Continuum Modeling in the Physical Sciences

Mathematical Modeling and Computation

About the Series

The SIAM series on Mathematical Modeling and Computation draws attention to the wide range of important problems in the physical and life sciences and engineering that are addressed by mathematical modeling and computation; promotes the interdisciplinary culture required to meet these large-scale challenges; and encourages the education of the next generation of applied and computational mathematicians, physical and life scientists, and engineers.

The books cover analytical and computational techniques, describe significant mathematical developments, and introduce modern scientific and engineering applications. The series will publish lecture notes and texts for advanced undergraduate- or graduate-level courses in physical applied mathematics, biomathematics, and mathematical modeling, and volumes of interest to a wide segment of the community of applied mathematicians, computational scientists, and engineers.

Appropriate subject areas for future books in the series include fluids, dynamical systems and chaos, mathematical biology, neuroscience, mathematical physiology, epidemiology, morphogenesis, biomedical engineering, reaction-diffusion in chemistry, nonlinear science, interfacial problems, solidification, combustion, transport theory, solid mechanics, nonlinear vibrations, electromagnetic theory, nonlinear optics, wave propagation, coherent structures, scattering theory, earth science, solid-state physics, and plasma physics.

E. van Groesen and Jaap Molenaar, *Continuum Modeling in the Physical Sciences*

Gerda de Vries, Thomas Hillen, Mark Lewis, Johannes Müller, and Birgitt Schönfisch, *A Course in Mathematical Biology: Quantitative Modeling with Mathematical and Computational Methods*

Ivan Markovsky, Jan C. Willems, Sabine Van Huffel, and Bart De Moor, *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*

R. M. M. Mattheij, S. W. Rienstra, and J. H. M. ten Thije Boonkkamp, *Partial Differential Equations: Modeling, Analysis, Computation*

Johnny T. Ottesen, Mette S. Olufsen, and Jesper K. Larsen, *Applied Mathematical Models in Human Physiology*

Ingemar Kaj, *Stochastic Modeling in Broadband Communications Systems*

Peter Salamon, Paolo Sibani, and Richard Frost, *Facts, Conjectures, and Improvements for Simulated Annealing*

Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook, *Credit Scoring and Its Applications*

Frank Natterer and Frank Wübbeling, *Mathematical Methods in Image Reconstruction*

Per Christian Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*

Michael Griebel, Thomas Dornseifer, and Tilman Neunhoeffer, *Numerical Simulation in Fluid Dynamics: A Practical Introduction*

Khosrow Chadan, David Colton, Lassi Päivärinta, and William Rundell, *An Introduction to Inverse Scattering and Inverse Spectral Problems*

Charles K. Chui, *Wavelets: A Mathematical Tool for Signal Analysis*

Continuum Modeling in the Physical Sciences

E. van Groesen
University of Twente
Enschede, The Netherlands

Jaap Molenaar
Wageningen University and Research Centre
Wageningen, The Netherlands

siam
Society for Industrial and Applied Mathematics
Philadelphia

10 9 8 7 6 5 4 3 2 1

Figure 4.3 (cathedral only) used with permission of Geoff Stearns, New York, New York.

Figure 5.2 and cover art used with permission of Bob Hawbaker, Cannon Falls, Minnesota.

Library of Congress Cataloging-in-Publication Data

Continuum modeling in the physical sciences / E. van Groesen, Jaap Molenaar.
p. cm. — (Mathematical modeling and computation)
Includes bibliographical references and index.
ISBN 978-0-898716-25-2 (alk. paper)
1. Mathematical physics—Mathematical models. 2. Mechanics—Mathematical models.
I. Groesen, E. W. C. van. II. Molenaar, J.

QC20.C66 2007
530.15—dc22

2006051255

Contents

Preface

The huge impact of mathematics on modern society is undoubtedly due to the power of mathematical modeling. Mathematics not only has an "unreasonable effectiveness" in the natural sciences, as concluded in 1960 by Wigner[1] but it also contributes to other fields in a remarkably effective way. This implies that mathematical modeling—the ability to apply mathematical concepts and techniques to real-life systems—has considerably expanded in the last few decades. It is impossible to deal with the many and varied aspects of mathematical modeling in one course or one textbook, and so in the present text we restrict ourselves to applications in the natural sciences and focus on an advanced level of modeling. This book grew out of lecture notes of a course for advanced undergraduates at the University of Twente. The original material, for a 10-week course, is greatly extended with extra topics and many examples so that lecturers can select those parts that best fit the audience. This book is intended to be used as a textbook, but we hope it will also be useful as a source of reference and inspiration for students and researchers alike.

Teaching mathematical modeling is a quite complicated challenge. On the one hand, one has to expose a great variety of general mathematical concepts, and on the other hand, one has to treat the principles of the field of application in some detail. It is this diversity of applicable techniques and possible applications that could seduce an author to present the subject as a long series of ingenious case studies, in which students can hardly discover any coherence. This approach could even disappoint the student, since having digested many particular models does not guarantee that one knows how to proceed when confronted with a new situation. To convince students of the power and beauty of modeling, we offer in this book an extensive exposition of general principles. Since students gain the most from a course if its structure is clearly highlighted, most chapters are devoted to central issues, such as dimensional analysis, conservation principles, balance laws, constitutive relations, stability, robustness, and variational methods. The core of these chapters will form the backbone of any course on mathematical modeling.

This book aims at applications of modeling techniques, and the relevant ideas and techniques are presented via examples and exercises. The book contains a multitude of classroom examples and exercises throughout the text, and several chapters contain a section of challenging problems. Furthermore, the last chapter is devoted to extensively worked-out case studies. The examples mostly stem from classical mechanics, wave phenomena, and continuous mechanics, showing the backgrounds of the authors. However, this does

[1] E.P. Wigner, *The unreasonable effectiveness of mathematics in the natural sciences*, Comm. Pure Appl. Math., 13, 1960, pp. 1–14.

not imply that the book could not be used to study the mathematical modeling of topics from other disciplines. On the contrary, we have tried to keep the treatment of modeling principles as general as possible.

Chapter 1 is devoted to dimensional analysis and scaling. It is fascinating to show how strong conclusions about a system can be drawn just by looking at the physical dimensions of the relevant quantities. Our hope is that the presented examples are so convincing that the reader will never start any modeling activity without first checking whether these techniques might be applied.

In Chapter 2 we introduce some basic elements of modeling, namely, conservation principles and constitutive relations. These notions are so general that any scientist must master them before modeling at all. They are first introduced in one dimension so that the reader becomes familiar with them in a natural way. The so-called transport theorem plays a central role in generalizing the conservation principles to more dimensions. This theorem allows us to deal with all kinds of quantities that may be transported (scalar, vector, and tensor like) on an equal footing.

In Chapter 3 we summarize the basics of differential equations. In our manner of presentation the analogies rather than the differences between ordinary and partial differential equations are emphasized.

Chapter 4 deals with stability and robustness. Stability is an essential aspect of any model analysis, since most models are used to control the systems under consideration. The related concept of robustness is important, since it provides the modeler with information about the sensitivity of the model with respect to perturbations of its parameters.

In Chapter 5 variational methods are discussed. These methods deserve an extensive treatment, since remarkably many problems in nature can be put in variational form, i.e., can be formulated as an optimization problem. We also point out how a variational formulation may yield a useful guideline for how to calculate the solution of a model numerically.

Chapter 6 is meant as a summarizing showcase, in which the ideas and techniques in the preceding chapters are applied to real-life problems. To that aim we extensively work out four advanced examples. The first deals with polymer dynamics. It nicely shows how a modeler may benefit by dimensional analysis. The second example concerns fiber spinning. It shows how a relatively simple system can lead to a hard stability analysis. The third example shows the modeling of water waves. It demonstrates the power of a variational approach. In the fourth example we study the transmittance of light through an optical fiber.

In writing this book we have benefitted from ideas and support (in various ways) from a number of persons. In particular, we would like to mention Dr. B.W. van de Fliert, Dr. Andonowati, Dr. M. Hammer, Dr. G. Klopman, and Dr. F.P.H. van Beckum, who helped us with comments on earlier versions. We give special acknowledgment to Mrs. T.J.M. Wolfs–van de Hurk for her precise correcting work and to Mrs. A. Megawati, who provided us with some figures. Finally, we would like to thank the students for their positive criticism on earlier drafts of our book.

E. van Groesen
Jaap Molenaar
Autumn 2006

Chapter 1
Dimensional Analysis and Scaling

1.1 Mathematical models

A mathematical model describes the behavior of a real-life system in terms of mathematical equations. These equations represent the relations between the relevant properties of the system under consideration. In these models we meet with *variables* and *parameters*. In variables, we discern between *dependent* and *independent*. For example, in mechanical systems one usually is interested in the positions of the different parts as functions of time, so in these systems the positions act as the dependent variables and time as the independent variable. Parameters are properties like masses, prescribed temperatures, currents, voltages, and friction coefficients. Parameters that can be influenced by the observer are referred to as *adjustable*. The other parameters act as constants in the model. For example, in atmospherical models used in weather forecasting one is interested in properties like temperature and humidity (the dependent variables) as functions of position and time (the independent variables). Important parameters are then the gravity field and the rotational speed of the earth, and these clearly belong to the class of *nonadjustable* parameters. The *solution* of a mathematical model is known if we can determine the relations between dependent and independent variables. Since the solution depends on the values of the adjustable parameters, mathematical models are a powerful tool with which to determine which values of the adjustable parameters yield specific required behavior.

1.2 Dimensions

If the variables or parameters in a model correspond to physical properties, they have physical dimensions. The *fundamental dimensions* used in this book are given in the following table:

dimension	symbol	MKS-unit
length	L	m (meter)
mass	M	kg (kilogram)
time	T	s (second)
temperature	Θ	$^\circ C$ (degree Celsius)
current	I	A (Ampere)

The dimension of any physical quantity can be expressed in terms of the fundamental dimensions. For most quantities this is clear from the definition. For example,

quantity	dimension
area	L^2
volume	L^3
velocity	L/T
acceleration	L/T^2
mass density	M/L^3
mechanical energy	ML^2/T^2
pressure	$M/(L\,T^2)$

In other cases the dimensionality of a quantity is deduced from the rule that all terms in a particular equation must have the same dimensionality. This rule is a consequence of the condition that the form of any equation in a mathematical model may not depend on the units used. For example, the dimension of force directly follows from the second law of Newton, which states that for a single mass, the mass times the acceleration equals the total force exerted on the mass. In standard notation, $F = ma$. So, the dimensionality of a force F, denoted as $[F]$, equals the dimensionality $[ma]$ of the product of mass m and acceleration a. Since $[ma] = [m][a]$, we conclude that $[F] = ML/T^2$. In this way we arrive at, e.g.,

quantity	dimension
force	ML/T^2
mechanical energy	ML^2/T^2
pressure	$M/(LT^2)$

For coefficients, the dimensionality may vary with specific choices made by the modeler. For example, if a frictional force is introduced with strength proportional to the velocity of the object, the constant of proportionality will have the dimension of the quotient of force and velocity. However, if the friction is assumed to be proportional to the velocity squared, the proportionality constant will have the dimension of force divided by velocity squared. See also Example 1.3c. For a *dimensionless quantity*, say, q, we have $[q] = 1$. Examples are angles and universal constants like π and e. In dimensional analysis, to be treated in §1.3, dimensionless quantities play a central role. There it is shown how they can be constructed. The existence of these so-called dimensionless numbers allows us to draw important conclusions about the system without solving the governing mathematical model.

Example 1.2a. Driven, damped, harmonic oscillator.
Consider a bead-spring system in one dimension under the influence of friction and a driving force. The position of the bead with mass m is denoted by its displacement u measured with respect to its equilibrium position. See Fig. **1.1**.

We are interested in u as a function of time t. So, u is the dependent and t the independent variable of the system. As for the notation, we shall use the convention $\dot{u} \equiv du/dt$, $\ddot{u} \equiv du^2/dt^2$, etc. The second law of Newton states that the *inertia force*, i.e., mass m times acceleration $\ddot{u}$, equals the sum of the forces exerted on the bead. These forces are the driving force F_d, which is taken harmonic with angular frequency ω and amplitude F_0;

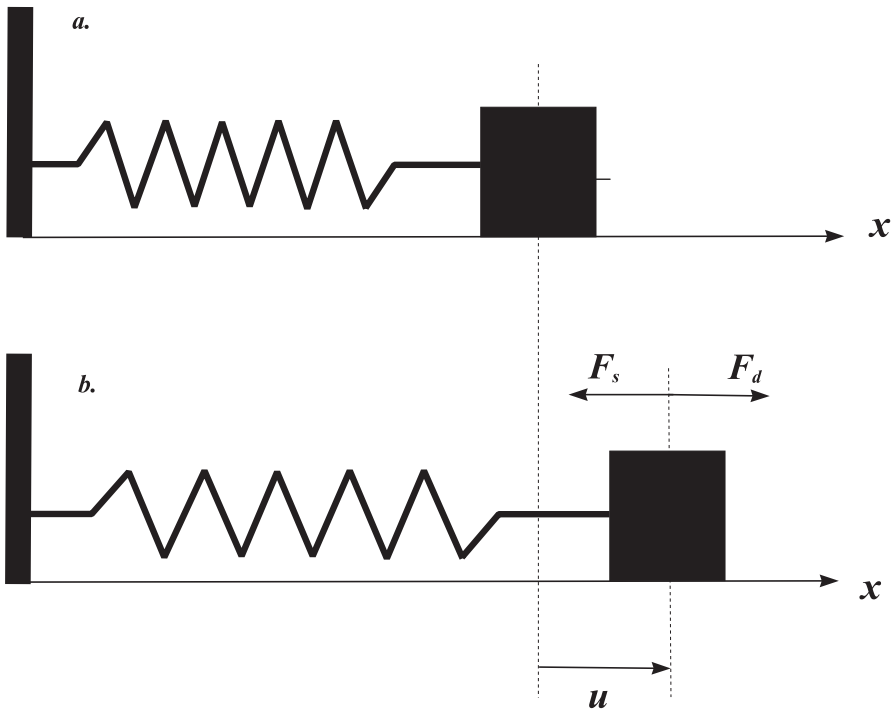

Figure 1.1. A vibrating mass attached to a spring is the prototype of harmonic motion if the spring response is linear, i.e., if the spring force F_s is linearly proportional to the deviation u (b) measured with respect to the equilibrium position (a). Apart from F_s, often an external driving force F_d and a friction force F_f apply. The latter usually is taken linearly proportional to the velocity of the mass but in the reverse direction.

