Appendix A
The Two-Body Problem

Consider two mass points with positions $r_i(t) \in \mathbb{R}^3$, $i = 1, 2$ and masses m_i, $i = 1, 2$. It is assumed that the point masses interact through a central potential $U = U(|r_1(t) - r_2(t)|)$ and that external forces are neglected. Thus, the system is closed. The explicit notation of time t is now omitted for the sake of a more compact presentation. Furthermore, we introduce with $p_i \in \mathbb{R}^3$, $i = 1, 2$ the point mass' momentum and the LAGRANGE function [1–5] of the system takes on the form

$$L(r_1, r_2, p_1, p_2) = \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} - U(|r_1 - r_2|) \, . \tag{A.1}$$

The moments p_i are replaced by

$$p_i = m_i \dot{r}_i, \quad i = 1, 2 \, , \tag{A.2}$$

and this yields for the LAGRANGE function (A.1)

$$L(r_1, r_2, \dot{r}_1, \dot{r}_2) = \frac{m_1}{2} \dot{r}_1^2 + \frac{m_2}{2} \dot{r}_2^2 - U(|r_1 - r_2|) \, , \tag{A.3}$$

where $\dot{r}_i$ denotes the time derivative of r_i. We note the following symmetries: the LAGRANGE function is (i) translational invariant, (ii) rotational invariant, and (iii) time invariant. We know from classical mechanics that each symmetry of the LAGRANGE function corresponds to a constant of motion (a quantity that is conserved throughout the motion) and, thus, results in a reduction of the dimensionality of the 12-dimensional phase space.

Let us demonstrate these symmetries: In order to prove translational invariance, we transform to center of mass coordinates which are defined as

$$R = \frac{m_1 r_1 + m_2 r_2}{m_1 + m_2} \qquad \text{and} \qquad r = r_2 - r_1 \, . \tag{A.4}$$

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

It is easily verified that we can express the original coordinates r_1 and r_2 with the help of (A.4) as

$$r_1 = R + \frac{m_2}{m_1 + m_2} r \qquad \text{and} \qquad r_2 = R - \frac{m_1}{m_1 + m_2} r \,. \tag{A.5}$$

The LAGRANGE function (A.3) is rewritten in these new coordinates (A.4) and this yields

$$\begin{aligned} L(r, R, \dot{r}, \dot{R}) &= \frac{M}{2}\dot{R}^2 + \frac{m}{2}\dot{r}^2 - U(|r|) \\ &\equiv L(r, \dot{r}, \dot{R}) \,, \end{aligned} \tag{A.6}$$

where we introduced the total mass M and the *reduced* mass m:

$$M = m_1 + m_2 \qquad \text{and} \qquad m = \frac{m_1 m_2}{m_1 + m_2} \,. \tag{A.7}$$

Obviously, the center of mass coordinate R plays in Eq. (A.6) the role of a *cyclic* coordinate: It does not appear explicitly in the LAGRANGE function. This means that the system is translational invariant and we can deduce from LAGRANGE's equations that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial \dot{R}} L = \frac{\partial}{\partial R} L = 0 \,, \tag{A.8}$$

and the center of mass momentum is conserved. Hence, we obtain that

$$\frac{\partial}{\partial \dot{R}} L = M\dot{R} = \text{const} \,, \tag{A.9}$$

with the solution

$$R(t) = At + B \,, \tag{A.10}$$

where $A, B \in \mathbb{R}^3$ are constants determined by the initial conditions of the problem. As a result, the center of mass moves along a straight line with constant velocity. We collect all results and reformulate the LAGRANGE function (A.6) as

$$\begin{aligned} L(r, \dot{r}) &= \frac{M}{2}A^2 + \frac{m}{2}\dot{r}^2 - U(|r|) \\ &\equiv \tilde{L}(r, \dot{r}) + \text{const} \,. \end{aligned} \tag{A.11}$$

Hence, the problem was reduced to a one-body problem with the LAGRANGE function $\tilde{L}(r, \dot{r})$. In what follows the tilde is omitted and the LAGRANGE function

$$L(r, \dot{r}) = \frac{m}{2}\dot{r}^2 - U(|r|) \,, \tag{A.12}$$

is now studied instead of Eq. (A.11). It is an effective one-body LAGRANGE function.

In the next step the effect of rotational invariance is investigated. Equation (A.12) resembles the LAGRANGE function of a particle of mass m which is located at position r and moves in the field of a central force $F \in \mathbb{R}^3$. This force points to the center of the coordinate system (or points from the center of the coordinate system to the particle). This situation is clearly invariant under a rotation of the coordinate system since $U = U(|r|)$ depends only on the modulus of r. Consequently, r is parallel to F for all $t \geq 0$. In such a case the vector of angular momentum $\ell \in \mathbb{R}^3$ is conserved, since

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell = \mathscr{M} = r \times F = 0 \,, \quad \to \ell = \text{const} \,, \tag{A.13}$$

where $\mathscr{M}$ is the *torque*. This allows us to arbitrarily rotate our coordinate system. We take advantage if this property and rotate it in such a way that

$$\ell = |\ell| e_z \,, \tag{A.14}$$

where e_z is the unit vector in z-direction. Moreover, since the angular momentum ℓ is given by

$$\ell = mr \times \dot{r} = \text{const} \,, \tag{A.15}$$

and because $\ell \| e_z$ we conclude that $r \perp e_z$. This allows us to set $z = 0$ which means that the whole motion of the point mass can be described in the $x - y$ plane. Rotational invariance led us to the conservation of angular momentum and this made the reduction from a three-dimensional problem to a two dimensional problem possible. The particular form (A.12) of the LAGRANGE function suggests the introduction of polar coordinates (ρ, φ):

$$L(\rho, \dot{\rho}, \dot{\varphi}) = \frac{m}{2}\left(\dot{\rho}^2 + \rho^2\dot{\varphi}^2\right) - U(\rho). \tag{A.16}$$

We solve now LAGRANGE's equations (A.6) on the basis of Eq. (A.16): The first step deals with the differential equation for the radius ρ

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial\dot{\rho}}L = m\ddot{\rho} = \frac{\partial}{\partial\rho}L = m\rho\dot{\varphi}^2 - \frac{\partial}{\partial\rho}U(\rho) \,, \tag{A.17}$$

thus

$$m\ddot{\rho} - m\rho\dot{\varphi}^2 + \frac{\mathrm{d}}{\mathrm{d}\rho}U(\rho) = 0\,. \tag{A.18}$$

The differential equation for the angle φ follows from

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial\dot{\varphi}}L = \frac{\mathrm{d}}{\mathrm{d}t}m\rho^2\dot{\varphi} = \frac{\partial}{\partial\varphi}L = 0\,, \tag{A.19}$$

which corresponds to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(m\rho^2\dot{\varphi}\right) = 0\,. \tag{A.20}$$

Equation (A.20) is trivially fulfilled since according to Eq. (A.15)

$$m\rho^2\dot{\varphi} = |\ell| = \text{const}\,. \tag{A.21}$$

However, we solve Eq. (A.21) for $\dot{\varphi}$

$$\dot{\varphi} = \frac{|\ell|}{m\rho^2}\,, \tag{A.22}$$

plug (A.22) into (A.18), and obtain

$$m\ddot{\rho} - \frac{|\ell|^2}{m\rho^3} + \frac{\mathrm{d}}{\mathrm{d}\rho}U(\rho) = 0\,. \tag{A.23}$$

We make use of the time invariance of the LAGRANGE function (A.16). This equation does not explicitly depend on time t and we have

$$\frac{\partial}{\partial t}L = 0\,. \tag{A.24}$$

This implies conservation of energy, as can easily be demonstrated. For this purpose, we regard the total time derivative of the LAGRANGE function L

$$\frac{\mathrm{d}}{\mathrm{d}t}L = \dot{\rho}\frac{\partial}{\partial\rho}L + \ddot{\rho}\frac{\partial}{\partial\dot{\rho}}L + \ddot{\varphi}\frac{\partial}{\partial\dot{\varphi}}L + \frac{\partial}{\partial t}L\,, \tag{A.25}$$

and solve for $\frac{\partial}{\partial t}L$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{\rho}\frac{\partial}{\partial\dot{\rho}}L + \dot{\varphi}\frac{\partial}{\partial\dot{\varphi}}L - L\right) = -\frac{\partial}{\partial t}L = 0\,. \tag{A.26}$$

Consequently

$$\dot{\rho}\frac{\partial}{\partial\dot{\rho}}L+\dot{\varphi}\frac{\partial}{\partial\dot{\varphi}}L-L=\text{const}\,, \tag{A.27}$$

which states the conservation of energy. We evaluate this expression with the help of Eq. (A.16). We obtain

$$\begin{aligned}\dot{\rho}\frac{\partial}{\partial\dot{\rho}}L+\dot{\varphi}\frac{\partial}{\partial\dot{\varphi}}L-L&=\frac{m}{2}\left(\dot{\rho}^2+\rho^2\dot{\varphi}^2\right)+U(\rho)\\&=\frac{m}{2}\dot{\rho}^2+\frac{|\ell|^2}{2m\rho^2}+U(\rho)\\&=E\,.\end{aligned} \tag{A.28}$$

Here we employed, in the second step, relation (A.22). In summary, time invariance resulted in:

$$\frac{m}{2}\dot{\rho}^2+\frac{|\ell|^2}{2m\rho^2}+U(\rho)=E\,. \tag{A.29}$$

This is a first order differential equation in ρ.

The necessary step required for a solution of the two-body problem can now be outlined: (i) Calculate $R(t)$ according to Eq. (A.10), (ii) solve Eq. (A.29) in order to obtain $\rho(t)$, (iii) plug $\rho(t)$ into Eq. (A.22) and solve for $\varphi(t)$, (iv) since $z(t)=0$, the original vectors $r_1(t)$, $r_2(t)$ can be constructed from $\rho(t)$ and $\varphi(t)$. All integration constants are uniquely determined by the initial conditions of the problem at hand.

From Eq. (A.29) we obtain

$$\dot{\rho}=\pm\sqrt{\frac{2}{m}\left(E-U(\rho)-\frac{|\ell|^2}{2m\rho^2}\right)}\,, \tag{A.30}$$

which results in an implicit equation for ρ

$$t=t_0+\int_{\rho_0}^{\rho}\mathrm{d}\rho'\frac{m\rho'}{\sqrt{2m\rho'^2\left[E-U(\rho')\right]-|\ell|^2}}\,, \tag{A.31}$$

where we defined $\rho_0\equiv\rho(t_0)$, t_0 is some initial time, and we neglected the negative root. Equation (A.31) defines t as a function of ρ, $t=t(\rho)$, which has to be inverted to, finally, obtain the required solution $\rho=\rho(t)$. Whether Eq. (A.31) can be solved analytically depends on the particular form of the potential $U(\rho)$. If Eq. (A.31) cannot be solved analytically one has to employ numerical approximations.

Finally, the angle φ can be expressed as a function of the radius ρ, i.e. $\varphi = \varphi(\rho)$. We get from Eqs. (A.22) and (A.30)

$$\frac{\mathrm{d}\varphi}{\mathrm{d}\rho} = \frac{\mathrm{d}\varphi}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\rho} = \pm\frac{|\ell|}{m\rho^2}\left[\frac{2}{m}\left(E - U(\rho) - \frac{|\ell|^2}{2m\rho^2}\right)\right]^{-\frac{1}{2}}, \tag{A.32}$$

integrate over ρ, and find the desired relation

$$\varphi = \varphi_0 \pm |\ell| \int_{\rho_0}^{\rho} \frac{\mathrm{d}\rho'}{\rho'\sqrt{2m\rho'^2\,[E - U(\rho')] - |\ell|^2}}, \tag{A.33}$$

where $\varphi_0 \equiv \varphi(t_0)$.

Appendix B
Solving Non-linear Equations: The NEWTON Method

We give a brief introduction into the solution of non-linear equations with the help of NEWTON's method.[1] We regard a differentiable function $F(x)$ and we would like to find the solution of the equation

$$F(x) = 0 \ . \tag{B.1}$$

The simplest approach is to transform the equation into an equation of the form

$$x = f(x) \ , \tag{B.2}$$

which is always possible. This equation could be solved iteratively by simply repeating

$$x_{t+1} = f(x_t) \ , \tag{B.3}$$

where we start with some initial value x_0. If this method converges, one can approximate the solution arbitrarily close, however, convergence is not guaranteed and will in fact depend on the transformation from Eqs. (B.1) to (B.2). A more advanced technique is the so called NEWTON method [6, 7]. It is based on the definition of $f(x)$ as

$$f(x) = x - \frac{F(x)}{F'(x)} \ , \tag{B.4}$$

which allows the iteration

$$x_{t+1} = x_t - \frac{F(x_t)}{F'(x_t)} \ . \tag{B.5}$$

[1] This method is also referred to as the NEWTON-RAPHSON method.

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

Here $F'(x)$ denotes the derivative of $F(x)$ with respect to x. The convergence behavior of the iteration (B.5) highly depends on the form of the function $F(x)$ and on the choice of the starting point x_0. The routine can be regarded as converged if $|x_{t+1} - x_t| < \epsilon$, where ϵ is the accuracy required.

If $F(x)$ is not differentiable one can use the *regula falsi* or employ *stochastic methods* which are discussed in the second part of this book. The iteration of the method known as *regula falsi* is [6, 7]

$$x_{t+1} = x_t - F(x_t)\frac{x_t - x_{t-1}}{F(x_t) - F(x_{t-1})} \ . \tag{B.6}$$

A more detailed discussion on methods to solve transcendental equations numerically can be found in any textbook on numerical methods, see for instance Refs. [8, 9]. We shall also briefly introduce the case of a non-linear system of equations of the form (B.1) where $F(x) \in \mathbb{R}^N$ and $x \in \mathbb{R}^N$. In this case the iteration scheme is given by

$$x_{t+1} = x_t - J^{-1}(x_t)F(x_t) \ , \tag{B.7}$$

where

$$J(x) = \nabla_x F(x) = \begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \frac{\partial F_1(x)}{\partial x_2} & \cdots & \frac{\partial F_1(x)}{\partial x_N} \\ \frac{\partial F_2(x)}{\partial x_1} & \frac{\partial F_2(x)}{\partial x_2} & \cdots & \frac{\partial F_2(x)}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_N(x)}{\partial x_1} & \frac{\partial F_N(x)}{\partial x_2} & \cdots & \frac{\partial F_N(x)}{\partial x_N} \end{pmatrix} \ . \tag{B.8}$$

is the JACOBI matrix of $F(x)$. We can also make use of the methods discussed in Chap. 2 to calculate numerically the derivatives in Eqs. (B.5) or (B.8).

