r=1}^{m}\sum_{s=1}^{m} Q_{rs}f_s x_s \;=\; \sum_{s=1}^{m} f_s x_s \sum_{r=1}^{m} Q_{rs} \;=\; \bar{f} \;=\; \phi\ ,$$

whereby the different notation indicates two different interpretations: (i) the term $\mathbf{1}\cdot\mathrm{W}\,\mathbf{x}^{\mathrm{t}}$ is the mean excess productivity of the population, which has to be compensated in order to avoid net growth and maintaining the population normalized, $\sum_{i=1}^{m} x_i = 1$, or (ii) $\phi(t)$ is an externally controllable *dilution flux* that is suggestive of considering a flowreactor (figure 4.7). It is straightforward to check that $\mathbb{S}_m$ is invariant under (5.34): if $\mathbf{x}(0) \in \mathbb{S}_m$ then $\mathbf{x}(t) \in \mathbb{S}_m$ for all $t > 0$. Equation (5.34) was introduced as a phenomenological equation describing the kinetics of *in vitro* evolution in a flowreactor under the constraint of constant population size. Here the aim is to relate conventional replication-mutation kinetics to multitype branching processes.

Some preliminary remarks setting the stage for the comparison are:

1. The linear differential equation

$$\frac{\mathrm{d}\mathbf{y}^{\mathrm{t}}}{\mathrm{d}t} \;=\; \mathrm{W}\,\mathbf{y}^{\mathrm{t}} \quad\text{and}\quad \mathbf{x}(t) \;=\; \frac{1}{\sum_{j=1}^{m} y_j(t)}\,\mathbf{y}(t) \tag{5.36}$$

with a positive or nonnegative irreducible matrix W fulfils for $\mathbf{y}(0) \in \mathbb{R}^m_{>0}$: (i) $\mathbf{y}(t) \in \mathbb{R}^m_{>0}$ and (ii) $\mathbf{x}(t) \in \mathbb{S}_m\, \forall\, t \geq 0$ and is a solution of (5.34).

[12] The selection-mutation equation (5.34) in the original formulation [55, 57] matrix contain also a degradation term $d_j x_j$ and the corresponding definition of the value matrix reads $\mathrm{W} = \{w_{ij} = Q_{ij}f_j - d_j\}$. In case all individuals follow the same death law, $d_j = d\,\forall\, j$ the parameter d can be absorbed in the population size conservation relation and need not be considered separately.

2. As noted in the references [150, 277] equation (5.36) can be obtained from (5.34) through the transformation

$$\psi(t) \;=\; \int_0^t \phi(\tau)\mathrm{d}\tau \quad \text{and} \quad \mathbf{y}(t) \;=\; \mathbf{x}(t)\, e^{\psi(t)} \;.$$

3. Accordingly, the nonlinear equation (5.34) is easy to solve and any equilibrium of this equation must satisfy

$$\mathrm{W}\,\boldsymbol{\xi}^{\mathrm{t}} \;=\; \varepsilon\,\boldsymbol{\xi}^{\mathrm{t}}$$

and therefore be a right eigenvector of W. By Perron-Frobenius there exists such a unique right eigenvector in $\mathbb{S}_m$, which we denoted by $\boldsymbol{\xi}$ and the corresponding eigenvalue ε is just the dominant eigenvalue of W. From the correspondence between equations (5.34) and (5.36) follows that all orbits of equations (5.34) converge to $\boldsymbol{\xi}$: $\lim_{t\to\infty} \mathbf{x}(t) = \boldsymbol{\xi}$.

4. Now we use the canonical way to associate difference and differential equations:

$$\mathbf{v}_{n+1} \;=\; \mathbf{F}(\mathbf{v}_n) \quad\Longleftrightarrow\quad \frac{\mathrm{d}\mathbf{v}^{\mathrm{t}}}{\mathrm{dt}} \;=\; \mathbf{F}(\mathbf{v})^{\mathrm{t}} \;-\; \mathbf{v}^{\mathrm{t}} \;. \tag{5.37}$$

Of course, such an unreflected *passage to continuous time* is not always justifiable, but for a generation length one the difference equation $\mathbf{v}_{n+1} - \mathbf{n} = \mathbf{F}(\mathbf{v}_n) - \mathbf{v}_n$ can be written as

$$\mathbf{v}(1) \;-\; (0) \;=\; \mathbf{F}\big(\mathbf{v}(0)\big) \;-\; \mathbf{v}(0) \;.$$

Provided we assume blending generations the change during the time interval $1/n$, $\mathbf{v}(1/n) - \mathbf{v}(0)$ can be approximated by $\big(\mathbf{F}\big(\mathbf{v}(0)\big) - \mathbf{v}(0)\big)/n$, or

$$\frac{\mathbf{v}(\Delta t) \;-\; \mathbf{v}(0)}{\Delta t} \;=\; \mathbf{F}\big(\mathbf{v}(0)\big) \;-\; \mathbf{v}(0) \;,$$

which in the limit $\Delta t \to 0$ yields the differential equation (5.37).

The relationship between branching processes and the mutation-selection equation 5.34 is sketched in figure 5.8. If we start out from the discrete multitype branching process $\vec{\mathcal{Z}}(n)$, then the expectation values $\mathbf{Y}(n) = E\big(\vec{\mathcal{Z}}(n)\big)$ satisfy $\mathbf{Y}(n)^{\mathrm{t}} = \mathrm{M}^n\mathbf{Y}(0)^{\mathrm{t}}$, where M is the mean matrix (5.23), and hence $\mathbf{Y}(n)$ is obtained by iteration from the difference equation $\mathbf{y}^{\mathrm{t}}_{n+1} = M\mathbf{y}^{\mathrm{t}}_n$. From here one can reach the Mutation-selection equation in two ways: (i) by first passing to continuous time as expressed by the differential equation $\mathrm{d}\mathbf{y}^{\mathrm{t}}/\,\mathrm{dt} = \mathrm{V}\mathbf{y}^{\mathrm{t}}$ with $\mathrm{V} = \mathrm{M} - \mathbb{I}$ followed by normalization, which yields

$$\frac{\mathrm{d}\mathbf{x}^{\mathrm{t}}}{\mathrm{dt}} \;=\; \mathrm{V}\,\mathbf{x}^{\mathrm{t}} \;-\; \mathbf{x}^{\mathrm{t}}(\mathbf{1}\cdot\mathrm{V}\mathbf{x}^{\mathrm{t}}) \;, \tag{5.38}$$

or (ii) in opposite sequence by first normalizing the difference equation

$$\mathbf{x}_{n+1}^{\mathrm{t}} \;=\; \frac{1}{\mathbf{1}\cdot \mathrm{M}\mathbf{x}_n^{\mathrm{t}}}\,\mathrm{M}\mathbf{x}_n^{\mathrm{t}}$$

on $\mathbb{S}_m$, and then passing to continuous time yields

$$\frac{\mathrm{d}\mathbf{x}^{\mathrm{t}}}{\mathrm{dt}} \;=\; \big(\mathrm{M}\mathbf{x}^{\mathrm{t}} \,-\, \mathbf{x}^{\mathrm{t}}(\mathbf{1}\cdot \mathrm{M}\mathbf{x}^{\mathrm{t}})\big)\,\frac{1}{\mathbf{1}\cdot \mathrm{M}\mathbf{x}^{\mathrm{t}}}\;. \tag{5.39}$$

Multiplication with the factor $\mathbf{1}\cdot \mathrm{M}\mathbf{x}^{\mathrm{t}}$, which is independent of i and always strictly positive on $\mathbb{S}_m$, results merely in a transformation of the time axis that is tantamount to a change in the velocity, and the orbits of (5.39) are the same as those of

$$\frac{\mathrm{d}\mathbf{x}^{\mathrm{t}}}{\mathrm{dt}} \;=\; \mathrm{M}\mathbf{x}^{\mathrm{t}} \,-\, \mathbf{x}^{\mathrm{t}}(\mathbf{1}\cdot \mathrm{M}\mathbf{x}^{\mathrm{t}})\;. \tag{5.39'}$$

Since $\mathrm{V} = \mathrm{M} - \mathbb{I}$, the two equations (5.34) and (5.39') are identical on $\mathbb{S}_m$.

Alternatively we begin now with a continuous Markovian multitype branching process $\vec{\mathcal{Z}}(t)$ for $t \geq 0$ and either reduce it by discretization to the discrete branching process $\vec{\mathcal{Z}}(n)$, or else we obtain $\mathbf{Y}(t)^{\mathrm{t}} = \mathrm{M}(t)\mathbf{Y}(0)^{\mathrm{t}}$ for the expectation values $\mathbf{Y}(t) = E\big(\vec{\mathcal{Z}}(t)\big)$, where $\mathrm{M}(t)$ is again the mean matrix with $\mathrm{M}(1) = \mathrm{M}$. The expectation value $\mathbf{Y}(t)$ is then the solution of the linear differential equation

$$\frac{\mathrm{d}\mathbf{y}^{\mathrm{t}}}{\mathrm{dt}} \;=\; \mathrm{A}\mathbf{y}^{\mathrm{t}} \;\text{ with }\; \mathrm{A} = \lim_{t\to+0}\frac{\mathrm{M}(t)-\mathbb{I}}{t)}t \tag{5.40}$$

as infinitesimal generator of the semigroup $\mathrm{M}(t)$, and $\mathrm{M}(t) = e^{\mathrm{A}t}$. Normalization leads to

$$\frac{\mathrm{d}\mathbf{x}^{\mathrm{t}}}{\mathrm{dt}} \;=\; \mathrm{A}\mathbf{x}^{\mathrm{t}} \,-\, \mathbf{x}^{\mathrm{t}}(\mathbf{1}\cdot \mathrm{A}\mathbf{x}^{\mathrm{t}}) \;\text{ on }\; \mathbb{S}_m\;. \tag{5.41}$$

This equation in general has a dynamics that is different from (5.39'), but the asymptotic behavior is the same, because A and $\mathrm{M} = e^{\mathrm{A}}$ have the same eigenvectors and accordingly $\mathbf{u}$ is the global attractor for both equations (5.39') and (5.41).

Three simple paths lead from branching processes to an essentially unique version of the mutation-selection equation (5.34) and the question whether or not such a reduction from a stochastic to a deterministic system is relevant. A superficial analysis may suggest that it isn't. Passing from the random variables $\mathcal{Z}_i(n)$ $(i = 1, \ldots, m)$ to the expectation values $E\big(\mathcal{Z}_i(n)\big)$ may be misleading because the variances grow too fast as can be easily verified for one-type branching. If μ and σ are the mean and the variance of a single individual in the first generation, $\mu = E\big(\mathcal{Z}(1)\big)$ and $\sigma^2 = \mathrm{var}\big(\mathcal{Z}(1)\big)$ then mean and variance of the n-generation grow in the supercritical case like

$$m^n \quad \text{and} \quad \sigma^2 \frac{m^n(m^n-1)}{m(m-1)} \;=\; \sigma^2 \sum_{k=n-1}^{2n-2} m^k \;,$$

respectively, and the ratio from the standard deviation and the mean converges to a positive constant,

$$\frac{\sqrt{\mathrm{var}\big(\mathcal{Z}(n)\big)}}{E\big(\mathcal{Z}(n)\big)} \;=\; \sqrt{\sum_{k=n-3}^{2n-4} m^k} \;.$$

Accordingly, the *window* of probable values of the random variable $\mathcal{Z}(n)$ is rather large. For a critical process the situation is still worse: the means remains constant whereas the variance grows to infinity (see figure 5.11). In case of multitype branching the situation is similar but the expressions for variance and correlations get rather complicated and again the second moments grow so fast that the averages tell precariously little about the process (see [119] for the discrete and [9] for the continuous process).

Normalization, however, changes the situation: The transition from expectation values to relative frequencies cancels the fluctuations or more precisely, if the process does not go to extinction, the relative frequencies of the random variables

$$\mathcal{X}_i \;=\; \frac{\mathcal{Z}_i}{\mathcal{Z}_1 + \ldots + \mathcal{Z}_m}$$

converge almost certainly to the value u_1 $(i = 1, \ldots, m)$, which are – at the same time the limits of the relative frequencies of the expectation values

$$x_i \;=\; \frac{y_i}{y_1 + \ldots + y_m} \;.$$

In this sense, the deterministic mutation-selection equation (5.34) yields a description of the stochastic evolution of the internal structure of the population, which is much more reliable than the dynamics of the unnormalized means. The qualitative features of the selection process condense the *variance free* part of the deterministic approach.

Finally, we mention other attempts to find stochastic solutions to the replication-mutation problem [127, 138, 151, 200].

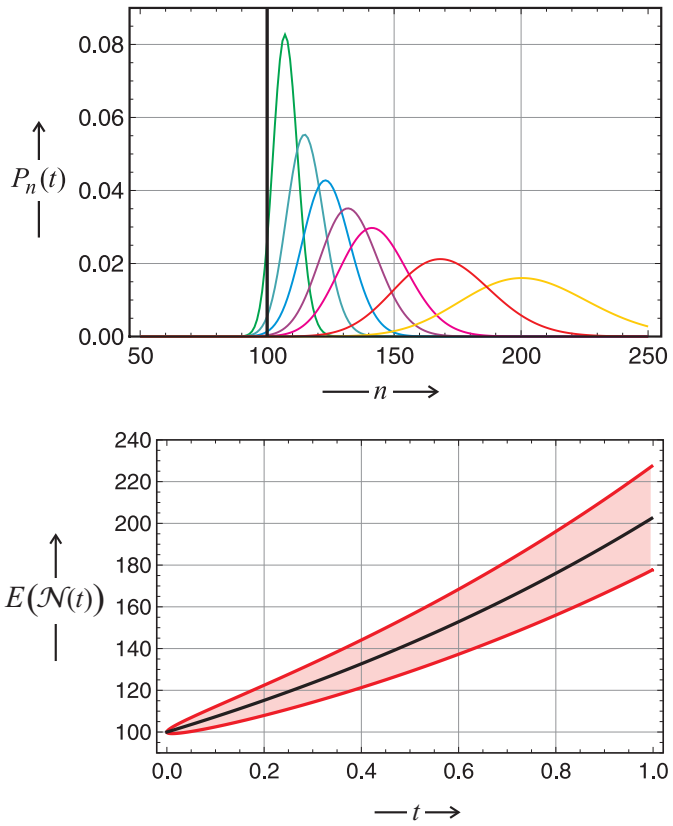

Fig. 5.9 A growing linear birth-and-death process. The two-step reaction mechanism of the process is $(\mathbf{X} \to 2\mathbf{X}, \mathbf{X} \to \varnothing)$ with rate parameters λ and μ, respectively. The growing or supercritical process is characterized by $\lambda > \mu$. The upper part shows the evolution of the probability density, $P_n(t) = \text{Prob}\,\mathcal{X}(t) = n$. The initially infinitely sharp density, $P(n,0) = \delta(n, n_0)$ becomes broader with time and flattens as the variance increases with time. In the lower part we show the expectation value $E\big(\mathcal{N}(t)\big)$ in the confidence interval $E \pm \sigma$. Parameters used: $n_0 = 100$, $\lambda = \sqrt{2}$, and $\mu = 1/\sqrt{2}$; sampling times (upper part): $t = 0$ (black), 0.1 (green), 0.2 (turquoise), 0.3 (blue), 0.4 (violet), 0.5 (magenta), 0.75 red), and 1.0 (yellow).

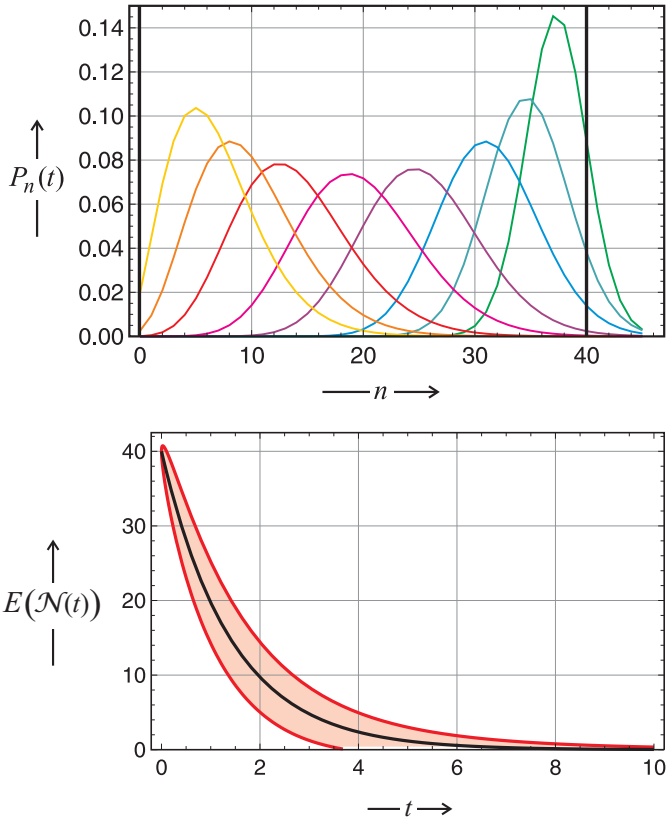

Fig. 5.10 A decaying linear birth-and-death process. The two-step reaction mechanism of the process is ($\mathbf{X} \to 2\mathbf{X}$, $\mathbf{X} \to \varnothing$) with rate parameters λ and μ, respectively. The decaying or subcritical process is characterized by $\lambda < \mu$. The upper part shows the evolution of the probability density, $P_n(t) = \text{Prob}\,\mathcal{X}(t) = n$. The initially infinitely sharp density, $P(n, 0) = \delta(n, n_0)$ becomes broader with time and flattens as the variance increases but then sharpens again as process approaches the absorbing barrier at $n = 0$. In the lower part we show the expectation value $E\big(\mathcal{N}(t)\big)$ in the confidence interval $E \pm \sigma$. Parameters used: $n_0 = 40$, $\lambda = 1/\sqrt{2}$, and $\mu = \sqrt{2}$; sampling times (upper part): $t = 0$ (black), 0.1 (green), 0.2 (turquoise), 0.35 (blue), 0.65 (violet), 1.0 (magenta), 1.5 red), 2.0 (orange), 2.5 (yellow), and $\lim_{t\to\infty}$ (black).

5.2.2 Birth-and-death processes

In section 3.2.5.2 we discussed the concept of birth-and-death processes in relation to the application of master equations in chemistry. Here, we shall come back to the original biological idea of birth and death of individuals but retain the close relation to master equations. In particular, we shall discuss the nature and influence of boundary conditions on birth-and-death processes, demonstrate the usefulness of first passage times and related concepts for finding straightforward answer to frequently asked questions, and present a collection of analytical results in table form [108].

