Graduate Texts in Physics

# Graduate Texts in Physics

Graduate Texts in Physics publishes core learning/teaching material for graduate- and advanced-level undergraduate courses on topics of current and emerging fields within physics, both pure and applied. These textbooks serve students at the MS- or PhD-level and their instructors as comprehensive sources of principles, definitions, derivations, experiments and applications (as relevant) for their mastery and teaching, respectively. International in scope and relevance, the textbooks correspond to course syllabi sufficiently to serve as required reading. Their didactic style, comprehensiveness and coverage of fundamental material also make them suitable as introductions or references for scientists entering, or requiring timely knowledge of, a research field.

Please view available titles in *Graduate Texts in Physics* on series homepage
http://www.springer.com/series/8431/

Josef Honerkamp

# Statistical Physics

An Advanced Approach with Applications

Third Edition

With collaboration of Björn Schelter

With 116 Figures

Springer

Josef Honerkamp
Fakultät für Mathematatilk und Physik
Albert-Ludwigs-Universität Freiburg
Freiburg
Germany

Additional material to this book can be downloaded from http://extras.springer.com

ISSN 1868-4513 ISSN 1868-4521 (electronic)
ISBN 978-3-642-28683-4 ISBN 978-3-642-28684-1 (eBook)
DOI 10.1007/978-3-642-28684-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012939644

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface to the Third Edition

I was very pleased that the publisher asked me for a further edition of this book on Statistical Physics. It deals not only with the topic of Statistical Mechanics but also with Data Analysis and it gives an thorough introduction into mathematics of random variables. Insofar it occupies a special position among physics textbooks on Statistical Physics.

Since meanwhile the physics of complex systems has received a greater importance, it was therefore obvious to expand that part of the book which deals with the Data Analysis. In particular the concept of Granger causality and the Unscented Kalman filter are discussed.

I thank my colleague Jens Timmer for some advice in this regard and I am very grateful to Dr. Björn Schelter for the formulation of the relevant sections.

Emmendingen
January 2012

Josef Honerkamp

# Preface to the Second Edition

The first edition of this book was appreciated very well, though, for some physicists, it seems to contain an unconventional collection of topics. But more and more complex systems are studied in physics, chemistry and biology so that knowledge and experience in data analysis and model building becomes essential. Physics students should learn more in statistical physics than only statistical mechanics, they should also become acquainted with the mathematical background of numerical methods for data analysis.

In this second edition I eliminated some misprints and some conceptual shortcomings, and I added some topics which are very useful and central for data analysis and model building: the Monte Carlo Markov Chain method, the estimation of realizations and of parameters in hidden systems such as hidden Markov systems and state space models, and I introduced also some basic facts about testing of statistical hypotheses and about classification methods. I hope that this enlargement will meet some needs or, at least, will arouse interest in the reader.

I would like to thank many colleagues and students for discussions and clarification, especially H.P. Breuer, J. Timmer, and H. Voss.

Emmendingen J. Honerkamp

# Contents

**Part II Analysis of Statistical Systems**

# Part I
# Modeling of Statistical Systems

# 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 $Y_N$, $N = 2, \ldots$ 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 $N \to \infty$ 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.

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_2, © Springer-Verlag Berlin Heidelberg 2012

## 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 $A \cap B = \overline{\bar{A} \cup \bar{B}}$.

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 $\{A_1, \ldots, A_N\}$ is given such that $A_i \cap A_j = \emptyset$ for $i \neq j$, and

$$A_1 \cup \cdots \cup A_N = \Omega. \tag{2.1}$$

  Such a set is called complete and disjoint.

*Examples.*

- When we throw a die the outcomes are the elementary events $\{i\}$, $i = 1, \ldots, 6$. 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

$$\{1\} \cup \cdots \cup \{6\} = \Omega \equiv \{1, \ldots, 6\}. \tag{2.2}$$

- 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 $A_i, i = 1, \ldots$ 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

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

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:

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

*Examples.* For the throw of a die, for example, we take $\mathcal{P}(\{i\}) = 1/6$ for $i = 1, \ldots, 6$. 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:

$$\lambda \to +\infty: \qquad P(\lambda) \to \mathcal{P}(\Omega) = 1, \tag{2.5a}$$

$$\lambda \to -\infty: \qquad P(\lambda) \to \mathcal{P}(\emptyset) = 0. \tag{2.5b}$$

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

$$\varrho(\lambda) = \frac{\mathrm{d}P(\lambda)}{\mathrm{d}\lambda}, \tag{2.6}$$

from which we get

$$P(\lambda) = \int_{-\infty}^{\lambda} \varrho(x)\,\mathrm{d}x. \tag{2.7}$$

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

$$\mathcal{P}(\{x_1 \leq x \leq x_2\}) = \int_{x_1}^{x_2} \varrho(x)\,\mathrm{d}x. \tag{2.8}$$

In particular, we have

$$\int_{-\infty}^{+\infty} \varrho(x)\,\mathrm{d}x = 1, \tag{2.9}$$

and also

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

If $\mathrm{d}x$ is small enough, $\varrho(x)\,\mathrm{d}x$ is the probability of the event $(x, x+\mathrm{d}x)$. In other words it is the probability of obtaining a value $x' \in (x,\ x + \mathrm{d}x)$ 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

$$\mathcal{P}(\{x|x \le \lambda\}) = P_X(\lambda) \tag{2.11}$$

and

$$\varrho_X(x) = \frac{\mathrm{d}P_X(x)}{\mathrm{d}x}. \tag{2.12}$$

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

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

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

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

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

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

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 $k = 0, 1, \ldots, n$ and the (discrete) probability density

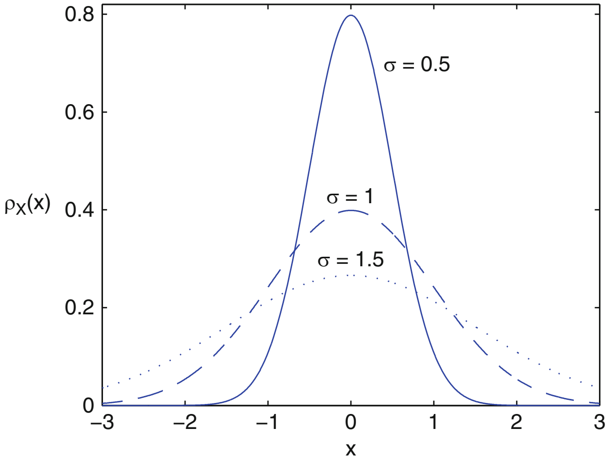


**Fig. 2.1** Density function of the Gaussian distribution for various values of $\sigma$

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

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

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

$$= \binom{n}{k} \Big(\frac{V_1}{V_2}\Big)^k \Big(\frac{V_2}{V_1+V_2}\Big)^n. \tag{2.18}$$

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 $k = 0, 1, \ldots$ and the discrete density

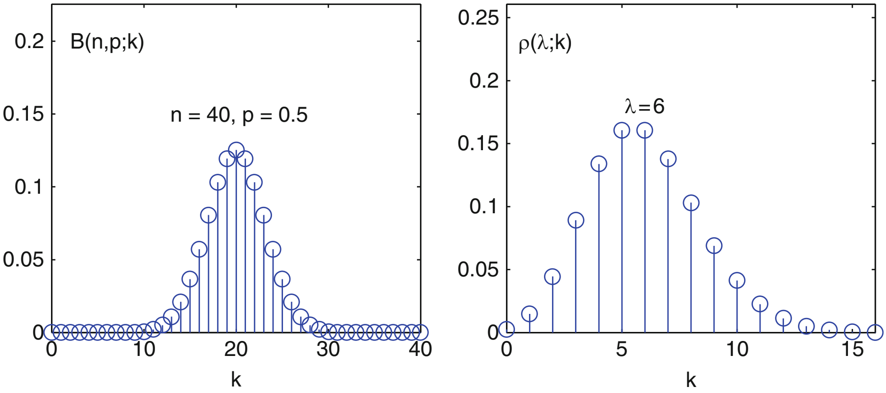


**Fig. 2.2** Densities of the binomial distribution (*left*) and the Poisson distribution (*right*)

$$\varrho(\lambda; k) = \frac{\lambda^k}{k!}\, \mathrm{e}^{-\lambda}, \qquad k = 0, 1, \ldots . \tag{2.19}$$

The density $\varrho(\lambda; k)$ results from $B(n, p; k)$ in the limit $p \to 0$, $n \to \infty$ and $p\,n = \lambda = \text{const.}$ $K$ 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 $\mathrm{d}t$ is just $p = \lambda\, \mathrm{d}t/T$. To see this, we divide the time interval $(0, T)$ into $n$ equal segments of length $\mathrm{d}t = T/n$. Then

$$p = \frac{\lambda}{T} \frac{T}{n} = \frac{\lambda}{n}, \tag{2.20}$$

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 $n \to \infty$ one takes $\mathrm{d}t \to 0$ and $p \to 0$ such that $p\,n = \lambda$ 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 $\lambda/T = \alpha\, \mathrm{s}^{-1}$. The probability that no decay occurs within a time interval of $T$ seconds is

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

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

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

## 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 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 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 $\varrho(x_1, \ldots, x_n)$ on $\mathbb{R}^n$. The probability that a realization of $X_1$ yields a value in the interval $(x_1, x_1 + \mathrm{d}x_1)$, and, similarly, that realizations of $X_2, \ldots, X_n$ yield values in the corresponding intervals is $\varrho(x_1, \ldots, x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_n$. Two examples will help to clarify this:

In a classical description, the momentum $\boldsymbol{p}$ 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))

$$\varrho(\boldsymbol{p}) = \frac{1}{A} \exp\left(-\beta \frac{\boldsymbol{p}^2}{2m}\right), \qquad \text{where} \qquad \beta = \frac{1}{k_\mathrm{B} T}. \tag{2.23}$$

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

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

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 $\mathsf{A}$ 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:

$$(\boldsymbol{p}, \boldsymbol{q}) = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N). \tag{2.25}$$

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))

$$\varrho(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{A} \mathrm{e}^{-\beta H(\boldsymbol{p}, \boldsymbol{q})}, \qquad \beta = \frac{1}{k_\mathrm{B} T}, \tag{2.26}$$

where again $T$ denotes the temperature, $k_{\mathrm{B}}$ Boltzmann's constant, and $A$ a normalization constant. Furthermore, $H(\boldsymbol{p}, \boldsymbol{q})$ 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,

$$\varrho'(x_1) = \int \mathrm{d}x_2 \ldots \mathrm{d}x_n \, \varrho(x_1, x_2, \ldots, x_n), \tag{2.27}$$

then $\varrho'(x_1)\,\mathrm{d}x_1$ is the probability of finding $X_1$ in the interval $[x_1, x_1 + \mathrm{d}x_1]$, irrespective of the outcome for the variables $X_2, \ldots, X_n$.

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

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

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

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

### 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

$$\varrho(\boldsymbol{p}, \boldsymbol{q} \mid T, V, N), \tag{2.30}$$

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

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

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

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

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.,

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

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

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

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

$$\bigcup_{i=1}^{N} B_i = \Omega \tag{2.35}$$

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

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

This should be compared with the formula

$$\varrho_{X_1}(x_1) = \int \mathrm{d}x_2 \, \varrho_{X_1,X_2}(x_1, x_2) = \int \mathrm{d}x_2 \, \varrho(x_1 \mid x_2) \, \varrho_{X_2}(x_2). \tag{2.37}$$

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.,

$$\mathcal{P}(A_1 \mid A_2) = \mathcal{P}(A_1). \tag{2.38}$$

In particular we have then also:

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

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

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

### Bayes' Theorem

From

$$\mathcal{P}(A, B) = \mathcal{P}(A \mid B)\mathcal{P}(B) = \mathcal{P}(B \mid A)\mathcal{P}(A) \tag{2.41}$$

it follows that

$$\mathcal{P}(B \mid A) = \frac{\mathcal{P}(A \mid B)\mathcal{P}(B)}{\mathcal{P}(A)}. \tag{2.42}$$

Hence $\mathcal{P}(B \mid A)$ can be determined from $\mathcal{P}(A \mid B)$, 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 $\mathcal{P}(B \mid A)$. 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

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

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

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

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

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

Hence, the results depend sensitively on the probability $\mathcal{P}(A)$, i.e., on the overall frequency of the illness and on $\mathcal{P}(B \mid \bar{A})$, 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

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

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

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

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

Second argument (using Bayes' theorem):

We have

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

Now

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

$$\mathcal{P}(M3 \mid A2) = 1, \tag{2.48b}$$

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

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

and therefore

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

Similarly, one obtains:

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

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

$$\langle H(\boldsymbol{X})\rangle = \int \mathrm{d}^n x\, H(\boldsymbol{x})\, \varrho(\boldsymbol{x}). \tag{2.51}$$

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

A particularly important moment is

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

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

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

- When $H(\boldsymbol{x})$ 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

$$\langle X^m \rangle = \int \mathrm{d}x\, x^m\, \varrho(x). \tag{2.53}$$

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

$$\mathrm{Var}(X) = \langle \big(X - \langle X \rangle\big)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2 \tag{2.54a}$$

$$= \int \mathrm{d}x\, (x - \langle X \rangle)^2\, \varrho(x). \tag{2.54b}$$

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. $\sqrt{\mathrm{Var}(X)}$ is also called the standard deviation.

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

$$\langle X \rangle = \mu, \tag{2.55}$$

and

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} \mathrm{d}x\, (x - \mu)^2\, \varrho(\mu, \sigma^2; x) = \sigma^2, \tag{2.56}$$

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:

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

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

$$\langle X_i X_j \rangle = \int \mathrm{d}^n x\, x_i x_j\, \varrho(x_1, \ldots, x_n). \tag{2.58}$$

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

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

$$= \int \mathrm{d}^n x\, (x_i - \mu_i)\, (x_j - \mu_j) \varrho(x_1, \ldots, x_n). \tag{2.59b}$$

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

$$\langle \boldsymbol{X} \rangle = \boldsymbol{\mu}, \tag{2.60}$$

and

$$\langle (\boldsymbol{X} - \boldsymbol{\mu})_i (\boldsymbol{X} - \boldsymbol{\mu})_j \rangle = \mathsf{A}_{ij}. \tag{2.61}$$

Hence, the matrix $\mathsf{A}$ 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:

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

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

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

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 $Y_j$ are both normally distributed, is this conclusion correct, because in this case the covariance matrix, and hence also the matrix $\mathsf{A}$ 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

$$G(k) \equiv \langle \mathrm{e}^{\mathrm{i}kX} \rangle = \int \mathrm{d}x\, \mathrm{e}^{\mathrm{i}kx}\, \varrho_X(x), \tag{2.64}$$

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:

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

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

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

with the cumulants

$$\kappa_1 = \mu = \langle X \rangle \tag{2.67a}$$

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

$$\kappa_3 = \langle X^3 \rangle - 3\langle X^2 \rangle \langle X \rangle + 2\langle X \rangle^3, \tag{2.67c}$$

and so on for higher cumulants.

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

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

i.e., one obtains for the Gaussian distribution

$$\kappa_1 = \mu \tag{2.69a}$$

$$\kappa_2 = \sigma^2, \tag{2.69b}$$

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),

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

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

$$G(k) = \mathrm{e}^{\mathrm{i}ka - |k|\gamma}. \tag{2.71}$$

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 $\varrho(\boldsymbol{r})$ we know the electric dipole moment

$$\boldsymbol{p} = \int \mathrm{d}^3 r \, \boldsymbol{r} \, \varrho(\boldsymbol{r}), \tag{2.72}$$

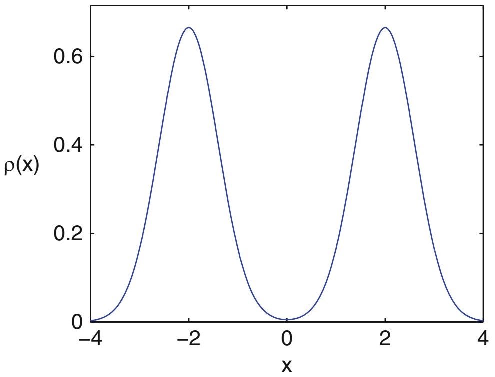


**Fig. 2.3** A density with two maxima (also called a bimodal distribution). The expectation value is zero, but $x = 0$ is never assumed

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

$$I_{ij} = \int \mathrm{d}^3 r\, m(\boldsymbol{r}) \left(-r_i r_j + \delta_{ij} \boldsymbol{r}^2\right). \tag{2.73}$$

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

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

and

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

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

$$\mathrm{Var}(K) = \lambda. \tag{2.76}$$

For the binomial distribution one obtains

$$\langle K \rangle = n\, p, \tag{2.77a}$$

$$\mathrm{Var}(K) = n\, p\, (1 - p). \tag{2.77b}$$

### 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

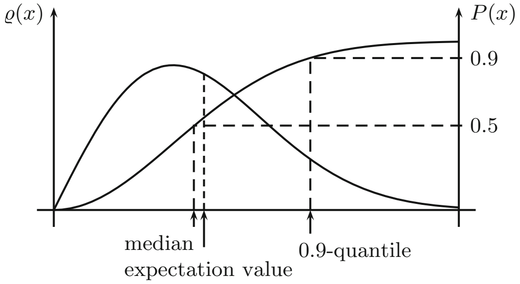


**Fig. 2.4** Median, expectation value and 0.9-quantile of a probability distribution $P(x)$, also shown for the density $\varrho(x)$

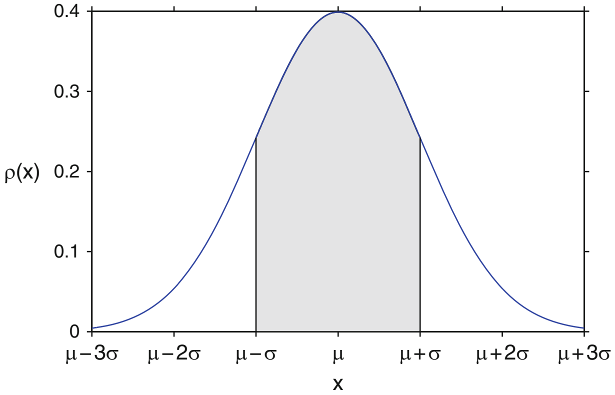


**Fig. 2.5** The area of the shaded region is 0.6827 times the total area under the curve, which is 1. The probability that the random variable $X$ assumes a value in the interval $[\mu - \sigma, \mu + \sigma]$ is therefore 0.6827 or 68.27%

$$P(x_\alpha) = \int_{-\infty}^{x_\alpha} \mathrm{d}x\, \varrho_X(x) = \alpha. \tag{2.78}$$

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:

$$\begin{array}{cc} \text{Interval} & \text{Percentage of values} \\ (\mu-\sigma,\mu+\sigma) & 68.27\% \\ (\mu-2\sigma,\mu+2\sigma) & 95.45\% \\ (\mu-3\sigma,\mu+3\sigma) & 99.73\%. \end{array} \tag{2.79}$$

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

$$A_1 \cup A_2 \cup \cdots \cup A_N = \Omega. \tag{2.80}$$

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

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

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 $k_{\mathrm{B}}$.

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 $0 \le \mathcal{P}(A_i) \le 1$ we always have $S \ge 0$.
- 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

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

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

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

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.,

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

we have

$$S = \ln N. \tag{2.85}$$

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

$$S[\varrho_X] = -k \int \mathrm{d}x \; \varrho_X(x) \, \ln\left(\frac{\varrho_X(x)}{\varrho_0}\right). \tag{2.86}$$

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 $1/\mathrm{d}x$, 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 $k = 1, \varrho_0 = 1$):

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

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

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

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

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.,

$$S[p|q] \leq 0. \tag{2.90}$$

This is easy to see: We use the inequality

$$\ln z \leq z - 1 \tag{2.91}$$

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

$$-\int \mathrm{d}x\, p(x)\, \ln\left(\frac{p(x)}{q(x)}\right) \leq \int \mathrm{d}x\, \big(q(x) - p(x)\big). \tag{2.92}$$

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 $k = 1$)

$$S[p|q] = \int \mathrm{d}x\, \big(p(x) - q(x)\big) - \int \mathrm{d}x\, p(x)\, \ln\left(\frac{p(x)}{q(x)}\right). \tag{2.93}$$

### 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

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

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

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

a short calculation yields

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

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

$$S_{12} \leq S_1 + S_2. \tag{2.97}$$

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

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

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

$$\langle g_i(X)\rangle \equiv \int \mathrm{d}x\; g_i(x)\, \varrho(x) = \eta_i, \qquad i = 1, \ldots, n. \tag{2.99}$$

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

$$\varrho(x) = \frac{1}{A}\, \mathrm{e}^{-\lambda_1 g_1(x) - \ldots - \lambda_n g_n(x)}. \tag{2.100}$$

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 $k = 1$)

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

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)

$$S[\varphi|\varrho] \leq 0, \tag{2.102}$$

and therefore

$$S[\varphi] = -\int \mathrm{d}x \, \varphi(x) \, \ln\left(\frac{\varphi(x)}{\varrho_0}\right) \tag{2.103}$$

$$\leq -\int \mathrm{d}x \, \varphi(x) \, \ln\left(\frac{\varrho(x)}{\varrho_0}\right) \tag{2.104}$$

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

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

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

$$\langle X \rangle = \eta. \tag{2.107}$$

We immediately find this density as

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

The normalization factor $A$ is given by

$$A = \int_0^\infty \mathrm{d}x \, \mathrm{e}^{-\lambda x} = \frac{1}{\lambda}, \tag{2.109}$$

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

$$\eta = \langle X \rangle = \int_0^\infty \mathrm{d}x \, x \, \lambda \, \mathrm{e}^{-\lambda x} \tag{2.110}$$

$$= -\lambda \frac{\partial}{\partial \lambda} \int_0^\infty \mathrm{d}x \, \mathrm{e}^{-\lambda x} = \frac{1}{\lambda}. \tag{2.111}$$

Therefore

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

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

$$\langle H(\boldsymbol{q}, \boldsymbol{p})\rangle = E, \tag{2.113}$$

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

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

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

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

In particular, we find

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

$\lambda$ follows from the supplementary condition:

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

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

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

and therefore, using (2.116),

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

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

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

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

$$Z = \alpha\, X, \tag{2.121}$$

and find

$$\varrho_Z(z) = \int \mathrm{d}x\, \delta(z-\alpha x)\, \varrho_X(x) = \frac{1}{|\alpha|}\varrho_X(\frac{z}{\alpha}). \tag{2.122}$$

Obviously

$$\langle Z\rangle = \alpha\langle X\rangle, \tag{2.123a}$$

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

because

$$\langle Z\rangle = \int \mathrm{d}z\, z\, \varrho_Z(z) = \int \mathrm{d}z\, z \int \mathrm{d}x\, \delta(z-\alpha x)\, \varrho_X(x) \tag{2.124}$$

$$= \alpha \int \mathrm{d}x\, x\, \varrho_X(x) = \alpha\langle X\rangle. \tag{2.125}$$

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

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

$$Z = f(X). \tag{2.126}$$

One obtains for the density function

$$\varrho_Z(z) = \int \mathrm{d}x\, \varrho_X(x)\delta(z - f(x)) \tag{2.127}$$

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

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 $x_i(z), i = 1, \ldots$.

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

$$1 = \int \mathrm{d}x\ \varrho_X(x) = \int \mathrm{d}z\ \left|\frac{\mathrm{d}x}{\mathrm{d}z}\right|\ \varrho_X(x(z)) = \int \mathrm{d}z\ \varrho_Z(z), \tag{2.129}$$

and we find, in agreement with (2.128),

$$\varrho_Z(z) = \varrho_X(x(z))\ \left|\frac{\mathrm{d}x}{\mathrm{d}z}\right|. \tag{2.130}$$

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

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

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

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

*Examples.*

(a) Let

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

With $\mathrm{d}z/\mathrm{d}x = -1/x$, and hence $|\mathrm{d}x/\mathrm{d}z| = |x| = \mathrm{e}^{-z}$, one obtains

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

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

(b) Let

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

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

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

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

$$Z = X_1 + X_2. \tag{2.137}$$

Then

$$\varrho_Z(z) = \int \mathrm{d}x_1 \, \mathrm{d}x_2 \, \delta(z - x_1 - x_2) \, \varrho(x_1, x_2) \tag{2.138}$$

$$= \int \mathrm{d}x_1 \, \varrho(x_1, z - x_1). \tag{2.139}$$

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

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

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

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

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

$$\langle Z \rangle = \langle X_1 \rangle + \langle X_2 \rangle, \tag{2.142}$$

$$\mathrm{Var}(Z) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2). \tag{2.143}$$

The first relation follows from

$$\langle Z \rangle = \int \mathrm{d}z \int \mathrm{d}x_1\, dx_2\, z\, \delta(z - x_1 - x_2)\, \varrho_1(x_1)\, \varrho_2(x_2) \tag{2.144}$$

$$= \int \mathrm{d}x_1\, \mathrm{d}x_2\, (x_1 + x_2)\, \varrho_1(x_1)\, \varrho_2(x_2) \tag{2.145}$$

$$= \langle X_1 \rangle + \langle X_2 \rangle. \tag{2.146}$$

The equation for $\mathrm{Var}(Z)$ 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

$$G_X(k) = \exp\left( \mathrm{i}k\kappa_1(X) - \frac{1}{2}k^2\kappa_2(X) - \ldots \right) \tag{2.147}$$

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 $\kappa_2(X) = \mathrm{Var}(X)$, (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

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

**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

$$Z = X_1\, X_2. \tag{2.149}$$

Then

$$\varrho_Z(z) = \int \mathrm{d}x_1\, \mathrm{d}x_2\, \delta(z - x_1 x_2)\, \varrho_{X_1}(x_1) \varrho_{X_2}(x_2) \tag{2.150}$$

$$= \int \mathrm{d}x_1\, \varrho_{X_1}(x_1)\, \frac{1}{|x_1|}\, \varrho_{X_2}\left(\frac{z}{x_1}\right). \tag{2.151}$$

### 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 $X_1, \ldots, X_n$ with standard normal distributions and set

$$Z = X_1^2 + \cdots + X_n^2. \tag{2.152}$$

The density distribution of $Z$ is given by

$$\varrho_Z(z) = \int \mathrm{d}x_1 \ldots \mathrm{d}x_n \, \delta(z - x_1^2 - \cdots - x_n^2) \, \varrho(x_1, \ldots, x_n), \tag{2.153}$$

with

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

We obtain

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

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

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

$$\langle \chi_n^2 \rangle = n, \tag{2.156a}$$

$$\mathrm{Var}(\chi_n^2) = 2n. \tag{2.156b}$$

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

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

The three components $v_i$ of the velocities of molecules in a gas at temperature $T$ are normally distributed with mean 0 and variance $\sigma^2 = k_\mathrm{B}T/m$ (cf. Sect. 2.2). Here, $m$ denotes the mass of a molecule and $k_\mathrm{B}$ 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

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

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

$$\left\langle \frac{m}{2} \boldsymbol{v}^2 \right\rangle = \frac{m}{2} \sigma^2 \left\langle \chi_3^2 \right\rangle = \frac{m}{2} \frac{k_\mathrm{B} T}{m} \cdot 3 = \frac{3}{2} k_\mathrm{B} T. \tag{2.159}$$

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

$$Z = \frac{Y_k / k}{Z_q / q} \tag{2.160}$$

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

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

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

$$T_q = \frac{Y}{\sqrt{Z_q / q}} \tag{2.162}$$

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

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

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 $z \to 0$, when $n$ exceeds 1 or 2, or when $k$ exceeds the value 2. For $z \to \infty$ 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.

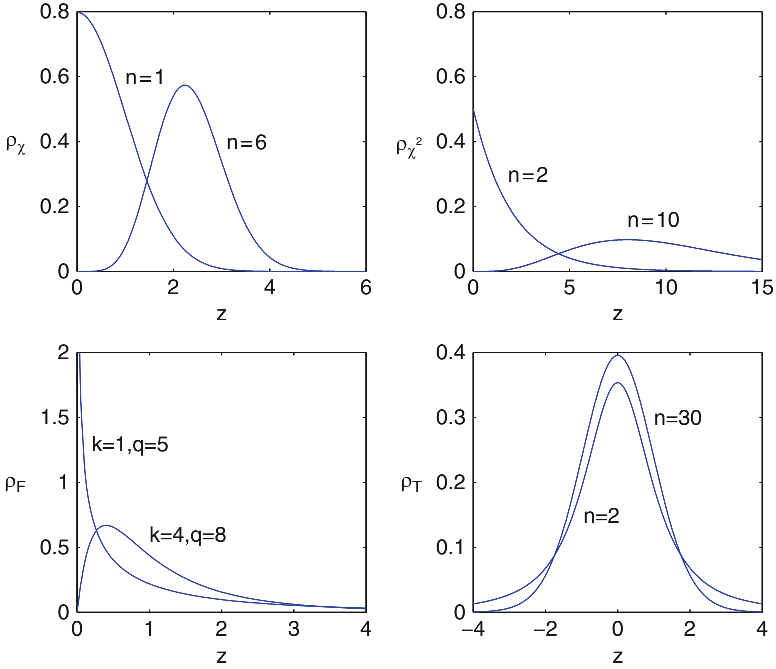


**Fig. 2.6** Density functions of the $\chi$-distribution (*upper left*), the $\chi^2$-distribution (*upper right*), the $F$-distribution (*lower left*) and the $t$-distribution (*lower right*)

### *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 $N \to \infty$. 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 $X_i, i = 1, \ldots, N$, be independent and identically distributed random variables. All cumulants shall exist and let

$$\langle X_i \rangle = 0, \tag{2.164a}$$

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

We set

$$Z_N = \frac{1}{\sqrt{N}}\big(X_1 + \cdots + X_N\big). \tag{2.165}$$

If follows that

$$\langle Z_N \rangle = 0, \tag{2.166a}$$

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

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

Thus for $N \to \infty$ 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 $\varrho(x) = 1$ for $x \in [0, 1]$, and

$$\langle X \rangle = \frac{1}{2}, \tag{2.167a}$$

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

Hence

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

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

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

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 $X, X_1, \ldots, X_N$, all having the same probability density. All moments and all cumulants shall exist. We set

$$Z_N = \frac{1}{N}\big(X_1 + \cdots + X_N\big). \tag{2.170}$$

Then

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

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

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

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

$$z_N = \frac{1}{N}\big(x_1 + \cdots + x_N\big). \tag{2.172}$$

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 $\mathrm{Var}(X)/N$. 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 $N \to \infty$ 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 $X, X_1, \ldots, X_N$ be independent and identically distributed random variables with a density $\varrho(x)$, and, furthermore, let

$$S_N = X_1 + \ldots + X_N \ . \tag{2.173}$$

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,

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

$$G(k) = \left\langle \mathrm{e}^{\mathrm{i}kX} \right\rangle = \mathrm{e}^{\mathrm{i}k\mu - |k|\gamma}, \quad \gamma > 0 \tag{2.175}$$

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

$$G_{S_N} = \mathrm{e}^{\mathrm{i}Nk\mu - N|k|\gamma}, \tag{2.176}$$

and therefore

$$Y = \frac{1}{N}(X_1 + \ldots + X_N) \tag{2.177}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$x = \gamma \tan r + \mu. \tag{2.184}$$

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

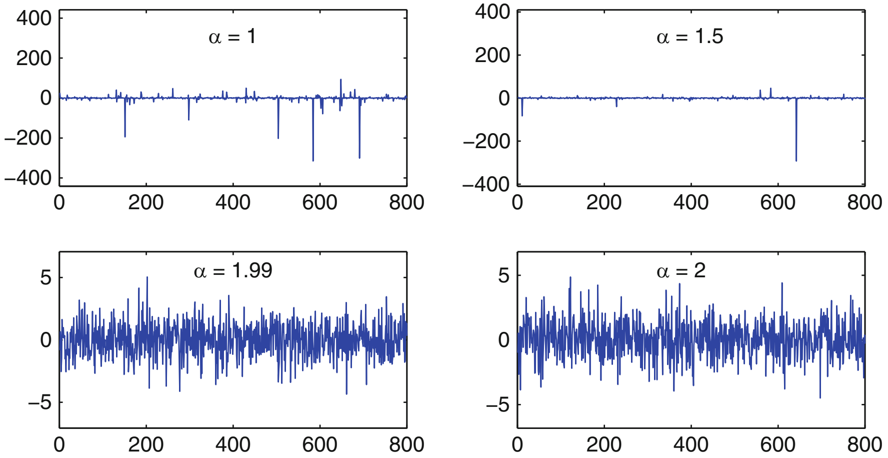


**Fig. 2.7** A series of realizations of a random variable with a stable density for four different values of the index $\alpha$

### 2.6.2 *The Renormalization Transformation*

There is a further way to characterize stable distributions: Let $\boldsymbol{X} = \{X_i\}_{i=-\infty}^{\infty}$ 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

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

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 familiy 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 $\boldsymbol{X} = \{X_i\}_{i=-\infty}^{\infty}$ is a fixed point of this group of transformations if the $X_i'$ resulting from $T_n \boldsymbol{X}$ 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

$$(T_2\varrho)(x) = \varrho_{X'}(x) = 2^{1/\alpha} \int \mathrm{d}y\, \varrho(2^{1/\alpha}x - y)\varrho(y). \tag{2.186}$$

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

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

with

$$(DT_2\eta)(x) = 2\,2^{1/\alpha} \int \mathrm{d}y\, \varrho^*(2^{1/\alpha}x - y)\eta(y). \tag{2.188}$$

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

$$(DT_2\phi_n)(x) = \lambda_n\phi_n(x). \tag{2.189}$$

Obviously, $\varrho^*(x)$ itself is an eigenfuction 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.

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

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

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

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

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

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)

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

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

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

For a more general distribution we also have

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

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

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

where

$$v_n' = \lambda_n v_n, \tag{2.197}$$

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 $\{\phi_n\}$,

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

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,

$$F[\varrho(x), t] = \ln\left(\int \mathrm{d}x\, \varrho(x) \mathrm{e}^{\mathrm{i}tx}\right), \tag{2.199}$$

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

$$\begin{aligned} F[\varrho'(x), t] &= \ln\left(\int \mathrm{d}x\, 2^{1/\alpha} \int \mathrm{d}y\, \varrho(2^{1/\alpha}x - y)\varrho(y)\mathrm{e}^{\mathrm{i}tx}\right) && (2.200) \\ &= 2F\left[\varrho(x), \frac{t}{2^{1/\alpha}}\right]. && (2.201) \end{aligned}$$

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 $F(v_1, \ldots, t)$ of these scale parameters and the variable $t$. Thus

$$F(v_1', v_2', v_3', \ldots, t) = 2F\left(v_1, v_2, v_3, \ldots, \frac{t}{2^{1/\alpha}}\right), \tag{2.202}$$

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

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

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

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

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:

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

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:

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

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 $N$ $(N = 2, \ldots)$ 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

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

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 $f(\mathrm{i}k) = \ln G(k)$ 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

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

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

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

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

$$f(t) = \ln(\cosh t). \tag{2.210}$$

**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

$$g(y) = \sup_t (ty - f(t)). \tag{2.211}$$

$g(y)$ 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

$$y = f'(t) \tag{2.212}$$

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

$$g(y) = t(y)y - f(t(y)) \tag{2.213}$$

and

$$\mathrm{d}g = y\,\mathrm{d}t + t\,\mathrm{d}y - f'(t)\,\mathrm{d}t = t\,\mathrm{d}y, \quad \text{i.e. also} \quad g'(y) = t(y). \tag{2.214}$$

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

$$H(p, q) = \sup_{\dot{q}} \Big( p\,\dot{q} - L(\dot{q}, q) \Big). \tag{2.215}$$

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

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

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

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

- For the exponential distribution follows from (2.209)

$$g(y) = my - 1 - \ln my, \tag{2.217}$$

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

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

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 $Y_N = (X_1 + \ldots + X_N)/N$, 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

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

with $a_N \to N$ for $N \to \infty$. The residual term $o(N)$ contains only contributions which increase sublinearly as a function of $N$. In the limit $N \to \infty$ the probability of an event being in $(y, y + \mathrm{d}y)$ 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. $N \to \infty$, the probability $\lim_{N\to\infty} \varrho_N(y)$ 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 $X, X_1, \ldots$ be identical and independent random variables with a density $\varrho(x)$. Furthermore, let the free energy function,

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

be finite for all $t$. Set

$$Y_N = \frac{1}{N} \sum_{i=1}^{N} X_i . \tag{2.221}$$

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 $f(t)$:

$$S(y) = \sup_t \, (ty - f(t)). \tag{2.222}$$

- 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 $\{x_1, \ldots, x_r\}$ with $x_1 < \ldots < x_r$, 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

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

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

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

$$= \int \mathrm{d}x\ \varrho_y(x) \left[ (x-\mu)^2/2\sigma^2 - (x-y)^2/2\sigma^2 \right]$$

$$= \frac{(y-\mu)^2}{2\sigma^2}. \tag{2.225}$$

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

$$\varrho_N(y) \propto \exp\left(-N(my - 1 - \ln my) + o(N)\right). \tag{2.226}$$

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

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

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)

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

where

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

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

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

$$+\varrho(-1)(1-y) \ln \left[ \frac{(1-y)/2}{1/2} \right] \tag{2.230}$$

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

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

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

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

$$\begin{aligned} \ln \varrho(y) = {} & N(\ln N - 1) - \frac{N}{2}(1+y) \ln \left(\frac{N}{2}(1+y)\right) \\ & - \frac{N}{2}(1-y) \ln \left(\frac{N}{2}(1-y)\right) - N \ln 2 + o(N) && (2.233) \\ = {} & -N \left[\frac{1+y}{2} \ln (1+y) + \frac{1-y}{2} \ln (1-y)\right] + o(N). && (2.234) \end{aligned}$$

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

# Chapter 3
# Random Variables in State Space: Classical Statistical Mechanics of Fluids

Using the concepts of probability theory and statistics introduced in Chap. 2, the problem of statistical mechanics can be stated as follows: Consider a system of $N$ particles with momenta and positions $\boldsymbol{p}_i, \boldsymbol{q}_i$, $i = 1, \ldots, N$. At each instant, the microscopic state $x = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ may be considered as a realization of a random variable in the $6N$-dimensional phase space. For this random variable the density function has to be determined under various external, macroscopically given conditions. Furthermore, we will consider only systems in a stationary state so that the density function may be assumed to be time independent.

In Sects. 3.1 and 3.2 the density function for different external, macroscopically fixed conditions will be formulated. The microcanonical, canonical and grand canonical system is introduced.

Section 3.3 introduces the different thermodynamic potentials and relates them to one another.

In Sect. 3.4, the susceptibilities, which characterize the response of a material to external changes, will be defined. Their relation to the variances and covariances of corresponding quantities of the macroscopic system will turn out to be a special case of the fluctuation–dissipation theorem which will be discussed in a more general form in Chap. 5. Relations among susceptibilities are discussed as well as the Maxwell relations.

While the ideal classical gas will already be discussed when introducing the various densities, the subsequent sections deal with the case where the interaction between the particles can no longer be neglected. Here equations of state for liquids can also be formulated. As gases and liquids together are referred to as ‘fluids’, one speaks in this context also of the theory of fluids.

First, in Sect. 3.5, the law of equipartition, useful for many approximate considerations, will be proven. Section 3.6 focuses on the particle density $N(\boldsymbol{r})$; the corresponding covariance function $\mathrm{Cov}(N(\boldsymbol{r})N(\boldsymbol{r}'))$ is the natural quantity in which the interaction between the particles of the statistical system becomes apparent. Therefore, this quantity, or equivalently the radial distribution function derived from it, is the central quantity in the theory of fluids. Its relation to experimentally

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_3, © Springer-Verlag Berlin Heidelberg 2012

accessible quantities and to the structure functions will be discussed as will its use in formulating the equations of state and the virial theorem.

Section 3.7 is dedicated to approximation methods. The virial expansion will be presented and its results compared with those of molecular dynamics. Ways of formulating integral equations for the radial distribution function will be discussed briefly, perturbation expansion will be introduced, and a generalized equation of state will be formulated, from which in the following section, Sect. 3.8, the equation of state for the van der Waals gas will be derived. This will allow us to discuss phase transitions in fluids, corresponding states, and critical behavior.

## 3.1 The Microcanonical System

We consider a closed system of $N$ particles in a volume $V$. As the system is closed the total internal energy has a fixed value $E$. (Here the internal energy of the system results only from the internal degrees of freedom.) A system where the macroscopic quantities $E$, $N$, and $V$ are given will be called a microcanonical system and the probability density for the microstate $x = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$, which we are going to determine in the following, is the microcanonical density.

The dynamics of the individual particles is given by the Hamiltonian function

$$H(x) \equiv H(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) \tag{3.1}$$

and as a constraint for $x$ one thus obtains

$$H(x) = E \ . \tag{3.2}$$

Fixing the energy to a sharp value is, of course, only an idealization. It is mathematically simpler and physically more realistic to determine the probability density in a domain

$$E \le H(x) \le E + \varDelta E \ . \tag{3.3}$$

It will turn out that the value of $\varDelta E$ has no actual significance.

As we do not have any information about this density distribution, it is most plausible to assume that all microstates within this domain are equally probable. This corresponds to a choice for the density in accordance with the maximum entropy principle (Sect. 2.4). It has the form

$$\varrho(x|E, V, N) = \begin{cases} \dfrac{1}{A} & \text{for } \{x|E \le H(x) \le E + \varDelta E\} \\ 0 & \text{otherwise} \ . \end{cases} \tag{3.4}$$

The denominator $A$ follows from the normalization condition:

$$A = \int \mathrm{d}x \ \left(\Theta\left(E + \varDelta E - H(x)\right) - \Theta\left(E - H(x)\right)\right) \ , \tag{3.5}$$

where $\Theta(x)$ is the Heaviside function,

$$\Theta(x) = \begin{cases} 1 & \text{for } x > 0 \,, \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

$A$ has the dimension $(pq)^{3N} = (\text{Js})^{3N}$.

For the entropy of the random variable belonging to the microstate $x$, we then obtain according to Sect. 2.4, (2.86), setting the constant $k$ equal to the Boltzmann constant $k_\text{B}$:

$$S = k_\text{B} \ln (A\varrho_0), \tag{3.7}$$

and now we have to decide about the value of $\varrho_0$. This cannot be done without anticipating some knowledge from quantum mechanics. The uncertainty relation of quantum mechanics suggests that one introduces cells with a volume proportional to $h^{3N}$ in phase space because the product of the standard deviations $\sigma_q \sigma_p$ for any component is of order $h$. Furthermore, if two cells differ only with respect to the exchange of single particles, they are quantum mechanically considered identical. Thus all $N!$ cells, related by a permutation of the particles, must be considered identical and therefore $N!h^{3N}$ is a natural volume in phase space for a microstate. Then the probability of finding a microstate $x$ in a certain cell $Z_i$ of the phase space (or in one of those which have to be considered identical) is

$$p_i = N! \int_{Z_i} \mathrm{d}x \, \frac{1}{A} = \frac{N!\,h^{3N}}{A} \,. \tag{3.8}$$

These probabilities all have the same value provided the cells lie within the domain of the energy shell $\{E \leq H(x) \leq E + \Delta E\}$; otherwise, $p_i = 0$.

The inverse

$$\Omega(E, V, N) \equiv \frac{A}{N!\,h^{3N}} \tag{3.9}$$

may be regarded as the number of microstates in the interval $(E, E + \Delta E)$, and $\Omega(E, V, N)$ is called the density of states (number of states per energy interval $\Delta E$).

For the entropy one now easily obtains (see Sect. 2.4)

$$S(E, V, N) = k_\text{B} \ln \Omega(E, V, N) \,. \tag{3.10}$$

Comparing with (3.7), we have obtained

$$\varrho_0 = \frac{1}{N!\,h^{3N}} \,. \tag{3.11}$$

*Example.* For an ideal gas the Hamiltonian function is given by

$$H(x) \equiv H(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m} \ , \tag{3.12}$$

and $\Omega(E, V, N)$ is easily determined. We introduce, as an auxiliary quantity,

$$\begin{aligned} \Omega'(E, V, N) &= \int \mathrm{d}x\ \Theta(E - H(x)) && (3.13)\\ &= V^N \int \mathrm{d}^{3N} p\ \Theta\big(E - H(p)\big) \ . && (3.14) \end{aligned}$$

The integral over the momenta represents the volume of a sphere in $3N$-dimensional space with radius $\sqrt{2mE}$; it is given by

$$\Omega'(E, V, N) = V^N \frac{\pi^{3N/2}}{\Gamma\left(\frac{3N}{2} + 1\right)} (2mE)^{3N/2} \ . \tag{3.15}$$

Thus to first order in $\Delta E$ one obtains

$$\begin{aligned} A = \Omega'(E + \Delta E, V, N) - \Omega'(E, V, N) &= E\, \frac{\partial \Omega'}{\partial E} \frac{\Delta E}{E} && (3.16)\\ = V^N \frac{\pi^{3N/2}}{\Gamma\left(\frac{3N}{2}\right)} (2mE)^{3N/2} \frac{\Delta E}{E} \ , && && (3.17) \end{aligned}$$

and finally

$$\Omega(E, V, N) = \frac{1}{h^{3N} N!} V^N \frac{\pi^{3N/2}}{\Gamma\left(\frac{3N}{2}\right)} (2mE)^{3N/2} \ , \tag{3.18}$$

where we have neglected the contribution $\Delta E/E$. (It is of order $10^{-x}$, where $x$ is of order 10; but $E^N$ is of order $10^x$ with $x \sim 10^{23}$!)

Using the expansion of the gamma function for large arguments,

$$\ln \Gamma(x) = x(\ln x - 1) + O(\ln x) \ , \tag{3.19}$$

yields for the entropy

$$\begin{aligned} S(E, V, N) = k_{\mathrm{B}} N \left[\ln V + \frac{3}{2} \ln\left(\frac{2\pi m E}{h^2}\right)\right] - k_{\mathrm{B}} N \left(\ln(N) - 1\right)\\ -k_{\mathrm{B}} \frac{3}{2} N \left(\ln\left(\frac{3}{2} N\right) - 1\right) + O(\ln N) \ . \end{aligned} \tag{3.20}$$

So for the entropy of the classical ideal gas we obtain

$$S(E, V, N) = k_{\mathrm{B}} N \left[ \ln\left(\frac{V}{N}\right) + \frac{3}{2} \ln\left(\frac{4\pi m E}{3N h^2}\right) + \frac{5}{2} \right] . \tag{3.21}$$

For a fixed volume per particle, $V/N$, and fixed energy per particle, $E/N$, the entropy is strictly proportional to the number of particles $N$, as it should be. However, had we not taken into account the indistinguishability of particles postulated by quantum mechanics, the factor $N!$ in (3.18) and thus the term $N \ln N$ in (3.20) would be missing. The entropy per particle would increase with the number of particles $N$ and for $N \to \infty$ even become infinite. This contradicts the laws of classical thermodynamics and was the origin of extensive discussions among physicists before the formulation of quantum mechanics. This contradiction, as well as others which would occur without the incorporation of the indistinguishability of particles, is known as Gibbs paradox (named after the American physicist Gibbs, who discussed the indistinguishability of particles). This Gibbs paradox should not be confused with the Gibbs phenomenon, which we will address in Sect. 9.3.

For gases under normal conditions one finds

$$\frac{S}{k_{\mathrm{B}} N} \sim 10 . \tag{3.22}$$

## 3.2 Systems in Contact

Up to now we have considered a closed, isolated system. Energy, number of particles and volume were fixed, externally given values, and, as the system was closed, these values remained unchanged.

We now want to bring two such systems into contact and thereby remove their isolation in various ways. In Sect. 3.2.1 we will assume contact in such a way that interactions of the particles allow energy to be exchanged. Such contact will be called thermal contact.

In Sect. 3.2.2 we will consider exchange of energy and volume, and, in Sect. 3.2.3, exchange of energy and particles.

For each case we will introduce new system variables – temperature, pressure, and chemical potential, respectively – and then also discuss systems where these variables are held fixed.

### *3.2.1 Thermal Contact*

We first consider two closed systems with $E^0, V, N$ and $E_B^0, V_B, N_B$ as given values for the system variables energy, volume, and particle number, respectively. We bring these systems into thermal contact, i.e., allow an exchange of energy, while the total system remains closed. Having established the contact we wait for a certain time until we know for sure that the total system is in a stationary equilibrium state.

The energies of the two systems, denoted by $e$ and $e_B$, will have changed. At each instant they can be considered as realizations of the random variables $E$ and $E_B$. However, because of energy conservation one always has $e+e_B = E_{\text{tot}} = E^0+E_B^0$.

We are interested in the probability density $\varrho(e)$ for the energy of the first system. To determine this we consider the number of microstates of the total system in which the first system has the energy $e$. This number is $\Omega(e, V, N)\, \Omega_B(E_{\text{tot}} - e, V_B, N_B)$, because each state of the first system can be combined with a distinct state of the second system.

In the closed total system all states are equally probable. Denoting the number of states of the total system by $A$, we therefore obtain

$$\varrho(e) = \frac{1}{A}\Omega(e, V, N)\Omega_B(E_{\text{tot}} - e, V_B, N_B) \ . \tag{3.23}$$

We will analyze this expression for the energy density of the first system and find that it may also be written as

$$\varrho(e) \propto \mathrm{e}^{-Ng(y)} \ , \quad y = \frac{e}{N} \ , \tag{3.24}$$

where $y$ is the energy per particle in the first system. Therefore, the random variable has the large deviation property. In the thermodynamic limit $N \to \infty$ the energy per particle will assume a unique value determined by the minimum of $g(y)$.

### Introduction of Temperature

First we determine the maximum of $\varrho(e)$ by setting the derivative of $k_{\mathrm{B}} \ln \varrho(e)$ with respect to $e$ equal to zero. Denoting the entropy of the microcanonical density by $S = k_{\mathrm{B}} \ln \Omega$ we have

$$k_{\mathrm{B}} \ln \varrho(e) = S(e, V, N) + S_B(E_{\text{tot}} - e, V_B, N_B) - k_{\mathrm{B}} \ \ln A \ , \tag{3.25}$$

and therefore

$$\frac{\partial k_{\mathrm{B}} \ln \varrho(e)}{\partial e} = \frac{\partial S(e, V, N)}{\partial e} - \left.\frac{\partial S_B(e_B, V_B, N_B)}{\partial e_B}\right|_{e_B = E_{\text{tot}} - e} = 0 \ . \tag{3.26}$$

In order to analyze this equation, we introduce

$$\frac{1}{T} = \frac{\partial S(e, V, N)}{\partial e} \tag{3.27}$$

and call $T$ the temperature. Similarly, we introduce the temperature $T_B$. Note that $k_\mathrm{B}T$ has the dimension of an energy. Thus (3.26) implies that the maximum of $\varrho(e)$ is determined by the solution of the equation

$$T(e, V, N) = T_B(E_\mathrm{tot} - e, V_B, N_B) \ , \tag{3.28}$$

that is, the most probable value of $e$ is that for which the temperatures of the two systems are equal.

*Remarks.*

- We have introduced a new system variable, the temperature, which can be attributed to each system. When the entropy function of the microcanonical density $S(E, V, N)$ is known, the temperature is given by (3.27). For the ideal gas, for instance, we have from (3.21)

$$S(E, V, N) = k_\mathrm{B}N \left[ \ln\left(\frac{V}{N}\right) + \frac{3}{2}\ln\left(\frac{4\pi m E}{3N h^2}\right) + \frac{5}{2} \right] \ , \tag{3.29}$$

  and therefore one obtains for the temperature

$$\frac{1}{T} = \frac{\partial S(E, V, N)}{\partial E} = k_\mathrm{B} N \frac{3}{2}\frac{1}{E} \ , \tag{3.30}$$

  or

$$T = \frac{E}{\frac{3}{2}k_\mathrm{B}N} \text{ or } E = \frac{3}{2}Nk_\mathrm{B}T \ . \tag{3.31}$$

  Of course, we still have to show how a variable like the temperature can be measured (see Sect. 7.2).
- The temperature is an intensive variable. These are system variables which mutually adjust when two systems enter corresponding contact.
- As the entropy increases with the number of possible states (cf. Sect. 2.4) and, in addition, the number of possible states generally increases with energy, the temperature will, in general, be positive. However, when there exists a maximum value for the energy, the number of states can decrease with growing energy. In this case it might happen that $T < 0$ (e.g., for the ideal paramagnetic crystal; see Sect. 4.4).
- One can show (see, e.g., Kittel (1980)) that energy and entropy flow from the system with initially higher temperature to the system with initially lower temperature.
- If two systems with slightly different temperatures are in thermal contact and there is an energy flow $\delta Q$ into one system, then the change of entropy is given by

$$S(e + \delta Q, V, N) - S(e, V, N) = \frac{\partial S}{\partial e}\delta Q \ , \tag{3.32}$$

i.e.,

$$\mathrm{d}S(e, V, N) = \frac{1}{T}\delta Q \ . \tag{3.33}$$

The change of energy for a system in thermal contact is therefore

$$\delta Q = T\, \mathrm{d}S(e, V, N) \ . \tag{3.34}$$

This quantity of energy which is exchanged in connection with a change of entropy is called heat (cf. Sect. 7.1).

### The Density of the Canonical System

We now come back to the analysis of formula (3.23) for the density $\varrho(e)$.

Let the energy $E_B$ of the second system be so large that the energy which will be exchanged in a contact is negligible compared to $E_B$. Then $T_B$ will also be practically unchanged, and the first system will assume the temperature $T_B$ as a consequence of the contact. A system that is large enough to determine the temperature of any other system it is in contact with, is called a heat bath. If the system variables $T, V, N$ for a system are given, one speaks of a canonical system.

The density $\varrho(e)$ of a canonical system has to be interpreted as $\varrho(e|T, V, N)$, i.e., the density of a canonical system. According to (3.23) we have

$$\varrho(e|T, V, N) = \frac{1}{A}\Omega(e, V, N)\Omega_B(E_{\mathrm{tot}} - e, V_B, N_B) \tag{3.35}$$

$$= \frac{1}{A}\Omega(e, V, N)\exp\left(\frac{1}{k_{\mathrm{B}}}S_B(E_{\mathrm{tot}} - e, V_B, N_B)\right) \ , \tag{3.36}$$

where $\Omega(e, V, N)$ is the number of cells, i.e., the number of microstates in the interval $(e, e + \mathrm{d}e)$, and $S_B(e, V, N)$ is the corresponding entropy.

The expression $S_B(E_{\mathrm{tot}} - e, V_B, N_B)$ can be expanded with respect to $e$ and one obtains

$$S_B(E_{\mathrm{tot}} - e, V_B, N_B) = S_B(E_{\mathrm{tot}}, V_B, N_B) - \frac{1}{T}e + O\left(\frac{1}{N_B}\right) \ , \tag{3.37}$$

so that, finally, we get the result

$$\begin{aligned} \varrho(e|T, V, N) &= \frac{1}{A'}\Omega(e, V, N)\mathrm{e}^{-\beta e} \\ &\propto \exp\left(-\beta e + S(e, V, N)/k_{\mathrm{B}}\right) \ , \end{aligned} \tag{3.38}$$

where $\beta = 1/k_{\mathrm{B}}T$.

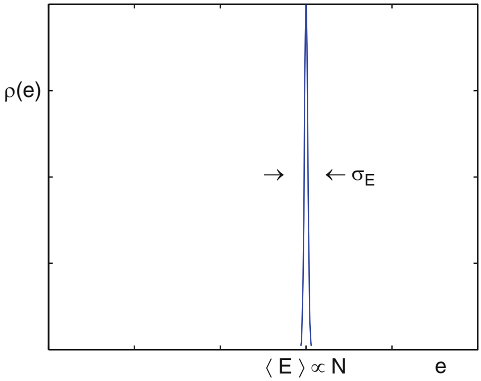


**Fig. 3.1** The energy density of a canonical system. To a good approximation this is the density of a Gaussian distribution with variance $\sigma_E^2 \sim N$

We note that if we introduce the energy per particle $Y_N = E/N$ as a random variable and take $s(y_N) = S(e, V, N)/N$ as the entropy per particle, the density $\varrho_{Y_N}(y)$ of $Y_N$ has, according to (3.38), the form

$$\varrho_{Y_N}(y) \propto \mathrm{e}^{-Ng(y)} \ , \ \ \text{where } g(y) = \beta y - s(y)/k_\mathrm{B} \ . \tag{3.39}$$

The sequence of random variables $\{Y_N, N \to \infty\}$ whose realizations are equal to the energy per particle, thus has the large deviation property.

We want to illustrate this further by determining the variance of the density $\varrho(e)$.

The value where $\varrho(e)$ assumes its maximum, will be denoted by $\hat{e}_0$, and we expand $\varrho(e)$ and $k_\mathrm{B} \ln \varrho(e)$ near $\hat{e}_0$: With $e = \hat{e}_0 + \eta$ we obtain

$$k_\mathrm{B} \ln \varrho(\hat{e}_0 + \eta) = k_\mathrm{B} \ln \varrho(\hat{e}_0) + \frac{1}{2}\eta^2 \lambda + \dots \ , \tag{3.40}$$

where

$$\lambda = \left.\frac{\partial^2 S(e, V, N)}{\partial e^2}\right|_{e=\hat{e}_0} = \frac{\partial}{\partial e}\left(\frac{1}{T}\right) = -\frac{1}{T^2}\frac{\partial T}{\partial e} \ . \tag{3.41}$$

This result shows the following:

- $\lambda$ is negative because $e$ increases with $T$; a similar result holds for $\lambda_B$. This implies that we really are dealing with a maximum.
- From $e = O(N)$ we find $\lambda = O(\frac{1}{N})$. The higher order terms in $\eta$ will also be of higher order in $1/N$. Up to terms of order $1/N$ the exponent of $\varrho(e)$ is thus a quadratic function of $\eta$, and $\varrho(e)$ corresponds to the density of a Gaussian distribution with variance $\sigma_E^2$, which is of the order $N$ (Fig. 3.1).

From this we find

$$\frac{\sigma_E}{\hat{e}_0} = O\left(\frac{1}{\sqrt{N}}\right) \ . \tag{3.42}$$

For $N \sim 10^{23}$ the ratio is therefore $\sigma_E/\hat{e}_0 \sim 10^{-12}$. This implies that, although the internal energy of a system which is in thermal contact with a second system is not fixed, in practice it assumes a fixed value $\hat{e}_0$, which can be derived from (3.28) and is practically identical to $\langle E \rangle$.

### The Boltzmann Density

Knowing the density $\varrho(e) \equiv \varrho(e \,|\, T, V, N)$ we now can calculate the density $\varrho(x|T, V, N)$ for a system in contact with a heat bath of temperature $T$. Using (3.4) and (3.9) one finds

$$\begin{aligned}
\varrho(x|T, V, N) &= \int \mathrm{d}e \, \varrho(x|e, V, N)\varrho(e|T, V, N) && (3.43)\\
&= \int_{H(x)-\Delta E}^{H(x)} \mathrm{d}e \, \frac{1}{A} \frac{1}{A'} \, \Omega(e, V, N)\mathrm{e}^{-\beta e} && (3.44)\\
&= \int_{H(x)-\Delta E}^{H(x)} \mathrm{d}e \, \frac{1}{A' h^{3N} N!} \, \mathrm{e}^{-\beta e} && (3.45)\\
&= \frac{1}{A' h^{3N} N!} \, \mathrm{e}^{-\beta H(x)} \Delta E && (3.46)\\
&= \frac{1}{A''} \, \mathrm{e}^{-\beta H(x)} \, . && (3.47)
\end{aligned}$$

This density is called the canonical density or Boltzmann density. We have already derived it in Sect. 2.4.5 proceeding from the maximum entropy principle as the density of maximal entropy in the situation where the energy $H(X)$ is not given but is a random variable with

$$\langle H(X) \rangle = E \, , \tag{3.48}$$

where $E$ turns out to be precisely the energy the system assumes according to (3.28) in a thermal contact. For the Boltzmann density we will also use the form

$$\varrho(x|T, V, N) = \frac{1}{h^{3N} N! Z} \, \mathrm{e}^{-\beta H(x)} \, , \tag{3.49}$$

where, in anticipation of quantum mechanical results, we have split off the factor $h^{3N} N!$ from the normalization factor $Z$. This remaining dimensionless normalization factor $Z$ is called the partition function. From the normalization condition we find

$$Z(T, V, N) = \int \mathrm{d}^{3N} p \, \mathrm{d}^{3N} q \, \frac{1}{h^{3N} N!} \, \mathrm{e}^{-\beta H(p,q)} \, . \tag{3.50}$$

### The Free Energy

Having determined the probability density for the microstates of a system with given values of temperature, volume, and number of particles, it now seems natural, from a probabilistic point of view, to define the generating functions of the moments and the cumulants. But as only the expectation values of macroscopic quantities like $H(x)$ are of interest in statistical mechanics, the partition function already plays the role which is played otherwise by the generating function of the moments: The repeated differentiation with respect to a parameter yields all moments. The variable $F(T, V, N)$, defined by

$$-\beta F(T,V,N) = \ln Z(T,V,N)\ , \tag{3.51}$$

so that

$$Z(T,V,N) = \mathrm{e}^{-\beta F(T,V,N)} \tag{3.52}$$

serves now as a generating function of the cumulants. We will call it the free energy. The prefactor $\beta = 1/k_{\mathrm{B}}T$ provides the free energy $F(T, V, N)$ with the dimension of an energy. For $E(T, V, N) \equiv \langle H(X)\rangle$ we thus obtain

$$\langle H(X)\rangle = \frac{1}{Z}\int d^{3N}p\, d^{3N}q\, \frac{1}{h^{3N}N!}\, H(p,q)\mathrm{e}^{-\beta H(p,q)} \tag{3.53}$$

$$= \frac{1}{Z}\left(-\frac{\partial}{\partial\beta}\right) Z \tag{3.54}$$

$$= -\frac{\partial}{\partial\beta}\ln Z(T,V,N), \tag{3.55}$$

and for $\mathrm{Var}(H(X))$ we have

$$\mathrm{Var}(H(X)) = \langle H^2\rangle - \langle H\rangle^2 \tag{3.56}$$

$$= \frac{1}{Z}\frac{\partial^2 Z}{\partial\beta^2} - \frac{1}{Z^2}\left(\frac{\partial Z}{\partial\beta}\right)^2 \tag{3.57}$$

$$= \frac{\partial^2}{\partial\beta^2}\ln Z. \tag{3.58}$$

For the entropy of the canonical density one obtains immediately from (2.86), setting $\varrho_0 = 1/(h^{3N}N!)$,

$$S(T,V,N) = -\frac{1}{T}\left(F(T,V,N) - E(T,V,N)\right) \tag{3.59}$$

with

$$E(T,V,N) = \langle H(X)\rangle\ . \tag{3.60}$$

Therefore, we also have

$$F(T, V, N) = E(T, V, N) - T\,S(T, V, N)\ . \tag{3.61}$$

Notice that the entropy of the canonical density $S(T, V, N)$ is a function entirely different from the entropy $S(E, V, N)$ of the microcanonical density. Nevertheless, as 'usual' in physics, we will use the same name 'S' for it.

In addition to the microcanonical system, for which the system variables $E, V, N$ are held fixed, we have now introduced a second system, the canonical system. It corresponds to an externally determined temperature (through the contact with a heat bath), volume, and particle number. The probability density of the microstates is given by (3.49), which we also may write as

$$\varrho(x|T, V, N) = \frac{1}{h^{3N} N!}\,\mathrm{e}^{\beta\left(F - H(x)\right)}\ . \tag{3.62}$$

*Example.* For the ideal gas we have

$$\begin{aligned} Z &= \frac{1}{N!h^{3N}} \int \mathrm{d}^{3N}q\,\mathrm{d}^{3N}p\ \exp\left(-\beta \sum_{i=1}^{N} \frac{p_i^2}{2m}\right) \\ &= \frac{1}{N!h^{3N}}\,V^N \int \mathrm{d}^{3N}p\ \exp\left(-\beta \sum_{i=1}^{N} \frac{p_i^2}{2m}\right) \\ &= \frac{1}{N!h^{3N}}\,V^N (2\pi m k_{\mathrm{B}} T)^{3N/2}\ , \end{aligned}$$

where we have used $\int_{-\infty}^{+\infty} \mathrm{d}x\,\mathrm{e}^{-x^2} = \sqrt{\pi}$. With help of

$$\lambda_{\mathrm{t}} = \sqrt{\frac{h^2}{2m\pi k_{\mathrm{B}} T}}\ , \tag{3.63}$$

which in Chap. 6 will be introduced as thermal de Broglie wavelength, we may also write

$$Z(T, V, N) = \frac{1}{N!}\left(\frac{V}{\lambda_{\mathrm{t}}^3}\right)^N\ . \tag{3.64}$$

Thus

$$\begin{aligned} -\beta F = \ln Z &= N \ln\left(\frac{V}{\lambda_{\mathrm{t}}^3}\right) - N(\ln N - 1) \\ &= N\left[\ln\left(\frac{V}{N\lambda_{\mathrm{t}}^3}\right) + 1\right]\ , \end{aligned}$$

i.e., with $v = V/N$:

$$F(T, N, V) = -k_{\mathrm{B}}TN\left[\ln\left(\frac{v}{\lambda_{\mathrm{t}}^3}\right) + 1\right] . \tag{3.65}$$

For $E(T, V, N) \equiv \langle H(X)\rangle$ one obtains

$$E(T, V, N) = -\frac{\partial}{\partial\beta}\ln Z(T, V, N) \tag{3.66}$$

$$= \frac{3}{2}Nk_{\mathrm{B}}T . \tag{3.67}$$

One can easily show that

$$S = -\frac{1}{T}(F - E) ,$$

and therefore

$$S(T, N, V) = k_{\mathrm{B}}N\left[\ln\left(\frac{v}{\lambda_{\mathrm{t}}^3}\right) + \frac{5}{2}\right] .$$

### 3.2.2 Systems with Exchange of Volume and Energy

After we have considered in the previous section the thermal contact of two systems and thereby introduced the temperature, we now want to allow not only the exchange of energy but also the exchange of volume for two systems in contact. Let $E^0$, $V^0$, $N$ and $E_B^0$, $V_B^0$, $N_B$ be the initial values for energy, volume, and number of particles, respectively, for two initially isolated systems. After these two systems come into contact and a time independent state of equilibrium has been established, we have to consider the energy $E, E_B$ and the volume $V, V_B$ as random variables with realizations $e, e_B, v, v_B$. Because of the conservation laws, we always have $e + e_B = E_{\mathrm{tot}} = E^0 + E_B^0$ and $v + v_B = V_{\mathrm{tot}} = V^0 + V_B^0$. Therefore, we now want to study the density of the random variables $V$ and $E$:

$$\varrho(e, v, N) = \frac{1}{A}\,\Omega(e, v, N)\Omega(E_{\mathrm{tot}} - e, V_{\mathrm{tot}} - v, N_B) . \tag{3.68}$$

This will again turn out to be the density of a Gaussian distribution, where the dispersion around the most probable values will be very small and indeed vanishes in the thermodynamic limit $N \to \infty$.

### Introduction of Pressure

These most probable values $\hat{e}, \hat{v}$ follow from the condition that the change of density (or rather the logarithm of the density) has to vanish to first order in d$e$ and d$v$. One finds

$$\mathrm{d}\big(k \ln \varrho(e,v)\big) = \left( \frac{\partial S(e,v,N)}{\partial e} - \left.\frac{\partial S(e_B, v_B, N_B)}{\partial e_B}\right|_{e_B = E_{\text{tot}} - e} \right) \mathrm{d}e$$
$$+ \left( \frac{\partial S(e,v,N)}{\partial v} - \left.\frac{\partial S(e_B, v_B, N_B)}{\partial v_B}\right|_{v_B = V_{\text{tot}} - v} \right) \mathrm{d}v \,. \qquad (3.69)$$

We introduce

$$p(e,v,N) = \frac{\partial S(e,v,N)}{\partial v}\, T(e,v,N)\,, \qquad (3.70)$$

and similarly $p_B(e_B, v_B, N_B)$. This new system variable will be called pressure; the justification for this will follow shortly. But already at this point it can be seen that the physical dimension of $p$ is energy/volume, or, equivalently, force/area.

As $e$ and $v$ can vary independently, the prefactors of d$e$ and d$v$ in (3.69) have to vanish separately. Therefore, the values $\hat{e}, \hat{v}$, for which the density $\varrho(e,v)$ assumes a maximum, are determined by the equations:

$$T(\hat{e}, \hat{v}, N) = T_B(\hat{e}_B, \hat{v}_B, N_B) \qquad (3.71a)$$

$$p(\hat{e}, \hat{v}, N) = p_B(\hat{e}_B, \hat{v}_B, N_B)\,. \qquad (3.71b)$$

Not only the temperatures but also the pressures of both systems adjust to each other. Hence, the pressure is also an intensive variable.

*Remarks.*

- If $S(E,V,N)$ is known for a system, the temperature and also the pressure are easy to calculate. For the ideal gas we have from (3.21)

$$S(E,V,N) = k_{\mathrm{B}} N \left[ \ln\left(\frac{V}{N}\right) + \frac{3}{2} \ln\left(\frac{4\pi m E}{3N h^2}\right) + \frac{5}{2} \right], \qquad (3.72)$$

and therefore

$$p(E,V,N) = T\, k_{\mathrm{B}} N \,\frac{1}{V} = \frac{E}{3/2 k_{\mathrm{B}} N}\, \frac{k_{\mathrm{B}} N}{V} = \frac{2}{3}\,\frac{E}{V}\,. \qquad (3.73)$$

- In the first equation of (3.73) we have already established the well known equation of state for the ideal gas, if we treat $T$ not as a function of $E$, but as an independent variable. In Sect. 3.3 we will see that $p(T, V, N)$ can also be calculated as

$$p(T, V, N) = -\frac{\partial F(T, V, N)}{\partial V} \tag{3.74}$$

and with (3.65) we also find immediately

$$p(T, V, N) = -\frac{\partial F(T, V, N)}{\partial V} = \frac{k_B T N}{V}\ . \tag{3.75}$$

- With increasing volume the number of possible microstates also increases, i.e., the pressure is always positive if the temperature is positive.
- One can show (see e.g., Kittel (1980)) that, for two systems being in contact, volume 'flows' into the system which initially has the larger pressure (i.e., that system increases its volume). This increase of volume leads to a decrease of pressure, while for the other system the pressure increases as the volume gets smaller until both pressures are equal.
- The second derivatives of $k_B \ln \varrho(e, v, N)$ at its maximum $(\hat{e}, \hat{v})$ are again of order $1/N$. This implies that $\varrho(e, v, N)$ is a two-dimensional Gaussian density. The relative dispersion around the maximum $(\hat{e}, \hat{v})$ is again of the order $1/\sqrt{N}$.
- When two systems of slightly different temperature and pressure are in contact and there is a flow of energy $\mathrm{d}E$ and volume $\mathrm{d}V$ into one system (i.e., this system increases its volume by $\mathrm{d}V$), then the change of entropy is

$$\mathrm{d}S = \frac{\partial S}{\partial E}\,\mathrm{d}E + \frac{\partial S}{\partial V}\,\mathrm{d}V = \frac{1}{T}\left(\mathrm{d}E + p\mathrm{d}V\right) \equiv \frac{\delta Q}{T}\ . \tag{3.76}$$

Hence, the change of energy under these conditions is

$$\mathrm{d}E = T\mathrm{d}S - p\mathrm{d}V\ . \tag{3.77}$$

Thus the pressure determines the change of energy which results from a change of volume. This change of energy is positive when the change of volume is negative, i.e., when the volume becomes smaller.

This agrees with our intuitive understanding and provides the justification for referring to this system variable as pressure.

Furthermore, consider, for instance, a gas in a box with a movable wall of area $A$ (Fig. 3.2). If this wall is slowly pressed inwards under the action of a force $F$ such that the mechanical pressure $p_{\text{mech}} = F/A$ is only infinitesimally larger than the counterpressure of the gas and any nonequilibrium states are avoided in this process, then the work done by the displacement of the wall by a distance $\mathrm{d}h$ is

$$\mathrm{d}W = F\mathrm{d}h = p_{\text{mech}}\,A\,\mathrm{d}h\ . \tag{3.78}$$

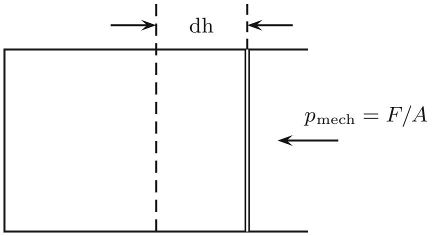


**Fig. 3.2** Change of volume $\mathrm{d}V = -A\,\mathrm{d}h$ by the displacement $\mathrm{d}h$ of a wall with area $A$

The energy $\mathrm{d}W$ is transferred to the gas. If we interpret $A\,\mathrm{d}h$ as the change of volume $-\mathrm{d}V$ of the gas, the respective change of energy of the gas is $\mathrm{d}E = -p_{\mathrm{mech}}\mathrm{d}V$. The pressure $p$ defined in (3.70) from the entropy is therefore identical to the mechanical pressure in (3.78) and to the counterpressure.

### The Density of the $T-p$ System

If $E_B$ and $V_B$ are large enough that under contact to another system the flow of energy and volume can be neglected, one speaks of a heat and volume bath. The system variables $T$ and $p$ of the smaller system can be regulated by bringing it into contact with such a bath. As a result the energy and the volume adjust. In this case one has to interpret the density $\varrho(e, v)$ as $\varrho(e, v|T, p, N)$, and it is given by

$$\begin{aligned}\varrho(e, v|T, p, N) &= \frac{1}{A}\,\Omega(e, v, N)\,\Omega_B(E_{\mathrm{tot}} - e, V_{\mathrm{tot}} - v, N_B) &(3.79)\\ &= \frac{1}{A}\,\Omega(e, v, N)\,\mathrm{e}^{S_B(E_{\mathrm{tot}}-e, V_{\mathrm{tot}}-v, N_B)/k_{\mathrm{B}}}\,. &(3.80)\end{aligned}$$

The entropy $S_B$ may be expanded in $e$ and $v$ and one obtains

$$\begin{aligned}S_B(E_{\mathrm{tot}} - e, V_{\mathrm{tot}} - v, N_B) = S_B(E_{\mathrm{tot}}, V_{\mathrm{tot}}, N_B) - \frac{1}{T}\,e - \frac{1}{T}\,pv\\ +O\left(\frac{1}{N_B}\right)\,, \qquad (3.81)\end{aligned}$$

and therefore

$$\varrho(e, v|T, p, N) = \frac{1}{A'}\,\Omega(e, v, N)\,\mathrm{e}^{-\beta(e+pv)}\,. \tag{3.82}$$

For a system in contact with a heat and a volume bath one thus obtains for the density of a microstate $x$

$$\begin{aligned}\varrho(x, V|T, p, N) &= \int_{H(x)-\Delta E}^{H(x)} \mathrm{d}e \int_{V}^{V+\Delta V} \mathrm{d}v\,\varrho(x|e, v, N)\,\varrho(e, v|T, p, N)\\ &= \frac{1}{A''}\,\mathrm{e}^{-\beta(H(x)+pV)}\,. &(3.83)\end{aligned}$$

Notice that the volume $V$ is an argument as well as the microstate $x$. In the derivation of (3.83) we have used the fact that $\varrho(x|e, v, N)$ is different from zero only for $(H(x) - \Delta E \leq e \leq H(x))$ and (now) for $(V \leq v \leq V + \Delta V)$.

This density can also be written as

$$\varrho(x, V|T, p, N) = \frac{1}{h^{3N} N!} \frac{1}{\Delta V} \frac{1}{Y'} \mathrm{e}^{-\beta(H(x)+pV)} . \tag{3.84}$$

Here $\Delta V$ is a reference volume ensuring that the expression has the correct physical dimension. The value of $\Delta V$ is not important, as it drops out in all calculations of measurable quantities. $Y'$ is called the partition function of the $T$–$p$ system. From the normalization conditions one obtains

$$\begin{aligned} Y'(T, p, N) &= \frac{1}{h^{3N} N! \, \Delta V} \int_0^\infty \mathrm{d}V \int \mathrm{d}^{3N} p \, \mathrm{d}^{3N} q \mathrm{e}^{-\beta(H(p,q)+pV)} && (3.85) \\ &= \frac{1}{\Delta V} \int_0^\infty \mathrm{d}V \, Z(T, V, N) \, \mathrm{e}^{-\beta p V} . && (3.86) \end{aligned}$$

$Y'$ may be represented as

$$Y'(T, p, N) = \mathrm{e}^{-\beta G(T,p,N)} , \tag{3.87}$$

and $G(T, p, N)$ is called the free enthalpy.

Setting $\varrho_0 = \dfrac{1}{h^{3N} N! \, \Delta V}$ one obtains from (2.86) the entropy of this density:

$$S(T, p, N) = -\frac{1}{T} \left(G(T, p, N) - E(T, p, N) - p \, V(T, p, N)\right) \tag{3.88}$$

with

$$E(T, p, N) = \langle H(X) \rangle, \tag{3.89a}$$

$$V(T, p, N) = \langle V \rangle . \tag{3.89b}$$

Notice that contrary to (3.60) the expectation value $\langle . \rangle$ now has to be taken with respect to the density $\varrho(x, V|T, p, N)$. For instance, for the volume we get

$$\begin{aligned} V(T, p, N) &= \langle V \rangle \\ &= \frac{1}{Y' \Delta V} \int_0^\infty \mathrm{d}V \int \mathrm{d}^{3N} p \, \mathrm{d}^{3N} q \; V \, \frac{1}{h^{3N} N!} \, \mathrm{e}^{-\beta(H(p,q)+pV)} \\ &= \frac{1}{Y' \Delta V} \int_0^\infty \mathrm{d}V \, Z(T, V, N) \, V \, \mathrm{e}^{-\beta p V} \\ &= -\frac{1}{\beta} \frac{1}{Y'} \frac{\partial Y'}{\partial p} = \frac{\partial}{\partial p} G(T, p, N) . && (3.90) \end{aligned}$$

Solving (3.88) for $G(T, p, N)$ one obtains

$$G(T, p, N) = E(T, p, N) + p\, V(T, p, N) - T\, S(T, p, N) \tag{3.91}$$

$$= F(T, p, N) + p\, V(T, p, N)\ . \tag{3.92}$$

*Example.* For the ideal gas we have

$$Z(T, V, N) = \frac{1}{N!} V^N \left( \frac{2\pi m k_{\mathrm{B}} T}{h^2} \right)^{3N/2} = \frac{1}{N!} \left( \frac{V}{\lambda_t^3} \right)^N\ , \tag{3.93}$$

and therefore

$$Y' = \mathrm{e}^{-\beta G(T,p,N)} = \frac{1}{\Delta V} \int_0^\infty \mathrm{d}V\, \mathrm{e}^{-\beta p V}\, \frac{1}{N!} V^N\, \lambda_t^{-3N} \tag{3.94}$$

$$= \lambda_t^{-3N}\, \frac{1}{(\beta p)^N} \frac{1}{\beta p \Delta V}\ , \tag{3.95}$$

where we have used

$$\int_0^\infty \mathrm{d}V V^N \mathrm{e}^{-\alpha V} = \left( -\frac{\mathrm{d}}{\mathrm{d}\alpha} \right)^N \int_0^\infty \mathrm{d}V \mathrm{e}^{-\alpha V} = N! \frac{1}{\alpha^{N+1}}\ . \tag{3.96}$$

Since the (dimensionless) term $\beta p \Delta V$ may be neglected, one obtains the following expression for the free enthalpy of an ideal gas:

$$G(T, p, N) = -k_{\mathrm{B}} T \left[ -N \ln(\beta p) - N \ln \lambda_t^3 \right] \tag{3.97}$$

$$= k_{\mathrm{B}} T N \ln \left( \frac{p \lambda_t^3}{k_{\mathrm{B}} T} \right)\ . \tag{3.98}$$

### 3.2.3 Systems with Exchange of Particles and Energy

Finally, we want to consider contact between two systems where an exchange of energy and particles is allowed. As the argument is similar to that in the previous sections, only the most important formulas will be given here.

- We define a new system variable

$$\mu(E, V, N) = -T(E, V, N)\, \frac{\partial S(E, V, N)}{\partial N}\ , \tag{3.99}$$

which is called chemical potential. For an ideal gas, for instance, one obtains

$$\mu(E, V, N) = k_{\mathrm{B}} T \left[ \ln \left( \frac{N}{V} \right) - \frac{3}{2} \ln \left( \frac{4\pi m E}{3h^2 N} \right) \right]\ . \tag{3.100}$$

Hence, the chemical potential depends only logarithmically on the density of particles $N/V$.

- The chemical potential is an intensive variable. For two systems in contact such that an exchange of particles is allowed, the chemical potentials become equal in stationary equilibrium. Until this equilibrium state is reached, there is a flow of particles from the system with higher chemical potential to the system with lower chemical potential.
- When two systems with slightly different temperature and chemical potential are in contact such that a quantity of energy $\mathrm{d}E$ and a quantity of particles $\mathrm{d}N$ flows into one system, the change of entropy for this system is

$$\mathrm{d}S = \frac{\partial S}{\partial E}\,\mathrm{d}E + \frac{\partial S}{\partial N}\,\mathrm{d}N = \frac{1}{T}\big(\mathrm{d}E - \mu\,\mathrm{d}N\big) \equiv \frac{\delta Q}{T}\,. \tag{3.101}$$

The change of energy for systems in such contact is therefore

$$\mathrm{d}E = T\mathrm{d}S \;+\; \mu\,\mathrm{d}N\,. \tag{3.102}$$

Thus the chemical potential determines the change of internal energy resulting from a change in the number of particles.

- One can define a ‘particle bath’. For a system in contact with a heat bath and a particle bath the variables $T$, $V$, and $\mu$ are fixed; the energy and the particle number adjust. The density for a microstate $x$ of an $N$-particle system where $N$ can vary from 0 to $\infty$ reads

$$\varrho(x, N|T, V, \mu) = \frac{1}{N!h^{3N}Y}\,\mathrm{e}^{-\beta(H(x)-\mu N)}\,. \tag{3.103}$$

This is also called the grand canonical or macrocanonical density. The grand canonical or macrocanonical partition function $Y$ is

$$Y(T, V, \mu) = \sum_{N=0}^{\infty}\int \mathrm{d}^{3N}p\,\mathrm{d}^{3N}q\;\frac{1}{N!h^{3N}}\,\mathrm{e}^{-\beta(H(p,q)-\mu N)} \tag{3.104}$$

$$= \sum_{N=0}^{\infty} Z(T, V, N)\,\mathrm{e}^{\beta\mu N}\,. \tag{3.105}$$

A corresponding variable $K(T, V, \mu)$ is introduced through the expression

$$Y(T, V, \mu) = \mathrm{e}^{-\beta K(T,V,\mu)}$$

or, equivalently,

$$K(T, V, \mu) = -k_{\mathrm{B}}T\ln Y(T, V, \mu)\,. \tag{3.106}$$

Again setting $\varrho_0 = 1/(h^{3N} N!)$ the entropy is obtained from (2.86) as

$$S(T,V,\mu) = -\frac{1}{T}\left(K(T,V,\mu) - E(T,V,\mu) + \mu N(T,V,\mu)\right)\,, \tag{3.107}$$

where

$$N(T,V,\mu) \equiv \langle N\rangle = k_{\mathrm{B}} T \frac{1}{Y}\frac{\partial Y}{\partial \mu} = -\frac{\partial K}{\partial \mu}\,. \tag{3.108}$$

Thereby one also obtains

$$K(T,V,\mu) = E(T,V,\mu) - TS(T,V,\mu) - \mu N(T,V,\mu)\,. \tag{3.109}$$

*Example.* For an ideal gas

$$Z(T,V,N) = \frac{1}{N!}\left(\frac{V}{\lambda_t^3}\right)^N \tag{3.110}$$

and one gets

$$Y(T,V,\mu) = \sum_{N=0}^{\infty} \frac{1}{N!}\left(\frac{V}{\lambda_t^3}\right)^N (\mathrm{e}^{\beta\mu})^N \equiv \mathrm{e}^{-\beta K}\,, \tag{3.111}$$

where

$$K(T,V,\mu) = -\frac{V}{\lambda_t^3}\,\mathrm{e}^{\beta\mu}\, k_{\mathrm{B}} T\,. \tag{3.112}$$

One observes that $K(T,V,\mu)$ is linear in $V$; this is also generally the case. This can be seen by the following argument: The intensive variable $p(T,V,\mu)$ cannot depend on the extensive variable $V$ because there is no other extensive variable available such that $V$ and this variable together can form an intensive quantity. Hence, $p(T,V,\mu) \equiv p(T,\mu)$, and as $p = -\frac{\partial K}{\partial V}$, we find always that

$$K = -p\,V. \tag{3.113}$$

From $N = -\partial K/\partial \mu$, one obtains

$$N = \frac{V}{\lambda_t^3}\,\mathrm{e}^{\beta\mu}\,, \tag{3.114}$$

and thus again [cf. (3.100)]

$$\mu(T,V,N) = k_{\mathrm{B}} T \ln \frac{\lambda_t^3}{v}\,. \tag{3.115}$$

## 3.3 Thermodynamic Potentials

In Sect. 3.2 we introduced the system variables temperature, pressure, and chemical potential and calculated the respective probability densities of the microstate $x = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ for situations where one or several of these intensive variables, in addition perhaps to other variables, are kept fixed externally.

In particular, we have represented the partition functions of these densities $Z, Y', Y$ in terms of the functions $F(T, V, N)$, $G(T, p, N)$, and $K(T, V, \mu)$. By definition these functions all have the physical dimension of an energy and they will play a central role. We will also refer to them as the Gibbs functions of the corresponding systems.

The starting point was the density of the microcanonical system, where the energy, the volume, and the number of particles are held fixed externally. We have calculated the entropy for this density and introduced the intensive variables by the following definitions:

$$\frac{1}{T} = \frac{\partial S}{\partial E}, \qquad \frac{p}{T} = \frac{\partial S}{\partial V}, \qquad \frac{\mu}{T} = -\frac{\partial S}{\partial N} \; . \tag{3.116}$$

The differential form $\mathrm{d}S$ of the entropy $S(E, N, V)$,

$$\mathrm{d}S = \frac{1}{T}\,\mathrm{d}E + \frac{p}{T}\,\mathrm{d}V - \frac{\mu}{T}\,\mathrm{d}N \; , \tag{3.117}$$

summarizes these relations in a simple and intelligible way, and the differential form $\mathrm{d}E$ of the internal energy $E = E(S, V, N)$ then reads

$$\mathrm{d}E = T\mathrm{d}S - p\mathrm{d}V + \mu\,\mathrm{d}N \; . \tag{3.118}$$

The differential form for $E$ will be called the Gibbs differential form.

We find that the system variables can be grouped into pairs $(T, S)$, $(p, V)$, $(\mu, N)$. The variables within each pair are conjugated with respect to energy (i.e., the product of their physical dimensions has the dimension of an energy). The first variable in each pair is an intensive variable, the second an extensive variable, i.e., it scales with the number of particles. Extensive variables, such as particle number, volume, and entropy, add when two systems come into contact.

From the fundamental form (3.118) one can read off the relations between the corresponding energy conjugated variables:

$$T = \frac{\partial E(S, V, N)}{\partial S} \; , \tag{3.119a}$$

$$p = -\frac{\partial E(S, V, N)}{\partial V} \; , \tag{3.119b}$$

$$\mu = \frac{\partial E(S, V, N)}{\partial N} \; . \tag{3.119c}$$

We will first show that the differential forms of the other Gibbs functions have the same relevance for the corresponding systems as the Gibbs form of the internal energy $E(S, V, N)$ does for the microcanonical system. For this reason, the Gibbs functions $F(T, V, N)$, $G(T, p, N)$, and $K(T, V, \mu)$ are called thermodynamic potentials. The internal energy $E(S, V, N)$ is therefore the Gibbs function or the thermodynamic potential of the microcanonical system.

From (3.61) we get for the free energy, the Gibbs function of the canonical system:

$$F(T, V, N) = -T\, S(T, V, N) + E(T, V, N)\,. \tag{3.120}$$

In principle, the microcanonical and the canonical system must be considered as different. In the microcanonical system, energy $E$ is fixed; in the canonical system, $E$ is a random variable with density $\varrho(e \mid T, V, N)$. When we observe a canonical system from time to time, we will always measure different realizations of this random variable.

How widely these measured values vary around a mean value depends, however, on the number of particles $N$. In Fig. 3.1, we illustrated that the density of, say, $\varrho(e)$ in a canonical system can be considered a good approximation as a normal distribution with an expectation value $\langle H(X)\rangle = \hat{e}_0 \propto N$ and a variance $\sigma^2 \propto N$. The variance itself is of order $N$ and in the following section we will associate the variance with a measurable system variable, a susceptibility.

The dispersion measured by the standard deviation $\sigma$ is, however, of the order $\sqrt{N}$, and the relative dispersion of the density around $\hat{e}_0$ is of the order $1/\sqrt{N}$. For particle numbers of the order $N \sim 10^{23}$, as occur in statistical mechanics, this is extremely small. For such a large number of particles, energy $E$ in a canonical system with a given temperature $T$ is practically equal to $\langle H(X)\rangle = e_0$, and the canonical system corresponds to a microcanonical system with fixed energy $e_0$, where $e_0$ is determined by

$$\frac{\partial S(e_0, V, N)}{\partial e_0} = \frac{1}{T}. \tag{3.121}$$

If we identify $\langle H(x)\rangle$, the expectation value of the energy in the canonical system, with energy $E$ of the microcanonical system, then

$$\mathrm{d}E = T\,\mathrm{d}S - p\,\mathrm{d}V + \mu\,\mathrm{d}N, \tag{3.122}$$

and thus

$$\mathrm{d}F = -T\,\mathrm{d}S - S\,\mathrm{d}T + \mathrm{d}E = -S\,\mathrm{d}T - p\,\mathrm{d}V + \mu\,\mathrm{d}N, \tag{3.123}$$

from which we can read off

$$S(T, V, N) = -\frac{\partial F(T, V, N)}{\partial T}, \tag{3.124a}$$

$$p(T,V,N) = -\frac{\partial F(T,V,N)}{\partial V}, \tag{3.124b}$$

$$\mu(T,V,N) = \frac{\partial F(T,V,N)}{\partial N}\ . \tag{3.124c}$$

Thus we have shown that the free energy is also a thermodynamic potential. For free enthalpy, we found (cf. (3.88)) after identifying $\langle H(x)\rangle$ with $E$ and $\langle V\rangle$ with $V$:

$$G(T,p,N) = E(T,p,N) - TS(T,p,N) + pV(T,p,N)\ , \tag{3.125}$$

and for the fundamental form, one obtains

$$\mathrm{d}G = -S\,\mathrm{d}T + V\,\mathrm{d}p + \mu\,\mathrm{d}N\ , \tag{3.126}$$

from which we can read off

$$S(T,p,N) = -\frac{\partial G(T,p,N)}{\partial T}, \tag{3.127a}$$

$$V(T,p,N) = \frac{\partial G(T,p,N)}{\partial p}, \tag{3.127b}$$

$$\mu(T,p,N) = \frac{\partial G(T,p,N)}{\partial N}\ . \tag{3.127c}$$

Finally, from (3.109) we find for $K(T,V,\mu)$

$$K(T,V,\mu) = E(T,V,\mu) - TS(T,V,\mu) - \mu N(T,V,\mu)\ , \tag{3.128}$$

and the fundamental form is

$$\mathrm{d}K = -S\,\mathrm{d}T - p\mathrm{d}V - N\,\mathrm{d}\mu\ , \tag{3.129}$$

from which we read off

$$S(T,V,\mu) = -\frac{\partial K(T,V,\mu)}{\partial T}, \tag{3.130a}$$

$$p(T,V,\mu) = -\frac{\partial K(T,V,\mu}{\partial V}), \tag{3.130b}$$

$$N(T,V,\mu) = -\frac{\partial K(T,V,\mu}{\partial \mu})\ . \tag{3.130c}$$

This conversion of the differential forms of the Gibbs functions suggests that the Gibbs functions itself are related by a Legendre transformation (see Sect. 2.7). This is the case, as we will show now.

We write the density $\varrho(e \mid T, V, N)$ in the form (3.39)

$$\varrho_{Y_N}(y) = \frac{1}{Z \Delta E} \exp\left(-N g(y \mid T, V, N)\right) , \qquad y = e/N , \tag{3.131}$$

where

$$g(y \mid T, V, N) = \beta y - s(y, V, N)/k_{\mathrm{B}} , \quad s = S/N . \tag{3.132}$$

For the partition function, we obtain from (3.131)

$$Z(T, V, N) = \frac{1}{\Delta E} \int_0^{\infty} \mathrm{d}y \ \mathrm{e}^{-Ng(y|T,V,N)} , \tag{3.133}$$

and one obtains for the free energy

$$-\beta F(T, V, N) = \ln \left[ \frac{1}{\Delta E} \int_0^{\infty} \mathrm{d}y \ \mathrm{e}^{-Ng(y|T,V,N)} \right] . \tag{3.134}$$

Now we consider the free energy $f = F/N$ per particle and take the thermodynamic limit $N \to \infty$, which leads us to

$$\beta f(T, V, N) = \lim_{N \to \infty} \frac{\beta F(T, V, N)}{N} \tag{3.135}$$

$$= - \lim_{N \to \infty} \left[ \frac{1}{N} \ln \left( \frac{1}{\Delta E} \int_0^{\infty} \mathrm{d}y \ \mathrm{e}^{-Ng(y|T,V,N)} \right) \right] \tag{3.136}$$

$$= \inf_y g(y \mid T, V, N) \tag{3.137}$$

$$= - \sup_y \left(-\beta y + s(y, V, N)/k_{\mathrm{B}}\right) . \tag{3.138}$$

Hence, the free energy per particle is the Legendre transform of the negative entropy $-s(y, V, N)$ of the microcanonical system. The entropy itself is concave, i.e., $-s(y, V, N)$ is convex, as required for a Legendre transformation. After multiplication by $N$ and $k_{\mathrm{B}} T$, we obtain, of course,

$$F(T, V, N) = - \sup_E \left(-E + TS(E, V, N)\right) = \inf_E \left(E - TS(E, V, N)\right) . \tag{3.139}$$

To find a minimum of $E - TS(E, V, N)$, we have to set the first derivative of this expression to zero. This leads to

$$1 - T \frac{\partial S}{\partial E} = 0, \tag{3.140}$$

from which we obtain some solution $E(T, V, N)$. Then the the free energy is the value of $E - TS(E, V, N)$ at this minimum:

$$\begin{aligned} F(T, V, N) &= E(T, V, N) - TS(E(T, V, N), V, N) \\ &= E(T, V, N) - TS(T, V, N), \end{aligned} \tag{3.141}$$

as known.

In the same manner, the thermodynamic potentials can also be converted to each other. To see this, we consider as an example the definition of $Y'$ in (3.86):

$$Y'(T, p, N) = \frac{1}{\Delta V} \int_0^\infty \mathrm{d}V \, Z(T, V, N) \, \mathrm{e}^{-p\beta V} \, . \tag{3.142}$$

The integral represents a Laplace transformation from the function $Z(T, V, N)$ to $Y'(T, p, N)$. We again identify the dependence on $N$ by introducing the free enthalpy $g$ per particle, the free energy $f$ per particle, and the volume $v$ per particle, and we obtain

$$-\beta g(T, p) = \frac{1}{N} \ln \left( \frac{1}{\Delta V} \int_0^\infty \mathrm{d}v \, \mathrm{e}^{-N(\beta f(T,V,N) + \beta p v)} \right) . \tag{3.143}$$

For $N \to \infty$ we obviously get

$$g(T, p) = -\sup_v \left(-pv - f(T, V, N)\right) , \tag{3.144}$$

or, equivalently,

$$G(T, p, N) = -\sup_V \left(-pV - F(T, V, N)\right) . \tag{3.145}$$

The free enthalpy can therefore be obtained from $F(T, V, N)$ by a Legendre transformation.

We can also determine the thermodynamic potential of the macrocanonical system from the free energy. We return to (3.105), from which we find

$$Y(T, V, \mu) = \mathrm{e}^{-\beta K(T,V,\mu)} = \sum_{N=0}^{\infty} \mathrm{e}^{V(\beta\mu n - \beta f'(T,V,N))} \, . \tag{3.146}$$

In this case we have introduced $n = N/V$ and the free energy per volume $f' = F/V$. As the sum now includes all possible numbers of particles, we have to define the thermodynamic limit by $V \to \infty$ and obtain eventually

$$K(T, V, \mu) = -\sup_N \left(\mu N - F(T, V, N)\right) . \tag{3.147}$$

Finally, the enthalpy $H(S, p, N)$ is a thermodynamic potential for a system where $S, p, N$ are given. It is

$$H(S, p, N) = -\sup_V (-pV - E(S, V, N)) . \tag{3.148}$$

Therefore, the fundamental form for the enthalpy reads:

$$\mathrm{d}H = T\mathrm{d}S + V\,\mathrm{d}p + \mu\,\mathrm{d}N . \tag{3.149}$$

The space of the macroscopic states, which in a mathematical approach might be introduced as a manifold (Straumann 1986), can be parametrized by various triplets of coordinates. We speak of triplets, because up to now we have considered only three independent macroscopic state variables. For magnetic systems or systems with several species of particles the number of coordinates increases. To each set of coordinates or triplet there corresponds a thermodynamic potential, from which the other system variables may be calculated.

These thermodynamic potentials are convex functions on this space. A Legendre transformation of the potentials represents a coordinate transformation and leads to a different thermodynamic potential.

*Remarks.*

- Motivated by the Legendre transform of $F(T, V, N)$ which leads to the free enthalpy, one may also introduce the so-called Landau free energy by

  $$\Lambda(p, T, V, N) = p\,V + F(T, V, N) . \tag{3.150}$$

  Then the free enthalpy can be determined by

  $$G(T, p, N) = \inf_V \Lambda(p, T, V, N) . \tag{3.151}$$

  The determination of the extremum of $\Lambda$ as a function of $V$ leads again to the condition that the first derivative of $\Lambda$ with respect to $V$ has to be zero, which is identical with the relationship $p = -\partial F/\partial V$. In Sect. 3.8, we will study the Landau free energy in a more complex situation, where $\Lambda$ may have more then one minimum. This happens when different phases can exist.
- The Gibbs fundamental form (3.118) specifies the ways in which this system can exchange energy: in the form of heat, $T\mathrm{d}S$, in the form of work, $-p\mathrm{d}V$, or in the form of chemical energy, $\mu\,\mathrm{d}N$. We have already referred to these forms of energy exchange in Sect. 3.2 and will return to them in Sect. 7.1.
- If heat is exchanged at constant pressure, from

  $$\mathrm{d}E = T\mathrm{d}S - p\mathrm{d}V \tag{3.152}$$

we also obtain

$$\delta Q \equiv T\mathrm{d}S = \mathrm{d}\,(E + pV) = \mathrm{d}H \ , \tag{3.153}$$

i.e., the change of enthalpy in processes which take place at constant pressure is equal to the exchanged quantity of heat.

- For an infinitesimal isothermal change of state we have

$$\mathrm{d}(E - TS) = -p\mathrm{d}V = \mathrm{d}W \ , \tag{3.154}$$

or, with $F = E - TS$,

$$\mathrm{d}F = \mathrm{d}W \ , \tag{3.155}$$

i.e., the change of the free energy in processes which take place at constant temperature is equal to the work done by the system (cf. Sect. 7.3).

## 3.4 Susceptibilities

Up to now we have determined the probability densities for the various systems. The systems differ with respect to the collection of system variables which are externally given. The first derivatives of the corresponding thermodynamic potentials lead to the corresponding energy conjugated system variables. However, in addition to such relations among the system variables of a gas or a fluid, one can also measure the so-called response functions or susceptibilities, which are a measure of the response of a system variable to a change in a second system variable. We now will introduce these response functions.

### *3.4.1 Heat Capacities*

The most important response functions are the heat capacities. They describe the quantity of heat $T\mathrm{d}S$ that must be added to a system in order to achieve an increase of the temperature by $\mathrm{d}T$. If these quantities of heat refer to a gramme, a mole, or a single particle of the substance in question, they are called specific heats. One still has to distinguish which system variables are kept constant.

1. When the number of particles and the volume $V$ are kept constant, the relevant heat capacity is

$$C_V = T\,\frac{\partial S(T,V,N)}{\partial T} = -T\,\frac{\partial^2 F(T,V,N)}{\partial T^2} \ . \tag{3.156}$$

On the other hand, for constant volume (and constant number of particles) we have

$$\mathrm{d}E = T\mathrm{d}S \ , \tag{3.157}$$

i.e., the heat absorbed is equal to the change of energy, and therefore

$$C_V = \frac{\partial E(T, V, N)}{\partial T} \; . \tag{3.158}$$

Being the second derivative of the free energy with respect to temperature, $C_V$ has to be related to a variance or covariance. To find this relation we consider

$$E(T, N, V) = -\frac{\partial}{\partial \beta} \ln Z \; , \tag{3.159}$$

and, using

$$\frac{\partial}{\partial T} = \frac{\partial \beta}{\partial T} \frac{\partial}{\partial \beta} = -\frac{1}{k_\mathrm{B} T^2} \frac{\partial}{\partial \beta} \; , \tag{3.160}$$

we obtain

$$\begin{aligned} \frac{\partial E(T, N, V)}{\partial T} &= \frac{1}{k_\mathrm{B} T^2} \frac{\partial^2}{\partial \beta^2} \ln Z = \frac{1}{k_\mathrm{B} T^2} \left( \frac{1}{Z} \frac{\partial^2 Z}{\partial \beta^2} - \frac{1}{Z^2} \left( \frac{\partial Z}{\partial \beta} \right)^2 \right) \\ &= \frac{1}{k_\mathrm{B} T^2} \left( \langle \mathcal{H}^2 \rangle - \langle \mathcal{H} \rangle^2 \right) \; . \end{aligned} \tag{3.161}$$

(In this section we use the symbol $\mathcal{H}$ to denote the Hamiltonian function, so as to avoid confusion with the enthalpy $H$.) Therefore

$$C_V = \frac{1}{k_\mathrm{B} T^2} \, \mathrm{Var}(\mathcal{H}(X)) \; , \tag{3.162}$$

where $X$ again stands for the random vector of positions and momenta. Hence, $C_V$ is proportional to the variance $\sigma^2_{\mathcal{H}}$ of the energy. Both $E(T, N, V)$ and $\sigma^2_{\mathcal{H}}$ are of the order $N$. As a further consequence we find that $C_V \geq 0$.

For an ideal gas $E = \frac{3}{2} k_\mathrm{B} N T$ and one obtains

$$C_V = \frac{3}{2} k_\mathrm{B} N \; . \tag{3.163}$$

2. When the number of particles and the pressure are kept constant, the relevant heat capacity is

$$C_p = T \, \frac{\partial S(T, p, N)}{\partial T} = -T \, \frac{\partial^2 G(T, p, N)}{\partial T^2} \; . \tag{3.164}$$

Since

$$\mathrm{d}H = T \mathrm{d}S + V \, \mathrm{d}p + \mu \, \mathrm{d}N \; , \tag{3.165}$$

the absorption of heat at constant pressure (and constant number of particles) results in a change of enthalpy, and $C_p$ can also be expressed in the form

$$C_p = \frac{\partial H(T, p, N)}{\partial T} \; . \tag{3.166}$$

Note that we consider here the enthalpy as a function of $T, p, N$, not of $S, p, N$. The enthalpy $H(T, p, N)$ can be written as

$$H(T, p, N) = -\frac{\partial}{\partial \beta} \ln Y' \; , \tag{3.167}$$

and with (3.160) one thus obtains

$$\frac{\partial H(T, p, N)}{\partial T} = \frac{1}{k_{\mathrm{B}} T^2} \frac{\partial^2}{\partial \beta^2} \ln Y' \; . \tag{3.168}$$

In analogy to (3.161) one finally gets

$$C_p = \frac{1}{k_{\mathrm{B}} T^2} \operatorname{Var}\bigl(\mathcal{H}(X) + p\, V\bigr) \; , \tag{3.169}$$

where the variance is now determined from the density $\varrho(x, V | T, p, N)$. From this expression for $C_p$ it follows immediately that $C_p \geq 0$.

For an ideal gas the enthalpy is

$$H(T, p, N) = E(T, p, N) + p\, V(T, p, N) \tag{3.170}$$

$$= \frac{3}{2} N k_{\mathrm{B}} T + N k_{\mathrm{B}} T = \frac{5}{2} N k_{\mathrm{B}} T \tag{3.171}$$

and therefore

$$C_p = \frac{5}{2} k_{\mathrm{B}} N \; . \tag{3.172}$$

Intuitively, one would also expect that $C_p > C_V$. To obtain an increase of temperature at constant pressure additional energy is needed to increase the volume.

### 3.4.2 Isothermal Compressibility

The isothermal compressibility is defined as

$$\kappa_T = -\frac{1}{V} \frac{\partial V(T, p, N)}{\partial p} \; . \tag{3.173}$$

From

$$V(T, p, N) = \frac{\partial G(T, p, N)}{\partial p} \tag{3.174}$$

one gets

$$\kappa_T = -\frac{1}{V}\,\frac{\partial^2 G(T, p, N)}{\partial p^2} = \frac{k_\mathrm{B} T}{V}\,\frac{\partial^2}{\partial p^2}\left(\ln Y'\right) \tag{3.175}$$

$$= \frac{1}{k_\mathrm{B} T V}\left(\langle V^2\rangle - \langle V\rangle^2\right), \tag{3.176}$$

i.e.,

$$\kappa_T = \frac{1}{k_\mathrm{B} T V}\,\sigma_V^2 \,. \tag{3.177}$$

Again we find $\kappa_T \geq 0$.

For an ideal gas the isothermal compressibility is

$$\kappa_T = -\frac{1}{V}\,\frac{\partial}{\partial p}\left(\frac{N k_\mathrm{B} T}{p}\right) = \left(\frac{N k_\mathrm{B} T}{p^2 V}\right) = \frac{1}{p}\,. \tag{3.178}$$

### 3.4.3 Isobaric Expansivity

The isobaric expansion coefficient (or isobaric expansivity)

$$\alpha = \frac{1}{V}\,\frac{\partial V(T, p, N)}{\partial T} = \frac{1}{V}\,\frac{\partial^2 G(T, p, N)}{\partial T\,\partial p} \tag{3.179}$$

measures the relative change of volume per unit temperature. Since it may also be written as a second derivative of the free enthalpy, it is possible to represent it as a covariance:

From

$$V(T, p, N) = \frac{\partial G(T, p, N)}{\partial p} = -\frac{1}{\beta}\,\frac{1}{Y'}\,\frac{\partial Y'}{\partial p} \tag{3.180}$$

and using

$$\frac{\partial}{\partial T} = \frac{\partial \beta}{\partial T}\,\frac{\partial}{\partial \beta} = -\frac{1}{k_\mathrm{B} T^2}\,\frac{\partial}{\partial \beta} \tag{3.181}$$

one obtains:

$$\alpha = \frac{1}{V}\,\frac{1}{k_\mathrm{B} T^2}\,\frac{\partial}{\partial \beta}\left(\frac{1}{\beta}\,\frac{1}{Y'}\,\frac{\partial Y'}{\partial p}\right) \tag{3.182}$$

$$= \frac{1}{V}\,\frac{1}{k_\mathrm{B} T^2}\left(-\frac{1}{\beta}\,\frac{1}{Y'^2}\,\frac{\partial Y'}{\partial \beta}\,\frac{\partial Y'}{\partial p} + \frac{1}{Y'}\,\frac{\partial}{\partial \beta}\left(\frac{1}{\beta}\,\frac{\partial Y'}{\partial p}\right)\right). \tag{3.183}$$

Since

$$\frac{1}{Y'}\frac{\partial}{\partial\beta}\left(\frac{1}{\beta}\frac{\partial Y'}{\partial p}\right) = \langle(\mathcal{H} + pV)\,V\rangle \tag{3.184}$$

and also

$$\frac{1}{Y'}\frac{\partial Y'}{\partial\beta} = -\langle(\mathcal{H} + pV)\rangle\;, \tag{3.185}$$

we obtain:

$$\alpha = \frac{1}{T}\frac{1}{k_{\mathrm{B}}TV}\,\mathrm{Cov}((\mathcal{H} + pV), V)\;. \tag{3.186}$$

where the covariance is determined with the density $\rho(x, V \mid T, p, N)$. For an ideal gas this yields:

$$\alpha = \frac{1}{V}\frac{\partial}{\partial T}\left(\frac{k_{\mathrm{B}}NT}{p}\right) = \frac{k_{\mathrm{B}}N}{pV} = \frac{1}{T}\;. \tag{3.187}$$

In general $\alpha$ is positive. But there are exceptions, e.g., the volume of water increases when the temperature decreases from 4°C to 0°C.

### 3.4.4 Isochoric Tension Coefficient and Adiabatic Compressibility

Finally, we give expressions for the isochoric tension coefficient

$$\beta = \frac{1}{p}\frac{\partial p(T, V, N)}{\partial T} = -\frac{1}{p}\frac{\partial^2 F(T, V, N)}{\partial V\,\partial T} \tag{3.188}$$

and the adiabatic compressibility

$$\kappa_S = -\frac{1}{V}\frac{\partial V(S, p, N)}{\partial p} = -\frac{1}{V}\frac{\partial^2 H(S, p, N)}{\partial p^2}\;. \tag{3.189}$$

The first coefficient measures the relative change of pressure of a system due to a change of temperature at constant volume. The second one describes the relative change of the volume of a system when the pressure is changed and the entropy is kept constant.

### 3.4.5 A General Relation Between Response Functions

We will conclude this section by deriving a general relation between response functions. One finds

$$C_p - C_V = VT\frac{\alpha^2}{\kappa_T}\;. \tag{3.190}$$

In an ideal gas, for example, we have $\alpha = \frac{1}{T}$ and $\kappa_T = \frac{1}{p}$ and therefore

$$C_p - C_V = VT\,\frac{p}{T^2} = \frac{pV}{T} = Nk_\mathrm{B}\ . \tag{3.191}$$

*Proof.* We consider

$$\delta Q = T\mathrm{d}S(T,p) = T\left(\frac{\partial S(T,p)}{\partial T}\,\mathrm{d}T + \frac{\partial S(T,p)}{\partial p}\,\mathrm{d}p\right) \tag{3.192}$$

$$= C_p\,\mathrm{d}T + T\,\frac{\partial S(T,p)}{\partial p}\,\left(\frac{\partial p(T,V)}{\partial T}\,\mathrm{d}T + \frac{\partial p(T,V)}{\partial V}\,\mathrm{d}V\right)\ . \tag{3.193}$$

If the volume is kept constant we find

$$\delta Q = C_V\,\mathrm{d}T = \left(C_p + T\,\frac{\partial S(T,p)}{\partial p}\,\frac{\partial p(T,V)}{\partial T}\right)\mathrm{d}T\ . \tag{3.194}$$

Now

$$\frac{\partial S(T,p)}{\partial p} = -\,\frac{\partial^2 G(T,p)}{\partial T\,\partial p} = -\,\frac{\partial V(T,p)}{\partial T} = -V\alpha\ . \tag{3.195}$$

On the other hand

$$\left(\frac{\partial p}{\partial T}\right)_V = -\,\frac{\left(\dfrac{\partial V(T,p)}{\partial T}\right)}{\left(\dfrac{\partial V(T,p)}{\partial p}\right)}\ , \tag{3.196}$$

which we can read off from $V(T,p)$ at constant $V$:

$$\mathrm{d}V = \frac{\partial V(T,p)}{\partial T}\,\mathrm{d}T + \frac{\partial V(T,p)}{\partial p}\,\mathrm{d}p = 0\ . \tag{3.197}$$

Therefore

$$\frac{\partial p(T,V)}{\partial T} = -\,\frac{\dfrac{\partial V(T,p)}{\partial T}}{\dfrac{\partial V(T,p)}{\partial p}} = -\,\frac{V\,\alpha}{-V\,\kappa_T} = \frac{\alpha}{\kappa_T}\ , \tag{3.198}$$

and thus

$$C_V = C_p - V\,T\,\frac{\alpha^2}{\kappa_T}\ , \tag{3.199}$$

hence proving (3.190).

*Remarks.*

- In formula (3.195) we made use of the fact that the second derivative of a thermodynamic potential can be interpreted in two different ways: On the one hand

$$\frac{\partial^2 G(T,p,N)}{\partial T\,\partial p} = \frac{\partial}{\partial p}\frac{\partial G(T,p,N)}{\partial T} = \frac{\partial}{\partial p}(-S(T,p,N))\,, \tag{3.200}$$

  and, on the other hand,

$$\frac{\partial^2 G(T,p,N)}{\partial T\,\partial p} = \frac{\partial}{\partial T}\frac{\partial G(T,p,N)}{\partial p} = \frac{\partial}{\partial T}V(T,p,N)\,. \tag{3.201}$$

  From this one obtains

$$\frac{\partial S(T,p,N)}{\partial p} = -\frac{\partial V(T,p,N)}{\partial T}\,. \tag{3.202}$$

  Relations of this kind, which are also called Maxwell relations, can be derived from all second derivatives of any thermodynamic potential.
- We have seen that the susceptibilities are related to the covariances of the quantities $H, V, pV$. This relation is often considered as a statement of the so-called fluctuation–dissipation theorem, since the covariances represent a measure of the fluctuations, and the susceptibilities are a measure of the dissipation. The connection is immediately evident. The thermodynamic potentials are the generating functions for the cumulants. Therefore the susceptibilities, being the second derivatives of the thermodynamic potentials, are exactly the covariance functions.

In Sect. 5.7 we will meet a more general fluctuation–dissipation theorem for stochastic processes.

## 3.5 The Equipartition Theorem

We consider a system of $N$ particles within the framework of classical physics. Let the microstate be given by $x = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$. If the Hamiltonian function $H(x)$ satisfies $H(x) \to \infty$ for a component $|x_i| \to \infty$ then the following relation holds for an arbitrary polynomial function $f(x)$:

$$\left\langle f(x)\,\frac{\partial H}{\partial x_i}\right\rangle = k_{\mathrm{B}}T\left\langle \frac{\partial f(x)}{\partial x_i}\right\rangle\,. \tag{3.203}$$

This relation is an immediate consequence of the identity

$$0 = \frac{1}{N!h^{3N}Z} \int \mathrm{d}^{3N}p\, \mathrm{d}^{3N}q\, \frac{\partial}{\partial x_i} \left( f(x)\, \mathrm{e}^{-\beta H(x)} \right) \tag{3.204}$$

$$= \left\langle \frac{\partial f(x)}{\partial x_i} \right\rangle - \beta \left\langle f(x)\, \frac{\partial H}{\partial x_i} \right\rangle . \tag{3.205}$$

In particular for $f(x) = x_j$ we obtain

$$\left\langle x_j\, \frac{\partial H}{\partial x_i} \right\rangle = \delta_{ij}\, k_{\mathrm{B}} T \ . \tag{3.206}$$

Let us look at two applications. First, we consider $H = H_{\mathrm{kin}} + V(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ with

$$H_{\mathrm{kin}} = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m} = \sum_{\alpha=1}^{3N} \frac{p_\alpha^2}{2m} \ , \tag{3.207}$$

where we have denoted by $p_\alpha$, $\alpha = 1, \ldots, 3N$ the $3N$ coordinates of the $N$ momenta $(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N)$.

Hence, we have for each $\alpha$:

$$p_\alpha\, \frac{\partial H}{\partial p_\alpha} = \frac{p_\alpha^2}{m} \tag{3.208}$$

from which it follows that

$$\sum_{\alpha=1}^{3N} \left\langle p_\alpha\, \frac{\partial H}{\partial p_\alpha} \right\rangle \equiv 2\, \langle H_{\mathrm{kin}} \rangle = 3N k_{\mathrm{B}} T \tag{3.209}$$

or

$$\langle H_{\mathrm{kin}} \rangle = \frac{3}{2} N k_{\mathrm{B}} T \ . \tag{3.210}$$

Second, when $V(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ is of the form

$$V(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) = \sum_{i=1}^{N} a_i \boldsymbol{q}_i^2 = \sum_{\alpha=1}^{3N} \tilde{a}_\alpha q_\alpha^2 \ , \tag{3.211}$$

one obtains

$$\sum_{\alpha=1}^{3N} \left\langle q_\alpha\, \frac{\partial H}{\partial q_\alpha} \right\rangle = 2 \sum_{\alpha=1}^{3N} \left\langle \tilde{a}_\alpha q_\alpha^2 \right\rangle \equiv 2\, \langle V \rangle \tag{3.212}$$

and thus, from (3.206),

$$\langle V \rangle = \frac{3}{2} N k_{\mathrm{B}} T \ . \tag{3.213}$$

We note that every canonical variable which appears quadratically in the Hamiltonian function contributes an amount of $\frac{1}{2} k_{\mathrm{B}} T$ to the average energy. Therefore we obtain, e.g., $\langle H \rangle = \frac{3}{2} N k_{\mathrm{B}} T$ for the classical ideal gas and we will get $\langle H \rangle = 3 N k_{\mathrm{B}} T$ from the contribution of the harmonic lattice vibrations to the internal energy of a solid. (See also the law of Dulong and Petit in Sect. 6.6.)

To each variable entering quadratically in the Hamiltonian function, one attributes a thermodynamic degree of freedom. The number of degrees of freedom is therefore $f = 3N$ for the ideal gas, and $f = 6N$ for the harmonic oscillator.

Hence, for a system with $f$ degrees of freedom one obtains the so-called equipartition theorem

$$\langle H \rangle = f \, \frac{1}{2} \, k_{\mathrm{B}} T \ . \tag{3.214}$$

For an ideal diatomic gas, taking into account the rotations, we find $f = (3 + 2)$ $N = 5N$, and when the vibrations are also taken into account $f = (3 + 2 + 2)$ $N = 7N$. The diatomic gas in the classical regime therefore has $C_V = \frac{7}{2} N k_{\mathrm{B}}$. However, a quantum-mechanical treatment of the ideal diatomic gas reveals that in most cases the vibrational excitations to not contribute to the specific heat at room temperature and therefore $C_V = \frac{5}{2} N k_{\mathrm{B}}$ (cf. Sect. 6.7).

Nonlinearities in the potential and quantum corrections will, of course, lead to deviations from this equipartition theorem.

## 3.6 The Radial Distribution Function

We first introduce the local density of particles as a random variable

$$N(\boldsymbol{r}) = \sum_{i=1}^{N} \delta(\boldsymbol{r} - \boldsymbol{Q}_i) \ . \tag{3.215}$$

Here $\boldsymbol{Q}_i$ are random variables whose realizations correspond to the positions of the individual molecules. That the quantity $N(\boldsymbol{r})$ really is something like a local particle density can easily be seen when we calculate the number of particles in a volume element $\mathrm{d}V$. In this case one finds for the number $n(\boldsymbol{r})$ of a realization of $N(\boldsymbol{r})$:

$$\int_{\mathrm{d}V} \mathrm{d}^3 r \, n(\boldsymbol{r}) = \sum_{i=1}^{N} \int_{\mathrm{d}V} \mathrm{d}^3 r \, \delta(\boldsymbol{r} - \boldsymbol{q}_i) = \sum_{i=1}^{N} I_{\boldsymbol{q}_i \in \mathrm{d}V} \ , \tag{3.216}$$

where $I_{\boldsymbol{q}_i \in dV} = 1$ if $\boldsymbol{q}_i$ is in $\mathrm{d}V$, and is otherwise zero. So all particles in $\mathrm{d}V$ are counted.

For a canonical system one has

$$\langle N(\boldsymbol{r})\rangle = \frac{1}{N!h^{3N}Z}\,N\int \mathrm{d}^{3N}p\int \mathrm{d}^{3N}q\,\delta(\boldsymbol{r}-\boldsymbol{q}_1)\,\mathrm{e}^{-\beta H(p,q)} \tag{3.217}$$

$$= N\varrho_1(\boldsymbol{r}) \tag{3.218}$$

with the marginal density

$$\varrho_1(\boldsymbol{r}) = \frac{1}{N!h^{3N}Z}\int \mathrm{d}^{3N}p\int \mathrm{d}^3q_2\ldots\mathrm{d}^3q_N\;\; \mathrm{e}^{-\beta H(p,q)}\Big|_{\boldsymbol{q}_1=\boldsymbol{r}}\;. \tag{3.219}$$

In a spatially homogeneous system

$$\varrho_1(\boldsymbol{r}) = \frac{1}{V} \tag{3.220}$$

and therefore one finds for the local density of particles

$$\langle N(\boldsymbol{r})\rangle = \frac{N}{V} \equiv n\;. \tag{3.221}$$

Next we consider the second moment

$$\big\langle N(\boldsymbol{r})N(\boldsymbol{r}')\big\rangle = \sum_{i,j=1}^{N}\big\langle\delta(\boldsymbol{r}-\boldsymbol{Q}_i)\,\delta(\boldsymbol{r}'-\boldsymbol{Q}_j)\big\rangle \tag{3.222}$$

$$= \sum_{i=1}^{N}\big\langle\delta(\boldsymbol{r}-\boldsymbol{Q}_i)\,\delta(\boldsymbol{r}'-\boldsymbol{Q}_i)\big\rangle \tag{3.223}$$

$$+\left\langle\sum_{i,j=1 i\neq j}^{N}\delta(\boldsymbol{r}-\boldsymbol{Q}_i)\,\delta(\boldsymbol{r}'-\boldsymbol{Q}_j)\right\rangle\;. \tag{3.224}$$

For the first term one obtains immediately

$$\sum_{i=1}^{N}\big\langle\delta(\boldsymbol{r}-\boldsymbol{Q}_i)\,\delta(\boldsymbol{r}'-\boldsymbol{Q}_i)\big\rangle = \delta(\boldsymbol{r}-\boldsymbol{r}')\,\sum_{i=1}^{N}\langle\delta(\boldsymbol{r}-\boldsymbol{Q}_i)\rangle \tag{3.225}$$

$$= \delta(\boldsymbol{r}-\boldsymbol{r}')\,\langle N(\boldsymbol{r})\rangle \tag{3.226}$$

$$= \frac{N}{V}\delta(\boldsymbol{r}-\boldsymbol{r}') = n\delta(\boldsymbol{r}-\boldsymbol{r}')\;, \tag{3.227}$$

and for the second term

$$\begin{aligned}\sum_{i,j=1 i\neq j}^{N} &\left\langle \delta(\boldsymbol{r}-\boldsymbol{Q}_i)\,\delta(\boldsymbol{r}'-\boldsymbol{Q}_j)\right\rangle \\ &= \frac{N(N-1)}{N!h^{3N}Z}\int \mathrm{d}^{3N}p\,\mathrm{d}^{3N}q\,\delta(\boldsymbol{r}-\boldsymbol{q}_1)\,\delta(\boldsymbol{r}'-\boldsymbol{q}_2)\,\mathrm{e}^{-\beta H(p,q)} \\ &= N(N-1)\,\varrho_2(\boldsymbol{r},\boldsymbol{r}') \qquad (3.228)\end{aligned}$$

with the marginal density

$$\varrho_2(\boldsymbol{r},\boldsymbol{r}') = \frac{1}{N!h^{3N}Z}\int \mathrm{d}^{3N}p \int \mathrm{d}^3q_3\ldots\mathrm{d}^3q_N\ \ \mathrm{e}^{-\beta H(p,q)}\big|_{\boldsymbol{q}_1=\boldsymbol{r}\boldsymbol{q}_2=\boldsymbol{r}'}\ . \qquad (3.229)$$

Hence, for the second moments of the local density of particles we find

$$\left\langle N(\boldsymbol{r})N(\boldsymbol{r}')\right\rangle = \frac{N}{V}\,\delta(\boldsymbol{r}-\boldsymbol{r}') + N(N-1)\,\varrho_2(\boldsymbol{r},\boldsymbol{r}')\ . \qquad (3.230)$$

For a spatially homogeneous system the second moment function $\langle N(\boldsymbol{r})N(\boldsymbol{r}')\rangle$ can depend only on $\boldsymbol{r}-\boldsymbol{r}'$. If in addition the system is isotropic, $\langle N(\boldsymbol{r})N(\boldsymbol{r}')\rangle$ can depend only on $|\boldsymbol{r}-\boldsymbol{r}'|$.

For large distances $|\boldsymbol{r}-\boldsymbol{r}'|$ the dependence between the two local densities should vanish such that

$$\left\langle N(\boldsymbol{r})N(\boldsymbol{r}')\right\rangle \to \langle N(\boldsymbol{r})\rangle\left\langle N(\boldsymbol{r}')\right\rangle = n^2 \quad \text{for} \quad |\boldsymbol{r}-\boldsymbol{r}'|\to\infty\ . \qquad (3.231)$$

This suggests that, in addition to the second moment function, one also introduces a radial distribution function given by

$$\begin{aligned} g_2(\boldsymbol{r},\boldsymbol{r}') &= \frac{\langle N(\boldsymbol{r})N(\boldsymbol{r}')\rangle - n\delta(\boldsymbol{r}-\boldsymbol{r}')}{\langle N(\boldsymbol{r})\rangle\,\langle N(\boldsymbol{r}')\rangle} && (3.232)\\ &= \frac{N(N-1)}{n^2}\varrho_2(\boldsymbol{r},\boldsymbol{r}') && (3.233)\\ &= \frac{(N-1)V}{n}\varrho_2(\boldsymbol{r},\boldsymbol{r}')\ . && (3.234)\end{aligned}$$

This function satisfies $g_2(\boldsymbol{r},\boldsymbol{r}')\to 1$ for $|\boldsymbol{r}-\boldsymbol{r}'|\to\infty$.

From their definition, the second moment function and the radial distribution function are closely related to the interactions between the particles. We shall clarify this point in more detail below.

For an ideal gas we get immediately from (3.229)

$$\varrho_2(\boldsymbol{r}, \boldsymbol{r}') = \frac{1}{V^2} \tag{3.235}$$

and therefore

$$\left\langle N(\boldsymbol{r})N(\boldsymbol{r}')\right\rangle = \frac{N}{V}\,\delta(\boldsymbol{r} - \boldsymbol{r}') + \frac{N(N-1)}{V^2}\;. \tag{3.236}$$

For the radial distribution function this leads to $g_2(\boldsymbol{r}, \boldsymbol{r}') = \frac{N-1}{N} \approx 1$.

Note that one always has

$$\int \mathrm{d}^3 r\, \mathrm{d}^3 r' \left\langle N(\boldsymbol{r})N(\boldsymbol{r}')\right\rangle = N^2\;. \tag{3.237}$$

Furthermore, we find

$$n g_2(\boldsymbol{r}, \boldsymbol{r}') = (N-1)\,\frac{\varrho_2(\boldsymbol{r}, \boldsymbol{r}')}{\varrho_1(\boldsymbol{r})} = (N-1)\,\varrho_2(\boldsymbol{r}'|\boldsymbol{r})\;, \tag{3.238}$$

where $\varrho_2(\boldsymbol{r}'|\boldsymbol{r})\,\mathrm{d}^3 r'$ is the probability of finding a particle in a region around $\boldsymbol{r}'$ under the condition that a particle is at $\boldsymbol{r}$. In particular, if $g_2(\boldsymbol{r}, \boldsymbol{r}')$ only depends on $|\boldsymbol{r} - \boldsymbol{r}'|$, one gets

$$\int \mathrm{d}^3 r'\, n g_2(|\boldsymbol{r} - \boldsymbol{r}'|) = \int \mathrm{d}r\, r^2\, 4\pi\, n g_2(r) = (N-1) \tag{3.239}$$

and therefore

$$n g_2(r) 4\pi r^2 \mathrm{d}r \tag{3.240}$$

is the average number of molecules in a spherical shell $(r, r + \mathrm{d}r)$ around a given molecule in the gas.

In a spatially homogeneous system the origin may be placed at the center of any particle picked at random. Obviously, for $r \to 0$ we must have $g_2(\boldsymbol{r}) \to 0$, as the molecules cannot overlap. $g_2(\boldsymbol{r})$ assumes its maxima at values which correspond to those distances from the reference particle where other particles are most likely to be found. Consequently, for a crystal, $g_2(\boldsymbol{r})$ not only depends on $|\boldsymbol{r}|$, but it displays sharp maxima for those vectors that are lattice vectors. In a liquid one would not expect such a sharp structure, but one will find rotational symmetry and a few peaks whose sharpness, however, decreases with increasing distance (Fig. 3.3).

We will now show how certain many-particle quantities can be reduced to two-particle quantities using the radial distribution function.

- Suppose that the interaction potential $V(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ can be represented as a sum of two-particle potentials

$$V(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) = \frac{1}{2} \sum_{i,j=1 i \neq j}^{N} V_2(\boldsymbol{q}_i - \boldsymbol{q}_j)\;; \tag{3.241}$$

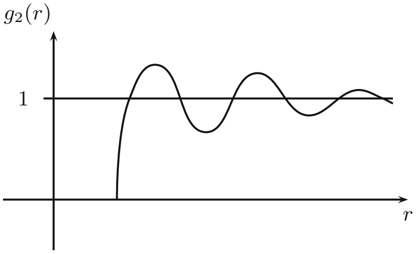


**Fig. 3.3** Typical form of the radial distribution function $g_2(r)$. The maxima are at those distances where neighboring particles are most likely to be found

then $\langle V(\boldsymbol{q}_1, \dots, \boldsymbol{q}_N) \rangle$ is determined by the radial distribution function. Indeed,

$$\langle V(\boldsymbol{q}_1, \dots, \boldsymbol{q}_N) \rangle = \frac{1}{2} \sum_{i,j=1 i \neq j}^{N} \left\langle V_2(\boldsymbol{q}_i - \boldsymbol{q}_j) \right\rangle \tag{3.242}$$

$$= \frac{1}{2} \frac{N(N-1)}{N! h^{3N} Z} \int \mathrm{d}^{3N} p \int \mathrm{d}^{3N} q \; V_2(\boldsymbol{q}_1 - \boldsymbol{q}_2) \, \mathrm{e}^{-\beta H(p,q)}$$

$$= \frac{1}{2} N(N-1) \int \mathrm{d}^3 q_1 \int \mathrm{d}^3 q_2 \; V_2(\boldsymbol{q}_1 - \boldsymbol{q}_2) \, \varrho_2(\boldsymbol{q}_1, \boldsymbol{q}_2)$$

$$= \frac{1}{2} \frac{N^2}{V} \int \mathrm{d}^3 q \; g_2(\boldsymbol{q}) \, V_2(\boldsymbol{q}) \;, \tag{3.243}$$

where we have used $(N-1) V \varrho_2(\boldsymbol{q}) = n g_2(\boldsymbol{q})$. Therefore, taking into account (3.210), one obtains for $\langle H \rangle$:

$$\langle H \rangle = \frac{3}{2} N k_{\mathrm{B}} T + \frac{N^2}{2V} \int \mathrm{d}^3 q \; g_2(\boldsymbol{q}) \, V_2(\boldsymbol{q}) \;. \tag{3.244}$$

In classical mechanics the virial is the time average of the quantity

$$\sum_{\alpha=1}^{3N} q_\alpha \frac{\partial V(\boldsymbol{q}_1 \,, \dots, \boldsymbol{q}_N)}{\partial q_\alpha} \;, \tag{3.245}$$

and a virial theorem holds in the form

$$2 \overline{H}_{\mathrm{kin}} = \overline{\sum_{\alpha=1}^{3N} q_\alpha \frac{\partial V}{\partial q_\alpha}} \;, \tag{3.246}$$

where $\overline{A}$ denotes the time average of some quantity $A$.

- In statistical mechanics the expectation value

$$\mathcal{V} = \left\langle \sum_{\alpha=1}^{3N} q_\alpha \frac{\partial V(\boldsymbol{q}_1, \dots, \boldsymbol{q}_N)}{\partial q_\alpha} \right\rangle \tag{3.247}$$

is called the virial of the potential $V(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$. If the potential can be written as a sum of two-particle potentials, then

$$\mathcal{V} = \left\langle \sum_{k=1}^{N} \boldsymbol{q}_k \cdot \nabla_k V \right\rangle \tag{3.248}$$

$$= \frac{1}{2} \sum_{i,j=1 i \neq j}^{N} \left\langle (\boldsymbol{q}_i - \boldsymbol{q}_j) \cdot \nabla V_2(\boldsymbol{q}_i - \boldsymbol{q}_j) \right\rangle \ . \tag{3.249}$$

All $N(N-1)$ terms in the sum in (3.249) now lead to the same contribution, thus

$$\begin{aligned} \mathcal{V} &= N(N-1)\frac{1}{2} \int \mathrm{d}^3 q_1 \int \mathrm{d}^3 q_2 \, (\boldsymbol{q}_1 - \boldsymbol{q}_2) \cdot \nabla V_2(\boldsymbol{q}_1 - \boldsymbol{q}_2) \, \varrho_2(\boldsymbol{q}_1, \boldsymbol{q}_2) \\ &= \frac{1}{2} \frac{N^2}{V} \int \mathrm{d}^3 q \, g_2(\boldsymbol{q}) \, \boldsymbol{q} \cdot \nabla V_2(\boldsymbol{q}) \ , \end{aligned} \tag{3.250}$$

where we have again used

$$(N-1) V \varrho_2(\boldsymbol{q}) = n g_2(\boldsymbol{q}) = \frac{N}{V} g_2(\boldsymbol{q}) \ . \tag{3.251}$$

Hence, the radial distribution function also determines the virial of the total many-particle potential.

To obtain a virial theorem in statistical mechanics we consider the state density of the canonical system. We consider the identity

$$Z(T, t^3 V, N) = \int \mathrm{d}^{3N} q \, \mathrm{d}^{3N} p \, \exp\left(-\beta \left(\sum_{\alpha=1}^{3N} \frac{p_\alpha^2}{2mt^2} + V(tq_1, \ldots, tq_{3N})\right)\right) .$$

That this equation holds can easily be seen by the coordinate transformation $p'_\alpha = \frac{1}{t}\, p_\alpha, q'_\alpha = t\, q_\alpha$, which leaves the measure $\mathrm{d}^{3N} q \, \mathrm{d}^{3N} p$ invariant but stretches each length by a factor $t$.

Taking the derivative with respect to $t$ and then setting $t = 1$ leads to

$$3V \frac{\partial Z(T, V, N)}{\partial V} = \left( 2\beta \left\langle H_{\mathrm{kin}} \right\rangle - \beta \left\langle \sum_{\alpha=1}^{3N} q_\alpha \frac{\partial V}{\partial q_\alpha} \right\rangle \right) Z \ , \tag{3.252}$$

or, with $\frac{1}{Z} \frac{\partial Z}{\partial V} = \frac{\partial}{\partial V} (-\beta F) = \beta p$

$$pV = \frac{2}{3} \left\langle H_{\mathrm{kin}} \right\rangle - \frac{1}{3} \left\langle \sum_{\alpha=1}^{3N} q_\alpha \frac{\partial V}{\partial q_\alpha} \right\rangle . \tag{3.253}$$

Using (3.250) and (3.210) and within the framework of statistical mechanics one obtains from the virial theorem the general equation of state in the form

$$pV = Nk_{\mathrm{B}}T - \frac{1}{6}\frac{N^2}{V}\int \mathrm{d}^3q\, g_2(\boldsymbol{q})\, \boldsymbol{q}\cdot\nabla V_2(\boldsymbol{q}) \,. \tag{3.254}$$

Hence, the radial distribution function plays a key role for the equations of state of real gases. It is even accessible by experiment. To explain this we first show that the radial distribution function in a spatially homogeneous system may also be written as

$$ng_2(\boldsymbol{r}) = \sum_{j=2}^{N}\left\langle \delta(\boldsymbol{r} - \boldsymbol{Q}_1 + \boldsymbol{Q}_j)\right\rangle \equiv (N-1)\left\langle \delta(\boldsymbol{r} - \boldsymbol{Q}_1 + \boldsymbol{Q}_2)\right\rangle \,. \tag{3.255}$$

This is easy to see: Since $\varrho_2(\boldsymbol{q}_1, \boldsymbol{q}_2)$ can only depend on $(\boldsymbol{q}_1 - \boldsymbol{q}_2)$, we get

$$\begin{aligned}&(N-1)\left\langle \delta(\boldsymbol{r} - \boldsymbol{Q}_1 + \boldsymbol{Q}_2)\right\rangle\\ &\quad = (N-1)\int \mathrm{d}^3q_1\, \mathrm{d}^3q_2\, \delta(\boldsymbol{r} - \boldsymbol{q}_1 + \boldsymbol{q}_2)\, \varrho_2(\boldsymbol{q}_1, \boldsymbol{q}_2)\\ &\quad = (N-1)V\varrho_2(\boldsymbol{r})\end{aligned} \tag{3.256}$$

and a comparison with (3.234) leads to the statement.

Now the Fourier transform of

$$\frac{1}{n}\left\langle N(0)N(\boldsymbol{r})\right\rangle = \delta(\boldsymbol{r}) + n\, g_2(\boldsymbol{r})$$

is easily calculated and we obtain

$$I(\boldsymbol{\kappa}) = \frac{1}{2\pi}\int \mathrm{d}^3r\, \mathrm{e}^{\mathrm{i}\boldsymbol{\kappa}\cdot\boldsymbol{r}} \sum_{j=1}^{N}\left\langle \delta(\boldsymbol{r} - \boldsymbol{Q}_1 + \boldsymbol{Q}_j)\right\rangle \tag{3.257}$$

$$= \frac{1}{2\pi}\sum_{j=1}^{N}\left\langle \mathrm{e}^{\mathrm{i}\boldsymbol{\kappa}\cdot\boldsymbol{Q}_1}\mathrm{e}^{-\mathrm{i}\boldsymbol{\kappa}\cdot\boldsymbol{Q}_j}\right\rangle \tag{3.258}$$

$$= \frac{1}{2\pi}\frac{1}{N}\sum_{i,j=1}^{N}\left\langle \mathrm{e}^{\mathrm{i}\boldsymbol{\kappa}\cdot\boldsymbol{Q}_i}e^{-\mathrm{i}\boldsymbol{\kappa}\cdot\boldsymbol{Q}_j}\right\rangle \,. \tag{3.259}$$

$I(\boldsymbol{\kappa})$ is called the static structure function. It can be measured experimentally in the quasi-elastic scattering of neutrons or X-rays at a momentum transfer of $\hbar\boldsymbol{\kappa}$. Thus the Fourier transform of the radial distribution function is a measurable quantity.

*Remarks.* The static structure function follows from the dynamical structure function $S(\boldsymbol{\kappa},\omega)$ for general inelastic scattering with momentum transfer $\hbar\boldsymbol{\kappa}$ and energy transfer $\hbar\omega$ by

$$I(\boldsymbol{\kappa}) = \hbar \int \mathrm{d}\omega\, S(\boldsymbol{\kappa},\omega)\ . \tag{3.260}$$

This dynamical structure function $S(\boldsymbol{\kappa},\omega)$ is proportional to the cross section for the inelastic scattering of a particle (neutron or photon) with momentum $\hbar\boldsymbol{k}$ and energy $E$ into the state $\hbar\boldsymbol{k}'$, $E'$:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E'} \propto S(\boldsymbol{\kappa},\omega), \quad \boldsymbol{\kappa} = \boldsymbol{k} - \boldsymbol{k}'\ ,\ E' - E = \hbar\omega\ . \tag{3.261}$$

A scattering process is called quasi-elastic if the energy transfer satisfies $|E' - E| \ll E$; for a given scattering angle $\boldsymbol{\kappa}$ is then independent of $E'$. If, then, for a given momentum transfer all photons or neutrons are registered regardless of their energies $E'$, this corresponds to the integration of the cross section over $\omega$ in (3.260) and therefore

$$I(|\boldsymbol{\kappa}|) \propto \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \equiv \int \mathrm{d}E'\, \frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E'}\ , \tag{3.262}$$

See also Berne and Pecora (1976).

## 3.7 Approximation Methods

In the previous section we have met a central quantity in the description of nonideal gases, namely the radial distribution function, and we have derived some general statements about the form of the equations of state in terms of this distribution function.

Now we will examine how systematic approximations allow us to calculate the radial distribution function or to obtain the equations of state directly. This is the subject of the statistical theory of fluids. A further approximation method, the mean field approximation, will be introduced in the context of spin systems in Sect. 4.4.

As a whole, however, within the framework of this book we can only convey a preliminary insight into the description of nonideal gases. For a more advanced treatment we refer to Barker and Henderson (1976).

### *3.7.1 The Virial Expansion*

For the ideal gas it was a straightforward matter to derive the equation of state (Sect. 3.2). We found that the pressure is linear in the particle density $n = N/V$, explicitly $p = k_\mathrm{B}Tn$. Here for nonideal gases an expansion with respect to powers of the particle density $n$ in the form

$$\frac{p}{nk_\mathrm{B}T} = \frac{pV}{Nk_\mathrm{B}T} = 1 + b(T)n + c(T)n^2 + \ldots \tag{3.263}$$

will be derived.

We proceed from the partition function of the macrocanonical system

$$Y = \sum_{N=0}^{\infty} z^N \, Z(T, V, N) \, . \tag{3.264}$$

Here $z = \mathrm{e}^{\beta\mu}$, also called fugacity, and $Z(T, V, N)$ is the partition function of the $N$-particle system. One obtains

$$\ln Y = -\beta K = \frac{pV}{kT} \tag{3.265}$$

$$= \ln\left(1 + zZ(T, V, 1) + z^2 Z(T, V, 2) + \ldots\right) \, . \tag{3.266}$$

Expanding the logarithm one finds

$$\ln Y = \frac{pV}{kT} = zZ_1 + z^2 Z_2 + z^3 Z_3 + \ldots \tag{3.267}$$

with

$$Z_1 = Z(T, V, 1) \, , \; Z_2 = Z(T, V, 2) - \frac{1}{2} \, Z^2(T, V, 1) \, , \quad \text{etc.} \tag{3.268}$$

Equation 3.267 would have the form of an equation of state if the fugacity $z$ were written as a function of $T, V, N$. In order to achieve this, we note that the determination of $N$

$$\begin{aligned} N &= -\frac{\partial K}{\partial \mu} = k_\mathrm{B}T \, \frac{\partial \ln Y}{\partial \mu} = k_\mathrm{B}T \, \frac{\partial z}{\partial \mu} \, \frac{\partial \ln Y}{\partial z} = z \, \frac{\partial \ln Y}{\partial z} \\ &= zZ_1 + 2z^2 Z_2 + 3z^3 Z_3 + \ldots \end{aligned} \tag{3.269}$$

also leads to a series expansion in $z$. We now make the ansatz

$$z = \frac{N}{Z_1} + \alpha N^2 + \beta N^3 + \ldots \tag{3.270}$$

for $z$ as a function of $T, V, N$. Inserting this into the expansion (3.269) we can determine the coefficients $\alpha, \beta, \ldots$ by comparing the coefficients of powers of $N$. In general this is a complicated procedure for which a systematic strategy should be developed. For the first two virial coefficients, however, we will go through this procedure explicitly:

The coefficients $\alpha$ and $\beta$ are determined from

$$N = \left(\frac{N}{Z_1} + \alpha N^2 + \beta N^3 + \ldots\right) Z_1 + 2\left(\frac{N}{Z_1} + \alpha N^2 + \ldots\right)^2 Z_2$$
$$+3\left(\frac{N}{Z_1} + \ldots\right)^3 Z_3 + \ldots\ . \tag{3.271}$$

The requirement that terms of order $N^2$ have to vanish on the right-hand side implies

$$\alpha Z_1 + 2\,\frac{Z_2}{Z_1^2} = 0\ , \tag{3.272}$$

and a similar requirement for the terms of order $N^3$ leads to

$$\beta Z_1 + 4\alpha\,\frac{Z_2}{Z_1} + 3\,\frac{Z_3}{Z_1^3} = 0\ . \tag{3.273}$$

Thus we find

$$\alpha = -2\,\frac{Z_2}{Z_1^3}\,, \quad \beta = -3\,\frac{Z_3}{Z_1^4} + 8\,\frac{Z_2^2}{Z_1^5}\ . \tag{3.274}$$

With these coefficients we may insert $z$, as given by (3.270), into (3.267), yielding

$$\frac{pV}{k_\mathrm{B}T} = \left(\frac{N}{Z_1} + \alpha N^2 + \beta N^3 + \ldots\right) Z_1 + \left(\frac{N}{Z_1} + \alpha N^2 + \ldots\right)^2 Z_2$$
$$+\left(\frac{N}{Z_1} + \ldots\right)^3 Z_3 + \ldots \tag{3.275}$$

$$= N + N^2\left(-\frac{2Z_2}{Z_1^2} + \frac{Z_2}{Z_1^2}\right)$$
$$+N^3\left(-\frac{3Z_3}{Z_1^3} + \frac{8Z_2^2}{Z_1^4} - \frac{4Z_2^2}{Z_1^4} + \frac{Z_3}{Z_1^3}\right)\,, \tag{3.276}$$

i.e., we obtain the equation of state in the form:

$$p = nk_\mathrm{B}T\ \left(1 + b(T)n + c(T)n^2 + \ldots\right) \tag{3.277}$$

with

$$b(T) = -\frac{Z_2 V}{Z_1^2}\,, \quad c(T) = -2\,\frac{Z_3 V^2}{Z_1^3} + 4\,\frac{Z_2^2 V^2}{Z_1^4}\ . \tag{3.278}$$

This is the virial expansion. $b(T)$ is called the second and $c(T)$ the third virial coefficient.

Let us discuss the expressions for these first nontrivial virial coefficients. First, we find (cp. 3.64)

$$Z_1 = Z(T, V, 1) = \frac{V}{\lambda_\mathrm{t}^3} \,. \tag{3.279}$$

where $\lambda_\mathrm{t}$ is the thermal de Broglie wavelength,

$$\lambda_\mathrm{t} = \sqrt{\frac{h^2}{2m\pi k_\mathrm{B} T}} \,,$$

already introduced in (3.63). The existence of a potential does not have any effect in this expression. Next, we have

$$\begin{aligned} Z(T, V, 2) &= \frac{1}{2h^6} \int \mathrm{d}^6 p \, \mathrm{d}^6 q \; \exp -\beta \left( \frac{p_1^2}{2m} + \frac{p_2^2}{2m} + V_2(\boldsymbol{q}_1 - \boldsymbol{q}_2) \right) \\ &= \frac{1}{2} \lambda_\mathrm{t}^{-6} \int \mathrm{d}^3 q_1 \, \mathrm{d}^3 q_2 \; \mathrm{e}^{-\beta V_2(\boldsymbol{q}_1 - \boldsymbol{q}_2)} \\ &= \frac{1}{2} \lambda_\mathrm{t}^{-6} V \int \mathrm{d}^3 q \; \mathrm{e}^{-\beta V_2(\boldsymbol{q})} \end{aligned} \tag{3.280}$$

and therefore

$$b(T) = -\frac{\left( Z(T, V, 2) - \frac{1}{2} Z^2(T, V, 1) \right) V}{Z^2(T, V, 1)} = -\frac{1}{2} \int \mathrm{d}^3 q \; f(\boldsymbol{q}) \tag{3.281}$$

with

$$f(\boldsymbol{q}) = \mathrm{e}^{-\beta V_2(\boldsymbol{q})} - 1 \,. \tag{3.282}$$

For $c(T)$ one obtains in the same manner

$$c(T) = -\frac{1}{3V} \int \mathrm{d}^3 q_1 \, \mathrm{d}^3 q_2 \, \mathrm{d}^3 q_3 \; f(\boldsymbol{q}_1 - \boldsymbol{q}_2) \, f(\boldsymbol{q}_1 - \boldsymbol{q}_3) \, f(\boldsymbol{q}_2 - \boldsymbol{q}_3) \,.$$

Similarly one can express all higher virial coefficients in terms of the function $f(\boldsymbol{q})$.

We now will calculate the virial coefficients for two potentials explicitly.

**Hard core potential.** We think of atoms as hard cores of radius $\sigma/2$, i.e., for $q \leq \sigma$ we have $V_2(\boldsymbol{q}) = \infty$, $f(\boldsymbol{q}) = -1$. For $q > \sigma$ we take $\beta V_2(\boldsymbol{q}) = \beta V_2(q) \ll 1$ such that $f(\boldsymbol{q}) = -\beta V_2(q)$ is a good approximation for $f(\boldsymbol{q})$.

Under these conditions we get

$$b(T) = -\frac{1}{2}\int \mathrm{d}^3q\, f(\boldsymbol{q}) \tag{3.283}$$

$$= \frac{1}{2}\int_0^{\sigma} \mathrm{d}q\, 4\pi\, q^2 + \frac{1}{2}\,\beta \int_{\sigma}^{\infty} \mathrm{d}q\, 4\pi q^2\, V_2(q) \tag{3.284}$$

$$= 2\pi\,\frac{\sigma^3}{3} + \frac{2\pi}{k_{\mathrm{B}}T}\int_{\sigma}^{\infty} \mathrm{d}q\, q^2\, V_2(q)\ , \tag{3.285}$$

i.e.,

$$b(T) = b_0 - \frac{a}{k_{\mathrm{B}}T}\ , \tag{3.286}$$

with

$$b_0 = 4\,\frac{4\pi}{3}\,\left(\frac{\sigma}{2}\right)^3 \tag{3.287}$$

being four times the volume of one particle, and

$$a = -2\pi \int_{\sigma}^{\infty} \mathrm{d}q\, q^2\, V_2(q)\ . \tag{3.288}$$

For an attractive potential ($V_2(q) < 0$ for $q \geq \sigma$) $a$ is positive.

**Lennard-Jones potential.** A frequently employed model for the interaction between atoms is the Lennard-Jones potential

$$V_2(q) = 4\varepsilon\left(\left(\frac{\sigma}{q}\right)^{12} - \left(\frac{\sigma}{q}\right)^{6}\right)\ . \tag{3.289}$$

Setting $x \equiv \dfrac{q}{\sigma}$ and $T^* = k_{\mathrm{B}}T/\varepsilon$ we obtain for this potential

$$b(T) = -\frac{1}{2}\int \mathrm{d}^3q\, \left(\mathrm{e}^{-\beta V_2(\boldsymbol{q})} - 1\right) = b_0\, b^*(T^*)\ , \tag{3.290}$$

where $b^*(T^*)$ has been introduced as a special function

$$b^*(T^*) = -3\int_0^{\infty} \mathrm{d}x\, x^2\, \left(\mathrm{e}^{-4(x^{-12}-x^{-6})/T^*} - 1\right)\ . \tag{3.291}$$

Equations 3.286 and 3.291 describe approximately the experimental dependence of the second virial coefficient $b(T)$ on temperature, $T$: For decreasing temperature $b(T)$ becomes negative (Fig. 3.4). Only if we compare the results for a very large range of temperature, do we find that different potentials lead to different predictions for $b(T)$. Thus a discrimination between different potentials on the basis of a comparison between theoretical and experimental results is difficult.

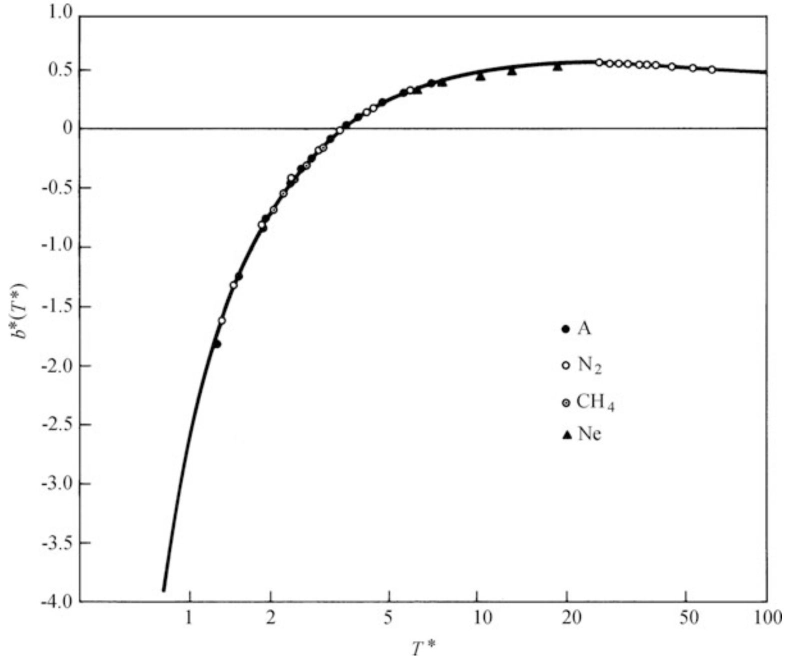


**Fig. 3.4** Reduced second virial coefficient $b^*(T^*)$ as a function of the reduced temperature $T^*$ for the Lennard-Jones potential (*solid line*) as well as some experimental data (From Hirschfelder et al. (1954). Reprinted by permission of John Wiley & Sons, Inc.)

In a similar way one can study $c(T)$ and one finds

$$c(T) = b_0^2 c^*(T^*) , \tag{3.292}$$

where $c^*(T^*)$ is again a universal function (i.e., independent of the parameters of the Lennard-Jones potential).

*Remarks.*

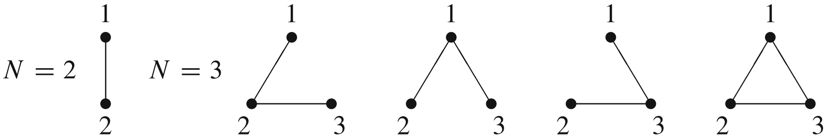
- The higher virial coefficients are obtained systematically by the cluster expansion of Ursell and Mayer (Ursell 1927, Mayer 1941, see also [Römer and Filk 1994]). One writes

$$\begin{aligned} Z(T,V,N) &= \frac{1}{N!\lambda_t^{3N}} \int d^3q_1 \ldots d^3q_N \, \exp\left(-\beta \sum_{i>j} V_2(\boldsymbol{q}_i - \boldsymbol{q}_j)\right) \\ &= \frac{1}{N!\lambda_t^{3N}} \int d^3q_1 \ldots d^3q_N \, \prod_{i>j}(1 + f_{ij}) \\ &= \frac{1}{N!\lambda_t^{3N}} \int d^3q_1 \ldots d^3q_N \left(1 + \sum_{i>j} f_{ij} + \sum_{i>jk>\ell} f_{ij} f_{k\ell} + \ldots\right) \end{aligned}$$

with

$$f_{ij} = \mathrm{e}^{-\beta V_2(\boldsymbol{q}_i - \boldsymbol{q}_j)} - 1 \, .$$

Keeping control of the large number of contributions is greatly facilitated when they are represented graphically. $Z(T, V, N)$, for instance, leads to contributions which for $N = 2$ and $N = 3$ may be represented as follows:



Each circle represents a particle, each line stands for a factor $f_{ij}$. One can show that the contributions to the virial coefficients may also be represented by such objects; in fact the only graphs that occur are those where each pair of points can be connected by at least two independent, nonintersecting paths.

- For $\beta \to 0$ all virial coefficients vanish, because $\mathrm{e}^{-\beta V_2(\boldsymbol{q})} - 1 \to 0$ for $\beta \to 0$ (except in cases, where we assume that $V_2(q) = \infty$ somewhere). In this limit all fluids behave like ideal gases. It may happen that the second virial coefficient also vanishes at a certain finite temperature. In this case the equation of state corresponds, apart from higher order corrections, to that of an ideal gas. This temperature is also called a $\theta$-point.
- The virial expansion is only useful when $n = N/V$ is sufficiently small that only a few virial coefficients are needed, since the determination of the higher coefficients becomes more and more difficult. One would thus like to know whether the equation of state obtained by taking into account only the first four to six virial coefficients also describes the liquid phase of a substance adequately. Essentially, liquids and gases differ only in their densities (whereas the solid state often shows crystalline structures). One might therefore expect that it is possible to determine an equation of state which is valid for fluids in general.

For the hard core gas, studies have shown how far one can get using the method of virial expansion. On the one hand, the virial expansion has been examined by taking into account the first five or six virial coefficients, and this expansion has then been extrapolated using a Padé approximation (Press et al. 2007). In a Padé approximation a polynomal series

$$f(x) = a_0 + a_1 x + \ldots + a_n x^n + O(x^{n+1}) \tag{3.293}$$

is replaced by a rational function

$$\overline{f}_{N,M}(x) = \frac{c_0 + c_1 x + \ldots + c_N x^N}{1 + d_1 x + \ldots + d_M x^M} \, , \tag{3.294}$$

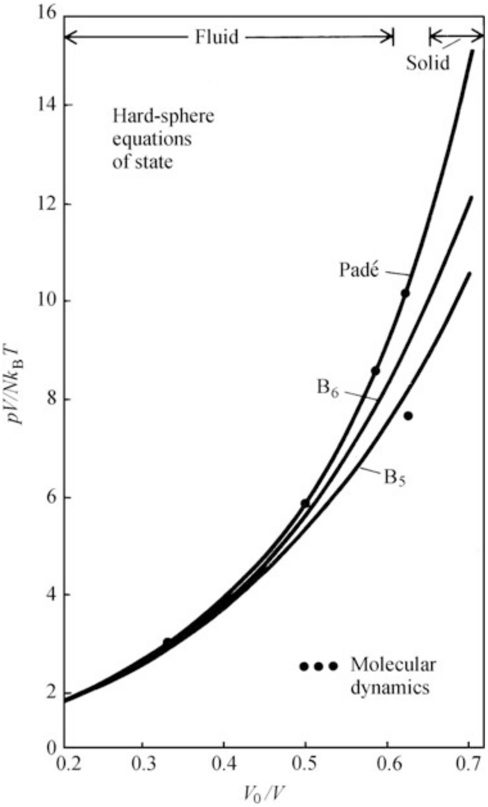


**Fig. 3.5** Graphical representation of $pV/Nk_BT$ as a function of $V_0/V$ for a hard core gas. $V_0$ is the volume of densest packing: $V_0 = N\sigma^3/\sqrt{2}$. The curves are: ($B_5$) virial expansion up to fifth order, ($B_6$) virial expansion up to sixth order, (Padé) Padé approximation. Results from molecular dynamics are indicated by small circles (From Ree and Hoover (1964))

such that the expansion of this function at $x = 0$ coincides with $f(x)$ up to order $x^{N+M}$ (see e.g. Bender and Orszag 1978; Press et al. 2007). For $N = M = 1$ one obtains, e.g.,

$$c_0 = a_0, \quad c_1 - c_0 d_1 = a_1, \quad -d_1 c_1 + d_1^2 c_0 = a_2 \,. \tag{3.295}$$

On the other hand, some points of the equation of state have been determined by a molecular dynamics calculation. In molecular dynamics calculations (see, e.g., Rahman 1964; Verlet 1968) one solves the equations of motion for some hundred particles numerically and regards the macroscopic state variables as the time average determined from the corresponding microscopic quantities.

In Fig. 3.5 the results of such an investigation are compared with those of a virial expansion including a Padé approximation.

Basically we may conclude that the extrapolation to all virial coefficients by Padé approximation yields a qualitatively satisfactory equation of state. Quantitative agreement, however, is not to be expected.

- In (3.254) we gave an equation of state for fluids which we now will write in the form

$$p = nk_{\mathrm{B}}T\left(1 - \frac{n}{6k_{\mathrm{B}}T}\int \mathrm{d}q\, q^2\, 4\pi\, g_2(q)\, q\, V'(q)\right) . \tag{3.296}$$

Expanding $g_2(q)$ in powers of $n$,

$$g_2(q) = g_{20}(q) + n\, g_{21}(q) + O(n^2) , \tag{3.297}$$

and comparing with (3.277) yields

$$b(T) = -\frac{1}{6k_{\mathrm{B}}T}\int \mathrm{d}q\, q^2\, 4\pi\, g_{20}(q)\, q\, V'(q) . \tag{3.298}$$

This result then has to be compared with

$$b(T) = -\frac{1}{2}\int dq\, q^2\, 4\pi\, \left(\mathrm{e}^{-\beta V_2(q)} - 1\right) . \tag{3.299}$$

which by partial integration yields

$$\begin{aligned} b(T) &= -\frac{1}{2}\int_0^\infty \mathrm{d}q\, 4\pi \left(\frac{\mathrm{d}}{\mathrm{d}q}\frac{q^3}{3}\right)\left(\mathrm{e}^{-\beta V_2(q)} - 1\right) & (3.300)\\ &= \frac{1}{6}\int \mathrm{d}q\, 4\pi\, q^2\, q\, \left(-\beta V_2'(q)\right)\, \mathrm{e}^{-\beta V_2(q)} & (3.301)\\ &= -\frac{1}{6k_{\mathrm{B}}T}\int \mathrm{d}q\, q^2\, 4\pi\, q\, V_2'(q)\, \mathrm{e}^{-\beta V_2(q)} . & (3.302)\end{aligned}$$

So we obtain $g_{20}(q) = \mathrm{e}^{-\beta V_2(q)}$. Thus, to first order in $n$, we can set $g_2(q) = \mathrm{e}^{-\beta V_2(q)}$, which is consistent with $g_2(q) \equiv 1$ for ideal gases.

### *3.7.2 Integral Equations for the Radial Distribution Function*

In the literature one can find various integrodifferential equations for the radial distribution function. The solutions of these equations reproduce the experimental results more or less satisfactory. We will briefly sketch the derivation of such integrodifferential equations, but otherwise refer to the literature (McQuarrie 1976; Balescu 1975).

We consider $\varrho_1(\boldsymbol{r})$ from (3.219), i.e.,

$$\varrho_1(\boldsymbol{r}) = \frac{1}{N!h^{3N}Z}\int \mathrm{d}^{3N}p \int \mathrm{d}^3q_2 \ldots \mathrm{d}^3q_N\;\; \mathrm{e}^{-\beta H(p,q)}\big|_{\boldsymbol{q}_1=\boldsymbol{r}} \tag{3.303}$$

for the Hamiltonian function

$$H(p,q) = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m} + \sum_{i>j} V_2(\boldsymbol{q}_i - \boldsymbol{q}_j) + \sum_{i=1}^{N} V_1(\boldsymbol{q}_i) \,. \tag{3.304}$$

Since an external potential $V_1(\boldsymbol{q})$ is now also present, $\varrho_1(\boldsymbol{r})$ is no longer independent of the position $\boldsymbol{r}$. Taking the derivative of $\varrho_1(\boldsymbol{r})$ in (3.303) with respect to $\boldsymbol{r}$ we get

$$\begin{aligned}\nabla_{\boldsymbol{r}}\varrho_1(\boldsymbol{r}) &= -\beta\nabla_{\boldsymbol{r}} V_1(\boldsymbol{r})\varrho_1(\boldsymbol{r}) - \beta\,\frac{1}{h^{3N}N!Z}\int \mathrm{d}^{3N}p \int \mathrm{d}^3 q_2 \ldots \\ &\quad \times \int \mathrm{d}^3 q_N\, \nabla_{\boldsymbol{r}} \sum_{j=2}^{N} V_2(\boldsymbol{r} - \boldsymbol{q}_j)\ \mathrm{e}^{-\beta H(p,q)}\Big|_{\boldsymbol{q}_1=\boldsymbol{r}} \\ &= -\beta\nabla_{\boldsymbol{r}} V_1(\boldsymbol{r})\varrho_1(\boldsymbol{r}) - \beta(N-1) \\ &\quad \times \int \mathrm{d}^3 q_2\, \nabla_{\boldsymbol{r}} V_2(\boldsymbol{r} - \boldsymbol{q}_2)\,\varrho_2(\boldsymbol{r},\boldsymbol{q}_2) \,.\end{aligned}$$

Using

$$N\varrho_1(\boldsymbol{r}) = n_1(\boldsymbol{r})$$

and

$$N(N-1)\varrho_2(\boldsymbol{r},\boldsymbol{r}') = n_2(\boldsymbol{r},\boldsymbol{r}') = n_1(\boldsymbol{r})n_1(\boldsymbol{r}')g_2(\boldsymbol{r},\boldsymbol{r}')$$

one obtains

$$\nabla_{\boldsymbol{r}} n_1(\boldsymbol{r}) = -\beta\nabla_{\boldsymbol{r}} V_1(\boldsymbol{r})n_1(\boldsymbol{r}) - \beta\int \mathrm{d}^3 q_2\, \nabla_{\boldsymbol{r}}\, V_2(\boldsymbol{r} - \boldsymbol{q}_2)\, n_2(\boldsymbol{r},\boldsymbol{q}_2) \,. \tag{3.305}$$

This equation for $n_1(\boldsymbol{r})$ is therefore not closed: On the right-hand side there appears the unknown quantity $n_2(\boldsymbol{r},\boldsymbol{q}_2)$. If one analogously derives a differential equation for this quantity, the next higher moment $n_3(\boldsymbol{r},\boldsymbol{r}',\boldsymbol{r}'')$ appears, etc. Hence, we never get a closed system of equations, but only an infinite hierarchy.