Appendix C
Numerical Solution of Linear Systems of Equations

We discuss briefly two of the most important methods to solve non-homogeneous systems of linear equations applying numerical methods. We consider a system of n equations of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \,, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \,, \\ &\vdots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \,, \end{aligned} \tag{C.1}$$

which is usually transformed into a matrix equation,

$$Ax = b \,. \tag{C.2}$$

The coefficients of the matrix $A = \{a_{ij}\}$ as well as the vector $b = \{b_i\}$ are assumed to be real valued and, furthermore, if

$$\sum_{i=1}^{n} |b_i| \neq 0 \,, \tag{C.3}$$

the problem (C.2) is referred to as non-homogeneous (inhomogeneous). The solution of non-homogeneous linear systems of equations is one of the central problems in numerical analysis, since numerous numerical methods, such as the finite difference approach to a boundary value problem, see Chap. 8, can be reduced to such a problem.

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

The solution of (C.2) is well defined as long as the matrix A is non-singular, i.e. as long as

$$\det(A) \neq 0 \,. \tag{C.4}$$

Then the unique solution of (C.2) can be written as

$$x = A^{-1} b \,. \tag{C.5}$$

However, the inversion of matrix A is very complex for $n \geq 4$ and one would prefer methods which are computationally more effective. Basically, one distinguishes between *direct* and *iterative* methods. Since a complete discussion of this huge topic would be too extensive, we will mainly focus on two methods.

In contrast to iterative procedures, direct procedures do not contain any methodological errors and can, therefore, be regarded as *exact*. However, these methods are often computationally very extensive and rounding errors are in many cases not negligible. As an example we will discuss the LU decomposition. On the other hand, many iterative methods are fast and rounding errors can be controlled easily. However, it is not guaranteed that an iterative procedure converges, even in cases where the system of equations is known to have unique solutions. Moreover, the result is an approximate solution. As an illustration for an iterative procedure we will discuss the GAUSS-SEIDEL method.

C.1 The *LU* Decomposition

The LU decomposition [6, 10] is essentially a numerical realization of GAUSSIAN elimination which is based on a fundamental property of linear systems of equations (C.2). This property states the system (C.2) to remain unchanged when a linear combination of rows is added to one particular row. This property is then employed in order to obtain a matrix in *triangular* form. It was demonstrated by DOOLITTLE and CROUT [6, 10, 11] that the GAUSSIAN elimination can be formulated as a decomposition of the matrix A into two matrices L and U:

$$A = LU \,. \tag{C.6}$$

Here, U is an *upper triangular* matrix and L is a *lower triangular* matrix. In particular, U is of the form

$$U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & & & \vdots \\ 0 & 0 & \dots & u_{nn} \end{pmatrix} \,, \tag{C.7}$$

and L is of the form

$$L = \begin{pmatrix} 1 & 0 & & \dots 0 \\ m_{21} & 1 & 0 & \dots 0 \\ m_{31} & m_{32} & 1 & \dots 0 \\ \vdots & & & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \dots 1 \end{pmatrix} . \tag{C.8}$$

The factorization (C.6) is referred to as *LU decomposition*. The corresponding procedure can be easily identified by equating the elements in (C.6). One can show that the following operations yield the desired result: For $j = 1, 2, \dots, n$ one computes

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} m_{ik} u_{kj} \qquad i = 1, 2, \dots, j \,, \tag{C.9}$$

$$m_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} m_{ik} u_{kj} \right) \quad i = j+1, j+2, \dots, n \,, \tag{C.10}$$

with the requirement that $u_{jj} \neq 0$. Note that in this notation we used the convention that the contribution of the sum is equal to zero if the upper boundary is less than the lower boundary. We rewrite Eq. (C.2) with the help of the LU decomposition (C.6)

$$Ax = LUx = b \,, \tag{C.11}$$

and by defining $y = Ux$, we retrieve a system of equations for the variable y:

$$Ly = b \,. \tag{C.12}$$

The particular form of L allows to solve the system (C.12) immediately by *forward substitution*. We find the solution

$$y_i = b_i - \sum_{k=1}^{i-1} m_{ik} y_k \,, \qquad i = 1, 2, \dots, n \,, \tag{C.13}$$

and the equation

$$Ux = y \,, \tag{C.14}$$

remains. It can solved by *backward substitution*:

$$x_i = \frac{1}{u_{ii}} \left(y_i - \sum_{k=i+1}^{n} u_{ik} x_k \right) \,, \quad i = n, n-1, \dots, 1 \,. \tag{C.15}$$

We note that this method can also be employed to invert the matrix A. The strategy is based on the relation

$$AX = I \,, \tag{C.16}$$

where $X = A^{-1}$ is to be determined and I is the n-dimensional identity. Equation (C.16) is equivalent to the following system of equations:

$$\begin{aligned} Ax_1 &= \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \,, \\ Ax_2 &= \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \,, \\ &\vdots \qquad\quad \vdots \\ Ax_n &= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \,, \end{aligned} \tag{C.17}$$

where the vectors x_i are the rows of the unknown matrix X, i.e. $X = (x_1, x_2, \ldots, x_n)$. The n equations of the system (C.17) can be solved with the help of the LU decomposition.

Furthermore, one can easily calculate the determinant of A using the LU decomposition. We note that

$$\det(A) = \det(LU) = \det(L)\det(U) = \det(U) \,, \tag{C.18}$$

since L and U are triangular matrices, the determinants are equal to the product of the diagonal elements, which yields $\det(L) = 1$. Hence we have

$$\det(A) = \det(U) = \prod_{i=1}^{n} u_{ii} \,. \tag{C.19}$$

In conclusion we remark that there are many specialized methods which have been designed particularly for matrices of specific forms, such as tridiagonal matrices, symmetric matrices, block-matrices, ... Such matrices commonly appear

in physics applications. For instance, we remember that the matrix we encountered in Sect. 8.2 within the context of a finite difference approximation of boundary value problems, was tridiagonal. These specialized methods are usually the first choice if one has a matrix of such a specific form because they are much faster and more stable than methods developed for matrices of more general form. Since a full treatment of these methods is beyond the scope of this book, we refer the interested reader to books on numerical linear algebra, for instance Refs. [10, 11].

C.2 The GAUSS-SEIDEL Method

The GAUSS-SEIDEL method is an iterative procedure to approximate the solution of non-homogeneous systems of linear equations [6, 12]. The advantage of an iterative procedure, in contrast to a direct approach, is that its formulation is in general much simpler. However, one might have problems with the convergence of the method, even in cases where a solution exists and is unique. We note that the GAUSS-SEIDEL method is of particular interest whenever one has to deal with *sparse* coefficient matrices.[1] This requirement is not too restrictive since most of the matrices encountered in physical applications are indeed sparse. As an example we remember the matrices arising in the context of a finite difference approach to boundary value problems, Sect. 8.2.

Again, we use Eq. (C.1) as a starting point for our discussion. It is a requirement of the GAUSS-SEIDEL method that *all* diagonal elements of A are non-zero. We then solve each row of (C.1) for x_i. This creates the following hierarchy

$$\begin{aligned}
x_1 &= -\frac{1}{a_{11}}\left(a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n - f_1\right) \ , \\
x_2 &= -\frac{1}{a_{22}}\left(a_{21}x_1 + a_{23}x_3 + \ldots + a_{2n}x_n - f_2\right) \ , \\
&\vdots \quad \vdots \\
x_n &= -\frac{1}{a_{nn}}\left(a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,n-1}x_{n-1} - f_n\right) \ ,
\end{aligned} \tag{C.20}$$

or in general for $i = 1, \ldots, n$

$$x_i = -\frac{1}{a_{ii}}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j - f_i\right) \ . \tag{C.21}$$

[1] A matrix A is referred to as sparse, when the matrix is populated primarily by zeros.

We note that Eq. (C.21) can be rewritten as a matrix equation

$$x = Cx + b \,, \tag{C.22}$$

where we defined the matrix $C = \{c_{ij}\}$ via

$$c_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}} & i \neq j \,, \\ 0 & i = j \,, \end{cases} \tag{C.23}$$

and the vector $b = \{b_i\}$ as

$$b_i = \frac{f_i}{a_{ii}} \,. \tag{C.24}$$

We recognize that Eq. (C.21) can be transformed into an iterative form with the help of a trivial manipulation

$$x_i = x_i - \left[x_i + \frac{1}{a_{ii}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j - f_i \right) \right] \,, \tag{C.25}$$

or

$$x_i^{(t+1)} = x_i^{(t)} - \Delta x_i^{(t)} \,, \tag{C.26}$$

where

$$\Delta x_i^{(t)} = x_i^{(t)} + \frac{1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(t+1)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(t)} - f_i \right) \,. \tag{C.27}$$

Equation (C.26) in combination with (C.27) produces a sequence of vectors

$$x^{(0)} \to x^{(1)} \to x^{(2)} \to \ldots \to x^{(m)} \,, \tag{C.28}$$

where $x^{(0)}$ is referred to as the initialization vector or trial vector. One can prove that if this sequence converges, it approaches the exact solution x arbitrarily close:

$$\lim_{t \to \infty} x^{(t)} = x \,. \tag{C.29}$$

We remark that if the terms $x_i^{(t+1)}$ on the right hand side of Eq. (C.27) are replaced by $x_i^{(t)}$ the method is referred to as the JACOBI method.

To terminate the GAUSS-SEIDEL method, we need an exit condition: One should terminate the iteration whenever:

- The approximate solution $x^{(t)}$ obeys the required accuracy ϵ or $\tilde{\epsilon}$, for instance

$$\max\left(|x_i^{(t)} - x_i^{(t-1)}|\right) \leq \epsilon \ , \tag{C.30}$$

  where ϵ is the absolute error, or

$$\max\left(\frac{|x_i^{(t)} - x_i^{(t-1)}|}{|x_i^{(t)}|}\right) \leq \tilde{\epsilon} \ , \tag{C.31}$$

  where $\tilde{\epsilon}$ is the relative error.
- When a maximum number of iterations is reached. This condition may be interpreted as an emergency exit which ensures that the iteration terminates even if the process is not convergent or has still not converged.

Let us discuss one final, however, crucial point of this section: In many cases the convergence of the GAUSS-SEIDEL method can be significantly improved by including a *relaxation parameter* ω to the iterative process. In this case the update routine (C.26) takes on the form

$$x_i^{(t+1)} = x_i^{(t)} - \omega \Delta x_i^{(t)} \ . \tag{C.32}$$

If the relaxation parameter ω obeys $\omega > 1$ one speaks of *over-relaxation*, if $\omega < 1$ of *under-relaxation* and if $\omega = 1$ the regular GAUSS-SEIDEL method is recovered. An appropriate choice of the relaxation parameter may fasten the convergence of the method significantly. The best result will certainly be obtained if the *ideal* value of ω, ω_i were known. Unfortunately, it is impossible to determine ω_i prior to the iteration in the general case. We remark the following properties:

- The method (C.32) is only convergent for $0 < \omega \leq 2$.
- If the matrix C is positive definite and $0 < \omega < 2$, the GAUSS-SEIDEL method converges for any choice of $x^{(0)}$ (OSTROWSKI-REICH theorem, [13]).
- In many cases, $1 \leq \omega_i \leq 2$. We note that this inequality holds only under particular restrictions for the matrix C [see Eq. (C.23)]. However, we note without going into detail, that these restrictions are almost always fulfilled when one is confronted with applications in physics.
- If C is positive definite and tridiagonal, the ideal value ω_i can be calculated using

$$\omega_i = \frac{2}{1 + \sqrt{1 - \lambda^2}} \ , \tag{C.33}$$

  where λ is the largest eigenvalue of C, Eq. (C.23).

- Since the calculation of λ is in many cases quite complex, one could employ the following idea: It is possible to prove that

$$\lim_{t\to\infty} \frac{|\Delta x^{(t+1)}|}{|\Delta x^{(t)}|} \to \lambda^2 \,. \tag{C.34}$$

  Hence, one may start with $\omega = 1$, perform t_0 ($20 < t_0 < 100$) iterations and then approximate ω_i with the help of Eq. (C.33) and

$$\lambda^2 \approx \frac{|\Delta x^{(t_0)}|}{|\Delta x^{(t_0-1)}|} \,. \tag{C.35}$$

  The iteration is then continued with the approximated value of ω_i until convergence is reached.

In conclusion we remark that numerous numerical libraries contain sophisticated routines to solve linear systems of equations. In many cases it is, thus, advisable to rely on such routines.

Appendix D
Fast Fourier Transform

Integral transforms are indispensable in modern mathematics and natural science because they can be employed to simplify complex mathematical problems. In this Appendix we will discuss the FOURIER transform as one prominent representative of integral transforms in general. Loosely speaking, the FOURIER transform is the unambiguous decomposition of a function $f(x)$ into its frequency components. Its applications range from the harmonic analysis of periodic signals to the solution of differential equations and the description of wave phenomena in classical mechanics [2, 4], electrodynamics [14–16], quantum mechanics [17–19], and many more. Here, we briefly discuss its numerical implementation, the *fast* FOURIER *transform* (FFT) and its applications in Computational Physics.

We start by recalling the concept of FOURIER series: It is asserted by FOURIER's theorem that every square-integrable, d-periodic function $f(x), f(x+d) = f(x)$, can be (uniquely) represented as[1]

$$f(x) = \sum_{n\in\mathbb{Z}} \hat{f}_n \exp\left(i\frac{2\pi nx}{d}\right), \tag{D.1}$$

where the complex coefficients $\hat{f}_n \in \mathbb{C}$ are related to $f(x)$ by the inverse transform

$$\hat{f}_n = \frac{1}{d}\int_0^d \mathrm{d}x f(x) \exp\left(-i\frac{2\pi nx}{d}\right). \tag{D.2}$$

[1] In other words, the plane waves $\exp(in2\pi x/d)$ with period d form a complete, orthonormal basis in the space of d-periodic, square integrable functions with the scalar product (10.10). We remark that this also applies to functions which are defined on a compact interval of length d [20].

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

The representation (D.1) of $f(x)$ is referred to as the FOURIER series of $f(x)$ and the coefficient (D.2) is the FOURIER coefficient of order n. Equation (D.1) is an unambiguous expansion of the function $f(x)$ into contributions which oscillate with an integer multiple of the frequency $2\pi/d$. There are numerous important properties, examples and applications of FOURIER series for which we refer to the literature [21–24].