5.2.2.1 The linear birth-and-death process

Reproduction of individuals is modeled by a simple duplication mechanism and death is represented by first order decay. In the language of chemical kinetics these two steps are:

$$\mathbf{A} + \mathbf{X} \xrightarrow{\lambda} 2\,\mathbf{X}\,, \tag{5.42a}$$

$$\mathbf{X} \xrightarrow{\mu} \mathbf{B}\,. \tag{5.42b}$$

The rate parameters for reproduction and extinction are denoted by λ and μ, respectively.[13] The material required for reproduction is assumed to be replenished as it is consumed and hence the amount of **A** available is constant and assumed to be included in the birth parameter: $\lambda = f \cdot [\mathbf{A}]$. The degradation product **B** does not enter the kinetic equation because reaction (5.42b) is irreversible. The stochastic process corresponding to equations (5.42) belongs to the class of linear birth-and-death processes with $w_n^+ = \lambda \cdot n$ and $w_n^- = \mu \cdot n$.[14] The master equation is of the form,

$$\frac{\partial P_n(t)}{\partial t} = \lambda\,(n-1)\,P_{n-1}(t) + \mu\,(n+1)\,P_{n+1}(t) - (\lambda+\mu)\,n\,P_n(t)\,, \tag{5.43}$$

and after introduction of the probability generating function $g(s,t)$ gives rise to the PDE

[13] Reproduction is to be understood a asexual reproduction here. Sexual reproduction, of course, requires two partners and gives rise to a process of order 2 (table 4.1).

[14] Here we use the symbols commonly applied in biology: $\lambda_{(n)}$ for birth, $\mu_{(n)}$ for death, and ν for immigration and ρ for emigration (tables 5.1 and 5.2). These notions were created especially for application to biological problems, in particular for problems in theoretical ecology. Other notions and symbols are common in chemistry: A birth corresponds to the production of a molecule, $f \equiv \lambda$, a death to its decomposition or degradation through a chemical reaction, $d \equiv \mu$. Influx and outflux are the proper notions for immigration and emigration.

$$\frac{\partial g(s,t)}{\partial t} \;-\; (s-1)\,(\lambda s-\mu)\,\frac{\partial g(s,t)}{\partial s} \;=\; 0\;. \tag{5.44}$$

Solution of this PDE yields different results for equal or different replication and extinction rate coefficients, $\lambda \neq \mu$ and $\lambda = \mu$, respectively. In the first case we substitute $\gamma = \lambda/\mu$ $(\neq 1)$ and $\eta(t) = \exp\bigl((\lambda-\mu)t\bigr)$, and find:

$$\begin{aligned} g(s,t) \;&=\; \left\{\frac{\bigl(\eta(t)-1\bigr)\;+\;\bigl(\gamma-\eta(t)\bigr)\,s}{\bigl(\gamma\eta(t)-1\bigr)\;+\;\gamma\,\bigl(1-\eta(t)\bigr)\,s}\right\}^{n_0} \quad \text{and} \\ P_n(t) \;&=\; \gamma^n \sum_{m=0}^{\min(n,n_0)} (-1)^m \binom{n_0+n-m-1}{n-m}\binom{n_0}{m} \times \\ &\qquad \times \left(\frac{1-\eta(t)}{1-\gamma\eta(t)}\right)^{n_0+n-m} \left(\frac{\gamma-\eta(t)}{\gamma\bigl(1-\eta(t)\bigr)}\right)^{m}\;. \end{aligned} \tag{5.45}$$

In the derivation of the expression for the probability distributions we expanded enumerator and denominator of the expression in the generating function $g(s,t)$, by using expressions for the sums $(1+s)^n = \sum_{k=0}^{n}\binom{n}{k}s^k$ and $(1+s)^{-n} = 1+\sum_{k=1}^{\infty}(-1)^k\,\frac{n(n+1)\ldots(n+k-1)}{k!}\,s^k$, multiply, order terms with respect to powers of s, and compare with the expansion of the generating function, $g(s,t) = \sum_{n=0}^{\infty} P_n(t)\,s^n$.

Computations of expectation value and variance are straightforward:

$$\begin{aligned} E\bigl(\mathcal{N}_X(t)\bigr) \;&=\; n_0\,e^{(\lambda-\mu)\,t} \quad \text{and} \\ \sigma^2\bigl(\mathcal{N}_X(t)\bigr) \;&=\; n_0\,\frac{\lambda+\mu}{\lambda-\mu}\,e^{(\lambda-\mu)\,t}\,\Bigl(e^{(\lambda-\mu)\,t}-1\Bigr) \end{aligned} \tag{5.46}$$

Illustrative examples of linear birth-and-death processes with growing ($\lambda > \mu$) and decaying ($\lambda < \mu$) populations are shown in figures 5.9 and 5.10, respectively.

In the degenerate case of *neutrality* with respect to growth, $\mu = \lambda$, the same procedure yields:

$$g(s,t) \;=\; \left(\frac{\lambda t \;+\; (1-\lambda t)\,s}{1+\lambda t \;+\; \lambda t\,s}\right)^{n_0}\;, \tag{5.47a}$$

$$P_n(t) \;=\; \left(\frac{\lambda t}{1+\lambda t}\right)^{n_0+n} \sum_{m=0}^{\min(n,n_0)} \binom{n_0+n-m-1}{n-m}\binom{n_0}{m}\left(\frac{1-\lambda^2t^2}{\lambda^2t^2}\right)^{m}\;, \tag{5.47b}$$

$$E\bigl(\mathcal{N}_X(t)\bigr) \;=\; n_0\,,\ \text{and} \tag{5.47c}$$

$$\sigma^2\bigl(\mathcal{N}_X(t)\bigr) \;=\; 2\,n_0\,\lambda t\;. \tag{5.47d}$$

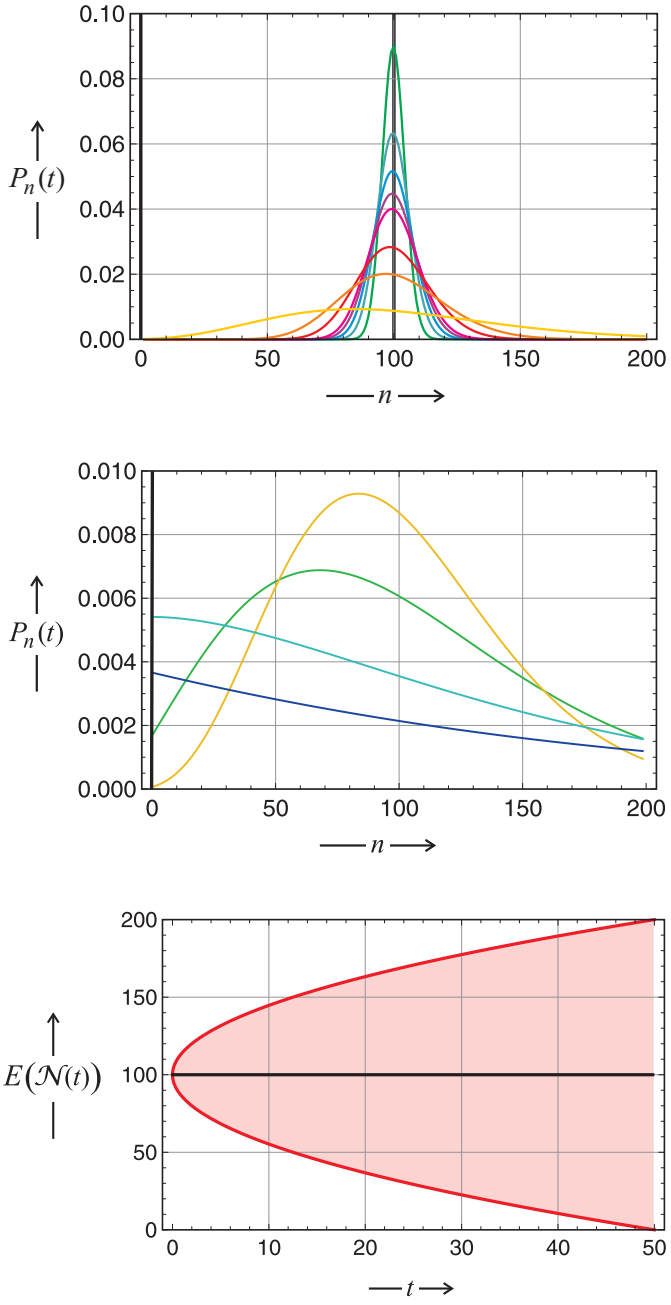

Fig. 5.11 Continued on next page.

Fig. 5.11 Probability density of a linear birth-and-death with equal birth and death rate. The two-step reaction mechanism of the critical process is ($\mathbf{X} \to 2\mathbf{X}$, $\mathbf{X} \to \varnothing$) with rate parameters $\lambda = \mu$. The upper and the middle part show the evolution of the probability density, $P_n(t) = \text{Prob}\big(\mathcal{X}(t) = n\big)$. The initially infinitely sharp density, $P(n,0) = \delta(n,n_0)$ becomes broader with time and flattens as the variance increases but then sharpens again as the process approaches the absorbing barrier at $n = 0$. In the lower part, we show the expectation value $E\big(\mathcal{N}(t)\big)$ in the confidence interval $E \pm \sigma$. The variance increases linearly with time and at $t = n_0/(2\lambda) = 50$ the standard deviation is as large as the expectation value. Parameters used: $n_0 = 100$, $\lambda = 1$; sampling times, upper part: $t = 0$ (black), 0.1 (green), 0.2 (turquoise), 0.3 (blue), 0.4 (violet), 0.49999 (magenta), 0.99999 (red), 2.0 (orange), 10 (yellow), and middle part: $t = 10$ (yellow), 20 (green), 50 (cyan), 100 (blue), and $\lim_{t\to\infty}$ (black).

Comparison of the last two expressions shows the inherent instability of this reaction system. The expectation value is constant whereas the fluctuations increase with time. The degenerate birth-and-death process is illustrated in figure 5.11. The case of steadily increasing fluctuations is in contrast to an equilibrium situation where both, expectation value and variance approach constant values. Recalling the Ehrenfest urn game, where fluctuations were negatively correlated with the deviation from equilibrium, we have here two uncorrelated processes, replication and extinction. The particle number n fulfils a kind of random walk on the natural numbers, and indeed in case of the random walk (see equation (3.75) in subsection 3.2.3.6 we had also obtained a constant expectation value $E = n_0$ and a variance that increases linearly with time, $\sigma^2(t) = 2\vartheta(t - t_0)$).

5.2.2.2 Sequential extinction times

A constant expectation value accompanied by a variance that increases with time has an easy to recognize consequence: At some critical time above which the standard deviation exceeds the expectation, $t_{\text{cr}} = n_0 \,/\, (2\lambda)$. From this instant on predictions on the evolution of the system based on the expectation value become obsolete. Then we have to rely on individual probabilities or other quantities. Useful in this context is the probability of extinction of all particles, which can be readily computed:

$$P_0(t) \;=\; \left(\frac{\lambda t}{1+\lambda t}\right)^{n_0} \quad . \tag{5.48}$$

Provided we wait long enough, the system will die out with probability one, since we have $\lim_{t\to\infty} P_0(t) = 1$. This seems to be a contradiction to the constant expectation value. As a matter of fact it is not: In almost all individual runs the system will go extinct, but there are very few cases of probability

measure zero where the particle number grows to infinity for $t \to \infty$. These rare cases are responsible for the finite expectation value.

Equation (5.48) can be used to derive a simple model for *random selection* [256]. We assume a population of n different species

$$\mathbf{A} + \mathbf{X}_j \xrightarrow{\lambda} 2\mathbf{X}_j \ , \ j = 1, \ldots, n \ , \tag{5.42a'}$$

$$\mathbf{X}_j \xrightarrow{\mu} \mathbf{B} \ , \quad j = 1, \ldots, n \ . \tag{5.42b'}$$

The probability joint distribution of the population is described by

$$P_{x_1 \ldots x_n} \ = \ P\big(\mathcal{X}_1(t) = x_1, \ldots, \mathcal{X}_n(t) = x_n\big) \ = \ P_{x_1}^{(1)} \cdot \ldots \cdot P_{x_n}^{(n)} \ , \tag{5.49}$$

wherein all probability distribution for individual species are given by equation (5.47b) and independence of individual birth events as well as death events allows for the simple product expression. In the spirit of Motoo Kimura's neutral theory of evolution [164] all birth and all death parameters are assumed to be equal, $\lambda_j = \lambda$ and $\mu_j = \mu$ for all $j = 1, \ldots, n$. For convenience we assume that every species is initially present in a single copy: $P_{n_j}(0) = \delta_{n_j,1}$. We introduce a new random variable that has the nature of a first passage time: $\mathcal{T}_k$ is the time up to the extinction of $n - k$ species and characterize it as *sequential extinction time*. Accordingly, n species are present in the population between $\mathcal{T}_n$, which fulfils $\mathcal{T}_n \equiv 0$ by definition, $\mathcal{T}_{n-1}$, $n-1$ species between $\mathcal{T}_{n-1}$ and $\mathcal{T}_{n-2}$, and eventually a single species between $\mathcal{T}_1$ and $\mathcal{T}_0$, which is the moment of extinction of the entire population. After $\mathcal{T}_0$ no particle $\mathbf{X}$ exists any more.

Next we consider the probability distribution of the sequential extinction times

$$H_k(t) \ = \ P(\mathcal{T}_k < t) \ . \tag{5.50}$$

The probability of extinction of the population is readily calculated: Since individual reproduction and extinction events are independent we find

$$H_0 \ = \ P_{0,\ldots,0} \ = \ P_0^{(1)} \cdot \ldots \cdot P_0^{(n)} \ = \ \left(\frac{\lambda t}{1 + \lambda t}\right)^n \ .$$

The event $\mathcal{T}_1 < t$ can happen in several ways: Either $\mathbf{X}_1$ is present and all other species have become extinct already, or only $\mathbf{X}_2$ is present, or only $\mathbf{X}_3$, and so on, but $\mathcal{T}_1 < t$ is also fulfilled if the whole population has died out:

$$H_1 \ = \ P_{x_1 \neq 0, 0, \ldots, 0} + P_{0, x_2 \neq 0, \ldots, 0} + P_{0,0,\ldots,x_n \neq 0} + H_0 \ .$$

The probability that a given species has not yet disappeared is obtained by exclusion since existence and nonexistence are complementary,

$$P_{x \neq 0} \ = \ 1 - P_0 \ = \ 1 - \frac{\lambda t}{1 + \lambda t} \ = \ \frac{1}{1 + \lambda t} \ ,$$

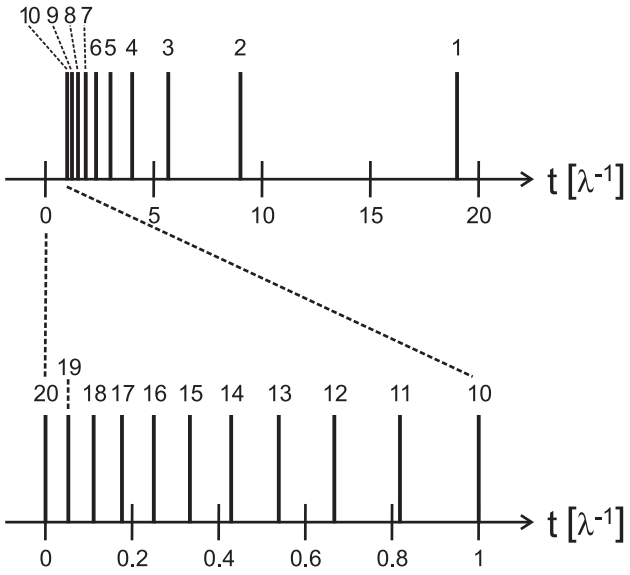

Fig. 5.12 The distribution of sequential extinction times $\mathcal{T}_k$. Shown are the expectation values $E(\mathcal{T}_k)$ for $n = 20$ according to equation(5.51). Since $E(\mathcal{T}_0)$ diverges, $\mathcal{T}_1$ is the extinction that appears on the average at a finite value. A single species is present above $\mathcal{T}_1$ and random selection has occurred in the population.

which yields the expression for the presence of a single species

$$H_1(t) \;=\; (n+\lambda t)\frac{(\lambda t)^{n-1}}{(1+\lambda t)^n}\;,$$

and by similar arguments a recursion formula is found for the extinction probabilities with higher indices

$$H_k(t) \;=\; \binom{n}{k}\frac{(\lambda t)^{n-k}}{(1+\lambda t)^n} \;+\; H_{k-1}(t)\;,$$

that eventually leads to the expression

$$H_k(t) \;=\; \sum_{j=0}^{k}\binom{n}{j}\frac{(\lambda t)^{n-j}}{(1+\lambda t)^n}\;.$$

The moments of the sequential extinction times are computed straightforwardly by means of a handy trick: H_k is partitioned into terms for the individual powers of λt, $H_k(t) = \sum_{j=0}^{k} h_j(t)$ and then differentiated with respect to time t

$$h_j(t) \;=\; \binom{n}{j}\frac{(\lambda t)^{n-k}}{(1+\lambda t)^n}\;,$$

$$\frac{\mathrm{d}h_j(t)}{\mathrm{dt}} \;=\; h_j' \;=\; \frac{\lambda}{(1+\lambda t)^{n+1}}\left(\binom{n}{j}(n-j)(\lambda t)^{n-j-1} \;-\; \binom{n}{j}j(\lambda t)^{n-j}\right)\;.$$

The summation of the derivatives is simple because $h_k' + h_{k-1}' + \ldots + h_0'$ is a telescopic sum and we find

$$\frac{\mathrm{d}H_k(t)}{\mathrm{dt}} \;=\; \binom{n}{k}(n-k)\,\lambda^{n-k}\,\frac{t^{n-k-1}}{(1+\lambda t)^{n+1}}\;.$$

Making use of the definite integral [109, p.338]

$$\int_0^\infty \frac{t^{n-k}}{(1+\lambda t)^{n+1}}\,\mathrm{dt} \;=\; \frac{\lambda^{-(n-k+1)}}{\binom{n}{k}\,k}\;,$$

we finally obtain for the expectation values of the sequential extinction times

$$E(\mathcal{T}_k) \;=\; \int_0^\infty \frac{\mathrm{d}H_k(t)}{\mathrm{dt}}\,t\;\mathrm{dt} \;=\; \frac{n-k}{k}\cdot\frac{1}{\lambda}\,,\; n \geq k \geq 1\;, \tag{5.51}$$

and $E(\mathcal{T}_0) = \infty$ (see figure). It is worth recognizing here another paradox of probability theory: Although extinction is certain, the expectation value for the time to extinction diverges. Similarly as the expectation values, we calculate the variances of the sequential extinction times:

$$\sigma^2(\mathcal{T}_k) \;=\; \frac{n(n-k)}{k^2(k-1)}\cdot\frac{1}{\lambda^2}\,,\; n \geq k \geq 2\;, \tag{5.52}$$

from which we see that the variances diverges for $k = 0$ and $k = 1$.