A closed system can obviously only be obtained if at some stage the moment of highest order is approximated by an expression containing only lower moments. Setting, e.g.,

$$n_2(\boldsymbol{r},\boldsymbol{r}') = n_1(\boldsymbol{r})n_1(\boldsymbol{r}') \,, \tag{3.306}$$

leads to the integrodifferential equation for $n_1(\boldsymbol{r})$:

$$\nabla_{\boldsymbol{r}} n_1(\boldsymbol{r}) = \left(-\beta\nabla_{\boldsymbol{r}} V_1(\boldsymbol{r}) - \beta\int \mathrm{d}^3 q_2\, \nabla_{\boldsymbol{r}} V_2(\boldsymbol{r} - \boldsymbol{q}_2)\, n_1(\boldsymbol{q}_2)\right) n_1(\boldsymbol{r})$$

or

$$n_1(\boldsymbol{r}) = \mathrm{e}^{-\beta\left(V_1(\boldsymbol{r}) - \int \mathrm{d}^3 q_2\, V_2(\boldsymbol{r}-\boldsymbol{q}_2)n_1(\boldsymbol{q}_2)\right)} \,. \tag{3.307}$$

A similar factorizing ansatz leading to an equation for $n_2(\boldsymbol{r}, \boldsymbol{r}')$ is

$$n_3(\boldsymbol{r}, \boldsymbol{r}', \boldsymbol{r}'') = \frac{n_2(\boldsymbol{r}, \boldsymbol{r}')n_2(\boldsymbol{r}', \boldsymbol{r}'')n_2(\boldsymbol{r}'', \boldsymbol{r})}{n_1(\boldsymbol{r})n_1(\boldsymbol{r}')n_1(\boldsymbol{r}'')} . \tag{3.308}$$

Equations which may be derived by this or similar assumptions are, e.g.,

- The Born-Green-Yvon (BGY) equation:

$$\begin{aligned} -k_\mathrm{B}T\nabla_1 \ln\big[g_2(\boldsymbol{r}_{12})\big] = {} & \nabla_1 V_2(\boldsymbol{r}_{12}) \\ & + n\int \mathrm{d}^3r_3\, \nabla_1 V_2(\boldsymbol{r}_{13})\, g_2(\boldsymbol{r}_{13})\, g_2(\boldsymbol{r}_{23}) , \end{aligned}$$

- The Percus–Yevick equation for $y(r) = \mathrm{e}^{\beta V_2(r)}\, g_2(r)$:

$$\begin{aligned} y(\boldsymbol{r}_{12}) = 1 + n \int \mathrm{d}^3r_3\, & \left(\mathrm{e}^{-\beta V_2(\boldsymbol{r}_{13})} - 1\right) y(\boldsymbol{r}_{13}) \\ & \times \left(\mathrm{e}^{-\beta V_2(\boldsymbol{r}_{23})}\, y(\boldsymbol{r}_{23}) - 1\right) , \end{aligned}$$

- the ‘hypernetted-chain equation’ (HNC)

$$\ln y(\boldsymbol{r}_{12}) = n\int \mathrm{d}^3r_3\, h(\boldsymbol{r}_{23})\, \left(h(\boldsymbol{r}_{13}) - \ln g_2(\boldsymbol{r}_{13}) - V_2(\boldsymbol{r}_{13})/k_\mathrm{B}T\right) ,$$

  with $h(r) = g_2(r) - 1$. In the last two equations $g_2(r)$ represents the radial distribution function of a macrocanonical system. From these equations one obtains to lowest order in $n$:

$$y \equiv 1 \qquad \text{etc.} \qquad g_2(r) = \mathrm{e}^{-\beta V_2(r)} . \tag{3.309}$$

- The form of the curve $g_2(r)$ as a function of the parameters $n$, $T$, calculated from the integral equations, may be compared with the molecular dynamics calculations for various potentials. Frequently one uses the hard core potential and the Lennard-Jones (12-6) potential (3.289).

  It turns out that the Percus–Yevick equation yields a function $g_2(r)$ and an $n$-dependence of $p/nk_\mathrm{B}T$ which display the best agreement with the data from molecular dynamics (McQuarrie 1976; Balescu 1975; Barker and Henderson 1976).

### 3.7.3 Perturbation Theory

Along with many other approximation methods, the formalism of a perturbation expansion, known from classical mechanics, is also applicable in statistical mechanics.

For the potential we write

$$V(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) = V^0(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) + V^1(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) \,, \tag{3.310}$$

where $V^0(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ is assumed to be a potential for which the partition function $Z_N^0$ is known. Let $\langle . \rangle_0$ be the expectation value which is derived from the density

$$\varrho_0(x) = \frac{1}{N! h^{3N} Z_N^0} \exp\left[-\beta \left(H_{\mathrm{kin}}(p, q) + V^0(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)\right)\right] \,. \tag{3.311}$$

Then

$$\begin{aligned} Z_N &= \int \mathrm{d}^{3N} p \int \mathrm{d}^{3N} q \, \frac{1}{N! h^{3N}} \, \mathrm{e}^{-\beta\left(H_{\mathrm{kin}}(p,q)+V^0(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_N)\right)} \, \mathrm{e}^{-\beta V^1(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_N)} \\ &= Z_N^0 \left\langle \mathrm{e}^{-\beta V^1(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_N)} \right\rangle_0 && (3.312) \\ &= Z_N^0 \left( 1 - \beta \left\langle V^1 \right\rangle_0 + \frac{\beta^2}{2} \left\langle \left(V^1\right)^2 \right\rangle_0 + \ldots \right) \,, && (3.313) \end{aligned}$$

where we have made a high-temperature expansion (i.e., an expansion in $\beta = 1/k_{\mathrm{B}}T$) in the last line. Finally, we obtain

$$F = F_0 + F_1 \tag{3.314}$$

where

$$F_0 = -k_{\mathrm{B}} T \, \ln Z_N^0(T, V, N) \tag{3.315}$$

and

$$\begin{aligned} F_1 &= -k_{\mathrm{B}} T \, \ln\left( 1 - \beta \left\langle V^1 \right\rangle_0 + \frac{\beta^2}{2} \left\langle \left(V^1\right)^2 \right\rangle_0 + \ldots \right) && (3.316) \\ &= \sum_{n=1}^{\infty} \frac{(-\beta)^{n-1}}{n!} \, \omega_n && (3.317) \\ &\equiv \left( \omega_1 - \frac{\omega_2}{2 k_{\mathrm{B}} T} + \ldots \right) && (3.318) \end{aligned}$$

with

$$\begin{aligned} \omega_1 &= \left\langle V^1 \right\rangle_0 && (3.319\mathrm{a}) \\ \omega_2 &= \left\langle \left(V^1\right)^2 \right\rangle_0 - \left\langle V^1 \right\rangle_0^2 \,, && (3.319\mathrm{b}) \end{aligned}$$

etc.

From

$$V^1(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N) = \sum_{i<j} v^1(\boldsymbol{q}_i - \boldsymbol{q}_j) \tag{3.320}$$

one finds

$$\omega_1 = \left\langle V^1 \right\rangle_0 = \frac{N(N-1)}{2} \left\langle v^1(\boldsymbol{q}_1 - \boldsymbol{q}_2) \right\rangle_0 \tag{3.321}$$

$$= \frac{1}{2} n^2 V \int \mathrm{d}^3 q \, v^1(\boldsymbol{q}) \, g_2^0(\boldsymbol{q}) \, , \tag{3.322}$$

where $g_2^0(\boldsymbol{q})$ is the radial distribution function for $V^0(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$.

In the calculation of $\left\langle \left( V^1 \right)^2 \right\rangle_0$ expressions of the form

$$\left\langle v^1(\boldsymbol{q}_1 - \boldsymbol{q}_2) v^1(\boldsymbol{q}_3 - \boldsymbol{q}_4) \right\rangle_0$$

occur. They can only be written in a compact form like $\langle v^1(\boldsymbol{q}_1 - \boldsymbol{q}_2) \rangle_0$ if the corresponding four-particle distribution function $g_4(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3, \boldsymbol{q}_4)$ is used. Such a four-particle distribution function, however, is difficult to calculate, thus revealing the limits of this method.

We now consider, as an example,

$$g_2^0(q) = \begin{cases} 0 \text{ for } q < \sigma \\ 1 \text{ for } q > \sigma \, . \end{cases} \tag{3.323}$$

Then

$$\omega_1 = \frac{1}{2} n^2 V \int_\sigma^\infty \mathrm{d}q \, q^2 \, 4\pi \, v^1(q) = -a \, \frac{N^2}{V} \tag{3.324}$$

with

$$a = -2\pi \int_\sigma^\infty \mathrm{d}q \, q^2 \, v^1(q) \, . \tag{3.325}$$

For an attractive potential ($v^1(q) < 0$ for $q \geq \sigma$) $a$ is positive. This yields, to a first approximation,

$$F = F_0 - a \, \frac{N^2}{V} \, , \tag{3.326}$$

and thus

$$p = -\frac{\partial F}{\partial V} = -\frac{\partial F_0}{\partial V} - a \, \frac{N^2}{V^2} = -\frac{\partial F_0}{\partial V} - \frac{a}{v^2} \, . \tag{3.327}$$

Instead of

$$p = -\frac{\partial F_0}{\partial V} \, , \tag{3.328}$$

which is the equation of state for the potential $V = V^0$, taking into account the potential $V^1$, we now obtain the equation of state

$$\left(p + \frac{a}{v^2}\right) = -\frac{\partial F_0}{\partial V} \; , \tag{3.329}$$

where $a$ is determined from the potential $v^1(\boldsymbol{q})$ according to (3.325). For the attractive two-particle potential $v^1(\boldsymbol{q})$ one finds to first order in the perturbation expansion a decrease of pressure.

A suitable choice of $g_2^0(\boldsymbol{q})$ together with a good approximation for $F_0$ and inclusion of the contributions of higher orders in a high-temperature expansion lead to equations of state which are an improvement over methods where the radial distribution function is calculated from integral equations.

## 3.8 The van der Waals Equation

In the previous section we obtained the equation of state

$$\left(p + \frac{a}{v^2}\right) = -\frac{\partial F_0}{\partial V} \; , \tag{3.330}$$

where $F_0$ is the free energy for a system with potential $V^0(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$. The derivative of this free energy with respect to volume can again be represented as a virial expansion

$$-\frac{\partial F_0}{\partial V} = n k_\mathrm{B} T \; (1 + b(T)n + \ldots) \; . \tag{3.331}$$

We now truncate this expansion after the term $b(T)n$ and set $b(T) = b_0$. For low densities and not too small temperatures this should be a good approximation. If, in addition, we replace the term $1 + b_0 n$ by its Padé approximant

$$1 + b_0 n \to \frac{1}{1 - b_0 n} = \frac{v}{v - b_0} \; , \tag{3.332}$$

we obtain an equation of state of the form

$$\left(p + \frac{a}{v^2}\right) (v - b_0) = k_\mathrm{B} T \; . \tag{3.333}$$

This is the van der Waals equation of state. For most fluids it agrees reasonably well with experiments. It allows the investigation of many phenomena in real fluids, in particular phase transitions.

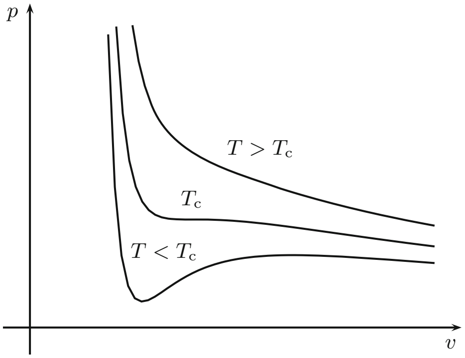


**Fig. 3.6** Isotherms of the van der Waals equation for various values of the temperature

The correction term $v \to v - b_0$, which occurs here as an approximation of the virial expansion, may be explained in another way:

Consider the hard core gas and let $v^0(q) = 0$ for $q \geq \sigma$, where $\sigma/2$ is the radius of the spheres which represent the molecules, and $v^0(q) = \infty$ for $q < \sigma$. Because of the latter assumption the integration over the positions in the canonical partition function

$$Z_N^0 = \frac{1}{N!} \lambda_t^{-3N} \int d^3q_1 \ldots d^3q_N \, e^{-\beta V^0(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_N)} \tag{3.334}$$

has for each integral a volume $V_0$ which is excluded from the integration. So we get

$$Z_N^0 = \frac{1}{N!} \lambda_t^{-3N} (V - V_0)^N \ . \tag{3.335}$$

Of course, $V_0 = N b_0$, where $b_0$ is of the order of the volume of a single molecule, and we obtain

$$F_0 = -k_B T N \ \ln(V - V_0) + \ (\text{terms independent of } V) \ , \tag{3.336}$$

and therefore

$$-\frac{\partial F_0}{\partial V} = \frac{k_B T N}{V - V_0} = \frac{k_B T}{v - b_0} \ . \tag{3.337}$$

## 3.8.1 The Isotherms

The isotherms for a van der Waals gas can easily be represented in a $p$–$v$ diagramm by considering $p$ as a function of $v$ for given $T$. When the temperature is below a critical value $T_c$, the isotherms exhibit two extremal points for $p$ as a function of $v$ (Fig. 3.6).

At $T = T_\mathrm{c}$ the two extrema merge and the curve exhibits a saddle point. This point is determined by

$$\frac{\partial p}{\partial v} = 0 \,, \qquad \frac{\partial^2 p}{\partial v^2} = 0 \,, \tag{3.338}$$

which together with the van der Waals equation of state in the form

$$p = \frac{k_\mathrm{B} T}{v - b_0} - \frac{a}{v^2} \tag{3.339}$$

leads to

$$\frac{-k_\mathrm{B} T}{(v - b_0)^2} + \frac{2a}{v^3} = 0 \tag{3.340}$$

and

$$\frac{2k_\mathrm{B} T}{(v - b_0)^3} - \frac{6a}{v^4} = 0 \,. \tag{3.341}$$

From these two equations for $p$ and $v$ we obtain as the critical temperature $k_\mathrm{B} T_\mathrm{c} = \frac{8}{27} \frac{a}{b_0}$, and the saddle point is at $v = v_\mathrm{c} = 3b_0$, $p = p_\mathrm{c} = \frac{a}{27 b_0^2}$. In particular, we find

$$\frac{k_\mathrm{B} T_\mathrm{c}}{p_\mathrm{c} v_\mathrm{c}} = \frac{8}{3} \approx 2.7 \,. \tag{3.342}$$

For real gases the values of this ratio is found, almost without exception, to be close to 3.4.

Given the free energy of the van der Waals gas as

$$F(T, V, N) = -N k_\mathrm{B} T \ln (V - V_0) - aN \frac{N}{V}, \tag{3.343}$$

one obtains for the Landau free energy (cp. 3.150) per particle

$$\lambda(p, T, v) = p\, v - k_\mathrm{B} T \ln (v - v_0) - \frac{a}{v} \,. \tag{3.344}$$

From the preceding discussion, one may easily derive that for $T < T_c$ an interval for the pressure exists, in which $\lambda(p, T, v)$ possesses two minima, say at $v_1(p, T)$ and at $v_2(p, T)$. Then two free enthalpy functions (and thus two chemical potentials) can be defined, each belongs to one phase. But only the state with the smaller minimum will correspond to a physical equilibrium state. The special value of $p$, for which $\mu_1(p, T) \equiv \lambda(p, T, v_1) = \lambda(p, T, v_2) \equiv \mu_2(p, T)$, is the just the pressure on the vapor pressure curve for the given temperature $T$.

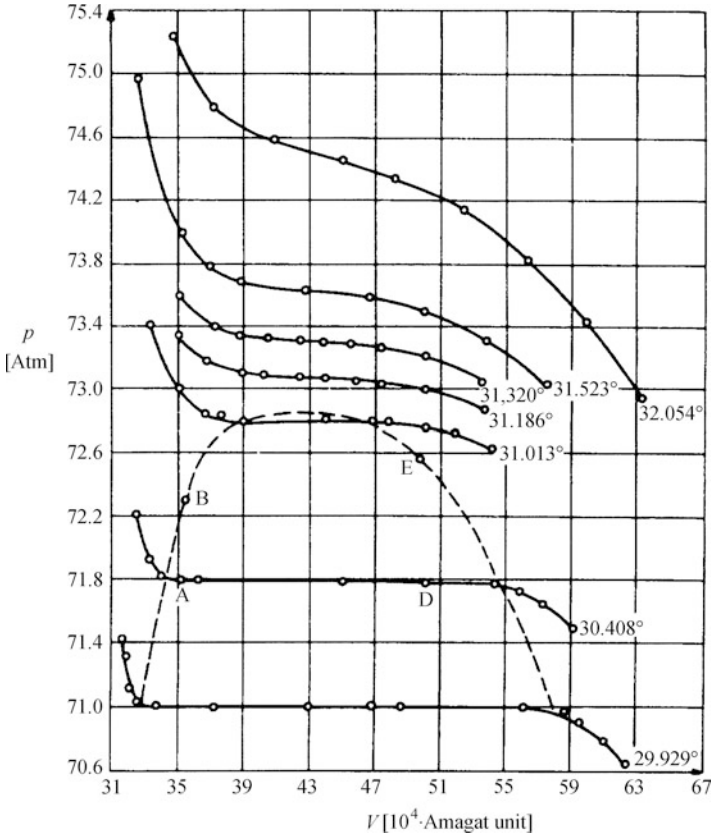


**Fig. 3.7** Isotherms for $CO_2$ at $T < T_c$ close to the critical point $(T_c, p_c, V_c)$. Temperatures are given in degrees Celsius. The Amagat is an old unit for volume and is equal to the volume of a mole of a substance at 0°C and 1 atm, approximately 22.4 L (From Michels et al. (1937))

### 3.8.2 The Maxwell Construction

Not all points on the isotherms for $T < T_c$ correspond to physically possible equilibrium states. Between the two extrema one finds, e.g., $\partial p/\partial v > 0$, while in general $\partial p/\partial v < 0$ has to hold, because a decrease of volume always accompanies an increase of pressure.

So we have to adjust the isotherms by hand. In this region we have to replace them by a horizontal line, as can be seen from the experimental data shown in Fig. 3.7.

This adjustment is called the Maxwell construction and it turns out that it reproduces the physical situation correctly: one replaces the isotherms for $T < T_c$ in the range between the extrema by a horizontal line such that the two shaded areas in the right-hand diagram of Fig. 3.8 are just equal. So we can identify three

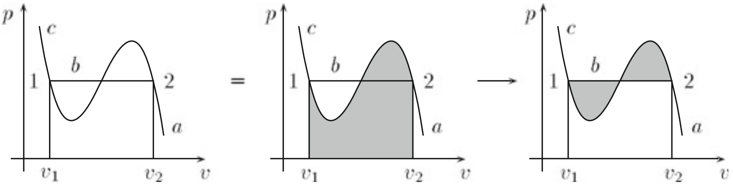


**Fig. 3.8** Maxwell construction for adjusting the isotherms of the van der Waals gas for $T < T_c$. The isotherm between 1 and 2 is replaced by a *horizontal* line such that both *shaded* areas in the picture on the *right* are equal

regions along the isotherms. Region $a$ corresponds to the gaseous phase, region $c$ to the liquid phase, and region $b$, i.e., the horizontal line obtained by the Maxwell construction, corresponds to the coexistence phase where gas and liquid coexist.

The condition that the two shaded areas have to be equal results from the following observation:

Consider a fluid in the liquid phase with values for $(p, v)$ in region $c$. When the volume $v$ is increased, the pressure falls and one reaches a state where the liquid phase and the gaseous phase start to coexist. Let this state correspond to point 1 (Fig. 3.8, left) at a pressure $p_0$. The chemical potential of the liquid is $\mu_1(p_0, T)$. After the liquid has evaporated completely, one reaches point 2, where all the fluid is in the gaseous phase. During this process the pressure remains constant. If we denote the chemical potential of the gas by $\mu_2(p_0, T)$, the relation $\mu_1(p_0, T) = \mu_2(p_0, T)$ had to be valid all the time, as the two substances (liquid and gas) can exchange particles. Denoting the number of molecules at point 1 (all being in the liquid phase) by $N_1$ and their number at point 2 (all in the gaseous phase) by $N_2$, we therefore also have

$$\mu_1 N_1 = \mu_2 N_2 \; . \tag{3.345}$$

Generally, the relation $\mu N = F + pV$ holds, as $K = -pV = E - TS - \mu N$, and therefore we find

$$F_1 - F_2 + p_0(V_1 - V_2) = 0 \; . \tag{3.346}$$

For the free energy of the van der Waals gas we may also write

$$F_1 - F_2 = -\int_{V_1}^{V_2} \mathrm{d}V \, \frac{\partial F}{\partial V} = \int_{V_1}^{V_2} \mathrm{d}V \, p(V) \; . \tag{3.347}$$

On the one hand, therefore, $F_1 - F_2$ is the area between $V_1$ and $V_2$ below the isotherms $p(V)$ (Fig. 3.8, middle), on the other hand, according to (3.346) $F_1 - F_2$ has to be equal to $p_0(V_2 - V_1)$, which is the area under the horizontal line from 1 to 2 at a height $p_0$.

These two areas have to be equal and therefore also the areas in the picture on the right-hand side of Fig. 3.8.

Energy is required, in the form of heat, to take one particle from the liquid phase to the vapor phase in the coexistence region. This heat of conversion heat, which is also called latent heat and which we will denote by $q$, follows from the Clausius–Clapeyron law (see e.g., Honerkamp and Römer 1993) as

$$q = T\Big(v_2(T,p) - v_1(T,p)\Big)\,\frac{\mathrm{d}p(T)}{\mathrm{d}T}\ . \tag{3.348}$$

Here $p(T)$ denotes the vapor pressure curve in a $p$–$T$ diagram, which results from the equation

$$\mu_1(p,T) = \mu_2(p,T)\ . \tag{3.349}$$

Hence, for $T \to T_{\mathrm{c}}$ the latent heat tends to zero, as may readily be observed in the $p$–$v$ diagram, because for $T \to T_{\mathrm{c}}$ the two volumes $v_1$ and $v_2$ become equal.

A phase transition with a nonzero latent heat is also called a phase transition of first order. The difference between $v_2$ and $v_1$ (i.e., between the volume of the fluid when completely in the gaseous phase and the volume of the fluid when completely in the liquid phase), which is present for $T < T_{\mathrm{c}}$, may also be interpreted as the difference between the first derivative with respect to pressure of the chemical potential in the gaseous phase and that in the liquid phase. In a phase transition of second order there are only the second derivatives of the chemical potential with respect to pressure which exhibit a difference. On the other hand, at $T = T_{\mathrm{c}}$ one speaks of a continuous transition.

### 3.8.3 Corresponding States

Introducing the variables

$$\tilde{p} = \frac{p}{p_{\mathrm{c}}}\ , \quad \tilde{v} = \frac{v}{v_{\mathrm{c}}}\ , \quad \tilde{T} = \frac{T}{T_{\mathrm{c}}} \tag{3.350}$$

one obtains the van der Waals equation in the form

$$\left(\tilde{p} + \frac{3}{\tilde{v}^2}\right)(3\tilde{v} - 1) = 8\tilde{T}\ . \tag{3.351}$$

This means that when the pressure, volume, and temperature are measured in units of their critical values, the equation of state assumes a universal form which is the same for all substances. Two fluids with the same values of $\tilde{p}, \tilde{v}$, and $\tilde{T}$ are said to be in corresponding states.

A similar universal form for the equation of state also results if one assumes that the potential for the interactions between the particles is of the form

$$V_2(r) = \varepsilon\, V_2(r/\sigma)\ , \tag{3.352}$$

for all fluids, where only the parameters $\varepsilon$ and $\sigma$ vary from one fluid to another.

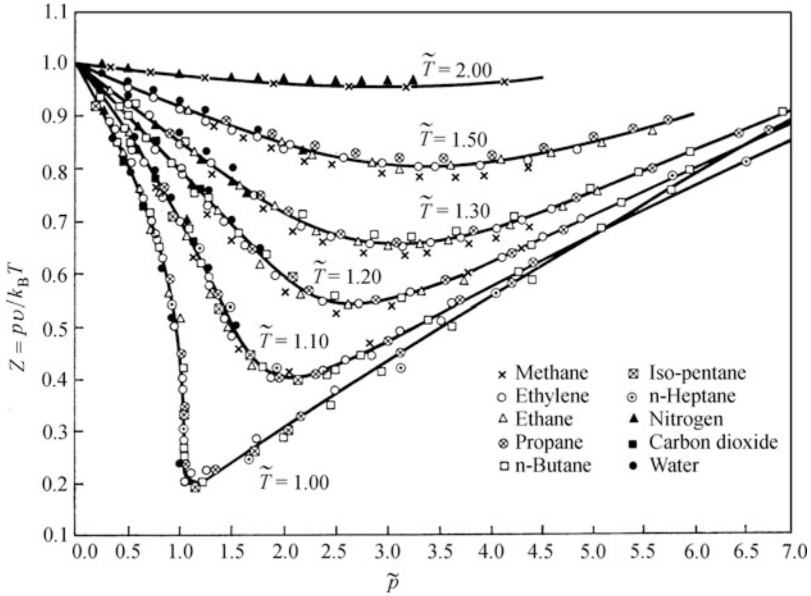


**Fig. 3.9** $Z = pv/k_BT$ as a function of $\tilde{p}$ for different temperatures $\tilde{T}$. The measured values for various fluids at a given temperature $\tilde{T}$ lie on a master curve and therefore confirm the validity of a universal equation of state in the variables $\tilde{p}, \tilde{v}, \tilde{T}$ (From Stanley (1971))

Thus one may suggest that also experimentally a universal equation of state can be confirmed. In order to study this the quantity

$$Z = \frac{pv}{k_BT} = Z_c \frac{\tilde{p}\tilde{v}}{\tilde{T}} \tag{3.353}$$

with

$$Z_c = \frac{p_c v_c}{k_B T_c} \tag{3.354}$$

has been examined for many fluids. In all cases $Z_c$ assumes almost the same value. Many measurements show that the experimental data for $Z$ as a function of $\tilde{p}$ at given temperature $\tilde{T}$ are on the same curve for many fluids (Fig. 3.9). (Note, that $\tilde{v}$ can also be regarded as a function of $\tilde{p}$ and $\tilde{T}$.) So indeed one can formulate the same equation of state,

$$Z = Z_c \frac{\tilde{p}\tilde{v}}{\tilde{T}} \equiv f(\tilde{p}, \tilde{T}) \; , \tag{3.355}$$

for all these fluids.

### 3.8.4 Critical Exponents

The state of a gas at $T = T_c$, $p = p_c$, and $v = v_c$ is called a critical state. At this point two phases, the gaseous and the liquid phase, exist simultaneously, but are no longer distinguishable. This phenomenon has already been mentioned in Sect. 2.7. Above $T_c$ the gas can no longer be liquified and there exists only one equilibrium state. Below $T_c$ there are two equilibrium states (phases) and a region of coexistence where both phases exist and where one can liquify the gas by decreasing the volume, i.e., one can change the relative portion of the phases in the total system.

In the critical state one finds extremely large density fluctuations. This leads to an opacity of the substance, which is also called critical opalescence. The critical state was first observed for ether in 1822 by C. Cagniard de la Tour. For water ($H_2O$) one finds $p_c = 217.7\,\text{atm}$, $T_c = 647.16\,\text{K} = 374°\text{C}$. The behavior of substances near the critical point is quite universal, i.e., large classes of substances show the same behavior close to the critical point.

We now want to examine the behavior of a van der Waals gas close to a critical point. In this context we will give a general characterization of the behavior at a critical point and also introduce the notion of a critical exponent.

At $T = T_c$ the van der Waals gas satisfies near $p = p_c$, $v = v_c$

$$\frac{p - p_c}{p_c^0} = \mathcal{D} \left| \frac{v - v_c}{v_c} \right|^{\delta} \text{sign}(v - v_c) \tag{3.356}$$

with the critical exponent $\delta = 3$ and $\mathcal{D} = -\frac{9}{16}$. Here $p_c^0$ denotes the pressure of a corresponding ideal gas, i.e., $p_c^0 = \frac{k_B T_c}{v_c}$. The critical exponent for the behavior of $p$ as a function of $v$ close to $p = p_c$ is therefore denoted by $\delta$. For a van der Waals gas we find $\delta = 3$. This is evident, since the critical point is just an inflection point. The coefficient is easily determined from the van der Waals equation of state.

We consider $n_{\text{liq}} - n_{\text{gas}}$ for $\varepsilon = \frac{T - T_c}{T_c} < 0$ and $\varepsilon \to 0$. Here $n_{\text{liq}}$ is the density of the liquid and $n_{\text{gas}}$ the density of particles in the gaseous phase. We write the behavior for $T \to T_c$ in the form

$$\frac{n_{\text{liq}} - n_{\text{gas}}}{2n_c} = \mathcal{B}\,(-\varepsilon)^{\beta} \,, \tag{3.357}$$

where $\beta$ is a second critical exponent. In a van der Waals gas we find $\mathcal{B} = 2$ and $\beta = \frac{1}{2}$.

For the specific heat $C_V$ at $v = v_c$ and $T \to T_c$ the critical exponents are defined according to

$$C_V \sim \begin{cases} (-\varepsilon)^{-\alpha'} & \text{for } T < T_c \,,\ \varepsilon < 0 \\ \varepsilon^{-\alpha} & \text{for } T > T_c \,,\ \varepsilon > 0 \,. \end{cases} \tag{3.358}$$

In a van der Waals gas the specific heat tends to a constant as $T \to T_c \pm 0$; therefore $\alpha$ and $\alpha'$ vanish.

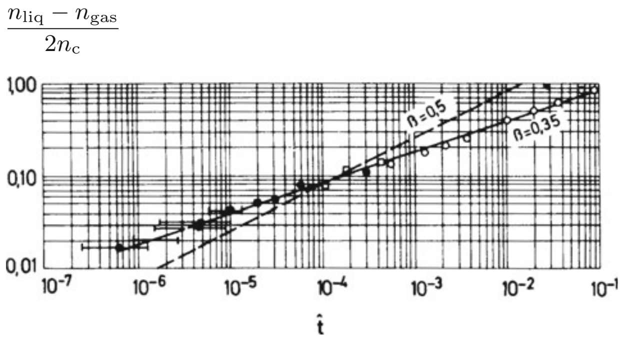


**Fig. 3.10** Temperature dependence of the density difference $(n_{\rm liq} - n_{\rm gas})/2n_{\rm c}$ for $CO_2$ in logarithmic coordinates. From the slope of the *straight* line one obtains in this case $\beta = 0.35$. The prediction for the van der Waals gas is $\beta = 0.5$, corresponding to the *dashed* line, where no data are actually found (From Heller (1967))

For the isothermal compressiblity one finds

$$\kappa_T = -\frac{1}{V}\frac{\partial V(T,p,N)}{\partial p} \ . \tag{3.359}$$

We consider

$$\left.\frac{1}{v_{\rm c}\kappa_T}\right|_{v=v_{\rm c}} = \left.\frac{\partial p(T,v,N)}{\partial v}\right|_{v=v_{\rm c}} \tag{3.360}$$

in the limit $T \to T_{\rm c} + 0$, i.e., we examine the way the slope of the isotherm tends to zero as $T \to T_{\rm c}$ for $T > T_{\rm c}$. Let $\kappa_T^0$ be the isothermal compressibility of the ideal gas, i.e., $\kappa_T^0 = 1/p_{\rm c}$. We write

$$\frac{\kappa_T}{\kappa_T^0} = C\,\varepsilon^{-\gamma}, \qquad \varepsilon = \frac{T - T_{\rm c}}{T_{\rm c}} \ . \tag{3.361}$$

For a van der Waals gas we get $C = \frac{4}{9}$ and the critical exponent $\gamma = 1$. The slope of the isotherm tends to zero linearly in $\varepsilon$ and therefore $\kappa_T$ diverges as $\varepsilon^{-1}$.

We have thus now defined the critical exponents

$$\alpha\ ,\ \alpha'\ ,\ \beta\ ,\ \gamma\ ,\ \delta\ . \tag{3.362}$$

For the van der Waals gas we have found the respective values $0,\ 0,\ \frac{1}{2},\ 1,\ 3$.

In a log–log plot one can easily extrapolate the critical exponents from the slope of the experimental data (see Fig. 3.10). More or less independent of the substance under consideration one finds:

$$\alpha,\ \alpha' < 0.4 \qquad \beta \approx 0.35 \qquad \gamma \approx 1.2 - 1.3 \qquad \delta \approx 4.2 - 4.4\ . \tag{3.363}$$

These values agree surprisingly well with the predictions from the van der Waals gas. Taking into account the simple structure of the van der Waals ansatz, one does not expect a better agreement.

Thus, one essential property of the critical exponents is that they can be measured easily. Furthermore, it turns out that certain relations among these quantities can be derived (Sect. 2.6). These are the so-called scaling relations, for example,

$$\alpha + \beta(1+\delta) = 2 \; , \qquad \alpha + 2\beta + \gamma = 2 \; . \tag{3.364}$$

In general, these are very well obeyed, because the critical exponents are even more universal than the equations of state for the reduced variables $\tilde{p}, \tilde{T}, \tilde{V}$. This implies that at the critical point the microscopic properties of a substance are no longer relevant. Long-range correlations become dominant and render the short-range local microstructure unimportant.

We will return to the subject of phase transitions in the framework of spin models in Sect. 4.4.

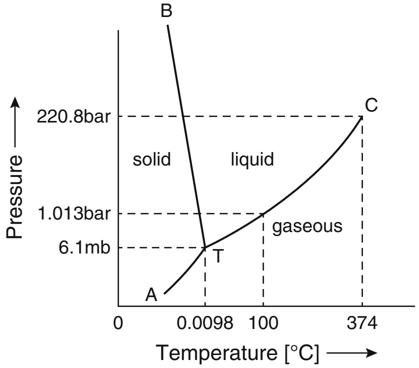
## 3.9 Some General Remarks about Phase Transitions and Phase Diagrams

A phase is homogeneous down to the molecular realm. A homogeneous mixture of two different chemical substances also represents a phase. For example, in a system composed of a variety of substances, there is only one gaseous phase, since gases are perfectly miscible. A sugar solution (sugar and water) also represents a phase. If this solution is cooled to a low enough temperature, the sugar begins to precipitate. Thus, two phases are obtained, a solid, pure sugar phase, and a sugar solution with a lower concentration of sugar.

An inhomogeneous system with several phases is called a multiple-phase system. Let us consider first a two-phase system composed of a single material, e.g., the system (water, water vapor). "Vapor" means the gaseous phase of a substance when this phase exchanges energy or mass with the liquid or solid phase, or when it can be transformed into another phase without too great a change in volume or temperature. Vapor is a real gas, so that relationships between state variables, i.e., state equations, will differ strongly from the equations of an ideal gas. However, if the vapor is separated from the other phases and progressively heated, the resulting superheated vapor will behave more like an ideal gas. In the two-phase system we are considering, vapor and another phase are in contact, which means that matter is continuously exchanged between the two phases. Molecules emerge out of their bound states in the solid or liquid phase into vapor, and in the other direction, molecules of vapor will be absorbed into the solid or liquid phase.

Thermodynamic equilibrium is established in a closed two-phase system if both phases are at the same temperature and pressure and if the number of particles

**Fig. 3.11** Phase diagram of water at lower pressures



of vapor and of the other phase have a constant statistical average, i.e. when the chemical potentials of both phases are the same. The vapor is then called saturated. Let $p$ and $T$ be the common pressure and temperature, $\mu(p, T)$ the chemical potential of water, and $\mu'(p, T)$ the chemical potential of water vapor. Then, in thermodynamic equilibrium,

$$\mu(p, T) = \mu'(p, T) . \tag{3.365}$$

Let us assume that these functions are given or have been calculated. In the $p$–$T$ plane, (3.365) defines a curve which contains all the points $(p, T)$ at which water and water vapor can exist in equilibrium (if this equation is satisfied identically for all $p$ and $T$, it is then no longer meaningful to talk of two distinguishable phases).

If there are three phases of a substance (e.g., if there is ice in addition to the water and water vapor), they can be in equilibrium if

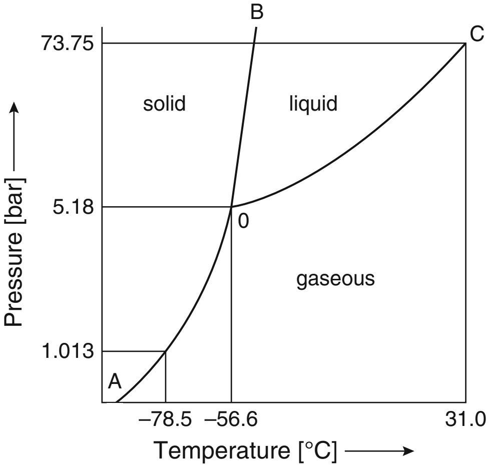
$$\mu(p, T) = \mu'(p, T) = \mu''(p, T) , \tag{3.366}$$

where $\mu''(p, T)$ represents the chemical potential of the third phase. Since we have two equations for two variables, they will normally determine a single point in the $p$–$T$ diagram, the so-called triple point.

In (Fig. 3.11), we show a typical phase diagram, namely, that of water. We see that, at T $=$ 100°C and $p = 1\,\text{atm} = 1.013\,\text{bar}$, water and water vapor are in thermodynamic equilibrium, while at lower pressures the equilibrium temperature has lower values, as we know from experience. (On a mountain, due to the lower pressure, water boils at temperatures under 100°C.) We also see that the solid phase, ice, is in equilibrium with the liquid phase at 0°C and the normal pressure 1.013 bar. If the pressure is raised, the temperature at which the two phases can be at equilibrium drops. Ice melts under the blade of an ice skate; glaciers can move slowly into valleys.

This behavior, where the melting point decreases with an increase in pressure, is rare. It occurs also in bismuth, but other materials always demonstrate the reverse

**Fig. 3.12** Phase diagram of $CO_2$ (not to scale)



behavior; the melting temperature increases with growing pressure. The triple point of water is at 0.0098°C and 4.58 Torr = 6.11 mbar. Since the triple point of water is uniquely defined and can be easily reproduced, it is used as a reference point for the temperature scale. Below 6.11 mbar there is no longer a liquid phase. Ice transforms directly to vapor when the temperature is raised. This process is called sublimation. This process is much better known in the case of dry ice (solid $CO_2$). If dry ice is heated at a pressure of 1.013 bar = 1 atm (Fig. 3.12), then at the temperature −78.5°C it changes directly into a gaseous state.

The last characteristic point in our phase diagram (Fig. 3.11) is the critical point C at $T_{\text{crit}} = 374°\text{C}$. Above this temperature, the chemical potentials of water and water vapor become identical as functions, that is, there is no longer any difference between the phases, and it no longer makes any sense to distinguish between them. This means the following: for $T < 374°\text{C}$, $\mu(p, T)$ is defined only above the curve TC and to the right of TB, and $\mu'(p, T)$, the chemical potential of water vapor, is defined only under the curve TC to the right of TA. For $T > 374°\text{C}$, $\mu(p, T) = \mu'(p, T)$, and they are both defined for all $p$. (This fact can also be formulated in another way. It is possible to define a chemical potential $\mu(p, T)$ for the substance $H_2O$ for all physical $p$ and $T$. This function of two variables is then continuous on the lines TA, and TC, but its derivatives can be discontinuous there.

The curve TC in Fig. 3.11 is called the vapor pressure curve. The pressure $p(T)$ on this curve is called the vapor pressure, because it is the pressure that vapor has at temperature $T$ when it is in equilibrium with the liquid phase. Figure 3.13 shows a phase diagram for $H_2O$ at higher pressures. We see that there can be several triple points; the different phases here are distinguished by their crystalline structures.

We ascertain from these diagrams that in one-phase regions, there are two variables which can be varied independently, while in a two-phase region, there is only one free variable. In a three-phase region, no variables are free. In an $m$-phase region there are thus $(3 - m)$ free variables.

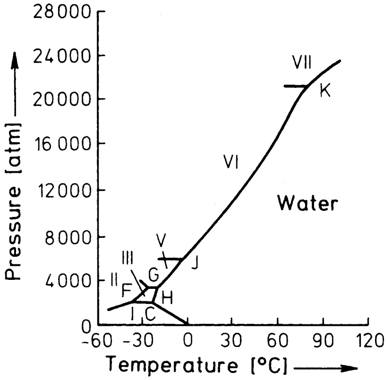


**Fig. 3.13** Phase diagram of water at higher pressures

We can generalize this statement for mixtures of $n$ substances. It can be shown that if $m$ phases are in equilibrium, there are

$$f = n + 2 - m \tag{3.367}$$

free variables. This is the Gibbs phase rule.

*Example.* Consider the two materials $H_2O$ and NaCl and their phases: ice, solid NaCl, saturated NaCl solution, and vapor. If all four phases are in equilibrium, there are no longer any free variables.

# Chapter 4
# Random Fields: Textures and Classical Statistical Mechanics of Spin Systems

If one associates a random variable with each point of a lattice in a one-, two-, or, in general, $n$-dimensional space the collection of these random variables is called a random field. The interdependence among the random variables of such a field will be significant in determining the properties of macroscopic quantities of such random fields. The interdependence will be restricted in Sect. 4.1 in such a way that we can formulate relatively simple models known as Markov fields. A second approach to simple models for random fields, consisting in the introduction of Gibbs fields, will prove to be equivalent. In Sect. 4.2, examples of Markov random fields will be presented. It will turn out that all classical spin models relevant in statistical mechanics are Markov random fields. Moreover, many two- or three-dimensional images, so-called textures, may also be interpreted as realizations of such fields.

Section 4.3 will provide many of the formulas which will be needed subsequently. The free energy as the generating function of the cumulants is introduced, as are the moments, covariances, susceptibilities, and the entropy. The relation between susceptibilites and covariances will emerge once more.

In Sect. 4.4, we will first treat two simple models for random fields. For the white noise field, where all random variables are considered as independent and which corresponds to the ideal paramagnetic system, all quantities can be calculated analytically. The same is true for the Ising model in one dimension.

In Sect. 4.5 we discuss solvable models with a phase transition: the Curie–Weiss model, the general model with summable coupling constants in the mean-field approximation, and the two-dimensional Ising model. It will easily be seen that the magnetization in the Curie–Weiss model possesses the large deviation property; the phenomenon of phase transitions can then be demonstrated most clearly and critical exponents can be determined. The connection between phase transitions and symmetry breaking will be discussed,as will the appearance of long-range order in the case of two equilibrium states. In connection with the Ising model, we will cast a glance at more general solvable models such as vertex models.

The models treated in Sect. 4.5 also serve to illustrate some approximations and phenomenological theories which will then be discussed in Sect. 4.6: the Landau

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_4, © Springer-Verlag Berlin Heidelberg 2012

free energy and Landau's phenomenological model for a phase transition of second order.

The topic of Sect. 4.7 is the renormalization group method, which was already introduced in Sect. 2.6 and which is known to be an important tool for investigating the properties of systems at a phase transition. Scaling properties are derived and their consequences for critical exponents will be discussed.

## 4.1 Discrete Stochastic Fields

We will begin by considering the discrete, Euclidean, finite or infinite space in one to three dimensions. Such a space will be denoted by $S$. The discrete points in this space are denoted by $\boldsymbol{r}$, or sometimes $\boldsymbol{r}_i, i = 1, \ldots, N$, if we think of these points as being listed in a sequence. Each point may also be characterized by its number $i$ in this list when the coordinates, given by $\boldsymbol{r}_i$, are unimportant. Thus we will denote a field by $\{X(\boldsymbol{r})\}$ or $\{X(\boldsymbol{r}_i), i = 1, \ldots, N\}$ or $\{X_i, i = 1, \ldots, N\}$.

Possible realizations of $X$ at the positions $\boldsymbol{r}$ or $\boldsymbol{r}_i$ are, for instance, gray-level intensities or the color code of a picture, the velocity of a fluid at position $\boldsymbol{r}$, or the orientation of a spin. Although $X$ may be vector-valued, this need not affect the notation for the moment.

The space of possible realizations at a single point will be denoted by $\mathcal{L}$. A realization of the total field is then an element in

$$\mathbb{F} = \underbrace{\mathcal{L} \otimes \mathcal{L} \otimes \ldots \otimes \mathcal{L}}_{N \text{ times}} . \tag{4.1}$$

When $\mathcal{L}$ has the dimension $M$, $\mathbb{F}$ is $M^N$ dimensional. We will call such a realization a pattern.

Two examples of random fields are the following.

(a) A solid may be considered as a system of atoms, vibrating around the points of a periodic three-dimensional lattice. The influence of the vibrations on the properties of the solid will be studied in Sect. 6.6. Here we want to investigate the magnetic properties of such a system of atoms each with a spin $s$. So we consider a lattice, i.e., a periodic arrangement of $N$ points in space, and at each point of this lattice there is an atom with spin $s$. (This, however, is not exactly the physical picture which emerges in deriving such models, see Sect. 6.9.) With respect to an external axis this spin can assume the values $\{-s, -s+1, \ldots, +s\}$. Hence we may consider the field of spins at the lattice points as a random field. The pattern is a spin configuration.

(b) Certain images, which are referred to as textures, may also be considered as realizations of random fields. Take, for example, a two-dimensional regular square lattice of 640 × 480 points. To each point is assigned a value 0 or 1, according to which it will be represented as black or white, respectively. The full

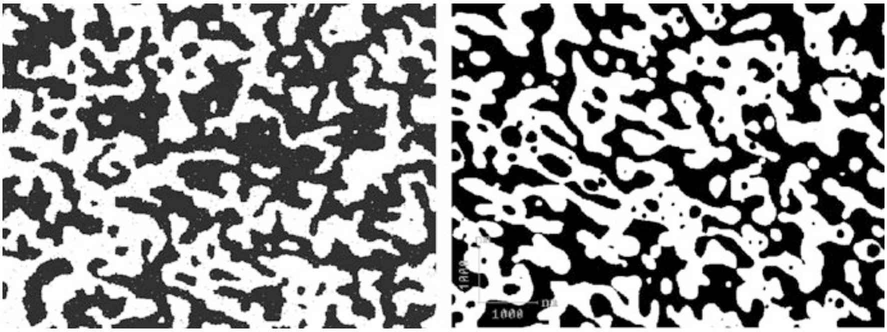


**Fig. 4.1** Typical images, which are called textures: (*Left*) a realization of a random field; (*Right*) image of a sectional view through a polymer mixture with two components

pattern corresponds to a black and white picture. For a color picture one assigns to each point a three-component color-vector, where the respective values of the components are in the interval $(0, 1, \dots, M), M = 2^n - 1, n = 2, \dots,$ and each component represents the intensity of one of the three basic colors. In a gray-level picture all three components are equal, i.e., one has $M$ different gray-levels.

The content of images is expressed in the correlations among the random variables for the single points or pixels. For pictures showing well-defined objects, these correlations vary strongly with the lattice site. If, however, the correlations are independent of the position, these images are called *textures*. In Fig. 4.1 two such textures are shown: on the left-hand side a realization of a random field, known as the Gibbs field, which we will introduce below, and on the right-hand side the image of a section through a mixture of two polymers.

### 4.1.1 Markov Fields

The properties of the random field are determined largely by the dependencies among the random variables of the field $\{X(\boldsymbol{r})\}$. These are described by the common densities

$$\varrho_n(x(\boldsymbol{r}_1), \dots, x(\boldsymbol{r}_n)) \equiv \varrho_n(x_1, x_2, \dots, x_n) \tag{4.2}$$

for all possible $1 \leq n \leq N$. Of course, from $\varrho_N(x_1, x_2, \dots, x_N)$ all densities $\varrho_n$ for $n < N$ can be derived.

To obtain simpler models for the common density one has to restrict the set of dependencies by further conditions. In order to formulate such conditions one first has to define for each point $\boldsymbol{r}_i$ in $S$ a neighborhood $\mathcal{N}_i$, i.e., the set of all points

$\neq \boldsymbol{r}_i$ which are considered neighbors of $\boldsymbol{r}_i$:

$$\mathcal{N}_i = \left\{ \boldsymbol{r} \mid \boldsymbol{r} \neq \boldsymbol{r}_i \text{ is neighbor of } \boldsymbol{r}_i \right\} . \tag{4.3}$$

The definition of neighborhood should be such that for $\boldsymbol{r}_{i'} \in \mathcal{N}_i$ also $\boldsymbol{r}_i \in \mathcal{N}_{i'}$.

*Examples.*

- On a one-dimensional lattice $S$ of points $\{r_i, i = 1, \ldots, N\}$ one may take, for example,

  $$\mathcal{N}_i = \{r_{i-1}, r_{i+1}\} \quad \text{for } 2 \leq i \leq N-1, \tag{4.4}$$

  i.e., both nearest neighbors to the right and left of $r_i$ belong to the neighborhood. Of course, one might also define the two next-to-nearest neighbors as part of the neighborhood. The boundary points $r_1$ and $r_N$ have fewer neighbors, only $r_2$ and $r_{N-1}$, respectively. But one can also define $r_1$ and $r_N$ as neighbors such that e.g. $\mathcal{N}_1 = \{r_N, r_2\}$.
- On a two-dimensional lattice $S$ of points $\{\boldsymbol{r}_i, i = 1, \ldots, N\}$ one may define as the neighborhood of $\boldsymbol{r}_i$:

  $$\mathcal{N}_i = \{\boldsymbol{r}_{i'} \,\big|\, |\boldsymbol{r}_{i'} - \boldsymbol{r}_i| \leq r \,, \, i \neq i'\} . \tag{4.5}$$

  All points within a distance $r$ from $\boldsymbol{r}_i$ belong to the neighborhood of $\boldsymbol{r}_i$. The point $\boldsymbol{r}_i$ itself is by definition not part of its neighborhood.

Apart from the notion of neighborhood we also need the concept of a clique. When a neighborhood $\mathcal{N}_i$ is defined for each point $i$ of a lattice, a clique consists of a set of points $\{i, i', \ldots\}$ which are mutual neighbors of each other. The simplest cliques are those consisting of a single lattice point $\{i\}$; the set of all these will be denoted by $\mathcal{C}_1$. $\mathcal{C}_2$ shall then be the set of all cliques which each contain two neighboring lattice points, and

$$\mathcal{C}_3 = \{\{i, i', i''\} \mid i, i', i'' \quad \text{are mutual neighbors}\} \tag{4.6}$$

is the set of all cliques containing three neighboring lattice points. Similarly one may define cliques of higher order. For limited neighborhoods there exist only cliques up to a certain degree. For a system of neighborhoods where each lattice point has four neighbors as in Fig. 4.2a, there are only the cliques shown in Fig. 4.2c–e. For a configuration where each lattice point has eight neighbors as in Fig. 4.2b, all possible cliques are shown in Fig. 4.2c–l.

Now let $X_S$ be the set of random variables defined on the space $S$, then $X_{\mathcal{N}_i}$ is the set of random variables defined on the points in the neighborhood of $\boldsymbol{r}_i$, $X_{S\setminus\{\boldsymbol{r}_i\}}$ is the set of all random variables on $S = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N\}$ with the exception of $X(\boldsymbol{r}_i) \equiv X_i$. Similarly we denote by $x_S, x_{S\setminus\{\boldsymbol{r}_i\}}$ the respective set of realizations.

Now we introduce the following definition: A stochastic field is a Markov field on the discrete space $S$ if

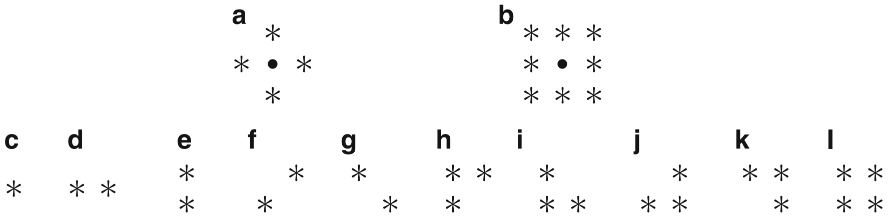


**Fig. 4.2** *Left*: System where the lattice point (●) has four neighbors (∗) in (**a**) and possible cliques in (**c**)–(**e**). *Right*: System with eight neighbors in (**b**) and possible cliques in (**c**)–(**l**)

$$\varrho(x_i \mid x_{S\setminus\{\boldsymbol{r}_i\}}) = \varrho(x_i \mid x_{\mathcal{N}_i})\ , \tag{4.7}$$

i.e., if the conditional probabilities for $X_i$ only depend on the realizations of the random variables in the neighborhood of $\boldsymbol{r}_i$.

### *4.1.2 Gibbs Fields*

Gibbs fields are stochastic fields whose common density has the following special structure

$$\varrho(x_S) = \frac{1}{Z}\mathrm{e}^{-\beta H(x_S)}\ . \tag{4.8}$$

Here $\beta$ is a factor which we might set equal to $1/k_{\mathrm{B}}T$, and in analogy to statistical mechanics $k_{\mathrm{B}}T$ is referred to as Boltzmann's constant times temperature. $Z$ is a normalization factor which, like in statistical mechanics, is generally called the partition function and which is determined from the normalization condition

$$Z = \sum_{x_S \in \mathbb{F}} \mathrm{e}^{-\beta H(x_S)}\ . \tag{4.9}$$

$H(x_S)$ is the Hamiltonian function or energy function, which for Gibbs fields should be of the form

$$H(x_S) = \sum_{i \in \mathcal{C}_1} V_i(x_i) + \sum_{\{i,j\} \in \mathcal{C}_2} V_{ij}(x_i, x_j) + \dots\ . \tag{4.10}$$

The energy function is thus a sum of one-point potentials, two-point potentials or pair-potentials, and higher-order potentials, i.e., of expressions which depend on cliques of definite order.

A Gibbs field is called homogeneous, if $V_i \equiv V_1$ is independent of $i$ and $V_{ij} \equiv V_2$ is independent of the position of the clique $(i, j) \in \mathcal{C}_2$. When only cliques with two or fewer lattice points are taken into account, the energy function $H$ can also be written as

$$H(x_S) = \sum_{i=1}^{N} V_1(x_i) + \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} V_2(x_i, x_j) \; . \tag{4.11}$$

This form of the density for a random field corresponds, of course, exactly to the canonical density for a classical fluid (cf. (3.5) and (3.6)), the only difference being that here the random variables are not positions and momenta but orientations of spins or certain intensities at some lattice point. We will show in Sect. 4.2 that such models can also be formulated in the context of the statistical mechanics of solids.

### *4.1.3 Equivalence of Gibbs and Markov Fields*

We have introduced Markov fields using the (local) Markov condition, and Gibbs fields by specifying the complete common density. Now, according to the Hammersley–Clifford theorem (Hammersley and Clifford 1971) each Markov field with a system of neighbors and the associated system of cliques is also a Gibbs field with the same system of cliques, and, vice versa, each Gibbs field is also a Markov field with the corresponding system of neighbors. Thus for any common density of the form (4.8) one finds that all conditional probabilities satisfy the Markov condition, and from the Markov condition for a random field follows that the common density has the form (4.8). A proof of this theorem can be found in many articles (Besag 1974). Here we will be content to show that a special Gibbs field is also a Markov field.

*Example.* We consider a Gibbs field with

$$V_1(x_i) = \frac{(x_i - \mu_i)^2}{2\sigma^2} \quad i = 1, \ldots, N \tag{4.12}$$

and

$$V_2(x_i, x_j) = J_{ij} \frac{(x_i - \mu_i)(x_j - \mu_j)}{2\sigma^2} \quad , \text{ for } i \neq j \; , \tag{4.13}$$

where $J_{ij} \neq 0$, if $j \in \mathcal{N}_i$. Without loss of generality, we set $\beta = 1$ and therefore obtain

$$\varrho(x_S) = \frac{\sqrt{\det \mathsf{B}}}{\sqrt{(2\sigma^2)^N}} \exp\left( -\frac{1}{2\sigma^2} \sum_{i,j=1}^{N} (x_i - \mu_i) \mathsf{B}_{ij} (x_j - \mu_j) \right) \; . \tag{4.14}$$

The matrix $\mathsf{B}$ has the matrix elements $\mathsf{B}_{ij}$ with

$$\mathsf{B}_{ij} = 1, \; i = j = 1, \ldots, N \tag{4.15}$$

$$\mathsf{B}_{ij} = J_{ij}, \; i \neq j, \; i, j = 1, \ldots, N. \tag{4.16}$$

We now consider the conditional probability

$$\varrho(x_i \mid x_{S\setminus\{i\}}) = \frac{\varrho(x_S)}{\varrho(x_{S\setminus\{i\}})} \ . \tag{4.17}$$

Writing out explicitly those terms in the numerator and denominator which contain $x_i$, we obtain

$$\varrho(x_i \mid x_{S\setminus\{i\}}) = \frac{\exp\left(-\frac{1}{2\sigma^2}\left((x_i-\mu_i)^2 + 2(x_i-\mu_i)A_i\right)\right)\mathrm{e}^{-\Sigma'}}{\left(\int \mathrm{d}x_i\ \exp\left(-\frac{1}{2\sigma^2}\left((x_i-\mu_i)^2 + 2(x_i-\mu_i)A_i\right)\right)\right)\mathrm{e}^{-\Sigma'}}$$

with

$$A_i = \sum_{j\in\mathcal{N}_i} J_{ij}(x_j-\mu_j) \ .$$

The terms $\mathrm{e}^{-\Sigma'}$ in the numerator and denominator do not contain any $x_i$ and mutually cancel.

Therefore, $\varrho(x_i|x_{S\setminus\{i\}})$ depends only on $\mathcal{N}_i$, and one also obtains the explicit form of $\varrho(x_i|\mathcal{N}_i)$. To within a numerical factor, the integral in the denominator yields $\exp(A_i^2/2\sigma^2)$, and so one finally obtains in this model

$$\varrho(x_i|\mathcal{N}_i) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu_i+\sum_{j\in\mathcal{N}_i} J_{ij}(x_j-\mu_j)\right)^2\right) \ , \tag{4.18}$$

where we have determined the prefactor from the normalization condition.

This calculation also conveys an idea of the general proof of the statement that each Gibbs field is also a Markov field.

## 4.2 Examples of Markov Random Fields

In this section we will introduce some of the main models of random fields.

### *4.2.1 Model with Independent Random Variables*

In the simplest model all random variables at different positions are mutually independent. Then $V_2 \equiv 0$ and

$$\varrho(x_S) = \prod_{i=1}^{N} \frac{1}{Z_1}\,\mathrm{e}^{-\beta V_1(x_i)} \equiv \prod_{i=1}^{N} \varrho_1(x_i) \tag{4.19}$$

with

$$\varrho_1(x_i) = \frac{1}{Z_1}\, \mathrm{e}^{-\beta V_1(x_i)} \; . \tag{4.20}$$

In a stochastic process this corresponds to white noise.

We may further assume that $V_1(x_i)$ is quadratic in $x_i$ such that $\varrho(x_i)$ is a Gaussian density; in this case one speaks also of (Gaussian) white noise. Only if $V_2 \neq 0$, i.e., in the presence of an interaction, is there a dependence among the random variables.

As an example of such a model with independent random variables we consider a solid consisting of atoms with spin $s$ due to its electrons. Each spin has a magnetic moment $\boldsymbol{\mu} = g\mu_\mathrm{B}\boldsymbol{S}$, where $g$ is the Landé factor, $\mu_\mathrm{B}$ Bohr's magneton, and $\boldsymbol{S}$ the spin operator. The solid may be in a constant external field $\boldsymbol{B}$. Let us assume that the energy function is simply

$$H = -\sum_{i=1}^{N} \boldsymbol{\mu}_i \cdot \boldsymbol{B} = -g\mu_\mathrm{B} \sum_{i=1}^{N} \boldsymbol{B} \cdot \boldsymbol{S}_i = -g\mu_\mathrm{B} B \sum_{i=1}^{N} (S_i)_z \; , \tag{4.21}$$

where $\boldsymbol{S}_i$ is the spin of the atom at lattice point $i$ and the index $i$ runs over all $N$ lattice sites. We have chosen the $z$-axis parallel to the $\boldsymbol{B}$ field. The eigenvalues of $(S_i)_z$ are $m_i = -s, \ldots, +s$ for each $i$ and thus there are $(2s + 1)^N$ possible configurations for the spin field $(S_1, \ldots, S_N)$.

This is a model for the ideal paramagnetic crystal, where here "ideal" means "free of interactions". The energy function for this random field may also be written as

$$H = -\sum_{i=1}^{N} \Theta x_i \; , \tag{4.22}$$

where we have set $\Theta = g\mu_\mathrm{B} B$ and $x_i = (S_i)_z$.

### 4.2.2 *Auto Model*

For

$$V_i(x_i) = x_i G_i(x_i) \tag{4.23}$$

$$V_{ij}(x_i, x_j) = -J_{ij} x_i x_j \, , \tag{4.24}$$

i.e.,

$$H(x_S) = \sum_{i=1}^{N} x_i G_i(x_i) - \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} J_{ij} x_i x_j \; , \tag{4.25}$$

one speaks of an *auto model*. $J_{ij}$ are called the interaction coefficients.

If $x_i$ only assumes the values $\{0, 1\}$ and $G_i(x_i) = \alpha_i$, this model is also known as the autologistic model. If the space $\mathcal{L}$ of the possible values of $x_i$ is equal to $\{-1, 1\}$, one obtains an Ising model, which we will study in more detail in Sect. 4.5. The generalization to the case in which $x_i$ assumes the values $\{0, \dots, M-1\}$ is also called the autobinomial model.

One finds for the conditional probability in this model

$$\varrho(x_i|\mathcal{N}_i) = \binom{M-1}{x_i} q^{x_i}(1-q)^{M-1-x_i} \tag{4.26}$$

with

$$q = \frac{\mathrm{e}^{\alpha_i + \sum_{j\in\mathcal{N}_i} J_{ij}x_j}}{1 + \mathrm{e}^{\alpha_i + \sum_{j\in\mathcal{N}_i} J_{ij}x_j}} \,. \tag{4.27}$$

For $M = 2$ we get

$$\varrho(x_i|\mathcal{N}_i) = q^{x_i}(1-q)^{1-x_i} \tag{4.28}$$

and therefore

$$\varrho(0|\mathcal{N}_i) = 1 - q = \frac{1}{1 + \exp\left(\alpha_i + \sum_{j\in\mathcal{N}_i} J_{ij}x_j\right)} \tag{4.29}$$

$$\varrho(1|\mathcal{N}_i) = q. \tag{4.30}$$

Setting $\alpha_i \equiv \alpha$ and, for example, $J_{ij} \equiv J$, these models are homogeneous.

### 4.2.3 *Multilevel Logistic Model*

In the multilevel logistic model, the random variable $X_i$ may assume $M$ possible values, e.g., $\mathcal{L} = \{1, \dots, M\}$. For the potentials one further specifies

$$V_1(x_i) = \begin{cases} \alpha_I & \text{if } x_i \in I \subset (1, \dots, M) \\ -\alpha_I & \text{otherwise} \end{cases} \tag{4.31}$$

and

$$V_2(x_i, x_j) = \begin{cases} J_\mathrm{c} & \text{if } j \in \mathcal{N}_i \text{ and } x_i = x_j \\ -J_\mathrm{c} & \text{otherwise}\,, \end{cases} \tag{4.32}$$

where $J_\mathrm{c}$ may depend on the relative positions of $i$ and $j$. In a two-dimensional lattice where each point has eight nearest neighbors there are four types of pairs for which one can choose different constants $J_\mathrm{c}$. This is illustrated in Fig. 4.3.

For $M = 2$ one again obtains the Ising model, which is anisotropic if one chooses $J_\mathrm{c} = J_1$ for the configuration of pairs (a) in Fig. 4.3 and $J_\mathrm{c} = J_2 \neq J_1$ for configuration (b) in Fig. 4.3. The cliques (c) and (d) are not present in the Ising

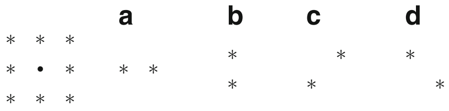


**Fig. 4.3** The eight neighbors of a point and the various orientations of pairs of neighbors

model, as there one usually only considers a system of neighborhoods with four neighbors (Fig. 4.2).

For general $M$ one speaks of a Potts model.

If $J_c < 0$, those patterns where the realizations of neighboring random variables assume predominantly the same value are more probable than other configurations.

### 4.2.4 Gauss Model

Finally, we include the Gauss model mentioned in the previous section. The realizations of $X_i$ are on the real axis and the common density reads

$$\varrho(x_S) = \frac{\sqrt{\det \mathsf{B}}}{\sqrt{(2\sigma^2)^N}} \exp\left( -\frac{1}{2\sigma^2} \sum_{i,j=1}^{N} (x_i - \mu_i)\mathsf{B}_{ij}(x_j - \mu_j) \right) . \tag{4.33}$$

Here, $\{\mu_i\}$ are given numbers and $\mathsf{B}$ should be a nonsingular matrix.

*Remarks.* Markov fields are suitable for the modeling of images, in particular, as we already have mentioned in Sect. 4.1 and shown in Fig. 4.1, those images which are called textures. Methods of generating realizations of Gibbs fields will be discussed in Sect. 5.5 in connection with the generation of realizations of stochastic processes.

On the other hand, the identification of a Gibbs model for a given image interpreting the image as a realization of a certain Gibbs field means the estimation of the parameters of this model. The parameters, whose number should of course be smaller than the number of data, represent a set of quantities characterizing the image. Image characterization is one of the tasks dealt with in image analysis. The problem of parameter estimation on the basis of a given set of data (for admittedly simpler situations) will be dealt with in part II of this book.

Gibbs fields as models for spin systems only cover the classical spin models. In Sect. 6.9 we will see how such classical spin systems appear as special cases of more general quantum spin systems.

## 4.3 Characteristic Quantities of Densities for Random Fields

In this section we will present the notation and equations which will be important for the later studies of explicit models and approximation methods. We start from the density for a random field $\{X_i\}$,

$$\varrho(\boldsymbol{x}) = \frac{1}{Z} \mathrm{e}^{-\beta H(\boldsymbol{x})} \quad \text{with} \quad H(\boldsymbol{x}) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} x_i x_j + \sum_{i=1}^{N} B_i x_i \ , \tag{4.34}$$

where the partition function $Z$ is given by

$$Z = \sum_{\{x_i\}} \mathrm{e}^{-\beta H(\boldsymbol{x})} \quad \text{or} \quad Z = \int \mathrm{d}^N x \, \mathrm{e}^{-\beta H(\boldsymbol{x})} \ , \tag{4.35}$$

depending on whether the $\{x_i\}$ assume discrete or continuous values, respectively. We will write the partition function as $Z(\beta, \{B_i\}, N)$.

In the framework of probability theory, the generating function of the cumulants would be defined as

$$f(\boldsymbol{\Theta}) = \ln \left\langle \mathrm{e}^{\boldsymbol{\Theta} \boldsymbol{X}} \right\rangle \tag{4.36}$$

$$= \ln \left[ \frac{1}{Z} \sum_{\{x_i\}} \mathrm{e}^{-\beta \left( -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} x_i x_j + \sum_{i=1}^{N} (B_i - \Theta_i / \beta) x_i \right)} \right] \tag{4.37}$$

$$= \ln Z(\beta, \{B_i - \Theta_i / \beta)\}, N) - \ln Z(\beta, \{B_i\}, N) \ . \tag{4.38}$$

In the framework of statistical mechanics, however, the generating function of the cumulants is the free energy

$$F(\beta, \{B_i\}, N) = -k_{\mathrm{B}} T \ln Z(\beta, \{B_i\}, N) \tag{4.39}$$

with $k_{\mathrm{B}} T = 1/\beta$. Since

$$\left. \frac{\partial}{\partial \Theta_i} f(\boldsymbol{\Theta}) \right|_{\boldsymbol{\Theta}=0} = \frac{\partial}{\partial B_i} F(\beta, \{B_k\}, N) \ , \tag{4.40}$$

the functions $F(\beta, \{B_i\}, N)$ and $f(\boldsymbol{\Theta})$ will, however, do the same job.

The most important quantities that we can derive from $F(\beta, \{B_i\}, N)$ or $f(\boldsymbol{\Theta})$ are:

- The first moment

$$M_i = \langle X_i \rangle = \int \mathrm{d}^N x \, x_i \, \frac{1}{Z} \mathrm{e}^{-\beta H(\boldsymbol{x})} = \frac{\partial}{\partial B_i} F(\beta, \{B_k\}, N) \tag{4.41}$$

or

$$m = \frac{1}{N} \sum_{i=1}^{N} M_i = \frac{1}{N} \sum_{i=1}^{N} \langle X_i \rangle = \frac{1}{N} \frac{\partial}{\partial B} F(\beta, B, N) \ , \tag{4.42}$$

where we have set all $B_i$ in $F(\beta, B, N)$ equal to $B$. In anticipation of later applications we will call $M_i$ the magnetic moment at site $i$ and $m$ the magnetization.

The magnetization $m$ is the expectation value of the random variable

$$Y_N = \frac{1}{N}\sum_{i=1}^{N} X_i , \tag{4.43}$$

i.e., $Y_N$ is a sum of random variables whose densities and mutual dependencies are determined by the common density $\varrho(\boldsymbol{x})$.

Statements about the properties of the random variable $Y_N$ may thus be considered as generalizations of the statements made in Sect. 2.5 concerning the central limit theorem. For the simplest case we required there that the random variables $X_i$ be independent and have identical distributions. In general, the properties of $Y_N$ are determined by the density $\varrho(\boldsymbol{x})$.

- The covariance matrix with elements

$$\mathsf{C}_{ij} = \mathrm{Cov}(X_i, X_j) = \int \mathrm{d}^N x \, (x_i x_j - M_i M_j)\frac{1}{Z}\mathrm{e}^{-\beta H(\boldsymbol{x})} \tag{4.44}$$

$$= \frac{1}{\beta}\frac{\partial^2}{\partial B_i \partial B_j} F(\beta, \{B_k\}, N). \tag{4.45}$$

- The susceptibility, defined by

$$\chi = \frac{\partial m}{\partial B} \; . \tag{4.46}$$

This quantity measures the sensitivity of the magnetization to a change of $B$. Obviously, we obtain from (4.42)

$$\chi = \frac{1}{N}\frac{\partial^2}{\partial B^2}F(\beta, B, N) = \frac{\beta}{N}\sum_{i,j=1}^{N}\mathrm{Cov}(X_i, X_j), \tag{4.47}$$

i.e., the susceptibility is again related to the covariance. It diverges for $N \to \infty$ if, for instance, the correlation between $X_i$ and $X_{i+\tau}$ does not decrease fast enough for large values of $r = |\tau|$, so that the sum $\sum_\tau \mathrm{Cov}(X_i, X_{i+\tau})$ does not exist.

- The Legendre transform of the free energy function $f(\boldsymbol{\Theta})$,

$$g(\boldsymbol{M}) = \sup_{\boldsymbol{\Theta}}(\boldsymbol{M}^T\boldsymbol{\Theta} - f(\boldsymbol{\Theta})) \; , \tag{4.48}$$

is identical to the Legendre transform of $F(\beta, B, N)$, given by

$$G(\beta, M, N) = MB(M) - F(\beta, B(M), N), \tag{4.49}$$

where $B(M)$ follows from solving the equation $M = \partial F/\partial B$ for $B$ ($M = Nm$, comp. (4.42) ).

- For the entropy we obtain

$$S(\beta, B, N) = k_{\mathrm{B}}(\beta\langle H(\boldsymbol{X})\rangle + \ln Z) = -\frac{\partial F(\beta, B, N)}{\partial T} \,. \tag{4.50}$$

## 4.4 Simple Random Fields

### *4.4.1 The White Random Field or the Ideal Paramagnetic System*

We now will study the model of a white random field, i.e., of an ideal paramagnetic crystal whose atoms have a spin $s = 1/2$. The common density of the random variables or spins reads

$$\varrho(\boldsymbol{x}) = \frac{1}{2^N Z} \exp\left(\beta\Theta \sum_{i=1}^{N} x_i\right), \tag{4.51}$$

where, for the paramagnetic crystal, $\Theta = g\mu_{\mathrm{B}} B/2$. For the partition function one obtains

$$Z(T, B, N) = 2^{-N} \sum_{x_1=\pm 1} \ldots \sum_{x_N=\pm 1} \exp\left(\beta\Theta \sum_{i=1}^{N} x_i\right) \tag{4.52}$$

$$= \left(\sum_{x=\pm 1} \frac{1}{2} \mathrm{e}^{\beta\Theta x}\right)^N \tag{4.53}$$

$$= (\cosh \beta\Theta)^N \,. \tag{4.54}$$

Following the methods of statistical mechanics one can derive from this partition function the free energy, the entropy, the magnetization, and the susceptibility.

For our present purposes, however, we will use the methods of Sect. 2.7. Taking into account also the prefactors, which appear for physical systems, the magnetization $m$ is now given by

$$m = \frac{1}{2} g\mu_B \langle Y_N \rangle, \tag{4.55}$$

where $Y_N$ is again the mean value of the random variables $X_i$ (cp. 4.42):

$$Y_N = \frac{1}{N} \sum_{i=1}^{N} X_i \,. \tag{4.56}$$

We will compute the density of this quantity for large $N$. The random variables $\{X_i\}$ of the field are independent and identically distributed with the density $\varrho(x) = \mathrm{e}^{\beta\Theta x}/(2\cosh\beta\Theta), x = \pm 1$. Hence, the free energy function of each of this random variables is

$$f(t) = \ln\langle \mathrm{e}^{tX}\rangle = \ln\cosh(t + \beta\Theta) - \ln\cosh(\beta\Theta). \tag{4.57}$$

Following Sect. 2.7 we have to calculate the Legendre transform $g(y)$ of $f(t)$. The density of $Y_N$ then assumes the form

$$\varrho_{Y_N}(y) \propto \mathrm{e}^{-Ng(y)}, \tag{4.58}$$

and it has a maximum at the value $y = y_0$ for which $g(y)$ is minimal. We do not have to determine $g(y)$ explicitly, since we only need this value $y_0$ for the calculation of the magnetization from (4.55). Now, according to (2.214), we have $g'(y) = t(y)$, where $t(y)$ follows from solving

$$y = f'(t) = \tanh(t + \beta\Theta) \tag{4.59}$$

for $t$. Therefore, $y_0$, being a zero of $g'(y)$, is also given by $y_0 = f'(0)$, and in the limit $N \to \infty$, where the minimum of $g(y)$ and $\langle Y_N\rangle$ become identical, we get

$$m = \frac{1}{2} g\mu_B \tanh\beta\Theta \equiv \frac{1}{2} g\mu_B \tanh(\beta g\mu_B B/2)\ . \tag{4.60}$$

*Remarks.*

- For small $\beta$, i.e., for high temperatures, we find $\beta\Theta \equiv \beta g\mu_{\mathrm{B}} B/2 \ll 1$. As $\tanh\beta\Theta = \beta\Theta + \ldots$ we obtain

$$m = \frac{\mu_B^2 g^2}{4kT} B + \mathcal{O}(B^2). \tag{4.61}$$

  For the magnetic susceptibility $\chi = \dfrac{\partial m}{\partial B}$ we therefore get, to a good approximation,

$$\chi = \frac{C}{T} \qquad \text{where} \qquad C = \frac{\mu_B^2 g^2}{4k}. \tag{4.62}$$

  $C$ is called the Curie constant and the behavior of $\chi$ as a function of $T$ is known as Curie's law for paramagnetic substances.
- For $B \to 0$ we also have $\Theta \to 0$ and thus $m \to 0$. There is no spontaneous magnetization, i.e., no magnetization for $B = 0$, in contrast to the well known case of a ferromagnet.

### 4.4.2 The One-Dimensional Ising Model

The Ising model in one dimension describes a Markov random field where the individual random variables are no longer independent. But the free energy and all other system variables can still be calculated analytically. In order to demonstrate this we will introduce the transfer matrix method, which has also been applied successfully to more general spin models.

The Hamiltonian function for the one-dimensional Ising model in an external field $B$ reads (cf. (4.25))

$$H = -J \sum_{i=1}^{N} x_i x_{i+1} - \mu B \sum_{i=1}^{N} x_i , \qquad x_i = \pm 1. \tag{4.63}$$

In general, one uses periodic boundary conditions, i.e., one defines $x_{N+1} = x_1$. The last term in the sum of interactions therefore reads $x_N x_1$. Thus we think of the lattice as being bent into a large circle such that $x_N$ and $x_1$ are also nearest neighbors.

The partition function can be written in the form

$$\begin{aligned} Z(T, B, N) &= \sum_{x_1=\pm 1} \cdots \sum_{x_N=\pm 1} \exp\left(\beta \sum_{i=1}^{N} (J x_i x_{i+1} + \mu B x_i)\right) && (4.64)\\ &= \sum_{x_1=\pm 1} \cdots \sum_{x_N=\pm 1} \langle x_1 | T | x_2 \rangle \ldots \langle x_N | T | x_1 \rangle , && (4.65) \end{aligned}$$

with

$$\langle x_i | T | x_{i+1} \rangle = \exp\left(\beta \left(J x_i x_{i+1} + \mu B (x_i + x_{i+1})/2\right)\right), \tag{4.66}$$

i.e.,

$$\begin{aligned} \langle + | T | + \rangle &= \mathrm{e}^{\beta (J + \mu B)}, \quad & \langle + | T | - \rangle &= \mathrm{e}^{-\beta J}, \\ \langle - | T | + \rangle &= \mathrm{e}^{-\beta J}, & \langle - | T | - \rangle &= \mathrm{e}^{\beta (J - \mu B)}. \end{aligned} \tag{4.67}$$

So we may write

$$Z(T, B, N) = \operatorname{tr}\left(T^N\right), \tag{4.68}$$

and $T$ is a 2×2 matrix with the elements given in (4.67). $T$ is also called the transfer matrix.

All we have to know are the eigenvalues of $T$, because if $T$ is known in the representation

$$T = U \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} U^{-1} \tag{4.69}$$

then

$$T^N = U \begin{pmatrix} \lambda_1^N & 0 \\ 0 & \lambda_2^N \end{pmatrix} U^{-1} \tag{4.70}$$

and therefore

$$Z(T,B,N) = \operatorname{tr}\left(T^N\right) = \lambda_1^N + \lambda_2^N = \lambda_1^N \left(1 + \left(\frac{\lambda_2}{\lambda_1}\right)^N\right). \tag{4.71}$$

The eigenvalues of a 2 × 2 matrix are easily determined as the roots of the quadratic equation

$$\begin{vmatrix} T_{11} - \lambda & T_{12} \\ T_{21} & T_{22} - \lambda \end{vmatrix} = 0, \tag{4.72}$$

i.e.,

$$T_{11}T_{22} - \lambda(T_{11} + T_{22}) + \lambda^2 - T_{12}T_{21} = 0. \tag{4.73}$$

The solutions are

$$\lambda_{1,2} = \frac{T_{11} + T_{22}}{2} \pm \frac{1}{2}\sqrt{(T_{11} - T_{22})^2 + 4T_{12}T_{21}}, \tag{4.74}$$

and

$$\lambda_1 = \mathrm{e}^{\beta J} \cosh(\beta\mu B) + \sqrt{\mathrm{e}^{2\beta J} \sinh^2(\beta\mu B) + \mathrm{e}^{-2\beta J}} \tag{4.75}$$

is obviously the larger eigenvalue, at least for finite values of $\beta$. However, for $B \equiv 0$ and $T \to 0$, i.e., $\beta \to \infty$, the contribution from the square root vanishes and therefore

$$\lambda_{1,2} = \mathrm{e}^{\beta J}. \tag{4.76}$$

We find

$$F_m = -k_\mathrm{B}T \ln Z(T,B,N) \tag{4.77}$$

$$= -Nk_\mathrm{B}T \ln\lambda_1 - k_\mathrm{B}T \ln\left(1 + \left(\frac{\lambda_2}{\lambda_1}\right)^N\right). \tag{4.78}$$

For large $N$ and for $\lambda_2 < \lambda_1$ the second term may be neglected and we obtain

$$\begin{aligned} F_m &= -Nk_\mathrm{B}T \ln\lambda_1 \qquad (4.79) \\ &= -Nk_\mathrm{B}T \ln\left(\mathrm{e}^{\beta J}\cosh(\beta\mu B) + \sqrt{\mathrm{e}^{2\beta J}\sinh^2(\beta\mu B) + \mathrm{e}^{-2\beta J}}\right) \\ &= -NJ - Nk_\mathrm{B}T \ln\left(\cosh(\beta\mu B) + \sqrt{\sinh^2(\beta\mu B) + \mathrm{e}^{-4\beta J}}\right). \end{aligned}$$

Thus we have determined the free energy, and everything else is easy to calculate. The above method, in which the partition function is represented as the trace of a matrix and the problem thereby reduced to the determination of the largest eigenvalue of this matrix, is called the transfer-matrix method and has also been used successfully for more general spin models on the lattice. Here, for the one-dimensional Ising model, this method is especially simple.

We obtain:

- For the magnetization:

$$\begin{aligned} m(T,B) &= -\frac{1}{V}\,\frac{\partial F_m(T,B)}{\partial B} \\ &= nk_{\mathrm{B}}T\,\beta\mu\,\frac{\sinh(\beta\mu B)+\frac{1}{\sqrt{\cdots}}\cosh(\beta\mu B)\sinh(\beta\mu B)}{\cosh(\beta\mu B)+\sqrt{\ldots}} \\ &= n\mu\,\sinh(\beta\mu B)\,\left(\sinh^2(\beta\mu B)+\mathrm{e}^{-4\beta J}\right)^{-1/2}, \end{aligned} \tag{4.80}$$

  i.e., for $B=0$ we also have $m=0$, independent of $\beta$. There is no spontaneous magnetization.

- For the susceptibility:

$$\begin{aligned} \chi &= \frac{\partial m(T,B)}{\partial B} \\ &= n\mu\,\beta\mu\,\cosh(\beta\mu B)\Big/\sqrt{\sinh^2(\beta\mu B)+\mathrm{e}^{-4\beta J}} \\ &\quad -\frac{1}{2}n\mu\,\sinh(\beta\mu B)\,\left(\sinh^2(\beta\mu B)+\mathrm{e}^{-4\beta J}\right)^{-3/2} \\ &\quad \cdot\,2\sinh(\beta\mu B)\,\cosh(\beta\mu B). \end{aligned} \tag{4.81}$$

  For $B=0$ one finds

$$\chi = n\mu^2\beta\,\mathrm{e}^{2\beta J}, \tag{4.82}$$

  i.e., $\chi$ is divergent for $T\to 0$, but exponentially. This does not correspond to a critical exponent, because the definition of such exponents assumed divergence according to a power law.

- For the covariance function (see also McCoy and Wu 1973):

$$\langle x_n x_{n'}\rangle = \frac{\sinh^2(\beta\mu B)+(\lambda_2/\lambda_1)^{|n-n'|}\mathrm{e}^{-4\beta J}}{\sinh^2(\beta\mu B)+\mathrm{e}^{-4\beta J}}, \tag{4.83}$$

  i.e., in particular,

$$\langle x_n x_n\rangle = \langle x_n^2\rangle = 1, \tag{4.84}$$

  and for $B\equiv 0$

$$\langle x_n x_{n'}\rangle = \left(\frac{\lambda_2}{\lambda_1}\right)^{|n-n'|} = \left(\frac{\sinh(\beta J)}{\cosh(\beta J)}\right)^{|n-n'|} \tag{4.85}$$

$$= \left(\tanh(\beta J)\right)^{|n-n'|}. \tag{4.86}$$

- For the energy $E$:

$$E = \frac{1}{Z}\,\mathrm{tr}\left(H\,\mathrm{e}^{-\beta H}\right) = -\frac{\partial}{\partial\beta}\,\ln\,Z(T,B,N), \tag{4.87}$$

i.e., for $B = 0$

$$E = -\frac{\partial}{\partial\beta}\,\ln\,Z\Big|_{B=0} = -\frac{\partial}{\partial\beta}\left(N\,\ln\,2 + N\,\ln\,\cosh(\beta J)\right) \tag{4.88}$$

$$= -J\,N\,\tanh(\beta J), \tag{4.89}$$

and therefore

$$C_B\Big|_{B=0} = \frac{\partial E(T, B=0)}{\partial T} = -JN\,\frac{1}{\cosh^2(\beta J)}\left(-\frac{1}{k_\mathrm{B}T^2}\,J\right) \tag{4.90}$$

$$= k_\mathrm{B}N\,\frac{J^2}{(kT)^2}\,\frac{1}{\cosh^2(J/kT)}. \tag{4.91}$$

- And, finally, for the entropy at $B = 0$:

$$\begin{aligned} S(T,B=0) &= -\frac{\partial}{\partial T}\left(-k_\mathrm{B}T\,\ln\,Z(T,B=0,N)\right) \\ &= k_\mathrm{B}\left[N\,\ln 2 + N\,\ln\,\cosh\left(\frac{J}{k_\mathrm{B}T}\right)\right. \\ &\qquad \left. -N\,\frac{J}{k_\mathrm{B}T}\,\tanh\left(\frac{J}{k_\mathrm{B}T}\right)\right]. \end{aligned} \tag{4.92}$$

For $T \to \infty$ one finds

$$S \to Nk_\mathrm{B}\,\ln 2\ . \tag{4.93}$$

For the limit $T \to 0$ we have to take into account (4.76), i.e., we get $F_m = -NJ - k_\mathrm{B}\,T\ln 2 + O(T^2)$, and thus for $T \to 0$

$$S \to k\,\ln 2\ . \tag{4.94}$$

This implies that the number of states of the system at $T = 0$ is just 2, which is plausible, since the system is in its ground state at $T = 0$, i.e.,

$$E = -N\,J, \tag{4.95}$$

and all spins are parallel. There are two possibilities for this configuration: either all spins have the magnetic quantum number $m_s = +1/2$, or all spins have $m_s = -1/2$.

## 4.5 Random Fields with Phase Transitions

### *4.5.1 The Curie–Weiss Model*

We consider a random field $\boldsymbol{X} = (X_1, \dots, X_N)$ with realizations $x_i \in \{-1, +1\}$ and energy function

$$H(\boldsymbol{x}) = -\frac{1}{2}\frac{J_0}{N}\sum_{i,j=1}^{N} x_i x_j - \Theta \sum_{i=1}^{N} x_i \ . \tag{4.96}$$

This model has the advantage that the energy function is only a function of

$$y_N = \frac{1}{N}\sum_{i=1}^{N} x_i . \tag{4.97}$$

To within a numerical factor, $y_N$ corresponds to the magnetization, i.e., we can also write

$$H(\boldsymbol{x}) \equiv H'(y_N) = -N\left(\frac{J_0}{2}y_N^2 + \Theta y_N\right) \tag{4.98}$$

and

$$\rho(\boldsymbol{x}) = \frac{1}{Z}\frac{1}{2^N}\mathrm{e}^{-\beta H(\boldsymbol{x})} \ . \tag{4.99}$$

Thus we can calculate the density of the random variable $Y_N$ and we find

$$\varrho_{Y_N}(y) = \sum_{\{x_i=\pm 1\}} \delta\left(y - \frac{1}{N}\sum_{i=1}^{N} x_i\right)\frac{1}{Z}\frac{1}{2^N}\ \exp N\left(\frac{\beta J_0}{2}y^2 + \beta\Theta y\right) \ .$$

In Sect. 2.7 we already determined the density for $Y_N$ under the condition that the random variables $\{X_i\}$ are independent (cf. (2.228)), and we obtained for large values of $N$

$$\frac{1}{2^N}\sum_{\{x_i=\pm 1\}} \delta\left(y - \frac{1}{N}\sum_{i=1}^{N} x_i\right) \propto \mathrm{e}^{-NS(y)} \tag{4.100}$$

with

$$S(y) = \frac{1+y}{2}\ln(1+y) + \frac{1-y}{2}\ln(1-y) \ . \tag{4.101}$$

Therefore we can write the density of the random variable $Y_N$ as

$$\varrho_{Y_N}(y) = \frac{1}{Z}\mathrm{e}^{-N\lambda(y)} \tag{4.102}$$

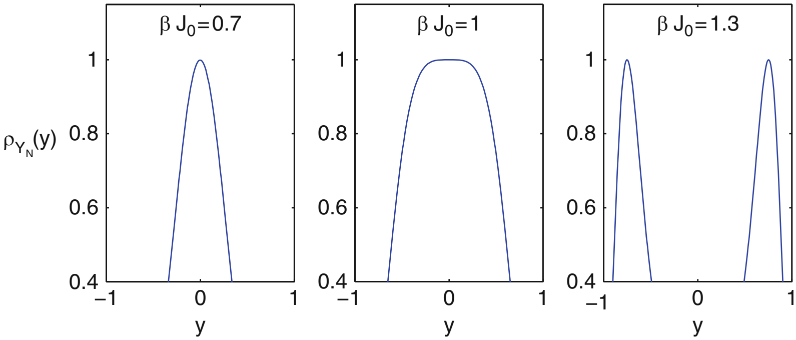


**Fig. 4.4** The density function $\varrho_{Y_N}(y)$ as a function of $y$ for $\Theta = 0$ and for $\beta J_0 < 1$ (*left*), $\beta J_0 = 1$ (*middle*), and $\beta J_0 > 1$ (*right*). For sufficiently large $\beta$ (low temperature) the density function exhibits two maxima, i.e., for $N \to \infty$ there are two equilibrium states

with

$$Z = \int \mathrm{d}y\, \mathrm{e}^{-N\lambda(y)} \tag{4.103}$$

and

$$\lambda(y) = S(y) - \left(\frac{\beta J_0}{2} y^2 + \beta\Theta y\right) . \tag{4.104}$$

For future use we note the expansion of $S(y)$ about $y = 0$:

$$S(y) = \frac{1}{2}y^2 + \frac{1}{12}y^4 + \dots , \tag{4.105}$$

so that

$$\lambda(y) = -\beta\Theta y + \frac{1}{2}(1 - \beta J_0)y^2 + \frac{1}{12}y^4 + \dots . \tag{4.106}$$

Figure 4.4 shows $\varrho_{Y_N}(y)$ for $\Theta = 0$ as a function of $y$ for the cases $\beta J_0 < 1$, $\beta J_0 = 1$, and $\beta J_0 > 1$. The general behavior of this density function obviously changes at $\beta J_0 = 1$. For $\beta J_0 < 1$ there is only one maximum $y_m$ and thus there is only one state in the limit $N \to \infty$, for which $y_m = 0$. On the other hand, for $\beta J_0 > 1$ there are two maxima at, say, $y = \pm y_m$. From the symmetry $H(-\boldsymbol{x}) = H(\boldsymbol{x})$ one can immediately deduce that there has to be a second maximum of the density if there is one at $y \neq 0$.

For $\beta J_0 > 1$ the system can therefore be in one of the two states, or even in a mixture of both. This will depend on the initial conditions. Hence, we find a magnetization $m \propto \pm y_m \neq 0$ even for $\Theta = 0$. This is the ferromagnetic phase.

At $\beta J_0 = 1$ there is a critical point, and exactly the phenomenon already discussed in Sect. 2.7 occurs. It is easy to show that close to $y = 0$ the density is of the form $p(y) \propto \exp(-\mathrm{const} \cdot y^4/12)$, i.e., there are large non-Gaussian fluctuations around $y = 0$. Such fluctuations can indeed be observed at critical points of real systems (cf. (3.8)).

The minimum (or minima) of $\lambda(y)$ are obtained from the equation

$$\lambda'(y_m) = 0, \quad \text{i.e.,} \quad \frac{1}{2}\ln\left(\frac{1+y_m}{1-y_m}\right) - \beta J_0 y_m - \beta\Theta = 0 \tag{4.107}$$

or

$$y_m = \tanh\left(\beta J_0 y_m + \beta\Theta\right) . \tag{4.108}$$

We thus find an implicit equation for $y_m$. In the next subsection we will obtain exactly the same equation for the magnetization in the framework of the mean field approximation, and there we will study the solutions of this equation.

The behavior of the covariance function $C_{ij}$ for large values of $r = |i - j|$ is also an indicator for critical points. In the ordered phase, one expects $C_{ij}$ to decrease comparatively slowly for large values of $r$, i.e., there is a long-range dependence between the random variables $\{X_i\}$. A measure for the long-range dependence is the quantity

$$\frac{1}{N}\sum_{i,j}\mathrm{Cov}(X_i X_j) = \sum_{\tau}\frac{1}{N}\sum_{i}\mathrm{Cov}(X_i X_{i+\tau}) . \tag{4.109}$$

The summation over $\tau$ will not converge for a long-range dependence and hence this quantity will be divergent in the limit $N \to \infty$. For the Curie–Weiss model one obtains for this quantity, excluding additive constants

$$\frac{1}{N}\sum_{i,j}\langle X_i X_j\rangle = N\frac{\int \mathrm{d}y\; y^2\, \exp -N\lambda(y)}{\int \mathrm{d}y\, \exp -N\lambda(y)} . \tag{4.110}$$

If $\lambda(y) \propto y^2$, the right-hand side is of the order $N^0 \equiv 1$, i.e., in the limit $N \to \infty$ the sum on the left-hand side converges. For $\lambda(y) \propto y^4$, at the critical point, the right-hand side is of the order $N^{1/2}$ and therefore the sum on the left-hand side diverges in the limit $N \to \infty$.

### 4.5.2 The Mean Field Approximation

We will now consider a model with the energy function

$$H = -\sum_{i,j=1}^{N} J(i-j)x_i\, x_j - \Theta\sum_{i=1}^{N} x_i , \tag{4.111}$$

where the parameters $\{J(k)\}$ are assumed to be summable, i.e., even for an infinite lattice

$$J_0 = \sum_{k} J(k) < \infty . \tag{4.112}$$

For $J(i-j) \equiv 0$ and $\Theta = g\mu_{\mathrm{B}} B/2$ we recover the model for the ideal paramagnetic crystal discussed in the previous section. In the mean field approximation the energy function $H$ is replaced by

$$H' = -J_0 m' \sum_{i=1}^{N} x_i - \Theta \sum_{i=1}^{N} x_i = -(J_0 m' + \Theta) \sum_{i=1}^{N} x_i . \tag{4.113}$$

The parameter $m'$ represents a mean field produced by all spins. To justify this approximation we consider the effective field felt by a spin at site $i$,

$$\Theta_i^{\mathrm{eff}} = \Theta + \sum_j J(i-j) x_j \ . \tag{4.114}$$

The energy function $H$ may then be written as

$$H = -\sum_{i=1}^{N} \Theta_i^{\mathrm{eff}} x_i . \tag{4.115}$$

The approximation now consists in the replacement of $\Theta_i^{\mathrm{eff}}$ by its expectation value and the requirement of homogeneity, i.e., $\Theta_i^{\mathrm{eff}}$ is replaced by $\langle \Theta_i^{\mathrm{eff}} \rangle = \theta^{\mathrm{eff}}$. Then

$$\theta^{\mathrm{eff}} = \left\langle \Theta_i^{\mathrm{eff}} \right\rangle = \Theta + \sum_j J(i-j) \langle x_j \rangle = \Theta + J_0 m' \tag{4.116}$$

so that (4.113) follows from (4.115). In particular, we obtain

$$m' = \langle x_j \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} \langle x_i \rangle \ . \tag{4.117}$$

According to (4.55), $m = g\mu_{\mathrm{B}} m'/2$ is now the magnetization of the system.

The model has therefore been reduced to an interaction-free model with independent random variables. In such a model one obtains from (4.60)

$$m' = \tanh\left(\beta \theta^{\mathrm{eff}}\right) \quad \text{or} \quad m' = \tanh\big(\beta(\Theta + J_0 m')\big), \tag{4.118}$$

and finally for the magnetization

$$m = \frac{1}{2} g\mu_B \tanh\Big(\frac{1}{2}\beta g\mu_{\mathrm{B}}(B + J_0' m)\Big), \tag{4.119}$$

with $J_0' = J_0/(\frac{1}{2} g\mu_{\mathrm{B}})^2$. Hence, the magnetization for $B = 0$ has to be determined self-consistently from this implicit equation. This is most easily done graphically

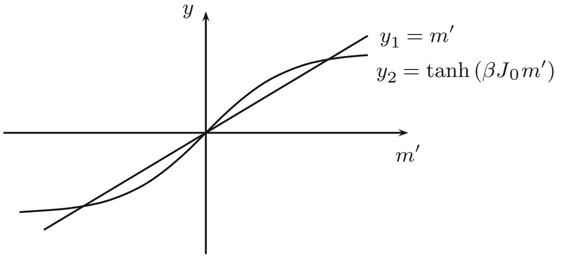


**Fig. 4.5** The intersection points of the curves for $y_1(m')$ and $y_2(m')$ correspond to the solutions of (4.118)

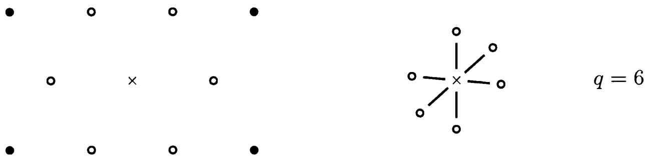


**Fig. 4.6** A triangular lattice (*left*) in two space dimensions and a cubic lattice (*right*) in three dimensions have the same number, $q = 6$, of nearest neighbors

by drawing the curve for $y_1 = m'$ and for $y_2 = \tanh(\beta J_0 m')$ and looking for the intersection point of the two curves (see Fig. 4.5).

This equation has a solution, other than $m' = 0$, if the slope of $y_2(m') = \tanh(\beta J_0 m')$ at $m' = 0$ is larger than 1. This is the case for

$$\beta J_0 > 1. \tag{4.120}$$

Thus in this approximation we obtain a critical temperature $T_c = J_0/k_B$ which only depends on $J_0$. If

$$J(i-j) \equiv J \neq 0 \qquad \text{for } i, j \text{ nearest neighbors,} \tag{4.121}$$

one therefore finds

$$J_0 = \sum_k J(k) = qJ, \tag{4.122}$$

where $q$ denotes the number of the nearest neighbors. Hence

$$T_c \propto J_0 \propto q, \tag{4.123}$$

and the critical temperature depends only on the number of nearest neighbors, but not on the dimension of the model under consideration. For instance, for a triangular lattice in two dimensions one obtains for $q$ the same value 6 as for a cubic lattice in three dimensions (cf. Fig. 4.6).

In this approximation, the critical temperature $T_c$ therefore has the same value for these two cases. This, however, contradicts experience, which tells us that the dimensionality of the lattice plays a significant role. For instance, the Ising model exhibits a phase transition in two dimensions but not in one.

Within the framework of the mean field approximation we may now determine explicitly many quantities of interest. We give some examples:

**The magnetic susceptibility.** This is given by

$$\chi = \frac{\partial m}{\partial B} = \frac{\partial m}{\partial \theta^{\text{eff}}} \frac{\partial \theta^{\text{eff}}}{\partial B}. \tag{4.124}$$

Since now $\theta^{\text{eff}} = \frac{1}{2} g\mu_{\text{B}}(B + J_0' m)$, we have

$$\frac{\partial \theta^{\text{eff}}}{\partial B} = \frac{1}{2} g\mu_{\text{B}}(1 + J_0' \chi) \tag{4.125}$$

and

$$\frac{\partial m}{\partial \theta^{\text{eff}}} = \chi_0 \Big/ \frac{1}{2} g\mu_B \ , \tag{4.126}$$

where $\chi_0$ is the susceptibility of an ideal paramagnetic crystal. We thus obtain

$$\chi = \chi_0(1 + J_0' \chi), \tag{4.127}$$

which finally leads to

$$\chi = \frac{\chi_0}{1 - J_0' \chi_0}. \tag{4.128}$$

In the regime $T > T_c$ and in the limit $B \to 0$ we find for the magnetization $m \to 0$ and thus also $\theta^{\text{eff}} \to 0$, and $\chi_0$ is equal to the susceptibility according to Curie's law: $\chi_0 = C/T$ with $C = (\frac{1}{2} g\mu_{\text{B}})^2 / k_{\text{B}}$. From $J_0' C = J_0/k_{\text{B}} = T_c$ one obtains in this case

$$\chi = \frac{\chi_0}{1 - J_0' \chi_0} = \frac{C/T}{1 - J_0' C/T} = \frac{C}{T - T_c} \ . \tag{4.129}$$

This is the Curie–Weiss law. It is of the same form as Curie's law with $T$ replaced by $T - T_c$.

For $T < T_c$ some calculation (see Diu et al. 1994) leads to

$$\chi = \frac{1}{2} \frac{C}{T_c - T}. \tag{4.130}$$

**Critical exponents.** The implicit equation to determine the magnetization now reads (cf. (4.119))

$$m = m_0 \tanh\left(\frac{1}{2} \beta g\mu_{\text{B}}(B + J_0' m)\right) \tag{4.131}$$

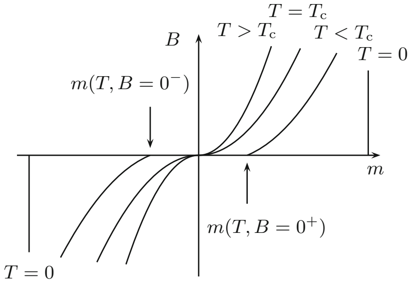


**Fig. 4.7** Isotherms for a spin system (which should be compared with the isotherms of a fluid in a $p$–$V$ diagram). In the $B$–$m$ diagram the magnetization $m$ is represented as a function of the external field $B$

with $m_0 = \frac{1}{2} g\mu_B$. For $T < T_c = J_0' C$ and $C = (g\mu_B)^2/4k_B$ we find a magnetization $m \neq 0$ even if $B = 0$. which corresponds to the $p$–$V$ diagram in the case of fluids. Like $p$, the magnetic field $B$ is an intensive variable, and the magnetization $m$ like $V$ is an extensive variable (Fig. 4.7).

For $T = T_c$ we obtain a curve of third order in the vicinity of $B = 0$ and $m = 0$, as we did for the isotherms in the van der Waals gas. For $T < T_c$ one again finds a horizontal branch on the trajectory for which, however, only the two end points $m(T, B = 0^+)$, $m(T, B = 0^-)$ are realized, depending on whether $B$ tends to zero from positive or negative values, respectively. For $T = 0$ this curve degenerates into two vertical lines at $M = \pm m_0$.

Let us consider the behavior of $B$ as a function of $m$ for $T = T_c$. In analogy to the discussion of the critical exponents for fluids, we parametrize this behavior as

$$\frac{\mu_B g B}{k_B T_c} = \mathcal{D} \left( \frac{m}{m_0} \right)^{\delta}, \qquad T = T_c. \tag{4.132}$$

We obtain $\delta = 3$, $\mathcal{D} = \frac{2}{3}$.

We now investigate how $m(T, B = 0)$ tends to zero for $T \to T_c$ in the regime $T < T_c$. This exactly corresponds to the problem of how $n_{\text{liq}} - n_{\text{gas}}$ or $n_{\text{liq}} - n_c$ tends to zero as $T \to T_c$ for $T < T_c$. Setting $\varepsilon = (T - T_c)/T_c$, we therefore parametrize this behavior as

$$\frac{m(T, B = 0)}{m_0} = \mathcal{B} \, (-\varepsilon)^{\beta}, \tag{4.133}$$

and after some calculation we obtain $\mathcal{B} = \sqrt{3}$ and $\beta = 1/2$.

The analogon to the isothermal compressibility $\kappa_T \propto \frac{\partial V}{\partial p}$ is the susceptibility

$$\chi_T = \left( \frac{\partial m}{\partial B} \right)_T. \tag{4.134}$$

For $T > T_c$ one can easily read off the behavior for $T \to T_c$ from (4.129):

$$\frac{\chi_T}{\chi_T^0} = \mathcal{C} \left( \frac{T - T_c}{T_c} \right)^{-\gamma}, \tag{4.135}$$

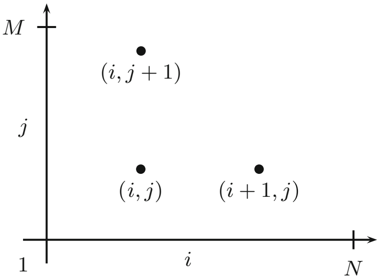


**Fig. 4.8** The two-dimensional square lattice on which we consider the Ising model

with $\mathcal{C} = 1$ and $\gamma = 1$. In the parametrization of the behavior as $T \to T_\text{c}$ we have introduced here the susceptibility $\chi_T^0 = C/T_\text{c}$ of an interaction-free spin system at $T = T_\text{c}$.

For $T < T_\text{c}$ one obtains

$$\frac{\chi_T}{\chi_T^0} = \mathcal{C}'(-\varepsilon)^{-\gamma'} \tag{4.136}$$

with $\gamma' = 1$ and $\mathcal{C}' = 1/2$.

### 4.5.3 The Two-Dimensional Ising Model

We now consider a two-dimensional square lattice. The spin variables will be denoted by $x_{i,j}$, where $i$ refers to the number of the column and $j$ to the number of the row. The energy function then reads

$$H = -E_1 \sum_{i,j} x_{i,j} x_{i,j+1} - E_2 \sum_{i,j} x_{i,j} x_{i+1,j} - g\mu_\text{B} B \sum_{i,j} x_{i,j} , \tag{4.137}$$

where $i$ runs from 1 to $N$ and $j$ form 1 to $M$ (see Fig. 4.8).

If we combine the spin configurations $\{x_{1,\alpha}, \ldots, x_{N,\alpha}\}$ of a row $\alpha$ to $\mu_\alpha$, the configuration of spins on the lattice may also be represented as $\{\mu_1, \ldots, \mu_M\}$.

We choose periodic boundary conditions, i.e., we set

$$\mu_{M+1} = \mu_1; \qquad x_{N+1,\alpha} = x_{1,\alpha} . \tag{4.138}$$

The energy function can now be written

$$H = \sum_{j=1}^{M} \Big( E(\mu_j, \mu_{j+1}) + E(\mu_j) \Big), \tag{4.139}$$

where

$$E(\mu_j) = -E_2 \sum_{i}^{N} x_{i,j} x_{i+1,j} - g\mu_\text{B} B \sum_{i}^{N} x_{i,j} , \tag{4.140a}$$

$$E(\mu_j, \mu_{j+1}) = -E_1 \sum_{i}^{N} x_{i,j} x_{i,j+1}, \tag{4.140b}$$

i.e., $E(\mu_j)$ describes the interaction within each row, and $E(\mu_j, \mu_{j+1})$ the interaction between the rows.

For the partition function one obtains

$$\begin{aligned} Z_{M,N} &= \sum_{\mu_1} \ldots \sum_{\mu_M} \exp -\beta \sum_{j=1}^{M} \Big( E(\mu_j, \mu_{j+1}) + \big( E(\mu_j) + E(\mu_{j+1}) \big)/2 \Big) \\ &= \sum_{\mu_1} \ldots \sum_{\mu_M} \langle \mu_1 | P | \mu_2 \rangle \langle \mu_2 | P | \mu_3 \rangle \ldots \langle \mu_M | P | \mu_1 \rangle \\ &= \operatorname{tr} \big( P^M \big), \end{aligned}$$

with

$$\langle \mu | P | \mu' \rangle = \exp -\beta \left( E(\mu, \mu') + \frac{E(\mu) + E(\mu')}{2} \right). \tag{4.141}$$

Here $P$ is again the transfer matrix. In contrast to the one-dimensional Ising model, where $P$ was a $2 \times 2$ matrix, the transfer matrix has now the size $2^N \times 2^N$, because $\mu$ represents $2^N$ different configurations, namely all the configurations within one row.

The problem of calculating the partition function has thus again been reduced to the determination of the largest eigenvalue $\lambda_{\max}$ of a matrix, in this case the $2^N \times 2^N$ matrix $P$. Denoting by $U$ the matrix which brings $P$ into diagonal form we obtain again

$$Z_{M,N} = \operatorname{tr} \big( U \; P \; U^{-1} \big)^M = \operatorname{tr} \begin{pmatrix} \lambda_1^M & & \\ & \ddots & \\ & & \lambda_{2^N}^M \end{pmatrix} = \sum_{i=1}^{2^N} \lambda_i^M, \tag{4.142}$$

where $\lambda_i$ are the eigenvalues of $P$. (It turns out that all $\lambda_i \geq 0$.) So we get

$$\ln Z_{M,N} = \ln \big( \lambda_1^M + \ldots \lambda_{2^N}^M \big) = \ln \lambda_{\max}^M (1 + \ldots), \tag{4.143}$$

and thus for $M \to \infty$

$$\lim_{M \to \infty} \frac{1}{M} \ln Z_{M,N} = \ln \lambda_{\max} + \ldots. \tag{4.144}$$

The determination of this eigenvalue is a lengthy procedure (see, e.g., Huang 1987). For $M = N \to \infty$ and $E_1 = E_2 = J$ one finally obtains for the free energy per lattice site:

$$F(T, B=0) = -k_{\mathrm{B}} T \Bigg[ \ln \Big( 2 \cosh (2\beta J) \Big) + \frac{1}{2\pi} \int_0^{\pi} \mathrm{d}\phi \, \ln \left( \frac{1}{2} \left( 1 + \sqrt{1 - \kappa^2 \sin^2 \phi} \right) \right) \Bigg], \quad (4.145)$$

where

$$\kappa = \frac{2 \sinh(2\beta J)}{\cosh^2(2\beta J)} \, . \quad (4.146)$$

Taking derivatives one obtains the internal energy and finally the specific heat, i.e., the heat capacity per lattice site

$$C(T, B=0) = k_{\mathrm{B}} \frac{2}{\pi} \big( \beta J \ \coth(2\beta J) \big)^2 \times \Big( 2\big( K(\kappa) - E(\kappa) \big) - (1 - \kappa')\big( \frac{\pi}{2} + K(\kappa)\kappa' \big) \Big), \quad (4.147)$$

where

$$K(\kappa) = \int_0^{\pi/2} \mathrm{d}\phi \, \big( 1 - \kappa^2 \sin^2 \phi \big)^{-1/2}$$

is the complete elliptic integral of first kind,

$$E(\kappa) = \int_0^{\pi/2} \mathrm{d}\phi \, \big( 1 - \kappa^2 \sin^2 \phi \big)^{1/2}$$

the complete elliptic integral of second kind, and

$$\kappa' = 2 \ \tanh^2(2\beta J) - 1, \quad \kappa^2 + \kappa^{'2} = 1 \, .$$

The function $K(\kappa)$ becomes singular at $\kappa = 1$, $\kappa' = 0$, and in the vicinity of $\kappa = 1$ one finds

$$E(\kappa) \approx 1, \qquad K(\kappa) \approx \ln \left( \frac{4}{\kappa'} \right), \quad (4.148)$$

i.e., the specific heat has a singularity at $\kappa = 1$. This signalizes a phase transition, and thus the critical temperature $T_{\mathrm{c}}$ or rather $\beta_{\mathrm{c}}$ is determined from

$$\frac{2 \sinh(2\beta_{\mathrm{c}} J)}{\cosh^2(2\beta_{\mathrm{c}} J)} = 1, \quad (4.149)$$

i.e.,

$$2 \sinh(2\beta_{\mathrm{c}} J) = \cosh^2(2\beta_{\mathrm{c}} J) = \sinh^2(2\beta_{\mathrm{c}} J) + 1, \quad (4.150)$$

or

$$\sinh(2\beta_{\mathrm{c}} J) = 1 \, . \quad (4.151)$$

Thus it is given by

$$\frac{J}{k_{\mathrm{B}}T_{\mathrm{c}}} = 0.4407. \tag{4.152}$$

For $T \to T_{\mathrm{c}}$ at $T < T_{\mathrm{c}}$ one finds

$$\frac{C(T)}{k_{\mathrm{B}}} = -0.4945 \ln \left| \frac{T_{\mathrm{c}} - T}{T_{\mathrm{c}}} \right| . \tag{4.153}$$

One does not obtain a power law behavior and hence, strictly speaking, there are no critical exponents $\alpha$ or $\alpha'$. In the literature, however, one often finds $\alpha = \alpha' = 0$ or $= O(\ln)$, which denotes a logarithmic dependence on the distance to the critical point.

For the magnetization one obtains

$$m(T) = \begin{cases} g\frac{\mu_B}{2}\left(1 - [\sinh(2\beta J)]^{-4}\right)^{1/8} & \text{for} \quad T < T_{\mathrm{c}} \\ 0 & \text{for} \quad T > T_{\mathrm{c}} \, , \end{cases} \tag{4.154}$$

i.e., in the limit $T \to T_{\mathrm{c}}$ for $T < T_{\mathrm{c}}$

$$m(T) = 1.224 \left( \frac{T_{\mathrm{c}} - T}{T_{\mathrm{c}}} \right)^{1/8} , \tag{4.155}$$

so the critical exponent $\beta$ is given by $\beta = 1/8 = 0.125$.

For the critical exponent $\gamma$, defined by the behavior of $\chi$ as $T \to T_{\mathrm{c}}$, $\chi \propto \left(\frac{T - T_{\mathrm{c}}}{T_{\mathrm{c}}}\right)^{-\gamma}$, one obtains $\gamma = 1.75$.

Note that no analytic solutions are known for the Ising model in three or more dimensions, but there are high temperature expansions which determine the critical point very accurately. The table of the critical exponents (Table 4.1) for the two-dimensional and the three-dimensional Ising model is taken from the review article (Wilson and Kogut 1974).

**Ferroelectrics** The two-dimensional Ising model represents a special case within a class of exactly solvable two-dimensional models in statistical mechanics. In this class one also finds models which describe ferroelectrics. Ferroelectrics are substances which below a critical temperature exhibit a macroscopic electric dipole moment, i.e., a polarization instead of a magnetization. This can happen for various ordered structures and, correspondingly, the ferroelectrics are subdivided into different groups.

One possible ordered structure is realized in barium titanate $\mathrm{BaTiO_3}$, where below a critical temperature the lattices of the $\mathrm{Ba^{2+}}$ ions and the $\mathrm{Ti^{4+}}$ ions are displaced relative to the $\mathrm{O^{2-}}$ lattice in such a way that a net electric dipole moment results.

**Table 4.1** Critical exponents from experiment and theory (From Wilson and Kogut (1974))

| Critical exponent | Mean field | 2-dim. Ising | 3-dim. Ising | Experiment |
|---|---|---|---|---|
| $\alpha$ | 0 | 0 | 0.12 | 0–0.1 |
| $\beta$ | 1/2 | 1/8 | 0.31 | 0.3–0.38 |
| $\gamma$ | 1 | 1.75 | 1.25 | 1.2–1.4 |

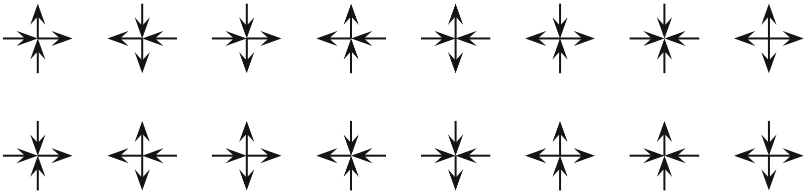


**Fig. 4.9** The sixteen possible configurations for the arrangements of H atoms on the bonds of a two-dimensional *square* lattice. An *arrow* points to the lattice point which is closer to the H atom

A different ordered structure is to be seen in $KH_2PO_4$ (potassium dihydrogen phosphate). In this case the H atoms are between the $PO_4$ complexes, but not in the center. This may be indicated by an arrow, pointing to the lattice site ($PO_4$ complex) which is closer to the H atom. In two dimensions this leads to the sixteen configurations (also called vertices) shown in Fig. 4.9.

The first eight vertices are special in that there are always an even number of arrows pointing inwards or outwards, i.e., close to a vertex are 0, 2, or 4 protons, while the other vertex configurations represent 1 or 3 protons close to the lattice site.

If all vertex configurations can occur, one speaks of a Sixteen-vertex model; if only the first eight vertex configurations are permitted, one has the eight-vertex model or Baxter model. In 1972, Baxter determined the free energy of this model analytically (Baxter 1984). The situation in which only the first six vertex configurations are allowed is called the six-vertex model or, in analogy to ice crystals, the ice model. In this case 2 protons belong to each vertex, i.e., it is electrically neutral.

One can define a certain statistical weight for each vertex configuration and write the partition function using a transfer matrix, as we did in the case of the Ising models. As Baxter has shown, the largest eigenvalue of the transfer matrix for the 8-vertex model can be determined analytically when special weights, depending on three free parameters, are assumed for the vertices.

***n*-vector models.** A further class of models which have been investigated in detail are the $n$-vector models:

$$H = -J \sum_{\boldsymbol{R}} \sum_{\boldsymbol{R}' \in \mathcal{N}_{\boldsymbol{R}}} \boldsymbol{S}(\boldsymbol{R}) \cdot \boldsymbol{S}(\boldsymbol{R}') \, , \tag{4.156}$$

where $\boldsymbol{S}(\boldsymbol{R})$ are now unit vectors in an $n$-dimensional space with components $S_a(\boldsymbol{R})$ such that

$$\sum_{a=1}^{n} S_a^2(\boldsymbol{R}) = 1 \quad \text{for all} \quad \boldsymbol{R} \; . \tag{4.157}$$

The lattice with lattice sites $\boldsymbol{R}$ may be in a $d$-dimensional space. For $n = 1$ one again obtains the Ising model; the case $n = 3$ corresponds to the classical Heisenberg model. Such a classical model for $n = 3$ results formally from a quantum mechanical model (Sect. 6.9) if one introduces the operator

$$\boldsymbol{s}(\boldsymbol{R}) = \frac{1}{\sqrt{S(S+1)}}\, \boldsymbol{S}(\boldsymbol{R}) \tag{4.158}$$

and therefore obtains for the Hamiltonian

$$H = -J\, S(S+1) \sum_{nn} \boldsymbol{s}(\boldsymbol{R}) \cdot \boldsymbol{s}(\boldsymbol{R}') \, . \tag{4.159}$$

Since

$$\sum_{a=1}^{3} S_a^2(\boldsymbol{R}) = S(S+1), \tag{4.160a}$$

$$\left[ S_a(\boldsymbol{R}),\, S_b(\boldsymbol{R}') \right] = \mathrm{i}\, \varepsilon_{abc}\, S_{\mathrm{c}}(\boldsymbol{R}) \delta_{\boldsymbol{R},\boldsymbol{R}'} \tag{4.160b}$$

one now finds for $\boldsymbol{s}(\boldsymbol{R})$

$$\sum_{a=1}^{3} s_a^2(\boldsymbol{R}) = 1, \tag{4.161}$$

$$\left[ s_a(\boldsymbol{R}),\, s_b(\boldsymbol{R}') \right] = \mathrm{i} \frac{1}{\sqrt{S(S+1)}}\, \varepsilon_{abc}\, s_{\mathrm{c}}(\boldsymbol{R}) \delta_{\boldsymbol{R},\boldsymbol{R}'} \; . \tag{4.162}$$

In the limit $S \to \infty$, $J \to 0$, such that $J\, S(S+1)$ remains finite, the operators $s_a(\boldsymbol{R})$ and $s_b(\boldsymbol{R})$ commute and $\boldsymbol{s}(\boldsymbol{R})$ may be represented by a unit vector.

The partition function now reads

$$Z = \prod_{\boldsymbol{R}} \int \mathrm{d}\Omega(\boldsymbol{R})\, \mathrm{e}^{-\beta H} \; . \tag{4.163}$$

Here, $\mathrm{d}\Omega(\boldsymbol{R})$ is the integration measure for an $n-1$-dimensional unit sphere at each lattice site, i.e., for $n = 3$ we have $\mathrm{d}\Omega = \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\varphi$.

For $n > 3$ there exists no magnetic system which can be described by such a model. In 1968, however, it was found that in the limit $n \to \infty$ the model can be solved exactly, also in three and more dimensions.

For the $n$-vector models extensive high-temperature expansions are known as well as all critical exponents. For the theoretical physicist these models serve as a laboratory where he can test approximation methods or study dependences on the lattice dimension or on the symmetry (see also Domb and Green 1974).

## 4.6 The Landau Free Energy

In the previous section we studied models where the density for the magnetization is known for large $N$, or where one can at least determine the characteristic variables like magnetization and entropy analytically. The change in the shape of the density at a critical point is a scenario which certainly may occur in more general models.

In order to study the conditions giving rise to this scenario, we consider a model with partition function

$$Z(\beta, \{\Theta\}, N) = \frac{1}{2^N} \sum_{\{x_i = \pm 1\}} \mathrm{e}^{-\beta H(\boldsymbol{x})}, \tag{4.164}$$

where

$$H(\boldsymbol{x}) = -\sum_{i,j=1}^{N} J_{ij} x_i x_j - \sum_{i=1}^{N} \Theta_i x_i . \tag{4.165}$$

Using the identity

$$\int \mathrm{d}^N y \, \exp\left(-\frac{1}{2}\boldsymbol{y}^T \mathsf{J}^{-1} \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{x}\right) = (2\pi)^{N/2} (\det \mathsf{J})^{1/2} \exp\left(\frac{1}{2}\boldsymbol{x}^T \mathsf{J} \boldsymbol{x}\right) \tag{4.166}$$

for a nonsingular matrix $\mathsf{J}$, the partition function may be put into the form

$$\begin{aligned} Z(\beta, \{\Theta\}, N) &= \frac{1}{2^N} \sum_{\{x_i = \pm 1\}} \int \mathrm{d}^N y \, \exp\left(-\frac{\beta}{2}\boldsymbol{y}^T \mathsf{J}^{-1} \boldsymbol{y} + \beta(\boldsymbol{y} + \boldsymbol{\Theta})^T \boldsymbol{x}\right) \\ &= \int \mathrm{d}^N y \, \mathrm{e}^{-\Lambda(y, \{\Theta_i\}, \beta)}, \end{aligned} \tag{4.167}$$

where the matrix $\mathsf{J}$ has to be chosen appropriately and

$$\Lambda(\boldsymbol{y}, \{\Theta_i\}, \beta) = \frac{\beta}{2}\boldsymbol{y}^T \mathsf{J}^{-1} \boldsymbol{y} + A(\boldsymbol{y}, \{\Theta_i\}, \beta) \tag{4.168}$$

with

$$e^{-\Lambda(\boldsymbol{y},\{\Theta_i\},\beta)} = \frac{1}{2^N} \sum_{\{x_i=\pm 1\}} e^{\beta(\boldsymbol{y}+\boldsymbol{\Theta})^T \boldsymbol{x}} \tag{4.169}$$

$$= \prod_{i=1}^{N} \cosh\left(\beta(y_i + \Theta_i)\right). \tag{4.170}$$

The quantity $\Lambda(\boldsymbol{y}, \{\Theta_i\}, \beta)$ is called the Landau free energy.

The form of this partition function resembles the form of the densities for the magnetization in the last section. Indeed, the partition function for the Curie–Weiss model or in the mean-field approximation also has the form

$$Z(\beta, \Theta, N) = \int \mathrm{d}y\, e^{-N\lambda(y,\Theta)} \,. \tag{4.171}$$

Hence, the Landau free energy $\Lambda(\boldsymbol{y}, \{\Theta_i\}, \beta)$ is the generalization of the quantity $N\lambda(y, \Theta)$.

Of course, $\Lambda$ will also be of the order $N$ and the approximation of the integral over $y$ using the saddle point method,

$$Z(\beta, \{\Theta\}, N) \propto e^{-\Lambda(\boldsymbol{y}_{\mathrm{m}},\{\Theta_i\},\beta)}, \tag{4.172}$$

will be exact in the limit $N \to \infty$. Here $\boldsymbol{y}_{\mathrm{m}}$ is a minimum of $\Lambda(\boldsymbol{y}, \{\Theta_i\}, \beta)$. For the free energy one obtains

$$F(\beta, \{\Theta\}, N) = k_{\mathrm{B}}T \inf_{\boldsymbol{y}} \Big(\Lambda(\boldsymbol{y}, \{\Theta_i\}, \beta)\Big) = k_{\mathrm{B}}T\, \Lambda(\boldsymbol{y}_{\mathrm{m}}, \{\Theta_i\}, \beta) \,. \tag{4.173}$$

The minimum $\boldsymbol{y}_{\mathrm{m}}$ is given by the solution of the equation

$$\frac{\partial \Lambda(\boldsymbol{y}, \{\Theta_i\}, \beta)}{\partial \boldsymbol{y}}|_{\boldsymbol{y}=\boldsymbol{y}_{\mathrm{m}}} = 0 \,, \tag{4.174}$$

under the condition

$$\frac{\partial^2 \Lambda(\boldsymbol{y}, \{B_i\}, \beta)}{\partial y_i \partial y_j}|_{\boldsymbol{y}=\boldsymbol{y}_{\mathrm{m}}} \geq 0. \tag{4.175}$$

Let us apply this procedure to the Curie–Weiss model with the partition function (cp. 4.98)

$$Z(\beta, \Theta, N) = \frac{1}{2^N} \sum_{\{x_i=\pm 1\}} \exp \beta N \left(\frac{J_0}{2} y_N^2 + \Theta y_N\right), \tag{4.176}$$

where $y_N$ denotes the mean value of the $\{x_i\}$. We obtain

$$\Lambda(y,\Theta,N) = N\left[\frac{\beta}{2J_0}y^2 - \ln\cosh\left(\beta(y+\Theta)\right)\right]. \tag{4.177}$$

The determination of the minimum of $\Lambda$ leads to the equation

$$\frac{\beta y}{J_0} = \beta\tanh\left(\beta(y+\Theta)\right). \tag{4.178}$$

Setting $y = J_0 y_{\mathrm{m}}$, this equation is identical to the implicit equation for the magnetization in the Curie–Weiss model (4.108). The second derivative of $\Lambda$ is

$$\frac{\partial^2\Lambda}{\partial y^2} = \beta^2\left[\frac{1}{\beta J_0} - \frac{1}{\cosh^2\left(\beta(y+\Theta)\right)}\right]. \tag{4.179}$$

Obviously, the second derivative at $y = 0$ and $\Theta = 0$ is positive for $\beta J_0 < 1$, i.e., above the critical point.

We also give the expansion of $\Lambda(y,\Theta,N)$ for small $y$:

$$\frac{\Lambda(y,\Theta,N)}{N} = \beta^2\left[\left(\frac{1}{\beta J_0} - 1\right)\frac{y}{2} - y\theta + u_0 y^4 + \ldots\right], \tag{4.180}$$

from which one again easily finds that the coefficient of the $y^2$ term changes its sign at the critical point determined by $(\beta J_0)^{-1} = 1$.

An appropriate ansatz for the form of the Landau free energy serves in general as a starting point for many phenomenological considerations.

We now present two important cases:

- If, instead of the lattice points $i$, one uses continuous space coordinates $\boldsymbol{r}$ to describe the positions of the random variables, the Landau free energy has to be considered as a functional $\Lambda[y(\boldsymbol{r}),\Theta(\boldsymbol{r}),\beta]$.

  In general, one assumes that close to a phase transition $\Lambda[\boldsymbol{y}(\boldsymbol{r}),\Theta(\boldsymbol{r}),\beta]$ can be expanded in powers of $y(\boldsymbol{r})$ and $\nabla y(\boldsymbol{r})$:

  $$\begin{aligned}\Lambda[y(\boldsymbol{r}),\Theta(\boldsymbol{r}),\beta] = {} & \frac{1}{2}(\nabla y(\boldsymbol{r}))^2 - y(\boldsymbol{r})\Theta(\boldsymbol{r}) + \frac{1}{2}r_0(\beta)y(\boldsymbol{r})^2 \\ & + s_0 y(\boldsymbol{r})^3 + u_0 y(\boldsymbol{r})^4 + \ldots.\end{aligned} \tag{4.181}$$

  This continuous version of the Landau free energy permits a particularly elegant formulation of an approximation for the covariance function $\mathrm{Cov}(X(\boldsymbol{r}),X(\boldsymbol{r}'))$. If one restricts $\Lambda$ to the quadratic terms in $\boldsymbol{y}$, the integral becomes a Gaussian integral, which can be evaluated explicitly, at least formally. One finds for the covariance function

$$\mathrm{Cov}(X(\boldsymbol{r}), X(\boldsymbol{r}')) = \frac{e^{|\boldsymbol{r}-\boldsymbol{r}'|/\xi}}{|\boldsymbol{r}-\boldsymbol{r}'|}, \tag{4.182}$$

where $\xi \propto \sqrt{1/r_0(\beta)}$ is the correlation length, which is a measure for the exponential decay of the correlation. If $r_0(\beta) = 0$, which may happen e.g. in the model we consider next, the correlation length becomes infinite. The covariance function now decreases only algebraically and there are long-ranged correlations among the random variables of the field.

- *Phenomenological Landau theory for a phase transition.* If one is only interested in global quantities like the magnetization $m$, one may set $B(\boldsymbol{r}) = B$ and similarly $y(\boldsymbol{r}) = y$, and one obtains

$$\Lambda(y, B, \beta) = -\beta y B + r_0(\beta)\frac{y^2}{2} + s_0 y^3 + u_0 y^4 + \ldots . \tag{4.183}$$

This ansatz represents a model where the behavior of the system at a phase transition is easily demonstrated. Choose $u_0 < 0, s_0 = 0, r_0 = a\tau(\beta)$, such that $\tau$ and therefore also $r$ can assume positive and negative values. The Landau free energy of the Curie–Weiss model (cf. (4.180)) has this structure. This ansatz leads to the same critical exponents in the vicinity of the critical point at $r_0 = 0$ as the Curie–Weiss model and the mean field approximation.

## 4.7 The Renormalization Group Method for Random Fields and Scaling Laws

In Sect. 2.6 we introduced the renormalization transformation and saw that random variables with stable densities are fixed points of these transformations. It seems obvious to define such transformations for random fields too. It will turn out that stable random fields are models for random fields at a phase transition.

### *4.7.1 The Renormalization Transformation*

We define the renormalization transformation as follows:

Consider a random field on a lattice in $d$ dimensions. We subdivide the total lattice into blocks of $b^d$ lattice points ($b = 2, 3, \ldots$), sum up the random variables within each block, and renormalize the resulting $X_i$'s in such a way that they assume the same possible values as the original random variables.

If the possible values of the random variables are $\{-1, 1\}$, as is the case for a spin, we may express such a transformation on the level of realizations as follows: Let $s$ be the sum of the realizations (spin) within one block, then define

$$s' = \begin{cases} +1, \text{ if } & s > 0 \\ -1, \text{ if } & s < 0 \\ s_1, \ \ \text{if} & s = 0, \end{cases} \tag{4.184}$$

where $s_1$ is the realization of the random variable, say, in the upper left corner of the block.

We characterize the densities by the coupling constants $\boldsymbol{K} = (K_1, K_2, \ldots)$ in front of the individual potential terms in the energy function. A renormalization transformation $T_b$ transforms a density with coupling constants $\boldsymbol{K}$ into a density with $\boldsymbol{K}' = T_b \boldsymbol{K}$. The fixed point density $\boldsymbol{K}^*$ then satisfies

$$\boldsymbol{K}^* = T_b \boldsymbol{K}^*. \tag{4.185}$$

As we did in Sect. 2.6, we can make a stability analysis and look for the eigenvectors and eigenvalues of $DT_b$, which is now a matrix with elements

$$(DT_b)_{\alpha\beta} = \left. \frac{\partial (T_b \boldsymbol{K})_\alpha}{\partial K_\beta} \right|_{\boldsymbol{K}=\boldsymbol{K}^*} . \tag{4.186}$$

Again we can expand every vector $\boldsymbol{K} \neq \boldsymbol{K}^*$ in terms of these eigenvectors and characterize the density by the coefficients $\{v_n\}$ of this expansion.

We expect that the density at the critical point corresponds to a fixed point of the renormalization transformation. The relevant scale parameters will be the magnetic field $B$ and the distance from the critical temperature $\varepsilon = (T - T_\mathrm{c})/T_\mathrm{c}$. If we consider the free energy $F(\varepsilon, B)$ as a functional of the density, we will find, in complete analogy to Sect. 2.6, a scaling relation of the form

$$F(\lambda^{a_1}\varepsilon, \lambda^{a_2} B) = \lambda F(\varepsilon, B). \tag{4.187}$$

We will not try to determine the critical exponents for given densities explicitly; see, however, Kadanoff (1966); Wilson (1971); Wegner (1972). It will be seen that even for undetermined parameters $a_1$ and $a_2$ we can say something about the critical behavior.

### 4.7.2 Scaling Laws

Before we study the consequences of the scaling relation, the generality of this relation will be illuminated by two statements about homogeneous functions.

- A function $f(\boldsymbol{r})$ is called homogeneous if the following relation holds

$$f(\lambda \boldsymbol{r}) = g(\lambda) f(\boldsymbol{r}). \tag{4.188}$$

In this case the function $g(\lambda)$ can only be of the form $g(\lambda) = \lambda^p$, with $p$ arbitrary, because from

$$f(\lambda\mu\boldsymbol{r}) = f(\lambda(\mu\boldsymbol{r})) = g(\lambda)g(\mu)f(\boldsymbol{r}) = g(\lambda\mu)f(\boldsymbol{r}) \tag{4.189}$$

one obtains immediately

$$g(\lambda\mu) = g(\lambda)g(\mu), \tag{4.190}$$

which is only possible if $g(\lambda) = \lambda^p$, where $p$ is arbitrary.

- A function $f(x, y)$ is called a generalized homogeneous function if

$$f(\lambda^{a_1}x, \lambda^{a_2}y) = \lambda f(x, y) . \tag{4.191}$$

Seemingly more general, one also might require

$$f(\lambda^{a_1}x, \lambda^{a_2}y) = \lambda^p f(x, y). \tag{4.192}$$

However, setting $\lambda = \mu^{1/p}$ one obtains again

$$f(\mu^{a_1/p}x, \mu^{a_2/p}y) = \mu f(x, y), \tag{4.193}$$

i.e., the same relation as above. Further equivalent relations are

$$f(\lambda x, \lambda^a y) = \lambda^p f(x, y), \tag{4.194}$$

and

$$f(\lambda^a x, \lambda y) = \lambda^p f(x, y). \tag{4.195}$$

Important is the appearance of two constants $a, p$ or $a_1, a_2$.

Let us now draw some consequences from the assumption

$$F(\lambda^{a_1}\varepsilon, \lambda^{a_2}B) = \lambda F(\varepsilon, B). \tag{4.196}$$

- We take the derivative with respect to $B$ and obtain

$$\lambda^{a_2}\frac{\partial F(\lambda^{a_1}\varepsilon, \lambda^{a_2}B)}{\partial\lambda^{a_2}B} = \lambda\frac{\partial F(\varepsilon, B)}{\partial B}, \tag{4.197}$$

i.e., the magnetization satisfies

$$\lambda^{a_2}m(\lambda^{a_1}\varepsilon, \lambda^{a_2}B) = \lambda m(\varepsilon, B) . \tag{4.198}$$

For $B = 0$ this leads to

$$m(\lambda^{a_1}\varepsilon, 0) = \lambda^{1-a_2}m(\varepsilon, 0). \tag{4.199}$$

Setting $\lambda^{a_1} = -1/\varepsilon$, we find

$$m(\varepsilon, 0) = (-1/\varepsilon)^{(1-a_2)/a_1} m(-1, 0). \tag{4.200}$$

So for the critical exponent $\beta$ we obtain

$$\beta = \frac{1-a_2}{a_1}. \tag{4.201}$$

The behavior in the limit $B \to 0$ at $\varepsilon = 0$ follows if we take $\lambda^{a_2} = 1/B$. We get

$$\delta = \frac{a_2}{1-a_2}. \tag{4.202}$$

Thus we have expressed the critical exponents $\beta$ and $\delta$ in terms of $a_1$ and $a_2$. All further exponents are also expressible in terms of $a_1$ and $a_2$, which will give rise to relations among the critical exponents.

- The derivative of (4.198) with respect to $B$ leads to the relation

$$\lambda^{2a_2} \chi(\lambda^{a_1}\varepsilon, \lambda^{a_2} B) = \lambda \chi(\varepsilon, B)\ . \tag{4.203}$$

After some calculation one obtains

$$\gamma = \gamma' = \frac{2a_2 - 1}{a_1} \tag{4.204}$$

and therefore

$$\gamma = \gamma' = \beta(\delta - 1). \tag{4.205}$$

This is called the Widom relation between the critical exponents $\beta$, $\gamma$, and $\delta$.

Taking the second derivative of (4.196) with respect to $\varepsilon$, one obtains

$$\lambda^{2a_1} C_B(\lambda^{a_1}\varepsilon, \lambda^{a_2} B) = \lambda C_B(\varepsilon, B), \tag{4.206}$$

from which one can derive for the critical exponent $\alpha$ in a similar way as above:

$$\alpha = \frac{2a_1 - 1}{a_1}. \tag{4.207}$$

So we find the relations

$$\alpha + \beta(1+\delta) = 2 \qquad \alpha + 2\beta + \gamma = 2\ . \tag{4.208}$$

- Setting $\lambda = |\varepsilon|^{-1/a_1}$ in (4.198) one obtains

$$m(\varepsilon, B) = |\varepsilon|^{(1-a_2)/a_1} m\left(\frac{\varepsilon}{|\varepsilon|}, \frac{B}{|\varepsilon|^{a_2/a_1}}\right) \tag{4.209}$$

or, if we express the coefficients $a_1, a_2$ in terms of the critical exponents,

$$m(\varepsilon, B) = |\varepsilon|^{\beta} m\left(\frac{\varepsilon}{|\varepsilon|}, \frac{B}{|\varepsilon|^{\delta\beta}}\right). \tag{4.210}$$

If we define

$$m' = |\varepsilon|^{-\beta} m(\varepsilon, B) \quad \text{and} \quad B' = |\varepsilon|^{-\beta\delta} B, \tag{4.211}$$

then (4.210) implies that $m'$ is a function only of $B'$, which depends on the sign of $\varepsilon$:

$$m' = f_{\pm}(B') \tag{4.212}$$

or, conversely

$$B' = F_{\pm}(m'), \tag{4.213}$$

i.e., $B' = B/|\varepsilon|^{\delta\beta}$, plotted as a function of $m' = |\varepsilon|^{-\beta} m(\varepsilon, B)$, follows a master curve $F_{\pm}$, independent of the temperature $T$. If the scaling relation did not hold, we would get different curves for different temperatures. The experimental data for $CrBr_3$ and for nickel confirm this consequence of the scaling law (Ho and Litster 1969; Kouvel and Comly 1968).

*Remark.* Corrections to the scaling behavior are also found when an irrelevant scaling parameter is taken into account in the scaling relation. In this case the scaling relation may be written as

$$F(v_1, v_2, v_3) = \lambda^{-1} F(\lambda^{a_1} v_1, \lambda^{a_2} v_2, \lambda^{a_3} v_3), \tag{4.214}$$

where $v_1, v_2$ again denote the relevant parameters and $v_3$ now denotes the irrelevant parameter. Setting, e.g., $\lambda^{a_1}|v_1| = 1$, in order to study the behavior for $v_1 \to 0$, we find

$$F(v_1, v_2, v_3) = |v_1|^{1/a_1} F\left(\pm 1, \frac{v_2}{|v_1|^{a_2/a_1}}, \frac{v_3}{|v_1|^{a_3/a_1}}\right). \tag{4.215}$$

Since we now have $a_3 < 0$, the ratio $v_3/|v_1|^{a_3/a_1}$ becomes very small in the limit $v_1 \to 0$. If $F$ can be expanded around $v_3 = 0$, at $v_2 = 0$ one obtains

$$F(v_1, 0, v_3) = |v_1|^{1/a_1} F(\pm 1, 0, 0) + |v_1|^{(1-a_3)/a_1} v_3 F'(\pm 1, 0, 0). \tag{4.216}$$

Sufficiently far away from the critical point this correction to the leading behavior $\propto |v_1|^{1/a_1}$ can be observed. This is indeed the case for the superfluid phase transition in $He^4$ (Ahlers 1973).

# Chapter 5
# Time-Dependent Random Variables: Classical Stochastic Processes

If one considers a random variable which depends on time, one is led to the concept of a stochastic process. After the definition of a general stochastic process in Sect. 5.1, we introduce the class of Markov processes. In Sect. 5.2 the master equation is formulated, an equation describing the time evolution of the probability density of a Markov process. In this context, the relevance of the master equation for the description of the dynamics of general open systems will be emphasized.

In Sect. 5.3, the reader will get to know important Markov processes and the corresponding master equations: the random walk, the Poisson process, the radioactive decay process, chemical reactions, reaction–diffusion systems, and, finally, scattering processes, in which connection the Boltzmann equation is introduced.

Analytic solutions for simple master equations will be derived in Sect. 5.4. It will become obvious that analytic solvability can only be an exceptional case, so that for most applications one has to rely on numerical methods for the simulation of stochastic processes. These will be explained in Sect. 5.5, as will the Monte Carlo method, with which realizations of random fields, such as spin systems or texture images, can be generated.

Actual diffusion processes will be defined in Sect. 5.6 and the corresponding equation for the time evolution of the probability density, the Fokker–Planck equation, will be discussed. For such processes the time evolution of the random variables themselves can be described by a stochastic differential equation; this is the Langevin equation. Stochastic differential equations with multiplicative noise and numerical methods for the simulation of stochastic differential equations are mentioned, but a thorough discussion of these topics is beyond the scope of this book.

In Sect. 5.7, the response function for a diffusion process will be introduced and the fluctuation–dissipation theorem connecting the response function and the covariance function will be discussed.

Section 5.8 will address more advanced topics. The important $\Omega$-expansion for master equations will lead either to the Fokker–Planck equation or to a deterministic process. This will make the differential equations in the theory of chemical reaction kinetics more intelligible. Furthermore we will illustrate how a factorization ansatz

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_5, © Springer-Verlag Berlin Heidelberg 2012

can approximately reduce a many particle problem to a single-particle problem. This corresponds to the mean field approximation, which has been met already in Chap. 4.

Finally, in Sect. 5.9, more general stochastic processes are considered. Self-similar processes, fractal Brownian motion, stable Levy processes, and autoregressive processes are defined and their relevance for modeling stochastic dynamical systems is indicated.

## 5.1 Markov Processes

Roughly speaking, a stochastic process is defined by a time-dependent random variable $Z(t)$. Let us suppose for the moment that we are dealing only with a discrete sequence of instants $t_i$, $i = 1, \ldots$. The stochastic process $\{ Z(t_i),\ i = 1, \ldots \}$ is then a sequence of random variables. The probability density may be different for each instant and the random variables at different instants are in general not independent, i.e., not only the probabilities for each instant $t_i$ but in addition all common probability densities will characterize the sequence of random variables $\{Z(t_i),\ i = 1, \ldots\}$.

In order to elucidate these remarks and also to take into account the case of continuous time, we denote the possible states by $z$, or rather $z_1$, $z_2, \ldots$, and consider

- The probability density $\varrho_1\big(z, t\big)$ that the state $z$ is present at time $t$;
- The conditional probability $\varrho_2\big(z_2, t_2 \mid z_1, t_1\big)$ that at time $t_2$ the state $z_2$ can be found, if at time $t_1 < t_2$ the state $z_1$ is present;
- The conditional probabilities

$$\varrho_n\big(z_n, t_n \mid z_{n-1}, t_{n-1},\ z_{n-2}, t_{n-2},\ \ldots,\ z_1, t_1\big)\ , \qquad n = 3, \ldots \tag{5.1}$$

that at time $t_n > t_{n-1}$ the state $z_n$ can be found, if at times $t_{n-1} > \ldots > t_1$ the respective states $z_{n-1}, \ldots, z_1$ are present.

The specification of all these probabilities defines a stochastic process. In general it therefore consists of infinitely many defining quantities. Of course, it is possible to define special processes which require fewer of these quantities by making certain assumptions about the probabilities. A very important assumption of this kind is the following:

All conditional probabilities satisfy the relation

$$\varrho_n\big(z_n, t_n \mid z_{n-1}, t_{n-1},\ \ldots,\ z_1, t_1\big) = \varrho_2\big(z_n, t_n \mid z_{n-1}, t_{n-1}\big)\ , \tag{5.2}$$

that means that the random variable $Z_n$ at time $t_n$ is independent of the random variables $Z_{n-2}, \ldots, Z_1$ at times $t_{n-2}, \ldots, t_1$, respectively, given a realization $z_{n-1}$ of the random variable at time $t_{n-1}$. Or, in other words, if at some time $t_{n-1}$ the state $z_{n-1}$ is given, then the probability that the system is in state $z_n$ at time $t_n$, is independent of the earlier history, i.e., of the states of the system before the instant $t_{n-1}$. One also speaks of a process without memory. A process with this property is called a Markov process and the assumption (5.2) is called a Markov assumption. This assumption has the consequence that all common distributions,

$$\varrho_n\big(z_n, t_n;\, z_{n-1}, t_{n-1};\, \dots,\, z_1, t_1\big)\;, \tag{5.3}$$

i.e., the probability that at time $t_1$ the value $z_1$, at $t_2$ the value $z_2$ etc. can be found, can be expressed by $\varrho_1(z, t)$ and $\varrho_2(z, t \mid z', t')$ alone. For instance, for $t_3 > t_2 > t_1$ we have

$$\begin{aligned}\varrho_3\big(z_3, t_3;\, z_2, t_2;\, z_1, t_1\big) &= \varrho_3\big(z_3, t_3 \mid z_2, t_2,\, z_1, t_1\big)\, \varrho_2\big(z_2, t_2;\, z_1, t_1\big)\\ &= \varrho_2\big(z_3, t_3 \mid z_2, t_2\big)\, \varrho_2\big(z_2, t_2 \mid z_1, t_1\big)\, \varrho_1\big(z_1, t_1\big)\;.\end{aligned} \tag{5.4}$$

The Markov property is an idealization that makes it possible to specify a stochastic process by only a few defining quantities, namely, by the conditional probability $\varrho(z, t \mid z', t')$, which is also called transition probability. If the transition probability depends only on the time difference $t - t'$, one speaks of a temporally homogeneous Markov process.

The Markov assumption leads to a certain consistency equation for the transition probability. Namely, from (5.4) one obtains by integrating over $z_2$:

$$\begin{aligned}\varrho_2\big(z_3, t_3;\, z_1, t_1\big) &= \int \mathrm{d}z_2\, \varrho_3\big(z_3, t_3;\, z_2, t_2;\, z_1, t_1\big) \qquad (5.5)\\ &= \int \mathrm{d}z_2\, \varrho_2\big(z_3, t_3 \mid z_2, t_2\big)\, \varrho_2\big(z_2, t_2 \mid z_1, t_1\big)\, \varrho_1\big(z_1, t_1\big)\;,\end{aligned}$$

and therefore the following identity has to hold:

$$\varrho_2\big(z_3, t_3 \mid z_1, t_1\big) = \int \mathrm{d}z_2\, \varrho_2\big(z_3, t_3 \mid z_2, t_2\big)\, \varrho_2\big(z_2, t_2 \mid z_1, t_1\big)\;. \tag{5.6}$$

This equation is also called Chapman–Kolmogorov equation. A similar identity has to be satisfied by $\varrho_1(z, t)$, because, if we multiply the Chapman–Kolmogorov equation by $\varrho_1\big(z_1, t_1\big)$ and integrate over $z_1$, we obtain

$$\varrho_1\big(z_3, t_3\big) = \int \mathrm{d}z_2\, \varrho_2\big(z_3, t_3 \mid z_2, t_2\big)\, \varrho_1\big(z_2, t_2\big)\;. \tag{5.7}$$

Thus the transition probability $\varrho_2\big(z_3, t_3 \mid z_2, t_2\big)$ also mediates between probability densities $\varrho_1(z, t)$ at different times. Therefore, it completely determines the evolution of the stochastic process, so that only the probability density $\varrho_1\big(z, t\big)$ at an initial instant $t_0$ has to be specified. If, for instance, $\varrho_1\bigl(z, t_0\bigr) = \delta(z - z_0)$, then $\varrho_1\bigl(z, t\bigr) = \varrho_2\bigl(z, t \mid z_0, t_0\bigr)$.

The density

$$\varrho_{\text{stat}}(z) = \lim_{t\to\infty} \varrho_1\bigl(z, t\bigr) \ , \tag{5.8}$$

if it exists, is called the stationary distribution. For large times the system approaches this stationary distribution. Of course, we also have

$$\varrho_{\text{stat}}(z) = \int \mathrm{d}z' \, \varrho_2\bigl(z, t \mid z', t'\bigr) \, \varrho_{\text{stat}}(z') \ , \tag{5.9}$$

independent of $t$ and $t'$.

A particularly simple Markov process results if we require

$$\varrho_2\bigl(z, t \mid z', t'\bigr) = \varrho_1\bigl(z, t\bigr) \ . \tag{5.10}$$

In this case the probability density $\varrho_2\bigl(z, t \mid z', t'\bigr)$ at a instant $t$ is even independent of the state which was present at an earlier instant $t'$. Hence, the random variables at different times are independent, and from $\varrho_2\bigl(z_1, t_1;\, z_2, t_2\bigr) = \varrho_2\bigl(z_1, t_1 \mid z_2, t_2\bigr)$ $\varrho_1\bigl(z_2, t_2\bigr)$ one obtains

$$\varrho_2\bigl(z_1, t_1;\, z_2, t_2\bigr) = \varrho_1\bigl(z_1, t_1\bigr) \, \varrho_1\bigl(z_2, t_2\bigr) \ . \tag{5.11}$$

If the density also satisfies $\varrho\bigl(z, t\bigr) \equiv \varrho(z)$, i.e., if it is independent of time, one speaks of independent and identically distributed random variables and writes $Z(t) \sim \mathrm{IID}(\mu, \sigma^2)$, where $\mu$ denotes the mean value and $\sigma^2$ the variance of the distribution $\varrho(z)$. (IID stands for 'independent and identically distributed'.)

If, furthermore, $\varrho(z)$ is the density of a normal distribution $\varrho_G(0, \sigma^2; z)$, the stochastic process is also called a Gaussian white noise (sometimes simply 'white noise'). For each instant $t$, $Z(t)$ is a normally distributed random variable and the realized value at one instant is independent of those at other instants. One also writes

$$Z(t) \propto \mathrm{WN}(0, \sigma^2) \ , \tag{5.12}$$

where W stands for 'white' and N for 'normal'.

The central limit theorem tells us (cf. Sect. 2.5) that, to a good approximation, a sum of random variables may be represented by a normally distributed random variable. Hence, if the sum of many fluctuating influences acts on a system, the total influence at each instant can be thought of as a Gaussian distributed random variable. The time correlations can be neglected if they only take place on time scales that are small with respect to the relevant time scales of the system. Therefore, the white noise is the stochastic standard process describing in a simple way the influence of fast subsystems whose fluctuations are noticeable only on a very short time scale.

In the following, we will denote a white noise by $\eta(t)$. Thus

$$\langle \eta(t) \rangle = 0 \, , \tag{5.13}$$

and the covariance of $\eta(t)$ and $\eta(t')$, known as two-time covariance, satisfies

$$\langle \eta(t)\eta(t') \rangle = \sigma^2 \delta_{tt'} \, , \tag{5.14}$$

or, if we consider time as a continuous variable,

$$\langle \eta(t)\eta(t') \rangle = \sigma^2 \delta(t - t') \, . \tag{5.15}$$

In a continuous-time process, the Fourier transform of the stationary two-time covariance function is

$$F(\omega) = \int \mathrm{d}\tau \langle \eta(t)\eta(t+\tau) \rangle \mathrm{e}^{\mathrm{i}\omega\tau} = \sigma^2, \tag{5.16}$$

i.e., it does not depend on $\omega$, which is a consequence of the white noise being uncorrelated in time. This is the source of the expression 'white' noise. For a 'red' noise $X(t)$, which we will discuss in Sect. 5.9, one obtains, in the case of a temporally homogeneous Markov process,

$$\langle X(t)\, X(t+\tau) \rangle = \frac{\sigma^2}{2m} \mathrm{e}^{-m|\tau|} \tag{5.17}$$

with a time constant $m^{-1}$. The Fourier transform now is

$$F(\omega) = \frac{\sigma^2}{\omega^2 + m^2} \, . \tag{5.18}$$

This noise is called 'red', because $F(\omega)$ increases as $\omega \to 0$. Figure 5.1 shows a typical realization of a white noise and a red noise.

*Example.* Let $X(t), t = 1, 2, \ldots$ be a stochastic process in discrete time, defined by the equation

$$X(t) = \alpha X(t-1) + \sigma\eta(t), \qquad 0 < \alpha < 1, \eta \sim WN(0, 1). \tag{5.19}$$

This is obviously a Markov process: Given $x(t-1)$, the distribution of $X(t)$ is known. Iterating this equation to

$$X(t) = \alpha^2 X(t-2) + \alpha\sigma\eta(t-1) + \sigma\eta(t), \tag{5.20}$$

again reveals the Markov character of the process: Once one realization in the past is given, the random variable at present does not depend on other random variables further in the past.

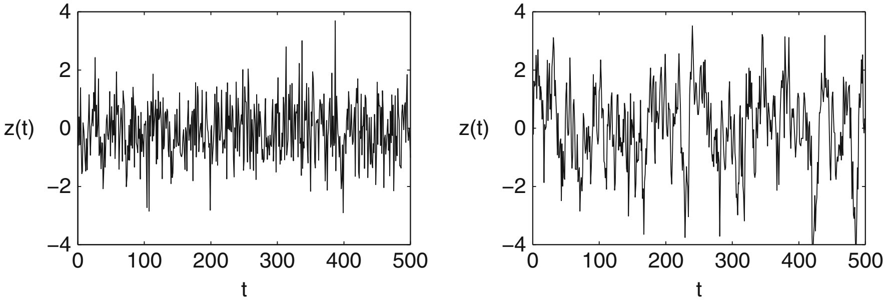


**Fig. 5.1** Realizations of a white noise (*left*) and a red noise (*right*). The time correlations for a red noise cause the evolution in time to be 'smoother' than it is for a white noise

## 5.2 The Master Equation

A transition probability obviously has to satisfy

$$\varrho_2\big(z, t + \tau \mid z', t\big) \to \delta(z - z') \qquad \text{for} \qquad \tau \to 0\,. \tag{5.21}$$

The next term in an expansion of the transition probability with respect to $\tau$, the so-called short-time behavior, is also essential for the properties of the process. We write the expansion in $\tau$ in the following form:

$$\varrho_2\big(z, t + \tau \mid z', t\big) = \big(1 - a(z', t)\tau\big)\,\delta(z - z') + \tau\, w(z, z', t) + O(\tau^2)\,. \tag{5.22}$$

From the normalization condition for $\varrho_2\big(z, t + \tau \mid z', t\big)$ it follows that, with $t > t'$, $\tau > 0$

$$1 = \int \mathrm{d}z\, \varrho_2\big(z, t + \tau \mid z', t\big) \tag{5.23}$$

$$= 1 - a(z', t)\tau + \tau \int \mathrm{d}z\, w(z, z', t) + O(\tau^2)\,, \tag{5.24}$$

i.e.,

$$a(z', t) = \int \mathrm{d}z\, w(z, z', t)\,. \tag{5.25}$$

The form (5.22) implies for the realization of a random variable $Z(t)$ that for each sufficiently small time step $\tau$ it keeps the same value with a certain probability or assumes a different value with a complementary probability. A typical trajectory therefore consists of lines $z_t$ = const., interrupted by jumps.

From the Chapman–Kolmogorov equation and (5.22) it follows that

$$\begin{aligned}\varrho_2(z,t+\tau\mid z',t') &= \int \mathrm{d}z''\,\varrho_2(z,t+\tau\mid z'',t)\,\varrho_2(z'',t\mid z',t')\\ &= (1-a(z,t)\tau)\,\varrho_2(z,t\mid z',t')\\ &\quad +\tau\int \mathrm{d}z''\,w(z,z'',t)\,\varrho_2(z'',t\mid z',t') + O(\tau^2)\;.\end{aligned} \tag{5.26}$$

Hence, in the limit $\tau \to 0$ one obtains:

$$\begin{aligned}\frac{\partial\varrho_2(z,t\mid z',t')}{\partial t} &= \int \mathrm{d}z''\,w(z,z'',t)\,\varrho_2(z'',t\mid z',t')\\ &\quad -\int \mathrm{d}z''\,w(z'',z,t)\,\varrho_2(z,t\mid z',t')\;,\end{aligned} \tag{5.27}$$

where we have used (5.25) to express $a(z,t)$ in terms of $w(z'',z,t)$.

Multiplying (5.27) by $\varrho_1(z',t')$ and integrating over $z'$ yields a similar equation for $\varrho_1(z,t)$:

$$\frac{\partial}{\partial t}\,\varrho_1(z,t) = \int \mathrm{d}z' w(z,z',t)\varrho_1(z',t) - \int \mathrm{d}z' w(z',z,t)\varrho_1(z,t)\;. \tag{5.28}$$

If the states are discrete, we may denote them with $n$ instead of $z$ and write $\varrho_n(t)$ and $w_{nn'}(t)$ instead of $\varrho_1(z,t)$ and $w(z,z',t)$, respectively. The equation for $\varrho_n(t)$ then reads

$$\dot{\varrho}_n(t) = \sum_{n'}\Big(w_{nn'}(t)\varrho_{n'}(t) - w_{n'n}(t)\varrho_n(t)\Big)\;. \tag{5.29}$$

The change in the probability of finding the state $n$ at time $t$ comprises a term representing the gain (transitions $n' \to n$) and a term representing the loss (transitions $n \to n'$). For a process which is homogeneous in time the transition rates $w(z,z',t)$ or $w_{nn'}$ are independent of $t$.

Notice that the physical dimension of the transition rates is $s^{-1}$. Multiplying all $w_{nn'}$ by the same factor $\alpha$ can be compensated by dividing the time $t$ by $\alpha$.

The equations (5.27)–(5.29) are called master equations. They describe the time dependence of the probability densities. The essential quantities for these equations are the transition rates $w(z,z',t)$ or $w_{nn'}(t)$.

If the states are discrete, they can be ordered in a row. Introducing the vector $\boldsymbol{\varrho}(t) = (\varrho_1(t),\ldots,\varrho_\alpha(t),\ldots)$, where $\alpha$ enumerates the states, the master equation may then be written in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\varrho_\alpha(t) = \sum_{\alpha'}\mathsf{V}_{\alpha\alpha'}\varrho_{\alpha'}(t)\;. \tag{5.30}$$

Here, the matrix $\mathsf{V}$ with elements $\mathsf{V}_{\alpha\alpha'}$ contains the transition rates $w_{n'n}$.

The conservation of the normalization implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{\alpha}\varrho_{\alpha}(t)=0\ ,\quad \text{i.e.,}\quad \sum_{\alpha}\mathsf{V}_{\alpha\alpha'}=0\ . \tag{5.31}$$

Hence, the sums of the elements in the individual columns of the matrix $\mathsf{V}$ have to vanish.

The general solution of such a master equation for a given initial value $\boldsymbol{\varrho}(0)$ is

$$\boldsymbol{\varrho}(t)=\mathrm{e}^{t\mathsf{V}}\boldsymbol{\varrho}(0)\ . \tag{5.32}$$

One can show easily that the matrix $\mathsf{V}$ is diagonalizable if there exists a stationary solution $\varrho_{\alpha}^{*}\neq 0$ for all states $\alpha$ such that

$$\mathsf{V}_{\alpha\alpha'}\varrho_{\alpha'}^{*}=\mathsf{V}_{\alpha'\alpha}\varrho_{\alpha}^{*}\quad\text{(no summation)}\ . \tag{5.33}$$

In this case the matrix $\overline{\mathsf{V}}$ with elements

$$\overline{\mathsf{V}}_{\alpha\alpha'}=(\varrho_{\alpha}^{*})^{-1/2}\mathsf{V}_{\alpha\alpha'}(\varrho_{\alpha'}^{*})^{1/2} \tag{5.34}$$

is a symmetric matrix and can therefore be diagonalized. Therefore $\mathsf{V}$ is also diagonalizable.

The symmetry of $\overline{\mathsf{V}}$ follows from (5.33):

$$\begin{aligned}\overline{\mathsf{V}}_{\alpha'\alpha}&=(\varrho_{\alpha'}^{*})^{-1/2}\mathsf{V}_{\alpha'\alpha}(\varrho_{\alpha}^{*})^{1/2}=(\varrho_{\alpha'}^{*})^{-1/2}\mathsf{V}_{\alpha'\alpha}\varrho_{\alpha}^{*}(\varrho_{\alpha}^{*})^{-1/2}\\&=(\varrho_{\alpha'}^{*})^{-1/2}\mathsf{V}_{\alpha\alpha'}\varrho_{\alpha'}^{*}(\varrho_{\alpha}^{*})^{-1/2}=(\varrho_{\alpha}^{*})^{-1/2}\mathsf{V}_{\alpha\alpha'}(\varrho_{\alpha'}^{*})^{1/2}=\overline{\mathsf{V}}_{\alpha\alpha'}\ .\end{aligned} \tag{5.35}$$

Condition (5.33) is identical to the condition

$$w_{nn'}\varrho_{n'}^{*}=w_{n'n}\varrho_{n}^{*} \tag{5.36}$$

for each pair $n, n'$. In general, this requirement is stronger than stationarity, which can be written as

$$\sum_{n'\neq n}w_{nn'}\varrho_{n'}^{*}=\sum_{n'\neq n}w_{n'n}\varrho_{n}^{*}\ . \tag{5.37}$$

Stationarity (5.37) means that the total current flowing into state $n$ is identical to the current flowing out of $n$. On the other hand, the requirement (5.36) implies in addition that the current from $n \to n'$ is identical to the current from $n' \to n$.

Equation 5.36 is called the condition of detailed balance. If it holds, the balance equation (5.37) is separately valid for each $n'$. It can be shown that such a property of the transition rates follows from the invariance of the process under time reversal (van Kampen 1985).

It is also possible to define the transition probability of a deterministic process, e.g., a process given by the differential equation

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \tag{5.38}$$

for $\boldsymbol{x} \in \mathbb{R}^n$. Then

$$\varrho_2\big(z, t \mid \boldsymbol{x}_0, t_0\big) = \delta\big(z - \boldsymbol{x}(t)\big) \ , \tag{5.39}$$

where $\boldsymbol{x}(t)$ is the solution of the differential equation for the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$. Obviously, in this form $\varrho_2\big(z, t \mid \boldsymbol{x}_0, t_0\big)$ may also be interpreted as a probability density. We obtain for the short-time behavior

$$\begin{aligned} \varrho_2\big(z, t + \tau \mid z', t\big) &= \delta\Big(z - \big(z' + \boldsymbol{f}(z')\tau\big)\Big) \\ &= \delta\big(z - z'\big) - \boldsymbol{f}(z') \cdot \frac{\partial}{\partial z}\,\delta\big(z - z'\big)\,\tau + O(\tau^2) \ , \end{aligned} \tag{5.40}$$

i.e., we now have

$$w(z, z') = -\boldsymbol{f}(z') \cdot \frac{\partial}{\partial z}\,\delta\big(z - z'\big) \ , \tag{5.41}$$

and therefore also

$$a(z') = \int \mathrm{d}^n z\, w(z, z') = -\int \mathrm{d}^n z\, \boldsymbol{f}(z') \cdot \frac{\partial}{\partial z}\,\delta\big(z - z'\big) = 0 \ . \tag{5.42}$$

From (5.28) we thus obtain the resulting master equation

$$\begin{aligned} \frac{\partial \varrho_1(z, t)}{\partial t} &= \int \mathrm{d}^n z' \Big(-\boldsymbol{f}(z') \cdot \frac{\partial}{\partial z}\,\delta\big(z - z'\big)\Big)\,\varrho_1(z', t) \\ &= -\frac{\partial}{\partial z} \cdot \big(\boldsymbol{f}(z)\varrho_1(z, t)\big) \ . \end{aligned} \tag{5.43}$$

This differential equation for the time evolution of a probability density based on a deterministic process is known as Liouville's equation. If $\varrho(p, q)$ is a density in phase space and $H(p, q)$ a Hamiltonian function, the Liouville equation is just

$$\frac{\partial}{\partial t}\,\varrho(p, q; t) = \big\{H, \varrho\big\} = \frac{\partial H}{\partial q}\,\frac{\partial \varrho}{\partial p} - \frac{\partial H}{\partial p}\,\frac{\partial \varrho}{\partial q} \ , \tag{5.44}$$

where $\{H, \varrho\}$ denotes the Poisson bracket.