the spring force F_s, which is linearly proportional to the displacement and reversely directed to it; and the frictional force F_f, which is assumed to be linearly proportional to the velocity and reversely directed to it. This leads to the balance of forces

$$m\ddot{u} = F_f + F_s + F_d = -c\dot{u} - ku + F_0 \sin \omega t.$$

The conventional form in which to write this *equation of motion* is

$$m\ddot{u} + c\dot{u} + ku = F_0 \sin \omega t. \tag{1.1}$$

Since m, c, k, and F_0 can all be influenced, they are adjustable parameters. Every term in this equation has the dimension of force, so ML/T^2. From this it follows that

$$[c] = \frac{ML/T^2}{L/T} = \frac{M}{T}\,, \quad [k] = \frac{ML/T^2}{L} = \frac{M}{T^2}\,, \quad [F_0] = \frac{ML}{T^2}\,.$$

The argument ωt of the sine function must be dimensionless, so $[\omega t] = 1$. We thus have

$$[\omega] = \frac{1}{T}\,.$$

A mathematical model in terms of an ordinary differential equation (ODE) is not yet complete if the initial values are left unspecified. This introduces two extra parameters into the system: initial position $u_0 \equiv u(t_0)$ and initial velocity $v_0 \equiv \dot{u}(t_0)$. The solution $u(t)$ thus depends on seven parameters, and we could write it as

$$u = u(t; m, c, k, F_0, \omega, u_0, v_0).$$

For such a simple system this is a huge number to handle, since in an experiment all these parameters could in principle be varied. In the following we show that such a system can essentially be described with fewer parameters, since it does not make sense to vary them all independently. □

The fact that the variables and parameters have physical dimensions can be fruitfully exploited. The techniques of *nondimensionalizing* and *scaling* are extremely powerful tools in analyzing the models. Their importance is fully appreciated only through examples, which account for the largest part of this chapter. The basic idea is to apply a transformation to the variables and parameters such that simplified equations result. It is often amazing how much structure is revealed simply by nondimensionalizing, without solving the model explicitly. Thanks to these techniques it is often known beforehand that the system depends not on all parameters separately but only on certain combinations. In an experimental situation it is of great importance to know how the system depends on the parameters, so this insight may save much time, cost, and energy.

In practice two methods are applied, *dimensional analysis* and *scaling*, each having its own merits. They are dealt with in the subsections below, respectively. Dimensional analysis fully exploits the information contained in the physical dimensions of the variables and parameters. Scaling has a more restricted scope and aims at a reduction of the number of parameters.

1.3 Dimensional analysis

Nondimensionalizing a mathematical model is a constructive way to formulate the model in terms of dimensionless quantities only. A big achievement is that dimensional analysis yields insight in the scaling relations of the system without using knowledge of any governing equation. An advantageous corollary is that the total number of variables and/or parameters is minimal. Reduction of the number of parameters is also the purpose of *scaling*, a technique to be dealt with in the next section. However, dimensional analysis is more general than scaling in that it is based on a transformation of both variables and parameters on the same footing, whereas in scaling only the variables are transformed. Another difference is that scaling starts from the governing equations, whereas dimensional analysis starts much more basically, namely, from the dimensions involved in the system, and it may even predict from them some quantitative features of the model without knowledge of the model equations. The basic idea of dimensional analysis is easily explained. Consider a system with scalar variables $x_1, \ldots, x_k$ and scalar parameters $p_1, \ldots, p_\ell$. So, the total number of quantities involved is $N = k + \ell$. Note that in the model, vectors, matrices, etc., may figure, but for this analysis all their components have to be treated separately. We now form the products

$$x_1^{r_1} \ldots x_k^{r_k}, \; p_1^{r_{k+1}} \ldots p_\ell^{r_N}$$

and ask for which choices of the r_i these products are dimensionless. The answer follows from replacing each x_i and p_i with its fundamental dimensions. If, say, m dimensions $d_1, \ldots, d_m$ are involved, the replacement gives rise to another type of product,

$$d_1^{s_1} \ldots d_m^{s_m},$$

with the numbers s_i, $i = 1, \ldots, m$, being linear functions of the r_j, $j = 1, \ldots, N$. The procedure is illustrated several times in the examples below. By requiring

$$s_i = 0, \quad i = 1, \ldots, m,$$

we obtain a set of m linear equations for the N unknowns $r_1, \ldots, r_N$. Note that the numbers r_j, $j = 1, \ldots, N$, are rational, since they are solutions of linear equations with rational coefficients. The rationality of these coefficients stems from the fact that in nature all measurable quantities turn out to have dimensions that are products of integer powers of the fundamental dimensions, as shown in the tables in §1.2. From linear algebra it follows that there are (at most) $N - m$ linearly independent solutions, corresponding to $N - m$ dimensionless quantities q_i, $i = 1, \ldots, (N - m)$. Buckingham formalized this in the following theorem.

Theorem (Buckingham).

Consider a system with variables $x_1, \ldots, x_k$ and parameters $p_1, \ldots, p_\ell$, in which m fundamental dimensions are involved. Then, $k + \ell - m$ dimensionless quantities q_i can be defined, which are products and quotients of the original variables and parameters. Each (scalar) model equation

$$f(x_1, \ldots, x_k, p_1, \ldots, p_\ell) = 0$$

between the x_i and p_i of a mathematical model can be replaced with a corresponding relation between the q_i:

$$f^*(q_1, \ldots, q_{k+\ell-m}) = 0\,.$$

Since Buckingham [6] denoted the dimensionless quantities by π_i, this theorem is often referred to as the π-theorem of Buckingham. We shall not follow his notation since it is no longer common in the literature. As follows from the construction of the q_i as solutions of an underdetermined set of linear equations, they are not uniquely defined by the procedure. If the procedure yields a set of q_i, we can apply a transformation, e.g., by taking algebraic or even functional combinations of them, obtaining another set of dimensionless quantities of the system. It is a matter of expertise, and partly of taste, to determine a convenient set of q_i for the system under consideration. If the number of variables and parameters is not small, the freedom of choice must be especially exploited with care.

We shall work out the nondimensionalizing procedure for a considerable number of examples, pointing out both the practical aspects of the technique and the insight it may yield about the behavior of the system without solving the equations explicitly.

Example 1.3a. Catapulting.
Let us start with an example in which the mathematics is very basic but the ideas behind dimensional analysis are clearly illustrated. A projectile with mass m is launched vertically. See Fig. **1.2**. At launching it has velocity v_0. Its trajectory, i.e., its vertical position z as a function of time t is assumed to be completely determined by the influence of gravity. The effect of friction due to the air is ignored here (but dealt with in Example 1.3e). The projectile will decelerate because of gravity until it reaches its highest position $z_{\max}$ at time $t_{\max}$. After that it falls back with increasing velocity and arrives on the earth at time t_{final}. Since we take v_0 such that $z_{\max}$ remains small compared to the Earth's radius, we may take the gravity field uniform with gravity constant g.

In this system the variables are z and t and the parameters are m, v_0, and g. The relevant physical dimensions are M, L, and T. So, $k = 2$, $\ell = 3$, and $m = 3$, and the theorem of Buckingham states that the system has two dimensionless quantities. All properties of the system can be expressed in only these two quantities. In this simple case the dimensionless quantities can be easily found from inspection of the dimensions: $[z] = L$, $[t] = T$, $[m] = M$, $[v_0] = L/T$, and $[g] = L/T^2$. An evident choice is

$$t^* = \frac{gt}{v_0}\,, \quad z^* = \frac{gz}{v_0^2}\,.$$

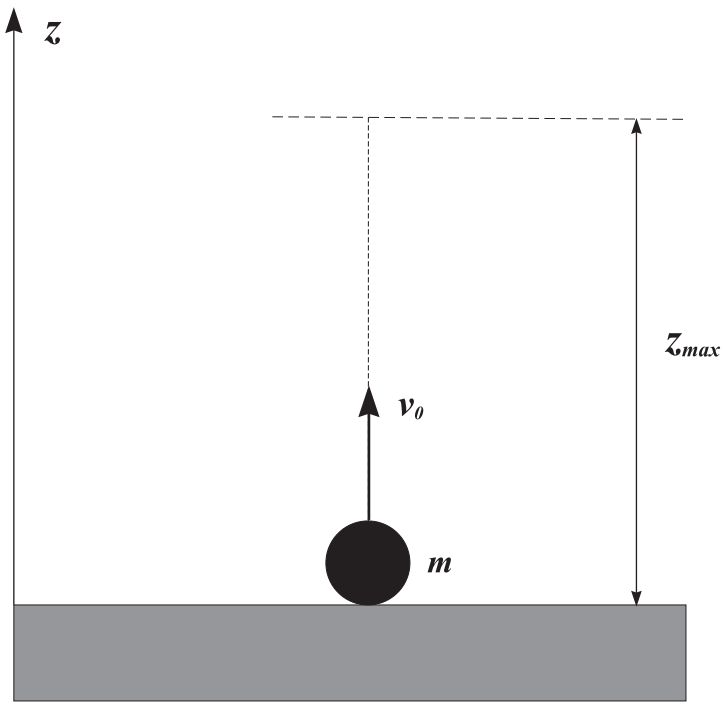

Figure 1.2. The main scaling characteristics of a mass m, launched with initial speed v_0, are easily predicted by dimensional analysis.

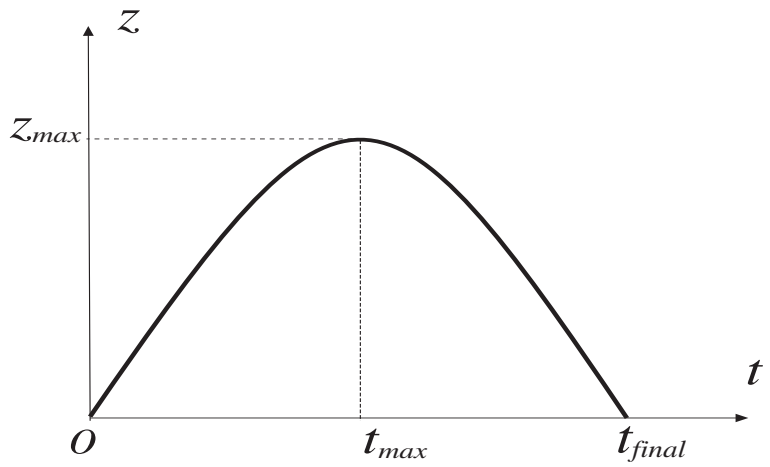

Figure 1.3. The height z of a mass, launched at speed v_0, as a function of time t. It will reach a maximum height z_{max} at time t_{max} and reach the earth's surface again at time t_{final}.

Note that the mass m is not present in t^* and z^*, since the physical dimension M is not present in one of the other variables and parameters. The Buckingham theorem yields that its motion is described by a relation between z^* and t^*. This immediately leads to the conclusion that the motion of the projectile is independent of its mass. From experimental evidence we know that the relation between z and t is more or less as sketched in Fig. **1.3**. The function $z(t)$ reaches a maximum $z_{\max}$ at $t_{\max}$ and vanishes at t_{final}. Since z^* and t^* are just scaled versions of z and t, z^* apparently can be written as an explicit function of t^*:

$$z^* = f^*(t^*). \tag{1.2}$$

The theorem does not specify any information about f^* but only ensures its existence and the insight that the form of f^* does not depend on any of the parameters m, v_0, and g separately. The latter property thus also holds for the dimensionless quantities $z^*_{\max}$, $t^*_{\max}$, and t^*_{final}. These are just fixed numbers, as shown in Exercise 1.3a. Using the relations between dimensional and dimensionless quantities, we have that

$$z_{\max} = \frac{v_0^2}{g} z^*_{\max} , \quad t_{\max} = \frac{v_0}{g} t^*_{\max} , \quad t_{\text{final}} = \frac{v_0}{g} t^*_{\text{final}} .$$

This yields the insight that $z_{\max}$ scales with v_0^2 and both $t_{\max}$ and t_{final} with v_0 for fixed value of g. We denote this as

$$z_{\max} \sim v_0^2 , \quad t_{\max} \sim v_0 , \quad t_{\text{final}} \sim v_0 .$$

So, launching with a twice-as-large velocity leads to a four-times-larger maximal height of the projectile. In the same way we conclude that

$$z_{\max} \sim \frac{1}{g} , \quad t_{\max} \sim \frac{1}{g} , \quad t_{\text{final}} \sim \frac{1}{g}$$

for a fixed value of v_0. So, catapulting on the moon, where g is (approximately six times) smaller than on the earth, enhances $z_{\max}$, $t_{\max}$, and t_{final} all by the same factor. □

Exercise 1.3a.

Check these conclusions on catapulting by explicitly solving the equation of motion

$$m\frac{d^2z}{dt^2} = -mg.$$

Show that f^* in (1.2) *has the explicit form as given in Fig.* **1.4**. *Calculate explicitly the values of $z^*_{\max}$, $t^*_{\max}$, and $^*_{\text{final}}$. Note that this function cannot be found from dimensional analysis only.*

Example 1.3b. Swinging pendulum.