For distinct birth parameters, $\lambda_1, \ldots, \lambda_n$, and different initial particle numbers, $x_1(0), \ldots, x_n(0)$, the expressions for the expectation values become considerably more complicated, but the main conclusion remains unaffected: $E(\mathcal{T}_1)$ is finite whereas $E(\mathcal{T}_0)$ diverges.

5.2.2.3 Boundaries of birth-and-death processes

One step birth-and-death processes have been studied extensively and analytical solutions are available in table form [108]. For transition probabilities at most linear in n, $w_n^+ = \nu + \lambda n$ and $w_n^- = \rho + \mu n$, one distinguishes birth (λ), death (μ), immigration (ν), and emigration (ρ) terms. Analytical solutions for the probability distributions were derived for all one step birth-and-death processes whose transitions probabilities are constant or maximally linear in the numbers of individuals n.

It is necessary, however, to consider also the influence of boundaries on these stochastic processes. For this goal we define an interval $[a, b]$ as domain of the stochastic variable $\mathcal{N}(t)$. Here we are dealing with classes of boundary conditions, *absorbing* and *reflecting boundaries*. In the former case, a particle that left the interval is not allowed to return, whereas the latter class of boundary implies that it is forbidden to exit from the interval. Boundary conditions can be easily implemented by *ad hoc* definitions of transition probabilities:

	Reflecting	Absorbing
Boundary at a	$w_a^- = 0$	$w_{a-1}^+ = 0$
Boundary at b	$w_b^+ = 0$	$w_{b+1}^- = 0$

The reversible chemical reaction with $w_n^- = k_1\, n$ and $w_n^+ = k_2\,(n_0 - n)$, for example, had two reflecting barriers at $a = 0$ and $b = n_0$. Among the examples we have studied so far we were dealing with an absorbing boundary in the replication-extinction process between $\mathcal{N} = 1$ and $\mathcal{N} = 0$ that is tantamount to the lower barrier at $a = 1$ fulfilling $w_0^+ = 0$: The state $n = 0$ is the end point or ω-limit of all trajectories reaching it.

Compared, for example, to an unrestricted random walk on positive and negative integers, $n \in \mathbb{Z}$, a chemical reaction or a biological process has to be restricted by definition, $n \in \mathbb{N}^0$, since negative particle numbers are not allowed. In general, the one step birth-and-death master Equ. (3.98),

$$\frac{\partial P_n(t)}{\partial t} \;=\; w_{n-1}^+\, P_{n-1}(t) \;+\; w_{n+1}^-\, P_{n+1}(t) \;-\; \Big((w_n^+ + w_n^-)\Big)\, P_n(t)\;,$$

is not restricted to $n \in \mathbb{N}^0$ and thus does not automatically fulfil the proper boundary conditions to model a chemical reaction. A modification of the equation at $n = 0$ is required, which introduces a proper boundary of the process:

$$\frac{\partial P_0(t)}{\partial t} \;=\; w_1^-\, P_1(t) \;-\; w_0^+\, P_0(t)\;. \tag{3.98'}$$

This occurs *naturally* if w_n^- vanishes for $n = 0$, which is always the case when the constant term referring to migration vanishes, $\nu = 0$. With $w_0^- = 0$ we only need to make sure that $P_{-1}(t) = 0$ and obtain equation (3.98'). This will be so whenever we take an initial state with $P_n(0) = 0\,\forall\, n < 0$, and it is certainly true for our conventional initial condition, $P_n(0) = \delta_{n,n_0}$ with $n_0 \geq 0$. By the same token we prove that the upper reflecting boundary for chemical reactions, $b = n_0$, fulfils the conditions of being *natural* too. Equipped with natural boundary conditions the stochastic process can be solved for the entire integer range, $n \in \mathbb{Z}$, and this is often much easier than with *artificial* boundaries. All the barriers we have encountered so far were natural.

An overview over a few selected birth-and-death processes is given in tables 5.1 and 5.2. Commonly, unrestricted and restricted processes are dis-

Table 5.1 Comparison of of results for some unrestricted processes. Data are taken from [108, pp.10,11]. Abbreviation and notations: $\gamma \equiv \lambda/\mu$, $\sigma \equiv e^{(\lambda-\mu)t}$, $(n, n_0) \equiv \min\{n, n_0\}$, and $I_n(x)$ is the modified Bessel function.

Process	λ_n	μ_n	$g_{n_0}(s,t)$	$P_{n,n_0}(t)$	Mean	Variance	Ref.
Poisson	ν	0	$s^{n_0}\, e^{\nu(s-1)\,t}$	$\frac{(\nu t)^{n-n_0}\, e^{\nu\, t}}{(n-n_0)!}$, $n \geq n_0$; $n_0 > (0,n)$	$n_0 + \nu t$	νt	[41]
Poisson	0	ρ	$s^{n_0}\, e^{\rho(1-s)\,t/s}$	$\frac{(\rho t)^{n-n_0}\, e^{\rho\, t}}{(n_0-n)!}$, $n \leq n_0$; $n_0 < (0,n)$	$n_0 - \rho t$	ρt	[41]
	ν	ρ	$s^{n_0}\, e^{-(\nu+\rho)t+(\nu s+\rho/s)t}$	$\binom{\nu}{\rho}^{(n-n_0)/2} I_{n_0-n}(2t\sqrt{\nu\rho})\, e^{-(\nu+\rho)t}$	$n_0 + (\nu-\rho)t$	$(\nu+\rho)t$	[126]
Birth	λ_n	0	$\Big(1 - e^{\lambda t}(1-1/s)\Big)^{-n_0}$	$\binom{n}{n_0}\, e^{-n_0\lambda\, t}(1-e^{-\lambda\, t})^{n-n_0}$, $n \geq n_0$; $n_0 > (0,n)$	$n_0\, e^{\lambda t}$	$n_0\, e^{\lambda t}(e^{\lambda t}-1)$	[11]
Death	0	μ_n	$\Big(1 - e^{-\mu t}(1-s)\Big)^{n_0}$	$\binom{n_0}{n}\, e^{-n\mu\, t}(1-e^{-\mu\, t})^{n_0-n}$, $n \leq n_0$; $n_0 < (0,n)$	$n_0\, e^{-\mu t}$	$n_0\, e^{-\mu t}(1-e^{-\mu t})$	[11]
	ν	μ_n	$\Big(1 - e^{-\mu t}(1-s)\Big)^{n_0} \times$	$\exp\Big(-\frac{\nu}{\mu}(1-e^{-\mu t})\Big) \times$	$n_0\, e^{-\mu t} +$	$\Big(\frac{\nu}{\mu} + n_0 e^{-\mu t}\Big) \times$	[41]
			$\times\, \exp\Big(\nu(s-1)(1-e^{-\mu t})/\mu\Big)$	$\times \sum_{k=0}^{(n,n_0)} \frac{e^{-\mu t k}(1-e^{-\mu t})^{n+n_0-2k}}{(n-k)!} \Big(\frac{\nu}{\mu}\Big)^{n-k}$	$+\frac{\nu(1-e^{-\mu t})}{\mu}$	$\times(1-e^{-\mu t})$	
Birth& Death	λ_n	μ_n	$\Big(\frac{(\sigma-1)+(\gamma-\sigma)s}{(\gamma\sigma-1)\gamma(1-\sigma)s}\Big)^{n_0}$	$\gamma^n \sum_{k=0}^{(n,n0)} (-1)^k \binom{n+n_0-k-1}{n-k}\binom{n_0}{k} \times$ $\times \Big(\frac{1-\sigma}{1-\gamma\sigma}\Big)^{n+n_0-k} \Big(\frac{1-\sigma/\gamma}{1-\sigma}\Big)^k$	$n_0\, \sigma$	$\frac{n_0\sigma(\gamma+1)(\sigma-1)}{\gamma-1}$	[11]
	λ_n	λ_n	$\Big(\frac{\lambda t+(1-\lambda t)s}{1+\lambda t-\lambda t\, s}\Big)^n$	$\Big(\frac{\lambda t}{1+\lambda t}\Big)^{n+n_0} \sum_{k=0}^{(n,n_0)} \binom{n_0}{k} \times$ $\times \binom{n+n_0-k-1}{n-k} \Big(\frac{1-\lambda^2 t^2}{\lambda^2 t^2}\Big)^k$	n_0	$2n_0\lambda\, t$	

Table 5.2 Comparison of of results for some restricted processes. Data are taken from [108, pp.16,17]. Abbreviation and notations used in the table are: $\gamma \equiv \lambda/\mu$, $\sigma \equiv e^{(\lambda-\mu)t}$, $\alpha \equiv (\nu/\rho)^{(n-n_0)/2} e^{(\nu+\rho)t}$; $I_n = I_{-n} \equiv I_n\big(2(\nu\rho)^{1/2}t\big)$ where $I_n(x)$ is a modified Bessel function; $G_n \equiv G_n(\xi_j,\gamma)$ where G_n is a Gottlieb polynomial, $\hat{G}_n \equiv G_n(\hat{\xi}_j,\gamma)$,
$G_n(x,\gamma) \equiv \gamma^n \sum_{k=0}^{n}(1-\gamma^{-1})^k \binom{n}{k}\binom{x-k+1}{k} = \gamma^n F(-n,-x,1,1-\gamma^{-1})$ where F is a hypergeometric function, ξ_j and $\hat{\xi}_j$ are the roots of $G_{u-l}(\xi_j,\gamma)=0$, $j=0,\ldots,u-l-1$ and $G_{u-l+1}(\hat{\xi}_j,\gamma) = \gamma\, G_{u-l}(\hat{\xi}_j,\gamma)$, $j=0,\ldots,u-l$, respectively; $H_n \equiv H_n(\zeta_j,\gamma)$, $\hat{H}_n \equiv H_n(\hat{\zeta}_j,\gamma)$, $H_n(x,\gamma) = G_n(x,\gamma^{-1})$, $H_{u-l}(\zeta_j,\gamma)=0$, $j=0,\ldots,u-l-1$ and $H_{u+l-1}(\hat{\zeta}_j,\gamma) = H_{u-l}(\hat{\zeta}_j,\gamma)/\gamma$, respectively.

λ_n	μ_n	Boundaries	$P_{n,n_0}(t)$	Ref.
ν	ρ	$u:$ abs; $l:-\infty$	$\alpha\big(I_{n-n_0} - I_{2u-n-n_0}\big)$	[41, 217]
ν	ρ	$u:+\infty$; $l:$ abs	$\alpha\big(I_{n-n_0} - I_{n+n_0-2l}\big)$	[41, 217]
ν	ρ	$u:$ refl; $l:-\infty$	$\alpha\left(I_{n-n_0} + \left(\frac{\nu}{\rho}^{1/2} I_{2u+l-n-n_0} + \left(1-\frac{\rho}{\nu}\right)\right)\cdot \sum_{j=2}^{\infty}\left(\frac{\nu}{\rho}\right)^{j/2} I_{2u-n-n_0+j}\right)$	[41, 217]
ν	ρ	$u:+\infty$; $l:$ refl	$\alpha\left(I_{n-n_0} + \left(\frac{\nu}{\rho}^{1/2} I_{n+n_0+l-2u} + \left(1-\frac{\rho}{\nu}\right)\right)\cdot \sum_{j=2}^{\infty}\left(\frac{\nu}{\rho}\right)^{j/2} I_{n+n_0-2l+j}\right)$	[41, 217]
ν	ρ	$u:$ abs; $l:$ abs	$\alpha\left(\sum_{k=-\infty}^{\infty} I_{n-n_0+2k(u-l)} - \sum_{k=0}^{\infty}\big(I_{n+n_0-2l+2k(u-l)} + I_{2l-n-n_0+2k(u-l)}\big)\right)$	[41, 217]
$\lambda(n-l+1)$	$\mu(n-l)$	$u:$ abs; $l:$ refl	$\gamma^{l-n}\sum_{k=0}^{u-l-1} G_{n_0-l}G_{n-l}\ \sigma^{\xi_k}\left(\sum_{j=0}^{u-l-1}\frac{G_j}{\gamma^j}\right)^{-1}$	[219, 265]
$\lambda(n-l+1)$	$\mu(n-l)$	$u:$ refl; $l:$ refl	$\gamma^{l-n}\sum_{k=0}^{u-l} \hat{G}_{n_0-l}\hat{G}_{n-l}\ \sigma^{\hat{\xi}_k}\left(\sum_{j=0}^{u-l}\frac{\hat{G}_j}{\gamma^j}\right)^{-1}$	[219, 265]
$\lambda(u-n)$	$\mu(u-n+1)$	$u:$ refl; $l:$ abs	$\gamma^{u-n}\sum_{k=0}^{u-l-1} H_{u-n_0}H_{u-n}\sigma^{-\zeta_k}\left(\sum_{j=0}^{u-l-1} H_j\gamma^j\right)^{-1}$	[219, 265]
$\lambda(u-n)$	$\mu(u-n+1)$	$u:$ refl; $l:$ refl	$\gamma^{u-n}\sum_{k=0}^{u-l} \hat{H}_{u-n_0}\hat{H}_{u-n}\sigma^{-\hat{\zeta}_k}\left(\sum_{j=0}^{u-l}\hat{H}_j\gamma^j\right)^{-1}$	[219, 265]

tinguished [108]. An unrestricted process is characterized by the possibility to reach all states $\mathcal{N}(t) = n$. A requirement imposed by physics demands that all changes in state space are finite for finite times, and hence the probabilities to reach infinity at finite times must vanish: $\lim_{n\to\pm\infty} P_{n,n_0} = 0$. The linear birth and death process in table 5.1 is unrestricted only in the positive direction and the state $\mathcal{N}(t) = 0$ is special because it represents an absorbing barrier. The restriction is here hidden and met by the condition $P_{n,n_0}(t) = 0 \;\forall\; n < 0$.

5.2.3 The Wright-Fisher and the Moran process

Here we shall introduce two common stochastic models in population biology, the *Wright-Fisher model* named after Sewall Wright and Ronald Fisher and the *Moran model* named after the Australian statistician Pat Moran. The Wright-Fisher model and the Moran model are stochastic models for evolution of allele distributions in populations with constant population size [21]. The first model [83, 304] also addressed as *beanbag* population genetics is presumably the simplest process for the illustration of genetic drift and definitely the most popular one [45, 67, 121, 197] deals with strictly separated generations, whereas the Moran process [222, 223] based on continuous time and overlapping generations is generally more appealing to statistical physicists. Both processes are introduced here for the simplest scenarios: haploid organisms, two alleles of the gene under consideration and no mutation. Extension to more complicated cases is readily possible. The primary question that was thought to be addressed by the two models is the evolution of populations in case of neutrality for selection.

5.2.3.1 The Wright-Fisher process

The Wright-Fisher process is illustrated in figure 5.13. A single reproduction event is modeled by a sequence of four steps: (i) A gene is randomly chosen from the gene pool of generation T containing exactly N genes distributed over M alleles, (ii) it is replicated, (iii) the original is put back into the gene pool T, and (iv) the copy is put into the gene pool of the next generation $T + 1$. The process is terminated when the next generation gene pool has exactly N genes. Since filling the gene pool of the $T + 1$ generation depends exclusively on the distribution of genes in the pool of generation T, and earlier gene distributions have no influence on the process the Wright-Fisher model is Markovian.

In order to simplify the analysis we assume two alleles A and B, which are present in a_T and b_T copies in the gene pool at generation T. Since the total number of genes is constant, $a_T + b_T = N$ and $b_T = N - a_T$, we are dealing

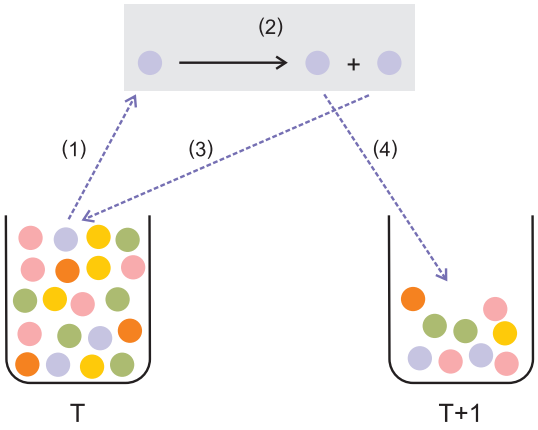

Fig. 5.13 The Wright-Fisher model of *beanbag* genetics. The gene pool of generation T contains N gene copies chosen from m alleles. Generation $T+1$ is built from generation T through ordered cyclic repetition of a four step event: (1) random selection of one gene from the gene pool T, (2) error-free copying of the gene, (3) putting back the original into gene pool T, and (4) placing the copy into the gene pool of the next generation $T+1$. The procedure is repeated until the gene pool $T+1$ contains exactly N genes. No mixing of generations is allowed.

with a single discrete variable, a_T, $T \in \mathbb{N}$. A new generation $T+1$ is produced from the gene pool at generation T through picking with replacement N times a gene. The probability to obtain $n = a_{T+1}$ alleles A in the new gene pool is given by the binomial distribution:

$$\text{Prob}\,(a_{T+1} = n) \;=\; \binom{N}{n} p_{\mathsf{A}}^{n}\, p_{\mathsf{B}}^{N-n} \;,$$

$p_{\mathsf{A}} = a_T/N$ and $p_{\mathsf{B}} = b_T/N = (N-a_T)/N$ with $p_{\mathsf{A}}+p_{\mathsf{B}} = 1$ are the individual probabilities of picking A or B, respectively. The transition probability from m alleles A at time T to n alleles A at time $T+1$ is simply given by[15,16]

$$W_{nm} \;=\; \binom{N}{n} \left(\frac{m}{N}\right)^{n} \left(1-\frac{m}{N}\right)^{N-n} \;. \tag{5.53}$$

Since the construction of the gene pool at generation $T+1$ is fully determined by the gene distribution at generation T, the process is Markovian.