Indeed, for $z = (p, q)$, $\dot{z} = \boldsymbol{f}$, i.e., $(\dot{p}, \dot{q}) = \Big(-\frac{\partial H}{\partial q}, \frac{\partial H}{\partial p}\Big)$, the Liouville equation can be written in the form

$$\frac{\partial \varrho(z, t)}{\partial t} = -\frac{\partial}{\partial z} \cdot \big(\boldsymbol{f}(z)\varrho(z, t)\big) \ , \tag{5.45}$$

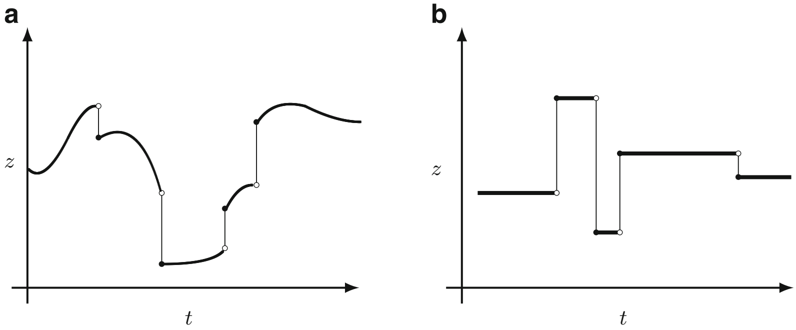


**Fig. 5.2** (**a**) Realization of a piecewise deterministic Markov process: the deterministic evolution according to $\dot{z} = f(z)$ is interrupted by jumps. The probability of a jump is determined by the transition probability $w(z, z')$. For $f \equiv 0$ one finds a trajectory such as that shown in (**b**)

since

$$\begin{aligned} -\frac{\partial}{\partial z} \cdot \boldsymbol{f}(z) &= -\frac{\partial}{\partial p}\left(-\frac{\partial H}{\partial q}\right) - \frac{\partial}{\partial q}\frac{\partial H}{\partial p} \\ &= \frac{\partial H}{\partial q}\frac{\partial}{\partial p} - \frac{\partial H}{\partial p}\frac{\partial}{\partial q}\,. \end{aligned} \tag{5.46}$$

An important combination of the two above-mentioned types of processes is the piecewise deterministic Markov processes. In this case the short-time behavior of the transition probability may be written in the form

$$\begin{aligned} \varrho_2\big(z, t + \tau \mid z', t\big) = (1 - a(z,t)\tau\big)\delta(z - z') + \tau\, w(z, z', t) \\ -\boldsymbol{f}(z', t) \cdot \frac{\partial}{\partial z}\,\delta(z - z')\tau + O(\tau^2) \end{aligned} \tag{5.47}$$

and the corresponding master equation reads

$$\begin{aligned} \frac{\partial}{\partial t}\,\varrho_1(z,t) &= -\frac{\partial}{\partial z} \cdot (\boldsymbol{f}(z,t)\,\varrho_1(z,t)) \\ &\quad + \int \mathrm{d}^n z'\, w(z, z', t)\,\varrho_1(z', t) - \int \mathrm{d}^n z'\, w(z', z, t)\,\varrho_1(z, t)\,. \end{aligned} \tag{5.48}$$

Thus a typical trajectory of such a process consists of pieces where $z(t)$ evolves deterministically according to $\dot{z} = \boldsymbol{f}(z)$. These pieces end with a jump, and a new deterministic evolution starts from the final point of this jump (Fig. 5.2).

Hence, we have developed a framework which allows the formulation of deterministic as well as special stochastic processes. The time evolution is dictated by the transition rates $w(z, z', t)$, which take over the role of the Hamiltonian in classical deterministic systems, and indeed the transition rates for a deterministic process also contain the Hamiltonian.

Therefore, the master equation is a generalization of Liouville's equation to open systems.

*Remarks.*

(a) When the master equation is used to describe transitions between quantum states, it is sometimes referred to as the Pauli equation. The transition probabilities $w_{nn'}$ are the usual matrix elements calculated in perturbation theory.

(b) Prigogine, Resibois, Montroll, and others (Prigogine and Resibois 1961; Montroll 1962) have also developed generalized master equations. The dwell times for the individual states, which for the standard master equation are distributed exponentially, can in these cases be chosen arbitrarily. For an excellent discussion of generalized master equations and more general stochastic processes, see also Montroll and West (1987).

Let us now look at two applications of the detailed balance equation:

(a) We consider a system with quantum states $\{|n\rangle\}$ at a temperature $T$ and in contact with a radiation field. Then the probability for a state $n$ is

$$\varrho_n^e = \frac{1}{Z}\,\mathrm{e}^{-\beta E_n}\ , \tag{5.49}$$

and in a transition $n \to n'$ with a transition rate $w_{n'n}$ a photon with energy $\hbar\omega = E_n - E_{n'}$ is emitted. From quantum mechanics it is known that

$$\frac{w_{n'n}}{w_{nn'}} = \frac{\big(n(\omega)+1\big)}{n(\omega)}\ , \tag{5.50}$$

where $n(\omega)$ is the average number of photons with frequency $\omega$ in a radiation field. This is a plausible relation, as the transition $n \to n'$ increases the number of photons in the radiation field by 1. From the equation of detailed balance one obtains, in addition,

$$w_{n'n}\,\varrho_n^e = w_{nn'}\,\varrho_{n'}^e\ , \tag{5.51}$$

i.e.,

$$\frac{w_{n'n}}{w_{nn'}} = \frac{\varrho_{n'}^e}{\varrho_n^e} = \mathrm{e}^{-\beta(E_{n'}-E_n)} = \frac{\big(n(\omega)+1\big)}{n(\omega)}\ , \tag{5.52}$$

and thus by solving for $n(\omega)$ we find

$$n(\omega) = \frac{1}{\mathrm{e}^{\beta\hbar\omega}-1}\ . \tag{5.53}$$

This is exactly the average number of particles in a photon gas. We will obtain the same result in a different way in Sect. 6.5.

(b) The equation of detailed balance may also be used to define a stochastic process for which the stationary density is given. If we call this also $\{\varrho_n^e\}$, the transition rates are chosen in such a way that, according to the detailed balance condition, their ratio is equal to the ratio of the given densities, i.e.,

$$\frac{w_{nn'}}{w_{n'n}} = \frac{\varrho_n^e}{\varrho_{n'}^e} \,. \tag{5.54}$$

We will show in Sect. 7.4 that each solution $\{\varrho_n(t)\}$ of the master equation of such a process tends towards $\{\varrho_n^e\}$.

In Sect. 5.5 we will use this method to numerically estimate the expectation values of macroscopic system variables in the statistical mechanics of equilibrium states. Furthermore, realizations of random fields can be generated by this method.

## 5.3 Examples of Master Equations

Up to now we have considered Markov processes in the framework of statistical mechanics. However, the concept of modeling a stochastic process as a Markov process extends much further.

Indeed, the prerequisites for this concept are merely the description of the states $\{z_1, \ldots\}$ of a system and the rates for the transitions between these states, i.e., the quantities $w_{n'n}$ for the transition $z_n \rightarrow z_{n'}$. These transition rates determine the transition probability $\varrho_2(z_{n'}, t + \tau \mid z_n, t)$ for sufficiently small time steps $\tau$, and due to the Markov condition this in turn determines the total stochastic process. Yet there are many different ways in which the state of a system may be defined.

- A system may consist of a particle and a state is characterized by the position of the particle. The stochastic process corresponds to jumps from place to place. In this case one speaks of a random walk. In the same way we can also describe e.g. the diffusion of pollutants in the atmosphere or the movement of electrons in a semiconductor.
- If the state is characterized by a particle number, e.g., the number of atoms, molecules, or living individuals of a particular kind, one can respectively describe radioactive decay, chemical processes, or even phenomena of population dynamics.

### 5.3.1 *One-Step Processes*

Particularly simple Markov processes result when the states can be labeled by a discrete one-dimensional variable $n$ and ordered in a linear chain such that the transition rates $w_{nn'}$ are nonzero only for transitions between neighboring states.

Setting $w_{n,n+1} = r(n+1)$, $w_{n,n-1} = g(n-1)$, and therefore also $w_{n-1,n} + w_{n+1,n} = r(n) + g(n)$, the master equation reads

$$\dot{\varrho}_n(t) = r(n+1)\,\varrho_{n+1}(t) + g(n-1)\,\varrho_{n-1}(t) - \big(r(n) + g(n)\big)\,\varrho_n(t)\ , \tag{5.55}$$

where $\varrho_n$ denotes the probability of state $n$.

Such processes are also called one-step processes. Using the operators $\mathsf{E}^k$, defined by

$$\mathsf{E}^k \varrho_n(t) = \varrho_{n+k}(t)\ , \quad k = 0, \pm 1, \dots\ , \tag{5.56}$$

the equation for the one-step process may also be written

$$\dot{\varrho}_n(t) = (\mathsf{E} - 1) r(n) \varrho_n(t) + (\mathsf{E}^{-1} - 1) g(n) \varrho_n(t)\ . \tag{5.57}$$

Important one-step processes are:

*The random walk.* Here $g \equiv r \equiv 1$. The master equation is

$$\dot{\varrho}_n(t) = \varrho_{n+1} + \varrho_{n-1} - 2\varrho_n \equiv \Big((\mathsf{E} - 1) + (\mathsf{E}^{-1} - 1)\Big)\varrho_n\ . \tag{5.58}$$

In this case $n$, with $-\infty < n < \infty$, denotes the position on a linear chain of points.

*Radioactive decay.* In this case $g = 0$, $r(n) = \gamma n$, i.e., the probability that one of $n$ particles decays within the time interval $(t, t + \mathrm{d}t)$ is $\gamma n \mathrm{d}t$. The master equation reads:

$$\dot{\varrho}_n(t) = \gamma(n+1)\,\varrho_{n+1} - \gamma n \varrho_n = (\mathsf{E} - 1)\gamma n\,\varrho_n\ . \tag{5.59}$$

Now, $n = 0, 1, \dots$ denotes the number of atoms. There are only transitions to a lower number and the rate depends linearly on $n$ (see Fig. 5.3a).

*The Poisson process.* In this case $g(n) = \lambda$, $r \equiv 0$.

$$\dot{\varrho}_n(t) = \lambda\,\varrho_{n-1} - \lambda\,\varrho_n = (\mathsf{E}^{-1} - 1)\lambda\,\varrho_n\ , \tag{5.60a}$$

$$\dot{\varrho}_0(t) = -\lambda\,\varrho_0\ . \tag{5.60b}$$

$n = 0, 1, \dots$ represents the number of events that have occurred up to time $t$. This number can only increase with time (Fig. 5.3b). The probability that in the interval $(t, t + \mathrm{d}t)$ one event occurs is equal to $\lambda \mathrm{d}t$.

In this case we had to write down the master equation for $n = 0$ separately. It follows formally from the general master equation for one-step processes if we introduce a fictitious state $n = -1$ and set $g(-1) = 0$. In the same way one should have defined for the radioactive decay $r(0) = 0$, but the function $r(n)$ takes care of this already. Thus there are no transitions between the fictitious state and the boundary state $n = 0$ (Fig. 5.4).

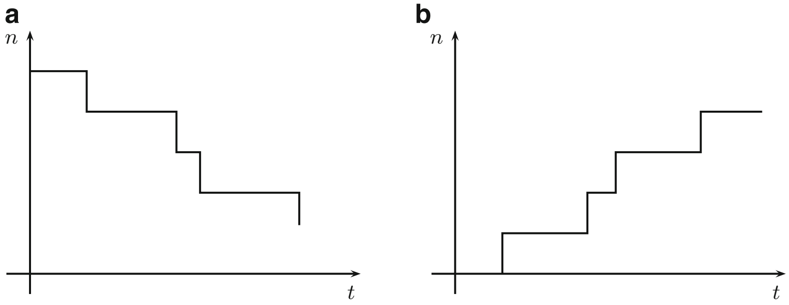


**Fig. 5.3** Realization of the stochastic processes (**a**) for the particle number in a radioactive substance and (**b**) for the number of events in a Poisson process

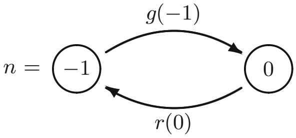


**Fig. 5.4** Setting $g(-1) = 0$ and $r(0) = 0$, there are no transitions between the state $n = 0$ and a fictitious state $n = -1$

Utilizing the rule $g(-1) = 0, r(0) = 0$, one may thus formally use the general form of the one-step process, even when there is a boundary at $n = 0$. In a similar manner, one can take care of other boundaries.

*Chemical reactions.* Suppose that in the chemical reaction $A \rightarrow X$ molecules of $A$ can change in molecules of $X$. Denoting by $n$ the number of $X$ molecules and $n_A$ the number of $A$ molecules we get

$$g(n) = w_{n+1,n} = w\, n_A = w\, \Omega c_A \ , \tag{5.61}$$

where $\Omega$ represents the volume occupied by the particles, $c_A = n_A/\Omega$ is the concentration, and $w$ a proportionality factor, known as the rate constant. A second reaction $X \rightarrow B$ is described by

$$r(n) = w_{n-1,n} = w'\, n \ , \tag{5.62}$$

with the rate constant $w'$. For the case where both reactions

$$A \rightarrow X, \qquad X \rightarrow B$$

occur, the master equation for $\varrho_n(t)$ is

$$\dot{\varrho}_n(t) = w\, c_A \Omega (\mathsf{E}^{-1} - 1)\, \varrho_n + w'\, (\mathsf{E} - 1)\, n\, \varrho_n \ . \tag{5.63}$$

The equations we have considered so far can be solved analytically, as we will show in the next section. The reason is that for these processes the transition rates $w_{n'n}$ depend at most linearly on $n$. As a consequence one obtains closed equations for the moments, and these are easily solved. As in any other field, linear models represent important examples for theory and applications, but in many cases only nonlinear transition rates will realistically describe the corresponding physical problem.

We will now introduce some important examples of such master equations:

### 5.3.2 Chemical Reactions

For the reaction

$$2X \to B \tag{5.64}$$

we denote by $n$ the number of $X$ molecules. Then

$$w_{n-2,n} = w'\, n(n-1)/\Omega \ , \tag{5.65}$$

because with $n$ particles given, there are $n(n-1)/2$ pairs which can form a $B$ molecule in one reaction.

Thus if we replace the chemical reaction $X \to B$ by $2X \to B$, e.g. in (5.63), we obtain the master equation

$$\dot{\varrho}_n(t) = w\, c_A \Omega (\mathsf{E}^{-1} - 1)\, \varrho_n(t) + \frac{w'}{\Omega}(\mathsf{E}^2 - 1)\, n(n-1)\, \varrho_n(t) \ , \tag{5.66}$$

which can no longer be solved analytically. Similarly, all master equations for chemical reactions in which more than one particle of each species take part, i.e., for which some of the stoichiometric coefficients are larger than 1, are nonlinear and therefore not solvable analytically.

*Remark.* In general, one obtains for the reaction $s\,X \to B$ the transition rate (van Kampen 1985)

$$w_{n-s,n} = \lambda' \Omega\, \frac{n(n-1)\cdots(n-s+1)}{\Omega^s} \ . \tag{5.67}$$

This may be seen as follows: Let $\omega$ be the volume within which the particles have to be present simultaneously for a reaction to take place. Then the probability $p_s$ that from a total of $n$ particles in $\Omega$ exactly $s$ particles are in $\omega$ is

$$p_s \propto \binom{n}{s} \left(\frac{\omega}{\Omega}\right)^s \left(1 - \frac{\omega}{\Omega}\right)^{n-s} \ . \tag{5.68}$$

For $\omega \ll \Omega$ one obtains from this expression the rate (5.67) to leading order in $\Omega$, to within a general factor (containing $\Omega$ itself).

### 5.3.3 Reaction–Diffusion Systems

Up to now, in describing chemical reactions, we have assumed a homogeneous spatial distribution of the molecules. In many cases, however, the variations of the concentrations can no longer be neglected. One has to subdivide the total volume of the reaction into small cells $\lambda$, ($\lambda = 1, 2, \ldots$) of volume $\Delta$, small enough that at least inside these cells the assumption of a homogeneous distribution is justified. In this case the stochastic quantities are no longer simply labeled by $n$, the number of molecules in $\Omega$, but by $\{n_\lambda,\ \lambda = 1, 2, \ldots\}$, the number of molecules inside the cells $\lambda = 1, 2, \ldots$, and therefore we have to consider the density $\varrho(n_1, n_2, \ldots; t)$. This characterization of a state of the system is called the occupation number representation, because a state is labeled by the occupation numbers of all cells. Of course, the sum of all $\{n_\lambda\}$ has to be equal to the total number $N$ of particles.

The number of molecules $n_\lambda$ in cell $\lambda$ may now change for one of two reasons: It may change, firstly, due to a chemical reaction, e.g., $A \to X$, $2X \to B$. The master equation for this process is

$$\begin{aligned}\dot{\varrho}\big(\{n_\lambda\}, t\big) = \Delta \sum_\lambda \Big[ & w c_A \left(\mathsf{E}_\lambda^{-1} - 1\right) \\ & + \frac{w'}{\Delta^2}\left(\mathsf{E}_\lambda^2 - 1\right) n_\lambda (n_\lambda - 1) \Big] \varrho\big(\{n_\lambda\}, t\big) \;; \end{aligned} \tag{5.69}$$

Secondly, it may change as a result of diffusion: during the interval $\mathrm{d}t$ a molecule passes from cell $\lambda$ to cell $\mu$. Hence, $n_\lambda$ decreases by 1 while $n_\mu$ increases by 1. We will denote the transition rate for this process by $w_{\mu\lambda} n_\lambda \, \Delta$.

The total master equation now reads

$$\begin{aligned}\dot{\varrho}\big(\{n_\lambda\}, t\big) = \Delta \Big[ & \sum_\lambda w c_A \left(\mathsf{E}_\lambda^{-1} - 1\right) + \frac{w'}{\Delta^2}\left(\mathsf{E}_\lambda^2 - 1\right) n_\lambda (n_\lambda - 1) \\ & + \sum_{\mu,\lambda} w_{\mu\lambda} \left(\mathsf{E}_\mu^{-1} \mathsf{E}_\lambda - 1\right) n_\lambda \Big] \varrho\big(\{n_\lambda\}, t\big) \;. \end{aligned} \tag{5.70}$$

Here $\varrho\big(\{n_\lambda\}, t\big)$ is the probability density for the stochastic process $\{N_\lambda(t)\}$. The subdivision into cells and the introduction of occupation numbers for each cell is, of course, only an auxiliary construction: Our real interest lies in the stochastic process $N(\boldsymbol{r}, t)$, from which $\{N_\lambda(t)\}$ may be obtained according to

$$N_\lambda(t) = \int_{\text{cell}\,\lambda} N(\boldsymbol{r}, t)\, \mathrm{d}^3 r \;, \tag{5.71}$$

and the probability density $\varrho\big[n(\boldsymbol{r}), t\big]$. $N(\boldsymbol{r}, t)$ is a time-dependent stochastic field and the probability density $\varrho\big[n(\boldsymbol{r}), t\big]$ is a functional of $n(\boldsymbol{r})$.

In Sect. 3.6 we have already examined the time-independent case, i.e., a field $N(\boldsymbol{r})$. We also considered the expectation values $\langle N(\boldsymbol{r})\rangle$ and $\langle N(\boldsymbol{r})N(\boldsymbol{r}')\rangle$, and for the case of a spatially homogeneous field we obtained

$$\langle N(\boldsymbol{r})\rangle = \frac{N}{V} = n\ . \tag{5.72}$$

Thus $N(\boldsymbol{r},t)$ is the local particle density, which may now be time dependent.

The master equation (5.70) makes a statement about the behavior of these quantities as a function of space and time, where space has been discretized. Of course, for many applications it is sufficient to derive and solve an equation for the expectation value of the local particle density, i.e., for

$$f(\boldsymbol{r},t) = \langle N(\boldsymbol{r},t)\rangle\ . \tag{5.73}$$

Also of interest, in addition to the expectation value $\langle N(\boldsymbol{r},t)\rangle$, is the so-called factorial cumulant,

$$\begin{aligned}\Big[\,N(\boldsymbol{r},t)\,N(\boldsymbol{r}',t)\Big] = {}& \langle N(\boldsymbol{r},t)\,N(\boldsymbol{r}',t)\rangle - \langle N(\boldsymbol{r},t)\rangle\langle N(\boldsymbol{r}',t)\rangle \\ & -\delta(\boldsymbol{r}-\boldsymbol{r}')\,\langle N(\boldsymbol{r},t)\rangle\ .\end{aligned} \tag{5.74}$$

To within the normalization, this is the time-dependent analog of the radial distribution function $g_2(\boldsymbol{r},\boldsymbol{r}')$. Indeed, the radial distribution function may be introduced as

$$\Big[\,N(\boldsymbol{r},t)\,N(\boldsymbol{r}',t)\Big] = \langle N(\boldsymbol{r},t)\rangle\langle N(\boldsymbol{r}',t)\rangle\,\Big(g_2(\boldsymbol{r},\boldsymbol{r}',t)-1\Big)\ . \tag{5.75}$$

### 5.3.4 Scattering Processes

Another important type of master equation results if one considers particles in a volume $\Omega$ that may also scatter among themselves.

Now it is the phase space which is subdivided into cells $\lambda$, $\lambda = 1,2,\ldots$, and we will again denote the number of particles in cell $\lambda$ by $n_\lambda$. When two particles scatter, particles from cells $\lambda$ and $\mu$ make a transition into cells $\lambda'$ and $\mu'$, i.e., the occupation numbers for the cells $\lambda$ and $\mu$ decrease by 1, while they increase for the cells $\lambda'$ and $\mu'$. Evidently, the rate of this transition from the state $z = \big(n_1,\ldots,n_\lambda,\ldots,n_\mu,\ldots\big)$ into the state $z' = \big(n_1,\ldots,n_\lambda-1,\ldots,n_\mu-1,\ldots,n_{\lambda'}+1,\ldots,n_{\mu'}+1,\ldots\big)$ is proportional to $n_\lambda\,n_\mu$, i.e.,

$$w_{z'z} = a_{\lambda'\mu'\lambda\mu}\,n_\lambda\,n_\mu\ , \tag{5.76}$$

and the master equation becomes

$$\dot{\varrho}(z,t) = \sum_{\lambda\mu\lambda'\mu'} a_{\lambda'\mu'\lambda\mu} \left( \mathsf{E}_\lambda^{-1} \mathsf{E}_\mu^{-1} \mathsf{E}_{\lambda'} \mathsf{E}_{\mu'} - 1 \right) n_\lambda\, n_\mu\, \varrho(z,t) \;. \tag{5.77}$$

This equation also contains nonlinear transition rates. Of course, it has to be completed by terms taking into account the deterministic movement of the particles.

In scattering processes the corresponding stochastic field should be denoted by $N(\boldsymbol{r}, \boldsymbol{p}; t)$, the local, time-dependent particle density in phase space, and an equation for

$$f(\boldsymbol{r}, \boldsymbol{p}; t) = \langle N(\boldsymbol{r}, \boldsymbol{p}; t) \rangle \tag{5.78}$$

leads under a certain approximation to the well-known Boltzmann equation (see van Kampen 1985; Gardiner 1985).

## 5.4 Analytic Solutions of Master Equations

In simple cases one can find analytic solutions to the master equation or to the equations derived from it. We now will present two methods which both transform the master equation into an equation which is easier to solve.

### *5.4.1 Equations for the Moments*

In many cases it is not the complete time evolution of $\varrho_n(t)$ which is of interest, but only the time evolution of the average particle number $\langle N(t)\rangle$ and the variance $\langle N^2(t)\rangle - \langle N(t)\rangle^2$. For one-step processes the equation for $\langle N(t)\rangle$ is easily derived from (5.55): For $\langle N(t)\rangle = \sum_n n\, \varrho_n(t)$ one obtains

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle N(t)\rangle = \sum_n n\, \dot{\varrho}_n(t) \tag{5.79}$$

$$\begin{aligned} &= \sum_n n \Big[ r(n+1)\, \varrho_{n+1}(t) - r(n)\, \varrho_n(t) \\ &\quad + g(n-1)\, \varrho_{n-1}(t) - g(n)\, \varrho_n(t) \Big] \;. \end{aligned} \tag{5.80}$$

Since

$$\sum_n n\, r(n+1)\, \varrho_{n+1}(t) = \sum_n (n-1)\, r(n)\, \varrho_n(t) \tag{5.81}$$

and

$$\sum_n n\, g(n-1)\, \varrho_{n-1}(t) = \sum_n (n+1)\, g(n)\, \varrho_n(t) \;, \tag{5.82}$$

we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle N(t)\rangle = -\langle r(N)\rangle + \langle g(N)\rangle \ . \tag{5.83}$$

Similarly, one finds, in general,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle N^k(t)\rangle = \Big\langle \big((N-1)^k - N^k\big)\, r(N)\Big\rangle + \Big\langle \big((N+1)^k - N^k\big)\, g(N)\Big\rangle \ . \tag{5.84}$$

The initial condition for $\langle N^k(t)\rangle$ follows from that for $\{\varrho_n(t), t = 0\}$:

$$\langle N^k\rangle\big|_{t=0} = \sum_n n^k\, \varrho_n(0) \ . \tag{5.85}$$

If $r(n)$ and $g(n)$ are linear in $n$, the right-hand sides of (5.83) and (5.84) are again moments of first and $k$-th order, respectively. Hence, one obtains a closed system of ordinary differential equations for the moments $\langle N^k(t)\rangle$.

On the other hand, if $r(n)$ or $g(n)$ are nonlinear, the right-hand side of (5.84) will contain moments of higher order than $k$. The result is an infinite hierarchy of equations and, in order to obtain a solvable, closed system of equations, one generally has to truncate this hierarchy by making some approximation.

### 5.4.2 The Equation for the Characteristic Function

For the characteristic function

$$G(z,t) = \sum_n z^n\, \varrho_n(t) \tag{5.86}$$

it holds that

$$\left(z\frac{\partial}{\partial z}\right)^k G(z,t) = \sum_n z^n\, n^k\, \varrho_n(t) \ , \tag{5.87}$$

and, therefore, for some polynomial, e.g. for $r(n)$, one also has

$$r\left(z\frac{\partial}{\partial z}\right) G(z,t) = \sum_n z^n\, r(n)\, \varrho_n(t) \ . \tag{5.88}$$

Furthermore

$$\sum_n z^n \Big(r(n+1)\, \varrho_{n+1}(t) - r(n)\, \varrho_n(t)\Big) = \left(\frac{1}{z} - 1\right) r\left(z\frac{\partial}{\partial z}\right) G(z,t) \ .$$

So one finds as the equation for $G(z,t)$

$$\frac{\partial}{\partial t}\,G(z,t) = \left[\left(\frac{1}{z}-1\right)\,r\left(z\,\frac{\partial}{\partial z}\right) + (z-1)\,g\left(z\,\frac{\partial}{\partial z}\right)\right]G(z,t)\;, \tag{5.89}$$

with

$$G(z,t=0) = \sum_n z^n\,\varrho_n(0)\;. \tag{5.90}$$

Once $G(z,t)$ has been determined, the moments follow from (5.87):

$$\left(z\,\frac{\partial}{\partial z}\right)^k\;G(z,t)\big|_{z=1} = \langle N^k(t)\rangle\;. \tag{5.91}$$

### 5.4.3 *Examples*

Let us apply these two methods to the examples of the previous section.

*The random walk.* $r(n) = g(n) = 1$. Let $\varrho_n(t=0) = \delta_{n,0}$, i.e., at $t=0$ the random walk starts at the position (state) $n=0$.

For the moments it follows from (5.84) that

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\langle N(t)\rangle = 0 \tag{5.92a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\langle N^2(t)\rangle = \langle -2N(t)+1\rangle + \langle 2N(t)+1\rangle = 2 \tag{5.92b}$$

and

$$\langle N^k(t)\rangle\big|_{t=0} = 0\;,\quad k=1,2,\ldots\;. \tag{5.93}$$

Therefore

$$\langle N(t)\rangle = 0\;, \tag{5.94a}$$

$$\langle N^2(t)\rangle = 2t\;. \tag{5.94b}$$

The equation for $G(z,t)$ is

$$\frac{\partial}{\partial t}\,G(z,t) = \left(z+\frac{1}{z}-2\right)\,G(z,t)\;, \tag{5.95}$$

where

$$G(z,0) = 1\;. \tag{5.96}$$

One obtains as the solution

$$G(z,t) = \exp\left(z + \frac{1}{z} - 2\right) t = \mathrm{e}^{-2t} \sum_{n=-\infty}^{+\infty} I_n(2t)\, z^n \ , \tag{5.97}$$

since $\exp\left(\frac{x}{2}(z + \frac{1}{z})\right)$ is the generating function of the modified Bessel function $I_n(x)$. From this solution one finds the moments

$$\langle N(t) \rangle = \left(z \frac{\partial}{\partial z}\right) \exp\left(z + \frac{1}{z} - 2\right) t \bigg|_{z=1} \tag{5.98}$$

$$= zt \left(1 - \frac{1}{z^2}\right) \exp\left(z + \frac{1}{z} - 2\right) t \bigg|_{z=1} = 0 \tag{5.99}$$

and after some calculation

$$\langle N^2(t) \rangle = \left(z \frac{\partial}{\partial z}\right)^2 \exp\left(z + \frac{1}{z} - 2\right) t \bigg|_{z=1} = 2t \ . \tag{5.100}$$

Both methods obviously yield the same results for the moments: The expectation value remains constant at $n = 0$, and the variance grows linearly in time. From (5.97) one obtains for $\varrho_n(t)$

$$\varrho_n(t) = \mathrm{e}^{-2t} I_n(2t) \ . \tag{5.101}$$

*Radioactive decay.* $g \equiv 0$, $r(n) = \gamma n$. Let $\varrho_n(t = 0) = \delta_{n,n_0}$. For $\langle N(t) \rangle$ one gets immediately

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle N(t) \rangle = -\gamma \langle N(t) \rangle \ , \qquad \langle N(t) \rangle|_{t=0} = n_0 \ , \tag{5.102}$$

i.e.,

$$\langle N(t) \rangle = n_0\, \mathrm{e}^{-\gamma t} \ . \tag{5.103}$$

For the variance one finds, after some calculation,

$$\mathrm{Var}\,(N(t)) = n_0 \left(\mathrm{e}^{-\gamma t} - \mathrm{e}^{-2\gamma t}\right) \ . \tag{5.104}$$

$G(z,t)$ now satisfies the equation

$$\frac{\partial}{\partial t} G(z,t) = \left(\frac{1}{z} - 1\right) \gamma z \frac{\partial}{\partial z} G(z,t) \tag{5.105}$$

$$G(z, t = 0) = z^{n_0} \ , \tag{5.106}$$

and the solution is

$$G(z,t) = \left(1 + (z-1)\mathrm{e}^{-\gamma t}\right)^{n_0} \tag{5.107}$$

$$= \left(1 - \mathrm{e}^{-\gamma t}\right)^{n_0} \left(1 + \frac{z\mathrm{e}^{-\gamma t}}{1 - \mathrm{e}^{-\gamma t}}\right)^{n_0} \tag{5.108}$$

$$= \sum_{n=0}^{n_0} \binom{n_0}{n} \mathrm{e}^{-n\gamma t} \, \left(1 - \mathrm{e}^{-\gamma t}\right)^{n_0 - n} \, z^n \ , \tag{5.109}$$

from which one finds

$$\varrho_n(t) = \binom{n_0}{n} \mathrm{e}^{-n\gamma t} \, \left(1 - \mathrm{e}^{-\gamma t}\right)^{n_0 - n} \ . \tag{5.110}$$

From this result one can also derive the expressions (5.103) and (5.104) for $\langle N(t)\rangle$ and Var $(N(t))$, respectively.

*The Poisson process.* $r \equiv 0, \ g(n) = \lambda$. Let $\varrho_n(t = 0) = \delta_{n,m}, \ m \geq 0$. One obtains

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle N(t)\rangle = \lambda \ , \qquad \langle N(t)\rangle|_{t=0} = m \ , \tag{5.111}$$

i.e.,

$$\langle N(t)\rangle = m + \lambda t \ . \tag{5.112}$$

Furthermore

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle N^2(t)\rangle = \lambda\langle 2N + 1\rangle \, , \qquad \langle N^2(t)\rangle\big|_{t=0} = m^2 \ , \tag{5.113}$$

hence,

$$\langle N^2(t)\rangle = (m + \lambda t)^2 + \lambda t \ , \tag{5.114}$$

and

$$\mathrm{Var}\,(N(t)) = \lambda t \ . \tag{5.115}$$

The equation for $G(z,t)$ reads

$$\frac{\partial}{\partial t} G(z,t) = (z-1)\,\lambda\, G(z,t) \ , \tag{5.116a}$$

$$G(z, t = 0) = z^m \ , \tag{5.116b}$$

therefore

$$G(z,t) = z^m \, \mathrm{e}^{\lambda t (z-1)} \tag{5.117}$$

$$= \mathrm{e}^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} z^{n+m} = \mathrm{e}^{-\lambda t} \sum_{n=m}^{\infty} \frac{(\lambda t)^{n-m}}{(n-m)!} z^n \ , \tag{5.118}$$

i.e.,

$$\varrho_n(t) = \frac{(\lambda t)^{n-m}}{(n-m)!}\,\mathrm{e}^{-\lambda t}\,, \quad n \geq m\,. \tag{5.119}$$

## 5.5 Simulation of Stochastic Processes and Fields

In Sect. 5.2 we wrote down the master equation for a piecewise deterministic Markov process $\{\boldsymbol{Z}(t)\}$, $\boldsymbol{Z} \in \mathbb{R}^n$ in the following form:

$$\frac{\partial}{\partial t}\,\varrho(z,t) = -\frac{\partial}{\partial z}\cdot \boldsymbol{f}\,\varrho(z,t) + \int \mathrm{d}^n z' \Big( w_{zz'}\varrho(z',t) - w_{z'z}\varrho(z,t)\Big)\,. \tag{5.120}$$

Such a stochastic process consists of a deterministic motion interrupted at random times by random jumps $z \to z'$.

The deterministic motion is described in this case by the differential equation

$$\dot{z} = \boldsymbol{f}(z)\,. \tag{5.121}$$

The probability for the interruption of this motion by jumps is determined by the transition rates $w_{z'z}$. We have seen that this concept of a piecewise deterministic Markov process describes – beyond Hamiltonian dynamics – the temporal development of open systems.

The master equation for this process is a partial integrodifferential equation for the density $\rho(z,t)$, whose solution would require enormous effort.

In many cases, however, one is not really interested in all the information contained in the time-dependent density. Often it is only the time dependence of expectation values of the random variables $\boldsymbol{Z}(t)$ which one wants to know. This suggests that one might simulate a sufficiently large sample of such a process and then estimate the quantities of interest on the basis of this sample. Thus one uses a computer to generate sufficiently many trajectories of such a process. The statistics of these trajectories will reflect the stochastic properties of $\boldsymbol{Z}(t)$ formulated by the master equation.

For a stochastic process with a discrete sequence of instants $t_i$ and discrete states $z = 1,\ldots,n$ such a simulation is most easy formulated. Let $t = 1,2,\ldots$, then the time development of the density can be written as

$$\boldsymbol{\varrho}(t+1) = \mathsf{M}\boldsymbol{\varrho}(t), \tag{5.122}$$

with $\mathsf{M} = \mathrm{e}^{\mathsf{V}}$ (see (6.199)). The elements of the matrix $\mathsf{M}$ then are conditional probabilities

$$\mathsf{M}_{ij} = \rho(i \mid j)\,, \quad i,j = 1,\ldots,n. \tag{5.123}$$

Now, given a state $j$ at time $t$, the state $i$ at time $t+1$ has to be chosen according to the probabilities $\{\rho(i \mid j), i = 1, \ldots, n\}$. This can be done in the following way: By adding up intervals of length $\rho(i \mid j)$ for $i = 1, \ldots, n$, starting at zero, one gets a subdivision of the unit interval $[0, 1]$. The limits of the $n$ intervals are given by $[0, y_1, \ldots, y_{n-1}, 1]$ with

$$y_i = \sum_{k=1}^{i} \rho(k \mid j). \tag{5.124}$$

Now if one realizes a random number $\xi$, equally distributed in $[0, 1]$, then the probability that it falls into the subinterval $[y_{i-1}, y_i]$ is just $\rho(i \mid j), i = 1, \ldots, n$, $y_0 = 0$. Hence the state realized at time $t+1$ is given by that value $i$ for which

$$y_{i-1} < \xi < y_i \,. \tag{5.125}$$

Very often, one of the conditional probabilities, e.g., the probability $\rho(j \mid j)$ for staying in the given state, is much larger than the others. To avoid the use (or calculation) of the division $[0, y_1, \ldots, y_{n-1}, 1]$ in each step, one may ask first whether there is a jump to a different state at all, and only when this is answered in the affirmative, then a decision is made into which state the jump shall occur. Because in most cases, no jump has to occur and since the first query is very fast, one may improve the speed of the code considerably.

Thus, in general, one has to set up an algorithm for the following three elements of the motion:

1. The deterministic motion for the case that no jump occurs;
2. The decision as to the instant at which a jump will occur, and,
3. When a jump does occur at time $t$ starting from $z(t)$, into which of the possible states $z'$ with $w_{z'z} \neq 0$ it will go.

In the following, we will describe the algorithms for these three elements. We suppose that we are dealing with a discrete set of states $z'$ which are accessible by a jump from $z$. We first present an algorithm for a fixed time step.

1. The deterministic motion according to the differential equation $\dot{z} = \boldsymbol{f}(z)$ has to be determined by standard numerical methods (Runge–Kutta etc.). We assume that the reader is familiar with these methods and do not go into further details. In particular, if $\boldsymbol{f} \equiv 0$, as in 'real' Markov processes, the deterministic evolution is extremely simple: $z(t) = \text{const}$.
2. Fundamental for the decision concerning the instant at which a jump $z(t) \to z'$ shall interrupt the deterministic evolution is the probability that for a given $z(t)$ at time $t$ exactly one jump occurs within the next time step up to $t + \mathrm{d}t$. This probability is

$$u_0(z, t)\,\mathrm{d}t = \sum_{z' \neq z} w_{z'z(t)}\,\mathrm{d}t \,. \tag{5.126}$$

Of course, $\mathrm{d}t$ has to be chosen small enough that $u_0(z, t)\,\mathrm{d}t$ is also sufficiently small compared to 1 (roughly of the order 0.01 to 0.1).

Hence, a jump should be realized with the probability $u_0(z,t)\,\mathrm{d}t$. For this purpose one generates a random number $\xi_1$ as a realization of a random variable uniformly distributed in the interval [0, 1]. With probability $u_0(z,t)\,\mathrm{d}t$ this random number is in the interval $[0,\ u_0(z,t)\,\mathrm{d}t]$. If, therefore, the generated random number $\xi_1$ lies within this interval, one decides in favor of a jump taking place. Otherwise the motion continues according to the deterministic equation.

3. If a jump does occur, one has to choose the state $z'$ in which to jump. Obviously,

$$y(z' \mid z) = \frac{w_{z'z}}{u_0(z)}\ , \quad z' \neq z \tag{5.127}$$

is the conditional probability that the jump will be into the given state $z' \neq z$, provided a jump should occur at all. We set $y(z \mid z) = 0$. From (5.126) we readily obtain

$$\sum_{z'} y(z' \mid z) = 1\ . \tag{5.128}$$

We arrange the states $z'$ for which $y(z' \mid z) \neq 0$ in a row and denote these states by $z'_\gamma$, where $\gamma = 1, 2, \ldots$. If we now set

$$y_\alpha = \sum_{\gamma=1}^{\alpha} y(z'_\gamma \mid z)\ , \tag{5.129}$$

then $y(z'_\alpha \mid z) = y_\alpha - y_{\alpha-1}$ for $\alpha > 1$ and $y(z'_1 \mid z) = y_1$. Hence, $y(z'_\alpha \mid z)$ is equal to the size of the intervals for a partition of [0, 1] according to

$$[0, y_1, \ldots, y_\alpha, \ldots, 1]\ . \tag{5.130}$$

Thus the target state for the jump, $z'_\alpha$, may be determined by generating a second random number $\xi_2$ as a realization of a random variable uniformly distributed on [0, 1], and taking the value $\alpha$ for which

$$y_{\alpha-1} \leq \xi_2 \leq y_\alpha\ . \tag{5.131}$$

This defines an algorithm to generate a trajectory of a stochastic process. The choice of the time step $\mathrm{d}t$ is, of course, essential both for the computer time used as well as for the accuracy.

Once $N$ trajectories, each starting from the same initial state at $t = 0$, have been generated, one has obtained $N$ realizations of the random variable $\boldsymbol{Z}(t)$ for each instant $t > 0$. From these samples for each instant $t$ one can estimate the required quantities such as expectation values or variances using the methods of Sect. 8.1.

There are many variants of this algorithm, and we refer to the literature, e.g., Binder (1979) and Honerkamp (1994), for more details. One particularly important variant, however, will be introduced here.

### Algorithm with a Stochastic Time Step

If the deterministic motion is sufficiently simple, one may determine the distribution of the intervals between two successive jumps. For this purpose we calculate the 'survival function' (also called the dwell time distribution) $V_t(\tau)$, which is the probability that no jump occurs in the interval $(t, t+\tau)$. It is given by

$$V_t(\tau) = \exp\left(-\int_t^{t+\tau} \mathrm{d}t'\, u_0(z(t'))\right) . \tag{5.132}$$

To see why, we first consider an interval of length $\tau$ which is small enough to regard $u(z(t))$ as constant on it. We subdivide the interval again into $n$ time steps $\mathrm{d}t$ such that $\mathrm{d}t = \tau/n$. The probability that no jump occurs within a time $\mathrm{d}t$ is

$$V_t(\mathrm{d}t) = \big(1 - u_0\, \mathrm{d}t\big) = \big(1 - u_0\, \frac{\tau}{n}\big) , \tag{5.133}$$

and the probability that no jump occurs within the $n$ time steps of the interval $\tau$ is

$$V_t(\tau) = \Big(V_t(\mathrm{d}t)\Big)^n = \Big(1 - u_0\, \frac{\tau}{n}\Big)^n . \tag{5.134}$$

Taking $\mathrm{d}t \to 0$ for fixed $\tau$, i.e., $n \to \infty$, one obtains

$$V_t(\tau) = \mathrm{e}^{-u_0 \tau} = \mathrm{e}^{-u_0(z(t))\,\tau} . \tag{5.135}$$

For a finite interval $(t, t+\tau)$, where now $u(z(t))$ is allowed to vary, we get

$$V_t(\tau) = \exp\left(-\int_t^{t+\tau} u_0(z(t'))\, \mathrm{d}t'\right) . \tag{5.136}$$

The interval between two successive jumps is also called the waiting time. From $V_t(\tau)$ now we can easily determine the cumulative distribution function $F_t(\tau)$ of the waiting time. It is simply

$$F_t(\tau) = 1 - V_t(\tau) , \tag{5.137}$$

because the survival function $V_t(\tau)$ can also be interpreted as the probability that the waiting time is greater than $\tau$. On the other hand the cumulative distribution $F_t(\tau)$ is just the probability that the waiting time is less than $\tau$.

Obviously $0 \leq F_t(\tau) \leq 1$, and the function $F_t$ maps the random variable 'waiting time' from the interval $(0, \infty)$ onto the interval $[0, 1]$. Thus one obtains a realization of this waiting time distribution by generating a random number $\xi$, uniformly distributed on the interval $[0, 1]$, and choosing $\tau$ such that $F(\tau) = \xi$. Thereby one obtains an equation for $\tau$:

$$V_t(\tau) = 1 - \xi \qquad \text{or} \qquad \ln V_t(\tau) = \ln(1 - \xi) , \tag{5.138}$$

or

$$\int_{t}^{t+\tau} \mathrm{d}t' \, u_0(z(t')) = -\ln(1-\xi) \; . \tag{5.139}$$

Hence, one first has to determine the deterministic movement $z(t')$ starting at $t' = t$, from this one obtains $u_0(z(t'))$, and finally one integrates over $t'$ according to (5.139). Solving this equation for $\tau$ at given $\xi$ yields a realization $\tau$ of the random time until the next jump. Even if the solution for $\tau$ has to be done numerically, this method is often faster than using a fixed time step $\mathrm{d}t$. (In this way one also circumvents the discussion about the correct choice for the time step $\mathrm{d}t$.)

Thus after each jump one has to determine a new random number $\tau$ according to the above described method and to integrate the deterministic equation $\dot{z} = \boldsymbol{f}(z)$ for this time $\tau$.

If, in particular, $z(t') =$ const. between the jumps, as is the case for 'real' Markov processes ($\boldsymbol{f} \equiv 0$), then $u_0(z)$ is also constant and therefore

$$\int_{t}^{t+\tau} \mathrm{d}t' \, u_0(z) = \tau \, u_0(z) \; , \tag{5.140}$$

where $\tau$ has to be determined from

$$\tau = -\frac{1}{u_0(z)} \ln(1-\xi) \; . \tag{5.141}$$

### Simulation of Stationary Quantities

After a sufficiently long time, the density $\varrho(z,t)$ of this process will be practically identical to the stationary density $\varrho^{\text{stat}}(z)$. During a certain time period the individual states $\{z\}$ will then be visited with a frequency proportional to their stationary probability $\varrho^{\text{stat}}(z)$. (We assume here that the process may be considered as ergodic during this time period.) The states $z$, generated by the simulation of the trajectory in the stationary regime, are therefore realizations of a random variable $\boldsymbol{Z}$ with a probability distribution $\varrho^{\text{stat}}(z)$.

However, successive realizations obtained in this way are not independent, because it is the transition probability $\varrho_2(z,t \mid z',t')$ which determines the process and thereby the simulation. To get independent realizations of a random variable with density $\varrho^{\text{stat}}(z)$ one simply lists only every $n$th number in the sequence of random numbers obtained from the simulation of a stationary trajectory, where $n$ is sufficiently large such that the correlations among these numbers have faded away. This correlation time is easily estimated by checking for which values of $\tau$ the covariance $\mathrm{Cov}\big(z(t)z(t+\tau)\big)$ is compatible with zero.

For other methods of generating independent random numbers see, for example, Honerkamp (1994).

### The Monte Carlo Method, Simulation of Random Fields

The fact that in the simulation of a trajectory after a sufficiently long time random numbers with the density $\varrho^{\text{stat}}(z)$ are generated may also be utilized to formulate algorithms generating high-dimensional random vectors with a given density. What one has to do is to construct a Markov process with the given density as its stationary density. In fact, this is fairly easy: One chooses transition rates satisfying the condition of detailed balance (Sect. 5.2).

Such methods for generating realizations of high-dimensional random vectors also allow the estimation of high-dimensional integrals, as they occur, for example, in the statistical mechanics of equilibrium systems (see Chaps. 3 and 4). They are known as Monte Carlo methods in the statistical mechanics of equilibrium states.

Let us consider the Ising model as an illustration of this method:
A state is denoted by

$$\boldsymbol{x} = (x_1, \ldots, x_n)\,, \quad x_i = \pm 1\,, \tag{5.142}$$

where $\{x_i\}$ represent the spin orientations on, say, a two-dimensional lattice. The density function for the canonical system is given by

$$\varrho(\boldsymbol{x}) = \frac{1}{Z}\,\mathrm{e}^{-\beta H(\boldsymbol{x})}\,, \tag{5.143}$$

where $Z$ is the partition function, and the energy function has the form (cf. (4.137))

$$H(\boldsymbol{x}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i} J\, x_i\, x_j\,. \tag{5.144}$$

The neighborhood $\mathcal{N}_i$ of a lattice point $i$ consists of the four nearest neighbor sites.

We will now construct a stochastic process $\boldsymbol{X}(t)$ whose stationary density $\varrho^{\text{stat}}(\boldsymbol{x})$ coincides with $\varrho(\boldsymbol{x})$. A simulation of this process, once the stationary state has been reached, will yield realizations of the random field $\boldsymbol{X}$ with the density (5.143). We may investigate such realizations by image processing methods, but in the statistical mechanics of equilibrium states we may also estimate expectations values such as

$$E = \langle H(\boldsymbol{x})\rangle = \sum_{\boldsymbol{x}} H(\boldsymbol{x})\,\varrho(\boldsymbol{x}) \tag{5.145}$$

from a sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of size $N$ by taking

$$\hat{E} = \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{x}_i)\,. \tag{5.146}$$

The problems encountered in the estimation of expectation values are addressed in Sect. 8.1.

To define the stochastic process we have to state the transition rates. Detailed balance requires

$$w_{\boldsymbol{x}\boldsymbol{x}'}\,\varrho(\boldsymbol{x}') = w_{\boldsymbol{x}'\boldsymbol{x}}\,\varrho(\boldsymbol{x}) \ , \tag{5.147}$$

i.e.,

$$\frac{w_{\boldsymbol{x}'\boldsymbol{x}}}{w_{\boldsymbol{x}\boldsymbol{x}'}} = \frac{\varrho(\boldsymbol{x}')}{\varrho(\boldsymbol{x})} = \mathrm{e}^{-\beta \Delta H} \ , \tag{5.148}$$

where

$$\Delta H = H(\boldsymbol{x}') - H(\boldsymbol{x}) \ . \tag{5.149}$$

Two frequently employed forms for $w_{\boldsymbol{x}\boldsymbol{x}'}$ are

- The ansatz introduced by (Metropolis et al. 1953):

$$w_{\boldsymbol{x}'\boldsymbol{x}}\,\mathrm{d}t = \begin{cases} \mathrm{e}^{-\beta\,\Delta H} & \text{if } \Delta H \geq 0 \\ 1 & \text{otherwise.} \end{cases} \tag{5.150}$$

  Hence, if the jump to $\boldsymbol{x}'$ leads to a smaller energy ($\Delta H < 0$), $\boldsymbol{x}'$ is always accepted. If $\boldsymbol{x}'$ has a larger energy, the acceptance of this state for the instant $t + \mathrm{d}t$ becomes increasingly improbable, the larger the energy difference.
- The ansatz of the so-called heat bath method:

$$w_{\boldsymbol{x}'\boldsymbol{x}}\,\mathrm{d}t = \frac{\mathrm{e}^{-\beta\,\Delta H}}{1 + \mathrm{e}^{-\beta\,\Delta H}} \ . \tag{5.151}$$

One can show that in both cases the condition of detailed balance is fulfilled.

Hence, one has to simulate a Markov process $\{\boldsymbol{x}(t)\}$ for which the transition rates are given by (5.150) or by (5.151).

It can be shown (Binder 1979) that the two steps,

1. Check whether there should be a jump at all and,
2. If yes, decide into which state,

can be exchanged, and therefore the steps,

1. Choose a possible state $\boldsymbol{x}'$ and
2. Check whether to jump into this state,

are also possible. This method has the advantage that the computational effort for the choice and the check is much smaller.

In physics this method is called the Monte Carlo method. In stochastics, an algorithm generating a spatial Markov field in the above manner is called a Metropolis sampler.

Summarizing, the algorithm thus consists of the following steps:

1. Start with an initial configuration $\boldsymbol{x}$.
2. A new configuration $\boldsymbol{x}'$ is obtained by selecting a component $x_i$ of the field $\boldsymbol{x}$ and by the assignment of a new value to this component. For this purpose one

chooses a $x_i^* \in \mathcal{L}$ (the space of all possible values of $x_i$), where all values in $\mathcal{L}$ are taken to be equally probable. One then sets $x_i' = x_i^*$ with probability $w_{\boldsymbol{x}'\boldsymbol{x}}\mathrm{d}t$.

3. The choice of a component $i$ may be done randomly such that every component occurs with the same probability. A simpler way is to proceed through the lattice systematically and to assign new configurations $\boldsymbol{x}'$ successively to all components. Such a passage through the lattice, by which $N$ new configurations are generated, is called a sweep. If the components are chosen randomly a collection of $N$ successive new configurations might also be called a sweep.
4. Make $M$ sweeps, where $M$ is a suitably chosen number.

*Remarks.*

- Because $\varrho(\boldsymbol{x})$ is formulated as a Gibbs field, i.e., a Markov field, $w_{\boldsymbol{x}'\boldsymbol{x}}\mathrm{d}t$ is a simple expression depending only on those terms in $H(\boldsymbol{x})$ that contain $x_i$.
- In the simulation of such processes starting from an initial configuration one has, of course, to estimate from when on the system is in equilibrium. Only then can a configuration $\boldsymbol{x}$ be considered as a realization of the random field $\boldsymbol{X}$ with the density $\varrho(\boldsymbol{x})$.

There are many variants of this method. One method using only the information of the conditional probabilities is known in stochastics as the Gibbs sampler. In this case, for each point $i$ the probability

$$\varrho_l = \varrho(x_i = l|\mathcal{N}_i) \tag{5.152}$$

is determined for every possible realization $l$ of $x_i$. Next one sets $x_i = l$ with probability $p_l$. This is repeated $M$ times.

Concerning the application of the Metropolis method to the two-dimensional Ising model and further related problems in statistical mechanics see, for example, Honerkamp (1994), Binder (1979), Binder and Heermann (1997), and Binder (1995, 1987). As an example we present a simulation of the isotropic Ising model (Sect. 4.5) at $\beta J = 0.5$ (the critical point is at $\beta J = 0.4407$). For the magnetization one obtains from Sect. 4.5 for $N \to \infty$ the exact result $m = \pm 0.9113$ and $\beta E = -0.8728$. The estimated values as a function of the number of sweeps are plotted in Fig. 5.5. Evidently, the stationary state is reached after about 200 sweeps, after which point the data can be used for an estimation of the magnetization and the energy. An analysis of the covariance function reveals that the realizations can be considered as independent after about 25 sweeps. From a time series consisting of every 25th data point one obtains

$$\hat{m} = 0.9124 \pm 0.0163, \qquad \beta\hat{E} = -0.8735 \pm 0.0144\ , \tag{5.153}$$

in good agreement with the exact results for $N \to \infty$. Longer observations of the stationary states lead to a longer time series and to a smaller error.

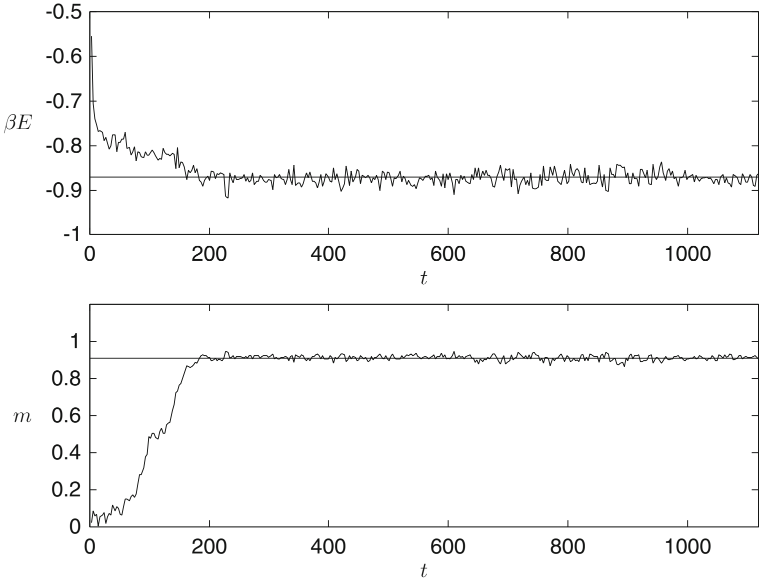


**Fig. 5.5** Estimation of the energy and the magnetization as a function of the number $t$ of sweeps on a 64 × 64 lattice. The lines represent the exact solution for $N \to \infty$

## 5.6 The Fokker–Planck Equation

In Sect. 5.2 we derived the master equation for Markov processes from a certain short-time behavior of the conditional probability $\varrho_2\big(z, t + \tau \mid z', t\big)$, formulated in a given expansion in $\tau$. A different requirement for the short-time behavior is the following:

$$\begin{aligned}
&\int \mathrm{d}z_2 (z_2 - z_1)_i \varrho(z_2, t + \tau \mid z_1, t) = A_i(z_1, t)\tau + o(\tau),\\
&\int \mathrm{d}z_2 (z_2 - z_1)_i (z_2 - z_1)_j \varrho(z_2, t + \tau \mid z_1, t) = D_{ij}(z_1, t)\tau + o(\tau),\\
&\int \mathrm{d}z_2 (z_2 - z_1)_{i_1} \dots (z_2 - z_1)_{i_n} \varrho(z_2, t + \tau \mid z_1, t) = o(\tau) \quad \text{for } n > 2\ .
\end{aligned} \tag{5.154}$$

Hence, the transition probability, i.e., the conditional probability when a state $z_1$ at time $t_1$ is given, shall in lowest order in $\tau$ be a normal distribution with means $A_i(z_1, t)$ and variances $D_{ij}(z_1, t)$ for the increments $(\boldsymbol{Z}_2 - \boldsymbol{Z}_1)_i$.

Under these conditions it is possible to show that the expansion of the transition probability for small $\tau$ has the form

$$\varrho_2\big(z, t+\tau \mid z', t\big) = \frac{1}{\sqrt{2\pi\tau}}\,\mathrm{e}^{-(z-z')^2/2\tau}\Big(a_0(z,z') + a_1(z,z')\tau + O(\tau^2)\Big)\ , \quad (5.155)$$

where the $a_i(z,z'), i = 0, 1, \ldots$ depend on the drift and diffusion terms. In this case, of course, we also have $\varrho_2\big(z, t+\tau \mid z', t\big) \to \delta(z - z')$ for $\tau \to 0$, but the limit is approximated in a completely different way compared to the one we required in the introduction of the master equation.

It turns out (see, e.g., van Kampen 1985; Honerkamp 1994) that such a short-time behavior yields an equation for the probability density of the form

$$\frac{\partial}{\partial t}\,\varrho(z,t) = \left[-\sum_{i=1}^{n}\frac{\partial}{\partial z_i}\,A_i(z,t) + \frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial z_i\,\partial z_j}\,D_{ij}(z,t)\right]\varrho(z,t)\ . \quad (5.156)$$

This is the Fokker–Planck equation.

Hence, like the master equation, this is a differential equation for the time-dependent density function of a Markov process. Yet, the time-dependent density function (and the conditional probability $\varrho_2(x, t \mid x', t')$) now describes a different type of stochastic processes. In the mathematical literature this is called a diffusion process. In contrast to those processes which are described by a master equation (and which in the physics literature are also referred to as diffusion processes), a characteristic feature of this process is that there are no jumps. These are in a way no longer resolved on the large length scale. This will be studied in more detail in Sect. 5.8. There we will also show that the Fokker–Planck equation arises as a certain limit of the master equation.

For the trajectories of a diffusion process one can formulate a stochastic differential equation. If the diffusion matrix $D_{ij}(z,t)$ does not depend on $z$, this can be done even without a lengthy discussion about additional concepts from stochastics. Indeed, one can show that the multivariate diffusion process $\boldsymbol{Z}(t) \in \mathbb{R}^n$, whose density function obeys the Fokker–Planck equation

$$\frac{\partial\varrho(z,t)}{\partial t} = \Big[-\sum_{i=1}^{n}\frac{\partial}{\partial z_i}\,A_i(z,t) + \frac{1}{2}\sum_{i,j=1}^{n}D_{ij}(t)\frac{\partial^2}{\partial z_i\,\partial z_j}\Big]\varrho(z,t)\ , \quad (5.157)$$

satisfies the differential equation

$$\dot{Z}_i(t) = A_i(\boldsymbol{Z},t) + \sum_{j=1}^{n} B_{ij}(t)\,\eta_j(t)\ , \quad i = 1, \ldots, n\ . \quad (5.158)$$

Here $\mathsf{B}$ is a matrix with the property $\mathsf{B}\mathsf{B}^T = \mathsf{D}$, where $\mathsf{D}$ is the matrix with elements $D_{ij}$. The random variables $\{\eta_j(t),\ j = 1, \ldots, n\}$ are independent 'white noise', i.e., stochastic processes which make the differential equation also a stochastic equation: For each realization of $\eta_j(t),\ j = 1, \ldots, n$, one obtains a realization of the process $Z_j(t)$. Such a stochastic differential equation is often called a Langevin equation.

For the increments

$$\mathrm{d}Z_i(t) = \int_t^{t+\mathrm{d}t} \mathrm{d}t'\, \dot{Z}_i(t') \tag{5.159}$$

one obtains from the differential equation up to order d$t$

$$\mathrm{d}Z_i(t) = A_i\big(\boldsymbol{Z}(t), t\big)\, \mathrm{d}t + \sum_{j=1}^{n} B_{ij}(t)\, \mathrm{d}W_j(t)\ , \tag{5.160}$$

where

$$\mathrm{d}W_j(t) = \int_t^{t+\mathrm{d}t} \mathrm{d}t'\, \eta_j(t') \tag{5.161}$$

is an integral over a stochastic process. It would lead us far beyond the scope of this book to enter into a discussion on general stochastic integrals (but see van Kampen 1985; Gardiner 1985; Honerkamp 1994). However, d$W_j(t)$ is an especially simple stochastic integral and with some intuition its properties are easily derived from the properties of $\{\eta_j(t)\}$. Obviously,

$$\langle \mathrm{d}W_j(t) \rangle = \int_t^{t+\mathrm{d}t} \mathrm{d}t'\, \langle \eta_j(t') \rangle = 0\ , \tag{5.162}$$

and

$$\langle \mathrm{d}W_i(t) \mathrm{d}W_j(t) \rangle = \int_t^{t+\mathrm{d}t} \mathrm{d}t_1 \int_t^{t+\mathrm{d}t} \mathrm{d}t_2\, \langle \eta_i(t_1) \eta_j(t_2) \rangle \tag{5.163}$$

$$= \int_t^{t+\mathrm{d}t} \mathrm{d}t_1\, \delta_{ij} = \delta_{ij}\, \mathrm{d}t\ . \tag{5.164}$$

Furthermore,

$$\langle \mathrm{d}W_i(t) \mathrm{d}W_j(t') \rangle = 0\ , \tag{5.165}$$

as long as the intervals $(t, t + \mathrm{d}t)$ and $(t', t' + \mathrm{d}t)$ are nonoverlapping, which is the case for successive time steps. Hence, for each $j$ the stochastic integral d$W_j(t)$ may be represented a random variable with zero mean and variance d$t$ which furthermore is gaussian, because according to (5.161) it is a sum of gaussian random variables. Hence we can write

$$\mathrm{d}W_j(t) = \eta_j\, \sqrt{\mathrm{d}t}\ , \tag{5.166}$$

where $\{\eta_j\}$ are independent standard normal random variables.

*Remarks.*

- Because the increments $dZ_i = (Z_2 - Z_1)_i$ as random variables are Gaussian in the lowest order in $\tau$ with means $A_i(Z_1, t)$ and variances $D_{ij}(Z_1, t)$, we can represent these as

  $$dZ_i = A_i(Z_1, t)\tau + B_{ij}\sqrt{\tau}\eta_j, \qquad \mathsf{B}\mathsf{B}^T = \mathsf{D}. \tag{5.167}$$

  Thus one could conclude directly from conditions (5.154) that such a type of stochastic differential equation might exist.
- If the diffusion matrix $\mathsf{D}$ also depends on the state $z$, the Langevin equation can be written in the form

  $$\dot{Z}(t) = a(Z, t) + b(Z, t)\eta(t) \,. \tag{5.168}$$

  In this case one speaks of a multiplicative noise, because the white noise is multiplied by a function of the stochastic variable. At this stage a problem arises, whose detailed examination would lead us far into the field of stochastic differential equations and stochastic integrals. If we were now to study the increments

  $$\mathrm{d}Z(t) = \int_t^{t+\mathrm{d}t} \mathrm{d}t' \, \dot{Z}(t') \,, \tag{5.169}$$

  we would encounter an integral over the product of two stochastic processes:

  $$I = \int_t^{t+\mathrm{d}t} \mathrm{d}t' \; b(Z(t'), t') \, \eta(t') \,. \tag{5.170}$$

  In the attempt to give a precise mathematical meaning to such expressions one finds several possibilities. Two of the possible definitions for such stochastic integrals are known as the Itô type stochastic integral and the Stratonovich type stochastic integral. When one formulates stochastic differential equations with multiplicative noise one always has to specify in which sense the product of two processes is to be understood.

### *5.6.1 Fokker–Planck Equation with Linear Drift Term and Additive Noise*

If the drift term $A_i(z, t)$ is linear in $z$, i.e.,

$$A_i(z, t) = -\sum_{j=1}^{n} A_{ij}(t) \, z_j \ , \tag{5.171}$$

all quantities of interest are easily calculated:

- The expectation value $\langle \boldsymbol{Z}(t)\rangle$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle Z_i(t)\rangle = -\sum_{j=1}^{n} A_{ij}(t)\,\langle Z_j(t)\rangle\ . \tag{5.172}$$

- For the covariance matrix $\Sigma(t)$ with elements

$$\Sigma_{ij}(t) = \langle Z_i(t)Z_j(t)\rangle - \langle Z_i(t)\rangle\langle Z_j(t)\rangle$$

  the following equation holds:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) = \mathsf{D} - \left(\mathsf{A}\Sigma + \Sigma\mathsf{A}^T\right)\ , \tag{5.173}$$

  where $\mathsf{A}$ and $\mathsf{D}$ are the matrices with elements $\mathsf{A}_{ij}$ and $\mathsf{D}_{ij}(t)$, respectively.
- The density $\varrho(z,t)$ is for all times the density of a Gaussian random vector and therefore $\varrho(z,t)$ is completely characterized by the expectation value $\langle \boldsymbol{Z}(t)\rangle$ and the covariance matrix $\Sigma(t)$. In such cases one also speaks of a Gaussian stochastic process.
- For the two-time covariance function $\mathsf{C}_{ij}(t_1,t_2) = \mathrm{Cov}(X_i(t_1)X_j(t_2))$ one obtains for $t_1 < t_2$, if the matrix $\mathsf{A}$ is time independent,

$$\mathsf{C}_{ij}(t_1,t_2) = \left[\mathrm{e}^{-\mathsf{A}(t_2-t_1)}\Sigma(t_1)\right]_{ij}\ , \tag{5.174}$$

  where $\Sigma(t_1)$ represents the variance at time $t_1$. In the stationary state, $\Sigma(t_1)$ has to be replaced by the stationary variance $\Sigma(\infty)$. Then $\mathsf{C}_{ij}(t_1,t_2)$ is only a function of $\tau = t_2 - t_1$. The spectral density $\tilde{\mathsf{C}}(\omega)$ in the stationary state, with matrix elements

$$\tilde{\mathsf{C}}_{ij}(\omega) = \int_{-\infty}^{\infty} \mathrm{d}\tau\,\mathsf{C}_{ij}(\tau)\mathrm{e}^{\mathrm{i}\omega\tau}\ , \tag{5.175}$$

  follows after some calculation as

$$\tilde{\mathsf{C}}(\omega) = (\mathsf{A} + \mathrm{i}\omega)^{-1}\mathsf{D}(\mathsf{A}^T - \mathrm{i}\omega)^{-1}\ . \tag{5.176}$$

  All higher common densities $\varrho(\boldsymbol{x}_1, t_1, \ldots)$ are determined solely by the expectation value and the two-time covariance function $\mathsf{C}_{ij}(\tau)$, in the same way as a single Gaussian random variable is determined by the expectation value and the variance alone.
- For the univariate case we set $A(z,t) = -mz, m \neq 0$, $D = \sigma^2$. The corresponding Langevin equation then reads

$$\dot{Z}(t) = -m\,Z(t) + \sigma\,\eta(t)\ , \tag{5.177}$$

and $\{Z(t)\}$ is called the Ornstein–Uhlenbeck process. In particular, for sufficiently large $t$ one obtains for the covariance of $Z(t)$ and $Z(t+\tau)$

$$\operatorname{Cov}\Big(Z(t)\, Z(t+\tau)\Big) = \frac{\sigma^2}{2m}\, \mathrm{e}^{-m\tau} \;, \tag{5.178}$$

i.e., $Z(t)$, defined by (5.177), is also a model for a red noise. In addition,

$$\operatorname{Var}\Big(Z(t)\Big) = \frac{\sigma^2}{2m} \;. \tag{5.179}$$

For the case $m = 0$ there is no longer a stationary state. The dissipation, represented by the term $-mZ(t)$, is now missing. Starting from $Z(0) = 0$ one gets

$$\operatorname{Var}(Z(t)) = \sigma^2 t \tag{5.180}$$

and

$$\begin{aligned} C(t_1, t_2) &= \operatorname{Cov}(Z(t_1)Z(t_2)) \\ &= \sigma^2 \min(t_1, t_2) = \frac{\sigma^2}{2}\,(|t_1| + |t_2| - |t_2 - t_1|) \;. \end{aligned} \tag{5.181}$$

This process is also called Brownian motion; it is the continuous time version of the random walk. The trajectories of a pollutant particle in air may be thought of as realizations of Brownian motion. If many such realizations are considered simultaneously, e.g. a cloud of such particles which do not disturb each other, the diameter is proportional to the standard deviation and therefore grows as the square root of time. Such motion was first studied by the Scottish botanist R. Brown. In 1827 he discovered that the microscopically small particles into which the pollen of plants decay in an aqueous solution are in permanent irregular motion.

The Langevin equation for Brownian motion reads

$$\dot{Z}(t) = \sigma\eta(t) \;, \quad \eta(t) \sim \mathrm{WN}(0,1) \;. \tag{5.182}$$

Hence, the infinitesimal increment $\mathrm{d}Z(t)$ is a Gaussian random variable and, therefore, so is the finite increment $Z(t_2) - Z(t_1)$, because being the sum of normal random variables, it is again normally distributed. Furthermore, the increments at different times are mutually independent. Hence, Brownian motion is a stochastic process with independent and normally distributed increments. We will meet various generalizations of this motion in Sect. 5.9.

*Remark.* The numerical integration of stochastic differential equations represents a wide field. Of special importance in this context is the fact that the treatment of the diffusion term determines the order of the numerical method, and that methods of higher order can only be formulated with a major effort. Here we briefly describe the simplest method for univariate processes with additive noise.

Consider the Langevin equation

$$\dot{Z}(t) = a(Z,t) + \sigma(t)\eta(t) \ . \tag{5.183}$$

Following what is known as the Euler method, we write for $z(t+h)$, given $z(t)$, the approximation

$$\bar{z}(t+h) = z(t) + a(z(t),t)h + \sigma(t)\sqrt{h}\varepsilon(t) \ . \tag{5.184}$$

Here, $\varepsilon(t)$ is for any $t$ a realization of a random variable with expectation value 0 and variance 1. In particular, it may be a random variable with a corresponding uniform distribution. Notice that the noise term is of the order $h^{1/2}$.

For the expectation value of a function $M(z(t))$ of $z(t)$ one obtains

$$\langle M(\bar{z}(t+h))\rangle - \langle M(z(t+h))\rangle = O(h^2) \ . \tag{5.185}$$

The single-step errors for the Euler method are therefore quadratic in the step size $h$. A more detailed treatment of the numerical integration of stochastic differential equations can be found in Honerkamp (1994) and Kloeden and Platen (1995).

## 5.7 The Linear Response Function and the Fluctuation–Dissipation Theorem

The properties of a stationary system are usually examined by exposing the system to a small external force and then measuring the reaction of the system to this disturbance. In particular, the changes of the expectation values $\langle X_i \rangle$ with time in such a situation are characteristic for the system.

Let $\{F_j(t)\}$ be the external forces. Then this characteristic property of the system is described by the repones function $R_{ij}(t-t')$, defined by the expansion of the expectation value in powers of $F$:

$$\langle X_i(t)\rangle_F = \langle X_i(t)\rangle_{F=0} + \int_0^t \mathrm{d}t' \sum_j R_{ij}(t-t')F_j(t') + O(F^2) \ . \tag{5.186}$$

Hence, the response function reflects the reaction of the system to an external disturbance in a linear approximation.

Of course, causality leads us to expect that $R_{ij}(\tau) = 0$ for $\tau < 0$. The Fourier transform

$$\chi_{ij}(\omega) = \int_0^\infty \mathrm{d}\tau \ R_{ij}(\tau)\mathrm{e}^{\mathrm{i}\omega\tau} \tag{5.187}$$

is a complex-valued matrix, whose elements are called susceptibilities. They measure the sensitivity of the expectation value of $X_i(t)$ to a change of the external force $F_j$.

For a diffusion process described by a Fokker–Planck equation one obtains (Gardiner 1985)

$$R_{ij}(t-t') = \Theta(t-t') \int \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{x}'\, x_i \varrho_2(\boldsymbol{x}, t \mid \boldsymbol{x}', t') \left( -\frac{\partial}{\partial x'_j} \right) \varrho_{\text{stat}}(\boldsymbol{x}') \,.$$

Furthermore, for a system where detailed balance holds one can derive a connection between the linear response function and the covariance functions $C_{ij}(t-t') = \mathrm{Cov}(X_i(t), X_j(t'))$ (Reichl 1980). Since the response function is determined by the dissipation in the system and the covariance function describes the characteristic fluctuations of the random variables, this connection is also called the fluctuation–dissipation theorem.

We will not derive the relation between response and covariance function in full generality but instead discuss it for the case where the drift term in the Fokker–Planck equation is linear in $\boldsymbol{x}$, i.e., has the form $A_i(\boldsymbol{x}, t) = -\sum_j \mathsf{A}_{ij} x_j$ (cf. (5.171)). For this case one readily obtains from (5.186) the linear response function:

$$\mathsf{R}(t-t') = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}\omega \frac{1}{\mathrm{i}\omega + \mathsf{A}} \mathrm{e}^{\mathrm{i}\omega(t-t')} \,, \tag{5.188}$$

i.e.,

$$\tilde{\mathsf{R}}(\omega) = \frac{1}{\mathrm{i}\omega + \mathsf{A}} \,. \tag{5.189}$$

Furthermore, the Fourier transform of the two-time covariance function $\tilde{\mathsf{C}}(\omega)$ in this case reads (cf. (5.176))

$$\tilde{\mathsf{C}}(\omega) = (\mathrm{i}\omega + \mathsf{A})^{-1} \mathsf{D} (-\mathrm{i}\omega + \mathsf{A}^T)^{-1} \,. \tag{5.190}$$

It is obvious that response function and covariance function are related to one another. In particular we find

$$\tilde{\mathsf{C}}(\omega) = \tilde{\mathsf{R}}(\omega) \mathsf{D} \tilde{\mathsf{R}}^T(-\omega) \,. \tag{5.191}$$

We will now discuss this dynamical version of the fluctuation–dissipation relation together with more simple versions in the context of an example.

We consider a Brownian particle of mass $m$ in an environment of many smaller particles, which act on the Brownian particle by stochastic forces. In addition, we assume that the Brownian particle is bound to a wall by a harmonic force, so that the particle can vibrate around an equilibrium position and simultaneously feels a frictional force as well as the above-mentioned stochastic forces. This stochastically driven, damped harmonic oscillator is the prototype of a statistical system with dissipation. When we consider the deviation $x(t)$ from the equilibrium position in only one dimension, the Langevin equation for this reads

$$m\ddot{x}(t) + \gamma\dot{x}(t) + m\omega_0^2 x(t) = \sigma\eta(t)\ . \tag{5.192}$$

Here $\eta(t)$ represents a white noise, $\sigma$ is the strength of the stochastic force, $\gamma$ is a measure for the strength of the damping, and $\omega_0$ denotes the natural frequency of the oscillator.

We introduce the momentum $p(t) = m\dot{x}(t)$ and find the equations of motion

$$\dot{x}(t) = \frac{1}{m} p(t) \tag{5.193a}$$

$$\dot{p}(t) = -\frac{\gamma}{m} p(t) - m\omega_0^2 x(t) + \sigma\eta(t)\ . \tag{5.193b}$$

Thus we have a Fokker–Planck equation for $\boldsymbol{x} = (x(t), p(t))$ with a linear drift term of the form (5.171), where

$$\mathsf{A} = \begin{pmatrix} 0 & -1/m \\ m\omega_0^2 & \gamma/m \end{pmatrix}\ , \quad \mathsf{D} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 \end{pmatrix}\ . \tag{5.194}$$

For the response function we obtain

$$R_{12}(t-t') = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega\ \frac{1}{m(\omega_0^2 - \omega^2 + \mathrm{i}\gamma\omega/m)}\ \mathrm{e}^{\mathrm{i}\omega(t-t')} \tag{5.195}$$

and, for $C_{ij}(t-t')$, by using the representations (5.175) and (5.176),

$$\begin{aligned} C_{11}(t-t') &= \mathrm{Cov}(X(t), X(t')) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega\ \frac{\sigma^2}{m^2|\omega_0^2 - \omega^2 + \mathrm{i}\gamma\omega/m|^2}\ \mathrm{e}^{\mathrm{i}\omega(t-t')} \\ &= \frac{\sigma^2}{2m\gamma}\frac{1}{\mathrm{i}\pi} P\int_{-\infty}^{\infty} \mathrm{d}\omega\ \frac{1}{\omega(\omega_0^2 - \omega^2 - \mathrm{i}\gamma\omega/m)} \cos\omega(t-t')\ , \end{aligned} \tag{5.196}$$

from which the relation can again be seen.

In Sect. 3.4 we already found relations between the susceptibilities and the covariances, which are also referred to as the fluctuation–dissipation theorem. Those were simpler relations following immediately from the definition of the time-independent susceptibilities. Of course, in some cases one obtains the time-independent form as a special case of the more general time-dependent form.

An even simpler version of such a relation results when we determine the stationary probability for the above-considered general dissipative system in a stochastic environment. One obtains

$$\varrho_{\text{stat}}(p,x) = C \exp\left[-\frac{\gamma}{m\sigma^2}p^2 - \frac{m\omega_0^2\gamma}{\sigma^2}x^2\right] , \tag{5.197}$$

because the variance for the stationary state follows from (5.173) as

$$\Sigma_{22} = \frac{m\sigma^2}{2\gamma} , \quad \Sigma_{11} = \frac{\sigma^2}{2\gamma\omega_0^2} , \quad \Sigma_{12} = \Sigma_{21} = 0 . \tag{5.198}$$

This result is reasonable: The stronger the friction is, the more sharply the distribution is peaked around $p = 0$ and $x = 0$. The opposite is true for the diffusion, whose strength is measured by $\sigma^2$. The distribution factorizes into two Gaussian distributions for $p$ and $x$.

If the environment is considered as a heat bath of temperature $T$, the stationary case describes a particle in equilibrium with its environment. In this case we have,

$$\frac{1}{2m}\langle P^2\rangle = \frac{1}{2}k_{\text{B}}T , \tag{5.199}$$

as shown in (3.210), and, since $\langle P^2\rangle/2m = \Sigma_{22}/2m = \sigma^2/4\gamma$, finally

$$\frac{\sigma^2}{\gamma} = 2k_{\text{B}}T . \tag{5.200}$$

The intensity of the noise is therefore determined by the temperature of the heat bath. One obtains from (5.197)

$$\varrho_{\text{stat}}(x,p) = C\,\mathrm{e}^{-H(p,x)/k_{\text{B}}T} \tag{5.201}$$

where

$$H(p,x) = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2x^2 , \tag{5.202}$$

which is the form known from statistical mechanics.

Condition (5.200), which is also called the Einstein relation, states a connection between the strength of the fluctuations, given by $\sigma^2$, and the strength of the dissipation $\gamma$. This is the fluctuation–dissipation relation in its simplest form.

## 5.8 Approximation Methods

The methods introduced in the previous sections for finding an analytic solution are only applicable to equations with linear transition rates. Even sophisticated generalizations of these methods, necessary for problems with more complicated boundary conditions (see e.g. van Kampen 1985; Honerkamp 1994), do not extend beyond the linear regime. For equations with nonlinear transition rates one has to rely on approximation methods.

### *5.8.1 The Ω Expansion*

In this section one of the most important approximation methods, the $\Omega$ expansion, will be introduced. Here $\Omega$ represents some size of the system which is relevant in the context of the problem and, in a certain sense, defines a macroscopic scale. Frequently, for instance, $\Omega$ denotes the volume of a system. An expansion of the terms in the master equation with respect to $1/\Omega$, where only the leading contributions in $1/\Omega$ are taken into account, means that one considers the motion only on a macroscopic scale and that smaller details are to be neglected.

#### **Introduction to the Concept in a Simple Example**

We begin by studying a simple form of this expansion, namely the case of the random walk. We thus consider the master equation

$$\dot{\varrho}_n = \alpha\varrho_{n+1} + \beta\varrho_{n-1} - (\alpha+\beta)\varrho_n \ , \tag{5.203}$$

for $-L \le n \le +L$ with appropriately modified boundary conditions near the boundary points $n = \pm L$. We choose $L \gg 1$, and $L$ now represents the above-mentioned size $\Omega$, i.e., we are interested in the motion on a length scale of the order $L$. A jump $n \to n \pm 1$, for instance, is small in comparison to $L$.

We introduce

$$x = \frac{n}{L} \ , \tag{5.204}$$

i.e., $n \to n \pm 1$ now becomes $x \to x \pm \frac{1}{L}$. For

$$\widetilde{\Pi}(x,t) = \varrho_n(t) \tag{5.205}$$

we get

$$\frac{\partial}{\partial t}\widetilde{\Pi}(x,t) = \alpha\widetilde{\Pi}(x+\frac{1}{L},t) + \beta\widetilde{\Pi}(x-\frac{1}{L},t) - (\alpha+\beta)\widetilde{\Pi}(x,t) \ . \tag{5.206}$$

We now expand $\widetilde{\Pi}(x\pm\frac{1}{L},t)$ with respect to $1/L$ and obtain an equation for $\widetilde{\Pi}(x,t)$:

$$\begin{aligned}\frac{\partial}{\partial t}\widetilde{\Pi}(x,t) = (\alpha-\beta)\frac{1}{L}\frac{\partial}{\partial x}\widetilde{\Pi}(x,t) + \frac{(\alpha+\beta)}{2}\frac{1}{L^2}\frac{\partial^2}{\partial x^2}\widetilde{\Pi}(x,t) \\ +O\left(\frac{1}{L^3}\right) \ .\end{aligned} \tag{5.207}$$

We first consider the case $\alpha = \beta$, i.e., the term of order $1/L$ vanishes. In addition, we simply set $\alpha = 1$. Defining $\tau = t/L^2$ and $\Pi(x,\tau) = \widetilde{\Pi}(x,t)$, we obtain, by neglecting terms of order $1/L^3$ and higher, an equation for $\Pi(x,\tau)$:

$$\frac{\partial}{\partial \tau}\, \Pi(x,\tau) = \frac{\partial^2}{\partial x^2}\, \Pi(x,\tau) \ . \tag{5.208}$$

Hence, instead of the variables $n, t$ we have introduced the variables

$$x = \frac{n}{L}\,, \qquad \tau = \frac{t}{L^2}\ . \tag{5.209}$$

Now $x$ and $\tau$ are of order 1, when $n$ is of the order $L$ and $t$ of the order $L^2$. On a large length and time scale the random walk can therefore be described by the density $\Pi(x,\tau)$ and by (5.208) for this density. This is the most simple Fokker–Planck equation (cf. (5.156)). One possible solution is

$$\Pi(x,\tau) = \frac{1}{\sqrt{2\pi\tau}}\, \exp\left(-\frac{x^2}{2\tau}\right)\ , \tag{5.210}$$

i.e., $\Pi(x,\tau)$ is the density of a Gaussian distribution with

$$\langle X \rangle = 0 \tag{5.211}$$

and

$$\mathrm{Var}(X) = \tau\ . \tag{5.212}$$

This is consistent with the results (5.94) that we obtained for the random walk before making the approximation.

Hence, when we observe a particle on a random walk 'from far away', i.e., on a large length scale, and 'for quite a while', i.e., on a large time scale, this jump process will look like diffusion. If one observes many particles moving simultaneously in this same way, one sees a particle cloud whose extension will increase in time $\tau$ as $\sqrt{\tau}$, as it is the case for the diffusion. A process which is described by a Fokker–Planck equation is therefore also called a diffusion process.

Now let $\alpha \neq \beta$. In this case the leading term on the right-hand side is of the order $1/L$. When we neglect all terms of higher order in $1/L$ and introduce

$$\tau = \frac{t}{L}\ , \tag{5.213}$$

we find as the equation for $\Pi(x,\tau) = \widetilde{\Pi}(x,t)$:

$$\frac{\partial}{\partial \tau}\, \Pi(x,\tau) = (\alpha - \beta)\, \frac{\partial}{\partial x}\, \Pi(x,\tau)\ . \tag{5.214}$$

But from (5.43), such an equation results for a deterministic motion of $x(\tau)$ according to

$$\dot{x}(\tau) = -(\alpha - \beta) \; . \tag{5.215}$$

It seems reasonable that the asymmetric random walk now contains such a motion, because the asymmetry of the transition rates will now lead to a tendency to favor one direction. However, superimposed on this deterministic motion will be fluctuations which are not visible in first order of $1/L$. So we would expect for the time-dependent random variable $N(t)$ with realizations $n(t)$ a motion which may be written in the form

$$N(t) = L\, x(\tau) + L^{1/2}\, \xi(\tau) \; , \quad \tau = \frac{t}{L} \; , \tag{5.216}$$

where $x(\tau)$ is determined by (5.215) and $\xi(\tau)$ represents the fluctuations. Then $\langle N(t) \rangle = O(L)$ and also $\mathrm{Var}(N(t)) = O(L)$, as we know from systems with many degrees of freedom (cf. Sect. 3.2). Instead of $N(t) = L\xi(\tau)$, as we used in (5.209), we should now insert the ansatz (5.216) into the master equation.

Hence, we must distinguish two cases. For the case $\alpha = \beta$ the term of order $1/L$ vanishes and one obtains a Fokker–Planck equation for the density function $\Pi(x, \tau)$, where $x = n/L$, and $\tau = t/L^2$. The stochastic process which we observe on these length and time scales is a diffusive process.

For $\alpha \neq \beta$ the term of order $1/L$ signals that there is a dominant deterministic motion, on which are superimposed fluctuations of the order $L^{1/2}$. Inserting the ansatz (5.216) into the master equation, one will obtain a deterministic equation for $x(\tau)$ and an equation for the density function of the fluctuations. This will now be investigated for the general case.

### Examination of the General Case

We again write $\Omega$ for the macroscopic system size and proceed from the master equation

$$\dot{\varrho}(X, t) = \int \Big( w_\Omega(X \mid X')\, \varrho(X', t) - w_\Omega(X' \mid X)\, \varrho(X, t) \Big)\, \mathrm{d}X' \; . \tag{5.217}$$

Following (van Kampen 1985), we require that the dependence of the transition rates on $X$, $X'$, and $\Omega$ can be represented in the form

$$w_\Omega(X \mid X') = f(\Omega) \left( \phi_0\left(\frac{X'}{\Omega}; r\right) + \frac{1}{\Omega}\, \phi_1\left(\frac{X'}{\Omega}; r\right) + \ldots \right) , \tag{5.218}$$

where $r = X - X'$. Setting $x = \frac{X}{\Omega}$ and $x' = \frac{X'}{\Omega}$ we obtain as the master equation for $\widetilde{\Pi}(x, t) = P(X, t)$

$$\begin{aligned}\frac{\partial \widetilde{\Pi}(x,t)}{\partial t} &= f(\Omega)\int\left(\phi_0\left(x-\frac{r}{\Omega};r\right)+\frac{1}{\Omega}\,\phi_1\left(x-\frac{r}{\Omega};r\right)+\ldots\right)\\ &\quad\times\widetilde{\Pi}\left(x-\frac{r}{\Omega},t\right)\,\mathrm{d}r \qquad (5.219)\\ &\quad-f(\Omega)\int\left(\phi_0(x,-r)+\frac{1}{\Omega}\,\phi_1(x,-r)+\ldots\right)\widetilde{\Pi}(x,t)\,\mathrm{d}r\ .\end{aligned}$$

Expansion of the right-hand side in powers of $\Omega^{-1}$ leads to

$$\begin{aligned}\frac{1}{f(\Omega)}\frac{\partial \widetilde{\Pi}(x,t)}{\partial t} &= \frac{1}{\Omega}\left[-\frac{\partial}{\partial x}\int \mathrm{d}r\, r\,\phi_0(x;r)\,\widetilde{\Pi}(x,t)\right]\\ &\quad+\frac{1}{\Omega^2}\left[-\frac{\partial}{\partial x}\int \mathrm{d}r\, r\,\phi_1(x;r)\,\widetilde{\Pi}(x,t)\right. \qquad (5.220)\\ &\quad\left.+\frac{1}{2}\frac{\partial^2}{\partial x^2}\int \mathrm{d}r\, r^2\,\phi_0(x;r)\,\widetilde{\Pi}(x,t)\right]+O(\Omega^{-3})\ .\end{aligned}$$

Setting

$$\alpha_{n,m}(x) = \int \mathrm{d}r\, r^n\,\phi_m(x;r) \qquad (5.221)$$

we therefore obtain

$$\begin{aligned}\frac{1}{f(\Omega)}\,\frac{\partial \widetilde{\Pi}(x,t)}{\partial t} &= \frac{1}{\Omega}\left[-\frac{\partial}{\partial x}\,\alpha_{1,0}(x)\,\widetilde{\Pi}(x,t)\right] \qquad (5.222)\\ &\quad+\frac{1}{\Omega^2}\left[-\frac{\partial}{\partial x}\,\alpha_{1,1}(x)\,\widetilde{\Pi}(x,t)+\frac{1}{2}\frac{\partial^2}{\partial x^2}\,\alpha_{2,0}(x)\;\widetilde{\Pi}(x,t)\right].\end{aligned}$$

Notice that the differential operators act on all terms to the right of them. Again we have to distinguish between the case where the term of order $1/\Omega$ is present and that where it vanishes because $\alpha_{1,0}(x)\equiv 0$.

We first consider the case

$$\alpha_{1,0}(x) = \int \mathrm{d}r\, r\,\phi_0(x;r)\equiv 0\ . \qquad (5.223)$$

Introducing $\tau = \Omega^{-2}f(\Omega)t$, one obtains in the limit $\Omega\to\infty$ for $\Pi(x,\tau) = \widetilde{\Pi}(x,t)$ the equation

$$\frac{\partial \Pi(x,\tau)}{\partial \tau} = \left[-\frac{\partial}{\partial x}\,\alpha_{1,1}(x)+\frac{1}{2}\,\frac{\partial^2}{\partial x^2}\,\alpha_{2,0}(x)\right]\Pi(x,\tau)\ . \qquad (5.224)$$

Equation 5.224 is a general Fokker–Planck equation. (In the case of the random walk $\phi_0\equiv 1,\ \phi_1\equiv 0$, $r$ takes on only the discrete values $r=\pm 1$, and therefore $\alpha_{1,1}\equiv 0,\ \alpha_{2,0}\equiv 2$.)

Hence, for large length and time scales the Fokker–Planck equation describes the stochastic process originally described by a master equation. We have assumed the validity of the representation (5.218) for the transition rates and have considered up to now the case (5.223). The term $\alpha_{1,1}(x)$ denotes the drift term, $\alpha_{2,0}(x)$ the diffusion term.

*Remark.* From

$$\begin{aligned} w_\Omega(X \mid X') = \frac{1}{2}\left(D(X') + \frac{1}{\Omega}A(X')\right)\delta_{X',X+1} \\ +\frac{1}{2}\left(D(X') - \frac{1}{\Omega}A(X')\right)\delta_{X',X-1} \ , \end{aligned} \tag{5.225}$$

one obtains immediately

$$\alpha_{1,1}(x) = A(x) \tag{5.226a}$$

$$\alpha_{2,0}(x) = D(x) \ , \tag{5.226b}$$

i.e., the Fokker–Planck equation with given drift term $A(x)$ and diffusion term $D(x)$ may also be obtained from a master equation with transition rates (5.225). Yet, this is only one master equation among many yielding the given Fokker–Planck equation in the limit $\Omega \to \infty$.

Now we turn to the case

$$\alpha_{1,0}(x) \equiv \int \mathrm{d}r\, r\, \phi_0(x,r) \not\equiv 0 \ . \tag{5.227}$$

Introducing a new time variable

$$\tau = f(\Omega)\frac{t}{\Omega} \tag{5.228}$$

into (5.222), we obtain for $\widetilde{\Pi}'(x,\tau) = \widetilde{\Pi}(x,t)$ up to order $1/\Omega$

$$\begin{aligned} \frac{\partial \widetilde{\Pi}'(x,\tau)}{\partial \tau} = \left[-\frac{\partial}{\partial x}\alpha_{1,0}(x)\,\widetilde{\Pi}'(x,t)\right] \\ +\frac{1}{\Omega}\left[-\frac{\partial}{\partial x}\alpha_{1,1}(x)\,\widetilde{\Pi}'(x,\tau) \ + \frac{1}{2}\frac{\partial^2}{\partial x^2}\alpha_{2,0}(x)\ \widetilde{\Pi}'(x,\tau)\right] . \end{aligned} \tag{5.229}$$

Without the term of order $1/\Omega$, this equation would be a master equation for a deterministic process with solution

$$\widetilde{\Pi}'(x,\tau) = \delta(x - c(\tau)) \ , \tag{5.230}$$

where $c(\tau)$ solves the differential equation

$$\dot{c}(\tau) = \alpha_{1,0}[c(\tau)] \equiv \alpha_{1,0}(x)|_{x=c(\tau)} \ . \tag{5.231}$$

The total stochastic process is therefore dominated by this deterministic process. Fluctuations to this process can be included by the ansatz

$$x(\tau) = c(\tau) + \Omega^{-1/2}\,\xi(\tau)\ , \tag{5.232}$$

i.e., since $X = \Omega\,x$ (cf. (5.218) and (5.219)),

$$X(t) = \Omega\,c(\tau) + \Omega^{1/2}\,\xi(\tau)\ , \quad \tau = f(\Omega)\frac{t}{\Omega}\ . \tag{5.233}$$

Here $\xi(\tau)$ is a time-dependent random variable. Inserting this ansatz into the master equation we obtain the deterministic equation (5.231) for $c(\tau)$, and an equation for the density

$$\Pi(\xi,\tau) = \widetilde{\Pi}'(x,\tau) \tag{5.234}$$

of the fluctuations $\xi(t)$.

We first observe that

$$\frac{\partial}{\partial x} = \frac{\partial\xi}{\partial x}\,\frac{\partial}{\partial\xi} = \Omega^{1/2}\frac{\partial}{\partial\xi} \tag{5.235}$$

and, for $x$ fixed, we get from (5.232)

$$\frac{\partial\xi}{\partial\tau} = -\Omega^{1/2}\,\dot{c}(\tau)\ . \tag{5.236}$$

From (5.229) we therefore obtain

$$\begin{aligned}\frac{\partial\Pi(\xi,\tau)}{\partial\tau} - \Omega^{1/2}\dot{c}(\tau)\,\frac{\partial\Pi(\xi,\tau)}{\partial\xi} &= -\Omega^{1/2}\,\alpha_{1,0}[c]\,\frac{\partial}{\partial\xi}\,\Pi(\xi,\tau)\\ &+\Big(-\alpha'_{1,0}[c]\,\frac{\partial}{\partial\xi}\,\xi\,+\,\frac{1}{2}\,\alpha_{2,0}[c]\,\frac{\partial^2}{\partial\xi^2}\Big)\,\Pi(\xi,\tau) + O(\Omega^{-1/2})\ ,\end{aligned} \tag{5.237}$$

where $\alpha'_{1,0}[c] = \frac{\partial}{\partial x}\alpha_{1,0}(x)\big|_{x=c(\tau)}$. From the terms of order $\Omega^{1/2}$ we therefore find indeed

$$\dot{c}(\tau) = \alpha_{1,0}[c] \tag{5.238}$$

as the law for the dominant deterministic motion. For the fluctuations around $c(\tau)$ we obtain a Fokker–Planck equation for the density $\Pi(\xi,\tau)$:

$$\frac{\partial\Pi(\xi,\tau)}{\partial\tau} = \Big(-\alpha'_{1,0}[c]\,\frac{\partial}{\partial\xi}\,\xi\,+\,\frac{1}{2}\,\alpha_{2,0}[c]\,\frac{\partial^2}{\partial\xi^2}\Big)\,\Pi(\xi,\tau)\ . \tag{5.239}$$

Notice that this Fokker–Planck equation only contains a linear drift term and a $\xi$-independent diffusion term. As we have seen in Sect. 5.6, it is easy to determine the behavior of the fluctuations in a linear diffusion process, especially for large times $\tau$

for which $c(\tau)$ is practically equal to the stationary solution $c^{\text{stat}}$. This is given by the solution of the equation

$$\alpha_{1,0}(c^{\text{stat}}) = 0 \, . \tag{5.240}$$

For the fluctuations $\xi(\tau)$ around the stationary solution we get from (5.172)

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\langle\xi(\tau)\rangle = \alpha'_{1,0}(c^{\text{stat}})\,\langle\xi(\tau)\rangle \, . \tag{5.241}$$

Thus we find that only when

$$\alpha'_{1,0}(c^{\text{stat}}) < 0 \, , \tag{5.242}$$

are the fluctuations damped, and only then is the stationary solution $c^{\text{stat}}$ stable (which also follows from a direct stability analysis of the differential equation for $c(\tau)$). This also implies that only for this case is the $\Omega$ expansion together with the ansatz (5.233), i.e.,

$$X(t) = \Omega c(\tau) + \Omega^k \xi(\tau), \quad k = \frac{1}{2}, \quad \tau = f(\Omega)\frac{t}{\Omega} \, , \tag{5.243}$$

adequate for the dynamics of the stochastic process.

In other cases the fluctuations will be of different order in $\Omega$. An $\Omega$ expansion with the ansatz (5.243) with a general exponent $k$, to be determined, might eventually lead then to a consistent picture of the structure of the stochastic motion (Malek Mansour et al. 1981). In the multivariate case, the different fluctuations can also be of different order in $\Omega$ (Breuer et al. 1995).

### 5.8.2 The One-Particle Picture

In Sect. 5.3 we have already met several master equations for stochastic processes in which many particles take part. Usually, one considers the number of particles $n_\lambda$ within one cell $\lambda$, where, depending on the problem, the cells represent volume elements in configuration space or in phase space. (If there are several species $A$, $B, \ldots$ of particles, the collection $z = \{n^A_\lambda, n^B_\lambda, \ldots, \ \lambda = 1, \ldots\}$ serves as a label for a momentary state $z$.)

This is known as the occupation number representation of a state in the system, and diffusion and scattering are easily described in this representation as a certain change of the occupation numbers of individual cells. Correspondingly simple is the formulation of a master equation for the density of the random variable $\boldsymbol{Z}$, which in the present context shall be denoted by $P(z;t)$.

Although this representation of a state in a system by a set of occupation numbers is complete, it has its disadvantages, in particular with respect to numerical methods. The vector $z$ for the numbers of particles in each cell generally has a very high dimension: Often a simulation utilizes more cells than particles and therefore many occupation numbers vanish, which is difficult to take into account properly in an algorithm.

This suggests that one might describe the process in a different picture, e.g. by tracking the stochastic paths of the individual particles.

To make this more explicit, we first consider the random walk of a single particle among the cells $\lambda = 1, \ldots$. Let $\varrho_\lambda(t)$ be the probability of finding the particle at time $t$ in cell $\lambda$, then the master equation for $\varrho_\lambda(t)$ reads

$$\dot{\varrho}_\lambda(t) = \sum_{\lambda'} \Big( w_{\lambda\lambda'} \varrho_{\lambda'}(t) - w_{\lambda'\lambda} \varrho_\lambda(t) \Big) \tag{5.244}$$

with properly chosen transition rates $w_{\lambda\lambda'}$. This master equation is orders of magnitude simpler than the one for $P(z;t)$, because $\lambda$, being the argument of $\varrho_\lambda(t)$, is just a number, whereas $z = (n_1, \ldots)$ is a vector of very high dimensionality.

A stochastic process which is described by (5.244) is easily to handle. From this equation for the stochastic process of a single particle we will now derive the master equation for the process with many independent particles in the occupation number representation, and we will ask under what conditions a general master equation can be reduced to a single-particle master equation.

So we now introduce the occupation number representation for the stochastic process in which $N$ particles are allowed to jump independently of each other (i.e., without interactions) from cell to cell according to (5.244). Obviously,

$$P(n_1, \ldots, n_\lambda, \ldots; t) = \begin{pmatrix} N \\ n_1 \, n_2 \ldots \end{pmatrix} \prod_\lambda \big(\varrho_\lambda(t)\big)^{n_\lambda} \, . \tag{5.245}$$

The multinomial factor $\binom{N}{n_1 \, n_2 \ldots} = \frac{N!}{n_1! n_2! \ldots}$ guarantees, amongst other things, that $P(n_1, \ldots; t)$ is normalized to 1, as is $\varrho_\lambda(t)$. We shall further insist that this factor vanishes unless $n_1 + n_2 + \cdots = N$.

The master equation for $P(n_1, \ldots; t)$ is now easily derived from the one for $\varrho_\lambda(t)$. We obtain

$$\begin{aligned}
\frac{\partial}{\partial t} P(n_1, \ldots; t) &= \sum_\lambda \begin{pmatrix} N \\ n_1 \ldots \end{pmatrix} \prod_{\lambda' \neq \lambda} \big(\varrho_{\lambda'}(t)\big)^{n_{\lambda'}} \, n_\lambda \big(\varrho_\lambda(t)\big)^{n_\lambda - 1} \, \dot{\varrho}_\lambda(t) \\
&= \sum_\lambda \begin{pmatrix} N \\ n_1 \ldots \end{pmatrix} \prod_{\lambda' \neq \lambda} \big(\varrho_{\lambda'}(t)\big)^{n_{\lambda'}} \, \big(\varrho_\lambda(t)\big)^{n_\lambda - 1} \, n_\lambda \\
&\quad \cdot \left( \sum_{\lambda''} w_{\lambda\lambda''} \varrho_{\lambda''}(t) - w_{\lambda''\lambda} \varrho_\lambda(t) \right) \\
&= \sum_{\lambda,\lambda''} w_{\lambda\lambda''} \left( \mathsf{E}_{\lambda''} \mathsf{E}_\lambda^{-1} - 1 \right) n_{\lambda''} \, P(n_1, \ldots; t) \, .
\end{aligned} \tag{5.246}$$

The last rearrangement may best be studied in an example. Suppose there are only two cells, $\lambda = 1, 2$. Then

$$P(n_1, n_2; t) = \frac{N}{n_1!\, n_2!} \left(\varrho_1(t)\right)^{n_1} \left(\varrho_2(t)\right)^{n_2} \tag{5.247}$$

and

$$\begin{aligned}
\frac{\partial}{\partial t} P\ (\ n_1, n_2; t)&\\
&= \frac{N}{n_1!\, n_2!} \left(\varrho_1(t)\right)^{n_1-1} \left(\varrho_2(t)\right)^{n_2} n_1 \left(w_{12}\varrho_2(t) - w_{21}\varrho_1(t)\right)\\
&\quad + \frac{N}{n_1!\, n_2!} \left(\varrho_1(t)\right)^{n_1} \left(\varrho_2(t)\right)^{n_2-1} n_2 \left(w_{21}\varrho_1(t) - w_{12}\varrho_2(t)\right)\\
&= \frac{N}{(n_1-1)!\,(n_2+1)!} \left(\varrho_1(t)\right)^{n_1-1} \left(\varrho_2(t)\right)^{n_2+1} (n_2+1)\, w_{12}\\
&\quad + \frac{N}{(n_1+1)!\,(n_2-1)!} \left(\varrho_1(t)\right)^{n_1+1} \left(\varrho_2(t)\right)^{n_2-1} (n_1+1)\, w_{21}\\
&\quad - \frac{N}{n_1!\, n_2!} \left(\varrho_1(t)\right)^{n_1} \left(\varrho_2(t)\right)^{n_2} \left(n_1 w_{21} + n_2 w_{12}\right)\\
&= \sum_{\lambda,\lambda'=1}^{2} w_{\lambda\lambda'} \left(\mathsf{E}_{\lambda'}\mathsf{E}_{\lambda}^{-1} - 1\right) n_{\lambda'}\, P(n_1, n_2; t)\ .
\end{aligned} \tag{5.248}$$

Hence, in the occupation number representation we have obtained a master equation for which the transition rate $w_{z'z}$ is linear in $z = (n_1, \ldots, n_\lambda, \ldots)$ (or rather a component of $z$). Both master equations describe the same stochastic process, which is simply the random walk of a particle among cells, here taken $N$-fold as the stochastic process of $N$ independent particles.

From these considerations we have learnt the following: Given a master equation,

$$\frac{\partial}{\partial t} P(z, t) = \sum_{z'} \left(w_{zz'} P(z', t) - w_{z'z} P(z, t)\right)\ , \tag{5.249}$$

in the occupation number representation, i.e., such that $z = (n_1, \ldots)$, then for $P(z, t)$ we may formulate the ansatz

$$P(z, t) \equiv P(n_1, \ldots; t) = \binom{N}{n_1 \ldots} \prod_\lambda \left(\varrho_\lambda(t)\right)^{n_\lambda} \tag{5.250}$$

and ask for the equation which the functions $\{\varrho_\lambda(t)\}$ have to satisfy in order for $P(z, t)$ to be a solution of the master equation (5.249).

If the master equation (5.249) has the form (5.246), we will find the master equation (5.244) for $\{\varrho_\lambda(t)\}$. This is the case in which the particles genuinely propagate independently, and the master equation (5.244) describes the motion of one of these particles. If there is an interaction among the particles, the transition rates $w_{z'z}$ are not of the form we found in (5.246). In such a case, one first has to find an approximation to bring them into this form and to reduce the problem to the simpler master equation (5.244) by making the ansatz (5.245). Such an approximation consists, for instance, in the replacement of all particle numbers occurring in $w_{z'z}$, except one, by their expectation values, which now have to be determined from a consistency condition. This corresponds to the mean field approximation.

## 5.9 More General Stochastic Processes

So far in this chapter we have addressed only Markov processes. In this section we will meet four large and more general classes of stochastic processes, which also play an important role for the description of many phenomena in complex systems.

### *5.9.1 Self-Similar Processes*

A stochastic process is called self-similar with index $H$, if for any $a > 0$ all distributions (i.e., also the two-time distribution and the multi-time distributions) of $X(at)$ are identical to the distributions of $a^H X(t)$. In such processes a change of the time scale can be compensated for by a change of the length scale (or, more precisely, the scale of $X$).

The class of self-similar processes was defined by Mandelbrot (1982), and it covers a great many processes. Brownian motion (Sect. 5.6), for instance, is self-similar with index $H = 1/2$, because

$$\langle X(at_1)X(at_2)\rangle = \min(at_1, at_2) = a\min(t_1, t_2) \tag{5.251}$$

$$= \left\langle a^{1/2}X(t_1)a^{1/2}X(t_2)\right\rangle \,, \tag{5.252}$$

and this covariance function determines all distributions.

In the following we assume $H > 0$. A self-similar process cannot be stationary, since $X(t)$, $X(at)$ and $a^H X(t)$ all have the same distribution and $a^H X(t)$ tends to $\infty$ for $a \to \infty$. However, to each self-similar process $X(t)$ with index $H$ one can associate a stationary process, namely a corresponding Ornstein–Uhlenbeck process by $Y(t) = \mathrm{e}^{-tH}X(\mathrm{e}^t)$ (Samorodnitzky and Taqqu 1994). Let $X(t)$ be the Brownian motion, for example; then one obtains for $Y(t) = \mathrm{e}^{-t/2}X(\mathrm{e}^t)$ the covariance

$$\langle Y(t_1)Y(t_2)\rangle = \exp\left(-\frac{1}{2}\,|t_1 - t_2|\right) \,, \tag{5.253}$$

already known to us from Ornstein–Uhlenbeck processes (cf. (5.178)).

### 5.9.2 Fractal Brownian Motion

In Sect. 5.6 we characterized Brownian motion by its independent and normally distributed increments. Now we will drop the assumption of independence.

As we have seen in Sect. 5.6, a Gaussian stochastic process $X(t)$ with $\langle X(t)\rangle = 0$ is determined only by its covariance function. For this we now require:

$$\begin{aligned} C(t_1, t_2) &\equiv \mathrm{Cov}(X(t_1), X(t_2)) \\ &= \frac{\sigma^2}{2}(|t_1|^{2H} + |t_2|^{2H} - |t_2 - t_1|^{2H}) \,, \end{aligned} \tag{5.254}$$

with $0 < H \leq 1$. For $H = 1/2$ we again obtain Brownian motion, for $H \neq 1/2$ the so-defined Gaussian process is called fractal Brownian motion. One can show that for $H = 1$ only the process $X(t) = tX(1)$ exists.

Fractal Brownian motion is self-similar with index $H$.

The increments

$$Y(t) = X(t+1) - X(t) \tag{5.255}$$

represent a stationary discrete-time process. Such a process is called fractal Gaussian noise. Hence, this is noise which is normally distributed, but for $H \neq 1/2$ it is not 'white'. For the covariance function one obtains (see, e.g., Samorodnitzky and Taqqu 1994)

$$\begin{aligned} C_j &\equiv \mathrm{Cov}(Y(t), Y(t+j)) \\ &= \frac{\sigma^2}{2}\left(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\right) \quad j = 1, 2, \ldots \,. \end{aligned} \tag{5.256}$$

For $H = 1/2$ we find, as expected, $C_j = 0$ for $j = 1, 2, \ldots$; for $H \neq 1/2$, however, we find for $j \to \infty$:

$$C_j \propto \sigma^2 H(2H-1) j^{2H-2}, \tag{5.257}$$

i.e., for $1/2 < H < 1$, $C_j$ decreases in the limit $j \to \infty$ so slowly that the $\sum_{j=1}^{\infty} C_j$ is divergent. In this case one speaks of a long-range dependence.

Fractal Gaussian noise is easy to simulate. We represent the covariance matrix with elements $K_{ij} = C_{j-i}, i, j = 1, \ldots, N$ according to the Cholesky method as

$$\mathsf{K} = \mathsf{R}\,\mathsf{R}^T \,, \tag{5.258}$$

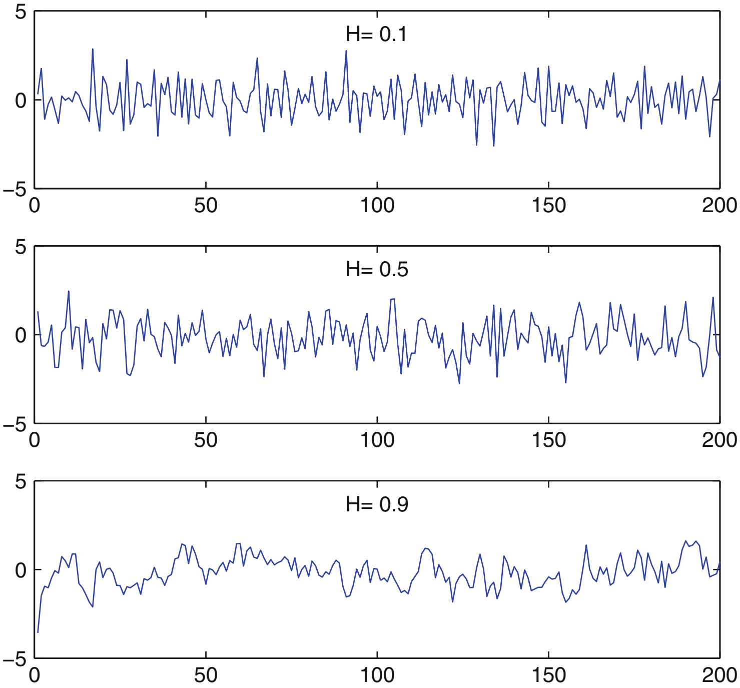


**Fig. 5.6** Fractal Gaussian noise for various indices $H$

where $\mathsf{R}$ is an upper triangular matrix. If we define

$$X_i = \sum_{k=1}^{N} R_{ik}\eta_k \ , \quad i = 1, \ldots, N \ , \tag{5.259}$$

where $\{\eta_k\}$ are independent standard normal random variables, then the random variables $\{X_i\}$ obviously have the required covariance matrix. Some realizations of fractal Gaussian noise are shown in Fig. 5.6.

### *5.9.3 Stable Levy Processes*

Having dropped the requirement of the independence of the increments when we introduced fractal Brownian motion, we now will retain this independence, but permit non-normal distributions for the increments. In fact, the increments will be random variables with a stable distribution for an index $\alpha$. In this case one speaks of $\alpha$-stable Levy processes. It is not difficult to see that these are self-similar with an index $H = 1/\alpha$.

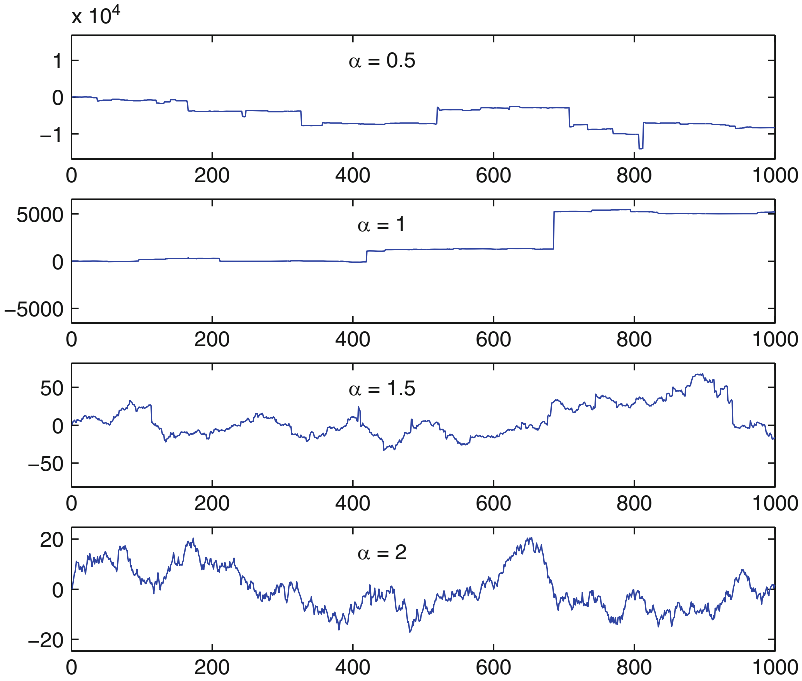


**Fig. 5.7** Realizations of Levy processes for various indices $\alpha$

Discrete-time Levy processes are easy to simulate, as one simply has to generate the realizations of $\alpha$-stable random variables (cf. Sect. 2.6) and sum them up. Various Levy processes are shown in Fig. 5.7. With decreasing $\alpha$ ever more very excessively large increments appear.

### *5.9.4 Autoregressive Processes*

A class of stochastic processes frequently discussed in time series analysis are the autoregressive processes. In this case one considers random variables $\{X(t)\}$ at discrete instants $t = \ldots, -1, 0, 1, \ldots$. The dependence among the random variables is described by an equation of the form

$$X(t) = \sum_{k=1}^{p} \alpha_k X(t-k) + \eta(t) + \sum_{k=1}^{q} \beta_k \eta(t-k) \tag{5.260}$$

$$\eta(t) \propto \mathrm{WN}(0, \sigma^2) \, .$$

The $\{\alpha_k, \beta_k\}$ are coefficients whose relevance will be discussed in a moment. A process which satisfies such an equation is also called an ARMA$(p, q)$ process, where the letters AR stand for ‘autoregressive’ and MA for ‘moving average’. Notice that for $p > 1$ the process no longer satisfies the Markov condition, because the probability of a realization $x$ at time $t$ is that of a Gaussian random variable with expectation value

$$\mathrm{E}(X(t)) = \sum_{k=1}^{p} \alpha_k x(t-k) \ , \tag{5.261}$$

i.e., for $p > 1$ the probability depends not only on the realization of $X(t)$ at a single earlier instant but on several earlier instants.

*Remark.* In Sect. 5.6 we studied stochastic differential equations for diffusion processes of the form

$$\dot{X}(t) = a(X(t), t) + b(t)\eta(t) \ , \tag{5.262}$$

and for the increment $\mathrm{d}X(t) = X(t+\mathrm{d}t) - X(t)$ we derived an expansion in powers of $\mathrm{d}t$. Up to order $\mathrm{d}t$ this was

$$\mathrm{d}X(t) = a(X(t), t)\mathrm{d}t + b(t)\sqrt{\mathrm{d}t}\,\eta_t \ , \tag{5.263}$$

were $\eta_t$ is a standard normal random variable. We now consider $\mathrm{d}t$ as a fixed time interval and choose our unit of time such that $\mathrm{d}t \equiv 1$. Furthermore, we set $b(t) = \sigma$ and $a(X(t), t) = (\alpha - 1)X(t)$, i.e., we consider a process with constant variance and a drift term linear in $X(t)$. With $\mathrm{d}X(t) = X(t+1) - X(t)$ we obviously obtain

$$X(t+1) = \alpha X(t) + \sigma\eta(t) \tag{5.264}$$

or

$$X(t) = \alpha X(t-1) + \sigma\eta(t) \ , \tag{5.265}$$

where the standard normal random variable $\eta_t$ has been interpreted respectively as $\eta(t+1)$ and as $\eta(t)$. Hence, in this manner we obtain an AR(1) process, which may be thought of as a discrete version of a linear stochastic differential equation of first order with constant variance. In a similar way, the higher AR processes are discrete versions of linear stochastic differential equations of higher order. The MA part of an ARMA process, however, is something new. Within the framework of stochastic processes, ARMA processes describe the dynamics of general linear discrete systems.

We introduce the shift operator $B$ by $B^k X(t) = X(t+k)$ and define what are known as characteristic polynomials $\alpha(z), \beta(z)$ from the respective coefficients $\{\alpha_k\}, \{\beta_k\}$:

$$\alpha(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{5.266}$$

$$\beta(z) = 1 + \sum_{k=1}^{q} \beta_k z^{-k} \ . \tag{5.267}$$

The ARMA$(p, q)$ process can now also be written in the form

$$\alpha(B)X(t) = \beta(B)\eta(t) \ , \quad \eta(t) \sim \mathrm{WN}(0, \sigma^2) \tag{5.268}$$

or

$$X(t) = \frac{\beta(B)}{\alpha(B)}\eta(t) \tag{5.269}$$

or

$$\eta(t) = \frac{\alpha(B)}{\beta(B)}X(t) \ . \tag{5.270}$$

The polynomials $\alpha(z), \beta(z)$ and their properties determine the ARMA process. It is known that each polynomial of $n$th order has $n$ zeros in $\mathbb{C}$:

Let $a_1, \ldots, a_p$ and $b_1, \ldots, b_q$ denote the zeros of $\alpha(z)$ and $\beta(z)$, respectively. We assume that $\alpha(z)$ and $\beta(z)$ have no common zeros, because if $a^* = b^*$ were a common zero of $\alpha(z)$ and $\beta(z)$, the factors $z - a^* = z - b^*$ would cancel in (5.268) or (5.270), and therefore we would be led to an equivalent process of order $(p - 1, q - 1)$.

We can now make the following statement:

If all roots $a_1, \ldots, a_p$ and $b_1, \ldots, b_q$ of the respective characteristic equations of $\alpha(z)$ and $\beta(z)$ lie inside the unit circle, there are practically no divergent realizations (Schlittgen and Streitberg 1987).

The remainder of this section presents three examples of ARMA(p,q) processes.

*Example 1.* For the process $X(t) = 0.8X(t-1) + \sigma\eta(t)$ we have

$$\alpha(z) = 1 - 0.8z^{-1} \ , \tag{5.271}$$

i.e., the zero of $\alpha$, $z = 0.8$, lies inside the unit circle. In simulations of this process we will not meet divergent realizations.

However, the AR(1) process with $\alpha_1 = 1$ corresponds to discrete Brownian motion, i.e., the random walk. Consistent with the above statement it is not stationary.

An ARMA$(p, q)$ process

$$X(t) = \sum_{i=1}^{p} \alpha_i X(t-i) + \sum_{i=1}^{q} \beta_i \eta(t-i) + \eta(t) \ , \tag{5.272}$$

$$\eta(t) \sim \mathrm{WN}(0, \sigma^2)$$

with $p > q$ may also be interpreted as a multivariate AR(1) process for a $p$-dimensional random vector $\boldsymbol{X}(t) = \big(X_1(t), \ldots, X_p(t)\big)$, with $X_1(t) \equiv X(t)$, for which

$$\boldsymbol{X}(t) = \mathsf{A}\,\boldsymbol{X}(t-1) + \mathbf{B}\eta(t)\ , \quad \eta(t) \sim WN(0, \sigma^2)\ , \tag{5.273}$$

where $\mathsf{A}$ is the following $p \times p$ matrix

$$\mathsf{A} = \begin{pmatrix} \alpha_1 & 1\,0 \ldots 0 \\ \alpha_2 & 0\,1 \ldots 0 \\ \vdots & \vdots \quad \ddots\, \vdots \\ \alpha_{p-1} & 0\,0 \ldots 1 \\ \alpha_p & 0\,0 \ldots 0 \end{pmatrix}\ , \tag{5.274}$$

and $\mathbf{B}$ the $p \times 1$ matrix

$$\mathbf{B} = \begin{pmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \tag{5.275}$$

with $\beta_j = 0$ for $j > q$.

*Example 2.* For an ARMA(2,1) process we obtain in the same way

$$\begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} = \begin{pmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{pmatrix} \begin{pmatrix} X_1(t-1) \\ X_2(t-1) \end{pmatrix} + \begin{pmatrix} 1 \\ \beta_1 \end{pmatrix} \eta(t)\ , \tag{5.276}$$

i.e.,

$$X_1(t) = \alpha_1 X_1(t-1) + X_2(t-1) + \eta(t) \tag{5.277a}$$

$$X_2(t) = \alpha_2 X_1(t-1) + \beta_1 \eta(t)\ . \tag{5.277b}$$

Inserting $X_2(t-1)$ from (5.277) into (5.277), we get for $X_1(t) \equiv X(t)$:

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \eta(t) + \beta_1 \eta(t-1)\ , \tag{5.278}$$

i.e., the ARMA(2,1) process in its standard form.

Using a similarity transformation

$$\mathsf{A}' = \mathsf{L}\,\mathsf{A}\,\mathsf{L}^{-1}, \quad \text{or} \quad \mathsf{A} = \mathsf{L}^{-1}\,\mathsf{A}'\,\mathsf{L} \tag{5.279}$$

and

$$\boldsymbol{X}'(t) = \mathsf{L}\,\boldsymbol{X}(t), \quad \mathbf{B}' = \mathsf{L}\,B\ , \tag{5.280}$$

(5.273) can be transformed into

$$\boldsymbol{X}'(t) = \mathsf{A}'\boldsymbol{X}'(t-1) + \mathbf{B}'\eta(t) , \tag{5.281}$$

and the initial random variable $X(t)$ of the univariate process (5.272) is represented as

$$X(t) = \left(\mathsf{L}^{-1}\boldsymbol{X}'(t)\right)_1 = \sum_{i=1}^{p}(\mathsf{L}^{-1})_{1i}\boldsymbol{X}'_i(t) . \tag{5.282}$$

The eigenvalues of the matrix $\mathsf{A}$ or $\mathsf{A}'$ will play an important role in the following. The eigenvalues of a matrix are easily determined when the matrix has the Frobenius form (5.274). For the characteristic polynomial one obtains immediately

$$(-1)^p\left(\lambda^p - \alpha_1\lambda^{p-1} - \ldots - \alpha_p\right) = 0 \tag{5.283}$$

or

$$(-1)^p\lambda^p\alpha(\lambda) = 0 , \tag{5.284}$$

where $\alpha(z)$ is the characteristic polynomial for the process defined in (5.266). We have seen that for a stationary process the zeros of this polynomial have to be inside the unit circle. As the zeros of $\alpha(\lambda)$ determine the eigenvalues $\lambda$, the eigenvalues for a stationary process also have to lie inside the unit circle.

Since in general the matrix $\mathsf{A}$ is not equivalent to a symmetric matrix, there may be complex eigenvalues. Suppose all eigenvalues are different and let $\lambda_1 + \mathrm{i}\mu_1$, $\lambda_1 - \mathrm{i}\mu_1$, $\lambda_3, \ldots$ denote the first three eigenvalues, then, using a suitable similarity transformation $\mathsf{L}$, we may bring $\mathsf{A}'$ into the form

$$\mathsf{A}' = \begin{pmatrix} \lambda_1 & \mu_1 & 0 & 0 \ldots \\ -\mu_1 & \lambda_1 & 0 & 0 \ldots \\ 0 & 0 & \lambda_3 & 0 \ldots \\ 0 & 0 & 0 & \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} . \tag{5.285}$$

A complex eigenvalue therefore leads to a 2 × 2 block on the diagonal, a real eigenvalue to a 1 × 1 block. In this representation the equations decouple to single equations (for a real eigenvalue), e.g.,

$$X'_3(t) = \lambda_3 X'_3(t-1) + B'_3\eta(t) , \tag{5.286}$$

or to pairs of equations (for a complex eigenvalue), e.g.,

$$X'_1(t) = \lambda_1 X'_1(t-1) + \mu_1 X'_2(t-1) + B'_1\eta(t) \tag{5.287a}$$

$$X'_2(t) = -\mu_1 X'_1(t-1) + \lambda_1 X_2(t-1) + B'_2\eta(t) . \tag{5.287b}$$

The individual equations of the form (5.286), and (5.287a) or (5.287b), which emerge through this decoupling, correspond to univariate and bivariate AR(1) processes, respectively.

Contributions to $X(t)$ of the form (5.286) are also called relaxators, i.e., $X'_3(t)$ is an example of such a relaxator. As the eigenvalues lie inside the unit circle, we also have $|\lambda_3| < 1$.

Pairs of equations, such as (5.287a) and (5.287b) which emerge in this decoupling describe linear damped oscillators driven by the noise $\eta(t)$. To see this, we consider the trajectory $\big(X'_1(t), X'_2(t)\big)$ in the complex plane with $Z(t) = X'_1(t) + \mathrm{i}X'_2(t)$. Using the polar representation for the complex eigenvalue,

$$\lambda_1 \pm \mathrm{i}\mu_1 = r\,\mathrm{e}^{\pm\mathrm{i}\varphi}\ , \tag{5.288}$$

we find the equation:

$$Z(t) = r\,\mathrm{e}^{-\mathrm{i}\varphi} Z(t-1)\ , \tag{5.289}$$

where we have disregarded the noise term for the moment. So we get

$$|Z(t)|^2 = r^2 |Z(t-1)|^2 \tag{5.290}$$

and

$$\arg\big(Z(t)\big) = \arg\big(Z(t-1)\big) - \varphi\ . \tag{5.291}$$

As $r < 1$ (the eigenvalues lie inside the unit circle), and still disregarding the driving noise $\eta(t)$, with each step $Z(t)$ comes closer to the origin and it changes by an angle $(-\varphi)$: the trajectory proceeds along a spiral curve towards the origin, like the trajectory of a damped linear oscillator in phase space. For the period we find

$$T = \frac{2\pi}{\varphi} = \frac{2\pi}{\arctan(\mu_1/\lambda_1)}\ , \tag{5.292}$$

and the damping rate $\tau$, defined by $r = \mathrm{e}^{-1/\tau}$, is

$$\tau = -2/\ln(\lambda_1^2 + \mu_1^2)\ . \tag{5.293}$$

Hence, without noise,

$$Z(t+n) = \mathrm{e}^{-n/\tau}\mathrm{e}^{-\mathrm{i}n\varphi} Z(t)\ , \tag{5.294}$$

and after one period $T$ the amplitude $|Z(t)|$ is smaller by the factor $\mathrm{e}^{-T/\tau}$. (From $\lambda_1^2 + \mu_1^2 < 1$ it follows that $\tau > 0$). Without the driving noise the oscillation will die out. The noise thus compensates the dissipation of energy, which is described by the fact that the eigenvalues of the matrix $\mathbf{A}$ are in magnitude smaller than 1. Hence, the stochastic oscillator is an open system.

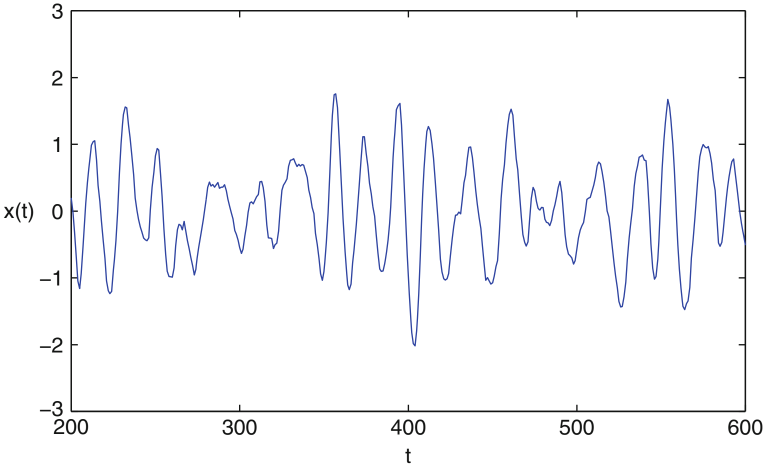


**Fig. 5.8** Part of a realization of an AR(2) process with $T = 20$

The parameters $\{\alpha_1, \ldots, \alpha_p\}$ of an ARMA$(p, q)$ process thus determine the content of oscillators and relaxators together with their properties, and the parameters $\{\beta_1, \ldots, \beta_q\}$ determine the relative strength of the stochastic drive. So we could also characterize an AR$(p)$ process by the characteristic parameters of the oscillators and relaxators.

*Example 3.* We consider a system consisting of one oscillator. From given parameters $(T, \tau)$ we get

$$r = \mathrm{e}^{-1/\tau} \qquad \text{and} \qquad \varphi = 2\pi/T \tag{5.295}$$

and therefore the eigenvalues $\lambda_1 \pm i\mu_1 = r\,\mathrm{e}^{\pm \mathrm{i}\varphi}$. According to (5.283), the polynomial $\alpha(\lambda)$ now reads

$$\begin{aligned}\alpha(\lambda) &= (\lambda - r\,\mathrm{e}^{\mathrm{i}\varphi})(\lambda - r\,\mathrm{e}^{-\mathrm{i}\varphi})\lambda^{-2} \\ &= 1 - 2r\cos\varphi\lambda^{-1} + r^2\lambda^{-2}\,,\end{aligned} \tag{5.296}$$

whence it follows from $\alpha(z) = 1 - \alpha_1 z^{-1} - \alpha_2 z^{-2}$ that

$$\alpha_1 = 2r\cos\varphi = 2\mathrm{e}^{-1/\tau}\cos(2\pi/T) \tag{5.297}$$

$$\alpha_2 = -r^2 = -\mathrm{e}^{-2/\tau}\,. \tag{5.298}$$

For instance, for $(T, \tau) = (20, 50)$ one obtains $\alpha_1 = 1.8645, \alpha_2 = -0.9608$. Figure 5.8 shows a realization of such an AR(2) process.

# Chapter 6
# Quantum Random Systems

In Chap. 3, we discussed the statistical properties of an $N$-particle system treated within the framework of classical mechanics. In this chapter, we will take the more fundamental point of view that the particles and their interactions have to be described quantum mechanically. In quantum mechanics, the state of a system has to be described by a vector in a Hilbert space. If the quantum state is not known completely, one has to introduce the density operator, which contains information about the possible quantum states and their probabilities.

In Sect. 6.1 the density operators for the various thermodynamical systems are introduced in close connection with the classical case.

After a derivation of the grand canonical thermodynamic potential for general ideal gases in Sect. 6.2, we will successively discuss the most important ideal quantum gases in Sects. 6.3–6.6. Such a discussion is part of the standard repertoire of any book on statistical mechanics.

The behavior of Fermi and Bose gases at low temperatures, the phenomenon of Bose-Einstein condensation, Planck's law of radiation, Kirchhoff's laws, Debye's theory of the specific heat in solids – these are central topics of statistical mechanics and they are all discussed very thoroughly.

Section 6.7 will extend such discussions to molecules with internal degrees of freedom. Based on the properties of electron gases, paramagnetic and diamagnetic properties of matter are considered in Sect. 6.8, and in Sect. 6.9 the concept of quasiparticles will be explained. Quantum spin systems, superfluidity, and the superfluid phase of He$^4$ will be briefly discussed within this framework.

## 6.1 Quantum-Mechanical Description of Statistical Systems

Up to now we have considered statistical systems within the framework of classical physics. On a microscopic level the states of the system were given by microstates, i.e., the points $x = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ in a $6N$-dimensional phase space, and observable quantities like the internal energy, $H(x)$, were functions on this

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_6, © Springer-Verlag Berlin Heidelberg 2012

phase space. The microstates were realizations of these random variables and it was possible to specify the probability densities of these variables for various external, macroscopic conditions.

For a quantum-mechanical description of the system, one has to distinguish first between the state of the system and the possible results of a measurement. The measured results continue to be positions, momenta, energies, etc., but in the framework of a quantum-mechanical description, a state of the system is a vector $|\psi\rangle$ in a Hilbert space and the components of this vector with respect to some basis are the wavefunctions $\psi(x)$, $\psi(p)$, or $\psi_E$, where now $\{x = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)\}$, $\{p = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N)\}$, or $\{E\}$ form a complete set of quantum numbers for a complete system of commuting observables. Observable quantities are no longer represented by functions in phase space but by self-adjoint operators in the Hilbert space, and their spectrum represents the possible results of a measurement.

Hence, we have to extend the notion of random variables. We still have to define the possible states and their probabilities, but these states are no longer points in phase space but states $\left\{|\psi_i\rangle, i = 1, \ldots\right\}$ in the Hilbert space.

A mathematical object which, like the probability density in the classical case, contains the information about possible states and their probabilities is the density operator

$$\varrho = \sum_i p(i)\, |\psi_i\rangle\langle\psi_i| \, , \tag{6.1}$$

where

$$\sum_i p(i) = 1. \tag{6.2}$$

Here the possible states $\left\{|\psi_i\rangle, i = 1, \ldots\right\}$ are not necessarily orthogonal. For a pure state, i.e., when the system is with certainty in a definite state $|\psi_{i'}\rangle$, we have $p(i') = 1$, $p(i) = 0$ otherwise, and the density operator

$$\varrho = |\psi_{i'}\rangle\langle\psi_{i'}| \tag{6.3}$$

is the projection operator onto the state $|\psi_{i'}\rangle$. A state which is not pure is called a mixed state.

The quantum-mechanical expectation value of an operator $A$ in a state $|\psi_i\rangle$ is $\langle\psi_i|A|\psi_i\rangle$, and the statistical expectation value is now

$$\langle A\rangle = \sum_i p(i)\, \langle\psi_i|A|\psi_i\rangle = \mathrm{Tr}\left(\sum_i p(i)\, |\psi_i\rangle\langle\psi_i|A\right) = \mathrm{Tr}\,(\varrho\, A). \tag{6.4}$$

*Note*: The equality

$$\mathrm{Tr}\ [|\psi_i\rangle\langle\psi_i|A] = \langle\psi_i|A|\psi_i\rangle$$

results as follows: Let $\{|\alpha\rangle, \alpha = 1, \ldots\}$ be a complete orthonormal system, then

$$\mathrm{Tr}\,[|\psi_i\rangle\langle\psi_i|A] = \sum_\alpha \langle\alpha|\psi_i\rangle\langle\psi_i|A|\alpha\rangle \tag{6.5}$$

$$= \sum_\alpha \langle\psi_i|A|\alpha\rangle\langle\alpha|\psi_i\rangle \tag{6.6}$$

$$= \langle\psi_i|A|\psi_i\rangle \ . \tag{6.7}$$

The probability densities are now replaced by the density operators for the various systems. Before we specify such density operators, we will first study some general properties of density operators.

- $\mathrm{Tr}\,(\varrho) = 1$, a fact which follows immediately from $\mathrm{Tr}\,(\varrho) = \sum_i p(i) = 1$.
- The density operator is hermitian, as can be seen from the definition (6.1).
- For every state $|\psi\rangle$ in the Hilbert space we have

$$\langle\psi|\varrho|\psi\rangle \ \geq \ 0 \ , \tag{6.8}$$

since

$$\langle\psi|\varrho|\psi\rangle = \sum_i p(i)\left|\langle\psi|\psi_i\rangle\right|^2 \geq 0. \tag{6.9}$$

- The density operator satisfies

$$\mathrm{Tr}\,(\varrho^2) \leq 1 \ , \tag{6.10}$$

and $\mathrm{Tr}\,(\varrho^2) = 1$, if and only if $\varrho$ represents a pure state. Since $\left|\langle\psi_i|\psi_j\rangle\right|^2 \leq 1$ by the Schwarz inequality, one finds

$$\mathrm{Tr}\,(\varrho^2) = \sum_{i,j} p(i)\,p(j)\left|\langle\psi_i|\psi_j\rangle\right|^2 \tag{6.11}$$

$$\leq \sum_{i,j} p(i)\,p(j) = 1. \tag{6.12}$$

The proof also implies immediately that $\mathrm{Tr}\,(\varrho^2) = 1$ only if $\left|\langle\psi_i|\psi_j\rangle\right|^2 = 1$ for all states for which $p(i)$ and $p(j)$ are different from zero. This is only possible if only one such state exists, i.e., the density operator is of the form

$$\varrho = |\psi_{i_0}\rangle\langle\psi_{i_0}|. \tag{6.13}$$

As in classical statistical mechanics we can introduce the entropy associated with a density. The entropy of a density operator is defined as

$$S = -k_{\mathrm{B}}\,\mathrm{Tr}\left(\varrho\ln\varrho\right). \tag{6.14}$$

For a complete set of eigenvectors $|\alpha\rangle$ of $\varrho$ one also obtains

$$\varrho = \sum_{\alpha} p_\alpha \, |\alpha\rangle\langle\alpha| \, , \tag{6.15}$$

where $\{p_\alpha\}$ are now the eigenvalues of $\varrho$. Therefore

$$-\ln\varrho = -\sum_{\alpha} (\ln p_\alpha) \, |\alpha\rangle\langle\alpha| \tag{6.16}$$

and

$$S = -k_{\mathrm{B}} \sum_{\alpha} p_\alpha \ln p_\alpha . \tag{6.17}$$

Finally, it is possible to define the relative entropy. The entropy of a density operator $\varrho$ relative to a density operator $\varrho'$ is

$$S[\varrho|\varrho'] = -k_{\mathrm{B}} \left[\mathrm{Tr}\,(\varrho\ln\varrho) - \mathrm{Tr}\,(\varrho\ln\varrho')\right]. \tag{6.18}$$

Again one can show:

$$S[\varrho|\varrho'] \le 0 \, , \tag{6.19}$$

and $S[\varrho|\varrho'] = 0$ only if $\varrho$ and $\varrho'$ are related by a unitary transformation.

*Remarks.*

- Let the free particles in a beam have momentum $\hbar\boldsymbol{k}$ and the eigenvalues for the $z$-component of the spin operator be $m_s = 1/2$ with probability $p_1$ and $m_s = -1/2$ with probability $p_2$. Then the density operator reads

$$\begin{aligned} \varrho &= p_1 \, \left|\boldsymbol{k},\tfrac{1}{2}\right\rangle\left\langle\boldsymbol{k},\tfrac{1}{2}\right| + p_2 \, \left|\boldsymbol{k},-\tfrac{1}{2}\right\rangle\left\langle\boldsymbol{k},-\tfrac{1}{2}\right| & (6.20)\\ &= |\boldsymbol{k}\rangle\langle\boldsymbol{k}| \left(p_1 \, \left|\tfrac{1}{2}\right\rangle\left\langle\tfrac{1}{2}\right| + p_2 \, \left|-\tfrac{1}{2}\right\rangle\left\langle-\tfrac{1}{2}\right|\right) & (6.21)\\ &= \varrho^{\boldsymbol{k}} \, \varrho^{\mathrm{spin}}. & (6.22) \end{aligned}$$

With the representation

$$\left|+\tfrac{1}{2}\right\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{6.23a}$$

$$\left|-\tfrac{1}{2}\right\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{6.23b}$$

$$\tag{6.23c}$$

one finds for the density operator in spin space:

$$\varrho^{\mathrm{spin}} = p_1 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + p_2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} p_1 & 0 \\ 0 & p_2 \end{pmatrix}. \tag{6.24}$$

The statistical expectation value is

$$\langle A\rangle = p_1 \left\langle \boldsymbol{k}, \tfrac{1}{2}\right| A \left|\boldsymbol{k}, \tfrac{1}{2}\right\rangle + \ p_2 \left\langle \boldsymbol{k}, -\tfrac{1}{2}\right| A \left|\boldsymbol{k}, -\tfrac{1}{2}\right\rangle . \tag{6.25}$$

On the other hand, the state

$$|\boldsymbol{k}, \alpha\rangle = |\boldsymbol{k}\rangle \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{6.26}$$

describes a beam of particles which are completely polarized in the $x$-direction. With

$$J_x = \frac{\hbar}{2}\sigma_x \ = \frac{\hbar}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{6.27}$$

as the $x$-component of the spin operator we get

$$S_x \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{\hbar}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} . \tag{6.28}$$

As

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}}\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right] \tag{6.29}$$

the expectation value is

$$\begin{aligned} \langle \boldsymbol{k}, \alpha | A | \boldsymbol{k}, \alpha\rangle = \frac{1}{2} \Big( & \left\langle \boldsymbol{k}, \tfrac{1}{2}\right| A \left|\boldsymbol{k}, \tfrac{1}{2}\right\rangle + \left\langle \boldsymbol{k}, -\tfrac{1}{2}\right| A \left|\boldsymbol{k}, \tfrac{1}{2}\right\rangle \\ & + \left\langle \boldsymbol{k}, \tfrac{1}{2}\right| A \left|\boldsymbol{k}, -\tfrac{1}{2}\right\rangle + \left\langle \boldsymbol{k}, -\tfrac{1}{2}\right| A \left|\boldsymbol{k}, -\tfrac{1}{2}\right\rangle \Big) . \end{aligned} \tag{6.30}$$

The system described by $|\boldsymbol{k}, \alpha\rangle$ is prepared completely and can be represented by a single wave function. If one expands the wave function with respect to the basis $\left|\boldsymbol{k}, \tfrac{1}{2}\right\rangle$, $\left|\boldsymbol{k}, -\tfrac{1}{2}\right\rangle$ the expectation value also contains the mixed terms $\left\langle \boldsymbol{k}, \pm\tfrac{1}{2}\right| A \left|\boldsymbol{k}, \mp\tfrac{1}{2}\right\rangle$, which do not appear in (6.25).

Depending on the physical preparation of the system one therefore gets a different expectation value.

- Consider $N$ different realizations $\{\boldsymbol{x}_i,\ i = 1, \ldots, N\}$ of a random vector $\boldsymbol{X}$ in $\mathbb{R}^M$ with $\langle \boldsymbol{X}\rangle = 0$.

  We may introduce a density matrix also in a finite dimensional vector space $\mathbb{R}^M$. Let $\{\boldsymbol{x}_i,\ i = 1, \ldots, N\}$ be the possible realizations of a random vector $\boldsymbol{X}$ in $\mathbb{R}^M$ with $\langle \boldsymbol{X}\rangle = 0$, and each may be realized with probability $1/N$. Then a density matrix in $\mathbb{R}^M$ may be defined by:

$$\varrho = \sum_{i=1}^{N} \frac{1}{N} \boldsymbol{x}_i \otimes \boldsymbol{x}_i \ , \tag{6.31}$$

with the matrix elements

$$\varrho_{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} x_{i\alpha}\, x_{i\beta} \ , \tag{6.32}$$

where $x_{i\alpha}$ is the $\alpha$-component of $\boldsymbol{X}$ in the $i$th realization.

$\boldsymbol{\varrho}$ is also the realization of a well-known estimator for the covariance matrix (see e.g. Honerkamp 1994).

Hence, we see that the density matrix corresponds in this case to the covariance matrix of a random vector $\boldsymbol{X}$ with a probability density $\varrho_X(\boldsymbol{x})$. (More precisely it corresponds to the estimator of this matrix.)

Thus one could take the following point of view: The actual random variable in the Hilbert space is the vector $|\psi\rangle$, and one should rather define a probability density $\mathcal{P}(\psi)$. In some recent publications this has indeed been done; see, e.g., Breuer and Petruccione (1995).

Given the Hamiltonian $H$ of a quantum mechanically system, the density operators of the various statistical systems are:

1. For the canonical system

$$\varrho = \frac{1}{Z}\, \mathrm{e}^{-\beta H} \qquad \text{with} \qquad Z = \mathrm{Tr}\, \mathrm{e}^{-\beta H} . \tag{6.33}$$

If $\{|\alpha\rangle, \alpha = 1, \ldots\}$ is a complete system of eigenvectors for a complete set of commuting observables containing the Hamiltonian operator $H$, then

$$\varrho = \sum_{\alpha} \frac{1}{Z}\, \mathrm{e}^{-\beta E_\alpha} |\alpha\rangle\langle\alpha| \ , \tag{6.34}$$

where $H|\alpha\rangle = E_\alpha|\alpha\rangle$. $\frac{1}{Z}\,\mathrm{e}^{-\beta E_\alpha}$ is called the Boltzmann factor.

2. For the $p$–$T$ system

$$\varrho = \frac{1}{Y'}\, \mathrm{e}^{-\beta(H+pV)} \qquad \text{with} \qquad Y' = \mathrm{Tr}\, \mathrm{e}^{-\beta(H+pV)} . \tag{6.35}$$

3. For the grand canonical system

$$\varrho = \frac{1}{Y}\, \mathrm{e}^{-\beta(H-\mu N)} \qquad \text{with} \qquad Y = \mathrm{Tr}\, \mathrm{e}^{-\beta(H-\mu N)} . \tag{6.36}$$

4. For the microcanonical system

$$\varrho = \frac{1}{\Omega(E)}\, \delta_\Delta(H-E) \qquad \text{with} \qquad \Omega(E) = \mathrm{Tr}\, \delta_\Delta(H-E) \ , \tag{6.37}$$

where $\delta_\Delta(H - E)$ is the projection operator onto the subspace of the Hilbert space spanned by the eigenfunctions of $H$ with energy eigenvalues in the interval $(E, E + \Delta E)$. For the above mentioned complete system of eigenvectors $\{|\alpha\rangle, \alpha = 1, \ldots\}$ one obtains, e.g.,

$$\varrho = \frac{1}{\Omega(E)} \sum_{\alpha E \leq E_\alpha \leq E + \Delta E} |\alpha\rangle\langle\alpha| . \tag{6.38}$$

These statements about the density operators may be justified in various ways. The form of the density operator for the microcanonical system is postulated with the same reasoning as in the classical case: All allowed energy eigenstates are equally probable. The other density operators can then be derived from this one. Equivalently, one may also fomulate the maximum entropy principle for the entropies of density operators and then derive the density operators from this principle. And, finally, one can consider the classical limit and show that the probability densities introduced in the previous sections result in this limit.

We note that all relations between the expectation values and the partition functions have the same form as in the classical case, for example,

- In the canonical system (cf. (3.55))

$$\begin{aligned} E(T, N, V) &= \langle H \rangle = \mathrm{Tr}\,(\varrho H) \\ &= \frac{1}{Z}\left(-\frac{\partial}{\partial\beta}\right) Z = -\frac{\partial}{\partial\beta} \ln Z(T, N, V)\ , \end{aligned} \tag{6.39}$$

- In the $p$–$T$ system (cf. (3.90))

$$\begin{aligned} V(T, p, N) &= \langle V \rangle = \mathrm{Tr}\,(\varrho V) \\ &= \frac{1}{Y'}\left(-\frac{1}{\beta}\right) \frac{\partial}{\partial p} Y' = \frac{\partial}{\partial p} G(T, p, N)\ , \end{aligned} \tag{6.40}$$

- And in the macrocanonical system (cf. (3.108))

$$\begin{aligned} N(T, V, \mu) &= \langle N \rangle = \mathrm{Tr}\,(\varrho N) \\ &= \frac{1}{Y}\frac{1}{\beta}\frac{\partial}{\partial\mu} Y = -\frac{\partial}{\partial\mu} K(T, V, \mu) . \end{aligned} \tag{6.41}$$

## 6.2 Ideal Quantum Systems: General Considerations

Having established some basic notions for statistical quantum systems in the previous section, we will now determine the thermodynamic potential for a general quantum-mechanical ideal gas and examine the classical approximation.

We consider an ideal gas of $N$ particles in a volume $V$. If these particles do not possess any internal degrees of freedom, the Hamiltonian reads

$$H = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m}. \tag{6.42}$$

The quantum-mechanical $N$-particle state may be written as the $N$-fold product of single-particle states, where this product has to be totally symmetrized for bosons and totally antisymmetrized for fermions.

Let $|\alpha\rangle$ be a complete orthonormal system of energy eigenstates for the single-particle Hamiltonian $H_1 = \frac{\boldsymbol{p}^2}{2m}$ such that

$$H_1|\alpha\rangle = \varepsilon_\alpha|\alpha\rangle\ , \quad \alpha = 1, \ldots . \tag{6.43}$$

Thus each particle can be in one of these single-particle states. Its contribution to the total energy then is $\varepsilon_\alpha$. We sometimes refer to the single-particle states as orbitals.

If the volume is a box of linear size $L$ in each direction, $\alpha$ corresponds to a momentum $\boldsymbol{p}$ or a wave vector $\boldsymbol{k}$, and a spin quantum number $m_s$. In this case

$$H_1\,|\boldsymbol{k}, m_s\rangle = \frac{\hbar^2 k^2}{2m}\,|\boldsymbol{k}, m_s\rangle \tag{6.44}$$

i.e.,

$$\varepsilon_\alpha \equiv \varepsilon(\boldsymbol{k}, m_s) = \frac{\hbar^2 k^2}{2m} \equiv \varepsilon(k)\ , \tag{6.45}$$

and

$$S_z\,|\boldsymbol{k}, m_s\rangle = m_s\,|\boldsymbol{k}, m_s\rangle\ , \tag{6.46}$$

where $S_z$ denotes the $z$-component of the spin operator. The real space wave function is

$$\psi_{\boldsymbol{k},m_s}(\boldsymbol{r}) = C\ \sin(k_1x_1)\sin(k_2x_2)\sin(k_3x_3)\ , \tag{6.47}$$

where $\boldsymbol{k} = (k_1, k_2, k_3)$, $\boldsymbol{r} = (x_1, x_2, x_3)$, $\quad 0 \le x_i \le L$. The boundary condition $\psi = 0$ on the walls of the box yields

$$k_iL = m_i\pi, \quad \text{with} \quad m_i = 1, 2, \ldots\ , \quad i = 1, 2, 3\ , \tag{6.48}$$

i.e.,

$$k_i = \frac{m_i\pi}{L}, \quad m_i = 1, 2, \ldots\ , \quad i = 1, 2, 3, \tag{6.49}$$

where the $\{m_i\}$ should not be confused with the spin quantum number $m_s$.

The $N$-particle state may also be labeled by the number of particles which are in each of the single-particle states $|\alpha\rangle$ , $\alpha = 1, \ldots$. Let this number be $n_\alpha$, then $|n_1, \ldots\rangle$ characterizes the $N$-particle state completely: For each single-particle state $\alpha = 1, \ldots$ the number of particles which are in this state has been specified. If the particles are fermions, only the values $n_\alpha = 0$ or 1 are possible, since, according to the Pauli principle, there can be at most one particle in each orbit. For bosons $n_\alpha$ can take on all integer values larger than or equal to 0. Furthermore, we have

$$N = \sum_\alpha n_\alpha , \tag{6.50}$$

and for the total internal energy

$$E = \sum_\alpha n_\alpha \varepsilon_\alpha \ . \tag{6.51}$$

A complete system of energy eigenstates of the $N$-particle system is therefore given by all $\{(n_1, n_2, \ldots)\}$ such that $\sum_{\alpha=1} n_\alpha = N$. This representation is also called occupation number representation.

Let us determine the grand canonical partition function of a macroscopic system. Because all the $N$-particle states are eigenfunctions to the energy and particle number operator, we have

$$Y(T, V, \mu) = \sum_{N=0}^{\infty} \sum_{\{(n_1,\ldots)\}\sum n_\alpha = N} \exp -\beta((\varepsilon_1 - \mu)n_1 + (\varepsilon_2 - \mu)n_2 + \ldots) \tag{6.52}$$

$$= \sum_{\{(n_1,\ldots)\}} \exp -\beta((\varepsilon_1 - \mu)n_1 + \ldots). \tag{6.53}$$

Rearranging the sums we obtain

$$Y(T, V, \mu) = \sum_{n_1} \exp[-\beta(\varepsilon_1 - \mu)n_1] \sum_{n_2} \exp[-\beta(\varepsilon_2 - \mu)n_2] \ \ldots \tag{6.54}$$

$$= \prod_\alpha \sum_{n_\alpha} \exp[-\beta(\varepsilon_\alpha - \mu)n_\alpha]. \tag{6.55}$$

For the case of fermions the sum over occupation numbers yields

$$\sum_{n_\alpha=0}^{1} \mathrm{e}^{-\beta(\varepsilon_\alpha - \mu)n_\alpha} = 1 + \mathrm{e}^{\beta\mu}\,\mathrm{e}^{-\beta\varepsilon_\alpha} = 1 + z\,\mathrm{e}^{-\beta\varepsilon_\alpha} \ , \tag{6.56}$$

where we have introduced the so-called fugacity $z = \mathrm{e}^{\beta\mu}$.

For bosons we find

$$\sum_{n_\alpha=0}^{\infty} \mathrm{e}^{-\beta(\varepsilon_\alpha - \mu)n_\alpha} = \frac{1}{1 - \mathrm{e}^{-\beta(\varepsilon_\alpha - \mu)}} = \frac{1}{1 - z\,\mathrm{e}^{-\beta\varepsilon_\alpha}} \ , \tag{6.57}$$

where we have assumed that $(\varepsilon_\alpha - \mu) > 0$. Thus we must have $\mu < \varepsilon_0$, where $\varepsilon_0$ is the energy of the single-particle ground state. If we set this energy to zero, we have $\mu < 0$ for bosons.

We have represented the single-particle state $|\alpha\rangle$ by $|\boldsymbol{k}, m_s\rangle$. The spin quantum number $m_s$ can now take on $g$ possible values (e.g., $g = 2$ for spin 1/2, i.e., $m_s = \pm 1$). Then we find

$$\ln Y(T, V, \mu) = \pm g \sum_k \ln\left(1 \pm z\,\mathrm{e}^{-\beta\varepsilon(k)}\right) \ , \tag{6.58}$$

where the upper sign refers to fermions and the lower one to bosons. For large $L$ the sum over the discrete $k$ values can be approximated by an integral according to

$$\sum_{\boldsymbol{k}} \rightarrow \frac{V}{(2\pi\hbar)^3} \int \mathrm{d}^3 p. \tag{6.59}$$

(The sum over, say, $k_1$ is actually a sum over $m_1$, since $k_1 = \frac{m_1 \pi}{L}$, $m_1 = 1, 2, \ldots$. Therefore, as $m_1 = \frac{L}{\pi} k_1$, we have

$$\sum_{m_1=1}^{\infty} \rightarrow \frac{L}{\pi} \int_0^{\infty} \mathrm{d}k_1 = \frac{L}{2\pi} \int_{-\infty}^{+\infty} \mathrm{d}k_1 = \frac{L}{2\pi\hbar} \int_{-\infty}^{+\infty} \mathrm{d}p_1. \tag{6.60}$$

In Sect. 6.4 this approximation of a sum by an integral will be made more rigorous using the Euler–MacLaurin sum formula (6.124).)

Hence, we finally obtain for $K = -k_\mathrm{B} T \ln Y$

$$K(T, V, \mu) = \mp g k_\mathrm{B} T \, \frac{V}{(2\pi\hbar)^3} \int \mathrm{d}^3 p \, \ln\left[1 \pm z \exp\left(-\beta \frac{p^2}{2m}\right)\right]. \tag{6.61}$$

This is the grand canonical thermodynamic potential for the ideal quantum gas (for particles without internal degrees of freedom). Proceeding from this potential we will derive all other quantities. But first we will derive some relations between the thermodynamic potential $K$ and other system variables, and we will use a certain approximation to obtain the grand canonical thermodynamical potential for a classical ideal gas.

### 6.2.1 Expansion in the Classical Regime

We expand the integrand in $K(T, V, \mu)$ with respect to the fugacity $z$. Since $\ln(1 + x) = x - \frac{x^2}{2} + \ldots$ we get

$$K(T,V,\mu) = -gk_\mathrm{B}T\frac{V}{(2\pi\hbar)^3}\left(\int \mathrm{d}^3p \quad \mathrm{e}^{-\beta\frac{p^2}{2m}}\,z \right.$$
$$\left. \mp\frac{1}{2}\int \mathrm{d}^3p \quad \mathrm{e}^{-2\beta\frac{p^2}{2m}}\,z^2 + \ldots\right). \tag{6.62}$$

Using

$$\int \mathrm{d}^3p\,\mathrm{e}^{-\alpha p^2} = 4\pi\int_0^\infty \mathrm{d}p\; p^2\,\mathrm{e}^{-\alpha p^2} = 4\pi\frac{\sqrt{\pi}}{4}\,\alpha^{-3/2} = \pi^{3/2}\,\alpha^{-3/2}, \tag{6.63}$$

we finally obtain

$$K(T,V,\mu) = -gk_\mathrm{B}TV\left[z\left(\frac{2\pi mk_\mathrm{B}T}{h^2}\right)^{3/2}\right.$$
$$\left.\mp z^2\left(\frac{2\pi mk_\mathrm{B}T}{h^2}\right)^{3/2}\frac{1}{2^{5/2}} + \ldots\right], \tag{6.64}$$

and therefore to first order in $z = \mathrm{e}^{\beta\mu}$

$$K(T,V,\mu) = -gk_\mathrm{B}T\frac{V}{\lambda_t^3}\mathrm{e}^{\beta\mu}, \tag{6.65}$$

where we have again introduced the thermal de Broglie wavelength

$$\lambda_\mathrm{t} = \sqrt{\frac{h^2}{2\pi mk_\mathrm{B}T}}. \tag{6.66}$$

This is of the order of magnitude of the wavelength of a particle with energy $k_\mathrm{B}T$. (Because from $E = \hbar^2k^2/2m = k_\mathrm{B}T$ and $\lambda = 2\pi/k$ one obtains $\lambda_\mathrm{t}^2 = (2\pi)^2/k^2 \approx h^2/(2\pi mk_\mathrm{B}T)$.) For room temperature and $m = m_{^4\mathrm{He}}$ one finds $\lambda_\mathrm{t} \approx 10^{-8}$cm.

To within the factor $g$ (equal to the number of spin orientations) $K(T,V,\mu)$ in (6.65) is exactly the grand canonical thermodynamical potential for the classical ideal gas (cf. (3.112)).

Hence, to a first approximation in $z$, we get also

$$N = -\frac{\partial K}{\partial\mu} = g\mathrm{e}^{\beta\mu}V\lambda_\mathrm{t}^{-3}. \tag{6.67}$$

Introducing $n = N/V$, the number of particles per unit volume, and $v = V/N$, the volume per particle, we find again (see also (3.115))

$$z = \mathrm{e}^{\beta\mu} = n\frac{\lambda_\mathrm{t}^3}{g} = \frac{\lambda_\mathrm{t}^3}{gv}. \tag{6.68}$$

The condition $z \ll 1$ therefore means that the volume available for one particle is large compared to the volume of a box of side length equal to the de Broglie wavelength. Under these conditions the particle may obviously be treated as classical, and therefore $z \ll 1$ can also be considered as the classical regime.

For a gas at atmospheric pressure the particle number density $n$ is of the order $10^{19}$ atoms per cm$^3$. At room temperature one finds for such a gas, with $\lambda_{\mathrm{t}} = 10^{-8}\,\mathrm{cm}$,

$$z \propto n\lambda_{\mathrm{t}}^3 \approx 10^{19} \cdot 10^{-24} = 10^{-5} \ll 1. \tag{6.69}$$

### 6.2.2 First Quantum-Mechanical Correction Term

When we also take into account the second term in the expansion (6.64), we obtain

$$K(T, V, \mu) = -gk_{\mathrm{B}}TV\lambda_{\mathrm{t}}^{-3}\left(z \mp \frac{z^2}{2^{5/2}} + \ldots\right) \tag{6.70}$$

and therefore, as $N = -\frac{\partial K}{\partial \mu}$ and $\frac{\partial z}{\partial \mu} = \beta z$,

$$N(T, V, \mu) = gV\lambda_{\mathrm{t}}^{-3}\left(z \mp \frac{z^2}{2^{3/2}} + \ldots\right), \tag{6.71}$$

or, with

$$u := \frac{\lambda_{\mathrm{t}}^3 N}{gV} = \frac{\lambda_{\mathrm{t}}^3 n}{g} = O(n), \tag{6.72}$$

$$u = z \mp \frac{z^2}{2^{3/2}} + O(z^3). \tag{6.73}$$

Solving this equation for $z$ we get $z = u \pm \frac{u^2}{2^{3/2}} + O(u^3)$, and therefore

$$-K(T, V, \mu) = pV = k_{\mathrm{B}}TN\frac{1}{u}\left(z \mp \frac{z^2}{2^{5/2}} + O(z^3)\right) \tag{6.74}$$

$$= Nk_{\mathrm{B}}T\left(1 \pm u\left(\frac{1}{2^{3/2}} - \frac{1}{2^{5/2}}\right) + O(n^2)\right), \tag{6.75}$$

and finally

$$pV = Nk_{\mathrm{B}}T\left(1 \pm \frac{\lambda_{\mathrm{t}}^3}{g2^{5/2}}n + O(n^2)\right). \tag{6.76}$$

Hence, the first correction term to the equation of state of an ideal gas is different for bosons and fermions. For fermions a higher pressure results, which is a consequence of the Pauli principle: Multiple occupation of states is not permitted and this leads to an effective repulsion.

The equation of state (6.76) has the form

$$p = k_{\mathrm{B}} T n (1 + b(T)n + c(T)n^2 + \ldots), \tag{6.77}$$

identical to the form obtained in Sect. 3.7 in the framework of the virial expansion for nonideal gases.

