

## Chapter 2

# Random Variables: Fundamentals of Probability Theory and Statistics

A fundamental concept for any statistical treatment is that of the random variable. Thus this concept and various other closely related ideas are presented at the beginning of this book. Section 2.1 will introduce event spaces, probabilities, probability distributions, and density distributions within the framework of the Kolmogorov axioms. The concept of random variables will then be introduced, initially in a simplified form. In Sect. 2.2 these concepts will be extended to multidimensional probability densities and conditional probabilities. This will allow us to define independent random variables and to discuss the Bayes' theorem. Section 2.3 will deal with characteristic quantities of a probability density, namely, the expectation value, variance, and quantiles. Entropy is also a quantity characterising a probability density, and because of its significance a whole section, Sect. 2.4, is devoted to entropy. In particular, relative entropy, i.e., the entropy of one density relative to another, will be introduced and the maximum entropy principle will be discussed. In Sect. 2.5, the reader will meet the calculus of random variables; the central limit theorem in a first simple version is proven, stressing the importance of the normal random variable; various other important random variables are also presented here.

Because many questions in statistical mechanics can be reduced to the formulation of limit theorems for properly normalized sums of random variables with dependencies given by a model, the subsequent sections present further discussion of the concept of a limit distribution.

Section 2.6 introduces renormalization transformations and the class of stable distributions as fixed points of such transformations. Domains of attraction are discussed and the distributions in such domains are characterized by their expansion in terms of eigenfunctions, which themselves are obtained by a stability analysis.

Finally, Sect. 2.7 addresses the large deviation property for a sequence of random variables  $Y_N$ ,  $N = 2, \dots$  and it is shown how the characteristic features of the density of  $Y_N$  can then be revealed. It can already be seen that the shape of the density of  $Y_N$  may, as a function of an external parameter, become bimodal so that in the thermodynamic limit  $N \rightarrow \infty$  not only one but two equilibrium states exist. Thus the phenomenon of different phases and of a phase transition can already be demonstrated on this level.

## 2.1 Probability and Random Variables

During the course of history many people devoted much thought to the subject of probability (Schneider 1986). For a long time people sought in vain to define precisely what is meant by probability. In 1933 the Russian mathematician A. N. Kolmogorov formulated a complete system of axioms for a mathematical definition of probability. Today this system is the basis of probability theory and mathematical statistics.

We speak of the probability of events. This means that a number is assigned to each event and this number should represent the probability of this event. Let us consider throwing a die as an example. In this case each possible event is an outcome showing a certain number of points, and one would assign the number  $1/6$  to each event. This expresses the fact that each event is equally likely, as expected for a fair die, and that the sum of the probabilities is normalized to 1.

The Kolmogorov system of axioms now specifies the structure of the set of events and formulates the rules for the assignment of real numbers (probabilities) to these events.

### 2.1.1 The Space of Events

We consider a basic set  $\Omega$ , whose elements consist of all possible outcomes of an experiment, irrespective of whether this experiment can actually be performed or is only imaginable. A single performance of this experiment is called a realization. It yields an element  $\omega$  in  $\Omega$ .

We now want to identify events as certain subsets of  $\Omega$ . One may think of events as the sets  $\{\omega\}$ , which contain one single element  $\omega$ , and as such represent elementary events.

However, one may also think of other sets which contain several possible outcomes, because the probability that the outcome of a realization belongs to a certain set of outcomes might also be interesting.

A more detailed mathematical analysis reveals that in general not all subsets of  $\Omega$  can be considered as events to which one can assign a probability. Only for certain subsets, which can be made members of a so-called Borel space, can one always consistently introduce a probability. Here, a Borel space is a set  $\mathcal{B}$  of subsets of  $\Omega$ , for which:

- $\Omega \in \mathcal{B}$
- If  $A \in \mathcal{B}$  then  $\bar{A} \in \mathcal{B}$ , where  $\bar{A}$  is the complement of  $A$
- If  $A, B \in \mathcal{B}$  then  $A \cup B \in \mathcal{B}$   
and, more generally, the union of countably many (i.e., possibly infinitely many) sets in  $\mathcal{B}$  also belongs to  $\mathcal{B}$ .

For sets of a Borel space the axioms imply immediately that  $\emptyset \in \mathcal{B}$ . Furthermore, for  $A, B \in \mathcal{B}$  we also have  $A \cap B \in \mathcal{B}$ , since  $A \cap B = \overline{\overline{A} \cup \overline{B}}$ .

Of course, a Borel space should be defined in such a way that all possible outcomes of an experiment are really contained in this space.

*Remarks.*

- We should distinguish between experimental outcomes and events. Each experimental outcome  $\omega$  is an event  $\{\omega\}$ , but not every event is an experimental outcome, because an event which does not correspond to an elementary event contains many outcomes. We want to assign a consistent probability not only to experimental outcomes, but to all events.
- Frequently, a set of events  $\{A_1, \dots, A_N\}$  is given such that  $A_i \cap A_j = \emptyset$  for  $i \neq j$ , and

$$A_1 \cup \dots \cup A_N = \Omega. \quad (2.1)$$

Such a set is called complete and disjoint.

*Examples.*

- When we throw a die the outcomes are the elementary events  $\{i\}$ ,  $i = 1, \dots, 6$ . Further examples of events are  $\{1, 2\}$  (points smaller than 3) or  $\{1, 3, 5\}$  (points are odd). Hence, not only the probability for the elementary outcome  $\{1\}$  may be of interest, but also the probability that the points are odd.

We have

$$\{1\} \cup \dots \cup \{6\} = \Omega \equiv \{1, \dots, 6\}. \quad (2.2)$$

- The Borel space may contain all intervals  $\{x_1 \leq x \leq x_2\}$  on the real axis. It then also contains all points and all open intervals, as well as the event  $\{x \leq \lambda\}$  for  $\lambda \in \mathbb{R}$ .

Subsets not belonging to this Borel space can only be defined by complicated mathematical constructions (see e.g. Dudley 1989). We do not want to pursue this any further, because such sets are not relevant in physical considerations.

### 2.1.2 Introduction of Probability

Having dealt with the space of events we can now introduce the notion of probability. To each event  $A$  in the space of events  $\mathcal{B}$  we assign a real number  $\mathcal{P}(A)$ , the probability of  $A$ . This assignment has to satisfy the following properties:

- $\mathcal{P}(A) \geq 0$  for all  $A \in \mathcal{B}$ ,
- $\mathcal{P}(\Omega) = 1$ ,
- Let  $A_i$ ,  $i = 1, \dots$  be countably many (i.e. possibly infinitely many) disjoint sets in  $\mathcal{B}$  with  $A_i \cap A_j = \emptyset$  for  $i \neq j$ , then

$$\mathcal{P}(\cup A_i) = \sum_i \mathcal{P}(A_i). \quad (2.3)$$

These three conditions are certainly necessary for  $\mathcal{P}(A)$  to be a probability. It was the achievement of Kolmogorov to show that these three requirements allow a complete and consistent definition of such an assignment for all events.

Note that we have introduced probability without saying what it means, i.e., how to measure it. We have merely introduced an assignment which has all properties one would expect of the notion of probability. In Part II we will deal with the measurement of probabilities and of quantities which are calculated from probabilities.

*Remarks.*

- From  $\mathcal{P}(A) + \mathcal{P}(\bar{A}) = \mathcal{P}(\Omega) = 1$  we find  $\mathcal{P}(A) \leq 1$ .
- It can be shown:  $\mathcal{P}(A_1) \leq \mathcal{P}(A_2)$ , if  $A_1 \subseteq A_2$ .
- More general, the following can be shown:

$$\mathcal{P}(A_1 \cup A_2) = \mathcal{P}(A_1) + \mathcal{P}(A_2) - \mathcal{P}(A_1 \cap A_2). \quad (2.4)$$

*Examples.* For the throw of a die, for example, we take  $\mathcal{P}(\{i\}) = 1/6$  for  $i = 1, \dots, 6$ . Hence we have, e.g.,  $\mathcal{P}(\{1, 3\}) = 1/3$ ,  $\mathcal{P}(\{1, 3, 5\}) = 1/2$ ,  $\mathcal{P}(\{2, 4, 5, 6\}) = 2/3$ ,  $\mathcal{P}(\{1, 3\}) < \mathcal{P}(\{1, 3, 5\})$ .

Next we will consider the Borel space containing the intervals and points on the real axis. We introduce the probabilities of all events by defining the probabilities  $\mathcal{P}(\{x \leq \lambda\})$  for the sets  $\{x \leq \lambda\}$  for all  $\lambda$ . The function  $\mathcal{P}(\{x \leq \lambda\})$  will be denoted by  $P(\lambda)$  and is called the probability distribution.

This function satisfies:

$$\lambda \rightarrow +\infty : \quad P(\lambda) \rightarrow \mathcal{P}(\Omega) = 1, \quad (2.5a)$$

$$\lambda \rightarrow -\infty : \quad P(\lambda) \rightarrow \mathcal{P}(\emptyset) = 0. \quad (2.5b)$$

When  $P(\lambda)$  is differentiable we also consider

$$\varrho(\lambda) = \frac{dP(\lambda)}{d\lambda}, \quad (2.6)$$

from which we get

$$P(\lambda) = \int_{-\infty}^{\lambda} \varrho(x) dx. \quad (2.7)$$

Using  $\varrho(x)$  we may now calculate the probability for any interval  $\{x_1 \leq x \leq x_2\}$  and represent it as

$$\mathcal{P}(\{x_1 \leq x \leq x_2\}) = \int_{x_1}^{x_2} \varrho(x) dx. \quad (2.8)$$

In particular, we have

$$\int_{-\infty}^{+\infty} \varrho(x) dx = 1, \quad (2.9)$$

and also

$$\mathcal{P}(\{x\}) = 0. \quad (2.10)$$

If  $dx$  is small enough,  $\varrho(x) dx$  is the probability of the event  $(x, x + dx)$ . In other words it is the probability of obtaining a value  $x' \in (x, x + dx)$  as the result of a realization (i.e., the performance of an experiment). The function  $\varrho(x)$  is referred to as the density function or density distribution. In physics and certain other fields the name ‘distribution function’ is also frequently used. In the mathematical literature, however, the latter name is reserved for the function  $P(\lambda)$ .

### 2.1.3 Random Variables

A discrete random variable is a collection of possible elementary events together with their probabilities. A ‘realization’ of a random variable yields one of the elementary outcomes and it does so with the probability which has been assigned to this elementary event.

When we throw a die the random variable ‘number of spots’ is realized. The possible realizations are the numbers 1–6, each with probability  $1/6$ .

Hence, to characterize a random variable one has to list the possible elementary events (realizations) together with their probabilities. Each realization will yield an outcome which in general is different from the previous one. Where these outcomes are numbers, they are also called random numbers.

If the possible realizations (outcomes) do not form a discrete set but a continuum in  $\mathbb{R}$ , the collection cannot contain the probabilities of all elementary events, but instead the distribution function  $P(\lambda)$  or the density  $\varrho(x)$ . If we denote the possible outcomes (realizations) by  $x$ , the random variable will be denoted by  $X$ , the corresponding distribution function by  $P_X(\lambda)$ , and the density by  $\varrho_X(x)$ . Hence, the random variable  $X$  is defined by the set of its possible realizations together with the probability density  $\varrho_X(x)$  or the probability distribution  $P_X(\lambda)$ . We have

$$\mathcal{P}(\{x|x \leq \lambda\}) = P_X(\lambda) \quad (2.11)$$

and

$$\varrho_X(x) = \frac{dP_X(x)}{dx}. \quad (2.12)$$

One may also consider functions  $Y = f(X)$  of random variables.  $Y$  is the random variable with the realizations  $y = f(x)$  given the realization  $x$  of  $X$ , and the

distribution function is

$$P_Y(\lambda) = \mathcal{P}(\{x | f(x) \leq \lambda\}). \quad (2.13)$$

In Sect. 2.5 we will see how to perform calculations with random variables.

In the mathematical literature (see e.g. Kolmogoroff 1933; Feller 1957) the concept of a random variable is often introduced in a more general way. One usually proceeds from a general basic set  $\Omega$ . The events of the Borel space  $\mathcal{B}$  should be measurable, but they do not have to be intervals of the real axis. The random variables are then defined as mappings  $X(\omega)$  of the outcomes  $\omega$  onto the real axis, and the sets  $\{x | x \leq \lambda\}$  are replaced by the sets  $A_\lambda = \{\omega | X(\omega) \leq \lambda\}$ ,  $\lambda \in \mathbb{R}$ , to which a probability is assigned. For this to be consistent, one has to require  $A_\lambda \in \mathcal{B}$ ,  $\lambda \in \mathbb{R}$ .

Having physical applications in mind, we have defined the basic set  $\Omega$  to be the real axis and, consequently, were able to choose the mapping as the identity. In this way, the concept of a random variable is very simple. However, the generalization of this concept is straightforward.

Some important random variables are the following:

- (a) Let the set of the possible realizations of  $X$  be the real numbers in the interval  $(a, b)$  with uniform probability. Then

$$\varrho_X(x) = \frac{1}{b-a}. \quad (2.14)$$

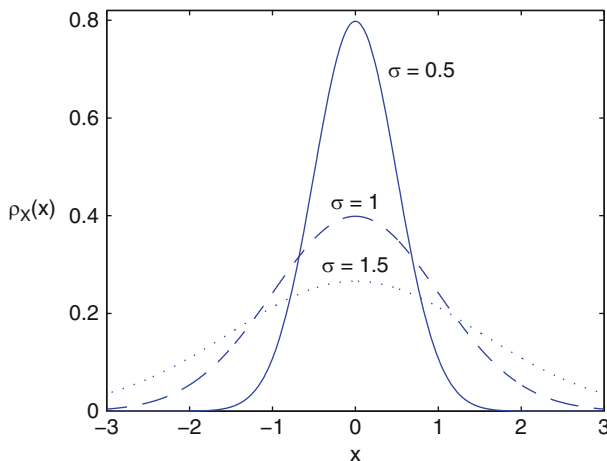
Almost every computer provides a more or less reliable random number generator which claims to yield uniformly and independently distributed random numbers on the interval  $(0, 1)$ .