[15] The notation applied here is the conventional way of writing transitions in physics: W_{nm} is the probability of the transition $n \leftarrow m$, whereas many mathematicians would write p_{mn} indicating $m \rightarrow n$.

[16] For doing actual calculations one has to recall the convention $0^0 = 1$ used in probability theory and combinatorics but commonly not in analysis where 0^0 is an indefinite expression.

In order to study the evolution of populations an initial state has to be specified. We assume that the number of alleles A has been n_0 at generation $T = 0$ and accordingly we are calculating the probability $P(n, T|\, n_0, 0)$, which we denote by $p_n(T)$. Since the Wright-Fisher model does not contain any interactions between alleles or mutual dependencies between processes involving alleles, the process can be modeled best by means of linear algebra. We define a probability vector $\mathbf{p}$ and a transition matrix, W:

$$\begin{aligned} \mathbf{p}(T) &= \big(p_0(T), p_1(T), \ldots, p_N(T)\big) \quad \text{and} \\ \mathrm{W} &= \begin{pmatrix} W_{00} & W_{01} & \cdots & W_{0N} \\ W_{10} & W_{11} & \cdots & W_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ W_{N0} & W_{N1} & \cdots & W_{NN} \end{pmatrix} = \begin{pmatrix} 1 & W_{01} & \cdots & 0 \\ 0 & W_{11} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & W_{N1} & \cdots & 1 \end{pmatrix} . \end{aligned} \tag{5.54}$$

Conservation of probability provides two conditions: (i) The probability vector has be normalized to a L^1-norm, $\sum_n p_n(T) = 1$, and (ii) it has to remain normalized in future generations, $\sum_n W_{nm} = 1$.[17] The evolution is now simply described by the matrix equation

$$\mathbf{p}(T+1)^{\mathrm{t}} = \mathrm{W} \cdot \mathbf{p}(T)^{\mathrm{t}} \quad \text{or} \quad \mathbf{p}(T)^{\mathrm{t}} = \mathrm{W}^T \cdot \mathbf{p}(0)^{\mathrm{t}} \; . \tag{5.55}$$

Equation (5.55) is identical with the matrix formulation of linear *difference equations*, $\mathbf{p}^{\mathrm{t}}_{k+1} = \mathrm{W} \cdot \mathbf{p}^{\mathrm{t}}_k$, which have been used in section 5.2.1.2 to discuss multitype branching, and which are presented and analyzed, for example, in the monograph [46, pp.179-216].

Solutions of (5.55) are known in the from of an analytical expression for the eigenvalues of the transition matrix W [75]:

$$\lambda_k = \binom{N}{k} \frac{k!}{N^k} ; \; k = 0, 1, 2, \ldots \; . \tag{5.56}$$

Although we do not have analytical expressions for the eigenvectors of transition matrix W at hand, the stationary state of the Wright-Fisher process can be deduced from the properties of a Markov chain by asking what the system might look like in the limit of an infinite number of generations when the probability density might adopt a stationary distribution $\bar{\mathbf{p}}$. If such a stationary state exists the density must fulfil the equation $\mathrm{W} \cdot \bar{\mathbf{p}} = \bar{\mathbf{p}}$ or in other words $\bar{\mathbf{p}}$ is a right eigenvector of W with the eigenvalue $\lambda = 1$.

By intuition we guess that a final absorbing state of the system must be either all B corresponding to $\bar{n} = 0$ and fixation of allele B or all A with $\bar{n} = N$ and fixation of allele A. Such a steady state would correspond to a probability density

$$\bar{\mathbf{p}} = (1 - \pi, 0, \ldots, 0, \pi) \; , \tag{5.57}$$

[17] A matrix W with this property is called a stochastic matrix.

which fulfils $\mathrm{W} \cdot \bar{\mathbf{p}} = \bar{\mathbf{p}}$ as is easily confirmed by inserting W from equation (5.54).

Next we compute the expected number of alleles A as a function of the generation number

$$\begin{aligned} \langle n(T+1)\rangle &= \sum_{n=0}^{N} n\, p_n(T+1) = \sum_{n=0}^{N} n \sum_{m=0}^{N} W_{nm}\, p_m(T) = \\ &= \sum_{m=0}^{N} p_m(T) \sum_{n=0}^{N} n\, W_{nm} = \sum_{m=0}^{N} m\, p_m(T) = \langle n(T)\rangle \ , \end{aligned} \tag{5.58}$$

where we have used the expectation value of the binomial distribution (2.34a) in the last line

$$\sum_{n=0}^{N} n\, W_{nm} = \sum_{n=0}^{N} n \binom{N}{n} \left(\frac{m}{N}\right)^n \left(1-\frac{m}{N}\right)^{N-n} = N\frac{m}{N} = m \ .$$

The expectation values of the numbers of alleles is independent of the generation T and this implies $\langle n(T)\rangle = \langle n(0)\rangle = n_0$. This result enables us to determine the probability π for the fixation of allele A. From equation (5.57) we deduce two possible states in the limit $T \to \infty$: (i) $n = N$ with probability π and (ii) $n = 0$ with probability $1 - \pi$ and accordingly we have

$$\lim_{T\to\infty} \langle n(T)\rangle = \pi N + (1-\pi)0 \implies n_0 = \pi N \ \text{ and } \ \pi = \frac{n_0}{N} \ . \tag{5.59}$$

Eventually, we found the complete expression for the stationary state of the Wright-Fisher process and the probability of fixation of allele A, which amounts to $\pi = n_0/N$.

5.2.3.2 The Moran process

The Moran process introduced by Pat Moran [222] is a continuous time process and deals with transitions that are defined for single events. As in the Wright-Fisher model we are dealing with two alleles, A and B, in a haploid population of population size N and the probabilities for choosing A or B are p_A and p_B, respectively. Unlike the Wright-Fisher model there is no defined previous generation from which a next generation is formed by sampling N genes and therefore overlapping generations make it difficult – if not impossible – to define generations unambiguously. The *event* in the Moran process is a *combined birth-and-death step*: Two genes are picked, one is copied and both template and copy are put back into the urn, and the second one is deleted (see figure 5.14). The probabilities are calculated form the state of the urn just before the event $p_\mathsf{A} = m(t)/N$ and $p_\mathsf{B} = \big(N - m(t)\big)/N$ where $n(t)$ is the number of alleles A, $N - m(t)$ the number of alleles B, and N is the constant total number of genes. After the event we have exactly n alleles

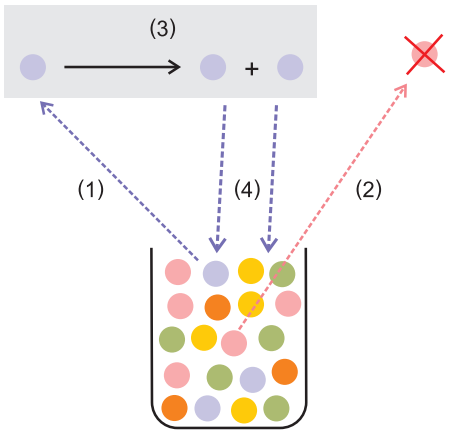

Fig. 5.14 The Moran process. The Moran process is a continuous time model for the same problem handled by the Wright-Fisher model (figure 5.13). The gene pool of a population of N genes chosen from m alleles is represented by the urn in the figure. Evolution proceeds via successive repetition of a four step process: (1) One gene is chosen from the gene pool at random, (2) a second gene is randomly chosen and deleted, (3) the first gene is copied, and (4) both genes, original and copy, are put back into the urn. The Moran process has overlapping generations and moreover the notion of generation is not well defined.

of type A and $N - n$ alleles of type B with $\Delta n = n - m = \pm 1, 0$ depending on the nature of the process. In particular, two different ways of picking two genes are commonly used in the literature: (i) In the more intelligible first counting one considers explicitly the reduction in numbers by one as a consequence of the first pick [235] and (ii) in the second procedure the changes introduced in the urn by picking the first gene are ignored in the second draw (see, e.g., [21]).[18] We shall present the (almost identical) results of both picking procedures here and start with the second perhaps easier to motivate count first.

Before the combined birth-death event we have m genes with allele A out of N genes in total. Because of the first pick the total number of genes and the number of the genes with allele A are reduced by one for the coupled second pick, $N \to N - 1$ and $m \to m - 1$, respectively. In case the first pick chose a B allele the changes in the numbers of genes were: $N \to N - 1$ and $N - m \to N - m - 1$. After the event the numbers have been change to n and $N - n$, respectively, and $n - m = 0, \pm 1$. Now we compute the probabilities for the four possible sequential draws and find:

(i) $\mathsf{A} + \mathsf{A}$: $p_{\mathsf{A}+\mathsf{A}} = \left(\frac{m}{N}\right)\left(\frac{m-1}{N-1}\right)$ contributing to $n = m$,

[18] It is fair to say that the same model results from an accurate but a little bit strange assumption: After the replication event the parent but not the offspring is put back into the pool from which the individual is chosen, which is doomed to die.

(ii) $\mathsf{A}+\mathsf{B}$: $p_{\mathsf{A}+\mathsf{B}} = \left(\frac{m}{N}\right)\left(1-\frac{m-1}{N-1}\right)$ contributing to $n = m+1$,
(iii) $\mathsf{B}+\mathsf{A}$: $p_{\mathsf{B}+\mathsf{A}} = \left(1-\frac{m}{N}\right)\left(\frac{m}{N-1}\right)$ contributing to $n = m-1$, and
(iv) $\mathsf{B}+\mathsf{B}$: $p_{\mathsf{B}+\mathsf{B}} = \left(1-\frac{m}{N}\right)\left(1-\frac{m}{N-1}\right)$ contributing to $n = m$.

It is readily verified that the four probabilities of the four possible events sum up to one: $p_{\mathsf{A}+\mathsf{A}} + p_{\mathsf{A}+\mathsf{B}} + p_{\mathsf{B}+\mathsf{A}} + p_{\mathsf{B}+\mathsf{B}} = 1$. The elements of the transition matrix can be written as

$$W_{nm} = \begin{cases} \frac{m}{N}\,\frac{m-N}{N-1} & \text{if } n = m+1 \\ \frac{m(m-1)+(N-m)(N-m-1)}{N(N-1)} & \text{if } n = m \\ \frac{m}{N}\,\frac{m-N}{N-1} & \text{if } n = m-1 \end{cases} . \tag{5.60}$$

We check easily that W is a stochastic matrix, $\sum_n W_{nm} = 1$. The transition matrix W of the Moran model is tridiagonal since only the changes $\Delta n = 0, \pm 1$ can occur.

In the slightly modified version of the model – procedure (ii) – we assume that the replicated individual – but nor the offspring – is returned to the pool from which the dying individual is chosen after the replication event. Then the elements of the transition matrix are:

$$W_{nm} = \begin{cases} \frac{m\,(m-N)}{N^2} & \text{if } n = m+1 \\ \frac{m^2+(N-m)^2}{N^2} & \text{if } n = m \\ \frac{m\,(m-N)}{N^2} & \text{if } n = m-1 \end{cases} . \tag{5.60'}$$

Clearly, $\sum_n W_{nm} = 1$ is fulfilled as in procedure (i).

The transition matrix $\mathrm{W} = \{W_{nm}\}$ has tridiagonal form and eigenvalues and eigenvectors are readily calculated [67, 222, 223]. The results for the different picking procedures are almost the same. For procedure (i) we find

$$\lambda_k = 1 - \frac{k(k-1)}{N(N-1)};\ k = 0, 1, 2, \dots , \tag{5.61}$$

and for procedure (ii) we get

$$\lambda_k = 1 - \frac{k(k-1)}{N^2};\ k = 0, 1, 2, \dots . \tag{5.61'}$$

For the Moran model the eigenvectors are the same for both procedures, (i) and (ii), and they are available in analytical form [223]. The first two eigenvectors belong to the doubly degenerate largest eigenvalue $\lambda_0 = \lambda_1 = 1$,

$$\begin{aligned} \boldsymbol{\zeta}_0 &= (1, 0, 0, 0, 0, \dots, 0)^{\mathrm{t}} \quad \text{and} \\ \boldsymbol{\zeta}_1 &= (0, 0, 0, 0, 0, \dots, 1)^{\mathrm{t}} , \end{aligned} \tag{5.62}$$

and they describe the long time behavior of the Moran process since stationarity indeed implies $\mathbf{p}(T+1)^{\mathrm{t}} = \mathbf{p}(T)^{\mathrm{t}} = \bar{\mathbf{p}}$, or $\mathrm{W}\,\bar{\mathbf{p}}^{\mathrm{t}} = \bar{\mathbf{p}}^{\mathrm{t}}$, and hence $\lambda = 1$. As in the Wright-Fisher model we are dealing here with twofold degeneracy and we recall that in such a case any properly normalized linear combination of the eigenvectors is a legitimate solution of the eigenvalue problem. Here we have to apply the L^1-norm and obtain

$$\boldsymbol{\eta} \;=\; \alpha\,\boldsymbol{\zeta}_0 \,+\, \beta\,\boldsymbol{\zeta}_1 \;\;\text{and}\;\; \alpha+\beta = 1\;,$$

and accordingly we find for the general solution of the stationary state

$$\boldsymbol{\eta} \;=\; (1-\pi, 0, 0, 0, 0, \ldots, \pi)^{\mathrm{t}}\;. \tag{5.63}$$

The interpretation of the stationary state, which is identical with the result for the Wright-Fisher process, is straightforward: The allele A goes into fixation in the population with probability π and it is lost with probability $1-\pi$, and the Moran model as well as the Wright-Fisher model provides a simple explanation for gene fixation by random drift. The calculation of the value for π that depends on the initial conditions,[19] which are again assumed to be $n(0) = n_0$, follows the same argumentation as for the Wright-Fisher model in equations (5.58) and (5.59) and from the generation independent expectation value $\langle n(T)\rangle = n_0$ we obtain:

$$\lim_{T\to\infty}\langle n(T)\rangle = \; N\,\pi \;=\; n_0 \;\;\text{and}\;\; \pi \,=\, \frac{n_0}{N} \tag{5.59'}$$

and the probability for the fixation of A finally is n_0/N. From the value of π follows immediately $\alpha = 1-\pi = (N-n_0)/N$ and $\beta = \pi = n_0/N$.

The third eigenvector belonging to the eigenvalue $\lambda_2 = 1 - 2/\big(N(N-1)\big)$ can be used to calculate the evolution towards fixation [21]:

$$\mathbf{p}(t) \;\approx\; \begin{pmatrix} 1-\frac{n_0}{N} \\ 0 \\ \vdots \\ 0 \\ \frac{n_0}{N} \end{pmatrix} + \frac{6n_0(N-n_0)}{N(N^2-1)} \begin{pmatrix} \frac{N-1}{2} \\ -1 \\ \vdots \\ -1 \\ \frac{N-1}{2} \end{pmatrix} \left(1-\frac{2}{N^2}\right)^T\;.$$

After sufficiently long time the probability density function becomes completely flat except at the two boundaries, $n = 0$ and $n = N$. We shall encounter the same form of the density for continuous time with the master equation and with the Fokker-Planck approximation (section 5.3).

[19] In the nondegenerate case stationary states do not depend on initial conditions but this is no longer true for linear combinations of degenerate eigenvectors: α and β, and π are functions of the initial state.

5.3 Master and Fokker-Planck equations in biology

In section 5.2.2 we used master equations to find solutions for simple birth-and-death processes. Here we consider more general models and start out from the standard Markov chain

$$\begin{aligned} P(n,t+1) &= \sum_m p_{nm} P(m,t) \quad \text{and} \\ P(n,t+1) - P(n,t) &= \sum_m p_{nm} P(m,t) - \sum_m p_{mn} P(n,t) \ , \end{aligned} \tag{5.64}$$

where we used the relation $\sum_m p_{mn} = 1$ in the last term on the right hand side. The two terms with $m = n$ can be omitted because of cancelation, and t that could be considered as an integer label for generations is now interpreted as time. Then the intervals Δt have to be taken sufficiently small that at most one sampling event occurs between t and $t + \Delta t$. Division by Δt yields

$$\frac{P(n,t+\Delta t) - (n,t)}{\Delta t} \ = \ \sum_m \left(\frac{p_{nm}}{\Delta t}\right) P(m,t) \ - \ \sum_m \left(\frac{p_{mn}}{\Delta t}\right) P(n,t) \ .$$

Instead of assuming that exactly one sampling event – including $n \to n$ where no actual transition occurs – happens per generation we consider now sampling events at *unit rate* such that one event *on average* takes place per generation. If t is sufficiently large, the by far most likely number of events that will have occurred is equal to t and we can expect that continuous time and discrete time processes will occur at large times.

The transition probability is replaced by the transition rate per unit time

$$p_{nm} \ = \ W(n|m)\,\Delta t \ + \ \mathcal{O}(\Delta t)^2 \quad \text{for} \quad n \neq m \ , \tag{5.65}$$

where the terms of order $(\Delta t)^2$ and higher express the probabilities that two or more events take place during the time interval Δt. Performing the limit $\Delta t \to 0$ yields the familiar master equation

$$\frac{\partial P(n,t)}{\partial t} \ = \ \sum_{m \neq n} \Big(W(n|m) P(m,t) \ - \ W(m|n) P(n,t) \Big) \ . \tag{4.28}$$

The only difference to the general form of the master equation is the assumption that the transition rates per unit time are rate parameters, which are independent of time. Accordingly, we can replace the conditional probabilities by the elements of a square matrix $W = \{W_{nm} = W(n|m)\}$.

For the purpose of illustration we consider two suitable examples: We derive solutions for the Moran model by means of a master equation. Analytical solutions of master equations are rare and therefore often approximations are made, which convert the master equations into Fokker-Planck equations. Our

example for diffusion in population space described by means of a Fokker-Planck equation is again the Moran model and, in particular, Motoo Kimura's solution for neutral evolution.