Consider the motion of a mathematical swing: this pendulum has mass m concentrated in a point at the end of a rigid rod of length ℓ. The motion is restricted to a vertical plane. See Fig. **1.5**. The position of the swinging pendulum is completely specified by the angle φ with the vertical. This is the independent variable, and time t is the dependent variable. Parameters are mass m, rod length ℓ, gravitational acceleration g, and the initial position $\varphi_0 = \varphi(0)$. For convenience we take the initial velocity vanishing. So, $k + \ell = 6$, and since the three fundamental dimensions M, L, and T are involved, the system has three dimensionless quantities. Since φ and φ_0 are already dimensionless, they form an obvious choice. To find the third, we form the products

$$t^{r_1}\ell^{r_2}m^{r_3}g^{r_4}.$$

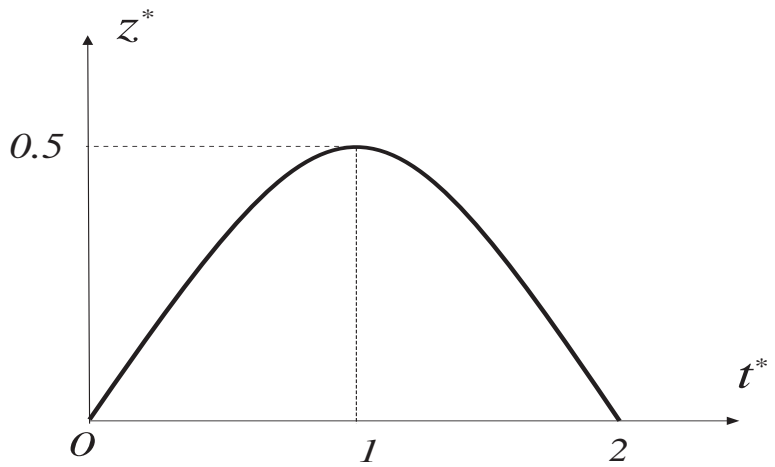

Figure 1.4. Explicit form of the dimensionless function f^* in (1.2). Note that this function is independent of the parameters m, g, and v_0 of the system. The dimensionless height z^* reaches a maximum value $z^*_{max} = 1/2$ at time $t^*_{max} = 1$ and hits the earth's surface again at time $t^*_{final} = 2$.

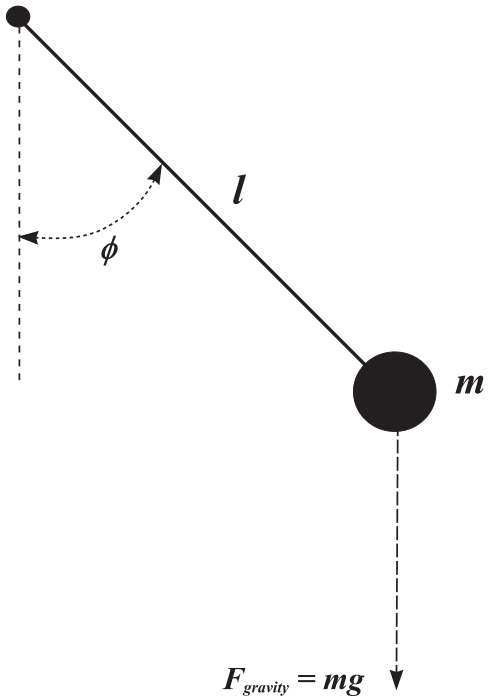

Figure 1.5. Swinging pendulum of mass m and length l. The motion is confined to a vertical plane, and the position of the pendulum can be indicated by the angle ϕ with the vertical.

The condition that this product must be dimensionless leads to the linear equations

$$\begin{aligned} r_1 - 2r_4 &= 0, \\ r_2 + r_4 &= 0, \\ r_3 &= 0. \end{aligned}$$

The choice $(r_1, r_2) = (1, 0)$ then yields

$$t^* = t\sqrt{\frac{g}{\ell}}.$$

Note that the mass m is not present in any of the dimensionless quantities φ, φ_0, and t^*. This implies that pendulum motion is independent from m. The movement of the pendulum is given by some relation between φ, φ_0, and t^*. With φ_0 constant and t^* monotonously increasing, we may write φ as an explicit function of t^*:

$$\varphi = f^*(t^*, \varphi_0).$$

This allows for a conclusion about the period of the system. One should realize that dimensional analysis as such does not reveal that φ is a periodic function of time. However, if

we take this for granted in view of the observations, we have that $f^*(t^* + \tau^*) = f^*(t^*)$, with τ^* the dimensionless period. Since $\tau^* = \tau\sqrt{g/\ell}$ and τ^* does not depend on any of the parameters, we find that τ scales with $\sqrt{\ell/g}$, so

$$\tau \sim \sqrt{\frac{\ell}{g}}. \qquad \square$$

Exercise 1.3b.

a. *Give the dimensionless form of the exact pendulum equation*

$$m\ell\ddot{\varphi} + mg\sin\varphi = 0.$$

b. *If* $|\varphi| \ll 1$*, the linearized pendulum equation*

$$m\ell\ddot{\varphi} + mg\varphi = 0$$

is a good approximation. Give its dimensionless form.

c. *Write the solution of the equation under* b *and check that the period indeed scales with* $\sqrt{\ell/g}$ *as derived in Example* 1.3b. *Determine how the period is influenced if the length is doubled and also when the pendulum is placed on the moon.*

Example 1.3c. Harmonic oscillator.
Here, we revisit the harmonic oscillator introduced in Example 1.2a. Setting the initial values at zero for convenience, the model equation

$$m\ddot{u} + c\dot{u} + ku = F_0\sin\omega t$$

has the two variables u and t and the five parameters m, c, k, F_0, and ω. So, $N = 7$ in this case. The fundamental dimensions involved are mass M, length L, and time T. Forming the products

$$u^{r_1}t^{r_2}m^{r_3}c^{r_4}k^{r_5}F_0^{r_6}\omega^{r_7}$$

and substituting the dimensions, we arrive at the products

$$L^{r_1}T^{r_2}M^{r_3}\left(\frac{M}{T}\right)^{r_4}\left(\frac{M}{T^2}\right)^{r_5}\left(\frac{ML}{T^2}\right)^{r_6}\left(\frac{1}{T}\right)^{r_7}.$$

Collecting powers of M, L, and T, we obtain the following three linear equations for the r_i:

$$\begin{aligned} &r_1 + r_6 = 0,\\ &r_2 - r_4 - 2r_5 - 2r_6 - r_7 = 0,\\ &r_3 + r_4 + r_5 + r_6 = 0. \end{aligned}$$

Here, we meet with three equations for seven unknowns, so four unknowns can be treated as free parameters. For example, we could take $r_1, \ldots, r_4$. The choices $(r_1, r_2, r_3, r_4) = (1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, and $(0, 0, 0, 1)$, respectively, yield the dimensionless quantities

$$u^* = \frac{uk}{F_0},\ t^* = \omega t,\ m^* = \frac{m\omega^2}{k},\ c^* = \frac{c\,\omega}{k}.$$

The dimensionless spring equation then reads as

$$m^*\ddot{u}^* + c^*\dot{u}^* + u^* = \sin t^*,$$

where the time derivative is taken with respect to t^*. □

Exercise 1.3c.

The approach used above for the driven spring system is based on the assumption $F_0 \neq 0$. Apply dimensional analysis to the case $F_0 = 0$ but with the initial position u_0 and initial velocity v_0 both nonvanishing.

Example 1.3d. Estimating the power of explosions.
Details of the strength of the first atomic bomb in 1945 were classified until the 1960s. However, the British physicist G.I. Taylor was able to give a very accurate estimate of the strength from dimensional analysis by using available film of the expansion of the mushroom shape of the explosion. His arguments proceed as follows (see, e.g., [31] and [3, 4]).

The basic appearance of the explosion is an expanding spherical fireball whose edge corresponds to a powerful shock wave, as sketched in Fig. **1.6**. Let R be the radius of the shock wave. It will depend on E, the energy released by the explosion; t, the time elapsed since the explosion; ρ, the initial and ambient air density, and p, the initial and ambient

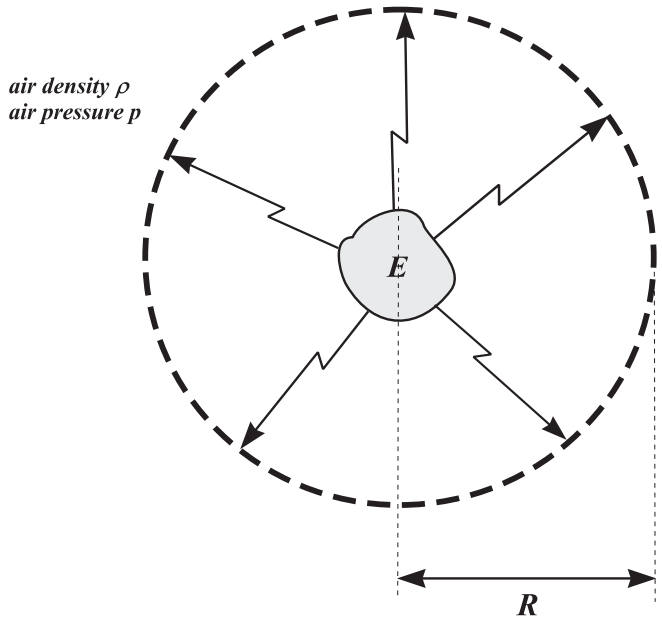

Figure 1.6. Sketch of a shock wave propagating from an explosion with energy E. Dimensional analysis shows that the energy E can be estimated from the propagation velocity of the front.

air pressure. In total we recognize five variables and parameters. Three dimensions are involved. Hence two dimensionless quantities can be found.

Exercise 1.3d.
The dimensions of the variables and parameters in Example 1.3c *can be looked up in the tables in §*1.2. *Show that from requiring the products*

$$R^{r_1} t^{r_2} E^{r_3} \rho^{r_4} p^{r_5}$$

to be dimensionless, the following dimensionless quantities can be derived:

$$q_1 = p \left(\frac{t^6}{E^2 \rho^3} \right)^{1/5}, \quad q_2 = \frac{R^5 \rho}{E t^2} .$$

The Buckingham theorem ensures that the motion of the shock front is governed by some relation between q_1 and q_2. Since q_1, which is essentially a scaled time, is monotonously increasing, we may write q_2 as an explicit function of q_1: $q_2 = f^*(q_1)$. Thus R can be expressed as

$$R = \left(\frac{E t^2 f^*(q_1)}{\rho} \right)^{1/5} . \tag{1.3}$$

From this relation we conclude in the first instance that R depends on t both via the prefactor of f^* and via q_1. This complicates the analysis. Taylor found a way out by first plotting measured values of R as a function of t in a double logarithmic plot. Taking logarithms of both sides of (1.3) we have

$$\log R = \frac{2}{5} \log t + \frac{1}{5} \log \left(\frac{E f^*(q_1)}{\rho} \right) .$$

The data turn out to lie nearly on a straight line. This suggested that $f^*(q_1)$ hardly depends on time so that it can be replaced with its initial value $f^*(0)$. Then, two unknown parameters still figure in the model: E and $f^*(0)$. Taylor estimated the value of $f^*(0)$ from performing an experiment under well-defined conditions, for which E was known and $R(t)$ measured. This led to the conclusion that $f^*(0) \approx 1$. Every shock wave is thus fairly well described by the model equation

$$\log R = \frac{2}{5} \log t + \frac{1}{5} \log \left(\frac{E}{\rho} \right) .$$

Since ρ is usually known, it is easy to estimate E from fitting this equation into measured $R(t)$ data. □

Example 1.3e. Estimating train speed from drop tracks.
Let us imagine we are traveling by train on a rainy day. Looking through the window we see rain drops attached to the glass following straight trajectories along the window downward. The angle of inclination of the drop paths appears to depend on the train speed. We wonder whether the speed can be estimated from this angle. To answer this question, the system

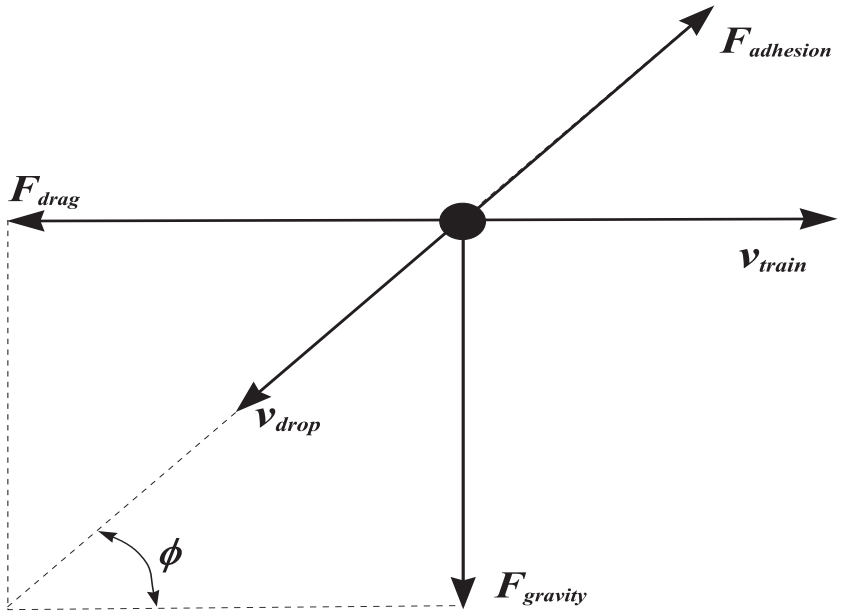

Figure 1.7. The path of a water droplet on a train window yields enough information to estimate the train speed, as follows from dimensional analysis.

of a drop sliding along a moving glass plate has to be modeled. See Fig. **1.7**. Three forces act on the droplet: the gravitational force with strength F_g, the drag force with strength F_d due to friction between the drop and surrounding air, and the adhesion force with strength F_a between drop and glass. F_g is directed vertically, and its strength is equal to mg with m the drop mass. F_d is directed horizontally, but its strength is not known beforehand. F_a is a friction force and directed in the reverse direction of the drop speed. This force influences the speed of the drop but not its direction. This implies that the angle of inclination φ of the drop trajectory is determined by F_g and F_d. From Fig. **1.7** we conclude that

$$\tan\varphi = \frac{F_g}{F_d}.$$

To estimate the train speed v_{train} from this relation, we must know how F_d depends on v_{train}. We investigate how dimensional analysis can help us to discover this relation. The friction between the drop and the passing air will depend on drop diameter $\mathcal{D}$ with $[\mathcal{D}] = L$, air density ρ with $[\rho] = M/L^3$, and air speed given by v_{train} with $[v_{\text{train}}] = L/T$. F_d is a force, so $[F_d] = ML/T^2$. The friction force exerted by a flow on an object moving through the flow is measured by the viscosity η with $[\eta] = M/LT$. It measures the internal friction. Common sense tells us that syrup has a larger viscosity than water. The value of η for a gas or fluid can be measured by dropping an object in the medium. Under influence of gravity it will initially accelerate. After some transient time its speed will become constant, since then the friction force comes into equilibrium with the gravitational force. This phenomenon is experienced by, e.g., parachutists. The viscosity can directly be deduced from

the equilibrium speed of a standardized object. In determining an expression for F_d the five quantities F_d, v_{train}, D, L, and η are involved. Since three physical dimensions play a role, the system has two dimensionless quantities.