- (b) The Gaussian or normal distribution: The set of possible realizations of  $X$  are all real numbers. The density is

$$\varrho_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (2.15)$$

Here  $\varrho_X(x)$  represents a normal curve with a maximum at  $x = \mu$  and a width characterized by  $\sigma$ . For larger values of  $\sigma$  the curve gets broader, but also flatter (see Fig. 2.1).  $\varrho_X(x)$  is the density function of the Gaussian distribution, which is also called the normal distribution  $N(\mu, \sigma^2)$ . For  $\mu = 0$  and  $\sigma = 1$  one also speaks of the standard normal distribution. Normal random variables (i.e., random variables with a normal distribution) will play an important role in subsequent discussions and we will frequently return to them.

- (c) The binomial distribution (also called the Bernoulli distribution): Let  $K$  be the discrete random variable with possible realizations  $k = 0, 1, \dots, n$  and the (discrete) probability density



**Fig. 2.1** Density function of the Gaussian distribution for various values of  $\sigma$

$$B(n, p; k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n. \quad (2.16)$$

Such a probability for a realization  $k$  occurs (naturally) in the following way: Let  $X$  be a random variable with the only two possible outcomes  $x_1$  (with probability  $p$ ) and  $x_2$  (with probability  $1-p$ ). Now consider an  $n$ -fold realization of  $X$ .  $K$  then represents the multiplicity of the occurrence of  $x_1$ . An example of this latter case is the following:

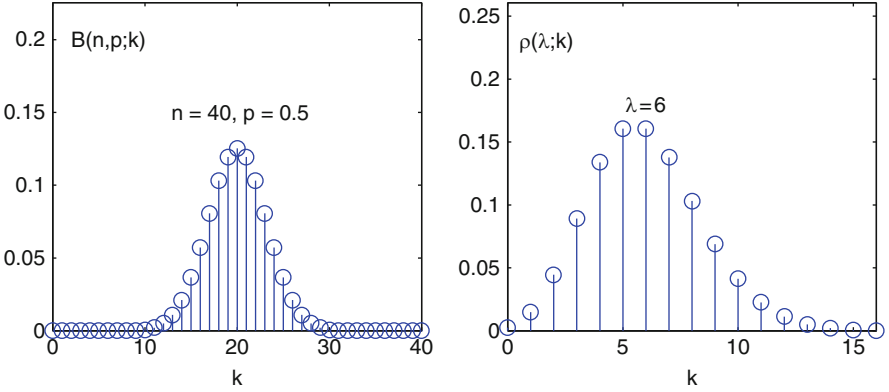
Consider two adjoining boxes of volumes  $V_1$  and  $V_2$ . Through a hole in a dividing wall particles of a gas can be exchanged. A particle will be in the volume  $V_1$  with probability  $p = V_1/(V_1+V_2)$  and in  $V_2$  with probability  $1-p = V_2/(V_1+V_2)$ . For a total of  $n$  particles we will find  $k$  particles in volume  $V_1$  with probability

$$B(n, p; k) = \binom{n}{k} \left( \frac{V_1}{V_1+V_2} \right)^k \left( \frac{V_2}{V_1+V_2} \right)^{n-k} \quad (2.17)$$

$$= \binom{n}{k} \left( \frac{V_1}{V_2} \right)^k \left( \frac{V_2}{V_1+V_2} \right)^n. \quad (2.18)$$

Of course, we should expect that  $B(n, p; k)$  has a maximum at  $k = np = n V_1/(V_1+V_2)$ . This will be confirmed later (see Fig. 2.2, left).

- (d) The Poisson distribution: Let  $K$  be the discrete random variable with possible realizations  $k = 0, 1, \dots$  and the discrete density



**Fig. 2.2** Densities of the binomial distribution (*left*) and the Poisson distribution (*right*)

$$\varrho(\lambda; k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots \quad (2.19)$$

The density  $\varrho(\lambda; k)$  results from  $B(n, p; k)$  in the limit  $p \rightarrow 0$ ,  $n \rightarrow \infty$  and  $pn = \lambda = \text{const}$ .  $K$  is equal to the number of events occurring within a time interval  $(0, T)$ , if the probability for the occurrence of one event within the time interval  $dt$  is just  $p = \lambda dt/T$ . To see this, we divide the time interval  $(0, T)$  into  $n$  equal segments of length  $dt = T/n$ . Then

$$p = \frac{\lambda}{T} \frac{T}{n} = \frac{\lambda}{n}, \quad (2.20)$$

and the probability that in  $k$  of these  $n$  segments one event occurs is then just given by  $B(n, p; k)$  of (2.16). For  $n \rightarrow \infty$  one takes  $dt \rightarrow 0$  and  $p \rightarrow 0$  such that  $pn = \lambda$  remains constant. The density function is shown in Fig. 2.2(right). It will turn out in Sect. 2.3 that  $\lambda$  is the average number of events in the time interval  $(0, T)$ .

Radioactive decay provides a physical example of the Poisson distribution. We consider a radioactive element with a radiation activity of  $\alpha$  Becquerel, i.e., on average  $\alpha$  decays occur within 1 s. Then we have  $\lambda/T = \alpha \text{ s}^{-1}$ . The probability that no decay occurs within a time interval of  $T$  seconds is

$$\varrho(\lambda = \alpha T; k = 0) = e^{-\alpha T}, \quad (2.21)$$

and the probability of just one decay within the time interval  $T$  is

$$\varrho(\lambda = \alpha T; k = 1) = \alpha T e^{-\alpha T}. \quad (2.22)$$



## 2.2 Multivariate Random Variables and Conditional Probabilities

### 2.2.1 Multidimensional Random Variables

In analogy to the definition of random variables one can introduce  $d$ -dimensional random vectors  $\mathbf{X} = (X_1, \dots, X_n)$  as  $n$  component random variables. In this case the basic space of possible outcomes is  $\mathbb{R}^n$ , and events are, among other things, domains in  $\mathbb{R}^n$  as cartesian products of events belonging to the Borel spaces of the components. The probability density is now a function  $\varrho(x_1, \dots, x_n)$  on  $\mathbb{R}^n$ . The probability that a realization of  $X_1$  yields a value in the interval  $(x_1, x_1 + dx_1)$ , and, similarly, that realizations of  $X_2, \dots, X_n$  yield values in the corresponding intervals is  $\varrho(x_1, \dots, x_n)dx_1 \dots dx_n$ . Two examples will help to clarify this:

In a classical description, the momentum  $\mathbf{p}$  of a particle with mass  $m$  in a gas may be considered as a (three-dimensional) random variable. For the density distribution one obtains (see (2.29))

$$\varrho(\mathbf{p}) = \frac{1}{A} \exp\left(-\beta \frac{\mathbf{p}^2}{2m}\right), \quad \text{where} \quad \beta = \frac{1}{k_B T}. \quad (2.23)$$

Here  $T$  represents the temperature,  $k_B$  is Boltzmann's constant, and  $A$  is a normalization constant. The density  $\varrho(\mathbf{p})$  in (2.23) is thus given by a three dimensional Gaussian distribution. The general  $n$ -dimensional (or 'multivariate') Gaussian distribution reads

$$\varrho(\boldsymbol{\mu}, \mathbf{A}; \mathbf{x}) = \frac{(2\pi)^{-n/2}}{(\det \mathbf{A})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})_i (\mathbf{A}^{-1})_{ij} (\mathbf{x} - \boldsymbol{\mu})_j\right). \quad (2.24)$$

Here we have used the summation convention, i.e. one has to sum over all indices appearing twice. The vector  $\boldsymbol{\mu}$  and the matrix  $\mathbf{A}$  are parameters of this distribution.

As our second example of a multivariate distribution we consider a gas of  $N$  classical particles characterized by the momenta and positions of all particles:

$$(\mathbf{p}, \mathbf{q}) = (\mathbf{p}_1, \dots, \mathbf{p}_N, \mathbf{q}_1, \dots, \mathbf{q}_N). \quad (2.25)$$

We will describe this state at each instant by a  $6N$ -dimensional random vector. When the volume and the temperature of the gas are specified, one obtains for the probability density in statistical mechanics (see (3.49))

$$\varrho(\mathbf{p}, \mathbf{q}) = \frac{1}{A} e^{-\beta H(\mathbf{p}, \mathbf{q})}, \quad \beta = \frac{1}{k_B T}, \quad (2.26)$$

where again  $T$  denotes the temperature,  $k_B$  Boltzmann's constant, and  $A$  a normalization constant. Furthermore,  $H(\mathbf{p}, \mathbf{q})$  is the Hamiltonian function for the particles in the volume  $V$ . This density is also called the Boltzmann distribution.

### 2.2.2 Marginal Densities

When one integrates over some of the variables of a multidimensional probability density, one obtains a probability density describing the probability for the remaining variables, *irrespective* of the values for those variables which have been integrated over. Let, for instance,

$$\varrho'(x_1) = \int dx_2 \dots dx_n \varrho(x_1, x_2, \dots, x_n), \quad (2.27)$$

then  $\varrho'(x_1) dx_1$  is the probability of finding  $X_1$  in the interval  $[x_1, x_1 + dx_1]$ , irrespective of the outcome for the variables  $X_2, \dots, X_n$ .

This may be illustrated for the case of the Boltzmann distribution (2.26). With the Hamiltonian function

$$H = \sum_{i=1}^N \frac{\mathbf{p}_i^2}{2m} + V(\mathbf{q}_1, \dots, \mathbf{q}_N), \quad (2.28)$$

one obtains, after taking the integral over  $\mathbf{p}_2, \dots, \mathbf{p}_N, \mathbf{q}_1, \dots, \mathbf{q}_N$ , the probability density (2.23) for a single particle, in this case in the form

$$\varrho'(\mathbf{p}_1) = \frac{1}{A'} \exp\left(-\beta \frac{\mathbf{p}_1^2}{2m}\right). \quad (2.29)$$

### 2.2.3 Conditional Probabilities and Bayes' Theorem

With the Boltzmann distribution (2.26) we have already met a distribution where certain given parameters need to be included explicitly, for instance, the temperature  $T$  and the volume  $V$ . The number of particles  $N$  may also be counted among these given parameters. In probability theory one writes  $A \mid B$  for an event  $A$  under the condition that  $B$  is given. So the probability  $\mathcal{P}(A)$  is then more precisely denoted by  $\mathcal{P}(A \mid B)$ , i.e., the probability of  $A$  when  $B$  is given.  $\mathcal{P}(A \mid B)$  is called the conditional probability.

This notion extends to the probability densities. The Boltzmann distribution can therefore be written as

$$\varrho(\mathbf{p}, \mathbf{q} \mid T, V, N), \quad (2.30)$$

or in words, the probability density for the positions and momenta at given temperature, volume, and number of particles. In the same way,

$$\varrho(p_x \mid p_y, p_z)$$

is the probability density for the  $x$ -component of the momentum of a particle under the condition that the  $y$ - and  $z$ -components are given.

One may form

$$\mathcal{P}(A, B) = \mathcal{P}(A \mid B)\mathcal{P}(B), \quad (2.31)$$

which is the joint probability for the occurrence of  $A$  and  $B$ . If  $B$  is an event in the same Borel space as  $A$ , then the joint probability  $\mathcal{P}(A, B)$  is equivalent to  $\mathcal{P}(A \cap B)$ .

One may also define the conditional probability by

$$\mathcal{P}(A \mid B) = \frac{\mathcal{P}(A, B)}{\mathcal{P}(B)}. \quad (2.32)$$

If the denominator  $\mathcal{P}(B)$  vanishes, it is also not meaningful to consider the conditional probability  $\mathcal{P}(A \mid B)$ .

Similarly, conditional densities might also be introduced by using the multivariate probability densities, e.g.,

$$\varrho(p_x \mid p_y, p_z) = \frac{\varrho(p_x, p_y, p_z)}{\varrho(p_y, p_z)}. \quad (2.33)$$

Example: Consider a fair die and  $B = \{2, 4, 6\}$ ,  $A = \{2\}$ . (Here  $A$  and  $B$  belong to the same Borel space.) Then

$$\mathcal{P}(A \mid B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)} = \frac{\mathcal{P}(A)}{\mathcal{P}(B)} = \frac{1/6}{1/2} = \frac{1}{3}. \quad (2.34)$$

The probability for the event  $\{2\}$ , given the number of points is even, is  $1/3$ . Obviously also  $\mathcal{P}(B \mid A) = 1$ .

We note that if

$$\bigcup_{i=1}^N B_i = \Omega \quad (2.35)$$

is a disjoint, complete partition of  $\Omega$  (such that  $B_i \cap B_j = \emptyset$  and the union of all  $B_i$  is equal to the total set  $\Omega$ ), then obviously

$$\mathcal{P}(A) = \mathcal{P}(A, \Omega) = \sum_{i=1}^N \mathcal{P}(A, B_i) = \sum_{i=1}^N \mathcal{P}(A \mid B_i)\mathcal{P}(B_i). \quad (2.36)$$

This should be compared with the formula

$$\varrho_{X_1}(x_1) = \int dx_2 \varrho_{X_1, X_2}(x_1, x_2) = \int dx_2 \varrho(x_1 | x_2) \varrho_{X_2}(x_2). \quad (2.37)$$

In the remainder of this section we describe two useful statements about conditional probabilities.

### Independence of Random Variables

Let  $A_1$  and  $A_2$  be two events (in the same, or possibly in different Borel spaces).  $A_1$  is said to be independent of  $A_2$  if the probability for the occurrence of  $A_1$  is independent of  $A_2$ , i.e.,

$$\mathcal{P}(A_1 | A_2) = \mathcal{P}(A_1). \quad (2.38)$$

In particular we have then also:

$$\mathcal{P}(A_1, A_2) = \mathcal{P}(A_1)\mathcal{P}(A_2). \quad (2.39)$$

If  $A_1$  is independent of  $A_2$ , then  $A_2$  is also independent of  $A_1$ : Statistical (in)dependence is always mutual.