### 6.2.3 Relations Between the Thermodynamic Potential and Other System Variables

From the expression (6.61) for the grand canonical potential $K$ one easily sees that again $K(T, V, \mu)$ is linear in $V$ (cp. (3.113)), and thus, as $p = -\frac{\partial K}{\partial V}$,

$$K = -p\,V. \tag{6.78}$$

Furthermore, for a nonrelativistic ideal gas, we will find

$$K = -\frac{2}{3}\,E. \tag{6.79}$$

For simplicity, we prove this only for the case of fermions. The proof for bosons is similar. The proof makes use of the energy–momentum relation for nonrelativistic particles. In Sect. 6.5 we will deal with the photon gas, which is a relativistic gas. The energy-momentum relation in that case reads $E = |\boldsymbol{p}|c$, and we will find a slightly different relation between $K$ and $E$:

*Proof.* We consider the integral in the expression for $K(T, V, \mu)$ and make the substitution $p^2 = x$, i.e., $2p\,\mathrm{d}p = \mathrm{d}x$, or $\mathrm{d}p = (2x^{1/2})^{-1}\mathrm{d}x$:

$$\int_0^\infty \mathrm{d}p\, p^2 \ln\left[1 + z\exp\left(-\beta\frac{p^2}{2m}\right)\right] = \frac{1}{2}\int_0^\infty \mathrm{d}x\, x^{1/2} \ln\left[1 + z\exp\left(-\beta\frac{x}{2m}\right)\right].$$

Partial integration yields:

$$\frac{1}{2}\frac{2}{3}x^{3/2}\ln\left[1 + z\exp\left(-\beta\frac{x}{2m}\right)\right]\Bigg|_0^\infty - \frac{1}{2}\frac{2}{3}\int \mathrm{d}x\, x^{3/2}\left[\frac{z\left(-\frac{\beta}{2m}\right)\exp\left(-\beta\frac{x}{2m}\right)}{1 + z\exp\left(-\beta\frac{x}{2m}\right)}\right].$$

Now $x^{3/2}\ln\left(1+ze^{-\beta\frac{x}{2m}}\right)\to 0$ for both limits $x\to 0$ and $x\to\infty$; resubstituting $p$ for $x$ the remaining integral can be written as

$$\frac{2}{3}\beta\int \mathrm{d}pp^2\,\frac{p^2}{2m}\left[\frac{z\exp\left(-\beta\frac{p^2}{2m}\right)}{1+z\exp\left(-\beta\frac{p^2}{2m}\right)}\right],$$

from which we finally obtain

$$K=-pV=-\frac{2}{3}\frac{gV}{(2\pi\hbar)^3}4\pi\int \mathrm{d}pp^2\,\frac{p^2}{2m}\,\frac{z\mathrm{e}^{-\beta\frac{p^2}{2m}}}{1+z\mathrm{e}^{-\beta\frac{p^2}{2m}}}=-\frac{2}{3}E,$$

as required. Together with $K=-p\,V$ we get the equation of state

$$p\,V=\frac{2}{3}E. \tag{6.80}$$

This is consistent with the equations for the classical ideal gas: $pV=Nk_{\mathrm{B}}T$, $E=\frac{3}{2}Nk_{\mathrm{B}}T$.

## 6.3 The Ideal Fermi Gas

We now consider the grand canonical potential for the ideal Fermi gas. In the previous section we found for this potential

$$K(T,V,\mu)=-gk_{\mathrm{B}}T\,\frac{V}{(2\pi\hbar)^3}\int \mathrm{d}^3p\,\ln\left[1+z\,\exp\left(-\beta\frac{p^2}{2m}\right)\right]. \tag{6.81}$$

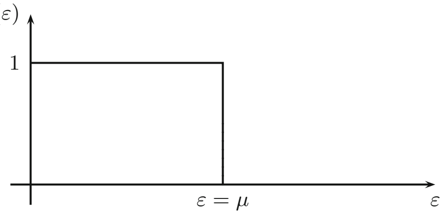
### *6.3.1 The Fermi–Dirac Distribution*

For the expectation values of the particle number and the energy we obtain

$$\begin{aligned}N(T,V,\mu)=\langle N\rangle&=-\frac{\partial K(T,V,\mu)}{\partial\mu}=-\frac{\partial z}{\partial\mu}\,\frac{\partial K(T,V,\mu)}{\partial z}\\&=\frac{gV}{(2\pi\hbar)^3}\int \mathrm{d}^3p\,\left[\frac{z\,\exp\left(-\beta\frac{p^2}{2m}\right)}{1+z\,\exp\left(-\beta\frac{p^2}{2m}\right)}\right],\end{aligned} \tag{6.82}$$

and, since $\dfrac{1}{Y}\dfrac{\partial}{\partial\beta}Y=-\langle H-\mu N\rangle$,

**Fig. 6.1** The Fermi–Dirac distribution at $T = 0$



$$E(T, V, \mu) = \langle H \rangle = -\frac{\partial}{\partial \beta} \ln Y + \mu N = -\frac{\partial}{\partial \beta} (-\beta K) + \mu N$$

$$= \frac{gV}{(2\pi\hbar)^3} \int \mathrm{d}^3 p \, \frac{p^2}{2m} \left[ \frac{z \exp\left(-\beta \frac{p^2}{2m}\right)}{1 + z \exp\left(-\beta \frac{p^2}{2m}\right)} \right]. \qquad (6.83)$$

Both expressions contain the quantity

$$\langle n_{p,m_s} \rangle := \frac{z \exp\left(-\beta \frac{p^2}{2m}\right)}{1 + z \exp\left(-\beta \frac{p^2}{2m}\right)}, \qquad (6.84)$$

which is called the average occupation number of the orbital $(\boldsymbol{k}, m_s)$ or $(\boldsymbol{p}, m_s)$. Thus

$$\langle N \rangle = \sum_{p,m_s} \langle n_{p,m_s} \rangle \ , \quad \langle H \rangle = \sum_{p,m_s} \frac{p^2}{2m} \langle n_{p,m_s} \rangle. \qquad (6.85)$$

This average occupation number $\langle n_{p,m_s} \rangle$ for an orbital, considered as a function $f(\varepsilon)$ of the energy $\varepsilon = \frac{p^2}{2m}$ of the orbital, is also called the Fermi–Dirac distribution.

We will now discuss this Fermi–Dirac distribution, written as

$$f(\varepsilon) = \frac{1}{1 + \mathrm{e}^{\beta(\varepsilon - \mu)}},$$

in more detail:

We have $0 \leq f(\varepsilon) \leq 1$, since $\mathrm{e}^{\beta(\varepsilon-\mu)}$ takes on values between 0 and $\infty$.

We consider $\mu > 0$ and examine the limit $T \to 0$ or $\beta = \frac{1}{k_{\mathrm{B}} T} \to \infty$. For $\varepsilon > \mu$ we obtain $f(\varepsilon) \to 0$, for $\varepsilon < \mu$ we get $f(\varepsilon) \to 1$. Hence, for $f(\varepsilon)$ as a function of $\varepsilon$ we find the dependence shown in Fig. 6.1. This means that at $T = 0$, all orbitals with an energy $\varepsilon(p) < \mu$ are occupied and all others are empty.

The orbitals for which $\varepsilon(p) < \mu$ are also called the orbitals of the Fermi sea, i.e., the Fermi sea is completely occupied at $T = 0$. The edge of the function $f(\varepsilon)$ at $\varepsilon = \mu$ is referred to as the Fermi edge. The surface in $\boldsymbol{k}$-space defined by $\hbar^2 k^2 / 2m = \mu$ is also called Fermi surface. For nonideal gases this surface is in general not spherical.

The Fermi momentum is the momentum $p_{\mathrm{F}}$ for which

$$\varepsilon(p_{\mathrm{F}}) = \mu, \tag{6.86}$$

and $\varepsilon_{\mathrm{F}} \equiv \varepsilon(p_{\mathrm{F}}) = \mu$ is called the Fermi energy. Finally, we define the Fermi temperature $T_{\mathrm{F}}$ through the equation

$$k_{\mathrm{B}} T_{\mathrm{F}} = \varepsilon_{\mathrm{F}}. \tag{6.87}$$

Fermi momentum and Fermi energy are determined from the equation for the particle number at $T = 0$:

$$N(T = 0, V, \mu) = g \frac{V}{(2\pi\hbar)^3} \int_{\varepsilon(p)<\mu} \mathrm{d}^3 p \tag{6.88}$$

$$= g \frac{V}{(2\pi\hbar)^3} 4\pi \int_0^{p_{\mathrm{F}}} \mathrm{d}p\, p^2 = \frac{gV}{(2\pi\hbar)^3}\, 4\pi\, \frac{p_{\mathrm{F}}^3}{3}\,. \tag{6.89}$$

So we obtain, with $v = V/N$ being the volume per particle,

$$p_{\mathrm{F}} = p_{\mathrm{F}}(v) = 2\pi\hbar \left( \frac{3N}{4\pi g V} \right)^{1/3} = \hbar \left( \frac{6\pi^2}{gv} \right)^{1/3} \tag{6.90}$$

and

$$\varepsilon_{\mathrm{F}} \equiv \varepsilon(p_{\mathrm{F}}) = \frac{p_{\mathrm{F}}^2}{2m} = \frac{\hbar^2}{2m} \left( \frac{6\pi^2}{gv} \right)^{2/3}. \tag{6.91}$$

Given $v$, the Fermi energy is the chemical potential at $T = 0$ and $\varepsilon_{\mathrm{F}} = \varepsilon_{\mathrm{F}}(v) \propto v^{-2/3}$.

The internal energy at $T = 0$ is

$$E(T = 0, V, \mu) = g \frac{V}{(2\pi\hbar)^3} \int_{\varepsilon(p)<\mu} \mathrm{d}^3 p\, \frac{p^2}{2m} \tag{6.92}$$

$$= \frac{gV}{(2\pi\hbar)^3} 4\pi \int_0^{p_{\mathrm{F}}} \mathrm{d}p\, \frac{p^4}{2m} = \frac{gV}{(2\pi\hbar)^3}\, 4\pi\, \frac{p_{\mathrm{F}}^5}{5}\, \frac{1}{2m}\,,$$

or

$$E(T = 0, V, \mu) = \frac{3}{5}\, N \varepsilon_{\mathrm{F}}(v). \tag{6.93}$$

Hence, the average energy per particle at $T = 0$ is $\frac{3}{5}\, \varepsilon_{\mathrm{F}}(v)$.

Both the Fermi temperature and the Fermi energy depend on $v$ or, equivalently, on the particle density $n$. For a classical gas of, say $^3$He atoms, with $n = 10^{19}\,\mathrm{cm}^{-3}$ we obtain $T_{\mathrm{F}} \approx \frac{1}{10}$ K. In this case the room temperature $T$ is much larger than the Fermi temperature: $T \gg T_{\mathrm{F}}$. This is obviously the classical regime. We will study cases where $T \ll T_{\mathrm{F}}$ in a moment.

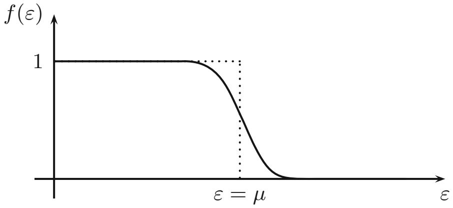


**Fig. 6.2** Fermi–Dirac distribution for $T > 0$ and $\mu > 0$ (*solid curve*) compared to $T = 0$ (*dotted line*)

In order to compare temperatures we may also introduce their ratio. A useful quantity is

$$\varrho_E \equiv \frac{\lambda_t^3}{gv} \approx \left(\frac{T_F}{T}\right)^{3/2} . \tag{6.94}$$

Thus in the classical regime, $\varrho_E$ is very small and $T \gg T_F$. The opposite limit corresponds to $\varrho_E > 1$ and therefore $T < T_F$. This regime is called the degenerate regime, and $\varrho_E$ is called the degree of degeneracy. At such temperatures the mean energies of the electrons is of the order of $\varepsilon_F$, that means, that mainly the lower energy levels are occupied.

For $T > 0$ the Fermi edge softens. If still $\mu > 0$, typically $f(\varepsilon)$ has the form shown in Fig. 6.2. There are now unoccupied orbitals in the Fermi sea, known as holes. On the other hand, orbitals above the Fermi surface are occupied. Each excitation of the ground state means that an electron is taken from an orbital below the Fermi energy into an orbital above the Fermi energy. This corresponds to the creation of a hole below the Fermi surface and an electron above the Fermi surface.

For $T > 0$ and $\mu < 0$ the Fermi–Dirac distribution is nearly zero for all $\varepsilon$. This is just the classical regime.

## 6.3.2 *Determination of the System Variables at Low Temperatures*

### The Chemical Potential

We consider the equation for $N(T, V, \mu)$ (cf. (6.82)):

$$N(T, V, \mu) = \frac{gV}{(2\pi\hbar)^3} \int \mathrm{d}^3 p \, \frac{1}{1 + z^{-1} \exp\left(\beta \frac{p^2}{2m}\right)} . \tag{6.95}$$

Using $p^2 = 2mk_B T x^2$ and $\lambda_t = \left(h^2/2\pi m k_B T\right)^{1/2}$ we also obtain

$$\frac{\lambda_t^3}{gv} = \frac{4}{\sqrt{\pi}} \int_0^\infty \mathrm{d}x \, \frac{x^2}{1 + z^{-1} \exp x^2} =: f_{\frac{3}{2}}(z), \tag{6.96}$$

where we have defined the function $f_{\frac{3}{2}}(z)$.

The classical approximation corresponds to $z \ll 1$, i.e., $z^{-1} \gg 1$, and to first order in $z$,

$$f_{\frac{3}{2}}(z) = \frac{4}{\sqrt{\pi}} \int_0^\infty \mathrm{d}x\, x^2\, \mathrm{e}^{-x^2} z = z\,, \tag{6.97}$$

from which we again obtain

$$\frac{\lambda_{\mathrm{t}}^3}{gv} = z \ll 1\,. \tag{6.98}$$

Let us now examine the case $z \gg 1$.

Since $\ln z = \beta\mu = \mu/k_{\mathrm{B}}T$, large values of $z$ imply small temperatures, and $T \to 0$ is equivalent to $z \to \infty$. Using a method of Sommerfeld (Huang 1987), for large values of $z$ one obtains the following expansion of $f_{\frac{3}{2}}(z)$:

$$f_{\frac{3}{2}}(z) = \frac{4}{3\sqrt{\pi}} \left( (\ln z)^{3/2} + \frac{\pi^2}{8} (\ln z)^{-1/2} + \ldots \right) + O(z^{-1}). \tag{6.99}$$

As $\ln z = \beta\mu = \frac{\mu}{k_{\mathrm{B}}T}$, we therefore get

$$\frac{\lambda_{\mathrm{t}}^3}{gv} = \frac{4}{3\sqrt{\pi}} \left( \left( \frac{\mu}{k_{\mathrm{B}}T} \right)^{3/2} + \frac{\pi^2}{8} \left( \frac{\mu}{k_{\mathrm{B}}T} \right)^{-1/2} + \ldots \right). \tag{6.100}$$

Let us solve this expansion for $\mu$ and determine $\mu = \mu(T, v)$. We know that in the limit $T \to 0$ we have $\mu = \varepsilon_{\mathrm{F}}$. We make the ansatz

$$\mu(T, v) = \varepsilon_{\mathrm{F}} \left( 1 + \alpha \left( \frac{k_{\mathrm{B}}T}{\varepsilon_{\mathrm{F}}} \right)^2 + \ldots \right) \tag{6.101}$$

and determine $\alpha$ by inserting this expression into (6.100) and comparing the terms of order $T^2$. We find $\alpha = -\frac{\pi^2}{12}$, and therefore, for the chemical potential of a Fermi gas, we obtain

$$\mu(T, v) = \varepsilon_{\mathrm{F}}(v) \left[ 1 - \frac{\pi^2}{12} \left( \frac{k_{\mathrm{B}}T}{\varepsilon_{\mathrm{F}}} \right)^2 + O(T^4) \right]. \tag{6.102}$$

Hence, $\mu(T, v)$ drops quadratically from a positive maximum value at $T = 0$. In the classical regime we found

$$\mu(T, v) = k_{\mathrm{B}}T \left[ \ln \left( \frac{N}{V} \right) - \frac{3}{2} \ln \left( \frac{2\pi m k_{\mathrm{B}}T}{h^2} \right) \right] \tag{6.103}$$

$$= -k_{\mathrm{B}}T \ln\left(\frac{v}{\lambda_{\mathrm{t}}^3}\right), \tag{6.104}$$

which is negative.

### The Internal Energy and the Specific Heat

The internal energy is first obtained as function of $T, V$, and $\mu$ (see (6.83)). Evaluating the integral with the previously mentioned method of Sommerfeld, one obtains the expansion of $E(T, V, \mu)$ for large $z = \exp(\beta\mu)$. Then inserting the expansion of $\mu(T, v)$, (6.102), we find $E(T, v, N)$ as

$$E(T, v, N) = \frac{3}{5}\, N\varepsilon_{\mathrm{F}}(v)\left[1 + \frac{5}{12}\pi^2\left(\frac{T}{T_{\mathrm{F}}}\right)^2 + O(T^4)\right], \tag{6.105}$$

and, therefore, for the specific heat,

$$C_V = \frac{\partial E(T, v, N)}{\partial T} = \frac{\pi^2}{2}\, Nk_{\mathrm{B}}\frac{T}{T_{\mathrm{F}}} + O\left((T/T_{\mathrm{F}})^3\right). \tag{6.106}$$

Hence, for $T \to 0$ the specific heat $C_V$ tends to zero, while for large temperatures, i.e., in the classical regime, $C_V = \frac{3}{2}k_{\mathrm{B}}N$. This result has a plausible explanation: When a Fermi gas is heated from $T = 0$ to the temperature $T$, the electrons that are excited to orbitals above the Fermi surface are mostly those whose energy is within a region $k_{\mathrm{B}}T$ below $\varepsilon_{\mathrm{F}}$. The number of these electrons is of the order $\frac{k_{\mathrm{B}}T}{\varepsilon_{\mathrm{F}}}N$, i.e., the total number multiplied by the ratio of the excitation energy to the Fermi energy. (For $k_{\mathrm{B}}T = \varepsilon_{\mathrm{F}}$, i.e., for $T = T_{\mathrm{F}}$, the number of electrons above the Fermi sea is of the same order as the number of electrons below the Fermi surface.) So the energy of the excitation is $\frac{k_{\mathrm{B}}T}{\varepsilon_{\mathrm{F}}}Nk_{\mathrm{B}}T = O(T^2)$, and therefore $C_V = O(T)$.

### The Pressure

The pressure is obtained from (6.80) and (6.105) as

$$p = \frac{2}{3}\frac{E}{V} = \frac{2}{5}\frac{\varepsilon_{\mathrm{F}}}{v}\left[1 + \frac{5\pi^2}{12}\left(\frac{T}{T_{\mathrm{F}}}\right)^2 + \ldots\right],$$

$$p \propto v^{-5/3}\ , \quad \text{i.e.,} \quad p \propto \left(\frac{N}{V}\right)^{5/3}. \tag{6.107}$$

As a consequence of the Pauli principle we find a zero-point pressure at $T = 0$.

### The Entropy

For the entropy,

$$S = \frac{1}{T}(E + pV - \mu N), \tag{6.108}$$

we obtain from (6.80), (6.105), and (6.102)

$$S = \frac{1}{T}\left(E + \frac{2}{3}\,E - \mu N\right) = N\,\varepsilon_{\mathrm{F}}\,\frac{\pi^2}{2}\,\frac{T}{T_{\mathrm{F}}^2} + O((T/T_{\mathrm{F}})^3). \tag{6.109}$$

So $S \to 0$ in the limit $T \to 0$. For $T = 0$ the system is in its ground state, i.e., all particles are in the Fermi sea. Therefore the entropy has to vanish in this limit.

### 6.3.3 Applications of the Fermi–Dirac Distribution

**Valence Electrons in a Metal**

Valence electrons in metallic atoms are very weakly bound and can propagate through the crystal lattice as quasi-free particles. These electrons may be considered as constituents of an electron gas. Alkali metals and also copper, silver, and gold contribute one electron per atom, the concentration of the electrons in the electron gas of these metals is therefore equal to the number of atoms, which is of the order

$$n = \frac{N}{V} = (1\text{–}10) \times 10^{22}\,\mathrm{cm}^{-3}. \tag{6.110}$$

For $g = 2$ we obtain

$$\varepsilon_{\mathrm{F}} = \frac{\hbar^2}{2m}\left(3\pi^2 n\right)^{2/3} = (2-8)\,\mathrm{eV} \tag{6.111}$$

and a Fermi temperature $T_{\mathrm{F}} = \varepsilon_{\mathrm{F}}/k_{\mathrm{B}}$:

$$T_{\mathrm{F}} = 20,000\text{–}80,000\,\mathrm{K}\,. \tag{6.112}$$

Hence, for a metal at room temperature we have $T \ll T_{\mathrm{F}}$, i.e., the electron gas in a metal is extremely degenerate.

(For the classical case we found $n = 10^{19}\,\mathrm{cm}^{-3}$ and $m$ was the mass of the He atom. For valence electrons the density is larger by a factor of $1,000$ and the mass is smaller by a factor of $4 \times 2,000$. Taking into account the exponent $2/3$ in the density this yields a factor $8 \times 10^5$ for $T_{\mathrm{F}}$. For the He gas we found $T_{\mathrm{F}} \approx 1/10\,\mathrm{K}$ and so we obtain 80,000 K for the valence electrons.)

Therefore we may apply the results for the degenerate Fermi gas. In particular, the heat capacity $C_V$ in metals at low temperature reveals a linear dependence on the temperature. In Sect. 6.5, dealing with phonons, we will see that lattice vibrations contribute a term of order $T^3$ to the heat capacity at low temperatures. Measuring $C_V$ as a function of $T$ and plotting $C_V/T$ against $T^2$ should lead to a straight line of the form

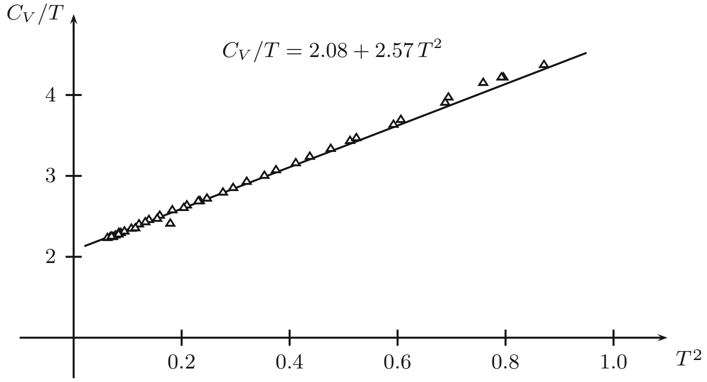


**Fig. 6.3** Experimental data for the heat capacity $C_V$ as a function of temperature $T$ for potassium (From Lien and Phillips 1964)

$$\frac{C_V}{T} = \gamma + AT^2, \tag{6.113}$$

where $\gamma$ should agree with the value calculated in (6.106). Indeed, such a straight line is found (Fig. 6.3) and in most cases $\gamma$ agrees with the theoretical value to within 20%. Typical values for $\gamma$ are 1 mJ/mol K$^2$.

In 1928 Sommerfeld laid the foundations of the modern theory of metals with basically such ideas in mind. At the beginning, however, it remained surprising that the ideal Fermi gas represents such an excellent model for the electrons in a metal, since one would expect the Coulomb repulsion among the electrons to be too large to be negligible.

As we know today, the long-range interaction among the electrons is screened by the polarization of the atomic cores. Thus, if the Coulomb interaction is taken into account, one may within a reasonable approximation introduce so-called quasi-particles, which correspond to the naked electrons including their polarization cloud. These quasi-particles behave like free particles, but they possess an effective mass $m^* \neq m_e$ (Sect. 6.9).

### Electrons in White Dwarf Stars

White dwarf stars are stars with a mass $M$ comparable to that of the sun, but with a radius $R$ which is closer to that of the earth. While for the sun $M = 2 \times 10^{33}$ g and $R = 7 \times 10^{10}$ cm, from which one obtains a density of about 1 g cm$^{-3}$, one finds for a typical white dwarf star like, e.g. Sirius B, the companion of Sirius,

$$M \approx 2 \times 10^{33}\,\text{g}, \qquad R \approx 2 \times 10^9\,\text{cm}\ (= 20{,}000\,\text{km}), \tag{6.114}$$

and for the density

$$\varrho = \frac{M}{V} = \frac{2 \times 10^{33}\,\text{g}}{\frac{4\pi}{3}\,(2 \times 10^{9}\,\text{cm})^3} \approx 0.7 \times 10^5\,\text{g}\,\text{cm}^{-3}\,. \tag{6.115}$$

Thus 1 $\text{cm}^3$ of this matter would weight approximately 1 tonne on earth. From this mass density we can deduce the particle density. one mole of hydrogen, for instance, corresponds to 1 g and contains $6 \times 10^{23}$ particles. Hence, if 1 $\text{cm}^3$ has a weight of $10^6$ g, it contains about $6 \times 10^{29}$ particles, which corresponds to a density of $n \approx 10^{30}\,\text{cm}^{-3}$.

The volume per particle is therefore $10^{-30}\,\text{cm}^3$, i.e., in the average one particle occupies a box with linear extension $10^{-10}\,\text{cm} = 0.01\,\text{Å}$. At such large densities the atomic structure is dissolved, the atoms are completely ionized and the electrons form a gas.

The density of this electron gas is larger by a factor of $10^8$ than the electron gas in a metal. Therefore, the Fermi energy is larger by a factor of $10^5$–$10^6$, i.e.,

$$\varepsilon_{\text{F}} = 10^5\text{–}10^6\,\text{eV}, \tag{6.116}$$

and, as $1\,\text{eV} \approx 10^4\,\text{K}$, we find for the Fermi temperature

$$T_{\text{F}} = 10^9\text{–}10^{10}\,\text{K}. \tag{6.117}$$

Of what we know from observations and from the theory of stars, the temperature inside a white dwarf star is of the order $10^7$ K, so again $T \ll T_{\text{F}}$, i.e., the electron gas in white dwarf stars is degenerate.

The pressure of the electron gas has to be compensated by the gravitational force among the He nuclei. However, if the mass of a star is too large the Fermi pressure cannot prevent a gravitational collapse. The limit at which this occurs is close to 1.4 times the solar mass. This is also called the Chandrasekhar limit, named after the Indo-american physicist Chandrasekhar.

White dwarf stars are quite common; there are about 0.001 per $(\text{light-year})^3$. On the average one would therefore expect one white dwarf star within every 10 light-years. The distance to Sirius B is just 8 light-years.

The nucleons in a white dwarf star also form a gas. However, since the mass of the nucleons is larger by a factor 2,000, the Fermi temperature $T_{\text{F}}$ for this gas is smaller by a factor of 2,000. Hence, for the nucleon gas in a white dwarf star we do not have $T \ll T_{\text{F}}$.

### Neutrons in a Neutron Star

Neutron stars have a density which is $10^9$ times larger than that of white dwarf stars, i.e., $n \approx 10^{39}\,\text{cm}^{-3}$ and $T_{\text{F}} \approx 10^{12}$ K. Under these conditions also atomic nuclei are dissociated and the neutrons form a degenerate Fermi gas.

## 6.4 The Ideal Bose Gas

The grand canonical thermodynamic potential for the ideal Bose gas is (cf. (6.58))

$$K(T,V,\mu) = gk_{\mathrm{B}}T\sum_{\boldsymbol{k}} \ln\left(1 - z\,\mathrm{e}^{-\beta\varepsilon(k)}\right). \tag{6.118}$$

From this we find for $N(T,V,\mu)$

$$N(T,V,\mu) = g\sum_{\boldsymbol{k}} \frac{z\,\mathrm{e}^{-\beta\varepsilon(k)}}{1 - z\,\mathrm{e}^{-\beta\varepsilon(k)}} = g\sum_{\boldsymbol{k}} \frac{1}{\mathrm{e}^{\beta(\varepsilon(k)-\mu)} - 1}. \tag{6.119}$$

### *6.4.1 Particle Number and the Bose–Einstein Distribution*

As we did in the Fermi case we introduce the average occupation number of the orbital $\boldsymbol{k}, m_s$. Now

$$\langle n_{\boldsymbol{k},m_s}\rangle = \frac{1}{\mathrm{e}^{\beta\left(\varepsilon(k)-\mu\right)} - 1}\,. \tag{6.120}$$

This quantity, considered as a function of $\varepsilon$,

$$f(\varepsilon) = \frac{1}{\mathrm{e}^{\beta(\varepsilon-\mu)} - 1}\,, \tag{6.121}$$

is called the Bose–Einstein distribution. Here $\varepsilon \geq \varepsilon_0$, where $\varepsilon_0$ is the ground state energy of the single-particle states. According to (6.48) this corresponds to $(m_1, m_2, m_3) = (1,1,1)$, i.e., it is nondegenerate. (Since only energy differences $\varepsilon - \mu$ are relevant, the ground state energy may be shifted such that $\varepsilon_0 = 0$.)

If we approximate the sum over the state labels $k$ in (6.119) by an integral, as we did in the previous sections, we obtain (cf. (6.59) and (6.60))

$$N(T,V,\mu) = \frac{gV}{(2\pi\hbar)^3}\int \mathrm{d}^3p\,\frac{1}{\mathrm{e}^{\beta\varepsilon(p)}\,z^{-1} - 1}\,. \tag{6.122}$$

We will see, however, that this approximation is now incorrect. Instead we have to treat the first term in the sum over states (6.119), which is the mean number of particles in the ground state, separately; the rest may then be replaced by an integral. Thus we obtain

$$N(T,V,\mu) = \frac{g}{\mathrm{e}^{\beta(\varepsilon_0-\mu)} - 1} + \frac{gV}{(2\pi\hbar)^3}\int \mathrm{d}^3p\,\frac{1}{\mathrm{e}^{\beta\varepsilon(p)}\,z^{-1} - 1}. \tag{6.123}$$

We will show that this is the proper replacement by using the Euler–MacLaurin sum formula. Let $f(n)$ be a function defined on the integers between $n = a$ and $n = b$, admitting a continuous extension to a function $f(x)$ on the real numbers. Then (see, e.g., Graham et al. 1994)

$$\sum_{n=a}^{n=b} f(n) = \int_a^b f(x)\mathrm{d}x + \frac{1}{2}(f(a) + f(b)) + \sum_{j=1}^{\infty}(-1)^j \frac{B_{2j}}{(2j)!}(f^{(2j-1)}(a) - f^{(2j-1)}(b)), \tag{6.124}$$

where $B_j$ are the Bernoulli numbers, $B_2 = 1/6$, $B_4 = -1/30, \ldots$, and $f^{(a)}(x)$ denotes the $a$-th derivative of $f(x)$. This is the Euler–MacLaurin sum formula. All terms on the right hand side apart from the integral may be neglected if they are much smaller than the integral. In general, this will be the case, as a single summand is usually not of the order of the total sum of all summands.

We now apply this formula to the sums in (6.119). The $k_1$-summation, for example, is actually a sum over $m_1$, since $k_1 = m_1\pi/L$ , $m_1 = 1, 2, \ldots$, and similarly for the $k_2$- and $k_3$-summations. Hence, the integral is of the order $L^3 = V \propto N$ (cf. (6.60)). The additional terms in (6.124) may therefore be neglected if they are smaller than $O(N)$. This was indeed the case in (6.59). For the summation in (6.119), however, we find:

The first term in the sum (6.119) may possibly be of the order $N$:
In the derivation of the thermodynamic potential in Sect. 6.2 we already noticed that for bosons $\mu < \varepsilon_0$ always holds, because otherwise the infinite sum over the occupation numbers does not converge. For any $\beta$ the value of $f(\varepsilon_0)$ can become arbitrarily large if $\mu$ comes close enough to $\varepsilon_0$. Hence, we may have $\beta(\varepsilon_0 - \mu) \ll 1$ and therefore also

$$f(\varepsilon_0) = \frac{1}{\mathrm{e}^{\beta(\varepsilon_0-\mu)} - 1} \approx \frac{1}{\beta(\varepsilon_0 - \mu)} \gg 1\ . \tag{6.125}$$

$f(\varepsilon_0)$ may even be of order $N$, namely when $\beta(\varepsilon_0 - \mu) = O(N^{-1})$. Since

$$f(\varepsilon_0) = \frac{z}{\mathrm{e}^{\beta\varepsilon_0} - z} = \frac{z}{1-z} \quad \text{and} \quad f(\varepsilon_0) = O(N) \tag{6.126}$$

this implies that the variable $z = \mathrm{e}^{\beta\mu}$ has to be close to 1 such that $1-z = O(N^{-1})$.

These findings reflect the possibility that, in the case of bosons, all particles may be in the lowest orbital. There is no exclusion principle as in the case of fermions. For each temperature we will therefore find such a region characterized by $\beta(\varepsilon_0 - \mu) = O(N^{-1})$.

The second term in the sum (6.119) can be at most of the order $N^{2/3}$:
To see how large the quantity

$$f(\varepsilon_1) = \frac{1}{e^{\beta(\varepsilon_1 - \mu)} - 1} \tag{6.127}$$

might be, we choose $\beta(\varepsilon_1 - \mu) \ll 1$, which allows us to make the approximation

$$f(\varepsilon_1) = \frac{1}{e^{\beta(\varepsilon_1 - \mu)} - 1} \approx \frac{1}{\beta(\varepsilon_1 - \mu)}, \tag{6.128}$$

and since

$$\beta(\varepsilon_1 - \mu) = \beta(\varepsilon_0 - \mu) + \beta(\varepsilon_1 - \varepsilon_0) \tag{6.129}$$

and $\varepsilon \propto p^2$, $|p| \propto \frac{1}{L} \propto \frac{1}{V^{1/3}}$ and $V = vN$, we finally get

$$(\varepsilon_1 - \varepsilon_0) = O(N^{-2/3}). \tag{6.130}$$

So $\varepsilon_1 - \mu$ can never become smaller than a term of order $N^{-2/3}$ and therefore $f(\varepsilon_1)$ never be larger than a term of order $N^{2/3}$.

We have thus seen that the first term has to be treated separately, whereas the summation over all other terms can be approximated by an integral. The lower limit of this integral is the value of $p$, which corresponds to the first excited state. However, the difference compared to the complete integration over all $p$ is of lower order in $N$ than $N$ itself. Therefore (6.123) is the correct approximation.

### 6.4.2 Bose–Einstein Condensation

Let us now examine the expression for $N(T, V, \mu)$. We set $\varepsilon_0 = 0$ and obtain from (6.123)

$$N = \frac{gz}{1-z} + \int_0^\infty \mathrm{d}p\, p^2 4\pi \frac{gV}{(2\pi\hbar)^3} \frac{1}{z^{-1}\, e^{\beta\varepsilon(p)} - 1}. \tag{6.131}$$

Making the substitution $p^2 = 2mk_B T x^2$ we encounter an integral similar to the one in the corresponding equation for $N(T, V, \mu)$ for the ideal Fermi gas. There we introduced the function $f_{\frac{3}{2}}(z)$. Now we define

$$\begin{aligned} h_{\frac{3}{2}}(z) = -f_{\frac{3}{2}}(-z) &= \frac{4}{\sqrt{\pi}} \int_0^\infty \mathrm{d}x\, \frac{x^2}{z^{-1}\, e^{x^2} - 1} \\ &\equiv \frac{4}{\sqrt{\pi}} \int_0^\infty \mathrm{d}x\, x^2 \frac{z\, e^{-x^2}}{1 - z\, e^{-x^2}}\,. \end{aligned} \tag{6.132}$$

Multiplying by $\lambda_\mathrm{t}^3/(gV)$ we can write (6.131) in the form

$$\frac{\lambda_\mathrm{t}^3}{gv} = \frac{\lambda_\mathrm{t}^3}{v}\frac{1}{N}\frac{z}{1-z} + h_{\frac{3}{2}}(z). \tag{6.133}$$

So we again find that, if $z/(1-z) \neq O(N)$, i.e., if $1-z$ is not of the order $\left(N^{-1}\right)$, we may neglect the first term in (6.133). Otherwise it yields a contribution of the order $\lambda_\mathrm{t}^3/v$, as does the term on the left-hand side.

We now want to solve this equation for $z$ to obtain $z = z(T, V, N)$ and thereby $\mu = \mu(T, V, N)$. For this we have to know something about the function $h_{\frac{3}{2}}(z)$. We observe:

- Expanding the denominator in the integral of (6.132) we obtain for $h_{\frac{3}{2}}(z)$:

$$h_{\frac{3}{2}}(z) = \sum_{j=1}^{\infty} \frac{z^j}{j^{3/2}}\ . \tag{6.134}$$

- $h_{\frac{3}{2}}(z)$ increases monotonously with a maximum at $z = 1$. At this point $h_{\frac{3}{2}}(1) = \zeta\left(\frac{3}{2}\right) = 2.612$, where $\zeta(x)$ is the Riemann zeta function defined by

$$\zeta(x) = \sum_{j=1}^{\infty} \frac{1}{j^x}\ . \tag{6.135}$$

Again we notice the relevance of the first term $\propto z/(1-z)$ in (6.133). Without this term we would always have to conclude that

$$\frac{\lambda_\mathrm{t}^3}{gv} \leq h_{\frac{3}{2}}(1) = \zeta\left(\frac{3}{2}\right),$$

which is obviously an unreasonable result. Instead we may now conclude:

- If $\lambda_\mathrm{t}^3/(gv) > h_{\frac{3}{2}}(1) = \zeta\left(\frac{3}{2}\right)$, (6.133) implies immediately that

$$\frac{\lambda_\mathrm{t}^3}{V}\frac{z}{1-z} \tag{6.136}$$

  has to supply a nonvanishing contribution, i.e., $z/(1-z)$ has to be of order $N$ and therefore $z$ close to 1, i.e., $1 - z = O\left(\frac{1}{N}\right)$.
- If $\lambda_\mathrm{t}^3/(gv) < h_{\frac{3}{2}}(1) = \zeta\left(\frac{3}{2}\right)$, (6.133) may be solved for $z$ even without the term $(\lambda_\mathrm{t}^3/V)z/(1-z)$. This will yield a value for $z$ such that $1-z$ is not $O\left(N^{-1}\right)$, and the term $(\lambda_\mathrm{t}^3/V)z/(1-z)$ may now be neglected compared to $\lambda_\mathrm{t}^3/(gv)$ and $h_{\frac{3}{2}}(z)$.

The limit case is determined by the equation

$$\frac{\lambda_{\mathrm{t}}^3}{gv} = \zeta\left(\frac{3}{2}\right), \qquad \text{i.e.,} \quad \left(\frac{h^2}{2\pi m k_{\mathrm{B}} T}\right)^{3/2} = gv\,\zeta\left(\frac{3}{2}\right). \tag{6.137}$$

For any given $v$ this defines a critical temperature

$$T_{\mathrm{c}} = T_{\mathrm{c}}(v) = \frac{h^2}{k_{\mathrm{B}}\, 2\pi m}\, \frac{1}{\zeta\left(\frac{3}{2}\right)^{2/3} (gv)^{2/3}}, \tag{6.138}$$

or, vice versa, for any given $T$ we obtain a critical volume per particle

$$v_{\mathrm{c}} = v_{\mathrm{c}}(T) = \left(\frac{h^2}{2\pi m k_{\mathrm{B}} T}\right)^{3/2} \frac{1}{g\zeta\left(\frac{3}{2}\right)}\,. \tag{6.139}$$

If

$$\frac{\lambda_{\mathrm{t}}^3}{gv} > \zeta\left(\frac{3}{2}\right), \tag{6.140}$$

or, equivalently, if

$$T < T_{\mathrm{c}}(v) \quad \text{or} \quad v < v_{\mathrm{c}}(T), \tag{6.141}$$

the number of particles in the ground state is of order $N$. This phenomenon is called Bose–Einstein condensation. However, the expression ‘condensation’ is not to be understood in a spatial sense but rather in relation to energy. The region $\lambda_{\mathrm{t}}^3/(gv) > \zeta\left(\frac{3}{2}\right)$, i.e., $T < T_{\mathrm{c}}(v)$ or $v < v_{\mathrm{c}}(T)$, is referred to as the condensation region.

In the condensation region we have

$$v < \lambda_{\mathrm{t}}^3\,, \tag{6.142}$$

i.e., the volume available for one particle is smaller than the volume of a box with sides equal to the thermal de Broglie wavelength. The uncertainty in position, which is at least equal to the thermal de Broglie wavelength, is therefore larger or equal to the average distance between atoms. The single-particle wave functions start to overlap.

We remark that in two dimensions, the function which replaces $h_{\frac{3}{2}}(z)$ is not finite for $z \to 1$. Hence, there is no need to take into account a term similar to (6.136) and there is therefore no condensation in two dimensions.

### Particle Number in the Condensation Region

Let us determine the number of particles $N_0$ in the ground state for $T < T_{\mathrm{c}}(v)$. Because the first term in (6.131) corresponds to the mean number of particles in the ground state, we find for the number of particles *not* in the ground state

$$N^* = gV\,\lambda_{\rm t}^{-3}(T)\,h_{\frac{3}{2}}(1) = gV\,\lambda_{\rm t}^{-3}(T)\,\zeta\left(\frac{3}{2}\right)\,. \tag{6.143}$$

For $T = T_{\rm c}(v)$ we have $N^* = N$, therefore

$$N = gV\,\lambda_{\rm t}^{-3}(T_{\rm c})\,\zeta\left(\frac{3}{2}\right), \tag{6.144}$$

i.e., as $\lambda_{\rm t}^{-1}(T) \propto T^{1/2}$,

$$\frac{N^*}{N} = \frac{\lambda_{\rm t}^{-3}(T)}{\lambda_{\rm t}^{-3}(T_{\rm c})} = \left(\frac{T}{T_{\rm c}}\right)^{3/2}, \tag{6.145}$$

and finally, since $N_0 = N - N^*$,

$$\frac{N_0}{N} = \left[1 - \left(\frac{T}{T_{\rm c}}\right)^{3/2}\right], \qquad \text{for} \quad T < T_{\rm c}. \tag{6.146}$$

For $T > T_{\rm c}$ the occupation of the ground state is of course so small that we may set $N_0 = 0$.

### 6.4.3 Pressure

We again consider the thermodynamic potential. From $K = -pV$ we obtain

$$\begin{aligned} -\,K(T,V,\mu) &= pV \qquad\qquad (6.147)\\ &= -gk_{\rm B}T\,\ln(1-z) - \frac{gk_{\rm B}T\,V}{(2\pi\hbar)^3}\int {\rm d}^3p\;\ln\left[1 - z\,\exp\left(-\beta\frac{p^2}{2m}\right)\right]\,. \end{aligned}$$

Again we have treated the contribution of the ground state separately, as we did for $N(T,V,\mu)$. Here, however, this would not have been necessary, because even if $1-z = O\left(N^{-1}\right)$ we now have $\ln(1-z) = O(\ln N)$, which for $N = 10^{23}$ gives $\ln N \approx 50 \ll N$.

Again we make the substitution $p^2 = 2mk_{\rm B}T\,x^2$ in the integral and obtain

$$p(T,V,\mu) = k_{\rm B}T\,\frac{g}{\lambda_{\rm t}^3}\,h_{\frac{5}{2}}(z) \tag{6.148}$$

with

$$h_{\frac{5}{2}}(z) = -\frac{4}{\sqrt{\pi}}\int_0^\infty {\rm d}x\,x^2\,\ln\left(1 - z\,{\rm e}^{-x^2}\right)\,. \tag{6.149}$$

The function $h_{\frac{5}{2}}(z)$ is closely related to $h_{\frac{3}{2}}(z)$, since

$$z\frac{\mathrm{d}}{\mathrm{d}z}\,h_{\frac{5}{2}}(z) = \frac{4}{\sqrt{\pi}}\int_0^\infty \mathrm{d}x\,x^2\,\frac{z\,\mathrm{e}^{-x^2}}{1-z\,\mathrm{e}^{-x^2}} = h_{\frac{3}{2}}(z)\;, \tag{6.150}$$

and the expansion with respect to $z$ now reads

$$h_{\frac{5}{2}}(z) = z + \frac{z^2}{2^{5/2}} + \frac{z^3}{3^{5/2}} + \ldots = \sum_{j=1}^{\infty}\frac{z^j}{j^{5/2}}. \tag{6.151}$$

We consider the case $T < T_\mathrm{c}$ or $v < v_\mathrm{c}$. Therefore we may set $z = 1$ and obtain

$$p(T,V,\mu) = k_\mathrm{B}T\,\frac{g}{\lambda_\mathrm{t}^3}\,\zeta\left(\frac{5}{2}\right) \propto (k_\mathrm{B}T)^{5/2}. \tag{6.152}$$

$p(T,V,\mu)$ is now independent of $v$ (and $\mu$, since $\mu$ is practically zero), and $p \to 0$ as $T \to 0$.

For the case $T > T_\mathrm{c}(v)$ or $v > v_\mathrm{c}(T)$ we proceed from the equation for $N(T,V,\mu)$,

$$N(T,V,\mu) = \frac{gV}{\lambda_\mathrm{t}^3}\,h_{\frac{3}{2}}(z)\;, \tag{6.153}$$

solve it with respect to $z$ and obtain

$$z = \frac{\lambda_\mathrm{t}^3}{gv} + \beta_2\left(\frac{\lambda_\mathrm{t}^3}{gv}\right)^2 + \ldots = \sum_{j=1}^{\infty}\beta_j\left(\frac{\lambda_\mathrm{t}^3}{gv}\right)^j\;, \tag{6.154}$$

with

$$\beta_2 = -\frac{1}{2^{3/2}}\;,\ \text{etc.} \tag{6.155}$$

This also determines the chemical potential for $T > T_\mathrm{c}(v)$ or $v > v_\mathrm{c}(T)$. Inserting this expansion in $p(T,V,\mu)$ yields

$$\frac{pV}{Nk_\mathrm{B}T} = \sum_{l=1}^{\infty}a_l\left(\frac{\lambda_\mathrm{t}^3}{gv}\right)^{l-1}\;, \tag{6.156}$$

with

$$a_1 = 1,\quad a_2 = -\frac{1}{2^{5/2}},\quad a_3 = -\left(\frac{2}{2^{5/2}} - \frac{1}{8}\right)\quad \ldots. \tag{6.157}$$

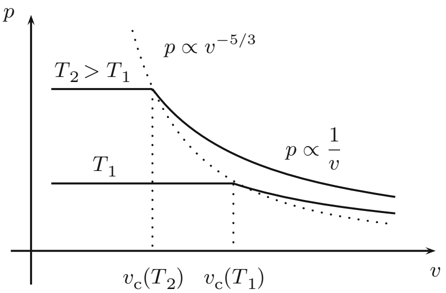


**Fig. 6.4** Isotherms for a Bose gas in a $p$–$v$ diagram. The region to the *left* of the *dotted curve* ($p \propto v^{-5/3}$) is the condensation region

Hence, in particular,

$$pV = Nk_BT\left(1 - \frac{\lambda_t^3}{gv}\frac{1}{2^{5/2}} - \dots\right), \tag{6.158}$$

cf. (6.76).

We now consider the isotherms in a $p$–$v$ diagram. For fixed $T$ and $v > v_c$ the pressure $p$ decreases as $\frac{1}{v} + O\left(\frac{1}{v^2}\right)$, while for $v < v_c(T)$ it is independent of $v$. So we obtain the curve plotted in Fig. 6.4.

As $v_c(T) \propto T^{-3/2}$ and $p \propto T^{5/2}$ for $v < v_c$, the pressure behaves as $p \propto v_c^{-5/3}$ for $v < v_c$. The points where the isotherms become a horizontal line lie on a curve given by $p \propto v^{-5/3}$.

If at constant temperature the volume of the gas gets smaller the pressure increases as long as $v > v_c$; for $v < v_c$ the pressure remains constant. More and more particles condense into the ground state. But as the momentum of the particles in the ground state is practically zero they do not contribute to the pressure.

So we may speak of a two-phase region, or, equivalently, of the coexistence of two phases in the condensation region: the condensed phase, where the number of particles in the ground state is of order $N$, and a normal phase, where some particles may also be in the ground state, but their number is not of order $N$.

Lowering $v$ leads to an increase of the particle number in the condensed phase at the expense of the number of particles in the normal phase. (In the expression $N_0/N = 1 - (T/T_c)^{3/2}$ the critical temperature $T_c = T_c(v)$ changes as $T_c \propto v^{-2/3}$, i.e., if $v$ becomes smaller $T_c$ gets larger and thus also $N_0$).

This two-phase region corresponds to the two-phase region that we already met in connection with phase transitions of first order in Sect. 3.8.

### 6.4.4 Energy and Specific Heat

In Sect. 6.2 we derived the relation

$$pV = \frac{2}{3}E \ . \tag{6.159}$$

From this we find, for $T > T_{\mathrm{c}}$ or $v > v_{\mathrm{c}}$,

$$E = \frac{3}{2}\, pV = \frac{3}{2}\, N k_{\mathrm{B}} T \left(1 - \frac{\lambda_{\mathrm{t}}^3}{gv}\, \frac{1}{2^{5/2}} - \dots\right), \tag{6.160}$$

from which the specific heat follows immediately as

$$C_V = \frac{\partial E}{\partial T} = \frac{3}{2}\, k_{\mathrm{B}} N \left(1 + \frac{1}{2}\, \frac{1}{2^{5/2}}\, \frac{\lambda_{\mathrm{t}}^3}{gv} + \dots\right), \tag{6.161}$$

since

$$\frac{\partial}{\partial T}\, T \lambda_{\mathrm{t}}^3 = -\frac{1}{2}\, \lambda_{\mathrm{t}}^3. \tag{6.162}$$

The classical results are obtained as the leading terms in $T$. The next to leading terms are smaller by a factor $\propto T^{-3/2}$ and become more relevant at lower temperatures. The energy decreases for lower temperatures while $C_V$ gets larger.

For $T < T_{\mathrm{c}}$ or $v < v_{\mathrm{c}}$ we obtain from (6.152)

$$E = \frac{3}{2}\, pV = \frac{3}{2}\, V k_{\mathrm{B}} T\, \frac{g}{\lambda_{\mathrm{t}}^3}\, \zeta\left(\frac{5}{2}\right) \propto (k_{\mathrm{B}} T)^{5/2}\ . \tag{6.163}$$

Using (cf. (6.137))

$$gV = N\, \frac{\lambda_{\mathrm{t}}^3(T_{\mathrm{c}})}{\zeta\left(\frac{3}{2}\right)}\ , \tag{6.164}$$

we may also write

$$E = \frac{3}{2}\, N k_{\mathrm{B}}\, \frac{\zeta\left(\frac{5}{2}\right)}{\zeta\left(\frac{3}{2}\right)}\, T \left(\frac{T}{T_{\mathrm{c}}}\right)^{3/2}\ . \tag{6.165}$$

For the specific heat we now obtain

$$C_V = \frac{\partial E}{\partial T} = \frac{15}{4} k_{\mathrm{B}} N\, \frac{\zeta\left(\frac{5}{2}\right)}{\zeta\left(\frac{3}{2}\right)} \left(\frac{T}{T_{\mathrm{c}}}\right)^{3/2}\ , \tag{6.166}$$

in particular, for $T = T_{\mathrm{c}}$,

$$C_V = \frac{15}{4} k_{\mathrm{B}} N\, \frac{\zeta\left(\frac{5}{2}\right)}{\zeta\left(\frac{3}{2}\right)} \approx 1.925 k_{\mathrm{B}} N > \frac{3}{2}\, k_{\mathrm{B}} N\ . \tag{6.167}$$

The dependence of $C_V$ on $T$ is shown in Fig. 6.5.

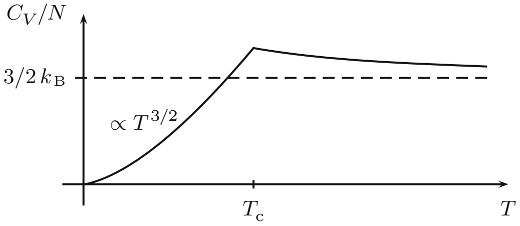


**Fig. 6.5** Specific heat $C_V$ for an ideal Bose gas as a function of temperature

### 6.4.5 Entropy

The entropy follows from

$$S = \frac{1}{T}\,(E - \mu N - K) = \frac{1}{T}\left(\frac{5}{2}\,pV - \mu N\right) \tag{6.168}$$

$$= \frac{5}{2}\,k_\mathrm{B} V \frac{g}{\lambda_\mathrm{t}^3} h_{\frac{5}{2}}(z) - k_\mathrm{B} N \ln z, \tag{6.169}$$

where we have used (6.148) and $\beta\mu = \ln z$. In the coexistence region we have $z = 1$ and therefore

$$\frac{S}{N} = \frac{5}{2}\,k_\mathrm{B} v \frac{g}{\lambda_\mathrm{t}^3}\zeta\left(\frac{5}{2}\right), \tag{6.170}$$

and with

$$gv = \frac{1}{\lambda_\mathrm{t}^{-3}(T_\mathrm{c})\,\zeta\left(\frac{3}{2}\right)} \tag{6.171}$$

we get

$$\frac{S}{N} = \frac{5}{2}\,k_\mathrm{B}\,\frac{\zeta\left(\frac{5}{2}\right)}{\zeta\left(\frac{3}{2}\right)}\left(\frac{T}{T_\mathrm{c}}\right)^{3/2} . \tag{6.172}$$

Hence, $S \to 0$ for $T \to 0$, as expected, because at $T = 0$ all particles are in the ground state.

### 6.4.6 Applications of Bose Statistics

The phenomenon of a Bose–Einstein condensation follows in a natural way for an ideal gas when the particles obey Bose statistics. Einstein predicted this phenomenon as early as 1924 and also gave the criterion for it to occur: when the thermal de Broglie wavelength $\lambda_\mathrm{t}$ becomes comparable with the interatomic distances.

Until recently, $^4$He gas was the only example quoted for the appearence of this phenomenon. In this case a superfluid phase occurs at 2.17 K and densities of

$n \approx 10^{22}\,\text{cm}^{-3}$. The large densities, however, imply that the He atoms are strongly interacting.

So one cannot expect all phenomena predicted for an ideal gas to really occur in this case. The phase transition to the superfluid state, the so-called $\lambda$-transition, is indeed of a different kind to that predicted for an ideal gas. The specific heat of $^4$He seems to have a logarithmic singularity at the critical temperature of 2.17 K and, for $T \to 0$, it decreases as $T^3$ instead of $T^{3/2}$.

On July 14th, 1995 both the New York Times and Science (Anderson et al. 1995) reported that two groups in the United States had succeeded in demonstrating Bose–Einstein condensation for a practically ideal gas. One group working at the National Institute of Standards and Technology in Boulder and the University of Colorado, was experimenting with $^{87}$Rb atoms; the second group from the Rice University in Houston,Texas, had demonstrated the condensation of $^7$Li atoms.

Actually, alkali atoms at low temperatures usually form a solid like other substances, but it is possible to maintain them in a metastable gaseous state even at temperatures of few nano-Kelvins. In these experiments the densities achieved were much smaller, in the range of some $10^{12}\,\text{cm}^{-3}$, because the gases were cooled by various techniques down to a temperature well below 170 nK. A gas of rubidium atoms, pretreated by laser cooling, was brought into a magnetic trap, where further cooling (known as evaporative cooling) and compression took place.

The Bose–Einstein condensation was confirmed by using a laser pulse to make a kind of snapshot of the velocity distribution after the magnetic field had been turned off and the atoms were escaping in all directions. The atoms condensed in the center of the trap spread out much more slowly, because of their small energy and therefore small velocity compared to the noncondensed atoms. This phenomenon sets in abruptly below a critical temperature (see also Collins 1995).

At present a number of experimental groups continue to investigate Bose–Einstein condensation with similar or even more sophisticated experimental techniques (Griffin et al. 1995). In 2001, the Nobel prize was given Eric A. Cornell, Wolfgang Ketterle, and Carl E. Wiemann for “the achievement of Bose-Einstein condensation in dilute gases of alkali atoms, and for early fundamental studies of the properties of the condensates”.

## 6.5 The Photon Gas and Black Body Radiation

After having established some general statements about Bose gases we will investigate a special, very important kind of Bose gas, namely the photon gas.

We consider a completely evacuated volume $V$. The walls of this volume have the temperature $T$ and they emit and absorb electromagnetic radiation. Hence, this cavity of volume $V$ contains only electromagnetic radiation, or, equivalently, making use of the notion of photons, a gas of photons. If this cavity is sufficiently large the statistical properties should be independent of any characteristic features of the walls and depend only on $T$ and $V$.

The theoretical framework for the concept of a photon is quantum electrodynamics. In this framework one postulates a Hamiltonian operator

$$H = \sum_{\boldsymbol{k},\alpha} \hbar\, \omega(\boldsymbol{k})\, a^{+}(\boldsymbol{k},\alpha)\, a(\boldsymbol{k},\alpha), \tag{6.173}$$

where $a^{+}(\boldsymbol{k},\alpha)$ is the creation operator for a photon with wave vector $\boldsymbol{k}$ and polarization $\alpha$ and $a(\boldsymbol{k},\alpha)$ the corresponding annihilation operator.

In a cubic box of volume $V = L^3$ we have

$$\boldsymbol{k} = \frac{2\pi}{L}\boldsymbol{m}, \quad \boldsymbol{m} = (m_1, m_2, m_3), \quad m_i = \pm 1, \ldots, \tag{6.174}$$

and the energy $\varepsilon(\boldsymbol{k},\alpha)$ of a photon in the state $(\boldsymbol{k},\alpha)$ is

$$\varepsilon(\boldsymbol{k},\alpha) \equiv \varepsilon(\boldsymbol{k}) = \hbar\omega(\boldsymbol{k}) = \hbar\, c|\boldsymbol{k}| = c|\boldsymbol{p}|, \tag{6.175}$$

with

$$\boldsymbol{p} = \hbar\boldsymbol{k}. \tag{6.176}$$

In general, for relativistic particles we have

$$E = \sqrt{m^2c^4 + c^2\boldsymbol{p}^2}, \tag{6.177}$$

and for photons the rest-mass $m$ is zero.

In addition, like any electromagnetic wave, each photon is characterized by a polarization vector $\boldsymbol{e}(\boldsymbol{k})$ with $\boldsymbol{e}(\boldsymbol{k}) \cdot \boldsymbol{k} = 0$. The vector $\boldsymbol{e}(\boldsymbol{k})$ lies thus in the plane orthogonal to $\boldsymbol{k}$. This plane is spanned by two base vectors, i.e., there are two possible polarizations, e.g., left circular and right circular. Hence, the index $\alpha$ takes the values 1 and 2.

A quantum-mechanical state of the entire photon gas may be characterized in an occupation number representation by the number of photons $n(\boldsymbol{k},\alpha)$ having wave vector $\boldsymbol{k}$ and polarization $\alpha$. Such a state is denoted by

$$|n(\boldsymbol{k}_1,\alpha_1), \ldots\rangle. \tag{6.178}$$

This state has particle number

$$N = \sum_{\boldsymbol{k},\alpha} n(\boldsymbol{k},\alpha) \tag{6.179}$$

and energy

$$E = \sum_{\boldsymbol{k},\alpha} \varepsilon(\boldsymbol{k}) n(\boldsymbol{k},\alpha) = \sum_{\boldsymbol{k},\alpha} \hbar\omega(\boldsymbol{k})\, n(\boldsymbol{k},\alpha)\ . \tag{6.180}$$

The fact that the rest-mass of the photon is zero implies that the number of particles $N$ in a photon gas cannot be held fixed, because the walls can emit and absorb photons however small the available energy is, and since $E \propto |\boldsymbol{k}|$ and $|\boldsymbol{k}| \propto L^{-1}$, the energy of the photons may become arbitrarily small for $L \to \infty$. Therefore the quantity conjugate to $N$, i.e., the chemical potential, is not defined. Only $T$ and $V$ may be held fixed; in principle $N$ is now a random quantity, for which we may determine, e.g., the expectation value; this we will do later.

If we consider the partition function for the canonial system without the supplementary condition that the number of particles is equal to $N$, we obtain

$$Y = \sum_{\{(n(\boldsymbol{k}_1,\alpha_1),\ldots)\}} \exp\left(-\beta\varepsilon(\boldsymbol{k}_1) n(\boldsymbol{k}_1,\alpha_1) - \beta\varepsilon(\boldsymbol{k}_2) n(\boldsymbol{k}_2,\alpha_2) - \ldots\right). \tag{6.181}$$

This is the expression to be expected when $T$ and $V$ are held fixed and the number of particles is arbitrary. Formally, this is also identical to the partition function of the grand canonical system, if we set $\mu = 0$.

Hence, the photon gas may be considered as a gas of bosons with two spin orientations ($g = 2$), where formally we have to set $\mu = 0$. So we obtain for the average occupation number:

$$\langle n(\boldsymbol{k},\alpha)\rangle = \frac{1}{\mathrm{e}^{\beta\varepsilon(\boldsymbol{k})} - 1}\ , \quad \varepsilon(\boldsymbol{k}) = \hbar\omega(\boldsymbol{k}) = \hbar c|\boldsymbol{k}|. \tag{6.182}$$

Let us now calculate the various system variables:

**Particle number.** The expectation value of the particle number is

$$N(T,V) = \sum_{\boldsymbol{k},\alpha} \langle n(\boldsymbol{k},\alpha)\rangle = g \sum_{\boldsymbol{k}} \frac{1}{\mathrm{e}^{\beta\varepsilon(\boldsymbol{k})} - 1}. \tag{6.183}$$

If we replace the sum over $\boldsymbol{k}$ by an integral according to

$$\sum_{\boldsymbol{k}} \to \frac{V}{(2\pi)^3} \int \mathrm{d}^3 k, \tag{6.184}$$

take $\omega = kc$ as the integration variable (i.e., $\mathrm{d}^3k = 4\pi k^2 \mathrm{d}k = (4\pi\omega^2/c^3)\,\mathrm{d}\omega$), and set $g = 2$, we obtain

$$N = \frac{V}{\pi^2 c^3} \int_0^\infty \mathrm{d}\omega\, \frac{\omega^2}{\mathrm{e}^{\beta\hbar\omega} - 1} \equiv \int_0^\infty \mathrm{d}\omega\, N(\omega)\ , \tag{6.185}$$

where

$$N(\omega)\,\mathrm{d}\omega = \frac{V}{\pi^2 c^3}\frac{\omega^2}{\mathrm{e}^{\beta\hbar\omega}-1}\mathrm{d}\omega \tag{6.186}$$

is the average number of photons in the interval $(\omega, \omega + \mathrm{d}\omega)$.

(Notice that we did not have to treat the contribution from the ground orbital separately. Since $\beta\varepsilon(k) \propto k \propto L^{-1} \propto V^{-1/3}$ the contribution from the lowest $\varepsilon(k)$ is at most

$$\frac{1}{\mathrm{e}^{\beta\varepsilon(k)}-1} = O(V^{1/3})\ . \tag{6.187}$$

Hence, there will be no condensate for particles.)

Finally, we get for the particle number:

$$\begin{aligned} N(T,V) &= \int_0^\infty \mathrm{d}\omega\, N(\omega) = \frac{V}{\pi^2 c^3}\int_0^\infty \mathrm{d}\omega\, \frac{\omega^2}{\mathrm{e}^{\frac{\hbar\omega}{k_\mathrm{B}T}}-1} \\ &= 0.244\left(\frac{k_\mathrm{B}T}{\hbar c}\right)^3 V. \end{aligned} \tag{6.188}$$

**Energy.** Similarly, we obtain for the energy

$$E(T,V) = \int_0^\infty \mathrm{d}\omega\,\hbar\omega\, N(\omega) = \int_0^\infty \mathrm{d}\omega\, E(\omega), \tag{6.189}$$

where

$$E(\omega)\,\mathrm{d}\omega = \frac{V}{\pi^2 c^3}\frac{\hbar\omega^3}{\exp\left(\frac{\hbar\omega}{k_\mathrm{B}T}\right)-1}\mathrm{d}\omega \tag{6.190}$$

is the average energy of the photon gas in the interval $(\omega, \omega + \mathrm{d}\omega)$.

Setting $\omega = 2\pi c/\lambda$, i.e., $\mathrm{d}\omega = \left|2\pi c/\lambda^2\right|\,\mathrm{d}\lambda$, we may also write

$$E = \int_0^\infty \mathrm{d}\lambda\, E(\lambda)\ , \tag{6.191}$$

where

$$E(\lambda)\,\mathrm{d}\lambda = \frac{V\hbar 16\pi^2 c}{\lambda^5}\frac{1}{\exp\left(\frac{2\pi\hbar c}{\lambda k_\mathrm{B}T}\right)-1}\mathrm{d}\lambda \tag{6.192}$$

is the average energy of the photon gas in the interval $(\lambda, \lambda + \mathrm{d}\lambda)$.

The relations (6.190) and (6.192) are also referred to as Planck's radiation law.

We observe:

- For low frequencies, i.e., $\hbar\omega \ll k_\mathrm{B}T$, we may write

$$\exp\left(\frac{\hbar\omega}{k_\mathrm{B}T}\right) \approx 1 + \frac{\hbar\omega}{k_\mathrm{B}T} \tag{6.193}$$

and therefore

$$E(\omega) = \frac{V}{\pi^2 c^3}\,\omega^2\,k_{\mathrm{B}}T. \tag{6.194}$$

There is no $\hbar$ in this formula for $E(\omega)$, which is proportional to $\omega^2$ in this limit. Each orbital contributes $k_{\mathrm{B}}T$ to the energy. This is the classical approximation, also known as the Rayleigh–Jeans formula.

- For $\hbar\omega \gg k_{\mathrm{B}}T$ we have

$$\mathrm{e}^{\hbar\omega/k_{\mathrm{B}}T} - 1 \approx \mathrm{e}^{\hbar\omega/k_{\mathrm{B}}T} \tag{6.195}$$

and therefore

$$E(\omega) = \frac{V\hbar}{\pi^2 c^3}\,\omega^3\,\mathrm{e}^{-\hbar\omega/k_{\mathrm{B}}T}. \tag{6.196}$$

This is referred to as Wien's radiation law.

- We now want to determine the frequency at which the energy $E(\omega)$ assumes its maximum. For this we have to compute the point where the function $x^3/(\mathrm{e}^x - 1)$ assumes its maximum, which we find to be $x = 2.822$. Hence the energy assumes its maximum value at $\omega_{\max}$, where

$$\frac{\hbar\,\omega_{\max}}{k_{\mathrm{B}}T} = 2.822. \tag{6.197}$$

The frequency corresponding to the maximum of $E(\omega)$ is therefore proportional to the temperature (Wien's displacement law). Similarly, we find for the value where $E(\lambda)$ is maximal:

$$\lambda_{\max}\,T = 2,880\,\mu\mathrm{m\,K}. \tag{6.198}$$

For $T = 6,000\,\mathrm{K}$ we get $\lambda_{\max} \approx 0.5\,\mu\mathrm{m} = 500\,\mathrm{nm}$, for $T = 300\,\mathrm{K}$ we find $\lambda_{\max} \approx 10\,\mu\mathrm{m}$ (Fig. 6.6).

Finally, since $\int_0^\infty \mathrm{d}x\, x^3/(\mathrm{e}^x - 1) = \pi^4/15$, we find

$$E(T,V) = \frac{V\hbar}{\pi^2 c^3}\,\frac{1}{\hbar^4}\,(k_{\mathrm{B}}T)^4 \int_0^\infty \mathrm{d}x\,\frac{x^3}{\mathrm{e}^x - 1} = \frac{4\sigma}{c}\,V\,T^4\,, \tag{6.199}$$

with the Stefan–Boltzmann constant

$$\sigma = \frac{\pi^2 k_{\mathrm{B}}^4}{60 c^2 \hbar^3} = 5.67 \times 10^{-8}\,\frac{\mathrm{W}}{\mathrm{m}^2\mathrm{K}^4}. \tag{6.200}$$

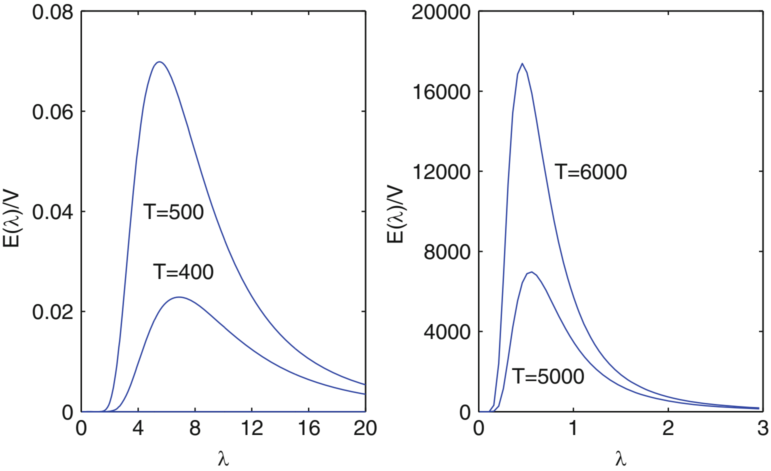


**Fig. 6.6** Planck spectrum $E(\lambda)/V$ for different ranges of the wavelength $\lambda$ in units of μm. The values for $E(\lambda)/V$ are given in units of $10^{-28}$ Ws/m$^4$

**Thermodynamic potential.** For the thermodynamic potential we obtain

$$K(T,V) = k_{\mathrm{B}}T\,\frac{2\,V}{(2\pi)^3}\,4\pi\int_0^\infty \mathrm{d}k\,k^2\,\ln\left[1-\exp\left(-\frac{\hbar\omega(k)}{k_{\mathrm{B}}T}\right)\right] \tag{6.201}$$

$$= \frac{Vk_{\mathrm{B}}T}{\pi^2c^3}\int_0^\infty \mathrm{d}\omega\,\omega^2\,\ln\left[1-\exp\left(-\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)\right]. \tag{6.202}$$

We could also have used this formula to derive the above expressions for $N(T,V)$ and $E(T,V)$ (after first introducing formally the chemical potential, building the corresponding derivatives and then setting the chemical potential again to zero). However, we will now relate the thermodynamic potential to the energy by partial integration:

$$K(T,V) = \frac{Vk_{\mathrm{B}}T}{3\pi^2c^3}\,(-)\int_0^\infty \mathrm{d}\omega\,\omega^3\,\frac{(-\hbar/k_{\mathrm{B}}T)}{\left(1-\mathrm{e}^{-\hbar\omega/k_{\mathrm{B}}T}\right)}\,\left(-\mathrm{e}^{-\hbar\omega/k_{\mathrm{B}}T}\right)$$

$$= -\frac{V\hbar}{3\pi^2c^3}\int_0^\infty \mathrm{d}\omega\,\omega^3\,\frac{1}{\mathrm{e}^{\hbar\omega/k_{\mathrm{B}}T}-1} = -\frac{1}{3}\,E(T,V)\,, \tag{6.203}$$

and since $K = -p\,V$, we now obtain for this relativistic gas (cf. (6.79) and (6.80))

$$p\,V = \frac{1}{3}\,E(T,V)\,. \tag{6.204}$$

Together with (6.199) this yields for the pressure:

$$p = \frac{4\sigma}{3c} T^4 \,, \tag{6.205}$$

which is independent of $V$, i.e., the isotherms in a $p$–$V$ diagram are horizontal lines. In analogy to the nonrelativistic Bose gas in the two phase region (6.152), where a condensate and a normal phase coexist, here we might call the vacuum a condensate. The photon gas may then be considered as a bose gas which is always in coexistence with its condensate, the vacuum. Compressing the gas by decreasing $V$ 'forces particles into the condensate', i.e., annihilates photons. This picture is consistent with the expectation value of the particle number $N$, which now corresponds to the number of particles $N^*$ not in the ground state. This expectation value is given by (6.188) and decreases as the volume decreases.

**Entropy.** For the entropy we get

$$S(T, V) = -\frac{\partial K}{\partial T} = \frac{16\sigma}{3c} V\, T^3. \tag{6.206}$$

In the limit $T \to 0$, energy, pressure, specific heat, and entropy thus all tend to zero. The system is in its ground state, which is now characterized by the total absence of photons. This corresponds to the vacuum state. It is unique and therefore its entropy vanishes.

**Specific heat.** For the specific heat the result is

$$C_V = \frac{\partial E(T, V)}{\partial T} = \frac{16\sigma}{c} V\, T^3 \,. \tag{6.207}$$

The electromagnetic radiation corresponding to a photon gas is also called heat radiation.

### 6.5.1 The Kirchhoff Law

We now consider a body in equilibrium with the heat radiation of its environment. We first assume that the body reflects, absorbs, and emits photons such that on the average the distribution of photons with respect to their frequencies and direction of wave vectors remains constant. Furthermore, the radiation should be isotropic and spatially homogeneous (i.e., the same in each volume element). The average energy of the photon gas in the interval $(\omega, \omega + \mathrm{d}\omega)$ is $E(\omega)$.

Let us investigate the radiation incident on an area element $\mathrm{d}A$ of the surface of the body.

For this we consider a volume element $\mathrm{d}V$ at a distance $r$ from the area element $\mathrm{d}A$. The connecting line between $\mathrm{d}V$ and $\mathrm{d}A$ shall make an angle $\theta$ with the

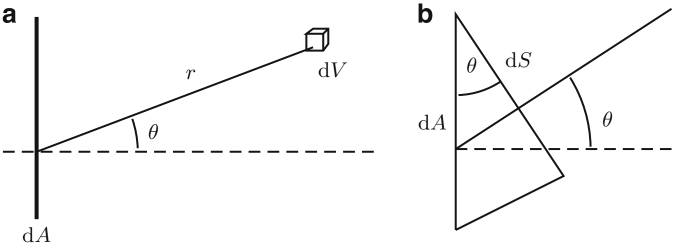


**Fig. 6.7** (**a**) Volume element d$V$ at a distance $r$ from the area element d$A$; (**b**) surface element d$S$ through which the radiation incident on d$A$ from d$V$ has to pass

normal to d$A$. So for the surface element d$S$ through which the radiation incident on d$A$ has to pass we find

$$\mathrm{d}S = \mathrm{d}A\,\cos\theta, \tag{6.208}$$

as shown in Fig. 6.7.

Now we have to take into account all volume elements d$V$ from which radiation may be incident on d$A$ in an infinitesimal time interval d$t$. We take the origin of the coordinate system to be in d$A$, i.e., $\mathrm{d}V = r^2\,\mathrm{d}r\,\mathrm{d}\Omega$, and if we are interested in the radiation coming from the solid angle d$\Omega$, we have to consider all volume elements $r^2\,\mathrm{d}r\,\mathrm{d}\Omega$ for which $0 \le r \le c\,\mathrm{d}t$.

Furthermore, from a volume element at a distance $r$, a fraction $\mathrm{d}S/4\pi r^2$ of the radiation $E(\omega)\mathrm{d}\omega\mathrm{d}V/V$ passes through the surface element $\mathrm{d}S = \mathrm{d}A\,\cos\theta$.

The incident energy per unit area and per unit time from a solid angle d$\Omega$ is therefore

$$E_{\mathrm{in}}\mathrm{d}\omega\,\mathrm{d}\Omega = \frac{1}{\mathrm{d}A\mathrm{d}t}\int_0^{c\mathrm{d}t} r^2\,\mathrm{d}r\,\mathrm{d}\Omega\,\frac{E(\omega)\,\mathrm{d}\omega}{V}\,\frac{\mathrm{d}S}{4\pi r^2} \tag{6.209}$$

$$= c\,\frac{E(\omega)}{4\pi V}\,\mathrm{d}\omega\,\cos\theta\mathrm{d}\Omega. \tag{6.210}$$

Let $A(\omega,\theta,T)$ with $0 \le A(\omega,\theta,T) \le 1$ be the absorption coefficient of the body, i.e. the fraction of the energy which is absorbed by the body. All radiation which is not absorbed is reflected (the percentage (1-A) 100% is called the albedo). With

$$e_0(\omega,T) := \frac{E(\omega)}{4\pi V} = \frac{1}{4\pi\pi^2c^3}\,\frac{\hbar\omega^3}{\mathrm{e}^{\hbar\omega/k_{\mathrm{B}}T}-1} \tag{6.211}$$

we obtain for the energy $E_A$ absorbed per unit time and per unit area from a solid angle d$\Omega$ and within the frequency interval $(\omega, \omega + \mathrm{d}\omega)$

$$E_A\mathrm{d}\omega\,\mathrm{d}\Omega = A(\omega,\theta,T)\,E_{in}\mathrm{d}\omega\,\mathrm{d}\Omega = A(\omega,\theta,T)\,c\,e_0(\omega)\,\mathrm{d}\omega\,\cos\theta\,\mathrm{d}\Omega. \tag{6.212}$$

Now let $I(\omega,\theta,T)\,\mathrm{d}\omega\,\mathrm{d}\Omega$ be the energy emitted per unit time and per unit area into the solid angle d$\Omega$ within the frequency interval $(\omega, \omega + \mathrm{d}\omega)$. In a state of equilibrium all absorbed radiation also has to be emitted, i.e.,

$$I(\omega,\theta,T)\,\mathrm{d}\omega\,\mathrm{d}\Omega = A(\omega,\theta,T)\,c\,e_0(\omega,T)\,\cos\theta\,\,\mathrm{d}\omega\,\mathrm{d}\Omega. \tag{6.213}$$

Thus, apart from the factor $A$, the dependence of the radiated energy with respect to the direction is given by $\cos\theta$. This is Lambert's law.

Of course, $I(\omega,\theta,T)$ and $A(\omega,\theta,T)$ depend on the nature of the walls. Independent of this nature, however, is the following ratio

$$\frac{I(\omega,\theta,T)}{A(\omega,\theta,T)} = c\,e_0(\omega)\,\cos\theta = \frac{c}{4\pi\pi^2c^3}\,\frac{\hbar\omega^3}{\mathrm{e}^{\hbar\omega/k_\mathrm{B}T}-1}\,\cos\theta. \tag{6.214}$$

Therefore the ratio of the emissivity $I(\omega,\theta,T)$, (with dimension energy per unit time and per unit area) and the absorption coefficient $A(\omega,\theta,T)$ of a body is universal and equal to $c\,e_0(\omega,T)\,\cos\theta$. This is Kirchhoff's law.

Let us consider the case where the light may also be scattered, i.e., the angle $\theta$ may change. We still have

$$\int \mathrm{d}\Omega\; I(\omega,\theta) = \int \mathrm{d}\Omega A(\omega,\theta)\,c\,e_0(\omega)\,\cos\theta. \tag{6.215}$$

If in the scattering process the frequency may also change (as is the case, e.g., for phosphorescence) Kirchhoff's law holds in the form

$$\int I(\omega,\theta)\,\mathrm{d}\Omega\,\mathrm{d}\omega = c\int e_0(\omega)\,A(\omega,\theta)\,\cos\theta\,\mathrm{d}\Omega\,\mathrm{d}\omega. \tag{6.216}$$

## 6.5.2 The Stefan–Boltzmann Law

A body for which $A(\omega,\theta) = 1$, i.e., one that absorbs all radiation, is called absolutely 'black'. Then the emissivity

$$I(\omega,\theta) = c\,e_0(\omega)\,\cos\theta \equiv I_S(\omega,\theta) \tag{6.217}$$

is independent of the nature of the body, i.e., it is the same for all black bodies.

A small hole in the wall of a cavity behaves as a black body, since all radiation entering through this hole is absorbed or reflected inside the cavity, but does not escape.

The emission through the hole is

$$\begin{aligned} I_0 &= \int_0^\infty \mathrm{d}\omega\mathrm{d}\Omega\, I_S(\omega,\theta) \\ &= c\int_0^\infty e_0(\omega)\,\mathrm{d}\omega \int_0^{\pi/2} \cos\theta\,\sin\theta\,\mathrm{d}\theta \int_0^{2\pi} \mathrm{d}\varphi \end{aligned}$$

$$= 2\pi \, \frac{1}{2} \, c \int_0^\infty e_0(\omega) \, \mathrm{d}\omega$$

$$= \pi c \, \frac{\hbar}{4\pi \pi^2 c^3} \int_0^\infty \mathrm{d}\omega \, \frac{\omega^3}{\mathrm{e}^{\frac{\hbar\omega}{k_\mathrm{B}T}} - 1} . \tag{6.218}$$

Using (cf. (6.189))

$$\frac{E}{V} = \frac{\hbar}{\pi^2 c^3} \int_0^\infty \mathrm{d}\omega \, \frac{\omega^3}{\mathrm{e}^{\hbar\omega/k_\mathrm{B}T} - 1} \tag{6.219}$$

we finally obtain (cf. (6.199))

$$I_0 = \frac{c}{4} \, \frac{E}{V} = \sigma T^4 . \tag{6.220}$$

This is the so-called Stefan–Boltzmann law.

In the remaining sections we consider some applications of the theory of the photon gas.

### 6.5.3 *The Pressure of Light*

We consider an incandescent lamp of 100 W, i.e., the emitted energy per second is at most 100 J/s. At a distance of one meter the incident energy per unit surface and per second is therefore, disregarding all (considerable) losses,

$$S_G = \frac{100}{4\pi} \, \frac{\mathrm{J}}{\mathrm{m}^2\mathrm{s}} . \tag{6.221}$$

The associated pressure is given by (see the discussion in the next paragraph)

$$p = \frac{S_G}{c} = \frac{100}{4\pi \times 3 \times 10^8} \, \frac{\mathrm{J}}{\mathrm{m}^3} \approx 2.6 \times 10^{-8} \, \frac{\mathrm{N}}{\mathrm{m}^2} , \tag{6.222}$$

i.e.,

$$p \approx 2.6 \times 10^{-8} \, \mathrm{Pa} = 2.6 \times 10^{-13} \, \mathrm{bar} , \tag{6.223}$$

(since 1 J = 1 N m, 1 Pa = 1 $\mathrm{Nm}^{-2} = 10^{-5}$ bar).

The radiation pressure generated by a incandescent lamp is therefore absolutely negligible.

To derive the relation $p = S_G/c$ between the incident energy per area element d$A$ and per time interval d$t$ and the pressure $p$ on the surface, we consider a volume of base area $A$ and height $c \, \mathrm{d}t$. Within this volume shall be $n$ photons per cm$^3$ with their momentum directed to the area. Hence, during the time interval d$t$ the number of photons hitting the area $A$ is $n A \, c \, \mathrm{d}t$ (Fig. 6.8).

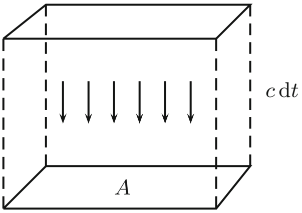


**Fig. 6.8** If there are $n$ photons per cm$^3$ inside the box of base area $A$ and height $c\,\mathrm{d}t$ with their momentum directed towards $A$, then $n\,A\,c\,\mathrm{d}t$ photons are incident on $A$ in the time interval $\mathrm{d}t$

Let us consider photons of frequency, say, $\nu$. Their momentum then is $h\nu/c$. If all photons are absorbed, the transmitted momentum is

$$F\,\mathrm{d}t = nAc\mathrm{d}t\,\frac{h\nu}{c}\,, \tag{6.224}$$

by which within $\mathrm{d}t$ the force $F$ is exerted. The pressure follows as

$$p = \frac{F}{A} = \frac{nch\nu}{c} = \frac{S_G}{c}\,, \tag{6.225}$$

because $nch\nu$ is the energy incident on the surface per unit time and unit area, which in turn is just equal to $S_G$.

As a second example of radiation pressure we consider the energy which is incident from the Sun on the outer atmosphere per unit area and unit time. This can be measured and amounts to

$$S = 1.36 \times 10^3\,\frac{\mathrm{J}}{\mathrm{m^2 s}} = 1.36\,\frac{\mathrm{kW}}{\mathrm{m^2}}. \tag{6.226}$$

This quantity is also called the solar constant. Roughly 30% of the radiation is reflected by the outer atmosphere of the Earth, i.e., on average about 1 kW/m$^2$ is incident on the surface of the Earth. This implies a radiation pressure of

$$p \approx \frac{10^3}{3 \times 10^8}\,\frac{\mathrm{J}}{\mathrm{m^3}} \approx 3 \times 10^{-6}\,\frac{\mathrm{N}}{\mathrm{m^2}} = 3 \times 10^{-11}\,\mathrm{bar}. \tag{6.227}$$

Thus the radiation pressure of the Sun on the Earth is also negligible.

Finally, we compare the pressure of a photon gas $p_\nu$ with the pressure $p_m$ of an ideal gas of molecules (atoms) at the same temperature. We have

$$p_\nu = \frac{4\sigma}{3c}\,T^4 \tag{6.228}$$

and

$$p_m = \frac{Nk_\mathrm{B}T}{V} = RTn_m\,, \tag{6.229}$$

where $R = k_\mathrm{B}\mathrm{A}$, $\mathrm{A}$ = Avogadro's number $= 6 \times 10^{23}$ particles/mol, is the gas constant and $n_m$ then is the number density in mol cm$^{-3}$. The value of the gas constant is $R = 8.3$ J/K mol, from which we obtain

$$\frac{p_\nu}{p_m} = \frac{4}{3}\frac{\sigma}{c}\frac{1}{Rn_m}T^3 \tag{6.230}$$

$$= \frac{4}{3}\frac{5.67\times10^{-8}}{3\times10^{8}}\frac{1}{8.3}\frac{T^3}{n_m}\frac{\mathrm{J}}{\mathrm{m^2sK^4}}\frac{1}{\mathrm{m\,s^{-1}}}\frac{\mathrm{K\,mol}}{\mathrm{J}} \tag{6.231}$$

$$\approx 3\times10^{-23}\frac{T^3}{n_m}\frac{1}{\mathrm{K^3}}\frac{\mathrm{mol}}{\mathrm{cm^3}}. \tag{6.232}$$

Taking $n_m = 1\,\mathrm{mol\,cm^{-3}}$, the temperature has to be $T \approx 3\times10^7\,\mathrm{K}$ for $p_\nu$ to be comparable with $p_m$. Although this corresponds to the temperatures inside the Sun, the density there is about $n_m \approx 100\,\mathrm{mol\,cm^{-3}}$ such that $p_\nu/p_m \approx 0.03$. The general estimate for stars is

$$\frac{p_\nu}{p_m} \le 0.5. \tag{6.233}$$

### 6.5.4 The Total Radiative Power of the Sun

From the solar constant $S$ (6.226) we may deduce the total radiation power $L$ of the Sun. With $R_{\mathrm{SE}} = 1.5\times10^{11}$ m being the distance between Sun and Earth we find

$$L = 4\pi\,R_{\mathrm{SE}}^2\,S \approx 12\times\left(1.5\times10^{11}\mathrm{m}\right)^2\;1.36\times10^3\frac{\mathrm{J}}{\mathrm{m^2s}} \tag{6.234}$$

$$\approx 4\times10^{26}\frac{\mathrm{J}}{\mathrm{s}} = 4\times10^{23}\,\mathrm{kW}. \tag{6.235}$$

Let $I_0$ be the energy emitted per unit time and unit area from the surface of the Sun. As $R_{\mathrm{S}} = 7\times10^8$ m is the radius of the Sun we obtain

$$I_0 = \frac{L}{4\pi R_{\mathrm{S}}^2} = \frac{R_{\mathrm{SE}}^2}{R_{\mathrm{S}}^2}\,S = \left(\frac{1.5\times10^{11}}{7\times10^8}\right)^2 S \tag{6.236}$$

$$\approx 4\times10^4\times1.36\times10^3\,\frac{\mathrm{J}}{\mathrm{m^2s}} \approx 6.2\times10^7\,\frac{\mathrm{J}}{\mathrm{m^2s}}. \tag{6.237}$$

According to the Stefan–Boltzmann law, $I_0 = \sigma T^4$, this is equivalent to a temperature of the surface of the Sun of

$$T = \left(\frac{I_0}{\sigma}\right)^{1/4} = \left(\frac{6.2\times10^7}{5.67\times10^{-8}}\,\frac{\mathrm{J/m^2s\,K^4}}{\mathrm{J/m^2s}}\right)^{1/4} \tag{6.238}$$

$$\approx \left(1,100\times10^{12}\,\mathrm{K}^4\right)^{1/4} \approx 5,750\,\mathrm{K}. \tag{6.239}$$

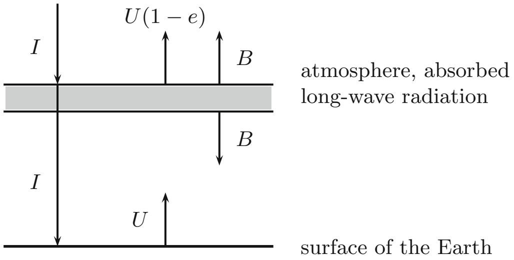


**Fig. 6.9** Radiation balance: short-wave radiation from the Sun ($\lambda < 4\,\mu$m) reaches the surface of the Earth with very little being absorbed, the atmosphere absorbs and reflects dominantly the long-wave part. A counter radiation $B$ results, leading to a higher temperature on the surface of the Earth

The disc of the Earth (as seen from the Sun) receives per unit time the energy

$$S_E = \pi R_E^2\, S\ , \tag{6.240}$$

where $R_E$ is the radius of the Earth. As we have assumed that 30% of this energy is reflected by the atmosphere of the Earth, about 70% is reradiated by the Earth, but in all directions. Hence, the Earth emits per unit time the energy

$$I_E = 0.7\,\frac{1}{4\pi R_E^2}\,S_E = \frac{0.7}{4}\,S. \tag{6.241}$$

If we use this value to calculate the temperature on the surface of the Earth according to the Stefan–Boltzmann law, $I_E = \sigma T_E^4$, we find

$$T_E^4 = \frac{0.7}{4}\,\frac{1.36\times 10^3\,\mathrm{J/m^2s}}{5.67\times 10^{-8}\,\mathrm{J/m^2sK^4}} \approx 0.04\times 10^{11}\mathrm{K}^4 = 40\times 10^8\mathrm{K}^4\ , \tag{6.242}$$

which yields

$$T_E \approx 250\,\mathrm{K}. \tag{6.243}$$

In reality, however, $T_E \approx 290\,$K. The discrepancy is explained by the so-called greenhouse effect, which plays a dominant role on Earth. The greenhouse effect denotes the influence of the atmosphere on the radiation and heat balance of the Earth. Like the glass panes of a greenhouse, water vapor and carbon dioxide in the atmosphere transmit the short-wave radiation from the Sun to the surface of the Earth with very little being absorbed. On the other hand, these gases absorb or reflect the long-wave radiation reradiated by the surface of the Earth. This leads to a heating of the Earth. To investigate this effect qualitatively, we consider the following radiation balance (Fig. 6.9).

Let $I$ be the total incident radiation, $U$ the radiation reradiated from the surface of the Earth, and $e$ the fraction of $U$ absorbed by the atmosphere, leading to the undirected radiation $B$ (counter radiation). The balance equation in equilibrium for the atmosphere is therefore

$$I = (1 - e)U + B \tag{6.244}$$

and on the surface

$$I = U - B. \tag{6.245}$$

Eliminating $B$ and solving for $U$ yields

$$U = \frac{I}{1 - e/2}. \tag{6.246}$$

Hence, the reradiated energy $U$ from the surface of the Earth increases with increasing absorbtion $e$ in the atmosphere and therefore, since $U = \sigma T_{\mathrm{E}}^4$, there is a rise in temperature $T_{\mathrm{E}}$ on the surface.

### *6.5.5 The Cosmic Background Radiation*

Another famous photon gas comprises the 3 K cosmic background radiation. During the first million years after the big-bang, the gas of photons and the gas of matter were in thermal equilibrium, both having the same temperature. At a time $t_d$, lying somewhere in the range

$$t_d = 600\text{–}2 \times 10^6 \text{ years}\ , \tag{6.247}$$

and at a temperature $T_\gamma(t_d)$ in the range

$$T_\gamma(t_d) \approx 100{,}000\text{–}2{,}000 \text{ K}\ , \tag{6.248}$$

a decoupling took place and the matter-dominated phase began. From that time on only the photon gas alone has to be considered as in thermal equilibrium. The temperature of the photon gas decreased further due to the expansion of the universe and today has reached a value of 2.7 K.

This implies that if the cosmos constitutes a kind of vessel in which there is a (background) photon gas of a temperature of 2.7 K, there should be a corresponding radiation with the characteristic $\omega$-dependence $\omega^3/(\mathrm{e}^{\hbar\omega\beta} - 1)$. In particular, this radiation should be distinguished by its isotropy and its independence of any local factors or circumstances. It was predicted by Gamov in 1948, and in 1965, Dicke, Peebles, Roll, and Wilkinson took up the idea. But before they were able to set up the experimental devices, the radiation was detected by Penzias and Wilson (Nobel prize, 1978) (1965).

At a 1977 American Physical Society meeting, a small anisotropy of this radiation was reported. Tiny temperature variations in the microwave sky were measured on board a Navy U2 aircraft. The interpretation of this anisotropy is a topic of current research. The spots of slightly different temperatures are believed to be primordial seeds produced in the "big bang" and their sizes are in agreement with theories that a large percentage of the matter in the universe is "dark matter", i.e., it does not interact with light except through gravity.

## 6.6 Lattice Vibrations in Solids: The Phonon Gas

In a crystalline solid the atoms may vibrate around an equilibrium position. We want to determine the contributions of these vibrations to the energy and specific heat of a solid.

We first consider $N$ atoms in a volume $V$ of the solid within the framework of classical physics. We denote the coordinates of the atoms by $(x_1, \ldots, x_{3N})$ and the coordinates of the equilibrium positions by $(\bar{x}_1, \ldots, \bar{x}_{3N})$, so that $q_i = x_i - \bar{x}_i$, $i = 1, \ldots, 3N$ are the displacements. Furthermore, let $m$ represent the mass of an atom. The kinetic energy is then given by

$$T = \frac{1}{2} m \sum_{i=1}^{3N} \dot{x}_i^2 = \frac{1}{2} m \sum_{i=1}^{3N} \dot{q}_i^2 \,, \tag{6.249}$$

and for the potential energy $V(x_1, \ldots, x_{3N})$ an expansion around the equilibrium position yields

$$\begin{aligned} V(x_1, \ldots, x_{3N}) &= V(\bar{x}_1, \ldots, \bar{x}_{3N}) \\ &\quad + \frac{1}{2} \sum \left. \frac{\partial^2 V}{\partial x_i \, \partial x_j} \right|_{\{x_k = \bar{x}_k\}} q_i \, q_j + O(q^3) \\ &\equiv \text{const} + \frac{1}{2} \sum_{i,j=1}^{3N} \alpha_{ij} \, q_i \, q_j + O(q^3), \end{aligned} \tag{6.250}$$

with

$$\alpha_{ij} = \left. \frac{\partial^2 V}{\partial x_i \, \partial x_j} \right|_{\{x_k = \bar{x}_k\}} . \tag{6.251}$$

For small displacements, the terms $O(q^3)$ in (6.250) may be neglected. The matrix $\boldsymbol{\alpha}$ with elements $\alpha_{ij}$, $i, j = 1, \ldots, 3N$ is symmetric and therefore possesses a complete system of orthonormal eigenvectors $\left\{\boldsymbol{\phi}^{(k)},\ k = 1, \ldots, 3N\right\}$; the associated eigenvalues will be denoted by $\{m\omega_k^2, k = 1, \ldots, 3N\}$. We know that $\omega_k^2 \geq 0$, since otherwise the equilibrium position would not be stable (there would be no minimum of $V$ at $x_i = \bar{x}_i$). Furthermore, for any degree of freedom related to a symmetry of the potential we get $\omega_k^2 = 0$. For the systems under consideration these are at least the three translational and the three rotational degrees of freedom corresponding to the translation and rotation of the entire solid.

Using the eigenvectors $\{\boldsymbol{\phi}^{(k)}\}$, we introduce the normal modes $\{q'_k, k = 1, \ldots, 3N\}$ by

$$q_i = \sum_{k=1}^{3N} \phi_i^{(k)} q'_k, \quad i = 1, \ldots, 3N \; . \tag{6.252}$$

This coordinate transformation brings the Hamiltonian function

$$H(\{p_k, q_k\}) = \sum_{k=1}^{3N} \left( \frac{{p_k}^2}{2m} + \frac{1}{2} \sum_{i,j=1}^{3N} \alpha_{ij}\, q_i\, q_j \right) \tag{6.253}$$

into the form

$$H = \sum_{k=1}^{3N} \left( \frac{{p'_k}^2}{2m} + \frac{1}{2} m\omega_k^2 {q'_k}^2 \right) \; . \tag{6.254}$$

Hence, the dynamics of the lattice vibrations may be represented in the harmonic approximation (i.e., neglecting the terms $O(q^3)$ in (6.250)) by $(3N - 6)$ uncoupled, i.e., noninteracting oscillators with frequencies $\{\omega_k\}, \omega_k \neq 0$. The frequencies of these oscillators are basically determined by the nature of the interatomic interactions (i.e., the eigenvalues of the matrix $\alpha_{ij}$).

We now relate the excitations of the various normal modes to the existence of quanta or quasi-particles by reinterpreting the Hamiltonian function as a Hamiltonian operator. Hence $p'_k$ and $q'_k$ become operators in a Hilbert space with commutation relations $[p_k, q_j] = \frac{\hbar}{\mathrm{i}} \delta_{kj}$. For the linear combinations

$$a_k^+ = \frac{1}{\sqrt{2m\hbar\omega_k}}\, p'_k + \mathrm{i} \sqrt{\frac{m\omega_k}{2\hbar}}\, q'_k \tag{6.255a}$$

$$a_k = \frac{1}{\sqrt{2m\hbar\omega_k}}\, p'_k - \mathrm{i} \sqrt{\frac{m\omega_k}{2\hbar}}\, q'_k \tag{6.255b}$$

this implies

$$[a_k, a_j^+] = \delta_{kj} \; , \tag{6.256}$$

and for the Hamiltonian operator one obtains

$$H = \sum_{k=1}^{3N} \left( \hbar\omega_k\, a_k^+ a_k + \frac{1}{2}\, \hbar\omega_k \right) . \tag{6.257}$$

Obviously, $a_k^+$ generates an excitation quantum with frequency $\omega_k$. Such a quantum is also called a phonon and it is generally referred to as a quasi-particle. If $|0\rangle$ denotes the ground state of the system with energy $E_0 = \sum_{k=1}^{3N} \frac{1}{2}\, \hbar\omega_k$, then

$$|n_1, \ldots, n_{3N}\rangle = \prod_{k=1}^{3N} (a_k^+)^{n_k} |0\rangle \tag{6.258}$$

is an eigenstate of $H$ with energy

$$E(n_1, \ldots, n_{3N}) = \sum_{k=1}^{3N} n_k \hbar\omega_k + E_0 \; . \tag{6.259}$$

Equation 6.258 represents a state containing $n_1$ phonons with frequency $\omega_1$, $n_2$ phonons with frequency $\omega_2, \ldots$.

In this model the concept of a quasi-particle presents itself in a particularly transparent form: Close to the ground state the Hamiltonian may be approximated by the Hamiltonian of uncoupled harmonic oscillators. The transition to a quantum-mechanical point of view leads to creation and annihilation operators for the excitation quanta, which also may be considered as particles, or, more cautiously, as quasi-particles. In this approximation these quasi-particles do not interact (there is no coupling term present in the Hamiltonian operator), and their energy results from the form of the Hamiltonian operator. The eigenstates of $H$ are of the form (6.258), where the quantum numbers can be interpreted as occupation numbers.

Hence, we find a similar situation to that met in the case of the photon gas. There, however, the frequency of a photon created, e.g., by $a^+(\boldsymbol{k}, \varepsilon)$, is known to be $\omega(\boldsymbol{k}) = |\boldsymbol{k}|c$. Furthermore, there exist infinitely many degrees of freedom in the case of photons, while in the case of phonons there are only $3N$ (or, more precisely, $3N-6$). On the other hand, phonons are also bosons, as more than one phonon may occupy a state $k$.

In order to determine further properties of phonons and to formulate the partition function for a gas of phonons, we have to make an ansatz for the number $f(\omega)\,\mathrm{d}\omega$ of frequencies $\{\omega_k\}$ in the frequency interval $(\omega, \omega + \mathrm{d}\omega)$. Obviously, there now has to be a maximum frequency $\omega_\mathrm{m}$, since the corresponding wavelength cannot be smaller than the distance between atoms in the lattice; a smaller wavelength would be meaningless for lattice vibrations.

Since only $3N$ frequencies are available, we must have

$$\int_0^{\omega_\mathrm{m}} f(\omega)\,\mathrm{d}\omega = 3N \; . \tag{6.260}$$

A proposal first made by Einstein is

$$f(\omega) = 3N\,\delta(\omega - \omega_E) \; , \tag{6.261}$$

where $\omega_E$ denotes a mean frequency. However, the consequences drawn from this ansatz do not agree with experimental results, in particular concerning the behavior of $C_V$ at low temperatures.

In 1912, Debye suggested the form

$$f(\omega) = \begin{cases} \dfrac{3V}{2\pi^2 c^3}\,\omega^2, & 0 \le \omega \le \omega_\mathrm{m} \\ 0\,, & \text{otherwise} \end{cases} \; , \tag{6.262}$$

where $c$ is the average sound velocity in the solid and $\omega_\mathrm{m}$ is determined from the normalization (6.260). From

$$3N = \int_0^{\omega_\mathrm{m}} f(\omega)\,\mathrm{d}\omega = \frac{3V}{2\pi^2 c^3}\,\frac{\omega_\mathrm{m}^3}{3} \tag{6.263}$$

it follows that

$$\omega_\mathrm{m} = c\left(\frac{6\pi^2}{v}\right)^{1/3}, \tag{6.264}$$

and thus we find for the minimum wavelength:

$$\lambda_\mathrm{m} = \frac{2\pi c}{\omega_\mathrm{m}} = \left(\frac{4\pi}{3}v\right)^{1/3} \approx 1.6 v^{1/3}\,. \tag{6.265}$$

The minimal wavelength is therefore of the order of the atomic distance, as expected.

The Debye form may be justified as follows: If the wavelength is large compared to the atomic distance in the lattice, i.e., at small frequencies, the lattice structure is of no great importance and we can use the continuum approximation for the elastic medium. According to this theory the displacement $\boldsymbol{q}(\boldsymbol{r},t)$ at the position $\boldsymbol{r}$ satisfies (Landau and Lifschitz 1959)

$$\varrho\,\frac{\partial^2 \boldsymbol{q}}{\partial t^2} = \mu\Delta\boldsymbol{q} + (\lambda+\mu)\nabla(\nabla\cdot\boldsymbol{q})\,, \tag{6.266}$$

where $\varrho$ denotes the density, and $\lambda, \mu$ are the Lamé elastic constants.

The solutions of this equation are

$$\boldsymbol{q} = \boldsymbol{a}\,\exp\left(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r} - \mathrm{i}\omega t\right)\,. \tag{6.267}$$

Here $\boldsymbol{k}$ denotes the direction of propagation of the wave and $\boldsymbol{a}$ the direction of the displacement: $\boldsymbol{a} \parallel \boldsymbol{k}$ means longitudinal polarization, $\boldsymbol{a} \perp \boldsymbol{k}$ transverse polarization.

For the frequency $\omega(\boldsymbol{k})$ one finds

$$\omega(\boldsymbol{k}) = \omega_\mathrm{t} = c_\mathrm{t}\,|\boldsymbol{k}|, \quad \text{for} \quad \boldsymbol{a} \perp \boldsymbol{k}\,, \tag{6.268a}$$

$$\omega(\boldsymbol{k}) = \omega_\ell = c_\ell\,|\boldsymbol{k}|, \quad \text{for} \quad \boldsymbol{a} \parallel \boldsymbol{k}\,, \tag{6.268b}$$

where $c_\mathrm{t} = \sqrt{\mu/\varrho}$ is the transverse sound velocity and $c_\ell = \sqrt{(2\mu+\lambda)/\varrho}$ the corresponding longitudinal velocity.

Hence, we are led to the following picture: The states are determined by the three-dimensional wave-vector $\boldsymbol{k}$ and the polarization. There are three possible polarizations (two transverse and one longitudinal; in the case of photons there are only two transverse polarizations). Introducing for simplicity a common sound velocity $c$, we always have $\omega = c|\boldsymbol{k}|$ as in the case of photons, but contrary to the case of photons there is now an upper limit for $|\boldsymbol{k}|$ or $\omega$, as there exists a minimum wavelength for lattice vibrations. A lower limit for $|\boldsymbol{k}|$ and thus an upper limit for the wavelength is given simply by the macroscopic length $L \propto V^{1/3}$ ($|\boldsymbol{k}|_{\text{min}} \propto 1/L$), as is the case for photons.

If we again assume $V$ to be large enough, so that the values for $\boldsymbol{k}$ may be considered as continuous, a summation over all states leads to

$$g \sum_{\boldsymbol{k}} \to 3 \frac{V}{(2\pi)^3} \int \mathrm{d}k 4\pi k^2 \tag{6.269}$$

so that we finally obtain for the number of frequencies in the interval $(\omega, \omega + \mathrm{d}\omega)$

$$f(\omega)\,\mathrm{d}\omega = \frac{3V}{(2\pi)^3} 4\pi\, k^2\, \mathrm{d}k\ , \tag{6.270}$$

and with $|\boldsymbol{k}| = \omega/c$,

$$f(\omega)\,\mathrm{d}\omega = \frac{3\,V}{2\pi^2 c^3} \omega^2\, \mathrm{d}\omega\ , \tag{6.271}$$

which is exactly the Debye form.

Hence, this ansatz is a direct consequence of the assumption that for large wavelengths, i.e., small frequencies, the classical wave-equation for lattice vibrations is valid. In particular, this implies that Debye's ansatz is adequate for small values of $\omega$. The lower the temperature in the solid, the more important are these thermal low frequency excitations. Hence, particularly at low temperatures, the predictions from Debye's ansatz should be in good agreement with the experimental data, e.g., for $C_V$.

*Remark.* The dispersion relation $\omega = c|\boldsymbol{k}|$ is also called the acoustic branch, because it results from the treatment of sound waves in a solid. If the distances between the equilibrium positions of neighboring atoms in a solid are not all equal (e.g. if there are two ions in each primitive cell of a Bravais lattice), one obtains further relations for $\omega$ as a function of $\boldsymbol{k}$. In this case a so-called optical branch may appear, where the name refers to the fact that the corresponding phonons can interact with electromagnetic waves. The exact form of this relation also has an influence on the optical properties of the solid (Ashcroft and Mermin 1976).

As was the case for photons, the average occupation number of the state $\boldsymbol{k}, \alpha$ is again given by

$$\langle n(\boldsymbol{k}, \alpha) \rangle = \frac{1}{\exp\left(\frac{\hbar\omega(\boldsymbol{k})}{k_{\mathrm{B}} T}\right) - 1}\ . \tag{6.272}$$

We will not discuss the expectation value of the number of phonons; anyhow, the letter $N$ is already occupied by the number of atoms in a solid. For the energy of a phonon gas, we obtain

$$E = \int_0^{\omega_\mathrm{m}} \mathrm{d}\omega\, f(\omega)\, \frac{\hbar\omega}{\mathrm{e}^{\hbar\omega/k_\mathrm{B}T} - 1} \tag{6.273}$$

$$= \frac{3V\hbar}{2\pi^2 c^3} \int_0^{\omega_\mathrm{m}} \mathrm{d}\omega\, \frac{\omega^3}{\mathrm{e}^{\hbar\omega/k_\mathrm{B}T} - 1} \tag{6.274}$$

$$= \frac{3V\hbar}{2\pi^2 c^3} \left(\frac{k_\mathrm{B}T}{\hbar}\right)^4 \int_0^{\hbar\omega_\mathrm{m}/k_\mathrm{B}T} \mathrm{d}t\, \frac{t^3}{\mathrm{e}^t - 1} \tag{6.275}$$

$$= \frac{3Vk_\mathrm{B}T}{2\pi^2 c^3} \left(\frac{k_\mathrm{B}T}{\hbar\omega_\mathrm{m}}\right)^3 c^3\, \frac{6\pi^2}{v} \int_0^{\hbar\omega_\mathrm{m}/k_\mathrm{B}T} \mathrm{d}t\, \frac{t^3}{\mathrm{e}^t - 1} \tag{6.276}$$

$$= 3k_\mathrm{B}TN\, D(x), \tag{6.277}$$

where

$$D(x) = \frac{3}{x^3} \int_0^x \mathrm{d}t\, \frac{t^3}{\mathrm{e}^t - 1}, \tag{6.278}$$

$$x = \frac{\hbar\omega_\mathrm{m}}{k_\mathrm{B}T} = \frac{T_\mathrm{D}}{T}, \qquad T_\mathrm{D} = \frac{\hbar\omega_\mathrm{m}}{k_\mathrm{B}}\ . \tag{6.279}$$

Here $D(x)$ is called the Debye function and $T_\mathrm{D}$ is the Debye temperature. For $D(x)$ one finds

$$D(x) = \begin{cases} 1 - \frac{3}{8}x + \frac{1}{20}x^2 + \ldots & \text{for} \quad x \ll 1\ , \quad \text{i.e., } T \gg T_\mathrm{D} \\ \frac{\pi^4}{5x^3} + O\left(\mathrm{e}^{-x}\right) & \text{for} \quad x \gg 1\ , \quad \text{i.e., } T \ll T_\mathrm{D} \end{cases}\ . \tag{6.280}$$

Hence, we obtain

- For $T \ll T_\mathrm{D}$,

$$E = 3Nk_\mathrm{B}T \left(\frac{\pi^4}{5}\, \frac{T^3}{T_\mathrm{D}^3} + O\!\left(\mathrm{e}^{-T_\mathrm{D}/T}\right)\right) \tag{6.281}$$

  and therefore

$$C_V = \frac{12}{5}\,\pi^4 Nk_\mathrm{B} \left(\frac{T}{T_\mathrm{D}}\right)^3 + \ldots\ . \tag{6.282}$$

  At low temperatures we thus observe a $T^3$-law. Responsible for this $T^3$-dependence of the specific heat and for the $T^4$-dependence of the energy is obviously the $\omega^2$-dependence of $f(\omega)$ for $\omega \to 0$.

- For $T \gg T_\mathrm{D}$,

$$E = 3Nk_\mathrm{B}T \left(1 - \frac{3}{8}\,\frac{T_\mathrm{D}}{T} + \frac{1}{20}\left(\frac{T_\mathrm{D}}{T}\right)^2 + \ldots\right) \tag{6.283}$$

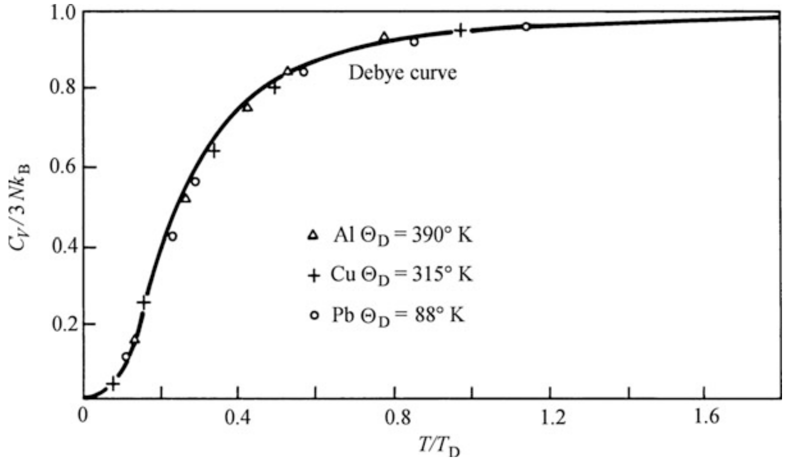


**Fig. 6.10** Illustration of the validity of the expression for the specific heat derived from the theory of Debye (From Mandl (1971), reprinted by permission of John Wiley & Sons, Inc.)

and therefore

$$C_V = 3Nk_B - \frac{3}{20}Nk_B\left(\frac{T_D}{T}\right)^2 + \ldots . \tag{6.284}$$

The relation $C_V = 3Nk$ corresponds to the law of Dulong and Petit. Hence, for large temperatures each quadratic term in the Hamiltonian function leads to a contribution of $\frac{1}{2}\,k_B T$ to the energy.

This result for large temperatures is even independent of the exact form of the frequency distribution and follows merely from its normalization. Indeed, for $T \gg T_D$ we have $\hbar\omega/k_B T < \hbar\omega_m/k_B T \equiv T_D/T \ll 1$ and therefore

$$E = \int_0^{\omega_m} d\omega\, f(\omega)\,\frac{\hbar\omega}{\hbar\omega/k_B T} = k_B T \int_0^{\omega_m} d\omega\, f(\omega) = 3Nk_B T\ . \tag{6.285}$$

For the full range of temperatures, the specific heat determined from (6.277) agrees well with the experiments (see Fig. 6.10).

The $T^3$-dependence of the specific heat at low temperatures is in excellent agreement with experimental data. A fit of (6.282) to the data leads to an estimate of the Debye temperature $T_D$ (Table 6.1). For many solids the Debye temperatures are in the range 100–300 K.

There are other procedures for determining the Debye temperature. One rather disputable practice is to infer from the experimental data for $C_V$ at given temperature $T$ a corresponding Debye temperature using the equation for $C_V$ which follows from (6.277). In this way one obtains a temperature-dependent Debye temperature

**Table 6.1** Debye temperatures for various solids (From Ashcroft and Mermin 1976)

| Solid | $T_{\mathrm{D}}\,(K)$ | Solid | $T_{\mathrm{D}}\,(K)$ |
|---|---|---|---|
| Na | 150 | Fe | 420 |
| K | 100 | Co | 385 |
| Cu | 315 | Ni | 375 |
| Ag | 215 | Al | 394 |
| Au | 170 | Ge | 360 |
| Be | 1,000 | Sn (grey) | 260 |
| Mg | 318 | Pb | 88 |
| Zn | 234 | Pt | 230 |
| Cd | 120 | C (diamond) | 1,860 |

$T_{\mathrm{D}}(T)$. A more reasonable method is to infer $T_{\mathrm{D}}$ from an overall fit of $C_V(T)$ to the data.

Hence, the values for $T_{\mathrm{D}}$ given in the literature differ to some extent depending on the method used to derive them (see also Ashcroft and Mermin 1976).

## 6.7 Systems with Internal Degrees of Freedom: Ideal Gases of Molecules

Up to now we have studied gases of monatomic species with only the kinetic energy contributing to the energy of the atom. For polyatomic molecules, and even for monatomic molecules when electronic excitations are included, one obtains a single-particle energy of the form

$$\varepsilon(\boldsymbol{k}, n) = \frac{\hbar^2 \boldsymbol{k}^2}{2m} + \varepsilon_{\mathrm{int}}(n) \; . \tag{6.286}$$

Here, $n$ stands for all quantum numbers other than the wave-vector $\boldsymbol{k}$, and $\varepsilon_{\mathrm{int}}(n)$ is the contribution of the additional internal degrees of freedom to the energy as a function of $n$.

Such quantum numbers and energy contributions include:

- The two atoms of a diatomic molecule may vibrate around an equilibrium separation. To a first approximation one obtains harmonic vibrations with a frequency $\omega$, where $\omega$ is determined by the interaction between the two atoms. $\varepsilon_{\mathrm{int}}(n)$ now corresponds to the energy spectrum of a harmonic oscillator

$$\varepsilon_{\mathrm{int}}(n) = \hbar\omega \left(n + \frac{1}{2}\right) \; , \quad n = 0, 1, \dots \; . \tag{6.287}$$

- Diatomic molecules also may rotate around the connecting axis as well as around the two orthogonal axes. Let the $z$-axis of a coordinate system be along the connecting line, then the moments of inertia are $T_{zz} = 0$, $T_{xx} = T_{yy} \equiv I = \mu r_0^2$, with $\mu$ the reduced mass and $r_0$ the equilibrium distance. From this follows for the energy of the rotational movement $E_{\text{rot}} = \frac{1}{2}\, I(\omega_1^2 + \omega_2^2)$, or, introducing the angular momentum $\boldsymbol{L} = I\,(\omega_1, \omega_2, 0)$,

$$E_{\text{rot}} = \frac{\boldsymbol{L}^2}{2I} \; . \tag{6.288}$$

A quantum-mechanical treatment then leads to

$$\varepsilon_{\text{int}}(\ell) = \frac{\hbar^2\, \ell(\ell+1)}{2I} \, , \quad \ell = 0, 1, \ldots \; . \tag{6.289}$$

- Finally, we can take into account the possibility of electronic excitations, or the potential presence of nuclear spin.

From (6.61) we obtain the grand canonical thermodynamical potential

$$K(T, V, \mu) = \mp \frac{gVk_{\text{B}}T}{(2\pi\hbar)^3} \int \mathrm{d}^3p \sum_n \ln\left[1 \pm z \exp -\beta\left(\frac{p^2}{2m} + \varepsilon_{\text{int}}(n)\right)\right] \, , \tag{6.290}$$

where the upper sign refers to fermions, the lower one to bosons. The evaluation of this expression, however, is in general quite difficult. As before, we will therefore confine ourselves to the case $z \ll 1$. Expanding the logarithm around $z = 0$ and keeping only the first term of this expansion we immediately obtain

$$K(T, V, \mu) = -k_{\text{B}}T\, \frac{zgV}{(2\pi\hbar)^3} \int \mathrm{d}^3p\, \exp\left(-\beta\frac{p^2}{2m}\right) \sum_n \mathrm{e}^{-\beta\varepsilon_{\text{int}}(n)} \tag{6.291}$$

$$= -\frac{zgV}{\lambda_{\text{t}}^3}\, Z_{\text{int}}\, k_{\text{B}}T, \tag{6.292}$$

where, as in (6.62) and (6.66), we have introduced the thermal de Broglie wavelength. The expression

$$Z_{\text{int}} = \sum_n \mathrm{e}^{-\beta\varepsilon_{\text{int}}(n)} \tag{6.293}$$

represents a kind of partition function for the internal degrees of freedom. Obviously, $Z_{\text{int}} \geq 1$, if the reference point for the energy is chosen in such a way that the lowest possible energy value is zero.

In complete analogy to the classical approximation we now get

$$N = -\frac{\partial K(T, V, \mu)}{\partial \mu} = \frac{gV}{\lambda_{\text{t}}^3}\, z\, Z_{\text{int}} \tag{6.294}$$

and therefore

$$z = \frac{\lambda_{\mathrm{t}}^3}{gv}\,\frac{1}{Z_{\mathrm{int}}}. \tag{6.295}$$

The condition

$$z \ll 1 \tag{6.296}$$

also implies

$$\lambda_{\mathrm{t}}^3 \ll Z_{\mathrm{int}} gv\;, \tag{6.297}$$

i.e., the thermal de Broglie wavelength is much smaller than the average distance between the particles. With respect to the translational motion we treat the gas in this approximation again as a classical gas. The internal degrees of freedom, however, are dealt with quantum-mechanically. If we further use $K = -pV$, we obtain from (6.292) and (6.294) the equation of state

$$pV = Nk_{\mathrm{B}}T, \tag{6.298}$$

which is therefore valid independently of the existence of internal degrees of freedom.

On the other hand, from (6.83) we get for the energy

$$E(T, V, \mu) = -\frac{\partial}{\partial\beta}\,(-\beta K) + \mu N \tag{6.299}$$

$$= -\frac{\partial}{\partial\beta}\left(\frac{zgV}{\lambda_{\mathrm{t}}^3}\,Z_{\mathrm{int}}\right) + \mu N, \tag{6.300}$$

and therefore

$$E(T, V, N) = \frac{3}{2}\,Nk_{\mathrm{B}}T - N\,\frac{\partial \ln Z_{\mathrm{int}}}{\partial\beta}\;, \tag{6.301}$$

since $\frac{\partial z}{\partial\beta} = \mu z$ and $-\frac{\partial}{\partial\beta}\,\lambda_{\mathrm{t}}^{-3} = \frac{3}{2}\,k_{\mathrm{B}}T\,\lambda_{\mathrm{t}}^{-3}$. Here we have also made use of (6.294), i.e., $(zgV/\lambda_{\mathrm{t}}^3)Z_{\mathrm{int}} = N$.

Next we have to study the partition function for the internal degrees of freedom $Z_{\mathrm{int}}$. In the rigid rotor-harmonic oscillator approximation one considers the angular motion of the diatomic molecule to be that of a rigid dumbbell of fixed internuclear distance, so that the Hamiltonian of the relative motion can be regarded as superposition of the rotational and vibrational Hamiltonians. Then the internal partition function can be written as a product of two individual contributions,

$$Z_{\mathrm{int}} = Z_{\mathrm{vib}}\,Z_{\mathrm{rot}}\;, \tag{6.302}$$

with

$$Z_{\mathrm{vib}} = \sum_{n=0}^{\infty} \exp\left(-\beta\hbar\omega\left(n + \frac{1}{2}\right)\right), \tag{6.303a}$$

$$Z_{\mathrm{rot}} = \sum_{\ell=0}^{\infty} (2\ell + 1)\,\exp\left(-\beta\frac{\hbar^2\ell(\ell+1)}{2I}\right)\;, \tag{6.303b}$$

where in $Z_{\text{rot}}$ we have also taken into account the $(2\ell + 1)$-fold degeneracy of a quantum state for a given quantum number $\ell$ for the angular momentum.

For the individual terms we obtain, firstly, neglecting the ground state energy

$$Z_{\text{vib}} = \sum_{n=0}^{\infty} \mathrm{e}^{-\beta\hbar\omega n} = \frac{1}{1 - \mathrm{e}^{-\beta\hbar\omega}}, \tag{6.304}$$

and therefore the contribution to the internal energy is

$$E_{\text{vib}} = -N \frac{\partial \ln Z_{\text{vib}}}{\partial \beta} = N\hbar\omega \left( \frac{1}{\mathrm{e}^{\beta\hbar\omega} - 1} \right) . \tag{6.305}$$

Introducing the temperature $\theta_{\text{v}}$, defined by $\hbar\omega = k\theta_{\text{v}}$, we get:

$$E_{\text{vib}} = Nk_{\text{B}}\theta_{\text{v}} \left( \frac{1}{\mathrm{e}^{\theta_{\text{v}}/T} - 1} \right), \tag{6.306}$$

i.e.,

$$E_{\text{vib}} = Nk_{\text{B}} \begin{cases} \theta_{\text{v}}\, \mathrm{e}^{-\theta_{\text{v}}/T} + \dots & \text{for } T \ll \theta_{\text{v}}, \\ T + \dfrac{1}{12} \dfrac{\theta_{\text{v}}^2}{T} + \dots & \text{for } T \gg \theta_{\text{v}} . \end{cases} \tag{6.307}$$

Thus the contribution of the vibrations to the specific heat is

$$C_{V,\text{vib}} = Nk_{\text{B}} \begin{cases} \left( \dfrac{\theta_{\text{v}}}{T} \right)^2 \mathrm{e}^{-\theta_{\text{v}}/T} + \dots & \text{for } T \ll \theta_{\text{v}}, \\ 1 - \dfrac{1}{12} \left( \dfrac{\theta_{\text{v}}}{T} \right)^2 + \dots & \text{for } T \gg \theta_{\text{v}} . \end{cases} \tag{6.308}$$

With two degrees of freedom for each oscillator we therefore obtain $\frac{1}{2}k_{\text{B}}$ per degree of freedom as the contribution of the vibrations to $C_V$ at sufficiently high temperatures ($T \gg \theta_{\text{v}}$).

Secondly, for the rotational contribution, we have

$$Z_{\text{rot}} = \sum_{\ell=0}^{\infty} (2\ell + 1) \exp\left( -\beta \frac{\hbar^2 \ell(\ell + 1)}{2I} \right) . \tag{6.309}$$

Defining the temperature $\theta_{\text{r}}$ by $k_{\text{B}}\theta_{\text{r}} = \hbar^2/I$, we obtain after some calculations for the contribution to the energy for $T \ll \theta_{\text{r}}$

$$E_{\text{rot}} = 3Nk_{\text{B}}\theta_{\text{r}}\, \mathrm{e}^{-\theta_{\text{r}}/T} + \dots , \tag{6.310}$$

and for $C_V$:

$$C_V = 3Nk_{\mathrm{B}}\left(\frac{\theta_{\mathrm{r}}}{T}\right)^2 \mathrm{e}^{-\theta_{\mathrm{r}}/T} + \dots \,. \tag{6.311}$$

For $T \gg \theta_{\mathrm{r}}$ the sum may be computed from the Euler–MacLaurin sum formula (6.124), where we can neglect the contributions from the upper limit and obtain

$$\sum_{\ell=0}^{\infty} f(\ell) = \int_0^{\infty} \mathrm{d}x\, f(x) + \frac{1}{2} f(0) - \frac{1}{12} f'(0) + \frac{1}{720} f'''(0) + \dots, \tag{6.312}$$

and therefore for $T \gg \theta_{\mathrm{r}}$:

$$Z_{\mathrm{rot}} = \frac{2T}{\theta_{\mathrm{r}}}\left(1 + \frac{1}{6}\frac{\theta_{\mathrm{r}}}{T} + \frac{1}{60}\left(\frac{\theta_{\mathrm{r}}}{T}\right)^2 + \dots\right) . \tag{6.313}$$

Hence, the contribution to the energy for $T \gg \theta_{\mathrm{r}}$ is

$$E_{\mathrm{rot}} = Nk_{\mathrm{B}}T - \frac{1}{6}Nk_{\mathrm{B}}\theta_{\mathrm{r}} - \frac{1}{180}Nk_{\mathrm{B}}\frac{\theta_{\mathrm{r}}^2}{T} + \dots \,, \tag{6.314}$$

and for the specific heat $C_V$ the contribution for $T \gg \theta_{\mathrm{r}}$ is

$$C_{V,\mathrm{rot}} = Nk_{\mathrm{B}} + O\left(\frac{1}{T^2}\right) . \tag{6.315}$$

Again we find $\frac{1}{2}k_{\mathrm{B}}$ per degree of freedom at large temperatures.

Finally, we investigate the electronic contribution. It is given by,

$$Z_{\mathrm{el}} = g_1 + g_2\,\mathrm{e}^{-\beta\varepsilon_2} \,, \tag{6.316}$$

where we have taken the ground state energy of the atom to be zero, i.e., $\varepsilon_2$ corresponds to the first excited state. $g_1$ and $g_2$ denote the respective degeneracies. Now, $\varepsilon_2$ is of the order of some eV. Setting $\varepsilon_2 = k_{\mathrm{B}}\theta_{\mathrm{el}}$ we obtain for the leading temperature-dependent term:

$$g_2\,\mathrm{e}^{-\theta_{\mathrm{el}}/T} \,, \tag{6.317}$$

which may almost always be neglected, since $\theta_{\mathrm{el}}/T \gg 1$.

We list some values for the specific temperatures $\theta_{\mathrm{r}}$, $\theta_{\mathrm{v}}$, and $\theta_{\mathrm{el}}$ for some molecules in Table 6.2.

$\theta_{\mathrm{r}}$ is obviously much smaller, and $\theta_{\mathrm{v}}$ in general much larger than room temperature. Hence, at this temperature only the rotational degrees of freedom are excited and therefore the total specific heat is $C_V = \frac{5}{2}Nk_{\mathrm{B}}T$.

*Remarks.*

- In the case of polyatomic molecules we can take care of the rotational degrees of freedom by treating the molecules as rigid bodies. Rigid bodies have three

**Table 6.2** Specific temperature (in K) of selected molecules

| Molecule | $\theta_r$ | $\theta_v$ | $\theta_{el}$ | Molecule | $\theta_r$ |
|---|---|---|---|---|---|
| $H_2$ | 85 | 5,958 | 129,000 | CO | 2.77 |
| $D_2$ | 43 | 4,210 | 129,000 | NO | 2.45 |
| $O_2$ | 2.1 | 2,228 | 11,000 | HCl | 15.02 |
| $I_2$ | 0.05 | 305 | 17,000 | HBr | 12.02 |
| $K_2$ | 0.08 | 132 |  | HI | 9.06 |

principal moments of inertia. If all three of them are equal, one speaks of a spherical top, if two are equal, this is known as a symmetric top, otherwise it is called an asymmetric top. The molecules $CH_4$ and $CCl_4$, for example, may be considered as spherical tops, $CH_3Cl$ and $NH_3$ as symmetric tops, and $H_2O$, and $NO_2$ as asymmetric tops.

The Hamiltonian operator for such a rigid body can be derived from the corresponding Lagrange function or the corresponding Hamiltonian function in classical mechanics (see e.g. Honerkamp and Römer 1993). For the symmetric top with $I_1 = I_2 \neq I_3$ one obtains, for example,

$$H = -\frac{\hbar^2}{2I_1}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2} - 2\frac{\cos\theta}{\sin^2\theta}\frac{\partial^2}{\partial\varphi\partial\psi} + \left(\frac{I_1}{I_3} + \frac{\cos^2\theta}{\sin^2\theta}\right)\frac{\partial^2}{\partial\psi^2}\right]. \tag{6.318}$$

Here, $\theta$, $\varphi$ and $\psi$ are the Euler angles describing the orientation of the rotator with respect to a fixed coordinate system. The eigenfunctions of $H$ are (see e.g. Dawydow 1965)

$$D^J_{MK}(\varphi,\psi,\theta) = \frac{1}{2\pi}\,\mathrm{e}^{\mathrm{i}(M\varphi+K\psi)}\,d^J_{MK}(\theta) \tag{6.319}$$

with corresponding eigenvalues

$$\varepsilon_{\text{int}} = \frac{\hbar^2}{2I_1}J(J+1) + \frac{\hbar^2}{2}\left(\frac{1}{I_3} - \frac{1}{I_1}\right)K^2\,. \tag{6.320}$$

$\{D^J_{MK}(\theta)\}$ denote the generalized spherical harmonics, where $K$ and $M$ may assume integer values between $-J$ and $+J$.

For a linear molecule the degree of freedom associated to the $\psi$-variable is absent, i.e., $K \equiv 0$ and one obtains

$$D^J_{M0} \propto Y_{JM}(\theta,\varphi) \quad \text{and} \quad \varepsilon_i = \frac{\hbar^2}{2I}J(J+1)\,. \tag{6.321}$$

In this case the generalized harmonic functions reduce to the spherical harmonics $\{Y_{JM}(\theta,\varphi)\}$ and the degeneracy is $(2J+1)$-fold.

For the spherical top ($I_1 = I_2 = I_3 \equiv I$) we have $\varepsilon_{\text{int}} = \frac{\hbar^2}{2I}\, J(J+1)$, as for the linear molecule, but now the degeneracy is $(2J+1)^2$-fold. In this case we therefore obtain

$$Z_{\text{rot}} = \sum_{J=0}^{\infty} (2J+1)^2 \exp\left(-\frac{\hbar^2}{2I}\,\beta\, J(J+1)\right), \tag{6.322}$$

and for high temperatures $T \gg \theta_{\text{r}}$ we get

$$Z_{\text{rot}} = \int_0^{\infty} \mathrm{d}J\, (4J^2) \exp\left(-\frac{\hbar^2}{2I}\,\beta\, J^2\right) \propto \left(\frac{T}{\theta_{\text{r}}}\right)^{3/2} . \tag{6.323}$$

So we find $C_V = \frac{3}{2}\, N k_{\text{B}}$ for $T \gg \theta_{\text{r}}$, i.e., again a contribution of $\frac{1}{2} k_{\text{B}}$ for each degree of freedom. (More detailed treatments may be found e.g. in McQuarrie (1976) and Diu et al. (1994).)

- If both nuclei are equal, the wavefunction has to be totally antisymmetric for fermionic nuclei and totally symmetric for bosonic nuclei. The entire wavefunction is a product

$$\psi = \psi_{\text{trans}}(\boldsymbol{x}_1,\boldsymbol{x}_2)\psi_{\text{rot}}(\boldsymbol{x}_1,\boldsymbol{x}_2)\psi_{\text{vib}}(\boldsymbol{x}_1,\boldsymbol{x}_2)\chi(1,2),$$

where $\chi(1,2)$ denotes the spin wavefunction of the nuclear spin. $\psi_{\text{trans}}$ and $\psi_{\text{vib}}$ are a priori symmetric under the replacement $\boldsymbol{x}_1 \leftrightarrow \boldsymbol{x}_2$, as $\psi_{\text{vib}}$ only depends on the displacement from the equilibrium position, which, however, is invariant under an exchange of the particles. Hence, only $\psi_{\text{rot}}$ and $\chi(1,2)$ have to be taken care of.

Consider the molecule $H_2$, for which the nuclei are fermions. If

- $\chi(1,2)$ is symmetric, $\psi_{\text{rot}}$ has to be antisymmetric: $J \equiv 1(2)$;
- $\chi(1,2)$ is antisymmetric, $\psi_{\text{rot}}$ has to be symmetric: $J \equiv 0(2)$.

The first case corresponds to $S = 1$ (ortho-state, the orientations of the spins are parallel, $\uparrow\uparrow$), and the second case to $S = 0$ (para-state, the orientations of the spins are antiparallel, $\uparrow\downarrow$). The respective sums over $J$ in $K_{\text{rot}}$ only run over the odd (even) values of $J$.

Therefore

$$Z_{\text{rot,para}} = \sum_{J=0,2,4,\ldots} (2J+1) \exp\left(-\frac{J(J+1)\theta_{\text{r}}}{2T}\right) \tag{6.324a}$$

$$Z_{\text{rot,ortho}} = 3 \sum_{J=1,3,5,\ldots} (2J+1) \exp\left(-\frac{J(J+1)\theta_{\text{r}}}{2T}\right). \tag{6.324b}$$

The typical time scales for the transformation from orthohydrogen into parahydrogen are quite large such that the $H_2$ gas may be regarded as a mixture of two different kinds of molecules which are not transformed into each other, and we are allowed to treat the particle numbers $N_\mathrm{p}$ and $N_\mathrm{o}$ as fixed. Under these conditions

$$C_{V,\mathrm{rot}} = \frac{N_\mathrm{o}}{N_\mathrm{o} + N_\mathrm{p}} C_{V,\mathrm{rot,ortho}} + \frac{N_\mathrm{p}}{N_\mathrm{o} + N_\mathrm{p}} C_{V,\mathrm{rot,para}} \ . \tag{6.325}$$

For an extensive discussion of ideal polyatomic gases see, e.g., McQuarrie (1976).

## 6.8 Magnetic Properties of Fermi Systems

Substances in external magnetic fields may display quite diverse behavior. Those that exhibit a magnetization in the opposite direction to the external field are called diamagnetic; their magnetic susceptibility is therefore negative. Among the materials with positive magnetic susceptibility one distinguishes between paramagnetic substances and ferromagnetic substances. The latter are characterized by a spontaneous magnetization.

Diamagnetic properties of a substance result from the influence of the external field on the motion of the charged particles in the substance. Electrons play the dominant role in this case, as they possess a considerably lower mass than the nuclei. Paramagnetic properties have their origin in a permanent magnetic moment of the electrons, atoms, or molecules.

### *6.8.1 Diamagnetism*

The Hamiltonian of a nonrelativistic electron in an external magnetic field $\boldsymbol{B}$ may be written as

$$H = \frac{1}{2m} \left( \boldsymbol{p} - \frac{e}{c} \boldsymbol{A} \right)^2 \ . \tag{6.326}$$

Here $\boldsymbol{A}$ is the vector potential related to $\boldsymbol{B}$ according to $\boldsymbol{B} = \nabla \times \boldsymbol{A}$. It has been proven by van Leeuwen that, in the framework of classical statistical mechanics, the partition function including the vector potential may be transformed into a partition function without the vector potential, so that diamagnetism cannot be explained classically. Landau has shown that it is the quantization of the orbits of charged particles in an external magnetic field which is responsible for diamagnetism. We briefly present this theory of Landau.

For a homogeneous magnetic field $\boldsymbol{B}$ along the $z$-axis the vector potential $\boldsymbol{A}$ may be chosen in the form

$$\boldsymbol{A} = (-By, 0, 0) \ . \tag{6.327}$$

The Hamiltonian operator now reads

$$H = \frac{1}{2m}\Big((p_x + \frac{e}{c}By)^2 + p_y^2 + p_z^2\Big) \tag{6.328}$$

and with the ansatz

$$\Psi(x, y, z) = \mathrm{e}^{\mathrm{i}(k_x x + k_z z)} f(y) \tag{6.329}$$

for the eigenfunction, one obtains, with $\omega_\mathrm{c} = eB/mc$ and $y_0 = \hbar k_x c/eB$, for $f(y)$ the eigenvalue equation

$$\Big[\frac{1}{2m}p_y^2 + \frac{m\omega_\mathrm{c}^2}{2}(y + y_0)^2\Big] f(y) = \varepsilon' f(y)\,, \tag{6.330}$$

which corresponds to the eigenvalue equation for a harmonic oscillator. The eigenstates are therefore characterized by the wave numbers $k_x, k_z$ and by the quantum numbers $j,\ j = 0, 1, 2, \ldots$ of a harmonic oscillator. The energy eigenvalues (also called Landau levels) read

$$\varepsilon(k_z, j, B) = \frac{\hbar^2 k_z^2}{2m} + \hbar\omega_\mathrm{c}\left(j + \frac{1}{2}\right) \tag{6.331}$$

and do not depend on $k_x$. Because of boundary conditions, again, $k_x = 2\pi n/L$ with $n = 1, \ldots, g$, where $L$ denotes the length of the edges of a cubic box containing the electron gas. The maximum number $g$ is determined by the condition that the shift $y_0$ must be less or equal to $L$. Obviously, $g$ is the degree of degeneracy, i.e., the number of states for given energy, and $g$ is given by

$$y_0 \equiv \frac{\hbar k_x c}{eB} = \frac{\hbar c}{eB}\frac{2\pi g}{L} = L\,, \tag{6.332}$$

hence

$$g = \frac{eB}{hc}L^2 = \frac{m\omega_\mathrm{c}}{h}L^2\,. \tag{6.333}$$

The movement of the electrons may be considered as a superposition of a free motion in $z$-direction and a circular motion on quantized circles in the $x$-$y$ plane. The degeneracy of the energy levels proportional to $L^2$ reflects the fact that the energy of the circular orbit in the $x$-$y$ plane does not depend on the location of its center.

This provides us with all the expressions needed to compute the particle number, energy, and magnetization. With $p = \hbar k_z$ we obtain

$$N(T, V, \mu) = \frac{2gL}{h}\int_{-\infty}^{\infty} \mathrm{d}p \sum_{j=0}^{\infty} \frac{1}{\mathrm{e}^{\beta(\varepsilon(p,j,B)-\mu)} + 1}\,, \tag{6.334}$$

and

$$E(T, V, \mu) = \frac{2gL}{h} \int_{-\infty}^{\infty} \mathrm{d}p \sum_{j=0}^{\infty} \frac{\varepsilon(p, j)}{\mathrm{e}^{\beta(\varepsilon(p,j,B)-\mu)} + 1} . \tag{6.335}$$

From these equations we can determine $E$ as a function of $T, B, N$, compute the magnetization from

$$m = -\frac{1}{V} \frac{\partial E(T, B, N)}{\partial B} , \tag{6.336}$$

and finally obtain the susceptibility at high temperatures as

$$\chi = \frac{\partial m}{\partial B}|_{B=0} = -\frac{1}{3} \frac{\mu_{\mathrm{B}}^2}{k_{\mathrm{B}} T v}, \tag{6.337}$$

where $\mu_{\mathrm{B}} = e\hbar/2mc$ is the Bohr magneton.

The reaction of the electron motion to the external field therefore leads to a negative contribution for the susceptibility. This shows that the motion of the electrons on quantized orbits causes a diamagnetic effect. We also find a characteristic $1/T$ dependence, which we have already seen several times in connection with susceptibilities (cf. Sect. 4.4). In Sect. 6.8.2 we will discuss the contribution of the magnetic moment of the spin and will find that this contribution to the susceptibility is positive.

The existence of the Landau levels, i.e., the quantization of the circular orbits in the $x$-$y$ plane for the motion of an electron in an external magnetic field, are the origin of two effects which we now briefly discuss.

**The de Haas–van Alphen effect.** As the degeneracy is proportional to the magnetic field, the individual Landau levels contain fewer and fewer electrons as the magnetic field decreases. This is observable at low temperatures, where only the lowest levels are occupied. We neglect motion in the $z$-direction, consider the case $T = 0$, and define $B_0$ by the requirement that for $T = 0$ the lowest Landau level is just filled with the $N$ particles. So we have

$$N = g \equiv \frac{eB_0}{hc} L^2 , \quad \text{i.e.,} \quad B_0 = \frac{hc}{e} \frac{N}{L^2} . \tag{6.338}$$

Consider first the case $B > B_0$. Then

$$E = N\mu_{\mathrm{B}} B = N\mu_{\mathrm{B}} B_0 x , \quad \text{with} \quad x = B/B_0 \tag{6.339}$$

and therefore

$$m = -\frac{1}{V} \frac{\partial E}{\partial B} = -\frac{N}{V} \mu_{\mathrm{B}} , \quad \chi = 0 . \tag{6.340}$$

If $B$ now changes to values smaller than $B_0$, the next higher Landau level has to be populated. The energy first increases, but then drops again for decreasing $B$. Although more and more electrons occupy the higher level, the energies of the levels also depend on the magnetic field $B$. As a whole we find a quadratic dependence

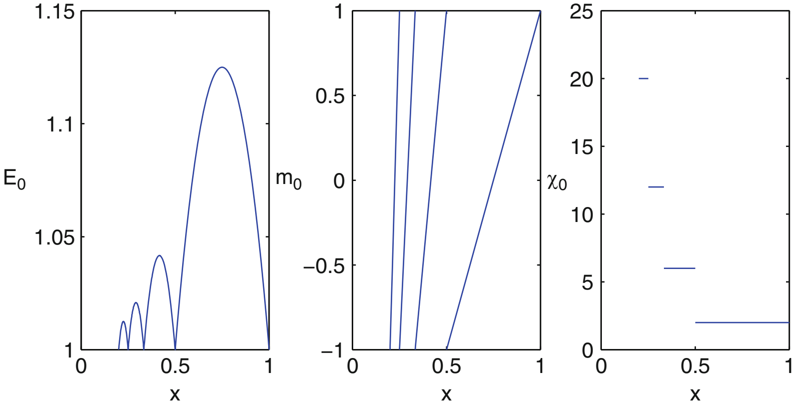


**Fig. 6.11** Dependence of the energy (*left*), magnetization (*center*), and susceptibility (*right*) on the external magnetic field $B$, ($x = B/B_0$, $E_0 = E/N\mu_B B_0$, $m_0 = m/n\mu_B$, $\chi_0 = \chi B_0/2n\mu_B$)

(Fig. 6.11). The magnetization first jumps to positive values, then falls off until the second Landau level is also filled. Now, the same phenomenon repeats itself if the $B$-field decreases further. Hence, the magnetization oscillates and the susceptibility grows in steps (Fig. 6.11). This effect was discovered in 1931 by de Haas and van Alphen, and the first simple explanation given by Peierls in 1933.

**The quantized Hall effect.** If moving charge carriers in a conducter are exposed to a magnetic field, a voltage, known as the Hall voltage, is induced orthogonal to the plane spanned by the current of the charge carriers and the magnetic field. This Hall voltage gives rise to a Hall current, which is superimposed on the original current.

The explanation of the Hall effect is simple: It originates from the Lorentz force $\boldsymbol{K}$ acting on moving charges in a magnetic field according to

$$\boldsymbol{K} = \frac{e}{c}\boldsymbol{v} \times \boldsymbol{B}. \tag{6.341}$$

Here, $\boldsymbol{v}$ is the velocity of the charge carriers. If we choose $\boldsymbol{B}$ orthogonal to $\boldsymbol{v}$ and take into account that the current $j$ in the absence of the magnetic field may be written as $j = env$, where $n$ is the particle density of the charge carriers, we obtain for the electric field strength $E_H$ which accompanies the Hall voltage $U_H$,

$$E_H = \frac{K}{e} = \frac{vB}{c} = \varrho_H j, \tag{6.342}$$

where

$$\varrho_H = \frac{B}{enc}, \tag{6.343}$$

denotes the classical Hall resistivity, which therefore should grow linearly with the external magnetic field.

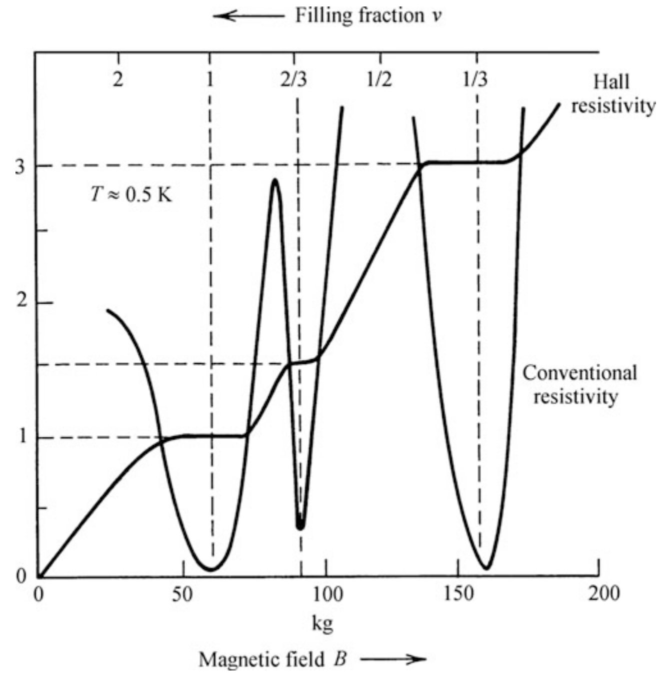


**Fig. 6.12** Quantized Hall effect: schematic representation of the experimental data. The Hall resistivity exhibits plateaus at filling fraction $\nu = 1, 2/3, 1/3, \ldots$, whereas the conventional resistivity drops dramatically at these values (From Huang (1987), reprinted by permission of John Wiley & Sons, Inc.)

However, in extremely thin films at low temperatures and for strong magnetic fields one observes a completely different behavior. With growing magnetic field, the Hall resistivity $\varrho_{\mathrm{H}}$, considered as a function of $B$, exhibits plateaus. These occur at values of $B$ for which the so-called filling fraction

$$\nu = \frac{hcn}{eB} \tag{6.344}$$

assumes the values $1, \frac{2}{3}, \frac{1}{3}, \ldots$ The value of $\varrho_{\mathrm{H}}$ in this case is

$$\varrho_{\mathrm{H}} = \frac{1}{\nu}\frac{h}{e^2} \, . \tag{6.345}$$

The conventional resistivity, which corresponds to the ratio of the applied voltage and the induced current with current density $j$, drops dramatically in the vicinity of these plateaus (Fig. 6.12).

This effect is called the quantized Hall effect. It was discovered by von Klitzing and his collaborators in 1980 (Klitzing et al. 1980). It provides a new method of measuring the Sommerfeld finestructure constants, and also makes it possible to establish the unit of the electric resistivity, 'Ohm', independent of place and time with the remarkable accuracy of some $10^{-9}$ (Kose and Melchert (1991)).

The value of $\varrho_{\mathrm{H}}$ for $\nu = 1$ is easily explained from the properties of the Landau levels: The lowest Landau level is completely filled. Due to the energy gap between the first and the second level, excitations with an energy smaller than this gap are not possible. The centers of the circular orbits of the electrons move like the particles of a free two-dimensional gas. The normal resistivity in this case is therefore small. With a particle density of

$$n \equiv \frac{N}{L^2} = \frac{eB}{hc} \tag{6.346}$$

one obtains for $\varrho_{\mathrm{H}}$ from (6.343)

$$\varrho_{\mathrm{H}} = \frac{B}{enc} = \frac{h}{e^2} \; . \tag{6.347}$$

The explanation of the plateaus, in particular for the rational values of $\nu$, i.e., the so-called fractional quantum Hall effect, goes beyond the scope of this book and, indeed, is in part still a subject of active research (Prange and Girvin 1987).

### *6.8.2 Paramagnetism*

To explain the paramagnetic effect, we again consider a nonrelativistic electron in an external magnetic field, but this time we only take into account the influence exerted by the magnetic moment of the spin. Therefore, we use the single-particle Hamiltonian operator

$$H = \frac{\boldsymbol{p}^2}{2m} - \mu_{\mathrm{B}} \boldsymbol{\sigma} \cdot \boldsymbol{B} \; . \tag{6.348}$$

Here, $\mu_B$ is the Bohr magneton and $\boldsymbol{\sigma}$ represent the Pauli spin matrices. For a $\boldsymbol{B}$-field in the $z$-direction the single-particle states are thus characterized by $\boldsymbol{p}, s$ with $s = \pm 1$, and the energy eigenvalues are

$$\varepsilon(\boldsymbol{p}, s) = \frac{\boldsymbol{p}^2}{2m} - s\mu_{\mathrm{B}} B \; . \tag{6.349}$$

As we are dealing with fermions, the occupation numbers $n(\boldsymbol{p}, s)$ can only assume the values 0 and 1. So we obtain

$$\langle n(\boldsymbol{p}, s)\rangle = \frac{1}{\mathrm{e}^{\beta(\varepsilon(\boldsymbol{p},s)-\mu)} + 1}, \tag{6.350}$$

$$E(T, V, z) = \sum_{\{\boldsymbol{p},s\}} \varepsilon(\boldsymbol{p}, s)\langle n(\boldsymbol{p}, s)\rangle, \tag{6.351}$$

and, after some calculation, for $T \to 0$,

$$\chi = \frac{3\mu_\mathrm{B}^2}{2\varepsilon_\mathrm{F} v}, \tag{6.352}$$

where $\varepsilon_\mathrm{F}$ is again the Fermi energy. For $k_\mathrm{B}T \gg \varepsilon_\mathrm{F}$ we get

$$\chi = \frac{\mu_\mathrm{B}^2}{k_\mathrm{B}T v}\,. \tag{6.353}$$

Hence, $\chi$ is always positive, i.e., we always find a paramagnetic effect. For large temperatures we again obtain a $1/T$ dependence.

The diamagnetic effect caused by the motion of electrons in an external magnetic field (see Sect. 6.8.1) and the paramagnetic effect due to their spins, as just discussed, superpose, giving rise to the susceptibility

$$\chi = \frac{2}{3}\frac{\mu_\mathrm{B}^2}{k_\mathrm{B}T v} \tag{6.354}$$

for $k_\mathrm{B}T \gg \varepsilon_\mathrm{F}$.

## 6.9 Quasi-Particles

In Sect. 6.6 we considered the ideal gas of phonons. We introduced phonons as the excitation quanta which emerge in the quantization of the Hamiltonian for the lattice vibrations of a crystal. We called them quasi-particles. Although they do not exist as free particles in a vacuum, they may otherwise be characterized like normal particles. The influence of the lattice vibrations on the thermal behavior of a solid can thus be described to a good approximation by the properties of an ideal gas of phonons.

This method of describing the low-energy excitations of an interacting many-particle system by a system of quasi-particles, may also be applied to other cases.

For example, one speaks of magnons in the context of the lowest excitations in a ferromagnet, or of excitons when considering the global excitations in an ionic crystal which might result from the excitation of a single ion. There are many more examples of quasi-particles as elementary excitations of many-particle systems (see the table given in Brenig (1992)).

As we did for the particle gases studied previously, we first have to identify in all cases the following quantities and relations which are relevant for the characterization of the gas of quasi-particles:

- The quantum numbers of the quasi-particle state are those quantities by which the single quasi-particles may be characterized. In many cases these are a wavevector $\boldsymbol{k}$, or momentum $\boldsymbol{p}$, often the polarization $\alpha = 1, 2$ or $\alpha = 1, 2, 3$, and further parameters which we shall subsume in $\alpha$.

- The energy $\varepsilon(\boldsymbol{k}, \alpha)$ of the single-particle state. These energy values $\{\varepsilon(\boldsymbol{k}, \alpha)\}$ correspond, of course, to the low-energy levels of the entire system. A quasi-particle therefore represents a collective excitation of the entire system, parametrized by $(\boldsymbol{k}, \alpha)$. Up to now we have met the relations $\varepsilon(\boldsymbol{k}, \alpha) = \hbar c|\boldsymbol{k}|$ ($c$ = velocity of light or sound).
- The possible occupation numbers $n(\boldsymbol{k}, \alpha)$ for the single-particle states. In the case of phonons, e.g., $n(\boldsymbol{k}, \alpha) = 0, 1, 2, \ldots$, i.e., the system of lattice vibrations can be multiply excited in the same 'mode' $(\boldsymbol{k}, \alpha)$.

A quantum-mechanical state for a gas of such particles can be described in the occupation number representation by $\mid n(\boldsymbol{k}_1, \alpha_1), n(\boldsymbol{k}_2, \alpha_2), \ldots\rangle$, where $(\boldsymbol{k}_1, \alpha_1), (\boldsymbol{k}_2, \alpha_2), \ldots$ enumerate the quasi-particle states. The particle number of such a state is given by

$$N = \sum_i n(\boldsymbol{k}_i, \alpha_i) \equiv \sum_{\boldsymbol{k},\alpha} n(\boldsymbol{k}, \alpha), \tag{6.355}$$

and for the energy we find

$$E(\{n(\boldsymbol{k}, \alpha)\}) = \sum_{\boldsymbol{k},\alpha} \varepsilon(\boldsymbol{k}, \alpha)\, n(\boldsymbol{k}, \alpha). \tag{6.356}$$

We now obtain for the partition function of a grand canonical system (cf. (6.53))

$$Y(T, V, \mu) = \sum_{\{n(\boldsymbol{k},\alpha)\}} \mathrm{e}^{-\beta[(E(\{n(\boldsymbol{k},\alpha)\})-\mu N(\{n(\boldsymbol{k},\alpha)\})]}. \tag{6.357}$$

The summation extends over all possible particle numbers for all states. If there is no conservation of the total particle number, as is the case for phonons and photons, we formally have to set $\mu = 0$.

The evaluation of this partition function then leads to the equations of state and to expressions for the response functions. In particular, we obtain for the average occupation number of the quasi-particle state $(\boldsymbol{k}', \alpha')$:

$$\langle n(\boldsymbol{k}', \alpha')\rangle = \frac{1}{Y} \sum_{\{n(\boldsymbol{k},\alpha)\}} n(\boldsymbol{k}', \alpha')\mathrm{e}^{-\beta[E(\{n(\boldsymbol{k},\alpha)\})-\mu N(\{n(\boldsymbol{k},\alpha)\})]}.$$

Written more explicitly, the numerator of the right-hand side is

$$\prod_{i,\boldsymbol{k}_i \neq \boldsymbol{k}'} \left[ \sum_{n(\boldsymbol{k}_i,\alpha_i)} \mathrm{e}^{-\beta(\varepsilon(\boldsymbol{k}_i,\alpha_i)-\mu)n(\boldsymbol{k}_i,\alpha_i)} \right] \sum_{n(\boldsymbol{k}',\alpha')} n(\boldsymbol{k}', \alpha')\mathrm{e}^{-\beta(\varepsilon(\boldsymbol{k}',\alpha')-\mu)n(\boldsymbol{k}',\alpha')}, \tag{6.358}$$

and the denominator $Y$ can also be written as

$$\prod_{i,\boldsymbol{k}_i\neq\boldsymbol{k'}}\left[\sum_{n(\boldsymbol{k}_i,\alpha_i)}\mathrm{e}^{-\beta(\varepsilon(\boldsymbol{k}_i,\alpha_i)-\mu)n(\boldsymbol{k}_i,\alpha_i)}\right]\sum_{n(\boldsymbol{k'},\alpha')}\mathrm{e}^{-\beta(\varepsilon(\boldsymbol{k'},\alpha')-\mu)n(\boldsymbol{k'},\alpha')}, \tag{6.359}$$

so that we finally obtain

$$\langle n(\boldsymbol{k'},\alpha')\rangle=\frac{\sum_n n\mathrm{e}^{-\beta(\varepsilon(\boldsymbol{k'},\alpha')-\mu)n}}{\sum_n \mathrm{e}^{-\beta(\varepsilon(\boldsymbol{k'},\alpha')-\mu)n}}. \tag{6.360}$$

For the Fermi gas, we get in this manner

$$\langle n(\boldsymbol{k'},\alpha')\rangle=\frac{\mathrm{e}^{-\beta(\varepsilon-\mu)}}{1+\mathrm{e}^{-\beta(\varepsilon-\mu)}}=\frac{1}{\mathrm{e}^{\beta(\varepsilon-\mu)}+1} \tag{6.361}$$

and for the Bose gas

$$\langle n(\boldsymbol{k'},\alpha')\rangle=\frac{\mathrm{e}^{-\beta(\varepsilon-\mu)}+2\mathrm{e}^{-\beta(\varepsilon-\mu)\cdot 2}+3\mathrm{e}^{-\beta(\varepsilon-\mu)\cdot 3}+\ldots}{1+\mathrm{e}^{-\beta(\varepsilon-\mu)}+\mathrm{e}^{-\beta(\varepsilon-\mu)\cdot 2}+\mathrm{e}^{-\beta(\varepsilon-\mu)\cdot 3}+\ldots}, \tag{6.362}$$

which, with $\gamma=\beta(\varepsilon-\mu)$ as shorthand notation, can be written as

$$\begin{aligned}\langle n(\boldsymbol{k'},\alpha')\rangle&=-\frac{\mathrm{d}}{\mathrm{d}\gamma}\log\left(1+\mathrm{e}^{-\gamma}+\mathrm{e}^{-2\gamma}+\mathrm{e}^{-3\gamma}+\ldots\right)\\&=\frac{\mathrm{d}}{\mathrm{d}\gamma}\log\left(1-\mathrm{e}^{-\gamma}\right)=\frac{1}{\mathrm{e}^{\beta(\varepsilon-\mu)}-1},\end{aligned} \tag{6.363}$$

as expected.

The total number of particles in the system is then given by

$$N(T,V,\mu)=\sum_{\{\boldsymbol{k},\alpha\}}\langle n(\boldsymbol{k},\alpha)\rangle, \tag{6.364}$$

and for the energy of the system we obtain

$$E(T,V,\mu)=\sum_{\{\boldsymbol{k},\alpha\}}\varepsilon(\boldsymbol{k},\alpha)\langle n(\boldsymbol{k},\alpha)\rangle\ . \tag{6.365}$$

The (possible) dependence of the energy on $\mu$ may be replaced by a dependence on $N$ by solving, as usual, the equation for $N(T,V,\mu)$ with respect to $\mu$ or $z=\mathrm{e}^{\beta\mu}$ and inserting the resulting expression for $\mu$ into $E(T,V,\mu)$.

Two equations shall be cited for the case of response functions. The heat capacity $C_V$ for constant volume is given by

$$C_V=\frac{\partial E(T,V,N)}{\partial T}=\sum_{\{\boldsymbol{k},\alpha\}}\varepsilon(\boldsymbol{k},\alpha)\frac{\partial\langle n(\boldsymbol{k},\alpha)\rangle}{\partial T}, \tag{6.366}$$

and for the magnetization of a system in a constant external magnetic field $B$ in $z$-direction we find:

$$m = -\frac{1}{V}\frac{\partial E(T,B,N)}{\partial B} = -\frac{1}{V}\sum_{\{\boldsymbol{k},\alpha\}} \varepsilon(\boldsymbol{k},\alpha)\frac{\partial\langle n(\boldsymbol{k},\alpha)\rangle}{\partial B} \ . \tag{6.367}$$

In the following two sections we shall briefly introduce two important phenomenological fields where the concept of quasi-particles is successful namely, magnetism and superfluidity. Further elaborations on this concept would lead us deep into the realm of quantum fluids and solid state physics and extend beyond the scope of this book.

### *6.9.1 Models for the Magnetic Properties of Solids*

In Chap. 4 we studied spin models as special examples of random fields. We saw that such models arise from general mathematical considerations, but we have not yet explained how such models can be derived on the basis of physical ideas about the interactions in a solid.

An important role in the description of the magnetic properties of solids are played by the Heisenberg models. We will first present these models and then discuss how they can be derived from physical arguments. We will see that the Hamiltonian operators of these models again describe the lowest-lying energy states of many-particle systems, which therefore allows the introduction of quasi-particles, now called magnons, as the excitation quanta.

We denote the spin operator at lattice site $\boldsymbol{R}$ by $\boldsymbol{S}(\boldsymbol{R})$. The eigenvalue of $\boldsymbol{S}^2(\boldsymbol{R})$ shall be $s(s+1)$ and the usual commutation relations for spin operators shall hold, e.g.,

$$[S^x(\boldsymbol{R}), S^y(\boldsymbol{R'})] = \mathrm{i}S^z(\boldsymbol{R})\delta_{\boldsymbol{R}\boldsymbol{R'}} \ . \tag{6.368}$$

The Hamiltonian operator for the isotropic Heisenberg model in an external field $B$ now reads

$$H = -\frac{1}{2}\sum_{\boldsymbol{R},\boldsymbol{R}'\boldsymbol{R}\neq\boldsymbol{R}'} J(\boldsymbol{R}-\boldsymbol{R}')\,\boldsymbol{S}(\boldsymbol{R})\cdot\boldsymbol{S}(\boldsymbol{R}') - g\mu_{\mathrm{B}}B\sum_{\boldsymbol{R}} S^z(\boldsymbol{R}) \ . \tag{6.369}$$

$J(\boldsymbol{R}-\boldsymbol{R}')$ are called the exchange coupling constants. Often they are chosen in such a way that

$$J(\boldsymbol{R}-\boldsymbol{R}') \neq 0 \qquad \text{only if } \boldsymbol{R} \text{ and } \boldsymbol{R}' \text{ are nearest neighbors.} \tag{6.370}$$

A further restriction results if we set all $J(\boldsymbol{R}-\boldsymbol{R}')$ to be equal to one constant $J$: In this case one writes

$$H = -\frac{1}{2} J \sum_{\mathrm{nn}} \boldsymbol{S}(\boldsymbol{R}) \cdot \boldsymbol{S}(\boldsymbol{R}') - g\mu_{\mathrm{B}} B \sum_{\boldsymbol{R}} S^z(\boldsymbol{R}) \, , \tag{6.371}$$

where 'nn' now stands for nearest neighbor, i.e., the sum extends over all $\boldsymbol{R}$, and for a given $\boldsymbol{R}$ over all nearest neighbors. The model is called isotropic because of the rotationally invariant coupling $\boldsymbol{S}(\boldsymbol{R}) \cdot \boldsymbol{S}(\boldsymbol{R}')$. An anisotropic variant for the case $\boldsymbol{B} = 0$ would be

$$H = -\frac{1}{2} \sum_{\mathrm{nn}} (J_x \, S^x(\boldsymbol{R}) S^x(\boldsymbol{R}') + J_y \, S^y(\boldsymbol{R}) S^y(\boldsymbol{R}') + J_z \, S^z(\boldsymbol{R}) S^z(\boldsymbol{R}')) \tag{6.372}$$

or

$$H = -\frac{1}{2} J \sum_{\mathrm{nn}} \boldsymbol{S}(\boldsymbol{R}) \cdot \boldsymbol{S}(\boldsymbol{R}') + A \sum_{\boldsymbol{R}} (S^z(\boldsymbol{R}))^2 \; . \tag{6.373}$$

Since the operator $\boldsymbol{S}(\boldsymbol{R})$ acts on a $(2s+1)$-dimensional space $\mathcal{H}^0$, the Hamiltonian $H$ may be represented by a matrix in the $(2s+1)^N$-dimensional space

$$\underbrace{\mathcal{H}^0 \otimes \mathcal{H}^0 \otimes \ldots \otimes \mathcal{H}^0}_{N} \; . \tag{6.374}$$

For spin $s = 1/2$ and $N = 10$ one is already confronted with a a $1{,}024 \times 1{,}024$ matrix. If only $J_z \neq 0$ in (6.372), such that the Hamiltonian only depends on the commuting operators $J_z(\boldsymbol{R})$, one speaks of an Ising model (named after the physicist Ising (1925), cf. Sect. 4.5).