Exercise 1.3e.

Check that we can choose for these quantities:

$$F_d^* = \frac{F_d}{\rho\, \mathcal{D}^2\, v_{\text{train}}^2}\,, \quad Re = \frac{\rho\, \mathcal{D}\, v_{\text{train}}}{\eta}.$$

The dimensionless *Reynolds number* Re plays an important role in all flow problems. It measures the ratio of the convective and viscous forces and is named after Osborne Reynolds, a researcher in fluid mechanics. We expect that for a still unknown function f^* it will hold that

$$F_d^* = f^*(Re).$$

For the drag force we thus have

$$F_d = \rho \mathcal{D}^2 v_{\text{train}}^2 f^*(Re).$$

From this we cannot deduce how F_d scales with v_{train} since Re also contains v_{train}. To answer this intricate question one has to determine the form of f^* from measurements. These data are given in Fig. **1.8**. Note that this form is universal and holds for all flows, thanks to the dimensionless formulation. The conclusion from these data is that

$$f^*(Re) \sim \begin{cases} \dfrac{1}{Re} & \text{if } Re < 10, \\ 1 & \text{if } Re > 100. \end{cases}$$

The range $10 < Re < 100$ is a transition region. To find the order of magnitude of Re for a moving train, we substitute some data. For air we have $\rho \approx 1.3$ kg/m^3 and $\eta \approx 1.5\ 10^{-5}$ kg/(m.s). The size of a droplet is $\mathcal{D} \approx 5.10^{-3}$ m. The velocity of the train varies from 0 to, say, 50 m/s. Substituting these numbers we find that $Re > 1000$ if $v_{\text{train}} > 2.5$ m/s (i.e., 10 km/h). This leads to the conclusion that for all relevant train speeds we have that

$$F_d = c_1 \rho \mathcal{D}^2 v_{\text{train}}^2$$

for some constant c_1 which does not depend on ρ, $\mathcal{D}$, and v_{train}. Eventually, we arrive at

$$v_{\text{train}} = \left(\frac{mg}{c_1 \rho \mathcal{D}^2 \tan\varphi} \right)^{1/2}.$$

This expression can be reduced a bit by noting that the drop is approximately a half sphere, so $m = \frac{1}{12}\pi \mathcal{D}^3$ if we take the density of water equal to unity. Since g and ρ hardly vary, we may write

$$v_{\text{train}} = c_2 \frac{\sqrt{\mathcal{D}}}{\sqrt{\tan\varphi}}.$$

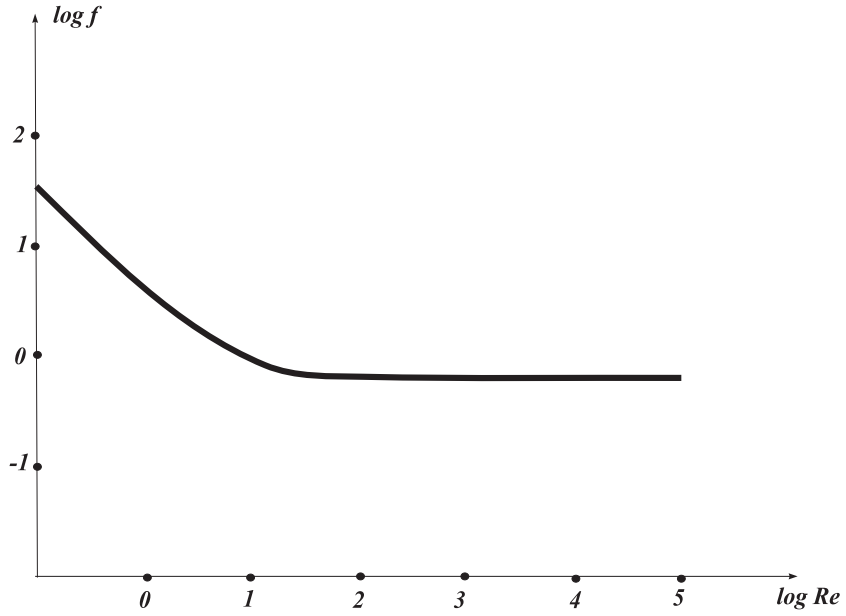

Figure 1.8. The drag force felt by an object in a flow as a function of the Reynolds number Re.

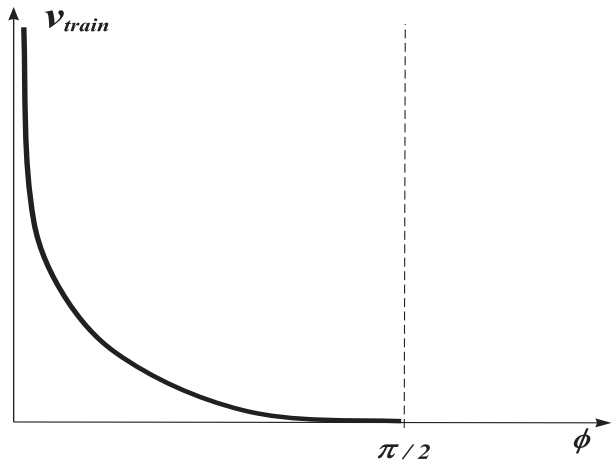

Figure 1.9. The dimensionless train velocity as a function of the angle ϕ indicated in Fig. **1.7**.

The value of c_2 can be determined from one observation. This is the relation we were aiming at. In Fig. **1.9** we plot the dimensionless velocity $v^*_{\rm train} \equiv v_{\rm train}/(c_2\sqrt{\mathcal{D}})$ as a function of φ. Note that the accuracy with which φ must be measured becomes increasingly important if φ becomes smaller and smaller and thus the train speed becomes increasingly higher. □

Example 1.3f. Ship modeling.
Let us model a ship of length ℓ sailing at constant speed v, as sketched in Fig. **1.10**. The motion of the ship transfers energy from the ship to the water as a result of viscous friction. This energy is used partly to induce surface waves and partly to overcome the internal friction of the turbulent motion of the water. In view of these effects, the acceleration of gravity g, the density of water ρ, and the viscosity η each will play a role, with dimensions $[g] = L/T^2$, $[\rho] = M/L^3$, and $[\eta] = M/LT$, respectively. If we assume that the ship is streamlined such that its height and width are not of importance, the system has five variables and parameters. Because three dimensions are involved, the number of dimensionless quantities is two. We can choose for these quantities

$$Fr = \frac{v}{\sqrt{\ell g}}, \quad Re = \frac{\rho \ell v}{\eta}.$$

Fr is called the *Froude number* after William Froude, a famous ship builder. Re is the Reynolds number that we already met in Example 1.3.e. Because real-life experiments are difficult to conduct for these systems, it is very attractive to perform experiments on (physical) models in which all sizes are scaled down by a certain factor. The conclusions from these experiments are valid for the original system only if both systems are described by the same dimensionless (mathematical) model. So, Fr and Re have to remain constant upon scaling. In practice, the values of g, ρ, and η can hardly be adjusted. To keep Fr constant, $v/\sqrt{\ell}$ may not change, and to keep Re constant, $v\ell$ must be preserved. The experimenter must be aware that these requirements can never be managed in the same experiment. This implies that in one experiment with scaled ship models, only a restricted aspect of the real situation can be studied. □

Exercise 1.3f.

a. *Check that we can indeed choose*

$$Fr = \frac{v}{\sqrt{\ell g}}, \quad Re = \frac{\rho \ell v}{\eta}$$

as dimensionless numbers.

b. *Why is it not possible to design the test facilities in such a way that both the Froude and the Reynolds number are preserved? Can you give an argument for why the*

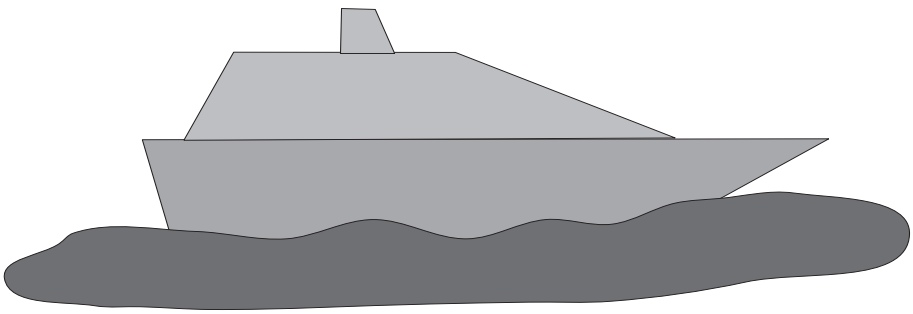

Figure 1.10. Study of scaled ship models is useful only if one realizes the consequences of dimensional analysis.

scaling of a ship is usually determined by the Froude number and not the Reynolds number? Will this be the same for testing an airplane in a (low-speed) wind tunnel?

c. *To conduct experiments on a ship 100 m long that sails with a maximum speed of 35 km/hr, one uses in a laboratory a model of the ship. Towing in the laboratory is restricted to velocities of at most 7 km/hr. What is the smallest scale of the model that can be used?*

1.4 Scaling

The aim of *scaling* is to reduce the number of parameters in a given model. So, a prerequisite of the technique of scaling is knowledge of the equations governing the system. Scaling does not necessarily yield dimensionless quantities. We introduce the technique via examples.

Example 1.4a. The scaled harmonic oscillator.
We apply the scaling technique to the harmonic oscillator introduced in Example 1.2a. As pointed out in that example and in Example 1.3f, this system has the two variables u and t and the seven parameters $m, c, k, F_0, \omega, u_0, v_0$. Consider the scaling of variables

$$\eta = \frac{u}{a}, \quad \tau = \frac{t}{b}$$

with the scaling factors a and b still unspecified. Substitution of this transformation into equation of motion (1.1) leads to

$$\frac{ma}{b^2}\ddot{\eta} + \frac{ca}{b}\dot{\eta} + ka\eta = F_0 \sin(\omega\tau),$$

where the time derivative is now with respect to τ. Dividing all terms by the factor ma/b^2 and choosing $b = \sqrt{m/k}$ and $a = F_0/k$ we obtain the reduced equation

$$\ddot{\eta} + c'\dot{\eta} + \eta = \sin(\omega'\tau)$$

with the new parameters $c' = c/\sqrt{km}$ and $\omega' = \omega\sqrt{m/k}$. The initial values u_0 and v_0 have to be scaled accordingly to u_0' and v_0'. Note that this reduction reveals that the harmonic oscillator depends on only four parameters (including the initial parameters). To study the behavior of the solution as a function of the parameters, it suffices to vary only the friction coefficient c' and the angular frequency ω' (apart from the initial values). We remark that the scaling procedure does not lead to a unique choice for the scaling factors a and b. An alternative choice is $b = 1/\omega$ and $a = F_0/(\omega^2 m)$, which yields the equation

$$\ddot{\eta} + c'\dot{\eta} + k'\eta = \sin(\tau)$$

with $c' = c/(\omega m)$ and $k' = k/(\omega^2 m)$. So, for a complete analysis of the system it suffices to vary only these c' and k' (apart from the initial values). □

Exercise 1.4a.

Compare the methods and results of scaling and dimensional analysis when applied to the driven, linear spring in Examples 1.3 *and* 1.3b.

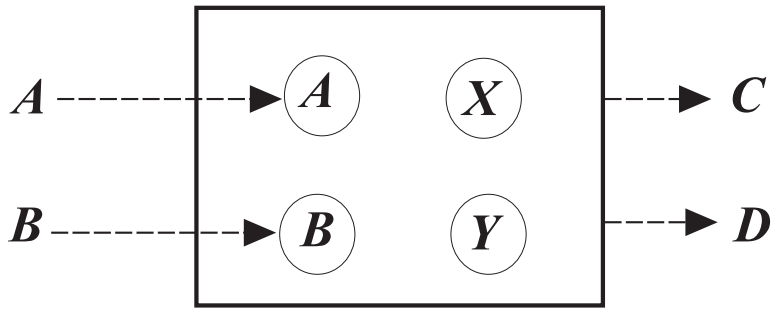

Figure 1.11. Sketch of the chemical reaction in Example 1.4b. The substances A and B are continuously supplied so that their concentrations are constant. The intermediate products X and Y are formed in reactions between A and B. C and D are the final products.

Example 1.4b. Chemical reaction.
Consider a hypothetical chemical reaction, the so-called Brusselator, with substances A, B, C, D, X, Y involved. The situation is sketched in Fig. **1.11**. C and D are produced from A and B with X and Y as intermediates. The reaction has the following irreversible stages:

$$\begin{aligned} A &\xrightarrow{k_1} X, \\ B + X &\xrightarrow{k_2} Y + C, \\ 2X + Y &\xrightarrow{k_3} 3X, \\ X &\xrightarrow{k_4} D. \end{aligned}$$

The capital letters denote reagents, while the constants k_i over the arrows indicate the reaction rates. It is assumed that A and B are excessively available so that the concentrations of A and B can be taken to be constant. We denote the concentrations of A, B, X, and Y by a, b, x, and y, respectively. The reaction equations for x and y are then

$$\begin{aligned} \dot{x} &= k_1 a - k_2 b x - k_4 x + k_3 x^2 y, \\ \dot{y} &= k_2 b x - k_3 x^2 y. \end{aligned}$$

□

Exercise 1.4b.