Similarly, the joint density of two independent random variables may be written as

$$\varrho_{X_1, X_2}(x_1, x_2) = \varrho_{X_1}(x_1) \varrho_{X_2}(x_2). \quad (2.40)$$

### Bayes' Theorem

From

$$\mathcal{P}(A, B) = \mathcal{P}(A | B)\mathcal{P}(B) = \mathcal{P}(B | A)\mathcal{P}(A) \quad (2.41)$$

it follows that

$$\mathcal{P}(B | A) = \frac{\mathcal{P}(A | B)\mathcal{P}(B)}{\mathcal{P}(A)}. \quad (2.42)$$

Hence  $\mathcal{P}(B | A)$  can be determined from  $\mathcal{P}(A | B)$ , if the a priori probabilities  $\mathcal{P}(A)$  and  $\mathcal{P}(B)$  are known.

This statement was first formulated by the English Presbyterian and mathematician Thomas Bayes (1702–1761) in an essay that was found after his death.

Bayes' theorem is a very useful relation for determining the a posteriori probabilities  $\mathcal{P}(B | A)$ . It has enormous numbers of applications, of which we merely give two examples here.

- (a) A company which produces chips owns two factories: Factory A produces 60% of the chips, factory B 40%. So, if we choose at random one chip from the company, this chip originates from factory A with a probability of 60%. We further suppose that 35% of the chips coming from factory A are defective, but only 25% of those coming from factory B.

Using Bayes' theorem one can determine the probability that a given defective chip comes from factory A. Let  $d$  be the event 'the chip is defective',  $A$  the event 'the chip comes from factory A', and  $B(= \bar{A})$  the event 'the chip comes from factory B'. From Bayes' theorem we then have

$$\mathcal{P}(A|d) = \frac{\mathcal{P}(d|A) \cdot \mathcal{P}(A)}{\mathcal{P}(d)} = \frac{\mathcal{P}(d|A) \cdot \mathcal{P}(A)}{\mathcal{P}(d|A) \cdot \mathcal{P}(A) + \mathcal{P}(d|B) \cdot \mathcal{P}(B)}. \quad (2.43)$$

Inserting the numbers  $\mathcal{P}(A) = 0.60$ ,  $\mathcal{P}(d|A) = 0.35$ ,  $\mathcal{P}(d|B) = 0.25$  yields a value of  $\mathcal{P}(A|d) = 0.68$ .

In the same manner we can determine the probability of having a certain illness when the test for this illness showed positive. Luckily enough, this is not as large as one might first expect. Let  $A$  be the event 'the illness is present' and  $B$  the event 'the test is positive'. The conditional probabilities  $\mathcal{P}(B|A)$  and  $\mathcal{P}(B|\bar{A})$  yield the probabilities that a test has been positive for a sick patient and a healthy patient, respectively.  $\mathcal{P}(B|A)$  is called the sensitivity,  $\mathcal{P}(\bar{B}|\bar{A}) = 1 - \mathcal{P}(B|\bar{A})$  the specificity of the test.

The probability  $\mathcal{P}(A)$  that the illness is present at all is in general of the order of magnitude 0.01 – 0.001. From this a surprisingly small value for the probability of being ill may result even if the test has been positive. In numbers: If we set e.g.  $\mathcal{P}(B|A) = 0.95$  and  $\mathcal{P}(B|\bar{A}) = 0.01$  one obtains

$$\mathcal{P}(A|B) \equiv \mathcal{P}(\text{patient is ill} | \text{test is positive}) \quad (2.44a)$$

$$= \frac{1}{1 + 0.0105 \frac{(1 - \mathcal{P}(A))}{\mathcal{P}(A)}} \quad (2.44b)$$

$$= \begin{cases} \frac{1}{1 + 10.5} \approx 0.087 & \text{for } \mathcal{P}(A) = \frac{1}{1000} \\ \frac{1}{1 + 1.04} \approx 0.490 & \text{for } \mathcal{P}(A) = \frac{1}{100}. \end{cases} \quad (2.44c)$$

Hence, the results depend sensitively on the probability  $\mathcal{P}(A)$ , i.e., on the overall frequency of the illness and on  $\mathcal{P}(B | \bar{A})$ , the probability that the test is positive for a healthy patient.

- (b) An amusing application of Bayes' theorem is in solving the following problem, which we cite from von Randow (1992):

You are taking part in a TV game show where you are requested to choose one of three closed doors. Behind one door a prize is waiting for you, a car, behind the other two doors are goats. You point at one door, say, number one. For the time being it remains closed. The showmaster knows which door conceals the car. With the words 'I'll show you something' he opens one of the other doors, say, number three, and a bleating goat is looking at the audience. He asks: 'Do you want to stick to number one or will you choose number two?'.

The correct answer is: It is more favorable to choose number 2, because the probability that the car is behind this door is  $2/3$ , whereas it is only  $1/3$  for door number 1.

Intuitively, most people guess that the probabilities for both remaining possibilities are equal and that the candidate has no reason to change his mind. In the above mentioned booklet one can read about the turmoil which was caused by this problem and its at first sight surprising solution after its publication in America and Germany.

We define (von Randow 1992):

$A1$ : the event that the car is behind door number 1, and similarly  $A2$  and  $A3$ .

$M1$ : the event that the showmaster opens door number 1 with a goat behind; similarly  $M2$  and  $M3$ .

As we have denoted the door the showmaster has opened as number 3 we are interested in

$$\mathcal{P}(A1 \mid M3) \quad \text{and} \quad \mathcal{P}(A2 \mid M3). \quad (2.45)$$

Are these probabilities equal or different?

First argument (not using Bayes' theorem):

$\mathcal{P}(A1 \mid M3)$  is independent of  $M3$ , because the showmaster acts according to the rule: Open one of the doors that the candidate has not chosen and behind which is a goat.

So

$$\mathcal{P}(A1 \mid M3) = \mathcal{P}(A1) = 1/3, \quad (2.46)$$

from which follows  $\mathcal{P}(A2 \mid M3) = 2/3$ .

Second argument (using Bayes' theorem):

We have

$$\mathcal{P}(A2 \mid M3) = \frac{\mathcal{P}(M3 \mid A2)\mathcal{P}(A2)}{\mathcal{P}(M3)}. \quad (2.47)$$

Now

$$\mathcal{P}(A2) = 1/3, \quad (2.48a)$$

$$\mathcal{P}(M3 | A2) = 1, \quad (2.48b)$$

$$\mathcal{P}(M3 | A1) = 1/2, \quad (2.48c)$$

$$\begin{aligned} \mathcal{P}(M3) &= \mathcal{P}(M3 | A1)\mathcal{P}(A1) + \mathcal{P}(M3 | A2)\mathcal{P}(A2) \\ &= 1/2 \cdot 1/3 + 1 \cdot 1/3 = 1/2, \end{aligned} \quad (2.48d)$$

and therefore

$$\mathcal{P}(A2 | M3) = 2/3. \quad (2.49)$$

Similarly, one obtains:

$$\mathcal{P}(A1 | M3) = 1/3. \quad (2.50)$$

One can simulate the game on a computer and will find that after  $N$  runs in approximately  $2N/3$  cases the goat is behind door 2, so that a change of the chosen door is indeed favorable.

## 2.3 Moments and Quantiles

### 2.3.1 Moments

Let  $\mathbf{X}$  be a random vector in  $\mathbb{R}^n$  with the distribution  $\varrho(\mathbf{x})$ . The expectation value of a function  $H(\mathbf{X})$  of the random variable  $\mathbf{X}$  is then defined as

$$\langle H(\mathbf{X}) \rangle = \int d^n x H(\mathbf{x}) \varrho(\mathbf{x}). \quad (2.51)$$

In the mathematical literature, the expectation value of  $H(\mathbf{X})$  is also written as  $E(H(\mathbf{X}))$ .

A particularly important moment is

$$\boldsymbol{\mu} \equiv E(\mathbf{X}) \equiv \langle \mathbf{X} \rangle = \int d^n x \mathbf{x} \varrho(\mathbf{x}), \quad (2.52)$$

which is the expectation value of the random variable itself. (Any possible outcome is multiplied by its probability, i.e., one forms  $\mathbf{x} \varrho(\mathbf{x}) d^n x$ , and then the sum is taken over all possible outcomes).

Some other important properties of moments and relations involving them are the following:

- When  $H(\mathbf{x})$  is a monomial of degree  $m$ , the expectation value is also called the  $m$ th moment of the distribution function. Hence, the  $m$ th moment for a scalar random variable is simply

$$\langle X^m \rangle = \int dx x^m \varrho(x). \quad (2.53)$$

- An important combination of moments is the variance. For a scalar random variable it is given by:

$$\text{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2 \quad (2.54a)$$

$$= \int dx (x - \langle X \rangle)^2 \varrho(x). \quad (2.54b)$$

Hence, the variance is the expectation value of the squared deviation of the random variable  $X$  from the expectation value  $\langle X \rangle$ . Therefore, the more scattered the realizations of  $X$  are around  $\langle X \rangle$ , the larger the variance is.  $\sqrt{\text{Var}(X)}$  is also called the standard deviation.

For the one-dimensional Gaussian distribution  $\varrho(\mu, \sigma^2; x)$  given in (2.15) one obtains

$$\langle X \rangle = \mu, \quad (2.55)$$

and

$$\text{Var}(X) = \int_{-\infty}^{\infty} dx (x - \mu)^2 \varrho(\mu, \sigma^2; x) = \sigma^2, \quad (2.56)$$

i.e., the parameter  $\sigma^2$  in (2.15) is identical to the variance of the normal distribution.

The higher moments of the normal distribution are easily calculated:

$$\langle (X - \mu)^k \rangle = \begin{cases} 0, & \text{if } k \text{ is odd,} \\ 1 \cdot 3 \cdot \dots \cdot (k-1) \cdot \sigma^k, & \text{if } k \text{ is even.} \end{cases} \quad (2.57)$$

- For a multivariate distribution we may define second moments with respect to different components, for example,

$$\langle X_i X_j \rangle = \int d^n x x_i x_j \varrho(x_1, \dots, x_n). \quad (2.58)$$

In analogy to the variance we now define a covariance matrix:

$$\text{Cov}(X_i, X_j) \equiv \sigma_{ij}^2 \equiv \langle (X - \boldsymbol{\mu})_i (X - \boldsymbol{\mu})_j \rangle \quad (2.59a)$$

$$= \int d^n x (x_i - \mu_i) (x_j - \mu_j) \varrho(x_1, \dots, x_n). \quad (2.59b)$$



For the multivariate normal distribution given in (2.24) we get

$$\langle \mathbf{X} \rangle = \boldsymbol{\mu}, \quad (2.60)$$

and

$$\langle (X - \mu)_i (X - \mu)_j \rangle = A_{ij}. \quad (2.61)$$

Hence, the matrix  $\mathbf{A}$  in the expression of the multivariate normal distribution given in (2.24) is the covariance matrix for the normal distribution.

- The correlation between two random variables  $X_i, X_j$  is obtained from the covariance by normalization:

$$\text{Cor}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} = \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}}. \quad (2.62)$$

When  $X_i$  and  $X_j$  are mutually independent, one obtains immediately

$$\text{Cor}(X_i, X_j) = \text{Cov}(X_i, X_j) \equiv 0. \quad (2.63)$$

On the other hand, if the correlation or the covariance of two random variables vanishes one can in general not conclude that they are statistically independent.

Only when  $X_i$  and  $X_j$  are both normally distributed, is this conclusion correct, because in this case the covariance matrix, and hence also the matrix  $\mathbf{A}$  in (2.24), is diagonal, and the total density function is the product of the density functions of the individual random variables.

- An important expectation value for a probability density is

$$G(k) \equiv \langle e^{ikX} \rangle = \int dx e^{ikx} \varrho_X(x), \quad (2.64)$$

which is called the characteristic function.  $G(k)$  is thus the Fourier transform of the density function. When all moments exist,  $G(k)$  can be expanded in a power series and the coefficients contain the higher moments:

$$G(k) = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \langle X^n \rangle. \quad (2.65)$$

The expansion of  $\ln G(k)$  with respect to  $k$  yields a power series, in which the so-called cumulants  $\kappa_n$  appear:

$$\ln G(k) = \sum_{n=1}^{\infty} \frac{(ik)^n}{n!} \kappa_n, \quad (2.66)$$

with the cumulants

$$\kappa_1 = \mu = \langle X \rangle \quad (2.67a)$$

$$\kappa_2 = \text{Var}(X) = \langle X^2 \rangle - \langle X \rangle^2 \quad (2.67b)$$

$$\kappa_3 = \langle X^3 \rangle - 3\langle X^2 \rangle \langle X \rangle + 2\langle X \rangle^3, \quad (2.67c)$$

and so on for higher cumulants.

It is now important to note that the Fourier transform of the Gaussian or normal distribution is

$$G(k) = \exp\left(i\mu k - \frac{1}{2}\sigma^2 k^2\right), \quad (2.68)$$

i.e., one obtains for the Gaussian distribution

$$\kappa_1 = \mu \quad (2.69a)$$

$$\kappa_2 = \sigma^2, \quad (2.69b)$$

and thus for a normal distribution all higher cumulants vanish!

- Individual moments, in particular the expectation value, need not be adequate characteristics of a distribution. For a distribution with two maxima, symmetric around  $x = 0$ , the expectation value is  $\mu = 0$ , although  $x = 0$  may never or seldom be assumed. (see Fig. 2.3). Similarly, for a broad or skew distribution the expectation value, for instance, is not a conclusive quantity.
- The moments may not always be finite. The Lorentz or Cauchy distribution (also called Breit–Wigner distribution),

$$\varrho(x) = \frac{1}{\pi} \frac{\gamma}{(x-a)^2 + \gamma^2}, \quad -\infty < x < \infty \quad (2.70)$$

decays so slowly at infinity that all moments diverge. This may also be seen from the characteristic function, for which one obtains

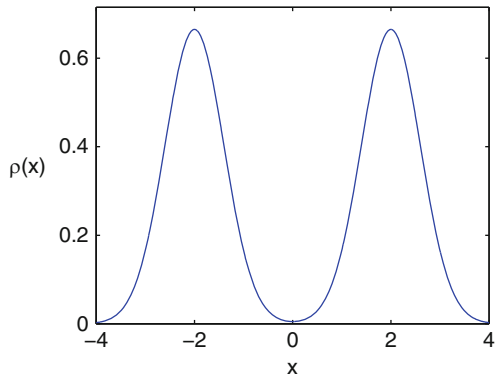
$$G(k) = e^{ika - |k|\gamma}. \quad (2.71)$$

This function has no power series expansion around  $k = 0$ .