5.3.1 The master equation of the Moran process

Revisiting the two-allele Moran model (section 5.2.3.2 and figure 5.14) we construct a master equation for the continuous time process and then make the approximations for large population sizes in the sense of a Fokker-Planck equation. We recall the probabilities for the different combinations of choosing genes from the pool and adopt the simpler to calculate procedure (ii) (section 5.2.3.2). Again we have a gene pool of N genes, exactly m alleles of type A and $N-m$ alleles of type B before the picking event. After the event the numbers have been changed to n and $N-n$, respectively:

(i) $\mathsf{A}+\mathsf{A}$: $p_{\mathsf{A}+\mathsf{A}} = \frac{m^2}{N^2}$ contributing to $n = m$,
(ii) $\mathsf{A}+\mathsf{B}$: $p_{\mathsf{A}+\mathsf{B}} = \frac{m(N-m)}{N^2}$ contributing to $n = m+1$,
(iii) $\mathsf{B}+\mathsf{A}$: $p_{\mathsf{B}+\mathsf{A}} = \frac{(N-m)m}{N^2}$ contributing to $n = m-1$, and
(iv) $\mathsf{B}+\mathsf{B}$: $p_{\mathsf{B}+\mathsf{B}} = \frac{(N-m)^2}{N^2}$ contributing to $n = m$.

These probabilities give rise to the same transition rates as before

$$\begin{aligned} W(n+1|n) &= \kappa\,\frac{m(N-m)}{N^2}\ , \\ W(n|n) &= \kappa\,\frac{m^2+(N-m)^2}{N^2}\ , \text{ and} \\ W(n-1|n) &= \kappa\,\frac{(N-m)m}{N^2}\ , \end{aligned} \tag{5.66}$$

where κ is a rate parameter. Apart from the two choices that don't change the composition of the urn we have only two allowed processes (see also equation (3.96)): (i) $n \to n+1$ with w_n^+ as transition probability and (ii) $n \to n-1$ with w_n^- as transition probability (see section 3.2.5.2), and moreover the analytical expressions are the same for both. Therefore we are dealing with a *symmetric one-step process*:

$$w_n^+ = w_n^- = \kappa\,\frac{n(n-N)}{N^2}\ . \tag{5.67}$$

It is of advantage to handle the selection case simultaneously and therefore we introduce a selective advantage for allele A in the form of a factor $(1+r)$ and then have for the reproduction of the fitter variant A[20]

[20] In population genetics the fitness parameter is conventionally denoted by s but use here r in order to avoid confusion with the auxiliary variable s.

$$w_n^+ \;=\; \kappa\,\frac{n(n-N)}{N^2}\,(1+r)\;, \tag{5.67'}$$

the process is no longer symmetric but can return to the neutral case by putting $r = 0$. The constant factor κ/N^2 can be absorbed in the time, which is measured in units $[N^2/\kappa]$. Then the master equation is of the form[21]

$$\begin{aligned}\frac{\partial P(n,t)}{\partial t} \;&=\; w_{n+1}^- P(n+1,t) \,+\, w_{n-1}^+ P(n-1,t) \,-\, (w_n^- + w_n^+)P(n,t) \;= \\ &=\; (n+1)(N-n+1)(1+r)\,P(n+1,t) + \\ &\quad +\,(n-1)(N-n-1)\,P(n-1,t) \,-\, n(N-n)(2+r)\,P(n,t)\;.\end{aligned} \tag{5.68}$$

An exact solution of the master equation (5.68) has been derived [135] for the neutral ($r = 0$) and the selective case ($r \neq 0$). It does not only provide an exact reference it gives also unambiguous answers to a number of open questions. The approach to find the analytical solution of equation (5.68) is the conventional one based on generating functions and partial differential equations as used for the solution of chemical master equations (section 4.3). We repeat the somewhat technical procedure here, because it has general applicability and one more example is quite illustrative.

First the usual probability generating function (2.24) is defined as

$$g(s,t) \;=\; \sum\nolimits_{n=0}^{N} s^n\,P(n,t) \tag{2.24'}$$

and the following PDE is obtained in the conventional way:

$$\frac{\partial g(s,t)}{\partial t} \;=\; (1-s)\big(1-(1+r)s\big)\frac{\partial}{\partial s}\left(Ng(s,t) - s\frac{\partial g(s,t)}{\partial s}\right)\;. \tag{5.69}$$

Equation (5.69) has to be solved now for a given initial condition, for example exactly n_0 alleles of type A at time $t = 0$:

$$P(n,0) \;=\; \delta_{n,n_0} \;\text{ or }\; g(s,0) \;=\; s^{n_0}\;. \tag{5.70}$$

From the definition of the probability generating function follow the boundary conditions

$$g(1,t) \;=\; 1 \;\text{ and} \tag{5.71a}$$

if $r = 0$ and therefore $w_n^+ = w_n^-$ and $\langle n(t)\rangle = n_0$, then

$$\left.\frac{\partial g(s,t)}{\partial s}\right|_{s=1} \;=\; n_0 \;\text{ or}\,, \tag{5.71b}$$

if $r \neq 0$ then $s = \sigma$ is a fixed point of the PDE (5.69) and therefore

[21] In order to make the transformation of the master equation into a Fokker-Planck equation in a more transparent way we change notation and write $P(n,t)$ instead of $P_n(t)$. Formally we consider n as a second variable and apply partial differentiation.

$$g(\sigma,t) \;=\; g(\sigma,0) \;=\; \sigma^{n_0}\;. \tag{5.71b'}$$

The beauty of this approach [135] is that the PDE (5.69) with the initial condition (5.70) and the boundary conditions (5.71) constitute a well defined problem in contrast to the stochastic diffusion equation used in population genetics, which requires separate *ad hoc* assumptions for the limiting gene frequencies $x = 0$ and $x = 1$ (see section 5.3.2 or [45, pp. 379-380]).

The neutral case: $r = 0$**.** A general solution of the master equation of the kind 5.68 is of the form

$$P(n,t) \;=\; \sum\nolimits_{k=0}^{N-1} \beta_k^{(n)}\, e^{\lambda_k t}\;, \tag{5.72}$$

since it can visualized as a system of $N+1$ first order linear differential equations with the constraint $\sum_{n=0}^{N} P(n,t) = 1$. The probability generating function $g(s,t)$ is just a combination of these probabilities with the weighting factors s^n and therefore it is suggestive to search for solutions among the linear combinations of the functions $\phi_n(s)e^{\lambda_n t}$ where $\psi_n(s)$ and λ_n are solutions of the eigensystem

$$\lambda_n\,\psi_n(s) \;=\; (1-s)^2\frac{\mathrm{d}}{\mathrm{d}s}\Big(N\psi_n(s) \;-\; s\,\frac{\mathrm{d}\psi_n(s)}{\mathrm{d}s}\Big)\;. \tag{5.73}$$

Solutions of (5.73) can be given in terms of hypergeometric functions ${}_2F_1(y)$ with $y = 1/(s-1)$, which will be dome in the discussion of the diffusion approximation in section 5.3.2. Here we present the direct derivation, which makes use of the polynomial character of the solutions.

The equation $\lambda = 0$ is fulfilled by the stationary solution (see paragraph *stationary solution*):

$$\lambda_0 = 0: \quad \psi_0(s) \;=\; \pi_0 \;+\; \pi_N\, s^N \;=\; \bar{g}(s)\;.$$

For $\lambda \neq 0$ we search for solutions that are polynomials in $(1-s)$ like

$$\psi(s) \;=\; \sum\nolimits_{k=0}^{N-1} a_k\,(1-s)^{k+1}\;, \tag{5.74}$$

because $s = 1$ is a double root of $\psi(s) = 0$. The first coefficient has to be zero, $a_0 = 0$, as the lowest term in the polynomial is the coefficient of $(1-s)^2$, a_1, and the other coefficients fulfil the recursion:

$$\big(\lambda \;+\; k(k+1)\big)\,a_k \;=\; k\,(k-N)\,a_{k-1}\,;\; k = 1,\ldots,N-1\;.$$

The relation for the first coefficient, $a_0 = 0$, implies that nontrivial solutions exist only if $\lambda = -n(n+1)$ for some integer n, and we make use of this integer to label the eigenvalues λ_n and the eigenfunctions $\psi_n(s)$:

$$\lambda_n \;=\; -\,n\,(n+1)\,;\; n=1,2,\ldots,N-1\;, \tag{5.75a}$$

$$\psi_n(s) \;=\; \sum\nolimits_{k=n}^{N-1} a_k^{(n)}\,(1-s)^{k+1}\;, \quad \text{and} \tag{5.75b}$$

$$\begin{aligned} a_n^{(n)} \;&=\; 1 \quad \text{and} \\ a_k^{(n)} \;&=\; \frac{k\,(N-k)}{n(n+1)-k(k+1)}\; a_{k-1}^{(n)}\,;\; k=n+1,\ldots,N-1\;. \end{aligned} \tag{5.75c}$$

The probability generating function can be expressed now in terms of these eigenfunctions

$$g(s,t) \;=\; \pi_0 \;+\; \pi_N s^N \;+\; \sum_{n=1}^{N-1} C_n\,\psi_n(s)\,e^{\lambda_n t}\;, \tag{5.76}$$

with the coefficients C_n to be determined from the initial conditions, e.g. from $g(s,0)=s^{n_0}$. The probabilities $P(n,t)$ follow then in the conventional way from the expansion of the probability generating function $g(s,t)$ in powers of the variable s and identification of coeffi ++++cients:

$$\begin{aligned} P(n,t) \;&=\; \pi_0\delta_{n,0} \;+\; \pi_N\delta_{n,N} \;+\; (-1)^n \sum_{k=n}^{N} \binom{n}{k}\,\alpha_{k-1}(t) \quad \text{where} \\ \alpha_{k-1}(t) \;&=\; \sum_{n=1}^{N-1} C_n\,a_k^{(n)}\,e^{\lambda_n t}\,;\; k=1,\ldots,N-1\;, \end{aligned} \tag{5.77}$$

with $\alpha_{-1}(t)\alpha_0(t)=0$. What remains still to be done is the derivation of compact expressions of the coefficients $a_k^{(n)}$ and C_n.

The recurrence relation (5.75c) allows also for direct computation of the coefficients:

$$a_k^{(n)} \;=\; \binom{k}{n}\,\frac{(1-N+n)_{k-n}}{(2n+2)_{k-n}}\;, \tag{5.75c'}$$

where the binomial coefficients $\binom{k}{n} = 0\,\forall\, k < n$ and $(x)_n$ is the rising Pochhammer symbol: $\Gamma(x+n)/\Gamma(x)$[22] The coefficients $a_k^{(n)}$ are zero except

[22] The definition of the Pochhammer symbol ambiguous [165, p. 414]. In the theory of special functions $(x)_n$ is used for the rising factorial

$$(x)_n \;\equiv\; x^{(n)} \;=\; x(x+1)(x+2)\cdots(x+n-1) \;=\; \frac{\Gamma(x+n)}{\Gamma(x)}\;,$$

whereas the same symbol is used in combinatorics for the falling factorial

$$(x)_n \;=\; x(x-1)(x-2)\cdots(x-n+1) \;=\; \frac{\Gamma(x+1)}{\Gamma(x-n+1)}\;,$$

The expression in terms of the Gamma function is unambiguous.

in the range $k \geq n$ and hence the relevant values fill an upper triangular $(N-1)\times(N-1)$ matrix $\mathrm{A} = \{a_{nk} = a_k^{(n)}\}$ with all diagonal elements being equal to unity, $a_n^{(n)} = 1$. In order to determine the coefficients C_n we apply the initial condition $g(s,0) = s^{n_0}$ and obtain from equations (5.75b) and (5.76) for $t = 0$:

$$s^{n_0} \;=\; \pi_0 \,+\, \pi_N s^N \,+ \sum_{n=1}^{N-1} C_n \sum_{k=1}^{N-1} a_k^{(n)}\,(1-s)^{k+1} \;\text{ and}$$

$$\sum_{n=1}^{N-1}\sum_{k=1}^{N-1} C_n a_k^{(n)}(1-s)^{k+1} \;=\; s^{n_0} - \pi_0 - \pi_N\, s^N \;=\; \sum_{k=1}^{N-1} b_k(1-s)^{k+1}\;,$$

where b_k results from a binomial expansion of the expression in the last equation

$$b_k \;=\; (-1)^k \left(\binom{N}{k+1}\pi_N \,-\, \binom{n_0}{k+1}\right)\;,$$

and the coefficients C_n are then calculated from the linear triangular system

$$\sum_{n=1}^{N-1} a_k^{(n)}\, C_n \;=\; b_k\,;\; k = 1,\ldots,N-1\;.$$

Alternatively, the C_k's can be calculated directly by means of a hypergeometric function

$$C_n \;=\; (-1)^{n+1}\, n_0\, \frac{(1-N)_n}{(n+1)_n}\, {}_3F_2(1-n_0,-n,n+1;2,1-N;1)\;, \qquad (5.75\mathrm{d})$$

which completes the exact solution of the neutral Moran master equation.

The selection case: $r \neq 0$. In the presence of selection we have $r \neq 0$ and the eigenvalue equation is change to

$$\lambda_n\,\sigma\,\psi_n(s) \;=\; (1-s)(\sigma-s)\frac{\mathrm{d}}{\mathrm{d}s}\left(N\psi_n(s) \,-\, s\frac{\mathrm{d}\psi_n(s)}{\mathrm{d}s}\right) \qquad (5.78)$$

with $\sigma = 1/(1+r)$. This ODE is known as Heun's equation [8]. The Heun polynomials, their eigenvalues have not yet been investigated as, for example, the hypergeometric functions and there are no explicit formulas for Heun's polynomials [135]. Nevertheless, knowledge of the results for the small r limit is often sufficient and then solutions of equation (5.78) can be obtained by perturbation theory on powers of r. Results in first order can be obtained by proper scaling from the solution of pure genetic drift ($r = 0$). A change in the auxiliary variable, $s \Rightarrow y = 1 - s/\sqrt{\sigma}$, is appropriate and leads to[23]

[23] The result for ε is easily obtained by making use of the infinite series for small x: $\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 \ldots$ and $1/\sqrt{1+x} = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{5}{16}x^3 \ldots$.

$$
\begin{aligned}
-\sqrt{\sigma}\,\lambda\,\psi \;&=\; \big(y^2-\epsilon(y-1)\big)\frac{\mathrm{d}}{\mathrm{d}y}\left(N\psi \,+\, (1-y)\frac{\mathrm{d}\psi}{\mathrm{ds}}\right) \quad \text{where} \\
\epsilon \;&=\; \sigma+\frac{1}{\sqrt{\sigma}}-2 \;=\; \frac{r^2}{4}+\mathcal{O}(r^3)\;.
\end{aligned}
\tag{5.79}
$$

Since the first non-vanishing term in the perturbation expansion of ϵ is $\propto r^2$, it is neglected in the first order perturbation calculation. After neglect of the $\mathcal{O}(\epsilon)$ term equation (5.79) takes on the same formal structure as the equation (5.73) for pure genetic drift that has been solved already and the probability generating function is now of the form

$$
\begin{aligned}
g(s,t) \;&=\; \pi_0 \,+\, \pi_N s^N \sum_{n=1}^{N-1} C_n^{(1)}\psi_n^{(1)}\, e^{-n(n+1)t/\sqrt{\sigma}} \,+\, \mathcal{O}(r^2) \quad \text{with} \\
\psi_n^{(1)} \;&=\; \sum_{k=1}^{N-1} a_k^{(n)} \left(1-\frac{s}{\sqrt{\sigma}}\right)^{k+1} \quad \text{with} \;\; \lambda_n^{(1)} \;=\; -\frac{n(n+1)}{\sqrt{\sigma}}\;.
\end{aligned}
\tag{5.80}
$$

The coefficients $a_k^{(n)}$ are the same as before and given in equation (5.75c') and the amplitudes $C_n^{(1)}$ are obtained again for the initial condition $g(s,0=s^{n_0}$ leading to

$$
b_k \;=\; (-1)^k \left(\binom{N}{k+1}\pi_N\sigma^{N/2} \,-\, \binom{n_0}{k+1}\sigma^{n_0/2}\right)\;.
$$

Second and higher order perturbation theory can be used to extend the range of validity of the approach but gives rise to quite clumsy expressions.

Another approximation is valid for large values of Ns and is base on the fact that then the term $s\,\partial g(s,t)/\partial s$ in comparable in size to $N\,g(s)$ only in the immediate neighborhood of $s=1$ and can be neglected therefore in the range $s\in[0,\sigma]$. The remaining approximate equation

$$
\sigma\,\frac{\partial g}{\partial t} \;=\; N(1-s)(\sigma-s)\frac{\partial g}{\partial s} \tag{5.81}
$$

can be solved exactly and yields

$$
g(s,t) \;=\; \left(\frac{(\sigma-s)e^{-Ns\,t} \,-\, \sigma(1-s)}{(\sigma-s)e^{-Ns\,t} \,-\, (1-s)}\right)^{n_0}\;. \tag{5.82}
$$

This equation was found to be a good approximation for the probability generating function for $Ns\geq 2$ on the interval $[0,\sigma]$ but equation (5.82) is not polynomial for $g(s,t$ and the determination of the probabilities $P(n,t)$ is numerically ill-conditioned except for small n. In particular, the expression for the probability of the loss of the allele A is very accurate:

$$P(0,t) \;=\; \left(\frac{1 \,-\, e^{-Nr\,t}}{1 \,+\, r \,-\, e^{-Nr\,t}}\right)^{n_0} \quad . \tag{5.83}$$

The stationary solution: $\lim t \,\to\, \infty$**.** The stationary solution of equation (5.69) fulfils the first order ODE

$$N\,\bar{g}(s) \,-\, s\,\frac{\mathrm{d}\bar{g}(s)}{\mathrm{ds}} \;=\; K \;=\; const.$$

that can be solved exactly and has the solution

$$\begin{aligned} \bar{g}(s) \;&=\; \pi_N\, s^N \,+\, \pi_0 \quad \text{with} \\ \pi_N \;&=\; \frac{n_0}{N} \quad \text{and} \quad \pi_0 \;=\; \frac{N-n_0}{N} \end{aligned} \tag{5.84}$$

in the neutral case, $r = 0$, where the two constants are determined by the two boundary conditions (5.71).

For the non-neutral condition, $r \neq 0$, the boundary condition (5.71b) has to replaced by (5.71b') and we obtain for the two constants:

$$\pi_N \;=\; \frac{1-\sigma^{n_0}}{1-\sigma^N} \quad \text{and} \quad \pi_0 \;=\; \frac{\sigma^{n_0}-\sigma^N}{1-\sigma^N} \;, \tag{5.84'}$$

where $\sigma = 1/(1+r)$ as before. The stationary probability can be calculated by comparison of coefficients:

$$\lim_{t\to\infty} P(n,t) \;=\; \bar{P}(n) \;=\; \pi_N\,\delta_{n,N} \,+\, \pi_0\,\delta_{n,0} \;, \tag{5.85}$$

where we can now identify π_N and π_0 as the total probability of fixation and the total probability for the loss of allele A, respectively.