Use scaling of (x, y, t) *to* (η, ξ, τ) *to show that these equations can be reduced to*

$$\begin{aligned} \dot{\eta} &= \alpha - (\beta + 1)\eta + \eta^2 \xi, \\ \dot{\xi} &= \beta \eta - \eta^2 \xi. \end{aligned}$$

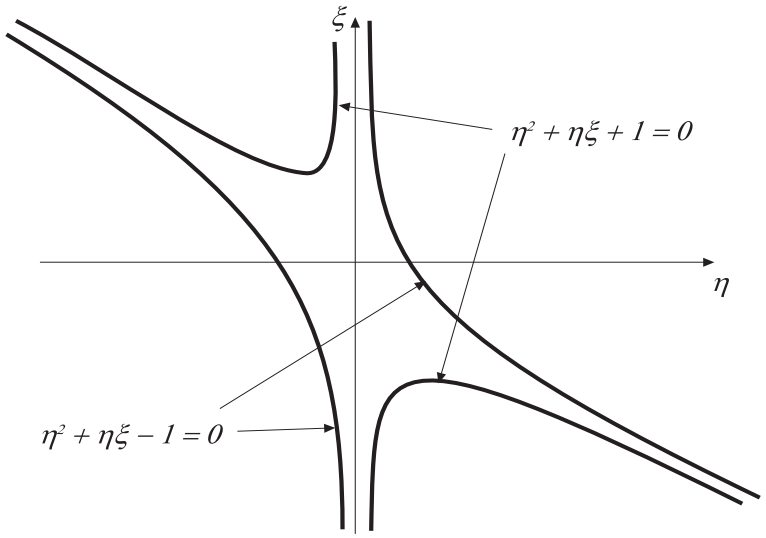

Figure 1.12. Solution sets of (1.5) after scaling.

Note that after this procedure the number of parameters is considerably reduced, since instead of the original parameters $a, b, k_1, \ldots, k_4$, the final set of equations contains only the two parameters α and β.

Exercise 1.4c.
Use the method of scaling to show that the equation

$$x^2 + axy + b = 0\,, \tag{1.4}$$

with variables x, y and parameters a, b, can be reduced to the equivalent equation

$$\eta^2 + \eta\xi + \mathrm{sign}(b) = 0 \tag{1.5}$$

with variables η, ξ. This implies that the structure of the solutions of (1.4) *can be caught without varying the parameters. It suffices to study* (1.5) *for the two cases sign$(b) = +1$ and -1. These curves are plotted in Fig.* **1.12***.*

1.5 Challenging problems

In this section we challenge the reader to apply the techniques of dimensional analysis and scaling to situations that are quite difficult to model. For application of dimensional analysis we need to know only the dimensions of the variables and parameters involved in the system. For scaling, one needs to start from the governing equations. In the present cases the reader must take these equations for granted; they will be derived in Chapter 2. Here, the focus is on the reduction of the models via scaling and not on the derivation of the models themselves.

1.5.1 The Prandtl–Blasius problem

The following description of flow above a flat plate is a simple model with which to study, for example, the following phenomena:

- The disturbance of the air by a thin airfoil of an airplane that cruises at constant altitude with constant speed U.
- The disturbance of a strong, uniform wind by vegetation, buildings, etc.
- The disturbance of water flowing over a rough surface.

In all these cases, the flow (air, water) near the surface experiences resistance from the presence of the airfoil, obstacles, etc. In an idealized way, we model the airfoil or the ground by a horizontal flat plate (standing still) with the air or water flowing over it, assuming that the flow is uniform before it reaches the plate. We choose coordinates such that the plate is in the half plane ($x \geq 0, -\infty < y < \infty,\ z = 0$). See Fig. **1.13**. Since the y-coordinate is irrelevant, we shall omit it. In front of the plate edge, where $x \leq 0$, the flow is uniform with velocity U in the positive x-direction. The fluid has mass density ρ with $[\rho] = M\,L^{-3}$ and viscosity η with $[\eta] = ML^{-1}T^{-1}$. Close to the plate the friction between plate and flow decelerates the flow. This friction causes the flow velocity near the

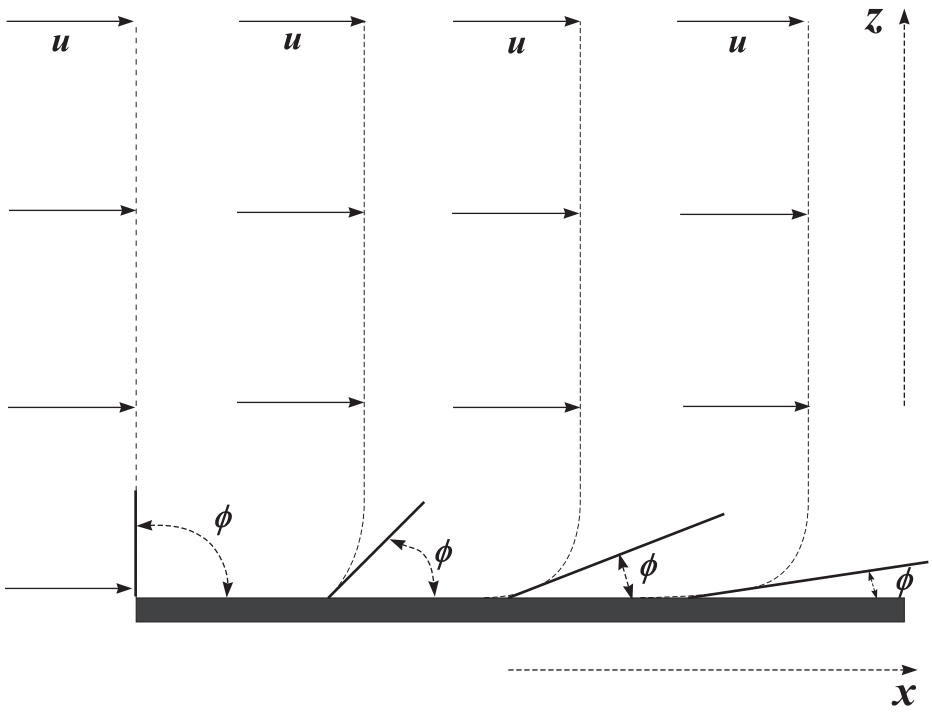

Figure 1.13. Time development of the velocity profile of air flow approaching and behind a trailing edge.

plate to drop. For increasing x, the boundary condition will approach the no-slip condition with the velocity at the plate exactly vanishing. For large x values, the velocity profile will become independent of x and approach a stationary profile, independent of x and thus t. We want to know how fast this convergence takes place as a function of time t and thus of position x. In Fig. **1.13** the velocity profiles above the plate are sketched for different x-positions. Let the velocity of the flow be denoted by (u, w) with u the velocity in the x-direction and w the velocity in the normal z-direction. The so-called shear rate is the variation of the horizontal velocity u in the normal direction. This is commonly denoted as $\dot{\gamma}$ (pronounced "gammadot"):

$$\dot{\gamma}(x, z) = \frac{\partial u}{\partial z}(x, z). \tag{1.6}$$

Its value at the plate is denoted as

$$\dot{\gamma}_0(x) = \dot{\gamma}(x, 0). \tag{1.7}$$

This quantity depends only on the distance x from the edge of the plate, and it is this dependence that we want to investigate. In Fig. **1.13** an angle $\varphi(x)$ is indicated. It is related to $\dot{\gamma}_0(x)$ via $\dot{\gamma}_0(x) = \tan\varphi(x)$. Far from the plate, where z is large, the flow is little influenced by the presence of the plate, and so there we may take $(u, w) = (U, 0)$.

In addition to the dependence on the distance x from the edge, the shear rate also will depend on the viscosity, the velocity U, and the density ρ. In the steady state we can generally write

$$\dot{\gamma}_0 = \dot{\gamma}_0(x; \eta, U, \rho).$$

In the following steps we want to find this relationship as precisely as possible. We first determine how far we can go with dimensional analysis in the first two exercises, whereas in the rest of this section we will apply scaling and use information from the governing equations.

a. *Determine two dimensionless variables from the set $\dot{\gamma}_0$, x, η, U, and ρ.*

b. *Show that for some function f it holds that*

$$\dot{\gamma}_0 = \frac{U^2\,\rho}{\eta}\, f\left(\frac{U\,x\,\rho}{\eta}\right).$$

To apply scaling we need information about the governing equations. The actual equations of motion are given by

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0, \tag{1.8}$$

$$\rho\left(u\,\frac{\partial u}{\partial x} + w\,\frac{\partial u}{\partial z}\right) = \eta\,\frac{\partial^2 u}{\partial z^2}. \tag{1.9}$$

The first equation (vanishing of the divergence of the velocity field) expresses the incompressibility of the flow. The second equation expresses the balance between the convection force (at the left-hand side) and the viscous friction force (at the right-hand side).

We apply the scalings

$$\bar{u} = \frac{u}{U}\ ,\quad \bar{w} = \frac{w}{W}\ ,\quad \bar{x} = \frac{x}{X}\ ,\quad \bar{z} = \frac{z}{Z}$$

with U given and X, Z, and W to be chosen later on.

c. *Show that in* (1.8) *and* (1.9) *the number of parameters reduces if the following two conditions are satisfied:*

$$\frac{X\,W}{Z\,U} = 1,\quad \frac{X\,\eta}{Z^2\,U\,\rho} = 1. \tag{1.10}$$

In the rest of this section we assume these relations hold. Given Z, the second condition determines X, after which W follows from the first condition. So, of the three parameters W, X, and Z, only Z is still free to be chosen. Note that the scaled variables are not necessarily dimensionless.

d. *Determine a scaling, i.e., choose Z, such that the scaled variables are dimensionless.*

Now, we take Z again as a free parameter. The nonuniqueness of the scaling can then be exploited to find an explicit expression for the stress. Note that this can be done without solving the equations explicitly. The strategy is as follows. The scaled shear rate $\bar{\dot{\gamma}}_0$ at the plate is defined as

$$\bar{\dot{\gamma}}_0 := \frac{\partial \bar{u}}{\partial \bar{z}} = \frac{Z}{U}\,\dot{\gamma}_0\ . \tag{1.11}$$

Since $\dot{\gamma}_0$ depends only on x, it must hold for some function h that

$$\dot{\gamma}_0(x) = h(x)\ . \tag{1.12}$$

e. *Use* (1.11), (1.12)*, and the fact that $x = X\,\bar{x}$ to find an expression for $\bar{\dot{\gamma}}_0(\bar{x})$ in terms of the function h. The resulting relation depends on Z. Since it must identically hold for any value of Z, we may draw a conclusion about the form of h. Show that for some positive constant c it must hold that*

$$h(x) = \frac{c}{\sqrt{x}}\ .$$

We emphasize that this argument can be applied only because Z was kept arbitrary. If Z were fixed, e.g., to make all quantities dimensionless, this relation could not be derived.

f. *Now, translating into the original, unscaled, quantities we show that*

$$\dot{\gamma}_0(x;\,\eta,\,U,\,\rho) = c\,\sqrt{\frac{U^3\,\rho}{\eta\,x}}\ ,$$

which is the relation we aimed at. Compare this with the result under b *and determine the explicit form of the function* f *mentioned there.*

1.5.2 Heat conduction in a bar

We consider heat conduction in a rod of length ℓ. The rod is assumed to be thermally isolated everywhere. Starting with a given temperature distribution $u(x, 0)$ over the rod, we are interested in the time evolution of the temperature profile $u(x, t)$. As will be explained in Example 2.3d in Chapter 2, the speed of the heat conduction along the rod is determined by the so-called thermal conductivity κ. The dimension of κ is L^2/T. As an initial profile we take a distribution which is everywhere vanishing except for a peak in the origin:

$$u(x, 0) = u_0\, \delta(x), \tag{1.13}$$

where $\delta(x)$ is the so-called delta function. Its definition is given in §3.3.3.

a. *Determine the dimension of the constant u_0. To that end, integrate the initial condition over some interval including $x = 0$ and use the properties of the delta function.*

The variables are thus u, x, and t, and the parameters are ℓ, κ, and u_0.

b. *Find three dimensionless quantities, choosing them such that they are just scalings of the variables u, x, and t.*

Now we assume that the rod is infinitely long. This implies that the parameter ℓ is no longer relevant and that the number of dimensionless quantities equals two.

c. *Show that these dimensionless quantities are appropriately chosen as $q_1 = x^2/(\kappa t)$ and $q_2 = u\sqrt{\kappa t}/u_0$.*

In view of the expected behavior of the system we may write

$$q_2 = f^*(q_1) \tag{1.14}$$

for some function $f^*(q_1)$. The important conclusion from dimensional analysis is that in an infinitely long rod the temperature is a function not of position x separately but of the quotient x^2/t. Note that this does not hold for a rod of finite length ℓ. To find the explicit form of f^*, one needs the governing equation. The equation that describes the time evolution of the temperature in the bar follows from conservation of heat. It will be derived in Example 2.3d of Chapter 2. The resulting partial differential equation (PDE), the so-called heat diffusion equation, reads as

$$\frac{\partial u}{\partial t} = \kappa\, \frac{\partial^2 u}{\partial x^2}\,. \tag{1.15}$$

d. *Show that the dimension of κ given above agrees with the dimension following from this equation.*

e. *Use* (1.14) *to rewrite the heat diffusion equation* (1.15) *in terms of* q_1, q_2, *and* f^*, *and show that this leads to the following ODE for* $f^*(q_1)$*:*

$$4q_1 \frac{\partial^2 f^*}{\partial q_1^2} + (q_1 + 2)\frac{\partial f^*}{\partial q_1} + \frac{1}{2} f^* = 0 \tag{1.16}$$

for $0 < q_1 < \infty$.

f. *Check that*

$$f^*(q_1) = c\, e^{-\frac{1}{4}q_1}$$

is a solution of (1.16) *for any constant* c.

g. *Which conservation principle can be invoked to determine* c*? Check that the dimensional solution reads as*

$$u(x,t) = \frac{u_0}{\sqrt{4\pi\kappa t}}\, e^{\frac{-x^2}{4\kappa t}} .$$

Thus, at each time t the spatial temperature distribution is a Gaussian distribution in x, with mean 0 and standard deviation determined by κt. The temperature decays with increasing time, since the initial peak of heat is conducted along the rod. This is sketched in Fig. **1.14** for different values of t.