The concept of FOURIER series can be generalized to the idea of the FOURIER transform of a square integrable function $f(x)$ by formally letting $d \to \infty$ [23]. The FOURIER transform relates the function $f(x)$ to its transform $\hat{f}(k)$, $k \in \mathbb{R}$, via[2]

$$f(x) = \int_{-\infty}^{\infty} \mathrm{d}k \hat{f}(k) \exp(ikx), \tag{D.3}$$

and the inverse transform is obtained as

$$\hat{f}(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}x f(x) \exp(-ikx). \tag{D.4}$$

The transform (D.3) and its inverse (D.4) can be used to considerably simplify mathematical problems. For instance, a linear differential equation for the function $f(x)$ is mapped onto a linear algebraic equation for $\hat{f}(k)$. The solution of the differential equation is then obtained by back-transforming the solution $\hat{f}(k)$ of the algebraic equation. Again, we refer to the literature for further applications and the various properties of the transforms (D.3) and (D.4). Instead, let us concentrate on the question of how to compute the FOURIER transform (D.2) numerically.

It appears to be reasonable to start with the concepts developed in Chap. 3.[3] For this purpose, we assume that the function $f(x)$ is solely known on a grid of N equidistant grid-points x_ℓ, $\ell = 0, \ldots, N-1$. In addition, we note that it is sufficient to limit our discussion to 2π-periodic functions. Thus, we can choose our grid-points to be $x_\ell = x_0 + \ell h$ where $x_0 = 0$ and $h = 2\pi/N$, so that $x_{N-1} = 2\pi(1-1/N)$.

Approximating the integral (D.2) with the help of the forward rectangular rule, Chap. 3, yields

$$\hat{f}_n = \frac{1}{N} \sum_{\ell=0}^{N-1} f_\ell \exp\left(-\frac{2\pi n \ell}{N}\right) + \mathcal{O}(h^2). \tag{D.5}$$

It follows from this equation that the coefficients $\hat{f}_n$ are periodic in n with period N due to the finite number of grid-points. Hence, the maximal number of distinct

[2] We work here with the asymmetric definition of the FOURIER transform. For other definitions, the pre-factors have to be adapted consistently.

[3] If $f(x)$ is not periodic we have to truncate the integral (D.4) and restrict the integration to a suitable finite interval so that the problem again reduces to the evaluation of Eq. (D.2).

coefficients is equal to the number of grid-points. The inversion of Eq. (D.5) follows directly from Eq. (D.1) and reads[4]

$$f_\ell = \sum_{n=0}^{N-1} \hat{f}_n \exp\left(i\frac{2\pi n\ell}{N}\right). \tag{D.6}$$

The transforms (D.5) and (D.6) are referred to as the *discrete* FOURIER *transform* (DFT) and its inverse, respectively. We cast these relations into matrix form by defining vectors $F = (f_0, \ldots, f_{N-1})^T$ and $\hat{F} = (\hat{f}_0, \ldots, \hat{f}_{N-1})^T$ together with the matrix W of elements:

$$W_{nm} = \omega_N^{nm}. \tag{D.7}$$

Here, $\omega_N = \exp\left(\frac{2\pi i}{N}\right)$ denotes the N-th root of unity. The transformation matrix W is known as the FOURIER matrix or DFT matrix and it is easy to prove that its inverse W^{-1} has the elements

$$\left(W^{-1}\right)_{nm} = \omega_N^{-nm}. \tag{D.8}$$

All this allows to rewrite Eqs. (D.5) and (D.6) in compact form:

$$\hat{F} = \frac{1}{N} W^{-1} F, \qquad \text{and} \qquad F = W\hat{F}. \tag{D.9}$$

Thus, we reduced the problem of numerically implementing the FOURIER transform (D.2) to the task of multiplying the $N \times N$ complex matrix W with the N-element vector F. This means that we have to perform N^2 complex multiplications and $N(N-1)$ complex additions. However, the symmetry $W_{nm} = W_{mn}$ already suggests that there is further room for improvement. In fact, there are methods that do much better and these algorithms are known as *fast* FOURIER *transform* (FFT) algorithms.

We limit our presentation to the version proposed by COOLEY and TUKEY [6, 25, 26] which is, with some variations, the most common algorithm. In its simplest form it is based on the observation that one can always split the FOURIER transform (D.5) into an even and an odd part

$$\hat{f}_n = \frac{1}{N} \sum_{\ell=0}^{N/2} f_{2\ell} \omega_N^{2n\ell} + \frac{\omega_N^n}{N} \sum_{\ell=0}^{N/2} f_{2\ell+1} \omega_N^{2n\ell}, \tag{D.10}$$

[4] It follows directly from the summation rule of the geometric series that

$$\sum_{m=0}^{N-1} \exp\left(i\frac{2\pi mn}{N}\right) = N\delta_{n0}.$$

provided that N is even.[5] Since $\omega_N^{2k} = \omega_{N/2}^n$, we can interpret Eq. (D.10) as the linear combination of two FOURIER transforms of length $N/2$. Denoting the FOURIER coefficients of the function values $f_{2\ell}$ on even grid-points by $\hat{A}_n$ and of the values $f_{2\ell+1}$ on odd grid-points by $\hat{B}_n$, we obtain for $n = 1, \ldots, N/2$

$$\hat{f}_n = \hat{A}_n + \omega_N^n \hat{B}_n. \tag{D.11}$$

We now make use of the property that the FOURIER coefficients are periodic, i.e. $\hat{A}_{n+N/2} = \hat{A}_n$, $\hat{B}_{n+N/2} = \hat{B}_n$, and that $\omega_N^{n+N/2} = -\omega_N^n$. Thus, we can calculate the remaining coefficients $\hat{f}_n$, $n = N/2 + 1, \ldots, N$ with the help of:

$$\hat{f}_{n+N/2} = \hat{A}_n - \omega_N^n \hat{B}_n. \tag{D.12}$$

Because of Eqs. (D.11) and (D.12) the N FOURIER coefficients can be computed as a linear combination of two FOURIER transforms of size $N/2$. The recursive application of the very same scheme to $\hat{A}_n$ and $\hat{B}_n$ constitutes the core of the FFT algorithm in its simplest variation [6]. It is also the efficiency of this algorithm that makes the FOURIER transform an attractive tool for numerical calculations [27]. In fact, there are several problems in this book where an algorithm based on FFT could have been evoked. Let us discuss two examples in more detail in order to illustrate this.

(i) In Chap. 9.3 we could have solved the stationary inhomogeneous heat equation for its FOURIER coefficients followed by the back-transform.[6] Denoting by $\hat{T}_n$ the FOURIER coefficients of the temperature $T(x)$ and by $\hat{\Gamma}_n$ the FOURIER coefficients of the heat source/drain, we obtain from Eq. (9.20)

$$\hat{T}_n = -\frac{\hat{\Gamma}_n}{(n\omega)^2}, \tag{D.13}$$

where $\omega = 2\pi/L$. Performing the inverse FFT on Eq. (D.13) and adding the homogeneous solution (9.4) immediately gives the required temperature profile $T(x)$. In a similar fashion, FFT could have been used for solving the partial differential equations discussed in Chap. 11. From the examples in this chapter, the time-dependent SCHRÖDINGER equation serves as our second application.

(ii) The Hamiltonian of a free point particle (for simplicity in one dimension) is diagonal in momentum space, $H = P^2/2m$, with the momentum operator (10.6). Given the position-space representation of the initial state $\psi(x, t)$, we can then

[5] This is not a limitation because we can always choose N to be even.

[6] Here we use the fact that the plane waves $\exp(in2\pi x/d)$ form a complete, orthonormal basis of the functions defined on a compact interval of length L [20].

compute the time evolved wave packet $\psi(x, t + \Delta t)$ for arbitrary $\Delta t > 0$ according to Eq. (10.17) as

$$\psi(x, t + \Delta t) = \frac{1}{2\pi\hbar} \int \mathrm{d}p \exp\left(-\frac{i\Delta t}{\hbar}\frac{p^2}{2m} + \frac{i}{\hbar}px\right) \hat{\psi}(p, t), \tag{D.14}$$

where $\hat{\psi}(p, t)$ is the momentum space representation of the initial state $\psi(x, t)$, i.e. its FOURIER transform with $k = p/\hbar$. Hence, the time evolution of the free wave packet is readily computed numerically with the help of the FFT and its inverse.

It is now certainly interesting to investigate whether or not a similar approach can be applied to solve Eq. (10.1) in the presence of a potential $V(x)$ which is diagonal in position space. Although this can be achieved by solving the full stationary eigenvalue problem (10.9) followed by the application of the eigenvector expansion (10.17), we present here a more efficient but approximate solution valid for small time steps Δt. In order to see this, we transform Eq. (D.14) into a slightly more compact form. Denoting by $\mathscr{F}$ the FOURIER transform operator, $\hat{\psi}(p) = \mathscr{F}\psi(x)$, we can write Eq. (D.14) as

$$\psi(x, t + \Delta t) = \mathscr{F}^{-1} U_{\Delta t} \mathscr{F} \psi(x, t), \tag{D.15}$$

where $U_{\Delta t} = \exp\left(-i\Delta t p^2/2\hbar m\right)$ is the unitary time evolution operator for the time interval Δt.[7] The correct result can not be obtained by multiplying Eq. (D.15) with the position-space time evolution of the potential $V_{\Delta t} = \exp(-i\Delta t V(x)/\hbar)$ because the operators V and P do not commute. However, by applying the BAKER-CAMPBELL-HAUSDORFF formula[8] [17, 19, 28], we can approximate the time evolution $\psi(x, t) \to \psi(x, t + \Delta t)$ for a small time step Δt by:

$$\psi(x, t + \Delta t) = V_{\Delta t} \mathscr{F}^{-1} U_{\Delta t} \mathscr{F} \psi(x, t) + \mathscr{O}(\Delta t^2). \tag{D.16}$$

An even better approximation is obtained by the symmetrized form

$$\psi(x, t + \Delta t) = \mathscr{F}^{-1} U_{\Delta t/2} \mathscr{F} V_{\Delta t} \mathscr{F}^{-1} U_{\Delta t/2} \mathscr{F} \psi(x, t) + \mathscr{O}(\Delta t^3). \tag{D.17}$$

This method, known as the *split operator technique* [29], is a frequently used method to numerically solve time dependent problems in quantum mechanics with the help of FFT.

[7] U_t is the momentum space representation of the free unitary time evolution operator $U = \exp(-itP^2/2\hbar m)$.

[8] The BAKER-CAMPBELL-HAUSDORFF formula states how the exponential function $\exp(X + Y)$ of two non-commuting operators X and Y can be expanded in terms of products of exponentials of their commutators [17, 28].

Appendix E
Basics of Probability Theory

E.1 Classical Definition

It is the aim of this Appendix to summarize the most important definitions and results from basic probability theory as required within this book. For a more in depth presentation we refer to the literature [30–34].

The classical probability $P(A)$ for an event A is defined by the number of favorable results n, divided by the number of possible results m,

$$P(A) = \frac{n}{m} \,. \tag{E.1}$$

For two events A and B we can deduce the following rules[1]

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) \,, \tag{E.2a}$$

$$P(Z) = 0 \quad \text{impossible event; } Z \ldots \text{zero element} \,, \tag{E.2b}$$

$$P(I) = 1 \quad \text{certain event; } I \ldots \text{identity element} \,, \tag{E.2c}$$

$$0 \leq P(A) \leq 1 \,, \tag{E.2d}$$

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \,, \tag{E.2e}$$

[1] Here we use the symbols $\vee$ and $\wedge$ to denote the Boolean operators OR and AND, respectively.

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

where $P(A|B)$ is the probability for the event A under the constraint that event B is true. Moreover, if $\overline{A}$ is the complementary event[2] to A we have

$$P(\overline{A}) = 1 - P(A) \ . \tag{E.3}$$

The statistical definition of the probability for an event A is given by:

$$P(A) = \lim_{m \to \infty} \frac{n}{m} \ . \tag{E.4}$$

E.2 Random Variables and Moments

A random variable is a functional which assigns to an event ω a real number x from the set of possible outcomes Ω: $x = X(\omega)$ [31].[3] Roughly speaking it is a variable whose value is assigned to the observation of some random process. The mean value of a discrete random variable X is defined by

$$\langle X \rangle = \sum_{\omega \in \Omega} X(\omega) P_\omega \ , \tag{E.5}$$

where P_ω is the probability for the event ω. For instance, in case of a dice-throw $X(\omega) \equiv n = 1, 2, \ldots, 6$.

We restrict ourselves now to discrete random variables and, thus, x can only take on discrete values. Furthermore, we introduce the function of random variables $Y = f(X)$ and define quite generally its mean value:

$$\langle f(X) \rangle \equiv \langle f \rangle = \sum_i f(x_i) P_i \ . \tag{E.6}$$

Note that

$$\langle 1 \rangle \equiv \sum_i P_i = 1 \ . \tag{E.7}$$

Moments of order k of a random variable X are defined by

$$m_k := \left\langle X^k \right\rangle \ , \tag{E.8}$$

[2] This means that $A \vee \overline{A} = I$ and $A \wedge \overline{A} = Z$.

[3] A more exact formulation will follow in the course of this Appendix.

and central moments are introduced via the relation

$$\mu_k := \left\langle (\Delta X)^k \right\rangle = \left\langle (X - \langle X \rangle)^k \right\rangle \ . \tag{E.9}$$

Of particular interest is the second central moment, the *variance*:

$$\text{var}\,(X) := \left\langle (X - \langle X \rangle)^2 \right\rangle = \left\langle X^2 \right\rangle - \langle X \rangle^2 \ . \tag{E.10}$$

Finally, the *standard deviation* σ is defined as the square root of the variance:

$$\sigma := \text{std}(X) = \sqrt{\text{var}\,(X)} \ . \tag{E.11}$$

We study now a discrete set of observations x_i where $i = 1, \ldots, N$. Then the sample mean value is given by

$$\overline{x} = \frac{1}{N} \sum_i x_i \ , \tag{E.12}$$

and the error (standard deviation) of $\overline{x}$ (*standard error*) can be determined from:

$$\text{var}\,(\overline{x}) = \text{var}\left(\frac{1}{N} \sum_i x_i \right) = \frac{\sigma^2}{N} \ . \tag{E.13}$$

We assumed here the x_i to be uncorrelated with the consequence that $\text{cov}\left(x_i, x_j\right) = \text{var}\,(x_i)\, \delta_{ij}$ [defined in Eq. (E.16)]. Therefore,

$$\text{standard error} = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}} \ , \tag{E.14}$$

where σ is the standard deviation of the observations as defined above.