5.3.2 Diffusion and neutral evolution

Again we choose the simple Moran model of pure genetic drift, $r = 0$, as an example. For large population sizes N it is appropriate to consider a new variable, $x \equiv n/N$, whereby $n = 0, 1, 2, \ldots$ is changed into $x = \frac{0}{N}, \frac{1}{N}, \frac{2}{N}, \ldots$, and in the limit $\lim n \,\to\, \infty$ the variable x becomes continuous. By this transformation the system space has become continuous on $x \in [0,1]$. Next, we make a Taylor expansion of the probabilities of the master equation

$$P(n{+}1,t) \;\Longrightarrow\; P\Big(x{+}\frac{1}{N},t\Big) = P(x,t) + \frac{1}{N}\frac{\partial P(x,t)}{\partial x} + \frac{1}{2N^2}\frac{\partial^2 P(x,t)}{\partial x^2} + \ldots$$

and obtain for the complete r.h.s. of equation (5.68):

$$\Big(x+\frac{1}{N}\Big)(1-x-\frac{1}{N})\left(P+\frac{1}{N}\frac{\partial P}{\partial x}+\frac{1}{2N^2}\frac{\partial^2 P}{\partial x^2}+\ldots\right)+$$
$$+\ \Big(x-\frac{1}{N}\Big)(1-x+\frac{1}{N})\left(P+\frac{1}{N}\frac{\partial P}{\partial x}-\frac{1}{2N^2}\frac{\partial^2 P}{\partial x^2}+\ldots\right)-$$
$$-\ 2x(1-x)\,P\ .$$

Insertion into (5.68) and expansion yields

$$\frac{\partial P(x,t)}{\partial t}\ =\ \frac{1}{N^2}\,\frac{\partial^2}{\partial x^2}\Big(x(1-x)\,P(x,t)\Big)\ +\ \mathcal{O}\Big(\frac{1}{N^3}\Big)\ .$$

The factor $1/N^2$ can be absorbed through a redefinition of time: $\tau \equiv 2t/N^2$ and in the limit $N \to \infty$ we obtain:

$$\frac{\partial P(x,\tau)}{\partial \tau}\ =\ \frac{1}{2}\,\frac{\partial^2}{\partial x^2}\Big(x(1-x)\,P(x,\tau)\Big)\ . \tag{5.86}$$

By means of the substitution the master equation (5.68) is changed into a standard Fokker-Planck equation with vanishing drift term: $A(x) = 0$ and $B(x) = x(1-x)$.

A consideration of the selection term, $r \neq 0$, as in equation (5.68), introduces a drift term into the Fokker-Planck equation

$$\frac{\partial P(x,\tau)}{\partial \tau}\ =\ -\frac{N\varrho}{2}\,\frac{\partial}{\partial x}\Big(x(1-x)\,P(x,\tau)\Big)\ +\ \frac{1}{2}\,\frac{\partial^2}{\partial x^2}\Big(x(1-x)\,P(x,\tau)\Big)\ , \tag{5.87}$$

with $\varrho = r/(1+r/2)$ and $\tau = (1+r/2)2t/N^2$. Now the equation contains a drift term $A(x) = N\varrho\ x(1-x)/2$.

The interpretation of the transformation to continuous space is straightforward: The transformed population is considered as a probability density $P(x,t)$, which migrates on a continuous state space, selection gives rise to a directed drift towards higher mean fitness of the population. The diffusion term describes the stochastic spreading.

Motoo Kimura [162, 163, 164] proposed the pure drift Fokker-Planck equation (5.86) for the stochastic evolution of a population of two alleles by random drift

$$\frac{\partial P(x,t)}{\partial t}\ =\ \frac{1}{4N}\,\frac{\partial^2}{\partial x^2}\Big(x(1-x)P(x,t)\Big)\ ,\ x\in]0,1[\ , \tag{5.86'}$$

which uses a slightly different transformation of the time axis. Kimura provides solution curves for the initial conditions $P(x,0) = \delta_{x,x_0}$:

$$\begin{aligned} P(x,t|x_0,0)\ =\ &\sum_{i=1}^{\infty} x_0(1-x_0)i(i+1)(2i+1)\,{}_2F_1(1-i,i+2,2,x_0)\cdot \\ &\cdot\ {}_2F_1(1-i,i+2,2,x)\,e^{-i(i+1)\,t/(4N)}\ , \end{aligned} \tag{5.88}$$

where ${}_2F_1$ is the conventional hypergeometric function. Despite the fact that equation (5.88) involves an infinite series convergence is commonly fast after a rather weakly converging first group of elements.

5.3.3 Comparison of Wright-Fisher and Moran models

A general comment of the often highly sophisticated analytical solutions of master and Fokker-Planck equations is appropriate at the end of this section:

5.4 Coalescent theory and backward equations

5.5 Stochastic modeling by numerical simulation

Chapter 6
Perspectives

Nothing in biology makes sense except in the light of evolution.
Theodosius Dobzhansky, 1972.

Abstract .

References

1. Abramowitz, M., Segun, I.A. (eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York (1965)
2. Abramson, M., Moser, W.O.J.: More birthday surprises. Amer. Math. Monthly **77**, 856–858 (1970)
3. Adams, W.J.: The Life and Times of the Central Limit Theorem, *History of Matheatics*, vol. 35, second edn. American Mathematical Society and London Mathematical Society, Providence, RI (2009). Articles by A. M. Lyapunov translated from the Russian by Hal McFaden.
4. Anderson, B.D.O.: Reverse-time diffusion equation models. Stochastic Processes and Applications **12**, 313–326 (1982)
5. Applebaum, D.: Lévy processes – From probability to finance and quantum groups. Notices Am. Math. Soc. **51**, 1336–1347 (2004)
6. Arfken, G.B., Weber, H.J.: Mathematical Methods for Physicists, fifth edn. Harcourt Academic Press, San Diego, CA (2001)
7. Arnold, L.: Stochastic Differential Equations. Theory and Applications. John Wiley & Sons, New York (1974)
8. Arscott, F.M.: Heun's equation. In: A. Ronveau (ed.) Heun's Differential Equations, pp. 3–86. Oxford University Press, New York (1955)
9. Athreya, K.B., Ney, P.E.: Branching Processes. Springer-Verlag, Heidelberg, DE (1972)
10. Bachelier, L.: Théorie de la spéculation. Annales scientifiques de l'É.N.S. 3^e série **17**, 21–86 (1900)
11. Bailey, N.T.J.: The Elements of Stochastic Processes with Application in the Natural Sciences. Wiley, New York (1964)
12. Bartholomay, A.F.: On the linear birth and death processes of biology as Markoff chains. Bull. Math. Biophys. **20**, 97–118 (1958)
13. Bartholomay, A.F.: Stochastic models for chemical reactions: I. Theory of the unimolecular reaction process. Bull. Math. Biophys. **20**, 175–190 (1958)
14. Bartholomay, A.F.: Stochastic models for chemical reactions: II. The unimolecular rate constant. Bull. Math. Biophys. **21**, 363–373 (1959)
15. Bazley, N.W., Montroll, E.W., Rubin, R.J., Shuler, K.E.: Studies in nonequilibrium rate processes: III. The vibrational relaxation of a system of anharmonic oscillators. J. Chem. Phys. **28**, 700–704 (1958). Erratum: *J.Chem.Phys.*, 29:1185-1186
16. Berry, R.S., Rice, S.A., Ross, J.: Physical Chemistry, second edn. Oxford University Press, New York (2000)

17. Biebricher, C.K., Eigen, M., William C. Gardiner, J.: Kinetics of RNA replication. Biochemistry **22**, 2544–2559 (1983)
18. Bienaymé, I.J.: Da la loi de Multiplication et de la durée des familles. Soc. Philomath. Paris Extraits **Ser. 5**, 37–39 (1845)
19. Billingsley, P.: Probability and Measure, third edn. Wiley-Interscience, New York (1995)
20. Billingsley, P.: Probability and Measure, anniversary edn. Wiley-Interscience, Hoboken, NJ (2012)
21. Blythe, R.A., McKane, A.J.: Stochastic models of evolution in genetics, ecology and linguistics. J. Stat. Mech,: Theor. Exp. (2007). P07018
22. Boas, M.L.: Mathematical Methods in the Physical Sciences, third edn. John Wiley & Sons, Hoboken, NJ (2006)
23. Boole, G.: An Investigation of the Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities. MacMillan, London (1854). Reprinted by Dover Publ. Co., New York, 1958
24. Born, M., Oppenheimer, R.: Zur Quantentheorie der Moleküle. Annalen der Physik **84**, 457–484 (1927). In German
25. Bouchaud, J.P., Georges, A.: Anomalous diffusion in disordered madia: Statistical mechanisms, models and physical applications. Physics Reports **195**, 127–293 (1990)
26. Brenner, S.: Theoretical biology in the third millenium. Phil. Trans. Roy. Soc. London B **354**, 1963–1965 (1999)
27. Brown, R.: A brief description of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants, and on the general existence of active molecules in organic and inorganic bodies. Phil. Mag., Series 2 **4**, 161–173 (1828). First Publication: The Edinburgh New Philosophical Journal. July-September 1828, pp.358-371.
28. Calaprice, A. (ed.): The Ultimate Quotable Einstein. Princeton University Press, Princeton, NJ (2010)
29. de Candolle, A.: Zur Geschichte der Wissenschaften und Gelehrten seit zwei Jahrhunderten nebst anderen Studien über wissenschaftliche Gegenstände insbesondere über Vererbung und Selektion beim Menschen. Akademische Verlagsgesellschaft, Leipzig, DE (1921). Deutsche Übersetzung der Originalausgabe *"Histoire des sciences et des savants depuis deux siècle"*, Geneve 1873, durch Wilhelm Ostwald.
30. Cao, Y., Gillespie, D.T., Petzold, L.R.: Efficient step size selection for the tau-leaping simulation method. J. Chem. Phys. **124**, 044,109 (2004)
31. Carter, M., van Brunt, B.: The Lebesgue-Stieltjes Integral. A Practical Introduction. Springer-Verlag, Berlin (2007)
32. Child, M.S.: Molecular Collision Theory. Dover Publications, Mineola, NY (1996)
33. Chung, K.L.: A Course in Probability Theory, *Probability and Mathematical Statistics*, vol. 21, second edn. Academic Press, New York (1974)
34. Chung, K.L.: Elementary Probability Theory with Stochastic Processes, 3rd edn. Springer-Verlag, New York (1979)
35. Cochran, W.G.: The distribution of quadratic forms in normal systems, with applications to the analysis of covariance. Math. Proc. Cambridge Phil. Soc. **30**, 178–191 (1934)
36. Conrad, K.: Probability distributions and maximum entropy. Expository paper, University of Connecticut, Storrs, CT (2005)
37. Cook, M., Soloveichik, D., Winfree, E., Bruck, J.: Programmability of chemical reaction networks. In: A. Condon, D. Harel, J.N. Kok, A. Salomaa, E. Winfree (eds.) Algorithimc Bioprocesses, *Natural Computing Series*, vol. XX, pp. 543–584. Springer-Verlag, Berlin (2009)

38. Cooper, B.E.: Statistics for Experimentalists. Pergamon Press, Oxford (1969)
39. Cortina Borja, M., Haigh, J.: The birthday problem. Significance **4**, 124–127 (2007)
40. Cover, T.M., Thomas, J.A.: Elements of Information Theory, second edn. John Wiley & Sons, Hoboken, NJ (2006)
41. Cox, D.R., Miller, H.D.: The Theory of Stochastic Processes. Methuen, London (1965)
42. Cox, R.T.: The Algebra of Probable Inference. The John Hopkins Press, Baltimore, MD (1961)
43. Craciun, G., Tang, Y., Feinberg, M.: Understanding bistability in complex enzyme-driven reaction networks. Proc. Natl. Acad. Sci. USA **103**, 8697–8702 (2006)
44. Crank, J.: The Mathematics of Diffusion. Clarendon Press, Oxford, UK (1956)
45. Crow, J.F., Kimura, M.: An Introduction to Population Genetics Theory. Sinauer Associates, Sunderland, MA (1970). Reprinted at *The Blackburn Press*, Caldwell, NJ, 2009
46. Cull, P., Flahive, M., Robson, R.: Difference Equations. From Rabbits to Chaos. Undergraduate Texts in Mathematics. Springer, New York (2005)
47. Darvey, I.G., Staff, P.J.: Stochastic approach to first-order chemical reaction kinetics. J. Chem. Phys. **44**, 990–997 (1966)
48. Demetrius, L., Schuster, P., Sigmund, K.: Polynucleotide evolution and branching processes. Bull. Math. Biol. **47**, 239–262 (1985)
49. Domingo, E., Parrish, C.R., John J, H. (eds.): Origin and Evolution of Viruses, second edn. Elsevier, Academic Press, Amsterdam, NL (2008)
50. Dudley, R.M.: Real Analysis and Probability. Wadsworth and Brooks, Pacific Grove, CA (1989)
51. Dyson, F.: A meeting with Enrico Fermi. How one intuitive physicist rescued a team from fruitless research. Nature **427**, 297 (2004)
52. Eddy, S.R.: What is Bayesian statistics? Nature Biotechnology **22**, 1177–1178 (2004)
53. Edwards, A.W.F.: Are Mendel's resulta really too close˙ Biological Reviews **61**, 295–312 (1986)
54. Ehrenfest, P., Ehrenfest, T.: Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. Z. Phys. **8**, 311–314 (1907)
55. Eigen, M.: Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften **58**, 465–523 (1971)
56. Eigen, M., McCaskill, J., Schuster, P.: The molecular quasispecies. Adv. Chem. Phys. **75**, 149–263 (1989)
57. Eigen, M., Schuster, P.: The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften **64**, 541–565 (1977)
58. Einstein, A.: Über die von der molekular-kinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annal. Phys. (Leipzig) **17**, 549–560 (1905)
59. Einstein, A.: Investigations on the Theory of the Brownian Movement. Dover Publications, New York (1956). Five original publications by Albert Einstein edited with notes by R. Fürth
60. Elliot, R.J., Anderson, B.D.O.: Reverse-time diffusions. Stochastic Processes and Applications **19**, 327–339 (1985)
61. Engl, H.W., Flamm, C., Kügler, P., Lu, J., Müller, S., Schuster, P.: Inverse problems in systems biology. Inverse Problems **25**, 123,014 (2009)
62. Evans, M., Hastings, N.A.J., Peacock, J.B.: Statistical Distributions, third edn. John Wiley & Sons, New York (2000)
63. Everett, C.J., Ulam, S.: Multiplicative systems I. Proc. Natl. Acad. Sci. USA **34**, 403–405 (1948)

64. Everett, C.J., Ulam, S.M.: Multiplicative systems in several variables I. Tech. Rep. LA-683, Los Alamos Scientific Laboratory (1948)
65. Everett, C.J., Ulam, S.M.: Multiplicative systems in several variables II. Tech. Rep. LA-690, Los Alamos Scientific Laboratory (1948)
66. Everett, C.J., Ulam, S.M.: Multiplicative systems in several variables III. Tech. Rep. LA-707, Los Alamos Scientific Laboratory (1948)
67. Ewens, W.J.: Mathematical Population Genetics. I. Theoretical Introduction, second edn. Interdisciplinary Applied Mathematics. Springer-Verlag, Berlin (2004)
68. Eyring, H.: The activated complex in chemical reactions. J. Chem. Phys. **3**, 107–115 (1935)
69. Feinberg, M.: Mathematical aspects of mass action kinetics. In: L. Lapidus, N.R. Amundson (eds.) Chemical Reactor Theory – A Review, pp. 1–78. Prentice Hall, Englewood Cliffs, NJ (1977)
70. Feinberg, M.: Lectures on Chemical Reaction Networks. Chemical Engineering & Mathematics. The Ohio State University, Columbus, OH (1979)
71. Feinberg, M.: Chemical oscillations, multiple equilibria, and reaction network structure. In: W.E. Stewart, W.H. Ray, C.C. Conley (eds.) Dynamics and Modelling of Reactive Systems, pp. 59–130. Academic Press, New York (1980)
72. Feinberg, M.: Chemical reaction network structure and the stability of complex isothermal reactors – II. Multiple steady states for networks of deficiency one. Chemical Engineering Science **43**, 1–25 (1988)
73. Feller, W.: The general form of the so-called law of the iterated logarithm. Trans. Amer. Math. Soc. **54**, 373–402 (1943)
74. Feller, W.: On the theory of stochastic processes, with particular reference to applications. In: The Regents of the University of California (ed.) Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, pp. 403–432. University of California Press, Berkeley, CA (1949)
75. Feller, W.: Diffusion processes in genetics. In: J. Neyman (ed.) Proc. 2nd Berkeley Symp. on Mathematical Statistics and Probability. University of Caifornia Press, Berkeley, CA (1951)
76. Feller, W.: An Introduction to Probability Theory and Its Application, vol. I, third edn. John Wiley & Sons, New York (1968)
77. Feller, W.: An Introduction to Probability Theory and Its Application, vol. II, second edn. John Wiley & Sons, New York (1971)
78. Fick, A.: Über diffusion. Annalen der Physik und Chemie **170 (4. Reihe 94)**, 59–86 (1855)
79. Field, R.J., Körös, E., Noyes, R.M.: Oscillations in chemical systems. II. Thorough analysis of temporal oscillations in the bromate-cerium-malonic acid system. J. Am. Chem. Soc. **94**, 8649–8664 (1972)
80. Field, R.J., Noyes, R.M.: Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction. J. Chem. Phys. **60**, 1877–1884 (1974)
81. Fisher, R.A.: Applications of "Student's" distribution. Metron **5**, 90–104 (1925)
82. Fisher, R.A.: Moments and product moments of sampling distributions. Proc. London Math. Soc. **Ser.2, 30**, 199–238 (1928)
83. Fisher, R.A.: The Genetical Theory of Natural Selection. Oxford University Press, Oxford, UK (1930)
84. Fisher, R.A.: Has Mendel's work been rediscovered? Annals of Science pp. 115–137 (1936)
85. Fisher, R.A.: The Design of Experiments, eighth edn. Hafner Publishing Company, Edinburgh, UK (1966)
86. Fisk, D.L.: Quasi-martingales. Trans. Amer. Math. Soc. **120**, 369–389 (1965)
87. Fisz, M.: Probability Theory and Mathematical Statistics, third edn. John Wiley & Sons, New York (1963)