- For distributions other than probability distributions moments can also be used for a global characterization. For instance, for a distribution of charges  $\varrho(\mathbf{r})$  we know the electric dipole moment

$$\mathbf{p} = \int d^3r \, \mathbf{r} \varrho(\mathbf{r}), \quad (2.72)$$

**Fig. 2.3** A density with two maxima (also called a bimodal distribution). The expectation value is zero, but  $x = 0$  is never assumed



and for a distribution of mass  $m(\mathbf{r})$  the moments of inertia

$$I_{ij} = \int d^3r m(\mathbf{r}) (-r_i r_j + \delta_{ij} r^2). \quad (2.73)$$

- For discrete probability distributions the moments are defined in an analogous way. For instance, for the Poisson distribution  $p(\lambda; k)$ ,

$$\langle K \rangle = \sum_{k=1}^{\infty} k p(\lambda; k) = \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \quad (2.74)$$

and

$$\langle K^2 \rangle = \sum_{k=1}^{\infty} k^2 p(\lambda; k) = \lambda^2 + \lambda. \quad (2.75)$$

For a Poisson random variable the variance is therefore equal to the mean:

$$\text{Var}(K) = \lambda. \quad (2.76)$$

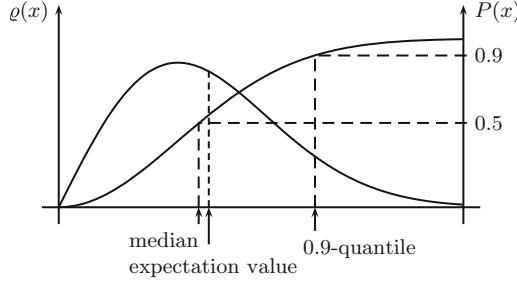
For the binomial distribution one obtains

$$\langle K \rangle = n p, \quad (2.77a)$$

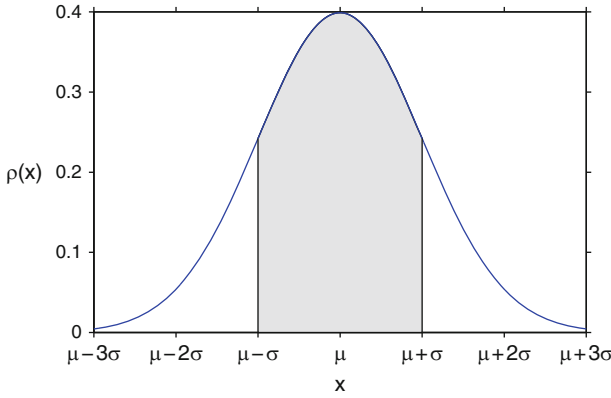
$$\text{Var}(K) = n p (1 - p). \quad (2.77b)$$

### 2.3.2 Quantiles

The  $\alpha$ -quantile for a probability distribution of a scalar random variable  $X$  is defined as the value  $x_\alpha$  for which



**Fig. 2.4** Median, expectation value and 0.9-quantile of a probability distribution  $P(x)$ , also shown for the density  $\varrho(x)$



**Fig. 2.5** The area of the shaded region is 0.6827 times the total area under the curve, which is 1. The probability that the random variable  $X$  assumes a value in the interval  $[\mu - \sigma, \mu + \sigma]$  is therefore 0.6827 or 68.27%

$$P(x_\alpha) = \int_{-\infty}^{x_\alpha} dx \varrho_X(x) = \alpha. \quad (2.78)$$

The probability of a realization yielding a value in the interval  $[-\infty, x_\alpha]$  is then  $\alpha$  or 100  $\alpha\%$ , the probability for a value  $x > x_\alpha$  is equal to  $(1 - \alpha)$  or  $(1 - \alpha)$  100%. The 1/2-quantile is also called the median (see Fig. 2.4).

The quantiles of the standard normal distribution can be found in a table for the distribution function or a table for the quantiles themselves. For instance  $x_{0.5} = 0$ ,  $x_{0.8413} = 1$ ,  $x_{0.9772} = 2$ . In this case symmetry reasons imply that  $x_{1-\alpha} = -x_\alpha$ , i.e., we also have  $x_{0.1587} = -1$ ,  $x_{0.0228} = -2$ . The interval  $(-1, 1)$  contains 84.13% – 15.87% = 68.27%, the interval  $(-2, 2)$  95.45% of the values (see Fig. 2.5). For a general normal distribution  $N(\mu, \sigma^2)$  one finds the following values:

Interval	Percentage of values	
$(\mu - \sigma, \mu + \sigma)$	68.27%	(2.79)
$(\mu - 2\sigma, \mu + 2\sigma)$	95.45%	
$(\mu - 3\sigma, \mu + 3\sigma)$	99.73%	

## 2.4 The Entropy

An important characteristic feature for a random variable is the entropy, which we will introduce in this section.

### 2.4.1 Entropy for a Discrete Set of Events

Let  $\{A_1, \dots, A_N\}$  be a complete, disjoint set of events, i.e.,

$$A_1 \cup A_2 \cup \dots \cup A_N = \Omega. \quad (2.80)$$

Furthermore, let  $\mathcal{P}$  be a probability defined for these events. We then define the entropy as

$$S = -k \sum_{i=1}^N \mathcal{P}(A_i) \ln(\mathcal{P}(A_i)). \quad (2.81)$$

Here  $k$  represents a factor which we set equal to 1 for the moment. In the framework of statistical mechanics  $k$  will be Boltzmann's constant  $k_B$ .

We observe:

- The entropy is defined for a complete, disjoint set of events of a random variable, irrespective of whether this partition of  $\Omega$  into events can be refined or not. If  $\Omega$  is the real axis, we might have, e.g.,  $N = 2$ ,  $A_1 = (-\infty, 0)$ ,  $A_2 = [0, \infty)$ .
- Since  $0 \leq \mathcal{P}(A_i) \leq 1$  we always have  $S \geq 0$ .
- If  $\mathcal{P}(A_j) = 1$  for a certain  $j$  and  $\mathcal{P}(A_i) = 0$  otherwise, then  $S = 0$ . This means that if the event  $A_j$  occurs with certainty the entropy is zero.
- If an event has occurred, then, as we will show in a moment,  $-\log_2 \mathcal{P}(A_j)$  is a good measure of the number of questions to be asked in order to find out that it is just  $A_j$  which is realized. In this context, 'question' refers to questions which can be answered by 'yes' or 'no', i.e., the answer leads to a gain of information of 1 bit. Hence, on average the required number of yes-or-no questions is

$$S' = - \sum_{j=1}^N \mathcal{P}(A_j) \log_2(\mathcal{P}(A_j)) = S + \text{const.} \quad (2.82)$$

The entropy is thus a measure of the missing information needed to find out which result is realized.

To show that  $-\log_2 \mathcal{P}(A_j)$  is just equal to the number of required yes-or-no questions, we first divide  $\Omega$  into two disjoint domains  $\Omega_1$  and  $\Omega_2$  such that

$$\sum_{A_i \in \Omega_1} \mathcal{P}(A_i) = \sum_{A_i \in \Omega_2} \mathcal{P}(A_i) = \frac{1}{2}. \quad (2.83)$$

The first question is now: Is  $A_j$  in  $\Omega_1$ ? Having the answer to this question we next consider the set containing  $A_j$  and multiply the probabilities for the events in this set by a factor of 2. The sum of the probabilities for this set is now again equal to 1, and we are in the same position as before with the set  $\Omega$ : We divide it again and ask the corresponding yes-or-no question. This procedure ends after  $k$  steps, where  $k$  is the smallest integer such that  $2^k \mathcal{P}(A_j)$  becomes equal to or larger than 1. Consequently,  $-\log_2 \mathcal{P}(A_j)$  is a good measure of the number of yes-or-no questions needed.

- If the probabilities of the events are equal, i.e.,

$$\mathcal{P}(A_i) = \frac{1}{N}, \quad (2.84)$$

we have

$$S = \ln N. \quad (2.85)$$

Any other distribution of probabilities leads to a smaller  $S$ . This will be shown soon.

The above observations suggest that the entropy may be considered as a lack of information when a probability density is given. On average it would require the answers to  $S$  yes-or-no questions to figure out which event has occurred. This lack is zero for a density which describes the situation where one event occurs with certainty. If all events are equally probable, this lack of information about which event will occur in a realization is maximal.

A less subjective interpretation of entropy arises when we think of it as a measure for uncertainty. If the probability is the same for all events, the uncertainty is maximal.

### 2.4.2 Entropy for a Continuous Space of Events

In a similar manner we define the entropy for a random variable  $X$ , where the space of events is a continuum, by

$$S[\varrho_X] = -k \int dx \varrho_X(x) \ln \left( \frac{\varrho_X(x)}{\varrho_0} \right). \quad (2.86)$$

When  $\varrho_X(x)$  has a physical dimension, the denominator  $\varrho_0$  in the argument of the logarithm cannot simply be set to 1. Since the physical dimension of  $\varrho_X(x)$  is equal to the dimension of  $1/dx$ , the physical dimension of  $\varrho_0$  has to be the same, in order that the argument of the logarithm will be dimensionless.

It is easy to see that a change of  $\varrho_0$  by a factor  $\alpha$  leads to a change of the entropy by an additive term  $k \ln \alpha$ . Such a change of  $\varrho_0$  only shifts the scale of  $S$ . Notice that we no longer have  $S \geq 0$ .

We now calculate the entropy for a Gaussian random variable  $N(\mu, \sigma^2)$ . We obtain (for  $k = 1, \varrho_0 = 1$ ):

$$S = \int dx \left( \frac{(x - \mu)^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \right) \varrho_X(x) \quad (2.87)$$

$$= \frac{1}{2} (1 + \ln(2\pi\sigma^2)). \quad (2.88)$$

The entropy increases with the width  $\sigma^2$  of the probability density, i.e., with the spreading around the expectation value. In this case we again find that the broader the distribution, the larger our ignorance about which event will occur in a realization, and the larger the entropy. Again, entropy means a lack of information or uncertainty.

### 2.4.3 Relative Entropy

The relative entropy of a density function  $p(x)$  with respect to a second density function  $q(x)$  is defined by

$$S[p|q] = -k \int dx p(x) \ln \left( \frac{p(x)}{q(x)} \right). \quad (2.89)$$

Obviously,  $p(x) \equiv q(x)$  if and only if  $S[p|q] = 0$ . However, while the entropy for a complete and disjoint set of events is positive semi-definite, the relative entropy of a density function  $p(x)$  with respect to a given density function  $q(x)$  is negative semi-definite, i.e.,

$$S[p|q] \leq 0. \quad (2.90)$$

This is easy to see: We use the inequality

$$\ln z \leq z - 1 \quad (2.91)$$

for  $z = \frac{q(x)}{p(x)}$ , multiply by  $p(x)$ , integrate over  $x$ , and obtain

$$-\int dx p(x) \ln \left( \frac{p(x)}{q(x)} \right) \leq \int dx (q(x) - p(x)). \quad (2.92)$$

Since both densities are normalized, the integrals on the right-hand side are equal, from which (2.90) follows.

#### 2.4.4 Remarks

The notion of entropy was first introduced into thermodynamics as an extensive quantity, conjugate to temperature. The revealing discovery of the connection between this quantity and the probability of microstates was one of the great achievements of L. Boltzmann, and the equation  $S = k \ln W$  appears on his tombstone. The introduction of entropy as a measure of the uncertainty of a density originated from Shannon (1948). Kullback and Leibler (1951) were the first to define the relative entropy, for which reason it is sometimes called Kullback–Leibler entropy. The relation between thermodynamics and information theory has been discussed extensively by Jaynes (1982).

Entropy and relative entropy may also be introduced as characteristic quantities for density functions which are not probability densities, for example, the mass density, charge density, etc. However, in these cases the densities are not necessarily normalized, and in order to obtain such a useful inequality as (2.90) one has to define the relative entropy by (see (2.92), setting  $k = 1$ )

$$S[p|q] = \int dx (p(x) - q(x)) - \int dx p(x) \ln \left( \frac{p(x)}{q(x)} \right). \quad (2.93)$$

#### 2.4.5 Applications

Using the inequality (2.90) satisfied by the relative entropy it will now be easy to see that a constant density distribution always has maximum entropy (compare with the statement in connection with (2.85) about the probability distribution (2.84)). Notice, however, that such a constant density distribution is only possible if  $\Omega$ , the set of possible outcomes, is a compact set, e.g., a finite interval.

Let  $q(x) \equiv q_0$  be the constant density on  $\Omega$  and  $\varrho(x)$  be an arbitrary density. The entropy of this density can also be written as

$$S[\varrho] = S[\varrho|q_0] - k \ln \left( \frac{q_0}{\varrho_0} \right). \quad (2.94)$$

From  $S[\varrho|q_0] \leq 0$  and  $S[\varrho|q_0] = 0$  for  $\varrho \equiv q_0$  follows:  $S[\varrho]$  is maximal for  $\varrho \equiv q_0$ .



As a second application we now consider two random variables  $X_1, X_2$ , their densities  $\varrho_{X_1}(x)$ ,  $\varrho_{X_2}(x)$ , and the joint density  $\varrho_{X_1, X_2}(x_1, x_2)$ . For the relative entropy

$$\begin{aligned} S[\varrho_{X_1, X_2} \mid \varrho_{X_1} \varrho_{X_2}] \\ = - \int dx_1 dx_2 \varrho_{X_1, X_2}(x_1, x_2) \ln \left[ \frac{\varrho_{X_1, X_2}(x_1, x_2)}{\varrho_{X_1}(x_1) \varrho_{X_2}(x_2)} \right] \end{aligned} \quad (2.95)$$

a short calculation yields

$$S[\varrho_{X_1, X_2} \mid \varrho_{X_1} \varrho_{X_2}] = S_{12} - S_1 - S_2, \quad (2.96)$$

where  $S_i$  is the entropy for the density  $\varrho_{X_i}(x)$ ,  $i = 1, 2$  and  $S_{12}$  the entropy of the joint density  $\varrho_{X_1, X_2}(x_1, x_2)$ . As the relative entropy is always smaller than or equal to zero, one always has

$$S_{12} \leq S_1 + S_2. \quad (2.97)$$

Hence, the entropy of the joint density is always smaller than or equal to the sum of the entropies of the single densities. Equality holds if and only if

$$\varrho_{X_1, X_2}(x_1, x_2) = \varrho_{X_1}(x_1) \varrho_{X_2}(x_2), \quad (2.98)$$

i.e., if the two random variables are independent: the entropies of independent random variables add up. For independent random variables the total entropy is maximal. Any dependence between the random variables reduces the total entropy and lowers the uncertainty for the pair of random variables, i.e. any dependency corresponds to an information about the pair of random variables. The relative entropy  $S[\varrho_{X_1, X_2} \mid \varrho_{X_1} \varrho_{X_2}]$  is also known as mutual information.