In the case of multiple random variables we can proceed as above. For instance, the expectation value of a function of two random variables is given by

$$\langle f(X, Y) \rangle := \sum_{i,j} f(x_i, y_j) P_{ij} \ , \tag{E.15}$$

and the covariance between two random variables:

$$\text{cov}\,(X, Y) := \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle = \langle XY \rangle - \langle X \rangle \langle Y \rangle \ . \tag{E.16}$$

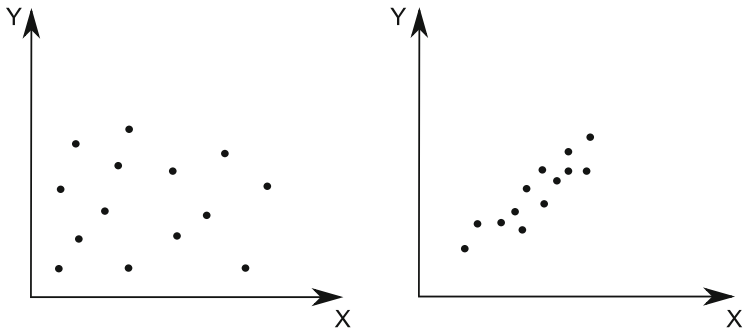

Fig. E.1 Uncorrelated (*left panel*) and positively correlated (*right panel*) variables X and Y

The value of the covariance together with its sign determines important properties of the random variables X and Y in their relation to each other:

$$\operatorname{cov}(X,Y) = \begin{cases} > 0 & \text{for } Y - \langle Y\rangle > 0 \Rightarrow X - \langle X\rangle > 0\ , \\ & \qquad \text{(positive linear correlation)} \\ < 0 & \text{for } Y - \langle Y\rangle > 0 \Rightarrow X - \langle X\rangle < 0\ , \\ < 0 & \text{for } X - \langle X\rangle > 0 \Rightarrow Y - \langle Y\rangle < 0\ , \\ & \qquad \text{(negative linear correlation)} \\ = 0 & \text{no linear dependence between } X \text{ and } Y\ . \end{cases} \tag{E.17}$$

Random variables whose covariance is zero are called *uncorrelated*. [This property was used in the derivation of Eq. (E.13).] To give an example, Fig. E.1 compares schematically uncorrelated and positively correlated random variables X and Y.

E.3 Binomial Distribution and Limit Theorems

The binomial distribution is given by

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}\ , \tag{E.18}$$

where $\binom{n}{k}$ is the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}\ . \tag{E.19}$$

For large values of n STIRLING's approximation can be applied to calculate an estimate of $n!$:

$$n! = n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi} \left[1 + \mathcal{O}(n^{-1})\right] . \tag{E.20}$$

Furthermore, it is easy to prove that the mean value and the variance of the binomial distribution are given by

$$\langle k \rangle = np , \tag{E.21}$$

$$\text{var}\,(k) = np(1-p) . \tag{E.22}$$

The DE MOIVRE-LAPLACE theorem states that for $\text{var}(k) \gg 1$

$$P(k|n,p) \approx g(k|k_0,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(k-k_0)^2}{2\sigma^2}\right] , \tag{E.23}$$

where $k_0 = \langle k \rangle$ and $\sigma = \sqrt{\text{var}\,(k)}$. We can also deduce that

$$P(k = np|n,p) = \frac{1}{\sqrt{2\pi np(1-p)}} \longrightarrow 0 , \tag{E.24}$$

for $n \to \infty$. From this, BERNOULLI's law of large numbers follows

$$P(|k/n - p| < \epsilon|n,p) \to 1 \quad \forall\, \epsilon > 0 . \tag{E.25}$$

E.4 POISSON Distribution and Counting Experiments

If the mean expectation value μ is independent of the number of experiments n, i.e. $np = \mu \equiv \text{const}$, it follows from Eq. (E.18) that

$$\lim_{n\to\infty} P\left(k\middle|n, p = \frac{\mu}{n}\right) = \exp(-\mu)\frac{\mu^k}{k!} =: P(k|\mu) . \tag{E.26}$$

The distribution $P(k|\mu)$ is called POISSON distribution. We obtain for the POISSON distribution:

$$\langle k \rangle = \mu , \tag{E.27}$$

$$\text{var}\,(k) = \mu . \tag{E.28}$$

It is important to note that counting experiments, as for instance radioactive decay, follow the Poisson statistics. A typical counting experiment observes within the time interval t (in average) μ events. This time interval is now divided into n subintervals with $\Delta t = t/n$. If the counting events can be assumed to be independent, the process follows a binomial distribution and we have $\mu = np$. This is equivalent to $p = \mu/n$. We return to the case of radioactive decay: We count μ signals within one minute which are uniformly distributed over the time interval. The experiment is now reduced to a time interval of one second and the probability of detecting a signal consequently reduces to $\mu/60$. For $p \ll 1$ but $np \not\gg 1$ the binomial distribution $P(k|n,p)$ can be approximated by $P(k|\mu)$ and we can use for large values of μ

$$P(k|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(k-\mu)^2}{2\sigma^2}\right] , \tag{E.29}$$

with

$$\sigma = \sqrt{\mu} \approx \sqrt{k} . \tag{E.30}$$

In most experimentally relevant cases is μ unknown and is approximated by:

$$\mu = k \pm \sqrt{k} . \tag{E.31}$$

E.5 Continuous Variables

We define the *cumulative distribution function* (cdf) [31, 32], $F(x)$, of a continuous variable x by[4]

$$F(x) := P(X \leq x|\mathscr{B}) , \tag{E.32}$$

where $\mathscr{B}$ is a generalized condition (*condition complex*). Moreover, we define the *probability density function* (pdf), $p(x)$ by

$$p(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) . \tag{E.33}$$

It follows that

$$p(x)\mathrm{d}x = [F(x+\mathrm{d}x) - F(x)] \stackrel{!}{=} P(x \leq X \leq x+\mathrm{d}x|\mathscr{B}) . \tag{E.34}$$

[4]For convenience we use here the notation $F(x)$ for the cumulative distribution function in contrast to the notation $P(x)$ used throughout the second part of this book.

Hence,

$$F(x) = \int_{-\infty}^{x} \mathrm{d}x' p(x') \ . \tag{E.35}$$

Note that the pdf is normalized

$$\int \mathrm{d}x' p(x') = F(\infty) = P(X \leq \infty | \mathscr{B}) = 1 \ , \tag{E.36}$$

and non-negative

$$p(x) \geq 0 \ . \tag{E.37}$$

E.6 BAYES' Theorem

We regard a set of discrete events A_i under the generalized condition $\mathscr{B}$. Then we have the normalization condition

$$\sum_i P(A_i|\mathscr{B}) = 1 \ , \tag{E.38}$$

and the marginalization rule

$$P(B|\mathscr{B}) = \sum_i P(B|A_i, \mathscr{B}) P(A_i|\mathscr{B}) \ . \tag{E.39}$$

BAYES' theorem [33, 35] for discrete variables follows from Eq. (E.2e) since $P(A \wedge B) = P(B \wedge A)$:

$$P(A|B, \mathscr{B}) = \frac{P(B|A, \mathscr{B}) P(A|\mathscr{B})}{P(B|\mathscr{B})} \ . \tag{E.40}$$

In case of continuous variables the above equations modify accordingly. The marginalization and BAYES' theorem for pdfs are then given by

$$P(B|\mathscr{B}) = \int \mathrm{d}x P(B|x, \mathscr{B}) p(x|\mathscr{B}) \ , \tag{E.41}$$

and

$$p(y|x, \mathscr{B}) = \frac{p(x|y, \mathscr{B}) p(y|\mathscr{B})}{p(x|\mathscr{B})} \ . \tag{E.42}$$

E.7 Normal Distribution

The normal distribution (GAUSS distribution) is defined by the pdf:

$$p(x) = \mathcal{N}(x|x_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-x_0)^2}{2\sigma^2}\right] . \tag{E.43}$$

The corresponding cdf

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \mathrm{d}x' \exp\left[-\frac{(x'-x_0)^2}{2\sigma^2}\right] \\ &= \Phi\left(\frac{x-x_0}{\sigma}\right) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x-x_0}{\sqrt{2\sigma^2}}\right) , \end{aligned} \tag{E.44}$$

follows. Here $\Phi(x)$ is given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{d}x' e^{-x'^2/2} , \tag{E.45}$$

and $\mathrm{erf}(x)$ is the error function [36, 37]:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \mathrm{d}x' e^{-x'^2} . \tag{E.46}$$

Furthermore, we obtain

$$\langle x \rangle = x_0 , \tag{E.47}$$

$$\mathrm{var}\,(x) = \sigma^2 . \tag{E.48}$$

E.8 Central Limit Theorem

Let S denote a random variable defined by

$$S = \sum_{i=1}^{N} c_i X_i , \tag{E.49}$$

where the X_i are independent and identically distributed random numbers with mean μ and variance σ^2 and

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} c_i^k = \mathrm{const} , \quad \forall k \in \mathbb{Z} . \tag{E.50}$$

Then,

$$p(S|N,\mathscr{B}) \approx \mathscr{N}\left[S|\,\langle S\rangle\,,\mathrm{var}\,(S)\right]\,, \tag{E.51}$$

with

$$\langle S\rangle = \mu\sum_{i=1}^{N} c_i\,, \tag{E.52}$$

and

$$\mathrm{var}\,(S) = \sigma^2\sum_{i=1}^{N} c_i^2\,, \tag{E.53}$$

for large values of N. The theorem of DE MOIVRE-LAPLACE is a special case of the central limit theorem, with the result that the X_i are binomial distributed.

E.9 Characteristic Function

The characteristic function $G(k)$ of a stochastic variable X is defined by [31, 32]

$$G(k) = \left\langle e^{ikX}\right\rangle = \int_I \mathrm{d}x e^{ikx} p(x)\,, \tag{E.54}$$

where I denotes the range of the pdf $p(x)$. It follows that

$$G(0) = 1 \qquad \text{and} \qquad |G(k)| \leq 1\,. \tag{E.55}$$

Expanding Eq. (E.54) in a Taylor series with respect to k yields

$$G(k) = \sum_m \frac{(ik)^m}{m!}\int_I \mathrm{d}x\, x^m p(x) \equiv \sum_m \frac{(ik)^m}{m!}\left\langle X^m\right\rangle\,. \tag{E.56}$$

Hence, the characteristic function is a *moment generating function*.

E.10 The Correlation Coefficient

We shall briefly define and discuss the correlation coefficient. Two random variables X and Y form a random vector $Z = (X, Y)$ which follows the pdf $p(Z) = p(X, Y)$ with the normalization

$$\int \mathrm{d}x\mathrm{d}y\, p(x,y) = 1\,. \tag{E.57}$$

The correlation coefficient r is now defined as

$$r = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \operatorname{var}(Y)}} , \tag{E.58}$$

where $\operatorname{cov}(X, Y)$ is the covariance (E.16) of X and Y while $\operatorname{var}(\cdot)$ denotes the variance (E.10) of the respective argument. It follows from the CAUCHY-SCHWARZ inequality that $0 \leq r^2 \leq 1$ and, therefore, $-1 \leq r \leq 1$.[5]

The random variables X and Y are said to be the stronger correlated the bigger r^2 becomes because for statistically independent (uncorrelated) variables we have $p(x, y) = q_1(x)q_2(y)$ with the consequence that $\operatorname{cov}(X, Y) = 0$ and, thus, $r = 0$.

The definition of the correlation coefficient is commonly motivated by the problem of linear regression: Suppose we have a set of data points Y associated with data points X. We would like to find a linear function $f(X) = a + bX$ which approximates the data points Y as good as possible. The problem may be stated as

$$\left\langle [Y - f(X)]^2 \right\rangle = \left\langle (Y - a - bX)^2 \right\rangle \to \min , \tag{E.60}$$

where a and b are real constants. This corresponds to GAUSS's method of minimizing the square of errors. We have

$$\frac{\partial}{\partial a} \left\langle (Y - a - bX)^2 \right\rangle = -2 \left\langle Y - a - bX \right\rangle = 0 , \tag{E.61}$$

and

$$\frac{\partial}{\partial b} \left\langle (Y - a - bX)^2 \right\rangle = -2 \left\langle (Y - a - bX)X \right\rangle = 0 . \tag{E.62}$$

Equations (E.61) and (E.62) result in:

$$a + b \langle X \rangle = \langle Y \rangle , \tag{E.63}$$

$$a \langle X \rangle + b \left\langle X^2 \right\rangle = \langle XY \rangle . \tag{E.64}$$

Both are easily solved for a and b and one obtains

$$a = \langle Y \rangle - b \langle X \rangle , \tag{E.65}$$

[5] One defines the scalar product between random variables $(X, Y) = \operatorname{cov}(X, Y)$ and therefore $\|X\|^2 = (X, X) = \operatorname{var}(X)$. The CAUCHY-SCHWARZ inequality reads

$$|(X, Y)|^2 \leq \| X \|^2 \| Y \|^2 , \tag{E.59}$$

and therefore $0 \leq r^2 \leq 1$.

where

$$b = \frac{\langle XY\rangle - \langle X\rangle \langle Y\rangle}{\langle X^2\rangle - \langle X\rangle^2} = \frac{\mathrm{cov}\,(X,Y)}{\mathrm{var}\,(X)} \,. \tag{E.66}$$

Thus, the linear function $f(X)$ which approximates the data points Y optimally is given by

$$\begin{aligned} f(X) &= \langle Y\rangle - \frac{\mathrm{cov}\,(X,Y)}{\mathrm{var}\,(X)}(X - \langle X\rangle) \\ &= \langle Y\rangle - r\sqrt{\frac{\mathrm{var}\,(Y)}{\mathrm{var}\,(X)}}(X - \langle X\rangle) \,, \end{aligned} \tag{E.67}$$

and it follows immediately for the squared error:

$$\left\langle [y - F(x)]^2 \right\rangle = \mathrm{var}\,(Y)\,(1 - r^2) \,. \tag{E.68}$$

Hence, the best result is achieved for $r = \pm 1$ in which case the association of the data points Y with the data points X is really linear while the worst result is found when $r = 0$ (no association what so ever).

E.11 Stable Distributions

A stable distribution is a distribution which reproduces itself [32]. In particular, consider two random variables X_1 and X_2 which are independent copies of the random variable X following the distribution p_X.[6]

The pdf p_X is referred to as a *stable distribution* if for arbitrary constants a and b the random variable $aX_1 + bX_2$ has the same distribution as the random variable $cX + d$ for some positive c and some $d \in \mathbb{R}$.