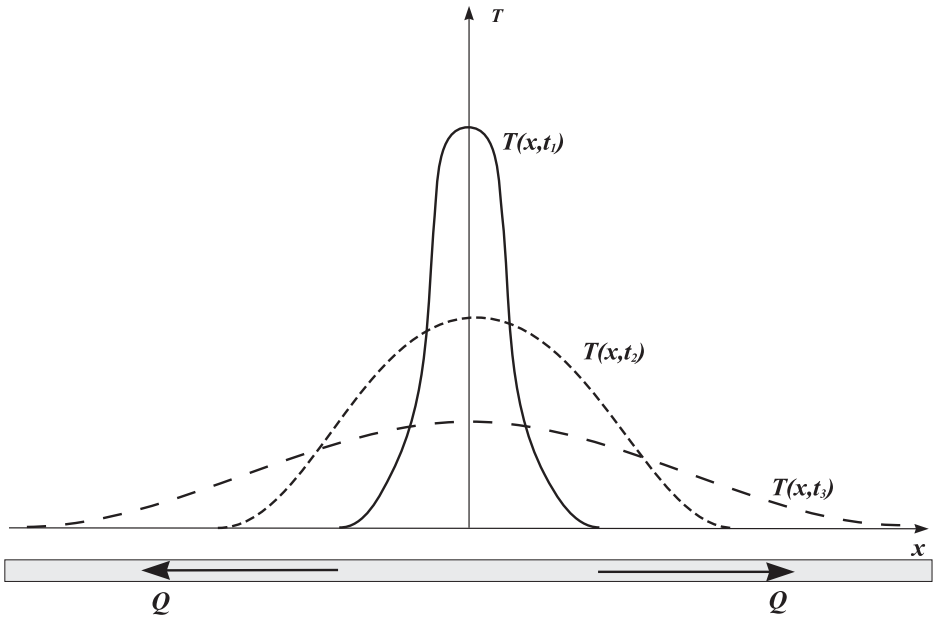

Figure 1.14. Temperature profiles for times $t_1 < t_2 < t_3$ in a long, thin rod.

1.5.3 Water waves

Consider a layer of fluid, for example, water, above a horizontal bottom. See Fig. **1.15**. When the fluid is set into motion, by whatever cause, the fluid particles will start to move, interacting with each other, and are influenced by gravity. When the upper surface of the fluid is *free*, this surface will also be deformed by the particle motions. Looking only at the surface, the deformation is often of a characteristic type, like waves that are caused by throwing a stone in a pond. In principle, the particles below the surface determine the surface elevation. However, when the fluid is incompressible, and when so-called irrotational flow is considered, it is possible to describe approximately the elevation of the surface without reference to the internal motion. We will not *derive* the governing equations from first principles, but we will investigate a postulated description of the phenomenon via dimensional analysis and scaling. We make the following assumptions:

- The fluid is incompressible with uniform mass density, which we set equal to unity for convenience.
- No forces act from above on the free surface; for instance, effects of wind are neglected.
- In the horizontal plane, the motion is uniform in one direction; that is, we consider *plane waves*. This implies that if we take the x-axis, say, as the direction of propagation, the motion does not depend on the transverse horizontal direction.

We take the z-axis in the vertical direction, opposite to the direction of gravity with its origin at the undisturbed water surface. The distance between bottom and undisturbed water surface is H, and g denotes the acceleration of gravity. The modeling focuses on the elevation $u(x, t)$ of the surface with respect to the undisturbed level $z = 0$.

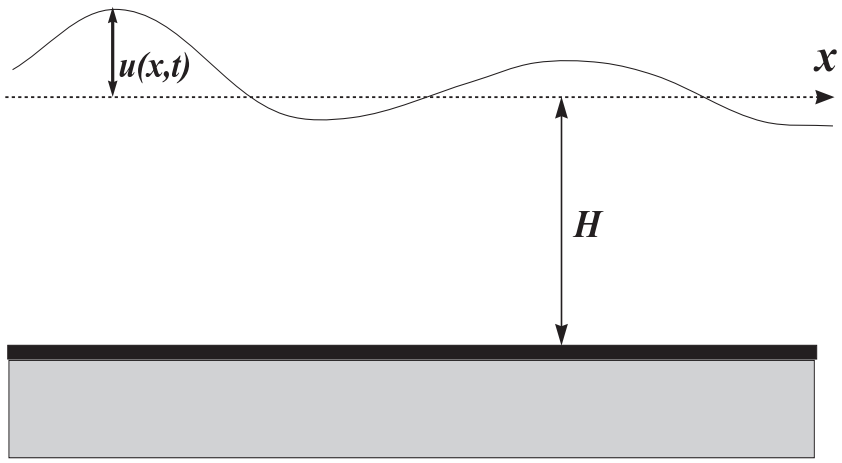

Figure 1.15. In the modeling of moving water waves one wants to find the elevation $u(x, t)$ of the surface with respect to the completely flat rest profile.

a. *With the (constant) mass density removed from consideration, the problem has three variables (u, x, t), two parameters (g, H), and two dimensions (length and time). Find three dimensionless quantities.*

A rough idea about wave propagation can be obtained by studying harmonic profiles, say, of the form

$$u(x,t) = a\cos\left(\frac{2\pi(x+Vt)}{\lambda}\right),$$

where a is the *wave amplitude* and λ is the *wave length*. This surface profile propagates at *wave velocity* V.

b. *Take these three quantities into account, together with the parameters g and H, and show that the problem is described by a relation between the normalized amplitude, the wave length, and the dimensionless velocity:*

$$f\left(\frac{a}{H}, \frac{\lambda}{H}, \frac{V}{\sqrt{gH}}\right) = 0.$$

c. *From the above result, try to explain the observation—which can be made at any coast when one looks at the waves running into the shore—that waves approach the coast perpendicularly, even when the coastal boundary is irregular.*

In 1895 Korteweg and de Vries [17, 24] published an equation for the elevation $u(x,t)$ that describes the surface elevation in a certain order of approximation. In the derivation of the equation it was assumed that the waves were "rather low" (small amplitude) and "rather long." This equation is known as the *Korteweg–de Vries equation* (KdV equation). This equation became famous when it was found to have special mathematical properties. We do not discuss them here but instead focus on the original purpose of this model. It reads as

$$\frac{\partial u}{\partial t} = -c\,\frac{\partial u}{\partial x} - \frac{cH^2}{6}\frac{\partial^3 u}{\partial x^3} - \frac{3c}{2H}u\frac{\partial u}{\partial x}, \tag{1.17}$$

where the parameter $c = \sqrt{gH}$ has been introduced since it plays an important role in the physical phenomenon, as we shall see later. This equation shows the time evolution of the free surface: at a fixed position, the time derivative of the elevation depends (in a complicated way) on the spatial derivatives of the elevation: *temporal variations and spatial variations are coupled*, which is characteristic for a PDE. Understanding the coupling would mean that the meaning of each of the three terms in the right-hand side should be clear, which at this moment is not possible. The equation is rather difficult in the sense that it is not easy to find explicit solutions. In the following we will try to interpret the various terms in the right-hand side and, particularly, try to understand how the underlying modeling assumptions of long and low waves are shown in this equation.

The starting point is rather characteristic. We perform a scaling of the variables without specifying at this moment the scaling factors; these will be determined, or chosen, at a later instance.

$$\bar{x} = \frac{x}{L},\ \bar{t} = \frac{t}{\tau},\ \bar{u}(\bar{x},\bar{t}) = \frac{u(x,t)}{a}.$$

Note that at this time we do not try to make the new variables $\bar{u}, \bar{x}, \bar{t}$ dimensionless. To simplify the notation and since confusion is unlikely, we drop the overhead bar, keeping in mind that the equations below are in terms of scaled quantities.

d. *Show that the scaling leads to the following form of the equation:*

$$\partial_t u + \alpha\, \partial_x u + \beta\, \partial_x^3 u + \gamma\, u\, \partial_x u = 0\,, \tag{1.18}$$

where

$$\alpha = \frac{\tau\, c}{L}, \quad \beta = \frac{\tau\, c\, H^2}{6\, L^3}, \quad \gamma = \frac{3\tau\, c\, a}{2H\, L}\,.$$

Here and often in the following, the spatial derivatives are indicated by $\partial_x u$ for $\frac{\partial u}{\partial x}$ and $\partial_t u$ for $\frac{\partial u}{\partial t}$, and similarly for higher-order derivatives. Observe that by the scaling we introduced three additional parameters (L, τ, and a), which together with H and g (or, equivalently, H and c) brings the total number of parameters to five. However, the equation shows that only specific combinations of these parameters (namely, α, β, and γ) play a role. We show how scaling can be used for various purposes.

To study the original KdV equation (1.17), it is sufficient to study (1.18). Therefore we would like to reduce the numbers of parameters in this equation as much as possible. This can be done by choosing the scaling coefficients in an appropriate way:

e. *Determine scaling coefficients such that the KdV equation* (1.18) *gets the following, parameterless form:*

$$\partial_t u + \partial_x u + \frac{1}{6}\, \partial_x^3 u + \frac{1}{6}\, u\, \partial_x u = 0\,. \tag{1.19}$$

This shows that by means of scaling, a lot of progress can be made. One needs only to study (1.19), and the results are directly applicable to various physical situations.

Scaling can be used for another purpose, as we shall show now. In particular we show how scaling arguments can give insight into the meaning and relative importance of the three terms in the right-hand side of (1.17) or, equivalently, of (1.18). As for comparisons in magnitude, it is necessary to make explicit that when dealing with the scaled variables, we assume these to be of unit order. So, in the following, x, t, and u and *all the derivatives of* u are considered to be of order unity. "Small" is then measured with respect to unity. Below, some limiting cases are considered.

Small amplitude waves

Since u is of order one, the value of a is a measure of the amplitude of the surface elevation: for large a the physical wave heights are large, whereas the limit $a \to 0$ means that waves of infinitesimal amplitude are considered. Since a appears only in the coefficient γ, we can simply take the limit $a \to 0$ by taking $\gamma \to 0$. The resulting equation is

$$\partial_t u + \alpha\, \partial_x u + \beta\, \partial_x^3 u = 0. \tag{1.20}$$

Note that this equation is linear. It can be solved by means of Fourier transformation techniques introduced in Chapter 3. This observation explains that the nonlinear term in the KdV equation describes effects that are due to the finiteness of the wave heights.

Long and short waves

The value of the parameter L determines the length of the physical spatial interval in which changes take place: for small L this interval is small, whereas for large L changes occur over a large physical interval. Small and large values of L correspond to short and long waves, respectively.

g. *If we apply the transformation $X := L\,x$, a function $f(x)$ transforms into a function $g(x)$, say, via*

$$f(x) = g(X) = g(Lx).$$

Relate the derivatives of f to those of g and see the effect of the value of L.

All three parameters α, β, and γ in (1.18) contain the parameter L, and therefore the effect of taking a limiting value of L is not immediately clear. Thus, we argue as follows. The appearance of L in α can easily be scaled away by taking $\tau = L/c$. Then

$$\alpha = 1,\ \beta = \frac{1}{6}\left(\frac{H}{L}\right)^2,\ \gamma = \frac{3}{2}\left(\frac{a}{H}\right). \tag{1.21}$$

h. *Show that if L is given the dimension of length, τ will have the dimension of time, and show that the variables x, t, and u and the parameters α, β, γ are dimensionless.*

i. *Observe that now L appears only in the coefficient β. Keeping all other parameters fixed, look at the limit for long waves, and explain that the third order spatial derivative in the KdV equation describes effects that are due to the length of the waves under consideration; the longer the waves, the less this term contributes. Find the equation obtained in the limit for infinite long waves of finite amplitude.*

Long waves with small amplitudes

j. *Consider the limit of infinitesimally small, infinitely long, waves. Show that by taking the limits $a \to 0,\ L \to \infty$ in* (1.18)*, the equation reduces to*

$$\partial_t u + \partial_x u = 0. \tag{1.22}$$

The solutions of this equation are easily found. Show that for arbitrary f it holds that

$$u(x,t) = f(x - t\,).$$

This solution represents that the profile f is translated into the direction of the positive x-axis, undisturbed in shape and at unit velocity. For this reason, this equation is called the translation equation.

k. *Show, by realizing that x and t are scaled variables here, that infinitesimal waves of long wave length propagate at speed $c = \sqrt{gH}$.*

Now, let us return to (1.18) with the scaling given by (1.21):

$$\partial_t u + \partial_x u + \frac{H^2}{6L^2}\,\partial_x^3 u + \frac{3\,a}{2H}\,u\,\partial_x u = 0. \tag{1.23}$$

In the reasoning of Korteweg and de Vries, the last two terms are improvements of the translation equation in the sense that (some) effects of the nonlinearity and dependence on wavelength are taken into account. Being corrections, the coefficients should be small compared to unity and so of order $\varepsilon \ll 1$. The KdV equation takes both effects into consideration. Let us assume that both effects are equally important, say, of the same order ε.

l. *Show that this means that the KdV equation describes waves of amplitude a and wavelength L such that*

$$\frac{a}{H} = \mathcal{O}(\varepsilon),\ \frac{H}{L} = \mathcal{O}(\sqrt{\varepsilon}).$$

Note that the quotients a/H and L/H are dimensionless and that all length scales that are relevant for measuring the two effects depend on the (only!) physical length H that is present in the model. Waves for which amplitude and wavelength are related in this way are said to satisfy the Boussinesq assumption.

m. *For experiments in a towing tank in hydrodynamic laboratories, wave heights (i.e., twice the amplitude) up to* 10% *of the depth are considered, for instance, waves of amplitude* $0.3\,m$ *on a depth of* $6\,m$*. Determine the order of the wavelength of waves that satisfy the Boussinesq assumption.*

Chapter 2

Conservation Principles and Constitutive Relations

Models in science usually contain three ingredients:

1. Basic laws.
2. Constitutive relations.
3. Conservation principles.

Basic laws express properties that are well established by measurements and extensively checked by comparing theory and observations. For example, in classical mechanics the laws of Newton govern the dynamics, and for electromagnetic phenomena the Maxwell equations act as such. A basic law is applicable to a whole class of systems. This is in contrast to a constitutive relation, which characterizes a single system.