88. Fisz, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik. VEB Deutscher Verlag der Wissenschaft, Berlin (1989). In German
89. Franklin, A., Edwards, A.W.F., Fairbanks, D.J., Hartl, D.L., Seidenfeld, T.: Ending the Mendel-Fisher Controversy. University of Pittburgh Press, Pittsburgh, PA (2008)
90. Galton, F.: The geometric mean in vital and social statistics. Proc. Roy. Soc. London **29**, 365–367 (1879)
91. Galton, F.: Natural Inheritance, second american edn. Macmillan & Co., London (1889). App. F, pp.241-248
92. Gardiner, C.W.: Handbook of Stochastic Methods, first edn. Springer-Verlag, Berlin (1983)
93. Gardiner, C.W.: Stochastic Methods. A Handbook for the Natural Sciences and Social Sciences, fourth edn. Springer Series in Synergetics. Springer-Verlag, Berlin (2009)
94. Gardiner, C.W., Zoller, P.: Quantum Noise – A Handbook of Markovian and Non-Markovien Quantum Stochastic Mehtods with Applications to Quantum Optics, first edn. Springer-Verlag, Berlin (2004)
95. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Baysian Data Analysis, second edn. Texts in Statistical Science. Chapman & Hall / CRC, Boca Raton, FL (2004)
96. George, G.: Testing for the independence of three events. Mathematical Gazette **88**, 568 (2004)
97. Georgii, H.: Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik, third edn. Walter de Gruyter GmbH & Co., Berlin (2007). In German. English translation: *Stochastics. Introduction to Probability and Statstics.* Walter de Gruyter GmbH & Co. Berlin (2008).
98. Gibbs, J.W.: Elementary Principles in Statistical Mechanics. Charles Scribner's Sons, New York (1902). Reprinted 1981 by Ox Bow Press, Woodbridge, CT
99. Gibbs, J.W.: The Scientific Papers of J. Willard Gibbs, Vol.I, Thermodynamics. Dover Publications, New York (1961)
100. Gihman, I.F., Skorohod, A.V.: The Theory of Stochastic Processes. Vol. I, II, and III. Springer-Verlag, Berlin (1975)
101. Gillespie, D.T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comp. Phys. **22**, 403–434 (1976)
102. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**, 2340–2361 (1977)
103. Gillespie, D.T.: Markov Processes: An Introduction for Physical Scientists. Academic Press, San Diego, CA (1992)
104. Gillespie, D.T.: A rigorous derivation of the chemical master equation. Physica A **188**, 404–425 (1992)
105. Gillespie, D.T.: Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. Phys. Rev. E **54**, 2084–2091 (1996)
106. Gillespie, D.T.: Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. **58**, 35–55 (2007)
107. Gillies, D.: Varieties of propensity. Brit. J. Phil. Sci. **51**, 807–853 (2000)
108. Goel, N.S., Richter-Dyn, N.: Stochastic Models in Biology. Academic Press, New York (1974)
109. Gradstein, I.S., Ryshik, I.M.: Tables of Series, Products, and Integrals, vol. 1. Verlag Harri Deutsch, Thun, DE (1981). In German and English. Translated from Russian by Ludwig Boll, Berlin
110. Gray, R.M.: Entropy and Information Theory, second edn. Springer, New York (2011)
111. Griffiths, A.J.F., Wessler, S.R., Caroll, J.B., Doebley, J.: An Introduction to Genetic Analysis, tenth edn. W. H. Freeman, New York (2012)

112. Grünbaum, B.: Venn diagrams and independent falilies of sets. Mathematics Magazine **48**, 12–23 (1975)
113. Grünbaum, B.: The construction of Venn diagrams. The College Mathematics Journal **15**, 238–247 (1984)
114. Gunawardena, J.: Chemical reaction network theory for *in-silico* biologists. Tech. rep., Bauer Center for Genomics Research at Harvard University, Cambridge, MA (2003)
115. Hájek, A.: Interpretations of probability. In: E.N. Zalta (ed.) The Stanford Encyclopedia of Philosophy, Winter 2012 edn. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford Universiy, Stanford, CA. World Wide Web URL: `http://plato.stanford.edu/entries/probability-interpret/` (2013). Retrieved January 23, 2013
116. Hamilton, J.D.: Time Series Analysis. Princeton University Press, Princeton, NJ (1994)
117. Hamilton, W.R.: On a general method in dynamics. Phil. Trans. Roy. Soc. London **II** for 1834, 247–308 (1834)
118. Hamilton, W.R.: Second essay on a general method in dynamics. Phil. Trans. Roy. Soc. London **I** for 1835, 95–144 (1835)
119. Harris, T.E.: Branching Processes. Springer-Verlag, Berlin (1963)
120. Harris, T.E.: The Theory of Branching Processes. Dover Publications, New York (1989)
121. Hartl, D.L., Clark, A.G.: Principles of Population Genetics, third edn. Sinauer Associates, Sunderland, MA (1997)
122. Hartman, P., Wintner, A.: On the law of the iterated logarithm. American Journal of Mathematics **63**, 169–173 (1941)
123. Haubold, H.J., Mathai, M.A., Saxena, R.K.: Mittag-Leffler functions and their applications. J. Appl. Math. **2011**, e298,628 (2011)
124. Haussmann, U.G., Pardoux, E.: Time reversal of diffusions. Annals Probability **14**, 1188–1205 (1986)
125. Hawkins, D., Ulam, S.: Theory of multiplicative processes I. Tech. Rep. LADC-265, Los Alamos Scientific Laboratory (1944)
126. Heathcote, C.R., Moyal, J.E.: The random walk (in continuous time) and its application to the theory of queues. Biometrika **46**, 400–411 (1959)
127. Heinrich, R., Sonntag, I.: Analysis of the selection equation for a multivariable population model. Deterministic anad stochastic solutios and discussion of the approach for populations of self-reproducing biochemical networks. J. theor. Biol. **93**, 325–361 (1981)
128. Heyde, C.C., Seneta, E.: Studies in the history of probability and statistics. xxxi. the simple branching porcess, a turning point test and a fundmanetal inequality: A historical note on I. J. Bienaymé. Biometrika **59**, 680–683 (1972)
129. Hinshelwood, C.N.: On the theory of unimolecular reactions. Proc. Roy. Soc. LOndon A **113**, 230–233 (1926)
130. Hocking, R.L., Schwertman, N.C.: An extension of the birthday problem to exactly k matches. The College Mathematics Journal **17**, 315–321 (1986)
131. Hogg, R.V., McKean, J.W., Craig, A.T.: Introduction to Mathematical Statistics, seventh edn. Pearson Education, Upper Saddle River, NJ (2012)
132. Hogg, R.V., Tanis, E.A.: Probability and Statistical Inference, eighth edn. Pearson – Prentice Hall, Upper Saddle River, NJ (2010)
133. Holdren, J.P., Lander, E., Varmus, H.: Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. President's Council of Advisors on Science and Technology, Washington, DC (2010)
134. Horn, F., Jackson, R.: General mass action kinetics. Archive for Rational Mechanics and Analysis **47**, 81–116 (1972)

135. Houchmandzadeh, B., Vallade, M.: An alternative to the diffusion equation in population genetics. Phys. Rev. E **83**, e051,913 (2010)
136. Houston, P.L.: Chemical Kinetics and Reaction Dynamics. The McGraw-Hill Companies, New York (2001)
137. Humphries, N.E., Queiroz, N., Dyer, J.R.M., Pade, N.G., Musyl, M.K., Schaefer, K.M., Fuller, D.W., Brunnschweiler, J.M., Doyle, T.K., Houghton, J.D.R., Hays, G.C., Jones, C.S., Noble, L.R., Wearmouth, V.J., Southall, E.J., Sims, D.W.: Environmental context explains Lévy and Brwonian movement patterns of marine predators. Nature **465**, 1066–1069 (2010)
138. Inagaki, H.: Selection under random mutations in stochastic Eigen model. Bull. Math. Biol. **44**, 17–28 (1982)
139. Ishida, K.: Stochastic model for bimolecular reaction. J. Chem. Phys. **41**, 2472–2478 (1964)
140. Itō, K.: Stochastic integral. Proc. Imp. Acad. Tokyo **20**, 519–524 (1944)
141. Itō, K.: On stochastic differential equations. Mem. Amer. Math. Soc. **4**, 1–51 (1951)
142. Jackson, E.A.: Perspectives of Nonlinear Dynamics, vol. 1. Cambridge University Press, Cambridge, UK (1989)
143. Jackson, E.A.: Perspectives of Nonlinear Dynamics, vol. 2. Cambridge University Press, Cambridge, UK (1989)
144. Jacobs, K.: Stochastic processes for Physicists. Understanding Noisy Systems. Cambridge University Press, Cambridge, UK (2010)
145. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**, 620–630 (1957)
146. Jaynes, E.T.: Information theory and statistical mechanics. II. Phys. Rev. **108**, 171–190 (1957)
147. Jaynes, E.T.: Probability Theory. The Logic of Science. Cambridge University Press, Cambridge, UK (2003)
148. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, *Probability and Mathematical Statistics. Applied Probability and Statistics*, vol. 1, second edn. John Wiley & Sons, New York (1994)
149. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, *Probability and Mathematical Statistics. Applied Probability and Statistics*, vol. 2, second edn. John Wiley & Sons, New York (1995)
150. Jones, B.L., Enns, R.H., Rangnekar, S.S.: On the theory of selection of coupled macromolecular systems. Bull. Math. Biol. **38**, 15–28 (1976)
151. Jones, B.L., Leung, H.K.: Stochastic analysis of a non-linear model for selection of biological macromolecules. Bull. Math. Biol. **43**, 665–680 (1981)
152. Joyce, G.F.: Forty years of *in vitro* evolution. Angew. Chem. Internat. Ed. **46**, 6420–6436 (2007)
153. Kassel, L.S.: Studies in homogeneous gas reactions I. J. Phys. Chem. **32**, 225–242 (1928)
154. Kendall, D.G.: Branching processes since 1873. J. of the London Mathematical Society **41**, 386–406 (1966)
155. Kendall, D.G.: The genalogy of genealogy: Branching processes before (an after) 1873. Bull. of the London Mathematical Society **7**, 225–253 (1975)
156. Kenney, J.F., Keeping, E.S.: Mathematics of Statistics, second edn. Van Nostrand, Princeton, NJ (1951)
157. Kenney, J.F., Keeping, E.S.: The k-Statistics. In Mathematics of Statistics. Part I, § 7.9, third edn. Van Nostrand, Princeton, NJ (1962)
158. Kesten, H., Stigum, B.P.: A limit theorem for multidimensional Galton-Watson processes. Annal. Math. Statistics **37**, 1211–1223 (1966)
159. Keynes, J.M.: A Treatise on Probability. MacMillan & Co., London (1921)
160. Khinchin, A.Y.: Über einen Satz der Wahrscheinlichkeitsrechnung. Fundamenta Mathematica **6**, 9–20 (1924). In German

161. Kim, S.K.: Mean first passage time for a random walker and its application to chemical knietics. J. Chem. Phys. **28**, 1057–1067 (1958)
162. Kimura, M.: Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA **41**, 144–150 (1955)
163. Kimura, M.: Diffusion models in population genetics. J. Appl. Prob. **1**, 177–232 (1964)
164. Kimura, M.: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK (1983)
165. Knuth, D.E.: Two notes on notation. Am. Math. Monthly **99**, 403–422 (1992)
166. Kolmogorov, A.N.: Über das Gesetz es interierten Logarithmus. Mathematische Annalen **101**, 126–135 (1929). In German
167. Kolmogorov, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, Berlin (1933). English translation: *Foundations of Probability*. Chelsea Publ. Co. New York, 1950
168. Kolmogorov, A.N., Dmitriev, N.A.: "Zur Lösung einer biologischen Aufgabe". Isvestiya Nauchno-Issledovatel'skogo Instituta Matematiki i Mekhaniki pri Tomskom Gosudarstvennom Universitete **2**, 1–12 (1938)
169. Kolmogorov, A.N., Dmitriev, N.A.: Branching stochastic processes. Doklady Akad. Nauk U.S.S.R. **56**, 5–8 (1947)
170. Kowalski, C.J.: Non-normal bivariate distributions with normal marginals. The American Statistician **27**, 103–106 (1973)
171. Laidler, K.J.: Chemical Kinetics, third edn. Addison Wesley, Boston, MA (1987)
172. Laidler, K.J., King, M.C.: The development of transition-state theory. J. Phys. Chem. **87**, 2657–2664 (1983)
173. Langevin, P.: Sur la théorie du mouvement Brownien. Comptes Rendues hebdomadaires des Séances de L'Academié des Sceinces **146**, 530–533 (1908)
174. Laplace, P.S.: Essai philosophique les probabililtés. Courcies Imprimeur, Paris (1814). English edition: A Philosophical Eaasay on Probabilties. Dover Publications, New York, 1951
175. Lauritzen, S.L.: Time series analysis in 1880: A discussion of contributions made by t. n, thiele. International Statistical Review **49**, 319–331 (1981)
176. Lee, P.M.: Bayesian Statistics, third edn. Hodder Arnold, London (2004)
177. Lévy, P.: Calcul de probabilités. Geuthier-Villars, Paris (1925). In French
178. Li, T., Kheifets, S., Medellin, D., Raizen, M.G.: Measurement of the instantaneous velocity of a Brownian particle. Science **328**, 1673–1675 (2010)
179. Limpert, E., Stahel, W.A., Abbt, M.: Log-normal distributions across the sciences: Keys and clues. BioScience **51**, 341–352 (2001)
180. Lin, S.H., Lau, K.H., Richardson, W., Volk, L., Eyring, H.: Stochastic model of unimolecular reactions and the RRKM theory. Proc. Nat. Acad. Sci. USA **69**, 2778–2782 (1972)
181. Lindeberg, J.W.: Über das Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. Annales Academiae Scientiarum Fennicae **16**, 1–23 (1920). In German.
182. Lindeberg, J.W.: Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift **15**, 211–225 (1922). In German
183. Lindemann, F.A.: Discussion on the radiation theory on chemical action. Trans. Farad. Soc. **17**, 598–606 (1922)
184. Liouville, J.: Note sur la théorie de la variation des constantes arbitraires. Journal de Mathématiques pure et appliquées **3**, 342–349 (1838). In French.
185. Liouville, J.: Mémoire sur l'intégration des équations différentielles du mouvement quelconque de points matériels. Journal de Mathématiques pure et appliquées **14**, 257–299 (1849). In French.

186. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atmospheric Sciences **20**, 130–141 (1963)
187. Lu, J., Engl, H.W., Rainer Machné, Schuster, P.: Inverse bifurcation analysis of a model for the mammalian G1/S regulatory module. Lecture Notes in Computer Science **4414**, 168Ű184 (2007)
188. Lu, J., Engl, H.W., Schuster, P.: Inverse bifurcation analysis: Application to simple gene systems. ABM – Algorithms for Molecular Biology **1**, e11 (2006)
189. Lyapunov, A.M.: Sur une proposition de la théorie des probabilités. Bull. Acad. Imp. Sci. St. Pétersbourg **13**, 359–386 (1900)
190. Lyapunov, A.M.: Nouvelle forme du théorème sur la limite des probabilités. Mem. Acad. Imp. Sci. St. Pétersbourg, Classe Phys. Math. **12**, 1–24 (1901)
191. Mahnke, R., Kaupužs, J., Lubashevsky, I.: Physics of Stochastic Processes. How Randomness Acts in Time. Wiley-VCh Verlag, Weinheim (Bergstraße), DE (2009)
192. Mallows, C.: Anothre comment on O'Cinneide. The American Statistician **45**, 257 (1991)
193. Mandelbrot, B.B.: The Fractal Geometry of Nature, updated edn. W. H. Freeman Company, New York (1983)
194. Marcus, R.A.: Unimolecular dissociations and free radical recombination reactions. J. Chem. Phys. **20**, 359–364 (1952)
195. Marcus, R.A.: Unimolecular reactions, rates and quantum state distributions of products. Phil. Trans. Roy. Soc. London A **332**, 283–296 (1990)
196. Marcus, R.A., Rice, O.K.: The kinetics of the recombination of methyl radical and iodine atoms. J. Phys. Colloid Chem. **55**, 894–908 (1951)
197. Maruyama, T.: Stochastic Problems in Population Genetics. Springer-Verlag, Berlin (1977)
198. Mathai, A.M., Saxena, R.K., Haubold, H.J.: A certain class of Laplace transforms with applications to reaction and reaction-diffusion equations. Astrophys. Space Sci. **305**, 283–288 (2006)
199. McAlister, D.: The law of the geometric mean. Proc. Roy. Soc. London **29**, 367–376 (1879)
200. McCaskill, J.S.: A stochastic theory of macromolecular evolution. Biol. Cybern. **50**, 63–73 (1984)
201. McKean, Jr., H.P.: Stochastic Integrals. John Wiley & Sons, New York (1969)
202. McQuarrie, D.A.: Kinetics of small systems. I. J. Chem. Phys. **38**, 433–436 (1962)
203. McQuarrie, D.A.: Stochastic approach to chemical kinetics. J. Appl. Prob. **4**, 413–478 (1967)
204. McQuarrie, D.A.: Mathematical Methods for Scientists and Engineers. University Science Books, Sausalito, CA (2003)
205. McQuarrie, D.A., Jachimowski, C.J., Russell, M.E.: Kinetics of small systems. II. J. Chem. Phys. **40**, 2914–2921 (1964)
206. Medvegyev, P.: Stochastic Integration Theory. Oxford University Press, New York (2007)
207. Meinhardt, H.: Models of Biological Pattern Formation. Academic Press, London (1982)
208. Meintrup, D., Schäffler, S.: Stochastik. Theorie und Anwendungen. Springer-Verlag, Berlin (2005). In German
209. Melnick, E.L., Tenenbein, A.: Misspecifications of the normal distribution. The American Statistician **36**, 372–373 (1982)
210. Mendel, G.: Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereins in Brünn **IV**, 3–47 (1866). In German
211. Merkle, M.: Jensen's inequality for medians. Statistics & Probability Letters **71**, 277–281 (2005)