For this case one can write down the characteristic function analytically. We will give a special case, the so called *symmetric* Lévy *distributions* [38]:

$$G_\alpha(k) = \exp\left(-\sigma |k|^\alpha\right) \,. \tag{E.69}$$

Here $\sigma > 0$ and $0 < \alpha \leq 2$. The pdf of such a distribution shows the asymptotic behavior

$$p_\alpha(x) \propto \frac{\alpha}{|x|^{1+\alpha}}, \qquad |x| \to \infty \,. \tag{E.70}$$

[6] Independent copies of a random variable, are random variables, which are independent and follow the same distribution as the original random variable.

The normal distribution follows from Eq. (E.69) for $\alpha = 2$. Moreover, we observe from Eq. (E.70) that the variance diverges for all $\alpha < 2$. However, the existence of the variance was the criterion for the validity of the central limit theorem formulated in Sect. E.8. We note that stable distributions reproduce themselves and are attractors for sums of independent identical distributed random variables. This is referred to as the *generalized central limit theorem*.

We remark, in conclusion, that for $\alpha = 1$ the CAUCHY distribution results from Eq. (E.69), and note that stable distributions are also referred to as LÉVY *α-stable distributions* [32].

Appendix F
Phase Transitions

F.1 Some Basics

In many systems transitions between different phases can be observed if an external parameter, such as the temperature or the particle density, changes. Familiar examples are the liquid-gaseous phase transition or the ferromagnetic-paramagnetic transition. The two phases exhibit different properties and often develop a different physical structure, like in disorder-to-order transitions. This suggests the introduction of an *order parameter* φ which is zero in one phase and takes on some finite value $\varphi \neq 0$ in the other one.[1] For instance, in the case of paramagnetic-ferromagnetic transitions the magnetization plays the role of the order parameter [43].

In order to classify phase transitions we briefly repeat some basics from statistical mechanics [39–42, 44, 45]. In a canonical ensemble the probability to find the system in micro-state r (as a function of the external parameters temperature T, volume V and number of particles N) is proportional to the BOLTZMANN-factor

$$P_r(T,V,N) = \frac{1}{Z(T,V,N)} \exp\left[-\beta E_r(V,N)\right] \ . \tag{F.1}$$

Here, $\beta = 1/(k_\mathrm{B}T)$, where k_B is the BOLTZMANN constant, T is the temperature, and E_r is the energy of micro-state r. The canonical partition function $Z(T,V,N)$,

$$Z(T,V,N) = \sum_r \exp\left[-\beta E_r(V,N)\right] \ , \tag{F.2}$$

[1] The choice of the order parameter may not be unique [39–42].

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

ensures the normalization of $P_r(T, V, N)$ and determines the free energy $F(T, V, N)$ according to

$$F(T, V, N) = -\frac{1}{\beta} \ln Z(T, V, N) \ . \tag{F.3}$$

The EHRENFEST classification [46] of phase transitions is based on the behavior of F near the transition point: If F is a continuous function of its variables at the transition point and its first derivative with respect to some thermodynamic variable is discontinuous we call it a *first order phase transition*. For instance, transitions from the liquid to the gaseous phase are classified as first order phase transitions because the density, which is proportional to the first derivative of the free energy with respect to the chemical potential, changes discontinuously at the boiling temperature $T = T_B$. We remark the following characteristics of first order phase transitions:

1. The transition involves a *latent heat* ΔQ: The system absorbs or releases energy. A familiar example is the *latent heat of fusion* in the case of melting or freezing.
2. Both phases can coexist at the transition point.
3. A metastable phase can be observed.

In a *second order phase transition* the first derivative of the free energy F with respect to some thermodynamic variable is continuous but the second derivative of F exhibits a discontinuity. For instance, in a *ferromagnetic phase transition* the magnetization (first derivative of F with respect to the external magnetic field B) changes continuously while the magnetic susceptibility χ (the second derivative of F with respect to B) is discontinuous at the CURIE temperature T_c [43].

The modern classification is based on the behavior of the order parameter near the critical point. The order parameter changes discontinuously for first order phase transitions while it changes continuously for second and higher order phase transitions. Second order transitions are typically related to spontaneous symmetry breaking, as for instance in the paramagnetic-ferromagnetic transition. Based on this observation, LANDAU developed a general description of second order phase transitions which we briefly discuss in the following section.

F.2 LANDAU Theory

We regard a second order phase transition characterized by the scalar order parameter φ [47]. Since $\langle\varphi\rangle$ changes continuously at $T = T_c$, it is convenient to define φ in such a way that $\langle\varphi\rangle\,|_{T\geq T_c} = 0$ while $\langle\varphi\rangle\,|_{T<T_c} \neq 0$.

For the free energy $F(T, h, \varphi)$, one chooses

$$F(T, h, \varphi) = F_0(T) + V\left[\frac{a(T - T_c)}{2}\varphi^2 + \frac{b}{4}\varphi^4 - h\varphi\right] , \tag{F.4}$$

where a and b are some material constants and h denotes the external field. This ansatz is motivated by the theory of the paramagnetic-ferromagnetic phase transition [43]. Thus, in equilibrium we have

$$\frac{\delta F}{\delta \varphi} = 0 , \tag{F.5}$$

which results in

$$a(T - T_c)\varphi + b\varphi^3 = h . \tag{F.6}$$

For $h = 0$ and $T < T_c$ we obtain

$$\langle \varphi_0 \rangle = \sqrt{\frac{a}{b}(T_c - T)} \sim (T_c - T)^{\gamma} , \tag{F.7}$$

where $\gamma = 1/2$ is called the *critical exponent*. For $T \geq T_c$ we have $\langle \varphi_0 \rangle = 0$. We now regard a weak external field h. The order parameter will change

$$\varphi = \langle \varphi_0 \rangle + \delta\varphi . \tag{F.8}$$

Again, we obtain for equilibrium:

$$\frac{\delta F}{\delta \varphi} = a(T - T_c)(\langle \varphi_0 \rangle + \delta\varphi) + b(\langle \varphi_0 \rangle + \delta\varphi)^3 - h = 0 . \tag{F.9}$$

Neglecting contributions of order $\mathcal{O}(\delta\varphi^2)$ yields for the susceptibility

$$\chi = \frac{\partial}{\partial h}\langle \varphi \rangle = \frac{\langle \delta\varphi \rangle}{h} \sim |T - T_c|^{\delta} , \tag{F.10}$$

where $\delta = -1$ is a second critical exponent. This is the CURIE-WEISS law [43]. Finally for $T = T_c$ we obtain from Eq. (F.6)

$$\varphi = \left(\frac{h}{b}\right)^{\frac{1}{3}} \sim h^{\frac{1}{\epsilon}} , \tag{F.11}$$

with the third critical exponent ϵ. The LANDAU theory is a mean-field approximation since local fluctuations of the order parameter are neglected.[2] Although the critical exponents obtained with LANDAU's approach deviate from experimental values, the theory is qualitatively correct. We remark that the critical exponents are universal (a property referred to as *universality* [40]) as they depend only on the dimensionality and the symmetry of the interaction.

[2]The extension to space dependent order parameters is referred to as GINZBURG-LANDAU theory [48].

Appendix G
Fractional Integrals and Derivatives in 1D

This section introduces briefly the common definitions and notations associated with fractional calculus in one dimension [49].

The RIEMANN-LIOUVILLE fractional integrals of order $\alpha \in \mathbb{C}$ $[\Re(\alpha) > 0]$, $I^{\alpha}_{a+}f(x)$ and $I^{\alpha}_{b-}f(x)$ on a finite interval $[a, b]$ on the real axis $\mathbb{R}$ are given by

$$I^{\alpha}_{a+}f(x) := \frac{1}{\Gamma(\alpha)} \int_a^x \mathrm{d}x' \, \frac{f(x')}{(x-x')^{1-\alpha}} \quad \text{for} \quad (x > a,\ \Re(\alpha) > 0), \tag{G.1a}$$

$$I^{\alpha}_{b-}f(x) := \frac{1}{\Gamma(\alpha)} \int_x^b \mathrm{d}x' \, \frac{f(x')}{(x'-x)^{1-\alpha}} \quad \text{for} \quad (x < b,\ \Re(\alpha) > 0)\ , \tag{G.1b}$$

where $\Gamma(x)$ denotes the Gamma function [36, 37], $\Re(\alpha)$ is the real part of α, and $f(x)$ is a sufficiently well behaved continuous, differentiable function for which the integrals in (G.1) exist. The corresponding RIEMANN-LIOUVILLE fractional derivatives $D^{\alpha}_{a+}f(x)$ and $D^{\alpha}_{b-}f(x)$ of order $\alpha \in \mathbb{C}$ $[\Re(\alpha) \geq 0]$ are defined by

$$\begin{aligned} D^{\alpha}_{a+}f(x) &:= \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^n (I^{n-\alpha}_{a+}f)(x) \\ &= \frac{1}{\Gamma(n-\alpha)} \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^n \int_a^x \mathrm{d}x' \, \frac{f(x')}{(x-x')^{\alpha-n+1}} \quad \text{for} \quad x > a\ , \end{aligned} \tag{G.2a}$$

and

$$\begin{aligned} D^{\alpha}_{b-}f(x) &:= \left(-\frac{\mathrm{d}}{\mathrm{d}x}\right)^n (I^{n-\alpha}_{b-}f)(x) \\ &= \frac{1}{\Gamma(n-\alpha)} \left(-\frac{\mathrm{d}}{\mathrm{d}x}\right)^n \int_x^b \mathrm{d}x' \, \frac{f(x')}{(x'-x)^{\alpha-n+1}} \quad \text{for} \quad x < b\ , \end{aligned} \tag{G.2b}$$

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

with $n = [\Re(\alpha)] + 1$. Here $[\Re(\alpha)]$ denotes the integer part of $\Re(\alpha)$. For $a \to -\infty$ and $b \to +\infty$ the RIEMANN-LIOUVILLE fractional integrals and derivatives are referred to as WEYL fractional integrals and derivatives. In what follows, they will be denoted by $I^{\alpha}_{\pm}$ and $D^{\alpha}_{\pm}$, respectively.

If $\alpha \in \mathbb{C}\ [\Re(\alpha) \geq 0]$ and $[a, b] \in \mathbb{R}$ is a finite interval, then the left- and right-sided CAPUTO fractional derivatives ${}^{C}D^{\alpha}_{a+}f(x)$ and ${}^{C}D^{\alpha}_{b-}f(x)$ are defined by

$$ {}^{C}D^{\alpha}_{a+}f(x) = D^{\alpha}_{a+}f(x) - \sum_{k=0}^{n-1} \frac{f^{(k)}(a)}{\Gamma(k-\alpha+1)}(x-a)^{k-\alpha} , \tag{G.3a} $$

and

$$ {}^{C}D^{\alpha}_{b-}f(x) = D^{\alpha}_{b-}f(x) - \sum_{k=0}^{n-1} \frac{(-1)^k f^{(k)}(b)}{\Gamma(k-\alpha+1)}(b-x)^{k-\alpha} , \tag{G.3b} $$

with

$$ n = \begin{cases} [\Re(\alpha)] + 1 & \alpha \notin \mathbb{N} , \\ \alpha & \alpha \in \mathbb{N}_0 . \end{cases} \tag{G.3c} $$

This is, however, equivalent to

$$ \begin{aligned} {}^{C}D^{\alpha}_{a+}f(x) &= \frac{1}{\Gamma(n-\alpha)} \int_a^x \mathrm{d}x' \, \frac{f^{(n)}(x')}{(x-x')^{\alpha-n+1}} \\ &= (I^{n-\alpha}_{a+} D^n f)(x) , \end{aligned} \tag{G.4a} $$

and

$$ \begin{aligned} {}^{C}D^{\alpha}_{b-}f(x) &= \frac{(-1)^n}{\Gamma(n-\alpha)} \int_x^b \mathrm{d}x' \, \frac{f^{(n)}(x')}{(x'-x)^{\alpha-n+1}} \\ &= (-1)^n (I^{n-\alpha}_{b-} D^n f)(x) . \end{aligned} \tag{G.4b} $$

The symmetric fractional integrals $I^{\alpha}_{|x|}$ and derivatives $\mathscr{D}^{\alpha}_{|x|}$ are referred to as RIESZ fractional integrals or derivatives and are of the form

$$ I^{\alpha}_{|x|} = \frac{I^{\alpha}_{+} + I^{\alpha}_{-}}{2\cos\left(\frac{\alpha\pi}{2}\right)} , \tag{G.5} $$

for $\alpha \in (0, 1)$ and

$$ \mathscr{D}^{\alpha}_{|x|} = \begin{cases} (-1)^{\frac{n}{2}} \frac{D^{\alpha}_{+} + D^{\alpha}_{-}}{2\cos(\alpha\pi/2)} & \text{for} \quad n = [\Re(\alpha)] + 1 \equiv 2k,\ k \in \mathbb{N}_0 , \\ (-1)^{\frac{n-1}{2}} \frac{D^{\alpha}_{+} - D^{\alpha}_{-}}{2\sin(\alpha\pi/2)} & \text{for} \quad n = [\Re(\alpha)] + 1 \equiv 2k+1,\ k \in \mathbb{N}_0 . \end{cases} \tag{G.6} $$

Appendix H
Least Squares Fit

H.1 Motivation

In numerous physics applications a set of corresponding data points (x_k, y_k) was measured or calculated and a set of certain parameters $\{\alpha_j\}$ characterizing a function $f(x_k, \{\alpha_j\})$ is to be determined in such a way that

$$\chi^2 = \sum_k c_k \left[y_k - f(x_k, \{\alpha_j\}) \right]^2 \to \min . \tag{H.1}$$

This is referred to as a *least squares fit* problem [6, 7]. Here, $c_k \geq 0$ are weights, which indicate the relevance of a certain data point (x_k, y_k) for the fitting routine, and $f(x, \{\alpha_j\})$ is referred to as the *model function*. Besides numerous applications within the context of experimentally obtained data points, we already came across such a problem in our discussion of data analysis in Chap. 19. Here it was of interest to determine the *experimental auto-correlation time* by fitting an exponential function to the measured auto-correlation coefficient $A(k)$, discussed in Sect. 19.3. Hence, we note that in many applications the parameters $\{\alpha_j\}$ can be associated with a physical property of interest.

We distinguish between two different cases: (i) the function $f(x_k, \{\alpha_j\})$ is a linear function of the parameters $\{\alpha_j\}$ and (ii) the function $f(x_k, \{\alpha_j\})$ is not linear in its parameters $\{\alpha_j\}$. It should be emphasized that in both cases the function does not need to be linear in x_k. This section will discuss methods for linear as well as non-linear least squares fits. However, before proceeding some comments on the data points $\{y_k\}$ seem to be required.