In the remainder of this section we address the maximum entropy principle. We are looking for the density function  $\varrho(x)$  which has maximum entropy and satisfies the supplementary conditions

$$\langle g_i(X) \rangle \equiv \int dx g_i(x) \varrho(x) = \eta_i, \quad i = 1, \dots, n. \quad (2.99)$$

Here  $g_i(x)$  are given functions and  $\eta_i$  are given real numbers.

From the proposition about the relative entropy proven above, one finds that the density function with maximum entropy satisfying the supplementary conditions (2.99) has the form

$$\varrho(x) = \frac{1}{A} e^{-\lambda_1 g_1(x) - \dots - \lambda_n g_n(x)}. \quad (2.100)$$

Here  $A$  is a normalization factor and  $\{\lambda_i\}$  may be calculated from  $\{\eta_i\}$ . With the help of this maximum entropy principle we can determine density functions.

*Proof.* For the density function (2.100) one obtains (with  $k = 1$ )

$$S[\varrho] = \ln(A\varrho_0) + \sum_{i=1}^n \lambda_i \eta_i, \quad (2.101)$$

where  $\varrho_0$  represents the factor which might be necessary for dimensional reasons. Let  $\varphi(x)$  be a second density satisfying the supplementary conditions (2.99). Then, according to (2.90)

$$S[\varphi|\varrho] \leq 0, \quad (2.102)$$

and therefore

$$S[\varphi] = - \int dx \varphi(x) \ln \left( \frac{\varphi(x)}{\varrho_0} \right) \quad (2.103)$$

$$\leq - \int dx \varphi(x) \ln \left( \frac{\varrho(x)}{\varrho_0} \right) \quad (2.104)$$

$$= \int dx \varphi(x) \left[ \ln(A\varrho_0) + \sum_{i=1}^n \lambda_i g_i(x) \right] \quad (2.105)$$

$$= \ln(A\varrho_0) + \sum_{i=1}^n \lambda_i \eta_i \equiv S[\varrho]. \quad (2.106)$$

Hence,  $\varrho(x)$  given by (2.100) is the density with maximum entropy.

Let us look at two examples. First we seek the density defined on  $[0, \infty)$  which has maximum entropy and satisfies the supplementary condition

$$\langle X \rangle = \eta. \quad (2.107)$$

We immediately find this density as

$$\varrho(x) = \frac{1}{A} e^{-\lambda x} \quad \text{for } x \geq 0. \quad (2.108)$$

The normalization factor  $A$  is given by

$$A = \int_0^\infty dx e^{-\lambda x} = \frac{1}{\lambda}, \quad (2.109)$$

and  $\lambda$  is determined by  $\eta$  according to

$$\eta = \langle X \rangle = \int_0^\infty dx x \lambda e^{-\lambda x} \quad (2.110)$$

$$= -\lambda \frac{\partial}{\partial \lambda} \int_0^\infty dx e^{-\lambda x} = \frac{1}{\lambda}. \quad (2.111)$$

Therefore

$$\varrho(x) = \frac{1}{\eta} e^{-x/\eta}. \quad (2.112)$$

As a second example we seek the density  $\varrho(\mathbf{q}, \mathbf{p})$ , defined on the  $6N$ -dimensional phase space for  $N$  classical particles, which has maximum entropy and satisfies the supplementary condition

$$\langle H(\mathbf{q}, \mathbf{p}) \rangle = E, \quad (2.113)$$

where  $H(\mathbf{q}, \mathbf{p})$  is the Hamiltonian function for the  $N$  particles. One obtains

$$\varrho(\mathbf{q}, \mathbf{p}) = \frac{1}{A} e^{-\lambda H(\mathbf{q}, \mathbf{p})}, \quad (2.114)$$

i.e., the Boltzmann distribution. We still have to determine  $A$  and  $\lambda$ . The former follows from the normalization condition

$$A = \int d^{3N} p d^{3N} q e^{-\lambda H(\mathbf{q}, \mathbf{p})}. \quad (2.115)$$

In particular, we find

$$-\frac{1}{A} \frac{\partial A}{\partial \lambda} = \langle H(\mathbf{q}, \mathbf{p}) \rangle. \quad (2.116)$$

$\lambda$  follows from the supplementary condition:

$$E = \langle H(\mathbf{q}, \mathbf{p}) \rangle = \frac{1}{A} \int d^{3N} p d^{3N} q H(\mathbf{q}, \mathbf{p}) e^{-\lambda H(\mathbf{q}, \mathbf{p})}. \quad (2.117)$$

The right-hand side yields a function  $f(\lambda, N, V)$ , which has to be equal to  $E$ . The resulting equation has to be solved for  $\lambda$  to obtain  $\lambda = \lambda(E, N, V)$ .

The meaning of  $\lambda$  becomes more obvious when we consider the entropy. We have

$$S[\varrho] = \ln(A\varrho_0) + \lambda E \quad (2.118)$$

and therefore, using (2.116),

$$\frac{\partial S[\varrho]}{\partial E} = \frac{\partial \lambda}{\partial E} \frac{\partial}{\partial \lambda} \ln(A\varrho_0) + \frac{\partial \lambda}{\partial E} E + \lambda = \lambda. \quad (2.119)$$

The quantity  $\lambda$  indicates the sensitivity of the entropy to a change in energy. In Chap. 3 on statistical mechanics we will introduce the temperature as being proportional to the inverse of this quantity  $\lambda$ , and we will consider a system of  $N$  particles in a volume  $V$ , for which the temperature, i.e. the parameter  $\lambda$ , is held fixed by contact with a heat bath. In this system, which will be called the canonical system, we will obtain the Boltzmann distribution as the probability density for the positions and momenta of the particles. In the present context it results from the requirement of maximum entropy under the supplementary condition  $\langle H \rangle = E$ .

This has a twofold significance: First, that the energy is not fixed, but the system may exchange energy with the environment (i.e. the heat bath), and second, that  $\langle H \rangle$  is independently given, which is equivalent to fixing the temperature in the canonical system. Both approaches to the Boltzmann distribution proceed from the same physical situation.

If we were looking for a system with maximum entropy which satisfies the supplementary conditions

$$\langle H \rangle = E \quad \text{and} \quad \langle H^2 \rangle = C, \quad (2.120)$$

we would construct a system where both  $\langle H \rangle$  and  $\langle H^2 \rangle$  are independently given. In the canonical system, however, one can determine  $\langle H^2 \rangle$  as a function of  $E, N, V$  or  $T, N, V$ .

## 2.5 Computations with Random Variables

### 2.5.1 Addition and Multiplication of Random Variables

Random variables can be added if their realizations can be added; they can be multiplied if the product of their realizations is meaningful. We may consider functions or mappings of random variables. The question then arises of how to determine the probability density of the sum, the product, and of the mapping.

**Multiplication of a random variable with some constant.** Let us first consider a random variable  $X$  with a density distribution  $\varrho_X(x)$ . We set

$$Z = \alpha X, \quad (2.121)$$

and find

$$\varrho_Z(z) = \int dx \delta(z - \alpha x) \varrho_X(x) = \frac{1}{|\alpha|} \varrho_X\left(\frac{z}{\alpha}\right). \quad (2.122)$$

Obviously

$$\langle Z \rangle = \alpha \langle X \rangle, \quad (2.123a)$$

$$\text{Var}(Z) = \alpha^2 \text{Var}(X), \quad (2.123b)$$

because

$$\langle Z \rangle = \int dz z \varrho_Z(z) = \int dz z \int dx \delta(z - \alpha x) \varrho_X(x) \quad (2.124)$$

$$= \alpha \int dx x \varrho_X(x) = \alpha \langle X \rangle. \quad (2.125)$$

In the same way one can derive the relation for  $\text{Var}(Z)$ .

**Function of a random variable.** Now let us take the more general case

$$Z = f(X). \quad (2.126)$$

One obtains for the density function

$$\varrho_Z(z) = \int dx \varrho_X(x) \delta(z - f(x)) \quad (2.127)$$

$$= \sum_i \frac{1}{|f'(x_i(z))|} \varrho_X(x_i(z)), \quad (2.128)$$

where  $\{x_i(z)\}$  are the solutions which result from solving the equation  $z = f(x)$  for  $x$ . There may be several solutions, which we denote by  $x_i(z)$ ,  $i = 1, \dots$

Apart from this complication, the transformation of the densities under a coordinate transformation  $x \rightarrow z(x)$  may also be obtained from the identity

$$1 = \int dx \varrho_X(x) = \int dz \left| \frac{dx}{dz} \right| \varrho_X(x(z)) = \int dz \varrho_Z(z), \quad (2.129)$$

and we find, in agreement with (2.128),

$$\varrho_Z(z) = \varrho_X(x(z)) \left| \frac{dx}{dz} \right|. \quad (2.130)$$

A similar relation holds for several dimensions. Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be an  $n$ -tuple of random variables and

$$Z_i = Z_i(\mathbf{X}), \quad i = 1, \dots, n \quad (2.131)$$

a mapping  $\mathbf{X} \rightarrow \mathbf{Z} = (Z_1, \dots, Z_n)$ . For the density function  $\varrho_{\mathbf{Z}}(\mathbf{z})$  one then finds

$$\varrho_{\mathbf{Z}}(\mathbf{z}) = \left| \frac{\partial(x_1, \dots, x_n)}{\partial(z_1, \dots, z_n)} \right| \varrho_{\mathbf{X}}(\mathbf{x}(\mathbf{z})). \quad (2.132)$$

*Examples.*

(a) Let

$$Z = -\ln X, \quad \varrho_X(x) = 1 \quad \text{for } x \in [0, 1]. \quad (2.133)$$

With  $dz/dx = -1/x$ , and hence  $|dx/dz| = |x| = e^{-z}$ , one obtains

$$\varrho_Z(z) = e^{-z} \varrho_X(x(z)). \quad (2.134)$$

Thus, with  $\varrho_X(x) = 1$ ,  $Z$  is exponentially distributed.

(b) Let

$$(Z_1, Z_2) = \sqrt{-2 \ln X_1} (\cos 2\pi X_2, \sin 2\pi X_2), \quad (2.135)$$

where  $X_1$  and  $X_2$  are independent and uniformly distributed in  $[0, 1]$ . Then

$$\varrho_{\mathbf{Z}}(\mathbf{z}) \equiv \left| \frac{\partial(x_1, x_2)}{\partial(z_1, z_2)} \right| \varrho_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-z_2^2/2}. \quad (2.136)$$

In this case  $(Z_1, Z_2)$  are also independent, each with a standard normal distribution.

**Addition of random variables.** Let  $X_1, X_2$  be two random variables with probability density  $\varrho(x_1, x_2)$ . We set

$$Z = X_1 + X_2. \quad (2.137)$$

Then

$$\varrho_Z(z) = \int dx_1 dx_2 \delta(z - x_1 - x_2) \varrho(x_1, x_2) \quad (2.138)$$

$$= \int dx_1 \varrho(x_1, z - x_1). \quad (2.139)$$

Three distinct cases are of interest. If  $X_1$  and  $X_2$  are both normally distributed, then, according to (2.24), the joint density  $\varrho(x_1, x_2)$  is an exponential function with an exponent quadratic in  $x_1$  and  $x_2$ . For  $Z = X_1 + X_2$  one may derive the density from (2.139). The integrand in (2.139), i.e.,  $\varrho(x_1, z - x_1)$ , and also the result of the integration are therefore exponential functions with quadratic exponents. This implies that  $\varrho_Z(z)$  is also of this form, and therefore  $Z$  is also a normally distributed random variable.

Hence, the sum of two normal random variables is always (even if they are mutually dependent) another normal random variable. More generally, every linear superposition of normal random variables is again a normal random variable.

Next, if  $X_1$  and  $X_2$  are independent with probability densities  $\varrho_{X_1}(x)$  and  $\varrho_{X_2}(x)$ , respectively, we find

$$\varrho_Z(z) = \int dx_1 \varrho_{X_1}(x_1) \varrho_{X_2}(z - x_1), \quad (2.140)$$

i.e., the density function for a sum of two independent random variables is the convolution of the individual density functions. For the characteristic function we obtain

$$G_Z(k) = G_{X_1}(k) G_{X_2}(k). \quad (2.141)$$

It is easy to show that for independent random variables  $X_1, X_2$

$$\langle Z \rangle = \langle X_1 \rangle + \langle X_2 \rangle, \quad (2.142)$$

$$\text{Var}(Z) = \text{Var}(X_1) + \text{Var}(X_2). \quad (2.143)$$

The first relation follows from

$$\langle Z \rangle = \int dz \int dx_1 dx_2 z \delta(z - x_1 - x_2) \varrho_1(x_1) \varrho_2(x_2) \quad (2.144)$$

$$= \int dx_1 dx_2 (x_1 + x_2) \varrho_1(x_1) \varrho_2(x_2) \quad (2.145)$$

$$= \langle X_1 \rangle + \langle X_2 \rangle. \quad (2.146)$$

The equation for  $\text{Var}(Z)$  can be derived similarly.

When all cumulants exist, we may use (2.141) for the characteristic function to prove the sum rules, because it is a direct consequence of

$$G_X(k) = \exp \left( ik\kappa_1(X) - \frac{1}{2}k^2\kappa_2(X) - \dots \right) \quad (2.147)$$

that the cumulants of a sum  $Z = X_1 + X_2$  are the sums of the cumulants. As  $\kappa_1(X) = \langle X \rangle$  and  $\kappa_2(X) = \text{Var}(X)$ , (2.142) and (2.143) follow.