212. Messiah, A.: Quantum Mechanics, vol. II. North-Holland Publishing Company, Amsterdam, NL (1970). Translated from the French by J. Potter
213. Metzler, R., Klafter, J.: The random walk's guide to anomalous diffusion: A fractional dynamics approach. Physics Reports **339**, 1–77 (2000)
214. Michaelis, L., Menten, M.L.: The kinetics of the inversion effect. Biochemische Zeitschrift **49**, 333–369 (1913)
215. Miller, R.W.: Propensity: Popper or Peirce? Brit. J. Phil. Sci. **26**, 123–132 (1975)
216. Mittag-Leffler, M.G.: Sur la nouvelle fonction $E_\alpha(x)$. C. R. Acad. Sci. Paris, Ser.II, **137**, 554–558 (1903)
217. Montroll, E.W.: Stochastic processes and chemical kinetics. In: W.M. Muller (ed.) Energetics in Metallurgical Phenomenon, vol. 3, pp. 123–187. Gordon & Breach, New York (1967)
218. Montroll, E.W., Shuler, K.E.: Studies in nonequilibrium rate processes: I. The relaxation of a system of harmonic oscillators. J. Chem. Phys. **26**, 454–464 (1956)
219. Montroll, E.W., Shuler, K.E.: The application of the theory of stochastic processes to chemical kinetics. Adv. Chem. Phys. **1**, 361–399 (1958)
220. Montroll, E.W., Weiss, G.H.: Random walks on lattices. II. J. Math. Phys. **6**, 167–181 (1965)
221. Moore, G.E.: Cramming more components onto intergrated circuits. Electronics **38**(8), 4–7 (1965)
222. Moran, P.A.P.: Random processes in genetics. Proc. Camb. Phil. Soc. **54**, 60–71 (1958)
223. Moran, P.A.P.: The Statistical Processes of Evolutionary Theroy. Clarendon Press, Oxford, UK (1962)
224. Nicolis, G., Prigogine, I.: Self-Organization in Nonequilibrium Systems. John Wiley & Sons, New York (1977)
225. Nolan, J.P.: Stable Distributions: Models for Heavy-Tailed Data. Birkhäuser, Boston, MA (2013). Unfinished manuscript. Online at `academic2.american.edu/∼jpnolan`.
226. Novitski, C.E.: On Fisher's criticism of Mendel's results with the garden pea. Genetics **166**, 1133–1136 (2004)
227. Novitski, C.E.: Revision of Fisher's analysis of Mendel's garden pea experiments. Genetics **166**, 1139–1140 (2004)
228. Nyman, J.E.: Another generalization of the birthday problem. Mathematics Magazine **48**, 46–47 (1975)
229. Øksendal, B.K.: Stochastic Differential Equations. An Introduction with Applications, sixth edn. Springer-Verlag, Berlin (2003)
230. Olbregts, J.: Termolecular reaction of nitrogen monoxide and oxygen. A still unsolved problem. Internat. J. Chem. Kinetics **17**, 835–848 (1985)
231. Park, S.Y., Bera, A.K.: Maximum entropy autoregressive conditional heteroskedasticy model. J. Econometrics **150**, 219–230 (2009)
232. Paschotta, R.: Field Guide to Laser Pulys Generation. SPIE Press, Bellingham, WA (2008)
233. Patrick, R., Golden, D.M.: Third-order rate constants of atmospheric importance. Internat. J. Chem. Kinetics **15**, 1189–1227 (1983)
234. Pearson, E.S., Wishart, J.: "Student's" Collected Papers. Cambridge University Press, Cambridge, UK (1942). Cambridge University Press for the Biometrika Trustees
235. Pearson, J.A.: Advanced Statistical Physics. University of Manchester, Manchester, UK (2009). URL: `http://www.joffline.com/`
236. Pearson, K.: On the criterion that a given system of deviations form the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 **50**(302), 157–175 (1900)

237. Peirce, C.S.: Vol.7: Science and philosophy and Vol.8: Reviews, correspondence, and bibliography. In: A.W. Burks (ed.) The Collected Papers of Charles Sanders Peirce, vol. 7-8. Belknap Press of Harvard University Press, Cambridge, MA (1958)
238. Penney, W.: Problem: Penney-Ante. J. Recreational Math. **2**(October), 241 (1969)
239. Philibert, J.: One and a half century of diffusion: Fick, Einstein, before and beyond. Diffusion Fundamentals **4**, 6.1–6.19 (2006)
240. Phillipson, P.E., Schuster, P.: Modeling by Nonlinear Differential Equations. Dissipative and Conservative Processes, *World Scientific Series on Nonlinear Science A*, vol. 69. World Scientific, Singapore (2009)
241. Plass, W.R., Cooks, R.G.: A model for energy transfer in inelasitc molecular collisions applicable at steady state and non-steady state and for an arbitrary distribution of collision energies. J. Am. Soc. Mass Spectrom. **14**, 1348–1359 (2003)
242. Popper, K.: The propensity interpretation of the calculus of probability and of the quantum theory. In: S. Körner, M.H.L. Price (eds.) Observation and Interpretation in the Philosophy of Physics: Proceedings of the Ninth Symposium of the Colston Research Society. Butterworth Scientific Publications, London (1957)
243. Popper, K.: The propensity theory of probability. Brit. J. Phil. Sci. **10**, 25–62 (1960)
244. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes. The Art of Scientific Computing. Cambridge University Press, Cambridge, UK (1986)
245. Price, R.: LII. an essay towards soliving a problem in the doctrine of chances. By the late Ref. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M.A. and F.R.S. Phil. Trans. Roy. Soc. London **53**, 370–418 (1763)
246. Protter, P.E.: Stochastic Intergration and Differential Equations, *Applications of Mathematics*, vol. 21, second edn. Springer-Verlag, Berlin (2004)
247. Qian, H., Elson, E.L.: Single-molecule enzymology: Stochastic Michaelis-Menten kinetics. Biophys. Chem. **101-102**, 565–576 (2002)
248. Rice, O.K., Ramsperger, H.C.: Theories of unimolecular gas reactions at low pressures. J. Am. Chem. Soc. **49**, 1617–1629 (1927)
249. Riley, K.F., Hobson, M.P., Bence, S.J.: Mathematical Methods for Physics and Engineering, second edn. Cambridge University Press, Cambridge, UK (2002)
250. Risken, H.: TheFokker-Planck Equation. Methods of Solution and Applications, 2nd edn. Springer-Verlag, Berlin (1989)
251. Robinett, R.W.: Quantum Mechanics. Classical Results, Modern Systems, and Visualized Examples. Oxford University Press, New York (1997)
252. Sagués, F., Epstein, I.R.: Nonlinear chemical dynamics. J. Chem. Soc.,Dalton Trans. **2003**, 1201–1217 (2003)
253. Schilling, M.F., Watkins, A.E., Watkins, W.: Is human height bimodal? The American Statistician **56**, 223–229 (2002)
254. Schlögl, F.: Chemical reaction models for non-equilibrium phase transitions. Z. Physik **253**, 147–161 (1972)
255. Schubert, M., Weber, G.: Quantentheorie. Grundlagen und Anwendungen. Spektrum Akademischer Verlag, Heidelberg, DE (1993). In German
256. Schuster, P., Sigmund, K.: Random selection - A simple model based on linear birth and death processes. Bull. Math. Biol. **46**, 11–17 (1984)
257. Selmeczi, D., Tolić-Nørrelykke, S., Schäffer, E., Hagedorn, P.H., Mosler, S., Berg-Sørensen, K., Larsen, N.B., Flyvbjerg, H.: Brownian motion after Einstein: Some new applications and new experiments. Lect. Notes Phys. **711**, 181–199 (2007)

258. Seneta, E.: Non-negative Matrices and Markov Chains, second edn. Springer-Verlag, New York (1981)
259. Seneta, E.: The central limit problem and lienear least squares in pre-revolutionary Russia: The background. Mathematical Scientist **9**, 37–77 (1984)
260. Seydel, R.: Practical Bifurcation and Stability Analysis. From Equilibrium to Chaos, *Interdisciplinary Applied Mathematics*, vol. 5, second edn. Springer-Verlag, New York (1994)
261. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**, 379–423 (1948)
262. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. Univ. of Illinois Press, Urbana, IL (1949)
263. Sharpe, M.J.: Transformations of diffusion by time reversal. The Annals of Probability **8**, 1157–1162 (1980)
264. Shuler, K.E.: Studies in nonequilibrium rate processes: II. The relaxation of vibrational nonequilibrium distributions in chemical reactions and shock waves. J. Phys. Chem. **61**, 849–856 (1957)
265. Shuler, K.E., Weiss, G.H., Anderson, K.: Studies in nonequilibrium rate processes. V. The relaxation of moments derived from a master equation. J. Math. Phys. **3**, 550–556 (1962)
266. Steffensen, J.F.: "deux problème du calcul des probabilités". Ann. Inst. Henri Poincaré **3**, 319–344 (1933)
267. Stepanow, S., Schütz, G.M.: The distribution function os a semiflexible polymer and random walks with constraints. Europhys. Letters **60**, 546–551 (2002)
268. Stevens, J.W.: What is Bayesian Statistics? What is ... ? Hayward Medical Communications, a division of Hayward Group Ltd., London (2009)
269. Stratonovich, R.L.: Introduction to the Theory of Random Noise. Gordon and Breach, New York (1963)
270. Stuart, A., Ord, J.K.: Kendall's Advanced Theory of Statistics. Volume 1: Distribution Theory, fifth edn. Charles Griffin & Co., London (1987)
271. Stuart, A., Ord, J.K.: Kendall's Advanced Theory of Statistics. Volume 2: Classical Inference and Relationship, fifth edn. Edward Arnold, London (1991)
272. Student: The probable error of a mean. Biometrika **6**, 1–25 (1908)
273. Swamee, P.K.: Near lognormal distribution. J. Hydrological Engineering **7**, 441–444 (2007)
274. Swetina, J., Schuster, P.: Self-replication with errors - A model for polynucleotide replication. Biophys. Chem. **16**, 329–345 (1982)
275. Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. Society for Industrial and Applied Mathematics, Philadelphia, PA (2005)
276. Thiele, T.N.: Om Anvendelse af midste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfædige Feijlkilder giver Feijlene en 'systenatisk' Karakter. Vidensk. Selsk. Skr. 5. rk., naturvid. og mat. Afd. **12**, 381–408 (1880). In Danish.
277. Thompson, C.J., McBride, J.L.: On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. Math. Biosci. **21**, 127–142 (1974)
278. Tolman, R.C.: The Principle of Statistical Mechanics. Oxford University Press, Oxford, UK (1938)
279. Tsukahara, H., Ishida, T., Mayumi, M.: Gas-phase oxidation of nitric oxide: Chemical kinetics and rate constant. Nitric Oxide: Biology and Chemistry **3**, 191–198 (1999)
280. Turing, A.M.: The chemical basis of morphogenesis. Phil. Trans. Roy. Soc. London B **237**(641), 37–72 (1952)

281. Uhlenbeck, G.E., Ornstein, L.S.: On the theory of the Brownian motion. Phys. Rev. **36**, 823–841 (1930)
282. Ullah, M., Wolkenhauer, O.: Family tree of Markov models in systems biology. IET Systems Biology **1**, 247–254 (2007)
283. Ullah, M., Wolkenhauer, O.: Stochastic Approaches for Systems Biology. Springer, New York (2011)
284. van den Berg, T.: Calibrating the Ornstein-Uhlenbeck-Vasicek model. Sitmo – Custom Financial Research and Development Services, `www.sitmo.com/article/calibrating-the-ornstein-uhlenbeck-model/` (2011)
285. van Kampen, N.G.: A power series expansion of the master equation. Can. Chem. Phys. **39**, 551–567 (1961)
286. van Kampen, N.G.: The expansion of the master equation. Adv. Chem. Phys. **34**, 245–309 (1976)
287. van Kampen, N.G.: Stochastic Processes in Physics and Chemistry, third edn. Elsvier, Amsterdam (2007)
288. Vasicek, O.: An equlibrium characterization of the term structure. J. Financial Economics **5**, 177–188 (1977)
289. Venn, J.: On the diagrammatic and mechanical representation of propositions and reasonings. The London, Edinburgh, and Diblin Philosophical Magazine and Journal of Science **9**, 1–18 (1880)
290. Venn, J.: Sybolic Logic. MacMillan, London (1881). Second edition, 1984. Reprinted by Lenox Hill Pub. & Dist. Co., 1971
291. Venn, J.: The Logic of Chance. An Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to its Logical Bearings and its Application to Moral and Social Science, and to Statistics, third edn. MacMillan and Co., London (1888)
292. Verhulst, P.: Notice sur la loi que la population pursuit dans son accroisement. Corresp. Math. Phys. **10**, 113–121 (1838)
293. Viswanathan, G.M., Raposo, E.P., da Luz, M.G.E.: Lévy flights and superdiffusion in the context of biological encounters and random searches. Physics of Life Reviews **5**, 133–150 (2008)
294. Vitali, G.: Sul problema della misura dei gruppi di pinti di una retta. Gamberini E. Parmeggiani, Bologna (1905)
295. Vitali, G.: Sui gruppi di punti e sulle funzioni di variabili reali. Atti dell'Accademia delle Science di Torino **43**, 75–92 (1908)
296. Volkenshtein, M.V.: Entropy and Information, *Progress in Mathematical Physics*, vol. 57. Birkhäuser Verlag, Basel, CH (2009). German version: W. Ebeling, Ed. Entropie und Information. Wissenschaftliche Taschenbücher, Band 306, Akademie-Verlag, Berlin 1990. Russian Edition: Nauka Publ., Moscow 1986
297. von Mises, R.: Über Aufteilungs- und Besetzungswahrscheinlichkeiten. Revue de la Faculté des Sciences de l'Université d'Istanbul, N.S. **4**, 145–163 (1938-1939). In German. Reprinted in *Selected Papers of Richard von Mises*, vol.2, American Mathematical Society, 1964, pp. 313-334
298. von Smoluchowski, M.: Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. Annal. Phys. (Leipzig) **21**, 756–780 (1906)
299. Waage, P., Guldberg, C.M.: Studies concerning affinity. J. Chem. Education **63**, 1044–1047 (1986). English translation by Henry I. Abrash
300. Watson, H.W., Galton, F.: On the probability of the extinction of families. J. Anthropological Institute of Great Britain and Ireland **4**, 138–144 (1875)
301. Weber, N.A.: Dimorphism of the African *oecophylla* worker and an anomaly (*hymenoptera formicidae*). Annals of the Entomological Society of America **39**, 7–10 (1946)

302. Wei, W.W.S.: Time Series Analysis. Univariate and Multivariate Methods. Addison-Wesley Publishing Company, Redwood City, CA (1990)
303. Williams, D.: Diffusions, Markov Processes and Martingales. Volume 1: Foundations. John Wiley & Sons, Chichester, UK (1979)
304. Wright, S.: Evolution in Mendelian populations. Genetics **16**, 97–159 (1931)
305. Zhabotinsky, A.M.: A history of chemical oscillations and waves. Chaos **1**, 379–386 (1991)

Glossary

The glossary provides short explanations for alphabetically ordered notions used frequently in the text.

Bijection: one-to-one correspondence

Closure:

Commensurability: Two non-zero real numbers a and b are commensurable if and only if a/b is rational number implying the existence of some real number γ such that $a = m\gamma$ and $b = n\gamma$ with m and n being integers.

Epigenetics, see genetics and epigenetics.

Force: The notion of force became quantifiable through the second law of Newtonian mechanics: $\mathbf{f} = m\mathbf{a}$, force is a vector and the acceleration $\mathbf{a} = d^2\mathbf{r}/dt^2$ of a particle caused by the force $\mathbf{f}$ is proportional to the particle's mass and points into the same direction. Generalized forces

Genetics and epigenetics: Genetic inheritance is dealing with transfer of biological information encoded on DNA molecules. In case of sexual reproduction it follows approximately the Mendelian rules.

Linearity: Linear functions are defined by a relation of the type $\mathbf{y}' = \mathrm{T} \cdot \mathbf{x}'$, or in components of the vectors $y_j = \sum_{i=1}^{n} t_{ji} x_i$, where T is called the transformation matrix.

Macroscopic level:

Mesoscopic level:

Microscopic level:

Neutrality Two genotypes are called *neutral* when selection is unable to distinguish between them.

Null hypothesis

Rank of a matrix: The rank of a matrix is the dimension of the row space.

Support

Notation

building blocks and degradation products	$\mathbf{A}, \mathbf{B}, \ldots,$
numbers of particles of $\mathbf{A}, \mathbf{B}, \ldots,$	$N_{\mathbf{A}}, N_{\mathbf{B}}, \ldots,$
concentrations of $\mathbf{A}, \mathbf{B}, \ldots,$	$[\mathbf{A}] = a, [\mathbf{B}] = b, \ldots,$
replicating molecular species	$\mathbf{I}_1, \mathbf{I}_2, \ldots,$
numbers of particles of $\mathbf{I}_1, \mathbf{I}_2, \ldots,$	$N_1, N_2, \ldots,$
concentrations of $\mathbf{I}_1, \mathbf{I}_2, \ldots,$	$[\mathbf{I}_1] = c_1, [\mathbf{I}_2] = c_2, \ldots,$
relative concentrations of $\mathbf{I}_1, \mathbf{I}_2, \ldots,$	$[\mathbf{I}_1] = x_1, [\mathbf{I}_2] = x_2, \ldots,$
partial sums of relative concentrations	$y_k = \sum_i x_i ,$
flow rate in the CSTR	$r ,$
influx concentration into the CSTR	$a_0 ,$
rate parameters	$d_i, k_i, f_i, \ldots \quad i = 1, 2, \ldots,$
global regulation flux	$\Phi(t) ,$
chain length of polynucleotides	$\nu ,$
superiority of the master sequence $\mathbf{I}_m$	$\sigma_m = \frac{f_m (1 - x_m)}{\sum_{i \neq m} f_i} ,$
population entropy	$S = \sum_i x_i \ln x_i .$

logical operators: $\forall$, $\rightarrow$, $\Longrightarrow$
scaling parameter: σ
support: 'supp', for example in $\bigcup_{\text{supp}}$
definition: ':='
vectors and matrices: transposition $^{\text{t}}$
linear span: 'span': $\text{span}(S) = \left\{\sum_{i=1}^{k} \lambda_i u_i | k \in \mathbb{N}, u_i \in S, \lambda_i \in \mathbf{K}\right\}$, S is a finite subset of a vector space $\mathcal{U}$ over a field $\mathbf{K}$.
concentration vectors: $\mathbf{x} = (a, b, \ldots) = ([\mathbf{A}], [\mathbf{B}], \ldots)$
reaction rate: $v\big(\mathbf{x}(t)\big)$, $\mathbf{v}\big(\mathbf{x}(t)\big)$

Index

Author Index