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

Suppose the points (x_k, y_k) stem from a measurement which has been repeated N-times. In this case for every value x_k we have N different values $\{y_k^j\}$ and we may use the arithmetic mean

$$\overline{y_k} = \frac{1}{N} \sum_j y_k^j \,, \tag{H.2}$$

instead of y_k in expression (H.1). We may also calculate the variance $\mathrm{var}\,(y_k)$ via[1]

$$\mathrm{var}\,(y_k) = \frac{1}{N} \sum_j (y_k^j - \overline{y_k})^2 \,. \tag{H.3}$$

If we assume that the data points y_k^j follow a normal distribution with mean $\langle y_k \rangle$ and variance $\mathrm{var}\,(y_k)$ we may proceed in the following way: The weights c_k are chosen as

$$c_k = \frac{1}{\mathrm{var}\,(y_k)} \,. \tag{H.4}$$

The resulting fit parameters $\{\alpha_j\}$ are then regarded as mean values of parameters where the variances $\mathrm{var}\,(\alpha_i)$ as well as the covariances $\mathrm{cov}\,(\alpha_i, \alpha_j)$ can be obtained from the matrix

$$N_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \alpha_i \partial \alpha_j} \,, \tag{H.5}$$

via inversion, i.e.

$$C = N^{-1} \,, \tag{H.6}$$

and

$$C_{ij} = \mathrm{cov}\,(\alpha_i, \alpha_j) \,. \tag{H.7}$$

The matrix C is commonly referred to as *covariance matrix*.

[1] In many cases one employs the *bias corrected variance* $\mathrm{var}\,(y_k)_B = \frac{N}{N-1} \mathrm{var}\,(y_k)$. For a detailed discussion of the bias corrected variance the interested reader is encouraged to consult a statistics textbook [45, 50–53].

H.2 Linear Least Squares Fit

In this particular case the model function $f(x_k, \{\alpha_j\})$ is defined as

$$f(x_k, \{\alpha_j\}) = \sum_j \alpha_j \varphi_j(x_k) \ , \tag{H.8}$$

where $\varphi_j(x_k)$ are linear independent basis functions, which do not have to be linear in x_k. The particular case of a *linear regression*, discussed in Sect. E.10, is included. Equation (H.8) specifies the model function $f(x_k, \{\alpha_j\})$ in (H.1) and this yields

$$\chi^2 = \sum_k c_k \left[y_k - \sum_j \alpha_j \varphi_j(x_k) \right]^2 \ , \tag{H.9}$$

which is supposed to tend to a minimum. We calculate

$$\frac{\partial \chi^2}{\partial \alpha_\ell} = -2 \sum_k c_k \varphi_\ell(x_k) \left[y_k - \sum_j \alpha_j \varphi_j(x_k) \right] \stackrel{!}{=} 0 \ , \tag{H.10}$$

and arrive at:

$$\sum_j \alpha_j \sum_k c_k \varphi_\ell(x_k) \varphi_j(x_k) = \sum_k c_k y_k \varphi_\ell(x_k) \ , \quad \forall \ell . \tag{H.11}$$

This equation can be reformulated as the linear equation

$$M\alpha = \beta \ , \tag{H.12}$$

where the vectors $\alpha = (\alpha_1, \alpha_2, \ldots)^T$ and $\beta = (\beta_1, \beta_2, \ldots)^T$ with

$$\beta_i = \sum_k c_k y_k \varphi_i(x_k) \ , \tag{H.13}$$

and the matrix M

$$M_{ij} = \sum_k c_k \varphi_i(x_k) \varphi_j(x_k) \ , \tag{H.14}$$

have been introduced.

Equation (H.12) can, for instance, be solved with the help of the methods discussed in Appendix C. It is also particularly simple to determine the covariances because we have

$$N_{ij} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial\alpha_i \partial\alpha_j} = M_{ij} \,, \tag{H.15}$$

and the covariances follow from Eqs. (H.6) and (H.7).

H.3 Nonlinear Least Squares Fit

Before we discuss the most general case of a completely arbitrary model function $f(x_k, \{\alpha_j\})$ we want to point out that it is in most cases of advantage to linearize the model function if at all possible. For instance, if the model function is an exponential function, it may be linearized by taking the data points $\ln(y_k)$ instead of y_k.

However, if this is not possible there are numerous alternatives to find a solution of the problem. For instance, the GAUSS-NEWTON method can be employed if the model function $f(x_k, \{\alpha_j\})$ and its derivatives with respect to the parameters α_j are known analytically. Another possibility is offered by the application of an deterministic optimization algorithm as they will be introduced in Appendix I. If even this method is not applicable, the methods of stochastic optimization, discussed in Chap. 20, might be an obvious choice.

We describe now the GAUSS-NEWTON method which is essentially a generalization of the NEWTON method presented in Appendix B. The GAUSS-NEWTON method is a method developed to minimize the expression (H.1) iteratively. The derivatives

$$\frac{\partial f(x_k, \{\alpha_j\})}{\partial\alpha_\ell} \,, \tag{H.16}$$

are assumed to be known analytically. This is an iterative algorithm and, thus, an iteration index is introduced and indicated by a superscript index n like in α_j^n. The algorithm is described by the following steps:

1. Choose a set of initial values $\{\alpha_j^0\}$ for the iteration.
2. Linearize the function $f(x_l, \{\alpha_j^n\})$ and insert the result into Eq. (H.1):

$$\chi^2 \approx \sum_k c_k \left\{ y_k - f(x_k, \{\alpha_j^n\}) - \sum_\ell \left[\frac{\partial f(x_k, \{\alpha_j\})}{\partial\alpha_\ell} \right]_{\{\alpha_j\}=\{\alpha_j^n\}} (\alpha_\ell - \alpha_\ell^n) \right\}^2 . \tag{H.17}$$

We introduce the following abbreviations for a more compact notation:

$$\mathrm{d}f_{k,\ell}^n = \left[\frac{\partial f(x_k, \{\alpha_j\})}{\partial \alpha_\ell}\right]_{\{\alpha_j\}=\{\alpha_j^n\}} , \tag{H.18}$$

and

$$f_k^n = f(x_k, \{\alpha_j^n\}) . \tag{H.19}$$

3. We have to solve

$$\frac{\partial \chi^2}{\partial \alpha_i} = -2 \sum_k c_k \mathrm{d}f_{k,i}^n \left[y_k - f_k - \sum_\ell \mathrm{d}f_{k,\ell}^n (\alpha_\ell - \alpha_\ell^n) \right] \stackrel{!}{=} 0 , \tag{H.20}$$

for all parameters $\{\alpha_j\}$. Therefore, we introduce vectors $\alpha = (\alpha_1, \alpha_2, \ldots)^T$, $\beta = (\beta_1, \beta_2, \ldots)^T$ with

$$\beta_i = \sum_k c_k (y_k - f_k^n) \mathrm{d}f_{k,i}^n , \tag{H.21}$$

and the matrix M with elements:

$$M_{ij} = \sum_k c_k \mathrm{d}f_{k,i}^n \mathrm{d}f_{k,j}^n . \tag{H.22}$$

This transforms Eq. (H.20) into a linear system of equations

$$M(\alpha - \alpha^n) = \beta , \tag{H.23}$$

which is solved for $\Delta\alpha^n = \alpha - \alpha^n$. Please note that α^n denotes the vector α after n iterations. The vector α^{n+1} for the next iteration step is guessed from:

$$\alpha^{n+1} = \alpha^n + \Delta\alpha^n . \tag{H.24}$$

4. The iteration is terminated if for all parameters the desired accuracy was achieved. For instance, the condition $|\alpha_j^{n+1} - \alpha_j^n| \leq \epsilon$ can be used with ϵ a small parameter. A criterion for the relative error can be formulated in analogue.

Some comments concerning the covariance matrix are in order: It is more complicated in the nonlinear case because we also have to consider the second partial derivatives of the model function $f(x_k, \{\alpha_\ell\})$. However, if these can for some reason be neglected we obtain, again, that $N_{ij} = M_{ij}$, as in Appendix, Sect. H.2. Another, more serious problem is found in the fact that the GAUSS-NEWTON method suffers from severe instability problems. However, a possible remedy was formulated by D. MARQUART [54] who suggested to multiply the diagonal elements

of the matrix M with a factor $(1 + \lambda)$ where $\lambda > 0$. A detailed analysis shows that one can choose λ sufficiently large and in such a way that the value of χ_n^2 decreases monotonically, i.e. $\chi_{n+1}^2 \leq \chi_n^2$ for all iteration steps n. However, an increase of λ decreases the convergence rate and more iterations are necessary until the required accuracy was obtained. It is therefore desirable to choose λ values in such a way that the error decreases monotonically but that, at the same time, a convergence rate is maintained which is as large as possible. A possible strategy is to start with some given value of λ and to reduce it after every iteration step by a constant rate. However, if at some point the error χ^2 increases, i.e. $\chi_{n+1}^2 > \chi_n^2$, then λ has again to be increased.

Appendix I
Deterministic Optimization

I.1 Introduction

We use the term *deterministic optimization* to distinguish these particular optimization methods from the stochastic optimization methods discussed in Chap. 20. There are numerous different deterministic methods designed to find the minimum (or maximum) of a given function $f(x)$, where x can be a vector. Roughly speaking, we can distinguish between methods which require the knowledge of the Hessian,[1] methods which need gradients only, and methods which are based on function values only. For instance, if the gradient of a function is known analytically one may exploit NEWTON's method, as it was introduced in Appendix B. Note that such an approach requires the Hessian of the function $f(x)$.

We plan to discuss here in some detail two specific methods, namely the method of *steepest descent* and the method of *conjugate gradients*. Both methods require the knowledge of the gradient of the function, however, the gradient can also be approximated with the help of finite differences (see Chap. 2). A discussion of additional methods is beyond the scope of this book and the interested reader is referred to the available literature [55].

However, before discussing these two methods in more detail, let us briefly consider the quadratic problem which can be solved analytically. In this case the function $f(x)$ can be written as

$$f(x) = \frac{1}{2} x^T A x - b^T x + c \,, \tag{I.1}$$

[1] The Hessian, or HESSE matrix, $H \in \mathbb{R}^{N \times N}$ of a function $f(x)$, $x \in \mathbb{R}^N$ is the Jacobian of the Jacobian $J(x)$ of $f(x)$ defined in Eq. (B.8). Thus, it is the matrix of second order partial derivatives of a function. It describes the local curvature of a function of many variables.

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

where $x \in \mathbb{R}^N, A \in \mathbb{R}^{N\times N}, b \in \mathbb{R}^N$ and $c \in \mathbb{R}$ where we restrict the discussion to real valued functions for reasons of simplicity. We demonstrate now that for symmetric and positive definite matrices A, i.e. $A^T = A$ and $x^T Ax > 0$ for all $x \neq 0$, the minimum of $f(x)$ is given by $x = A^{-1}b$. The gradient of $f(x)$ is readily evaluated and is given by[2]:

$$\nabla f(x) = \frac{1}{2}A^T x + \frac{1}{2}Ax - b \; . \tag{I.2}$$

This immediately yields the desired result:

$$Ax = b \; . \tag{I.3}$$

It follows that $x = A^{-1}b$ is a minimum because we assumed A to be positive definite. It is possible to solve the optimization problem even if A is not symmetric by inverting the symmetrized matrix $\left(A + A^T\right)/2$. Finally, the linear equation (I.3) can be solved with the methods discussed in Appendix C.

I.2 Steepest Descent

The most simple gradient based method is the method of *steepest descent* [6]. It is based on the rather straight forward idea of moving in each iteration step into the opposite direction of the gradient, i.e. *downhill*. Hence, we may formulate it mathematically in the following way: Let x_n be the current position of our search for the minimum. Then we choose

$$x_{n+1} = x_n - \alpha_n \nabla f(x_n) \; , \tag{I.4}$$

where the step-size in direction of the negative gradient, α_n, has to be determined in an additional step. The step-size should be chosen in such a way that we reach the line minimum in direction $\nabla f(x_n)$:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha_n} f[x_{n+1}(\alpha_n)] = -\nabla f(x_{n+1}) \cdot \nabla f(x_n) \stackrel{!}{=} 0 \; . \tag{I.5}$$

[2] We remember from vector analysis that

$$\nabla_x \left(x^T Ax\right) = \underbrace{\nabla_x \left(x^T A\right)}_{=A} x + \underbrace{\nabla_x \left(x^T A^T\right)}_{=A^T} x = (A + A^T)x \; .$$

Hence, we observe that for an optimal choice of α_n the search directions are orthogonal. In practice α_n is estimated with the help of a separate minimization technique, such as bisection. This technique has already been used in our discussion of the shooting methods in Chap. 10.

We provide an example which is supposed to make the method more transparent and to help in the discussion of its caveats: We want to determine the global minimum of the function

$$f(x, y) = \cos(2x) + \sin(4y) + \exp(1.5x^2 + 0.7y^2) + 2x \; . \tag{I.6}$$

Its gradient is easily evaluated

$$\frac{\partial f(x, y)}{\partial x} = -2\sin(2x) + 3x\exp(1.5x^2 + 0.7y^2) + 2 \; , \tag{I.7}$$

and

$$\frac{\partial f(x, y)}{\partial y} = 4\cos(4y) + 1.4y\exp(1.5x^2 + 0.7y^2) \; . \tag{I.8}$$

We define the algorithm *steepest descent* with the following steps:

1. Choose some initial values x_0 and y_0.
2. Calculate the gradient $\nabla f(x_n, y_n)$ in iteration step n.
3. Determine α_n in such a way that

$$f[x_{n+1}(\alpha_n), y_{n+1}(\alpha_n)] \to \min \; , \tag{I.9}$$

which is equivalent to

$$g(\alpha_n) := \nabla f[x_{n+1}(\alpha_n), y_{n+1}(\alpha_n)] \cdot \nabla f(x_n, y_n) = 0 \; . \tag{I.10}$$

This is achieved by a bisection technique similar to the one employed in Sect. 10.3,

a. Set $\alpha_n^a = 0$ and chose α_n^b arbitrary.
b. Increase α_n^b until $g(\alpha_n^a)g(\alpha_n^b) < 0$.
c. Define

$$\alpha_n^c = \frac{\alpha_n^a + \alpha_n^b}{2} \; , \tag{I.11}$$

and determine $g(\alpha_n^c)$.