Finally, for two *dependent* random variables  $X_1, X_2$ , (2.142) still holds, which is not necessarily true for (2.143). In this case

$$\text{Var}(Z) = \text{Var}(X_1) + 2\text{Cov}(X_1, X_2) + \text{Var}(X_2). \quad (2.148)$$

**Multiplication of independent random variables.** Let  $X_1, X_2$  be two independent random variables with probability densities  $\varrho_{X_1}(x)$  and  $\varrho_{X_2}(x)$ . We set

$$Z = X_1 X_2. \quad (2.149)$$

Then

$$\varrho_Z(z) = \int dx_1 dx_2 \delta(z - x_1 x_2) \varrho_{X_1}(x_1) \varrho_{X_2}(x_2) \quad (2.150)$$

$$= \int dx_1 \varrho_{X_1}(x_1) \frac{1}{|x_1|} \varrho_{X_2}\left(\frac{z}{x_1}\right). \quad (2.151)$$

### 2.5.2 Further Important Random Variables

Having learnt how to calculate with random variables, we can now construct some important new random variables by combining some of those that we have already met.

First, we consider  $n$  independent random variables  $X_1, \dots, X_n$  with standard normal distributions and set

$$Z = X_1^2 + \dots + X_n^2. \quad (2.152)$$

The density distribution of  $Z$  is given by

$$\varrho_Z(z) = \int dx_1 \dots dx_n \delta(z - x_1^2 - \dots - x_n^2) \varrho(x_1, \dots, x_n), \quad (2.153)$$

with

$$\varrho(x_1, \dots, x_n) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(x_1^2 + \dots + x_n^2)\right). \quad (2.154)$$

We obtain

$$\varrho_Z(z) = \frac{1}{2^{n/2} \Gamma(n/2)} z^{n/2-1} e^{-z/2}, \quad (2.155)$$

where  $\Gamma(x)$  is the gamma function ( $\Gamma(x+1) = x\Gamma(x)$ ,  $\Gamma(1/2) = \sqrt{\pi}$ ,  $\Gamma(N+1) = N!$  for  $N = 0, 1, \dots$ ).

This random variable  $Z$  occurs so frequently that it bears its own name:  $\chi_n^2$  with  $n$  degrees of freedom. One finds

$$\langle \chi_n^2 \rangle = n, \quad (2.156a)$$

$$\text{Var}(\chi_n^2) = 2n. \quad (2.156b)$$

One is equally likely to encounter the random variable  $\sqrt{Z} = \chi_n$  with density

$$\varrho_{\chi_n}(z) = \frac{1}{2^{n/2-1} \Gamma(n/2)} z^{n-1} e^{-z^2/2}. \quad (2.157)$$

The three components  $v_i$  of the velocities of molecules in a gas at temperature  $T$  are normally distributed with mean 0 and variance  $\sigma^2 = k_B T/m$  (cf. Sect. 2.2). Here,  $m$  denotes the mass of a molecule and  $k_B$  Boltzmann's constant. Therefore,  $n = 3$  and  $v_i = \sigma X_i$ , where  $X_i$  is a random variable with a standard normal distribution. For the absolute value of the velocity  $v = \sqrt{v_1^2 + v_2^2 + v_3^2}$  we obtain the density

$$\varrho(v) = \frac{1}{\sigma} \varrho_{\chi_3}\left(\frac{v}{\sigma}\right) = \sqrt{\frac{2m^3}{\pi(k_B T)^3}} v^2 \exp\left(-\frac{mv^2}{2k_B T}\right). \quad (2.158)$$



This is also called the Maxwell–Boltzmann distribution. Furthermore, we find

$$\left\langle \frac{m}{2} v^2 \right\rangle = \frac{m}{2} \sigma^2 \langle \chi_3^2 \rangle = \frac{m}{2} \frac{k_B T}{m} \cdot 3 = \frac{3}{2} k_B T. \quad (2.159)$$

Next, consider two  $\chi^2$ -distributed random variables  $Y_k$  and  $Z_q$  with  $k$  and  $q$  degrees of freedom, respectively. The ratio

$$Z = \frac{Y_k/k}{Z_q/q} \quad (2.160)$$

is a so-called  $F_{k,q}$ -distributed random variable. For the density one obtains

$$\begin{aligned} \varrho_{F_{k,q}}(z) &= \left(\frac{k}{q}\right)^{k/2} \frac{\Gamma(1/2(k+q))}{\Gamma(1/2k)\Gamma(1/2q)} \\ &\times z^{k/2-1} \left(1 + \frac{k}{q}z\right)^{-(k+q)/2}. \end{aligned} \quad (2.161)$$

Finally, let  $Y$  be a random variable with a standard normal distribution and  $Z_q$  be a  $\chi^2$ -distributed random variable with  $q$  degrees of freedom. The ratio

$$T_q = \frac{Y}{\sqrt{Z_q/q}} \quad (2.162)$$

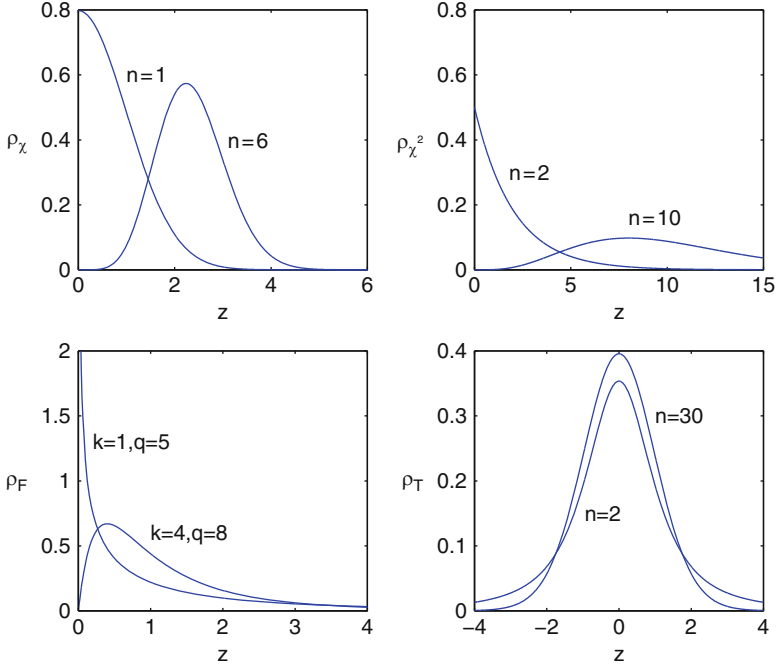
defines a  $t$ -distributed random variable with  $q$  degrees of freedom. The density

$$\varrho_{T_q}(z) = \frac{1}{\sqrt{q}} \frac{\Gamma(1/2 + q/2)}{\Gamma(1/2)\Gamma(q/2)} \left(1 + \frac{z^2}{q}\right)^{-(q+1)/2} \quad (2.163)$$

is also called the density function of the Student  $t$ -distribution (after the pseudonym ‘Student’ assumed by the English statistician W. S. Gosset).

*Remark.* Above we have introduced the probability densities of some frequently occurring random variables. In the computation of characteristic quantities, in particular the  $\alpha$ -quantiles for general  $\alpha$ , which are often required in practice, one encounters special functions like the incomplete beta function. Here we do not want to deal with such calculations, since quantities such as the  $\alpha$ -quantiles can be found from any statistics software package.

However, we do want to introduce the graphs of some densities in Fig. 2.6. As can be seen from the formulas, the densities of the  $\chi$ -,  $\chi^2$ -, and  $F$ -distributions tend to zero for  $z \rightarrow 0$ , when  $n$  exceeds 1 or 2, or when  $k$  exceeds the value 2. For  $z \rightarrow \infty$  these functions decrease exponentially or as a power law. For large values of  $q$  the density of the  $t$ -distribution strongly resembles the normal distribution.



**Fig. 2.6** Density functions of the  $\chi$ -distribution (*upper left*), the  $\chi^2$ -distribution (*upper right*), the  $F$ -distribution (*lower left*) and the  $t$ -distribution (*lower right*)

### 2.5.3 Limit Theorems

In this subsection we consider sums of  $N$  independent and identically distributed random variables and investigate the properties of the densities for such sums as  $N \rightarrow \infty$ . The resulting propositions are called limit theorems. They play an important role for all complex systems which consist of many subsystems and where the characteristic quantities of the total system result from sums of the corresponding quantities of the subsystems.

**The central limit theorem.** Let  $X_i, i = 1, \dots, N$ , be independent and identically distributed random variables. All cumulants shall exist and let

$$\langle X_i \rangle = 0, \quad (2.164a)$$

$$\text{Var}(X_i) = \sigma^2, \quad i = 1, \dots, N. \quad (2.164b)$$

We set

$$Z_N = \frac{1}{\sqrt{N}}(X_1 + \dots + X_N). \quad (2.165)$$

It follows that

$$\langle Z_N \rangle = 0, \quad (2.166a)$$

$$\text{Var}(Z_N) = \frac{1}{N} \sum_{i=1}^N \text{Var}(X_i) = \sigma^2, \quad (2.166b)$$

furthermore, all higher moments and cumulants decrease at least as fast as  $N^{-1/2}$  for  $N \rightarrow \infty$ .

Thus for  $N \rightarrow \infty$  the random variable  $Z_N$  is a Gaussian random variable with mean 0 and variance  $\sigma^2$ . Because of its far-reaching significance this statement is also called the ‘central limit theorem’. It has been proven for many different and more general conditions (see e.g. Gardiner 1985).

So, according to the central limit theorem, we may describe the total influence resulting from a superposition of many stochastic influences by a Gaussian random variable. For this reason one often assumes that the measurement errors are realizations of Gaussian random variables.

We will make frequent use of the central limit theorem. A first simple application is the following: Suppose the random number generator of a computer provides us with random numbers  $x$  which are uniformly distributed on the interval  $[0, 1]$ . Then  $q(x) = 1$  for  $x \in [0, 1]$ , and

$$\langle X \rangle = \frac{1}{2}, \quad (2.167a)$$

$$\text{Var}(X) = \int_0^1 x^2 dx - \langle X \rangle^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \quad (2.167b)$$

Hence

$$X' = (X - \frac{1}{2})\sqrt{12}\sigma \quad (2.168)$$

is a random number with vanishing mean value and variance  $\sigma^2$ , uniformly distributed in  $[-\frac{1}{2}\sqrt{12}\sigma, \frac{1}{2}\sqrt{12}\sigma]$ . If we select  $N$  such numbers and set

$$Z_N = \frac{1}{\sqrt{N}}(X'_1 + \dots + X'_N), \quad (2.169)$$

then  $Z_N$  is approximately a Gaussian random variable with variance  $\sigma^2$  and mean 0. For  $N = 12$  this approximation is already quite good.

**The mean of random variables.** Consider the independent random variables  $X, X_1, \dots, X_N$ , all having the same probability density. All moments and all cumulants shall exist. We set

$$Z_N = \frac{1}{N}(X_1 + \dots + X_N). \quad (2.170)$$

Then

$$\langle Z_N \rangle = \frac{1}{N} \sum_{i=1}^N \langle X_i \rangle = \langle X \rangle, \quad (2.171a)$$

$$\text{Var}(Z_N) = \frac{1}{N} \text{Var}(X), \quad (2.171b)$$

and all higher moments and cumulants decrease like  $1/N^2$  or faster in the limit  $N \rightarrow \infty$ .

As an application, consider  $N$  independent realizations  $x_1, \dots, x_N$  of the random variable  $X$  and form the expectation value

$$z_N = \frac{1}{N} (x_1 + \dots + x_N). \quad (2.172)$$

Each  $x_i$  may also be thought of as a realization of  $X_i$ , where each random variable  $X_i$  is a ‘copy’ of  $X$ ; therefore  $z_N$  is a realization of  $Z_N$ . For  $N$  large enough the higher cumulants are negligible and thus  $Z_N$  may be regarded as a Gaussian random variable with expectation value  $\langle X \rangle$  and variance  $\text{Var}(X)/N$ . For larger values of  $N$  the realization  $z_N$  of the mean value scatter less and less around  $\langle X \rangle$ , and the distribution of  $Z_N$  is better and better approximated by a Gaussian distribution. For  $N \rightarrow \infty$  the support of the density for the random variables  $Z_N$ , i.e., the domain where the density is larger than any arbitrarily small  $\varepsilon$ , shrinks to the value  $\langle X \rangle$ .

Thus, by forming the mean value of  $N$  realizations of a random variable  $X$  one obtains a good ‘estimator’ for  $\langle X \rangle$ . This estimator gets better and better for larger values of  $N$ . This is the origin of the casual habit of using the expressions ‘mean value’ and ‘expectation value’ as synonyms, although the expectation value is a quantity which is derived from a probability density, while the mean value always refers to the mean value of realizations. In Part II we will make the concept of estimators more precise.

Above we have considered two differently normalized sums of  $N$  independent and identically distributed random variables with finite variance. In the first case the limit distribution is again a normal distribution, in the second case it is concentrated around a point. These are two typical scenarios which occur frequently in statistical physics. There, however, we mostly deal with dependent random variables, and the dependence is described by the models of the interactions among the different subsystems.

Sums of random variables will be further considered in the next two sections.

## 2.6 Stable Random Variables and Renormalization Transformations

In Sect. 2.5.3 we identified the normal distribution as the limit of a large class of distributions. Now we will become acquainted with other classes of random variables that all have prominent distributions as limit distributions.

### 2.6.1 Stable Random Variables

We first introduce the notion of a stable distribution. Let  $X, X_1, \dots, X_N$  be independent and identically distributed random variables with a density  $\varrho(x)$ , and, furthermore, let

$$S_N = X_1 + \dots + X_N. \quad (2.173)$$

We define the density  $\varrho(x)$  as *stable* if there exist constants  $c_N > 0$  and  $d_N$  for any  $N \geq 2$  such that  $S_N$  has the same density as  $c_N X + d_N$ . The density  $\varrho(x)$  is called *strictly stable* if this statement is true for  $d_N = 0$ .