To illustrate the difference between basic laws and constitutive relations, let us consider the second law of Newton. This basic law states that the dynamics of a single mass m is governed by $F = ma$ with F the total force exerted on the mass and a its acceleration. To proceed with the modeling, the force F has to be specified. It will usually stem from the interaction with other particles or the gravitational or electromagnetic field. The formulas describing these interactions form the constitutive relation and are typical for the system under consideration. In some cases the constitutive relations are well known, and formulas for them have long been available. For example, for the gravitational force between masses, Newton discovered a highly reliable expression. It states that this force is proportional to the inverse of the distance squared and parallel to the line connecting the centers of gravity. In many cases, only restricted knowledge is available about the dependence of the interaction forces on the positions and velocities of the constituting particles. For example, for the friction of a mass sliding along a wall, only approximating expressions are known. It is a standard strategy to assume that the mathematical expression for such a force belongs to a certain class of functions containing a number of parameters. The values of the parameters are then estimated from experiments.

Conservation principles, also referred to as balance laws, are as fundamental as the basic laws but are of a different nature. Based on accurate observations and careful analysis, it has been found that in many systems, conservation of total mass, momentum, and/or energy

holds. So, if the amount of mass, momentum, or energy in some part of the system increases, it must decrease in the rest of the system. This insight leads to balance equations. These quantities are not always conserved separately. For example, transition of mass into energy and vice versa is possible, as nuclear fusion and fission show.

In this chapter we concentrate on the mathematical formulation of conservation principles. Meanwhile, we meet with several commonly used constitutive relations when modeling daily life problems, such as car traffic and heat conduction. Since we are also interested in conservation principles for continuous media, we first point out how a continuous description of matter arises from a molecular approach via averaging procedures. To get the reader acquainted with the ideas behind balance laws and constitutive equations, we start with simple, one-dimensional systems. Next, the balance laws are dealt with in more dimensions. We shall show that all these laws follow from one balance principle, the so-called transport theorem. A classical introduction to flow problems is [18]. General introductions to the art of mathematical modeling are given in [9, 29, 30, 25, 26, 15, 10].

2.1 Discrete versus continuous models

An important aspect of mathematical modeling is the choice of the level at which the system is studied. On a molecular level, matter consists of discrete particles, atoms and molecules, and one is interested in the detailed dynamics of all these particles. Consequently, a reliable model will consist of equations of motion for all the interaction forces. Such a model is of a *discrete* nature. In many cases, however, one is interested only in the properties of matter as it is observed on the macro level of human size. Then one wants to ignore the detailed dynamics on the micro level, since this manifests itself on the macro level only in an average sense. In that approach, matter is conveniently described as a continuum, characterized by properties such as density, stiffness, viscosity, and a constitutive equation relating the stress, i.e., the internal forces, to the strain, i.e., the internal deformations. The averaging procedure to obtain the properties on the macro level from the molecular structure may be extremely difficult to apply in numerical calculations. In the case of mass, the averaging procedure itself is easily understood as follows. From a microscopic point of view, it is not meaningful to speak about the mass in a particular point; it may happen that no particle is present at exactly that position. In continuum mechanics, the density of mass in a specific point is read as the amount of mass contained in a box around that point divided by the volume of the box. The size of the box must be so large that it contains a considerable number of particles so that the average mass density does not depend on the detailed structure of the particles. On the other hand, the box must be small compared to the size of the system so that it can be treated as a point from a macroscopic perspective. In the literature, the resulting function of position (and time) is referred to in many ways, e.g., as *mass density, mass distribution, mass profile,* or *mass field*. We shall use all these terms synonymously.

The car flow problem in Example 2.1a provides a nice example of the averaging procedure for mass. On the level of individual cars, one deals with a discrete model in which cars are counted individually. When driving a car ourselves, we are highly conscious of the discrete nature of road traffic. However, when observing road traffic from a great distance, for example, from a plane or in satellite photos, introducing a continuous car distribution is quite natural. The car flow example has the advantage that in the simplest version only one

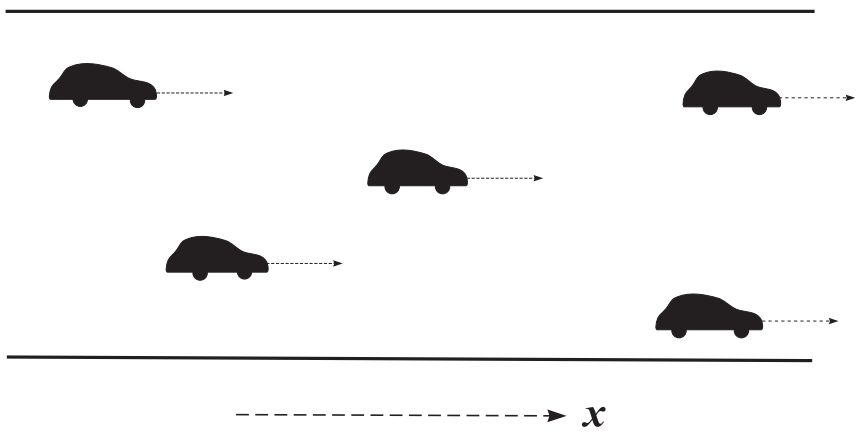

Figure 2.1. Sketch of one-way traffic without side streets.

spatial dimension is involved. It serves as an illustration of both conservation of mass and the transition from a discrete to a continuous model.

Example 2.1a. Car flow.
Consider a long, one-lane freeway with cars driving in one direction at different speeds. See Fig. **2.1**. We derive a model for the dynamics of the car traffic, assuming that no cars leave the freeway. The position along the road is denoted by x. Let $N(x,t)$ be the number of cars at time t that are fully included in the interval $[x-\frac{1}{2}L, x+\frac{1}{2}L]$ for some interval length L. Cars just crossing the boundaries are not counted. Then, N attains integer values, and a model in terms of $N(x,t)$ would be discrete. Next, we want to consider the traffic as a continuous fluid with density ρ. This density follows from N via averaging. To that end choose a characteristic length L. It must be so large that an arbitrary interval of length L always contains a considerable number of cars. On the other hand, it must be short in comparison to the road length. Then, we introduce the car density $\bar{\rho}(x,t)$ via

$$\bar{\rho}(x,t) = \frac{N(x,t)}{L}.$$

It is clear that this density depends on the interval length L. If L is very short, the density will wildly fluctuate, whereas for very large L the density will vary little. The specific choice of L depends on in how much detail we want to study the traffic variations.

Besides mass density, an essential quantity in a continuous model is the velocity density. In the first instance we could define it as the average of the velocities v_i of the $N(x,t)$ cars that at time t are completely included in the interval $[x-\frac{1}{2}L, x+\frac{1}{2}L]$. So,

$$\bar{v}(x,t) = \frac{1}{N(x,t)} \sum_{i=1}^{N(x,t)} v_i.$$

For car flow these density definitions may be satisfying, since the fluctuations in time are not very fast. For other systems, e.g., gas molecules, the movements of the individual particles

are so capricious that the density variations in mass and velocity take place at a much smaller time scale than we are interested in. To average out such fast fluctuations, we choose a time T, which is characteristic of the time scale in which we are interested. The time-averaged densities are defined as

$$\rho(x,t) := \frac{1}{T}\int_{t-\frac{1}{2}T}^{t+\frac{1}{2}T} \bar{\rho}(x,t')\,dt'$$

and

$$v(x,t) = \frac{1}{T}\int_{t-\frac{1}{2}T}^{t+\frac{1}{2}T} \bar{v}(x,t')\,dt'.$$

It should be realized that, in general, L and T cannot be chosen independently. □

Exercise 2.1a.

Determine for the car flow problem in Example 2.1a *realistic values for the averaging parameters L and T.*

An important aspect of the averaging procedure is that it usually yields densities that are smooth functions of x and t. In the following we assume that the densities are differentiable at most times and most positions of the system. Exceptions will be met in systems with, e.g., shocks and phase boundaries. Then, the relevant densities may exhibit localized jumps, which have to be treated with extra care. In Example 2.2b we meet with such a situation.

2.2 Mass and heat balances in one dimension

Before dealing with conservation principles in general, it is illustrative to work out mass conservation in one dimension. One could appropriately think of the car flow problem introduced in Example 2.1a, but the resulting formulas are generally valid. The key idea is that the balance of mass flows leads to a relation between $\rho(x,t)$ and $v(x,t)$. In a one-dimensional system, the mass in an arbitrary interval $[a,b]$ is given by

$$m_{a,b}(t) = \int_a^b \rho(x,t)dx\,. \tag{2.1}$$

Changes in $m(t)$ are due to possible mass flux through the boundaries at $x=a$ and $x=b$. In a formula this reads as

$$\frac{d}{dt}m_{a,b}(t) = \rho(a,t)v(a,t) - \rho(b,t)v(b,t) = -[Q(b,t) - Q(a,t)]. \tag{2.2}$$

Here, we introduce the notation

$$Q = \rho\, v.$$

Q is referred to as the *mass flux*, since it measures the amount of mass passing position x per unit of time. Rewriting the right-hand side of (2.2) and substituting (2.1) at the left-hand side we may write

$$\frac{d}{dt}\int_a^b \rho(x,t)dx = -\int_a^b \frac{\partial Q}{\partial x}(x,t)dx.$$

Interchanging differentiation and integration—which is allowed if ρ and v are smooth functions—we find

$$\int_a^b \left\{ \frac{\partial \rho}{\partial t} + \frac{\partial Q}{\partial x} \right\} dx = 0. \tag{2.3}$$

This expresses mass conservation in integral or *global* form. The differential or *local* form is obtained from the following lemma.

Lemma 2.2a (Lagrange).

Let f be a continuous function defined on an interval $[A, B]$. If it holds that

$$\int_a^b f(x)dx = 0$$

for each subinterval $[a, b] \subset [A, B]$, then f vanishes identically: $f(x) = 0$ for each $x \in [A, B]$.

Exercise 2.2a.

a. *Prove this lemma.*

b. *Generalize this lemma to functions of more variables.*

Applying this lemma, we obtain the differential or *local* form of conservation of mass:

$$\frac{\partial \rho}{\partial t} + \frac{\partial Q}{\partial x} = 0. \tag{2.4}$$

A more concise notation, which we shall regularly use in the following, is

$$\partial_t \rho + \partial_x Q = 0. \tag{2.5}$$

So, the time derivative of the mass density has to be equal to minus the spatial derivative of the mass flux.

In the context of the car flow model, crossing roads (side streets) act as sinks and sources of cars. See Fig. **2.2**. The conservation equations (2.3) and (2.5) have to be adjusted if mass sources or sinks are involved. Let the local creation or annihilation of mass be described by a mass flux density $S(x, t)$. So, the system has locally a mass source or sink, and S specifies the amount of mass added or removed per unit of time and per unit of length. If $S > 0$, mass is added to the system; if $S < 0$, the system has locally a mass sink. In the presence of sources/sinks $S(x, t)$, the global form of the conservation equation becomes

$$\int_a^b \{\partial_t \rho + \partial_x Q - S\} dx = 0$$

and the local form

$$\partial_t \rho + \partial_x Q = S.$$

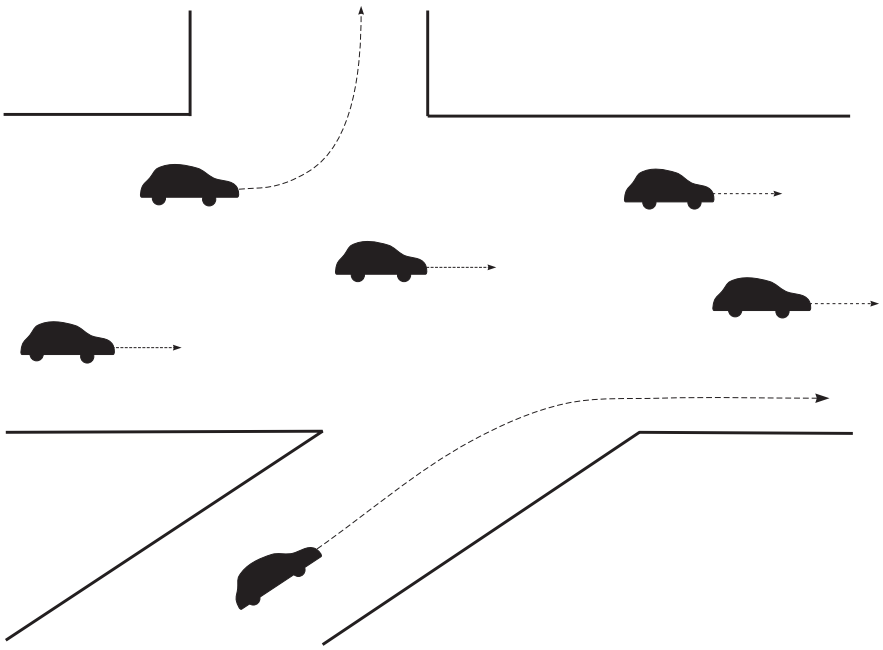

Figure 2.2. Sketch of one-way traffic. The side streets act as sources and sinks of cars.

Until now a conservation law has been specifically derived for mass. In fact, the resulting equations apply to any scalar quantity that is conserved. In the next example we show this for heat.

Example 2.2a. Heat conduction in one dimension.
Diffusion is important in many physical processes: dye will diffuse in water, smoke and pollutants will diffuse in air, heat will be conducted in solids, etc. Although the physical appearances may be different, the governing equations are quite similar and have a characteristic form. As a typical example, we will consider heat conduction. In §1.5.2 we introduced this system to show how dimensional analysis can be used to analyze its behavior. In a solid, the molecules may vibrate, and the larger the amplitudes of these thermal vibrations, the more energy is stored in these motions. The amount of (thermal) energy is measured by the temperature, and the specific heat c acts as the constant of proportionality. With u the temperature and ρ the mass density of the solid, the combination $\rho\, c\, u$, which has the dimension of energy per unit volume, is called the *heat* H. Heat is conducted via collisions and/or electromagnetic interactions between the molecules. We denote the corresponding heat flux Q.