Why are such Hamiltonians for spin systems formulated? What can be explained by these models? To find an answer to these questions we first consider a very small lattice of atoms: the hydrogen molecule, i.e., a system of two protons and two electrons. The protons shall be fixed at the positions $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, the position vectors of the electrons shall be $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. Then the Hamiltonian reads

$$H = H_1 + H_2 + H', \tag{6.375}$$

where

$$H_i = \frac{p_i^2}{2m_e} - \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{R}_i|} \, , \qquad i = 1, 2 \, , \tag{6.376}$$

$$H' = \frac{e^2}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + \frac{e^2}{|\boldsymbol{R}_1 - \boldsymbol{R}_2|} - \frac{e^2}{|\boldsymbol{r}_1 - \boldsymbol{R}_2|} - \frac{e^2}{|\boldsymbol{r}_2 - \boldsymbol{R}_1|} \, . \tag{6.377}$$

In the so-called Heitler–London approximation (see e.g. McQuarrie 1983) one obtains for the electron states corresponding to the two lowest energy eigenvalues

$$\psi_{\mathrm{s/t}}(\boldsymbol{r}_1, \boldsymbol{r}_2) = (\phi_2(\boldsymbol{r}_1)\phi_1(\boldsymbol{r}_2) \pm \phi_1(\boldsymbol{r}_1)\phi_2(\boldsymbol{r}_2))\chi_{\mathrm{s/t}} \, . \tag{6.378}$$

Here, $\phi_i(\boldsymbol{r})$ are the ground state eigenfunctions of

$$H_i = \frac{p^2}{2m_e} - \frac{e^2}{|\boldsymbol{r} - \boldsymbol{R}_i|}, \tag{6.379}$$

i.e., they describe the single-electron states with smallest eigenvalue in the potential of a proton at position $\boldsymbol{R}_i$. $\chi_{\mathrm{s/t}}$ are the eigenfunctions of $(\boldsymbol{S}_1 + \boldsymbol{S}_2)^2$ and $(\boldsymbol{S}_1 + \boldsymbol{S}_2)_z$, where $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ are the spin operators for the electrons. $\chi_{\mathrm{s}}$ describes a singlet state and is an antisymmetric function, while $\chi_{\mathrm{t}}$ describes a triplet state and is symmetric. Taking into account the Pauli principle, the corresponding real space wavefunction has to be antisymmetric or symmetric, respectively.

Using these approximate solutions one obtains for the energy eigenvalues:

$$E_{\mathrm{s/t}} = \langle \psi_{\mathrm{s/t}} | H | \psi_{\mathrm{s/t}} \rangle \tag{6.380}$$

$$= \int \mathrm{d}^3 r_1 \, \mathrm{d}^3 r_2 \, (\phi_2(\boldsymbol{r}_1)\phi_1(\boldsymbol{r}_2) \pm \phi_1(\boldsymbol{r}_1)\phi_2(\boldsymbol{r}_2))^* \tag{6.381}$$

$$\times H(\phi_2(\boldsymbol{r}_1)\phi_1(\boldsymbol{r}_2) \pm \phi_1(\boldsymbol{r}_1)\phi_2(\boldsymbol{r}_2)) \,. \tag{6.382}$$

It turns out that $E_{\mathrm{s}} > E_{\mathrm{t}}$. Despite the absence of an explicit spin–spin interaction term in the Hamiltonian, the total spin of the two electrons has an influence on the energy eigenvalues due to the Pauli principle.

The first excited states of the electrons are of the order 1 eV $\simeq$ 10,000 K, i.e., at normal temperatures they are of no importance. Neglecting, therefore, the states of higher energy one is dealing with only two energy levels, $E_{\mathrm{s}}$, $E_{\mathrm{t}}$, but four states, which may be characterized by the total spin.

To reproduce these energy levels, one can define an 'effective' Hamiltonian operator $H_{\mathrm{eff}}$ in spin space by

$$H_{\mathrm{eff}} = -J \boldsymbol{S}_1 \cdot \boldsymbol{S}_2, \tag{6.383}$$

because now

$$H_{\mathrm{eff}} = -\frac{J}{2}\left((\boldsymbol{S}_1 + \boldsymbol{S}_2)^2 - \boldsymbol{S}_1^2 - \boldsymbol{S}_2^2\right) \tag{6.384}$$

$$= -\frac{J}{2}\left(s(s+1) - \frac{3}{4} - \frac{3}{4}\right) \tag{6.385}$$

$$= \begin{cases} -\frac{J}{2}\left(\frac{1}{2}\right) = -\frac{1}{4}J & \text{for } s = 1 \\ -\frac{J}{2}\left(-\frac{3}{2}\right) = +\frac{3}{4}J & \text{for } s = 0, \end{cases} \tag{6.386}$$

i.e., $H_{\mathrm{eff}}$ describes an energy difference between the singlet and the triplet state given by $J$, and we may choose $J = E_{\mathrm{s}} - E_{\mathrm{t}}$. The splitting of the energy eigenvalues is therefore controlled by the exchange integral $J$. If $J$ were zero, the ground state would be fourfold degenerate.

Heisenberg generalized this idea to a system of $N$ atoms fixed to the sites of a crystal lattice, in order to reproduce the energy spectrum for the lowest lying states of this system by the energy spectrum of a spin Hamiltonian, such as (6.372). The Heisenberg model therefore describes the splitting of the ground state in a crystal, which for $J \equiv 0$ would be $(2s+1)^N$-fold degenerate.

Due to the isotropy in spin space (no interaction among the spins of the electrons) one thus obtains the isotropic Heisenberg model. If, e.g., a spin–orbit coupling were also included, this would give rise to anisotropic effects in the model Hamiltonian. The spin quantum number $s$ of the model is determined by the spin $s$ of the ground state of the atom with respect to its valence electrons.

The justification for Heisenberg's generalization is highly nontrivial and requires many complicated approximations (see also Herring 1966).

In some cases both the ground state of the Hamiltonian as well as the first excited states can be computed explicitly. For instance, in the Heisenberg model of a ferromagnet the ground state $|F\rangle$ is the state for which all spins point upwards. The superpositions of states where just one spin points downwards, more precisely

$$|\boldsymbol{k}\rangle = \frac{1}{\sqrt{2NS}} \sum_{\boldsymbol{R}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}}\, S_-(\boldsymbol{R})|F\rangle, \tag{6.387}$$

where

$$S_-(\boldsymbol{R}) = S_1(\boldsymbol{R}) - \mathrm{i}S_2(\boldsymbol{R}),$$

turn out to be the first excited eigenstates of the Hamiltonian. The state $|\boldsymbol{k}\rangle$ describes a spin wave or magnon. Magnons are therefore quasi-particles like phonons. In the spin-wave approximation the solid is considered as an ideal gas of magnons. Of course, this approximation is justified only for low temperatures (see Ashcroft and Mermin 1976).

The application of spin Hamiltonians of the Heisenberg type has been quite fruitful in the description of magnetic, nonconducting systems. In general, one distinguishes magnetic systems of the following types:

### Ferromagnets

These exhibit a macroscopic magnetization at sufficiently low temperatures, even when the external field $\boldsymbol{B}$ has been turned off, i.e., below a critical temperature $T_\mathrm{c}$ the interaction causes the spins to align predominantly in one direction. The tendency to disorder due to thermal agitation is overridden by the tendency to order due to the interaction. This magnetization at $\boldsymbol{B} = 0$ is also called spontaneous magnetization, and the critical temperature is referred to as the Curie temperature. For ferromagnets the exchange coupling in the Hamiltonian is positive, $J > 0$. Some examples of ferromagnets are listed in Table 6.3.

Ferromagnets exist also in two or even in one space dimension. In this case the interaction between different planes, respectively chains of atoms is much smaller

**Table 6.3** Examples of ferromagnets

| Conductors | | Insulators which may be described by Heisenberg models | |
|---|---|---|---|
| Substance | $T_c$ | Substance | $T_c$ |
| Fe | 1,043 | $CrBr_3$ | 37 |
| Co | 1,388 | EuO | 77 |
| Ni | 627 | EuS | 16.5 |

**Table 6.4** Examples of antiferromagnets ($T_c$ is the Néel temperature)

| Space dimension | Material | $T_c$ | Type |
|---|---|---|---|
| $d = 3$ | MnO | 122 | |
| | FeO | 198 | |
| | NiO | 600 | Insulator |
| | Cr | 311 | |
| $d = 2$ | $RbFeF_4$ | | Insulator |
| $d = 1$ | $CsCoCl_3$ | | Insulator |
| | $CsNiCl_3$ | | Insulator |

than the interaction within the plane or chain. Examples for ferromagnets in two space dimensions are $CrBr_3$ or $K_2CuF_4$ and in one space dimension $CsCuCl_3$, $CsNiF_3$.

## Antiferromagnets

Antiferromagnets (see Table 6.4) also exhibit an ordered structure for $T < T_c$, but it is not accompanied by a spontaneous magnetization. This structure is most easily visualized as two mutually penetrating sublattices, where both sublattices have a spontaneous magnetization, but in opposite directions. The exchange constant in the Hamiltonian is negative, $J < 0$.

## Ferrimagnets

If the spins associated with the different sublattices are not of the same size, one can imagine an ordered structure along a chain of the following kind $\uparrow \downarrow \uparrow \downarrow \uparrow$, so that there is always a net spontaneous magnetization present.

For the modeling of ferrimagnetic structures one has to use a generalized Heisenberg Hamiltonian with different spin operators. Table 6.5 lists examples of ferrimagnetic substances.

*Remarks.*

- The description of magnetic conductors, such as Ni or Fe, is complicated and remains a subject of present-day research.

**Table 6.5** Examples for ferrimagnets

| Substance | $T_c$ |
|---|---|
| $Fe_3O_4$ | 858 |
| $CoFe_2O_4$ | 793 |
| $Y_3Fe_4O_{12}$ | 560 |

- The domain structure in ferromagnets (known as Weiss domains) is a consequence of the hitherto neglected magnetic dipole interactions among the spins, which energetically favors the splitting into domains. Although this interaction is much weaker than the exchange interaction, it is long-ranged. In the presence of sufficiently many spins it therefore becomes important. The hysteresis effect is also explained only by this interaction (see also Ashcroft and Mermin 1976).

### 6.9.2 Superfluidity

Superfluidity denotes the flow of a fluid without viscosity. This phenomenon can be observed, e.g., in liquid $^4$He at low temperatures (1–2 K). That there exists a liquid phase of He at these low temperatures at all is a phenomenon which itself is worthy of discussion. In the present context, however, we will study superfluidity as an example of the successful application of the concept of quasi-particles.

The lowest excitations of liquid $^4$He are found by neutron scattering at low temperatures. Their energy is similar to the energy transferred from the neutrons to the system. The same holds for the momentum. From these experiments one obtains a dependence of the energy $\varepsilon(k)$ on $k$ as shown in Fig. 6.13. Close to $k = 0$ one finds the usual phonon spectrum, $\varepsilon(k) = \hbar kc$, and close to a value $k_0$ $\left(k_0 = 1.92 \pm 0.01\,\text{Å}^{-1}\right)$, the energy $\varepsilon(k)$ exhibits a minimum. Excitations with $k \approx k_0$ may be treated as special quasi-particles, which are called rotons. A minimum energy $\varDelta$ is required to create a roton. The roton mass $m^*$ can be found by expanding $\varepsilon(k)$ around $k_0$ in the form

$$\varepsilon(k) = \varDelta + \frac{\hbar^2(k-k_0)^2}{2m^*} + \dots \,. \tag{6.388}$$

For temperatures below a critical temperature $T_\lambda$, the fluid may be described approximately as an ideal gas of two kinds of quasi-particles, the phonons and the rotons. The heat capacity of each gas is easily determined. For phonons one obtains

$$C_{\text{phonon}} = k_B N \frac{2\pi^2 v (k_B T)^3}{16\hbar^3 c^3} \tag{6.389}$$

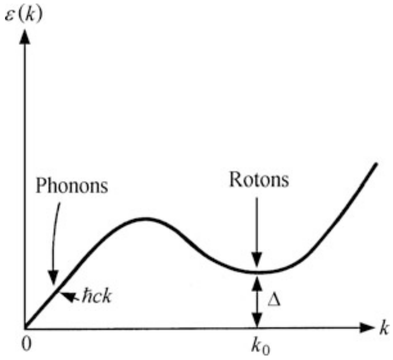


**Fig. 6.13** Energy spectrum of the elementary excitations in liquid $^4$He (From Huang (1987), reprinted by permission of John Wiley & Sons, Inc.)

and for rotons

$$C_{\text{roton}} = k_{\mathrm{B}} N \frac{2\sqrt{m^*} k_0^2 \Delta^2 v \mathrm{e}^{-\Delta/k_{\mathrm{B}}T}}{(2\pi)^{3/2} \hbar (k_{\mathrm{B}} T)^{3/2}} \,. \tag{6.390}$$

The heat capacity of the fluid is given by the sum of these two terms. For temperatures below 0.7 K the phonon part dominates; above this temperature the roton part is larger. The experimental results agree quite well with these estimates.

The gas of quasi-particles constitutes the so-called ‘normal’ component of the superfluid. The superfluid component, however, is a pure superfluid. The total fluid may be treated in a two-fluid model as a mixture of both phases. At low temperatures the normal component consists almost exclusively of phonons, whose number drops further with decreasing temperature. An object with sufficiently small velocity can be pulled through such a fluid without any friction. The dissipation may be neglected, because on the one hand the energy transfer by scattering with phonons is negligible due to their small number, and on the other hand there is not enough energy present to generate new phonons.

For descriptions of the effects observed in superfluids and a more detailed theoretical treatment we refer to textbooks on statistical mechanics (see e.g. Huang 1987).

If the flow of charge carriers is free of viscosity, one speaks of superconductivity. In superconductors, like $^3$He, at sufficiently low temperatures, a current flows practically without resistance. In this case too, it is the properties of quasi-particles that play the dominant role.

# Chapter 7
# Changes of External Conditions

In the previous chapters we have dealt with systems in thermodynamic equilibrium, i.e., with systems being and remaining in a state of maximum entropy. The external conditions were kept constant and the systems were adapted to these external conditions. It was for this situation that we formulated the equations of state expressing the relations among the system variables.

We now want to discuss what happens when the external conditions are modified. We will to study how the system reacts to such a modification with a change of state, i.e., a change of its state variables.

## 7.1 Reversible State Transformations, Heat, and Work

Let $\{|n\rangle\}$ be the possible microscopic states of a system and $\varrho_n$ be the probability that a system is in the state $|n\rangle$. For a canonical system in thermal equilibrium we have

$$\varrho_n = \frac{1}{Z}\,\mathrm{e}^{-\beta E_n}\ , \tag{7.1}$$

where $E_n$ is the energy of the state $|n\rangle$, and

$$Z = \sum_n \mathrm{e}^{-\beta E_n} \tag{7.2}$$

denotes the partition function. The entropy is given by

$$S = -k_\mathrm{B} \sum_n \varrho_n\ \ln \varrho_n , \tag{7.3}$$

and for the energy of the system, which is equal to the expectation value of the Hamiltonian, we obtain

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_7, © Springer-Verlag Berlin Heidelberg 2012

$$E = \sum_n E_n \varrho_n . \tag{7.4}$$

We now consider infinitesimal transformations of the state (the particle number should remain constant). Extrinsic system variables, such as the volume $V$ or the temperature $T$, may change. This leads to a change in energy, $\mathrm{d}E$, which we may decompose as

$$\mathrm{d}E = \sum_n \mathrm{d}E_n \varrho_n + \sum_n E_n \mathrm{d}\varrho_n , \tag{7.5}$$

i.e., either the energy eigenvalues for the individual states $|n\rangle$ change, or the probabilities $\varrho_n$, or both. We will investigate each term separately. The term

$$\delta Q = \sum_n E_n \, \mathrm{d}\varrho_n \tag{7.6}$$

is called heat, and the term

$$\delta W = \sum_n \mathrm{d}E_n \, \varrho_n \tag{7.7}$$

is referred to as work.

Let us first examine the heat term. A change of $\mathrm{d}\varrho_n$, as in $\delta Q$, also has an influence on the entropy:

$$\mathrm{d}S = -k_\mathrm{B} \sum_n \mathrm{d}\varrho_n \, \ln \varrho_n - k_\mathrm{B} \sum_n \varrho_n \, \mathrm{d}(\ln \varrho_n). \tag{7.8}$$

However, since

$$\mathrm{d}(\ln \varrho_n) = \frac{1}{\varrho_n} \, \mathrm{d}\varrho_n , \tag{7.9}$$

and because $\sum_n \varrho_n = 1$, we get:

$$\sum_n \varrho_n \, \mathrm{d}(\ln \varrho_n) = \sum_n \mathrm{d}\varrho_n = \mathrm{d} \sum_n \varrho_n = 0. \tag{7.10}$$

The second term in (7.8) thus vanishes, leaving us with

$$\mathrm{d}S = -k_\mathrm{B} \sum_n \mathrm{d}\varrho_n \, \ln \varrho_n . \tag{7.11}$$

For a system in equilibrium (i.e., a state of maximum entropy) relation (7.1) holds and therefore

$$\ln \varrho_n = -\frac{1}{k_\mathrm{B} T} \, E_n - \ln Z , \tag{7.12}$$

hence

$$\mathrm{d}S = \frac{1}{T} \sum_n E_n \, \mathrm{d}\varrho_n = \frac{\delta Q}{T} , \tag{7.13}$$

i.e., the term $\delta Q$ may also be expressed through $\mathrm{d}S$ as

$$\delta Q = T\,\mathrm{d}S. \tag{7.14}$$

The contribution $\delta W$ is due to a change of the energy eigenvalues, while the probabilities remain fixed. The eigenvalues of the Hamiltonian depend, among others, on the volume of the system, i.e., the energy eigenvalues in general change if the volume of the system or other external extensive parameters are modified. For a change in volume, for example, we obtain

$$\delta W = \sum_n \frac{\partial E_n}{\partial V}\,\varrho_n\,\mathrm{d}V = -p\,\mathrm{d}V, \tag{7.15}$$

with

$$p = -\sum_n \frac{\partial E_n}{\partial V}\,\varrho_n. \tag{7.16}$$

We can thus conclude that, when one or more external state variables of a system are changed infinitesimally, the change in energy can be written in the form

$$\mathrm{d}E = \delta Q + \delta W. \tag{7.17}$$

$\delta Q$ is the contribution correlated with a change in entropy, while $\delta W$ is determined by the change of the energy eigenvalues of the system. We will soon discuss state transformations for which one of these two contributions vanishes.

Heat $\delta Q$ and work $\delta W$ are energy contributions which the system exchanges with its environment. They are defined as positive, if they represent an increase of the system energy. If work is done by the system, $\delta W$ is negative. Furthermore, an exchange of heat, $\delta Q$, always accompanies an exchange of entropy, $\mathrm{d}S$.

Infinitesimal transformations of the state of a system, which thereby remains in equilibrium, are called reversible. For such transformations of systems in equilibrium the contribution $\delta Q$ to the energy change is related to the change in entropy by (7.14), because $\ln \varrho_n$ is given by (7.12). Furthermore, work done by the system only results in a change of the energy eigenvalues, not in a change of the probabilities $\varrho_n$.

An infinitesimal transformation of the state of a system through nonequilibrium states is called irreversible. In this case, the relation $\delta Q = T\mathrm{d}S$ does not hold, but instead

$$\mathrm{d}S > \frac{\delta Q}{T} = \sum_n \frac{E_n}{T}\mathrm{d}\varrho_n, \tag{7.18}$$

because the system always tends towards its equilibrium state, i.e., the state of maximum entropy, and therefore entropy is created spontaneously. After the change of state, the system will contain, in addition to the exchanged entropy $\mathrm{d}S_{\mathrm{ex}}$ related to $\delta Q$, also created entropy $\mathrm{d}S_{\mathrm{c}} > 0$.

Furthermore, for an irreversible state transformation due to a change in the external extensive parameters, i.e., one for which $\delta W$ is not zero, the probabilities $\varrho_n$ also change, again because the system spontaneously tends towards the state of maximum entropy and therefore the entropy increases.

### 7.1.1 Finite Transformations of State

Up to now we have discussed infinitesimal transformations of state which start from equilibrium states. Now we will combine many of these transformations in succession to obtain finite transformations of state.

A finite process will be called quasi-static if the system is transformed from an initial state to a final state through a continuous succession of equilibrium states. During such a process the probability distribution is at any moment given by the maximum entropy principle. In practice, the change of the macroscopic system variables has to be so slow that the deviation from equilibrium can be neglected, and it must take place in so many steps that there is always enough time to reach a new equilibrium state before the next change of state follows.

A quasi-static process may therefore also be called a reversible process.

Of course, such a reversible process is the ideal limiting case of a general irreversible process, where the individual changes follow each other so rapidly that there is not enough time for the system to relax into an equilibrium state.

Let us investigate some important transformations of state:

**Heating** The simplest process consists of the reversible 'heating' of a system. No work is done and energy is added only in the form of heat, i.e., the volume is kept constant and for each temperature (which changes during the heating process) we have

$$\mathrm{d}E = \delta Q = T\,\mathrm{d}S \quad \text{and} \quad \delta W = 0. \tag{7.19}$$

The energy needed to change the temperature by $\mathrm{d}T$ is

$$C_V = \frac{\mathrm{d}E}{\mathrm{d}T} = T\frac{\mathrm{d}S}{\mathrm{d}T}. \tag{7.20}$$

**Adiabatic process** A process is called adiabatic if no heat is exchanged during the whole process. An adiabatic reversible process is a particularly simple reversible process. In this case the probabilities $\varrho_n$ remain constant, and only the energy eigenvalues change. Therefore, $\delta Q = 0$ and $\mathrm{d}E = \delta W$.

Let us determine the adiabatic reversible expansion of an ideal classical gas. In this case $E = \frac{3}{2}Nk_\mathrm{B}T$, i.e., $\mathrm{d}E = \frac{3}{2}Nk_\mathrm{B}\,\mathrm{d}T$. On the other hand,

$$\mathrm{d}E = \delta W = -p\,\mathrm{d}V = -\frac{Nk_\mathrm{B}T}{V}\,\mathrm{d}V, \tag{7.21}$$

and therefore

$$\frac{3}{2}\frac{\mathrm{d}T}{T}+\frac{\mathrm{d}V}{V}=0, \tag{7.22}$$

or

$$\mathrm{d}\ \ln(T^{3/2}V)=0. \tag{7.23}$$

Hence, we obtain:

$$T^{3/2}V=\text{const.} \qquad \text{or} \qquad pV^{5/3}=\text{const.}. \tag{7.24}$$

For the change in energy we find

$$\Delta E=E_2-E_1=\frac{3}{2}\,Nk_\mathrm{B}\,(T_2-T_1) \quad \text{where} \quad T_2^{3/2}V_2=T_1^{3/2}V_1. \tag{7.25}$$

**Reversible isothermic expansion** Another important reversible process is the reversible isothermic expansion of an ideal gas in contact with a heat bath (infinitesimally realizable by a small external negative pressure such that the gas may expand slowly). In this case the energy eigenvalues $E_n$ and the probabilities change in such a way that throughout the process

$$\varrho_n=\frac{1}{Z}\,\mathrm{e}^{-\beta E_n} \tag{7.26}$$

remains valid.

Since $E=\frac{3}{2}\,Nk_\mathrm{B}T$, the constancy of temperature implies $\mathrm{d}E=0$, and therefore

$$\mathrm{d}E=\delta Q+\delta W=0, \tag{7.27}$$

where now

$$\delta W=-\int_{V_1}^{V_2}p\,\mathrm{d}V=-Nk_\mathrm{B}T\int_{V_1}^{V_2}\frac{\mathrm{d}V}{V}=-Nk_\mathrm{B}T\ \ln\left(\frac{V_2}{V_1}\right). \tag{7.28}$$

$\delta W$ is negative because the gas loses the energy $|\delta W|$. On the other hand, the gas gains energy in the form of heat

$$\delta Q=\int_{S_1}^{S_2}T\,\mathrm{d}S=T(S_2-S_1). \tag{7.29}$$

For an ideal gas $S=Nk_\mathrm{B}\ \ln V+f(N,T)$, and we obtain

$$\delta Q=Nk_\mathrm{B}T\ \ln\left(\frac{V_2}{V_1}\right). \tag{7.30}$$

Hence we have confirmed that the energy received from the heat bath in the form of heat is exactly balanced by the loss of energy due to the work done.

**Spontaneous expansion** Finally, we examine an irreversible process: The process of spontaneous expansion. Consider an isolated system of volume $V_2$, which has no exchange whatsoever with its environment. The gas is initially in a subvolume $V_1$, separated by some partition from the remaining volume. If this partition is suddenly removed, the gas will expand and occupy the total volume $V_2$. We now compare the initial state of the system, state 1, where the partition has just been removed but where the gas still occupies the volume $V_1$, with state 2, where the gas has expanded to fill the whole volume $V_2$ and is again in equilibrium. The first state is obviously an extremely unlikely state and was only realized because the system had been prepared this way using the partition. The system tends towards the most likely state, the equilibrium state. Although there is no exchange with the environment, one of the system variables, the entropy, changes. For an ideal gas

$$\Delta S = S_2 - S_1 = N k_{\mathrm{B}} \ln \left( \frac{V_2}{V_1} \right) . \tag{7.31}$$

Hence, entropy is created. In this case, however, the increase of entropy is not due to exchange, but solely due to creation.

Of course, the energy is conserved, i.e.,

$$E = \sum_n E_n \varrho_n^1 = \sum_n E_n \varrho_n^2 , \tag{7.32}$$

where $\varrho_n^1$ denotes the probability density for the state 1, which is not in equilibrium, and $\varrho_n^2$ denotes the probability density for the equilibrium state 2. Of all possible densities, $\{\varrho_n^2\}$ is simply the one possessing maximum entropy under the supplementary condition (7.32).

## 7.2 Cyclic Processes

We first examine the reversible isothermal expansion of a gas using a $p$–$V$ diagram and a $T$–$S$ diagram. The initial and final state of this process correspond to the points 1 and 2 (Fig. 7.1).

The work done in this process may be represented by the area under the curve in the $p$–$V$ diagram. In order to be able to do this work, energy has to be brought into the system in the form of heat. The amount of heat $\delta Q = T(S_2 - S_1)$ is equal to the area under the horizontal isotherm 1 → 2 in the $T$–$S$ diagram.

A simple reversal of this process, i.e., the process 2 → 1, would lead us back to the initial state 1. However, the work previously done by the system now has to be done on the system, and the entropy has to be returned completely. Hence, in order to produce some net work done by the system we have to return to state 1 along a different path in the $p$–$V$ diagram, for instance, by an adiabatic change of state 2 → 3, followed by an isothermal reversible compression 3 → 4, and finally by

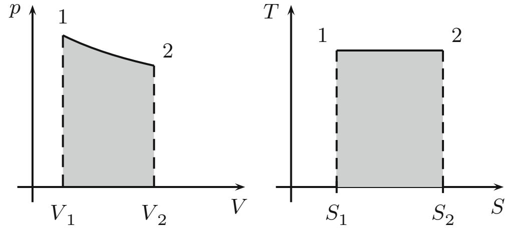


**Fig. 7.1** Isothermal expansion (1 → 2) on a $p$–$V$ and a $T$–$S$ diagram

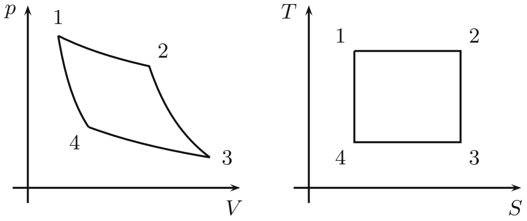


**Fig. 7.2** Cyclic processes on a $p$–$V$ diagram and a $T$–$S$ diagram. The variables $T, S, V$ and $p$ are brought back from state 2 to their values in state 1 by an adiabatic expansion followed by an isothermal compression and an adiabatic compression

an adiabatic compression 4 → 1 back to 1. In doing this we run through a cyclic process, which may be repeated arbitrarily many times (see Fig. 7.2).

For each cycle, the entropy $\Delta S = S_2 - S_1$ added to the system during the isothermal expansion has to be exported again. This, however, may be done at a lower temperature, for which only part of the work done before has to be returned.

Let us compute the contributions of the work done and the heat in more detail. Proceeding from the Gibbs fundamental form

$$\mathrm{d}E = T\,\mathrm{d}S - p\,\mathrm{d}V, \tag{7.33}$$

the energy balance along the cycle reads

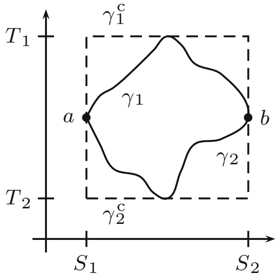
$$0 = \oint \mathrm{d}E = \oint T\,\mathrm{d}S - \oint p\,\mathrm{d}V, \tag{7.34}$$

where $\oint$ denotes the integral around the complete cycle. In particular,

$$W = \oint p\,\mathrm{d}V \tag{7.35}$$

is the work done, i.e., the area enclosed by the cycle in a $p$–$V$ diagram represents the work done during such a process. Furthermore,

**Fig. 7.3** An arbitrary cycle with $T_1$ and $T_2$ as the maximum and minimum temperatures (*solid line*) and the corresponding Carnot cycle (*dotted curve*)



$$Q_1 = \int_{1\to 2} T\,\mathrm{d}S = T_1(S_2 - S_1) \tag{7.36}$$

is the quantity of heat absorbed by the system during the isothermal expansion, and

$$Q_2 = \int_{3\to 4} T\,\mathrm{d}S = -T_2(S_2 - S_1) \tag{7.37}$$

corresponds to the quantity of heat emitted during the isothermal compression. The area enclosed by the cycle in a $T$–$S$ diagram is equal to $Q_1 + Q_2 = W$.

Finally, for the efficiency $\eta$, defined as the ratio of the work done to the heat absorbed, $Q_1$, we obtain

$$\eta = \frac{W}{Q_1} = \frac{Q_1 + Q_2}{Q_1} = 1 + \frac{Q_2}{Q_1} = 1 - \frac{T_2}{T_1}. \tag{7.38}$$

The above described cycle is known as a Carnot cycle and the efficiency $\eta$ is called Carnot efficiency.

It is easy to prove that the Carnot efficiency is the maximum efficiency achievable for a process which operates between the temperatures $T_1$ and $T_2$. For this we consider an arbitrary reversible cycle along the lines $\gamma_1$ and $\gamma_2$ in the $T$–$S$ diagram, where $T_1$ and $T_2$ denote the maximum and minimum temperatures, respectively.

For this cycle shown in Fig. 7.3 (solid line)

$$Q_1 = \int_{a,\gamma_1}^{b} T\,\mathrm{d}S \tag{7.39}$$

is the heat absorbed by the system. If we construct the corresponding Carnot cycle, (Fig. 7.3, dotted line) the heat absorbed in this cycle is

$$Q_1^{\mathrm{c}} = \int_{a,\gamma_1^{\mathrm{c}}}^{b} T\,\mathrm{d}S \tag{7.40}$$

As the integrals agree with the areas under the curves $\gamma_1$ and $\gamma_1^{\mathrm{c}}$, respectively, it is evident that

$$Q_1 \leq Q_1^{\mathrm{c}}. \tag{7.41}$$

Similarly,

$$Q_2 = \int_{b,\gamma_2}^{a} T\,\mathrm{d}S \qquad \text{and} \qquad Q_2^{\mathrm{c}} = \int_{b,\gamma_2^{\mathrm{c}}}^{a} T\,\mathrm{d}S \tag{7.42}$$

are the quantities of heat emitted during the two processes. Obviously,

$$|Q_2| \geq |Q_2^{\mathrm{c}}|. \tag{7.43}$$

Therefore,

$$\frac{|Q_2|}{Q_1} \geq \frac{|Q_2^{\mathrm{c}}|}{Q_1^{\mathrm{c}}}, \tag{7.44}$$

and we find for the efficiency of the process $(\gamma_1 \cup \gamma_2)$

$$\eta = 1 - \frac{|Q_2|}{Q_1} \leq 1 - \frac{|Q_2^{\mathrm{c}}|}{Q_1^{\mathrm{c}}} = 1 - \frac{T_2}{T_1}. \tag{7.45}$$

Finally let us consider an irreversible Carnot cycle. Then entropy is created and must be transferred to the environment. This e.g. increases $|Q_2|$ and the efficiency $\eta = 1 - (|Q_2|/Q_1)$ becomes smaller than the Carnot efficiency.

*Remarks.*

- One may also proceed around a cycle in the reverse direction. In this case work is done on the system, which is now used to withdraw energy from a heat bath at lower temperature and to add this energy to a heat bath at higher temperature. This is the principle of a heat pump.

  This process is also never completely reversible in reality. Hence, the notion of irreversibility does not refer to the direction of a process, but to the creation of entropy. It is the creation of entropy which cannot be reversed, since entropy cannot be annihilated. Therefore, the entropy created has to be transferred to the environment if the system is to return to its original state after a cycle.
- An engine operating around the reversible Carnot cycle may also be used as a device for measuring the temperature. As (cp. (7.36) and (7.37))

  $$\frac{Q_1}{T_1} + \frac{Q_2}{T_2} = 0, \tag{7.46}$$

  we also have

  $$T_1 = T_2 \frac{Q_1}{|Q_2|}. \tag{7.47}$$

  Hence, the measurement of ratios of temperatures has been reduced to the measurements of energies (heat). The definition of the unit for the temperature scale is a matter of convention: To a particular, well-defined state of a certain substance a value is assigned for the temperature. This special state is, following a convention from 1954, the triple point of water, the point where ice, liquid water, and vapor are simultaneously in equilibrium in a completely closed vessel. This triple point exists at a pressure of $6.1 \times 10^2$ Pa and a temperature defined to be $T_0 = 273.16\,\mathrm{K}$ (the numerical value has historical reasons). Remember, the temperature at the melting point of water at the standard pressure $1,013 \times 10^2$ Pa is defined as 0°C, the triple point is then at 0.0098°C.

For an ideal gas at this temperature $T_0$, we must have

$$pV = N k_{\mathrm{B}} T_0. \tag{7.48}$$

Measuring $p$,$V$, and $N$, one determines the value of Boltzmann's constant $k_{\mathrm{B}}$.

- The so-called 'principal laws' (or simply 'laws') of thermodynamics should not remain unmentioned. They summarize some of the most relevant results about energy and entropy within the framework of statistical mechanics. In phenomenological thermodynamics they form the starting point from which many conclusions may be derived.

  The first law is essentially the conservation of energy: 'Energy cannot be created or annihilated'. The second law, however, states: 'Entropy may be created, but can never be annihilated'. A decrease of entropy in one system always implies an export of entropy beyond the borders of the system.

  If entropy is again interpreted as a lack of information, the second law may also be stated as: Information never increases by itself, a decrease of the lack of information can only be accomplished by an import of information.

  Finally, the third law states that the entropy vanishes at the absolute zero of temperature. This statement is also a consequence of the results we found in the previous chapters. More precisely, we may say that at $T = 0$ the system is in its ground state, and thus we obtain for the entropy

$$S = k_{\mathrm{B}} \ln g, \tag{7.49}$$

  where $g$ is the degeneracy of the ground state. Indeed, for $g = 1$ we find $S = 0$, and even if $g \neq 1$ we never will have $S = O(N)$. Therefore, the entropy at $T = 0$ is practically equal to zero compared to the entropy at $T \neq 0$, where it is of the order $N$.

## 7.3 Exergy and Relative Entropy

In the discussion of the Carnot cycle we found that the energy absorbed by a system during a cycle can only partly be transformed into work done by the system. On the other hand, the same amount of energy, added to the system in the form of, say, an electric current, may be transformed into work completely. The various types of energy are therefore different with respect to the relative amount which may be transformed into work. The maximum fraction of an energy form which (in a reversible process) can be transformed into work is called exergy. The remaining part is called anergy, and this corresponds to the waste heat. A heat engine, i.e., an engine which transforms heat into work, will inevitably generate a certain amount of waste heat. During an irreversible realization of the process, which is more or less always the case, even more waste heat is generated than this theoretically unavoidable amount, and part of the exergy is not transformed into work but into anergy.

Let us discuss the exergy in more detail. Consider a system with volume $V$, energy $E$, and entropy $S$. We now ask the question: How much work can be done by this system if it is brought into contact with a heat and volume bath of temperature $T_B$ and pressure $p_B$. This contact partly leads to an exchange of heat with the bath, and partly to work being done. Exactly how much work is done depends on the realization of the contact, but there is an upper limit for the fraction of energy $E$ of the system which can be transformed into work, and this we now shall determine.

Let $\mathrm{d}E$, $\mathrm{d}V$, and $\mathrm{d}S$ be the respective changes of energy, volume, and entropy of the system, and $\mathrm{d}E_B$, $\mathrm{d}V_B$, and $\mathrm{d}S_B$ shall be the corresponding quantities for the bath. Then the following balance relations hold:

$$\mathrm{d}V + \mathrm{d}V_B = 0 \tag{7.50a}$$

$$\mathrm{d}S + \mathrm{d}S_B \geq 0, \tag{7.50b}$$

where the equality sign holds for a reversible process, and

$$\mathrm{d}E + \mathrm{d}E_B = -\mathrm{d}W, \tag{7.51}$$

where $\mathrm{d}W$ is the work done in an infinitesimal step of the process. Due to its size the bath remains in thermal equilibrium, therefore

$$\mathrm{d}S_B = \frac{1}{T_B}\,(\mathrm{d}E_B + p_B\,\mathrm{d}V_B), \tag{7.52}$$

and thus also

$$\mathrm{d}E_B = T_B\,\mathrm{d}S_B - p_B\,\mathrm{d}V_B. \tag{7.53}$$

Hence, the work done in an infinitesimal step of the process is

$$\begin{aligned}
\mathrm{d}W &= -\mathrm{d}E - \mathrm{d}E_B = -\mathrm{d}E - T_B\mathrm{d}S_B + p_B\,\mathrm{d}V_B &\quad (7.54)\\
&= -\mathrm{d}E + T_B\,\mathrm{d}S - p_B\,\mathrm{d}V - T_B(\mathrm{d}S_B + \mathrm{d}S) &\quad (7.55)\\
&\leq -\mathrm{d}E + T_B\,\mathrm{d}S - p_B\,\mathrm{d}V, &\quad (7.56)
\end{aligned}$$

where again the equality sign refers to a reversible process, for which $\mathrm{d}S + \mathrm{d}S_B = 0$.

We now introduce the quantity

$$\Lambda = E - E_0 - T_B(S - S_0) + p_B(V - V_0)\ , \tag{7.57}$$

where $E_0$, $S_0$, and $V_0$ are the values of the state variables of the system after it has settled into equilibrium through the contact with the bath, i.e., after is has assumed the values $T_B$ for the temperature and $p_B$ for the pressure. Obviously, in this case $\Lambda = 0$. According to (7.56), the decrease $-\mathrm{d}\Lambda$ of $\Lambda$ before the system has reached the equilibrium state satisfies

$$\mathrm{d}W \leq -\mathrm{d}\Lambda \, . \tag{7.58}$$

It thus corresponds to the maximum work done, i.e., the work done in a reversible step. Hence, $\Lambda \geq 0$, and $\Lambda$ is exactly the amount of energy of the system which can be transformed into work if the system is brought into contact with a bath of temperature $T_B$ and pressure $p_B$.

Therefore $\Lambda$ corresponds to the exergy. This quantity represents a kind of 'work-account' of a system, i.e., the maximum amount of energy which can be transformed into work. Of course, one always has to specify the kind of bath the system is brought into contact with in order to achieve this gain in work. In the above case we considered a bath with temperature $T_B$ and pressure $p_B$. If we only permit the contact with a heat bath and allow no exchange of volume, we find

$$\Lambda = E - E_0 - T_B(S - S_0) \, . \tag{7.59}$$

This expression resembles the free energy $F = E - TS$ (in the same way as $\Lambda$ in (7.57) resembles the free enthalpy). However, we have to bear in mind that, in contrast to $F$, the exergy depends on the energies and entropies of two states of the system, but only on one temperature, the temperature $T_B$, which finally results for the system in equilibrium with the heat bath.

We will now show that, up to a factor, the exergy corresponds to the relative entropy, i.e.,

$$\Lambda = k_{\mathrm{B}} T_0 \, \mathrm{Tr}\, (\varrho \ln \varrho - \varrho \ln \varrho_0) \, . \tag{7.60}$$

We start by noticing that the above expression is equivalent to

$$\Lambda = k_{\mathrm{B}} T_0 \Big[ \mathrm{Tr}\, (\varrho \ln \varrho) - \mathrm{Tr}\, (\varrho_0 \ln \varrho_0) - \mathrm{Tr}\, (\varrho - \varrho_0) \ln \varrho_0 \Big]. \tag{7.61}$$

With

$$\varrho_0 = \frac{1}{Z_0} \, \mathrm{e}^{-\beta_0 (H + p_0 V)} = \mathrm{e}^{\beta_0 (F_0 - H - p_0 V)} \tag{7.62}$$

it follows in particular that

$$\mathrm{Tr}\, ((\varrho - \varrho_0) \ln \varrho_0) = \mathrm{Tr}\, \Big( (\varrho - \varrho_0) \beta_0 (F_0 - H - p_0 V) \Big) \tag{7.63}$$

$$= \frac{1}{k_{\mathrm{B}} T_0} \Big( -(E - E_0) - p_0 (V - V_0) \Big), \tag{7.64}$$

and therefore, since $k_{\mathrm{B}} \, \mathrm{Tr}\, (\varrho \ln \varrho) = -S$,

$$\Lambda = -T_0 (S - S_0) + (E - E_0) + p_0 (V - V_0) \, , \tag{7.65}$$

which agrees with (7.57). (Note that $T_0 = T_B$ and $p_0 = p_B$ in equilibrium.)

In order to have a measure for the irreversibility of a process, one introduces the exergy efficiency by

$$\zeta = \frac{\Lambda_{\text{out}}}{\Lambda_{\text{in}}} \,, \tag{7.66}$$

where $\Lambda_{\text{out}}$ is the work done by the system, i.e., the exergy emerging from the system, and $\Lambda_{\text{in}}$ is the exergy entering the system.

For a reversible Carnot cycle, we therefore obtain $\zeta = 1$, and for a irreversible process, $\zeta < 1$ is a measure for the exergy transformed into anergy, i.e., the exergy wasted.

For further applications of the notion of exergy, see, e.g., Baehr (1996).

## 7.4 Time Dependence of Statistical Systems

Using the master equation introduced in Sect. 5.2, we now want to describe the evolution in time of a microcanonical system with an energy in the interval $(E, E + \Delta E)$. The state $|n\rangle$ is now the realization of a random variable and the set of functions $\{\varrho_n(t),\, n = 1, \ldots\}$ represents for each time $t$ a probability: $\varrho_n(t)$ denotes the probability that the system is in the state $|n\rangle$ at time $t$. This, of course, is identically zero if $E_n$ is not in the interval $(E, E + \Delta E)$. The transition rates $w_{nn'}$ correspond exactly to those quantities which are computed in quantum mechanics, e.g., by using Fermi's golden rule. The invariance under time reversal, which is fundamental in the framework of quantum mechanics, implies (cf. van Kampen 1985) that the matrix of these transition rates is symmetric, i.e.,

$$w_{nn'} = w_{n'n} \qquad \text{for} \qquad E_n,\, E_{n'} \in (E, E + \Delta E)\,. \tag{7.67}$$

Of course, for a canonical system one obtains different transition rates. We shall denote them by $w^C_{nn'}$. It can be shown that they satisfy the following relation:

$$w^C_{nn'}\, \mathrm{e}^{-\beta E_{n'}} = w^C_{n'n}\, \mathrm{e}^{-\beta E_n}. \tag{7.68}$$

These two relations for the transition rates may be combined as

$$w_{nn'}\, \varrho^{\mathrm{e}}_{n'} = w_{n'n}\, \varrho^{\mathrm{e}}_{n}, \tag{7.69}$$

where $\varrho^{\mathrm{e}}_n$ is the equilibrium probability distribution known from statistical mechanics, i.e., for a microcanonical system

$$\varrho^{\mathrm{e}}_n = \begin{cases} \text{const.} \neq 0, \text{ if } E_{n'} \in (\,E,\; E + \Delta E\,) \\ 0 \qquad\qquad\quad \text{otherwise}\,, \end{cases} \tag{7.70}$$

and for a canonical system

$$\varrho_n^{\mathrm{e}} = \frac{1}{Z}\,\mathrm{e}^{-\beta E_n}\ . \tag{7.71}$$

Relation (7.69) is nothing other than the equation of detailed balance, which we already have met in Sect. 5.2. It guarantees that the equilibrium distributions $\{\varrho_n^{\mathrm{e}},\ n = 1,\ldots\}$, computed in the statistical mechanics of equilibrium systems, are stationary solutions of the master equation. This follows, because the master equation implies for the density $\{\varrho_n^{\mathrm{stat}},\ n = 1,\ldots\}$ the condition

$$\sum_{n'} w_{nn'}\,\varrho_{n'}^{\mathrm{stat}} = \sum_{n'} w_{n'n}\,\varrho_n^{\mathrm{stat}}\ , \tag{7.72}$$

and $\{\varrho_n^{\mathrm{e}},\ n = 1,\ldots\}$ satisfies this condition even term by term.

Hence, the equilibrium densities are also stationary solutions of the master equation.

*Remark.* The interpretation given makes it plausible that the validity of the condition of detailed balance is related to invariance under time reversal. A simple system for which neither detailed balance nor the invariance under time reversal holds is the following: Consider a system with three states. The stationary probabilities for the states are equal, and the transition rates are $w_{12} > w_{21}$, $w_{23} > w_{32}$, $w_{31} > w_{13}$. Hence, detailed balance is not fulfilled. But we also see that the cycle $1 \to 2 \to 3 \to 1$ is more likely than the cycle $1 \to 3 \to 2 \to 1$. Invariance under time reversal does not hold: A video film of such a process, running backwards, would show the less probable cycle more frequently.

We shall now prove that every solution $\varrho_n(t)$ of the master equation in the limit $t \to \infty$ tends to the stationary solution $\{\varrho_{n'}^{\mathrm{stat}}\}$, for which (7.72) holds. This is the equilibrium distribution resulting from the master equation.

For this purpose we consider a system with nonequilibrium probability density $\{\varrho_n(t),\ n = 1,\ldots\}$. These probabilities shall satisfy a master equation which also admits a stationary density $\{\varrho_n^{\mathrm{stat}},\ n = 1,\ldots\}$ as a solution.

We study the relative entropy

$$S_{\mathrm{rel}}(t) = -k_{\mathrm{B}} \sum_n \varrho_n(t)\,\ln\left(\frac{\varrho_n(t)}{\varrho_n^{\mathrm{stat}}}\right). \tag{7.73}$$

It not only satisfies $S_{\mathrm{rel}}(t) \le 0$, which every relative entropy does, but we will show that, furthermore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,S_{\mathrm{rel}}(t) \ge 0, \tag{7.74}$$

i.e., the relative entropy increases monotonically or remains constant.

*Proof.* We write $S_{\mathrm{rel}}(t)$ in the form

$$S_{\mathrm{rel}}(t) = -k_{\mathrm{B}} \sum_n \varrho_n^{\mathrm{stat}}\, f\left(\frac{\varrho_n(t)}{\varrho_n^{\mathrm{stat}}}\right)\ , \tag{7.75}$$

with $f(x) = x \ln x$. $f(x)$ is a convex function for $x > 0$, i.e., $f''(x) > 0$ for $x > 0$.

Defining $x_n = \varrho_n(t)/\varrho_n^{\text{stat}}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} S_{\text{rel}}(t) = -k_{\text{B}} \sum_n f'(x_n) \sum_{n'} (w_{nn'}\, \varrho_{n'} - w_{n'n}\, \varrho_n) \tag{7.76}$$

$$= -k_{\text{B}} \sum_{nn'} w_{nn'}\, \varrho_{n'}^{\text{stat}} \left(f'(x_n)\, x_{n'} - f'(x_{n'})\, x_{n'}\right) . \tag{7.77}$$

For arbitrary $\{\Psi_n\}$ we obtain from (7.72)

$$\sum_{nn'} w_{nn'}\, \varrho_{n'}^{\text{stat}}\, \Psi_n = \sum_{nn'} w_{n'n}\, \varrho_n^{\text{stat}}\, \Psi_n \tag{7.78}$$

$$= \sum_{nn'} w_{nn'}\, \varrho_{n'}^{\text{stat}}\, \Psi_{n'} , \tag{7.79}$$

and therefore also for $\Psi_n = f(x_n) - x_n\, f'(x_n)$

$$0 = -k \sum_{nn'} w_{nn'}\, \varrho_{n'}^{\text{stat}} \left(f(x_n) - f(x_{n'}) - x_n\, f'(x_n) + x_{n'}\, f'(x_{n'})\right) . \tag{7.80}$$

Adding this equation to (7.77) yields

$$\frac{\mathrm{d}}{\mathrm{d}t} S_{\text{rel}}(t) = -k_{\text{B}} \sum_{nn'} w_{nn'}\, \varrho_{n'}^{\text{stat}} \left(f(x_n) - f(x_{n'}) + f'(x_n)(x_{n'} - x_n)\right) . \tag{7.81}$$

Now, for $0 \le \delta \le 1$,

$$\begin{aligned} f(x_{n'}) = {} & f(x_n) + f'(x_n)(x_{n'} - x_n) \\ & + \frac{1}{2}\, f''(x_n + \delta(x_{n'} - x_n))(x_{n'} - x_n)^2 , \end{aligned}$$

and thus, since $f''(x) > 0$ for all $x > 0$,

$$\begin{aligned} & f(x_n) - f(x_{n'}) + f'(x_n)(x_{n'} - x_n) \\ & \quad = -\frac{1}{2}\, f''(x_n + \delta(x_{n'} - x_n))(x_{n'} - x_n)^2 \; < \; 0 . \end{aligned}$$

Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t} S_{\text{rel}}(t) > 0 , \tag{7.82}$$

unless $x_n = x_{n'}$ for all pairs $(n, n')$ for which $w_{nn'} \neq 0$, which would also imply that $x_n$ assumes the same value for all states which are accessible by transitions from some given state. In general this set includes all states, since otherwise the set of states decomposes into disjoint subsets among which there are no transitions, and the system will always remain in one of these subsets. This case will not be considered here.

Hence, $S_{\text{rel}}(t)$ increases permanently, but can never become positive. It therefore has to tend towards a limit value for which $\frac{\mathrm{d}}{\mathrm{d}t} S_{\text{rel}}(t) = 0$, and thus $x_n = \varrho_n(t)/\varrho_n^{\text{stat}}$ has the same value for all $n$ in this limit, i.e.,

$$\lim_{t\to\infty} \varrho_n(t) = \alpha\, \varrho_n^{\text{stat}} \,. \tag{7.83}$$

Normalization tells us that $\alpha = 1$ and therefore $\lim_{t\to\infty} \varrho_n(t) = \varrho_n^{\text{stat}}$.

We have shown therefore that for large times every solution $\varrho_n(t)$ tends towards the stationary solution $\varrho_n^{\text{stat}}$.

In Sect. 7.3 we have seen that the relative entropy agrees with the exergy to within a constant factor. We found

$$\Lambda(t) = -k_{\mathrm{B}} T_0\, S_{\text{rel}}(t) \,, \tag{7.84}$$

where $T_0$ is the temperature of the system after it has reached equilibrium. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t} \Lambda(t) \leq 0 \,, \tag{7.85}$$

where the equality only holds for the equilibrium state. We found for an isolated system

$$\Lambda(t) = -T_0(S(t) - S_0); \tag{7.86}$$

for a system which is brought into in contact with a heat bath of temperature $T_0$

$$\Lambda(t) = E(t) - E_0 - T_0(S(t) - S_0); \tag{7.87}$$

and for a system approaching equilibrium by contact with a heat and volume bath

$$\Lambda(t) = E(t) - E_0 - T_0(S(t) - S_0) + p_0(V(t) - V_0) \,. \tag{7.88}$$

If, therefore, for systems not in equilibrium, we define

- The entropy $S(t) = -k_{\mathrm{B}} \sum_n \varrho_n(t) \,\ln \varrho_n(t)$
- The free energy $F(t) = \sum_n E_n\, \varrho_n(t) - T_0\, S(t)$
- The free enthalpy $G(t) = F(t) + p_0\, V(t)$,

where $T_0$ and $p_0$ are the corresponding state variables after the system has finally reached equilibrium, then the monotonic decrease of $\Lambda(t)$ until equilibrium implies that the entropy increases monotonically, the free energy and the free enthalpy decrease monotonically, until in equilibrium they assume their maximum and minimum values, respectively.

We conclude that, provided the temporal behavior of a system can be described by a master equation, one can derive the statement of the second law of thermodynamics, i.e., that the entropy can only increase or remain constant, but never decrease.

*Remark.* A function $H(t)$ with $H(t) \geq 0$ and $\mathrm{d}H(t)/\mathrm{d}t \leq 0$ is called a Lyapunov function. The properties of such a function imply that it converges to a constant for $t \rightarrow \infty$. The exergy, or the negative relative entropy, is therefore a Lyapunov function.

From the relative entropy we can derive a second Lyapunov function. Following (Jiu-li et al. 1984), the time derivative of $S_{\mathrm{rel}}$ may also be written as

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\mathrm{d}S_{\mathrm{c}}}{\mathrm{d}t} + \frac{\mathrm{d}S_{\mathrm{ex}}}{\mathrm{d}t} \,, \tag{7.89}$$

where

$$\frac{\mathrm{d}S_{\mathrm{c}}}{\mathrm{d}t} = \frac{1}{2}\sum_{nn'}[w_{nn'}\varrho_{n'}(t) - w_{n'n}\varrho_n(t)] \ln\left(\frac{w_{nn'}\varrho_{n'}(t)}{w_{n'n}\varrho_n(t)}\right) \tag{7.90}$$

corresponds to the entropy production and

$$\frac{\mathrm{d}S_{\mathrm{ex}}}{\mathrm{d}t} = -\frac{1}{2}\sum_{nn'}[w_{nn'}\varrho_{n'}(t) - w_{n'n}\varrho_n(t)] \ln\left(\frac{w_{nn'}}{w_{n'n}}\right) \tag{7.91}$$

to the flow of entropy. In the stationary case, i.e., when $\rho_n(t) \equiv \rho_n^{\mathrm{e}}$ satisfies detailed balance, both contributions vanish by construction. In general one obtains for the entropy production alone

$$\frac{\mathrm{d}S_{\mathrm{c}}}{\mathrm{d}t} \geq 0 \,. \tag{7.92}$$

Furthermore, it can be shown that in all cases

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}S_{\mathrm{c}}}{\mathrm{d}t}\right) \leq 0 \,, \tag{7.93}$$

i.e., the time derivative of the entropy production is also a Lyapunov function.

Now consider the case where a system is forced to remain in a state of nonequilibrium by some supplementary condition such that entropy is continuously produced. This nonequilibrium state must be of such a form that the production of entropy is minimum, because, according to (7.93), this quantity gets smaller and smaller and will assume its minimum value in the stationary nonequilibrium state.

This is called the principle of minimum entropy production (see also Nicolis and Prigogine 1977).

# Chapter 1
# Statistical Physics: Is More than Statistical Mechanics

It is a general belief that physics is a science where fundamental laws of nature are established. One expects these laws to have a deterministic character. A student being introduced to physics first encounters the fundamental theories, for instance Newtonian mechanics, electrodynamics, and the theory of relativity, and sees in these the ideals of a scientific theory.

Entering quantum mechanics and its statistical interpretation the student is confronted – often for the first time – with concepts of probability theory. For practical applications of quantum mechanics this is of no major relevance; he or she learns to get the solution by solving a differential equation.

But in statistical mechanics it seems that statistics has to be taken serious. However, when one inspects textbooks on statistical mechanics to see how they present the concepts of probability theory, one finds, with the exception of a few recent mathematically inclined books, that they get along with only a few of these notions. And, where mentioned at all, they are presented in such a limited and specialized way that the reader can merely guess their significance for other applications. However, a physicist who is confronted with the interpretation of experimental results for a thesis, or who wishes to analyze the effects of complex systems in his or her later profession, will almost inevitably find that this cannot be done without the application of statistical methods. Limited information about the dynamics of a system and uncontrollable influences on the system often mean that is can only be considered as a statistical system. And apart from that, when analyzing experimental data, one always has to take into account that any measured value has an uncertainty expressed by its error. One should not only expect limited knowledge but one should also know how to deal with it. One has to be acquainted with mathematical tools allowing one to quantify uncertainties both in the knowledge about a system to be modeled as well as in the knowledge that can be inferred from experimental data of a system. These tools are probability theory and statistics.

Statistics plays a special role in the description and analysis of complex systems; a physicist will need statistics not only in the context of statistical mechanics. For this reason the present book 'Statistical Physics – An Advanced Approach with Applications' introduces more concepts and methods in statistics than are needed

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_1, © Springer-Verlag Berlin Heidelberg 2012

for statistical mechanics alone. In particular, it includes methods from mathematical statistics and stochastics which are important in the analysis of data. It is hoped that the reader will learn to appreciate the usefulness and indispensability of statistical concepts and methods for the modeling and the analysis of general complex systems.

The six chapters (Chaps. 2–7) that constitute Part I of this book, 'Modeling of Statistical Systems', will present statistical mechanics from a probabilistic point of view so that one can recognize the general character of the concepts and methods used. Chapter 2 introduces important fundamental notions and methods which will be relevant for any study of a statistical system, for example, random variables, the large deviation property and renormalization transformations. In the subsequent chapters the reader will meet more and more complex random variables or densities. Random variables in the state space of classical mechanics are the background for the statistical mechanics of classical fluids in Chap. 3, random variables on a lattice the background for spin systems and images in Chap. 4. With time dependent random variables one can formulate (Chap. 5) the dynamics of statistical systems. The case where the random variable is a quantum state of a system leads to quantum statistics. This is treated in Chap. 6, where we will concentrate on ideal gases and systems that can be successfully described as an ideal gas of quasiparticles. This is traditionally the central part of statistical mechanics. The very interesting topic of quantum fluids and quantum stochastic processes, however, is beyond the scope of this introduction to statistical physics. In Chap. 7 the connection to thermodynamics is established.

Part II is concerned with the analysis of statistical systems, a topic which does not traditionally belong to theoretical physics. But really it should! In three chapters the concept of an estimator is introduced and developed. Typical estimators for inferring parameters or distributions from experimental data, e.g. least squares, maximum likelihood, and regularization estimators for inverse problems, are given a sound statistical basis. There is a great ignorance among many scientists about these methods and a sharp contrast between the effort that goes into generating data on the one hand and analysing it on the other. The student should learn these methods and the corresponding concepts in a theoretical course. One should be careful attach enough importance to this: The average physicist is more likely to be confronted with an inverse problem than with the need to solve a problem by renormalization group methods.

The selection of topics for a textbook is always a difficult problem. Opinions on what a student really needs to know are, apart from an undisputed core, quite diverse. The selection made in this book results from many years of experience in collaboration with colleagues both in physics and in various neighboring disciplines which physicists are traditionally likely to enter. This book aims to provide a broad and precise fundamental knowledge in statistical physics, designed to prepare the student for his or her work in a modern and continuously evolving research landscape.

# Part II
# Analysis of Statistical Systems

# Chapter 8
# Estimation of Parameters

In Chap. 2 we introduced random variables, their corresponding densities, and quantities characterizing these densities. We have also seen how to compute with random variables.

Up to now, we have avoided a discussion about the interpretation of the probability density. Of course, what we always had in mind is the so-called relative frequency interpretation of the probability density: One assumes that many realizations of the random variable exist. Then a histogram where the relative frequencies of the measured values in so-called bins, i.e., small intervals, are plotted against the intervals themselves, should converge with increasing number of realizations to the probability density.

But one can also conceive situations for which repeatable realizations do not exist. In this case the probability density has to be interpreted as something like a belief function. It seems reasonable to postulate the applicability of the Kolmogorov axioms also for such apparently subjective functions, since these axioms really just represent the rules for common sense guessing.

In this context, however, it will almost always taken for granted that frequent realizations of a random variable are possible, i.e., that the relative frequency interpretation will be appropriate. Only in the case of an unknown systematic error will we have to interpret a density as a belief function (cf. the final remark at the end of Sect. 8.2).

Therefore, we will assume in the following that $N$ observations have been made, and that their results may be considered as realizations of a random variable. The random character stems from the fact that every measurement is affected at least by an experimental error. This error can be modeled by a random variable, whose probability density has to be chosen consistent with the statistics of the experimental error. We thus always have in mind that an observation $x$ is a realization of a random variable $X$, and writing

$$X = \langle X \rangle + \eta, \tag{8.1}$$

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_8, © Springer-Verlag Berlin Heidelberg 2012

we decompose this random variable into its mean, which could be viewed as the true value of the observed quantity, and its deviation from the mean, $\eta$, which carries the statistical properties of $X$. The standard deviation $\sigma = \sqrt{\text{Var}(\eta)}$ is therefore often called the standard error of the observation.

In the first Sects. 8.1–8.4 we will assume that the observations are obtained under constant external conditions. A first task then is to find, for example, the true value $\langle X \rangle$ and the standard error from a number of observations. We will also illustrate in Sect. 8.4 how one can deduce a density from the statistical properties of the observations. In Sect. 8.5 we then treat the case where one parameter (or more) are changed during the experiment and where one is interested in how the experimental result depends on this parameter. The information provided by the observations will then be used to establish a model of the dependence.

## 8.1 Samples and Estimators

A set of $N$ realizations of a random variable $X$ is called a sample of size $N$. The determination of quantities characterizing the density $\varrho_X(x)$, such as the mean $\langle X \rangle$, the variance, or even the density itself, on the basis of such a sample is referred to as an 'estimation'. In this context we will only be concerned with so-called point estimations, where, instead of whole densities or distributions, only single parameters are estimated.

How well the estimated values agree with the true values depends, of course, on the type of sample and the estimation method.

Sample theory is a subject of mathematical statistics, which we cannot enter here in more detail, but see, e.g., Hartung et al. (1986). For the moment we will assume that each realization $x_i$ from the sample $\{x_1, \ldots, x_n\}$ of a random variable may also be considered as a realization of a random variable $X_i$, where the $X_i$ all have the same density as $X$; i.e., the random variables $X_i$ are all 'copies' of the random variable $X$. Furthermore, we will first assume that the $X_i$ are mutually independent, i.e., the realization $x_i$ is independent of the other realizations of $X$. Under these conditions one also speaks of a random sample.

An estimator $Z$ for a quantity to be estimated from the realizations of the $\{X_i\}$ is a function of these $\{X_i\}$, i.e., it is also a random variable whose realizations depend on the form of this function as well as the realizations of the $\{X_i\}$. The density of $Z$ will also depend on the functional form and should have the property that the realizations of $Z$ scatter as little as possible around the quantity to be estimated.

These statements are illuminated more clearly by the first example in Sect. 2.5: The random variable

$$Z = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{8.2}$$

may serve as an estimator for the expectation value $\langle X\rangle$. Indeed, as we have shown in Sect. 2.5.3, the density of $Z$ has the mean value $\langle X\rangle$ and the variance $\frac{1}{N}\mathrm{Var}(X)$. With $N\to\infty$ the density of $Z$ approaches a Gaussian distribution with a smaller and smaller variance around $\langle X\rangle$. For sufficiently large values of $N$, there is a large probability that any realization of $Z$ will be close to $\langle X\rangle$ (we will make this statement more precise when we introduce the concept of confidence intervals).

The quantity to be estimated, for example, $\langle X\rangle$, will be denoted in the following by $\theta$ and the corresponding estimator by $\hat{\Theta}$. Therefore, $\hat{\Theta}\equiv\hat{\Theta}(X_1,\dots,X_N)$ is a random variable with realizations $\hat{\theta}(x_1,\dots,x_N)$, and $\theta$ is a number.

We now describe some particular types of estimator and their application.

An estimator $\hat{\Theta}(X_1,\dots,X_N)$ for $\theta$ is called *unbiased* if

$$\langle\hat{\Theta}(X_1,\dots,X_N)\rangle=\theta. \tag{8.3}$$

Estimators which are not unbiased are said to possess a bias, defined by

$$\mathrm{Bias}\,(\hat{\Theta})=\langle\hat{\Theta}\rangle-\theta, \tag{8.4}$$

which is then different from zero.

Hence, $Z$, as defined in (8.2), is an unbiased estimator for $\langle X\rangle$, because

$$\langle\hat{\Theta}\rangle=\langle Z\rangle=\frac{1}{N}\sum_{i=1}^{N}\langle X_i\rangle=\langle X\rangle\equiv\theta. \tag{8.5}$$

An estimator $\hat{\Theta}(X_1,\dots,X_N)$ for $\theta$ is called *consistent*, if the probability that the distance between a realization of $\hat{\Theta}$ and the true value $\theta$ is larger than an $\varepsilon>0$, tends to zero for increasing sample size, i.e., if

$$\mathcal{P}\left(|\hat{\theta}(x_1,\dots,x_N)-\theta|>\varepsilon\right)\to 0\quad\text{for}\quad N\to\infty. \tag{8.6}$$

Thus, the more observations are available, the better is the estimation of $\theta$.

The quantity $Z$, as defined in (8.2), is also a consistent estimator for $\langle X\rangle$, because $\mathrm{Var}(Z)=\frac{1}{N}\,\mathrm{Var}(X)$.

The *mean squared error*, MSE, is defined by

$$\mathrm{MSE}(\hat{\Theta},\theta)=\left\langle\left(\hat{\Theta}-\theta\right)^2\right\rangle. \tag{8.7}$$

We have

$$\mathrm{MSE}(\hat{\Theta},\theta)=\mathrm{Var}(\hat{\Theta})+\left(\mathrm{Bias}(\hat{\Theta})\right)^2, \tag{8.8}$$

since

$$\left\langle \left(\hat{\Theta} - \theta\right)^2 \right\rangle = \langle \hat{\Theta}^2 \rangle - 2\theta \, \langle \hat{\Theta} \rangle + \theta^2 \tag{8.9a}$$

$$= \left( \langle \hat{\Theta}^2 \rangle - \langle \hat{\Theta} \rangle^2 \right) + \left( \langle \hat{\Theta} \rangle - \theta \right)^2 . \tag{8.9b}$$

*Examples.* We will examine three estimators for the variance of a random variable. We first examine the estimator

$$\hat{\Theta}_1(X_1, \ldots, X_N) = \frac{1}{N-1} \sum_{i=1}^{N} \left( X_i - \hat{X}_N \right)^2 , \tag{8.10}$$

where $\hat{X}_N$ is the estimator for the expectation value $\langle X \rangle$,

$$\hat{X}_N \equiv Z = \frac{1}{N} \sum_{i=1}^{N} X_i . \tag{8.11}$$

In order to determine $\langle \hat{\Theta}_1 \rangle$, we first consider an individual summand in (8.10):

$$\begin{aligned} \left\langle \left( X_i - \hat{X}_N \right)^2 \right\rangle &= \left\langle \, \left( \, (X_i - \langle X \rangle) - (\hat{X}_N - \langle X \rangle) \, \right)^2 \right\rangle \\ &= \mathrm{Var}(X_i) - 2 \left\langle (X_i - \langle X \rangle) \, (\hat{X}_N - \langle X \rangle) \right\rangle + \mathrm{Var}(\hat{X}_N) \\ &= \mathrm{Var}(X) - 2 \, \frac{1}{N} \, \mathrm{Var}(X) + \frac{1}{N} \, \mathrm{Var}(X) \\ &= \frac{N-1}{N} \, \mathrm{Var}(X), \end{aligned} \tag{8.12}$$

since

$$\begin{aligned} \left\langle (X_i - \langle X \rangle) \, (\hat{X}_N - \langle X \rangle) \right\rangle &= \frac{1}{N} \sum_{j=1}^{N} \left\langle (X_i - \langle X \rangle) \, (X_j - \langle X \rangle) \right\rangle \\ &= \frac{1}{N} \left\langle \left( X_i - \langle X \rangle \right)^2 \right\rangle \\ &= \frac{1}{N} \, \mathrm{Var}(X). \end{aligned} \tag{8.13}$$

Therefore,

$$\langle \hat{\Theta}_1(X_1, \ldots, X_N) \rangle = \frac{1}{N-1} \, \frac{N-1}{N} \sum_{i=1}^{N} \mathrm{Var}(X) = \mathrm{Var}(X). \tag{8.14}$$

Hence, this estimator is unbiased.

Had we made the same calculation for the estimator

$$\hat{\Theta}_2(X_1,\ldots,X_N) = \frac{1}{N}\sum_{i=1}^{N}\big(X_i - \hat{X}_N\big)^2, \tag{8.15}$$

we would have obtained:

$$\langle\hat{\Theta}_2(X_1,\ldots,X_N)\rangle = \frac{N-1}{N}\,\mathrm{Var}(X), \tag{8.16}$$

and therefore

$$\mathrm{Bias}(\hat{\Theta}_2) = -\frac{1}{N}\,\mathrm{Var}(X). \tag{8.17}$$

On the other hand, suppose the expectation value $\langle X\rangle$ is known, then

$$\hat{\Theta}_3(X_1,\ldots,X_N) = \frac{1}{N}\sum_{i=1}^{N}\big(X_i - \langle X\rangle\big)^2 \tag{8.18}$$

is an unbiased estimator for the variance of $X$, as may easily be seen from the computation (8.12).

It can be shown (e.g. Honerkamp 1994) that $\hat{\Theta}_1$ is also a consistent estimator. In particular, when $X$ is normally distributed, we find

$$\mathrm{Var}\big(\hat{\Theta}_1(X_1,\ldots,X_N)\big) = 2\,\frac{\big(\mathrm{Var}(X)\big)^2}{(N-1)}. \tag{8.19}$$

### 8.1.1 Monte Carlo Integration

Numerical methods play an important role in statistical physics. When the realizations of random variables are simulated, the names for the methods are often given the name Monte Carlo, alluding to the famous casino.

In Monte Carlo integration the integral to be calculated is interpreted as the expectation value of a certain function of some random variable. This expectation value is then estimated by some sample. There are various versions of Monte Carlo integration, see e.g. Binder (1979) and Honerkamp (1994). These methods differ according to how the integral to be calculated is represented as an expectation value.

Let $\boldsymbol{x}$ be an arbitrary vector in $\mathbb{R}^n$ and $G$ some domain in $\mathbb{R}^n$. The integral

$$I = \int_G \mathrm{d}\boldsymbol{x}\; f(\boldsymbol{x}) \tag{8.20}$$

shall be determined. To do this we proceed as follows:

We can reformulate the integral as

$$I = \int_G \mathrm{d}\boldsymbol{x}\, \frac{f(\boldsymbol{x})}{\varrho(\boldsymbol{x})}\, \varrho(\boldsymbol{x}) = \langle F(\boldsymbol{X}) \rangle, \tag{8.21}$$

with

$$F(\boldsymbol{X}) = \frac{f(\boldsymbol{X})}{\varrho(\boldsymbol{X})}\, I_G(\boldsymbol{X}), \tag{8.22}$$

where $I_G(\boldsymbol{x}) = 1$, if $x \in G$, and $I_G(\boldsymbol{x}) = 0$ otherwise. The density $\varrho(\boldsymbol{x})$ has to be chosen suitably. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ be independent random variables, distributed according to $\varrho(\boldsymbol{x})$. Then an estimator for $I$ is given by

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} F(\boldsymbol{X}_i), \tag{8.23}$$

with variance

$$\mathrm{Var}(\hat{I}) = \frac{1}{N}\, \mathrm{Var}\big(F(\boldsymbol{X})\big), \tag{8.24}$$

where

$$\mathrm{Var}\big(F(\boldsymbol{X})\big) = \int_G \mathrm{d}\boldsymbol{x} \left( \frac{f(\boldsymbol{x})}{\varrho(\boldsymbol{x})} \right)^2 \varrho(\boldsymbol{x}) - I^2. \tag{8.25}$$

The density $\varrho(\boldsymbol{x})$ may be chosen in various ways, for example:

- $\varrho(\boldsymbol{x}) =$ const. in $G$. This corresponds to the simplest method. In general it is not difficult to generate random numbers which are uniformly distributed in some domain. This method is called the standard method.
- $\varrho(\boldsymbol{x})$ is large where $f(\boldsymbol{x})$ is large, and small where $f(\boldsymbol{x})$ is small. In this case, however, one has to know how to generate random numbers distributed with respect to such a density. The integrals appearing in statistical mechanics often contain a density function such as

$$\varrho(\boldsymbol{x}) = \frac{1}{Z}\, \mathrm{e}^{-\beta H(x)}. \tag{8.26}$$

  This method of Monte Carlo integration, which is also referred to as the method of importance sampling, therefore finds its natural application in this field. The essential advantage of this method is that the variance of $\hat{I}$ for a suitably chosen density may be much smaller than the variance of an estimator based on a constant density.

An alternative way of determining $I$ is to extend the spatial dimension by one and rewrite the integral in the form

$$\begin{aligned} I &= \int_G \mathrm{d}\boldsymbol{x} \int_0^{f(\boldsymbol{x})} \mathrm{d}y \\ &= \int_G \mathrm{d}\boldsymbol{x} \int_0^{\max\left(f(\boldsymbol{x})\right)} \mathrm{d}y \, \frac{\varTheta\left(f(\boldsymbol{x}) - y\right)}{\varrho(\boldsymbol{x}, y)} \, \varrho(\boldsymbol{x}, y) \\ &= \langle F(\boldsymbol{X}, Y) \rangle, \end{aligned} \tag{8.27}$$

with

$$F(\boldsymbol{X}, Y) = \frac{\varTheta\left(f(\boldsymbol{X}) - Y\right)}{\varrho(\boldsymbol{X}, Y)}, \tag{8.28}$$

where $\varrho(x, y)$ is some suitable density and $\varTheta(x)$ denotes the Heaviside step function (cf. (3.6)). A realization $\hat{\imath}$ of the estimator $\hat{I}$ is now given by

$$\hat{\imath} = \frac{1}{N} \sum_{i=1 \, y_i \leq f(\boldsymbol{x}_i)}^{N} \frac{1}{\varrho(\boldsymbol{x}_i, y_i)}, \tag{8.29}$$

where $\{(\boldsymbol{x}_i, y_i), \, i = 1, \ldots, N\}$ are realizations with respect to $\varrho(\boldsymbol{x}, y)$. For the variance we obtain

$$\mathrm{Var}(\hat{I}) = \frac{1}{N} \left[ \int_G \mathrm{d}\boldsymbol{x} \int_0^{f(\boldsymbol{x})} \mathrm{d}y \, \frac{1}{\varrho(\boldsymbol{x}, y)} - I^2 \right], \tag{8.30}$$

which may be estimated similarly. This is called the scoring method.

The three estimators for the integral $I$ presented above are all unbiased by construction, but they differ with respect to their variance. For practical applications one usually favors the estimator which in a given time leads to the estimation with smallest variance. The faster the corresponding random numbers can be generated, the larger the sample size that can be chosen.

*Examples.*

- We consider the function $y = f(x)$, given by the parametrization $y = \cos\varphi$, $x = \sin\varphi$, $0 \leq \varphi \leq \pi/2$. It describes the upper right quarter of a circle with radius 1 around the origin. The area below the curve of this function is equal to $\pi/4$. We want to estimate the integral yielding this area using the scoring method.

  We choose the density for which $\varrho(x, y) = 1$ in the unit quadrant $0 \leq x, y \leq 1$, generate $N$ random numbers uniformly distributed in this unit quadrant (Fig. 8.1)

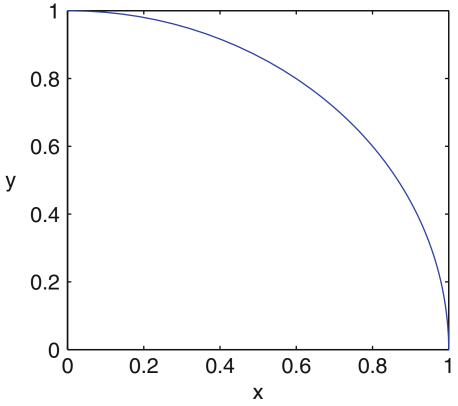


**Fig. 8.1** Unit quadrant and quarter of the unit *circle*

and count those numbers for which $x^2 + y^2 \leq 1$. Let their number be $N'$. The estimation for $\pi/4$ is now

$$\hat{\imath} = \frac{N'}{N}. \tag{8.31}$$

For the variance we find

$$\mathrm{Var}(\hat{I}) = \frac{1}{N}\left(I - I^2\right). \tag{8.32}$$

- As a second example, we wish to determine the integral

$$I = \frac{1}{\sqrt{2\pi}} \int_{-b}^{b} \mathrm{d}x\, \mathrm{e}^{-x^2/2}. \tag{8.33}$$

Let us consider two approaches:

1. We generate random numbers uniformly distributed in the interval $[-b, b]$. The standard method now yields

$$\hat{\imath} = \frac{2b}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_i^2/2}. \tag{8.34}$$

2. We choose $\varrho(x) = \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-x^2/2}$, i.e., we use random numbers normally distributed around $x = 0$ with variance 1 and rewrite the integral $I$ in the form

$$I = \int_{-\infty}^{\infty} g(x)\, \varrho(x)\, \mathrm{d}x \tag{8.35}$$

where

$$g(x) \equiv \frac{f(x)}{\varrho(x)} = \begin{cases} 1 & \text{in } [-b, b]\,, \\ 0 & \text{otherwise.} \end{cases} \tag{8.36}$$

Now

$$\hat{\imath} = \frac{1}{N} \sum_{i=1}^{N} g(x_i) = \frac{N'}{N}, \tag{8.37}$$

where $N'$ counts the normal random numbers in the interval $[-b, b]$.

In order to compare the variances of the above two methods, we have to compute, according to (8.25), the two integrals

$$2b \int_{-b}^{b} \frac{1}{2\pi}\, \mathrm{e}^{-x^2}\, \mathrm{d}x \tag{8.38}$$

and

$$\int_{-b}^{b} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-x^2/2}\, \mathrm{d}x. \tag{8.39}$$

For sufficiently large $b$ the first integral is larger than the second and method (2) leads to a smaller variance. The critical value for $b$ is roughly

$$b_0 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, \mathrm{d}x \Big/ 2 \int_{-\infty}^{\infty} \frac{1}{2\pi}\, \mathrm{e}^{-x^2}\, \mathrm{d}x = \sqrt{\pi}. \tag{8.40}$$

For further methods of variance reduction, see Rubinstein (1981) and Hengartner and Theodorescu (1978).

Compared to other integration methods (Press et al. 2007), Monte Carlo integration is competitive only for higher dimensional integrals. For more than about 20 dimensions Monte Carlo integration is the only possible method.

## 8.2 Confidence Intervals

An estimator $\hat{\Theta}(X_1, \ldots, X_N)$ for a quantity $\theta$ generally yields a different number for each sample $\{x_1, \ldots, x_N\}$. We have already seen that the realizations of a consistent estimator $\hat{\Theta}$ show less scatter around $\theta$ as the sample becomes larger. We can make such a statement more precise by specifying an interval such that the true value of the quantity $\theta$ lies within this interval with a probability of $(1 - \alpha)\, 100\%$. We shall see in a moment how to choose $\alpha$ suitably.

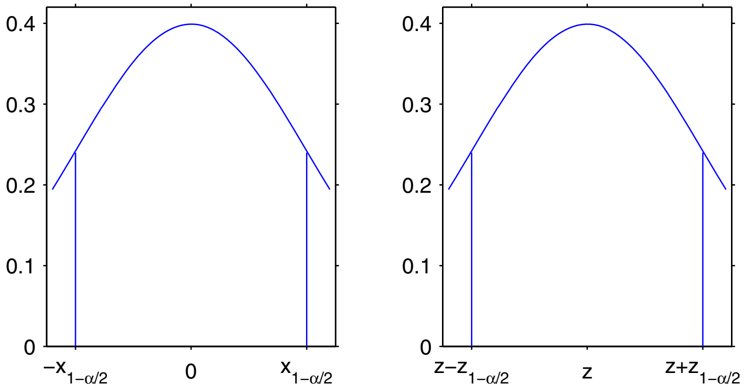


**Fig. 8.2** *Left*: $Y$ has approximately a standard normal distribution, i.e., $100\,(1-\alpha)\%$ of the realizations of $Y$ lie in the interval $(-x_{1-\alpha/2}, x_{1-\alpha/2})$. *Right*: The true value therefore lies within the interval $(z - z_{1-\alpha/2}, z + z_{1-\alpha/2})$, where $z_{1-\alpha/2} = \sigma_X x_{1-\alpha/2}/\sqrt{N}$, with a probability of $100\,(1-\alpha)\%$

Let us again begin by studying the case where $Z$ is the estimator of $\langle X \rangle$ as defined in (8.2). If the density of $Z$ were exactly a Gaussian distribution with mean $\langle X \rangle$ and variance $\frac{1}{N}\,\mathrm{Var}(X) \equiv \frac{1}{N}\,\sigma_X^2$, the quantity

$$Y = \sqrt{N}\,\frac{Z - \langle X \rangle}{\sigma_X} \tag{8.41}$$

would be standard normally distributed. For such a random variable, $(1-\alpha)\,100\%$ of the realizations are inside the interval $(-x_{1-\alpha/2}, x_{1-\alpha/2})$, where $x_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution, and $x_{\alpha/2} = -x_{1-\alpha/2}$ is the $\alpha/2$-quantile. Hence, also $(1-\alpha)\,100\%$ of the realizations $z$ of the estimator $Z$ fall within the interval (cf. Fig. 8.2, left):

$$-x_{1-\alpha/2} \leq \sqrt{N}\,\frac{z - \langle X \rangle}{\sigma_X} \leq x_{1-\alpha/2}. \tag{8.42}$$

As (8.42) is equivalent to

$$z - \frac{1}{\sqrt{N}}\,\sigma_X\, x_{1-\alpha/2} \leq \langle X \rangle \leq z + \frac{1}{\sqrt{N}}\,\sigma_X\, x_{1-\alpha/2}, \tag{8.43}$$

we may also say that for a given realization $z$ the quantity to be estimated, $\langle X \rangle$, lies inside the interval

$$\left[ z - \frac{1}{\sqrt{N}}\,\sigma_X\, x_{1-\alpha/2},\ z + \frac{1}{\sqrt{N}}\,\sigma_X\, x_{1-\alpha/2} \right] \tag{8.44}$$

with a probability of $(1-\alpha)\,100\%$ (Fig. 8.2, right). This is called the $(1-\alpha)\,100\%$ confidence interval for the estimator $\hat{\Theta}(X_1, \ldots, X_N) = Z$, defined by (8.2). An alternative notation for a confidence interval is $z \pm \frac{1}{\sqrt{N}} \sigma_X x_{1-\alpha/2}$.

Usually $\alpha$ is chosen such that $x_{1-\alpha/2} = 1, 2$, or 3, i.e., the result is stated in the form $z \pm \frac{1}{\sqrt{N}} k\, \sigma_X, k = 1, 2$, or 3. One speaks in this case of a one-, two-, or three-sigma error, and $(1 - \alpha)\, 100\%$ is $68\%, 95\%$, or $99\%$, respectively. For example, there is a 68% probability that the true value lies within the one-sigma confidence interval. In the representation of experimental data or estimated quantities the size of the confidence intervals is usually marked by an error bar of the appropriate size.

In general, such a statement holds for every unbiased estimator $\hat{\Theta}(X_1, \ldots, X_N)$ whose density is asymptotically normal, so that one can define a quantity $Y$ as in (8.41), namely

$$Y = \frac{\hat{\Theta} - \theta}{\sigma_{\hat{\Theta}}} \tag{8.45}$$

which, for large values of $N$, may also be regarded as normally distributed.

In many cases $\sigma^2_{\hat{\Theta}}$ is not known explicitly and an estimated value has to be inserted instead. If the size $N$ of the sample is sufficiently large (say, $N > 30$), this is a good approximation, as we shall see in a moment. For smaller samples one has to be more precise. Let us discuss this for $\theta = \langle X \rangle$, where $\sigma^2_{\hat{\Theta}} = \sigma^2_X / N$.

According to (8.10), the estimator for $\sigma^2_X$ is $\hat{\Theta}_1$. The quantity

$$Y' = \frac{(N-1)\, \hat{\Theta}_1}{\sigma^2_X}, \tag{8.46}$$

which we may define using this estimator, is $\chi^2$-distributed with $N - 1$ degrees of freedom, since it may also be written as the sum of squares of $N - 1$ independent standard normal random numbers

$$Z_i = \frac{1}{\sigma_X \sqrt{i(i+1)}} (X_1 + \ldots + X_i - i X_{i+1}), \quad i = 1, \ldots, N-1. \tag{8.47}$$

Hence, the quantity

$$T = \frac{Y}{\sqrt{Y'/(N-1)}} = \sqrt{N}\, \frac{Z - \langle X \rangle}{\sqrt{\hat{\Theta}_1}}, \tag{8.48}$$

which we obtain when we replace $\sigma_X$ in $Y$ (cf. (8.42)) by its estimation, is a quotient of the form (2.162) and thus a $t$-distributed random number with $N - 1$ degrees of freedom. For small sample sizes $N$, one now considers this quantity $T$ instead of the quantity $Y$ given by (8.41), and therefore in (8.42) and the subsequent equations one has to replace $x_{1-\alpha/2}$ by the $(1-\alpha/2)$-quantile $t_{1-\alpha/2}$ of a $t$-distributed random variable.

As already remarked, for large $N$, say, larger than about 30, this makes almost no difference, since the density of a $t$-distribution becomes closer and closer to a standard normal distribution for increasing $N$.

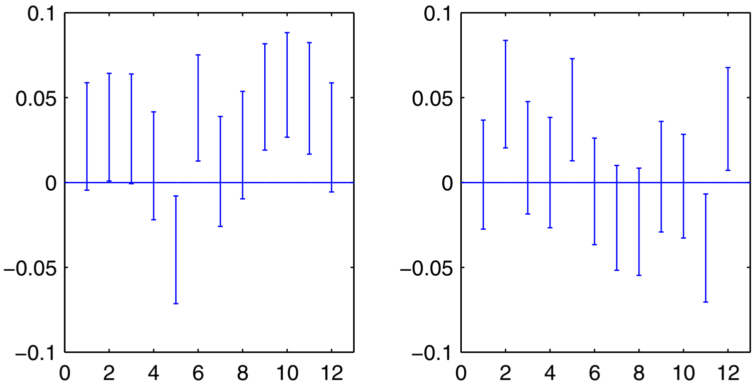


**Fig. 8.3** Two examples of 12 measurements of some quantity. In each case about one third of the error bars giving the one-sigma interval do not contain the true value, denoted by the *straight line*

*Remarks.*

- Because the true value lies within the one-sigma interval with 68% probability, the result of, say, 12 measurements of some quantity has to resemble those in Fig. 8.3: About one third of the error bars do not contain the true value, denoted by the straight line.
- Up to now we have considered only the confidence interval resulting from a statistical error. In many cases, however, a systematic error is also present. If its size is known, it may easily be taken into account in the computation of the expectation value. If it is not known, its influence on the expectation value and the confidence interval may also be determined using the concepts of probability theory: For this purpose one defines a suitable probability density based on knowledge about the systematic error. This density is not to be interpreted in the sense of relative frequency, but rather in the sense of a belief function. As long as the definition of such a function is also based on the Kolmogoroff axioms, we may include them in our computations in the same way as the densities with the relative frequency interpretation. The systematic error is then treated as a further random variable and the variances and confidence intervals may be calculated according to the laws of probability theory.

## 8.3 Propagation of Errors

In many cases, the results of measurements are used for subsequent calculations to compute other quantities. For instance, the values of certain parameters of a mathematical model often have to be deduced from experimental data, as we will discuss in Sect. 8.5. The uncertainty of the data has to show up again in the uncertainty of the quantities derived from them. If we consider the data as realizations of random variables, the quantities derived from them are also random variables. Both their expectation values and their variances may be determined.

We denote the outcomes of the measurements by $x_i, i = 1, \dots, n$ and the quantity to be derived from these values by $y$. Of course, we no longer assume that the random variables $X_i, i = 1, \dots, n$, which we associate to each measurement $x_i$, are copies of one specific random variable. The $\{X_i\}$ in general belong to different measurable quantities and usually differ in mean and variance. We assume, however, that the mean value, $\langle X_i \rangle$, is the true value of the quantity of the $i$th measurement, so that

$$X_i = \langle X_i \rangle + \eta_i, \tag{8.49}$$

where the random variable $\eta_i$ represents the experimental error.

In the context of the problem another quantity $y$ may be defined as a function of the measurable quantities. We have to formulate such a function as

$$y = F(\langle X_1 \rangle, \dots, \langle X_n \rangle), \tag{8.50}$$

because such a relation is usually derived from theoretical arguments, and thus only holds for the true values. We are looking for an estimator of $y$.

An obvious choice is

$$\hat{Y} = F(\hat{X}_1, \dots, \hat{X}_n), \tag{8.51}$$

where $\hat{X}_i$ is an unbiased estimator for $\langle X_i \rangle$, i.e., $\langle \hat{X}_i \rangle = \langle X_i \rangle$. However, only if the relation $F$ is linear may we conclude that

$$\langle \hat{Y} \rangle \equiv \langle F(\hat{X}_1, \dots, \hat{X}_n) \rangle = F(\langle \hat{X}_1 \rangle, \dots, \langle \hat{X}_n \rangle) \equiv F(\langle X_1 \rangle, \dots, \langle X_n \rangle), \tag{8.52}$$

i.e., that $\hat{Y}$ is an unbiased estimator for $y$.

In general $\langle \hat{Y} \rangle$ will differ from $F(\langle X_1 \rangle, \dots, \langle X_n \rangle)$. The bias may be computed as follows: We take

$$\hat{X}_i = \langle \hat{X}_i \rangle + E_i = \langle X_i \rangle + E_i, \tag{8.53}$$

with $\langle E_i \rangle = 0$ and $\operatorname{Cov}(E_i, E_j) = \sigma_{ij}^2$. We now obtain

$$\hat{Y} = F(\langle X_1 \rangle, \dots, \langle X_n \rangle) + \sum_{i=1}^{n} E_i \frac{\partial F}{\partial \langle X_i \rangle} + \frac{1}{2} \sum_{i,j=1}^{n} E_i E_j \frac{\partial^2 F}{\partial \langle X_i \rangle \partial \langle X_j \rangle} + \dots, \tag{8.54}$$

and therefore approximately

$$\langle \hat{Y} \rangle = F(\langle X_1 \rangle, \ldots, \langle X_n \rangle) + \frac{1}{2} \sum_{i,j=1}^{n} \sigma_{ij}^2 \frac{\partial^2 F}{\partial \langle X_i \rangle \partial \langle X_j \rangle}. \tag{8.55}$$

If $\hat{x}_i$ is a realization of $\hat{X}_i$, we take as an estimation of $y$ the realization of $\hat{Y}$, i.e., we set

$$\hat{y} = F(\hat{x}_1, \ldots, \hat{x}_n), \tag{8.56}$$

but we have to take into account a bias, which may be approximated by

$$\mathrm{Bias}(\hat{Y}) = \frac{1}{2} \sum_{i,j=1}^{n} \sigma_{ij}^2 \frac{\partial^2 \hat{y}}{\partial \hat{x}_i \partial \hat{x}_j}. \tag{8.57}$$

The uncertainty of this estimation, measured by the standard error $\sigma_{\hat{Y}}$ follows from the variance of $\hat{Y}$. Proceeding from (8.54), again neglecting the terms $\langle E_i E_j E_k \rangle$ as well as the terms of higher order, and replacing the expectation values $\langle X_i \rangle$ by the measured values $\hat{x}_i$, we finally obtain

$$\sigma_{\hat{Y}}^2 = \sum_{i=1}^{n} \sigma_{ii}^2 \left( \frac{\partial \hat{y}}{\partial \hat{x}_i} \right)^2 + \sum_{i,j=1 i \neq j}^{n} \sigma_{ij}^2 \left( \frac{\partial \hat{y}}{\partial \hat{x}_i} \right) \left( \frac{\partial \hat{y}}{\partial \hat{x}_j} \right). \tag{8.58}$$

### *8.3.1 Application*

We want to determine the volume $V$ of a rectangular parallelepiped by measuring the length of the edges. Suppose we have found the values $\hat{a}, \hat{b}, \hat{c}$ with measurement errors $\sigma_a, \sigma_b, \sigma_c$, then we obtain for the volume

$$\hat{v} = \hat{a}\hat{b}\hat{c}, \tag{8.59}$$

and for uncorrelated errors of measurement we find for the variance $\sigma_v^2$, i.e. for the square of the measurement error of $V$:

$$\sigma_v^2 = \sigma_a^2 \big(\hat{b}\hat{c}\big)^2 + \sigma_b^2 \big(\hat{a}\hat{c}\big)^2 + \sigma_c^2 \big(\hat{a}\hat{b}\big)^2. \tag{8.60}$$

Therefore,

$$\frac{\sigma_v^2}{\hat{v}^2} = \frac{\sigma_a^2}{\hat{a}^2} + \frac{\sigma_b^2}{\hat{b}^2} + \frac{\sigma_c^2}{\hat{c}^2}. \tag{8.61}$$

In this case the bias vanishes. This would not be true for correlated errors.

## 8.4 The Maximum Likelihood Estimator

Up to now we have discussed estimators for the expectation value or the variance of a random variable. Sometimes we also want to estimate other parameters of a density. Let $\varrho(x \mid \theta)$ be a family of densities parametrized by $\theta$, where $\theta$ represents one or more parameters. The question that we now want to address is how to formulate an estimator for $\theta$, given an independent sample $\{x_1, \ldots, x_N\}$.

The probability that a realization yields a value in the interval $[x, x + \mathrm{d}x]$ is, for some given value of $\theta$, just $\varrho(x \mid \theta)\,\mathrm{d}x$ and the probability for an independent sample $\{x_1, \ldots, x_N\}$ is then determined by

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \varrho(x_i \mid \theta). \tag{8.62}$$

$\mathcal{L}(\theta)$ is also called the likelihood or likelihood function of the sample.

For a sample with realizations of dependent random variables the likelihood can also be stated if the joint density $\varrho(x_1 \ldots, x_N \mid \theta)$ is known. In this case we simply have $\mathcal{L}(\theta) = \varrho(x_1, \ldots, x_N \mid \theta)$.

The likelihood is thus determined from the probability density: One inserts a realization and considers the resulting quantity as a function of the parameter. The expression ‘likelihood’ refers to the fact that this function originates from a probability density.

The maximum likelihood estimate (ML estimate) $\hat{\theta}_{ML}$ for the parameter $\theta$ is defined as the value of $\theta$ which maximizes the likelihood. The probability for the sample under consideration is maximum for this value.

Often one considers the log likelihood function

$$L(\theta) = \ln\big(\mathcal{L}(\theta)\big) = \sum_{i=1}^{N} \ln\big(\varrho(x_i \mid \theta)\big), \tag{8.63}$$

and maximizes this function with respect to $\theta$. This yields the same value for $\theta$, since the logarithm is a monotonic function.

The ML estimate $\hat{\theta}_{\mathrm{ML}}$ depends on the sample and may be regarded as the realization of an estimator $\hat{\Theta}_{\mathrm{ML}}(X_1, \ldots, X_N)$. Hence, $\hat{\theta}_{\mathrm{ML}}$ is the realization of a random variable and in order to find the properties of the estimator we have to study its distribution.

With help of Bayes' theorem and under the assumption that the a-priori probability distribution for the parameter $\theta$ is given by a constant density, say $\varrho_\Theta(\theta) = \text{const.}$, the likelihood can be considered to be proportional to the a posteriori probability distribution $\varrho(\theta \mid x_1, \ldots, x_N)$, because

$$\varrho(\theta \mid x_1, \ldots, x_N) = \frac{\varrho(x_1, \ldots, x_N \mid \theta)\varrho_\Theta(\theta)}{\varrho(x_1, \ldots, x_N)} \propto \mathcal{L}(\theta)\,. \tag{8.64}$$

The maximum likelihood estimate $\hat{\theta}_{\mathrm{ML}}$ is then just the maximum of the probability distribution for $\theta$, given the data $x_1, \ldots, x_N$. Of course, this definition does work only, if there exists exactly one maximum. If, furthermore, this a posteriori distribution for $\theta$ near its maximum can be approximated by a normal distribution, then the variance of the likelihood estimator $\hat{\Theta}_{\mathrm{ML}}$ can be shown to be given by the negative inverse of the second derivative of $L(\theta)$ at $\hat{\theta}_{\mathrm{ML}}$:

$$\sigma^2_{\hat{\Theta}_{\mathrm{ML}}} = -\frac{1}{\partial^2 L(\theta)/\partial\theta^2}|_{\hat{\theta}_{\mathrm{ML}}} \,. \tag{8.65}$$

The use of the ML estimator is quite common and in many cases yields good results for the estimation. However, there may also be problems connected with its application:

- There may exist several maxima, or even none, as already considered as possible.
- The ML estimator is not always unbiased. For example, the ML estimator for the variance turns out to be the biased estimator $\hat{\Theta}_2$ that we met in Sect. 8.1.

However, in general the bias is of order $1/N$, where $N$ is the size of the sample. Furthermore, if there is only one maximum, this estimator can be shown to be consistent, i.e., the estimation error becomes smaller with increasing sample size. For a detailed discussion of the properties of the ML estimator see Lehmann (1991).

*Examples.*

- We consider the random variable $K$ from example (c) in Sect. 2.1.3, i.e., we are given a random variable $X$ with possible realizations $x_1$ and $x_2$ and the respective probabilities $p$ and $(1-p)$. Then the density

$$B(n, p; k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{8.66}$$

is equal to the probability of obtaining in $n$ realizations of $X$ the event $x_1$ exactly $k$ times. Suppose $p$ is unknown, and an experiment with $n$ realizations yields $k$ times the result $x_1$. Intuitively we would expect for $p$:

$$\hat{p} = \frac{k}{n}. \tag{8.67}$$

The same result also follows for the ML estimator. Take, for example, $n = n_1$ and a single realization $k_1$ for $K$, then

$$L(p) = \ln B(n_1, p; k_1) = \ln \binom{n_1}{k_1} + k_1 \ln p + (n_1 - k_1) \ln(1-p), \tag{8.68}$$

and from

$$\frac{\partial}{\partial p}\,\ln B(n_1,p;k_1)=\frac{k_1}{p}-\frac{(n_1-k_1)}{(1-p)}=0 \tag{8.69}$$

we obtain

$$k_1(1-p)=(n_1-k_1)p \qquad \text{or} \qquad p=\frac{k_1}{n_1}. \tag{8.70}$$

- Suppose we measure the lifetime of certain objects (elementary particles, lightbulbs, etc.) For the density function we assume the form

$$\varrho(t \mid \tau)=\frac{1}{\tau}\,\mathrm{e}^{-t/\tau}, \quad t>0, \tag{8.71}$$

i.e., $\varrho(t \mid \tau)\,\mathrm{d}t$ is the probability that we observe a lifetime in the interval $[t,t+\mathrm{d}t]$. The parameter $\tau$ is also the expectation value of the lifetime.

From a series of data $t_1,\ldots,t_N$ for the lifetimes we obtain the likelihood function

$$\mathcal{L}(\tau)=\prod_{i=1}^{N}\frac{1}{\tau}\,\mathrm{e}^{-t_i/\tau} \tag{8.72}$$

or

$$L(\tau)=\ln\bigl(\mathcal{L}(\tau)\bigr)=\sum_{i=1}^{N}\bigl(-\ln(\tau)-t_i/\tau\bigr). \tag{8.73}$$

The maximum $\hat{\tau}$ of $L(\tau)$ is determined from

$$\frac{\partial\,L(\tau)}{\partial\tau}=\sum_{i=1}^{N}\left(-\frac{1}{\tau}+\frac{t_i}{\tau^2}\right)=0, \tag{8.74}$$

from which we obtain

$$\hat{\tau}=\frac{1}{N}\sum_{i=1}^{N}t_i\,. \tag{8.75}$$

For this value of $\tau$ we also have

$$\left.\frac{\partial^2 L(\tau)}{\partial\tau^2}\right|_{\hat{\tau}}=\left.\left(N\frac{1}{\tau^2}-2\frac{1}{\tau^3}\sum_{i=1}^{N}t_i\right)\right|_{\hat{\tau}}=\left.\frac{N}{\tau^3}\left(\tau-\frac{2}{N}\sum_{i=1}^{N}t_i\right)\right|_{\hat{\tau}}=-\left.\frac{N}{\tau^2}\right|_{\hat{\tau}}, \tag{8.76}$$

which is negative as required for a maximum. The quantity $\hat{\tau}$ in (8.75) therefore represents a maximum likelihood estimate for the expectation value of the lifetime of the objects. It is identical to the usual estimate, the mean value. Furthermore the variance of the estimator is $\hat{\tau}^2/N$, i.e., as usual of the order of $1/N$.

- Suppose we measure some quantity with different experimental devices; then we will in general find different results. If the deviation $\sigma_i$ of the experimental error is known for each device and if this error in each case is normally distributed, then the density for the distribution of the data around the true mean value $\lambda$ for the $i$th device is given by

$$\varrho_i(\lambda, \sigma_i^2; x) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2\sigma_i^2}(x-\lambda)^2\right). \tag{8.77}$$

The likelihood function for the experimental data $x_i$ obtained with the experimental devices $i = 1, \ldots, N$ then reads:

$$\mathcal{L}(\lambda) = \prod_{i=1}^{N} \varrho_i(\lambda, \sigma_i^2; x_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2\sigma_i^2}(x_i-\lambda)^2\right). \tag{8.78}$$

It is maximum for

$$\lambda = \hat{\lambda} = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right)^{-1}, \tag{8.79}$$

i.e., for the weighted average of the data. The larger the deviation of the data for one device, the less the value obtained with this device enters in the final result $\hat{\lambda}$ for the quantity to be determined.

The variance of $\lambda$ may be now computed also according to the laws of error propagation (cf. Sect. 8.3). One obtains

$$\sigma_\lambda^2 = \left(\sum_{j=1}^{N} \frac{1}{\sigma_j^2}\right)^{-2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} = \left(\sum_{j=1}^{N} \frac{1}{\sigma_j^2}\right)^{-1}. \tag{8.80}$$

As an illustration, suppose that the value of some fundamental constant $\alpha$ has been determined independently by various experimental methods. After all imaginable systematic errors have been taken into account, the six results shown together with their statistical errors in Fig. 8.4(left) have been obtained. The ‘world average’ may then be determined from (8.79) and (8.80), and it is represented in Fig. 8.4 by the error bar on the right.

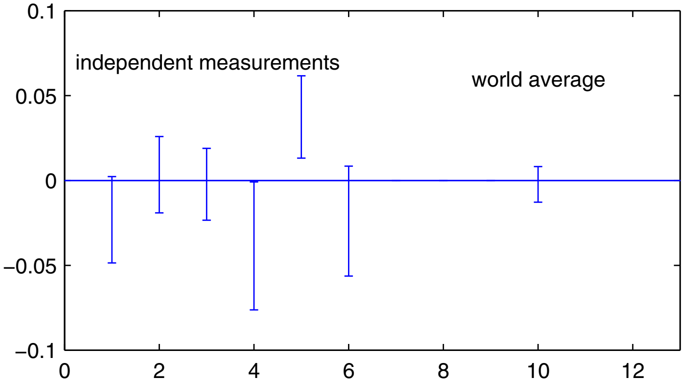


**Fig. 8.4** Simulation of outcomes of different independent experiments and the resulting average, denoted as 'world average'. The corresponding errors are given as one-sigma intervals

- In some cases the maximum likelihood estimator may be useless, as can be seen from the following example: We attempt to estimate the parameter $\mu$ of the Cauchy density (cf. (2.70), we take $\gamma = 1$)

$$\varrho(x \mid \mu) = \frac{1}{\pi} \frac{1}{(x-\mu)^2 + 1} \tag{8.81}$$

on the basis of a sample $\{x_1, \ldots, x_N\}$ using the maximum likelihood estimator.

Maximizing the log likelihood

$$L(\mu) = -\sum_{i=1}^{N} \ln\left((x_i - \mu)^2 + 1\right) - N \ln \pi \tag{8.82}$$

leads to the equation

$$f(\mu) \equiv \sum_{i=1}^{N} \frac{\mu - x_i}{(x_i - \mu)^2 + 1} = 0. \tag{8.83}$$

If we can assume that $(x_i - \mu)^2 \ll 1$ is satisfied for all $i$, we obtain as an approximate solution

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i . \tag{8.84}$$

This approximation is obviously valid when all $x_i$ are close to each other and therefore also close to $\mu$. For the case $N = 2$ we can easily see what happens if the observed values $x_i$ are farther apart. For $|x_1 - x_2| < 2$ the function $f(\mu)$ has only one zero close to $(x_1 + x_2)/2$, for $|x_1 - x_2| > 2$, however, there are three zeros and $L(\mu)$ has two maxima (Fig. 8.5).

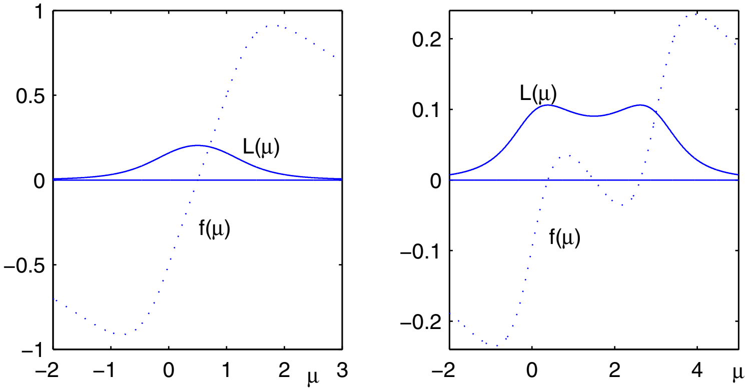


**Fig. 8.5** The function $f(\mu)$ and the log likelihood $L(\mu)$ for $|x_1-x_2|<2$ (*left*) and $|x_1-x_2|>2$ (*right*). In the second case the maximum likelihood estimator is useless, because there is more than one maximum

## 8.5 The Least-Squares Estimator

A situation that frequently occurs in physics is the following:

In the course of an experiment one parameter $x$ is changed systematically, i.e., the experiment is performed at different settings $x_i,\ i=1,\dots,N$ for this parameter, and for the value $x_i$ the result of the measurement is $y_i$.

The observed data $\{y_1,\dots,y_N\}$ can be regarded as realizations of random variables $\{Y_1,\dots,Y_N\}$ which, however, do not have the same distributions, in contrast to the case for a random sample. The mean and the variance of $Y_i$ will be functions $f$ and $\sigma^2$ of $x_i$:

$$\langle Y_i\rangle = f(x_i) \tag{8.85a}$$

$$\mathrm{Var}(Y_i) = \sigma^2(x_i) \equiv \sigma_i^2. \tag{8.85b}$$

The random variables $\{Y_i\}$ may even depend on each other such that in general the covariances

$$\mathrm{Cov}(Y_i,Y_j) = \mathsf{C}_{ij} \tag{8.86}$$

are nonzero for $i \neq j$.

In the following we write

$$Y_i = f(x_i) + E_i \tag{8.87}$$

for such a case, where $E_i$ are random variables whose realizations correspond to the errors of measurement, and $f(x_i)$ represents the mean value of $y_i$ for the parameter value $x_i$. For $E_i$ we therefore assume $\langle E_i \rangle = 0$. Furthermore, $\text{Var}(E_i) = \sigma_i^2$, or more generally

$$\text{Cov}(E_i, E_j) = \mathsf{C}_{ij}. \tag{8.88}$$

The particular context often suggests the use of certain parametric functions $f(x)$ and $\sigma(x)$ as models for the dependence on the parameter $x$. A very simple model would be:

$$f(x) = f(x \mid a, b) = ax + b, \quad \sigma^2(x) = \sigma_0^2, \; \mathsf{C}_{ij} = 0 \quad \text{for} \quad i \neq j, \tag{8.89}$$

where $a$, $b$, and $\sigma_0$ are parameters still to be determined. This model implies that the observed data depend linearly on the values $x$ apart from the errors of measurement, which themselves are assumed to be independent of the setting $x$. In this case one speaks of a constant absolute error of measurement.

A more general parametric model for $f(x)$ is

$$f(x \mid a_1, \ldots, a_M) = \sum_{\alpha=1}^{M} a_\alpha X_\alpha(x), \tag{8.90}$$

where $\{X_\alpha(x)\}$ are given functions which may be nonlinear. The model is linear, however, with respect to the parameters $\{a_\alpha\}$.

A more general model for the errors of measurement is

$$\sigma_i = \sigma(x_i) = \sigma_0 f(x_i) = \sigma_0 \langle Y_i \rangle, \; \mathsf{C}_{ij} = 0 \quad \text{for} \quad i \neq j \tag{8.91}$$

i.e., the error of measurement is proportional to the measured value. In this case one speaks of a constant relative error of measurement.

Equation 8.87 together with some parametrization of the functions $f(x)$ and $\sigma(x)$ such as (8.90) and (8.91), respectively, constitutes a model for the dependence of the measured quantities on $x$. In order to specify this model completely, one has to determine the parameters. This can be done by estimating them on the basis of the data $\{y_1, \ldots, y_n\}$.

In the more general case one may want to estimate the matrix elements $\{\mathsf{C}_{ij}\}$ and the parameters appearing in an ansatz such as (8.90) for $f(x)$. Here, however, we will only be concerned with an estimator for the parameters $a_1, \ldots, a_M$ of (8.90) assuming that $\mathsf{C}_{ij}$ is known. Of course, we also assume that the number $N$ of data is larger than the number $M$ of parameters to be estimated.

The most appropriate estimator for the parameters $\{a_\alpha\}$ is the least-squares estimator. If we combine the $\{a_\alpha\}$ into a vector $\boldsymbol{a} = (a_1, \ldots, a_M)$, this estimate $\hat{\boldsymbol{a}}$ is defined as the set of values for the parameters for which the quadratic form

$$q^2(\boldsymbol{a}) = \sum_{i,j=1}^{N} \big(y_i - f(x_i \mid \boldsymbol{a})\big)\big(\mathsf{C}^{-1}\big)_{ij}\big(y_j - f(x_j \mid \boldsymbol{a})\big) \tag{8.92}$$

is minimum. If the errors of measurement are independent, the weighted sum of the squared deviations reduces to

$$q^2(\boldsymbol{a}) = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(y_i - f(x_i \,|\, \boldsymbol{a})\right)^2 . \tag{8.93}$$

*Remark.* If the errors of measurement $\{e_i\}$ are realizations of Gaussian random variables $\{E_i\}$, then the random variables $Y_i$ are also Gaussian with

$$\langle Y_i \rangle = f(x_i \,|\, \boldsymbol{a}) \tag{8.94}$$

and

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(E_i, E_j) = \mathsf{C}_{ij} , \tag{8.95}$$

and for the joint density we obtain from (2.24)

$$\begin{aligned} \varrho(y_1, \ldots, y_N \,|\, \boldsymbol{a}) = {} & \frac{1}{\sqrt{(2\pi)^N \, \det \mathsf{C}}} \\ & \times \exp\left[ -\frac{1}{2} \sum_{i,j} \left(y_i - f(x_i \,|\, \boldsymbol{a})\right) \left(\mathsf{C}^{-1}\right)_{ij} \left(y_j - f(x_j \,|\, \boldsymbol{a})\right) \right] . \end{aligned} \tag{8.96}$$

The likelihood function of the sample results if we replace the variables in the density by the special values of the sample (which shall be denoted here by the same symbols). So we get

$$\begin{aligned} L(\boldsymbol{a}) &= \ln \varrho(y_1, \ldots, y_N \,|\, \boldsymbol{a}) \\ &= -\frac{1}{2} \sum_{i,j} \left(y_i - f(x_i \,|\, \boldsymbol{a})\right) \left(\mathsf{C}^{-1}\right)_{ij} \left(y_j - f(x_j \,|\, \boldsymbol{a})\right) \\ &\quad - \frac{1}{2} \ln \det(\mathsf{C}) - \frac{N}{2} \ln(2\pi) . \end{aligned} \tag{8.97a}$$

For given $\mathsf{C}$ the maximization of $L$ is equivalent to the minimization of the quadratic form in (8.97), which is now identical to the quantity $q^2(\boldsymbol{a})$ in (8.92). In this case the least-squares estimator is equal to the maximum likelihood estimator.

We now will turn to the solution of the optimization problem in order to find an explicit expression $\{\hat{A}_\alpha\}$ for the least-squares estimator. For simplicity we shall assume that the random variables $\{Y_i\}$ are mutually independent.

We minimize $q^2(\boldsymbol{a})$ as given in (8.93) for given $\{\sigma_i\}$ and for a given model (8.90). Since the parameters to be determined appear only linearly in these models, $q^2(\boldsymbol{a})$ is a quadratic form:

$$q^2(\boldsymbol{a}) \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left( y_i - \sum_{\alpha=1}^{M} a_\alpha X_\alpha(x_i) \right)^2 \tag{8.98a}$$

$$= \sum_{i=1}^{N} \left( b_i - \sum_{\alpha=1}^{M} K_{i\alpha} a_\alpha \right)^2 , \tag{8.98b}$$

with

$$K_{i\alpha} = \frac{1}{\sigma_i} X_\alpha(x_i), \quad b_i = \frac{y_i}{\sigma_i} \quad i = 1, \ldots, N, \quad \alpha = 1, \ldots, M . \tag{8.99}$$

The minimum of this quadratic form is easily determined and one is led to a linear system of equations for $\{a_\alpha\}$. Alternatively we may also introduce the $N \times M$ matrix $\mathsf{K}$ with components $\mathsf{K}_{i\alpha}$, regarded as a mapping from the $M$-dimensional space of parameters into the $N$-dimensional space of the realizations of $\{Y_i\}$, $(N > M)$, and reformulate the minimization problem of $q^2(\boldsymbol{a})$ in the following form: Find a vector $\boldsymbol{a}$ such that the norm

$$q^2(\boldsymbol{a}) = \|\boldsymbol{b} - \mathsf{K}\boldsymbol{a}\|^2 \tag{8.100}$$

is minimum. Here, $\boldsymbol{b} = (b_1, \ldots, b_N)$ is the $N$-dimensional vector with components $b_i = y_i/\sigma_i$ and the norm is the standard quadratic norm in $\mathbb{R}^N$.

The matrix $\mathsf{K}$ maps the $M$-dimensional space of parameters, $\mathbb{R}^M$, onto an $M$-dimensional subspace of $\mathbb{R}^N$. Although the data $\boldsymbol{b}$ are in $\mathbb{R}^N$, they are not necessarily in the image of $\mathbb{R}^M$. The difference $\boldsymbol{e} = \boldsymbol{b} - \mathsf{K}\boldsymbol{a}$ assumes its minimum norm for just that vector $\boldsymbol{a}$ for which $\mathsf{K}\boldsymbol{a}$ equals the projection of $\boldsymbol{b}$ onto the image of $\mathbb{R}^M$ in $\mathbb{R}^N$ (see Fig. 8.6).

From numerical mathematics (see, e.g., Press et al. 2007) it is known that this projection operator is given by the pseudo-inverse

$$\mathsf{K}^+ = \left(\mathsf{K}^{\mathrm{T}}\mathsf{K}\right)^{-1}\mathsf{K}^{\mathrm{T}}. \tag{8.101}$$

From the singular value decomposition of the $N \times M$ matrix $\mathsf{K}$ one can find an orthonormal system of basis vectors $\{\boldsymbol{v}_\gamma,\ \gamma = 1, \ldots, M\}$ in $\mathbb{R}^M$ and a corresponding orthonormal system $\{\boldsymbol{u}_\gamma,\ \gamma = 1, \ldots, M\}$ in the image of $\mathbb{R}^M$ in $\mathbb{R}^N$ such that

$$\mathsf{K} = \sum_{\gamma=1}^{M} w_\gamma\, \boldsymbol{u}_\gamma \otimes \boldsymbol{v}_\gamma . \tag{8.102}$$

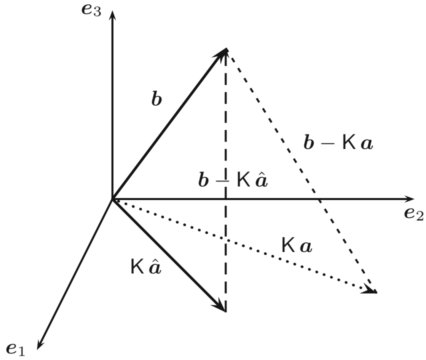


**Fig. 8.6** Let the image space of $\mathsf{K}$ be spanned by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. The difference vector $\boldsymbol{b} - \mathsf{K}\boldsymbol{a}$ has minimum norm if $\mathsf{K}\boldsymbol{a}$ is equal to the projection of $\boldsymbol{b}$ onto the image space, i.e., if $\boldsymbol{a} = \hat{\boldsymbol{a}}$

Here, the tensor product $\boldsymbol{u}_\gamma \otimes \boldsymbol{v}_\gamma$ is a matrix with

$$\boldsymbol{u}_\gamma \otimes \boldsymbol{v}_\gamma \cdot \boldsymbol{a} = \boldsymbol{u}_\gamma(\boldsymbol{v}_\gamma \cdot \boldsymbol{a})$$

and

$$\boldsymbol{a} \cdot \boldsymbol{u}_\gamma \otimes \boldsymbol{v}_\gamma = (\boldsymbol{a} \cdot \boldsymbol{u}_\gamma)\boldsymbol{v}_\gamma .$$

(Sometimes the tensor product is also written in the form $\boldsymbol{u}_\gamma\, \boldsymbol{v}_\gamma^{\mathrm{T}}$ instead of $\boldsymbol{u}_\gamma \otimes \boldsymbol{v}_\gamma$.) The diagonal elements $\{w_\gamma,\ \gamma = 1, \ldots, M\}$ are called singular values, the quantities $\{w_\gamma^2,\ \gamma = 1, \ldots, M\}$ are also the eigenvalues of the $M \times M$ matrix

$$\mathsf{K}^{\mathrm{T}}\mathsf{K} \equiv \mathsf{D} = \sum_{\gamma=1}^{M} w_\gamma^2\, \boldsymbol{v}_\gamma \otimes \boldsymbol{v}_\gamma . \tag{8.103}$$

For $\mathsf{K}^+$ we now obtain

$$\mathsf{K}^+ = \sum_{\gamma=1}^{M} \frac{1}{w_\gamma} \boldsymbol{v}_\gamma \otimes \boldsymbol{u}_\gamma \tag{8.104}$$

and also have

$$\mathsf{K}^+\mathsf{K} = \mathsf{I} \quad \text{in } \mathbb{R}^M \tag{8.105}$$

and

$$\mathsf{K}\mathsf{K}^+ = \sum_{\gamma=1}^{M} \boldsymbol{u}_\gamma \otimes \boldsymbol{u}_\gamma = \mathsf{P}, \tag{8.106}$$

where $\mathsf{P}$ is the projection operator in $\mathbb{R}^N$ onto the image of $\mathbb{R}^M$.

The minimum of $q^2(\boldsymbol{a})$ is therefore assumed for

$$\hat{\boldsymbol{a}} = \mathsf{K}^+\boldsymbol{b} = \sum_{\gamma=1}^{M} \frac{1}{w_\gamma} \boldsymbol{v}_\gamma \left(\boldsymbol{u}_\gamma \cdot \boldsymbol{b}\right) \qquad \text{with} \qquad b_i = \frac{y_i}{\sigma_i}. \tag{8.107}$$

Hence, the explicit form of the least-squares estimator (LS estimator) is

$$\hat{A}_\alpha = \sum_{i=1}^{N} \left(\mathsf{K}^+\right)_{\alpha i} \frac{Y_i}{\sigma_i}, \quad \alpha = 1, \ldots, M. \tag{8.108}$$

In the remainder of this section we detail some relevant properties of least-squares estimators.

**Bias** From

$$\langle Y_i \rangle = f(x_i \,|\, \boldsymbol{a}) \equiv \sigma_i \sum_{\beta=1}^{M} K_{i\beta} a_\beta \tag{8.109}$$

and (8.105) it follows that

$$\langle \hat{A}_\alpha \rangle = \sum_{i=1}^{N} \left(\mathsf{K}^+\right)_{\alpha i} \frac{\langle Y_i \rangle}{\sigma_i} = a_\alpha, \quad \alpha = 1, \ldots, M. \tag{8.110}$$

The LS estimator is thus unbiased.

**Covariance** With $\mathrm{Cov}(Y_i, Y_j) = \sigma_i^2\, \delta_{ij}$ we find for the covariances of the parameter estimators

$$\begin{aligned} \mathrm{Cov}(\hat{A}_\alpha, \hat{A}_\beta) &= \sum_{i=1}^{N} \left(\mathsf{K}^+\right)_{\alpha i} \left(\mathsf{K}^+\right)_{\beta i} = \left(\mathsf{K}^+ \left(\mathsf{K}^+\right)^{\mathrm{T}}\right)_{\alpha\beta} \\ &= \sum \frac{1}{w_\gamma^2} \left(\boldsymbol{v}_\gamma \otimes \boldsymbol{v}_\gamma\right)_{\alpha\beta} \equiv \left(\mathsf{D}^{-1}\right)_{\alpha\beta}. \end{aligned} \tag{8.111}$$

From this result we draw the following two conclusions:

Firstly, since

$$(\mathsf{D})_{\alpha\beta} = (\mathsf{K}^{\mathrm{T}}\mathsf{K})_{\alpha\beta} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} X_\alpha(x_i)\, X_\beta(x_i), \tag{8.112}$$

with $\sigma_i \equiv \sigma_0$ or $\sigma_i \equiv \sigma_0 \langle Y_i \rangle$, the quantity $\mathrm{Cov}(\hat{A}_\alpha, \hat{A}_\beta) = (\mathsf{D}^{-1})_{\alpha\beta}$ is also proportional to $\sigma_0^2$, i.e., the error of estimation is proportional to the error of measurement.

(We cannot expect that the errors of estimation tend to zero for $N \to \infty$, because the realizations of $Y_i$ were obtained for different $x_i$.)

Secondly, if $\{Y_i, i = 1, \ldots, N\}$ are Gaussian random variables, so also are the LS estimators $\{\hat{A}_\alpha\}$. The expectation values and covariance matrix are given by (8.110) and (8.111). The joint density function for the realizations of $\{\hat{A}_\alpha\}$ is therefore the $M$-dimensional Gaussian density

$$\varrho_{\hat{A}_1,\ldots,\hat{A}_M}(\boldsymbol{a}) = \frac{(2\pi)^{-M/2}}{(\det \mathsf{D}^{-1})^{1/2}} \exp\left[-\frac{1}{2}\sum_{\alpha,\beta=1}^{M}(\boldsymbol{a} - \langle\hat{\boldsymbol{A}}\rangle)_\alpha \mathsf{D}_{\alpha\beta}(\boldsymbol{a} - \langle\hat{\boldsymbol{A}}\rangle)_\beta\right].$$

The hypersurfaces in $\boldsymbol{a}$-space for a given value of $\varrho(\boldsymbol{a})$ are ellipsoids centered around $\langle\hat{\boldsymbol{A}}\rangle$. In particular, the boundary of the region inside which $100p\%$ of all vectors $\boldsymbol{a}$ are to be expected is described by the equation

$$\Delta = \sum_{\alpha,\beta=1}^{M} \delta a_\alpha \mathsf{D}_{\alpha\beta}\delta a_\beta = \sum_{\gamma=1}^{M} w_\gamma^2(\boldsymbol{v}_\gamma \cdot \delta\boldsymbol{a})^2 = \sum_{\gamma=1}^{M}\frac{(\boldsymbol{v}_\gamma \cdot \delta\boldsymbol{a})^2}{\frac{1}{w_\gamma^2}}, \tag{8.113}$$

with $\delta\boldsymbol{a} = \boldsymbol{a} - \langle\hat{\boldsymbol{A}}\rangle$, where the constant $\Delta$ still has to be determined for a given $p$. The confidence region for the confidence level $100p\%$ around $\hat{\boldsymbol{a}}$ will be an equally large region, whose boundary is also described by (8.113), where now, however, $\delta\boldsymbol{a} = \boldsymbol{a} - \hat{\boldsymbol{a}}$. This confidence region therefore corresponds to the interior of an ellipsoid around $\hat{\boldsymbol{a}}$, whose principal axes are directed towards $\boldsymbol{v}_\gamma$ and have the length $\sqrt{\Delta}/w_\gamma$ (Fig. 8.7).

How do we determine $\Delta \equiv \Delta(p)$? By definition, the interior of the ellipsoid $E(\Delta)$, whose boundary is characterized by (8.113), has the volume $p$, i.e., (with $x_\alpha = \delta a_\alpha$)

$$\int_{E(\Delta)} \frac{(2\pi)^{-M/2}}{\left[\det(\mathsf{D}^{-1})\right]^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\sum_{\alpha,\beta=1}^{M} x_\alpha \mathsf{D}_{\alpha\beta} x_\beta\right) \mathrm{d}^M x = p. \tag{8.114}$$

A transformation to the principal axes $\boldsymbol{v}_\gamma$ and a rescaling by the factor $w_\gamma$ yields for the left hand side of (8.114):

$$\begin{aligned} &\Omega_M (2\pi)^{-M/2}\int_0^{\sqrt{\Delta}} r^{M-1}\,\mathrm{e}^{-r^2/2}\,\mathrm{d}r \\ &\qquad = \frac{2\pi^{M/2}}{\Gamma(M/2)}(2\pi)^{-M/2}\int_0^{\sqrt{\Delta}} r^{M-1}\,\mathrm{e}^{-r^2/2}\,\mathrm{d}r && (8.115)\\ &\qquad = P(M/2, \Delta/2). && (8.116)\end{aligned}$$

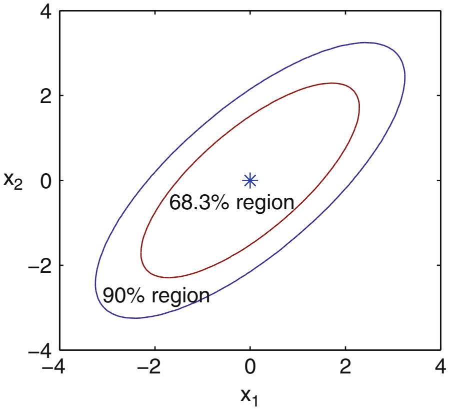


**Fig. 8.7** Confidence regions in two-dimensional space. The true value lies with 68.3% probability inside the 68.3% region and with 90% inside the 90% region

**Table 8.1** $\Delta(p)$ for various values of $100p\%$ and $M$

| $M$ $100p\%$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 68.3 | 1.00 | 2.30 | 3.53 | 4.72 | 5.89 | 7.04 |
| 90 | 2.71 | 4.61 | 6.25 | 7.78 | 9.24 | 10.6 |
| 95.4 | 4.00 | 6.17 | 8.02 | 9.70 | 11.3 | 12.8 |
| 99 | 6.63 | 9.21 | 11.3 | 13.3 | 15.1 | 16.8 |
| 99.73 | 9.00 | 11.8 | 14.2 | 16.3 | 18.2 | 20.1 |
| 99.99 | 15.1 | 18.4 | 21.1 | 23.5 | 25.7 | 27.8 |

Here, $\Omega_M = 2\pi^{M/2}/\Gamma(M/2)$ is the surface of a sphere in an $M$-dimensional space and the incomplete gamma function $P(a,x)$ is defined (Press et al. 2007) by

$$P(a,x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1}\,\mathrm{e}^{-t}\,\mathrm{d}t. \tag{8.117}$$

Hence, $\Delta(p)$ is determined by the equation

$$P(M/2, \Delta/2) = p. \tag{8.118}$$

Table 8.1 lists the values $\Delta$ for selected values of $M$ and $p$.

On the other hand, the standard deviation $s_\alpha$ to be derived from

$$s_\alpha^2 \equiv \mathrm{Cov}(\hat{A}_\alpha, \hat{A}_\alpha) = \mathrm{Var}(\hat{A}_\alpha) \tag{8.119}$$

is the confidence interval for $\hat{a}_\alpha$, *irrespective* of the results for the other $\hat{a}_\beta$, $\beta \neq \alpha$, because, e.g.,

$$\begin{aligned} \mathrm{Var}(\hat{A}_1) \equiv \langle(\hat{A}_1 - \langle\hat{A}_1\rangle)^2\rangle &= \int \mathrm{d}^M a\,(a_1 - \langle\hat{A}_1\rangle)^2 \varrho_{\hat{A}_1,\ldots,\hat{A}_M}(a_1,\ldots,a_M) \\ &= \int \mathrm{d}a_1\,(a_1 - \langle\hat{A}_1\rangle)^2 \varrho_{\hat{A}_1}(a_1)\,, \end{aligned} \tag{8.120}$$

with

$$\varrho_{\hat{A}_1}(a_1) = \int \mathrm{d}a_2 \ldots \mathrm{d}a_M \varrho_{\hat{A}_1,\ldots,\hat{A}_M}(a_1, \ldots, a_M), \tag{8.121}$$

which is the density distribution for $\hat{A}_1$ *irrespective* of the realizations for the other $\hat{A}_\beta$, $\beta \neq \alpha$,

If $\{y_i\}$ are normally distributed, then so are $\{\hat{a}_\alpha\}$, and with 68% probability the true value lies within the interval

$$[\hat{a}_\alpha - s_\alpha, \hat{a}_\alpha + s_\alpha]. \tag{8.122}$$

Similarly, one may define confidence intervals for other confidence levels.

**Distribution of $q^2(\hat{a})$** The weighted sum of the squared deviations,

$$q^2(\hat{\boldsymbol{a}}) = \boldsymbol{e}^2 \quad \text{with} \quad \boldsymbol{e} = \boldsymbol{b} - \mathsf{K}\hat{\boldsymbol{a}} \tag{8.123}$$

is the realization of a $\chi^2$-distributed random variable with $N - M$ degrees of freedom if $\{Y_i, i = 1, \ldots, N\}$ are independent Gaussian random variables.

This can be shown as follows: The singular value decomposition supplies us with a basis $\{\boldsymbol{u}_\gamma,\ \gamma = 1, \ldots, M\}$ of the image of the parameter space. If we extend this basis by $\{\boldsymbol{u}_\gamma,\ \gamma = M + 1, \ldots, N\}$ to a complete basis of $\mathbb{R}^N$, the vector $\boldsymbol{e}$, which is orthogonal to the image of the parameter space in $\mathbb{R}^N$, may be expanded with respect to this basis as

$$\boldsymbol{e} = \sum_{\gamma=M+1}^{N} e'_\gamma \boldsymbol{u}_\gamma, \tag{8.124}$$

with

$$e'_\gamma = \boldsymbol{u}_\gamma \cdot \boldsymbol{e}, \quad \gamma = M + 1, \ldots, N. \tag{8.125}$$

Since $\boldsymbol{u}_\gamma \cdot \mathsf{K} = 0$ for $\gamma = M + 1, \ldots, N$, we also have

$$e'_\gamma = \boldsymbol{u}_\gamma \cdot \boldsymbol{b}. \tag{8.126}$$

The components $\{e'_\gamma,\ \gamma = M + 1, \ldots, N\}$ are therefore realizations of the random variable

$$E'_\gamma = \sum_{i=1}^{N} (\boldsymbol{u}_\gamma)_i \left( \frac{Y_i}{\sigma_i} - (\mathsf{K}\hat{\boldsymbol{A}})_i \right) \equiv \sum_{i=1}^{N} (\boldsymbol{u}_\gamma)_i \, \frac{Y_i}{\sigma_i}, \tag{8.127}$$

$$\gamma = M + 1, \ldots, N \tag{8.128}$$

and satisfy

$$\langle E'_\gamma \rangle = 0, \quad \text{and} \quad \mathrm{Cov}(E'_{\gamma'}, E'_{\gamma''}) = \delta_{\gamma'\gamma''}. \tag{8.129}$$

Hence, $E'_\gamma, \gamma = M+1, \ldots, N$ are uncorrelated and thus, being Gaussian, also independent.

Since

$$q^2(\hat{\boldsymbol{a}}) = \boldsymbol{e}^2 = \sum_{\gamma=M+1}^{N} e'^2_\gamma, \tag{8.130}$$

we see that $q^2(\hat{\boldsymbol{a}})$ is the sum of $N-M$ squares of independent standard normal random variables and therefore the realization of a corresponding $\chi^2$-distributed random variable.

**Estimation of $\boldsymbol{\sigma_0^2}$** If the value for $\sigma_0$ in the model for the error, $\sigma_i = \sigma_0$, is unknown, it can be estimated from the data. An unbiased estimator for $\sigma_0^2$ is

$$\hat{S}^2 = \frac{1}{N-M}\left(\boldsymbol{Y} - \mathsf{K}\hat{\boldsymbol{A}}\right)^2, \tag{8.131}$$

where $\mathsf{K}$ and $\hat{\boldsymbol{A}}$ now have to be determined for the value $\sigma_0 = 1$. The realizations of $\hat{S}^2$ can also be written as:

$$\hat{s}^2 = \frac{\sigma_0^2}{N-M}\, q^2(\hat{\boldsymbol{a}}). \tag{8.132}$$

We have seen above that $q^2(\hat{\boldsymbol{a}})$ is the realization of a $\chi^2$-distributed random variable with $N-M$ degrees of freedom, and since for such a random variable the expectation value is equal to the number of degrees of freedom (cf. Sect. 2.5), we get

$$\langle \hat{S}^2 \rangle = \frac{\sigma_0^2}{N-M}\,(N-M) = \sigma_0^2. \tag{8.133}$$

**Ill-posed problems** The lengths of the principal axes within the confidence regions are proportional to $1/w_\gamma$. Hence, if a singular value is very small, the confidence region is elongated in the corresponding direction and the associated estimated value is determined only poorly. If some of the singular values happen to be zero, the matrices $\mathsf{D}$ and $\mathsf{K}$ are singular. Let $M'$ be the number of nonvanishing singular values, then

$$\mathsf{K}_{i\alpha} = \sum_{\gamma=1}^{M'} w_\gamma (\boldsymbol{u}_\gamma)_i (\boldsymbol{v}_\gamma)_\alpha, \tag{8.134}$$

and the vector $\hat{\boldsymbol{a}}$ with components

$$\hat{\boldsymbol{a}}_\alpha = \sum_{\gamma=1}^{M'} \frac{1}{w_\gamma} (\boldsymbol{v}_\gamma)_\alpha (\boldsymbol{u}_\gamma \cdot \mathbf{b}) + \sum_{\gamma=M'+1}^{M} \lambda_\gamma \boldsymbol{v}_\gamma, \tag{8.135}$$

minimizes $q^2(\boldsymbol{a})$, where $\lambda_\gamma,\ \gamma = M' + 1, \ldots, M$ may be arbitrary. Evidently there are now many solutions to the minimization problem.

In practice this is also the case when some of the $w_\gamma$ are very small (compared to the largest value $w_1$). In this case the matrix $\mathsf{K}$ is called ill-conditioned, and the problem of fitting $\{a_\alpha\}$ to the data is called also ill-conditioned, or also sometimes called 'ill-posed' (cf. Sect. 11.4).

Hence, we can see from the singular values $w_\gamma$ whether the problem of minimizing $q^2(\boldsymbol{a})$ is 'ill-posed'. It is evident from (8.107) that the reciprocal values of $w_\gamma$ enter into the determination of $\hat{a}_\alpha$. If some $w_\gamma$ become too small, small changes in $b_i$ may cause large changes in $\hat{a}_\alpha$. However, since in general $b_i$ are measured data and therefore subject to uncertainties, this case may lead to large standard errors (cf. (8.111)) for $\hat{a}_\alpha$ such that their determination becomes useless. Furthermore, a computation with single precision may yield completely different values compared to the same computation with double precision.

Problems in which matrices with extremely widely spread singular values occur are therefore called 'ill-posed'. In this case the least-squares estimator is not a good estimator for the parameters and the information from the data is obviously not sufficient for a given model (8.90) to determine the parameters with an adequately small confidence region.

There have been many attempts to determine $\hat{a}_\alpha$ with acceptable standard errors, even though the matrix $\mathsf{K}$ is ill-conditioned. One possibility consists in finding a proper regularization (Miller 1974; Delves and Mohamed 1985); another is to set equal to zero those singular values which are too small, and to determine the $\{\hat{a}_\alpha\}$ only within the subspace of eigenvectors belonging to the remaining, nonvanishing singular values (Press et al. 2007). Other approaches introduce special principles for the determination of $\hat{a}_\alpha$, e.g., the maximum entropy principle (Papoulis 1984; Levine and Tribus 1979).

All these methods use certain additional information about the estimated values. The estimators defined by these methods can thus be interpreted as so-called Bayes estimators, which exploit a priori knowledge about the distribution of the parameters (cf. Chap. 11).

# Chapter 9
# Signal Analysis: Estimation of Spectra

In Chap. 5 we introduced stochastic processes. When a stochastic processes is observed, the result of a measurement can only be recorded for discrete times, e.g., $t, t + \Delta t, t + 2\Delta t, \ldots$. The time step $\Delta t$ is called the sampling time and $1/\Delta t$ the sampling frequency. A sampling frequency of, say, 1 kHz implies that a value is registered 1,000 times a second, i.e., each millisecond. If we choose the unit of time such that $\Delta t = 1$, the sampling times are $t, t + 1, \ldots$ or $t = 1, \ldots, N$. Hence, we will always write for a time series $Y(t), t = 1, \ldots, N$. The finite time series is the result of an observation, and therefore we should denote the values by small letters $y(t)$, since they are to be considered as realizations of random variables. For later purposes, however, we will use the random variable $Y(t)$ associated with each observation $y(t)$.

Observations of stochastic processes thus always result in time series. This suggests that one also considers discrete-time processes of the form $\{Y(t); t = \ldots, -1, 0, 1, \ldots\}$. The sequence of these random variables may in principle be infinitely long. Observations, however, always consist of a finite number of realizations.

In this chapter we will present methods for the transformation of time series. These transformations often serve as a first inspection or a visualization of the data. In the center is the estimation of the spectrum of the process which gives rise to a given time series. In the first section we will discuss the spectrum of such stochastic processes itself, which will serve as a guide for the estimation of the spectrum when given only the times series.

## 9.1 The Spectrum of a Stochastic Process

Björn Schelter

The interdependences of the random variables of a process $\{Y(t)\}$ for different times constitute an essential property of the process. A measure for these interdependences

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_9, © Springer-Verlag Berlin Heidelberg 2012

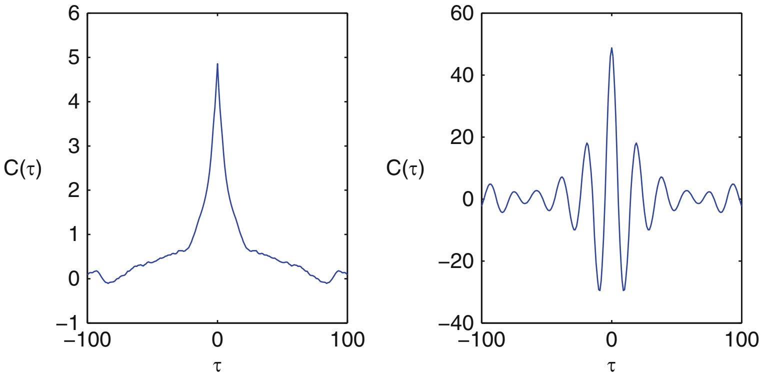


**Fig. 9.1** Two typical covariance functions $C(\tau)$

is the covariance (see e.g. Sects. 5.1 or 5.6, in the following we always assume that $\langle Y(t)\rangle \equiv 0$)

$$\begin{aligned} C(t,\tau) &= \langle Y(t+\tau)Y(t)\rangle \qquad (9.1) \\ &= \int \mathrm{d}y_1\,\mathrm{d}y_2\; y_1 y_2\; \varrho_2(y_1,t+\tau;y_2,t) \\ &= \int \mathrm{d}y_1\,\mathrm{d}y_2\; y_1 y_2\; \varrho_2(y_1,t+\tau|y_2,t)\,\varrho_1(y_2,t)\;. \end{aligned}$$

For a temporally homogeneous process the conditional probability $\varrho_2(y_1,t+\tau|y_2,t)$ does not depend on $t$, and if the system is in its stationary state, $\varrho_1(y_2,t)$ is identical to $\varrho^{\mathrm{stat}}(y_2)$. The covariance $C(t,\tau) \equiv C(\tau)$ then depends only on $\tau$.

Obviously, in this case we also have

$$C(-\tau) = \langle Y(t)Y(t-\tau)\rangle = \langle Y(t+\tau)Y(t)\rangle = C(\tau)\;. \qquad (9.2)$$

The covariance $C(\tau)$, considered as a function of $\tau$, is also referred to as the covariance function. A white noise, for instance, with variance $\sigma^2$ has the covariance function $C(\tau) = \sigma^2\delta_{\tau,0}$, i.e., $C(\tau)$ vanishes for $\tau \neq 0$. In other cases we may see from the behavior of $C(\tau)$ for $\tau > 0$ how the dependences decrease with increasing time difference.

Typical shapes of the function $C(\tau)$ are shown in Fig. 9.1. Thus it might happen that $C(\tau)$ falls off exponentially (or like a sum of exponentially decreasing terms), but $C(\tau)$ may also oscillate with a decreasing amplitude.

For a process at discrete times, $\{Y(t), t = \ldots,-1,0,1,\ldots\}$, the covariance function is only defined for integer values of $\tau$. Its Fourier transform has the form

$$\tilde{C}(\omega) = \sum_{\tau=-\infty}^{+\infty} C(\tau)\mathrm{e}^{\mathrm{i}\omega\tau}\;. \qquad (9.3)$$

The Fourier transform of the process $\{Y(t)\}$ with discrete times reads

$$\tilde{Y}(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{t=-\infty}^{\infty} Y(t)\mathrm{e}^{-\mathrm{i}\omega t}, \quad \omega \in [-\pi, \pi], \tag{9.4}$$

and the inverse is

$$Y(t) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \mathrm{d}\omega \tilde{Y}(\omega)\mathrm{e}^{\mathrm{i}\omega t} . \tag{9.5}$$

Then one may show that

$$\langle \tilde{Y}(\omega)\tilde{Y}^*(\omega')\rangle = \frac{1}{2\pi} \sum_{t,t'} \langle Y(t)\, Y(t')\rangle \mathrm{e}^{-\mathrm{i}\omega t}\mathrm{e}^{\mathrm{i}\omega' t'} \tag{9.6}$$

$$= \langle \left|\tilde{Y}(\omega)\right|^2\rangle \delta(\omega - \omega') = \tilde{C}(\omega)\delta(\omega - \omega') . \tag{9.7}$$

$\tilde{C}(\omega)$ is called the (also power spectrum or spectral density function) spectrum of the process $\{Y(t)\}$.

Thus the spectrum can be calculated either by the Fourier transformation of the covariance function or by building the second moments of the Fourier transform of the stochastic process in time.

Analogous relations can be found for a stochastic processes for which the time variable is continuous. The Fourier-transformation of the covariance function is now defined as in (5.175)

$$\tilde{\mathsf{C}}(\omega) = \int_{-\infty}^{\infty} \mathrm{d}\tau\, \mathsf{C}(\tau)\mathrm{e}^{\mathrm{i}\omega\tau} . \tag{9.8}$$

Introducing the Fourier transform of a stochastic process with continuous time variable as

$$\tilde{Y}(\omega) = \frac{1}{\sqrt{2\pi}} \int\limits_{t=-\infty}^{\infty} Y(t)\mathrm{e}^{-\mathrm{i}\omega t} dt, \quad \omega \in [-\infty, \infty], \tag{9.9}$$

the spectrum $\tilde{\mathsf{C}}(\omega)$ can also be read off from the second moment of the Fourier transformed stochastic process

$$\langle \left|\tilde{Y}(\omega)\tilde{Y}(\omega')\right|\rangle = \langle \left|\tilde{Y}(\omega)\right|^2\rangle \delta(\omega - \omega') = \tilde{\mathsf{C}}(\omega)\delta(\omega - \omega'). \tag{9.10}$$

We note here, that the rigorous mathematical definition of the Fourier transformation for stochastic processes is complicated and we refer the interested reader to Brockwell and Davies (1987) for more details on this aspect.

*Examples.* As an example for a continuous stochastic process we consider a linear first order Langevin process , the model for the 'red' noise, already discussed in Sect. 5.6.1

$$\dot{X}(t) = aX(t) + \eta(t) \text{ with } \eta(t) \propto \mathrm{WN}(0, \sigma^2) . \tag{9.11}$$

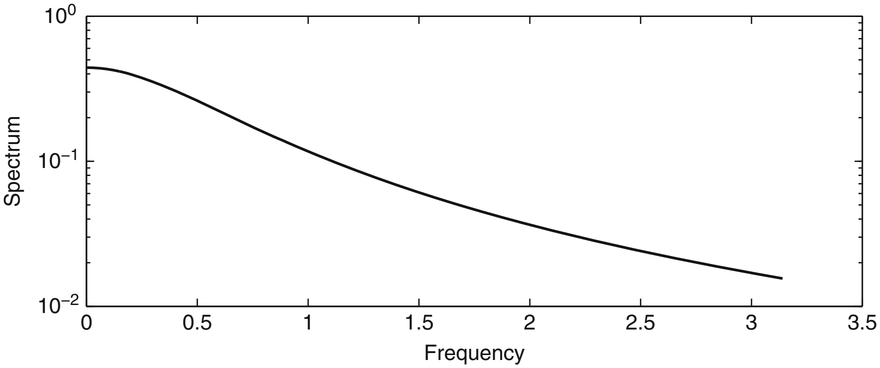


**Fig. 9.2** Spectrum of a Langevin process of first order with $\alpha = 0.6$ and $\sigma^2 = 1$

Applying the Fourier transformation to all terms in this equation, especially with

$$\tilde{\dot{X}}(\omega) = \frac{1}{\sqrt{2\pi}} \int\limits_{t=-\infty}^{\infty} \dot{X}(t) \mathrm{e}^{-\mathrm{i}\omega t} dt \tag{9.12}$$

$$= \frac{\mathrm{i}\omega}{\sqrt{2\pi}} \int\limits_{t=-\infty}^{\infty} X(t) \mathrm{e}^{-\mathrm{i}\omega t} dt = \mathrm{i}\omega \tilde{X}(\omega) \tag{9.13}$$

we obtain

$$\mathrm{i}\omega \tilde{X}(\omega) = \alpha \tilde{X}(\omega) + \tilde{\eta}(\omega)\,. \tag{9.14}$$

$\tilde{X}(\omega)$ represents a stochastic process in the frequency domain. Considering the second moment of this stochastic process, we find

$$\tilde{C}(\omega) = \langle \left|\tilde{X}(\omega)\right|^2 \rangle = \frac{1}{\omega^2 + \alpha^2} \langle |\tilde{\eta}(\omega) \rangle|^2\,. \tag{9.15}$$

By using $C(\tau) = \sigma^2 \delta_{\tau,0}$ for a white noise we get

$$\langle |\tilde{\eta}(\omega)|^2 \rangle = \sigma^2\,, \tag{9.16}$$

and finally for the spectrum of the solution of the linear Langevin equation ((9.11), compare (5.18))

$$\tilde{C}(\omega) = \frac{\sigma^2}{\omega^2 + \alpha^2}\,. \tag{9.17}$$

This spectrum is shown in (Fig. 9.2).

As an example for a spectrum of a discrete stochastic process we consider the spectrum of an autoregressive moving average process, discussed in (5.260) in Sect. 5.9.4

$$\begin{aligned} X(t) &= \sum_{k=1}^{p} \alpha_k X(t-k) + \eta(t) + \sum_{k=1}^{q} \beta_k \eta(t-k) \\ \eta(t) &\propto \mathrm{WN}(0,\sigma^2)\ , \end{aligned} \tag{9.18}$$

To this end, the autoregressive moving average process should be rewritten as

$$X(t) - \sum_{k=1}^{p} \alpha_k X(t-k) = \eta(t) + \sum_{k=1}^{q} \beta_k \eta(t-k) \tag{9.19}$$

Applying the Fourier transformation on both sides of the equation we obtain

$$(1 - \sum_{k=1}^{p} \alpha_k \mathrm{e}^{-\mathrm{i}\omega k})\tilde{X}(\omega) = (1 + \sum_{k=1}^{q} \beta_k \mathrm{e}^{-\mathrm{i}\omega k})\tilde{\eta}(\omega) \tag{9.20}$$

so that also

$$\left|(1 - \sum_{k=1}^{p} \alpha_k \mathrm{e}^{-\mathrm{i}\omega k})\right|^2 \langle\left|\tilde{X}(\omega)\right|^2\rangle = \left|(1 + \sum_{k=1}^{q} \beta_k \mathrm{e}^{-\mathrm{i}\omega k})\right|^2 \langle|\tilde{\eta}(\omega)|^2\rangle\ . \tag{9.21}$$

As again $\langle|\tilde{\eta}(\omega)|^2\rangle = \sigma^2$, we get for the spectrum

$$\tilde{C}(\omega) = \langle\left|\tilde{X}(\omega)\right|^2\rangle = \frac{\left|\sum_{k=0}^{q} \beta_k \mathrm{e}^{-\mathrm{i}\omega k}\right|^2}{\left|\sum_{k=0}^{p} \alpha_k \mathrm{e}^{-\mathrm{i}\omega k}\right|^2}\,\sigma^2 = \frac{\left|\beta(\mathrm{e}^{-\mathrm{i}\omega})\right|^2}{|\alpha(\mathrm{e}^{-\mathrm{i}\omega})|^2}\,\sigma^2 \tag{9.22}$$

with

$$\alpha(z) = 1 - \sum_{k=1}^{p} \alpha_k z^k$$

and

$$\beta(z) = 1 + \sum_{k=1}^{q} \beta_k z^k\ .$$

The autoregressive part in the signal causes peaks in the spectrum (see Fig. 9.3) as the denominator can obtain values close to zero, while the moving average part in the signal leads to dips in the spectrum, as the numerator can obtain values close to zero. The order of the respective parts decides about the number of peaks and dips in the spectrum. This characteristics of the autoregressive moving average processes becomes important when we discuss filters below.

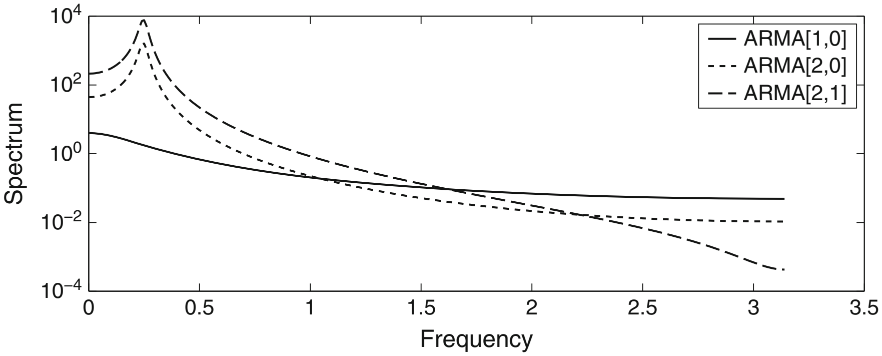


**Fig. 9.3** Spectrum of various ARMA processes, ARMA[1,0] with $a_1 = 0.9$, ARMA[2,0] with $a_1 = 1.9$, and $a_2 = -0.96$, and ARMA[2,1] with $a_1 = 1.9$, $a_2 = -0.96$, and $b_1 = 1.2$. For all spectra $\sigma^2 = 1$

Please note that the spectrum of the AR(1) looks very similar to the spectrum of the red noise for small $\omega$. This is not a surprise: As shown in Sect. 5.9, the AR(1) process is the discrete version of the linear stochastic differential equation for the red noise; for small frequences i.e. large wavelength the discretisation of the time does not matter any more. This can also be seen formally: The spectrum of the AR(1) process reads

$$\tilde{C}(\omega) = \frac{\sigma^2}{|1-\alpha_1 \mathrm{e}^{-\mathrm{i}\omega}|^2} = \frac{\sigma^2}{1-2\alpha_1\cos\omega + \alpha_1^2} \tag{9.23}$$

and for small $\omega$ we can approximate $\cos\omega$ by $1-\frac{1}{2}\omega^2$, thus the denominator is also a linear function of $\omega^2$ as the in the expression of the spectrum of the red noise.

As it becomes important later on, we include here also the discussion of multivariate spectra. If the autoregressive moving average process was an $n$-dimensional vector-valued process $\boldsymbol{X}(t)$ with components $(X_1(t), X_2(t), \ldots, X_n t)^T$, the vector autoregressive moving average process can be defined as

$$\begin{aligned} \boldsymbol{X}(t) &= \sum_{k=1}^{p} \underline{\alpha}_k \boldsymbol{X}(t-k) + \boldsymbol{\eta}(t) + \sum_{k=1}^{q} \underline{\beta}_k \boldsymbol{\eta}(t-k) \\ \boldsymbol{\eta}(t) &\propto \mathrm{WN}(\mathbf{0}, \boldsymbol{\Sigma})\,. \end{aligned} \tag{9.24}$$

The $\underline{(\cdot)}$ stresses that $(\cdot)$ is a matrix, i.e.

$$\underline{\alpha}_k = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \vdots & & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{pmatrix}_k$$

and analogously for $\underline{\beta}_k$. A similar derivation as for the spectrum of an one-dimensional autoregressive moving average process leads to the so-called spectral matrix for the $n$-dimensional vector autoregressive moving average process

$$\tilde{\mathbf{C}}(\omega) =$$

$$\Big(\mathbf{1}-\sum_{k=1}^{p}\underline{\alpha}_k \mathrm{e}^{-\mathrm{i}\omega k}\Big)^{-1}\Big(\mathbf{1}+\sum_{k=1}^{p}\underline{\beta}_k \mathrm{e}^{-\mathrm{i}\omega k}\Big)\boldsymbol{\Sigma}\Big[\Big(\mathbf{1}+\sum_{k=1}^{p}\underline{\beta}_k \mathrm{e}^{-\mathrm{i}\omega k}\Big)\Big(\mathbf{1}-\sum_{k=1}^{p}\underline{\alpha}_k \mathrm{e}^{-\mathrm{i}\omega k}\Big)^{-1}\Big]^H , \tag{9.25}$$

where $(\cdot)^H$ denotes Hermitean transposition and $\mathbf{1}$ the $n$-dimensional identity matrix.

Considering autoregressive processes without the moving average part, the spectral matrix reads

$$\tilde{\mathbf{C}}(\omega) = \Big(\mathbf{1} - \sum_{k=1}^{p}\underline{\alpha}_k \mathrm{e}^{-\mathrm{i}\omega k}\Big)^{-1}\boldsymbol{\Sigma}\Big[\Big(\mathbf{1} - \sum_{k=1}^{p}\underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\Big)^{-1}\Big]^H . \tag{9.26}$$

We note that the autoregressive coefficients enter the spectrum and spectral matrix only inversely.

## 9.2 The Fourier Transform of a Time Series and the Periodogram

Let us now consider a given time series $\{Y(t), t = 1, \ldots, N\}$. In order to estimate the spectrum of the underlying process from these observations we first have to define an estimator for the covariance function. A simple estimator is

$$\hat{C}_N(\tau) = \frac{1}{N}\sum_{t=1}^{N-|\tau|} Y(t)Y(t+\tau) , \quad \tau = 0, \pm 1, \ldots, N-1 . \tag{9.27}$$

The next step would normally be to write down the Fourier transform of this estimator. However, we will first introduce a Fourier transform for the time series $\{Y(t), t = 1, \ldots, N\}$ itself. It is defined as

$$\tilde{Y}_N(\omega) = \frac{1}{\sqrt{N}}\sum_{t=1}^{N} Y(t)\mathrm{e}^{-\mathrm{i}\omega t} . \tag{9.28}$$

The inverse of this Fourier transformation is

$$Y(t) = \frac{1}{\sqrt{N}}\sum_{l=1}^{N} \tilde{Y}_N(\omega_l)\mathrm{e}^{\mathrm{i}\omega_l t} , \tag{9.29}$$

with

$$\omega_l = \frac{2\pi l}{N} \; . \tag{9.30}$$

In the framework of Fourier transformation we only need to know $\tilde{Y}_N(\omega)$ for $\omega = \omega_l, l = 1, \ldots, N$. This sounds plausible, because the Fourier transformation can only be an invertible mapping, if we think of it as a mapping from the set of $N$-tuples (whose elements are denoted by $\{Y(t); t = 1, \ldots, N\}$) into the set of $N$-tuples (with elements $\{\tilde{Y}_N(\omega_l),\ l = 1, \ldots, N\}$).

Notice that the Fourier representation (9.29) defines $Y(t)$ for all $t$. However, it satisfies the relation $Y(t + N) = Y(t)$.

Let us investigate which of the $N$ numbers of the set $\{\tilde{Y}_N(\omega_l),\ l = 1, \ldots, N\}$) are independent. From

$$\tilde{Y}_N(\omega \pm 2\pi) = \tilde{Y}_N(\omega) \tag{9.31}$$

we find

$$\tilde{Y}_N(\omega_{l \pm N}) = \tilde{Y}_N(\omega_l) \; . \tag{9.32}$$

Since $Y(t)$ is real, we obtain

$$\tilde{Y}^*_N(\omega_l) = \tilde{Y}_N(-\omega_l) = \tilde{Y}_N(\omega_{-l}) \tag{9.33}$$

and therefore also

$$\tilde{Y}^*_N(\omega_l) = \tilde{Y}_N(\omega_{N-l}) \; . \tag{9.34}$$

In particular,

$$\tilde{Y}_N(0) \equiv \tilde{Y}_N(\omega_0) = \tilde{Y}_N(\omega_N) \tag{9.35}$$

is real.

If $N$ is even, then from

$$\tilde{Y}^*_N(\omega_{N/2}) = \tilde{Y}_N(-\omega_{N/2}) = \tilde{Y}_N(\omega_{N/2}) \tag{9.36}$$

we also find that $\tilde{Y}_N(\omega_{N/2})$ is real. Hence, in this case,

$$\tilde{Y}_N(\omega_0), \tilde{Y}_N(\omega_1), \ldots, \tilde{Y}_N(\omega_{N/2-1}), \tilde{Y}_N(\omega_{N/2}) \tag{9.37}$$

represent $2+2(N/2-1) = N$ independent values of the set $\{\tilde{Y}_N(\omega_l),\ l = 1, \ldots, N\}$.

If $N$ is odd, we have

$$\tilde{Y}_N(\omega_{(N+1)/2}) = \tilde{Y}^*_N(\omega_{(N-1)/2}) \; , \tag{9.38}$$

and

$$\tilde{Y}_N(\omega_0), \ldots, \tilde{Y}_N(\omega_{(N-1)/2})$$

represent $1 + 2(\frac{N-1}{2}) = N$ independent values.

Let us consider for simplicity only the case $N$ even. In this case the time series $\{Y(t), t = 1, \ldots, N\}$ is mapped into its Fourier transform with frequencies $\omega_l = 2\pi l/N, l = 0, \ldots, N/2$.

We observe two characteristic features of this Fourier transform: Firstly, since the time series is finite we also obtain the Fourier transform only for discrete values of frequencies. The distance between two neighboring points on the frequency axis is

$$\omega_{l+1} - \omega_l = \frac{2\pi}{N} \ , \tag{9.39}$$

i.e., it becomes smaller for increasing length $N$ of the time series. In the limit $N \to \infty$ we obtain a continuum of frequencies in the interval $[0, \pi]$.

Secondly, the maximum frequency $\omega$ appearing in the independent set of Fourier components is

$$\omega_c = \omega_{N/2} = \frac{2\pi l}{N}|_{l=N/2} = \pi \tag{9.40}$$

or

$$f_c \equiv \frac{\omega_c}{2\pi} = \frac{1}{2} \ . \tag{9.41}$$

This result holds for a sampling time $\Delta t$ equal to 1. For general $\Delta t$ we find

$$f_c = \frac{1}{2\Delta t} = \frac{1}{2} f \ , \tag{9.42}$$

where $f$ denotes the sampling frequency. $f_c$ is also called the Nyquist frequency (after the Swedish-American electrical engineer H. Nyquist, born 1889). All signals with a frequency higher than $f_c$ appear, under the sampling with frequency $f = \frac{1}{\Delta t}$, as a signal with a frequency smaller than $f_c$ (see the remark at the end of this section and Fig. 9.5).

We now define a so-called periodogram by

$$\tilde{I}_N(\omega) = |\tilde{Y}_N(\omega)|^2 \ . \tag{9.43}$$

It thus is easily computed for an arbitrary time series by a Fourier transformation. Furthermore, we only have to determine the periodogram for frequencies $f < f_c$. For frequencies between $f_c$ and $2f_c$, i.e., for $\pi < \omega < 2\pi$, the contributions follow from relation (9.34), which corresponds to a reflection of the periodogram in the line $f = f_c$.

It turns out that the periodogram agrees with the discrete Fourier transform of the estimator of the covariance function, defined as

$$\tilde{\hat{C}}_N(\omega) = \sum_{\tau=-(N-1)}^{N-1} \hat{C}_N(\tau) \mathrm{e}^{\mathrm{i}\omega\tau} \tag{9.44}$$

because:

$$\hat{I}_N(\omega_l) \equiv \tilde{Y}_N(\omega_l)\tilde{Y}^*_N(\omega_l) \tag{9.45}$$

$$= \frac{1}{N}\sum_{t=1}^{N}\sum_{t'=1}^{N} Y(t)\mathrm{e}^{-\mathrm{i}\omega_l t}Y(t')\mathrm{e}^{\mathrm{i}\omega_l t'} \tag{9.46}$$

$$= \sum_{\tau=-(N-1)}^{N-1} \frac{1}{N}\sum_{t=1}^{N-|\tau|} Y(t)Y(t+\tau)\mathrm{e}^{\mathrm{i}\omega_l \tau} \tag{9.47}$$

$$= \sum_{\tau=-(N-1)}^{N-1} \hat{C}_N(\tau)\mathrm{e}^{\mathrm{i}\omega_l \tau} \tag{9.48}$$

$$= \tilde{\hat{C}}_N(\omega_l)\ . \tag{9.49}$$

This relation between $\tilde{\hat{C}}_N(\omega_l)$ and the periodogram is known as the Wiener–Khinchin theorem.

Hence, the periodogram is an estimator for the spectrum of the process realized by the time series $\{Y(t), t = 1, \ldots, N\}$. This estimator may therefore be determined in two ways: Either we first estimate the covariance function and compute its Fourier transform according to (9.44), or we first transform the time series and form the product according to (9.43).

Estimators are used to determine the expectation value of a stochastic quantity on the basis of a sample. As we saw in Sect. 8.1, there is a whole catalog of desirable properties for an estimator, and in general there are several estimators for the same quantity which meet these requirements to different degrees.

It turns out that the periodogram $\hat{I}_N(\omega_l) = \tilde{\hat{C}}_N(\omega_l)$ is not a good estimator for the spectrum $\tilde{C}(\omega)$, because it is not a consistent estimator. The variance of $\hat{I}_N(\omega_l)$ does not tend towards zero with increasing length $N$ of the time series, but is actually independent of $N$ (see e.g. Schlittgen and Streitberg 1987).