For example, the normal distribution is stable: A sum of normal random variables is again normally distributed. But the Cauchy distribution, which we met in Sect. 2.3, with its density and generating function,

$$\varrho(x) = \frac{1}{\pi} \frac{\gamma^2}{(x - \mu)^2 + \gamma^2}, \quad (2.174)$$

$$G(k) = \langle e^{ikX} \rangle = e^{ik\mu - |k|\gamma}, \quad \gamma > 0 \quad (2.175)$$

is also stable. The sum of  $N$  Cauchy random variables is again a Cauchy random variable, because in this case the characteristic function of  $S_N$  is

$$G_{S_N} = e^{iNk\mu - N|k|\gamma}, \quad (2.176)$$

and therefore

$$Y = \frac{1}{N} (X_1 + \dots + X_N) \quad (2.177)$$

is again Cauchy distributed with the same parameters. The densities of the normal distribution and the Cauchy distribution differ with respect to their behavior for large  $|x|$ . The moments of the Cauchy distribution do not exist.

Thus the normal distribution and the Cauchy distribution are two important stable distributions, the first one with  $c_N = N^{1/2}$ , the second one with  $c_N = N$ . One can now prove the following statement (Feller 1957; Samorodnitzky and Taqqu 1994): The constant  $c_N$  can in general only be of the form

$$c_N = N^{1/\alpha} \quad \text{with} \quad 0 < \alpha \leq 2. \quad (2.178)$$

The quantity  $\alpha$  is called the index or the characteristic exponent of the stable density. For the Cauchy distribution we find  $\alpha = 1$ ; for the normal distribution  $\alpha = 2$ .

A stable density with index  $\alpha = 1/2$  is

$$\varrho(x) = \begin{cases} \frac{1}{\sqrt{2\pi}x^3} e^{-1/2x} & \text{for } x > 0, \\ 0 & \text{for } x < 0. \end{cases} \quad (2.179)$$

For such random variables  $\{X_i\}$  with  $\alpha = 1/2$ ,

$$Y = \frac{1}{N^2}(X_1 + \dots + X_N) \quad (2.180)$$

is again a random variable with the density given in (2.179).

For a stable density  $\varrho(x)$  with exponent  $\alpha \neq 1$  one can always find a constant  $\mu$  such that  $\varrho(x - \mu)$  is strictly stable. For  $\alpha = 1$  this shift of the density is unnecessary as the Cauchy distribution is strictly stable even for  $\mu \neq 0$ .

The generating function of strictly stable densities is

$$G(k) = e^{-|k|^\alpha \gamma}, \quad (2.181)$$

with some scale parameter  $\gamma > 0$ . Thus for  $\alpha < 2$  one obtains for  $x \rightarrow \infty$

$$|x|^{1+\alpha} \varrho(x) \rightarrow \text{const.} \neq 0. \quad (2.182)$$

The stable densities with characteristic exponents  $\alpha < 2$  do not have a finite variance.

More generally, stable densities may be characterized not by three but by four parameters: In addition to the index  $\alpha$ , the scale parameter  $\gamma$ , and the shift parameter  $\mu$ , one has the skewness  $\beta$ , which we now meet for the first time. The skewness  $\beta$  measures the deviation from symmetry. For  $\beta = 0$  we have  $\varrho(-x) = \varrho(x)$ . As we want to deal here only with strictly stable densities, for which  $\beta = 0$  for all  $\alpha \in (0, 2]$ , we give no further details concerning this parameter or its role in the characteristic function.

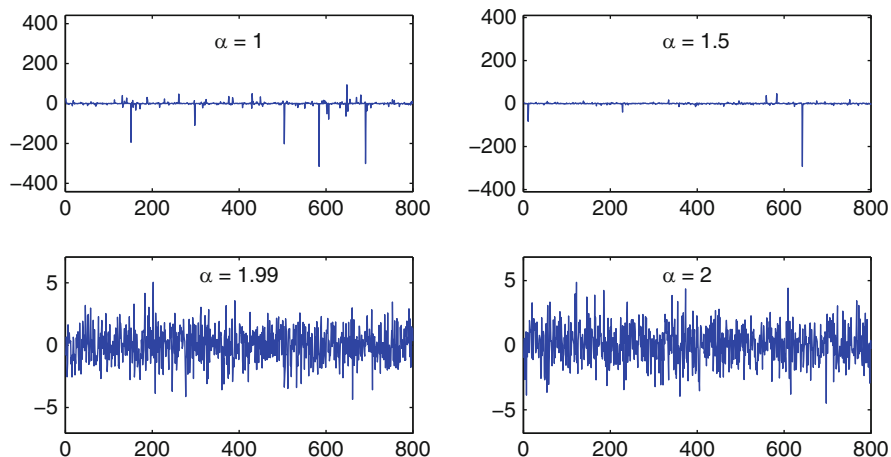
*Remark.* A realization  $x$  of a random variable with a strictly stable density for an index  $\alpha$  and a scale parameter  $\gamma = 1$  can be constructed as follows (Samorodnitzky and Taqqu 1994): Take a realization  $r$  of a uniformly distributed random variable in the interval  $[-\pi/2, \pi/2]$  and, independently, a realization  $v$  of an exponential random variable with mean 1. Then set

$$x = \frac{\sin(\alpha r)}{(\cos r)^{1/\alpha}} \left( \frac{\cos((1-\alpha)r)}{v} \right)^{(1-\alpha)/\alpha}. \quad (2.183)$$

A series of such realizations is represented in Fig. 2.7 for various values of  $\alpha$ . For decreasing  $\alpha$  the larger deviations become larger and more frequent. A realization  $x$  of a Cauchy random variable ( $\alpha = 1$ ) with scale parameter  $\gamma$  and shift parameter  $\mu$  is more easily constructed: Take a realization  $r$  of a random variable uniformly distributed in  $[-\pi/2, \pi/2]$  and set

$$x = \gamma \tan r + \mu. \quad (2.184)$$

Special constructions also exist for  $\alpha = 2^{-k}$ ,  $k \geq 1$ .



**Fig. 2.7** A series of realizations of a random variable with a stable density for four different values of the index  $\alpha$

### 2.6.2 The Renormalization Transformation

There is a further way to characterize stable distributions: Let  $X = \{X_i\}_{i=-\infty}^{\infty}$  be a sequence of independent and identically distributed random variables with density  $\varrho(x)$ . We consider the transformation  $T_n$ ,  $n \geq 1$ , for which

$$X'_i = (T_n X)_i = \frac{1}{n^\delta} \sum_{j=i-n}^{(i+1)n-1} X_j. \quad (2.185)$$

In this transformation the random variables are thus combined into blocks of length  $n$ . The random variables within each block are summed up and this sum is renormalized by a power  $\delta$  of the length  $n$  of this block. This transformation is called a renormalization transformation. The family of transformations  $\{T_n, n \geq 1\}$  form a semi-group, i.e.  $T_{mn} = T_m T_n$ . This semi-group is also called renormalization group. A sequence  $X = \{X_i\}_{i=-\infty}^{\infty}$  is a fixed point of this group of transformations if the  $X'_i$  resulting from  $T_n X$  have the same density  $\varrho(x)$  as the  $X_i$ .

A sequence of independent strictly stable random variables with characteristic exponent  $\alpha$  is obviously a fixed point for  $\{T_n, n \geq 1\}$  with  $\delta = 1/\alpha$ . Therefore, such stable densities appear as the limit of sequences of densities, which result from successive applications of the transformation  $T_n$  with  $n$  fixed (or a single transformation  $T_n$  with increasing  $n$ ) to a given sequence of random variables with a given density. Under successive transformations all densities with finite variance, i.e.,  $\alpha = 2$ , approach the normal distribution. This corresponds to the central limit theorem.

Hence, stable densities have a domain of attraction of densities. For the transformation with index  $\alpha$  all densities with the asymptotic behavior (2.182) belong to the domain of attraction of the stable density with exponent  $\alpha$ .

Suppose we are given a density which belongs in the above sense to the domain of attraction of a stable density with index  $\alpha$ . If the ‘wrong’ transformation is applied to this density, i.e., a transformation with index  $\beta \neq \alpha$ , then the limit is either not a density or there is a drift towards a density which is concentrated around one point. There are also densities which do not belong to the domain of attraction of any stable density.

### 2.6.3 Stability Analysis

We now want to examine the behavior of densities close to a fixed point. For  $n = 2$  the renormalization transformations may also easily be formulated on the level of densities. For  $\delta = 1/\alpha$  one obtains

$$(T_2 \varrho)(x) = \varrho_{X'}(x) = 2^{1/\alpha} \int dy \varrho(2^{1/\alpha}x - y) \varrho(y). \quad (2.186)$$

The stable density representing the fixed point will be denoted by  $\varrho^*(x)$ . Let  $\varrho(x) = \varrho^*(x) + \eta(x)$ . For the deviation  $\eta(x)$  the transformation  $T_2$  leads to

$$\eta' = T_2(\varrho^* + \eta) - T_2 \varrho^* = DT_2 \eta + \mathcal{O}(\eta^2) \quad (2.187)$$

with

$$(DT_2 \eta)(x) = 2 \int dy \varrho^*(2^{1/\alpha}x - y) \eta(y). \quad (2.188)$$

Let  $\phi_n(x)$  denote the eigenfunctions of  $DT_2$  and  $\lambda_n$  the corresponding eigenvalues. Then

$$(DT_2 \phi_n)(x) = \lambda_n \phi_n(x). \quad (2.189)$$

Obviously,  $\varrho^*(x)$  itself is an eigenfunction with eigenvalue 2. We set  $\phi_0 = \varrho^*(x)$ ,  $\lambda_0 = 2$ .

Let  $v_n$  be the coefficients of an expansion of the deviation  $\eta(x)$  with respect to the eigenfunctions  $\phi_n$ , i.e.

$$\eta(x) = \sum_{n=1}^{\infty} v_n \phi_n(x). \quad (2.190)$$

*Remark.* For  $\alpha = 2$ , i.e., for the densities with finite variance, the eigenfunctions and eigenvalues are simply

$$\phi_n(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} H_n\left(\frac{x}{\sigma}\right) \quad \text{and} \quad \lambda_n = (\sqrt{2})^{2-n}. \quad (2.191)$$



Here  $\{H_n\}$  denote the Hermite polynomials. In particular, the first polynomials are

$$\begin{aligned} H_1(x) &= x, & H_2(x) &= x^2 - 1, \\ H_3(x) &= x^3 - 3x, & H_4(x) &= x^4 - 6x^2 + 3. \end{aligned} \quad (2.192)$$

When a density  $\varrho(x)$  belonging to the domain of attraction of the normal distribution is approximated by the density of the normal distribution with the same variance and the same mean value, the difference  $\eta$  can be represented as (cf. Papoulis 1984)

$$\eta(x) \equiv \varrho(x) - \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \sum_{n=3}^{\infty} v_n H_n\left(\frac{x}{\sigma}\right), \quad (2.193)$$

and the coefficients  $\{v_n\}$  are proportional to the moments  $\{\mu_n\}$  of the density  $\varrho(x)$ . One obtains, for example,

$$v_3 = \frac{1}{3!\sigma^3} \mu_3, \quad v_4 = \frac{1}{4!\sigma^4} (\mu_4 - 3\sigma^4). \quad (2.194)$$

For a more general distribution we also have

$$v_1 = \frac{1}{\sigma} \mu_1, \quad v_2 = \frac{1}{2!\sigma^2} (\mu_2 - \sigma^2). \quad (2.195)$$

In a linear approximation the deviation  $\eta(x)$  changes under a renormalization transformation according to

$$\eta'(x) = (DT_2\eta)(x) = \sum_{n=1}^{\infty} v_n' \phi_n(x), \quad (2.196)$$

where

$$v_n' = \lambda_n v_n, \quad (2.197)$$

i.e. the coefficients  $\{v_n\}$  are the characteristic parameters of the density  $\varrho(x)$ , which in a linear approximation transform covariantly under a renormalization transformation. We will also call them scale parameters.

For the density  $\varrho(x)$  to belong to the domain of attraction of the density  $\varrho^*(x)$  under the renormalization transformation  $T_2$ , the eigenvalues  $\lambda_n$  obviously have to satisfy  $\lambda_n < 1$ , unless  $v_n = 0$ . In physics, those parameters  $v_n$  for which  $\lambda_n > 1$  are called relevant parameters. If  $\lambda_n = 1$  one speaks of marginal parameters, and if  $\lambda_n < 1$  they are called irrelevant parameters.

The subspace of the space of parameters  $\{v_n\}$  for which all relevant parameters vanish is called a ‘critical surface’ in physics. So this space is identical to the domain of attraction of the stable density  $\varrho^*(x)$ .

### 2.6.4 Scaling Behavior

For a given renormalization transformation we now examine the transformation properties of a density which does not belong to the domain of attraction of the corresponding stable density, but which is close to this domain in the following sense: There shall exist an expansion of this density with respect to the eigenfunctions  $\{\phi_n\}$ ,

$$\varrho(x) = \varrho^*(x) + \sum_{n=1} v_n \phi_n(x), \quad (2.198)$$

such that the relevant parameters, which we take to be  $v_1$  and  $v_2$  without loss of generality, are supposed to be small. If they were to vanish,  $\varrho(x)$  would belong to the domain of attraction.

The generating function of the cumulants,

$$F[\varrho(x), t] = \ln \left( \int dx \varrho(x) e^{itx} \right), \quad (2.199)$$

is now considered as a functional of  $\varrho(x)$  and a function of  $t$ . Let  $\varrho'(x) = (T_2\varrho)(x)$ , then

$$F[\varrho'(x), t] = \ln \left( \int dx 2^{1/\alpha} \int dy \varrho(2^{1/\alpha}x - y) \varrho(y) e^{itx} \right) \quad (2.200)$$

$$= 2F \left[ \varrho(x), \frac{t}{2^{1/\alpha}} \right]. \quad (2.201)$$

The functional  $F[\varrho(x), t]$  transforms covariantly under the renormalization transformation. As the densities may equivalently be characterized by their scale parameters  $\{v_n\}$  and  $\{v'_n\}$ ,  $F$  can also be considered as a function  $F(v_1, \dots, t)$  of these scale parameters and the variable  $t$ . Thus

$$F(v'_1, v'_2, v'_3, \dots, t) = 2F \left( v_1, v_2, v_3, \dots, \frac{t}{2^{1/\alpha}} \right), \quad (2.202)$$

or, taking  $v'_n = \lambda_n v_n$ ,  $\lambda = 2$ ,  $\lambda_n = \lambda^{a_n}$ ,  $\lambda^{a_t} = 2^{1/\alpha}$ ,

$$F(\lambda^{a_1} v_1, \lambda^{a_2} v_2, \lambda^{a_3} v_3, \dots, \lambda^{a_t} t) = \lambda F(v_1, v_2, v_3, \dots, t). \quad (2.203)$$

For densities close to the fixed point of the renormalization transformation the irrelevant scale parameters  $v_3, \dots$  will be small. To a good approximation  $F$  can be considered as independent of these parameters and one obtains the scaling relation:

$$F(\lambda^{a_1} v_1, \lambda^{a_2} v_2, \lambda^{a_t} t) = \lambda F(v_1, v_2, t). \quad (2.204)$$

This behavior holds for all densities that are close to the domain of attraction of the corresponding stable density. In this sense it may be called universal.