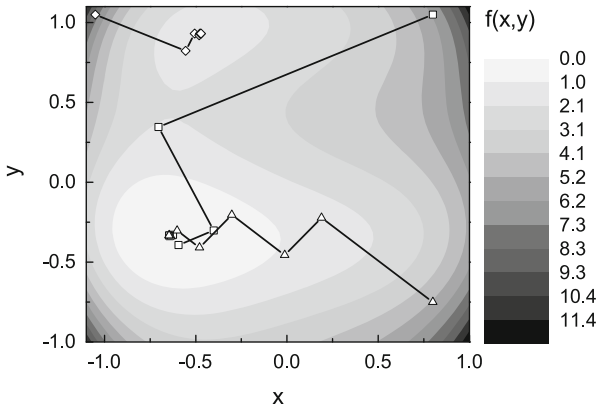

Fig. I.1 Iteration sequence of the method of steepest descent for three different starting points

d. If $g(\alpha_n^a)g(\alpha_n^c) < 0$, set $\alpha_n^b = \alpha_n^c$ and return to step c. Otherwise, set $\alpha_n^a = \alpha_n^c$ and return to step c.
e. The bisection is terminated if $|g(\alpha_n^a)| < \epsilon$, with ϵ some required accuracy for the bisection part.

4. Check whether $|f(x_{n+1}, y_{n+1}) - f(x_n, y_n)| \leq \eta$ with η some required accuracy. Return to step 2 for the next iteration step if the algorithm is not converged.

The above algorithm was executed for the function $f(x, y)$ given by Eq. (I.6) for three different starting points, $(0.8, -0.75)$, $(0.8, 1.05)$, and $(-1.05, 1.05)$. The function $f(x, y)$ as well as the iteration sequence towards the minimum for all three starting points is illustrated in Fig. I.1.

We note the following properties of the method: First of all it is a rather slow method due to the orthogonality of subsequent search directions. Moreover, as we observe from Fig. I.1, we can only find the local minimum closest to the starting point and not the global minimum of the function $f(x, y)$. The convergence rate is also highly affected by the choice of the initial position. However, it is a very simple method which works in spaces of arbitrary dimension.

I.3 Conjugate Gradients

The method of *conjugate gradients* [6, 55] is based on the definition of N orthogonal search directions $\{\psi_i\}$ in an N dimensional space. In contrast to steepest descent it is designed in such a way that we take only *one* step in each search direction and the minimum is found after at most N steps, if the function $f(x)$ is of the quadratic form (I.1). In the more general case, however, it will take more steps but will, nevertheless, be much more efficient than the method of steepest descent. Let us formulate the method for a general function $f(x)$.

We approximate the function $f(x)$, with $x \in \mathbb{R}^N$, in the vicinity of the reference point x_n in the n-th iteration step up to second order and name the resulting function $\hat{f}(x)$:

$$\begin{aligned}\hat{f}_n(x) &:= f(x_n) + \nabla f(x_n) \cdot (x - x_n) + \frac{1}{2}(x - x_n) \cdot [\Delta f(x_n)(x - x_n)] \\ &\equiv f(x_n) - b_n^T(x - x_n) + \frac{1}{2}(x - x_n)^T A_n(x - x_n) \ . \end{aligned} \tag{I.12}$$

Here, A_n denotes the Hessian[3] at position x_n and b_n is the negative gradient at x_n. In particular, for a quadratic function $f(x)$ the equality $\hat{f}(x) = f(x)$ holds. We now write the minimum $\hat{x}$ of $f(x)$ as a linear combination of search directions $\{\psi_i\}$ with coefficients λ_i and the initial point x_0:

$$\hat{x} = x_0 + \sum_{i=0}^{M} \lambda_i \psi_i \ . \tag{I.13}$$

Note that in the quadratic case (I.1) this sum will be restricted to $M = N - 1$. At each iteration instance we have the relation

$$x_{n+1} = x_n + \lambda_n \psi_n \ , \tag{I.14}$$

together with the goal

$$x_M \stackrel{!}{=} \hat{x} \ . \tag{I.15}$$

Let us define now a couple of useful quantities. The deviation from the minimum at iteration step $n + 1$, δ_{n+1}, is given by

$$\begin{aligned}\delta_{n+1} &= x_{n+1} - \hat{x} \\ &= x_n + \lambda_n \psi_n - \hat{x} \\ &= \delta_n + \lambda_n \psi_n \ . \end{aligned} \tag{I.16}$$

We define, furthermore, the residual

$$\begin{aligned}r_{n+1} &:= -\nabla \hat{f}_n(x_{n+1}) \\ &= b_n - A_n(x_{n+1} - x_n) \\ &= b_n - \lambda_n A_n \psi_n \ , \end{aligned} \tag{I.17}$$

[3] Note that the Hessian is always symmetric for real valued functions $f(x)$ due to the symmetry of second order derivatives.

where we employed that

$$\frac{1}{2}\nabla\left[(x-x_n)^T A_n (x-x_n)\right] = A_n(x-x_n)\ . \tag{I.18}$$

Finding the minimum of the quadratic approximation $\hat{f}(x)$ of $f(x)$ around x_n is equivalent to the condition

$$r_{n+1} = 0\ . \tag{I.19}$$

In particular, we have to find the product $\lambda_n \psi_n$ in such a way that $r_{n+1} = 0$. Of course, we could invert the Hessian A_n in order to obtain this result. However, this would be too expensive from a computational point of view. The idea is to apply the *ideal* search strategy for quadratic functions to $\hat{f}_n(x)$ in order to obtain x_{n+1}. Hence, the method of conjugate gradients executes packages of N steps, where each package solves the quadratic problem around x_n, until the minimum of the original function $f(x)$ has been found. Therefore, we have to generalize the relations (I.14), (I.16), and (I.17) for iterations within step n.

We have, in particular, for every iteration step n

$$x_{n+1} = x_n + \sum_{\ell=0}^{N-1} \lambda_n^\ell \psi_n^\ell\ , \tag{I.20}$$

together with the definitions

$$x_n^{\ell+1} = x_n^\ell + \lambda_n^\ell \psi_n^\ell\ , \tag{I.21}$$

where $x_{n+1} \equiv x_n^N$. Furthermore, we define the deviation

$$\delta_n^{\ell+1} = x_n^{\ell+1} - x_{n+1} = \delta_n^\ell + \lambda_n^\ell \psi_n^\ell\ , \tag{I.22}$$

and the residual

$$\begin{aligned} r_n^{\ell+1} &= -\nabla \hat{f}_n(x_n^{\ell+1}) \\ &= b_n - A_n(x_n^{\ell+1} - x_n)\ . \end{aligned} \tag{I.23}$$

In contrast to relation (I.17), Eq. (I.23) features the difference $(x_n^{\ell+1} - x_n)$ rather than $(x_{n+1} - x_n)$. We insert the recurrence (I.21) and obtain

$$\begin{aligned} r_n^{\ell+1} &= b_n - A_n(x_n^\ell - x_n) - \lambda_n^\ell A_n \psi_n^\ell \\ &= r_n^\ell - \lambda_n^\ell A_n \psi_n^\ell\ . \end{aligned} \tag{I.24}$$

Hence, in contrast to relation (I.17) Eq. (I.24) defines a recurrence relation. Again, we want to choose the search directions ψ_n^ℓ and the step length λ_n^ℓ in such a way that we find the minimum as quickly as possible. Suppose we already knew the search direction ψ_n^ℓ. The line minimum in this direction is then given by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda_n^\ell}\hat{f}_n(x_n^{\ell+1}) &= \nabla\hat{f}(x_n^{\ell+1})\cdot\psi_n^\ell \\ &= -r_n^{\ell+1}\cdot\psi_n^\ell \\ &= -(r_n^\ell - \lambda_n^\ell A_n\psi_n^\ell)^T\psi_n^\ell \\ &= -(r_n^\ell)^T\psi_n^\ell + \lambda_n^\ell(\psi_n^\ell)^T A_n\psi_n^\ell \\ &\stackrel{!}{=} 0\,, \end{aligned} \tag{I.25}$$

and we have

$$\lambda_n^\ell = \frac{(r_n^\ell)^T\psi_n^\ell}{(\psi_n^\ell)^T A_n\psi_n^\ell}\,. \tag{I.26}$$

Hence, the remaining unknown quantities in our algorithm are the search directions ψ_n^ℓ. So far, the only information we obtained is that the search direction ψ_n^ℓ is orthogonal to the residual $r_n^{\ell+1}$, see Eq. (I.25).

However, we also know that

$$\begin{aligned} 0 &= A_n\,(x_{n+1} - x_n) - b_n \\ &= A_n\sum_{\ell=0}^{N-1}\lambda_n^\ell\psi_n^\ell - b_n\,, \end{aligned} \tag{I.27}$$

and therefore

$$0 = \left(\psi_n^k\right)^T A_n\sum_{\ell=0}^{N-1}\lambda_n^\ell\psi_n^\ell - \left(\psi_n^k\right)^T b_n\,, \tag{I.28}$$

for arbitrary k. A sufficient condition to ensure the validity of relation (I.28) is to impose A_n-*orthogonality*:

$$\left\langle\psi_n^k \,\middle|\, \psi_n^\ell\right\rangle_{A_n} \equiv (\psi_n^k)^T A_n\psi_n^\ell = \delta_{k,\ell}\left\langle\psi_n^k \,\middle|\, \psi_n^k\right\rangle_{A_n}\,. \tag{I.29}$$

We note that $\left\langle\psi_n^k \,\middle|\, \psi_n^\ell\right\rangle_A$ constitutes indeed a scalar product since A_n is positive definite in the neighborhood of a minimum.

Let us briefly demonstrate that the choice (I.29) fulfills Eq. (I.28). First of all we note that we obtain from Eq. (I.24)

$$r_n^{\ell+1} = b_n - \sum_{k=0}^{\ell} \lambda_n^k A_n \psi_n^k \,, \tag{I.30}$$

and, therefore, we derive the coefficients λ_n^ℓ from Eq. (I.26) in the convenient form:

$$\lambda_n^\ell = \frac{b_n^T \psi_n^\ell}{\langle \psi_n^\ell \mid \psi_n^\ell \rangle_{A_n}} \,. \tag{I.31}$$

The condition of orthogonality (I.29) is used to rewrite Eq. (I.28) as

$$0 = \lambda_n^k \langle \psi_n^k \mid \psi_n^k \rangle_{A_n} - \left(\psi_n^k\right)^T b_n \,, \tag{I.32}$$

which together with Eq. (I.31) proves the equality (I.28). Hence, the strategy is clear: We choose an initial direction ψ_n^0 and then construct the further directions in such a way that they fulfill A_n-orthogonality (I.29). Before discussing the construction of search directions in more detail we observe the following property:

$$(\psi_n^k)^T r_n^\ell = (\psi_n^k)^T b_n - \sum_{m=0}^{\ell-1} \lambda_n^m \langle \psi_n^k \mid \psi_n^m \rangle_{A_n} = \begin{cases} (\psi_n^k)^T b_n & \text{for } k \geq \ell \,, \\ 0 & \text{else.} \end{cases} \tag{I.33}$$

This means that all search directions ψ_n^k for $k \leq \ell - 1$ are orthogonal to the residual r_n^ℓ, or in other words, all residuals r_n^ℓ are orthogonal (in the classical sense) to all previous search directions.

We shall now briefly outline the resulting update algorithm for search directions: Let $\{\varphi_n^\ell\}$ be a set of linear independent vectors that span our search space for $\hat{f}_n(x)$.[4] We write the search direction ψ_n^k as

$$\psi_n^k = \varphi_n^k + \sum_{\ell=0}^{k-1} \beta_n^{k\ell} \psi_n^\ell \,, \tag{I.34}$$

together with

$$\psi_n^0 = \varphi_n^0 \,. \tag{I.35}$$

[4] In principle these linear independent vectors $\{\varphi_n^\ell\}$ do not need to depend on the index n, i.e. on the actual position x_n. However, we consider here the most general case as will soon become clear.

The expansion coefficients $\beta_n^{\ell k}$ can be determined recursively by imposing A_n-orthogonality for all $\ell < k$:

$$\begin{aligned} 0 &= \left\langle \psi_n^k \,\middle|\, \psi_n^\ell \right\rangle_{A_n} \\ &= \left\langle \varphi_n^k \,\middle|\, \psi_n^\ell \right\rangle_{A_n} + \sum_{m=0}^{k-1} \beta_n^{km} \left\langle \psi_n^m \,\middle|\, \psi_n^\ell \right\rangle_{A_n} \\ &= \left\langle \varphi_n^k \,\middle|\, \psi_n^\ell \right\rangle_{A_n} + \beta_n^{k\ell} \left\langle \psi_n^\ell \,\middle|\, \psi_n^\ell \right\rangle_{A_n} , \end{aligned} \tag{I.36}$$

and, therefore:

$$\beta_n^{k\ell} = -\frac{\left\langle \varphi_n^k \,\middle|\, \psi_n^\ell \right\rangle_{A_n}}{\left\langle \psi_n^\ell \,\middle|\, \psi_n^\ell \right\rangle_{A_n}} . \tag{I.37}$$

This procedure is known as the GRAM-SCHMIDT *conjugation* [6, 12].

Now, the question arises how one should choose the basis vectors φ_n^ℓ and whether or not it is advantageous to choose the φ_n^ℓ as a function of n. A particularly clever choice is to take the residuals, i.e.

$$\varphi_n^\ell = r_n^\ell . \tag{I.38}$$

In this case we have for $\ell < k$

$$\begin{aligned} \beta_n^{k\ell} &= -\frac{\left\langle r_n^k \,\middle|\, \psi_n^\ell \right\rangle_{A_n}}{\left\langle \psi_n^\ell \,\middle|\, \psi_n^\ell \right\rangle_{A_n}} \\ &= -\frac{(r_n^k)^T A_n \psi_n^\ell}{\left\langle \psi_n^\ell \,\middle|\, \psi_n^\ell \right\rangle_{A_n}} \\ &= -\frac{(r_n^k)^T}{\left\langle \psi_n^\ell \,\middle|\, \psi_n^\ell \right\rangle_{A_n}} \left[\frac{r_n^\ell - r_n^{\ell+1}}{\lambda_n^\ell} \right] , \end{aligned} \tag{I.39}$$

where we used recurrence (I.24). We now calculate with the help of Eq. (I.34)

$$(r_n^k)^T (r_n^\ell) = (r_n^k)^T \psi_n^\ell - (r_n^k)^T \sum_{m=0}^{\ell-1} \beta_n^{\ell m} \psi_n^m = 0 , \tag{I.40}$$

for $\ell < k$ due to the orthogonality of the search direction and the residuals, see Eq. (I.33). Hence, we obtain for all $\ell < k$

$$\begin{aligned}\beta_n^{k\ell} &= \frac{1}{\lambda_n^{k-1}} \frac{(r_n^k)^T r_n^k \delta_{\ell+1,k}}{\left\langle \psi_n^{k-1} \,\middle|\, \psi_n^{k-1} \right\rangle_{A_n}} \\ &= \frac{(r_n^k)^T r_n^k}{(r_n^{k-1})^T r_n^{k-1}} \delta_{\ell,k-1} \ . \end{aligned} \tag{I.41}$$

Hence, the name *conjugated gradients*.