*Example.* Let $\{Y(t), t = \ldots, -1, 0, 1, \ldots\}$ be a white noise with variance $\sigma^2$. Then

$$C(\tau) = \langle Y(t)Y(t+\tau)\rangle = \sigma^2\delta_{\tau,0} \tag{9.50}$$

and

$$\tilde{C}(\omega) = \sum_{\tau=-\infty}^{+\infty} C(\tau)\mathrm{e}^{\mathrm{i}\omega\tau} = \sigma^2\ . \tag{9.51}$$

Hence, the spectrum is independent of frequency (which is the reason for the expression 'white' noise). We now consider a realization of this process for a finite number of instants, i.e., a time series $\{y(t), t = 1, \ldots, N\}$, where each $y(t)$ is a realization of a normal random variable with variance $\sigma^2$. We compute the periodogram $\hat{I}_N(\omega_l)$ according to (9.43). The result is represented in Fig. 9.4 for

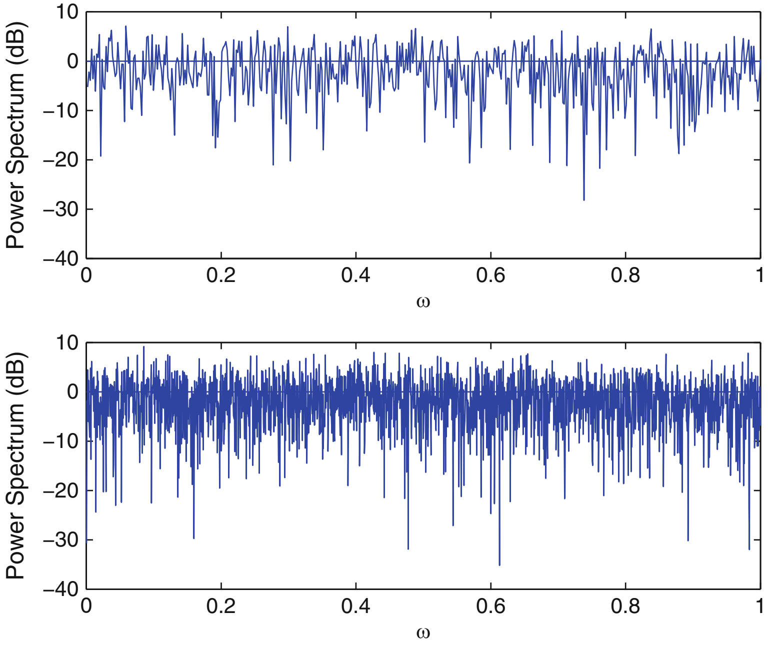


**Fig. 9.4** Periodograms of time series of a white noise. The length of the time series is $N = 1{,}028$ (*above*) and $N = 4{,}096$ (*below*). The *horizontal line* corresponds in each case to the spectrum of the process. The frequency is given in units of the Nyquist frequency and shown is not $\hat{I}_N$ itself, but $10 \log_{10} \hat{I}_N$ (See remark at the end of Sect. 9.4)

$N = 1{,}024$ and $N = 4{,}096$. Obviously the variance does not decrease for larger values of $N$.

This can also be seen by direct analytical calculation. The Fourier transform of the white noise at some frequency

$$\tilde{\epsilon}_N(\omega) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \epsilon(t) \mathrm{e}^{-\mathrm{i}\omega t} = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \epsilon(t) \big[ \cos(\omega t) + \mathrm{i} \sin(\omega t) \big] \,. \tag{9.52}$$

is a complex number, where the real part

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} \epsilon(t) \cos(\omega t) \tag{9.53}$$

as well as the imaginary part

$$\mathrm{i} \, \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \epsilon(t) \sin(\omega t) \tag{9.54}$$

are normal random variables with mean zero and variance $\sigma^2/2$ as $\sin^2(\omega t) + \cos^2(\omega t) = 1$. The periodogram at some frequency therefore is, as a sum of 2 Gaussian distributed random variables, a $\chi^2$ distributed random variable with 2 degrees of freedom and with mean $\sigma^2$. The distribution is independent on $N$.

For a very general class of stochastic processes one will find a similar result using the central limit theorem stating that the sum of independently distributed random variable with finite first and second order moments is asymptotically Gaussian distributed. As most processes actually follow a distribution with finite first and second order moment, the central limit theorem guarantees hat the periodogram is $\chi^2$ distributed with 2 degrees of freedom times the true spectrum.

It should be mentioned here that the above derivation is only true for $\omega \neq, 0, \pi, 2\pi, \ldots$. In cases where $\omega$ assumes one of these values we find a $\chi^2$ distribution with one degree of freedom, as can be seen from the sum $1/\sqrt{N}\ \sum_{t=1}^{N} \epsilon(t)\mathrm{e}^{-\mathrm{i}\omega t}$ in those cases.

We may construct a consistent estimator of the spectrum by starting, however, with the periodogram. Consistency can be achieved by applying a special so-called smoothing filter on it. Before we discuss this we will introduce filters in the next section, and we will take up the subject of spectral estimation again in Sect. 9.4.

*Remark.* If we consider the time series resulting from sampling the continuous signals $\sin((\omega + 2m\pi)t), m = 0, \pm 1, ..$ with a sampling time $\Delta t = 1$, we find that all these signals yield the same discrete time series (Fig. 9.5).

In a general signal all Fourier contributions of this type are present. The discrete time series cannot resolve these contributions and its Fourier transform is therefore falsified. But as these Fourier contributions in general decrease rapidly as a function of the frequency, this effect is most important just below the Nyquist frequency, as here the influence of the contributions from just above the Nyquist frequency is largest.

This effect is also referred to as the 'aliasing effect'. This aliasing effect can be reduced if, before the actual sampling, the contributions from higher frequencies are suppressed by an electronic device, known as an anti-aliasing filter.

## 9.3 Filters

The Fourier transformation is a special, frequency dependent transformation of a process or a time series. Other linear transformations include the so-called filters. We will first discuss filters for stochastic processes. If we denote the transformed quantities by $Z(t)$, then the most general filter, applied to a time series $\{Y(t)\}$, is given by

$$Z(t) = \sum_{k=1}^{p} \alpha_k Z(t-k) + \sum_{k=-s}^{q} \beta_k Y(t-k)\ . \qquad (9.55)$$

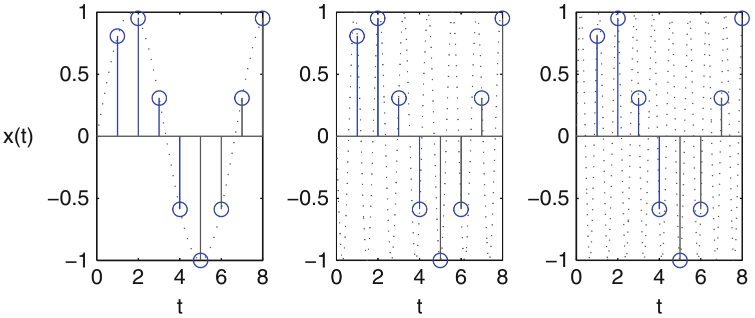


**Fig. 9.5** The signals $\sin \omega t$, $\sin(\omega - 2\pi)t$, and $\sin(\omega + 2\pi)t$ yield the same time series for a sampling time $\Delta t = 1$. In a general signal all Fourier contributions of this type are present. The discrete time series thus cannot resolve these contributions and its Fourier transform is therefore falsified (aliasing)

If $s = 0$, only values of $Y(t)$ from the present and the past have an influence on the value of $Z(t)$. In this case the filter is called causal. An ARMA process (Sect. 5.9) is thus a filtered white noise, i.e., $Y(t) = \eta(t) \propto \mathrm{WN}(0,1)$.

Using the formal shift operator $B$ defined by

$$B^k Y(t) = Y(t+k) \ , \tag{9.56}$$

the filter may also be written as

$$\alpha(B) Z(t) = \beta(B) Y(t) \ , \tag{9.57}$$

where $\alpha$ and $\beta$ now are polynomials of the form

$$\alpha(B) = 1 - \alpha_1 B^{-1} - \ldots - \alpha_p B^{-p} \tag{9.58}$$

$$\beta(B) = \beta_{-s} B^s + \ldots + \beta_q B^{-q} \ . \tag{9.59}$$

Hence, formally we may also write

$$Z(t) = \frac{\beta(B)}{\alpha(B)} Y(t) \ . \tag{9.60}$$

A filter should be stable. That is, a process $Y(t)$ which always takes finite values is transformed into a process $Z(t)$ which is finite in the same sense. The realizations of $Z(t)$ therefore may not diverge for increasing $t$. It can be shown that a filter is stable if all zeros of the polynomial $\alpha(z)$ are inside the unit circle. As we have seen (Sect. 5.9), this is also the condition for the corresponding ARMA process to be stable.

### *9.3.1 Filters and Transfer Functions*

Let us investigate how the Fourier transforms of $\{Y(t)\}$ and $\{Z(t)\}$ are related. For this purpose we first consider

$$\sum_t B^{-k} Z(t) \mathrm{e}^{-\mathrm{i}\omega t} = \sum_t Z(t-k) \mathrm{e}^{-\mathrm{i}\omega(t-k)} \mathrm{e}^{-\mathrm{i}\omega k} \tag{9.61}$$

$$= \left(\mathrm{e}^{-\mathrm{i}\omega}\right)^k \tilde{Z}(\omega) \, . \tag{9.62}$$

If we write in general

$$Z(t) = \sum_{t'} h(t') Y(t-t') \, , \tag{9.63}$$

this immediately leads to

$$\tilde{Z}(\omega) = \tilde{h}(\omega) \tilde{Y}(\omega) \, , \tag{9.64}$$

with

$$\tilde{h}(\omega) = \sum_{t'} h(t') \mathrm{e}^{-\mathrm{i}\omega t'} \, . \tag{9.65}$$

The inverse of this relation reads

$$h(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\omega \tilde{h}(\omega) \mathrm{e}^{\mathrm{i}\omega t} \, . \tag{9.66}$$

The function $\tilde{h}(\omega)$ is called the transfer function. We find from (9.60) to (9.62)

$$\tilde{h}(\omega) = \frac{\beta\left(\mathrm{e}^{\mathrm{i}\omega}\right)}{\alpha\left(\mathrm{e}^{\mathrm{i}\omega}\right)} \, . \tag{9.67}$$

The transfer function may also be represented as

$$\tilde{h}(\omega) = r(\omega) \mathrm{e}^{\mathrm{i}\phi(\omega)}, \quad r(\omega) \geq 0 \, . \tag{9.68}$$

Here $r(\omega)$ is called the gain of the filter and $\phi(\omega)$ the phase. The gain amplifies or weakens the corresponding Fourier components of $\tilde{Y}(\omega)$, depending on whether $r(\omega)$ is larger or smaller than 1, respectively. Setting $r(\omega) = 0$ for $\omega \geq \omega_0$ one speaks of an (ideal) low-pass filter, which permits only frequencies lower than $\omega_0$ to pass through. We will see that in reality the best one can expect is that $r(\omega)$ falls off sharply at $\omega = \omega_0$ such that $r(\omega)$ may be more or less neglected for $\omega \geq \omega_0$. A high-pass filter is defined correspondingly. A band-pass filter only transmits frequencies within a band $\omega_1 \leq \omega \leq \omega_2$; only in this interval is $r(\omega)$ essentially different from 0. In a band-stop filter the gain vanishes for frequencies inside such an interval, which is then also called a stop-band, while the gain is close to 1 outside this interval.

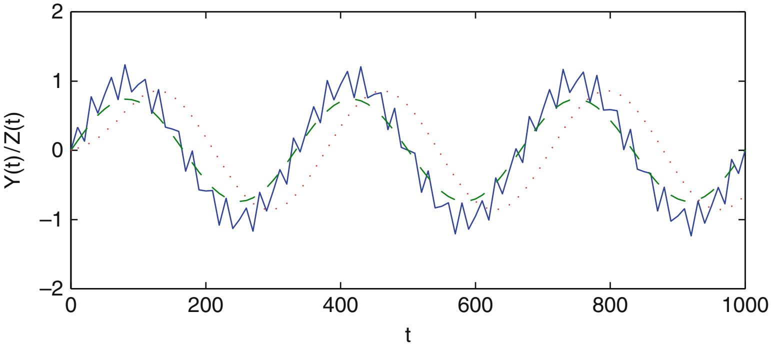


**Fig. 9.6** Time series (*solid line*) filtered with a low-pass filter $\tilde{h}(\omega)$ $(\cdots)$, exhibiting a phase such that a time shift results. The same time series filtered with $\tilde{h}^*(\omega)\tilde{h}(\omega)$ (——) has an identically vanishing phase and therefore no time shift. The smaller gain, however, is revealed in a weaker amplitude

A low-pass filter is also always a smoothing filter, because it filters out the high frequencies. For instance, the moving average,

$$Z(t) = \frac{1}{2p+1}[Y(t+p)+ \\ \ldots + Y(t+1) + Y(t) + Y(t-1) + \ldots + Y(t-p)] \,, \tag{9.69}$$

is a simple smoothing filter. However, it has a bias e.g. at local maxima of the function, because $Z(t)$ can be considered as the fit of the function within the moving window $(t-p, t+p)$ by a constant and at a local maximum this constant will certainly lie below the maximum. A better approximation would be a fit within the window by a polynomial of order, say, two or four. This will lower the bias considerably. Such a filter is called a Sawitzky-Golay smoothing filter (see also Press et al. 2007).

A phase $\phi(\omega) \neq 0$ changes the phase relation between different frequency contributions. However, a phase $\phi(\omega) \propto \omega$ only leads to a time shift of the total signal. Indeed, with $\tilde{h}(\omega) = \mathrm{e}^{-\mathrm{i}\omega d}$ we obtain

$$Z(t) = \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \tilde{Z}_N(\omega_l)\mathrm{e}^{\mathrm{i}\omega_l t} = \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \mathrm{e}^{-\mathrm{i}\omega_l d}\, \tilde{Y}_N(\omega_l)\mathrm{e}^{\mathrm{i}\omega_l t} \tag{9.70}$$

$$= \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \tilde{Y}_N(\omega_l)\mathrm{e}^{\mathrm{i}\omega_l (t-d)} = Y(t-d) \,. \tag{9.71}$$

Time shifts may be avoided if one uses the filter with the transfer function

$$\tilde{h}_2(\omega) = \tilde{h}^*(\omega)\tilde{h}(\omega) = |r(\omega)|^2 \,, \tag{9.72}$$

which obviously has the phase $\phi(\omega) \equiv 0$ (see Fig. 9.6).

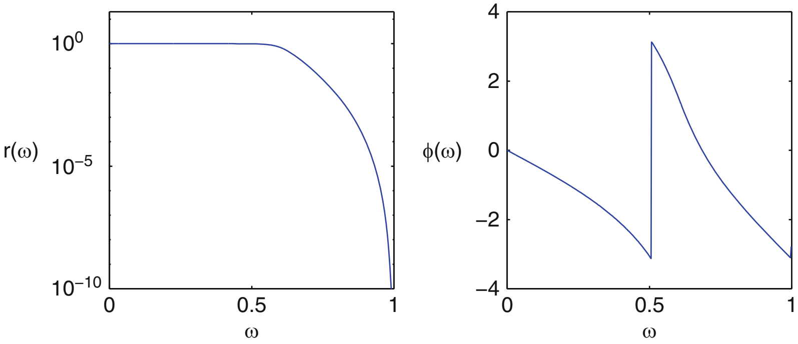


**Fig. 9.7** Typical dependence of gain (*left*) and phase (*right*) of a low-pass filter (Nyquist frequency = 1)

On the other hand, filtering with the transfer function $\tilde{h}^*(\omega)$ implies that

$$Z(t) = \frac{\beta(B^{-1})}{\alpha(B^{-1})} Y(t) \, , \tag{9.73}$$

and therefore

$$Z(t) = \alpha_1 Z(t+1) + \ldots + \beta_0 Y(t) + \beta_1 Y(t+1) + \ldots \, , \tag{9.74}$$

i.e., the filtering is backwards in time. Such a filter is called acausal. It can only be used off-line, i.e., when the total time series $Y(t)$ is at hand, not when it is generated.

Figure 9.7 shows the gain $r(\omega)$ and the phase $\phi(\omega)$ of a typical low-pass filter.

We distinguish the following cases:

- When $\alpha(z) \equiv 1$, i.e., $p = 0$, the filter is called an FIR filter (FIR stands for Finite Impulse Response). Other names for such a filter include: ‘nonrecursive’, ‘moving average’, and ‘all-zero’ (the transfer function $\tilde{h}(\omega)$ has only zeros, no poles). Certain symmetry relations among the coefficients $h(t)$ of an FIR filter may lead to a linear phase. If, e.g., $h(-t) = \pm h(t)$, then $\tilde{h}(\omega)$ is real or purely imaginary. In either case the phase is constant. If the coefficients $h(0), \ldots, h(N)$ of a causal FIR filter satisfy the relation $h(N-t) = \pm h(t)$, we obtain for the phase
$$\phi(\omega) = -\frac{N}{2}\omega + \text{const} \, . \tag{9.75}$$
- When $\alpha(z) \neq 1$, i.e., $p \neq 0$, the filter is called an IIR filter (IIR stands for Infinite Impulse Response). Other names are: ‘recursive’ and ‘autoregressive’. If in addition $q = 0$, this filter is also referred to as an ‘all-pole’ filter (the transfer function possesses only poles, no zeros).

  If $p$ and $q$ are both different from zero, the filter is a general IIR filter, and it is called an autoregressive moving-average filter.

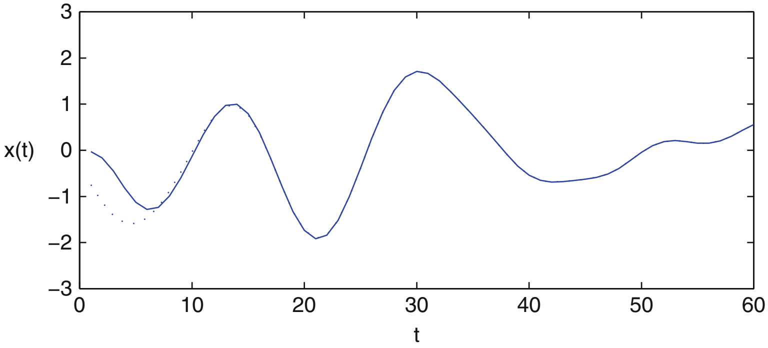


**Fig. 9.8** Application of an IIR filter of order 3 to an AR(2) process, with two different sets of initial values. The transient time is clearly larger than the order of the filter

In the filtering of time series one encounters the problem of startup transients. For $t = 1$, e.g., we need the 'initial values' $Z(0), Z(-1), \ldots, Z(1-p)$ and $Y(0), Y(-1), \ldots, Y(1-q)$, which are not given. Often the filtered time series is defined as if these initial values were equal to zero. A different choice, however, would lead to different values for $Z(t)$, at least for a certain transient time. For an FIR filter this transient time is in any case surpassed for $t > q$; for an IIR filter this transient time in principle never ends, although the influence of the initial values becomes smaller and smaller with increasing time and is soon negligible (Fig. 9.8).

## *9.3.2 Filter Design*

For given polynomials $\beta(z)$ and $\alpha(z)$, the properties of a filter can easily be analysed. But in practice one is usually confronted with the reverse problem. One has an idea about the ideal shape of the gain and the phase of the filter and wants to determine appropriate polynomials $\beta(z)$ and $\alpha(z)$. Of course, for a finite order of the polynomials the ideal shape can only be approximated. The different types of filter differ with respect to the features emphasized in this approximation.

The construction of a filter with certain qualities is treated extensively in many textbooks on electrical engineering and signal analysis (Rabiner and Gold 1975; Oppenheimer and Schafer 1989). Here we will present only a brief survey.

### FIR Filter Design

FIR filters are particularly simple to design. In addition, they are always stable; it is easily arranged for them to have a linear phase (choose $h(N-t) = \pm h(t)$); and also the hardware may be constructed without major effort. The disadvantage, however, is that a very high filtering order $q$ is often required to obtain the desired features. As an example we will demonstrate how to construct a low-pass filter with

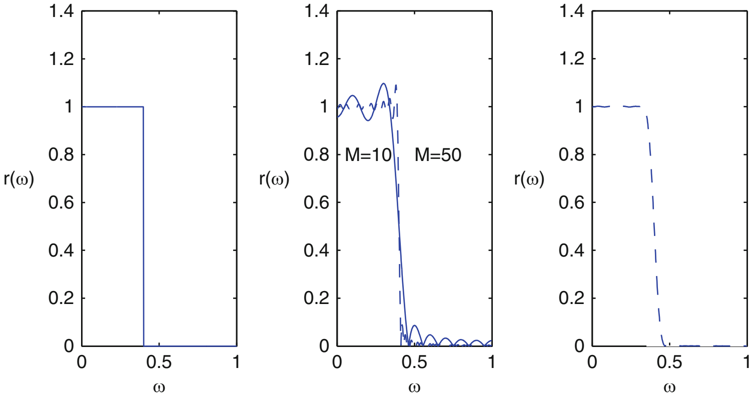


**Fig. 9.9** Ideal gain $r(\omega)$ (*left*), gain of the filter $\tilde{h}'(\omega)$ for various values of $M$: demonstration of the Gibbs phenomenon (*middle*) and approximation by $\tilde{h}''(\omega)$ (*right*) (Nyquist frequency = 1)

a cut-off frequency at $\omega = \omega_0$, i.e., with an idealized shape of the gain as shown in Fig. 9.9(left). Then

$$h(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{h}(\omega)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega = \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0} \mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega \tag{9.76}$$

$$= \frac{\omega_0}{\pi}\operatorname{sinc}\left(\frac{\omega_0}{\pi}t\right), \tag{9.77}$$

where the sinc function is defined by

$$\operatorname{sinc}(x) = \frac{\sin \pi x}{\pi x}\ . \tag{9.78}$$

In this form $h(t)$ thus corresponds to an IIR filter. A finite filter may be obtained by taking

$$h'(t) = h(t)w(t)\ , \tag{9.79}$$

where $w(t)$ is a so-called window function, which vanishes for $|t|$ larger than a suitably chosen $t_0 = M$. We first choose a simple rectangular window, i.e.,

$$w(t) \equiv w_M^{\mathrm{R}}(t) = \begin{cases} 1 \text{ for} & |t| < M \\ 0 \text{ otherwise}\ . \end{cases} \tag{9.80}$$

We thereby obtain a filter whose transfer function $\tilde{h}'(\omega)$ differs, of course, from $\tilde{h}(\omega)$ (Fig. 9.9, middle). This cutting off of the filter $h(t)$ has the effect that ripples appear in the gain in the vicinity of the cut-off frequency $\omega_0$. For larger values of $M$ these ripples do not get smaller, but they move closer together. The amplitude of the last ripple before the discontinuitiy of the ideal filter always amounts to about 9% of the height of this discontinuity.

The appearance of these ripples is known as the Gibbs phenomenon. A smoothing of the ripples may be achieved by letting a smoothing filter act on $\tilde{h}'(\omega)$, which involves a convolution. For $h'(t)$ this implies multiplication by a second window function, or, in other words, the choice of a different window function right from the beginning. Window functions which have proved to be suitable are

- The Hamming window, defined by

$$w_M(t) = \alpha + (1-\alpha)\cos\left(\frac{2\pi t}{2M+1}\right), \quad -M \leq t \leq M, \tag{9.81}$$

  with, e.g., $\alpha = 0.5$,
- The Bartlett window, defined by

$$w_M(t) = 1 - \frac{|t|}{M}, \quad -M \leq t \leq M\ . \tag{9.82}$$

Figure 9.9(right) shows the gain of the filter

$$h''(t) = h(t)w_M(t)\ , \tag{9.83}$$

where $w_M(t)$ is the Hamming window. Obviously, the ripples have almost vanished at the cost of a reduction of the edge steepness.

There are further methods to approximate a given shape of the gain function for an FIR filter. One possibility is to minimize the integral over the quadratic discrepancy between ideal and actual shape of the gain function (least-squares FIR filter). The discrepancy may be additionally weighted by a frequency-dependent function. One can also minimize the maximum of this discrepancy over the whole frequency range. This leads to a so-called equiripple filter, where all ripples have the same height.

### IIR Filter Design

For the construction of an IIR filter one has to approximate the given shape of the gain function $|\tilde{h}(\omega)|$ by a rational function $|\beta(\mathrm{e}^{\mathrm{i}\omega})/\alpha(\mathrm{e}^{\mathrm{i}\omega})|$. In comparison to an FIR filter, a much lower order of the polynomials $\beta(z)$ and $\alpha(z)$ is in many cases sufficient to obtain the necessary quality for the approximation. The phase often shows a nonlinear dependence; however, in an off-line filtering one can always achieve a zero-phase filtering by using the transfer function $\tilde{h}^*(\omega)\tilde{h}(\omega)$.

The specifications of IIR filters may be given in several ways:

- One may state only the order of the filter and the lower and upper frequencies of the pass-bands. (i.e. the frequency ranges for which $r(\omega)$ should be as close to 1 as possible.)

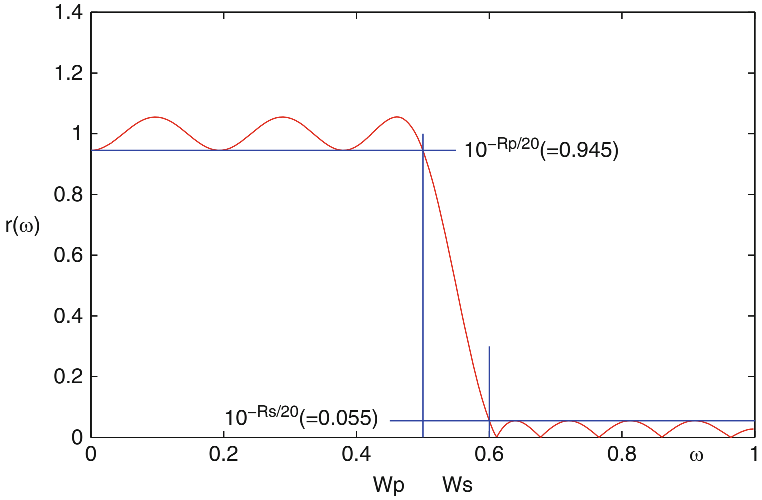


**Fig. 9.10** Parametrization of the amplitudes of the ripples in the pass-band and the stop-band by $Rp$ and $Rs$, respectively, and of the transient range by $Wp, Ws$ for a low-pass filter (Nyquist frequency = 1)

- One may specify the maximum amplitudes of the ripples (parametrized by two numbers $Rp, Rs$ for the pass-band and the stop-band, respectively) as well as the maximum size of the transient range $[Wp, Ws]$ (Fig. 9.10). Hence, four parameters characterize each switch between a pass-band and a stop-band.
- One may also specify the general shape of the gain function and the phase function and make a suitable fit for the function $\tilde{h}(\omega)$.

Other types of IIR filters are distinguished by the way the adaptation to a given specification is realized (Fig. 9.11).

- Butterworth filters exhibit an optimal flat shape of $r(\omega)$ in the pass-band and the stop-band. This leads, however, to a weaker steepness of the jumps in frequency space and a nonlinear phase dependence.
- In Tschebyshev filters emphasis is placed on a good approximation in the stop-band (type I filter) or the pass-band (type II filter), while the respective other band may have ripples of size equal to some given amplitude.
- The elliptic filters have a transient range which is as narrow as possible. On the other hand, ripples of a certain given amplitude are permitted in both the stop-band and the pass-band.
- Bessel filters exhibit a linear dependence of the phase in those ranges where $r(\omega)$ is large. With respect to the time evolution after the filtering this leads to a minimum ‘overshooting’ for steplike shapes of the signal. However, this advantage is accompanied by a very poor steepness of the jumps.

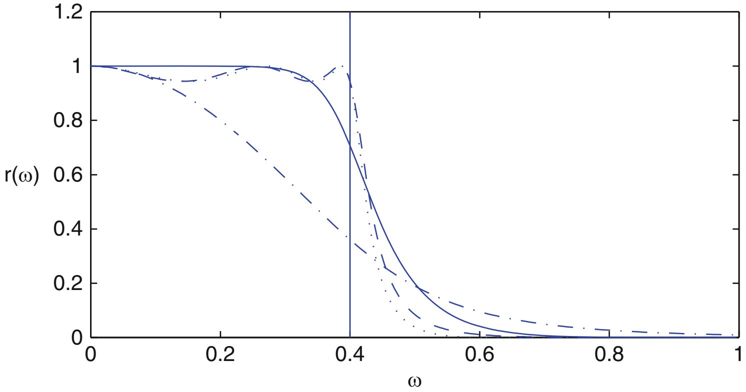


**Fig. 9.11** Dependence of the gain $r(\omega)$ on the frequency (Nyquist frequency = 1) for various IIR filters of fifth order: Butterworth filter (*solid line*), Tschebyshev filter (---), elliptic filter (...) and Bessel filter (-·-·) (Nyquist frequency = 1)

## 9.4 Consistent Estimation of Spectra

Having discussed different types of filters in the previous section, in particular smoothing filters, we now come back to the estimation of the spectrum of a time series.

One method extensively studied in the literature (Schlittgen and Streitberg 1987; Priestley 1981) to obtain a consistent estimator for the spectral density simply involves smoothing the periodogram by some window function. This method has also been applied in the previous section for smoothing the ripples in an FIR filter.

For the smoothing we introduce some weight function or spectral window $K_M(\omega)$, depending on some extra parameter $M = M(N)$, where $N$ denotes again the length of the time series.

The estimator for the spectrum $\widetilde{C}(\omega)$ of a time series of length $N$ is defined by

$$\widehat{\widetilde{C}}(\omega_j) = \sum_{|k|\leq M} K_M(\omega_{k+j})\widehat{I}_N(\omega_{k+j}) \,, \tag{9.84}$$

where $\widehat{I}_N(\omega)$ is the periodogram defined in Sect. 9.2. For example, the simplest smoothing is achieved by the moving average

$$\widehat{\widetilde{C}}(\omega_j) = \frac{1}{2M+1} \sum_{|k|<M} \widehat{I}_N(\omega_{k+j}) \,. \tag{9.85}$$

(The corresponding window function which in this case multiplies the covariance function is called the Daniell window.) The conditions for the smoothing filter, which are to be imposed in order for the estimate $\hat{\hat{C}}(\omega_j)$ to be consistent, are (Brockwell and Davies 1987)

$$M \to \infty \qquad \text{and} \quad M/N \to 0 \quad \text{as} \quad N \to \infty \tag{9.86}$$

$$K_M(-\omega_k) = K_M(\omega_k); \quad K_M(\omega_k) \geq 0 \quad \text{for all}\, k, \tag{9.87}$$

$$\sum_{|k| \leq M} K_M(\omega_k) = 1, \tag{9.88}$$

and

$$\sum_{|k| \leq M} K_M^2(\omega_k) \to 0 \quad \text{for} \quad N \to \infty\,. \tag{9.89}$$

For the variances we obtain

$$\mathrm{Var}(\hat{\hat{C}}(\omega_j)) \propto \hat{\hat{C}}(\omega_j)^2 \sum_{|k| \leq M} K_M^2(\omega_k)\,, \tag{9.90}$$

which for $N \to \infty$ tends to zero because of (9.89). Thus consistency is achieved.

Furthermore, one can show that $\hat{\hat{C}}(\omega_j)$ is unbiased in the limit $N \to \infty$ and that

$$\mathrm{Cov}\Big(\hat{\hat{C}}(\omega), \hat{\hat{C}}(\omega')\Big) \to 0 \quad \text{for} \quad N \to \infty\,, \tag{9.91}$$

provided that $\omega,\ \omega'$ are not too close to each other, i.e., provided $|\omega - \omega'|$ is larger than the width of the window. Hence, the estimators for different frequencies are also uncorrelated if the frequencies are sufficiently far apart in the above-mentioned sense.

Figure 9.12 shows different estimates of the spectrum of an AR(2) process. The exact spectrum is shown in each subfigure as a solid line. In the two figures in the left column, the window width $M$ is increased as $\sqrt{N}$ ($M \approx 0.4\sqrt{N}$). The variance obviously decreases with $N$. In the middle column, $M$ is held fixed when $N$ is increased. The factor $\sum_{|k| \leq M} K_M^2(\omega_k)$ then stays constant, and the variance does not decrease. In the right column, a higher value of $M$ is chosen, which leads to better smoothing (compare it with the other upper figures where the same $N$ is chosen). The smoothing is, however, too strong for the peak, where a large bias is observed. Only if we enlarge the sample size $N$, the peak is again reconstructed appropriately.

There is a trade-off between unbiasedness and smoothness. The discussion of criteria for an optimum choice of the parameter $M$, for given $N$, goes beyond the scope of this textbook. In the literature, one also finds methods for a local, i.e., frequency dependent choice of the smoothing parameter $M$. In the neighborhood of peaks, a smaller $M$ is preferred; outside of the peaks, stronger smoothing is achieved with higher values of $M$.

Though the estimator of the spectrum discussed so far is asymptotically unbiased, the bias for a finite time series may be large. The bias can be reduced considerably

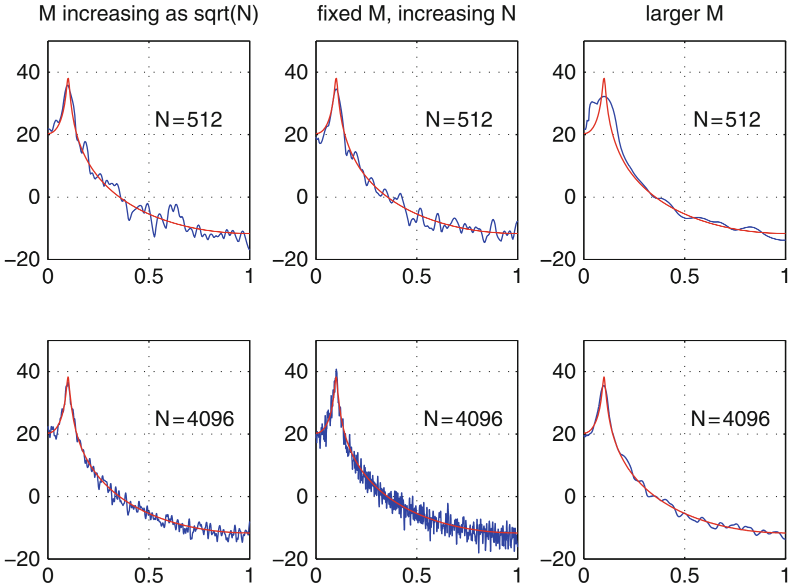


**Fig. 9.12** In the two figures in the *left* column, the window width $M$ is increased as $\sqrt{N}$ ($M \approx 0.4\sqrt{N}$). The variance obviously decreases with $N$. In the *middle* column, $M$ is held fixed when $N$ is increased. The variance does not decrease. In the *right* column, a higher value of $M$ is chosen, which leads to better smoothing. The smoothing is, however, too strong for the peak, where a large bias is observed. Only if we enlarge the sample size $N$, the peak is again reconstructed appropriately. The exact spectrum is shown in each subfigure as a *solid line*

by a procedure which is called "taper" or "data windowing". This is achieved by multiplying the data $\{Y(t), t = 1, \ldots, N\}$ by a window function $\{w(t), t = 1, \ldots, N\}$. Using such data windowing, one may construct the modified Fourier transform

$$\tilde{Y}'(\omega) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} w(t) Y(t) \mathrm{e}^{-\mathrm{i}\omega t} \tag{9.92}$$

and the modified periodogram

$$I'_N(\omega_i) = \frac{1}{W} |\tilde{Y}'(\omega_i)|^2 \,, \tag{9.93}$$

where

$$W = \sum_{t=1}^{N} w(t)^2, \tag{9.94}$$

which has to be smoothed again by a spectral window $K_M(\omega)$ to get a consistent estimator for the spectrum $\widetilde{C}(\omega_i)$.

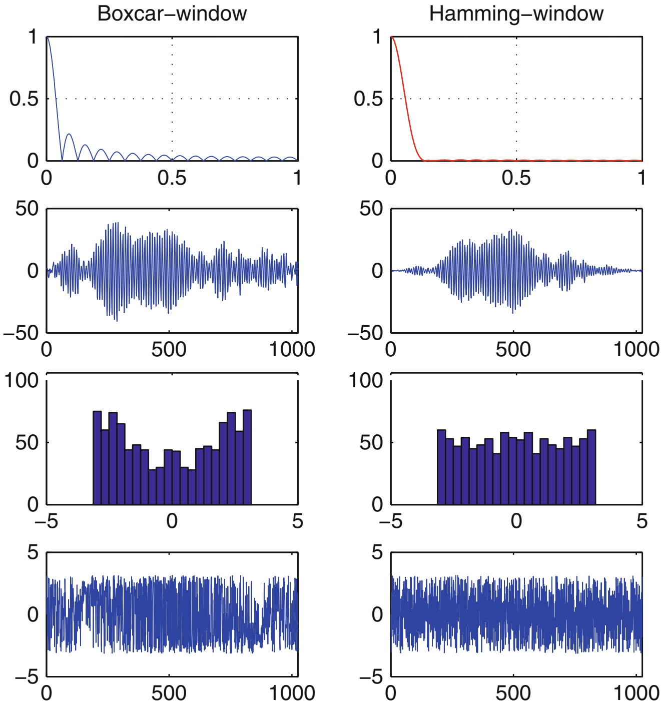


**Fig. 9.13** Fourier transform of the boxcar window (*upper left*) and of the Hamming window (*upper right*). Below these are correspondingly presented: the windowed times series, the histogram of the phases of the Fourier transform of the windowed data, and the times series of the phases

That such data windowing decreases the bias can be seen as follows: The observed times series has to be regarded as a segment which has been cut out of a much longer realization of a stochastic process. Thus, taking the observed time series as it is, corresponds to multiplying the longer realization by a so-called boxcar data window $\{w(t) \equiv 1,\ t = 1, \ldots, N,\ w(t) = 0, \text{ otherwise}\}$. The Fourier transform of the time series thus is a convolution of the exact Fourier transform (of the longer realization) with the Fourier transform of the boxcar window. In Fig. 9.13 the Fourier transform of the boxcar window is shown in the upper left subfigure. One observes that this window will mix the frequencies over a long range. In the left column of this figure, furthermore, are shown the time series itself and the histogram of the phases of the Fourier transform of the time series. According to the theory, these phases should be equally distributed, but they are apparently not, because of this mixing. The time series of these phases in the last subfigure below also shows this strange behavior. In the right column of Fig. 9.13 the same is shown for the Hamming data window (see (9.81)), now given by

$$w(t) = \alpha - (1-\alpha)\cos\left(2\pi \frac{t-1}{N-1}\right), \quad 1 \le t \le N , \tag{9.95}$$

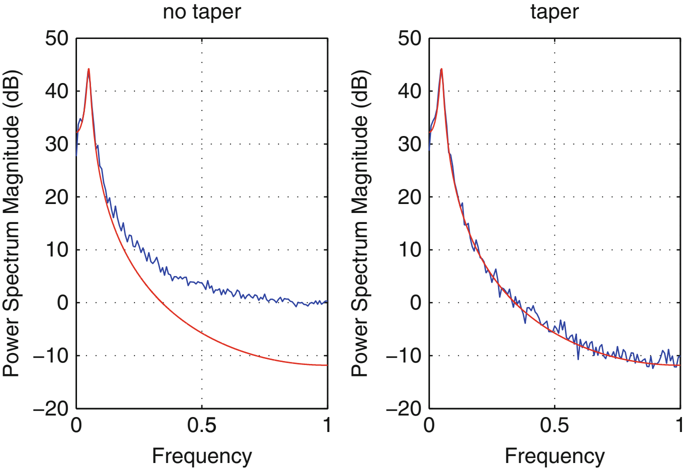


**Fig. 9.14** Estimate of the spectrum of an AR(2) process without a data window (*left*) and with a Hamming window as a taper

so that the location of the maximum of the window is now, e.g. for odd $N$, at $t = (N + 1)/2$. There are no ripples for higher frequencies, the "windowed data" look like the original data fading in and fading out, but the distribution of phases of the Fourier transform of the time series is comparable with a constant distribution, and also the time series of the phases is inconspicuous.

Thus already the Fourier transform and the periodogram of the time series are heavily deformed because of the ugly Fourier transform of the boxcar window. This is another version of the Gibbs phenomenon. Because one cannot avoid any window, one has to use a window with better properties in the frequency domain, e.g., the Hamming window. Then the bias of the periodogram and of the estimate of the spectrum is also reduced (see Fig. 9.14).

The ugly Fourier transform of the boxcar window is also the reason for the sidelobes of the spectrum, which appear in the spectrum, e.g., of a deterministic time series with a definite frequency, if the sampling window does not cover a complete number of cycles. The Fourier transform of the finite time series is the convolution of the Fourier transform of the boxcar window with the Fourier transform of the harmonic wave (one contribution only). The influence of the Fourier transform of the boxcar window on the spectrum then vanishes only if a complete number of cycles fits into the sampling window.

Other possible ways of constructing consistent estimators for the spectral density $\widetilde{C}(\omega)$ are the following:

- The given time series is split into $K$ successive segments, which are now regarded as $K$ realizations of a time series of length $N/K$. For each segment one computes the periodogram and then takes the average over the resulting $K$ periodograms.

The variance so obtained is smaller by a factor of $1/\sqrt{K}$. On the other hand, one obtains the spectral density only for the points

$$\omega_j = \frac{2\pi j}{N\Delta t/K} \,, \quad j = 0, 1, \ldots, \frac{N}{2K} \,. \tag{9.96}$$

The different segments may also be overlapping.

- Further methods include the maximum entropy method (MEM) or MUSIC. Details may be found, e.g., in the manual of MATLAB 5, Signal Processing Toolbox.
- In Sect. 5.9 we met ARMA$(p, q)$ processes. The spectrum of these ARMA processes can be computed explicitly. If a given time series is fitted to an ARMA process, this also determines the spectrum of the time series.

*Remarks.*

- For the variance one always obtains

$$\mathrm{Var}\big(\widehat{\widetilde{C}}(\omega)\big) \propto \widetilde{C}^2(\omega) \,. \tag{9.97}$$

Therefore it is advantageous to study $\ln \widehat{\widetilde{C}}(\omega)$ rather than $\widehat{\widetilde{C}}(\omega)$. The distribution of $\ln \widehat{\widetilde{C}}(\omega)$ is closer to a normal distribution. It can be shown that

$$\mathrm{E}\big(\ln \widehat{\widetilde{C}}(\omega)\big) = \ln \widetilde{C}(\omega) \,, \tag{9.98}$$

and that the variance is given by (Brockwell and Davies 1987)

$$\mathrm{Var}\big(\ln \widehat{\widetilde{C}}(\omega)\big) \approx g^2 \sum_{|k| \leq M} K_M^2(\omega_k) \,. \tag{9.99}$$

where $g = 1$ for no data windowing and

$$g^2 = \frac{q_4}{q_2^2} \tag{9.100}$$

with

$$q_J = \frac{1}{N} \sum_{t=1}^{N} w^J(t) \,. \tag{9.101}$$

for the data window $\{w(t), t = 1, \ldots, N\}$ (Bloomfield 1976).

Hence, the approximate 95% confidence bounds for $\ln \widetilde{C}(\omega_k)$ are given by

$$\ln \widehat{\widetilde{C}}(\omega_k) \pm 1.96 \Big( g^2 \sum_{|k| \leq M} K_M^2(\omega_k) \Big)^{1/2} \,.$$

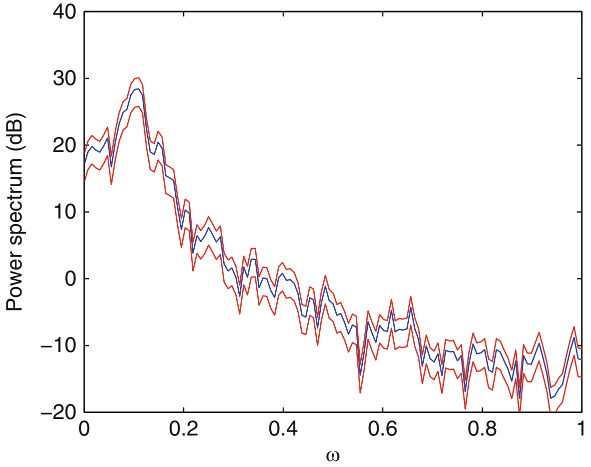


**Fig. 9.15** Estimation of the spectrum of an AR(2) process and the 95% confidence intervals

Recall that for the standard normal random variable the probability for a realization to be in the interval [−1.96, 1.96] is just 0.95.

Figure 9.15 shows the spectrum of an AR(2) process as well as the confidence intervals.

In applications one often uses the logarithm to the base 10 and introduces the quantity

$$n(\omega) = 10 \log_{10} \left( \frac{\widehat{\widehat{C}}(\omega)}{C_0} \right) . \tag{9.102}$$

Here $C_0$ represents some reference quantity such that $n$ is dimensionless. However, in order emphasize that $n$ is given by the logarithm of the ratio of 2 physical quantities with the same dimension it has been given the dimension decibel (dB, a dB is a tenth of a bel, named after A. G. Bell).

- Since the autocovariance function $\widehat{C}(k)$ and the spectrum $\widehat{\widehat{C}}(\omega)$ are connected by a Fourier transformation, $\widehat{C}(k)$ can also be calculated by first transforming the time series $\{Y(t),\ t = 1, \ldots, N\}$ (using the fast algorithm FFT), determining the periodogram, and finally calculating the autocovariance function via an inverse Fourier transformation (FFT) according to

$$\widehat{C}(k) = \frac{1}{\sqrt{2N-1}} \sum_{p=-(N-1)}^{N-1} \widehat{I}_N(\omega'_p) \mathrm{e}^{i\omega'_p k}, \quad k = 0, \pm 1, \ldots, \pm (N-1), \tag{9.103}$$

with $\omega'_p = 2\pi p/(2N-1)$. Here one requires the periodogram at the points $\omega'_p = 2\pi p/(2N-1)$ instead of $\omega_p = 2\pi p/N$, since $\widehat{C}(k)$ is to be defined at the $(2N-1)$ points $k = -(N-1), \ldots, (N-1)$. The time series $\{Y(t),\ t = 1, \ldots, N\}$ must be completed by setting $\{Y(t) = 0 \text{ for } t = N+1, \ldots, 2N-1\}$.

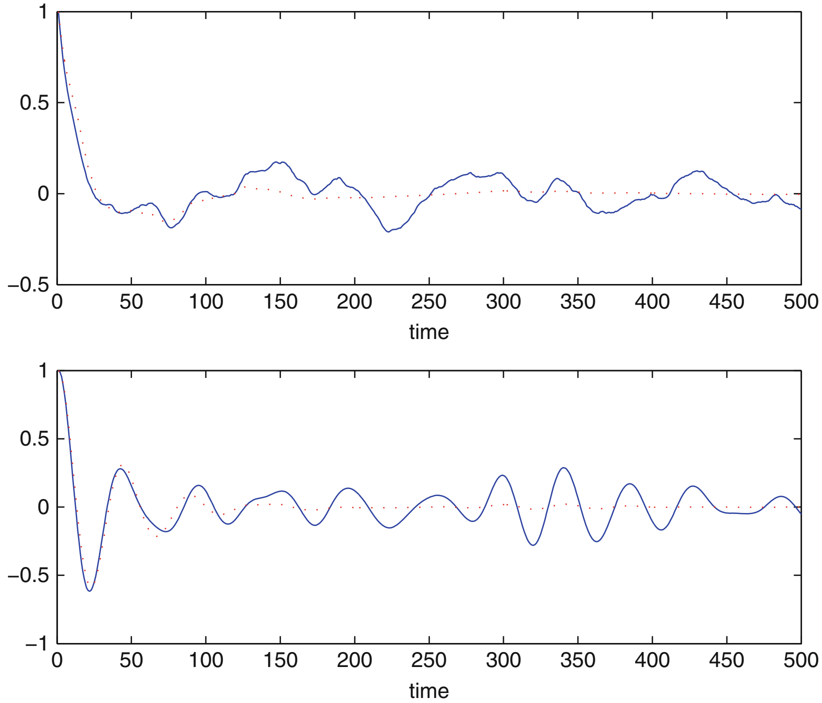


**Fig. 9.16** Estimations of the covariance function of an AR(1) process (*above*) and an AR(2) process (*below*), obtained by the inverse Fourier transform of the periodogram (*solid line*) and by the inverse Fourier transform of the periodogram, smoothed by some spectral window (*broken line*)

This procedure is a particularly fast way to determine the autocovariance function for a long time series. Figure 9.16 shows the covariance function, estimated in this way, for an AR(1) process (above) and an AR(2) process (below) together with their corresponding exact analytical results (dotted lines), which is

$$C(\tau) = \frac{\sigma^2}{1-\alpha_1} \mathrm{e}^{\tau \ln \alpha_1} \tag{9.104}$$

for the AR(1) process $Y(t) = \alpha_1 Y(t-1) + \sigma\eta(t),\ \eta \sim WN(0,1)$, and

$$C(\tau) = \frac{\sigma^2}{1 - 2r^2\cos^2\phi + r^4\cos 2\phi} \mathrm{e}^{\tau \ln r} \cos \tau\phi \tag{9.105}$$

for the AR(2) process $Y(t) = \alpha_1 Y(t-1) + \alpha_2 Y(t-2) + \sigma\eta(t),\ \eta \sim WN(0,1), \alpha_1 = 2r\cos\phi, \alpha_2 = -r^2$, where $T = 2\pi/\phi$ is the period of the stochastic oscillator. The determination of the estimation errors is very cumbersome (see, e.g., Brockwell and Davies 1987), but it is obvious, that they are comparably large when the covariance function is nearly zero.

On the other hand, if we calculate the inverse Fourier transform of the smoothed periodogram, then we obtain an estimate of the covariance function, which is shown as the broken line in Fig. 9.16. The fluctuations for larger values of $\tau$, where the covariance function is actually nearly zero, are now strongly damped. This can be understood easily: Smoothing in the frequency domain by a convolution with a spectral window means multiplication of $\widehat{C}(\tau)$ by a function $\lambda(\tau)$, the Fourier transform of the spectral window, and this multiplication by a factor, which turns out to be less than one, damps the fluctuating estimation errors. Thus we understand that these fluctuations are responsible for the problems of the periodogram and their damping smoothes the periodogram and makes it into a decent estimator for the spectrum.

## 9.5 Cross-Spectral Analysis

Björn Schelter

Filtering is closely related to cross-spectral analysis, which is concerned with the estimation of interactions between different processes. We have characterized the filters by the relation between the input and the output of the filter. In this chapter we briefly generalize this concept to investigate relations between processes in general.

To this end, we utilize the spectral matrix

$$\tilde{\mathbf{C}}(\omega) = \left[\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\right]^{-1} \Sigma \left[\left(\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\right)^{H}\right]^{-1} . \tag{9.106}$$

for an $n$-dimensional vector autoregressive process of order $p$ (cf. Sect. 9.1)

$$\begin{aligned} \boldsymbol{X}(t) &= \sum_{k=1}^{p} \underline{\alpha}_k \boldsymbol{X}(t-k) + \boldsymbol{\eta}(t) \\ \boldsymbol{\eta}(t) &\propto \mathrm{WN}(\mathbf{0}, \Sigma). \end{aligned} \tag{9.107}$$

Hermitean transposition is denoted by $(\cdot)^H$.

This concept can be generalized to arbitrary stationary stochastic processes $\mathbf{Y}(t)$ defining the spectral matrix as

$$\tilde{\mathbf{C}}(\omega) = \langle \tilde{\mathbf{Y}}(\omega) \tilde{\mathbf{Y}}^{H}(\omega) \rangle = \begin{pmatrix} \tilde{C}_{11}(\omega) & \dots & \tilde{C}_{1n}(\omega) \\ \vdots & & \vdots \\ \tilde{C}_{n1}(\omega) & \dots & \tilde{C}_{nn}(\omega) \end{pmatrix} \tag{9.108}$$

with

$$\tilde{\mathbf{Y}}(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{t=-\infty}^{\infty} \mathbf{Y}(\mathbf{t}) \mathrm{e}^{-\mathrm{i}\omega t} = \frac{1}{\sqrt{2\pi}} \sum_{t=-\infty}^{\infty} \begin{pmatrix} Y_1(t) \\ \vdots \\ Y_n(t) \end{pmatrix} \mathrm{e}^{-\mathrm{i}\omega t} , \quad \omega \in [-\pi, \pi] . \tag{9.109}$$

The off-diagonal elements of the spectral matrix $\tilde{C}_{ij}(\omega)$ with $i \neq j$, also called the cross-spectrum , quantify the strength and the type of the interaction between the different processes. The normalized quantity

$$\mathrm{Coh}_{ij}(\omega) = \frac{\left|\tilde{C}_{ij}(\omega)\right|}{\sqrt{\tilde{C}_{ii}(\omega)\tilde{C}_{jj}(\omega)}} \in [0, 1] \ , \ i, j = 1, \ldots, n \ , \ i \neq j \tag{9.110}$$

is called the coherence, while the argument

$$\phi_{ij}(\omega) = \arg(\tilde{\mathbf{C}}_{ij}(\omega)) \ \in [-\pi, \pi) \ , \ i, j = 1, \ldots, n \ , \ i \neq j \tag{9.111}$$

is the phase spectrum between the components $i$ and $j$ of the vector-process $\mathbf{Y}(t)$.

The coherence spectrum quantifies the strength of the interaction. As a normalized value between [0, 1], it assumes a value of 1 if the two components of the process are perfectly linearly related. A value of zero is assumed if there is no linear interaction between the processes.

The phase spectrum provides information about the type of interaction. A pure delay between the two processes $Y_j(t) = Y_i(t - \tau)$ leads to a phase spectrum $\phi(\omega) = \tau\omega$. Filters typically show a nonlinear phase spectrum as shown in the following example.

*Example.* For the coherence between a first order IIR filtered signal $X(t)$ and the input signal Z(t)

$$X(t) = aX(t-1) + Z(t) \tag{9.112}$$

we get

$$\tilde{X}_1(\omega) = a\mathrm{e}^{\mathrm{i}\omega}\tilde{X}_1(\omega) + \tilde{X}_2(\omega) \tag{9.113}$$

$$\tilde{X}_2(\omega) = Z(\omega) \tag{9.114}$$

and we obtain for the spectral matrix

$$\tilde{\mathbf{C}}(\omega) = \begin{pmatrix} \frac{\langle |Z(\omega)|^2 \rangle}{|1 - a\mathrm{e}^{\mathrm{i}\omega}|^2} & \frac{\langle |Z(\omega)|^2 \rangle}{1 - a\mathrm{e}^{\mathrm{i}\omega}} \\ \frac{\langle |Z(\omega)|^2 \rangle}{1 - a\mathrm{e}^{-\mathrm{i}\omega}} & \langle |Z(\omega)|^2 \rangle \end{pmatrix} . \tag{9.115}$$

The coherence between $X(t)$ and $Z(t)$ is equal to one for all frequencies and the phase spectrum reads

$$\phi_{XZ}(\omega) = \arctan\left[\frac{a \sin\omega}{1 - a\cos\omega}\right]. \tag{9.116}$$

*Remarks.*

- If the delay between the processes is zero, the phase spectrum is zero. This is only possible if $\tilde{C}_{ij}(\omega)$ is real-valued. Based on this knowledge, it has been suggested that the cross-spectrum should be separated into a real and an imaginary contribution

  $$\tilde{C}_{ij}(\omega) = \mathcal{R}\left\{\tilde{C}_{ij}(\omega)\right\} + \mathcal{I}\left\{\tilde{C}_{ij}(\omega)\right\} . \tag{9.117}$$

  The real-valued part originates from non-delayed interactions between the two components $i$ and $j$ of the vector-process $\mathbf{Y}(t)$. Example, common noise influences occurring in several applications will be reflected in the real-part. As identical noise contributions onto the two components are not reflecting a true interaction between the processes, it was suggested to analyze only the imaginary part, defining the imaginary coherence (Nolte et al. 2004)

  $$\text{ICoh}_{ij}(\omega) = \frac{\left|\mathcal{I}\left\{\tilde{C}_{ij}(\omega)\right\}\right|}{\sqrt{\tilde{C}_{ii}(\omega)\tilde{C}_{jj}(\omega)}} . \tag{9.118}$$

  This imaginary coherence enables the investigation of interactions that are subject to a nonzero time lag. This is particularly interesting when real-world signals are analyzed, in which one actually expects input/output relations with a certain time lag. Coherence might potentially be influenced by common noise contribution as they occur for instance in cross-talk during data acquisition.
- The spectral matrix consists of the spectra on the diagonal and the cross-spectra as the off-diagonal entries. In Sects. 9.2 and 9.4 we have discussed various ways to estimate the spectrum. For the cross-spectrum between the components $i$ and $j$ of $Y(t)$, the estimation procedure works similar to spectral estimation. Similar to (9.46), we define the cross-periodogram between components $i$ and $j$ by

  $$\tilde{I}_{ij,N}(\omega) = \tilde{Y}_{i,N}(\omega)\tilde{Y}^{H}_{j,N}(\omega) \quad i,j = 1,\ldots,n \tag{9.119}$$

  with $(\cdot)^H$ denoting hermitean transposition. Based on the definition of the cross-periodogram, the consistent estimation works as for the spectra.
- In cross-spectral analysis it is essential that consistent estimators for the spectral matrix are derived. If instead of the cross-spectral and auto-spectral estimates, the cross- and auto-periodograms were used in (9.118), the coherence would always be one.

For an $n$-dimensional autoregressive process

$$\boldsymbol{X}(t) = \sum_{k=1}^{p} \underline{\alpha}_k \boldsymbol{X}(t-k) + \boldsymbol{\eta}(t) \tag{9.120}$$
$$\boldsymbol{\eta}(t) \propto \mathrm{WN}(\mathbf{0}, \Sigma)$$

of order $p$ it is possible to decide whether components $i$ and $j$ interact directly by checking for $\alpha_{ij,k} \neq 0$ for any $k$. Whenever more than two processes are involved and enter the analysis it is, however, an interaction can also only be an indirect one. In such a case the interaction from component $j$ onto $i$ might be mediated by some component $l$, i.e. $\alpha_{il,k}\alpha_{lj,k} \neq 0$. A corresponding autoregressive process could for instance be

$$\begin{aligned} X_1(t) &= \alpha_{12,1} X_2(t-1) + \eta_1(t), \\ X_2(t) &= \alpha_{23,1} X_3(t-1) + \eta_2(t), \\ X_3(t) &= \alpha_{33,1} X_3(t-1) + \eta_3(t). \end{aligned} \tag{9.121}$$

Here $\alpha_{13,1} = \alpha_{31,1} = 0$, but $\alpha_{12,1}\alpha_{23,1} \neq 0$. The influence from process $X_3(t)$ onto $X_1(t)$ is mediated by $X_2(t)$.

An example, which shows a quite different structure of dependence is

$$\begin{aligned} X_1(t) &= \alpha_{11,1} X_1(t-1) + \eta_1(t), \\ X_2(t) &= \alpha_{22,1} X_2(t-1) + \eta_2(t), \\ X_3(t) &= \alpha_{31,1} X_1(t-1) + \alpha_{32,1} X_2(t-1) + \eta_3(t). \end{aligned} \tag{9.122}$$

Both $X_1$ and $X_2$ exert influence on $X_3$ (and on itself) at the next time step, but there is no interaction between $X_1$ and $X_2$; the coherence $X_1(t)$ and $X_2(t)$ is zero. This situation is called ‘marrying parents of a joint child’: the two ‘parents’ $X_1$ and $X_2$ share a common child.

A useful notion in such situations is the so called partial spectral matrix, which for an $n$-dimensional autoregressive process (9.120) reads

$$\tilde{\mathbf{PC}}(\omega) = \big(\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\big)^H \Sigma^{-1} \big(\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\big). \tag{9.123}$$

Here, the coefficient matrices of the vector autoregressive process enter non-inversely. The normalized quantity is the so-called partial coherence

$$\mathrm{PCoh_{ij}}(\omega) = \frac{\left|\tilde{\mathrm{PC}}_{ij}(\omega)\right|}{\sqrt{\tilde{\mathrm{PC}}_{ii}(\omega)\tilde{\mathrm{PC}}_{jj}(\omega)}}, \; i, j = 1, \ldots, n, \; i \neq j. \tag{9.124}$$

The partial coherence simplifies if the covariance matrix of the driving noise

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & 0 & \dots & 0 \\ 0 & \sigma_{22}^2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{nn}^2 \end{pmatrix}$$

is diagonal. This assumption is often fulfilled. We note that it presents only a simplification of the argument and is not crucial for the conclusions drawn next; for details about the general case we refer the interested reader to Dahlhaus (2000).

For a diagonal noise covariance matrix $\Sigma$ we get for the partial coherence

$$\mathrm{PCoh}_{\mathrm{ij}}(\omega) = \frac{\left|\tilde{\mathrm{PC}}_{ij}(\omega)\right|}{\sqrt{\tilde{\mathrm{PC}}_{ii}(\omega)\tilde{\mathrm{PC}}_{jj}(\omega)}}\ , \ i, j = 1, \dots, n\ ,\ i \neq j \tag{9.125}$$

$$= \frac{\left|\sum_{m=1}^{n}(\delta_{mi} - \sum_{k=1}^{p} \alpha_{mi,k} \mathrm{e}^{-\mathrm{i}\omega k})\frac{1}{\sigma_{mm}^2}(\delta_{mj} - \sum_{k=1}^{p} \alpha_{mj,k} \mathrm{e}^{\mathrm{i}\omega k})\right|}{\sqrt{\sum_{m=1}^{n}\left|\delta_{mi} - \sum_{k=1}^{p} \alpha_{mi,k} \mathrm{e}^{\mathrm{i}\omega k})\frac{1}{\sigma_{mm}}\right|^2 \sum_{m=1}^{n}\left|\delta_{mj} - \sum_{k=1}^{p} \alpha_{mj,k} \mathrm{e}^{\mathrm{i}\omega k})\frac{1}{\sigma_{mm}}\right|^2}}$$

$$i, j = 1, \dots, n\ ,\ i \neq j\ .$$

The partial coherence does not vanish if

1. $\alpha_{ij,k}$ or $\alpha_{ji,k}$ are not zero for at least one $k$, i.e. there is a direct interaction between the two processes. In this case also the coherence function does not vanish,
   or
2. $\alpha_{mi,k}$ and $\alpha_{mj,k}$ are both nonzero for at least one $k$ and one $m$, i.e. components $i$ and $j$ exert an influence on a process $m$. This happens in the model which is called 'marrying parents of a joint child effect', but now the coherence function vanishes.

The case in which the interaction is mediated by another process as in (9.121) the partial coherence function vanishes, but the coherence function is nonzero.

Thus by means of a coherence and partial coherence analysis one can discriminate between the three different settings: direct interaction between processes $i$ and $j$ (Coherence and partial coherence are nonzero), 'marrying parents of a joint child effect' (only partial coherence is nonzero). In the case where process $i$ and $j$ interact only mediated by another process (i.e. there is no additional 'marrying' effect) only the coherence is nonzero.

The partial phase spectrum can be defined via the argument

$$P\phi_{ij}(\omega) = \arg(\tilde{\mathrm{PC}}_{ij}(\omega))\ .$$

If no autoregressive representation is available for the processes, partial coherence and partial phase spectrum can be defined for arbitrary processes by using the following definitions of the spectral matrix

$$\tilde{\mathbf{C}}(\omega) = \langle \tilde{\mathbf{Y}}(\omega)\tilde{\mathbf{Y}^H}(\omega)\rangle \,, \tag{9.126}$$

where $\tilde{\mathbf{Y}}(\omega)$ denotes the Fourier transform of $\mathbf{Y}(t)$, and the inverse, the partial spectral matrix is

$$\tilde{\mathbf{PC}}(\omega) = \tilde{\mathbf{C}}^{-1}(\omega)\,. \tag{9.127}$$

Estimation of the partial spectral matrix can numerically efficiently be performed by inverting the spectral matrix for every frequency $\omega$.

We like to note that the coherence as well as the partial coherence are symmetric measures of the strength of interaction, as $\mathrm{Coh}_{ij}(\omega) = \mathrm{Coh}_{ji}(\omega)$ and $\mathrm{PCoh}_{ij}(\omega) = \mathrm{PCoh}_{ji}(\omega)$. The phase spectrum changes its sign, i.e. $\phi_{ij}(\omega) = -\phi_{ji}(\omega)$ and $P\phi_{ij}(\omega) = -P\phi_{ji}(\omega)$. This is in particular remarkable as the coefficient matrices of the autoregressive process enter the definition for the partial spectral matrix (9.123). The not necessarily symmetric coefficient matrices are symmetricized in the equation (9.123), rendering the investigation towards the direction of information flow impossible. This will be discussed further in the Section on Granger causality.

## 9.6 Frequency Distributions for Nonstationary Time Series

In the previous sections we defined the spectrum of a stationary stochastic process $\{Y(t)\}$ and introduced the periodogram as a first, albeit inconsistent, estimator. In Sect. 9.4 we then derived a consistent estimator for the spectrum by filtering the periodogram.

In the case of a nonstationary time series we may proceed as follows: We split the time series into small segments such that the time dependence within each segment can be neglected. For each segment we estimate the spectrum and thereby obtain a first insight into the time dependence of the spectrum. The shorter these segments are, the better fulfilled is the assumption of stationarity within each segment. However, the number of frequency points on which the estimation of the spectrum is based naturally becomes smaller. If the segments are chosen to be strongly overlapping, one obtains a relatively smooth time dependence of the spectrum. If the time-dependent spectrum of a nonstationary time series is determined by this method, one also speaks of a spectrogram. Figure 9.17d shows such a spectrogram. The original signal is a sine curve with an oscillating frequency between 1,000 and 4,000 Hz. Figure 9.17a shows the frequency as a function of time, and in Fig. 9.17b, c two segments from the signal are reproduced.

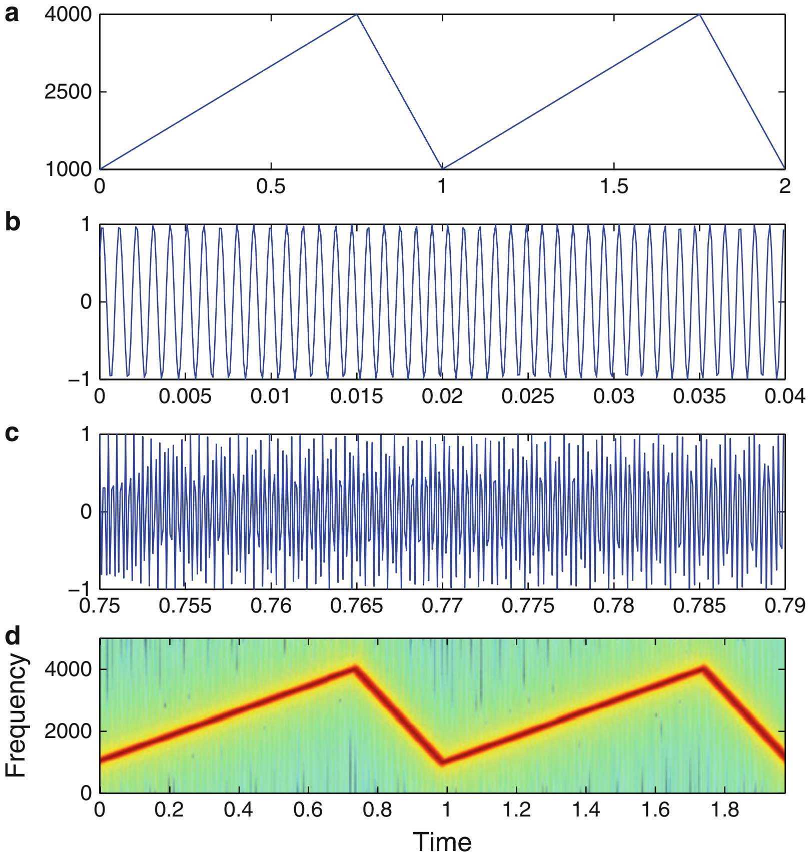


**Fig. 9.17** Frequency of the signal as a function of time (**a**); two segments from the time series (**b, c**); spectrogram (**d**): The frequency dependence is plotted vertically, the intensity is a measure for the height of the spectrum. The maximum of the spectrum varies in time in the same way as the frequency (Even if a white noise with variance 1 is superimposed on the time series, this variation of the maximum will be still evident)

The spectrogram may be considered as a special case of a more general time–frequency distribution. In order to introduce such a generalized time-frequency distribution, we define the covariance for a stochastic process $\{Y(t)\}$ in a symmetrized form by

$$C(t,\tau) = \langle Y(t-\tau)Y(t+\tau)\rangle \ . \tag{9.128}$$

In the stationary case, which we always have in mind as the limiting case, we might replace $t \to t+\tau$ to obtain $C(t,\tau) = \langle Y(t)Y(t+2\tau)\rangle$. The symmetrized definition therefore consists only in a different normalization of $\tau$.

The covariance now also depends on the time $t$. Nevertheless, we again take the Fourier transform with respect to $\tau$:

$$W(t,\omega) = \sum_{\tau=-\infty}^{\infty} C(t,\tau)\,\mathrm{e}^{2\mathrm{i}\omega\tau} . \tag{9.129}$$

The function $W(t,\omega)$ is called the Wigner–Ville spectrum. For a stationary process it is identical, to within a constant factor, to the spectrum $\tilde{C}(\omega)$.

For a (finite) time series $\{Y(t), t = 1,\ldots,N\}$ we may define the following estimator for the Wigner–Ville spectrum:

$$\hat{W}_N(t,\omega) = \sum_{\tau} Y(t-\tau)Y(t+\tau)\mathrm{e}^{2\mathrm{i}\omega\tau} . \tag{9.130}$$

Here $\hat{W}_N(t,\omega)$ is the Wigner–Ville distribution of the time series $\{Y(t), t = 1,\ldots,N\}$. The Wigner–Ville distribution is regarded as the fundamental time–frequency distribution; it corresponds to the periodogram for stationary processes. Similar to the way in which we construct consistent estimators from the periodogram by filtering, where we may choose among various possibilities, here too we can introduce further frequency distributions by a convolution of the Wigner–Ville distribution with some kernel $\tilde{\Phi}(t-t',\omega-\omega')$.

We therefore define the general time-dependent frequency distribution for a stochastic process by

$$\hat{\tilde{C}}_N(t,\omega) = \sum_{t'}\sum_{\omega'} \tilde{\Phi}(t-t',\omega-\omega')\hat{W}_N(t',\omega') . \tag{9.131}$$

A kernel whose effect is simply the averaging of $\hat{W}_N(t,\omega)$ over all instants $t = 1,\ldots,N$ of the time series, as might be advantageous for time series of stationary processes, just leads back to the estimator $\hat{\tilde{C}}_N(\omega)$ for the spectrum $\tilde{C}(\omega)$, i.e., the periodogram, because this may also be written as

$$\hat{\tilde{C}}_N(\omega) = \widehat{I}_N(\omega) = \frac{1}{2\pi}\sum_{t,\tau}\frac{1}{N}Y(t-\tau)Y(t+\tau)\,\mathrm{e}^{2\mathrm{i}\omega\tau} . \tag{9.132}$$

The choice $\tilde{\Phi}(t-t',\omega-\omega') = \frac{1}{N}\delta(\omega-\omega')$ thus yields the periodogram. Taking

$$\tilde{\Phi}(t-t',\omega-\omega') = \frac{1}{2\pi}\frac{1}{N}K_M(\omega-\omega') , \tag{9.133}$$

we can construct a consistent estimator for the spectrum of stationary processes. With

$$\tilde{\Phi}(t-t',\omega-\omega') = w^R(t-t')K_M(\omega-\omega') , \tag{9.134}$$

where $w^R(t-t')$ denotes the window function for the segmentation of the time series into individual sections, we again obtain the spectrogram.

An ansatz for the kernel which has proven to be useful for many practical purposes is

$$\tilde{\Phi}(t-t', \omega-\omega') = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(t-t')^2}{2\sigma_1^2} - \frac{(\omega-\omega')^2}{2\sigma_2^2}\right) , \tag{9.135}$$

where $\sigma_1$ and $\sigma_2$ are two parameters. $\hat{\hat{C}}_N(t,\omega)$ is then also called the smoothed pseudo-Wigner distribution.

We may also take the Fourier transformation of the kernel $\tilde{\Phi}$ or the Wigner–Ville distribution with respect to both arguments. This leads to the so-called ambiguity function $A(\xi,\tau)$, which may also be written as

$$A(\xi,\tau) = \sum_{t=1}^{N} Y(t-\tau)Y(t+\tau)\mathrm{e}^{\mathrm{i}\xi t} . \tag{9.136}$$

In terms of the Fourier transform $\Phi(\xi,\tau)$ of the kernel function $\tilde{\Phi}$ we also have the relation

$$\hat{\hat{C}}_N(t,\omega) = \sum_{\xi}\sum_{\tau} \Phi(\xi,\tau)A(\xi,\tau)\mathrm{e}^{-\mathrm{i}\xi t}\mathrm{e}^{-2\mathrm{i}\tau\omega} . \tag{9.137}$$

For this representation, if $\Phi(\xi,\tau) = N^{-1}\delta(\xi)$, we recover the periodogram; $\Phi \equiv 1$ corresponds to the Wigner distribution.

## 9.7 Filter Banks and Discrete Wavelet Transformations

In Sect. 9.3 we wrote the filtering of a time series $\{X(t)\}$ in the form

$$Y(t) = h(B)X(t) = \sum_k h_k B^{-k} X(t) = \sum_k h_k X(t-k) . \tag{9.138}$$

Here $B$ denotes the shift operator defined by $B^k X(t) = X(t+k)$. There we also discussed various filters, characterized by the function $h(B)$ or $h(z), z \in \mathcal{C}$. For a causal FIR filter, $h(z)$ was a polynomial in $z^{-1}$.

The inverse of such an FIR filter, however, constitutes always an IIR filter, unless $h(z)$ has the trivial form $h(z) = z^{-l}$, corresponding to a simple shift of time $Y(t) = X(t-l)$. A reconstruction of the time series $X(t)$ from the filtered time series $Y(t)$ is therefore only possible, if at all, with an IIR filter. If we insist that the FIR property also applies to the reconstruction of the time series $X(t)$, we have to look at so-called filter banks, i.e., a set of filters which may be organized into a matrix. Indeed, for an $M \times M$-matrix $\tilde{\mathsf{H}}(z)$ whose elements are polynomials in $z$, it is possible to have an inverse matrix $\tilde{\mathsf{H}}^{-1}(z)$ which is also a polynomial in $z$, provided the determinant of $\tilde{\mathsf{H}}(z)$ satisfies

$$\det \tilde{\mathsf{H}}(z) = \text{const}\, z^{-l} \;, \tag{9.139}$$

where $l$ is arbitrary. The reason is that the matrix elements of the inverse of a matrix are given by

$$(\tilde{\mathsf{H}}^{-1})_{ij}(z) = \frac{(j,i)\text{th cofactor of } \tilde{\mathsf{H}}(z)}{\det \tilde{\mathsf{H}}(z)} \;. \tag{9.140}$$

The $(j,i)$th cofactor of $\tilde{\mathsf{H}}(z)$ is a polynomial in $z$ if $\tilde{\mathsf{H}}(z)$ is. As the denominator depends on $z$ only in the form $z^l$, the inverse matrix $\tilde{\mathsf{H}}^{-1}(z)$ is also a polynomial in $z$.

To make use of this property we consider the $z$-transform

$$X(z) = \sum_t X(t) z^{-t} \tag{9.141}$$

of $X(t)$. We do not specify the limits for this summation. For those values of $t$ for which $X(t)$ is not defined by the series we set $X(t) = 0$.

We now introduce the $M$-phase resolution of $X(z)$ as an $M$-dimensional vector $\boldsymbol{X}(z)$ with components

$$X_\gamma(z) = \sum_t X(Mt - \gamma) z^{-t}, \quad \gamma = 0, \ldots M-1 \;. \tag{9.142}$$

Such a vector $\boldsymbol{X}(z)$ with $M$ components $X_\gamma(z), \gamma = 0, \ldots, M-1$ can be constructed from any time series $X(t)$. On the other hand, if this vector $\boldsymbol{X}(z)$ is given, we can recover the $z$-transform of $X(z)$ by

$$X(z) = \sum_{\gamma=0}^{M-1} X_\gamma(z^M) z^\gamma \;. \tag{9.143}$$

For $M = 2$, for example, $\boldsymbol{X}(z) = (X_0(z), X_1(z))$, where

$$X_0(z) = \sum_t X(2t) z^{-t} \quad \text{and} \quad X_1(z) = \sum_t X(2t-1) z^{-t} \;, \tag{9.144}$$

and again

$$X(z) = X_0(z^2) + X_1(z^2) z \tag{9.145}$$

$$= \sum_t X(2t) z^{-2t} + \sum_t X(2t-1) z^{-2t+1} = \sum_t X(t) z^{-t} \;. \tag{9.146}$$

We now consider the following filter on $\boldsymbol{X}(z)$:

$$\boldsymbol{Y}(z) = \tilde{\mathsf{H}}(z) \boldsymbol{X}(z) \;, \tag{9.147}$$

where $\tilde{\mathsf{H}}(z)$ is an $M \times M$ matrix. Hence, if $\tilde{\mathsf{H}}(z)$ is a polynomial in $z$, and if the determinant of $\tilde{\mathsf{H}}(z)$ is of the form (9.139), then the inverse transformation

$$\boldsymbol{X}(z) = \tilde{\mathsf{H}}^{-1}(z)\boldsymbol{Y}(z) \tag{9.148}$$

also represents an FIR filtering. In the following we want to examine for $M = 2$ the filtering with the matrix $\tilde{\mathsf{H}}$ with elements

$$H_{\gamma\alpha}(z) = \sum_{k=0}^{n} c_\gamma(2k+\alpha)z^{-k} \ , \tag{9.149}$$

where the coefficients $c_\gamma(k)$ are suitably chosen.

Before deriving conditions for these coefficients, let us have a closer look at the factor equations (9.147) and (9.148).

**The filter $\tilde{\mathsf{H}}(z)$** For $M = 2$ we obtain, with $H_{\gamma\alpha}(z)$ in the form (9.149),

$$\begin{aligned} Y_\gamma(z) &= \sum_t \left( \sum_{k=0}^{n} c_\gamma(2k)X(2t)z^{-t-k} + \sum_{k=0}^{n} c_\gamma(2k+1)X(2t-1)z^{-t-k} \right) \\ &= \sum_t \left( \sum_{l=0}^{2n+1} c_\gamma(l)X(2t-l) \right) z^{-t} \ . \end{aligned} \tag{9.150}$$

In general we also obtain

$$Y_\gamma(t) = \sum_{l=0}^{2n+1} c_\gamma(l)X(2t-l) = \sum_m c_\gamma(2t-m)X(m) \ . \tag{9.151}$$

The summation limits for $m$ result from the convention that all coefficients $c_\gamma(\ldots)$ which have not been defined before shall be equal to zero.

With the elements $H_{\gamma\alpha}(z)$ of the filter matrix $\tilde{\mathsf{H}}(z)$ we construct the filters

$$C_\gamma(z) = \sum_{\alpha=0}^{M-1} H_{\gamma\alpha}(z^2)z^{-\alpha} \ . \tag{9.152}$$

From (9.149) we obviously have

$$C_\gamma(z) = \sum_{\alpha=0}^{1}\sum_{k=0}^{n} c_\gamma(2k+\alpha)z^{-2k-\alpha} \tag{9.153}$$

$$= \sum_{k'=0}^{2n+1} c_\gamma(k')z^{-k'}, \quad \gamma = 0, 1 \ . \tag{9.154}$$

It turns out that each component $Y_\gamma(z)$ of $\boldsymbol{Y}(z)$ may also be obtained by the following operation: Filtering of $X(t)$ with $C_\gamma(z)$, i.e.,

$$Y'_\gamma(z) = C_\gamma(z)X(z) \ , \tag{9.155}$$

followed by a so-called 'down-sampling'. Here, down-sampling of a time series means that only the values at the instants $t = 0 \bmod M$ are taken into account. For $M = 2$, for example, this means that we only select the values for even times $t$ from the series $Y'_\gamma(t)$ to obtain the series $Y_\gamma(t)$. Hence, the filtering of a time series $X(t)$ of length $N$ leads to two time series $(Y_0(t), Y_1(t))$, each having length $N/2$ ($N$ shall be even). We thus get

$$Y'_\gamma(z) = \sum_{k'} c_\gamma(k')z^{-k'} \sum_t X(t)z^{-t} \tag{9.156}$$

$$= \sum_t \sum_{k'} c_\gamma(k')X(t-k')z^{-t} \ , \tag{9.157}$$

and 'down-sampling' ($t \to 2t$) leads to $Y_\gamma(z)$ or $Y_\gamma(t)$. Analogously, for general $M$ we obtain $M$ time series of length $N/M$, with possible boundary effects if $N$ is not divisible by $M$.

**The inverse filter $\tilde{\mathsf{H}}^{-1}(z)$** The vector $\boldsymbol{X}(z)$ may be represented as a sum of $M$ contributions $X^{(\kappa)}(z)$. Indeed, each of these contributions results if we only take into account one component $Y_\kappa(z)$ in (9.148). For a given value of $\kappa, \kappa = 0, \ldots, M-1$ we therefore set

$$X_\gamma^{(\kappa)}(z) = (\tilde{\mathsf{H}}^{-1})_{\gamma\kappa}(z)Y_\kappa(z) \ , \quad \gamma = 0, \ldots, M-1 \ . \tag{9.158}$$

This defines $M$ contributions to $X_\gamma(z)$. According to (9.143) we construct from these

$$X^{(\kappa)}(z) = \sum_{\gamma=0}^{M-1} X_\gamma^{(\kappa)}(z^M)z^\gamma \ , \tag{9.159}$$

and thereby also find a splitting of $X(z)$ into $M$ contributions $\{X^{(\kappa)}(z), \kappa = 0, \ldots, M-1\}$:

$$X(z) = \sum_{\kappa=0}^{M-1} X^{(\kappa)}(z) \ . \tag{9.160}$$

If we write $(\tilde{\mathsf{H}}^{-1})_{\gamma\kappa}(z)$ in the form

$$(\tilde{\mathsf{H}}^{-1})_{\gamma\kappa}(z) = \sum_{k=0}^{n} d_\kappa(2k+\gamma)z^k \ , \tag{9.161}$$

we get

$$X^{(\kappa)}(z) = \sum_k \left( d_\kappa(2k) z^{2k} + d_\kappa(2k+1) z^{2k+1} \right) Y_\kappa(z^2) \tag{9.162}$$

$$= \sum_{l=0}^{2n+1} d_\kappa(l) z^l \sum_t Y_\kappa(t) z^{-2t} \tag{9.163}$$

$$= \sum_m \left( \sum_t d_\kappa(2t-m) Y_\kappa(t) \right) z^{-m} \tag{9.164}$$

$$\equiv \sum_m X^{(\kappa)}(m) z^{-m} \ . \tag{9.165}$$

Hence, in our example $M = 2$, the time series $X(m)$ may be reconstructed from the time series $Y_0(m), Y_1(m)$ by

$$X(m) = \sum_t \left( d_0(2t-m) Y_0(t) + d_1(2t-m) Y_1(t) \right) \ . \tag{9.166}$$

Equations 9.160 and 9.166 represent a segmentation of the time series $X(t)$ which will give rise to various approximations as well as methods for its investigation. This will be discussed after we have studied the possible form of the matrix $\tilde{\mathsf{H}}(z)$.

From the representation (9.161) for $(\tilde{\mathsf{H}}^{-1})_{\gamma\kappa}(z)$ we obtain the reconstruction filter in the form

$$D_\kappa(z) = \sum_{l=0}^{2n+1} d_\kappa(l) z^l , \quad \kappa = 0, 1 , \tag{9.167}$$

such that now $X^\kappa(z) = D_\kappa(z) Y_\kappa(z^2)$ (cf. (9.165)). In this form the filter $D_\kappa(z)$ is acausal, since it is a polynomial in $z$, not in $1/z$. However, if we write $D_\kappa(z) = z^{2n+1} D'_\kappa(z)$, with

$$D'_\kappa(z) = \sum_{l=0}^{2n+1} d_\kappa(l) z^{l-2n-1} = \sum_{m=0}^{2n+1} d_\kappa(2n+1-m) z^{-m} \ , \tag{9.168}$$

the filter $D_\kappa(z)$ has been factorized into a causal filter $D'_\kappa(z)$ and a second filter $z^{2n+1}$, which only causes a time shift. Apart from the latter, a reconstruction filter therefore consists of a causal filter, characterized by the coefficients $\{d_\kappa(2n+1-m)\}$.

This completes the discussion of the filters $\tilde{\mathsf{H}}(z)$ and $\tilde{\mathsf{H}}^{-1}(z)$.

**Two explicit classes of filter banks** The condition that the filter bank $\tilde{\mathsf{H}}(z)$ has a determinant which is a monomial of the form $cz^l$ still admits many possibilities. We will now consider two explicit classes of filter banks for the case $M = 2$.

Suppose, first, that $\tilde{\mathsf{H}}(z)$ be of the form

$$\tilde{\mathsf{H}}(z) = A_0 S_l \Lambda(z) S_{l-1} \Lambda(z) \ldots S_1 \Lambda(z) \ldots S_0 \ , \tag{9.169}$$

with

$$A_0 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \ , \quad \Lambda(z) = \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \tag{9.170}$$

and

$$S_j = \begin{pmatrix} a_j & b_j \\ b_j & a_j \end{pmatrix} \ , \tag{9.171}$$

for arbitrary parameters $a_j, b_j$. ($S_j$ does not depend on $z$.)

Evidently, the determinant of such a matrix has the required form $c\, z^l$. We will see that the symmetry of the matrices $S_j$ leads to a symmetry of the filters $C_\gamma(z)$, so that they have a linear phase. Furthermore, the length of these filters for different $\gamma$ is equal by construction. The matrix $A_0$ to the left of $S_l$ in (9.169) is a matter of convention.

Secondly, let $\tilde{\mathsf{H}}(z)$ be of the form

$$\tilde{\mathsf{H}}(z) = \Lambda(-1) R_l \Lambda(z) R_{l-1} \Lambda(z) \ldots R_1 \Lambda(z) \ldots R_0 \ , \tag{9.172}$$

with

$$\Lambda(z) = \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \quad \text{and} \quad R_j = \begin{pmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{pmatrix} . \tag{9.173}$$

The matrix $\Lambda(-1)$ to the left of $R_l$ in (9.172) is again convention. The matrices $R_j$ are obviously orthonormal, and therefore

$$\tilde{\mathsf{H}}^{-1}(z) = \tilde{\mathsf{H}}^T(z^{-1}) \ . \tag{9.174}$$

From the representations (9.149) and (9.161) we obtain in this case for the coefficients of the filters $C_\gamma$ and $D_\gamma$:

$$c_\gamma(m) = d_\gamma(m) \ . \tag{9.175}$$

### 9.7.1 Examples of Filter Banks

1. Let $M = 2$ and

$$\tilde{\mathsf{H}}(z) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} . \tag{9.176}$$

This is the only filter bank with a linear phase which is also orthogonal. We have

$$Y_0(z) = \sum_{t=0}^{N/2} \Big( X(2t) + X(2t-1) \Big) z^{-t} \ ,$$

$$Y_1(z) = \sum_{t=0}^{N/2} \Big( X(2t) - X(2t-1) \Big) z^{-t} \ ,$$

and therefore

$$C_0(z) = 1 + z^{-1}, \quad \text{and} \quad C_1(z) = 1 - z^{-1} \ . \tag{9.177}$$

$C_0$ is a low-pass filter, $C_1$ a high-pass filter. This filter bank is also known as a Haar filter bank (after A. Haar, who first described it in his thesis of 1909).

The inverse matrix of $\tilde{\mathsf{H}}(z)$ is easily written down. Reconstruction of $X(z)$, taking into account only the contribution $Y_0(z)$, leads to

$$X_0^{(0)}(z) = \frac{1}{2} Y_0(z)$$

$$X_1^{(0)}(z) = \frac{1}{2} Y_0(z) \ ,$$

and thus

$$X^{(0)}(z) = X_0^{(0)}(z^2) + X_1^{(0)}(z^2) z = \frac{1}{2}(1+z) Y_0(z^2) \ ,$$

and for $X_k^{(1)}(z)$

$$X_0^{(1)}(z) = \frac{1}{2} Y_1(z)$$

$$X_1^{(1)}(z) = -\frac{1}{2} Y_1(z) \ ,$$

i.e.,

$$X^{(1)}(z) = X_0^{(1)}(z^2) + X_1^{(1)}(z^2) z = \frac{1}{2}(1-z) Y_1(z^2) \ .$$

We obtain a decomposition of $X(t)$ into a low-frequency and a high-frequency part (Fig. 9.18).

2. Let $M = 2$ and

$$\tilde{\mathsf{H}}(z) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos\,\theta_1 & \sin\theta_1 \\ -\sin\theta_1 & \cos\theta_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \begin{pmatrix} \cos\theta_2 & \sin\theta_2 \\ -\sin\theta_2 & \cos\theta_2 \end{pmatrix} \ . \tag{9.178}$$

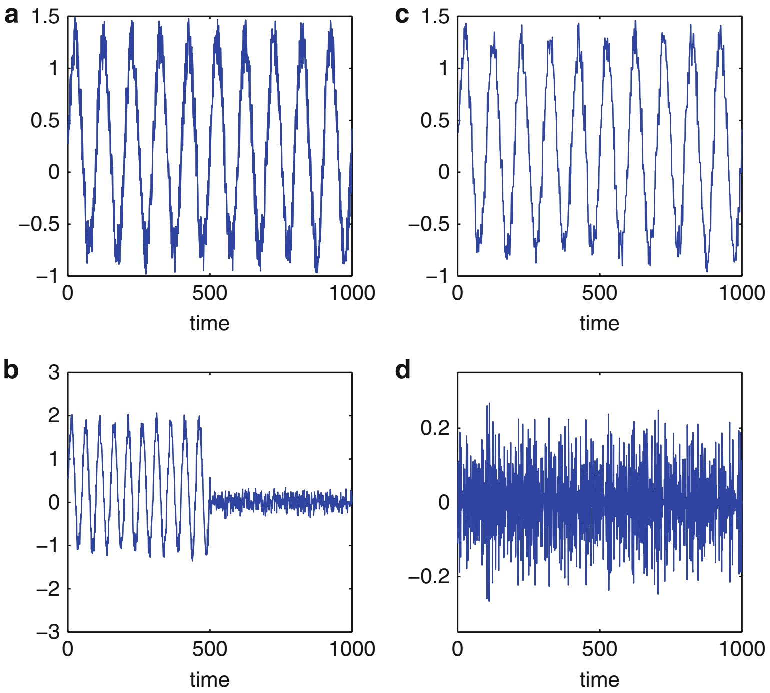


**Fig. 9.18** Decomposition of a time series $X(t)$, represented in (**a**), into two filtered parts $Y_0(t)$ and $Y_1(t)$, each having half the length (**b**), as well as a low-frequency part $X^{(0)}$ in (**c**) and a high-frequency part $X^{(1)}$ in (**d**)

Comparison with the form

$$\tilde{\mathsf{H}}(z) = \begin{pmatrix} c_0(0) + c_0(2)/z & c_0(1) + c_0(3)/z \\ c_1(0) + c_1(2)/z & c_1(1) + c_1(3)/z \end{pmatrix} \tag{9.179}$$

leads to the coefficients

$$\begin{aligned} c_0(0) &= \cos\theta_1 \cos\theta_2 \,, & c_0(1) &= \cos\theta_1 \sin\theta_2 \\ c_0(2) &= -\sin\theta_1 \sin\theta_2 \,, & c_0(3) &= \sin\theta_1 \cos\theta_2 \end{aligned}$$

and $c_1(k) = (-1)^k c_0(3-k)$. As a consequence of orthogonality we obtain

$$\sum_{k=0}^{3} c_0^2(k) = 1; \quad \sum_{k=0}^{3} c_1^2(k) = 1; \quad \sum_{k=0}^{3} c_0(k) c_1(k) = 0 \,. \tag{9.180}$$

3. Let $M = 2$ and

$$\tilde{\mathsf{H}}(z) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} a & b \\ b & a \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \begin{pmatrix} a' & b' \\ b' & a' \end{pmatrix} \tag{9.181}$$

$$= \begin{pmatrix} (a+b)(a'+b'/z) & (a+b)(b'+a'/z) \\ (a-b)(a'-b'/z) & (a-b)(b'-a'/z) \end{pmatrix} . \tag{9.182}$$

Now

$$C_0(z) = (a+b)a' + (a+b)b'z^{-1} + (a+b)b'z^{-2} + (a+b)a'z^{-3},$$
$$C_1(z) = (a-b)a' - (a-b)b'z^{-1} + (a-b)b'z^{-2} - (a-b)a'z^{-3} .$$

While

$$C_0(z) \equiv c_0(0) + c_0(1)z^{-1} + c_0(2)z^{-2} + c_0(3)z^{-3}$$

satisfies the symmetry relation $c_0(k) = c_0(3-k), k = 0, 1$, we find for

$$C_1(z) \equiv c_1(0) + c_1(1)z^{-1} + c_1(2)z^{-2} + c_1(3)z^{-3}$$

the relation $c_1(k) = -c_1(3-k), k = 0, 1$.

Filter banks which are frequently used or intensively studied often carry the names of their inventors. Thus one finds Daubechies filter banks of various order (after I. Daubechies), several Coiflets filter banks (after R. Coifman) and Meyer filter banks (after Y. Meyer). Other filter banks are named ‘symlets’ or ‘morlets’. Since these filter banks play an important role for the discrete wavelet transformations, they are sometimes also referred to as wavelets. For a more detailed discussion see e.g. Strang and Nguyen (1996) or the Matlab Wavelet Toolbox (Misiti et al. 1996).

Having decomposed a time series $X(t)$ into $M$ components $Y_\kappa(t)$, or $X^{(\kappa)}$, $\kappa = 0, \ldots, M-1$, each of these components may now of course be decomposed again by the same procedure. This may lead finally to an entire cascade of decompositions of the time series $X(t)$ into different types of summands. A particularly useful and clear decomposition for $M = 2$ is the following: After each decomposition the so-obtained low-frequency part is decomposed again. For instance, if we write

$$X(t) = X^{(0)}(t) + X^{(1)}(t) \equiv a1(t) + d1(t) , \tag{9.183}$$

where $a1(t)$ shall denote the low-frequency part $X^{(0)}(t)$, then $a1(t)$ is decomposed further

$$a1(t) = a1^{(0)}(t) + a1^{(1)}(t) \equiv a2(t) + d2(t) , \tag{9.184}$$

and, eventually,

$$a2(t) = a3(t) + d3(t) . \tag{9.185}$$

Hence, in this case $X(t)$ is a sum of the form

$$X(t) = a3(t) + d3(t) + d2(t) + d1(t) \; . \tag{9.186}$$

Each level yields a new high-frequency part such that in the end $X(t)$ is represented as a sum of all these high-frequency contributions $d1, d2, \ldots$, which are also referred to as details, plus the remaining low-frequency contribution from the last decomposition, i.e., in the above example, $a3(t)$. The individual low-frequency parts $a1, a2, \ldots$ are also called approximations. In Fig. 9.19 such a decomposition is shown for the case of five successive levels.

This type of decomposition is called a discrete wavelet transformation. It can have various advantages:

1. A clever choice for the filter bank and the degree of decomposition may split the time series $X(t)$ into different frequency contributions in such a way that essential characteristics become evident (cf. Fig. 9.20).
2. Those contributions in the high-frequency parts, like $X^{(1)}(t)$, $X^{(01)}(t)$, which are smaller than some given threshold may be set to zero. This has two effects:
   - A smoothing of $X(t)$. This strategy is used for the denoising of signals as illustrated in Fig. 9.21.
   - A compression of the signal. If $X(t)$ in (9.186) consists of $1,024$ data points ($t = 1, \ldots, 1,024$), then $X^{(0)}(t)$ may be determined from 512, $X^{(00)}(t)$ from 256, and $X^{(000)}(t)$ from 128 data points. (One should recall that $Y_0(t)$, from which e.g. $X^{(0)}(t)$ can be determined, contains only half the number of data points.) If none of the high-frequency contributions were taken into account, we would have achieved a compression by a factor of $2^3$.

## 9.8 Wavelets

The filter banks presented in the previous section permit the decomposition of a time series into various contributions which reflect the temporal correlations in different frequency bands.

We will now develop such a 'multiresolution' of a process from a completely different point of view. We will find, however, that this leads back to filter banks again.

### *9.8.1 Wavelets as Base Functions in Function Spaces*

Let $\mathcal{L}^2$ be the set of real-valued functions $\{f(t)\}$ which are square integrable with respect to the scalar product

$$\langle f, g \rangle = \int \mathrm{d}t \; f(t) g(t) \; . \tag{9.187}$$

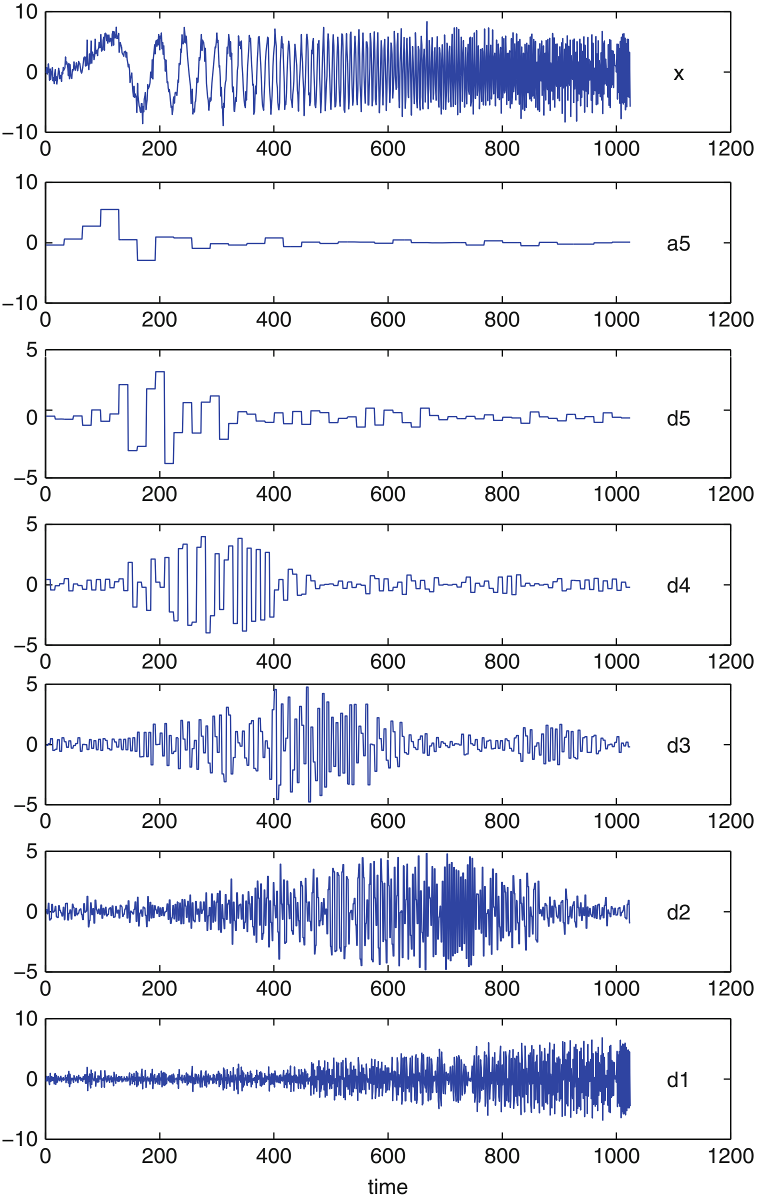


**Fig. 9.19** Decomposition of a time series $X(t)$ (*top plot*) in five levels such that the signal is represented as the sum of the approximation $a5(t)$ of the fifth level (*second plot*) and the details $d1, \ldots, d5$ of the levels 1 to 5 (*third* to *seventh plot*). The different details represent the signal in different frequency bands. The approximation $a5(t)$ corresponds to an extreme smoothing of the signal

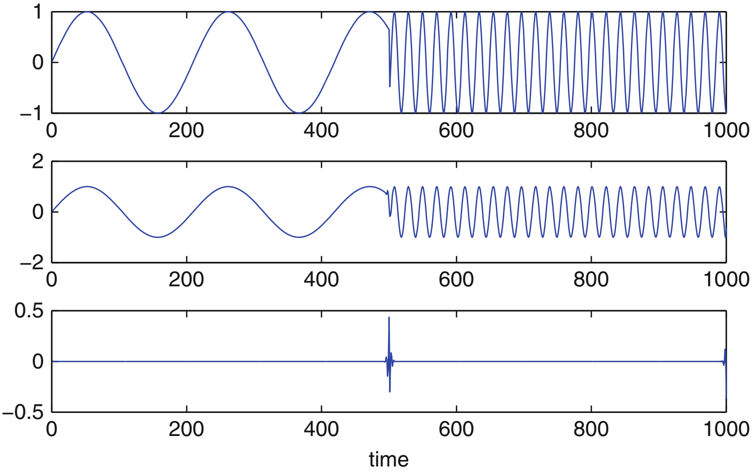


**Fig. 9.20** Signal with a sudden frequency change (*top*), its first approximation (*center*), and its first detail (*bottom*), determined with a Daubechies filter of order one. A strong evidence of this change is seen in the first detail function

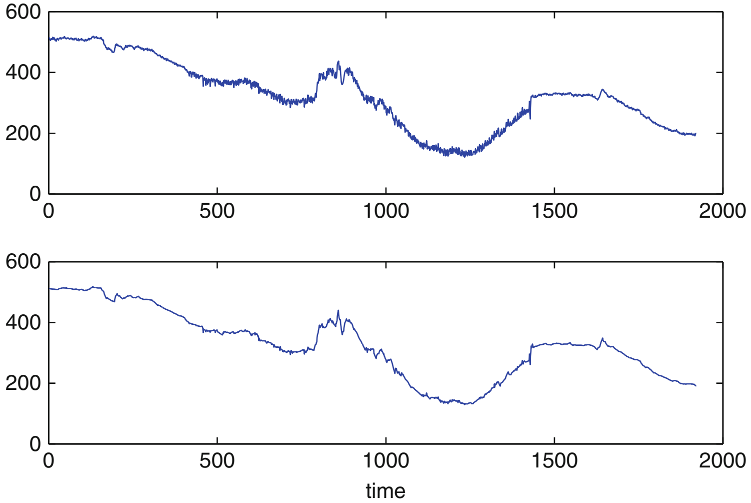


**Fig. 9.21** Signal (*top*) and denoised signal (*bottom*), smoothed by a Daubechies filter of order five with a five-level decomposition (Example taken from Matlab5 Wavelet Toolbox)

We consider a sequence of subspaces $V_j$, $j = 0, 1, \ldots$ of $\mathcal{L}^2$ for which

$$V_0 \subset V_1 \subset \ldots V_j \subset V_{j+1} \subset \ldots \, . \tag{9.188}$$

The component of a function $f(t)$ in $V_j$, which will be denoted by $f_j(t)$, will in general only be an approximation. For $j \to \infty$ the function spaces $\{V_j\}$ converge to $\mathcal{L}^2$, and therefore $f_j(t)$ will converge to $f(t)$. Let us now look at the spaces $W_j, j = 0, \ldots$, which are the respective augmentations of the function spaces and are defined by

$$V_1 = V_0 \oplus W_0, \quad \text{and in general} \quad V_{j+1} = V_j \oplus W_j \ . \tag{9.189}$$

We will write $w_j(t)$ to denote the contribution of the function $f(t)$ in $W_j$, $j = 0, 1, \ldots$, and $\phi_0(t)$ will be the contribution in $V_0$. Then $f(t)$ may be written as

$$f(t) = \phi_0(t) + w_0(t) + w_1(t) + \ldots + w_j(t) + \ldots \ . \tag{9.190}$$

This represents a multiresolution.

Let us now investigate how to choose the sequence of spaces $\{V_j\}$ suitably.

If $g(t)$ is a function in $V_j$, then we require that

1. The functions $g(t-k), k = \pm 1, \pm 2, \ldots$ are also functions in $V_j$ (and therefore also in $V_{j+1}$),
2. The functions $g(2t-k), k = 0, \pm 1, \pm 2, \ldots$ are functions in $V_{j+1}$,
3. There exists a function $\phi(t)$ such that $\{\phi_k(t) \equiv \phi(t-k), k = 0, \pm 1, \ldots\}$ is a complete and orthonomal basis in $V_0$.

We will always deal with functions which are different from zero only in a finite interval. For the function $g(2t-k)$ this interval is shifted by $k$ and smaller by a factor of 2 than the interval of $g(t)$. We say that the scale of $g(2t)$ is smaller by a factor of 2.

We will show shortly that these requirements can be fulfilled by giving an explicit example. First, however, we want to elaborate some consequences of these assumptions:

*The base functions in $V_j, j = 1, 2, \ldots$, and the scaling function* If $\{\phi_k(t) = \phi(t-k), k = 0, \pm 1, \ldots\}$ is a basis of $V_0$, then

$$\{\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k), \ k = 0, \pm 1, \ldots\} \tag{9.191}$$

is a basis in $V_j$. By assumption we know that $\phi(t)$ in $V_0$ implies $\phi(2t-k)$ in $V_1$, etc., i.e., $\phi(2^j t - k)$ in $V_j$ for all $k$. Since $\{\phi(t-k)\}$ represents a basis in $V_0$, $\{\phi(2t-k)\}$ represents a basis in $V_1$. Normalization leads to the prefactor.

Hence, the function $\phi(t)$ determines all base functions. It is therefore the central function for the multiresolution and is also referred to as scaling function.

Since $V_0 \subset V_1$, the function $\phi(t)$ can be expanded with respect to the basis of $V_1$, i.e.,

$$\phi(t) = \sqrt{2} \sum_k h(k)\phi(2t-k) \tag{9.192}$$

with suitable coefficients $\{h(k)\}$. This equation is called the dilation equation. Its solution determines the scaling function and thus the complete base of the multiresolution. The functions $\{\phi(2t - k)\}$ constitute an orthonormal basis, i.e., they satisfy

$$\int \mathrm{d}t\ \phi(t)\phi(t-m) = \delta_{m0}\ . \tag{9.193}$$

If we replace both factors in this integrand by the representation (9.192), we can derive the following conditions for the coefficients $\{h(k)\}$:

$$\sum_k h(k)h(k-2m) = \delta_{m0}\ . \tag{9.194}$$

These conditions for the coefficients $\{h(k)\}$ will later be identified with the conditions for the coefficients of a filter bank.

*The base functions in* $W_j$, $j = 1, 2, \ldots$, *and the wavelet* We will construct a basis in $W_0$. Any base function $w(t)$ in $W_0$ has an expansion with respect to the base functions in $V_1$:

$$w(t) = \sqrt{2}\sum_k g(k)\phi(2t-k)\ , \tag{9.195}$$

and since $w(t)$ is orthogonal to the base functions of $V_0$ we obtain for the coefficients $\{g(k)\}$ the relations

$$\int \mathrm{d}t\, w(t)\phi(t-m) \equiv \sum_k h(k)g(k-2m) = 0\ . \tag{9.196}$$

Furthermore, we want the functions $\{w(t-k), k = 0, \pm 1, \ldots\}$ to form an orthonormal basis, which yields in addition

$$\int \mathrm{d}t\, w(t)w(t-m) = \sum_k g(k)g(k-2m) = \delta_{m0}\ . \tag{9.197}$$

If these conditions are satisfied, we can also state the following: The functions $\{w_{jk}, k = 0, \pm 1, \ldots\}$, defined by

$$w_{jk} = 2^{j/2}\, w(2^j t - k), \quad j = 0, 1, \ldots, \quad k = 0, \pm 1, \ldots\ , \tag{9.198}$$

form, for any given $j$ $(j = 1, \ldots)$, a basis in $W_j$, and together with the base functions of $V_0$ they form an orthonormal basis of the entire space $\mathcal{L}^2$. The value of $j$ is called the scale of the base function $w_{jk}$ of $W_j$.

The function $w(t)$ is named 'wavelet'. It determines this orthogonal basis, and it itself is determined by the scale function $\phi(t)$ and the coefficients $\{g(k)\}$, satisfying the conditions (9.194), (9.196) and (9.197).

Thus we have defined a framework where we can construct a complete orthonormal basis in the function space $\mathcal{L}^2$ whose elements may be transformed into each other purely by shifts and dilations. The expansion of an arbitrary function with respect to this basis in the form

$$f(t) = \sum_k f_{0k}\phi(t-k) + \sum_{j,k} g_{jk} w_{jk}(t) \tag{9.199}$$

therefore describes a decomposition into contributions which describe the function on different scales.

The wavelet components $\{f_{0k}, g_{jk}\}$ of a function $f(t)$ are to be determined from

$$f_{0k} = \int \mathrm{d}t\ f(t)\phi(t-k), \quad g_{jk} = \int \mathrm{d}t\ f(t) w_{jk}(t)\ . \tag{9.200}$$

In some sense the wavelet coefficients measure the agreement of the function $f(t)$ with the scale function $\phi(t-k)$ or the wavelet $w_{jk}(t)$ for all scales $j$ and for all points $k$.

In order to complete this framework we still have to show:

- How to determine the coefficients $\{h(k), g(k)\}$. We shall see that each orthogonal filter bank provides such a set of coefficients, i.e., each orthogonal filter bank is associated with a wavelet and thus a special decomposition of the function space $\mathcal{L}^2$;
- How to solve the dilation equation and determine the scale function and, therefore, also the original wavelet $w(t)$;
- How the coefficients $\{f_{0m}, g_{jk}\}$ of a given function $f(t)$ are obtained most efficiently, e.g. by some recursive method.

We shall address the last question first. It will turn out that this leads us back to the filter banks, and the problem of determining the coefficients, $\{h(k), g(k)\}$, will solve itself in the process.

### 9.8.2 Wavelets and Filter Banks

If we take $\phi_{1m}(t) = \sqrt{2}\phi(2t-m)$ as base functions in $V_1$ we obtain from the dilation equation (9.192) for $\phi(t-k)$

$$\begin{aligned}\phi(t-k) &= \sqrt{2}\sum_m h(m)\phi(2t-2k-m) \\ &= \sum_m h(m-2k)\phi_{1m}(t)\ .\end{aligned} \tag{9.201}$$

Similarly from the wavelet equation (9.195) we get

$$\begin{aligned} w(t-k) &= \sqrt{2}\sum_m g(m)\phi(2t-2k-m) \\ &= \sum_m g(m-2k)\phi_{1m}(t) \; . \end{aligned} \tag{9.202}$$

Multiplying both equations by $f_1(t) \in V_1$, and noticing that $\phi(t-k)$ and $w(t-k)$ are base elements in $V_0$ and $W_0$, respectively, we obtain

$$f_{0k} \equiv \int f_1(t)\phi(t-k) = \sum_m h(m-2k)f_{1m} \tag{9.203a}$$

$$g_{0k} \equiv \int f_1(t)w(t-k) = \sum_m g(m-2k)f_{1m} \; . \tag{9.203b}$$

From these equations we can determine the lowest coefficients $\{f_{0m}, g_{0m}\}$ from $\{f_{1m}\}$. In an analogous way we obtain the recursion formulas

$$\begin{aligned} f_{jk} &= \sum_m h(m-2k)f_{(j+1)m} \\ g_{jk} &= \sum_m g(m-2k)f_{(j+1)m} \end{aligned} \tag{9.204}$$

as well as the inverse relation:

$$f_{(j+1)k} = \sum_l h(k-2l)f_{jl} + \sum_l g(k-2l)g_{jl} \; . \tag{9.205}$$

Here, $\{f_{0k}, g_{0k}, g_{1k}, \ldots, g_{jk}\}$ are the wavelet components needed to represent a function $f_{j+1}(t) \in V_{j+1}$ according to (9.199).

Let us compare (9.204) and (9.205) with the respective (9.151) and (9.166), which we write down again for convenience:

$$Y_\gamma(t) = \sum_m c_\gamma(2t-m)X(m) \; , \tag{9.206}$$

$$X(m) = \sum_t \Big(d_0(2t-m)Y_0(t) + d_1(2t-m)Y_1(t)\Big) \; . \tag{9.207}$$

Taking

$$h(2n+1-m) = d_0(m) = c_0(m), \quad m = 0, \ldots, 2n+1 \; , \tag{9.208a}$$

$$g(2n+1-m) = d_1(m) = c_1(m), \quad m = 0, \ldots, 2n+1 \; , \tag{9.208b}$$

where $2n + 1$ is the length of the filter $\{C_\gamma, D_\gamma\}$ (cf. (9.154)), the equations (9.204) for $j = 0$ agree with (9.206). The same holds for (9.205) and (9.207). In this case, $X(m)$ corresponds to the wavelet coefficient $f_{1m}$, $Y_0(m)$ to the coefficient $f_{0m}$, and $Y_1(m)$ to the coefficient $g_{0m}$.

The set of coefficients $\{h(k), g(k)\}$ therefore corresponds to the coefficients of the reconstruction filter (cf. (9.168)). For this choice of $\{h(k), g(k)\}$, the conditions (9.194), (9.196), and (9.197) become conditions for $\{c_0(k), c_1(k)\}$ which according to (9.180) are satisfied for an orthogonal filter bank.

In this way we may derive from a set of filters $\{c_0(k), c_1(k)\}$ belonging to an orthogonal filter bank a set of coefficients $\{h(k), g(k)\}$ for the definition of a scale function and a wavelet. For the wavelet components one finds exactly the same relations (9.204) as hold among the low-frequency part of a certain decomposition level and the contributions from the next larger scale (next smaller $j$).

Proceeding from the expansion of a function $f(t)$

$$f(t) = \sum_{k=1}^{N} X(k)\phi(t-k) \, , \tag{9.209}$$

we may study not only function spaces with base functions on smaller scales, but also ones with larger scales. We now consider the decompositions $V_0 = V_{-1} \oplus W_{-1}$, $V_{-1} = V_{-2} \oplus W_{-2}$, etc., and

$$w_{jk} = 2^{-j/2} w(2^{-j}t - k), \quad j = 0, 1, \ldots, \quad k = 0, \pm 1, \ldots \tag{9.210}$$

are the base functions on these larger scales. It is this sequence of decompositions which we need here, because if we use (9.205) to write

$$f(t) \equiv \sum_{k=1}^{N} f_{0k}\phi(t-k) \tag{9.211}$$

$$= \sum_{k=1}^{N} \left( \sum_{l} h(2l-k) f_{-1l} + \sum_{l} g(2l-k) g_{-1l} \right) \phi(t-k) \, , \tag{9.212}$$

we finally obtain with the aid of the dilation equation and the wavelet equation

$$f(t) = \sum_{l=1}^{N} f_{-1l} \frac{1}{\sqrt{2}} \phi(2^{-1}t - l) + \sum_{l=1}^{N} g_{-1k} \frac{1}{\sqrt{2}} w(2^{-1}t - k) \, , \tag{9.213}$$

i.e., the corresponding decomposition into $V_{-1} \oplus W_{-1} (= V_0)$.

We should add two important remarks:

- The time series $X(t), t = 1, \ldots, N$ does not correspond to the values of $f(t)$ at fixed, integer arguments $t$, but to the expansion coefficients in (9.209). If $f(t)$ is

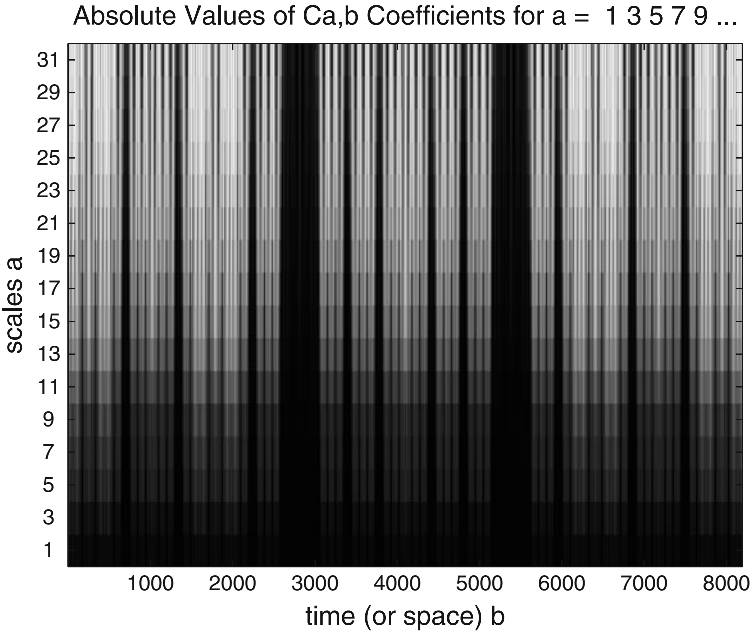


**Fig. 9.22** Wavelet coefficients of a self-similar Koch curve. One clearly sees that for certain shift parameters $b$ the coefficients are comparatively large (of high intensity) for all scales $a$

known for $t = 1, \ldots, N$, and if we want to determine the wavelet coefficients of $f(t)$ from these values, we first have to solve the system of equations obtained from (9.209) after inserting $f(t)$ for $t = 1, \ldots, N$ with respect to $X(t)$, $t = 1, \ldots, N$. The resulting values for $X(t)$ now represent the expansion coefficients $\{f_{0t}\}$, from which we obtain the other coefficients by recursion.

- We may alternatively introduce the wavelet base functions in the form

$$w_{a,b}(t) = \frac{1}{\sqrt{a}} w\left(\frac{t-b}{a}\right) , \quad a > 0 , \ b \in \mathbb{R} . \tag{9.214}$$

The parameters $a, b$ now can vary continuously. In this case the wavelet components of a function $f(t)$ may be defined as

$$c_{a,b} = \int \mathrm{d}t \ f(t) w_{a,b}(t) . \tag{9.215}$$

For $b = ak$ and $a = 2^j$ we recover the wavelets defined in (9.210). In general $\{c_{a,b}\}$ will represent an over-determined set of wavelet components. However, visualizing this set for some range of $a$ and $b$ often exhibits characteristic features of the function $f(t)$ (Fig. 9.22).

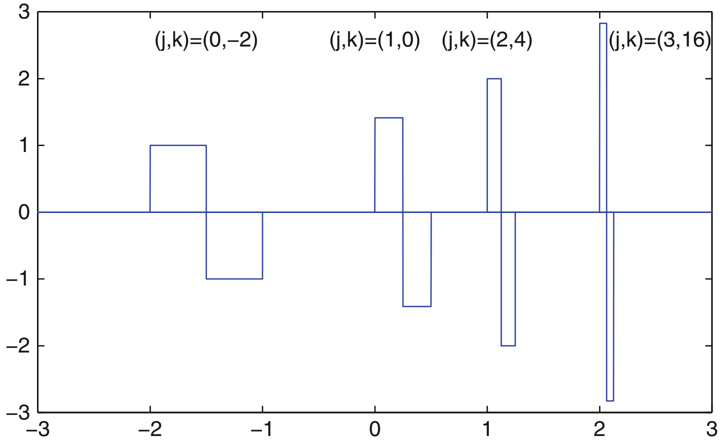


**Fig. 9.23** Haar wavelets for various parameters $(j, k)$

### 9.8.3 *Solutions of the Dilation Equation*

A full discussion of the various methods of solving the dilation equation would go far beyond the scope of this book. For selected filter banks, however, we will state the solution.

The simplest wavelet results from the simplest filter bank. For the Haar filter bank we have

$$\tilde{\mathsf{H}}(z) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} , \tag{9.216}$$

i.e.,

$$(c_0(0), c_0(1)) = \frac{1}{\sqrt{2}}(1, 1) \tag{9.217a}$$

$$(c_1(0), c_1(1)) = \frac{1}{\sqrt{2}}(1, -1) \; . \tag{9.217b}$$

With these coefficients we obtain for the dilation equation

$$\phi(t) = \phi(2t) + \phi(2t - 1) \tag{9.218}$$

and for the wavelet

$$w(t) = \phi(2t) - \phi(2t - 1) \; . \tag{9.219}$$

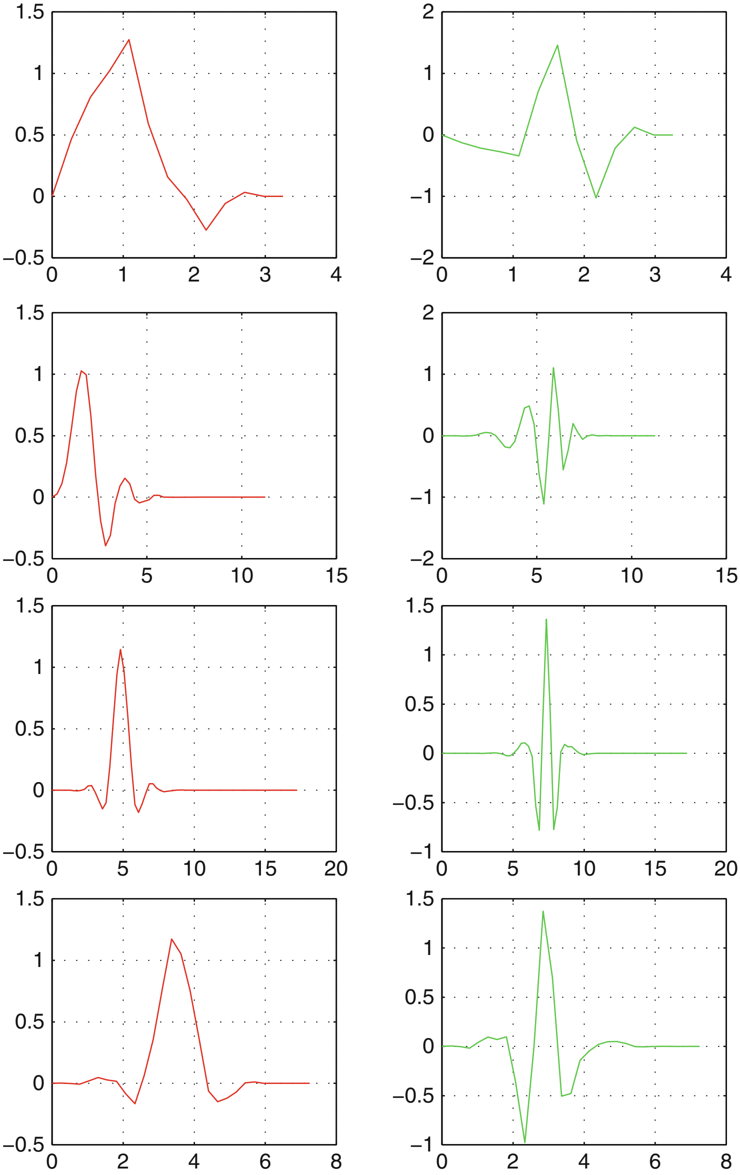


**Fig. 9.24** Scale function (*left*) and wavelet function (*right*) for various wavelets. From *top* to *bottom*: Daubechies of order 2, Daubechies of order 6, Coiflet of order 3, symmetric wavelet of order 4

In this case the solution of the dilation equation is quite simple. We find

$$\phi(t) = \begin{cases} 1 & \text{for} \quad 0 < t < 1 \\ 0 & \text{otherwise}\,. \end{cases} \tag{9.220}$$

Hence, $\phi(2t)$ reproduces the function $\phi(t)$ in the interval $(0, 1/2)$, and $\phi(2t - 1)$ reproduces it in the interval $(1/2, 1)$. The wavelet $w(t)$ and the base functions $w_{jk}(t) = 2^{j/2} w(2^j t - k)$ now are step functions, as shown in Fig. 9.23.

Next we consider the filter bank of the Daubechies filter of order 2 with

$$(c_0(0), c_0(1), c_0(2), c_0(3)) = \frac{1}{4\sqrt{2}}(1 + \sqrt{3}, 3 + \sqrt{3}, 3 - \sqrt{3}, 1 - \sqrt{3}) \ ,$$

$$(c_1(0), c_1(1), c_1(2), c_1(3)) = \frac{1}{4\sqrt{2}}(1 - \sqrt{3}, -3 + \sqrt{3}, 3 + \sqrt{3}, -1 - \sqrt{3}) \ .$$

Now

$$\begin{aligned} h &= (h(0), h(1), h(2), h(3)) = \frac{1}{4\sqrt{2}}(1 - \sqrt{3}, 3 - \sqrt{3}, 3 + \sqrt{3}, 1 + \sqrt{3}) \\ &= (-0.1294, 0.2241, 0.8365, 0.4830), \\ g &= (g(0), g(1), g(2), g(3)) = \frac{1}{4\sqrt{2}}(-1 - \sqrt{3}, 3 + \sqrt{3}, -3 + \sqrt{3}, 1 - \sqrt{3}) \\ &= (-0.4830, 0.8365, -0.2241, -0.1294) \ . \end{aligned}$$

The dilation equation for this case is not so easy to solve. We show the results for the wavelet in Fig. 9.24, which also includes further examples.

# Chapter 10
# Estimators Based on a Probability Distribution for the Parameters

## 10.1 Bayesian Estimator and Maximum a Posteriori Estimator

Given a model with a set $\boldsymbol{\theta}$ of $p$ parameters, the probability of the observable quantities $y_{1\ldots N}$ is $\rho(y_{1\ldots N}|\boldsymbol{\theta})$. We may treat the set of parameters $\boldsymbol{\theta}$ also as a random quantity. Let us denote the prior probability density, i.e., the probability one may assume without knowledge of the data, by $\pi(\boldsymbol{\theta})$, whereas $\rho(\boldsymbol{\theta}|y_{1\ldots N})$, i.e., the density of $\boldsymbol{\theta}$, given the data $y_{1\ldots N}$, may be called the a posteriori density.

Then according to the Bayesian theorem, this a posteriori probability density is given by

$$\rho(\boldsymbol{\theta}|y_{1\ldots N}) = \frac{\rho(y_{1\ldots N}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\rho(y_{1\ldots N})} \,. \tag{10.1}$$

Based on this density, we may define two slightly different estimators of the parameters $\boldsymbol{\theta}$ of the model.

- The Bayesian estimator $\widehat{\boldsymbol{\theta}}_B$ for $\boldsymbol{\theta}$ is defined by the expectation value of $\boldsymbol{\theta}$ with respect to the density $\rho(\boldsymbol{\theta}|y_{1\ldots N})$:

$$\widehat{\boldsymbol{\theta}}_B = \int d^p\,\theta\;\boldsymbol{\theta}\;\rho(\boldsymbol{\theta}|y_{1\ldots N}) \,. \tag{10.2}$$

- The maximum a posteriori estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is defined by

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_\theta \rho(\theta|y_{1\ldots N}) \equiv \arg\max_\theta(\rho(y_{1\ldots N}|\theta)\pi(\theta)). \tag{10.3}$$

  For $\pi(\theta) \propto$ *const.*, the MAP estimator becomes identical to the maximum-likelihood estimator.

*Example.* Let $Y$ be identically distributed with expectation value $\theta$ and variance $\sigma^2$, and $\{y_{1\ldots N}\}$ be an independent sample from which the expectation value has to be

DOI 10.1007/978-3-642-28684-1_10, © Springer-Verlag Berlin Heidelberg 2012

estimated. The common estimator for the expectation value, which needs no prior knowledge, is

$$\widehat{\Theta}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i \ . \tag{10.4}$$

$\widehat{\Theta}_N$ is an unbiased and consistent estimator for $\theta$, i.e., $< \widehat{\Theta}_N, > = 0$ and $Var(\widehat{\Theta}_N) \equiv \sigma^2/N$ converges to zero for $N \to \infty$.

Now assume that, according to our prior knowledge, $\theta$ is "near" $\theta_0$. Mathematically this can be formulated as the assumption, that $\theta$ is a realization of an normally distributed random variable $\Theta$ with expectation value $\theta_0$ and a variance $\sigma_0^2$, i.e., with density

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(\theta-\theta_0)^2} \ . \tag{10.5}$$

Because $\{Y_i\}$ are independent and normally distributed, we obtain for the density $\rho(y_{1\ldots N}|\theta)$

$$\rho(y_{1\ldots N}|\theta) \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\theta)^2} \tag{10.6}$$

$$\propto e^{-\frac{1}{2}\frac{(\overline{y}-\theta)^2}{\sigma^2/N}+\text{terms, not dependent on } \theta} \tag{10.7}$$

$$\text{with} \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i , \tag{10.8}$$

and consequently

$$\rho(\theta|y_{1\ldots N}) \propto \rho(y_{1\ldots N}|\theta)\,\pi(\theta) \propto e^{-\frac{1}{2\pi\sigma_0^2}(\theta-\theta_0)^2} e^{-\frac{1}{2}\frac{(\overline{y}-\theta)^2}{\sigma^2/N}} \propto e^{-\frac{1}{2}\frac{(\theta-\hat{\theta}_B)^2}{\widehat{\Sigma}_B^2}} , \tag{10.9}$$

whereby $\widehat{\theta}_B$ and $\widehat{\Sigma}_B$ can be read off from

$$N\frac{(\overline{y}-\theta)^2}{\sigma^2} + \frac{(\theta-\theta_0)^2}{\sigma_0^2} = \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\theta^2 - 2\theta\left(\frac{\overline{y}}{\sigma^2/N}+\frac{\theta_0}{\sigma_0^2}\right)\right.$$
$$\left. \times\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1} + \ldots\right);$$

thus

$$\widehat{\Sigma}_B^2 = \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1} \tag{10.10}$$

$$\equiv \frac{\frac{\sigma^2}{N}}{1+\frac{\sigma^2}{N\sigma_0^2}} \tag{10.11}$$

$$\equiv \frac{\sigma_0^2}{1+\frac{N\sigma_0^2}{\sigma^2}}, \tag{10.12}$$

$$\widehat{\theta}_B = \frac{1}{1+\frac{\sigma^2}{N\sigma_0^2}}\overline{y} + \frac{1}{1+\frac{\sigma_0^2 N}{\sigma^2}}\theta_0 \tag{10.13}$$

$$\equiv \overline{y} + \frac{1}{1+\frac{\sigma_0^2 N}{\sigma^2}}(\theta_0 - \overline{y}) \tag{10.14}$$

$$\equiv \theta_0 + \frac{1}{1+\frac{\sigma^2}{N\sigma_0^2}}(\overline{y} - \theta_0)\,. \tag{10.15}$$

Hence the random variable $\theta|y_{1\ldots N}$ is a normal random variable with expectation value $\widehat{\theta}_B$ and variance $\widehat{\Sigma}_B^2$. The Bayesian estimator in this case is identical to the MAP estimator. The expectation value and variance differ from the prior values of $\theta_0$ and $\sigma_0^2$ because of the influence of the data. Furthermore, compared to the values $\overline{y}$ and $\sigma^2/N$ for the common estimator, one observes that the variance is reduced at the price of a bias.

On the other hand, the larger the prior variance $\sigma_0^2$, the smaller the deviation of mean and variance from their common values. That means that the case of no prior knowledge can also be simulated by a normal prior distribution with a very large variance. We will always have this in mind when we want to construct a prior distribution with no knowledge at all for a variable with infinite range. A uniform distribution then does not exist, but a normal distribution with a very large variance will simulate the lack of information.

## 10.2 Marginalization of Nuisance Parameters

Within the context of an Bayesian estimator, one may skip the estimation of parameters which are of no interest. Such parameters are called also nuisance parameters.

Consider a particular problem, where two parameters $\theta_1$ and $\theta_2$ are involved, but only, say $\theta_1$, is of interest. Then we may marginalize

$$\rho(\theta_1, \theta_2|y_{1\ldots N}) = \frac{\rho(y_{1\ldots N}|\theta_1, \theta_2)\rho_1(\theta_1, \theta_2)}{\rho(y_{1\ldots N})} \tag{10.16}$$

by integrating with respect to $\theta_2$:

$$\rho(\theta_1|y_{1\ldots N}) = \int d\theta_2 \rho(\theta_1, \theta_2|y_{1\ldots N}) = \int d\theta_2 \frac{\rho(y_{1\ldots N}|\theta_1, \theta_2)\rho_1(\theta_1, \theta_2)}{\rho(y_{1\ldots N})}\,. \tag{10.17}$$

If this integral can be computed analytically, we obtain a reduction of dimensionality and a more tractable estimator by such a marginalization.

*Example.* We consider the general linear model,

$$Y_i = \sum_{j=1}^{M} \underline{K}_{ij} X_j + \epsilon_i \,, \qquad \epsilon_i \sim N(0, \sigma^2)\,,\, i = 1, \ldots, N, \tag{10.18}$$

where the matrix $\underline{K}$ may still depend on some parameters $\{\omega\}$. Two special, particularly simple cases are, e.g., the model

$$Y_i = A \cos \omega t_i + B \sin \omega t_i + \epsilon_i, \quad i = 1, \ldots, N, \tag{10.19}$$

so that $\underline{K}$ is the $N \times M$ matrix (with $M = 2$)

$$\underline{K} = \begin{pmatrix} \cos \omega t_1 & \sin \omega t_1 \\ \cos \omega t_2 & \sin \omega t_2 \\ \vdots & \vdots \\ \cos \omega t_N & \sin \omega t_N \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix}, \tag{10.20}$$

and the model

$$Y_i = a_0 + a_1 t_i + \epsilon_i \tag{10.21}$$

so that $\underline{K}$ is the matrix

$$\underline{K} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}. \tag{10.22}$$

The likelihood function is given by

$$\rho(y_{1\ldots N} | \omega, \mathbf{x}, \sigma) = (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \mathbf{e}^T \cdot \mathbf{e}} \tag{10.23}$$

where

$$\mathbf{e} = \mathbf{y} - \underline{K}\mathbf{x} \tag{10.24}$$

and the a posteriori probability reads, e.g., in the first special example

$$\rho(\omega, \mathbf{x}, \sigma | y_{1\ldots N}) = \frac{\rho(y_{1\ldots N} | \omega, \mathbf{x}, \sigma) \rho_\omega(\omega) \rho_{\mathbf{x}}(\mathbf{x}) \rho_\sigma(\sigma)}{\rho(y_{1\ldots N})}\,. \tag{10.25}$$

Now we have to formulate our prior knowledge about the parameters. Let us take

$$\rho_\omega(\omega) = const. \tag{10.26}$$

That means we have no knowledge at all about the value of the $\{\omega\}$ set. With the same argument, we may assume uniform priors for the values of $\mathbf{x}$. An assumption about the prior knowledge of the variance of the noise has to take into account that $\sigma > 0$. We may assume a uniform prior for $\log \sigma$, so that

$$\rho_\sigma(\sigma) \propto \frac{1}{\sigma}. \tag{10.27}$$

As already remarked, taking these prior distributions as uniform, means that we introduce normal distributions with a very large variance so that the influence of their mean and variance can be neglected.

Let us now discuss different marginalizations.

*Marginalization with respect to $\sigma$* The integral over $\sigma$ reads

$$\rho(\omega, \mathbf{x}|y_{1\ldots N}) \propto \int_0^\infty d\sigma \sigma^{-N-1} e^{-\frac{1}{2\sigma^2}\mathbf{e}^T\cdot\mathbf{e}} \tag{10.28}$$

$$\propto \frac{1}{(\mathbf{e}^T \cdot \mathbf{e})^{N/2}} = \frac{1}{[(\mathbf{y} - \underline{K}\mathbf{x})^T(\mathbf{y} - \underline{K}\mathbf{x})]^{N/2}}\ . \tag{10.29}$$

Thus, given $\omega$, the a posteriori density becomes maximal when

$$\chi^2 = (\mathbf{y} - \underline{K}\mathbf{x})^T(\mathbf{y} - \underline{K}\mathbf{x}) \tag{10.30}$$

becomes minimal: The MAP estimator coincides with the least-squares estimator.

*Marginalization with respect to* $\mathbf{x}$ On the other hand, the integral over $\mathbf{x}$ leads to

$$\rho(\omega, \sigma|y_{1\ldots N}) \propto \frac{1}{\sigma}(2\pi\sigma^2)^{-N/2} \int d\mathbf{x} e^{-\frac{1}{2\sigma^2}\mathbf{e}^T\cdot\mathbf{e}} \tag{10.31}$$

$$\propto \frac{1}{\sigma}(2\pi\sigma^2)^{-\frac{N-M}{2}} \frac{1}{\sqrt{det\underline{K}^T\underline{K}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\cdot\mathbf{y}-\mathbf{f}^T\cdot\mathbf{f})}, \tag{10.32}$$

where

$$\mathbf{f} = \underline{K}(\underline{K}^T\underline{K})^{-1}\underline{K}^T\mathbf{y}\ . \tag{10.33}$$

Now, for a given $\omega$, the maximum of this a posteriori density with respect to $\sigma$ is achieved at

$$\hat{\sigma} = \text{arg min}_\sigma \left[\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - \mathbf{f}^T\mathbf{f}) + (N - M + 1)\ln\sigma\right]\ . \tag{10.34}$$

Setting the derivative of the argument with respect to $\sigma$ equal to zero, we obtain

$$-\frac{1}{\sigma^3}(\mathbf{y}^T\mathbf{y} - \mathbf{f}^T\mathbf{f}) + (N - M + 1)\frac{1}{\sigma} = 0\,, \tag{10.35}$$

($M = 2$ in our case), hence

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{f}^T\mathbf{f}}{N - M + 1} \,. \tag{10.36}$$

The estimator for the variance, obtained in this way, is the discrepancy $\mathbf{y}^T\mathbf{y} - \mathbf{f}^T\mathbf{f}$ divided by $N - M + 1$.

*Marginalization with respect to $\sigma$ and* $\mathbf{x}$ If we finally marginalize both with respect to $\sigma$ and $\mathbf{x}$, we obtain (see, e.g., Fitzgerald and Ó Ruanaidh 1996)

$$\rho(\omega | y_{1\ldots N}) \propto \frac{1}{\sqrt{\det(\underline{K}^T\,\underline{K})}} \frac{1}{(\mathbf{y}^T\mathbf{y} - \mathbf{f}^T\mathbf{f})^{(N-M)/2}} . \tag{10.37}$$

Let us evaluate this density for the model

$$Y_i = A \cos \omega t_i + B \sin \omega t_i + \epsilon_i \tag{10.38}$$

where

$$\underline{K} = \begin{pmatrix} \cos\omega t_1 & \sin\omega t_1 \\ \vdots & \vdots \\ \cos\omega t_N & \sin\omega t_N \end{pmatrix} , \tag{10.39}$$

and consequently

$$\underline{K}^T\,\underline{K} = \begin{pmatrix} \cos\omega t_1 & \ldots & \cos\omega t_N \\ \sin\omega t_N & \ldots & \sin\omega t_N \end{pmatrix} \cdot \begin{pmatrix} \cos\omega t_1 & \sin\omega t_1 \\ \vdots & \vdots \\ \cos\omega t_N & \sin\omega t_N \end{pmatrix} \tag{10.40}$$

$$= \begin{pmatrix} \cos\omega {t_1}^2 + \ldots & \cos\omega t_1 \sin\omega t_1 + \ldots \\ \sin\omega t_1 \cos\omega t_1 + \ldots & \sin\omega {t_1}^2 + \ldots \end{pmatrix} \tag{10.41}$$

$$\approx \frac{N}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} , \tag{10.42}$$

and

$$\underline{K}^T\mathbf{y} = \begin{pmatrix} y_1 \cos\omega t_1 + \ldots + y_N \cos\omega t_N \\ y_1 \sin\omega t_1 + \ldots + y_N \sin\omega t_N \end{pmatrix} . \tag{10.43}$$

Then, because of

$$\mathbf{f}^T\mathbf{f} = \mathbf{y}^T\,\underline{K}(\underline{K}^T\,\underline{K})^{-1}\underline{K}^T\,\underline{K}(\underline{K}^T\,\underline{K})^{-1}\underline{K}\mathbf{y} = \mathbf{y}^T\,\underline{K}(\underline{K}^T\,\underline{K})^{-1}\underline{K}^T\mathbf{y}, \tag{10.44}$$

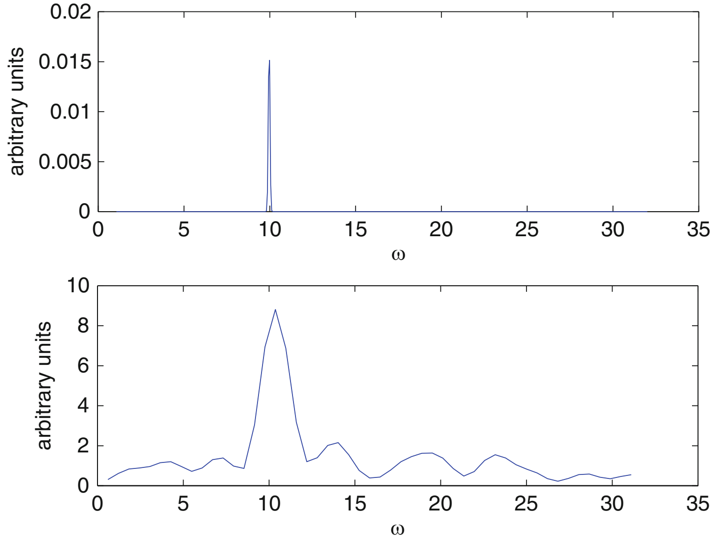


**Fig. 10.1** *Above*: The a posteriori density $\rho(\omega|y_{1\dots N})$ given a signal $y_{1\dots N}$ with $\omega = 10$ and $\sigma = 1$. *Below*: The power spectrum of this signal

one obtains

$$\mathbf{y}^T\mathbf{y} - \mathbf{f}^T\mathbf{f} = \mathbf{y}^T\mathbf{y} - \frac{2}{N}\left[\left(\sum_{i=1}^{N} y_i \cos\omega t_i\right)^2 + \left(\sum_{i=1}^{N} y_i \sin\omega t_i\right)^2\right]$$

$$= \sum_{i=1}^{N} y_i^2 - 2C(\omega), \tag{10.45}$$

where

$$C(\omega) = \frac{1}{N}\left|\sum_{i=1}^{N} y_i e^{i\omega t_i}\right|^2 \tag{10.46}$$

can also be interpreted as the sum over the power spectrum of the time series. This sum contains only one term because there is power in only one frequency. Hence we obtain

$$\rho(\omega|y_{1\dots N}) \propto \frac{1}{|\sum_{i=1}^{N} y_i^2 - 2C(\omega)|^{(N-2)/2}}\,. \tag{10.47}$$

This density exhibits a maximum for a frequency $\omega$, for which $2C(\omega)$ is nearest to $\sum_{i=1}^{N} y_i^2$.

Figure 10.1 shows $\rho(\omega|y_{1\dots N})$ for $N = 100$ in the upper subplot, given a signal generated according to (13.87) with $\omega = 10$ and $\sigma = 1$. In the lower subplot, the power spectrum of the signal is presented. The peak here is much broader.

Thus because of the detailed model under which the data are analyzed, the estimation of $\omega$ is much better than by an estimation of the spectrum.

*Remark.* Parseval's theorem, on the other hand, states that for the Fourier transform

$$\tilde{y}_k = \sum_{i=1}^{N} y_i e^{i\omega_k t_i}, \qquad \omega_k = \frac{2\pi k}{N}, k = 0 \ldots \frac{N}{2}, \tag{10.48}$$

one obtains in general

$$2\sum_{k=1}^{\frac{N}{2}} C(\omega_k) = \frac{2}{N}\sum_{k=1}^{\frac{N}{2}} |\tilde{y}_k|^2 \tag{10.49}$$

$$= \frac{2}{N}\sum_{k=1}^{\frac{N}{2}}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j e^{i\omega_k t_i} e^{-i\omega_k t_j} \tag{10.50}$$

$$= \sum_{i=1}^{N} y_i^2 \tag{10.51}$$

because of

$$\frac{2}{N}\sum_{k=1}^{N/2} e^{i\omega_k(t_i - t_j)} = \delta_{ij}\,. \tag{10.52}$$

Thus, if exactly only one frequency $\omega_k$ is present, then also $2C(\omega_k) = \sum_{i=1}^{N} y_i^2$, the maximum of $\rho(\omega|y_{1\ldots N})$, will be infinitely sharp.

## 10.3 Numerical Methods for Bayesian Estimators

In the previous section, we introduced the probability density for the set $\theta$ of parameters, given the data $y_{1\ldots N}$,

$$\rho(\theta|y_{1\ldots N}) \propto \rho(y_{1\ldots N}|\theta)\pi(\theta). \tag{10.53}$$

One cannot always find the maximum of this expression with respect to $\theta$. But one may try to draw random numbers according to this density to infer the mean and variance from such a sample. This corresponds to the Bayesian estimator instead of the MAP estimator.

If one knows the density of $Y|y_{1\ldots N}$ explicitly, one can pick up the idea of Sect. 5.5 where the Monte Carlo method and the Gibbs sampler were introduced. The underlying idea is to construct a stochastic process, so that its stationary probability distribution is identical to that given in (10.53). Here we have to define a stochastic process in parameter space, and there are basically two different strategies to do that:

*The Metropolis–Hastings method:* One defines a transition $\theta' \to \theta$ by

$$\theta' = \theta + \zeta \tag{10.54}$$

where $\zeta$ is a realization of a random variable with a given proposal density $\rho_P(\zeta)$, which usually can be chosen to be normally distributed with zero mean and a variance which should comparable to the expected variance of the estimation of $\theta$.

This transformation will be accepted with probability

$$w_{\theta'\theta} dt = \min\left(1, \frac{\rho(\theta'|y_{1\ldots N})}{\rho(\theta|y_{1\ldots N})}\right). \tag{10.55}$$

This means that one draws a random number $r$ from a uniform distribution in [0, 1] and accepts $\theta'$ as the new state in parameter space if

$$r < \frac{\rho(\theta'|y_{1\ldots N})}{\rho(\theta|y_{1\ldots N})}. \tag{10.56}$$

Otherwise, the state will remain at $\theta$.

This choice of $w_{\theta'\theta} dt$ obviously obeys the principle of detailed balance because for $\rho(\theta'|y_{1\ldots N})/\rho(\theta|y_{1,\ldots,N}) < 1$, e.g.,

$$w_{\theta'\theta} dt = \frac{\rho(\theta'|y_{1,\ldots,N})}{\rho(\theta|y_{1,\ldots,N})}, \tag{10.57}$$

and

$$w_{\theta\theta'} dt = \min\left(1, \frac{\rho(\theta|y_{1,\ldots,N})}{\rho(\theta'|y_{1,\ldots,N})}\right) = 1, \tag{10.58}$$

hence

$$\frac{w_{\theta'\theta}}{w_{\theta\theta'}} = \frac{\rho(\theta'|y_{1,\ldots,N})}{\rho(\theta|y_{1,\ldots,N})}, \tag{10.59}$$

as required for the principle of detailed balance (see (5.36)). In the same way, one shows detailed balance for $\rho(\theta'|y_{1\ldots N})/\rho(\theta|y_{1,\ldots,N}) > 1$.

*The Gibbs sampler:* The task of drawing samples from a multivariate density $\rho(\theta|y_{1\ldots N})$ can be broken down to drawing successive samples from densities of smaller dimensions, e.g., univariate samples. For example, for sampling $\rho(\theta_1, \theta_2)$, one draws samples from $\rho(\theta_1|\theta_2)$ and $\rho(\theta_2|\theta_1)$ consecutively, where, e.g., $\theta_1$ in $\rho(\theta_2|\theta_1)$ is the realization of $\theta_1|\theta_2$ in the preceding step.

In general, if $\theta = (\theta_1, \ldots, \theta_n)$ is the parametric vector, one consecutively generates realizations of $\theta_i|\theta_{\setminus i}, y_{1\ldots N}$ for $i = 1, \ldots, n$, doing this $N$ times to obtain $N$ samples of the parameter vector $\theta|y_{1\ldots N}$. Here, $\theta_{\setminus i}$ means

$(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_N)$. We will see that these conditional densities $\rho(\theta_i|\theta_{\setminus i}, y_{1\ldots N})$ can easily be derived in a large class of models.

This method of obtaining realizations of the random vector $\theta|y_{1\ldots N}$ after a transition time is due to Geman and Geman (1984).

*Examples.*

- We consider a random vector $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)$ with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and covariance matrix $C$. Then the density reads

$$\rho(\theta) = \frac{1}{2\pi det(C)} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})C^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})} \,. \tag{10.60}$$

To find $\rho(\theta_1|\theta_2)$, we look at the quadratic form in the exponent

$$\begin{aligned}(\boldsymbol{\theta} - \boldsymbol{\mu})C^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) &= \theta_1^2(C^{-1})_{11} + 2\theta_1 \left[(C^{-1})_{12}(\theta_2 - \mu_2) - \mu_1(C^{-1})_{11}\right] + \ldots \\ &= (C^{-1})_{11} \left[\theta_1^2 - 2\theta_1 \left[\mu_1 - \frac{(C^{-1})_{12}}{(C^{-1})_{11}}(\theta_2 - \mu_2)\right] + \ldots\right. \end{aligned} \tag{10.61}$$

Hence we conclude that $\rho(\theta_1|\theta_2)$ has the form

$$\rho(\theta_1|\theta_2) \propto e^{-\frac{1}{2\sigma^2_{\Theta_1|\theta_2}}(\theta_1 - E(\Theta_1|\theta_2))^2} \tag{10.62}$$

with

$$\sigma^2_{\Theta_1|\theta_2} = Var(\Theta_1|\theta_2) = \frac{1}{(C^{-1})_{11}}, \tag{10.63}$$

$$E(\Theta_1|\theta_2) = \mu_1 - \frac{(C^{-1})_{12}}{(C^{-1})_{11}}(\theta_2 - \mu_2) \tag{10.64}$$

$$\equiv \mu_1 + \frac{C_{12}}{C_{22}}(\theta_2 - \mu_2). \tag{10.65}$$

A realization of $\Theta_1|\theta_2$ can be drawn as a realization of

$$\Theta_1|\theta_2 = \mu_1 + \frac{C_{12}}{C_{22}}(\theta_2 - \mu_2) + \frac{1}{\sqrt{(C^{-1})_{11}}}\eta_1, \quad \eta_1 \propto N(0, 1), \tag{10.66}$$

and similarly a realization of $\Theta_2|\theta_1$

$$\Theta_2|\theta_1 = \mu_2 + \frac{C_{21}}{C_{11}}(\theta_1 - \mu_1) + \frac{1}{\sqrt{(C^{-1})_{22}}}\eta_2, \quad \eta_2 \propto N(0, 1)\,. \tag{10.67}$$

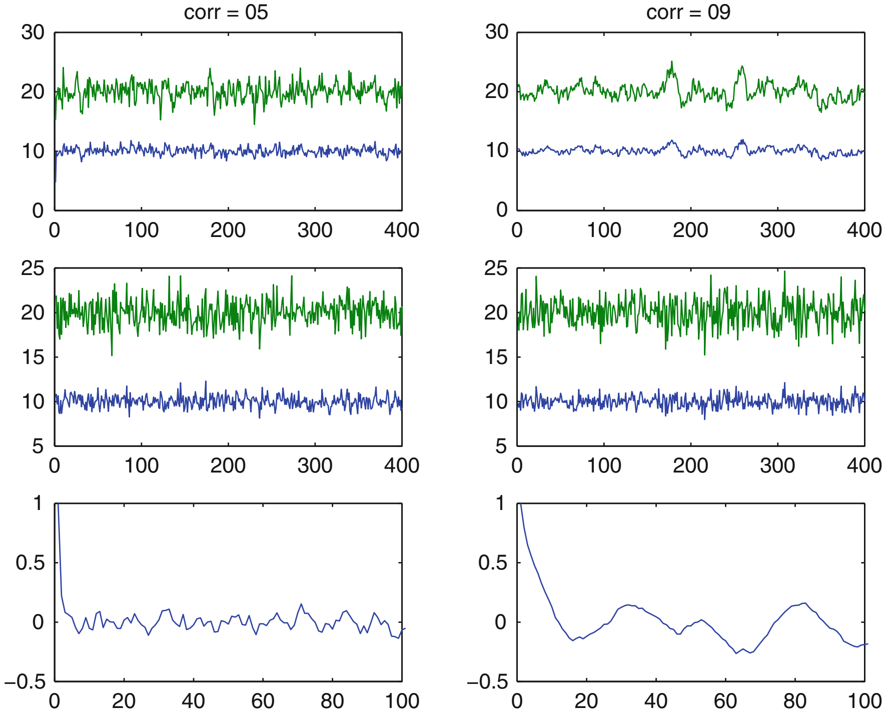


**Fig. 10.2** *Above*: Realization of $\theta_1, \theta_2$ with a correlation of 0.5 (*left*) and 0.9 (*right*), generated by a Gibbs sampler using (10.66) and (10.67). For correlation 0.9 between $\theta_1$ and $\theta_2$, the consecutive realizations of $\theta_1$ and $\theta_2$ also appear dependent. *Middle*: Realization of $(\theta_1, \theta_2)$ for correlation 0.5 and 0.9, respectively, generated by using the Choleski decomposition, where no dependence of consecutive realizations can be observed. The *lowest* subplots show the correlation function $\mathrm{Corr}(\theta_1^{(i)}\theta_1^{(i+\tau)})$ of the consecutive realizations of $\theta_1^{(i)}$ as function of $\tau$, generated by a Gibbs sampler. Whereas this function immediately drops down for correlation 0.5 between $\theta_1$ and $\theta_2$, the correlation between consecutive realizations is evident for correlation 0.9 between $\theta_1$ and $\theta_2$

Starting with some $\theta_1^{(1)}$, one draws a random variable $\theta_2^{(1)}$ with mean $\mu_2 + \frac{C_{21}}{C_{11}}(\theta_1 - \mu_1)$ and variance $\frac{1}{(C^{-1})_{22}}$. In this manner, one constructs a process

$$\theta_1^{(1)} \rightarrow \theta_2^{(1)} \rightarrow \theta_1^{(2)} \rightarrow \theta_2^{(2)} \tag{10.68}$$

which can be paired to samples $(\theta_1^{(k)}, \theta_2^{(k)})$, which are realizations of $\rho(\theta_1, \theta_2)$. Figure 10.2 shows a series of such generated normally distributed random numbers in the upper left panel, where the correlation between $\theta_1$ and $\theta_2$ is set to 0.5. In the left middle subplot, realization of a series of the same couple of random variables is shown, generated by the common method using the Cholesky decomposition. Obviously, one detects no great difference. In the right subplots, the same is done for random variables with correlation 0.9. Now one finds strong deviations in the upper right panel compared to a commonly generated series of such normally generated random numbers. The consecutively generated numbers are not independent as one may also see from the lowest subplots: for correlation 0.9, there is still a non negligible correlation between consecutively generated numbers, whereas for correlation 0.5 between the two random variables, this correlation falls off strongly with distance.

Thus one has to control whether the random numbers generated are really independent. If this is not the case as for correlation 0.9 in the example, one should find a rough estimate of the correlation distance. If this is $N_0$, then one should only take each $N_0$th random number and throw away all others generated in between. Following this procedure, one also gets a series of random numbers which is visually indistinguishable from one commonly generated.

- We will now estimate the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ in the model

$$\mathbf{y} = \underline{K}\boldsymbol{\theta} + \mathbf{e} \qquad \mathbf{e}_i \sim N(0, \sigma^2). \tag{10.69}$$

Examples are:

- For the linear model $Y_i = \theta_1 + \theta_2 t_i + \epsilon_i$, we have

$$\underline{K} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix}. \tag{10.70}$$

- For $Y_i = \theta_1 \cos \omega t_i + \theta_2 \sin \omega t_i + \epsilon_i$,

$$\underline{K} = \begin{pmatrix} \cos \omega t_1 & \sin \omega t_1 \\ \vdots & \vdots \\ \cos \omega t_N & \sin \omega t_N \end{pmatrix}. \tag{10.71}$$

- An AR(p) model with $y(t) = \theta_1 y(t-1) + \ldots + \theta_p y(t-p) + e(t)$ can also be written as $\mathbf{e} = \mathbf{y} - \underline{K}\boldsymbol{\theta}$:

$$\begin{pmatrix} e(p+1) \\ \vdots \\ e(N) \end{pmatrix} = \begin{pmatrix} y(p+1) \\ \vdots \\ y(N) \end{pmatrix} - \begin{pmatrix} y(p) & \ldots & y(1) \\ \vdots & & \vdots \\ y(N-1) & \ldots & y(N-p) \end{pmatrix} \cdot \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix},$$

so that

$$\mathbf{y} = \begin{pmatrix} y(p+1) \\ \vdots \\ y(N) \end{pmatrix}, \underline{K} = \begin{pmatrix} y(p) & \ldots & y(1) \\ \vdots & & \vdots \\ y(N-1) & \ldots & y(N-p) \end{pmatrix} \tag{10.72}$$

and

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} . \tag{10.73}$$

In all of these models, the a posteriori density $\rho(\boldsymbol{\theta}\,|y_{1\ldots N})$ can be written as

$$\rho(\boldsymbol{\theta}\,|y_{1\ldots N}) \propto \frac{1}{\sigma}(2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\underline{K}\boldsymbol{\theta})^T(\mathbf{y}-\underline{K}\boldsymbol{\theta})} . \tag{10.74}$$

To estimate the parameters with a Gibbs sampler, we have to calculate the densities $\rho(\theta_i\,|\theta_{\setminus i}, y_{1\ldots N})$. One obtains

$$\rho(\theta_i\,|\theta_{\setminus i}, y_{1\ldots N}) \propto \frac{1}{\sigma}(2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}Q} \tag{10.75}$$

with

$$Q = (\underline{K}^T\underline{K})_{ii}\theta_i^2 - 2\theta_i \left[ (\underline{K}^T\mathbf{y})_i - \sum_{j\neq i}(\underline{K}^T\underline{K})_{ij}\theta_j \right] + \ldots \tag{10.76}$$

$$= (\underline{K}^T\underline{K})_{ii} \left[ \theta_i^2 - 2\theta_i \left[ \frac{(\underline{K}^T\mathbf{y})_i - \sum_{j\neq i}(\underline{K}^T\underline{K})_{ij}\theta_j}{(\underline{K}^T\underline{K})_{ii}} \right] + \ldots \right] \tag{10.77}$$

Hence

$$E(\theta_i\,|\theta_{\setminus i}, y_{1\ldots N}) = \frac{(\underline{K}^T\mathbf{y})_i - \sum_{j\neq i}(\underline{K}^T\underline{K})_{ij}\theta_j}{(\underline{K}^T\underline{K})_{ii}}, \tag{10.78}$$

$$Var(\theta_i\,|\theta_{\setminus i}, y_{1\ldots N}) = \frac{1}{(\underline{K}^T\underline{K})_{ii}}. \tag{10.79}$$

In Figs. 10.3 and 10.4 realizations of $(\theta_1, \theta_2)$ are shown for the linear model and the AR(2) model, respectively. One observes that after a short transition time, the random numbers fluctuate about the exact values indicated by a line.

*Remark.* Finally, if there are some missing data, say $y(t)$, $t = n_1 + 1\ldots n_2$, one may reconstruct these by the following method: Let us start with some initial estimates. For these data $y(t)$, $t = 1, \ldots, n_1, n_1 + 1, \ldots, n_2, \ldots, N$, one may estimate the parameters of an AR(p) model. Given these parameters, one may determine upgraded estimates of the missing data, as explained below. Then, taking these into account, one may again estimate the parameters, and so on, until the procedure converges.

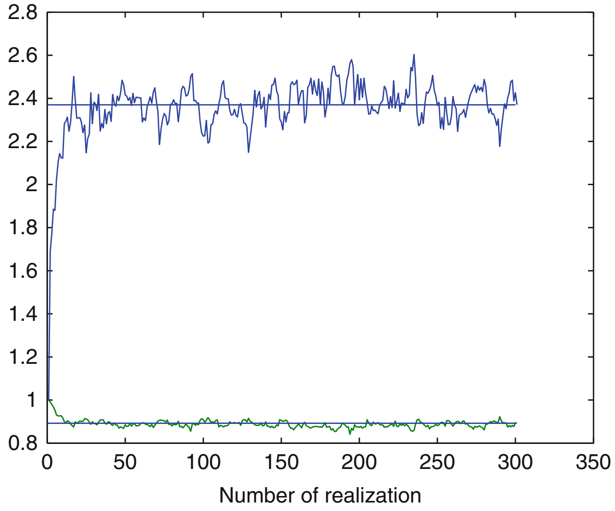


**Fig. 10.3** A realization of $(\theta_1, \theta_2)$ for the linear model. The *lines* indicate the exact values

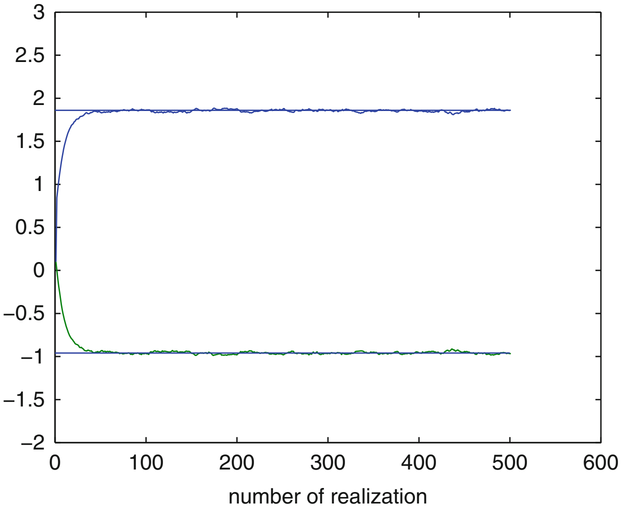


**Fig. 10.4** A realization of $(\theta_1, \theta_2)$ for an AR(2) model. The *lines* indicate the exact values

The estimation of the missing data for given parameters of the AR(p) model proceeds as follows. One writes the residuals $e(t)$, $t = 1, \ldots, N$ as $\mathbf{e} = \underline{G} \cdot \mathbf{w}$, i.e., as

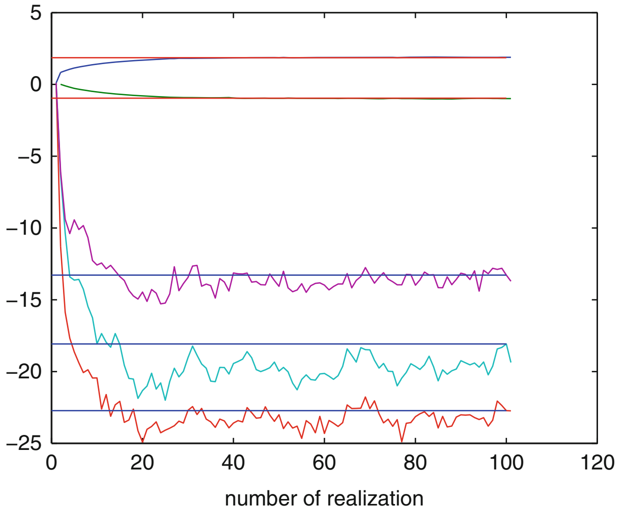


**Fig. 10.5** A realization of $(\theta_1, \theta_2)$ and of estimations of three values which were missing in the time series of an AR(2) model. The *lines* indicate the exact values

$$\begin{pmatrix} e(1) \\ \\ \vdots \\ e(p+1) \\ \vdots \\ \\ e(N) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ -\theta_1 & 1 & 0 & \dots & \dots & 0 \\ \vdots & & & & & \vdots \\ -\theta_p & -\theta_{p-1} & \dots & 1 & \dots & 0 \\ 0 & -\theta_p & \dots & & .. & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & & -\theta_2 & -\theta_1 & 1 \end{pmatrix} \cdot \begin{pmatrix} y(1) \\ \\ \vdots \\ y(p+1) \\ \vdots \\ \\ y(N) \end{pmatrix} . \tag{10.80}$$

Then one may decompose $\mathbf{w}$ with $\mathbf{w}^T = (y(1) \dots y(N))$ as

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{x} \\ \mathbf{w}_2 \end{pmatrix}, \quad \begin{aligned} \mathbf{w}_1^T &= \quad (y(1) \dots y(n_1)), \\ \mathbf{x}^T &= (y(n_1+1) \dots y(n_2)), \\ \mathbf{w}_2^T &= (y(n_2+1) \dots y(N)) . \end{aligned} \tag{10.81}$$

Then also

$$\mathbf{e} = \underline{G}\mathbf{w} = \underline{K}\mathbf{x} - \mathbf{d} \tag{10.82}$$

with

$$K_{ij} = G_{ij}, \quad \text{for} \quad i = 1, \dots, N, \quad j = n_1 + 1, \dots, n_2, \tag{10.83}$$

and

$$d_i = -\sum_{j=1}^{n_1} G_{ij}(w_1)_j - \sum_{j=n_2+1}^{N} G_{ij}(w_2)_j . \tag{10.84}$$

Then **e** is again in the form of a general linear model with **x** as the vector of unknowns and **y** as the vector of (known) data.