From such a scaling law one can easily determine the behavior of  $F$  (and other derived quantities) close to the domain of attraction (the critical surface) of the stable density corresponding to the renormalization transformation. We will come back to this point in Sect. 4.7, where we will consider renormalization transformations in the context of random fields, in particular for spin systems.

*Remark.* For those densities in the domain of attraction of the normal distribution,  $F$  can be explicitly represented by an expansion in cumulants:

$$F[\varrho(x), t] = \sum_{n=1}^{\infty} \frac{(it)^n}{n!} \kappa_n. \quad (2.205)$$

The cumulants  $\{\kappa_n\}$  transform in the same way as the  $\{v_n\}$ , i.e.  $\kappa'_n = 2^{1-n/2} \kappa_n$ , because the cumulants of a sum of random variables is the sum of the cumulants and the renormalization by the factor  $2^{-1/2}$  produces a further factor  $2^{-n/2}$ . Then also  $\kappa'_n (\lambda^{a_t} t)^n \equiv 2^{1-n/2} \kappa_n (\sqrt{2}t)^n = 2\kappa_n t^n$ , hence, in this case there exists an easier way to derive the scaling relations:

$$F(\lambda^{a_1} \kappa_1, \lambda^{a_2} \kappa_2, \lambda^{a_3} \kappa_3, \dots, \lambda^{a_t} t) = \lambda F(\kappa_1, \kappa_2, \kappa_3, \dots, t). \quad (2.206)$$

It is of the same form as (2.203) with  $\{v_n\}$  now replaced by  $\{\kappa_n\}$ . Note, however, that the two sets of covariant parameters are easily transformed into each other.

## 2.7 The Large Deviation Property for Sums of Random Variables

In Sect. 2.5 we have learnt the arithmetic of random variables and investigated the behavior of densities of  $N$  independent random variables for large values of  $N$ . We have formulated a first version of the central limit theorem. In Sect. 2.6 we studied special classes of random variables which can also be seen as limits of a sequence of properly normalized sum of  $N$  ( $N = 2, \dots$ ) random variables.

A sum of random variables represents a prototype of a macrovariable for a statistical system. In systems having many degrees of freedom, quantities describing the total system are often represented by sums of quantities pertaining to the single components or degrees of freedom. The kinetic energy of the total system is composed of the kinetic energies of the single constituents; the magnetization of a spin system is the mean value of all magnetic moments. Every extensive quantity is a sum of corresponding quantities for the subsystems.

In this section we will study the density of such sums for large  $N$ . We first introduce a special property for such a sequence which will turn out to be very relevant in statistical systems and, whenever this property is met, some strong statements can be made about the density of the macrovariable.

In order to be able to introduce this property and to make the statements about the density we first introduce two new concepts.

**The free energy function.** Let  $X$  be a random variable with density  $\varrho(x)$ . Then

$$f(t) = \ln \langle e^{tX} \rangle = \ln \left[ \int dx e^{tx} \varrho(x) \right] \quad (2.207)$$

is called the free energy function. This name for  $f(t)$  refers to the fact that in statistical mechanics this function is closely related to the free energy of thermodynamics.  $f(t)$  is the generating function of the cumulants, if they exist. In Sect. 2.3, formula (2.66), we introduced  $f(ik) = \ln G(k)$  as such a generating function. But  $f(t)$  is real and it is easy to show that it is also a strictly convex function, i.e., the second derivative always obeys  $f''(t) > 0$ , unless the density  $\varrho(x)$  is concentrated around a point.

We give some examples:

- For the normal distribution  $X \sim N(\mu, \sigma^2)$  one finds

$$f(t) = \mu t + \frac{1}{2} \sigma^2 t^2. \quad (2.208)$$

- For the exponential distribution  $\varrho(x) = m e^{-mx}$  we have

$$f(t) = -\ln \left( \frac{m-t}{m} \right), \quad t < m. \quad (2.209)$$

- For a random variable with discrete realizations  $\{-1, +1\}$  and  $\varrho(\pm 1) = 1/2$  one obtains

$$f(t) = \ln (\cosh t). \quad (2.210)$$

**The Legendre transform.** Let  $f(t)$  be a strictly convex function, then the Legendre transform  $g(y)$  of  $f(t)$  is defined on  $[0, \infty)$  by

$$g(y) = \sup_t (ty - f(t)). \quad (2.211)$$

$g(y)$  is again strictly convex.

Hence, in order to write down  $g(y)$  explicitly one first has to determine the supremum; it is found at  $t = t(y)$ , where  $t(y)$  follows from solving the equation

$$y = f'(t) \quad (2.212)$$

for  $t$ . The convexity of  $f(t)$  guarantees that  $t(y)$  exists. Thereby one obtains

$$g(y) = t(y)y - f(t(y)) \quad (2.213)$$

and

$$dg = y dt + t dy - f'(t) dt = t dy, \quad \text{i.e. also} \quad g'(y) = t(y). \quad (2.214)$$

In this way the Hamiltonian function  $H(p, q)$  of a classical mechanical system is the Legendre transform of the Lagrange function  $L(\dot{q}, q)$ :

$$H(p, q) = \sup_{\dot{q}} \left( p \dot{q} - L(\dot{q}, q) \right). \quad (2.215)$$

The Lagrange function is convex with respect to the argument  $\dot{q}$ , since  $L(\dot{q}, q) = m\dot{q}^2/2 + \dots$

Let us determine the Legendre transforms for the above mentioned examples.

- For the normal distribution  $N(\mu, \sigma^2)$  one obtains from (2.208)

$$g(y) = \frac{(y - \mu)^2}{2\sigma^2}, \quad (2.216)$$

- For the exponential distribution follows from (2.209)

$$g(y) = my - 1 - \ln my, \quad (2.217)$$

- And for a random variable with discrete realizations  $\{-1, +1\}$  and  $\varrho(\pm 1) = 1/2$  one finds

$$g(y) = \frac{1+y}{2} \ln(1+y) + \frac{1-y}{2} \ln(1-y). \quad (2.218)$$

The convexity of  $f(t)$  and  $g(y)$  is easily verified for each case.

Armed with these preparations we are now able to introduce the central notion, the large deviation property.

We consider a sequence  $Y_N$  of random variables with densities  $\varrho_N(y)$ . We may think of them as the densities for  $Y_N = (X_1 + \dots + X_N)/N$ , where  $\{X_i\}$  are random variables with a density  $\varrho(x)$ . However, any other sequence is also possible.

We say that such a sequence has the large deviation property, if the densities for  $\varrho_N(y)$  obey

$$\varrho_N(y) = e^{-a_N S(y) + o(N)}, \quad (2.219)$$

with  $a_N \rightarrow N$  for  $N \rightarrow \infty$ . The residual term  $o(N)$  contains only contributions which increase sublinearly as a function of  $N$ . In the limit  $N \rightarrow \infty$  the probability of an event being in  $(y, y + dy)$  should therefore be arbitrarily small for almost all  $y$ . For large  $N$  a significant probability remains only for minima of  $S(y)$ .

The function  $S(y)$  therefore plays an essential role for the densities  $\varrho_N(y)$  for large  $N$ . In the so-called thermodynamic limit, i.e.  $N \rightarrow \infty$ , the probability  $\lim_{N \rightarrow \infty} \varrho_N(y)$  is different from zero only at the absolute minimum  $y_{\min}$  of the function  $S(y)$ .

We will see that in models of real statistical systems such a value  $y_{\min}$  corresponds to the equilibrium state and that the function  $S(y)$  corresponds to the negative of the entropy, and we know that the entropy assumes its maximum for an equilibrium state.

But we can already see the following: If the function  $S(y)$  assumes its absolute minimum at two (or more) values, one also obtains two (or more) possible equilibrium states. In this case one speaks of two phases. Which phase or which mixture of phases is realized depends on the initial conditions and/or boundary conditions.

If  $S(y)$  depends on a parameter and if for a certain value of this parameter the minimum splits into two minima, this value is called a critical point. The splitting is called a phase transition. Hence this phenomenon can already be described at this stage.

The determination of the function  $S(y)$  is, of course, of utmost importance. An example where this quantity is particularly easy to calculate is the following.

Let  $X, X_1, \dots$  be identical and independent random variables with a density  $\varrho(x)$ . Furthermore, let the free energy function,

$$f(t) = \ln \langle e^{tX} \rangle = \ln \left( \int dx e^{tx} \varrho(x) \right), \quad (2.220)$$

be finite for all  $t$ . Set

$$Y_N = \frac{1}{N} \sum_{i=1}^N X_i. \quad (2.221)$$

Under these conditions the sequence  $\varrho_{Y_N}(y)$  has the large deviation property. Indeed, (2.219) holds with  $a_N = N$ , and we find that

- The function  $S(y)$  is the Legendre transform of  $f(t)$ :

$$S(y) = \sup_t (ty - f(t)). \quad (2.222)$$

- The function  $S(y)$  is the negative relative entropy  $S[\varrho_y(x) \mid \varrho(x)]$ , where  $\varrho(x)$  is the density of  $X$  and  $\varrho_y(x)$  follows from the density of  $\varrho(x)$  after a shift of the expectation value to  $y$ .

If the realizations of  $X$  assume only discrete values in a finite set  $\{x_1, \dots, x_r\}$  with  $x_1 < \dots < x_r$ , then  $S(y)$  is finite and continuous in the interval  $(x_1, x_r)$ , while  $S(y) = \infty$  for  $y$  outside  $(x_1, x_r)$ .

For a proof of these statements we refer to the literature (Ellis 1985; Shwartz and Weiss 1995). However, we want to illustrate them for the above-mentioned examples.

- Let  $\varrho(x)$  be the normal distribution  $N(\mu, \sigma^2)$ , i.e.,  $f(t)$  is given by (2.208) and its Legendre transform by (2.216). As expected, one finds

$$S(y) = \frac{(y - \mu)^2}{2\sigma^2}. \quad (2.223)$$

The same result may be obtained by forming the negative relative entropy:

$$-S[\varrho_y(x) \mid \varrho(x)] = \int dx \varrho_y(x) \ln \left[ \frac{e^{-(x-y)^2/2\sigma^2}}{e^{-(x-\mu)^2/2\sigma^2}} \right] \quad (2.224)$$

$$\begin{aligned} &= \int dx \varrho_y(x) [(x-\mu)^2/2\sigma^2 - (x-y)^2/2\sigma^2] \\ &= \frac{(y-\mu)^2}{2\sigma^2}. \end{aligned} \quad (2.225)$$

- For large values of  $N$ , the mean value  $Y_N$  of  $N$  exponential random variables has a density (cf. (2.217))

$$\varrho_N(y) \propto \exp(-N(my - 1 - \ln my) + o(N)). \quad (2.226)$$

The sum  $Z = NY_N$  of  $N$  exponential random variables is also called a gamma distributed random variable. One obtains for its density

$$\varrho_Z(z) = \frac{(mz)^{N-1}}{(N-1)!} e^{-mz}, \quad (2.227)$$

which is in accordance with (2.226).

- For the discrete random variable with the possible realizations  $\{-1, 1\}$  and  $\varrho(\pm 1) = 1/2$  one finds according to (2.218)

$$\varrho_N(y) = \frac{1}{2^N} \sum_{\{x_i = \pm 1\}} \delta\left(y - \frac{1}{N} \sum_{i=1}^N x_i\right) \propto e^{-NS(y)}, \quad (2.228)$$

where

$$S(y) = \frac{1+y}{2} \ln(1+y) + \frac{1-y}{2} \ln(1-y). \quad (2.229)$$

One obtains the same results by forming the negative relative entropy:

$$\begin{aligned} -S[\varrho_y(x) \mid \varrho(x)] &= \varrho(1)(1+y) \ln \left[ \frac{(1+y)/2}{1/2} \right] \\ &\quad + \varrho(-1)(1-y) \ln \left[ \frac{(1-y)/2}{1/2} \right] \end{aligned} \quad (2.230)$$

$$= \frac{1+y}{2} \ln(1+y) + \frac{1-y}{2} \ln(1-y). \quad (2.231)$$

In this case we may use the Bernoulli distribution for an explicit calculation of  $q(y)$  and thus also  $S(y)$ . It gives us the probability that  $N$  realizations of  $X$  yield  $q$  times the value 1 and therefore  $y = (q - (N - q))/N = 2q/N - 1$ ,

$$q(y) = \binom{N}{q} \left(\frac{1}{2}\right)^q \left(\frac{1}{2}\right)^{(N-q)} = \binom{N}{\frac{N}{2}(1+y)} 2^{-N}. \quad (2.232)$$

Using Stirlings formula  $\ln N! = N(\ln N - 1) + o(N)$  we obtain

$$\begin{aligned} \ln q(y) &= N(\ln N - 1) - \frac{N}{2}(1+y) \ln \left(\frac{N}{2}(1+y)\right) \\ &\quad - \frac{N}{2}(1-y) \ln \left(\frac{N}{2}(1-y)\right) - N \ln 2 + o(N) \end{aligned} \quad (2.233)$$

$$= -N \left[ \frac{1+y}{2} \ln(1+y) + \frac{1-y}{2} \ln(1-y) \right] + o(N). \quad (2.234)$$

In the next chapter we will make use of the representation of the density given in (2.219).



Statistical Physics

An Advanced Approach with Applications

Honerkamp, J.

2012, XIV, 554 p., Hardcover

ISBN: 978-3-642-28683-4