We are now in a position to describe the algorithm for the method of *conjugated gradients*:

1. Choose an initial position x_0.
2. Determine the vector b_n and the matrix A_n for a given position x_n.
3. Perform the following N steps in order to calculate x_{n+1}:

   a. Set

$$\psi_n^0 = r_n^0 = b_n \quad \text{and} \quad \lambda_n^0 = \frac{b_n^T \psi_n^0}{\left\langle \psi_n^0 \,\middle|\, \psi_n^0 \right\rangle_{A_n}} \ , \tag{I.42}$$

   as well as

$$x_{n+1} = x_n + \lambda_n^0 \psi_n^0 \ . \tag{I.43}$$

   b. Calculate for $k = 1, \ldots, N-1$ the residuals,

$$r_n^k = r_n^{k-1} - \lambda_n^{k-1} A_n \psi_n^{k-1} \ , \tag{I.44}$$

   the new search directions

$$\psi_n^k = r_n^k + \frac{(r_n^k)^T r_n^k}{(r_n^{k-1})^T r_n^{k-1}} \psi_n^{k-1} \ , \tag{I.45}$$

   the step lengths

$$\lambda_n^k = \frac{b_n^T \psi^k}{\left\langle \psi_n^k \,\middle|\, \psi_n^k \right\rangle_{A_n}} \ , \tag{I.46}$$

   and, finally, the modified positions

$$x_{n+1} = x_n + \lambda_n^k \psi_n^k \ . \tag{I.47}$$

4. If $|f(x_{n+1}) - f(x_n)| < \epsilon$, with ϵ some required accuracy, terminate the iteration, otherwise return to step 2. In case of a convex function $f(x)$ terminate also after N steps.

Strictly speaking, this algorithm is only valid for convex functions because we note that one might get into trouble whenever a position is reached at which the Hessian is not positive definite. It is therefore desirable to exclude the Hessian from the algorithm. This can be achieved by an algorithm developed by FLETCHER and REEVES [56]. Based on our previous discussion the generalization is rather obvious: If we do not want to use the Hessian explicitly, we have to determine the step length λ_n^ℓ by minimizing $f(x_n^\ell + \lambda_n^\ell \psi_n^\ell)$ for a given search direction ψ_n^ℓ numerically. The residuals are then taken to be the exact gradient of the function $f(x_n^\ell)$ rather than of $\hat{f}_n(x_n^\ell)$. The next search direction ψ_n^{k+1} is then determined via

$$\psi_n^{k+1} = -\nabla f(x_n^{k+1}) + \frac{\|\nabla f(x_n^{k+1})\|^2}{\|\nabla f(x_n^k)\|^2} \psi_n^k \,. \tag{I.48}$$

Hence, we have the following algorithm (FLETCHER-REEVES *algorithm*):

1. Choose an initial position x_0.
2. Perform the following N steps in order to calculate x_{n+1}:

   a. Set

   $$\psi_n^0 = -\nabla f(x_n) \,. \tag{I.49}$$

   b. Calculate for $k = 0, \ldots, N-1$ λ_n^k by minimizing $f(x_n^k + \lambda_n^k \psi_n^k)$, the new position $x_n^{k+1} = x_n^k + \lambda_n^k \psi_n^k$, and the new search direction via

   $$\psi_n^k = -\nabla f(x_n^{k+1}) + \frac{\|\nabla f(x_n^{k+1})\|^2}{\|\nabla f(x_n^k)\|^2} \psi_n^k \,. \tag{I.50}$$

3. If $|f(x_{n+1}) - f(x_n)| < \epsilon$, with ϵ some required accuracy, terminate the iteration, otherwise return to step 2.

The resulting sequence of steps towards the minimum for the same function and initial conditions as were used for Fig. I.1 is illustrated in Fig. I.2. In comparing Figs. I.1 and I.2 we note immediately that the search strategy developed for the method of conjugate gradients superbly outperforms the search strategy of the method of steepest descent. In particular, if the ratio between the gradient in x and y direction is large, a strategy of orthogonal search directions is disadvantageous. This particular case is illustrated in Fig. I.3 for both, steepest descent and conjugate gradients. Here we investigate the convex function

$$f(x, y) = x^2 + 10y^2 \,, \tag{I.51}$$

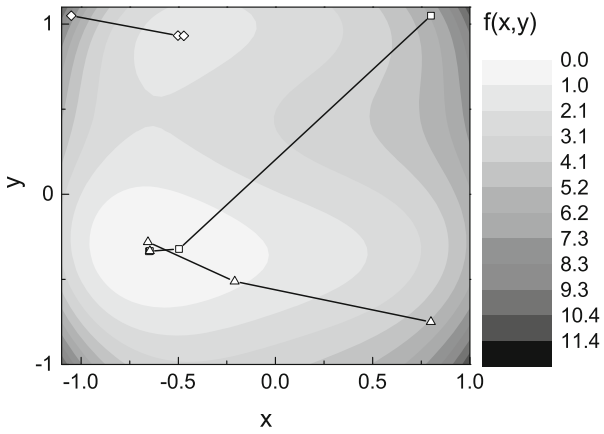

Fig. I.2 Iteration sequence of the method of conjugated gradients for three different starting points

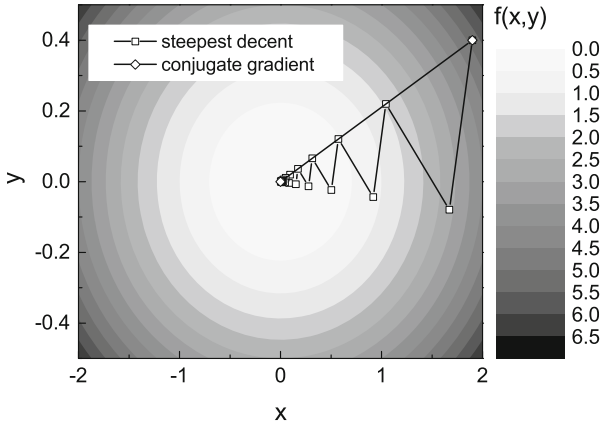

Fig. I.3 Comparison of the iteration sequence between the method of steepest descent and the method of conjugated gradients

together with an initial position $(x_0, y_0) = (1.9, 0.4)$. The resulting sequence of points towards the minimum is illustrated in Fig. I.3. In the case of steepest descent the sequence approaches the minimum rather slowly since subsequent search directions have to be orthogonal to each other in the classical sense. The advantage of conjugate gradients is that A_n-orthonormality accelerates the convergence towards the minimum. In this example we reach it within two steps and a required absolute accuracy of $\eta = 10^{-7}$.

As a final remark we note that also the method of conjugate gradients will only find the *local minimum* closest to the initial position. Hence, the outcome of the method highly depends on the choice of x_0. Moreover, the calculation of the gradients may be very tedious and time-consuming from a numerical point of view.

References

1. Arnol'd, V.I.: Mathematical Methods of Classical Mechanics, 2nd edn. Graduate Texts in Mathematics, vol. 60. Springer, Berlin/Heidelberg (1989)
2. Fetter, A.L., Walecka, J.D.: Theoretical Mechanics of Particles and Continua. Dover, New York (2004)
3. Scheck, F.: Mechanics, 5th edn. Springer, Berlin/Heidelberg (2010)
4. Goldstein, H., Poole, C., Safko, J.: Classical Mechanics, 3rd edn. Addison-Wesley, Menlo Park (2013)
5. Fließbach, T.: Mechanik, 7th edn. Lehrbuch zur Theoretischen Physik I. Springer, Berlin/Heidelberg (2015)
6. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C++, 2nd edn. Cambridge University Press, Cambridge (2002)
7. Ralston, A., Rabinowitz, P.: A First Course in Numerical Analysis, 2nd edn. Dover, New York (2001)
8. Jacques, I., Judd, C.: Numerical Analysis. Chapman and Hall, London (1987)
9. Burden, R.L., Faires, J.D.: Numerical Analysis. PWS-Kent Publishing Comp., Boston (1993)
10. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis, 2nd edn. Springer, Berlin/Heidelberg (1993)
11. Westlake, J.R.: A Handbook of Numerical Matrix Inversion and Solution of Linear Equations. Wiley, New York (1968)
12. Hazewinkel, M. (ed.): Encyclopaedia of Mathematics. Springer, Berlin/Heidelberg (1994)
13. Varga, R.S.: Matrix Iterative Analysis, 2nd edn. Springer Series in Computational Mathematics, vol. 27. Springer, Berlin/Heidelberg (2000)
14. Jackson, J.D.: Classical Electrodynamics, 3rd edn. Wiley, New York (1998)
15. Greiner, W.: Classical Electrodynamics. Springer, Berlin/Heidelberg (1998)
16. Griffiths, D.J.: Introduction to Electrodynamics, 4th edn. Addison-Wesley, Menlo Park (2013)
17. Sakurai, J.J.: Modern Quantum Mechanics. Addison-Wesley, Menlo Park (1985)
18. Sakurai, J.J.: Advanced Quantum Mechanics. Addison-Wesley, Menlo Park (1987)
19. Ballentine, L.E.: Quantum Mechanics. World Scientific, Hackensack (1998)
20. Courant, R., Hilbert, D.: Methods of Mathematical Physics, vol. 1. Wiley, New York (1989)
21. Schücker, T.: Distributions, Fourier Transforms and Some of Their Applications to Physics. Lecture Notes in Physics, vol. 37. World Scientific, Hackensack (1991)
22. Kammler, D.W.: A First Course in Fourier Analysis. Cambridge Univerity Press, Cambridge (2008)
23. Hansen, E.W.: Fourier Transforms: Principles and Applications. Wiley, New York (2014)
24. Bernatz, R.: Fourier Series and Numerical Methods for Partial Differential Equations. Wiley, New York (2010)
25. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**, 297–301 (1965). doi:10.1090/S0025-5718-1965-0178586-1
26. Nussbaumer, H.J.: Fast Fourier Transform and Convolution Algorithms. Springer Series in Information Sciences, vol. 2. Springer, Berlin/Heidelberg (1981)
27. Frigo, M., Johnson, S.G.: FFTW, The Fastest Fourier Transform in the West,Vers. 3.3.4 (2015). http://www.fftw.org
28. Bakhturin, Y.A.: Campell-Hausdorff formula. In: Hazewinkel, M. (ed.) Encyclopaedia of Mathematics. Springer, Berlin/Heidelberg (1994)
29. McLachlan, R.I., Quispel, G.R.W.: Splitting methods. Acta Numer. **11**, 341–434 (2002). doi:10.1017/S0962492902000053
30. Rotar, V.: Probability Theory. World Scientific, Singapore (1998)
31. Papoulis, A., Pillai, S.: Probability, Random Variables and Stochastic Processes. McGraw Hill, New York (2001)
32. Breuer, H.P., Petruccione, F.: Open Quantum Systems, chap. 1. Clarendon Press, Oxford (2010)

33. von der Linden, W., Dose, V., von Toussaint, U.: Bayesian Probability Theory. Cambridge University Press, Cambridge (2014)
34. Klenke, A.: Probability Theory. Universitext. Springer, Berlin/Heidelberg (2014)
35. Lee, P.M.: Bayesian Statistics: An Introduction. Wiley, New York (2012)
36. Abramovitz, M., Stegun, I.A. (eds.): Handbook of Mathemathical Functions. Dover, New York (1965)
37. Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W.: NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge (2010)
38. Applebaum, D.: Lévy Processes and Stochastic Calculus. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (2006)
39. Yeomans, J.M.: Statistical Mechanics of Phase Transitions. Clarendon Press, Oxford (1992)
40. Cardy, J.: Scaling and Renormalization in Statistical Physics. Cambridge Lecture Notes in Physics. Cambridge University Press, Cambridge (1996)
41. Fließbach, T.: Statistische Physik. Lehrbuch zur Theoretischen Physik IV. Springer, Berlin/Heidelberg (2010)
42. Pathria, R.K., Beale, P.D.: Statistical Mechanics, 3rd edn. Academic, San Diego (2011)
43. White, R.M.: Quantum Theory of Magnetism, 3rd edn. Springer Series in Solid-State Sciences. Springer, Berlin/Heidelberg (2007)
44. Mandl, F.: Statistical Physics, 2nd edn. Wiley, New York (1988)
45. Keener, R.W.: Theoretical Statistics. Springer, Berlin/Heidelberg (2010)
46. Jaeger, G.: The Ehrenfest classification of phase transitions: introduction and evolution. Arch. Hist. Exact Sci. 51–81 (1998). doi:10.1007/s004070050021
47. Landau, L.D., Lifshitz, E.M.: Course of Theoretical Physics, vol. 5: Statistical Physics. Pergamon Press, London (1963)
48. Ter Haar, D.: Collected Papers of L. D. Landau, p. 546. Pergamon Press, London (1965)
49. Kilbas, A.A., Srivastava, H.M., Trujillo, J.J.: Theory and Applications of Fractal Differential Equations. Elsevier, Amsterdam (2006)
50. Iversen, G.P., Gergen, I.: Statistics. Springer Undergraduate Textbooks in Statistics. Springer, Berlin/Heidelberg (1997)
51. Wilcox, R.R.: Basic Statistics. Oxford University Press, New York (2009)
52. Monahan, J.F.: Numerical Methods of Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2011)
53. Wood, S.: Core Statistics. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge (2015)
54. Marquart, D.: An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. **11**, 431–441 (1963). doi:10.1137/0111030
55. Avriel, M.: Nonlinear Programming: Analysis and Methods. Dover, New York (2003)
56. Fletcher, R., Reeves, C.M.: Function minimization by conjugate gradients. Comput. J. **7**, 149–154 (1964). doi:10.1093/comjnl/7.2.149

Index

B.A. Stickler, E. Schachinger, *Basic Concepts in Computational Physics*,
DOI 10.1007/978-3-319-27265-8

D

E

F

M

N

O

P