Figure 10.5 shows the results of a Gibbs sampler for estimating the two parameters and three missing values of an AR(2) model.

# Chapter 11
# Identification of Stochastic Models from Observations

Having studied methods for analyzing a signal that was interpreted as a realization of a stochastic process, we now turn to some basics of model-based data analysis. We shall formulate the questions which may be addressed for a given set of data assuming a suitable model or a suitable structure of a model.

## 11.1 Parameter Identification for Autoregressive Processes

Björn Schelter

In several applications it is of importance to be able to identify parameters of stochastic processes from measured data. Among others the Bayes estimation or the Maximum Likelihood estimation discussed in the previous chapter presents a promising approach for this. As there are too many possible scenarios for which parameter estimation would be of quite some interest, we need to restrict ourselves to discussing few examples and demonstrating the underlying principles thereupon. First, we will therefore discuss autoregressive processes these processes play an important role in spectral estimation (Sect. 9.1) and will also play an important role in the following chapter discussing Granger causality (Sect. 11.2).

Autoregressive models are represented by

$$\begin{aligned} X(t) &= \sum_{k=1}^{p} \alpha_k X(t-k) + \eta(t), \qquad (11.1) \\ \eta(t) &\propto \mathrm{WN}(\mathbf{0}, \sigma^2). \end{aligned}$$

In order to identify there parameters one needs to tackle two issues. Firstly, the order $p$ needs to be determined, secondly, the parameters $\alpha_k$ and $\sigma^2$ need to be estimated. Both challenges can be addressed using the concept of the best linear predictor, which can be defined for random variables in general.

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_11, © Springer-Verlag Berlin Heidelberg 2012

**The best linear predictor.** Given a random variable $X_0$, which may depend on some other random variables $X_1, X_2, X_3, \ldots, X_n$. For simplicity let us assume, that all random variables have zero mean. A linear predictor for $X_0$ is the superposition

$$\hat{X}_0 = \sum_{i=1}^{n} \alpha_i X_i . \tag{11.2}$$

In order to get the best linear prediction for $X_0$ one has to choose the parameters $\alpha_i$ such that

$$\langle (X_0 - \hat{X}_0)^2 \rangle = \left\langle \left( X_0 - \sum_{i=1}^{n} \alpha_i X_i \right)^2 \right\rangle \tag{11.3}$$

is minimal. This leads to the equation

$$\left\langle \left( X_0 - \sum_{j=1}^{n} \alpha_j X_j \right) X_i \right\rangle = 0, \quad \text{for } i = 1, \ldots, n. \tag{11.4}$$

By using the covariance matrix

$$C_{ij} = \langle X_i X_j \rangle, \ \ i, j = 0, \ldots, n \tag{11.5}$$

(11.4) finally reads

$$\sum_{j=1}^{n} \alpha_j C_{ji} = C_{0i}, \ \ i = 1, \ldots, n. \tag{11.6}$$

From this system of linear equations, knowm as Yule–Walker equations, the parameters $\alpha_j$ can be deduced. The best linear predictor then reads

$$\hat{X}_0 = \sum_{i,r=1}^{N} X_i (\mathbf{C}^{-1})_{ir} C_{r0}. \tag{11.7}$$

For the prediction error we get

$$P = \langle (X_0 - \hat{X}_0)^2 \rangle = C_{00} - \sum_{l=1}^{n} \alpha_l C_{l0}. \tag{11.8}$$

**Application to a times series.** Given a time series one may try to predict $X(t)$ on the basis of the past of the process, say $X(t-1), \ldots, X(t-n)$. For each $n$ one may determine the corresponding parameters, which we will now call $\alpha_j^n$, $j = 1, \ldots, n$. If the time series can ever be described by a linear model, we expect, that for every

$j$ the $\alpha_j^n$ converge with increasing $n$ to a fixed value $\hat{\alpha}_j$ and that at a given value of $n$ the last parameter $\hat{\alpha}_n^n$ will become compatibel with zero. That means that for this $n$ there is no dependence from $X(t-n)$ anymore so that the time series can modelled satisfactory by an AR(p)-model with $p = n - 1$.

For a simulated AR(p)-model this scenarium can be demonstrated impressively. Thus, in this way the order $p$ as well as the corresponding parameters $\alpha_j^p$, $j = 1, \ldots, p$ can be estimated whenever an autoregressive model is taken into account (see e.g. Honerkamp 1994).

There is a deep connection between the parameters $\alpha_n^n$, $n = 1, \ldots$ and a quantity which is called the partial correlation function, defined as follows:

Let us assume that for a given time series the $X(t-n), \ldots, X(t-1)$ exert influence on $X(t)$. Then the best linear predictor for X(t) is

$$\hat{X}(t) \mid X(t-n), \ldots, X(t-1) = \alpha_1 X(t-1) + \ldots + \alpha_n X(t-n) \tag{11.9}$$

where the $\alpha_j$, $j = 1, \ldots, n$ have to be determined by the above mentioned procedure.

We will also be interested in the best linear predictor for $X(t-n-1)$

$$\hat{X}(t-n-1) \mid X(t-n), \ldots, X(t-1) = \alpha_1 X(t-n) + \ldots + \alpha_n X(t-1) \tag{11.10}$$

which is nothing else then the best backcasting of $X(t-n-1)$ from the set $X(t-n), \ldots, X(t-1)$. Because the times series is invariant with respect to time reversal, as we have always assumed, the set of parameters $\alpha_j$, $j = 1, \ldots, n$ will turn out be the same as for the above introduced best linear predictor for $X(t)$. In the residuals

$$r^{\leftarrow}(t) = X(t) - \hat{X}(t) \tag{11.11}$$

and

$$r^{\rightarrow}(t-n-1) = X(t-n-1) - \hat{X}(t-n-1) \tag{11.12}$$

now all dependencies of the set $X(t-1), \ldots, X(t-n)$ are eliminated.

The partial covariance function is now defined as

$$pcov_n = \mathrm{cov}(r^{\leftarrow}(t), r^{\rightarrow}(t-n-1)),\ n = 1, \ldots \tag{11.13}$$

and then the partial correlation function $\Phi_n$ as usual by normalisation.

The scenarium is illustrated as

$$\ldots, X(t-n-1), \underbrace{X(t-n), \ldots, X(t-1)}_{\text{exerts influence}}, X(t), X(t+1), \ldots \tag{11.14}$$

and by increasing $n$ one will observe that at a given value the partial correlation function $\Phi_n$ will become compatibel with zero because the residuals become uncorrelated: $X(t-n-1)$ has no influence on $X(t)$ and $X(t)$ does not play any role in an backcasting of $X(t-n-1)$. That reminds one of the parameter $\alpha_n^n$, and indeed, one can show that (Lütkepohl and Krätzig 2004)

$$\Phi_n = \alpha_n^n. \tag{11.15}$$

*Example.* For $n = 1$ one obtains from (11.6) with $C_{ij} = C(i-j) = C(j-i)$

$$\alpha_1^1 = \Phi_n = \frac{C(1)}{C(0)}. \tag{11.16}$$

For $n = 2$ (11.6) reads explicitely

$$\begin{aligned} \alpha_1 C(0) + \alpha_2 C(1) &= C(1) \\ \alpha_1 C(1) + \alpha_2 C(0) &= C(2) \end{aligned} \tag{11.17}$$

which leads to

$$\alpha_2^2 = \Phi_2 = \frac{C(2)C(0) - C^2(1)}{C^2(0) - C^2(1)}. \tag{11.18}$$

Here we would like to briefly mention that in Sect. 9.5 we have introduced the partial coherence and the partial phase spectrum. Both quantities can be motivated using best linear predictors as well. This is why they are called *partial* quantities. The mathematical theory can be found in Dahlhaus (2000).

As an illustrative example the AR[2] process

$$x(t) = 1.9x(t-1) - 0.991x(t-2) + \eta(t) \tag{11.19}$$

is considered. The parameters correspond to a relaxation time of 220 s and a frequency of 0.3 Hz representing a duration of one period of $T = 21$ s. The autocorrelation and partial autocorrelation functions for this process simulated with 1,000 data points are shown in Fig. 11.1a, b. While the autocorrelation function decays rather slowly, the partial autocorrelation function is compatible with zero for lags larger than two. This indicates that the simulated process was an autoregressive process of order two. Here, compatible with zero means that it is below the dashed lines, the lower one indicating a pointwise 95% significance level the higher one a simultaneous 95% significance level.

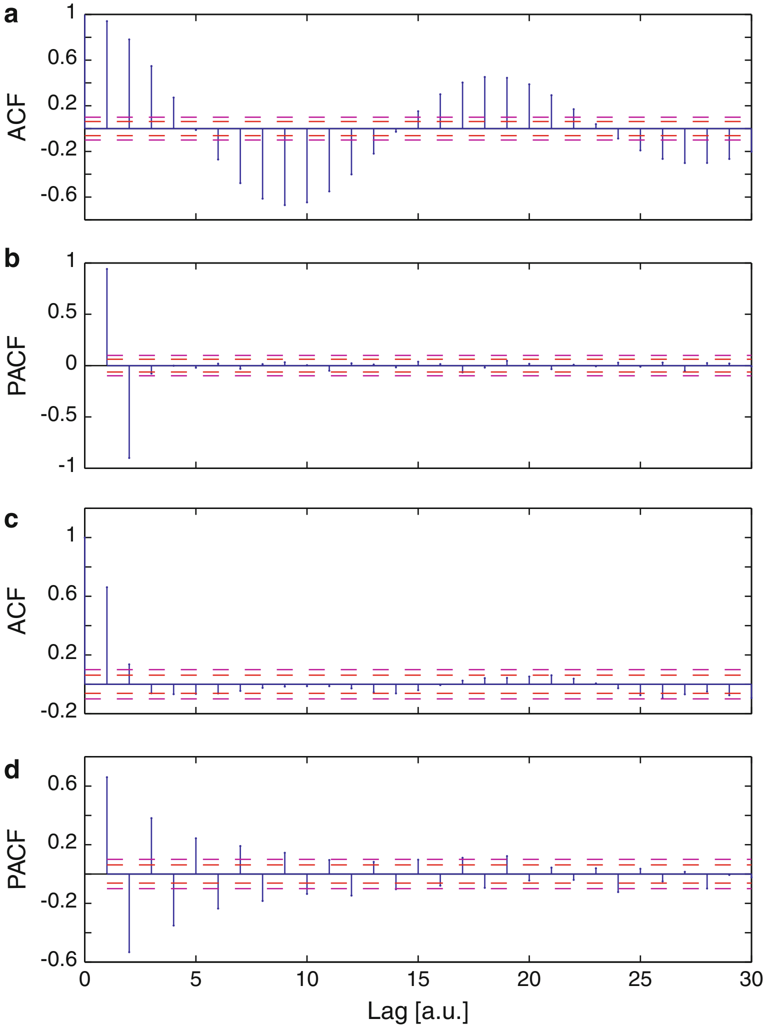


**Fig. 11.1** Autocorrelation ($C$) and partial autocorrelation $\Phi_n$ functions for an auto-regressive and a moving-average process. (**a**) $C$ for the AR process, (**b**), $\Phi_n$ for the AR process, (**c**) $C$ for the MA process, and (**d**) $\Phi_n$ for the MA process. The *vertical dashed lines* indicate the pointwise and the simultaneous 95% significance levels. The order of the processes can be estimated from the $\Phi_n$ for auto-regressive processes and from the $C$ for moving-average processes

*Remark.* A pointwise significance level is valid only for a single choice of the parameter, in this case, for a single time lag. Usually one is interested whether or not there is an effect for any of the time lags. Using the same significance level would result in false positive conclusions. This can be intuitively understood: eventually the (partial) autocorrelation function has more than one single chance to cross the significance level. In other words, if $g$ tests are performed the probability to cross the $\alpha$-significance level at least once is

$$p = 1 - (\alpha)^g. \tag{11.20}$$

For instance, if $g = 50$ which would correspond to evaluations of the (partial) autocorrelation function at 50 values of $\tau$, the probability $p$ would be almost 80% for $\alpha = 95\%$. Thus, the probability to find a significant (partial) autocorrelation function by chance is 80% and not the desired 5%.

A simultaneous significance level indeed corrects for this effect. One possible method which was utilized here, is the so-called Bonferroni correction. The Bonferroni correction is based on the idea that $\alpha$ has to be substituted by $\alpha_{\text{corr}}$ such that a desired $p = 1 - \alpha$ holds for all evaluations at certain time lags. For each individual test a higher $\alpha$-value has to be used. The Bonferroni corrected $\alpha$-value can be achieved by choosing $\alpha_{\text{corr}} = 1 - (1 - \alpha)/g$. In the above example this would lead to $\alpha_{\text{corr}} = 99.83\%$. The corrected $\alpha_{\text{corr}}$-significance level corresponds to a higher critical value as shown in Fig. 11.1.

In contrast to the results of an autoregressive process, for a moving-average process

$$x(t) = 1.8\eta(t-1) + 0.8\eta(t-2) + \eta(t) \tag{11.21}$$

the role of the autocorrelation and partial autocorrelation functions changes if the order has to be estimated. In other words, the autocorrelation function is compatible with zero for orders higher than the order of the process (Fig. 11.1c, d). This behavior is expected since the moving-average in contrast to the autoregressive process has no memory beyond the highest time lag. The partial autocorrelation function that does make explicit usage of auto-regressive processes becomes non-zero for various time lags. Thus, for autoregressive processes the natural choice for the analysis is the partial correlation function while for moving average processes autocorrelation functions are superior. This is because the autoregressive processes are causally influenced by their past values and moving average processes by the noise realization up to some finite order. The above mentioned inversion of moving-average processes substantiates the fact that a moving-average process has a slowly decaying partial autocorrelation function since the moving-average process can be interpreted as an infinite autoregressive one.

The parameter estimation can be generalized to multivariate autoregressive processes

$$\boldsymbol{X}(t) = \sum_{k=1}^{p} \underline{\alpha}_k \boldsymbol{X}(t-k) + \boldsymbol{\eta}(t), \tag{11.22}$$

$$\boldsymbol{\eta}(t) \propto \mathrm{WN}(\boldsymbol{0}, \varSigma).$$

based on Yule-Walker equations. Parameter identification for multivariate autoregressive processes is a key to Granger causality inference as discussed in the following chapter.

## 11.2 Granger Causality

Björn Schelter

In this Section, we introduce the concept of Granger-causality. In fact, investigating Granger-causality provides information about the linear predictability. Providing a definition of causality is extremely difficult and goes beyond the scope of the book. We use the term causality throughout this section, basically focussing on the operational definition of predictability.

In several applications it is of particular interest to determine the causal interactions between components of a multivariate network. In the Sect. 9.5 on cross-spectral analysis we have discussed partial coherence analysis that enables to infer whether an interaction between two components is a direct one or it is mediated by third processes. This is a first step towards causal inference but causality implies in a common sense that there is a cause and an effect. In terms of a differential equation

$$\dot{Y} = f(Y, X; \theta)$$

$X(t)$ would be considered as causal for $Y(t)$ as a change in $X(t)$ would result in a change in $Y(t)$. This in turn would again not provide the complete picture of causality in some cases as in $f(Y, X; \theta)$ the information of possibly influential third processes is missing. For causality we assume here that two ingredients are essential. Firstly, there must be a cause and an effect with the cause preceding the effect; secondly, the cause needs to influence the effect directly.

If we need to decide about causality based on finite measurements we have to face even more challenges. Firstly, based on a passive observation it is not possible to probe a possible reaction of $Y(t)$ given a change in $X(t)$ as we cannot alter $X(t)$. Secondly, the process has been observed with a particular sampling rate. If the true causal interaction acts faster than the time scale of the sampling rate, the interaction would appear as a correlation rather than a causal interaction, as we cannot determine the important temporal relation between cause and effect; a cause must occur before the effect. Thirdly, it is almost impossible to know which potentially important third processes would need to be taken into account. In some cases it is even impossible to observe all relevant processes although prior information might be available about these potentially interesting processes.

Especially this last issue is of importance. Missed observations of important processes can lead to false positive conclusions about the network structure. To motivate this, consider the following vector autoregressive process $\mathbf{X}(t)$ with components

$$\begin{aligned} X_1(t) &= \eta_1(t), \\ X_2(t) &= \eta_2(t), \\ X_3(t) &= \beta X_1(t-4) + \eta_3(t), \\ X_4(t) &= \gamma X_1(t-3) + \delta X_2(t-2) + \eta_4(t), \end{aligned} \tag{11.23}$$

where $\eta_v(t)$, $v = 1, \dots, 4$ are independently Gaussian distributed with mean zero and covariance matrix $\Sigma$ (cf. Eichler 2006). Component $X_1$ is causal for $X_3$ and $X_4$. From (11.23), additionally $X_2$ causes $X_4$ with respect to $\mathbf{X}_{\{1,2,3,4\}}$.

Imagine now, that only the three-dimensional subprocess $\mathbf{X}_{\{2,3,4\}}$ is observed one might expect in the first place that $X_2$ remains causal for $X_4$ while $X_3$ becomes isolated from the other processes. Calculations, however, show that for $\mathbf{X}_{\{2,3,4\}}$ the autoregressive representation is given by

$$\begin{aligned} X_2(t) &= \eta_2(t), \\ X_3(t) &= \tfrac{\beta\gamma}{1+\gamma^2} X_4(t-1) - \tfrac{\beta\gamma\delta}{1+\gamma^2} X_2(t-3) + \eta_3(t) - \tfrac{\beta\gamma}{1+\gamma^2}\eta_4(t-1) + \tfrac{\beta}{1+\gamma^2}\eta_1(t-4), \\ X_4(t) &= \delta X_2(t-2) + \eta_4(t) + \gamma\eta_1(t-3). \end{aligned} \tag{11.24}$$

As a result we obtained an autoregressive moving average process, where the moving average part contains unobserved contributions at various the lags. The autoregressive part contains interactions which were not present in the original system. This leads to a false positive conclusion about the directed interactions. If we could identify the lagged correlations in the moving average part, we would be able to demonstrate that important processes were not measured. We would thus get a hint, that the interaction structure which we obtain is actually containing spurious interactions.

Alternatively, one could identify unobserved components by excluding other processes in the analysis. For instance an analysis of only the bivariate subprocess $\mathbf{X}_{\{2,3\}}$ would result in two uncorrelated noise processes following a similar derivation.

In other words, missed important third influences can already for an autoregressive process change the resulting causal network. As one can hardly guarantee that all important processes are included in the analysis in applications, it would be desirable to identify unobserved processes. The two alternative approaches to identify missing processes mentioned above would enable us to do this. But it is often impossible or at least very difficult to identify autoregressive moving average processes from measured signals. Also to exclude further processes from the analysis is challenging as the possible number of exclusions increases as a factorial function. Therefore, it is quite challenging to overcome this challenge in actual applications. It should, however, be kept in mind when making interpretations about the causal interaction structure.

Without giving a formal definition, we have used the autoregressive process above and discussed causality based upon it. This has become common practice as an operational definition for causality has been suggested by Granger (1969). In contrast to the above derivations in which we discussed causality based on the coefficients of the autoregressive process, Granger based his operational definition of causality on the residual variances of the driving noise term of the autoregressive models.

### 11.2.1 Granger Causality in the Time Domain

Granger causality in the time domain is defined based on the variance of the residuals. This residual variance of the fitted autoregressive process for the component $i$ given all the other components is

$$\mathrm{var}\big(X_i(t)\,|\,\widetilde{\mathbf{X}}\big) = \mathrm{var}\big(\eta_i(t)\big) = \sigma_{ii}\,. \tag{11.25}$$

Thereby $\widetilde{\mathbf{X}} = \{\mathbf{X}(t-u), u \in \mathbb{N}\}$ denotes the past values of $\mathbf{X}(t)$. By fitting the entire multivariate autoregressive process

$$\boldsymbol{X}(t) = \sum_{k=1}^{p} \underline{\alpha}_k \boldsymbol{X}(t-k) + \boldsymbol{\eta}(t), \qquad \boldsymbol{\eta}(t) \propto \mathrm{WN}(\mathbf{0}, \Sigma). \tag{11.26}$$

to the data the best linear predictor as defined in Sect. 11.1 in (11.5) for $X_i(t)$ has been found. The obtained residual variance should be compared to the residual variance of

$$\mathbf{X}_{\backslash j}(t) = \sum_{k=1}^{p} \underline{\tilde{\alpha}}_k \boldsymbol{X}_{\backslash j}(t-k) + \boldsymbol{\eta}_{\backslash j}(t) \tag{11.27}$$

$$\boldsymbol{\eta}_{\backslash j}(t) \propto \mathrm{WN}(\mathbf{0}, \tilde{\Sigma}). \tag{11.28}$$

where the $j$-component has been removed denoted by $(\cdot)_{\backslash j}$. The residual noise process $\boldsymbol{\eta}_{\backslash j}(t)$ is again a white noise process with mean zero and covariance matrix $\tilde{\Sigma}$. Thus, the mean square prediction error for predicting $X_i(t)$ from the past values of all process $\mathbf{X}_{\backslash j}$ but $X_j(t)$ is given by

$$\mathrm{var}\big(X_i(t)\,|\,\widetilde{\mathbf{X}}_{\backslash j}\big) = \mathrm{var}\big(\boldsymbol{\eta}_{\backslash j,i}(t)\big) = \sigma_{\backslash j,ii}\,. \tag{11.29}$$

In general, the mean square prediction error in (11.29) will be larger than that in (11.25) since adding more information – in this case more processes – can only improve prediction. The two variances will be identical if and only if the best linear predictor of $X_i(t)$ based on the full past $\widetilde{\mathbf{X}}$ does not depend on the past values of $X_j$.

In Granger (1969) this is used as a definition for causality by introducing the quantity

$$\gamma_{\backslash j,i} = \ln\left(\sigma_{\backslash j,ii}/\sigma_{ii}\right).$$

If $\gamma_{\setminus j,i} > 0$ than $X_j(t)$ is Granger causal for $X_i(t)$. All the other components have been taken into account by fitting the entire autoregressive process. In other words $\gamma_{\setminus j,i} > 0$ is only the case, if the causal influence is a direct one.

It can now be proven that the following definitions of Granger causality are equivalent (Eichler 2006). Let $i$ and $j$ be two components of $\mathbf{X}(t)$. Then $X_j$ is *Granger-causal* for $X_i(t)$ with respect to $\mathbf{X}(t)$ if one of the equivalent conditions

- $\left|\mathrm{var}\big(X_i(t)|\widetilde{\mathbf{X}}\big)\right| < \left|\mathrm{var}\big(X_i(t)|\widetilde{\mathbf{X}}_{\setminus j}\big)\right|$;
- $\gamma_{\setminus j,i} \neq 0$
- $\underline{\alpha}_{ij,k} \neq 0$ for one $k \in \mathbb{N}$

holds. Statistical tests on either of these quantities provide the information of the Granger causal interactions. This enables an investigation of direct directed interactions based on measurements in actual applications.

As it is often meaningful to consider spectral information in the data, Granger causality has also been defined in the frequency domain. This will briefly be discussed in the next section.

### *11.2.2 Granger-Causality in the Frequency Domain*

In many applications, the time series of interest are characterized in terms of their frequency properties, especially if oscillatory or almost oscillatory signals have to be analyzed. It is, therefore, important to examine the relationships among multiple time series also in the frequency domain. The frequency domain analysis of stationary multivariate time series $\mathbf{X}(t)$ is based on the partial spectral representation of the autoregressive model for $\mathbf{X}(t)$ (cf. (9.127))

$$\widetilde{\mathbf{PC}}(\omega) = \left(\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\right)^H \Sigma^{-1} \left(\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\right). \tag{11.30}$$

The partial spectral matrix is a hermitean matrix and does therefore not provide any information about Granger causality although the coefficient matrices $\underline{\alpha}_k$ themselves do. Breaking the symmetry was thus used to define an asymmetric quantity, the so called partial directed coherence

$$\left|\pi_{i\leftarrow j}(\omega)\right| = \left|\frac{\left(\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\right)_{ij}}{\sqrt{\sum_k \left|\left(\mathbf{1} - \sum_{k=1}^{p} \underline{\alpha}_k \mathrm{e}^{\mathrm{i}\omega k}\right)_{ik}\right|^2}}\right| \in [0,1], \tag{11.31}$$

with $()_{ik}$ denoting the $i,k$ entry of the matrix, that enables the investigation of Granger causality in the frequency domain (Schelter et al. 2006b). With the normalization, the partial directed coherence indicates the relative strength of the effect of $X_j$ on $X_i$ as compared to the strength of the effect of $X_j$ on other processes. Thus, partial directed coherence ranks the relative interaction strengths with respect to a given signal source.

A detailed justification for this can be found in Baccala and Sameshima (2001) and a more in-depth discussion including an analysis of the corresponding statistics can be found in Schelter et al. (2006a,b). Alternative measures for Granger causality can also be found in Schelter et al. (2006b).

*Remark.* In order to estimate Granger causality from measured signals, one usually first fits an autoregressive process to the data; estimation procedures as those introduced in Sect. 11.1 are utilized to determine the coefficients, noise covariance as well as the order of the process. Once the coefficients and noise covariance are estimated, the estimators are plugged into the equations for Granger causality in the time or frequency domain.

In few cases a direct estimation of Granger causality is possible without the need to first estimate the coefficients of the autoregressive model (Jachan et al. 2009). As these procedures rely on several assumptions and are numerically quite demanding, it is, however, common practice to base the estimation on the fitting of the AR-model.

*Example.* To illustrate the performance of partial directed coherence in detecting causal influences, the following three-dimensional vector autoregressive process of order $p = 1$

$$\begin{aligned} X_1(t) &= 0.5\,X_1(t-1) + \varepsilon_1(t) \\ X_2(t) &= 0.7\,X_2(t-1) + 0.6\,X_1(t-1) + 0.4\,X_3(t-1) + \varepsilon_2(t) \\ X_3(t) &= 0.5\,X_3(t-1) + 0.3\,X_2(t-1) + \varepsilon_3(t) \end{aligned} \tag{11.32}$$

is investigated. The covariance matrix of the noise $\mathbf{\Sigma}$ is set to the identity, the number of simulated data points is $N = 10{,}000$ for each component of the vector autoregressive process. The structure of the autoregressive model can be summarized by the network in Fig. 11.2a. In this graph, a directed edge from $X_j$ to $X_i$ is drawn if $X_j$ Granger causes $X_i$. The partial directed coherence spectra $\left|\pi_{i\leftarrow j}(\omega)\right|$ can be obtained from the $i$th column and $j$th row of Fig. 11.2b. Process $X_1$ has an indirect influence on $X_3$ that is mediated by $X_2$ since, for all times $t$, the present value of $X_3(t)$ is influenced by $X_2(t-1)$ and $X_2(t)$ is influenced by $X_1(t-1)$. If $X_2$ would be blocked somehow, changes in $X_1$ would no longer affect $X_3$.

Interestingly, we have investigated a situation here, which could not have been resolved by coherence or partial coherence analysis as introduced in Sect. 9.5. Coherence would have detected a connection between all processes as there is always an interaction present, although the one between $X_1$ and $X_3$ is indirect. Partial coherence which in principle is able to infer whether or not an interaction

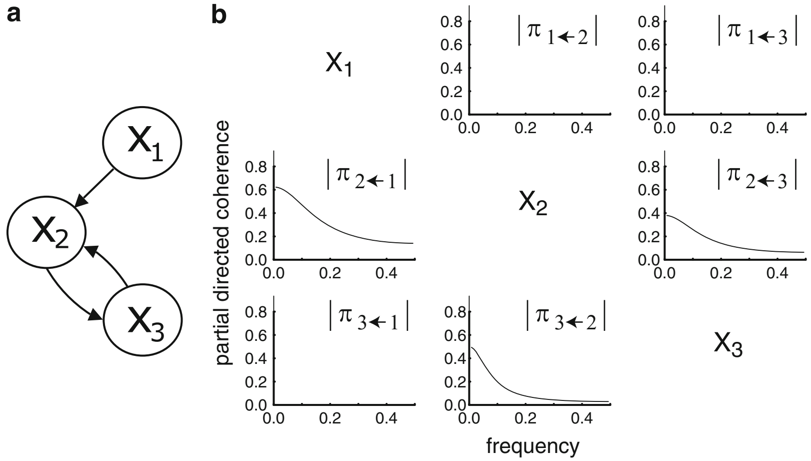


**Fig. 11.2** Graph (without self-loops) summarizing the causal influences for the example of a three-dimensional VAR[1] (**a**). Corresponding partial directed coherence (off-diagonal) (**b**). The simulated causal influences are reproduced correctly by the estimated partial directed coherence

is a direct one would have failed here because of marrying parents of a joint child, $X_1$ and $X_3$ are parents of $X_2$. Thus, partial coherence would have detected a spurious direct interaction between $X_1$ and $X_3$. As also the coherence between these two processes is significant, it would have been impossible to unmask the indirect interaction. Granger causality applying partial directed coherence is able to detect this. We emphasize here that also time domain Granger causality concepts would have detected the true network structure.

## 11.3 Hidden Systems

In Chaps. 4 and 5, we met several models for stochastic fields and processes. We first want to add an essential point concerning such models. Most systems that are described by models are not observed directly. For each type of model, therefore one has to formulate a so-called observation equation, which relates the observable quantities $y_{1\ldots N} = (y_1, y_2, \ldots, y_N)$ to the system quantities $x_{1\ldots M} = (x_1, \ldots, x_M)$ described by the model. Such an equation will also contain the unavoidable error of measurement or observation. It may be formulated as a relation between the random variables $\{X_i\}$ and $\{Y_j\}$, e.g., in the form

$$Y_i = \sum_{j=1}^{M} A_{ij} X_j + E_i, \tag{11.33}$$

where $E_i$ represents the error of measurement, which in many cases can be regarded as normally distributed. However, a relationship between the observables and the system quantities may also be defined by a suitable conditional probability $\varrho(y_{1\ldots N}\,|\,x_{1\ldots M})$.

Examples of models expressing such relationships between system quantities and observables are

- A hidden Markov model: The time dependence of the state $z(t)$ is described by a master equation (Sect. 5.2). In this case the observation equation can often be written in the form

  $$y(t) = f(z(t)) + \sigma(t)\eta(t), \quad \eta(t) \propto \mathrm{WN}(0,1). \tag{11.34}$$

  Hence, the observable $f(z(t))$ depends on the state $z(t)$. The measured quantity is $y(t)$ which includes a superimposed noise term.

  For discrete-time state transformations, the model for the time dependence may also be an ARMA model, e.g., an AR(2) model of the form

  $$z(t) = \alpha_1 z(t-1) + \alpha_2 z(t-2) + \sigma_1 \eta(t), \quad \eta(t) \propto \mathrm{WN}(0,1). \tag{11.35}$$

  In Sect. 5.9.4 we have seen that every ARMA(p,q) model can be written as a multivariate AR(1) model:

  $$\boldsymbol{X}(t) = \mathsf{A}\,\boldsymbol{X}(t-1) + \mathbf{B}\eta(t), \quad \eta(t) \sim WN(0,\sigma^2), \tag{11.36}$$

  This system equation together with an observation equation constitutes what is known as a state space model.
- Suppose the time dependence of a state $z(t)$ of a system is described by a set of differential equations. Even in this case the states are usually not observed directly, and an additional relationship between the state variables $z(t)$ and the measured quantities $y(t)$ has to be formulated in the form of an observation equation (11.34).
- Let $\boldsymbol{x}$ be a spatial random field, defined by a Gibbs function (Chap. 4), for example,

  $$\varrho(\boldsymbol{x}) = \mathrm{e}^{-\beta H(\boldsymbol{x})} \quad \text{with} \quad H(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i x_i + \sum_{i=1}^{N} \sum_{j \in N_i} J_{ij} x_i x_j\,. \tag{11.37}$$

  A possible observation equation for this case is

  $$y_i = \sum_j K_{ij} x_j + \varepsilon_i\,. \tag{11.38}$$

  In the theory of image processing, such relationships describe the connection between the “true” image and the observed image. Images are usually corrupted by noise.

- Let $x(\tau)$ be a distribution function, e.g., for the size of particles or density of matter. The measurement of such distributions can, in nearly all cases, be done only indirectly by measuring, say, the intensity of scattered or absorbed light. Then an equation such as

$$y(t) = \int \mathrm{d}\tau \, \mathrm{K}(t, \tau) \, x(\tau), \tag{11.39}$$

  represents the relationship between the observable $y(t)$ and the quantity $x(\tau)$ which is not measurable directly. To infer the function $x(\tau)$ from the function $y(t)$ involves solving an integral equation of the first kind. This inference of the function $x(\tau)$ based on the data $\{y(t_i),\ i = 1, \dots, N\}$, taking into account the error of measurements $\{\sigma(t_i)\,\varepsilon(t_i)\}$, is called an inverse problem.

  Discretizing the integral and displaying the experimental errors one obtains, choosing $M$ points $\tau_j, j = 1, \dots, M$ and setting $x(\tau_j) = x_j$

$$y_i \equiv y(t_i) = \sum_{j=1}^{M} \mathrm{K}_{ij} x_j + \sigma_i \, \varepsilon_i, \tag{11.40}$$

  with $y_i \equiv y(t_i)$, $\sigma_i \equiv \sigma(t_i)$, $\varepsilon_i \equiv \varepsilon(t_i)$, and $\mathrm{K}_{ij} \equiv \mathrm{K}(t_i - \tau_j)\, m_j$, where $m_j$ is a factor following from the discretization.

  Thus an observation equation of the kind (11.38) emerges. But now there is no equation for the system quantities $x_{1\dots M}$. But we will see in the next section that one cannot solve such an inverse problem without a substitute for the system equation.

Within such a framework, three fundamental questions arise in model-based data analysis:

1. Estimation of the hidden realization: How can we estimate the realization $x_{1\dots M}$ of the fundamental process from the given data $y_{1\dots N}$ and a given model structure? This question arises in a hidden Markov model, where one is interested in the realization $x_{1\dots N}$ of the unobservable process $X(t)$, given the observation $y_{1\dots N}$; this question arises in any kind of the inverse problem, e.g., in image restoration, where one wants to estimate the true image $x_{1\dots M}$ lying behind the image $y_{1\dots N}$ which is corrupted by noise. This problem also occurs frequently in space technology, where the position of a space ship has to be determined on the basis of the received signals. Methods for estimating hidden realizations are introduced in this chapter.
2. Parameter identification: How can we estimate the parameters of a model from the given data $y_{1\dots N}$ and a given model structure? This includes the estimation of the parameters of the fundamental model as well as of the parameters of the observation equation. This will be discussed in Chap. 12.
3. Model selection: How can we decide, on the basis of the data, which model structure is most appropriate? Methods for the problem of model selection will be discussed in Chap. 13.

## 11.4 The Maximum a Posteriori Estimator for the Inverse Problem

In this section we introduce methods for estimating a distribution function $x(\tau)$, which is related to an experimentally accessible quantity $y(t)$ by an integral equation of the first kind. The most general estimator for this inverse problem is the maximum a posteriori estimator, already introduced in Chap. 10. First of all we will adapt this estimator to the inverse problem.

From the Bayes rule (Sect. 2.2), we obtain the probability of a realization $x_{1 \ldots M}$ under the condition that a certain set of data $y_{1 \ldots N}$ is given:

$$\varrho(x_{1 \ldots M} | y_{1 \ldots N}) = \frac{\varrho(y_{1 \ldots N} | x_{1 \ldots M})\, \varrho(x_{1 \ldots M})}{\varrho(y_{1 \ldots N})}. \tag{11.41}$$

For a given set of data, the distribution $\varrho(y_{1 \ldots N})$ is a constant with respect to $x_{1 \ldots M}$, while $\varrho(y_{1 \ldots N} | x_{1 \ldots M})$ depends on the parameters of the observation equation, and $\varrho(x_{1 \ldots M})$ depends on the parameters of the process. $\varrho(x_{1 \ldots N} | y_{1 \ldots N})$ is the a posteriori distribution.

The maximum a posteriori estimator (MAP estimator) is equal to that "path" $\widehat{x}_{1 \ldots N}$ for which the a posteriori distribution $\varrho(x_{1 \ldots M} | y_{1 \ldots N})$ is maximum. Hence,

$$\widehat{x}_{1 \ldots N} = \arg \max_{x_{1 \ldots M}} \{\varrho(y_{1 \ldots N} | x_{1 \ldots M})\, \varrho(x_{1 \ldots M})\}. \tag{11.42}$$

If we represent the conditional density $\varrho(y_{1 \ldots N} \mid x_{1 \ldots M})$, with which the observation equation is defined, in the form

$$\varrho(y_{1 \ldots N} \mid x_{1 \ldots M}) = \mathrm{e}^{-U_B(y_{1 \ldots N} | x_{1 \ldots M})}, \tag{11.43}$$

and if we represent the prior probability $\varrho(x_{1 \ldots M})$, which enters in the formulation of the model, in the form

$$\varrho(x_{1 \ldots M}) = \mathrm{e}^{-\beta\, U_M(x_{1 \ldots M})}, \tag{11.44}$$

then we can also write the MAP estimator

$$\hat{x}_{1 \ldots N} = \arg \min_{x_{1 \ldots M}} \big\{U_B(y_{1 \ldots N} \mid x_{1 \ldots M}) + \beta\, U_M(x_{1 \ldots M})\big\}. \tag{11.45}$$

Here we have extracted a factor $\beta$ from the energy function of the prior density, which we may use to regulate the variance of the distribution of paths or images $x_{1 \ldots M}$ around the most probable realization. If the data are not taken into account, then the most probable realization $x_{1 \ldots M}$ is that which minimizes $U_M(x_{1 \ldots M})$. By minimizing $U_B + \beta\, U_M$, we also take into account the data, and the parameter $\beta$ controls the relative weight with which this prior knowledge enters into the estimation. The smaller $\beta$ is, the larger the variance of the prior distribution of the paths or images $x_{1\ldots M}$, and the greater the difference between the estimation $\hat{x}_{1\ldots N}$ on the basis of the data. The most probable prior realization $x^0_{1\ldots M}$ is given by

$$x^0_{1\ldots M} = \arg\min_{x_{1\ldots M}} U_M(x_{1\ldots M}), \tag{11.46}$$

i.e., the less the significance of prior knowledge in the estimation. For $\beta = 0$, this prior knowledge does not enter at all.

Use of the MAP estimator to solve the inverse problem implies that $\hat{x}_{1\ldots N}$ and $\hat{x}(\tau)$ are defined, respectively, by

$$\hat{x}_{1\ldots M} = \arg\min_{x_{1\ldots M}} \left\{ \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} \left( y_i - \sum_{j=1}^{M} \mathrm{K}_{ij} x_j \right)^2 + \beta\, U_M(x_{1\ldots M}) \right\}$$

and

$$\hat{x}(\tau) = \arg\min_{\{x(\tau)\}} \left\{ \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} \left( y_i - \int \mathrm{d}\tau\; \mathrm{K}(t_i, \tau)\, x(\tau) \right)^2 + \beta\, U_M\big(x(\tau)\big) \right\} .$$

For a complete definition of the estimator we still have to specify

- The prior energy function $U_M(x_{1\ldots M})$, i.e., the prior knowledge about the probabilities $\varrho(x_{1\ldots M})$ for the individual paths $x_{1\ldots M}$, irrespective of the data;
- The parameter $\beta$, which regulates the strength of the variance in this probability distribution and therefore determines how much the prior knowledge enters into this estimation.

### 11.4.1 The Least Squares Estimator as a Special MAP Estimator

At first sight, the choice $\beta = 0$ seems to be favorable, since in this case the prior energy function does not have to be specified at all. For the MAP estimator, all $x_{1\ldots M}$ now have the same probability (i.e., $\varrho(x_{1\ldots M})$ is a constant). This also means that no prior information whatsoever enters into the estimation. The MAP estimator has been reduced to the least-squares estimator.

The least-squares estimator was introduced in Sect. 8.5. We have seen that the variance of this estimator is proportional to the inverse of the singular values of the matrix, and therefore becomes larger, the smaller these singular values. The usefulness of the least-squares estimator thus depends strongly on the conditioning of the matrix $\{\mathsf{K}_{ij}\}$ (see Fig. 11.3).

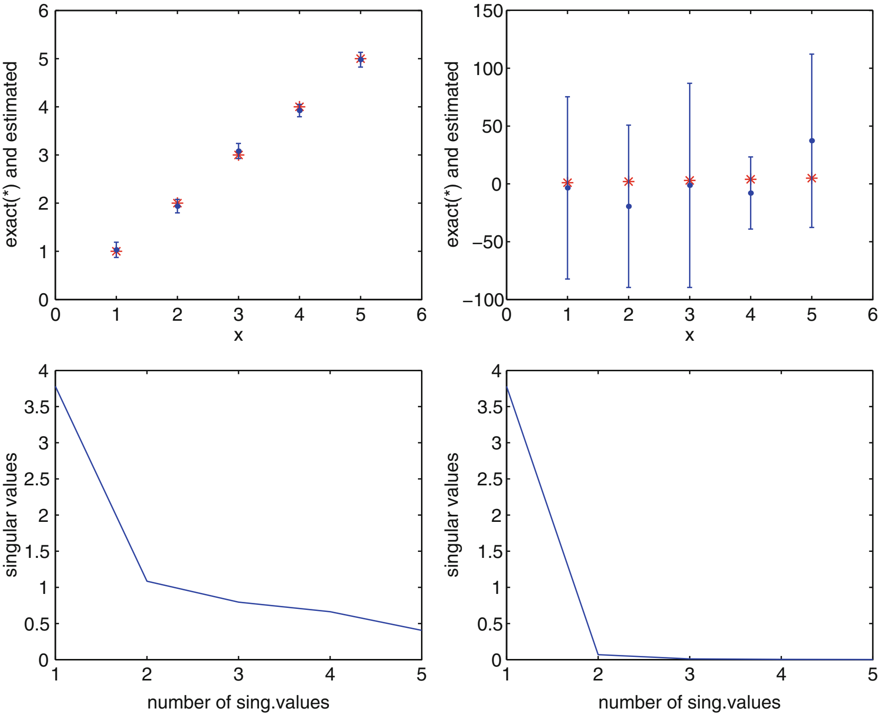


**Fig. 11.3** True values, marked by an asterisk, and estimated values with confidence interval (error bars) for a problem with a well-conditioned matrix (*left*) and with an ill-conditioned matrix (*right*). The ordered series of singular values is shown for each case in the *bottom* subfigure

If this matrix has been obtained from an observation equation of the form

$$y(t_i) = \int \mathrm{d}\tau\, \mathsf{K}(t_i - \tau)\, x(\tau) + \varepsilon(t_i), \quad i = 1, \dots, N, \tag{11.47}$$

by discretizing the integral, and therefore

$$\mathsf{K}_{ij} \sim \mathsf{K}(t_i - \tau_j), \quad i = 1, \dots, N, \quad j = 1, \dots, M, \tag{11.48}$$

then this matrix $\mathsf{K}$ with elements $\mathsf{K}_{ij}$ is certainly ill-conditioned, and the conditioning will become worse, the larger the number of points $\tau_j$, $j = i, \dots, M$, for which $x(\tau_j)$ shall be determined within the approximation of this discretization. Indeed, the larger the value of $M$, the closer neighboring $\tau_j$'s (i.e., $\tau_j$ and $\tau_{j+1}$) move together, and the more similar are the columns $\{\mathsf{K}_{ij}\ , i = 1, \dots, N\}$ and $\{\mathsf{K}_{i,j+1}\ , i = 1, \dots, N\}$ of the matrix. But this implies that the matrix $\mathsf{K}$ becomes "more singular" and therefore the singular values are smaller.

Hence, the least-squares estimator is in general inappropriate for the deconvolution problem. The inverse problem, which corresponds to the solution of an integral equation, is therefore also called an "ill-posed" problem. (For a mathematical definition and treatment of ill-posed problems see also Tikhonov and Arsenin 1977; Groetsch 1984; Morozov 1984). Without prior knowledge, it is not possible to estimate the function $x(\tau)$ on the basis of a finite set of data.

### *11.4.2 Strategies for Choosing the Regularization Parameter*

Having chosen a prior energy functional $U_M[x(\tau)]$, we can first determine the MAP estimator $\hat{x}_{1\ldots M}\big|_\beta$ for a given $\beta$ according to

$$\hat{x}_{1\ldots M}\Big|_\beta = \arg\min_{x_{1\ldots M}} \left\{ \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} \left( y_i - \sum_{j=1}^{M} \mathrm{K}_{ij} x_j \right)^2 + \beta\, U_M(x_{1\ldots M}) \right\} .$$

We now may vary the parameter $\beta$ and investigate the various resulting $\hat{x}_{1\ldots M}\big|_\beta$. In practice we will often find a region of $\beta$ values where we expect the "optimal value" for $\beta$. However, first we have to define what has to be understood by an "optimal value" for $\beta$. This we do by formulating a strategy based on mathematical criteria to determine this parameter. Two possible criteria shall be presented:

**The discrepancy method.** This method is based on the assumption that the realization of the observational noise $\varepsilon_i$ in the observed value $y_i$ is of the order $\sigma_i$, and therefore the discrepancy for each observed value $y_i$, given by

$$\frac{1}{\sigma_i^2} \left( y_i - \sum_{j=1}^{M} \mathrm{K}_{ij}\, x_j \right)^2 , \tag{11.49}$$

should be of order 1.

Therefore, $\beta$ is chosen such that for the total discrepancy one gets

$$\mathrm{D}(\hat{x}_{1\ldots M}\big|_\beta \mid y_{1\ldots M}) \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left( y_i - \sum_{j=1}^{M} \mathrm{K}_{ij}\, \hat{x}_j\big|_\beta \right)^2 = N. \tag{11.50}$$

In Fig. 11.4, the total discrepancy for some problem is shown as function of $\beta$. Obviously, the least-squares solution belongs to a much smaller total discrepancy. For large enough values of $M$, this can even become zero.

**The self-consistent method.** Let us suppose for the moment that the true solution $x_{1\ldots M}$ is known, as, for instance, in the simulation of data $y_{1\ldots N}$ for known $x_{1\ldots M}$

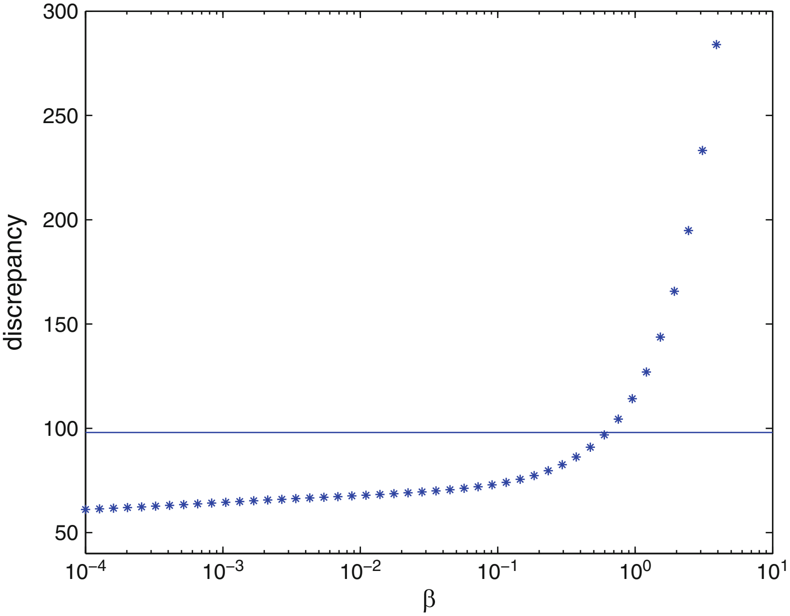


**Fig. 11.4** Typical course of the total discrepancy as a function of the regularization parameter $\beta$. The value of $N$, the number of data points, is indicated by the *straight line*

according to the observation equation. Obviously, now we may evaluate the mean square discrepancy

$$\mathrm{D}(\beta) \equiv \left\langle (x_{1\ldots M} - \hat{X}_{1\ldots M}\big|_{\beta,Y_{1\ldots N}})^2 \right\rangle \tag{11.51}$$

of the estimator $\hat{X}_{1\ldots M}\ |_{\beta,Y_{1\ldots N}}$, where $X_{1\ldots M}$ and $Y_{1\ldots N}$ have been denoted by large letters to indicate that they have to be regarded as random variables.

Now we might choose $\beta$ such that $\mathrm{D}(\beta)$ is minimum. Hence, the optimal value $\beta^*$ would be determined by

$$\frac{\partial}{\partial\beta}\,\mathrm{D}(\beta)\Big|_{\beta=\beta^*} = \frac{\partial}{\partial\beta}\left\langle \left(x_{1\ldots M} - \hat{X}_{1\ldots M}\big|_{\beta,Y_{1\ldots N}}\right)^2 \right\rangle\Big|_{\beta=\beta^*} = 0.$$

In practice, however, the true solution $x_{1\ldots M}$ is never known. Nevertheless, instead of the true solution we may insert the "optimal estimation" $\hat{x}_{1\ldots M}\ |_{\beta^*,y_{1\ldots N}}$, i.e., the one determined with the parameter $\beta^*$, which is a candidate for an optimal parameter, and thereby obtain a criterion for $\beta^*$:

$$\frac{\partial}{\partial\beta}\,\mathrm{D}'(\beta^*,\beta)\,\Big|_{\beta=\beta^*} = \frac{\partial}{\partial\beta}\left\langle \left(\hat{x}_{1\ldots M}\big|_{\beta^*,y_{1\ldots M}} - \hat{X}_{1\ldots M}\big|_{\beta,Y_{1\ldots N}}\right)^2 \right\rangle\Big|_{\beta=\beta^*} = 0. \tag{11.52}$$

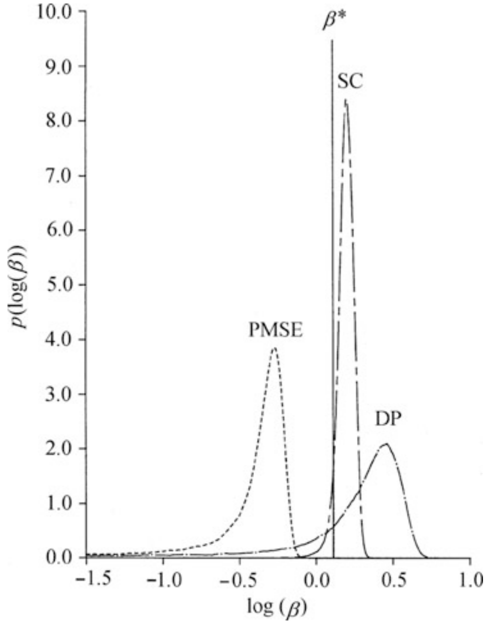


**Fig. 11.5** Distributions of the regularization parameter for various strategies (*DP* discrepancy method, *SC* self-consistent method, *PMSE* predictive mean square signal error) (From Honerkamp and Weese (1990))

This expectation value $\mathrm{D}'(\beta^*, \beta)$ first has to be computed explicitly as a function of $\beta$ and $\beta^*$. This is possible for those cases where $X_{1\ldots M}\ \left|_{\beta, Y_{1\ldots N}}\right.$ depends linearly on the random variables $Y_{1\ldots N} = \{Y_1, Y_2, \ldots, Y_N\}$. Even in the general case, however, we may in principle estimate the expectation value for all pairs $(\beta, \beta^*)$ by the following steps: First we choose a $\beta^*$ and determine a solution $\hat{x}_{1\ldots M}\left|_{\beta^*, y_{1\ldots N}}\right.$, and with this solution we generate a sample $\{y^{(1)}_{1\ldots N}, y^{(2)}_{1\ldots N}, \ldots\}$ of data using the observation equation. For each element $y^{(\alpha)}_{1\ldots N}$ of this sample, we use the parameter $\beta$ to generate a realization of $X_{1\ldots N}\left|_{\beta, Y_{1\ldots N}}\right.$, and finally we estimate the expectation value from all of these realizations. This, however, is extremely time-consuming, and one should always try to find at least an approximate expression for $\mathrm{D}'(\beta^*, \beta)$.

The quality of such strategies for determining the regularization parameter $\beta$ may be investigated by Monte Carlo methods:

For $x_{1\ldots M}$ given, one simulates a sample of data sets $\{y^{(1)}_{1\ldots N}, y^{(2)}_{1\ldots N}, \ldots\}$, from which the quantity $\mathrm{D}(\beta)$ may be estimated. It is then minimized with respect to $\beta$ yielding the optimal value $\beta^*$. For the different strategies one also determines the corresponding regularization parameter for each data set $y^{(\alpha)}_{1\ldots N}$, $\alpha = 1, 2, \ldots$ of the sample. For each strategy one thereby obtains a distribution of $\beta$. The closer the maximum of this distribution comes to $\beta^*$ and the narrower the distribution, the better and more stable is the corresponding strategy. Figure 11.5 shows $\beta^*$, together with the distributions, $p(\ln(\beta))$, for the discrepancy method, the self-consistent method, and a further method which will not be described here. It can be seen that the self-consistent method performs best.

*Remark.* Until to now we have assumed a linear relationship of the form

$$y(t) = \int \mathrm{d}\tau \; \mathsf{K}(t,\tau)\, x(\tau) \tag{11.53}$$

between the measured quantity $y(t)$ and the distribution $x(\tau)$ to be determined. As a consequence, the form of $\varrho(y_{1\dots N} \mid x_{1\dots M})$ turned out to be relatively simple – it was bilinear in $x_{1\dots M}$ – and therefore the constitutive equation for the estimator $\hat{x}_{1\dots M}\big|_{\beta}$ was linear. If there is a nonlinear dependence on $x(\tau)$, as, for example, in light scattering, where a relationship of the form

$$y(t) = a \left( \int \mathrm{d}\tau \; \mathsf{K}(t,\tau)\, x(\tau) \right)^2 + b \tag{11.54}$$

constitutes the direct problem, the prescription for constructing the estimator does not have to be changed. The computation of the minimum, however, will be numerically much more elaborate, as will be the implementation of the self-consistent method for determining the regularization parameter (Honerkamp et al. 1993).

### *11.4.3 The Regularization Method*

This section is dedicated to the choice of the prior energy function $U_M(x_{1\dots M})$. In general, it is not possible to formulate a precise model for the distribution $x(\tau)$ or $x_{1\dots M}$. In most cases, one would like the function $x(\tau)$ or the path $x_{1\dots M}$ to satisfy certain conditions, and $U_M(x_{1\dots M})$ is then constructed so that those paths $x_{1\dots M}$ for which these conditions hold have the smallest energy (i.e., the largest probability).

Five frequently used energy functions are the following:

1.
$$U_M(x_{1\dots M}) \;=\; \sum_{i=1}^{M} x_i^2 \tag{11.55}$$

or

$$U[x(\tau)] = \int \mathrm{d}\tau \; (x(\tau))^2 \, . \tag{11.56}$$

In these cases, it is assumed that all $x_i$ are mutually independent and normally distributed. The most probable configuration is $x_i = 0$ for all $i$.

2. $$U_M(x_{1\ldots M}) = \sum_{i=1}^{M-1} \left(\frac{x_{i+1} - x_i}{\varepsilon}\right)^2 \qquad (11.57)$$

with a suitably chosen $\varepsilon$, or

$$U_M[x(\tau)] = \int \mathrm{d}\tau \left(x'(\tau)\right)^2 . \qquad (11.58)$$

Now one assumes a correlation between the $x_i$. $U_M[x(\tau)]$ is minimum for $x'(\tau) \equiv 0$.

The probability of configurations or paths $\{x(\tau)\}$ becomes larger, as the average steepness of $x(\tau)$ becomes smaller. If there were no data, $\hat{x}(\tau) = \text{const}$ would be the most probable configuration. Taking the data into account in the functional to be minimized leads to solutions for which the path varies only slightly, predominantly such that $x_{i+1} - x_i < \varepsilon$.

3. Another possible ansatz is

$$U_\mathrm{M}[x(\tau)] = \int \mathrm{d}\tau \ \Psi\big(x'(\tau)\big), \qquad (11.59)$$

where $\Psi\big(x'(\tau)\big)$ may be some nonlinear function of the first derivative of $x(\tau)$, which, however, does not increase without limit for $|x'| \to \infty$. In this case, sufficiently large fluctuations in the data may lead to jumps in the estimation. Therefore, if one expects a piecewise constant function $x(\tau)$ occasionally interrupted by jumps, an energy function of this type is a good choice.

4. Let $x_0(\tau)$ be some given distribution, or $x^0_{1\ldots M}$ be some path. The relative entropy of $x(\tau)$ with respect to $x_0(\tau)$ is defined as (see Sect. 2.4)

$$S[x \mid x_0] = -\int \mathrm{d}\tau \ x(\tau) \ln \left(\frac{x(\tau)}{x_0(\tau)}\right) . \qquad (11.60)$$

In particular, $S[x_0 \mid x_0] = 0$, and $S[x \mid x_0] < 0$ for all $x(\tau) \neq x_0(\tau)$. Hence, for some suitable constant $c > 0$,

$$U_M[x(\tau)] = -c \ S[x \mid x_0] \qquad (11.61)$$

is an energy function which assumes its minimum for $x(\tau) \equiv x_0(\tau)$ or $x_{1\ldots M} \equiv x^0_{1\ldots M}$. The use of this energy function in treating the inverse problem is called the maximum entropy method.

5. For the choice

$$U_M[x(\tau)] = \int \mathrm{d}\tau \ \left(\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} x(\tau)\right)^2 \qquad (11.62)$$

or

$$U_M(x_{1\ldots M}) = \sum_{i=2}^{M-1} \left( \frac{x_{i+1} - 2x_i + x_{i-1}}{\varepsilon} \right)^2 , \tag{11.63}$$

with some suitable constant $\varepsilon$, one obviously expects the solution to have small second derivatives, i.e., the solution should be as smooth as possible. Of course, in analogy to (11.59), also the more general ansatz

$$U_{\mathrm{M}}[x(\tau)] = \int \mathrm{d}\tau \; \Psi\big(x''(\tau)\big), \tag{11.64}$$

e.g.,

$$U_{\mathrm{M}}[x(\tau)] = \int \mathrm{d}\tau \frac{(x''(\tau))^2}{\sqrt{1 + (\alpha x''(\tau))^2}} \tag{11.65}$$

with some appropriately chosen parameter $\alpha$ would be possible, which is appropriate when one expects a piecewise smooth function $x(\tau)$, occasionally interrupted by sharp peaks (Roths et al. 2000).

If we can write $U_{\mathrm{M}}$ in the form

$$U_{\mathrm{M}}[x(\tau)] = \int \mathrm{d}\tau \;\; [\mathcal{L}\, x(\tau)]^2 , \tag{11.66}$$

where $\mathcal{L}$ is some differential operator, then

$$U_{\mathrm{M}}(x_{1\ldots M}) = \big(\boldsymbol{\mathcal{L}}\, \boldsymbol{x}\big)^2 \tag{11.67}$$

with $\boldsymbol{x} = (x_1, \ldots, x_M)$, and where $\boldsymbol{\mathcal{L}}$ is a matrix determined by the discretization of the differential operator $\mathcal{L}$. If we further write $\boldsymbol{b} = (y_1/\sigma_1, \ldots, y_N/\sigma_N)$, $\mathsf{K}'_{ij} = \mathsf{K}_{ij}/\sigma_i$, then also

$$\hat{x}_{1\ldots M} \;|_{\beta} = \arg \min_{x_{1\ldots M}} \left\{ (\boldsymbol{b} - \mathsf{K}'\boldsymbol{x})^2 + \beta\, (\boldsymbol{\mathcal{L}}\boldsymbol{x})^2 \right\} , \tag{11.68}$$

and the minimum may be obtained analytically: From

$$V(\boldsymbol{x}) = (\boldsymbol{b} - \mathsf{K}'\, \boldsymbol{x})^2 + \beta\, (\boldsymbol{\mathcal{L}}\, \boldsymbol{x})^2 \tag{11.69}$$

$$= \boldsymbol{b}^2 - 2\, \boldsymbol{b}^{\mathrm{T}}\, \mathsf{K}'\, \boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\, (\mathsf{K}'^{\mathrm{T}}\mathsf{K}' + \beta\, \boldsymbol{\mathcal{L}}^{\mathrm{T}}\boldsymbol{\mathcal{L}})\, \boldsymbol{x} , \tag{11.70}$$

we now get

$$\hat{x}_{1\ldots M} \;|_{\beta} = \mathsf{K}'^{+}(\beta)\, \boldsymbol{b}, \tag{11.71}$$

with

$$\mathsf{K}'^{+}(\beta) = \left(\mathsf{K}'^{\mathrm{T}}\mathsf{K}' + \beta \boldsymbol{\mathcal{L}}^{\mathrm{T}}\boldsymbol{\mathcal{L}}\right)^{-1} \mathsf{K}'^{\mathrm{T}}. \tag{11.72}$$

Hence, for $\beta = 0$ we obtain

$$\mathsf{K}'(\beta = 0) = \left(\mathsf{K}'^{\mathrm{T}}\mathsf{K}'\right)^{-1} \mathsf{K}'^{\mathrm{T}}, \tag{11.73}$$

i.e., $\mathsf{K}'^{+}(\beta = 0)$ is just the pseudo-inverse of the matrix $\mathsf{K}'$ (see Sect. 8.5).

When the operator $\mathcal{L}$ represents the identity, i.e., $\boldsymbol{\mathcal{L}}$ is the identity matrix, we obtain from the singular value decomposition of $\mathsf{K}'$,

$$\mathsf{K}' = \sum_{j=1}^{M} w_j\, \boldsymbol{u}_j \otimes \boldsymbol{v}_j, \tag{11.74}$$

where $w_j$ denote the singular values and $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ the eigenvectors of $\mathsf{K}'^{\mathrm{T}}\mathsf{K}'$ and $\mathsf{K}'\mathsf{K}'^{\mathrm{T}}$, respectively, the corresponding representation of $\mathsf{K}'^{+}(\beta)$:

$$\mathsf{K}'^{+}(\beta) = \sum_{j=1}^{M} \frac{w_j}{w_j^2 + \beta}\, \boldsymbol{v}_j \otimes \boldsymbol{u}_j. \tag{11.75}$$

If $\mathsf{K}'$ is ill-conditioned, i.e., some of the singular values $w_j$ are very small, the prefactors $w_j/(w_j^2 + \beta)$ are not as large for $\beta \neq 0$ as they are for $\beta = 0$. Therefore, the divergency of the prefactors of $\mathsf{K}'^{+}(\beta)$ for $w_j \to 0$ is avoided if we choose $\beta \neq 0$.

This is also called a regularization, and the estimation of $x_{1\ldots M}$, which we have presented in the present context as a Bayes estimation, has also been introduced as a regularization procedure independent of any Bayes estimation (Tikhonov and Arsenin 1977; Groetsch 1984; Morozov 1984). The parameter $\beta$ is also called a regularization parameter.

### *11.4.4 Examples of the Estimation of a Distribution Function by a Regularization Method*

We discuss an example taken from rheology. Linear viscoelastic materials, for example, polymer melts, are materials that are characterized by a spectrum of relaxation times which correspond to different relaxation processes in the material. A density function $h(\tau)$ ($h(\tau) \geq 0$) specifies the relative weight with which a relaxation process with a characteristic relaxation time $\tau$ contributes to an experimentally accessible quantity. This function is also called a relaxation-time spectrum.

According to the theory of linear viscoelastic materials, a linear relationship exists between the response variables of the system that describes its behavior with respect to deformations and the spectrum $h(\tau)$. If the material is exposed to an oscillating shear deformation, two so-called dynamical moduli $G'(\omega)$ and $G''(\omega)$ may be measured (essentially these are stress components measured per unit area). They are related to the relaxation-time spectrum $h(\tau)$ by

$$\begin{aligned} G'(\omega) &= \int \mathrm{d}\ln\tau\, K'(\omega,\tau)\, h(\tau), \\ G''(\omega) &= \int \mathrm{d}\ln\tau\, K''(\omega,\tau)\, h(\tau), \end{aligned} \tag{11.76}$$

where

$$\begin{aligned} K'(\omega,\tau) &= \frac{(\omega\tau)^2}{1+(\omega\tau)^2}, \\ K''(\omega,\tau) &= \frac{\omega\tau}{1+(\omega\tau)^2}. \end{aligned} \tag{11.77}$$

Hence, from the spectrum $h(\tau)$, i.e., a density distribution, two measurable quantities $G'(\omega)$ and $G''(\omega)$ are derived.

To demonstrate the quality of the maximum entropy method and the regularization procedure, we will now determine the spectrum not from some experimental data, but instead we will postulate a certain spectrum, from which we will generate a set of data by simulations, and then we will try to reconstruct this spectrum. We choose a bimodal spectrum of the form

$$h(\tau) = A_1 \exp\left(-\frac{\left(\ln\frac{\tau}{\tau_a}\right)^2}{b_1^2}\right) + A_2 \exp\left(-\frac{\left(\ln\frac{\tau}{\tau_b}\right)^2}{b_2^2}\right), \tag{11.78}$$

with $A_1 = 10, \tau_a = 5\times10^{-3}, b_1 = 1; A_2 = 0.5, \tau_b = 50, b_2 = 1$.

On a logarithmic $\tau$-scale, $h(\tau)$ corresponds to the superposition of two Gaussians with their maxima at $\tau_a = 5\times10^{-3}$ and $\tau_b = 50$, where the first maximum is larger than the second by a factor of 20.

The data are now simulated according to

$$\begin{aligned} g_i'^{\sigma} &= G'(\omega_i)\,(1+\sigma_0\eta_i'), \\ g_i''^{\sigma} &= G''(\omega_i)\,(1+\sigma_0\eta_i''), \qquad i = 1,\ldots,N, \end{aligned} \tag{11.79}$$

where

$$\omega_i = \omega_{\min}\left(\frac{\omega_{\max}}{\omega_{\min}}\right)^{\frac{i-1}{N-1}},\ i = 1,\ldots,N, \tag{11.80}$$

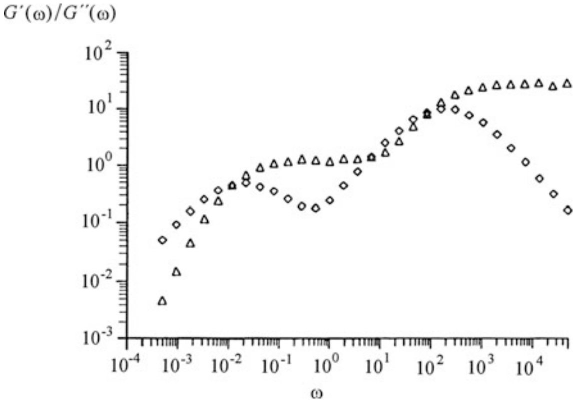


**Fig. 11.6** Simulated data for $G'(\omega)$ ($\triangle$) and $G''(\omega)$ ($\diamond$) from the spectrum given by (11.78) and shown as a *solid line* in Fig. 11.7 (From Honerkamp (1994), reprinted by permission of Wiley-VCH, Weinheim)

$$\omega_{\min} = 5 \times 10^{-4}, \quad \omega_{\max} = 5 \times 10^{4}, \quad N = 30. \tag{11.81}$$

Here, $G'(\omega_i)$ and $G''(\omega_i)$ have been computed from (11.76), where the integrals have been replaced by summations, e.g.,

$$G'(\omega_i) = \sum_{\alpha=1}^{M} m_\alpha K'(\omega_i, \tau_\alpha) h_\alpha \tag{11.82}$$

with

$$h_\alpha = h(\tau_\alpha), \quad \tau_\alpha = \tau_a \left(\frac{\tau_b}{\tau_a}\right)^{\frac{\alpha-1}{M-1}} \tag{11.83}$$

and

$$m_\alpha = \frac{\ln \frac{\tau_b}{\tau_a}}{M-1}, \quad \alpha = 1, \ldots, M. \tag{11.84}$$

The quantities $\eta_i'$ and $\eta_i''$ in (11.79) are both standard normal random numbers. The form of the simulation of the errors of measurement in (11.79) is equivalent to assuming a constant relative error $\sigma_0$. The value was $\sigma_0 = 0.03$, i.e., we assumed a relative error of 3%.

The simulated measurement data are presented in Fig. 11.6. Figure 11.7 shows the given spectrum $h(\tau)$ as a solid line and the reconstructed values with error bars obtained from the estimation by a maximum entropy method.

We notice fairly good agreement, which is even more remarkable as the heights of the two maxima differ enormously.

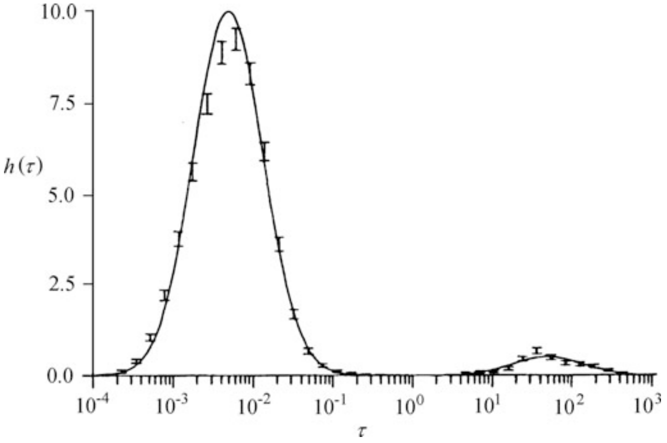


**Fig. 11.7** A given spectrum (*solid line*) and the reconstruction (values with error bars) by the maximum entropy method (From Honerkamp (1994), reprinted by permission of Wiley-VCH, Weinheim)

Figure 11.8 displays the solution of the same problem using the regularization method. Figure 11.8a shows the values for $\mathcal{L} = 1$ and Fig. 11.8b the values for $\mathcal{L} = d^2/(d \ln \tau)^2$. In both cases, we used the self-consistent method for determining the regularization parameter. Obviously, the regularization method for $\mathcal{L} = 1$ has difficulties resolving the smaller peak. This is not surprising since this method favors solutions with weights as small as possible (the additional functional is $\sum_\alpha m_\alpha (h_\alpha)^2$). The reconstruction of the larger peak leads to a certain neglect of the smaller one. Concerning the application of these methods to real data sets, we refer to the literature (Honerkamp and Weese 1989, 1993).

## 11.5 Estimating the Realization of a Hidden Process

In this section we discuss a method how to estimate the realization of a hidden process. Two very important algorithms will be introduced, in Sect. 11.5.1 the Viterbi algorithm for the estimation of the realization $x_{1\ldots N} = (x_1, \ldots, x_N)$, given the observations $y_{1\ldots N} = (y_1, \ldots, y_N)$ in a hidden Markov process, and in Sect. 11.5.2 the Kalman filter for the estimation of the process $x(t)$, $t = 1, \ldots, N$, given the observations $y(t)$, $t = 1, \ldots, N$ in a state space model.

### *11.5.1 The Viterbi Algorithm*

We consider a Markov process with discrete states $x_t \in \{1, \ldots, N\}$ for $t = 1, \ldots, N$ and transition probabilities $\rho(x'_t \mid x_t)$. The observation equation may be defined by the conditional probability density $\rho(y_{1\ldots N} \mid x_{1\ldots N})$ with

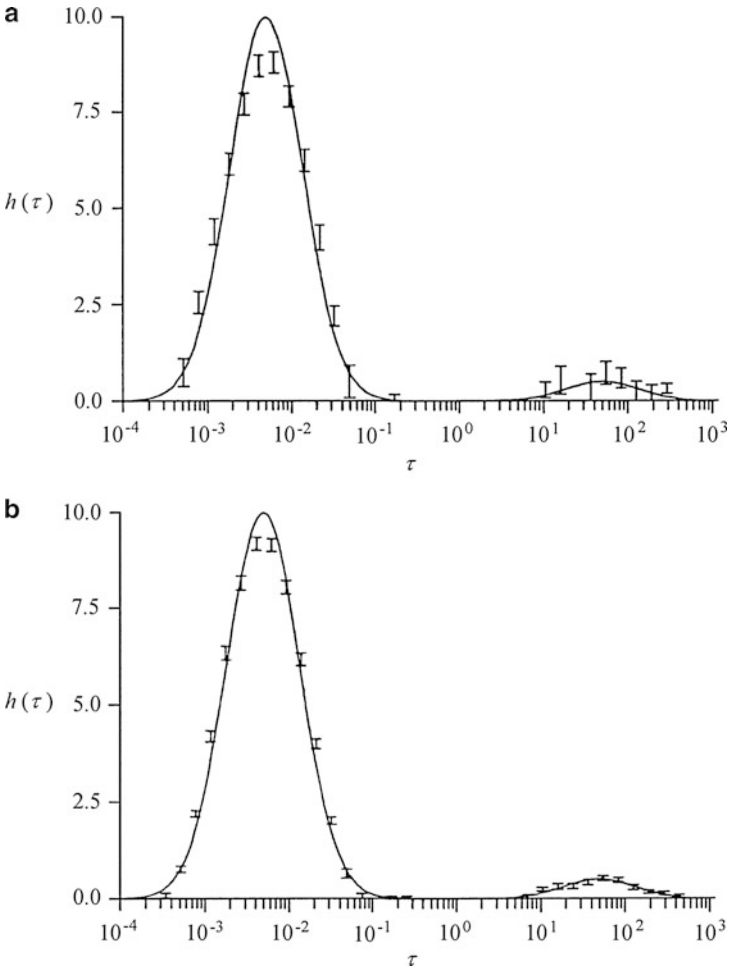


**Fig. 11.8** The given spectrum (*solid line*) and the reconstruction (values with error bars) by the regularization method, (**a**) with $\mathcal{L} = 1$, (**b**) with $\mathcal{L} = \mathrm{d}^2/(\mathrm{d}\ln\tau)^2$ (From Honerkamp (1994), reprinted by permission of Wiley-VCH, Weinheim)

$$\rho(y_{1\ldots N} \mid x_{1\ldots N}) = \prod_{t=1}^{N} \rho(y_t \mid x_t), \tag{11.85}$$

i.e., the probability for the observation $y_t$ may depend only on the state $x_t$. The Markov property of the system dynamics asserts that, furthermore,

$$\rho(x_{1\ldots N}) = \rho(x_1) \prod_{t=2}^{N} \rho(x_t \mid x_{t-1}). \tag{11.86}$$

The MAP estimator for $x_{1\ldots N}$ is defined by the maximum of $\rho(x_{1\ldots N} \mid y_{1\ldots N})$, i.e.,

$$\widehat{x}_{1\ldots N} = \arg\max_{x_{1\ldots N}} \rho(x_{1\ldots N} \mid y_{1\ldots N}) = \arg\max_{x_{1\ldots N}} \Big[\rho(y_{1\ldots N} \mid x_{1\ldots N})\, \rho(x_{1\ldots N})\Big]. \quad (11.87)$$

Let us define the negative logarithm of the conditional density $\rho(x_{1\ldots N} \mid y_{1\ldots N})$ as the length $L(x_{1\ldots N})$ of the path $x_{1\ldots N}$,

$$L(x_{1\ldots N}) = -\ln \Big[\rho(y_{1\ldots N} \mid x_{1\ldots N})\, \rho(x_{1\ldots N})\Big]; \quad (11.88)$$

then, the MAP estimator $\widehat{x}_{1\ldots N}$ is given by the shortest path with

$$\widehat{x}_{1\ldots N} = \arg\min_{x_{1\ldots N}} L(x_{1\ldots N}). \quad (11.89)$$

Instead of searching in the $m^N$-dimensional space of the paths $\{x_{1\ldots N}\}$ for the minimum, one can find it iteratively with the Viterbi algorithm:

One introduces the shortest path, which ends in state $\alpha$ at time $t$. We call it $x^*_{1\ldots t}\big|_{x_t=\alpha}$. Its length is $L(x^*_{1\ldots t}\big|_{x_t=\alpha})$.

For $t = 1$, obviously, $x^*_1\big|_{x_1=\alpha}$ for all $\alpha$ is given by $x^*_1 = \alpha$ itself. The length is

$$L(x^*_1|_{x_1=\alpha}) = -\ln\, \rho(y_1|x_1 = \alpha) - \ln \rho(x_1 = \alpha). \quad (11.90)$$

Now let $\{x^*_{1\ldots t-1}\big|_{x_{t-1}=\beta}\}$ be given for all $\beta = 1, \ldots, m$. An extension of the path from $(t-1)$ to $t$ to $x_t = \alpha$ passes through some state $x_{t-1} = \beta$ at time $t-1$, and the path $x^*_{1\ldots t}\big|_{x_t=\alpha}$ is shortest for that $x_{t-1} = \beta$ for which

$$\begin{aligned} L(x_{1\ldots t}|_{x_t=\alpha, x_{t-1}=\beta}) = &-\ln\, \rho(y_t \mid x_t = \alpha) - \ln\, \rho(x_t = \alpha \mid x_{t-1} = \beta) \\ &+ L(x^*_{1\ldots t-1}|_{x_{t-1}=\beta}), \end{aligned} \quad (11.91)$$

is a minimum. Thus

$$L(x^*_{1\ldots t}|_{x_t=\alpha}) = \min_{\beta} L(x_{1\ldots t}|_{x_t=\alpha, x_{t-1}=\beta}) \quad (11.92)$$

is the length of the shortest path with $x_t = \alpha$. For this path, the state $\beta$, which is assumed at $t-1$, is given by

$$\beta = \arg\min_{\beta'} L(x_{1\ldots t}|_{x_t=\alpha, x_{t-1}=\beta'}). \quad (11.93)$$

We will calculate this for every $\alpha$ and denote this by $\beta = ptr(x_t = \alpha)$.

Having determined in this way the shortest path $x^*_{1\ldots t}\big|_{x_t=\alpha}$ and $ptr(x_t = \alpha)$ for all $t$ until $t = N$, one may finally minimize with respect to $\alpha$ to find the shortest path at all. Obviously, by

$$x^*_N = \arg\min_{\alpha} L(x^*_{1\ldots N}|_{x_N=\alpha}), \quad (11.94)$$

the state $x^*_N$ at time $t = N$ with the shortest path independent of the final state is determined. The states of the whole shortest path can now be reconstructed by using

$$x^*_{t-1} = ptr(x^*_t) \tag{11.95}$$

recursively, starting with $x_{N-1} = ptr(x^*_N)$.

In this estimation of the shortest path, the information about the whole data set $y_{1\ldots N}$ is used for every $x_t$. Thus we may also call this estimation $x^*_{1\ldots t} \mid y_{1\ldots N}$. Another estimation would be $x^*_{1\ldots t} \mid y_{1\ldots t}$, i.e., an estimation for $x_{1\ldots t}$ on the basis only of $y_{1\ldots t}$. This could be formulated by choosing

$$x^*_t = \arg\min_\alpha L(\,x^*_{1\ldots t}|_{x_t=\alpha}), \tag{11.96}$$

for every $t$ as in (11.94) for $t = N$, thus not using the record $x^*_{t-1} = ptr(x^*_t)$ for $t = N, N-1, \ldots$.

*Example.* We consider a hidden Markov model with two states and $a_{12} \equiv \rho(x_t = 1 \mid x_{t-1} = 2) = 0.1$, $a_{21} \equiv \rho(x_t = 2 \mid x_{t-1} = 1) = 0.05$, $\mu_1 = 1$, $\mu_2 = 2$. Then

$$\begin{aligned} -\ln\rho(y_t \mid x_t = \alpha) &= \frac{1}{2\sigma_\alpha^2}(y_t - \mu_\alpha)^2, \quad \alpha = 1, 2 \\ \ln\rho(\alpha \mid \beta) &= \ln a_{\alpha\beta}. \end{aligned} \tag{11.97}$$

Figure 11.9 shows the observation $\{y_t\}$ and the exact hidden process $\{x_t\}$ for two different values $\sigma_1 = \sigma_2 = 0.1$ (left) and 0.3 (right) in the upper subplot. The subplots in the middle show the estimates $x^*_{1\ldots t}|y_{1\ldots t}$ and the lower the estimates $x^*_{1\ldots t}|y_{1\ldots N}$. For $\sigma_1 = \sigma_2 = 0.1$, no error can be observed, whereas for $\sigma_1 = \sigma_2 = 0.3$, some larger fluctuations are interpreted as transitions in the estimate $x^*_{1\ldots t}|y_{1\ldots t}$.

### 11.5.2 The Kalman Filter

The Viterbi algorithm should also be applicable for estimating a hidden AR process, which usually is formulated in terms of a state space model (see Sect. 11.3):

$$\boldsymbol{X}(t) = \mathsf{A}\,\boldsymbol{X}(t-1) + \mathsf{B}\,\boldsymbol{\eta}(t), \qquad \boldsymbol{\eta}(t) \sim \mathrm{WN}(\mathbf{0}, 1), \tag{11.98}$$

$$\boldsymbol{Y}(t) = \mathsf{C}\,\boldsymbol{X}(t) + \boldsymbol{\epsilon}(t), \qquad \boldsymbol{\epsilon}(t) \sim \mathrm{WN}(\mathbf{0}, \mathsf{R}). \tag{11.99}$$

$\mathsf{A}, \mathsf{B}, \mathsf{C}$ may be time-dependent matrices, $\mathsf{B}$ and $\mathsf{C}$ not necessarily quadratic, $\mathsf{R}$ is the covariance matrix of the observational noise $\boldsymbol{\epsilon}$). Then, with $\mathsf{Q} = \mathsf{B}\mathsf{B}^T$,

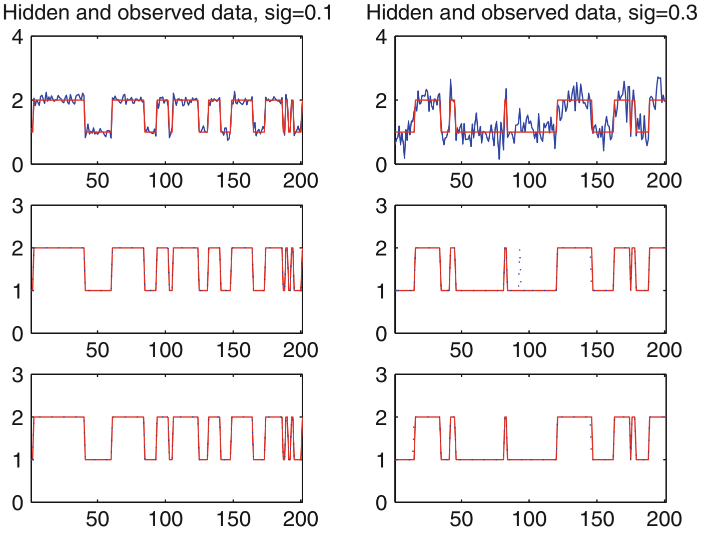


**Fig. 11.9** Realization and exact hidden process (*above*), exact and estimate $x^*_{1\dots t} \mid y_{1\dots t}$(*middle*), and exact and estimate $x^*_{1\dots t} \mid y_{1\dots N}$(*below*); for $\sigma_1 = \sigma_2 = 0.1$ (*left*) and $\sigma_1 = \sigma_2 = 0.3$ (*right*)

$$-\ln \rho(\boldsymbol{y}_t, t \mid \boldsymbol{x}_t, t) = \frac{1}{2}(\boldsymbol{y}_t - \mathsf{C}\,\boldsymbol{x}_t)^T \mathsf{R}^{-1}(\boldsymbol{y}_t - \mathsf{C}\,\boldsymbol{x}_t) \tag{11.100}$$

$$-\ln \rho(\boldsymbol{x}_t, t \mid \boldsymbol{x}_{t-1}, t-1) = \frac{1}{2}(\boldsymbol{x}_t - \mathsf{A}\,\boldsymbol{x}_{t-1})^T \mathsf{Q}^{-1}(\boldsymbol{x}_t - \mathsf{A}\,\boldsymbol{x}_{t-1}), \tag{11.101}$$

and in the same way as in the hidden Markov process, one should be able to define the estimates $\boldsymbol{x}^*_{1\dots t} \mid \boldsymbol{y}_{1\dots t}$ and $\boldsymbol{x}^*_{1\dots t} \mid \boldsymbol{y}_{1\dots N}$.

This is the case. Before we do this, we will state the result.

**The forward Kalman filter.** The estimation procedure can be seen to consist of two parts. The first is the so-called prediction prior to the observation $\boldsymbol{Y}(t)$. Here $\boldsymbol{X}(t)$ is estimated on the basis of the $\boldsymbol{Y}(t-1), \dots$ and of $\hat{\boldsymbol{X}}(t-1)$. This estimator function is called $\tilde{\boldsymbol{X}}(t)$, whereas $\hat{\boldsymbol{X}}(t)$ is the estimate of $\boldsymbol{X}(t)$ on the basis of $\boldsymbol{Y}(t), \boldsymbol{Y}(t-1), \dots$. From (11.98) and (11.99), we get

$$\tilde{\boldsymbol{X}}(t) = \mathsf{A}(t-1)\hat{\boldsymbol{X}}(t-1). \tag{11.102}$$

The expected observation $\tilde{\boldsymbol{Y}}(t)$ for known $\tilde{\boldsymbol{X}}(t)$ is then

$$\tilde{\boldsymbol{Y}}(t) = \mathsf{C}(t)\tilde{\boldsymbol{X}}(t). \tag{11.103}$$

The second step is the correction of $\tilde{\boldsymbol{X}}(t)$ because of observation $\boldsymbol{Y}(t)$. This correction is assumed to be linear in the forecast error for the observations $\boldsymbol{Y}(t) - \tilde{\boldsymbol{Y}}(t)$. Thus one eventually obtains for the estimate $\hat{\boldsymbol{X}}(t)$

$$\hat{\boldsymbol{X}}(t) = \tilde{\boldsymbol{X}}(t) + \mathsf{K}(t)\left[\boldsymbol{Y}(t) - \tilde{\boldsymbol{Y}}(t)\right], \tag{11.104}$$

where $\mathsf{K}(t)$, the so-called *Kalman gain factor*, is a matrix yet to be calculated.

Having calculated the matrix $\mathsf{K}(t)$, (11.102)–(11.104) then allow recursive determination of $\hat{\boldsymbol{X}}(t)$. Furthermore, the prediction error

$$\hat{P}_{\alpha\beta}(t) \equiv \langle (\boldsymbol{X}(t) - \hat{\boldsymbol{X}}(t))_\alpha \; (\boldsymbol{X}(t) - \hat{\boldsymbol{X}}(t))_\beta \rangle \tag{11.105}$$

can be calculated at each time.

The quantities $\hat{\mathsf{P}}(t)$ and $\mathsf{K}(t)$ can be determined recursively prior to the observations. If $\hat{\mathsf{P}}(t-1)$ is given, then a matrix $\tilde{\mathsf{P}}(t)$ is calculated as

$$\tilde{\mathsf{P}}(t) = \mathsf{A}(t-1)\hat{\mathsf{P}}(t-1)\mathsf{A}^T(t-1) + \mathsf{B}(t-1)\mathsf{Q}(t)\mathsf{B}^T(t-1). \tag{11.106}$$

With the help of $\tilde{\mathsf{P}}(t)$,

$$\mathsf{K}(t) = \tilde{\mathsf{P}}(t)\mathsf{C}^T(t)\;[\mathsf{C}(t)\tilde{\mathsf{P}}(t)\mathsf{C}^T(t) + \mathsf{R}(t)]^{-1} \tag{11.107}$$

is obtained, and then

$$\hat{\mathsf{P}}(t) = \tilde{\mathsf{P}}(t) - \mathsf{K}(t)\mathsf{C}(t)\tilde{\mathsf{P}}(t). \tag{11.108}$$

As initial values, one must specify $\hat{\mathsf{P}}(1)$ and, for (11.102), $\hat{\boldsymbol{X}}(1)$, for example, $\hat{\boldsymbol{X}}(1) = 0$ and $\hat{\mathsf{P}}(1) = c\mathbf{1}$ with appropriately chosen $c$. All of the $\hat{\mathsf{P}}(t)$ and $\mathsf{K}(t)$ for $t > 1$ can then be determined from (11.106) to (11.108), and all of the $\hat{\boldsymbol{X}}(t)$ from (11.102) to (11.104).

Until now we have considered two estimators for $\boldsymbol{X}(t)$, namely, $\tilde{\boldsymbol{X}}(t)$ as an estimator for given $\boldsymbol{Y}(t-1), \boldsymbol{Y}(t-2), \ldots$, and $\hat{\boldsymbol{X}}(t)$ as an estimator for given $\boldsymbol{Y}(t), \boldsymbol{Y}(t-1), \ldots$. These estimators are called forward Kalman filters because they can be calculated proceeding forward in time.

**The backward Kalman Filter.** Given the observation $\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(N)$, one might ask for the best estimator for $\boldsymbol{X}(t)$ if all of the observations $\boldsymbol{Y}(t)$ from $t = 1$ up to $t = N$ are taken into account. If we call this estimator $\check{\boldsymbol{X}}(t)$, then clearly

$$\check{\boldsymbol{X}}(t)\Big|_{t=N} = \hat{\boldsymbol{X}}(t)\Big|_{t=N}, \tag{11.109}$$

and for the error matrix

$$\check{\mathsf{P}}(t) = \Big\langle \big[\boldsymbol{X}(t) - \check{\boldsymbol{X}}(t)\big]\big[\boldsymbol{X}(t) - \check{\boldsymbol{X}}(t)\big]^T \Big\rangle, \tag{11.110}$$

it similarly holds that

$$\check{\mathsf{P}}(t)\Big|_{t=N} = \hat{\mathsf{P}}(t)\Big|_{t=N}. \tag{11.111}$$

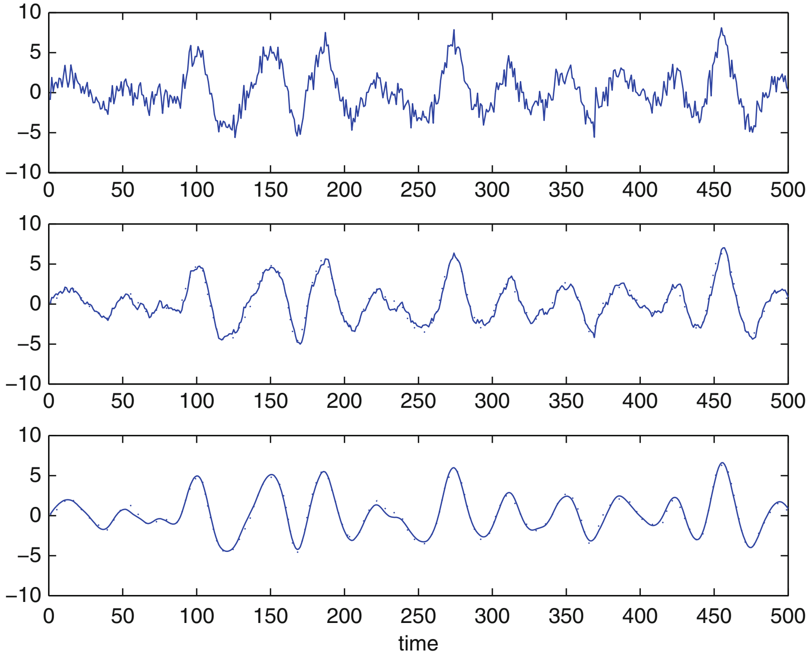


**Fig. 11.10** Realization of a hidden AR(2) process (*above*), exact and estimate $\hat{x}(t)$(*middle, solid line*), and exact and estimate $\check{x}(t)$(*below, solid line*). The true realization $x(t)$ of the hidden process is given by the *broken line*

The estimates $\check{\boldsymbol{X}}(t)$ for $t < N$ can now be determined recursively for $t = N - 1$, $N - 2, \ldots$ through

$$\check{\boldsymbol{X}}(t) = \hat{\boldsymbol{X}}(t) + \check{\mathsf{K}}(t)\big[\check{\boldsymbol{X}}(t+1) - \tilde{\boldsymbol{X}}(t+1)\big], \quad t = N-1, \ldots, \tag{11.112}$$

with

$$\check{\mathsf{K}}(t) = \hat{\mathsf{P}}(t)\mathsf{A}^T\,\tilde{\mathsf{P}}^{-1}(t+1), \tag{11.113}$$

and for $\check{\mathsf{P}}(t)$, one obtains

$$\check{\mathsf{P}}(t) = \hat{\mathsf{P}}(t) + \check{\mathsf{K}}(t)\big[\check{\mathsf{P}}(t+1) - \tilde{\mathsf{P}}(t+1)\big]\check{\mathsf{K}}^T(t). \tag{11.114}$$

Thus while $\hat{\boldsymbol{X}}(t)$ takes into account only the observations $\boldsymbol{Y}(t')$ for $t' = 1, \ldots, t$, for the estimation of $\check{\boldsymbol{X}}(t)$ all observations $\boldsymbol{Y}(t')$, $t' = 1, \ldots, N$, are taken into account. Thus the latter estimate is clearly superior (see Fig. 11.10).

A remarkable property of the Kalman filter is that the estimates $\hat{x}(t)$ and $\check{x}(t)$ are very robust. Even when the parameters of the model used in the construction of the Kalman filters differ much from the true ones (as one can check, e.g., in simulations), the estimates are only slightly deteriorated (see Fig. 11.11).

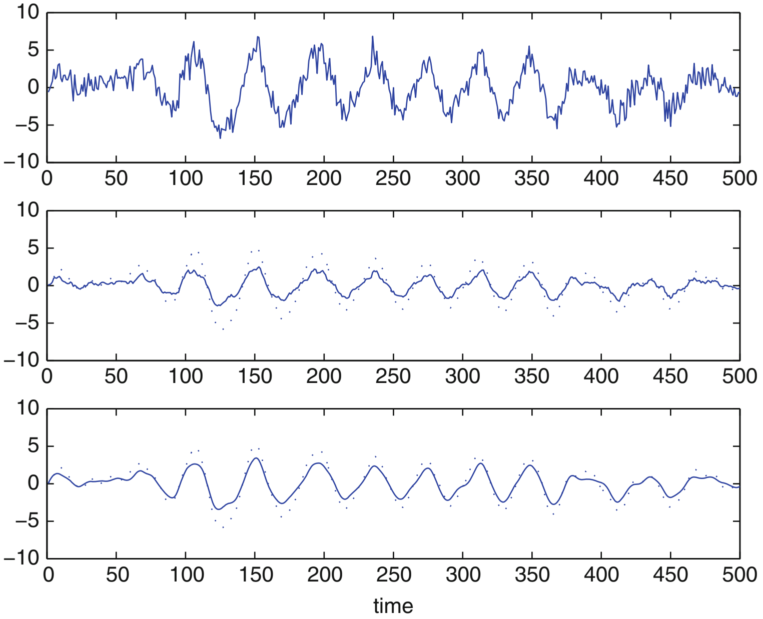


**Fig. 11.11** Realization of a hidden AR(2) process (*above*), estimate $\hat{x}(t)$(*middle, solid line*), and estimate $\check{x}(t)$(*below, solid line*). The true realization $x(t)$ of the hidden process is given by the *broken line*. The parameters used in the simulation are $T = 40$, $\tau = 50$ for the period and damping constant of the AR(2) model, whereas in constructing the Kalman filter $T = 20$, $\tau = 10$ are chosen

**The extended Kalman filter.** Until now, we have assumed linear system equations and observation equations. The Kalman filter can also easily be formulated for nonlinear equations of the form

$$\boldsymbol{X}(t) = \boldsymbol{f}\big(\boldsymbol{X}(t-1)\big) + \mathsf{B}(t-1)\boldsymbol{\eta}(t), \tag{11.115}$$

$$\boldsymbol{Y}(t) = \boldsymbol{g}\big(t, \boldsymbol{X}(t)\big) + \boldsymbol{\epsilon}(t), \tag{11.116}$$

where the assumptions given in (11.98)–(11.99) about the noise terms is made. The predictions prior to the observation $\boldsymbol{Y}(t)$ are

$$\tilde{\boldsymbol{X}}(t) = \boldsymbol{f}\big(\hat{\boldsymbol{X}}(t-1)\big), \tag{11.117}$$

$$\tilde{\boldsymbol{Y}}(t) = \boldsymbol{g}\big(t, \tilde{\boldsymbol{X}}(t)\big), \tag{11.118}$$

and for the estimate $\hat{\boldsymbol{X}}(t)$, one again assumes

$$\hat{\boldsymbol{X}}(t) = \tilde{\boldsymbol{X}}(t) + \mathsf{K}(t)[\boldsymbol{Y}(t) - \tilde{\boldsymbol{Y}}(t)]. \tag{11.119}$$

For the determination of $\tilde{\mathsf{P}}(t)$, $\mathsf{K}(t)$, and $\hat{\mathsf{P}}(t)$, the same algorithm as in the linear case is obtained, only the meaning of the matrices $\mathsf{A}$ and $\mathsf{C}$ has changed. These matrices are now given by

$$A_{\alpha\beta}(t-1) = \left.\frac{\partial f_\alpha(\boldsymbol{X})}{\partial X_\beta}\right|_{\boldsymbol{X}=\hat{\boldsymbol{X}}(t-1)}, \tag{11.120}$$

and

$$C_{\alpha\beta} = \left.\frac{\partial g_\alpha(t,\boldsymbol{X})}{\partial X_\beta}\right|_{\boldsymbol{X}=\tilde{\boldsymbol{X}}(t)}. \tag{11.121}$$

If time is a continuous instead of a discrete parameter, the Kalman filter can be defined similarly (see, for example, Gelb 1974). Now, the system and observation equations are

$$\dot{\boldsymbol{X}}(t) = \boldsymbol{f}\big(\boldsymbol{X}(t),t\big) + \boldsymbol{\eta}(t), \quad \boldsymbol{\eta}(t) \sim WN(\mathbf{0},\mathsf{Q}(t)), \tag{11.122}$$

$$\boldsymbol{Y}(t) = \boldsymbol{g}\big(\boldsymbol{X}(t),t\big) + \boldsymbol{\epsilon}(t), \quad \boldsymbol{\epsilon}(t) \sim WN(\mathbf{0},\mathsf{R}(t)), \tag{11.123}$$

$$E\big(\boldsymbol{\epsilon}(t),\boldsymbol{\eta}(t')\big) = \mathbf{0}. \tag{11.124}$$

The estimate of the state vector then is

$$\dot{\boldsymbol{X}}(t \mid t) = \boldsymbol{f}\big(\hat{\boldsymbol{X}}(t),t\big) + \mathsf{K}(t)\big[\boldsymbol{Y}(t) - \boldsymbol{g}(\hat{\boldsymbol{X}}(t),t)\big] \tag{11.125}$$

and

$$\mathsf{K}(t) = \mathsf{P}(t)\mathsf{C}^T\big(\boldsymbol{X}(t),t\big)\mathsf{R}^{-1}(t), \tag{11.126}$$

where the matrix $\mathsf{C}$ again contains the elements

$$C_{\alpha\beta} = \left.\frac{\partial g_\alpha(\boldsymbol{X},t)}{\partial X_\beta}\right|_{\boldsymbol{X}=\hat{\boldsymbol{X}}(t)}. \tag{11.127}$$

The equation for the error matrix is

$$\dot{\mathsf{P}}(t) = \mathsf{A}\,\mathsf{P}(t) + \mathsf{P}(t)\mathsf{A}^T + \mathsf{Q}(t) - \mathsf{P}(t)\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C}\,\mathsf{P}(t) \tag{11.128}$$

with

$$(\mathsf{A})_{\alpha\beta} = \big(\mathsf{A}(\hat{\boldsymbol{X}}(t),t)\big)_{\alpha\beta} = \left.\frac{\partial f_\alpha(\boldsymbol{X},t)}{\partial X_\beta}\right|_{\boldsymbol{X}=\hat{\boldsymbol{X}}(t)}. \tag{11.129}$$

*Application* (following Gelb 1974).
Consider a vertically falling body in the gravitational field of the earth. Suppose that $x$ is the height of the body above the surface of the earth. The equation of motion is

$$\ddot{x} = \alpha - g \tag{11.130}$$

with the deceleration because of the frictional force

$$\alpha = \frac{\rho}{2\beta}\,\dot{x}^2. \tag{11.131}$$

$\rho$ is the density of the atmosphere,

$$\rho = \rho_0 e^{-x/k}, \tag{11.132}$$

$k$ is a given constant, and $\beta$ is the so-called ballistic constant, which we view as unknown.

With $x_1 = x$, $x_2 = \dot{x}$, $x_3 = \beta$, we obtain for the equation of motion of the system

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} x_2 \\ \alpha - g \\ 0 \end{pmatrix}, \qquad \alpha = \alpha(\boldsymbol{x}). \tag{11.133}$$

Measurements of the height of the body are given by

$$y(t) = x_1(t) + \epsilon(t), \qquad \epsilon(t) \sim N(0, r). \tag{11.134}$$

The initial values are $\mathbf{x}(0) = \boldsymbol{\mu}$, and

$$\mathsf{P}(0) = \begin{pmatrix} p_{110} & 0 & 0 \\ 0 & p_{220} & 0 \\ 0 & 0 & p_{330} \end{pmatrix}. \tag{11.135}$$

With the Kalman filter, the quantities $x_1(t)$, $x_2(t)$, and $x_3(t)$ can be estimated.

*Remark.* For $\mathsf{R} \to \mathbf{0}$, the observation becomes exact. Then, if $\mathsf{C}^{-1}(t)$ exists, it follows from (11.107) that $\mathsf{K}(t) = \mathsf{C}^{-1}(t)$. From (11.104),

$$\hat{\boldsymbol{X}}(t) = \tilde{\boldsymbol{X}}(t) + \mathsf{C}^{-1}(t)[\boldsymbol{Y}(t) - \mathsf{C}\tilde{\boldsymbol{X}}(t)] = \mathsf{C}^{-1}(t)\boldsymbol{Y}(t) \tag{11.136}$$

then becomes consistent with (11.99).

For $\mathsf{R} \neq \mathbf{0}$, the Kalman filter is more effective than a simple estimation

$$\hat{\boldsymbol{X}}(t) = \mathsf{C}^{-1}(t)\boldsymbol{Y}(t), \tag{11.137}$$

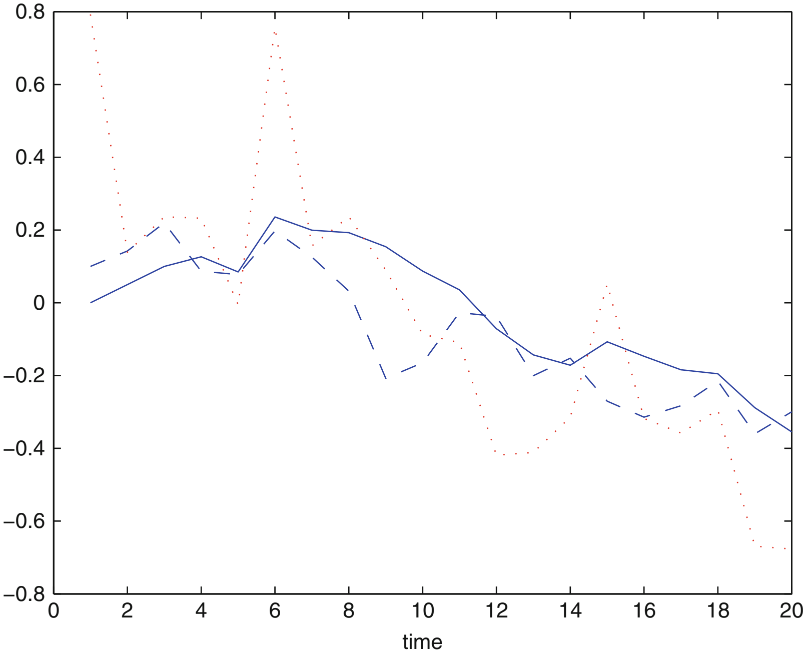


**Fig. 11.12** Realizations of a hidden AR(1) process with $C = 1$ in the observation equation: True realization $x(t)$ of the hidden process (*solid line*), Kalman estimate (*broken line*), and the naive simple estimate $\hat{x}(t) = C^{-1}y(t) \equiv y(t)$ (*dotted line*) (Obviously the Kalman estimate is much better)

that is, the prediction error with the Kalman error is smaller than in this estimate (11.137) (see Fig. 11.12) and the Kalman filter always becomes more effective than this simple estimate with growing variance $\mathsf{R}(t)$, that is, with growing uncertainty in the observations.

**Proof of the forward and backward Kalman filter.** For convenience we will assume for this proof that system and observation variable are univariate. As can easily be shown the log likelihood can be written as

$$L(x_{1\ldots t-1} \mid x_{t-1} = \beta_{t-1}) = \frac{1}{2}(\beta_{t-1} - \hat{x}_{t-1})^T \hat{\mathsf{P}}^{-1}(t-1)(\beta_{t-1} - \hat{x}_{t-1})$$
$$+ \text{terms independent on } \beta_{t-1}. \tag{11.138}$$

Then

$$\begin{aligned}L(x_{1,\ldots,t} \mid x_t = \alpha_t, x_{t-1} = \beta_{t-1}) &= \frac{1}{2}(y_t - \mathsf{C}x_t)^T \mathsf{R}^{-1}(y_t - \mathsf{C}x_t) \mid_{x_t=\alpha_t} \\ &+ \frac{1}{2}(x_t - \mathsf{A}\beta_{t-1})^T \mathsf{Q}^{-1}(x_t - \mathsf{A}\beta_{t-1}) \mid_{x_t=\alpha_t} \\ &+ \frac{1}{2}(\beta_{t-1} - \hat{x}_{t-1})^T \hat{\mathsf{P}}^{-1}(t-1)(\beta_{t-1} - \hat{x}_{t-1}). \end{aligned} \tag{11.139}$$

Finding the minimum of this expression with respect to $\beta_{t-1}$, we write the second and third terms as

$$\begin{aligned}L' = &\frac{1}{2}\Big(\beta_{t-1} - (\mathsf{A}^T\mathsf{Q}^{-1}\mathsf{A} + \hat{\mathsf{P}}^{-1}(t-1))^{-1}(\mathsf{A}^T\mathsf{Q}^{-1}x_t + \hat{\mathsf{P}}^{-1}(t-1)\hat{x}_{t-1})\Big)^T \\ &\times \Big(\mathsf{A}^T\mathsf{Q}^{-1}\mathsf{A} + \hat{\mathsf{P}}^{-1}(t-1)\Big) \\ &\times \Big(\beta_{t-1} - (\mathsf{A}^T\mathsf{Q}^{-1}\mathsf{A} + \hat{\mathsf{P}}^{-1}(t-1))^{-1}(\mathsf{A}^T\mathsf{Q}^{-1}x_t + \hat{\mathsf{P}}^{-1}(t-1)\hat{x}_{t-1})\Big) \\ &+ \frac{1}{2}x_t^T\mathsf{Q}^{-1}x_t + \frac{1}{2}\hat{x}_{t-1}^T\hat{\mathsf{P}}^{-1}(t-1)\hat{x}_{t-1} \\ &- \frac{1}{2}\Big(\mathsf{A}^T\mathsf{Q}^{-1}x_t + \hat{\mathsf{P}}^{-1}(t-1)\hat{x}_{t-1}\Big)^T \Big(\mathsf{A}^T\mathsf{Q}^{-1}\mathsf{A} + \hat{\mathsf{P}}^{-1}(t-1)\Big)^{-1} \\ &\times \Big(\mathsf{A}^T\mathsf{Q}^{-1}x_t + \hat{\mathsf{P}}^{-1}(t-1)\hat{x}_{t-1}\Big). \end{aligned} \tag{11.140}$$

From this expression, one can read off the value of $\beta_{t-1}$ for which the minimum is achieved. We will need this value for the Kalman smoothing or backward filter. Now we are interested only in the value at the minimum. After some lengthy calculation, but using only the identities $\mathsf{A}^{-1}\mathsf{B}^{-1} = (\mathsf{BA})^{-1}$ and $\mathsf{A}^{-1} - \mathsf{B}^{-1} = \mathsf{A}^{-1}(\mathsf{B} - \mathsf{A})\mathsf{B}^{-1}$, we get for $L'$ at the minimum

$$L'_{\min} = \frac{1}{2}(x_t - \mathsf{A}\hat{x}_{t-1})^T \tilde{\mathsf{P}}^{-1}(t)(x_t - \mathsf{A}\hat{x}_{t-1}) \tag{11.141}$$

with

$$\tilde{\mathsf{P}}(t) = \mathsf{A}\hat{\mathsf{P}}(t-1)\mathsf{A}^T + \mathsf{Q}. \tag{11.142}$$

Then the log likelihood can be written as

$$\begin{aligned}L &= \min_{\beta_{t-1}} L(x_{1\ldots t} \mid x_t = \alpha_t, x_{t-1} = \beta_{t-1}) \\ &= \frac{1}{2}(y_t - \mathsf{C}x_t)^T\mathsf{R}^{-1}(y_t - \mathsf{C}x_t) \\ &\quad + \frac{1}{2}(x_t - \mathsf{A}\hat{x}_{t-1})^T\tilde{\mathsf{P}}^{-1}(t)(x_t - \mathsf{A}\hat{x}_{t-1}) \mid_{x_t=\alpha_t} . \end{aligned} \tag{11.143}$$

To find the minimum with respect to $\alpha_t$, we write $L$ with the same procedure as above as

$$\begin{aligned}
L = \frac{1}{2}&\Big(x_t - (\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t))^{-1}(\mathsf{C}^T\mathsf{R}^{-1}y_t + \tilde{\mathsf{P}}^{-1}(t)\mathsf{A}\hat{x}_{t-1})\Big)^T \\
&\times \Big(\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t)\Big) \\
&\times \Big(x_t - (\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t))^{-1}(\mathsf{C}^T\mathsf{R}^{-1}y_t + \tilde{\mathsf{P}}^{-1}(t)\mathsf{A}\hat{x}_{t-1})\Big) \\
&+ \frac{1}{2}y_t^T\mathsf{R}^{-1}y_t + \frac{1}{2}\hat{x}_{t-1}^T\mathsf{A}^T\tilde{\mathsf{P}}^{-1}(t)\mathsf{A}\hat{x}_{t-1} \\
&- \frac{1}{2}\Big(\mathsf{C}^T\mathsf{R}^{-1}y_t + \tilde{\mathsf{P}}^{-1}(t)\mathsf{A}\hat{x}_{t-1}\Big)^T\Big(\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t)\Big)^{-1} \\
&\times \Big(\mathsf{C}^T\mathsf{R}^{-1}y_t + \tilde{\mathsf{P}}^{-1}(t)\mathsf{A}\hat{x}_{t-1}\Big). \qquad (11.144)
\end{aligned}$$

which again can be written as

$$L = \frac{1}{2}(x_t - \hat{x}_t)^T\hat{\mathsf{P}}^{-1}(t)(x_t - \hat{x}_t) + \text{terms independent of } x_t \qquad (11.145)$$

with

$$\hat{\mathsf{P}}^{-1}(t) = \mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t) \qquad (11.146)$$

and

$$\hat{x}_t = \Big(\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t)\Big)^{-1}\Big(\mathsf{C}^T\mathsf{R}^{-1}y_t + \tilde{\mathsf{P}}^{-1}(t)\mathsf{A}\hat{x}_{t-1}\Big). \qquad (11.147)$$

Therefore, we obtain the equation of the Kalman filter. From (11.146), we get

$$\hat{\mathsf{P}}(t) = \Big(\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t)\Big)^{-1} = \tilde{\mathsf{P}}(t) - \mathsf{K}(t)\mathsf{C}\tilde{\mathsf{P}}(t) \qquad (11.148)$$

with the Kalman gain

$$\mathsf{K}(t) = \Big(\mathsf{C}^T\mathsf{R}^{-1}\mathsf{C} + \tilde{\mathsf{P}}^{-1}(t)\Big)^{-1}\mathsf{C}^T\mathsf{R}^{-1} = \tilde{\mathsf{P}}(t)\mathsf{C}^T\Big(\mathsf{C}\tilde{\mathsf{P}}(t)\mathsf{C}^T + \mathsf{R}\Big)^{-1}, \qquad (11.149)$$

and the equation for $x_t$ can also be written as

$$x_t = \mathsf{A}\hat{x}_{t-1} + \mathsf{K}(t)(y_t - \mathsf{C}\mathsf{A}\hat{x}_{t-1}). \qquad (11.150)$$

Thus one has arrived at the formula for the Kalman filter by minimizing the log likelihood at each time $t$. In this way, one finds the path which is the shortest, given the data $y_{1\ldots t}$ for $i = 1, \ldots, t$.

However, we could also ask for the shortest path, given all the data $y_{1\ldots N}$ for $i = 1, \ldots, N$. Then one has to determine for every given $x_t$ that value of $x_{t-1}$ for which the path until time $t$ is shortest. This value can be read off from (11.140). Calling it $\check{x}_{t-1}$ we obtain, setting $x_t$ already equal $\check{x}_t$,

$$\begin{aligned}\check{x}_{t-1} &= (\mathsf{A}^T\mathsf{Q}^{-1}\mathsf{A} + \hat{\mathsf{P}}^{-1}(t-1))^{-1}(\mathsf{A}^T\mathsf{Q}^{-1}\check{x}_t + \hat{\mathsf{P}}^{-1}(t-1)\hat{x}_{t-1}) \\ &= \hat{x}_{t-1} + \check{\mathsf{K}}(t-1)(\check{x}_t - \mathsf{A}\hat{x}_{t-1}) \qquad (11.151)\end{aligned}$$

with

$$\check{\mathsf{K}}(t-1) = \Big(\mathsf{A}^T\mathsf{Q}^{-1}\mathsf{A} + \hat{\mathsf{P}}^{-1}(t-1)\Big)^{-1}\mathsf{A}^T\mathsf{Q}^{-1} = \hat{\mathsf{P}}(t-1)\mathsf{A}^T\tilde{\mathsf{P}}^{-1}(t). \qquad (11.152)$$

This is a backward recurrence equation: We get $\check{x}_{t-1}$ given $\check{x}_t$ and $\hat{x}_{t-1}$. Thus we have first to determine the $\{\hat{x}_t\}$ for $t = 1, \ldots, N$. Then also $\check{x}_N \equiv \hat{x}_N$ and $\check{\mathsf{P}}(N) \equiv \hat{\mathsf{P}}(N)$, because for $t = N$, all data are known in any case, and from (11.151), one may the infer the $\check{x}_t$ for $t = N-1, \ldots, 1$.

To determine the variance

$$\check{\mathsf{P}}(t) = \langle (x_t - \check{x}_t)^2 \rangle, \qquad (11.153)$$

one writes

$$x_t - \check{x}_t = (x_t - \hat{x}_t) - (\check{x}_t - \hat{x}_t) \qquad (11.154)$$

and because of

$$\begin{aligned}\check{x}_t - \hat{x}_t &= \check{\mathsf{K}}(t)(\check{x}_{t+1} - \tilde{x}_{t+1}) \\ &= -\check{\mathsf{K}}(t)\Big((x_{t+1} - \check{x}_{t+1}) - (x_{t+1} - \tilde{x}_{t+1})\Big) \qquad (11.155)\end{aligned}$$

one obtains

$$\check{\mathsf{P}}(t) = \hat{\mathsf{P}}(t) + \check{\mathsf{K}}(t)(\check{\mathsf{P}}(t+1) - \tilde{\mathsf{P}}(t+1))\check{\mathsf{K}}^T(t). \qquad (11.156)$$

### *11.5.3 The Unscented Kalman Filter*

Björn Schelter

The is an alternative for the extended Kalman filter for nonlinear systems (11.116)

$$\mathbf{X}(t) = \mathbf{f}\big(\mathbf{X}(t-1)\big) + \mathbf{B}(t-1)\boldsymbol{\eta}(t) \tag{11.157}$$

$$\mathbf{Y}(t) = \mathbf{g}\big(t, \mathbf{X}(t)\big) + \boldsymbol{\epsilon}(t), \tag{11.158}$$

for the which the optimal predictions prior to the observation $\mathbf{Y}(t)$ read

$$\tilde{\mathbf{X}}(t) = \mathbf{f}\big(\hat{\mathbf{X}}(t-1)\big) \tag{11.159}$$

$$\tilde{\mathbf{Y}}(t) = \mathbf{g}\big(t, \tilde{\mathbf{X}}(t)\big). \tag{11.160}$$

While the extended Kalman filter relies on the linearization of the nonlinear function $\mathbf{f}\big(\mathbf{X}(t-1)\big)$, the unscented Kalman filter (UKF) is motivated by the fact that it is easier to approximate a distribution by a Gaussian than to approximate an arbitrary nonlinear function by linearization (Julier et al. 1995, 2000; Voss et al. 2004). can be accomplished by taking into account the values of $\mathbf{f}(\mathcal{X}_i)$, $i = 1, \ldots, 2d$ where $\mathcal{X}_i$, $i = 1, \ldots, 2d$ are the so-called $\sigma$-points, which originate from the vicinity of the $d$-dimensional vector $\hat{\mathbf{X}}(t-1)$. The resulting $\mathbf{f}(\mathcal{X}_i)$, $i = 1, \ldots, 2d$ are averaged to obtain the new optimal prediction. In this way, the full nonlinearity of $\mathbf{f}(\cdot)$ is captured. Of course, the choice of the $\sigma$-points is crucial.

For the linear Kalman filter, the matrix $\mathbf{P}(t)$ consisting of (cf. (11.105))

$$\hat{P}_{\alpha\beta}(t) \equiv \langle (\mathbf{X}(t) - \hat{\mathbf{X}}(t))_\alpha \; (\mathbf{X}(t) - \hat{\mathbf{X}}(t))_\beta \rangle \tag{11.161}$$

quantifies the deviation of the estimated $\hat{\mathbf{X}}(t)$ and the true $\mathbf{X}(t)$. Therefore, the matrix $\hat{\mathbf{P}}(t)$ is a sensible measure for the uncertainty and should be used to define the $d$-dimensional $\sigma$-points

$$\mathcal{X}_i(t) = \hat{\mathbf{X}}(t) + \big(\sqrt{d\,\mathbf{P}(t)}\big)_i \quad i = 1, \ldots, d \tag{11.162}$$

$$\mathcal{X}_{i+d}(t) = \hat{\mathbf{X}}(t) - \big(\sqrt{d\,\mathbf{P}(t)}\big)_i \quad i = 1, \ldots, d \tag{11.163}$$

around the $d$-dimensional $\hat{\mathbf{X}}(t)$. Please note that the square root refers to the square root of the matrix. The notation $(\cdot)_i$ denotes here the $i$th column or row of $(\cdot)$.

As an illustrative example, assume that

$$\mathbf{P} = \begin{pmatrix} P_1^2 & 0 & 0 \\ 0 & P_2^2 & 0 \\ 0 & 0 & P_3^2 \end{pmatrix} \tag{11.164}$$

which would lead to the set of $\sigma$-points

$$\sqrt{3}\left\{ \begin{pmatrix} P_1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ P_2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ P_3 \end{pmatrix}, -\begin{pmatrix} P_1 \\ 0 \\ 0 \end{pmatrix}, -\begin{pmatrix} 0 \\ P_2 \\ 0 \end{pmatrix}, -\begin{pmatrix} 0 \\ 0 \\ P_3 \end{pmatrix} \right\} \tag{11.165}$$

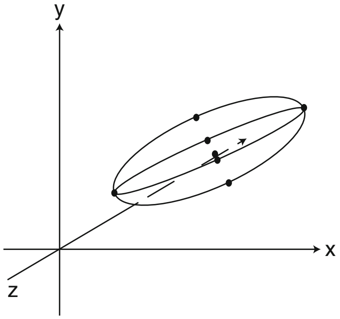


**Fig. 11.13** Schematic drawing of the location of the mean and the $\sigma$-points for a three dimensional example

if $\hat{\mathbf{X}}(t)$ is zero. The $\sigma$-points are, thus, exactly the $1\sigma$ quantiles of a multivariate Gaussian distribution centered around $\hat{\mathbf{X}}(t)$, the best estimator for $\mathbf{X}(t)$ (Fig. 11.13).

With

$$\mathcal{X}_i(t) = \mathbf{f}\big(\mathcal{X}_i(t-1)\big), \tag{11.166}$$

we set now for the optimal prediction prior to the observation $\mathbf{Y}(t)$:

$$\tilde{\mathbf{X}}(t) = \frac{1}{2d}\sum_{i=1}^{2d}\mathcal{X}_i(t). \tag{11.167}$$

The corresponding $\mathbf{P}_{\mathbf{XX}}(t)$ then results in

$$\mathbf{P}_{\mathbf{XX}}(t) = \frac{1}{2d}\sum_{i=1}^{2d}\big(\mathcal{X}_i(t) - \hat{\mathbf{X}}(t)\big)\big(\mathcal{X}_i(t) - \hat{\mathbf{X}}(t)\big)'. \tag{11.168}$$

By doing this we get now for the prediction up to the second order in a Taylor expansion (Julier and Uhlmann 1996)

$$\tilde{\mathbf{X}}(t) = \mathbf{f}\big(\hat{\mathbf{X}}(t-1)\big) + \frac{\nabla_{\mathbf{X}}'\mathbf{P}_{\mathbf{XX}}(t-1)\nabla_{\mathbf{X}}}{2}\mathbf{f}\big(\mathbf{X}(t-1)\big)\Big|_{\mathbf{X}=\hat{\mathbf{X}}} \tag{11.169}$$

and

$$\mathbf{P}_{\mathbf{XX}}(t) = \nabla\mathbf{f}\big(\hat{\mathbf{X}}(t-1)\big)\mathbf{P}_{\mathbf{XX}}(t-1)\big(\nabla\mathbf{f}\big(\hat{\mathbf{X}}(t-1)\big)\big)'. \tag{11.170}$$

This presents a more accurate results than the one that we would obtain for the extended Kalman filter for which we have

$$\tilde{\mathbf{X}}(t) = \mathbf{f}\big(\hat{\mathbf{X}}(t-1)\big) \tag{11.171}$$

and also

$$\mathbf{P_{XX}}(t) = \nabla\mathbf{f}\big(\hat{\mathbf{X}}(t-1)\big)\mathbf{P_{XX}}(t-1)\big(\nabla\mathbf{f}\big(\hat{\mathbf{X}}(t-1)\big)\big)' \tag{11.172}$$

but with $\mathbf{P_{XX}}(t-1)$ as

$$\mathbf{P_{XX}}(t-1) = \big\langle\big(\mathbf{X}(t-1) - \hat{\mathbf{X}}(t-1)\big)\big(\mathbf{X}(t-1) - \hat{\mathbf{X}}(t-1)\big)'\big\rangle. \tag{11.173}$$

Thus, by replacing the prior estimates $\tilde{\mathbf{X}}(t)$, $\tilde{\mathbf{Y}}(t)$ and the covariances $\tilde{\mathbf{P}}_{\mathbf{XX}}$, $\tilde{\mathbf{P}}_{\mathbf{XY}}$, $\tilde{\mathbf{P}}_{\mathbf{YY}}$ of the ordinary Kalman filter by

$$\tilde{\mathbf{X}}(t) = \frac{1}{2d}\sum_{i=1}^{2d}\tilde{\mathcal{X}}_i(t), \quad \text{with } \tilde{\mathcal{X}}_i(t) = \mathbf{f}\big(\mathcal{X}_i(t-1)\big) \tag{11.174}$$

$$\tilde{\mathbf{Y}}(t) = \frac{1}{2d}\sum_{i=1}^{2d}\tilde{\mathcal{Y}}_i(t), \quad \text{with } \tilde{\mathcal{Y}}_i(t) = \mathbf{g}\big(\mathcal{X}_i(t)\big) \tag{11.175}$$

$$\tilde{\mathbf{P}}_{\mathbf{XX}}(t) = \frac{1}{2d}\sum_{i=1}^{2d}\big(\tilde{\mathcal{X}}_i(t) - \tilde{\mathbf{X}}(t)\big)\big(\tilde{\mathcal{X}}_i(t) - \tilde{\mathbf{X}}(t)\big)' \tag{11.176}$$

$$\tilde{\mathbf{P}}_{\mathbf{XY}}(t) = \frac{1}{2d}\sum_{i=1}^{2d}\big(\tilde{\mathcal{X}}_i(t) - \tilde{\mathbf{X}}(t)\big)\big(\tilde{\mathcal{Y}}_i(t) - \tilde{\mathbf{Y}}(t)\big)' \tag{11.177}$$

$$\tilde{\mathbf{P}}_{\mathbf{YY}}(t) = \frac{1}{2d}\sum_{i=1}^{2d}\big(\tilde{\mathcal{Y}}_i(t) - \tilde{\mathbf{Y}}(t)\big)\big(\tilde{\mathcal{Y}}_i(t) - \tilde{\mathbf{Y}}(t)\big)' + \mathbf{R} \tag{11.178}$$

we obtain the uncented Kalman filter (Voss et al. 2004). The equations remain unchanged. Computationally the unscented Kalman filter is slower as the ordinary Kalman filter as all $\sigma$-points have to be kept track of. This, however, is much faster and much more robust than calculating the Jacobian for the extended Kalman filter, in particular in those cases where an analytical derivative cannot be obtained.

The unscented Kalman filter as discussed above makes use of the $\sigma$-points only. The unscented Kalman filter can potentially be biased in cases where heavy tail distributions are present or on cases where multimodal distributions play an important role. Practically, the unscented Kalman filter can often also be applied also in these cases.

### *11.5.4 The Dual Kalman Filter*

Björn Schelter

This chapter is concerned with the estimation of parameters of the state space model

$$\mathbf{X}(t) = \mathbf{A}\,\mathbf{X}(t-1) + \boldsymbol{\eta}(t), \qquad \boldsymbol{\eta}(t) \sim \mathrm{WN}(0, \mathbf{Q_X}) \tag{11.179}$$

$$\mathbf{Y}(t) = \mathbf{C}\,\mathbf{X}(t) + \boldsymbol{\epsilon}(t), \qquad \boldsymbol{\epsilon}(t) \sim \mathrm{WN}(0, \mathbf{R}), \tag{11.180}$$

in which the parameter matrices $\mathbf{A}$ are assumed to be a function of time $\mathbf{A}(t)$. A straight forward way to approach this, is to augment the hidden process by a difference equation for the parameters

$$\mathbf{A}(t) = \mathbf{A}(t-1) + \boldsymbol{\Xi}(t), \tag{11.181}$$

$$\mathbf{X}(t) = \mathbf{A}\,(t-1)\;\mathbf{X}(t-1) + \boldsymbol{\eta}(t), \qquad \boldsymbol{\eta}(t) \sim \mathrm{WN}(0, \mathbf{Q}_\mathbf{X}) \tag{11.182}$$

$$\mathbf{Y}(t) = \mathbf{C}\,\mathbf{X}(t) + \boldsymbol{\epsilon}(t), \qquad \boldsymbol{\epsilon}(t) \sim \mathrm{WN}(0, \mathbf{R}). \tag{11.183}$$

For higher order vector autoregressive processes we use (5.273) and (5.274) to rewrite the model as a vector autoregressive model of order one.

The entries in the matrix $\boldsymbol{\Xi}$ are all considered to be independent random variables with zero mean and variance $\sigma_{ij}^{\xi}$. Thus, by the model (Kitagawa and Gersch 1996)

$$a_{ij}(t) = a_{ij}(t-1) + \xi_{ij}(t)$$

the capability of a change of the parameter $a_{ij}(t)$ within one time step is formulated.

To estimate the time-dependent parameters $\mathbf{a}(t)$ together with the hidden process $\mathbf{X}(t)$ we can use the dual Kalman filter (Wan and Nelson 1997). The idea of the dual Kalman filter is an estimation in two steps. In the first step the parameters $\mathbf{A}(t)$ are assumed to be known. Equation (11.182) and (11.183) boil down to a standard linear state space model

$$\mathbf{X}(t) = \mathbf{A}(t-1)\,\mathbf{X}(t-1) + \mathbf{j}(t) \tag{11.184}$$

$$\mathbf{Y}(t) = \mathbf{C}\mathbf{X}(t) + \boldsymbol{\epsilon}(t), \tag{11.185}$$

in which the vector $\mathbf{A}(t-1)$ determines the time varying entries of the coefficient matrices of the VAR$[p]$ process.

In the second step, we use a second state space model, for which we assume that the $\mathbf{X}(t)$ are known parameters. Formally, (11.182) and (11.183) can be summarized in one single observation equation by substituting $\mathbf{X}(t)$ in (11.182) by (11.183) leading to

$$\mathbf{A}(t) = \mathbf{A}(t-1) + \boldsymbol{\Xi}(t) \tag{11.186}$$

$$\begin{aligned} \mathbf{Y}(t) &= \mathbf{C}\,[\mathbf{A}\,(t-1)\;\mathbf{X}(t-1) + \boldsymbol{\eta}(t)] + \boldsymbol{\epsilon}(t), \\ &\qquad \boldsymbol{\epsilon}(t) \sim \mathrm{WN}(0, \mathbf{R}), \boldsymbol{\eta}(t) \sim \mathrm{WN}(0, \mathbf{Q}_\mathbf{X}). \end{aligned} \tag{11.187}$$

Iterating these two steps leads to a convergence to the solution of the full state space model of the (11.182)–(11.183).

The dual Kalman filter consists of one linear Kalman filter for the process state space (11.184), (11.185) and another one for the parameter state space (11.186) and (11.187). The process Kalman filter calculates the optimal estimates $\mathbf{X}(t)$ given the

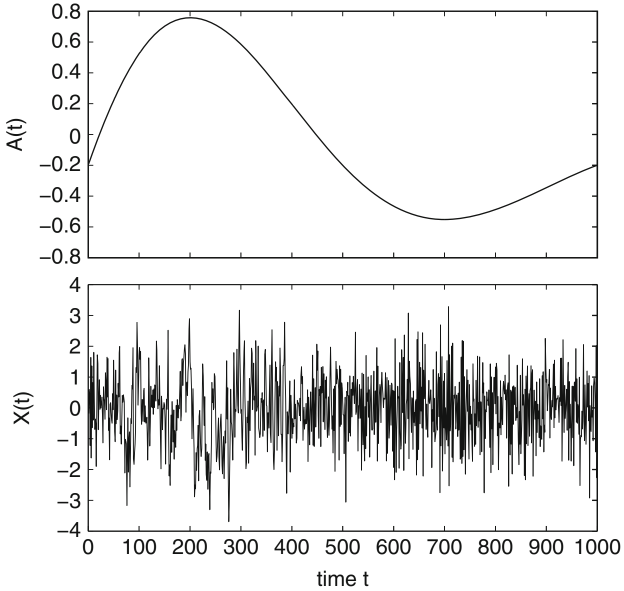


**Fig. 11.14** A realization of a nonstationary AR[1] process is shown over the time together with the time-dependent AR coefficient $A(t)$ (cf. Sommerlade et al. 2012)

time-dependent parameters $\mathbf{A}(t-1)$. The parameter Kalman filter computes the optimal estimates for $\mathbf{A}(t)$ given the process state vectors $\mathbf{X}(t-1)$.

**Improvement of the dual Kalman filter.** The dual Kalman filter so far described neglects the estimation error for $\mathbf{A}\,(t-1)$ given all observations up to time $t-1$ in the process state space (11.184) and (11.185). The variance of the random variable $\mathbf{X}(t) = \mathbf{A}\,(t-1)\,\mathbf{X}\,(t-1)$ given the observations $\{\mathbf{Y}(1), \ldots, \mathbf{Y}(t-1)\}$ is underestimated, because it neglects the variance of $\mathbf{A}\,(t-1)$. Instead of considering $\mathbf{A}\,(t-1)$ as exact, it can be treated as a random variable and the nonlinear components of the random variable $\mathbf{X}((t) = \mathbf{A}\,(t-1)\,\mathbf{X}(t-1)$ can be aproximated by a first order Taylor expansion. The results can be found in Sommerlade et al. (2012).

**Choosing the smoothness prior.** Until now, we assumed that the variances of $\boldsymbol{\Xi}$, here called $\mathbf{Q}_A$, the variances $\mathbf{Q}_X$ and $\mathbf{R}$ of the two state space models were known. In the following, we demonstrate the importance of an adequate choice of $\mathbf{Q}_A$ (Kitagawa and Gersch 1996), while $\mathbf{Q}_X$ and $\mathbf{R}$ are still assumed to be given.

To demonstrate the effect of the choice of $Q_A$ using the dual Kalman filter and dual smoothing filter, a one dimensional AR[1] process has been simulated with a time varying parameter $A(t) = -0.2 + 1.5 \sin\left(\frac{2\pi t}{1{,}000}\right) \exp\left(-0.002(t-1)\right)$. The variance of the system noise $\eta_(t)$ has been chosen $Q_X = 1$. In Fig. 11.14,

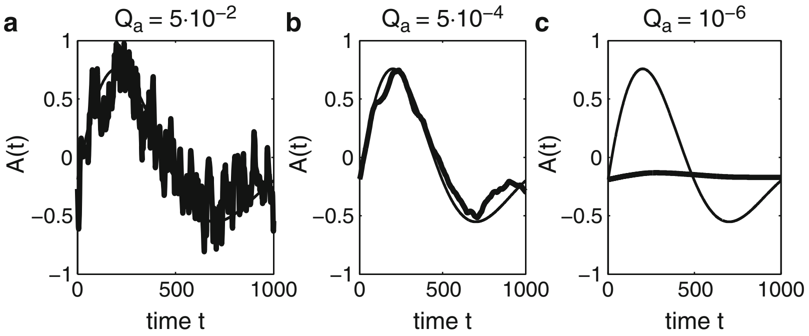


**Fig. 11.15** The time-dependent parameter $A(t)$ of an AR[1] process $X(t) = A(t)X(t-1) + \eta(t)$ with $\eta(t) \sim$ WN(0, 1) are presented in *black*. The time-dependent parameters are estimated by the dual smoothing filter. System noise was set to $Q_X = 1$, the observation noise to $R = 0.5$ and three different parameter noise variances $Q_A = 5 \cdot 10^{-2}, 5 \cdot 10^{-4}, 1 \cdot 10^{-6}$ are used for the dual smoothing filter. The results of the estimation are plotted as the *thick black lines* (cf. Sommerlade et al. 2012)

a realization of the process $X(t)$ and the time-depending AR coefficient $A(t)$ are shown.

The observation of the time-dependent AR[1] process $x_t$ is contaminated with observation noise $\eta(t) \sim \mathrm{N}(0, R = 0.5)$. Since the variance of the AR process depends on $A(t)$ and thus on time, the signal to noise ratio lies between 2 and 4.5.

In Fig. 11.15, the time-dependent parameter $A(t)$ (thin black line) of a one dimensional AR[1] process estimated (thick black line) by the dual smoothing filter. The process and observation noise variances have been set to the true values. The influence of different parameter noise variances is demonstrated. In (a), the parameter noise variance $Q_A = 5 \cdot 10^{-2}$ is chosen too large and the estimator for $A(t)$ fluctuates around the true parameter curve. In (c), the $Q_A = 1 \cdot 10^{-6}$ is very small and the Kalman smoother results cannot follow the true parameter values. In (b), the estimated parameters capture well the true parameters with a parameter noise variance of $Q_A = 5 \cdot 10^{-4}$. This example demonstrates that the parameter $A(t)$ can only be estimated well if the parameters $Q_A$, $Q_X$ and $R$ are well known.

If such parameter are unknown, they have to be estimated from the observations. This topic is addressed in the next chapter.

# Chapter 12
# Estimating the Parameters of a Hidden Stochastic Model

Having discussed the algorithms for estimating the realization of the hidden process $\{x_t\}$ from the observations $\{y_t\}$, we will now introduce some methods for estimating the parameters in the model by which one wants to interpret the observations. In the first case, this will be a hidden Markov model, in the second case, a state space model will be assumed. In each case, all of the parameters will be collected in the vector $\boldsymbol{\theta}$. As an estimator we use the maximum-likelihood estimator, so that we have to find the maximum of the likelihood $\rho(y_{1\ldots N}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. In general, the likelihood is the sum over all possible paths $x_{1\ldots N}$ of the hidden process

$$\begin{aligned} L = \rho(y_{1\ldots N} \mid \boldsymbol{\theta}) &= \sum_{x_{1\ldots N}} \rho(y_{1\ldots N}, x_{1\ldots N} \mid \boldsymbol{\theta}) \\ &= \sum_{x_{1\ldots N}} \rho(y_{1\ldots N} \mid x_{1\ldots N})\, \rho(x_{1\ldots N} \mid \boldsymbol{\theta})\,. \end{aligned} \tag{12.1}$$

This is a sum over $m^N$ terms, when $m$ is the number of discrete states. In nearly every case, this summation cannot be done.

However, the special dependence structure of the random variables $\{X_t, Y_t\}$ in the hidden Markov model is such that the likelihood of the hidden system is of the form

$$\begin{aligned} \rho(y_{1\ldots N}, x_{1\ldots N}|\boldsymbol{\theta}) &= \rho(y_{1\ldots N}|x_{1\ldots N}, \boldsymbol{\theta})\rho(x_{1\ldots N}|\boldsymbol{\theta}) \\ &= \prod_{t=2}^{N} \rho(y_t|x_t, \boldsymbol{\theta})\rho(x_t|x_{t-1}, \boldsymbol{\theta})\, \rho(y_1|x_1, \boldsymbol{\theta})\rho(x_1|\boldsymbol{\theta}). \end{aligned} \tag{12.2}$$

This will allow calculating the likelihood recursively for any given set $\Theta$ of parameters, as we will show. Thus the estimate of the parameters can be found by one of the known maximization routines.

This step of brute force maximization can furthermore be avoided and can be replaced by an iterative determination of another quantity. The key for a precise

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_12, © Springer-Verlag Berlin Heidelberg 2012

formulation of such an iterative method is the expectation maximization method (EM method, Dempster et al. 1977). This method will be introduced in Sect. 12.1 and in the following sections it will be applied to the hidden Markov model and to the state space model.

## 12.1 The Expectation Maximization Method (EM Method)

With the EM method, one can construct a sequence of parametric vectors $\{\boldsymbol{\theta}^{(k)}\}$ that converge toward the maximum of the likelihood. The definition of $\boldsymbol{\theta}^{(k+1)}$ for given $\boldsymbol{\theta}^{(k)}$ consists of two steps:

- In the first step, an expectation value is calculated, namely, the expectation of the function $\ln \rho(y_{1\ldots N}, x_{1\ldots N} \mid \boldsymbol{\theta})$ with respect to the density $\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)})$, that is

$$\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \sum_{x_{1\ldots N}} \Big[ \ln \rho(y_{1\ldots N}, x_{1\ldots N} \mid \boldsymbol{\theta}) \Big] \rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(\boldsymbol{k})}) \,. \tag{12.3}$$

- In the second step, the function $\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\boldsymbol{k})})$ will be maximized with respect to $\boldsymbol{\theta}$, and the value of $\boldsymbol{\theta}$ at the maximum will be called $\boldsymbol{\theta}^{(k+1)}$:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) \,. \tag{12.4}$$

We will show that this leads to maximization of the log likelihood $\mathrm{L}(\boldsymbol{\theta} \mid y_{1\ldots N}) = \ln \rho(y_{1\ldots N} \mid \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

First we will show that always

$$\mathrm{L}(\boldsymbol{\theta}^{(k+1)} \mid y_{1\ldots N}) \geq \mathrm{L}(\boldsymbol{\theta}^{(k)} \mid y_{1\ldots N}) \,. \tag{12.5}$$

*Proof.* Because of

$$\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}) = \frac{\rho(x_{1\ldots N}, y_{1\ldots N} \mid \boldsymbol{\theta})}{\rho(y_{1\ldots N} \mid \boldsymbol{\theta})}, \tag{12.6}$$

$$\mathrm{L}(\boldsymbol{\theta} \mid y_{1\ldots N}) \equiv \ln \rho(y_{1\ldots N} \mid \boldsymbol{\theta}) = \ln \rho(x_{1\ldots N}, y_{1\ldots N} \mid \boldsymbol{\theta}) - \ln \rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}). \tag{12.7}$$

We multiply this equation by $\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)})$ and sum over all paths. Then we obtain

$$\mathrm{L}(\boldsymbol{\theta} \mid y_{1\ldots N}) = \mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) - \mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) \tag{12.8}$$

with

$$\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \sum_{x_{1\ldots N}} \ln \rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta})\, \rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)}). \tag{12.9}$$

On the other hand,

$$\begin{aligned}\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = &\sum_{x_{1\ldots N}} \ln \rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)})\, \rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)}) \\ &+ \mathrm{S}[\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}]\,,\end{aligned} \tag{12.10}$$

where $\mathrm{S}[\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}]$ is the relative entropy of the density $\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)})$ given the density $\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta})$:

$$\mathrm{S}[\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}] = -\sum_{x_{1\ldots N}} \ln \left( \frac{\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)})}{\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta})} \right) \rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)}). \tag{12.11}$$

Because the first term of $\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ in (12.10) does not depend on $\boldsymbol{\theta}$ and because the relative entropy obeys: $\mathrm{S}[\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}] \leq 0$, we can conclude that

$$\mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) \leq \mathrm{B}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})\,. \tag{12.12}$$

Now we construct

$$\begin{aligned}\mathrm{L}(\boldsymbol{\theta} \mid y_{1\ldots N}) - \mathrm{L}(\boldsymbol{\theta}^{(k)} \mid y_{1\ldots N}) = &[\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) - \mathrm{A}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})] \\ &+ [\mathrm{B}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) - \mathrm{B}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})].\end{aligned} \tag{12.13}$$

For

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}), \tag{12.14}$$

we now have

$$\mathrm{A}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) \geq \mathrm{A}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}), \tag{12.15}$$

$$\mathrm{B}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) \geq \mathrm{B}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}), \tag{12.16}$$

and therefore,

$$\mathrm{L}(\boldsymbol{\theta}^{(k+1)} \mid y_{1\ldots N}) \geq \mathrm{L}(\boldsymbol{\theta}^{(k)} \mid y_{1\ldots N})\,, \tag{12.17}$$

which completes the proof.

Thus on each sequence of parametric vectors $(\ldots, \boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}, \ldots)$ which is constructed with the EM method, the likelihood always increases or, at least, stays constant. Let us denote by $\boldsymbol{\theta}^*$ the parametric vector to which the sequence converges. One may start such a sequence at any point in a region around $\boldsymbol{\theta}^*$. Thus $\boldsymbol{\theta}^*$ is also a maximum of the likelihood.

Now, the likelihood of the hidden system is of the form

$$\begin{aligned}\rho(y_{1\ldots N}, x_{1\ldots N}|\boldsymbol{\theta}) &= \rho(y_{1\ldots N}|x_{1\ldots N}, \boldsymbol{\theta})\rho(x_{1\ldots N}|\boldsymbol{\theta}) \\ &= \prod_{t=2}^{N} \rho(y_t|x_t, \boldsymbol{\theta})\rho(x_t|x_{t-1}, \boldsymbol{\theta})\, \rho(y_1|x_1, \boldsymbol{\theta})\rho(x_1|\boldsymbol{\theta}), \quad (12.18)\end{aligned}$$

so that the log likelihood is a sum of terms such as $\ln \rho(y_t|x_t, \boldsymbol{\theta})$ or $\ln \rho(x_t|x_{t-1}, \boldsymbol{\theta})$, which contain only a few of the arguments $\{x_t\}$ and $\{y_t\}$. Thus in calculating the expectation value of such terms with respect to the density $\rho(x_{1\ldots N} \mid y_{1\ldots N}, \boldsymbol{\theta}^{(k)})$, the sum over all other arguments means a marginalization and we obtain e.g. contributions of the type

$$\begin{aligned}T_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \sum_{x_{1\ldots N}} \ln \rho(x_{t+1}|x_t, \boldsymbol{\theta})\, \rho(x_{1\ldots N}|y_{1\ldots N}, \boldsymbol{\theta}^{(k)}) \\ &= \sum_{x_{t+1}, x_t} \ln \rho(x_{t+1}|x_t, \boldsymbol{\theta})\, \rho(x_t, x_{t+1}|y_{1\ldots N}, \boldsymbol{\theta}^{(k)}) \qquad (12.19)\end{aligned}$$

and

$$\begin{aligned}T_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \sum_{x_{1\ldots N}} \ln \rho(y_t|x_t, \boldsymbol{\theta})\, \rho(x_{1\ldots N}|y_{1\ldots N}, \boldsymbol{\theta}^{(k)}) \\ &= \sum_{x_t} \ln \rho(y_t|x_t, \boldsymbol{\theta})\, \rho(x_t|y_{1\ldots N}, \boldsymbol{\theta}^{(k)})\,. \qquad (12.20)\end{aligned}$$

Thus the most important quantity is therefore the density $\rho(x_t, x_{t+1}|y_{1\ldots N}, \boldsymbol{\theta}^{(k)})$, which is the probability density for $X_t$ and $X_{t+1}$ for given $y_{1\ldots N}$ and given parameters. From this, one may also find the quantity $\rho(x_t|y_{1\ldots N}, \boldsymbol{\theta}^{(k)})$ by marginalization with respect to $x_{t+1}$.

When the state space is discrete, the probability $\Psi_{ji}(t) \equiv \rho(x_t = i, x_{t+1} = j|y_{1\ldots N}, \boldsymbol{\theta})$ can be estimated for given data and a given parameter set $\boldsymbol{\theta}$, as we will show in the following subsection.

For a continuous state space as in the usual state space model, this is not a good strategy because estimating a continuous density functions is notoriously difficult. One should then better use the fact that the properties of the density $\rho(\boldsymbol{x}_t|\boldsymbol{y}_{1\ldots N}, \Theta^{(i)})$ are determined by the solution of the Kalman smoothing filter: We had

$$\check{\boldsymbol{x}}(t) = \arg\max_{\boldsymbol{x}_t} \rho(\boldsymbol{x}_t|\boldsymbol{y}_{1\ldots N}, \boldsymbol{\theta}^{(k)})\,. \qquad (12.21)$$

The maximum of the density can also be identified from the expectation value and therefore also

$$\check{x}_i(t) =< X_i(t) >= \int d\boldsymbol{x}\, x_i\, \rho(\boldsymbol{x}|\boldsymbol{y}_{1\ldots N}, \boldsymbol{\theta}^{(k)}) \,. \tag{12.22}$$

Furthermore, the second moments $X_i(t)X_j(t)$ can be written as

$$< X_i(t)X_j(t) >= \mathrm{Cov}(X_i(t), X_j(t)) - < X_i >< X_j >, \tag{12.23}$$

and the covariance matrix can be identified from the matrix $\hat{\mathsf{P}}(t)$ determined in the Kalman smoothing filter:

$$< (X_i(t) - \check{x}_i(t))(X_j(t) - \check{x}_j(t))^T >= (\check{\mathsf{P}})_{ij}(t) \,. \tag{12.24}$$

Finally, the task of finding the expectation value of $\ln \rho(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})$ may also lead to the problem of calculating the moment $< X_i(t)X_j(t-1) >$ or $< X_i(t)X_j(t+1) >$. The key to such calculations is the observation that the forecast error $\boldsymbol{X}(t) - \hat{\boldsymbol{x}}(t)$ is independent of $\boldsymbol{\eta}(t+1)$ because $\hat{\boldsymbol{x}}(t)$ is only the estimation of $\boldsymbol{X}(t)$ for given data $\boldsymbol{y}_{1\ldots t}$ and the noise $\boldsymbol{\eta}(t+1) = \boldsymbol{X}(t+1) - \mathsf{A}\boldsymbol{X}(t)$ is independent of $\boldsymbol{X}(t)$. Thus $< (\boldsymbol{X}(t) - \hat{\boldsymbol{x}}(t))\boldsymbol{X}(t+1)^T >$ can be calculated easily with the help of $\boldsymbol{X}(t+1) = \mathsf{A}\boldsymbol{X}(t) + \boldsymbol{\eta}(t+1)$, leading to

$$< (\boldsymbol{X}(t) - \hat{\boldsymbol{x}}(t))\boldsymbol{X}(t+1)^T >= \hat{\mathsf{P}}(t)\mathsf{A}^T + (\check{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t))\hat{\boldsymbol{x}}(t)^T\mathsf{A}^T. \tag{12.25}$$

If we write the expression $< (\boldsymbol{X}(t) - \check{\boldsymbol{x}}(t))\boldsymbol{X}(t+1)^T >$ in

$$< \boldsymbol{X}(t)\boldsymbol{X}(t+1)^T >=< (\boldsymbol{X}(t) - \check{\boldsymbol{x}}(t))\boldsymbol{X}(t+1)^T > +\check{\boldsymbol{x}}(t)\check{\boldsymbol{x}}(t+1)^T \tag{12.26}$$

as

$$<(\boldsymbol{X}(t)-\check{\boldsymbol{x}}(t))\boldsymbol{X}(t+1)^T >=< \Big(\boldsymbol{X}(t)-\hat{\boldsymbol{x}}(t)-(\check{\boldsymbol{x}}(t)-\hat{\boldsymbol{x}}(t))\Big)\boldsymbol{X}(t+1)^T> \tag{12.27}$$

and use $\check{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t) = \check{\mathsf{K}}(t)(\check{\boldsymbol{x}}(t+1) - \tilde{\boldsymbol{x}}(t+1)) - \check{\mathsf{K}}(t)\boldsymbol{X}(t+1) + \check{\mathsf{K}}(t)\boldsymbol{X}(t+1)$, we get

$$\begin{aligned}
&< (\boldsymbol{X}(t) - \check{\boldsymbol{x}}(t))\boldsymbol{X}(t+1)^T > \\
&=< \big(\boldsymbol{X}(t) - \hat{\boldsymbol{x}}(t) - \check{\mathsf{K}}(t)\check{\boldsymbol{x}}(t+1) + \check{\mathsf{K}}(t)\tilde{\boldsymbol{x}}(t+1) \\
&\quad +\check{\mathsf{K}}(t)\boldsymbol{X}(t+1) - \check{\mathsf{K}}(t)\boldsymbol{X}(t+1)\big)\boldsymbol{X}(t+1)^T > \\
&=< (\boldsymbol{X}(t) - \hat{\boldsymbol{x}}(t))\boldsymbol{X}(t+1)^T > \\
&\quad +\check{\mathsf{K}}(t) < (\boldsymbol{X}(t+1) - \check{\boldsymbol{x}}(t+1))\boldsymbol{X}(t+1)^T > \\
&\quad -\check{\mathsf{K}}(t) < (\boldsymbol{X}(t+1) - \tilde{\boldsymbol{x}}(t+1))\boldsymbol{X}(t+1)^T >
\end{aligned}$$

$$\begin{aligned}&= \hat{\mathsf{P}}(t)\mathsf{A}^T + (\check{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t))\hat{\boldsymbol{x}}(t)^T\mathsf{A}^T + \check{\mathsf{K}}(t)\check{\mathsf{P}}(t+1)\\ &\quad -\check{\mathsf{K}}(t)\tilde{\mathsf{P}}(t+1) - \check{\mathsf{K}}(t)(\check{\boldsymbol{x}}(t+1) - \tilde{\boldsymbol{x}}(t+1))\tilde{\boldsymbol{x}}(t+1)^T\\ &= \check{\mathsf{K}}(t)\check{\mathsf{P}}(t+1),\end{aligned} \tag{12.28}$$

where for the last equality we have used that $\hat{\mathsf{P}}(t)\mathsf{A}^T = \check{\mathsf{K}}(t)\tilde{\mathsf{P}}(t+1)$ (see (11.113)) and $\check{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t) - \check{\mathsf{K}}(t)(\check{\boldsymbol{x}}(t+1) - \tilde{\boldsymbol{x}}(t+1)) = 0$ and also $\mathsf{A}\hat{\boldsymbol{x}}(t) = \tilde{\boldsymbol{x}}(t+1)$. Thus we get finally

$$< \boldsymbol{X}(t)\boldsymbol{X}(t+1)^T >= \check{\mathsf{K}}(t)\check{\mathsf{P}}(t+1) + \check{\boldsymbol{x}}(t)\check{\boldsymbol{x}}(t+1)\,. \tag{12.29}$$

In the following sections, we will use these considerations to estimate the parameters in the hidden Markov model and in the state space model.

## 12.2 Estimating the Parameters of a Hidden Markov Model

The most important quantity for calculating the likelihood for a hidden Markov model are the probabilities

$$\alpha_i(t) = \rho(y_{1\ldots t}, x_t = i|\boldsymbol{\theta}) \tag{12.30}$$

and

$$\alpha_i^{(t)}(t) = \frac{\rho(y_{1\ldots t}, x_t = i|\boldsymbol{\theta})}{\rho(y_{1\ldots t}|\boldsymbol{\theta})} = \rho(x_t = i|y_{1\ldots t}, \boldsymbol{\theta})\,. \tag{12.31}$$

We get

$$\sum_i \alpha_i(t) = \sum_i \rho(y_{1\ldots t}, x_t = i|\boldsymbol{\theta}) = \rho(y_{1\ldots t}|\boldsymbol{\theta}) \tag{12.32}$$

and especially

$$L = \sum_i \rho(y_{1\ldots N}, x_N = i|\boldsymbol{\theta}) = \sum_i \alpha_i(t = N)\,. \tag{12.33}$$

The conditional dependence structure of the hidden Markov model is such that the $\{\alpha_j(t)\}$ can be calculated iteratively. For $t = 1$ we have (not indicating the dependence on the parametric set $\boldsymbol{\theta}$ for a while)

$$\alpha_i(1) = \rho(y_1|x_1 = i)\rho(x_1 = i), \tag{12.34}$$

and with given $\alpha_j(t-1)$ for all $j$, we find $\alpha_i(t)$ by

$$\alpha_i(t) = \sum_j \rho(y_t|x_t = i)\rho(x_t = i|x_{t-1} = j)\alpha_j(t-1), \quad t = 2, \ldots, N\,. \tag{12.35}$$

The $\{\alpha_j(t)\}$ decrease quickly with increasing time $t$. To avoid a numerical underflow, one has to deal with the corresponding quantities $\alpha_i^{(t)}(t)$ which will be derived in each step by a normalization:

### 12.2.1 The Forward Algorithm

In the first step, $\alpha_i^{(1)}(1)$ can be obtained from $\alpha_i(1)$ by multiplying by a factor $\pi^{(1)}$ so that

$$\sum_i \alpha_i^{(1)}(1) = \sum_i \pi^{(1)} \rho(y_1, x_1 = i) = \sum_i \pi^{(1)} \alpha_i(1) = 1, \tag{12.36}$$

from which we conclude that

$$\pi^{(1)} = \frac{1}{\sum_i \alpha_i(1)} = \frac{1}{\rho(y_1)}\,. \tag{12.37}$$

In the next step, we first define

$$\alpha_i^{(1)}(2) = \pi^{(1)} \alpha_i(2) \equiv \sum_j \rho(y_2|x_2 = i)\rho(x_2 = i|x_1 = j)\alpha_j^{(1)}(1), \tag{12.38}$$

and for normalization, we multiply $\alpha_i^{(1)}(2)$ by a factor $\pi^{(2)}$ to get

$$\alpha_i^{(2)}(2) = \pi^{(2)}\pi^{(1)}\alpha_i(2)\,. \tag{12.39}$$

Normalization of $\alpha_i^{(2)}(2)$ leads to

$$\pi^{(1)}\pi^{(2)} = \frac{1}{\sum_i \alpha_i(2)} = \frac{1}{\rho(y_1, y_2)}\,. \tag{12.40}$$

Thus by normalizing in each step, we obtain the probabilities $\alpha_i^{(t)}(t) = \alpha_i(t)/\rho(y_{1\ldots t}) = \rho(x_t = i|y_{1\ldots t}, \boldsymbol{\theta})$, and we pick up normalization factors $\{\pi^{(i)}\}$, so that

$$\pi^{(1)}\pi^{(2)}\ldots\pi^{(N)} = \frac{1}{\sum_i \alpha_i(N)} = \frac{1}{\rho(y_{1\ldots N})}, \tag{12.41}$$

and we obtain thus for the log likelihood

$$\ln L = -\sum_{j=1}^{N} \ln \pi^{(j)}\,. \tag{12.42}$$

Hence the likelihood can be calculated easily for each parametric set $\boldsymbol{\theta}$, and by a numerical minimization routine one may find the minimum of the negative log likelihood with respect to $\boldsymbol{\theta}$.

### 12.2.2 The Backward Algorithm

There is, however, another method, by which one may calculate analytically the estimators for the parameters. For this one has to introduce the posterior probabilities

$$\beta_i(t) = \rho(y_{t+1\ldots N}|x_t = i). \tag{12.43}$$

Then, because of

$$\begin{aligned}\beta_i(t-1) &= \rho(y_{t\ldots N}|x_{t-1} = i)\\ &= \sum_j \rho(y_{t+1}\ldots N|x_t = j)\rho(y_t|x_t = j)\rho(x_t = j|x_{t-1} = i),\end{aligned} \tag{12.44}$$

we obtain the recursion relationship

$$\beta_i(t-1) = \rho(y_{t\ldots N}|x_{t-1} = i) = \sum_j \beta_j(t)\rho(y_t|x_t = j)\rho(x_t = j|x_{t-1} = i)\,. \tag{12.45}$$

On the other hand,

$$\beta_i(N-1) = \rho(y_N|x_{N-1} = i) = \sum_j \rho(y_N|x_N = j)\rho(x_N = j|x_{N-1} = i), \tag{12.46}$$

thus $\beta_j(N) = 1$ for all $j$. With this initial condition, one finds all $\beta_j(t), t = N-1, \ldots, 1$ for all $j$ from the recursion relationship. Again an underflow problem arises which can be avoided by a normalization in each step in the same manner as before.

### 12.2.3 The Estimation Formulas

Then we may also write

$$\begin{aligned}\rho(x_t = i|y_{1\ldots N})\rho(y_{1\ldots N}) &= \rho(y_{1\ldots N}, x_t = i)\\ &= \rho(y_{t+1\ldots N}|y_{1\ldots t}, x_t = i)\rho(y_{1\ldots t}, x_t = i),\\ &= \rho(y_{t+1\ldots N}|x_t = i)\rho(y_{1\ldots t}, x_t = i),\end{aligned} \tag{12.47}$$

where the last equation is true because $y_{t+1\dots N}$ is independent of $y_{1\dots t}$, given $x_t$. Therefore the probability $\rho(x_t = i\,|\,y_{1\dots N})$ can also be written as

$$\rho(x_t = i\,|\,y_{1\dots N}) = \frac{\beta_i(t)\alpha_i(t)}{\rho(y_{1\dots N})}\,. \tag{12.48}$$

This probability should not be confused with

$$\rho(x_t = i\,|\,y_{1\dots t}) = \frac{\alpha_i(t)}{\rho(y_{1\dots t})}\,. \tag{12.49}$$

Furthermore we define

$$\begin{aligned}
&\rho(x_t = i, x_{t-1} = j\,|\,y_{1\dots N})\rho(y_{1\dots N})\\
&= \rho(x_t = i, x_{t-1} = j, y_{1\dots N})\\
&= \rho(y_{t+2\dots N}|x_{t+1} = j)\rho(y_{t+1}|x_{t+1} = j)\rho(x_{t+1} = j\,|x_t = i)\rho(y_{1\dots t}, x_t = i)\,.
\end{aligned} \tag{12.50}$$

Hence

$$\Psi_{ji}(t) = \frac{\beta_j(t+1)\rho(y_{t+1}|x_{t+1} = j)\rho(x_{t+1} = j\,|x_t = i)\alpha_i(t)}{\rho(y_{1\dots N})} \tag{12.51}$$

is the probability $\rho(x_t = i, x_{t+1} = j\,|\,y_{1\dots N})$ that $x_t = i$ and $x_{t+1} = j$ for given $y_{1\dots N}$ (see Fig. 12.1), and from this one obtains by marginalization

$$\Psi_i(t) = \rho(x_t = i\,|\,y_{1\dots N}) = \sum_j \rho(x_t = i, x_{t+1} = j\,|\,y_{1\dots N}) = \sum_j \Psi_{ji}(t), \tag{12.52}$$

again the probability that $x_t = i$ for given $y_{1\dots N}$ which is identical to the quantity given in (12.48) and determined there in a different way.

Taking into account again the dependence on the parameter set and considering this in the $k$th iteration of the EM method, we should write, e.g., $\Psi_{ji}(t)$ as

$$\Psi_{j\,i\,|\boldsymbol{\theta}^{(k)}}(t) \equiv \rho(x_t = i, x_{t+1} = j\,|\,y_{1\dots N}, \boldsymbol{\theta}^{(k)}). \tag{12.53}$$

With these densities then, we have to calculate the expectation value of $\ln \rho(x_{t+1} = j \mid x_t = i, \boldsymbol{\theta})$, and the parametric set $\boldsymbol{\theta}$ may now consist of the parameters $a_{ji} \equiv \rho(x_{t+1} = j \mid x_t = i, \boldsymbol{\theta})$ and $\{\mu_i\}$ and $\{\sigma_i\}$ contained in $\rho(y_t \mid x_t = i) \equiv \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_t-\mu_i)^2}{2\sigma_i^2}}$.

Then we obtain

$$T_1(a_{ji} \mid \boldsymbol{\theta}^{(k)}) = \sum_t \Psi_{ji}(t) \ln a_{ji} \tag{12.54}$$

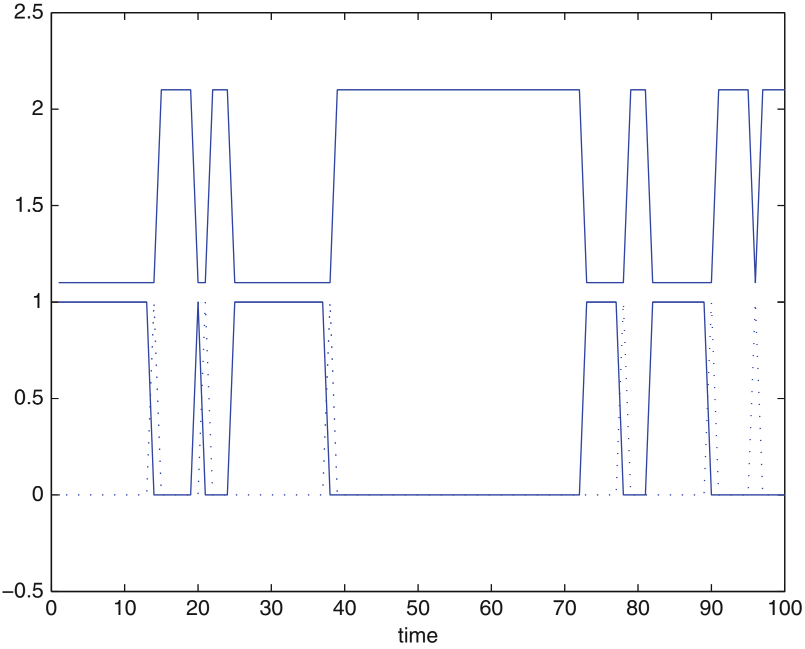


**Fig. 12.1** Realization of an exact hidden process (*above*) and some of the functions $\Psi_{ij}(t)$ (*below*: $\Psi_{21}(t)$ *dotted line*, $\Psi_{11}(t)$ *solid line*)

and

$$T_2(\mu_i, \sigma_i \mid \boldsymbol{\theta}^{(k)}) = \sum_t \Psi_i(t) \left( -\frac{(y_t - \mu_i)^2}{2\sigma_i^2} - \ln \sigma_i + \text{const} \right). \tag{12.55}$$

With this result, the expectation step is finished. The maximization can be done analytically:

Derivation of $T_2$ with respect to the parameters $\mu_i, \sigma_i$ leads to

$$\hat{\mu}_i = \frac{\sum_t \Psi_i(t) y_t}{\sum_t \Psi_i(t)}, \tag{12.56}$$

$$\hat{\sigma}_i^2 = \frac{\sum_t \Psi_i(t)(y_t - \hat{\mu}_i)^2}{\sum_t \Psi_i(t)}. \tag{12.57}$$

This looks plausible: the data at time $t$ contribute to the estimates of state $i$ according to the weight of the probability that state $i$ is realized at time $t$.

To estimate the transition probability the condition $\sum_j a_{ji} = 1$ has to be taken into account with the help of Lagrange multipliers. One defines

$$\tilde{T}_1(a_{ji}) = T_1(a_{ji}) + \sum_i \lambda_i \sum_j (a_{ji} - 1) \; . \tag{12.58}$$

Setting the derivatives with respect to the parameters $a_{ji}$ and the Lagrange multipliers $\lambda_i$ equal to zero, leads to

$$\hat{a}_{ji} = \frac{\sum_t \Psi_{ji}(t)}{\sum_t \Psi_j(t)} \tag{12.59}$$

with an interpretation similar to that for the other estimators.

Thus the new parameters can be calculated analytically and can be used for the next iterative step until convergence is achieved.

## 12.3 Estimating the Parameters in a State Space Model

We write the state space model again as (see Sect. 11.5.2)

$$\boldsymbol{X}(t) = \mathsf{A}\boldsymbol{X}(t-1) + \boldsymbol{\eta}(t), \qquad \boldsymbol{\eta}(t) \sim \mathrm{WN}(\mathbf{0}, \mathsf{Q}), \tag{12.60}$$

$$\boldsymbol{Y}(t) = \mathsf{C}\boldsymbol{X}(t) + \boldsymbol{\epsilon}(t), \qquad \boldsymbol{\epsilon}(t) \sim \mathrm{WN}(\mathbf{0}, \mathsf{R}) \; . \tag{12.61}$$

With a similarity transformation (see Sect. 5.9) one can always obtain $\mathsf{Q} = 1$. For $t = 1$, we will set $\boldsymbol{X}(1) \sim N(\boldsymbol{x}^{(1)}, \mathsf{P}^{(1)})$.

The set of parameters then contains the elements of the matrices $\mathsf{R}$, $\mathsf{C}$, $\mathsf{A}$, $\boldsymbol{x}^1$, and $\mathsf{P}^1$. We collect all of these parameters into a vector $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = (\mathrm{A}_{\alpha\beta}, \mathrm{C}_{i\alpha}, \mathrm{R}_{ij}, \mathrm{P}^{(1)}_{\alpha\beta}, x^{(1)}_{\alpha}) \; . \tag{12.62}$$

Now

$$\rho(\boldsymbol{x}_1 \mid \boldsymbol{\theta}) = (2\pi)^{-\frac{m}{2}} (\det \mathsf{P}^{(1)})^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\boldsymbol{x}(1)-\boldsymbol{x}^{(1)})^{\mathrm{T}} (\mathsf{P}^{(1)})^{-1} (\boldsymbol{x}(1)-\boldsymbol{x}^{(1)})}$$

$$\rho(\boldsymbol{x}_{2\ldots N} \mid \boldsymbol{x}_1, \boldsymbol{\theta}) \propto \prod_{t=2}^{N} e^{-\frac{1}{2}\left(\boldsymbol{x}(t)-\mathsf{A}\boldsymbol{x}(t-1)\right)^{\mathrm{T}}\left(\boldsymbol{x}(t)-\mathsf{A}\boldsymbol{x}(t-1)\right)}$$

$$\rho(\boldsymbol{y}_{1\ldots N} \mid \boldsymbol{x}_{1\ldots N}, \boldsymbol{\theta}) \propto \prod_{t=1}^{N} (\det \mathsf{R})^{-\frac{1}{2}} \, e^{-\left(\boldsymbol{y}(t)-\mathsf{C}\boldsymbol{x}(t)\right)^{\mathrm{T}} \mathsf{R}^{-1} \left(\boldsymbol{y}(t)-\mathsf{C}\boldsymbol{x}(t)\right)} \tag{12.63}$$

Therefore building $\ln \rho(\boldsymbol{y}_{1\ldots N}, \boldsymbol{x}_{1\ldots N} \mid \boldsymbol{\theta})$, we obtain, up to unimportant constants,

$$\ln \rho(\boldsymbol{x}(t) \mid \boldsymbol{x}(t-1), \boldsymbol{\theta}) = -\frac{1}{2} \sum_{t=2}^{N-1} \left(\boldsymbol{x}(t) - \mathsf{A}\boldsymbol{x}(t-1)\right)^{\mathrm{T}} \left(\boldsymbol{x}(t) - \mathsf{A}\boldsymbol{x}(t-1)\right), \tag{12.64}$$

$$\ln \rho(\boldsymbol{y}(t) \mid \boldsymbol{x}(t), \boldsymbol{\theta}) = -\frac{1}{2} N \ln(\det \mathsf{R}) \\ -\frac{1}{2} \sum_{t=1}^{N} \big(\boldsymbol{y}(t) - \mathsf{C}\boldsymbol{x}(t)\big)^{\mathrm{T}} \mathsf{R}^{-1} \big(\boldsymbol{y}(t) - \mathsf{C}\boldsymbol{x}(t)\big), \tag{12.65}$$

and

$$\ln \rho(\boldsymbol{x}(1) \mid \boldsymbol{\theta}) = -\frac{1}{2} \ln(\det \mathsf{P}^{(1)}) \\ -\frac{1}{2} \big(\boldsymbol{x}(1) - \boldsymbol{x}^{(1)}\big)^{\mathrm{T}} (\mathsf{P}^{(1)})^{-1} \big(\boldsymbol{x}(1) - \boldsymbol{x}^{(1)}\big) . \tag{12.66}$$

We have to calculate the expectation values of these terms according to

$$< O > \big|_{y_{1\ldots N}, \boldsymbol{\theta}^{(k)}} = \sum_{\boldsymbol{x}_{1\ldots N}} O \, \rho(\boldsymbol{x}_{1\ldots N} \mid \boldsymbol{y}_{1\ldots N}, \boldsymbol{\theta}^{(k)}), \tag{12.67}$$

so that the following moments arise

$$\mathsf{Q}^{(1)} = \sum_{t=1}^{N} \boldsymbol{y}^{\mathrm{T}}(t) \, \boldsymbol{y}(t) \big|_{y_{1\ldots N}, \boldsymbol{\theta}^{(k)}} , \tag{12.68}$$

which could be regarded as the zero-th moment, and

$$\mathsf{Q}^{(2)} = \sum_{t=1}^{N} < \boldsymbol{X}(t) > \big|_{y_{1\ldots N}, \boldsymbol{\theta}^{(k)}} \boldsymbol{Y}^{\mathrm{T}}(t) \tag{12.69}$$

$$\mathsf{Q}^{(3)} = \sum_{t=2}^{N} < \boldsymbol{X}(t) \, \boldsymbol{X}^{\mathrm{T}}(t) > \big|_{y_{1\ldots N}, \boldsymbol{\theta}^{(k)}}, \tag{12.70}$$

$$\mathsf{Q}^{(4)} = \sum_{t=2}^{N} < \boldsymbol{X}(t-1) \, \boldsymbol{X}^{\mathrm{T}}(t) > \big|_{y_{1\ldots N}, \boldsymbol{\theta}^{(k)}}, \tag{12.71}$$

$$\mathsf{Q}^{(5)} = \sum_{t=2}^{N} < \boldsymbol{X}(t-1) \, \boldsymbol{X}^{\mathrm{T}}(t-1) > \big|_{y_{1\ldots N}, \boldsymbol{\theta}^{(k)}} . \tag{12.72}$$

With the help of the formula for moments, derived in Sect. 12.1, we obtain

$$\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = -\frac{1}{2} N \; \ln \det \mathsf{R} - \frac{1}{2} \ln \det \mathsf{P}^{(1)}$$

$$\begin{aligned}
&-\frac{1}{2}\,\mathrm{Tr}\big(\mathsf{R}^{-1}(\mathsf{Q}^{(1)} - \mathsf{C}\,\mathsf{Q}^{(2)} - (\mathsf{Q}^{(2)})^{\mathrm{T}}\mathsf{C}^{\mathrm{T}} + \mathsf{C}\,\mathsf{Q}^{(3)}\mathsf{C}^{\mathrm{T}}\big)\\
&-\frac{1}{2}\,\mathrm{Tr}\big(\mathsf{Q}^{(3)} - \mathsf{A}\,(\mathsf{Q}^{(4)}) - (\mathsf{Q}^{(4)})^{\mathrm{T}}\mathsf{A}^{\mathrm{T}} + \mathsf{A}\,\mathsf{Q}^{(5)}\mathsf{A}^{\mathrm{T}}\big)\\
&-\frac{1}{2}\,\mathrm{Tr}\big((\mathsf{P}^{(1)})^{-1}(\boldsymbol{x}(0) - \boldsymbol{x}^{(1)})(\boldsymbol{x}(0) - \boldsymbol{x}^{(1)})^{T}\big)\,.
\end{aligned}$$

The maximum of $\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ with respect to the parameters $\mathbf{A}, \mathbf{C}, \mathbf{R}$ can be determined analytically. One obtains

$$\widehat{\mathsf{A}} = (\mathsf{Q}^{(4)})^{\mathrm{T}}(\mathsf{Q}^{(5)})^{-1} \tag{12.73}$$

and

$$\widehat{\mathsf{C}} = (\mathsf{Q}^{(2)})^{\mathrm{T}}(\mathsf{Q}^{(3)})^{-1}. \tag{12.74}$$

To determine $\widehat{\mathsf{R}}$, we use the formulas

$$\frac{\partial \ln \det \mathsf{R}}{\partial R_{rs}} = \frac{\partial \mathrm{Tr} \ln \mathsf{R}}{\partial R_{rs}} = \mathrm{Tr}\big(\mathsf{R}^{-1}\frac{\partial \mathsf{R}}{\partial R_{rs}}\big) \tag{12.75}$$

and

$$\frac{\partial \mathsf{R}^{-1}}{\partial R_{rs}} = -\mathsf{R}^{-1}\frac{\partial \mathsf{R}}{\partial R_{rs}}\mathsf{R}^{-1}\,. \tag{12.76}$$

Then by setting the derivative with respect to $R_{rs}$ equal to zero, one arrives at the equation

$$N\mathsf{R}^{-1} - \mathsf{R}^{-1}\Big(\mathsf{Q}^{(1)} - \mathsf{C}\,\mathsf{Q}^{(2)} - (\mathsf{Q}^{(2)})^{\mathrm{T}}\mathsf{C}^{\mathrm{T}} + \mathsf{C}\,\mathsf{Q}^{(3)}\mathsf{C}^{\mathrm{T}}\Big)\mathsf{R}^{-1} = 0 \tag{12.77}$$

and thus, also already using the update $\widehat{\mathsf{C}}$ for $\mathsf{C}$,

$$\begin{aligned}
\widehat{\mathsf{R}} &= \frac{1}{N}\Big(\mathsf{Q}^{(1)} - \widehat{\mathsf{C}}\mathsf{Q}^{(2)} - (\mathsf{Q}^{(2)})^{\mathrm{T}}\widehat{\mathsf{C}}^{\mathrm{T}} + \widehat{\mathsf{C}}\mathsf{Q}^{(3)}\widehat{\mathsf{C}}^{\mathrm{T}}\Big) && (12.78)\\
&= \frac{1}{N}\sum_{t=1}^{N}\big(\boldsymbol{y}(t) - \widehat{\mathsf{C}}\,\check{\boldsymbol{x}}(t)\big)^{\mathrm{T}}\big(\boldsymbol{y}(t) - \widehat{\mathsf{C}}\,\check{\boldsymbol{x}}(t)\big) + \widehat{\mathsf{C}}\check{\mathsf{P}}(t)\widehat{\mathsf{C}}^{\mathrm{T}}. && (12.79)
\end{aligned}$$

In a similar way, one obtains

$$\widehat{\mathsf{P}}^{(1)} = \check{\mathsf{P}}(1), \tag{12.80}$$

and

$$\widehat{\boldsymbol{x}}^{(1)} = \check{\boldsymbol{X}}(1)\,. \tag{12.81}$$

These new matrices $\widehat{\mathsf{A}}, \widehat{\mathsf{C}}, \widehat{\mathsf{R}}, \widehat{\mathsf{P}}^{(1)}, \widehat{\boldsymbol{x}}^{(1)}$ constitute the new parameter set $\boldsymbol{\theta}^{(k+1)}$, and one iterative step is finished in this way.

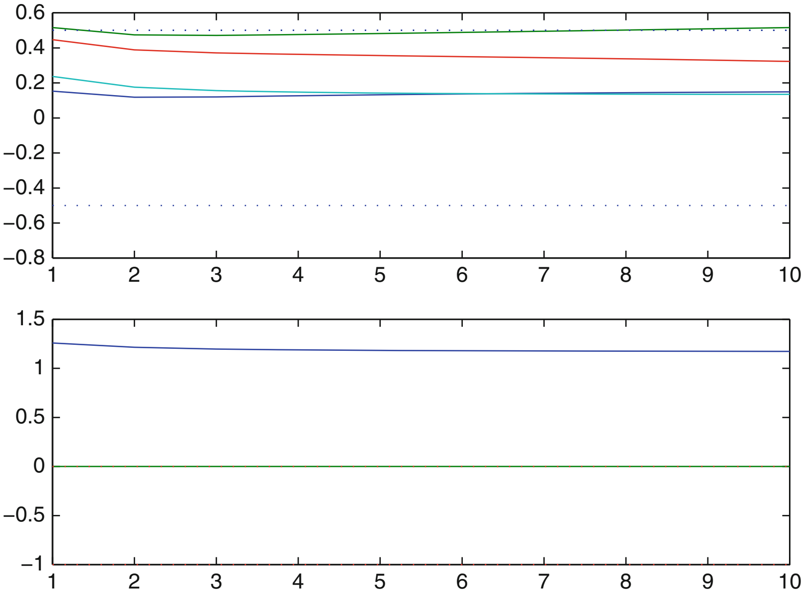


**Fig. 12.2** Estimates of the parameters obtained in the iteration of the EM method. *Above*: The parameters of the matrix $\mathsf{A}$; *below*: the same for the parameters of matrix $\mathsf{C}$

To judge the estimates of the parameters, one has to take into account that, even after the variance $\mathsf{Q}$ of the noise has been made equal to the identity matrix by a similarity transformation, the system equation is still invariant under a transformation by an orthogonal matrix. This leaves $\mathsf{Q}$ invariant but changes matrix $\mathsf{C}$. Considering the singular value decomposition (see Sect. 8.4)

$$\mathsf{C} = \mathsf{U}\mathsf{D}\mathsf{V}^T, \tag{12.82}$$

where $\mathsf{U}$ and $\mathsf{V}$ are orthogonal matrices and $\mathsf{D}$ is a diagonal matrix, one may introduce as standard form for $\mathsf{C}$ that for which the matrix $\mathsf{V}$ is the identity, namely,

$$\mathsf{C} = \mathsf{U}\mathsf{D}. \tag{12.83}$$

*Example.* We simulate a state space model with ($\mathrm{A}_{11} = \mathrm{A}_{21} = \mathrm{A}_{22} = -\mathrm{A}_{12} = 0.5$, $\mathrm{C}_{11} = -1, \mathrm{C}_{12} = 0, \mathrm{R} = 1$). In Fig. 12.2 is shown how the estimations of the parameters converge to values that are near the exact ones.

The estimations tend quickly toward the exact values even if the initial values are far away. The convergence rate, however, slows down near the convergence point.

*Remark.* In the literature the methods for estimating parameters in pure AR model or ARMA models is discussed Brockwell and Davies (1987). Because of the ubiquitous existence of the observational noise, these methods play a minor role in applications. We will therefore not go into details. They can, however, be derived as special cases of the methods explained here. One has only to take into account that the observational equation then reads

$$\rho(\boldsymbol{y}_{1\ldots N} \mid \boldsymbol{x}_{1\ldots N}) = \delta(\boldsymbol{y}_1 - \boldsymbol{x}_1)\ldots\delta(\boldsymbol{y}_N - \boldsymbol{x}_N). \tag{12.84}$$

Then one obtains, e.g., for the estimate of parameter $A$ in the univariate AR(1) model

$$X(t) = AX(t-1) + \sigma\eta(t), \qquad \eta \sim N(0,1), \tag{12.85}$$

according to (12.73),

$$\hat{A} = \frac{\sum_t y(t-1)y(t)}{\sum_t y(t-1)y(t-1)}. \tag{12.86}$$

# Chapter 13
# Statistical Tests and Classification Methods

## 13.1 General Comments Concerning Statistical Tests

### *13.1.1 Test Quantity and Significance Level*

Since it is often necessary to make hypotheses about the properties of random variables, it is important to have a well-defined procedure to test these hypotheses. Typical hypotheses are

$$H_0 : \quad \{y_1, \ldots, y_N\} \quad \text{is a sample of a random variable } Y \\ \text{with density } \rho_Y(y),$$

or

$$H_0 : \quad \{y_1, \ldots, y_N\} \quad \text{is a sample of some random variable } Y \\ \text{with expectation value } \mu.$$

Thus, given realizations of a random variable together with a hypothesis $H_0$ about the random variable, one should know whether the hypothesis is compatible with the realizations that have been obtained.

For example, suppose that we generate $N$ random numbers using a random number generator that is supposed to yield uniformly distributed independent random numbers. From these realizations, we can estimate the mean value and the variance, and we can ask whether the estimated values are consistent with the assumption that the generator delivers uniformly distributed random variables by comparing mean and variance. We can also exhibit the frequency distribution in the interval $[0, 1]$ and ask whether this distribution is acceptable as a uniform distribution. For this purpose, we must define a measure for the compatibility of the observed quantity with the hypothesis, and one must define what one will regard as compatible and what not.

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1_13, © Springer-Verlag Berlin Heidelberg 2012

Such a measure for a test is called a test quantity or test statistic. This is a random variable $C(Y_1, \ldots, Y_N)$ which under the null hypothesis $H_0$ possesses a specific distribution $\rho(c \mid H_0)$. The value of $C$, obtained from the sample $y_1, \ldots, y_N$ to be tested, should be a very probable value if the hypothesis is valid and a very improbable one, if the hypothesis is false.

To be more precise, one may introduce an $(1-\alpha)$-quantile $c_{1-\alpha}$ for the distribution of $\rho(c \mid H_0)$, so that $c > c_{1-\alpha}$ with $100\alpha\%$ probability. If we choose, e.g., $\alpha = 0.05$, then with only 5% probability a realization of $C$ will be larger than $c_{0.95}$, if the hypothesis $H_0$ is valid. We will call $\alpha$ the significance level. Now, if the realization of $C$ is smaller than $c_{1-\alpha}$, we accept the hypothesis; otherwise, we decide to reject the hypothesis.

Because the rejection interval is on one side of the acceptance interval, the test introduced so far is called a one-sided test. In many cases, a two-sided test makes more sense: the introduction of an interval $\{c \mid c_{\alpha/2} < c < c_{1-\alpha/2}\}$ as an acceptance interval. The $H_0$ hypothesis is then rejected if the realization of $C$ for a sample $y_1, \ldots, y_N$ is either smaller than $c_{\alpha/2}$ or larger than $c_{1-\alpha/2}$. The probability for this is also $100\alpha\%$ (see Fig. 13.1).

Thus, testing a hypothesis means finding a test quantity together with its probability distribution when the hypothesis is valid, choosing a significance level, and formulating with it the rules for rejecting and accepting the hypothesis.

Now, if the null hypothesis is valid, the probability that a realization falls into the rejection interval is $\alpha$. Thus, even if the null hypothesis is valid it can also be rejected, and an error is made with probability $100\alpha\%$. This error is called the error of the first kind.

If the value of $C$ falls within the acceptance interval, then the null hypothesis is not rejected. That may of course also be an error, since the same value of $C$ may also be probable if an alternative null hypothesis, say, $H_1$ is true. If the density function $\rho(c \mid H_1)$ is known, this error can be determined. The probability that under the hypothesis $H_1$ the realization of $C$ lies in the interval $(c_{\alpha/2}, c_{1-\alpha/2})$ is

$$\beta = \int_{c_{\alpha/2}}^{c_{1-\alpha/2}} dc\, \rho(c \mid H_1), \tag{13.1}$$

as can easily be seen from Fig. 13.1, below. $\beta$ is called the error of the second kind. A decrease in $\alpha$ shifts the right limit of the acceptance interval further to the right, and thus increases $\beta$. The "goodness" of a test quantity $C$ is thus evidenced by the extent of the overlap of the density functions of $C$ under $H_0$ and $H_1$. On the other hand, an increase in the size of the random sample narrows the distributions and thus improves the disentanglement of the tests.

In many cases, however, one does not know the density $\rho(c \mid H_1)$ for alternative hypotheses, and thus an estimation of the error of second kind is not possible.

There is an alternative formulation of the test. For given data one may, determine the realization of $C$ and also the corresponding confidence interval. Then one asks whether the hypothesis $H_0$ is compatible with this confidence interval. If, e.g., $H_0$ means that the data are realizations of a random variable with mean $\mu_0$, then

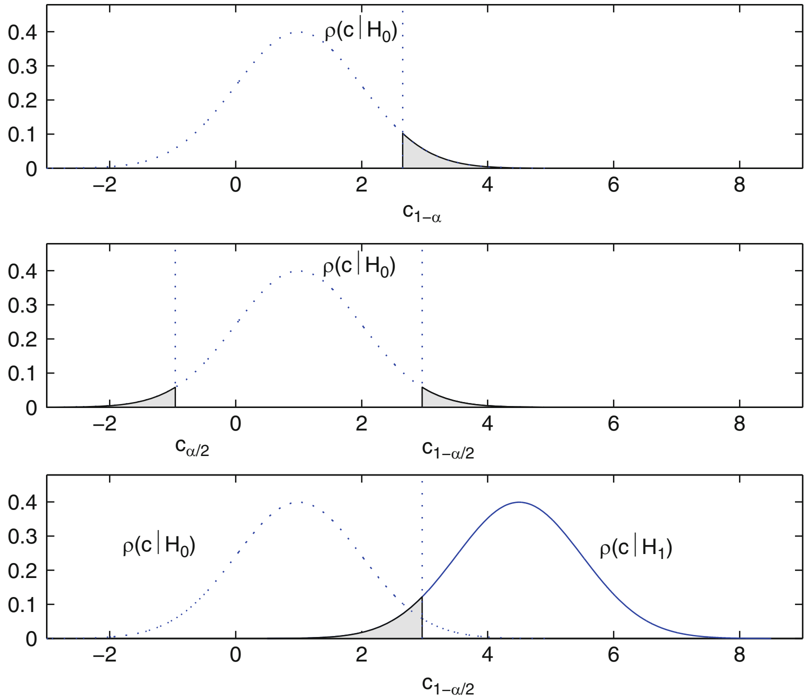


**Fig. 13.1** Acceptance and rejection intervals in the one-sided test (*above*) and in the two-sided test (*middle*). *Below* the density of the test quantity for an alternative hypothesis is also shown, and the shaded area corresponds to the error of the second kind concerning the alternative hypothesis $H_1$

one accepts the hypothesis if $\mu_0$ lies within the confidence interval; otherwise, one rejects this hypothesis (see Fig. 13.2).

There is a further alternative formulation: In a one-sided test, one can take the realized value of a test quantity as the critical value $c_{1-p}$ and calculate the corresponding value of $p$. The smaller the $p$-value, the less probable is it that the null hypothesis is justified. For $p < 0.05$, one usually rejects the hypothesis.

*Examples.*

1. Let $H_0$ be the hypothesis that given random numbers are independent realizations of a random variable which is normally distributed with mean value $\langle X \rangle$ equal to a given value $\mu_0$. Thus, we test the null hypothesis

$$H_0 : \qquad \langle X \rangle = \mu_0. \tag{13.2}$$

The test quantity can be chosen as the estimator of the mean

$$C = \hat{X}_N = \frac{1}{N}(X_1 + \ldots + X_N), \tag{13.3}$$

and its distribution $\rho(c \mid H_0)$ can easily be given. Here a two-sided test is in place.

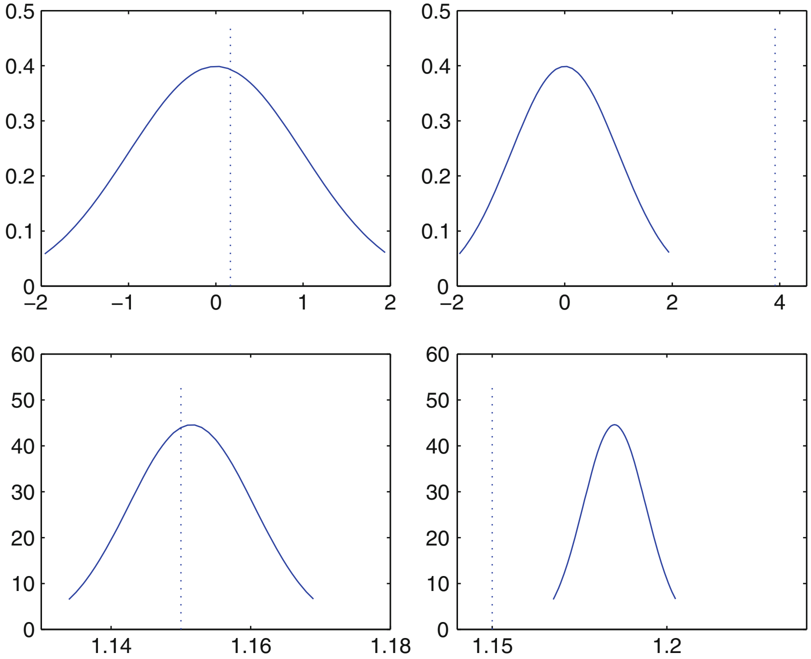


**Fig. 13.2** *Above*: Density of the test quantity, shown from $c_{\alpha/2}$ to $c_{1-\alpha/2}$, for the test of the hypothesis that the mean is $\mu_0$. The *vertical line* denotes the realization of $c$ for the given sample. *Left*: A case where the hypothesis is valid. *Right*: A case where the hypothesis has to be rejected. *Below*: The confidence interval around the realization $c$ of $C$, denoted by a Gaussian density which has a maximum at $c$ and which is drawn only from one end of the confidence interval to the other. The *vertical line* denotes the value $\mu_0$. *Left*: A case where the hypothesis is valid. *Right*: A case where the hypothesis has to be rejected

If $\sigma_X$ is known, then one can construct the quantity

$$C = \frac{\hat{X}_N - \mu_0}{\sigma_X}\sqrt{N}, \tag{13.4}$$

which has a normal standard distribution for sufficiently large $N$ under the null hypothesis. The test using this test quantity $C$ is called the $z$-test.

However, we also wish to elaborate the more difficult case in which $\sigma_X$ is not known. Then we construct

$$C = \frac{\hat{X}_N - \mu_0}{S_N}\sqrt{N}, \tag{13.5}$$

where $S_N$ is the estimator for the variance. Now $C$ is a $t$-distributed random variable with $N - 1$ degrees of freedom. The hypothesis $H_0$ must be rejected with a significance level $\alpha$ if the realization of $C$ lies outside of the interval $(-t_{N-1,\alpha/2}, t_{N-1,1-\alpha/2})$, that is, if

$$\mid t \mid > t_{N-1,1-\alpha/2}. \tag{13.6}$$

Alternatively we could determine the confidence interval of $\hat{X}_N$. If the hypothesis is valid, namely, that the given random numbers are independent realizations of a random variable which is normally distributed, then the confidence interval is given by $[\hat{X}_N - S_N/\sqrt{N}, \hat{X}_N + S_N/\sqrt{N}]$ and one may ask whether $\mu_0$ lies in this interval.

Had we wished to test the hypothesis

$$\langle X \rangle < \mu_0, \tag{13.7}$$

then it would have to be rejected with significance level $\alpha$ if

$$t \geq t_{N-1,1-\alpha}, \tag{13.8}$$

since with probability $\alpha$, the mean $\langle X \rangle$ is significantly greater than $\mu_0$.

### 13.1.2 Empirical Moments for a Test Quantity: The Bootstrap Method

In many cases the distribution of the test quantity $C$ cannot be given analytically. Thus also the mean and variance of $C$ are not available, and a confidence interval cannot be formulated.

A method known as bootstrap can, however, provide an approximate estimate of the confidence interval which in most applications is sufficient for the decision.

Given the data $\boldsymbol{y} = y_1, \ldots, y_N$, one may define a bootstrap sample $\boldsymbol{y}^\star$ consisting of $N$ data values drawn with replacement from $\boldsymbol{y} = y_1, \ldots, y_N$. Thus $\boldsymbol{y}^\star$ contains some of the $\{y_i\}$ twice or multiple times, some not. One may now construct, in this way, $M$ of these bootstrap samples and regard these samples $\boldsymbol{y}^{\star 1}, \ldots \boldsymbol{y}^{\star M}$ as surrogates of $M$ samples drawn from the density underlying the data sample $\boldsymbol{y} = \{y_1, \ldots, y_N\}$.

Given these bootstrap samples, one may estimate the mean and standard deviation of the test quantity $C(y_1, \ldots, y_N)$ by

$$\mathrm{mean}_B(C) = \frac{1}{M} \sum_{i=1}^{M} C(\boldsymbol{y}^{\star i}) \tag{13.9}$$

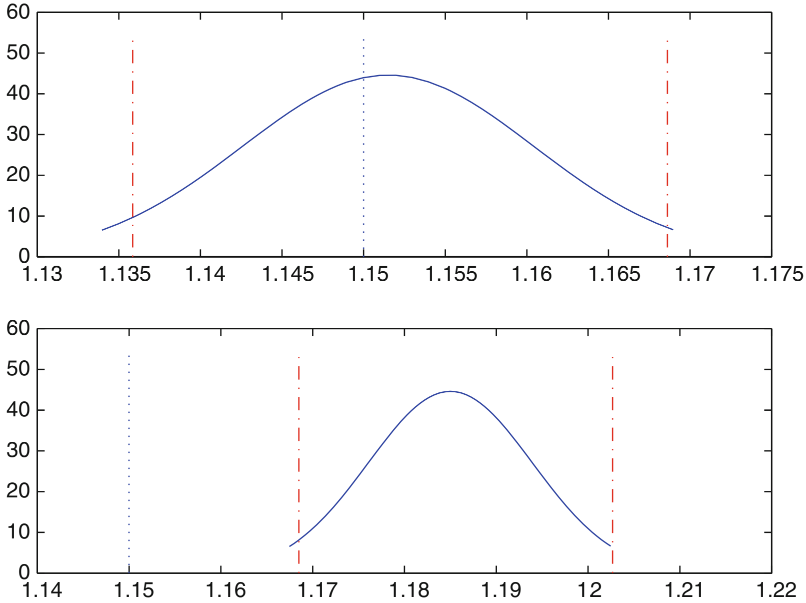


**Fig. 13.3** Test of the bootstrap method for a test statistic, for which the confidence intervals can be determined analytically (denoted by the ends of the distribution) and for which they are determined by the bootstrap method (*dashed dotted lines*). The two intervals do not differ very much. The *dotted vertical line* denotes the value of the mean given in the hypothesis. *Above*: A case where the hypothesis has to be accepted. *Below*: A case where the hypothesis has to be rejected

and

$$\mathrm{std}_B(C) = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(C(\boldsymbol{y}^{\star i}) - \mathrm{mean}_B(C))^2}\,. \tag{13.10}$$

In this way one also gets a confidence interval of $c$. One may test these bootstrap estimates in cases where the estimates for mean and standard variation for the test statistic are analytically calculable as, e.g., for the test statistic $C$ ='mean()'. In Fig. 13.3 this is shown for a typical data set $\boldsymbol{y} = y_1, \ldots, y_N$ with $N = 20$ and with $M = 100$.

*Remarks.*

- Given an estimator $\hat{\Theta}_N(X_1, \ldots, X_N)$ for a quantity $\theta$, one can also use the bootstrap method to find an estimate for the standard deviation of this estimator.
- Constructing the bootstrap samples means that one approximates the density of the random variable, say $X$, which is realized from the sample $\boldsymbol{y} = y_1, \ldots, y_N$, by the empirical density

$$\rho_X(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - y_i). \tag{13.11}$$

It turns out that this approximation is good enough to estimate the empirical standard deviation $\mathrm{std}(C)$, which is the estimate of the standard deviation of $C$ on the basis of the sample. It is shown by Efron (Efron and Tibshirani 1993), that under very mild conditions, $\mathrm{std}_B(C)$ given in (13.10) tends to this empirical standard deviation for $M \to \infty$. In praxis it suffices to choose $M$ between 60 and 200.

- A forerunner of the bootstrap method is the jackknife method for estimating the standard deviation and bias of test quantities or for estimates by constructing further samples from a given sample. Given $\boldsymbol{y} = y_1, \ldots, y_N$, the jackknife samples are defined by

$$\boldsymbol{y}_{(i)} = \{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N\}. \tag{13.12}$$

The $i$th jackknife estimator is then

$$\hat{\Theta}_{(i)} = \hat{\Theta}(\boldsymbol{y}_{(i)}), \tag{13.13}$$

and the jackknife estimate

$$\hat{\Theta}_{(.)} = \frac{1}{N}\sum_{i=1}^{N} \hat{\Theta}_{(i)}. \tag{13.14}$$

The jackknife can be viewed as an approximation of the bootstrap (Efron and Tibshirani 1993).

*Example.* We test the hypothesis that the data are realizations of a random variable with a symmetrical density function. As a test quantity, we use

$$C = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{Y}_N)^3, \tag{13.15}$$

with $\hat{Y}_N$ the mean of the data. We do not know the distribution of $C$ under the hypothesis but can easily determine the confidence interval by the bootstrap method.

### 13.1.3 The Power of a Test

The error of the second kind of a test depends crucially on the shape of the densities $\rho(c \mid H_0)$ and $\rho(c \mid H_1)$ because the overlap of these densities determines the probability $\beta$ for an acceptance of $H_0$ when $H_1$ is valid. The power $\beta'$ of a test is simply related to the error of the second kind, it is the probability for a rejection of $H_0$ when $H_1$ is valid.

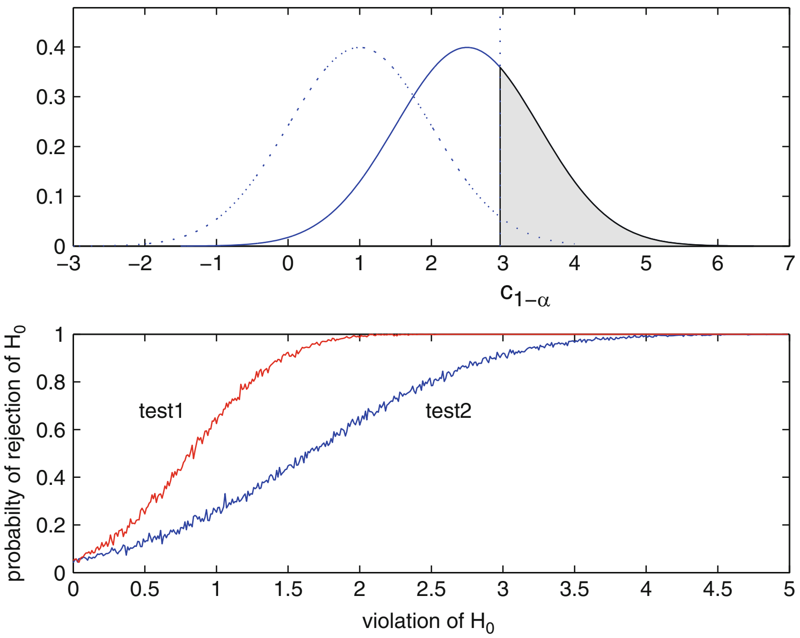


**Fig. 13.4** *Above*: Density distribution of the test quantity given the hypothesis $H_0$ and $H_1$, respectively, the shaded area is the power $\beta'$, the probability of a rejection at testing $H_0$, if $H_1$ is true. *Below*: A test is more efficient, the faster the power, i.e., the number of rejections grows with increasing violation of the hypothesis $H_0$

This probability is equal to the probability that the realization of $C$ is larger than the given threshold $c_{1-\alpha}$ for the hypothesis $H_0$ (Fig. 13.4, above), thus

$$\beta' \equiv \int_{c_{1-\alpha}}^{\infty} \mathrm{d}c\rho(c \mid H_1) \equiv 1 - \beta. \tag{13.16}$$

Of course, one may improve the power of the test easily by lowering the variance of the test quantity, e.g., by increasing the number $N$ of observations. Given $H_1$, one may thus also ask for the number $N$ of observations for which a certain power is achieved. Usual goals are a power of 0.8 or 0.9, respectively, 80% or 90%, i.e., 20% or 10% probability for an error of the second kind.

In many cases, the $H_0$ hypothesis can be violated continuously as, e.g., for the hypothesis that the mean has some value $\mu$. One may then discuss all hypotheses $H_1(\mu')$ that the mean is $\mu' \neq \mu$ and may study the densities under these hypotheses and the power in the dependence on $\mu'$. For $H_1 = H_0$, respectively, $\mu' = \mu$, the power is just $\alpha$. For a test with always higher power, the chance of rejecting $H_0$ rises more quickly for $H_1(\mu')$ with increasing $\mu'$ (Fig. 13.4, below) than for a test with generally lower power.

## 13.2 Some Useful Tests

### 13.2.1 The z- and the t-Test

It is often of interest to establish whether the mean values of two sets of measurements can be viewed as statistically equivalent. Thus, if we consider the sets of measurements (i.e., the random samples) $x_1, \ldots, x_{N_1}$ and $y_1, \ldots, y_{N_2}$ of size $N_1$ and $N_2$, respectively, then the null hypothesis to be tested is that these are samples of the random variables $X$ and $Y$, respectively, with

$$\langle X \rangle = \langle Y \rangle. \tag{13.17}$$

First we assume that the variances $\sigma_X^2$ and $\sigma_Y^2$ are equal ($\sigma_X^2 = \sigma_Y^2 = \sigma^2$). Let $\hat{X}_{N_1}$ be the estimate of $\langle X \rangle$ and $\hat{Y}_{N_2}$ the estimate of $\langle Y \rangle$. Then, under the null hypothesis,

$$\Delta = \hat{X}_{N_1} - \hat{Y}_{N_2} \tag{13.18}$$

is a normally distributed random variable with mean value zero and variance

$$\mathrm{Var}(\Delta) = \sigma^2 \left( \frac{1}{N_1} + \frac{1}{N_2} \right). \tag{13.19}$$

Such a test, where $\sigma^2$ is assumed to be known, is called a z-test.

However, in most cases, $\sigma^2$ is not known. Instead, one could use $s_X^2$ or $s_Y^2$ with, for example,

$$S_X^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (X_i - \bar{X}_{N_1})^2. \tag{13.20}$$

It is more useful to use a weighted average of $s_X^2$ and $s_Y^2$, namely

$$\begin{aligned} S_\Delta^2 &= \frac{1}{N_1 + N_2 - 2} [(N_1 - 1) S_X^2 + (N_2 - 1) S_Y^2] \\ &= \frac{1}{N_1 + N_2 - 2} \left( \sum_{i=1}^{N_1} (X_i - \bar{X}_{N_1})^2 + \sum_{i=1}^{N_2} (Y_i - \bar{Y}_{N_2})^2 \right). \end{aligned} \tag{13.21}$$

The quantity

$$t = \frac{\Delta / \sqrt{\mathrm{Var}(\Delta)}}{\sqrt{S_\Delta^2 / \sigma^2}} = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \frac{\Delta}{S_\Delta} \tag{13.22}$$

is then a $t$-distributed random variable with $f = N_1 + N_2 - 2$ degrees of freedom.

If the variances $\sigma_X^2$ and $\sigma_Y^2$ are not equal, then they must be estimated as $s_X^2$ and $s_Y^2$ according to (13.20). One constructs the test quantity $t$ as

$$t = \frac{\Delta}{\left(\frac{S_X^2}{N_1} + \frac{S_Y^2}{N_2}\right)^{1/2}}. \tag{13.23}$$

It can be shown that $t$ is approximately $t$-distributed with the number of degrees of freedom $f$ obtained from

$$f = \frac{\left(\frac{s_X^2}{N_1} + \frac{s_Y^2}{N_2}\right)^2}{\frac{1}{N_1 - 1}\left(\frac{s_X^2}{N_1}\right)^2 + \frac{1}{N_2 - 1}\left(\frac{s_Y^2}{N_2}\right)^2}. \tag{13.24}$$

Here $f$ is not necessarily an integer; the $t$ distribution is defined for all real numbers $f \geq 0$ (Hartung et al. 1986).

If the realization of the quantity $t$ in both cases falls outside the interval

$$(t_{f;\alpha/2}, t_{f;1-\alpha/2}) = (-t_{f;1-\alpha/2}, t_{f;1-\alpha/2}), \tag{13.25}$$

that is, if, for example, in the case of equal variance, $\Delta$ is outside the interval

$$\left(-t_{f;1-\alpha/2}\sqrt{\frac{N_1 + N_2}{N_1 N_2}} s_\Delta, \quad t_{f;1-\alpha/2}\sqrt{\frac{N_1 + N_2}{N_1 N_2}} s_\Delta\right), \tag{13.26}$$

then the hypothesis of equal mean values must be rejected with significance level $\alpha$.

### 13.2.2 Test for the Equality of the Variances of Two Sets of Measurements, the $F$-Test

If a given quantity is measured by two different methods, then the mean values of these measurements must be equal (unless one or both of the methods are plagued by systematic errors). However, the variances may be different, since they are a measure of the quality of the method. Suppose that $N_1$ and $N_2$ are the sizes of the random samples and $s_1^2$ and $s_2^2$ the corresponding empirical variances, as calculated from the estimator function (13.20). Then,

$$Y_i = \frac{S_i^2(N_i - 1)}{\sigma_i^2}, \qquad i = 1, 2, \tag{13.27}$$

is $\chi^2$ distributed with $N_i - 1$ degrees of freedom. Under the null hypothesis that $\sigma_1 = \sigma_2$, it follows that (see (2.160))

$$F \equiv \frac{Y_1/(N_1-1)}{Y_2/(N_2-1)} = \frac{S_1^2}{S_2^2} \tag{13.28}$$

is an $F$-distributed random variable with $N_1 - 1$ and $N_2 - 1$ degrees of freedom.

From (2.161) one obtains for the density function

$$p_F(m,n;z) = \left(\frac{m}{n}\right)^{\frac{1}{2}m} \frac{\Gamma(\frac{1}{2}(m+n))}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} z^{\frac{1}{2}m-1} \left(1+\frac{m}{n}z\right)^{-\frac{1}{2}(m+n)} \tag{13.29}$$

with $m = N_1 - 1, n = N_2 - 1$.

Once again an $F$ value $F_{m,n;\gamma}$ for given $\gamma$ is defined through the relationship

$$\gamma = \int_0^{F_{m,n;\gamma}} p_F(m,n;z)\mathrm{d}z. \tag{13.30}$$

By interchanging $S_1^2$ and $S_2^2$, it is easy to see that the relationship

$$F_{m,n;\gamma} = \frac{1}{F_{n,m;1-\gamma}} \tag{13.31}$$

holds. If the null hypothesis that $\sigma_1^2 = \sigma_2^2$ is to be rejected with a significance level $\alpha$, then the value of $F$ obtained from the two sets of measurements must lie outside the interval

$$F_{m,n;\alpha/2} \leq F \leq F_{m,n;1-\alpha/2}. \tag{13.32}$$

### *13.2.3 The $\chi^2$-Test*

This test checks the null hypothesis that a set of measurements can be taken as a sample of a random variable with a given distribution. For this purpose, one divides the possible realizations (if they are not already discrete) into $k$ classes. For example, if the random numbers lie in the interval [0, 1], then one divides this interval into $k$ subintervals. The frequency of a result from class $s$ is denoted $Y_s$. The theoretical probability that the result of a measurement belongs to the class $s$ is determined by the given density function $p_s$. The theoretical frequency of a result from class $s$ is then $n_s = Np_s$, where $N$ is the size of the random sample. Clearly,

$$\sum_{s=1}^{k} p_s = 1, \quad \sum_{s=1}^{k} Y_s = N. \tag{13.33}$$

$Y_s$ is a Poisson-distributed random variable with mean value $n$ and variance $n_s$. For large $n_s$, the distribution of $Y_s$ is approximately normal and therefore the distribution of

$$Z_s = \frac{Y_s - n_s}{\sqrt{n_s}} \tag{13.34}$$

is approximately a standard normal distribution. The quantity

$$V^2 = \sum_{s=1}^{k} Z_s^2 = \sum_{s=1}^{k} \frac{(Y_s - n_s)^2}{n_s} \tag{13.35}$$

is then the sum of squares of (approximately) Gaussian-distributed random variables. However, since

$$\sum_{s=1}^{k} Y_s = N, \tag{13.36}$$

it also holds that

$$\sum_{s=1}^{k} Z_s \sqrt{n_s} = 0; \tag{13.37}$$

that is, only $k-1$ of the $\{Z_s\}$ are independent. The distribution function of a random variable $V^2$, which is a sum of $k$ Gaussian random variables of which only $k-1$ are independent, is a $\chi^2$ distribution with $k-1$ degrees of freedom (see Knuth 1969; Frodesen et al. 1979). For the distribution function of the $\chi$ distribution, one obtains (see (2.157))

$$\begin{aligned} P_\chi(k-1;\chi) &= \int_0^\chi p_\chi(k-1;y)\mathrm{d}y \\ &= \int_0^\chi \frac{1}{\Gamma(\frac{k-1}{2})2^{\frac{1}{2}(k-1)-1}} y^{k-2} e^{-y^2/2}\mathrm{d}y. \end{aligned} \tag{13.38}$$

In terms of the incomplete Gamma function

$$P(a,x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} \mathrm{d}t, \quad a > 0, \tag{13.39}$$

(13.38) can also be written as

$$P_\chi(k-1;\chi) = P\left(\frac{k-1}{2}, \frac{1}{2}\chi^2\right), \tag{13.40}$$

and from

$$1 - \alpha = P\left(\frac{k-1}{2}, \frac{1}{2}\chi^2_{k-1;1-\alpha}\right), \tag{13.41}$$

one can calculate the $(1-\alpha)$ quantile $\chi^2_{k-1;1-\alpha}$. If $V^2 > \chi^2_{k-1;1-\alpha}$, then the hypothesis should be rejected with a significance level $\alpha$. However, the hypothesis is correct with a probability of $100\alpha\%$. If, on the other hand, one evaluates

$$\alpha' = 1 - P\left(\frac{k-1}{2}, \frac{V^2}{2}\right), \tag{13.42}$$

then the validity of the hypothesis is less probable for smaller $\alpha'$.

If a set of measurements is not to be compared with a theoretical distribution but rather with a second independent set of measurements, then one considers the variable

$$V^2 = \sum_{s=1}^{k} \frac{(Y_s - Y_s')^2}{Y_s + Y_s'}, \tag{13.43}$$

where $Y_s'$ is the frequency of occurrence of a result in class $s$ in the second set of measurements. Under the hypothesis that both sets of measurements can be viewed as two random samples of a single random variable, $V^2$ is again $\chi^2$ distributed, with $k-1$ $(k)$ degrees of freedom when the sum $Y_s$ is (is not) equal to the sum $Y_s'$ (Press et al. 2007).

### 13.2.4 The Kolmogorov–Smirnov Test

This test has the same purpose as the $\chi^2$-test. The advantage of this test lies in the fact that in a comparison with a continuous theoretical distribution it is not necessary to construct discrete classes first since here one measures the maximum difference between the cumulative distribution functions.

Suppose that

$$\bar{P}_N(x) = \frac{1}{N}(\text{number of } x_i \text{ with } x_i < x) \tag{13.44}$$

is the empirical distribution function and $P(x)$ is the theoretical one. Both distributions increase monotonically with increasing $x$ from zero to one. The Kolmogorov–Smirnov test measures the difference between the two distributions with the help of the two measures

$$D_N^+ = \max_x[\bar{P}_N(x) - P(x)], \tag{13.45}$$

$$D_N^- = \max_x[P(x) - \bar{P}_N(x)]. \tag{13.46}$$

$D_N^+$ measures the greatest deviation when $\bar{P}_N(x)$ is larger than $P(x)$, and $D_N^-$ measures the same when $\bar{P}_N$ is smaller than $P(x)$ (Fig. 13.5). These measures are introduced because, as with the $\chi^2$-distribution, the distributions of the $D_N^\pm$ for each $N$ can be calculated from general considerations. (For the derivation of the distributions for $D_N^\pm$, see Knuth (1969)). Denoting the density of this distribution

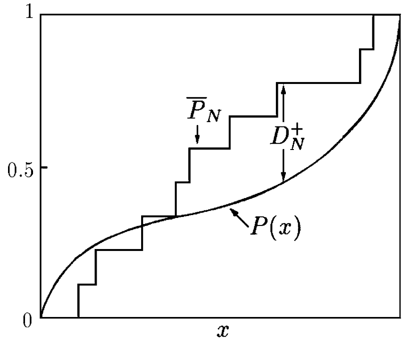


**Fig. 13.5** A theoretical and an empirical distribution function together with the measure $D_N^+$ (From Honerkamp (1994), reprinted by permission of Wiley-VCH, Weinheim)

by $p_{KS}(N;x)$, we define the quantity $d_{N;\gamma}$ through

$$\gamma = P_{KS}(N;d_{N;\gamma}) = \int_0^{d_{N;\gamma}} p_{KS}(N;x)dx. \tag{13.47}$$

For a given $\gamma$, the values of $d_{N;\gamma}$ can again be found in tables. For $N \geq 100$, $P_{KS}(N;\lambda)$ can be represented as (Frodesen et al. 1979; Press et al. 1992a)

$$P_{KS}(N;\lambda) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2N^2\lambda^2}. \tag{13.48}$$

$d_{N;1-\alpha}$ can be determined this way in a test with significance level $\alpha$. The hypothesis is rejected if $D_N^+$ or $D_N^-$ is larger than $d_{N;1-\alpha}$.

In the comparison of two sets of measurements, the factor $N$ in the quantities $D_N^{\pm}$, $\rho(N;\gamma)$, and $P(N;\gamma)$ should be replaced by

$$\frac{N_1N_2}{N_1+N_2}, \tag{13.49}$$

where $N_i$ is the size of the random sample of the $i$th set of measurements (Press et al. 2007).

### *13.2.5 The F-Test for Least-Squares Estimators*

We consider two models of the type

$$Y = \sum_{\alpha=0}^{M^k} a_\alpha X_\alpha(x), \qquad k = 1,2, \tag{13.50}$$

and especially two models with $M = M_1$ and $M_2 < M_1$, respectively. We call the model with the $(M_1 + 1)$ parameters $\{a_0, \ldots, a_{M_1}\}$ the **large model**; the model, in which $a_{M_2+1} = \ldots = a_{M_1} = 0$, the **small model**.

For given data $\{(x_i, y_i, \sigma_i)\}_{i=1}^{N}$, one may estimate the parameters $\{a_\alpha\}$ within each model by the *least-squares* method. By this method, one determines the minimum of the discrepancy

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left( y_i - \sum_{\alpha=0}^{M_k} a_\alpha X_\alpha(x_i) \right)^2 . \tag{13.51}$$

We will call the minima $(\chi^2_{M_1})_{\min}$, respectively, $(\chi^2_{M_2})_{\min}$. They are realizations of $\chi^2$-distributed random variables with $(N - M_1 - 1)$, respectively, $(N - M_2 - 1)$ degrees of freedom.

We will test the hypothesis

$$a_{M_2+1} = \ldots = a_{M_1} = 0 . \tag{13.52}$$

If this is valid, then the small model explains the data as well as the large model.

The test quantity will be constructed from the minimum discrepancies $(\chi^2_{M_1})_{\min}$ and $(\chi^2_{M_2})_{\min}$. Obviously, the value of $(\chi^2_{M_1})_{\min}$ is always smaller than $(\chi^2_{M_2})_{\min}$, because the larger the number of parameters, the better the data can be fitted. But the question is whether the minimum discrepancy $(\chi^2_{M_2})_{\min}$ within the model with $M_2$ parameters can be significantly lowered by introducing further parameters $a_{M_2+1}, \ldots, a_{M_1}$. Thus we expect that the difference between the minimum discrepancies will play a role in a test.

We choose as a test quantity

$$F = \frac{\left((\chi^2_{M_2})_{\min} - (\chi^2_{M_1})_{\min}\right)/(M_1 - M_2)}{(\chi^2_{M_1})_{\min}/(N - M_1 - 1)} . \tag{13.53}$$

We can easily determine the density of this test quantity. The numerator is the difference of two $\chi^2$-distributed random variables and therefore is also $\chi^2$-distributed; the number of degrees of freedom is equal to the difference of the numbers of the individual random variables. Therefore the test quantity $F$ is an $F$-distributed random variable with $M_1 - M_2$ and $N - M_1 - 1$ degrees of freedom.

### *13.2.6 The Likelihood-Ratio Test*

The parameters $\boldsymbol{\theta}$ may have been estimated with the maximum-likelihood estimator. First of all, for these estimates $\hat{\boldsymbol{\theta}}$, one can show (Lehmann 1991) that

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{N\to\infty}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{13.54}$$

where $\boldsymbol{\theta}_0$ are the exact parameters and the elements of the matrix $\boldsymbol{\Sigma}$ are given by

$$N\, \Sigma^{-1}_{\alpha\beta} = -\frac{\partial^2}{\partial\theta_\alpha\, \partial\theta_\beta}\, \mathrm{L}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{13.55}$$

with L as the log likelihood. That means that the maximum-likelihood estimates $\hat{\boldsymbol{\theta}}$ are asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and covariance $\boldsymbol{\Sigma}/N$, i.e., the estimator is unbiased and consistent.

This result can be used to determine the distribution of the quantity $2\,(\mathrm{L}(\hat{\boldsymbol{\theta}}) - \mathrm{L}(\boldsymbol{\theta}_0))$:

Expansion of $\mathrm{L}(\boldsymbol{\theta})$ as function of $\boldsymbol{\theta}$ around $\hat{\boldsymbol{\theta}}$ up to second order leads to

$$\begin{aligned} \mathrm{L}(\boldsymbol{\theta}) = \mathrm{L}(\hat{\boldsymbol{\theta}}) + \sum_\alpha \frac{\partial}{\partial\theta_\alpha} \mathrm{L}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_\alpha \\ -\frac{1}{2} N(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \ldots, \end{aligned} \tag{13.56}$$

and because of $\frac{\partial}{\partial\theta_\alpha}\mathrm{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$ according to the definition of $\hat{\boldsymbol{\theta}}$, we obtain for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

$$2\,(\mathrm{L}(\hat{\boldsymbol{\theta}}) - \mathrm{L}(\boldsymbol{\theta}_0)) = N(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\ . \tag{13.57}$$

Furthermore a singular-value decomposition of $\boldsymbol{\Sigma}$ reads

$$\boldsymbol{\Sigma} = \sum_{\gamma=1}^{r} w_\gamma^2\, \boldsymbol{u}_\gamma \otimes \boldsymbol{u}_\gamma, \tag{13.58}$$

where $r$ is equal to the number of parameters $\theta_\alpha$ , $\alpha = 1, \ldots, r$ and the $\{\boldsymbol{u}_\gamma\}$ are the eigenvectors of $\boldsymbol{\Sigma}$. Then the right-hand side of (13.57) can also be written as

$$N\,(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) = \sum_{\gamma=1}^{r} Z_\gamma^2 \tag{13.59}$$

with

$$Z_\gamma = \frac{\sqrt{N}}{w_\gamma}\, \boldsymbol{u}_\gamma^{\mathrm{T}}\, (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \tag{13.60}$$

$$\text{and} \qquad < Z_\gamma > = 0\,, \tag{13.61}$$

$$\begin{aligned} < Z_\gamma\, Z_{\gamma'} > &= N\, \frac{1}{w_\gamma}\, \boldsymbol{u}_\gamma^{\mathrm{T}} < (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^{\mathrm{T}} > \boldsymbol{u}_{\gamma'}\, \frac{1}{w'_\gamma} \\ &= \delta_{\gamma\gamma'}. \end{aligned} \tag{13.62}$$

Hence the right-hand side of (13.57) can also be written as the sum of the squares of $r$ independent standard normally distributed random variables $Z_\gamma$, and thus it is a $\chi^2$-distributed random variable with $r$ degrees of freedom:

$$2\,(\mathrm{L}(\hat{\boldsymbol{\theta}}) - \mathrm{L}(\boldsymbol{\theta}_0)) \sim \chi^2_r. \tag{13.63}$$

Now we are ready to discuss to consider two models $M_1(\boldsymbol{\theta}_1)$ and $M_2(\boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ is a parameter set with $r_1$ parameters, and $\boldsymbol{\theta}_2$ a parameter set with $r_2$ parameters. Let $r_1 > r_2$ so that $M_1(\boldsymbol{\theta}_1)$ is the *large model*, from which the small model $M_2(\boldsymbol{\theta}_2)$ follows by expressing the $r_2$ parameters of $\boldsymbol{\theta}_2$ in terms of the parameters of $\boldsymbol{\theta}_1$. One may also regard $M_2(\boldsymbol{\theta}_2)$ as a parameterization of $M_1(\boldsymbol{\theta}_1)$.

Given the data $\{y_1, \ldots, y_N\}$, we will test the following hypothesis:

$H_0$: Model $M_2$ explains the data as well as model $M_1$.

If this hypothesis $H_0$ is accepted, then model $M_2(\hat{\boldsymbol{\theta}}_2)$ is preferable because it can explain the data with fewer parameters.

The test quantity is

$$C = 2\,(\mathrm{L}(\hat{\boldsymbol{\theta}}_1) - \mathrm{L}(\hat{\boldsymbol{\theta}}_2)). \tag{13.64}$$

Obviously, we always have $\mathrm{L}(\hat{\boldsymbol{\theta}}_1) \geq \mathrm{L}(\hat{\boldsymbol{\theta}}_2)$ because a larger log likelihood can be achieved with a larger number of parameters. Furthermore, for the distribution of the test quantity, one obtains according to (13.63)

$$C \equiv 2\,(\mathrm{L}(\hat{\boldsymbol{\theta}}_1) - \mathrm{L}(\hat{\boldsymbol{\theta}}_2)) \sim \chi^2_{r_1 - r_2}\,. \tag{13.65}$$

Because the difference of the two log likelihood functions corresponds to the logarithm of the ratio of 2 likelihood functions, this test is also called the likelihood-ratio test.

## 13.3 Classification Methods

It is frequently possible to recognize different signals or objects as different realizations or manifestations of a single signal or cause in spite of large variabilities. For example, we recognize the sound “a” whether it is spoken by a woman, a man, or a child. We usually recognize letters of the alphabet no matter who wrote them.

If we imagine that the quantifiable characteristics of a signal or object are combined into a feature vector $\boldsymbol{x}$, then we can “explain” the variability in the signal or object as representing different realizations of a random vector $\boldsymbol{X}$. The variance of this random quantity is a measure of the variability in the manifestation of the signal or object.

Instead of speaking of a signal or an object, we will speak of a class, and each class is associated with a random vector $\boldsymbol{X}$. Realizations of $\boldsymbol{X}$ are observed, and since as a rule one has several classes and therefore several random vectors $\boldsymbol{X}_\alpha$, $\alpha = 1, \ldots, M$, a typical classification problem consists of deciding with which class $\alpha$

to associate an observation $\boldsymbol{x}$. In other words, one has to decide of which random variable $\boldsymbol{X}_\alpha$ the observation $\boldsymbol{x}$ is a realization.

Thus, for example, one might associate different written symbols with different letters or with the same letter or recognize different sounds as distinct or not distinct.

Therefore, we observe only realizations of random variables or of feature vectors. Structuring of observations often means a classification, namely, the association of a feature vector with the random vector that represents that class.

Three points should immediately be made:

- Errors will inevitably be made in this classification. We will calculate the probability of such errors.
- The problem of how to collect the observations into a feature vector $\boldsymbol{x}$ in a particular situation is an important and often difficult problem. How this is done, of course, influences the quality of the classification, that is, the frequency of errors. Here we assume that feature vectors are given.
- We are interested in a disjoint and exhaustive classification, that is, each object or signal $\boldsymbol{x}$ should be associated with exactly one class. Other classification methods such as nondisjoint or hierarchical ones will not be considered here.

### *13.3.1 Classifiers*

In the introduction to this chapter, we characterized each class $\alpha$ by a random vector. Suppose that $\rho(\alpha)$ is the probability of class $\alpha$ and $\rho(\boldsymbol{x} \mid \alpha)$ is the density distribution for a given class $\alpha$, that is, the density for $\boldsymbol{X}_\alpha$. We assume that these densities are unimodal, that is, that they have a single maximum. Suppose furthermore that $\rho(\boldsymbol{x})$ is the density for the random variable $\boldsymbol{X}$, whose realizations we observe independently of the class. Then according to the Bayesian theorem,

$$\rho(\alpha \mid \boldsymbol{x}) = \frac{\rho(\boldsymbol{x} \mid \alpha)\rho(\alpha)}{\rho(\boldsymbol{x})}. \tag{13.66}$$

However, $\rho(\alpha \mid \boldsymbol{x})$ is the probability of class $\alpha$ for a given $\boldsymbol{x}$, and one is inclined to associate with an observation vector $\boldsymbol{x}$ the class $\alpha$ for which $\rho(\alpha \mid \boldsymbol{x})$ is a maximum. If we define

$$d_\alpha(\boldsymbol{x}) = -\ln[\rho(\alpha \mid \boldsymbol{x})], \tag{13.67}$$

then the following decision rule might be established:

$$\min_\alpha\{d_\alpha(\boldsymbol{x})\} = d_\beta(\boldsymbol{x}) \quad \rightarrow \quad \boldsymbol{x} \in \text{class } \beta. \tag{13.68}$$

Because of (13.66), $d_\alpha(\boldsymbol{x})$ can be changed into a different form. Since $\rho(\boldsymbol{x})$ does not depend on $\alpha$, this function can be replaced with any other $\alpha$-independent function $\rho_0(\boldsymbol{x})$ provided that it has the same physical dimension as $\rho(\boldsymbol{x})$. Therefore, one can also write

$$d_\alpha(\boldsymbol{x}) = -\ln\left(\frac{\rho(\boldsymbol{x} \mid \alpha)\rho(\alpha)}{\rho_0(\boldsymbol{x})}\right). \tag{13.69}$$

We will usually set $\rho_0(\boldsymbol{x}) \equiv 1$ and ignore the physical dimension.

$d_\alpha(\boldsymbol{x})$ is also called the *classification function* or simply the *classifier*. If this function is known, then, according to rule (13.68), the classification problem is solved. (This procedure is also called *discrimination analysis*.) One should, however, also be interested in the classification error.

The probabilities $\rho(\alpha)$ and the densities $\rho(\boldsymbol{x} \mid \alpha)$ are the main quantities in the classification. Of course, these quantities are usually not known. In the next sections, we discuss how classifiers can be determined from a sequence of independent observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Here we first study some simple examples of classifiers and discuss the classification error.

### Examples of Classifiers

Suppose that the $\{\rho(\alpha),\ \alpha = 1, \ldots, M\}$ are all equal and that the $\{\rho(\boldsymbol{x} \mid \alpha)\}$ are normal distributions $N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$. Then up to a constant,

$$d_\alpha(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_\alpha)\boldsymbol{\Sigma}_\alpha^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_\alpha) + \frac{1}{2}\ln(\det \boldsymbol{\Sigma}_\alpha). \tag{13.70}$$

Thus the classifier $d_\alpha(\boldsymbol{x})$ is a quadratic function of $\boldsymbol{x}$.

If all covariance matrices $\boldsymbol{\Sigma}_\alpha$ are equal, $\boldsymbol{\Sigma}_\alpha \equiv \boldsymbol{\Sigma}$, then in (13.69) $\rho_0(\boldsymbol{x}) = \exp(-\frac{1}{2}\boldsymbol{x} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{x})$ could also be chosen, so that for the $\{d_\alpha(\boldsymbol{x})\}$ one also obtains linear functions

$$d_\alpha(\boldsymbol{x}) = \boldsymbol{v}_\alpha \cdot \boldsymbol{x} + a_\alpha, \qquad \alpha = 1, \ldots, m, \tag{13.71}$$

which are characterized only by vectors $\{\boldsymbol{v}_\alpha\}$ and parameters $\{a_\alpha\}$.

Furthermore, if all covariance matrices are of the form $\boldsymbol{\Sigma}_\alpha = \sigma^2\mathbf{1}$, then the classifiers can be written in the form

$$d_\alpha(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_\alpha)^2. \tag{13.72}$$

In this way each classifier is characterized entirely by the mean value vector $\boldsymbol{\mu}_\alpha$ of the class, and an observation vector $\boldsymbol{x}$ is associated with class $\alpha$ if it lies closest to its mean value vector.

As a rule, the density $\rho(\boldsymbol{x} \mid \alpha)$ is not that of a normal distribution, and $d_\alpha(\boldsymbol{x})$ may be a complicated nonlinear function. To present a relatively general parametrization of a nonlinear function, let us consider once again the linear classifier

$$d_\alpha = \sum_{i=1}^{p} v_{\alpha i} x_i + a_\alpha, \quad \alpha = 1, \ldots, m\,, \tag{13.73}$$

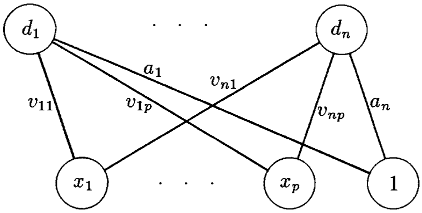


**Fig. 13.6** Representation $\{x_i\} \to \{d_\alpha\}$ of (13.73), shown as a neural network (From Honerkamp (1994), reprinted by permission of Wiley-VCH, Weinheim)

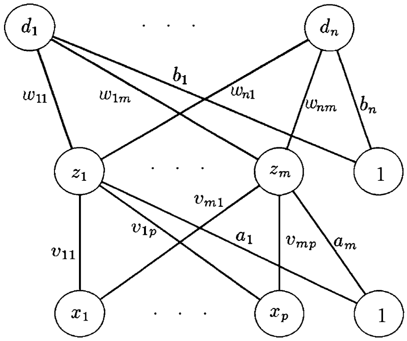


**Fig. 13.7** A neural network that represents a nonlinear map (From Honerkamp (1994), reprinted by permission of Wiley-VCH, Weinheim)

and let us consider the following illustration of the map $\{x_i\} \to \{d_\alpha\}$ (see Fig. (13.6)):

Consider two layers of points (which will also be called neurons). The $(p + 1)$ entry points are indicated by the values $x_i$ and a "1"; these are networked with the exit points, which take on the values $d_\alpha$, $\alpha = 1, \ldots, m$. The connections carry the weights $v_{\alpha i}$ and $a_\alpha$, that is, the coefficients of the classifiers. The network of points (or neurons) is thus a graphical depiction of the linear representation.

Now consider a nonlinear representation that corresponds to the graph in Fig. 13.7.

Then

$$z_\alpha = f_0\left(\sum_{i=1}^{p} v_{\alpha i} x_i + a_\alpha\right), \qquad \alpha = 1, \ldots, m', \tag{13.74}$$

and

$$d_\alpha = f_0\left(\sum_{\beta=1}^{m'} w_{\alpha\beta} z_\beta + b_\alpha\right), \qquad \alpha = 1, \ldots, m. \tag{13.75}$$

Here $f_0$ is a sigmoidal function, for example,

$$f_0(x) = \tanh(x). \tag{13.76}$$

$d_\alpha(\boldsymbol{x})$ is a nonlinear function of $\boldsymbol{x}$, characterized by the parameters $\{v_{\alpha i}, w_{\alpha\beta}, a_\alpha, b_\alpha\}$, that is, by the weights of the connections among the neurons. Such a neural network therefore represents a parametrization of a nonlinear function. It is often possible to interpret nonlinear maps that arise in practice as neural networks.

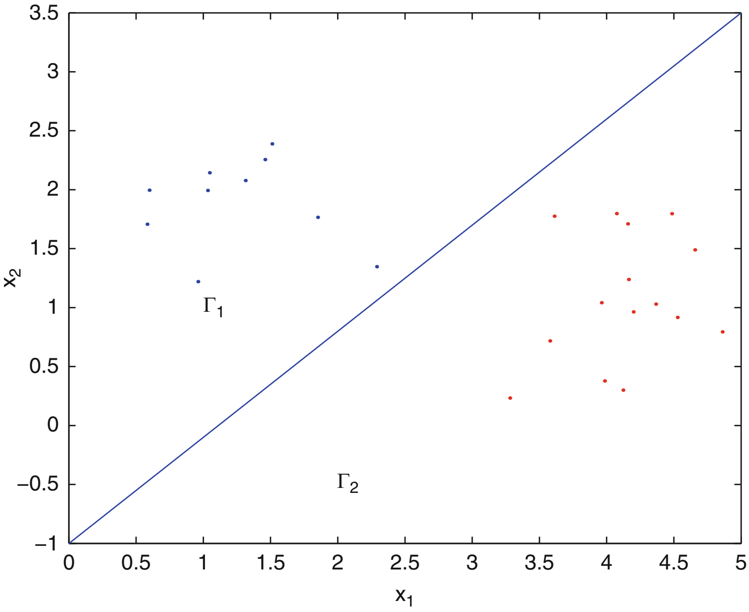


**Fig. 13.8** A linear classifier can lead only to a linear equation for a *borderline*, which then is only a hyperplane (a *straight line* in two dimensions)

*Remarks.* ***L***inear classifiers produce linear borderlines. The classification rule (13.68) decomposes the $p$-dimensional space of the feature vectors into regions $\Gamma_\alpha$ so that $\boldsymbol{x} \in \Gamma_\alpha$ is identical to the statement that $\boldsymbol{x}$ is associated with the class $\alpha$. On the border between, say, class 1 and class 2, $d_1(\boldsymbol{x}) = d_2(\boldsymbol{x})$, and for a linear classifier, this constitutes a linear equation for $\boldsymbol{x}$, which means that the border is part of a hyperplane in the $p$-dimensional space of the feature vectors (Fig. 13.8). Hence, a nonlinear borderline between two classes can be achieved only by nonlinear classifiers.

**Classification errors.** Let $\Gamma_\alpha$ be the region of class $\alpha$, so that $\boldsymbol{x} \in \Gamma_\alpha$ is identical to the statement that $\boldsymbol{x}$ is associated with the class $\alpha$. Then

$$E_{1\alpha} = \int_{\mathcal{R}^p - \Gamma_\alpha} \rho(\boldsymbol{x}' \mid \alpha)\, d^p x' \tag{13.77}$$

is the probability that a feature vector $\boldsymbol{x}$ belongs to class $\alpha$ but does not lie in $\Gamma_\alpha$ and is therefore not associated with class $\alpha$. Thus, $E_{1\alpha}$ is the probability of a classification error. We call this type of error an *error of the first kind*, in analogy with errors of the first kind in tests of statistical hypotheses. (Such a test can also be viewed as a classification problem.) Conversely,

$$E_{2\alpha} = \sum_{\substack{\beta=1 \\ \beta\neq\alpha}}^{M} \int_{\Gamma_\alpha} \rho(\boldsymbol{x}' \mid \beta)\, d^p x' \, \frac{\rho(\beta)}{\sum_{\beta'\neq\alpha} \rho(\beta')} \tag{13.78}$$

is the probability that a feature vector $\boldsymbol{x}$ is associated with class $\alpha$ though it does not belong to the class $\alpha$ . This is called an *error of the second kind.*

### 13.3.2 Estimation of Parameters that Arise in Classifiers

In Sect. 13.3.1, we introduced the classifiers

$$d_\alpha(\boldsymbol{x}). \tag{13.79}$$

These are characterized by parameters such as, for example, mean values and variances for normally distributed densities $\rho(\boldsymbol{x} \mid \alpha)$. If these parameters are not known, they must be determined “by experience”. Here we wish to consider the situation in which we have at our disposal a sequence of *independent* observation vectors

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \tag{13.80}$$

and the correct class association of each observation vector $\boldsymbol{x}_i,\ i = 1, \ldots, N$ is known. Given this information, we wish to estimate the parameters in the classifiers.

Subsequent feature vectors $\boldsymbol{x}_{N+1}, \ldots$ can then be classified. The observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ thus represent a training set of feature vectors (learning with teacher).

The following three scenarios cover most cases:

- If it can be assumed that the densities $\rho(\boldsymbol{x} \mid \alpha)$ are those of a normal distribution, $d_\alpha(\boldsymbol{x} \mid \alpha)$ has the form (13.70), and from the feature vectors that belong to class $\alpha$, only the mean value $\boldsymbol{\mu}_\alpha$ and the covariance matrix $\boldsymbol{\Sigma}_\alpha$ have to be estimated. This is easy (see Chap. 8).
- If the feature components $\{x_i, i = 1, \ldots, p\}$ have discrete values, then the probabilities $\rho(\boldsymbol{x} \mid \alpha)$ can often be estimated just by counting the corresponding events in the training set and finally normalizing. For such a direct estimation of the probabilities, it is helpful to assume that all of the feature components are independent:

$$\rho(\boldsymbol{x} \mid \alpha) = \prod_{i=1}^{p} \rho(x_i \mid \alpha)\,. \tag{13.81}$$

If a probability $\rho(x_i \mid \alpha)$ is small, it may happen that there is no corresponding event in the training set and the estimation of this probability would amount to zero. This would lead to zero estimations for all of those probabilities in which $\rho(x_i \mid \alpha)$ appears as a factor. To avoid that, one may introduce the a priori probability $\rho_a(x_i \mid \alpha)$. Then, if $c$ is a possible value of $x_i$, $n_c$ the number of events $x_i = c \mid \alpha$, and $n$ the number of cases with class $\alpha$ in the training set, then one may estimate

$$\hat{\rho}(x_i \mid \alpha) = \frac{n_c + m\rho_a(x_i \mid \alpha)}{n + m}, \tag{13.82}$$

where $m$ is a number that measures the weight of the a priori estimate. For $m = 0$, we get the usual estimate based only on the number of events in the training set;, for $m \to \infty$, the estimate becomes the a priori estimate. A usual choice for $m$ is $m = 1/\rho_a(x_i \mid \alpha)$, so that

$$\hat{\rho}(x_i \mid \alpha) = \frac{n_c + 1}{n + 1/\rho_a(x_i \mid \alpha)}. \tag{13.83}$$

- In the more general case, the parameters must be chosen so that for all $\{\boldsymbol{x}_i\}$ that belong to class $\alpha$, $d_\alpha(\boldsymbol{x}_i)$ is smaller than $d_\beta(\boldsymbol{x}_i)$ for $\beta \neq \alpha$. One may formulate this condition by introducing the function $\boldsymbol{F}$ which assigns to the vector $\boldsymbol{d}$ with components $d_\alpha(\boldsymbol{x}, \alpha = 1, \ldots, M)$ a class vector $\boldsymbol{y}$ with components $y_\alpha = 1$ if $\alpha = \arg\min_\beta d_\beta$ and $y_\alpha = 0, \alpha = 1, \ldots, M$ else. Let us call such a vector $\boldsymbol{y}^\alpha$ to indicate which component of the vector is nonzero. Then we have to find a map $\boldsymbol{F} \circ \boldsymbol{d}$ which, on the training set, reproduces the map

$$\boldsymbol{y}_i \equiv \boldsymbol{y}^{\alpha_i} = (\boldsymbol{F} \circ \boldsymbol{d})(\boldsymbol{x}_i), \qquad i = 1, \ldots, N, \tag{13.84}$$

where $\alpha_i$ is the class of the feature vector $\boldsymbol{x}_i$. Thus $\boldsymbol{F} \circ \boldsymbol{d}$ is a map from the $p$-dimensional feature space into the $M$-dimensional space of class vectors $\boldsymbol{y}_i \equiv \boldsymbol{y}^{\alpha_i}$. One may formulate such a map by a neuronal net and may adapt the parameters in such a model for the map with the techniques used in connection with neural nets (see, e.g., Matlab Neural Network Toolbox).

### 13.3.3 Automatic Classification (Cluster Analysis)

In the previous, section we assumed that we have at our disposal a sequence of independent feature vectors whose class association we know ("learning with a teacher"). Here we begin with the assumption that we know only the sequence of observation vectors, but that we do not know their class association. However, we assume that we know the total number of classes. Thus, here we wish to associate a sequence of independent observation vectors

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \tag{13.85}$$

with classes $\alpha$, $\alpha = 1, \ldots, m$, and thereby develop a procedure that will allow us to associate every future observation vector with one of these classes. This is called *cluster analysis*, or, since there is no prior knowledge of class association, it is called *automatic classification* or also "learning without a teacher."

This problem is very elegantly formulated as a problem with incomplete data which one can handle with the EM algorithm introduced in Sect. 12.1. To show this, let us change the notation: We will in the following denote the observation vectors $\boldsymbol{x}_i$, $i = 1, \ldots, N$ by $\boldsymbol{y}_i$, $i = 1, \ldots, N$, and the unknown classes by $x_i$, $i = 1, \ldots, N$. Thus the complete data set would be $(\boldsymbol{y}_i, x_i)$, $i = 1, \ldots, N$, but only the set $\boldsymbol{y}_i, i = 1, \ldots, N$ is observed. We will assume some model for the probability $\rho(\boldsymbol{y}|x, \boldsymbol{\theta})$ and for $\rho(x|\boldsymbol{\theta})$, $x = 1, \ldots, M$, where $\boldsymbol{\theta}$ is a collection of parameters. If one could estimate these parameters $\boldsymbol{\theta}$ from the incomplete data set $\boldsymbol{y}_i$, $i = 1, \ldots, N$, one would get complete knowledge about the classifiers $\rho(\boldsymbol{y}|x, \boldsymbol{\theta})$ and could use them for classification.

Now, as an estimate of $\boldsymbol{\theta}$, we may use the maximum-likelihood estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}'} \rho(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N|\boldsymbol{\theta}'), \tag{13.86}$$

and now

$$\begin{aligned}\rho(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N|\boldsymbol{\theta}) &= \sum_{x_1,\ldots,x_N} \rho(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N|x_1, \ldots, x_N, \boldsymbol{\theta})\rho(x_1, \ldots, x_N|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} \rho(\boldsymbol{y}_i|x_i, \boldsymbol{\theta})\rho(x_i|\boldsymbol{\theta})\,. \end{aligned} \tag{13.87}$$

This is an especially simple type of problem, in which one determines the maximum of the likelihood with help of the EM algorithm because compared to the hidden Markov models, here, the $\{x_i\}$ are independent.

*Example.* Let

$$\rho(\boldsymbol{y}, x|\boldsymbol{\theta}) \equiv \rho(\boldsymbol{y}|x, \boldsymbol{\theta})\,\rho(x|\boldsymbol{\theta}) = N(\boldsymbol{\mu}(x), \sigma^2\mathbf{1})\frac{1}{M}, \tag{13.88}$$

where $M$ is the number of classes and $N(\mu, \sigma^2)$ means the normal density with mean $\mu$ and variance $\sigma^2$. Thus the parameter vector $\boldsymbol{\theta}$ contains the $M$ vectors $\boldsymbol{\mu}(x), x = 1, \ldots, M$ and $\sigma$.

In general,

$$\rho(x|\boldsymbol{y}, \boldsymbol{\theta}) = \frac{\rho(\boldsymbol{y}, x|\boldsymbol{\theta})}{\rho(\boldsymbol{y})} \propto \rho(\boldsymbol{y}, x|\boldsymbol{\theta}), \tag{13.89}$$

thus in this example, using normalization, we get

$$\rho(x = \alpha|\boldsymbol{y}, \boldsymbol{\theta}) = \frac{\exp\Big(-(\boldsymbol{y} - \boldsymbol{\mu}(\alpha))^2/(2\sigma^2)\Big)}{\sum_\beta \exp\Big(-(\boldsymbol{y} - \boldsymbol{\mu}(\beta))^2/(2\sigma^2)\Big)}\,. \tag{13.90}$$

Then we are prepared to calculate the function $\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ of Sect. 12.1 and obtain

$$\begin{aligned}\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{N} \sum_{x_i=1}^{M} \ln \rho(\boldsymbol{y}_i, x_i | \boldsymbol{\theta}) \rho(x_i | \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)}) \\ &= \sum_{i=1}^{N} \sum_{\alpha=1}^{M} \left[ -\frac{(\boldsymbol{y}_i - \boldsymbol{\mu}(\alpha))^2}{2\sigma^2} + \ln \frac{1}{\sqrt{2\pi\sigma^2}} \right] \rho(x_i = \alpha | \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)}) \,.\end{aligned} \tag{13.91}$$

The maximum of $\mathrm{A}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$ can easily be determined by setting the derivative, e.g., with respect to $\boldsymbol{\mu}(\alpha)$, to zero. Thus one obtains, e.g.,

$$\boldsymbol{\mu}^{(k+1)}(\alpha) = \frac{\sum_{i=1}^{N} \boldsymbol{y}_i \rho(x_i = \alpha | \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{N} \rho(x = \alpha | \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)})} \tag{13.92}$$

and

$$(\sigma^2)^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\alpha=1}^{K} (\boldsymbol{y}_i - \boldsymbol{\mu}(\alpha))^2 \rho(x_i = \alpha | \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)}). \tag{13.93}$$

# Appendix: Random Number Generation for Simulating Realizations of Random Variables

In statistical physics experimental results and observations are considered as realizations of random variables. The problems that follow are related to the simulation of experimental results or observations. For this purpose we will employ a machine which provides us with realizations of random variables. Of course, this machine will be a computer, and from what has been said in Sect. 2.5, we need only a prescription for generating realizations of random variables uniformly distributed in the interval $[0, 1]$, because given such random variables we can generate the realizations of any other random variable with appropriate transformations.

Naturally, a computer, being a deterministic machine, can only generate pseudo-random numbers, i.e., a sequence of numbers which are 'for all practical purposes' mutually independently and 'sufficiently' uniformly distributed on the interval $[0, 1]$. How well these two conditions are satisfied depends on the prescription that generates this sequence. Since applications often require millions of random numbers, the random number generating algorithm must be fast. A frequently used algorithm introduced by Lehmer (1951) is based on the linear congruence method in the form

$$r_{n+1} = (a r_n + b) \bmod m. \tag{A.1}$$

Here $a$, $b$, and $m$ are appropriately chosen integer parameters. Starting from an initial value $r_0$, the recursion formula (A.1) yields a sequence of integers $(r_0, \ldots, r_n, \ldots)$ in the interval $[0, m-1]$.

Upon division by $m$, one obtains the random number $\eta_n$ in the interval $[0, 1]$:

$$\eta_n = r_n / m. \tag{A.2}$$

The same initial value $r_0$ will always lead to the identical sequence of random numbers. If $r_0$ is made time dependent, e.g. by linking it to the internal computer clock, one always obtains new and essentially unpredictable sequences.

The quality of the random number generator depends strongly on the choice of the parameters $m$, $a$ and $b$. Since the numbers in the sequence $(\ldots, r_n, \ldots)$ are

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1, © Springer-Verlag Berlin Heidelberg 2012

repeated after at most $m$ numbers, $m$ should be chosen to be large (perhaps equal to the largest integer that the computer can handle), and $a$ and $b$ should not be factors of $m$.

The choice of $b$ is not critical – in particular, one may choose $b = 0$. However, the statistical properties of the random number generator depend strongly on $a$ and its relation to $m$.

With even the most skilful choice of $a$, $b$, and $m$, one will discover certain regularities in the pseudorandom numbers generated in this manner. The points

$$y_n = (\eta_n, \eta_{n+1}, \ldots, \eta_{n+s-1}) \tag{A.3}$$

in $s$-dimensional space with $s \geq 2$ lie on a small number of hypersurfaces. Various general nonlinear recursion formulas have been developed in order to avoid these correlations (Eichenauer et al. 1988; Niederreiter 1989). However, these nonlinear recursion formulas require more computer time.

If the correlations in the sequence are completely unimportant for the particular purpose at hand, so-called quasirandom numbers can be used (Stroud 1971; Davis and Rabinowitz 1975). The recursion formulas for the generation of these random numbers are constructed in such a way that the desired distribution is increasingly well approximated as the length of the sequence increases.

Every standard computer system offers a reasonably good random number generator which provides us with independent realizations of random numbers uniformly distributed in $[0, 1]$. In principle, one should check such a random number generator before it is used for serious applications (see Knuth 1969; Rubinstein 1981 for tests of independence and uniformity). In the problems and exercises that follow we will assume that the available generator will satisfy our needs.

# Problems

## Chapter 2

**2.1.** Let $F$ be a Borel space and $\mathcal{P}$ be a probability distribution on $F$ which satisfies the Kolmogorov axioms. Show that

(a) For $A, B \in F$ also $A \cap B \in F$.
(b) $\mathcal{P}(A) \leq \mathcal{P}(B)$, if $A \subseteq B$.
(c) $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$.

**2.2.** Write a program which generates a binomial distribution using the following prescription: Take $n$ random numbers uniformly distributed in $[0,1]$, divide the interval into two segments of length $p$ and $q$, and determine the number $k$ of random numbers in the segment of length $p$. Repeat this step sufficiently often and construct a histogram. Compare the result for suitably chosen parameters with a Poisson distribution.

**2.3.** Let $K$ be a random number distributed according to $P(k)$ and with values in $\{0, 1, 2, 3, \ldots\}$. The generating function $G(z)$ for such a discrete distribution is defined by

$$G(z) \equiv \sum_{k=0}^{\infty} z^k P(k).$$

(a) Show that $G(z)$ is defined for all complex $z$ with $|z| \leq 1$ and that the moments (if they exist) satisfy

$$\langle K^n \rangle = \left( z \frac{\mathrm{d}}{\mathrm{d}z} \right)^n G(z) \bigg|_{z=1} .$$

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1, © Springer-Verlag Berlin Heidelberg 2012

(b) Determine the generating function for the Poisson distribution $P(\lambda; k)$ and show that

$$\langle K(K-1)(K-2)\ldots(K-n+1)\rangle = \lambda^n .$$

(c) Conclude that in particular

$$\langle K\rangle = \lambda, \quad \mathrm{Var}(K) = \lambda .$$

**2.4.** Simulate the 'goat problem' mentioned in Sect. 2.2.

**2.5.** A company which produces chips owns two factories: Factory A produces 60% of the chips, factory B 40%. Hence, if we choose one of the company's chips at random, it originates from factory A with a probability of 60%. Furthermore, we suppose that 35% of the chips from factory A are defective and 25% from factory B. We find that a randomly chosen chip is defective. In Sect. 2.2 we determined the probability that this chip originates from factory A. Simulate this situation $n$ times and compare the relative frequency of a defective chip from factory A with the theoretical probability.

**2.6.** Consider a discrete probability distribution $P(i, j)$ on a base set $\Omega$, consisting of pairs $(i, j)$, where $i, j = 1, 2, 3, \ldots$. We define the distributions for the random variables $I$ and $J$ by

$$P_1(i) \equiv \sum_j P(i, j), \quad \text{and } P_2(j) \equiv \sum_i P(i, j).$$

The corresponding entropies of these distributions are defined by

$$S \equiv -\sum_{i,j} P(i, j) \ln P(i, j)$$

$$S_1 \equiv -\sum_i P_1(i) \ln P_1(i)$$

$$S_2 \equiv -\sum_j P_2(j) \ln P_2(j).$$

Show that these entropies are additive, i.e.,

$$S = S_1 + S_2,$$

if and only if the random variables $I$ and $J$ are statistically independent, i.e., if $P(i, j) = P_1(i)P_2(j)$.

**2.7.** Use the maximum entropy method to determine the distribution of a real random number which has maximum entropy under the condition that the mean $\langle X\rangle = \mu$ and the variance $\mathrm{Var}(X) = \sigma^2$ are known.

**2.8.** (a) Given a random variable $X$ uniformly distributed in the interval $[0, 1]$, determine the density $\varrho_Y(y)$ of the random variable $Y = X^n$ for $n > 1$.

(b) Given a random variable $X$ uniformly distributed in the interval $[0, 1]$, determine the density of the random variable $Y = -\ln X$.

(c) Given two independent random variables $X_1$ and $X_2$, both uniformly distributed in the interval $[0, 1]$, show that

$$Y_1 = \sqrt{-2\ln X_1}\cos 2\pi X_2,$$

and

$$Y_2 = \sqrt{-2\ln X_1}\sin 2\pi X_2$$

define two independent standard normal random variables.

**2.9.** Let $I$ and $J$ be two independent random variables with Poisson distributions $P(\lambda_1; i)$ and $P(\lambda_2; j)$, respectively. Determine the distribution of the random variable $K = I + J$ by

(a) Using the general expression for the distribution of the sum of random variables, and

(b) Using the corresponding relation between the characteristic functions.

**2.10.** Consider two dependent real random variables $X_1$ and $X_2$ together with their sum $Z = X_1 + X_2$. Show that

$$\langle Z\rangle = \langle X_1\rangle + \langle X_2\rangle,$$

but, in general,

$$\mathrm{Var}(Z) \neq \mathrm{Var}(X_1) + \mathrm{Var}(X_2),$$

and instead

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

**2.11.** Consider a real random variable $X$ with the density

$$\varrho_X(x) = \frac{1}{\pi}\frac{1}{1 + (x-\mu)^2}$$

(Cauchy distribution). Let $X_1, X_2, \ldots, X_N$ be mutually independent copies of the random variable $X$. Show that the random variable

$$Z = \frac{1}{N}\sum_{i=1}^{N} X_i$$

has the same density as $X$, i.e.,

$$\varrho_Z(z) = \varrho_X(z).$$

**2.12.** Entropy and temperature:
Consider the random variables

$$X_i = \frac{P_i^2}{2m}, \qquad Y_N = \frac{1}{N}\sum_{i=1}^{N} X_i,$$

where the densities and expectation values of $X_i$ shall be

$$\varrho_{X_i}(x) = \frac{1}{Z}\mathrm{e}^{-\beta x} = \beta \mathrm{e}^{-\beta x}, \quad \langle X_i \rangle_\varrho = \beta.$$

$Y_N$ has the *large deviation* property:

$$\varrho_{Y_N}(y) \propto \mathrm{e}^{-Ng(y)}.$$

Determine $g(y)$ and interpret this function.

**2.13.** Renormalization group transformations:
Realize a sequence of $2^{13}$ random numbers distributed with respect to a stable density of index $\alpha$, with $0 < \alpha \leq 2$. From this set of random numbers construct new sets of random numbers of length $2^{12}$, $2^{11}$, ..., by successive applications of the renormalization group transformation $T_2$:

$$X_i' = (T_2X)_i = \frac{1}{n^{1/\beta}} \sum_{j=in}^{(i+1)n-1} X_j,\ 0 < \beta \leq 2.$$

By making histograms, visualize the distributions of $X_i$ and of the renormalized random variables. Investigate the cases $\beta = \alpha$ and $\beta \neq \alpha$.

**2.14.** Legendre transformation:
Write a numerical routine for the determination of the Legendre transform of a function $f(t)$. Check this routine for the following functions:

(a) The free energy function of the normal distribution $N(\mu, \sigma^2)$: $f(t) = \mu t + \sigma^2 t^2/2$.
(b) The free energy function of the exponential distribution: $f(t) = \log \lambda/(\lambda - t)$ for $t \in ]-\infty, \lambda[$.
(c) The free energy function of the symmetric two-point distribution: $f(t) = \log\cosh t$.

**2.15.** Eigenfunctions of the linearized renormalization group transformation
Let $DT_2$ be the linearization of the renormalization group transformation $T_2$ at the point $\varrho^*$. Furthermore, let $\Phi_i$ be the eigenfunctions with eigenvalues $\lambda_i$ of this linear operator,

$$(DT_2)\Phi_i = \lambda\Phi_i .$$

We consider the special case $\alpha = 2$ and the density of the standard normal distribution $\varrho^*(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$. Show that the eigenfunctions and eigenvalues are given by

$$\Phi_n(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}H_n(x) \quad \text{and} \quad \lambda_n = 2^{1-n/2},\ n = 0, 1, \ldots ,$$

where $H_n$ denote the Hermite polynomials

$$H_n(x) = (-1)^n \mathrm{e}^{x^2/2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\mathrm{e}^{-x^2/2}.$$

**2.16.** Expansion of the deviation from a stable density with respect to eigenfunctions of the linearized renormalization group transformation:

Consider the random variable $X$ with density $\varrho(x)$, expectation value $\mathrm{E}[X] = \mu$, and variance $\mathrm{Var}[X] = \sigma^2$. Examine the deviation of the density of this random variable from the density $\varrho^*(x)$ of the normal distribution with mean $\mu$ and variance $\sigma^2$.

For this purpose, first generalize the result of Problem 2.7 to general $\mu$ and $\sigma^2$ by a translation and a rescaling of $X$. Expand the difference $\eta = \varrho - \varrho^*$ with respect to the eigenfunctions. Determine the first two nonvanishing expansion coefficients as a function of the moments of $X$.

Finally, consider explicitly the random variable $X$ which is uniformly distributed in the interval $[-0.5, 0.5]$ ($\mu = 0$, $\sigma^2 = \frac{1}{12}$). Plot the difference $\eta(x)$ and the first nonvanishing contribution to the expansion of $\eta(x)$ in terms of the eigenfunctions.

## Chapter 3

**3.1.** Consider an ideal gas consisting of $N$ particles in a volume $V$. If the particles are regarded as distinguishable, one obtains from the number $\Omega(E, V, N)$ of microstates the following expression for the entropy (cf. Sect. 3.1):

$$\begin{aligned} S(E,V,N) &= k_\mathrm{B} \ln \Omega(E,V,N) \\ &= Nk_\mathrm{B}\left\{ \ln\left[\frac{V}{\mathrm{h}^3}\left(\frac{4\pi m E}{3N}\right)^{3/2}\right] + \frac{3}{2}\right\} . \end{aligned}$$

This expression, however, cannot be correct for the following reason. From phenomenological thermodynamics it is known that the entropy has to be a homogeneous function of the extensive state variables, i.e., if $E$, $V$, and $N$ are increased by a factor of $\lambda$ (such that the intensive state variables, like energy per particle or volume per particle remain unchanged), then the entropy also increases by the same factor:

$$S(\lambda E, \lambda V, \lambda N) = \lambda S(E, V, N).$$

The above formula for the entropy violates this condition. Show that one obtains an expression which is homogeneous as a function of the extensive state variables, if the number $\Omega$ of states is decreased by a factor of $1/N!$. What is the physical reason behind this?

**3.2.** Compute the canonical partition function for a system of $N$ uncoupled harmonic oscillators with frequency $\omega$ and mass $m$ both classically and quantum mechanically. Show that the resulting expressions coincide for the case $\hbar\omega/k_\mathrm{B}T \ll 1$. Also determine the entropy and the internal energy $E$ from the quantum mechanical partition function.

**3.3.** (a) Show that in general:

$$\frac{\kappa_S}{\kappa_T} = \frac{C_V}{C_p},$$

where

$$\kappa_S \equiv -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_S$$

is the adiabatic compressibility and

$$\kappa_T \equiv -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T$$

the isothermal compressibility. $C_V$ and $C_p$ denote the heat capacities for constant volume $V$ and constant pressure $p$, respectively.

(b) In the same manner, show that

$$\frac{\chi_S}{\chi_T} = \frac{C_M}{C_B},$$

where now

$$\chi_S \equiv \left(\frac{\partial M}{\partial B}\right)_S$$

is the adiabatic susceptibility and

$$\chi_T \equiv \left(\frac{\partial M}{\partial B}\right)_T$$

the isothermal susceptibility. $C_M$ and $C_B$ denote the heat capacities for constant magnetization $M$ and constant magnetic field $B$, respectively.

**3.4.** Chemical equilibrium:
A vessel of volume $V$ contains a mixture of two ideal gases consisting of molecules of type $A$ and type $B$, which may react with each other according to the equation

$$2A \longleftrightarrow B.$$

In the following we will neglect the internal degrees of freedom of the molecules and consider the vessel as closed, i.e., the volume $V$ as well as the internal energy $E$ remain constant. Show that in thermal equilibrium

$$\frac{N_A^2}{N_B} = V\left(\frac{2\pi k_\mathrm{B}T}{\mathrm{h}^2}\frac{m_A^2}{m_B}\right)^{3/2}.$$

This is known as the *law of mass action*. Here, $N_A$ and $N_B$ denote the number of molecules of type $A$ and type $B$, respectively, and $m_A$ and $m_B$ are their respective masses.

**3.5.** Consider a classical system of $N$ identical particles with an interaction given by the potential

$$V(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N) = \sum_{1\le i<j\le N} V_2(\boldsymbol{r}_i,\boldsymbol{r}_j).$$

The potential $V_2$, which describes the pair interaction, is symmetric. In thermal equilibrium the $N$-particle probability density is

$$\varrho_N(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N) = \frac{1}{\lambda^{3N}N!}\exp\{(F - V(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N))/k_\mathrm{B}T\},$$

where $F$ denotes the free energy and $\lambda$ the thermal wavelength. For $n = 1, 2, \ldots,$ $N-1$ we introduce the marginal distributions

$$\varrho_n(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n) = \frac{N!}{(N-n)!}\int \mathrm{d}^3r_{n+1}\ldots d^3r_N\,\varrho_N(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N).$$

Show that the following hierarchy of equations holds:

$$\frac{\partial\varrho_n(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n)}{\partial\boldsymbol{r}_i} + \frac{1}{k_\mathrm{B}T}\frac{\partial V_n(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n)}{\partial\boldsymbol{r}_i}\varrho_n(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n)$$
$$+\frac{1}{k_\mathrm{B}T}\int \mathrm{d}^3r\,\frac{\partial V_2(\boldsymbol{r}_i,\boldsymbol{r})}{\partial\boldsymbol{r}_i}\varrho_{n+1}(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n,\boldsymbol{r}) = 0,$$

where

$$V_n(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n) \equiv \sum_{1 \le i < j \le n} V_2(\boldsymbol{r}_i, \boldsymbol{r}_j).$$

What happens for the cases $n = 1$ and $n = N$?

**3.6.** The virial expansion for the equation of state of a real gas reads

$$pV = Nk_{\mathrm{B}}T \sum_{l=1}^{\infty} a_l(T) \left(n\lambda^3\right)^{l-1},$$

where $n$ denotes the particle density and $\lambda$ the thermal wavelength. The interaction between the particles is described by a potential $u(r)$, where $r$ is the distance between the particles. The first two virial coefficients are

$$a_1 = 1, \qquad a_2 = -\frac{2\pi}{\lambda^3} \int_0^{\infty} \left(\mathrm{e}^{-u(r)/k_{\mathrm{B}}T} - 1\right) r^2 \, \mathrm{d}r.$$

Compute the second virial coefficient and the corresponding approximation for the equation of state for the following potential:

$$u(r) = \begin{cases} +\infty\,, \; r < a \\ -u_0\,, \;\; a < r < b \\ 0\,, \qquad b < r. \end{cases}$$

**3.7.** Two fixed electric dipoles with dipole moments $\boldsymbol{m}$ and $\boldsymbol{m}'$ are in thermal equilibrium at a temperature $T$. Let the distance vector between these two dipoles be $\boldsymbol{R}$. For the case of high temperatures (i.e., $mm'/k_{\mathrm{B}}TR^3 \ll 1$), determine the average force $\langle \boldsymbol{F} \rangle$ through which the two dipoles mutually interact. Show in particular that the absolute value of this force is proportional to $R^{-7}$.

## Chapter 4

**4.1.** Consider a system of $N$ spin-1/2 particles, fixed at the positions of the lattice sites of a crystal lattice. The system shall be in a homogeneous magnetic field $\boldsymbol{B} = B\boldsymbol{e}_z$ ($\boldsymbol{e}_z$ is the unit vector in $z$-direction). We are only interested in the $N$ spin degrees of freedom and neglect all other degrees of freedom as well as the mutual couplings among the spins and the couplings between the spins and the lattice degrees of freedom. In this approximation the Hamiltonian operator of the system reads

$$H = -B\gamma \sum_{i=1}^{N} S_z^{(i)},$$

where $S_z^{(i)}$ is the $z$-component of the spin operator for the particle at lattice site $i$. $\gamma$ denotes the gyromagnetic ratio.

(a) What determines the microstate of the system? Compute the number $\Omega(E, B, N)$ of microstates with total energy $E$ for fixed $B$ and $N$. From $\Omega(E, B, N)$ determine the entropy by using Stirling's formula. Make a rough drawing of the entropy as a function of $E$. Where does the entropy assume its maximum value?

(b) Compute and sketch the following quantities (cf. Sect. 4.4):

1. The temperature $T$ as a function of the energy $E$.
2. The energy $E$ as a function of the temperature $T$.
3. The specific heat

$$C = \frac{\partial E}{\partial T}$$

as a function of the temperature.

(c) Let $N_\pm$ be the average number of spins in the eigenstate $\left|\pm\frac{1}{2}\right\rangle$ of $S_z$. Show that

$$\frac{N_-}{N_+} = \mathrm{e}^{-\varepsilon/k_\mathrm{B}T},$$

where $\varepsilon \equiv \gamma B$. Compute and sketch the magnetization

$$M \equiv \frac{\gamma}{2}(N_+ - N_-).$$

(d) From part (b) of this problem it follows that, in a certain region, the temperature becomes negative. What property of the system under consideration is responsible for this behavior? Think of possible physical conditions under which one could realize such a system.

**4.2.** Landau theory of phase transitions and critical indices:
Consider a magnetizable material whose free energy $F$ close to the critical temperature $T_\mathrm{c}$ has the following form as a function of magnetization $M$:

$$F(T, M) = F_0(T) + L_2(T) \cdot M^2 + L_4(T) \cdot M^4,$$

where

$$L_k(T) = l_{k0} + l_{k1} \cdot (T - T_\mathrm{c}) + l_{k2} \cdot (T - T_\mathrm{c})^2 + \dots .$$

The constants $l_{kj}$ satisfy

$$l_{20} = 0 \qquad \text{and} \qquad l_{kj} > 0 \qquad \text{otherwise.}$$

Furthermore, we assume that

$$C_{0M} \equiv -T \left( \frac{\partial^2 F_0}{\partial T^2} \right)_M \neq 0$$

is nonsingular. Setting

$$\varepsilon \equiv \frac{T - T_\mathrm{c}}{T_\mathrm{c}}$$

we can define the critical indices $\alpha$, $\alpha'$, $\beta$, $\delta$, $\gamma$, and $\gamma'$ by the singular behavior close to $\varepsilon = 0$:

$$\begin{aligned}
&C_H(\varepsilon, H = 0) \propto \varepsilon^{-\alpha} && \text{for } T > T_\mathrm{c}, \\
&C_H(\varepsilon, H = 0) \propto (-\varepsilon)^{-\alpha'} && \text{for } T < T_\mathrm{c}, \\
&M(\varepsilon, H = 0) \propto (-\varepsilon)^{\beta} && \text{for } T < T_\mathrm{c}, \\
&M(0, H) \propto H^{1/\delta} && \text{for } T = T_\mathrm{c}, \\
&\chi_T(\varepsilon, H = 0) \propto \varepsilon^{-\gamma} && \text{for } T > T_\mathrm{c}, \\
&\chi_T(\varepsilon, H = 0) \propto (-\varepsilon)^{-\gamma'} && \text{for } T < T_\mathrm{c}.
\end{aligned}$$

$C_H$ denotes the heat capacity for fixed magnetic field $H$ and

$$\chi_T \equiv \left( \frac{\partial M}{\partial H} \right)_T .$$

(a) Discuss the free enthalpy $G = F - HM$ for fixed magnetic field $H$ as a function of the magnetization $M$ above and below the critical temperature. Which condition determines the magnetization if the magnetic field is given?
(b) Evaluate the critical indices $\beta$, $\gamma$, $\gamma'$, and $\delta$.
(c) Evaluate the critical indices $\alpha$ and $\alpha'$. Determine the jump $\Delta C_H$ of the heat capacity $C_H$ at the critical point:

$$\Delta C_H \equiv \lim_{\varepsilon \to 0+} C_H - \lim_{\varepsilon \to 0-} C_H .$$

**4.3.** Cliques and neighborhoods:
Consider a two dimensional triangular lattice with a 6-neighborhood as shown in Fig. P.1. Determine the cliques.

**4.4.** The energy spectrum:
Discuss the various energy states for an auto logistic model and for an $M$-level model. For both models consider the case $M = 3$ corresponding to three possible spin orientations.

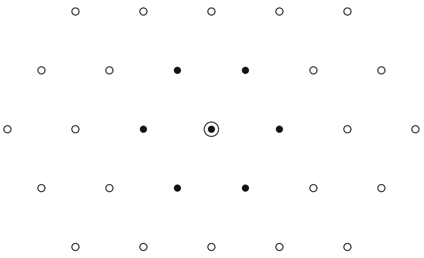


**Fig. P.1** Triangular lattice with a two dimensional 6-neighborhood

**4.5.** The Curie–Weiss model at the phase transition:
Show that the density $\varrho_Y(y)$ in the Curie–Weiss model in the vicinity of $y = 0$ for $\beta J_0 = 1$ can be written in first approximation as

$$\varrho_Y(y) = \mathrm{e}^{-y^4/4}.$$

Realize a random variable with this density using the rejection method and generate a time series as a sequence of such independent realizations. Compare the result with a Gaussian white noise.

**4.6.** Divergent covariance at the phase transition:
A measure for fluctuations is the following quantity derived from the covariance matrix:

$$\frac{1}{N}\sum_{i,j=1}^{N} \mathrm{Cov}(X_i, X_j).$$

Show that in the Curie–Weiss model close to the phase transition, i.e., at $\beta J_0 = 1$, this quantity diverges as $N^{1/2}$, while it remains finite for $|\beta J_0 - 1| > \varepsilon > 0$ even in the limit $N \to \infty$.

**4.7.** Density of states:
Apply the method for rewriting the partition function described in Sect. 4.6 to the Curie–Weiss model and show that the density of states may be written in the form

$$Z(\beta, \Theta, N) = \int \mathrm{d}y\ \mathrm{e}^{-N\lambda(y,\Theta,\beta)}.$$

## Chapter 5

**5.1.** Realize a white noise.

**5.2.** In the same way as in the previous problem, simulate the stochastic processes for

(a) Radioactive decay
(b) The Poisson process
(c) The chemical reaction $A \to X, 2X \to B$.

**5.3.** Consider a harmonic oscillator interacting with a radiation field. Due to this interaction there will be transitions between the energy eigenstates $|n\rangle$, $n = 0, 1, 2, \ldots$. The probability that a jump $n-1 \to n$ occurs within one time unit shall be $\beta n$, and the probability of a jump $n \to n-1$ shall be $\alpha n$, where $\beta < \alpha$. Simulate this system in a similar way to the chemical reaction in Problem 3.4 and verify that the stationary distribution for the energy levels is given by

$$\varrho(n) \propto \left(\frac{\beta}{\alpha}\right)^n .$$

**5.4.** Consider the stochastic process $X(t)$ for $t = 0, 1, \ldots$, defined by the equation

$$X(t) = \alpha X(t-1) + \sigma\eta(t), \quad X(0) = 0.$$

Take the values $\alpha = 0.8$, $\sigma = 1$, and take $\eta(t)$ to be a white noise.

(a) Determine the densities $\varrho(x,t)$ and $\varrho(x,t \,|\, x', t-1)$ and also the covariances $\langle X(t)\, X(t-\tau)\rangle$. Why is this process a model for a red noise?
(b) Realize this process and compare the time series so obtained with the time series of a white noise.

**5.5.** Simulation of fractal Brownian motion:
To generate realizations of fractal Brownian motions in time, one needs long-range correlated random numbers. Seek information about so-called *midpoint displacement* methods (Saupe 1988, Sect. 2.3.2) and the associated problems.

Such processes are more easily generated by taking a diversion through frequency space. One makes use of the fact that the spectrum $S(\omega)$ of fractal Brownian motion has the form

$$S(\omega) \propto \frac{1}{\omega^\beta}, \qquad \beta = 2H + 1.$$

Therefore, the Fourier transform of a realization of fractal Brownian motion is given by

$$f(\omega) = \mathrm{N}(0, \frac{C}{\omega^{\beta/2}}) + \mathrm{iN}(0, \frac{C}{\omega^{\beta/2}}),$$

with some constant $C$. Based on this formula, write a program for an algorithm to simulate fractal Brownian motion.

Compare your results with the discussion in Sect. 2.4.3 of Saupe (1988).

**5.6.** AR($p$) processes. Decomposition into oscillators and relaxators:
An AR($p$) process in a vector space $V$

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \eta_t \qquad \eta_t \sim \mathrm{WN}(0, \sigma^2)$$

has the characteristic polynomial

$$\alpha(z) = 1 - \sum_{k=1}^{p} \alpha_k z^k .$$

The characteristic polynomial has $k$ zeros in $\mathbb{C}$, which are either complex conjugated pairs or real.

Represent the process as an AR(1) process $Z_t$ in the vector space $V^n$ and find a base transformation such that the components of $Z_t$ with respect to this base are either univariate (each component by itself) or bivariate (two components) AR(1) processes. What is the relation to the zeros (absolute value and phase) of the characteristic polynomial?

Write a program which generates realizations of $AR(2)$ processes and makes them visible and audible, and where the absolute value and the phase of the complex zeros (oscillator) or the values of the two real zeros (relaxators) may be varied interactively.

## Chapter 6

**6.1.** The Hamiltonian operator for a spin-1/2 particle in a magnetic field $\boldsymbol{B}$ reads

$$H = -\mu_{\mathrm{B}} \boldsymbol{\sigma} \cdot \boldsymbol{B} ,$$

where $\boldsymbol{\sigma}$ is the Pauli spin operator and $\mu_{\mathrm{B}}$ the Bohr magneton. For the $z$-axis along the direction of the magnetic field, determine the density matrix

$$\varrho = \frac{1}{Z} \mathrm{e}^{-\beta H} , \qquad Z \equiv \mathrm{Tr}\, \mathrm{e}^{-\beta H} ,$$

first, in the basis where $\sigma_z$ is diagonal and, second, in the basis where $\sigma_x$ is diagonal. Evaluate the expectation value $E = \langle H \rangle = \mathrm{Tr}\,(\varrho H)$.

**6.2.** This problem concerns the ideal Fermi gas of spin-1/2 particles. We introduce the following notations:

1. $V$ denotes the volume, $n \equiv N/V$ is the particle density, $\mu$ is the chemical potential, and $\mu_0$ is the chemical potential at $T = 0$.

2. The Fermi distribution:

$$f(\varepsilon) = \frac{1}{\mathrm{e}^{\beta(\varepsilon-\mu)} + 1}.$$

3. $D(\varepsilon)$ denotes the single-particle density of states, i.e., for an arbitrary function $g(\varepsilon)$ of the single-particle energy $\varepsilon = \varepsilon(\boldsymbol{p}) = \boldsymbol{p}^2/2m$ we have

$$\int 2\frac{V\mathrm{d}^3 p}{(2\pi\hbar)^3}\, g(\varepsilon(\boldsymbol{p})) \equiv \int_0^\infty \mathrm{d}\varepsilon\, D(\varepsilon)\, g(\varepsilon).$$

(a) Determine the chemical potential $\mu$ and the internal energy $E$ of an ideal Fermi gas at low temperatures up to order $T^4$.

(b) Show that the heat capacity $C_V$ of an ideal Fermi gas at low temperatures is given by the expression

$$C_V = \frac{1}{3}\pi^2 k_\mathrm{B}^2 T D(\mu_0).$$

(c) The energy of an electron in an external magnetic field $\boldsymbol{B}$ is given by

$$\varepsilon = \frac{1}{2m}\boldsymbol{p}^2 - \boldsymbol{\mu}_\mathrm{e} \cdot \boldsymbol{B}$$

(to avoid a confusion with the chemical potential $\mu$ we denote the magnetic moment of the electron by $\boldsymbol{\mu}_\mathrm{e}$). The paramagnetic susceptibility $\chi$ of an ideal electron gas due to the spin of the electron in the limit of a weak magnetic field is given by

$$\chi = \lim_{B\to 0}\left(\frac{M}{VB}\right),$$

where $V$ denotes the volume and $M$ the magnetization. Show that for arbitrary temperatures the following formula holds:

$$\chi = \frac{\mu_\mathrm{e}^2}{V}\int_0^\infty \mathrm{d}\varepsilon\, f(\varepsilon) D'(\varepsilon).$$

From this expression derive the following limiting cases:

$$\lim_{T\to 0}\chi = \frac{3}{2}n\mu_\mathrm{e}^2/\mu_0, \quad \chi \overset{T\to\infty}{\longrightarrow} n\mu_\mathrm{e}^2/k_\mathrm{B}T.$$

(d) Write a program to compute the internal energy $E$ of the ideal Fermi gas as a function of temperature and compare the result with the approximate solution of part (a).

**6.3.** Show that an ideal Bose gas in two dimensions has no Bose–Einstein condensation.

**6.4.** Let $n_j$ be the number of photons in a certain radiation mode $j = (\boldsymbol{k}, \alpha)$. Prove that

$$\langle N_j \rangle = \frac{1}{\mathrm{e}^{\beta\hbar\omega_j} - 1}$$

$$\frac{\mathrm{Var}(N_j)}{\langle N_j \rangle^2} = 1 + \frac{1}{\langle N_j \rangle}.$$

Furthermore, let $p_j(n_j)$ be the probability of finding exactly $n_j$ photons in the mode $j$. Show that $p_j$ may be written as

$$p_j(n_j) = \frac{\langle n_j \rangle^{n_j}}{\big(1 + \langle n_j \rangle\big)^{n_j+1}}.$$

Interpret these results.

**6.5.** We consider an ultra-relativistic and highly degenerate ideal Fermi gas with spin 1/2. For this purpose we take the approximate single-particle dispersion relation in the form

$$\varepsilon(\boldsymbol{p}) = c\,|\boldsymbol{p}|$$

($c$ is the velocity of light). Determine the chemical potential $\mu$ and the internal energy $E$ for low temperatures up to order $T^2$, and from this result derive the heat capacity up to order $T$. By what factor does this result differ from that for the nonrelativistic case.

**6.6.** We consider a semiconductor with a completely filled valence band and an empty conduction band at $T = 0$. Let $E_\mathrm{g}$ be the energy gap between the uppermost level of the valence band and the lowest level of the conduction band. At finite temperatures, electrons will be excited from the valence band into the conduction band. The conductivity of the semiconductor results both from the excited electrons (in the conduction band) and the holes (in the valence band). We denote the density (number per unit volume) of electrons in the conduction band by $n$ and the density of holes in the valence band by $p$. Furthermore, we assume that these electrons and holes behave like free particles with effective masses $m_\mathrm{e}$ and $m_\mathrm{h}$, respectively. If we define the uppermost level of the valence band as the energy zero, we find the following approximate single-particle dispersion relations:

$$\varepsilon(\boldsymbol{p}) = \begin{cases} E_\mathrm{g} + \frac{1}{2m_\mathrm{e}}\,\boldsymbol{p}^2 & \text{for electrons in the conduction band,} \\ -\frac{1}{2m_\mathrm{h}}\,\boldsymbol{p}^2 & \text{for holes in the valence band.} \end{cases}$$

We also assume that $E_g - \mu \gg k_B T$ and $\mu \gg k_B T$. Show that under these conditions

$$n = p = 2\left\{\frac{2\pi\sqrt{m_e m_h}k_B T}{h^2}\right\}^{3/2} e^{-E_g/2k_B T}.$$

From this result, estimate the density $n$ for $E_g = 0.7\,\mathrm{eV}$ at room temperature by assuming that $m_e \approx m_h \approx m$ ($m$ denotes the mass of free electrons).

**6.7.** Dispersion relations in a solid:

(a) Suppose certain lattice vibrations in a Debye solid to have a dispersion relation of the form

$$\omega = Ak^s$$

($\omega$ is the frequency and $k$ the wave number of the lattice vibrations; $A$ is a constant). Show that the corresponding contribution to the specific heat at low temperatures is proportional to $T^{3/s}$.

(b) Show that the contribution of phonons with the dispersion relation

$$\omega = Ak$$

in an $n$-dimensional Debye solid at low temperatures is proportional to $T^n$.

**6.8.** Consider the specific heat of a solid as a function of temperature: $C_V = C_V(T)$. We have $C_V(0) = 0$ and for $T \longrightarrow \infty$ the specific heat $C_V(T)$ tends towards $C_V(\infty) = 3Nk_B$ (the Dulong–Petit law). Compute the area $\mathcal{A}$ enclosed by the graph of the function $C_V(T)$ and a line parallel to the $T$-axis through $C_V(\infty)$:

$$\mathcal{A} = \int_0^\infty \mathrm{d}T\ \{C_V(\infty) - C_V(T)\}.$$

What is the physical significance of this area $\mathcal{A}$?

**6.9.** Compute the contribution of the internal degrees of freedom (vibrations and rotations) to the heat capacity of the gas $NH_3$ at 300 K. Use the following data: Principal moments of inertia:

$$I_1 = 4.44 \times 10^{-40}\,\mathrm{cm^2 g},$$
$$I_2 = I_3 = 2.816 \times 10^{-40}\,\mathrm{cm^2 g}.$$

Normal modes:

$$\bar{\nu}_1 = \bar{\nu}_2 = 3336\,\mathrm{cm^{-1}}, \qquad \bar{\nu}_3 = \bar{\nu}_4 = 950\,\mathrm{cm^{-1}},$$
$$\bar{\nu}_5 = 3414\,\mathrm{cm^{-1}}, \qquad \bar{\nu}_6 = 1627\,\mathrm{cm^{-1}}.$$

The frequencies of the normal modes are given in terms of inverse wavelengths: $\bar{\nu} \equiv \lambda^{-1} = \nu/c$ ($c$ is the velocity of light).

**6.10.** Consider an ideal Bose gas of particles with one internal degree of freedom. For simplicity, assume that apart from the ground state (energy 0) each particle may be in exactly one excited state (energy $\varepsilon^*$). Hence, taking into account the translational degrees of freedom, the single-particle dispersion relation reads

$$\varepsilon(\boldsymbol{p}) = \begin{cases} \boldsymbol{p}^2/2m & \text{for the ground state} \\ \boldsymbol{p}^2/2m + \varepsilon^* & \text{for the excited state.} \end{cases}$$

Write down an implicit equation for the condensation temperature $T_\mathrm{c}$ of the gas. Using this equation, show that for $\varepsilon^*/k_\mathrm{B}T_\mathrm{c}^0 \gg 1$ the following relation holds

$$\frac{T_\mathrm{c}}{T_\mathrm{c}^0} \approx 1 - \frac{2}{3\zeta(3/2)}\,\mathrm{e}^{-\varepsilon^*/k_\mathrm{B}T_\mathrm{c}^0},$$

where $T_\mathrm{c}^0$ denotes the condensation temperature of the ideal Bose gas *without* internal degrees of freedom.

**6.11.** Consider an ideal gas of $N$ diatomic molecules with an electric dipole moment $\boldsymbol{\mu}$. Ignore the induced dipole moments and assume that the electric field acting on the molecules is equal to a homogeneous, constant, external field $\boldsymbol{E}$. Show that the absolute value of the electric polarization $\boldsymbol{P}$ in thermal equilibrium is given by

$$P = \frac{N}{V}\mu\left\{\coth\left(\frac{\mu E}{k_\mathrm{B}T}\right) - \frac{k_\mathrm{B}T}{\mu E}\right\}$$

($E \equiv |\boldsymbol{E}|, \mu \equiv |\boldsymbol{\mu}|$). In addition, prove that for $\mu E \ll kT$ the dielectric constant $\varepsilon$ of the gas is approximately given by

$$\varepsilon = 1 + 4\pi\frac{N}{V}\frac{\mu^2}{3k_\mathrm{B}T}.$$

## Chapter 7

**7.1.** Figure P.2 shows the so-called Diesel circle on a $p$–$V$ diagram. The associated process will be performed with an ideal gas as a working substance. Let $Q_\mathrm{H}$ be the amount of heat absorbed and $Q_\mathrm{L}$ the amount of heat emitted. Determine the efficiency

$$\eta_\mathrm{Diesel} \equiv 1 - \frac{Q_\mathrm{L}}{Q_\mathrm{H}}$$

of the Diesel process. Represent the Diesel process on a $T$–$S$ diagram.

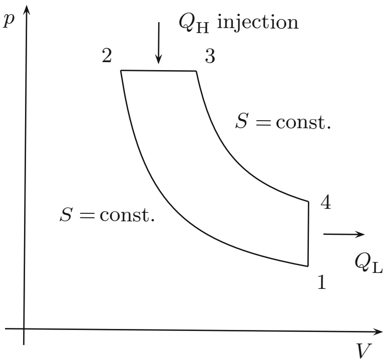


**Fig. P.2** The $p$–$V$ diagram for a diesel process

**7.2.** This problem examines the Joule–Thomson experiment, which is performed with a real gas. The molecules of the gas are subject to a pair interaction described by the potential $V_2(r)$, $r = |\boldsymbol{r}_i - \boldsymbol{r}_j|$.

(a) Show that for the Joule–Thomson experiment the change in temperature as a function of pressure has in general the following form:

$$\delta \equiv \left(\frac{\partial T}{\partial p}\right)_H = \frac{1}{C_p}\left[T\left(\frac{\partial V}{\partial T}\right)_p - V\right].$$

$H$ denotes the enthalpy and $C_p$ the heat capacity for fixed pressure. $\delta$ is called the Joule–Thomson coefficient. What is the physical significance of this coefficient?

(b) We now treat the real gas up to the second order in a virial expansion (cf. Problem 3.6). Show that in this approximation the Joule–Thomson coefficient may be expressed in terms of the second virial coefficient $a_2(T)$ in the following way:

$$\delta = \frac{N}{C_p}\left(T\,\frac{\mathrm{d}B}{\mathrm{d}T} - B\right),$$

where

$$B(T) \equiv \lambda^3 a_2(T) = 2\pi \int_0^\infty \left(1 - \mathrm{e}^{-V_2(r)/kT}\right) r^2\,\mathrm{d}r.$$

$N$ denotes the particle number and $\lambda$ the thermal wavelength.

(c) Compute the Joule–Thomson coefficient for the following cases:

1. $$V_2(r) = \frac{\alpha}{r^n}, \qquad \alpha = \text{const.} > 0, \quad n > 0$$

2.
$$V_2(r) = \begin{cases} +\infty, \ r < a \\ -u_0 \ , \ a < r < b, \quad u_0 > 0 \\ 0, \qquad b < r. \end{cases}$$

Discuss the temperature dependence of $\delta$ for both cases and interpret the result with respect to the Joule–Thomson experiment.

## Chapter 8

**8.1.** Show that for the case of a *known* expectation value $\mu = \langle X \rangle$ of a real random variable $X$ the quantity

$$\hat{\theta}_3(X_1, \ldots, X_N) = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2$$

is an unbiased estimator for the variance.

**8.2.** Let $X$ be a Gaussian real random variable with density

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

Determine the maximum likelihood estimator for the parameters $\mu$ and $\sigma^2$ for a sample of size $N$.

**8.3.** Estimate the expectation value $n$ times for given data $y_1, \ldots, y_N$ and determine the one-sigma confidence interval. Show that the true value is inside the confidence interval for only roughly two third of these $n$ cases

**8.4.** In Sect. 8.5 we derived the least squares estimator for a general parametric model of the form

$$f(x \mid a_1, \ldots, a_M) = \sum_{\alpha=1}^{M} a_\alpha X_\alpha(x)$$

using the singular value decomposition of the $N \times M$ matrix

$$\mathsf{K}_{i\alpha} = \frac{1}{\sigma_i} X_\alpha(x_i), \quad i = 1, \ldots, N, \quad \alpha = 1, \ldots, M.$$

($N$ denotes the number of data pairs and $M$ the number of parameters of the model). Consider the special case of a linear model with $M = 2$:

$$f(x \,|\, a_1, a_2) = a_1 x + a_2.$$

Write down the matrix $\mathsf{K}$ explicitly and construct the matrix $\mathsf{D} \equiv \mathsf{K}^T\mathsf{K}$ and the pseudoinverse $\mathsf{K}^+ \equiv (\mathsf{K}^T\mathsf{K})^{-1}\mathsf{K}^T$. Determine the least squares estimator for the parameters $a_1$ and $a_2$. In addition, determine the base vectors $\mathbf{v}_\gamma$ and $\mathbf{u}_\gamma$ introduced in Sect. 8.5 as well as the singular values $w_\gamma$ ($\gamma = 1, 2$) for the decomposition

$$\mathsf{K} = w_1\, \mathbf{u}_1 \otimes \mathbf{v}_1 + w_2\, \mathbf{u}_2 \otimes \mathbf{v}_2.$$

Try again to visualize again the geometric meaning of this construction.

## Chapter 9

**9.1.** Filters:
A filter may be regarded as a mapping from one time series onto another. A filter is called stable if this mapping is bounded, i.e., if the image of a time series with finite norm has also finite norm. It is called invertible if the mapping is injective. How are stability and invertability related to the positions of zeros and poles?

**9.2.** Simulate a time series, say, an AR(2) process (see Sect. 5.9.4, Example 3) of $N = 1024$ data points, and determine the absolute values and phases of the Fourier transform. Replace the phases by random values, equally distributed in $[0, 2\pi]$, and subject to the constraint derived in Sect. 9.1, so that the time series, obtained by the inverse Fourier transform, is real. What a meaning has this time series?

**9.3.** Time-dependent spectrogram:
Generate three AR(2) processes (see Sect. 5.9.4), each of length $4,098$ and with damping rate 200, and with period 5, 10, and 30 respectively. Concatenate these time series and determine the time-dependent spectrogram of the result. Furthermore: Record some spoken words or other sounds and discuss the time-dependent spectrograms.

**9.4.** Generate a harmonic wave with 8 and with 8.5 cycles within $1,024$ data points. Calculate the periodogram and visualize it. Why do the second periodogram show sidelobes, the first one from a time series with a whole number of cycles not? Interpret the time series as a multiplication of an infinite time series and a boxcar window, and calculate the Fourier transform of it as the convolution of the Fourier transforms of the factors.

## Chapter 10

**10.1.** Write a program, by which, for a given realization of an AR(2) process, the parameters of the model are estimated with the Monte Carlo Markov Chain method as explained in Sect. 10.3. Generate the sample by simulating an AR(2) model with period $T = 20$ and damping rate $\tau = 50$ (see Sect. 5.9.4, Example 3).

**10.2.** Estimate coefficients and frequency of a model $y(t) = a_1 \sin(\omega t) + a_2 \cos(\omega t) + \sigma\eta(t)$, with $\eta(t) \sim \mathrm{WN}(0, 1)$. Use the Gibbs sampler for the estimation of the coefficients $a_1, a_2$, and the Metropolis-Hastings method for the estimation of the frequency $\omega$.

## Chapter 11

**11.1.** You may gain some experience with the regularization method as follows:

(a) Least squares estimator with ill-conditioned matrices:
Use a bimodal spectrum as given in (11.78) and generate a set of data as outlined in Sect. 11.4.4. Reconstruct the spectrum with the least squares method and estimate the confidence intervals of the estimated values of $h_\alpha = h(\tau_\alpha)$, $\alpha = 1, \ldots, M$ for M = 10,30,50. Correlate the growth of the size of these intervals with the change in the condition of the matrix with elements $K'(\omega_i, \tau_\alpha)$.

(b) Choose a specific energy functional and estimate the spectrum for a given regularization parameter $\lambda$. Plot the result including the standard confidence intervals. Discuss the dependence of the result on the regularization parameter $\lambda$.

(c) Determine the regularization parameter using a particular strategy. Do this not only for a single generated data set, but for several different sets and get an impression of the variation of $\lambda$. Estimate the mean and the variance of the distribution of $\lambda$. Repeat the same with some other strategy for determining the regularization parameter and compare the means and variances.

(d) Solve parts (b) and (c) for a different energy functional. Examine the changes and try to understand them.

**11.2.** Generate a random $N \times M$ matrix and construct from it, by manipulating the singular values, a well-conditioned and an ill-conditioned matrix $K_{i,j}$. Use these matrices to define a linear regression model, simulate data by choosing some parameters and try to estimate these parameters, given the simulated data. Study the estimates and their confidence intervals in dependence of the condition of the matrix $K_{i,j}$ .

**11.3.** Simulate a hidden Markov model and apply the Viterbi algorithm to the observations $\{y(t), t = 1, \ldots, N\}$. Calculate the estimate based on $\{y(t'), t' = 1, \ldots, t\}$ and the estimate based on $\{y(t'), t' = 1, \ldots, N\}$.

**11.4.** Write a subroutine for a Kalman filter and a state space model (for a univariate variables $x$ and $y$ only) and use these to show that for the observation equation $y(t) = x(t) + \epsilon(t)$ the Kalman filter estimate is, for given $\{y(t), t = 1, \ldots, N\}$, a better approximation to $\{x(t), t = 1, \ldots, N\}$ than the naive simple estimate $\{y(t), t = 1, \ldots, N\}$ (see Fig. 11.12).

## Chapter 12

**12.1.** Formulate a hidden Markov model with two hidden states, simulate some data points, and estimate the parameters, given the data.

**12.2.** A casino manager uses from time to time an unfair die, so the casino can be in two different states (fair die, unfair die). Observable are only the points after each throw. Formulate a hidden Markov model for this situation. Simulate data and estimate the parameters of the model, given the data. Plot the probability, in dependence on time, that the casino is in the unfair state.

## Chapter 13

**13.1.** For the test statistic 'mean' the standard deviation can also be estimated analytically. Use normally distributed random numbers with zero mean and standard deviation equal 2 as data and determine the bootstrap results for mean and standard deviation. Vary the size of the bootstrap sample from 50 to 500 and compare the results.

**13.2.** Two different sorts of scripts can be generated in the following way: One defines a vocabulary, e.g., the numbers from 1 to 10, and for every sort a different probability distributions for every word of the vocabulary. Given a collection of $n_1$ scripts of sort 1 and $n_2$ scripts of sort 2, estimate the probability of each word of each sort (learning with teacher). Use the Bayes' classificator to determine to which sort newly generated scripts belong. Try to apply this method to realistic scripts.

**13.3.** Write a program for estimating with the help of the EM algorithm the mean and standard deviation of three clusters of normally distributed random two-dimensional data.

# Hints and Solutions

## Chapter 2

**2.1.** (a) Use $A \cap B = \overline{\bar{A} \cup \bar{B}}$.
(b) Write $B$ as the disjoint union of $A$ and something else.
(c) Write $A \cup B$ and $A$ as a disjoint union.

**2.2.** See Sect. 2.2.

**2.3.** The generating function is $G(z) = \mathrm{e}^{\lambda(z-1)}$.

**2.4.** Simulation of the occurrence of an event with probability $q, 0 \leq q \leq 1$ may be performed as follows: Generate a random number $r$, uniformly distributed in [0, 1], and check whether $r \leq q$. If that is the case, the event has occurred.

**2.5.** See Sect. 2.2.

**2.6.** Express the difference $S - S_1 - S_2$ in terms of the relative entropy and use the inequality $x \ln x - x \ln y - x + y \geq 0$.

**2.7.** $\varrho(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-x^2/2\sigma^2}$.

**2.8.** (a) $\varrho_Y(y) = \frac{1}{n} y^{1/n-1}$,
(b) $\varrho_Y(y) = \mathrm{e}^{-y}$,
(c) $\varrho(y_1, y_2) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y_1^2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y_2^2}$.

**2.9.** (a) $P(k) = \frac{\lambda_1+\lambda_2}{k!} \mathrm{e}^{-(\lambda_1+\lambda_2)}$,
(b) $G_I(z) = \mathrm{e}^{\lambda_1(z-1)}$, $G_K(z) = \mathrm{e}^{(\lambda_1+\lambda_2)(z-1)}$.

**2.10.** Use $\varrho_Z(z) = \int \mathrm{d}x_1\, \mathrm{d}x_2\, \delta(z - x_1 - x_2)\varrho(x_1, x_2)$.

**2.11.** Use the characteristic function of the Cauchy distribution.

**2.12.** $g(y) = \beta y - 1 \ln \beta y$; see the discussion in Sect. 3.3.

**2.13.** For $\beta \neq \alpha$ the density function shrinks to zero.

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1, © Springer-Verlag Berlin Heidelberg 2012

**2.14.** Use a minimization subroutine in the computer language you prefer.

**2.15.**
$$(DT_2)\Phi_i(x) = 2\sqrt{2}\int \mathrm{d}y\ \varrho^*(\sqrt{2}x - y)\Phi_i(y).$$

Use partial integration to obtain

$$(DT_2)\Phi_i(x) = -\frac{1}{\sqrt{2}}\frac{\mathrm{d}}{\mathrm{d}x}(DT_2)\Phi_{i-1}(x).$$

Then starting with $(DT_2)\Phi_0(x) = 2\Phi_0(x)$, the rest follows by induction.

**2.16.** See Sect. 2.6.

## Chapter 3

**3.1.** The particles are considered to be indistinguishable. See Sect. 3.1.

**3.2.** Classical result: $Z_N = Z_1^N = (k_\mathrm{B}T/\hbar\omega)^N$. Quantum mechanical result: $Z_N = Z_1^N = (2\sinh(\hbar\omega/2k_\mathrm{B}T))^{-N}$.

**3.3.** Use $(\partial x/\partial y)_z = -(\partial z/\partial y)_x(\partial z/\partial x)_y^{-1}$ and apply this to $(\partial V/\partial p)_S$.

**3.4.** Write the entropy of the entire system,

$$S = S_A + S_B,$$

as a function of the internal energies $E_A$ and $E_B$, the volume $V$, and the number of particles $N_A$, $N_B$. Find the maximum of the entropy $S$ under the supplementary conditions

$$E_A + E_B = E = \text{const.} \qquad \text{and} \qquad N_A + 2N_B = \text{const.}$$

**3.5.** Proceed from

$$\frac{\partial V(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)}{\partial \boldsymbol{r}_i} = \frac{\partial V_n(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n)}{\partial \boldsymbol{r}_i} + \frac{\partial}{\partial \boldsymbol{r}_i}\sum_{j=n+1}^{N} V_2(\boldsymbol{r}_i,\boldsymbol{r}_j),$$

multiply with $1/\lambda^{3N}(N-n)!$ and integrate over $\boldsymbol{r}_{n+1},\ldots,\boldsymbol{r}_N$.

**3.6.** $(pV/Kk_\mathrm{B}T) = 1 - (2\pi/3)n\left(\mathrm{e}^{u_0/k_\mathrm{B}T}(b^3 - a^3) - b^3\right)$.

**3.7.** Handle the problem according to classical statistics. The interaction energy $\Phi$ of the two dipoles may be written as

$$\Phi = -\frac{1}{R^5}\left\{3(\boldsymbol{m}\cdot\boldsymbol{R})(\boldsymbol{m}'\cdot\boldsymbol{R}) - \boldsymbol{R}^2(\boldsymbol{m}\cdot\boldsymbol{m}')\right\}.$$

## Chapter 4

**4.1.** (a) $S(E) = -k_{\mathrm{B}} N \left[ (\frac{1}{2} + \frac{E}{\varepsilon N}) \ln (\frac{1}{2} + \frac{E}{\varepsilon N}) + (\frac{1}{2} - \frac{E}{\varepsilon N}) \ln (\frac{1}{2} - \frac{E}{\varepsilon N}) \right]$ with $\varepsilon = \gamma B$. $S(E)$ has a maximum at $E = 0$.

(b) $k_{\mathrm{B}} T/\varepsilon = \ln \left( \frac{1-2E/\varepsilon N}{1+2E/\varepsilon N} \right)^{-1}$,
$E = -\frac{N\varepsilon}{2} \tanh (\varepsilon/2k_{\mathrm{B}}T)$,
$C = N k_{\mathrm{B}} (\varepsilon/2k_{\mathrm{B}}T)^2 (\cosh(\varepsilon/2k_{\mathrm{B}}T))^{-2}$.

(c) $M = \frac{N\gamma}{2} \tanh (\varepsilon/2k_{\mathrm{B}}T)$.

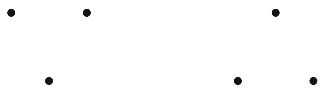
(d) There must be a finite upper limit for the energy of the system as is the case for spin systems. Nuclear and electron spin systems can be prepared in such a way that they have negative temperature (see, e.g., Purcell and Pound 1951; Abragam and Proctor 1957, 1958; Klein 1956; Ramsey 1956).

**4.2.** (a) Given $H$, the magnetization $M$ has to be determined from $\partial G/\partial M = 2L_2 M + 4L_4 M^3 - H = 0$ (see Sect. 4.6).

(b) $\beta = 1/2, \gamma = \gamma' = 1, \delta = 3$.

(c) $\alpha = \alpha' = 0$, $\Delta C_H = -T_c l_{21}^2 / 2 l_{40}$.

**4.3.** The cliques of e.g. $\mathcal{C}_3$ are



There are no cliques of type $\mathcal{C}_4$.

**4.4.** If $x_i$ can be $1, 2$ or $3$, then in the auto-model $J_{ij} x_i x_j$ can take the values $1, 2, 3, 4, 6, 9$ times $-J_{ij}$. In the 3-level model, $J_{ij} x_i x_j$ can take the values $J_c, -J_c$, and $J_c$ may depend on the relative positions of $i$ and $j$.

**4.5.** The rejection method: One has to look for a density $\varrho'(x)$ for which random variables are easily generated and such that with a constant $C$

$$C\varrho'(x) \geq \varrho(x)$$

holds. Generate pairs $\{(x_i, y_i)\}$ so that $x_i$ is chosen according to the density $\varrho'(x)$, and $y_i$ according to a uniform density in $[0, C\varrho'(x_i)]$, and accept each of them if $y_i \leq \varrho(x_i)$. Then the accepted $\{x_i\}$ possess the desired density.

Take here $\varrho'(x)$ as the standard normal density. With $\varrho(x) = \frac{1}{Z} \mathrm{e}^{-y^4/4}$ then choose e.g. $C\varrho'(x) = \frac{1}{Z} \mathrm{e}^{1/4} \mathrm{e}^{-x^2/2}$.

**4.6.** See Sect. 4.5.

**4.7.** See Sect. 4.5.

## Chapter 5

**5.1.** Generate a sample of normally distributed random variables.

**5.2.** See Sect. 5.5.

**5.3.** See Sect. 5.5.

**5.4.** $\varrho(x,t) = \mathrm{N}(0, \sigma^2/(1-\alpha^2))$, $\varrho(x,t|x',t-1) = \mathrm{N}(x', \sigma^2)$, see Sect. 5.1.

**5.5.** Generate pairs $(a_k, b_k)$ of normal random numbers with zero mean and variance $\propto k^{-\beta/2}$. Take $\{a_k + \mathrm{i}b_k\}$ as the Fourier components of $X(t)$. In (Saupe 1988) the polar decomposition of a complex random variable is used.

**5.6.** See Sect. 5.9.

## Chapter 6

**6.1.** $\langle H\rangle = \mu_{\mathrm{B}} B \tanh(\mu_{\mathrm{B}} B/k_{\mathrm{B}} T)$.

**6.2.** (a) $\mu = \mu_0 \left(1 - \frac{\pi^2}{12}(k_{\mathrm{B}}T/\mu_0)^2 - \frac{\pi^4}{80}(k_{\mathrm{B}}T/\mu_0)^4 + \ldots\right)$,

$E = \frac{3}{5}N\mu_0 \left(1 + \frac{5\pi^2}{12}(k_{\mathrm{B}}T/\mu_0)^2 - \frac{\pi^4}{16}(k_{\mathrm{B}}T/\mu_0)^4 + \ldots\right)$.

(b) Use $C_V = (\partial E/\partial T)_{V,N}$.

(c) Determine $M = \mu_e(N_+ - N_-)$ in terms of $D(\varepsilon)$ and $f(\varepsilon)$.

**6.3.** Show that the particle number $N$ is given by

$$N = \frac{A}{\lambda^2}\int_0^\infty \frac{\mathrm{d}x}{z^{-1}\mathrm{e}^x - 1}.$$

$A$ denotes the area of the gas and $\lambda$ the thermal wavelength. Examine the behavior of the integral for $z \to 1$ and compare this with the behavior of $h_{\frac{3}{2}}(z)$ in that limit.

**6.4.** Determine the probability of the state $\{n_j\}$: $p(\{n_j\}) = \prod_j p_j(n_j)$.

One obtains that the ratio $p(n_j+1)/p(n_j) = \mu/(1+\mu)$, $\mu = \langle N_j\rangle$ is independent of $n_j$ in contrast to the classical Poisson distribution (Boltzmann gas). Also, the relative variance $\mathrm{Var}(N_j)/\langle N_j\rangle^2$ of the ideal Bose gas is larger than that for a classical Boltzmann gas.

**6.5.** $C_V = (2\pi^{7/3}k_{\mathrm{B}}/3^{1/3}\mathrm{h}cn^{1/3})Nk_{\mathrm{B}}T$, where $N$ denotes the particle number and $n$ the particle density.

**6.6.** Calculate $N = n + p$, $n$ and $p$ in terms of $\mu$. Then find $\mu$ and substitute. $n = p = 3.3 \times 10^{13}\mathrm{cm}^{-3}$.

**6.7.** Find $g(\omega)$ defined by $g(\omega)\mathrm{d}\omega = \mathrm{d}^n k$ ($n$ = dimension of the solid) and study the energy $E$ as integral over the frequency contributions.

**6.8.** Write $C_V$ as $\partial E/\partial T$ and discuss the expansion of $E(T)$ for large $T$. The area $\mathcal{A}$ corresponds to the zero point energy of the solid.

**6.9.** $C_V^{\text{rot}} = \frac{3}{2}Nk_{\text{B}}$, $C_V^{\text{vib}} \approx 0.47Nk_{\text{B}}$.

**6.10.** The implicit equation reads

$$N = \frac{V}{\lambda^3(T_{\text{c}})}\left(h_{\frac{3}{2}}(1) + h_{\frac{3}{2}}(\mathrm{e}^{-\varepsilon^*/k_{\text{B}}T_{\text{c}}})\right).$$

The limit $\varepsilon \to \infty$ leads to an equation for $T_{\text{c}}^0$.

**6.11.** Use classical statistics to solve this problem and treat the molecule as a rigid rotator whose axis coincides with the directions of the dipole moment. The following relation holds between the electric displacement $\boldsymbol{D}$, the electric field strength $\boldsymbol{E}$, and the polarization $\boldsymbol{P}$:

$$\boldsymbol{D} = \varepsilon\boldsymbol{E} = \boldsymbol{E} + 4\pi\boldsymbol{P}.$$

## Chapter 7

**7.1.** Proceed from

$$\mathrm{d}Q = \mathrm{d}E + p\,\mathrm{d}V \quad \text{or} \quad \mathrm{d}Q = \mathrm{d}H + V\,\mathrm{d}p$$

and make use of the fact that $C_V = \mathrm{d}E/\mathrm{d}T$ and $C_p = \mathrm{d}H/\mathrm{d}T$ may be considered independent of temperature.

**7.2.** (a) Use that the enthalpy is constant. $\delta$ measures the temperature difference achieved during an expansion. $\delta = 0$ for an ideal gas.

(b) Use the equation of state in the form $V = (Nk_{\text{B}}/p)(T + \frac{N}{V}B(T)V)$ and calculate $(\partial V/\partial T)_p$.

(c) For the computation of $B(T)$ for the first case, make a partial integration and express the remaining integral by a $\Gamma$-function.

## Chapter 8

**8.1.** Use that $\langle (X_i - \mu)^2 \rangle = \text{Var}(X)$.

**8.2.** $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$.

**8.3.** Compare Sect. 8.1.

**8.4.** $$\mathsf{K} = \begin{pmatrix} \frac{x_1}{\sigma_1} & \frac{1}{\sigma_1} \\ \cdots & \cdots \\ \frac{x_N}{\sigma_N} & \frac{1}{\sigma_N} \end{pmatrix}.$$

With $S = \sum_i (1/\sigma_i^2),\ S_x = \sum_i (x_i/\sigma_i^2),\ S_y = \sum_i (y_i/\sigma_i^2),\ S_{xy} = \sum_i (x_i y_i/\sigma_i^2)$ one obtains

$$\hat{a}_1 = \frac{S\,S_{xy} - S_x S_y}{S\,S_{xx} - S_x^2}, \qquad \hat{a}_2 = \frac{S_y\,S_{xx} - S_x S_{xy}}{S\,S_{xx} - S_x^2}.$$

Furthermore $w_{1,2} = \sqrt{(S_{xx} + S \pm E)/2}$ with $E = \sqrt{(S_{xx} - S)^2 + 4S_x^2}$.

## Chapter 9

**9.1.** See Sect. 9.3 and compare with Sect. 5.9.

**9.2.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**9.3.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**9.4.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

## Chapter 10

**10.1.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**10.2.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

## Chapter 11

**11.1.** (a) For determination of the confidence intervals see Sect. 8.5. For the condition of the matrix use the singular value decomposition and determine the ratio of the largest and smallest singular value.

(b) One should observe that the confidence intervals increase with decreasing regularization parameter and that there is an optimum value of the regularization parameter for which there is the best compromise between the reconstruction of the spectrum the and size of the confidence intervals. This optimum value may also be determined by minimization of a discrepancy between the true spectrum and the reconstructed one.

(c) From such Monte Carlo simulations one gets an idea of the power of a strategy for the determination of the regularization parameter. The best strategy is the one that produces a distribution of the regularization parameter with the smallest variance and with a mean closest to the optimum regularization parameter.

(d) The proper choice of the energy functional depends on ones expectations for the spectrum. For spectra like those chosen in (a), the square of the second derivate will be a good choice. For spectra that are piecewise constant and interrupted by jumps, an energy functional depending on the first derivative (second example in Sect. 11.4.3) will do better.

**11.2.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**11.3.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**11.4.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

## Chapter 12

**12.1.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**12.2.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

## Chapter 13

**13.1.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**13.2.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

**13.3.** See http://webber.physik.uni-freiburg.de/HonStatphys.htm

# References

Abragam, A., Proctor, W.: Experiments on Spin Temperature, Phys. Rev. **106**, 160 (1957)

Abragam, A., Proctor, W.: Spin Temperature, Phys. Rev. **109**, 1441 (1958)

Ahlers, G.: Thermodynamics and experimental tests of static scaling and universality near the superfluid transition in He$^4$ under pressure. Phys. Rev. A **8**(1), 530–568 (1973)

Anderson, M., Ensher, J., Matthews, M., Wieman, C., Cornell, E.: Observation of Bose–Einstein condensation in a dilute atomic vapor. Science **269**, 198–201 (1995)

Ashcroft, N.W., Mermin, N.: Solid State Physics. Holt–Saunders, Fort Worth (1976)

Baccala, L.A., Sameshima, K.: Partial directed coherence: a new concept in neural structure determination. Biol. Cybern. **84**, 463–474 (2001)

Baehr, H.: Thermodynamik, 9th edn. Springer, Berlin/Heidelberg (1996)

Balescu, R.: Equilibrium and Nonequilibrium Statistical Mechanics. Wiley, New York (1975)

Barker, J., Henderson, D.: What is a liquid? Understanding the states of matter. Rev. Mod. Phys. **48**, 487 (1976)

Baxter, R.: Exactly Soluble Models in Statistical Mechanics. Academic, New York (1984)

Bender, C.M., Orszag, S.A.: Advanced Mathematical Methods for Scientists and Engineers. McGraw-Hill, New York (1978)

Berne, B., Pecora, R.: Dynamic Light Scattering. Wiley, New York (1976)

Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B **36**, 192–236 (1974)

Binder, K.: Theory and technical aspects of Monte Carlo simulation. In: Binder, K. (ed.) Monte Carlo Methods in Statistical Physics. Springer, Berlin/Heidelberg (1979)

Binder, K.: Applications of the Monte-Carlo Method in Statistical Physics. Springer, Berlin/Heidelberg (1987)

Binder, K.: The Monte Carlo Method in Condensed Matter Physics. Springer, Berlin/Heidelberg (1995)

Binder, K., Heermann, D.: Monte Carlo Simulation in Statistical Physics. Springer, Berlin/Heidelberg (1997)

Bloomfield, P.: Fourier Analysis of Time Series: An Introduction. Wiley, New York (1976)

Brenig, W.: Statistische Theorie der Wärme. Springer, Berlin/Heidelberg (1992)

Breuer, H.-P., Petruccione, F.: Stochastic dynamics of quantum jumps. Phys. Rev. E **52**(1), 428–441 (1995)

Breuer, H., Huber, W., Petruccione, F.: The macroscopic limit in a Stochastic Reaction–Diffusion process. Europhys. Lett. **30**(2), 69–74 (1995)

Brockwell, P.J., Davies, R.A.: Time Series: Theory and Methods. Springer, New York (1987)

Collins, G.: Gaseous BoseEinstein Condensate Finally Observed, Phys. Today **48**, 17-20 (1995)

Dahlhaus, R.: Graphical interaction models for multivariate time series. Metrika **51**, 157–172 (2000)

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1, © Springer-Verlag Berlin Heidelberg 2012

Davis, P., Rabinowitz, P.: Methods of Numerical Integration. Academic, New York (1975)

Dawydow, A.S.: Quantum Mechanics. Pergamon Press, Oxford (1965)

Delves, L., Mohamed, J.: Computational Methods for Integral Equations. Cambridge University Press, Cambridge (1985)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM-Algorithm. J. R. Stat. Soc. B **39**, 1 (1977)

Diu, B., Guthmann, C., Lederer, D., Roulet, B.: Grundlagen der Statistischen Physik. de Gruyter, Berlin (1994)

Domb, C., Green, M.: Phase Transitions and Critical Phenomena, vol. 3. Academic, New York (1974)

Dudley, R.: Real Analysis and Probability. Wadsworth and Brooks/Cole, Pacific Grove (1989)

Efron, B., Tibshirani, R.: An Introduction to Bootstrap. Chapman & Hall, New York (1993)

Eichenauer, J., Grothe, H., Lehn, J.: Marsaglia's lattice test and non-linear congruential pseudo random number generators. Metrika **35**, 241 (1988)

Eichler, M.: Graphical modeling of dynamic relationships in multivariate time series. In: Schelter, B., Winterhalder, M., Timmer, J. (eds.) Handbook of Time Series Analysis. Wiley, New York (2006)

Ellis, R.S.: Entropy, Large Deviations, and Statistcal Mechanics. Springer, Berlin/Heidelberg (1985)

Feller, W.: An Introduction to Probability Theory, vol. 2. Wiley, New York (1957)

Fitzgerald, W.J., Ó Ruanaidh, J.J.K.: Numerical Bayesian Methods Applied to Signal Processing. Springer, New York/Berlin/Heidelberg (1996)

Frodesen, A.G., Skjeggestad, O., Tofte, H.: Probability and Statistics in Particle Physics. Universitätsforlaget, Bergen/Oslo/Tromso (1979)

Gardiner, C.: Handbook of Stochastic Methods. Springer, Berlin/Heidelberg (1985)

Gelb, A. (ed.): Applied Optimal Estimation. MIT-Press, Cambridge (1974)

Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**(6), 721–741 (1984)

Graham, R.L., Knuth, D.E., Patashnik, O.: Concrete Mathematics, 2nd edn. Addison, Reading (1994)

Granger, C.J.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrics **37**, 424–438 (1969)

Griffin, A., Snoke, D., Stringari, S. (eds.): Bose–Einstein Condensation. Cambridge University Press, Cambridge (1995)

Groetsch, C.W.: The Theory of Tikhonovs Regularisation for Fredholm Equations of the First Kind. Pitman, Boston (1984)

Hammersley, J.M., Clifford, P.: Markov Field on Finite Graphs and Lattices. unpublished (1971)

Hartung, J., Elpelt, B., Klösener, K.: Statistik. Oldenbourg, Munich/Vienna (1986)

Heller, P.: Experimental investigations of critical phenomena, Rep. Progr. Phys. **30**(II), 731 (1967)

Hengartner, W., Theodorescu, R.: Einführung in die Monte-Carlo Methode. Carl Hanser, Munich/Vienna (1978)

Herring, C.: Direct exchange. In: Rado, G., Suhl, H. (eds.) Magnetism, vol. 2B. Academic, New York (1966)

Hirschfelder, J., Curtiss, C., Bird, R.: Molecular Theory of Gases and Liquids. Wiley, New York (1954)

Ho, J., Litster, J.: Magnetic equation of state of $CrBr_3$ near the critical point. Phys. Rev. Lett. **22**(12), 603–606 (1969)

Honerkamp, J.: Stochastic Dynamical Systems. VCH, Weinheim (1994)

Honerkamp, J., Römer, H.: Theoretical Physics: A Classical Approach. Springer, Berlin/Heidelberg (1993)

Honerkamp, J., Weese, J.: Determination of the Relaxation Spectrum by a Regularization Method. Macromolecules **22**, 4372–4377 (1989)

Honerkamp, J., Weese, J.: Tikhonovs regularization method for Ill-posed problems: a comparision of different methods. Contin. Mech. Thermodyn. **2**, 17–30 (1990)

Honerkamp, J., Weese, J.: A nonlinear regularization method for the calculation of the relaxation spectrum. Rheol. Acta **32**, 32–65 (1993)

Honerkamp, J., Maier, D., Weese, J.: A nonlinear regularization method for the analysis of photon correlation spectroscopy data. J. Chem. Phys. **98**(2), 865–872 (1993)

Huang, K.: Statistical Mechanics, 2nd edn. Wiley, New York (1987)

Ising, E.: Beitrag zur theorie des ferromagnetismus. Z. Phys. **31**, 253–258 (1925)

Jachan, M., Henschel, K., Nawrath, J., Schad, A., Timmer, J., Schelter, B.: Inferring direct directed-information flow from multivariate nonlinear time series. Phys. Rev. E **80**, 011138 (2009)

Jaynes, E.: Papers on Probability, Statistics and Statistical Physics. Reidel, Dordrecht (1982)

Jiu-li, L., Van den Broeck, C., Nicolis, G.: Stability criteria and fluctuations around nonequilibrium states. Z. Phys. B Condens. Matter **56**, 165–170 (1984)

Julier, S., Uhlmann, J.: A general method for approximating nonlinear transformations of probability distributions. Technical Report, Department of Engineering Science, University of Oxford (1996)

Julier, S.J., Uhlmann, J.K., Durrant-Whyte, H.F.: A new approach for filtering nonlinear systems. In: Proceedings of the American Control Conference, Seattle, vol. 3, pp. 1628–1632 (1995)

Julier, S.J., Uhlmann, J.K., Durrant-Whyte, H.F.: A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Trans. Autom. Control **45**, 477–482 (2000)

Kadanoff, L.: Scaling Laws for Ising Models Near Tc, Physics **2**, 263 (1966)

Kitagawa, G., Gersch, W.: Smoothness Priors Analysis of Time Series. Lecture Notes in Statistics, vol. 116. Springer, New York (1996)

Kittel, C.: Thermal Physics. Freeman, San Francisco (1980)

Klein, M.J.: Phys. Rev. **104**, 589 (1956)

Klitzing, K.V., Dorda, G., Pepper, M.: New method for high-accuracy determination of the fine-structure constant based on quantized hall resistance. Phys. Rev. Lett. **45**, 494–497 (1980)

Kloeden, P., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin/Heidelberg (1995)

Knuth, D.E.: Seminumerical Algorithms, vol. 2: The Art of Computer Programming. Addison, Reading (1969)

Kolmogoroff, A. Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin/Heidelberg (1933)

Kose, V., Melchert, F.: Quantenmaße in der Elektrischen Meßtechnik. VCH, Weinheim (1991)

Kouvel, J., Comly, J.: Magnetic equation of state for nickel near its curie point. Phys. Rev. Lett. **20**(22), 1237–1239 (1968)

Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)

Landau, L.D., Lifschitz, E.M.: Theory of Elasticity. Pergamon, London (1959)

Lehmann, E.: Theory of Point Estimation. Wadsworth & Brooks/Cole, New York (1991)

Lehmer, D.H.: Mathematical methods in large-scale computing units. In: Nr. 141. Proceedings of the 2nd Symposium on Large-Scale Digital Calculating Machinery. Harvard University Press, Cambridge (1951)

Levine, R.D., Tribus, M. (eds.): The Maximum Entropy Formalism. MIT Press, Cambridge (1979)

Lien, W.H., Phillips, N.E.: Low-temperature heat capacities of potassium, rubidium and cesium. Phys. Rev. A **133**(5A), 1370 (1964)

Lütkepohl, H., Krätzig, M.: Applied Time Series Econometrics. Cambridge University Press, Cambridge (2004)

Malek Mansour, M., Van den Broeck, C., Nicolis, G., Turner, J.: Asymptotic properties of Markovian master equations. Ann. Phys. **131**, 283–313 (1981)

Mandelbrot, B.: The Fractal Geometry of Nature. Freeman, San Francisco (1982)

Mandl, F.: Statistical Physics. Wiley, New York (1971)

Mayer, J.E.: The Statistical Mechanics of Condensing Systems. I , J. Chem. Phys. 5, 67 (1937)

McCoy, B., Wu, T.: The Two-Dimensional Ising Model. Harvard University Press, Cambridge (1973)

McQuarrie, D.A.: Statistical Mechanics. Harper & Row, New York (1976)

McQuarrie, D.A.: Quantum Chemistry. Oxford University Press, Oxford (1983)

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087 (1953)

Michels, A., Blaisse, B., Michels, C.: Proc. R. Soc. A **160**, 358 (1937)

Miller, G.F.: In: Delves, L.M., Walsh, J. (eds.) Numerical Solution of Integral Equations. Clarendon, Oxford (1974)

Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.-M.: Wavelet Toolbox, for Use with MatLab. The MathWorks, Natick (1996)

Montroll, E.: Some remarks on the integral equations of statistical mechanics In: Cohen, E. (ed.) Fundamental Problems in Statistical Mechanics, p. 230. North-Holland, Amsterdam (1962)

Montroll, E., West, B.: On an enriched collection of Stochastic processes. In: Montroll, E., Lebowitz, J. (eds.) Fluctuation Phenomena, Chap. 2, p. 62–206. North-Holland, Amsterdam (1987)

Morozov, V.A.: Methods for Solving Incorrectly Posed Problems. Springer, New York (1984)

Nicolis, G., Prigogine, I.: Self-Organization in Non-equilibrium Systems. Wiley, New York (1977)

Niederreiter, H.: Recent trends in random number and random vector generation. Ann. Oper. Res., Proc. 5th Int. Conf. Stoch. Programm. Ann. Arbor. **35**, 241 (1989)

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M.: Identifying true brain interaction from EEG data using the imaginary part of coherency. Clin. Neurophysiol. **115**, 2292–2307 (2004)

Oppenheimer, A., Schafer, R.: Discrete-Time Signal Processing. Prentice-Hall, Englewood Cliffs (1989)

Papoulis, A.: Probability, Random Variables, and Stochastic Processes. McGraw-Hill, New York (1984)

Penzias, A.A.,Wilson, R.: A Measurement of Excess Antenna Temperature at 4080 mc/s., Astrophys. J. **142**, 419-421 (1965)

Prange, R., Girvin, S. (eds.): The Quantum Hall Effect. Springer, Berlin/Heidelberg (1987)

Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical Recipes. Cambridge University Press, Cambridge (2007)

Priestley, M.: Spectral Analysis and Time Series. Academic, New York (1981)

Prigogine, I., Resibois, P.: On the kinetics of the approach to equilibrium, Physica **27**(629) (1961)

Purcell, E., Pound, R.: A Nuclear Spin System at Negative Temperature,, Phys. Rev. **81**, 279 (1951)

Rabiner, L., Gold, B.: Theory and Application of Digital Signal Processing. Prentice-Hall, Englewood Cliffs (1975)

Rahman, A: Correlations in the Motion of Atoms in Liquid Argon, .: Phys. Rev. A **136**, A405 (1964)

Ramsey, N.: Thermodynamics and Statistical Mechanics at Negative Absolute Temperatures, Phys. Rev. **103**, 20 (1956)

Ree, F.H., Hoover, W.G.: Fifth and Sixth Virial Coefficients for Hard Spheres and Hard Disks , J. Chem. Phys. **40**, 939 (1964)

Reichl, L.E.: A Modern Course in Statistical Physics. Arnold, London (1980)

Römer, H., Filk, T.: Statistische Mechanik. VCH, Weinheim (1994)

Roths, T., Maier, D., Friedrich, C., Marth, M., Honerkamp, J.: Determination of the relaxation time spectrum from dynamic moduli using an edge preserving regularization method. Rheol. Acta **39**, 163–173 (2000)

Rubinstein, R.Y.: Simulation and the Monte-Carlo Method. Wiley, New York (1981)

Samorodnitzky, G., Taqqu, M.: Stable Non-Gaussian Random Processes. Chapman & Hall, New York/London (1994)

Saupe, D.: Algorithms for random fractals. In: Peitgen, H., Saupe, D. (eds.) The Science of Fractal Images. Springer, Berlin/Heidelberg (1988)

Schelter, B., Winterhalder, M., Eichler, M., Peifer, M., Hellwig, B., Guschlbauer, B., Lücking, C.H., Dahlhaus, R., Timmer, J.: Testing for directed influences among neural signals using partial directed coherence. J. Neurosci. Method **152**, 210–219 (2006a)

Schelter, B., Winterhalder, M., Timmer, J. (eds.): Handbook of Time Series Analysis. Wiley-VCH, Berlin (2006b)

Schlittgen, R., Streitberg, H.: Zeitreihenanalyse. R. Oldenbourg, Munich (1987)

Schneider, I. (ed.): Die Entwicklung der Wahrscheinlichkeitsrechnung von den Anfängen bis 1933. Wissenschaftliche Buchgesellschaft, Darmstadt (1986)

Shannon, C.: A Mathematical Theory of Communication. Bell Syst. Technol. J. **27**, 379–423 (1948)

Shwartz, A., Weiss, A.: Large Deviations for Performance Analysis. Chapman & Hall, New York/London (1995)

Sommerlade, L., Thiel, M., Platt, B., Plano, A., Riedel, G., Grebogi, C., Timmer, J., Schelter, B. Inference of Granger causal time-dependent influences in noisy multivariate time series. J. Neurosci. Method **203**, 173–185 (2012)

Stanley, H.E.: Phase Transition and Critical Phenomena. Clarendon Press, Oxford (1971)

Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley-Cambridge Press, Wellesley (1996)

Straumann, N.: Thermodynamik. Springer, Berlin/Heidelberg (1986)

Stroud, A.: Approximate Calculation of Multiple Integrals. Prentice Hall, Englewood Cliffs (1971)

Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-Posed Problems. Wiley, New York (1977)

Ursell, H.D.: The evaluation of Gibb's phase-integral for imperfect gases, Proc. Cambridge Phil. Soc. **23**, 685 (1927)

Verlet, L.: Computer "Experiments" on Classical Fluids. II. Equilibrium Correlation Functions, Phys. Rev. **165**, 201 (1968)

van Kampen, N.G.: Stochastic Processes in Physics and Chemistry. North-Holland, Amsterdam (1985)

von Randow, G.: Das Ziegenproblem. Rowohlt, Reinbek (1992)

Voss, H., Timmer, J., Kurths, J.: Nonlinear dynamical system identification from uncertain and indirect measurements. Int. J. Bifurc. Chaos **6**, 1905–1933 (2004).

Wan, E.A., Nelson, A.T.: Neural dual extended Kalman filtering: applications in speechenhancement and monaural blind signal separation. In: IEEE Proceedings in Neural Networks for Signal Processing, Munich, pp. 466–475.

Wegner, F.: Corrections to Scaling Laws, Phys. Rev. B **5**, 4529 (1972)

Wilson, K.G.: Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture, Phys. Rev. B **4**, 31743184 (1971)

Wilson, K.G., Kogut, J.: The renormalization group and the $\epsilon$ expansion. Phys. Rep. **12 C**, 75–200 (1974)

# Index

J. Honerkamp, *Statistical Physics*, Graduate Texts in Physics 38,
DOI 10.1007/978-3-642-28684-1, © Springer-Verlag Berlin Heidelberg 2012