Consider a thin rod positioned along the x-axis and let $u(x,t)$ denote the temperature density. See Fig. **2.3**. The heat $H(t)$ in an interval $[a,b]$ at time t is given by

$$H(t) = \int_a^b \rho\, c\, u\, dx.$$

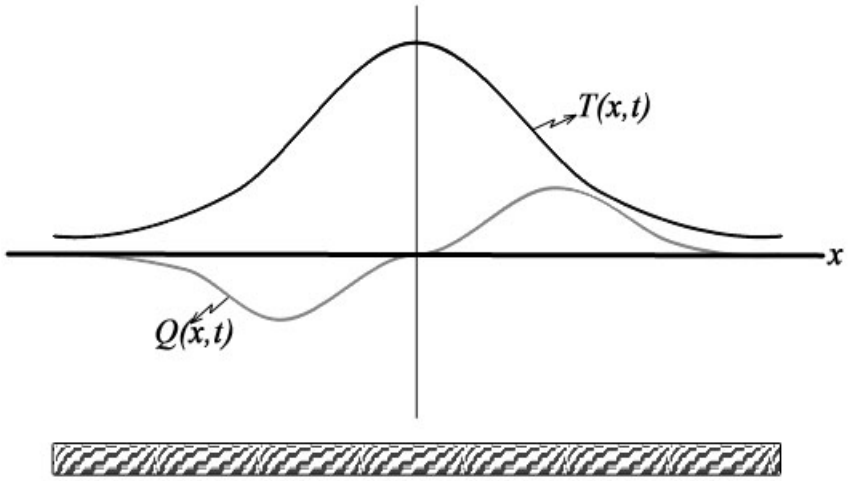

Figure 2.3. Temperature and heat flux profiles in a long, thin rod.

Changes in $H(t)$ can be caused by two effects:

- Energy is produced locally, for instance, by heating the rod from outside or by chemical reactions inside the rod. We denote this energy source by a density $S(x, t)$, the supply of heat per unit of length and per unit of time.
- Heat leaves/enters the interval through the endpoints.

Taking both causes into account, the change of heat per unit of time in $[a, b]$ is given by

$$\frac{dH}{dt} = \frac{d}{dt}\int_a^b \rho\, c\, u\, dx = -\,[Q]_{x=a}^{x=b} + \int_a^b S\, dx.$$

Using the arguments given in §2.2 for car flow, this can be rewritten as a global balance law for heat:

$$\int_\Omega \{\partial_t(\rho\, c\, u) + \partial_x Q - S\}\, dx = 0.$$

If the integrand is continuous, we thus find

$$\partial_t(\rho\, c\, u) + \partial_x Q = S, \tag{2.6}$$

the local heat balance law. These balance equations are nothing but special cases of the so-called transport theorem, which we derive in §2.4 in a very general setting. □

The differential form of the conservation equation is not applicable in regions where the densities involved are not continuous. As mentioned above, this will be the case if somewhere in the system sudden changes occur. Examples are shock fronts and phase

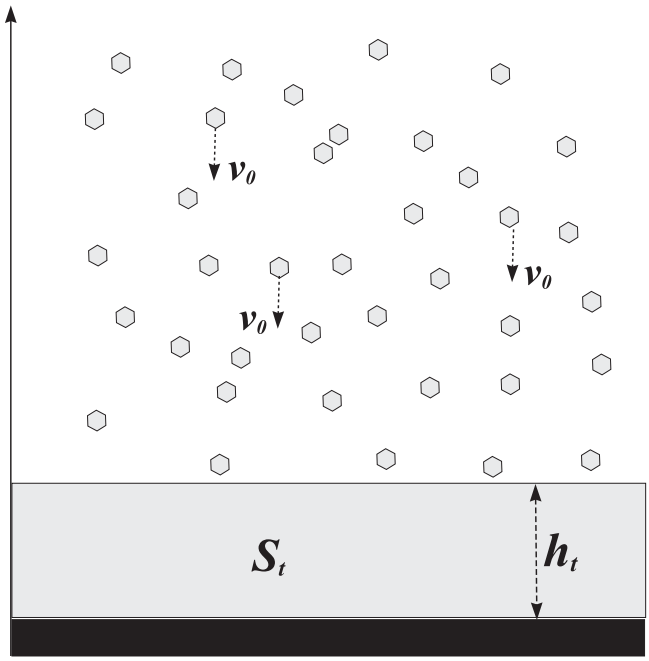

Figure 2.4. Sedimentation: precipitating particles form a growing layer.

boundaries. In these regions the conservation principle has to be dealt with carefully. The integral formulation then still applies. In the following example it is shown how the integral form of conservation of mass allows for the determination of the position of a shock front.

Example 2.2b. Front tracking for falling snow.
Consider snow falling and forming a layer on the earth. See Fig. **2.4**. We want to find the increase of the thickness of the layer with time. In an initial approach, we make the following modeling assumptions:

- While it is falling, the snow is uniformly distributed in the air with density ρ_0, say, and it falls with constant downward velocity $-v_0$ with $v_0 > 0$.
- Within the layer the snow has vanishing velocity and is uniformly distributed with density $\rho_1 > \rho_0$.

We take the z-axis vertically, with the bottom at $z = 0$, and denote the layer thickness by $h(t)$. Further, $\rho(z)$, $v(z)$ are the density and velocity of the snow at height z, respectively. The system shows discontinuous changes in density and velocity at height $z = h(t)$. This is typical for a shock. The evolution of the layer thickness $h(t)$ follows from conservation of mass across the shock front. Consider the global conservation equation

$$\frac{d}{dt}\int_a^b \rho\, dz = -\,[Q]_a^b\,,$$

where a and b are points below and above the front, respectively, and $Q := \rho\, v$ is the snow flux. Splitting up the integral, we obtain for the left-hand side

$$\frac{d}{dt}\left\{\int_a^{h(t)} \rho\, dx + \int_{h(t)}^{b} \rho\, dx\right\} = (\rho_1 - \rho_0)\frac{dh}{dt}\ .$$

In this formulation the discontinuous change in ρ causes no problem. We find

$$\frac{dh}{dt} = -\frac{Q_b - Q_a}{\rho_1 - \rho_0}\ . \tag{2.7}$$

Note that as an integration interval, any interval around the shock front can be taken. Taking a very narrow interval we obtain the general formula

$$\frac{dh}{dt} = -\frac{[Q]}{[\rho]}, \tag{2.8}$$

where $[Q]$ and $[\rho]$ are the jumps in the flux and the mass density over the shock front, respectively. This formula is generally applicable. We thus conclude that in these cases we meet with a differential quotient evaluated across the front. Applying this formula to the snow layer for which $Q_a = 0,\ Q_b = -\rho_0 v_0$, we obtain that the evolution of the snow front is given by

$$h(t) = \frac{\rho_0 v_0}{\rho_1 - \rho_0}\, t. \tag{2.9}$$

So, the layer thickness grows linearly in time, as expected. □

In the next exercise we consider a similar front taking the fall velocities velocity dependent.

Exercise 2.2b. *Sedimentation.*

We study the sedimentation of identical particles suspended in a fluid. Due to gravity they tend to move downward, but their speed will depend on the local situation. Let $c(z,t)$ denote the concentration and $v(z,t)$ the fall velocity in the fluid, due to gravity. See also Fig. **2.4**.

a. *Derive the law of mass conservation for this system in a z-range not containing a shock.*

b. *When the concentration of particles increases, the fall velocity may reduce. A simple model for this effect is*

$$v = v_0(1 - \alpha\, c)$$

for some constants v_0 and α. For which value c_{max} of the concentration is the sediment maximally packed? Apply the conservation law derived in a *to this case and find the governing equation for the concentration. Note that this sedimentation model is very similar to the traffic model in Example* 2.3a *in the next section.*

c. *In a layer at the bottom we have that $c = c_{max}$. Let us suppose that initially a layer is not yet present and that the concentration is uniform: $c(z,t) = c_0 < c_{max}\ \forall z > 0$. Argue that the concentration profile above the layer that will develop if time evolves remains uniform and that the growth rate of the layer thickness is given by an equation similar to* (2.9).

d. *For a nonuniform initial profile the situation is not that simple. The concentration and thus the velocity above the layer must then be calculated by solving the conservation equation derived under* b. *Apart from an in initial profile, this requires specification of a boundary condition. Since the position of the top of the layer is not fixed, one could prescribe the concentration quite far above the bottom. A possible choice is* $c(H,t) = \beta$ *for all times, with* H *big and* β *very small or vanishing. Let us take an initial profile that includes an initial layer of thickness* h_0*:* $c(z,0) = c_{max}$ *if* $z \le h_0$ *and* $c(z,0) = c_0(z)$ *if* $z > h_0$*. The function* $c_0(z)$ *must satisfy the boundary condition, so* $c_0(H) = \beta$*. It also holds that* $c_0(z) < c_{max},\ z > h_0$.

To calculate the position $h(t)$ *of the front (i.e., the thickness of the layer) as a function of time, one has to solve the conservation equation derived in* b *and* (2.8) *simultaneously. Could you outline a numerical scheme to perform this? For possible numerical schemes to solve partial differential equations (PDEs), see, e.g.,* [23].

2.3 Constitutive relations in one dimension

Basic laws and conservation relations alone do not provide a complete mathematical model. They have to be supplemented with constitutive relations. Such a relation characterizes the system under consideration. Its formulation may require a lot of expertise of the modeler in the specific field. If all aspects of a real-life system were taken into account, its constitutive relation probably would become too complicated to handle. Fortunately, one is always interested in only certain aspects, and these may guide the modeler's choice. The choice of a constitutive relation is not unique. Since some aspects are ignored for the time being, the modeler should always be aware that possible shortcomings of the model, which come about when model predictions and measured data are compared, might stem from the approximations made at this stage.

To illustrate these concepts, we come back to the car flow problem. There, a constitutive relation is an equation that relates the mass density ρ to the velocity v and so effectively to the flux density $Q = \rho v$. Two ways to model this relation are discussed in the following examples.

Example 2.3a. Car flow for nonanticipating drivers.
Let us assume that a driver will react only to the intensity of the traffic in his or her near environment. Then, the velocity will depend on only the local density. A simple model for v is a linear dependence on ρ:

$$v = \beta\,(\rho_m - \rho). \tag{2.10}$$

The parameters β and ρ_m in general are not known in advance but must be found from fitting the equation to measured data. Here, they are simple to estimate, since they have a clear interpretation. Relation (2.10) implies that the velocity vanishes if the density reaches its maximum value ρ_m. Then the cars occupy all available space on the road, as in a traffic jam. Furthermore, $\beta\,\rho_m$ is the maximal car velocity that can be achieved on an empty road, where $\rho = 0$. Substituting $Q(\rho) = \beta\,\rho\,(\rho_m - \rho)$ in the local conservation law (2.4), we find

$$\partial_t\,\rho + V(\rho)\,\partial_x\,\rho = 0. \tag{2.11}$$

The so-called density velocity V is defined by

$$V(\rho) = \frac{d\,Q}{d\,\rho} = \beta\,(\rho_m - 2\,\rho) \tag{2.12}$$

and is sketched in Fig. **2.5**. It is interesting to observe that the density velocity $V(\rho)$ is always smaller than the car velocity. The car velocity v is positive for all ρ, whereas the density velocity V is negative for $\rho > \frac{1}{2}\,\rho_m$ and positive for $\rho < \frac{1}{2}\rho_m$. The case $V < 0$ corresponds to heavy traffic and the case $V > 0$ to light traffic. □

Exercise 2.3a.

a. *Show that any car density that is uniform in space, i.e., independent of position, is also constant in time. Give the interpretation of this observation.*

b. *Assume that the car velocity is equal to a constant value v_0 and thus independent of both t and x. Let the car density be observed at $t = 0$ and given by a function r:*

$$\rho(x, 0) = r(x).$$

Determine the density at later times, and give the interpretation of the result (e.g., on an infinitely long road).

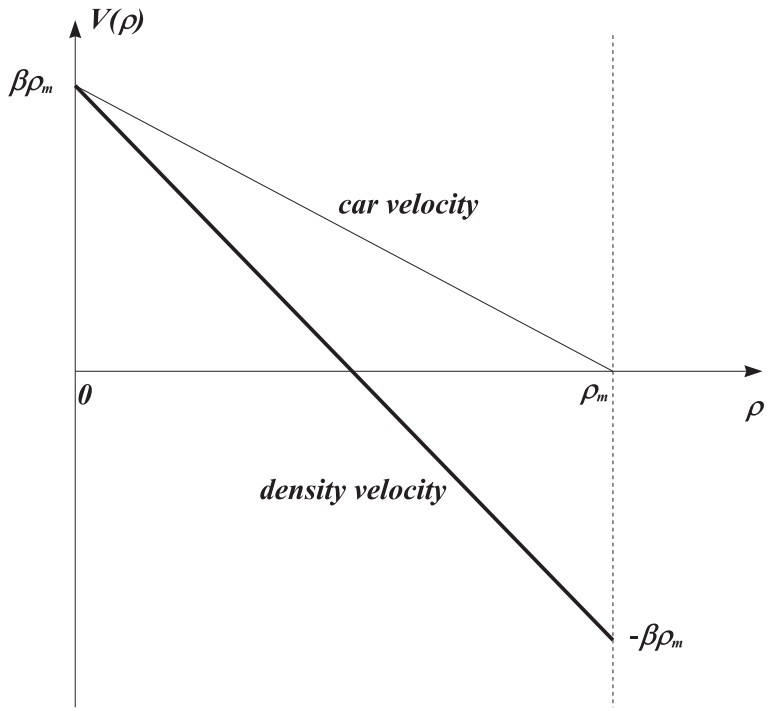

Figure 2.5. Car and density velocities, defined in (2.12), for the highway traffic model.

c. *Now investigate a situation in which the density is equal to a constant ρ_0 (uniform in space and constant in time) but with a small perturbation η that may depend on space and time:*

$$\rho(x,t) = \rho_0 + \eta(x,t).$$

For ease of interpretation we assume that (initially) the perturbation is a "localized hump," a small increase of car density on part of the road. Then the equation for η is approximately the linearized equation (ignoring higher-order terms in η)

$$\partial_t \eta + V_0 \partial_x \eta = 0, \tag{2.13}$$

where $V_0 := V(\rho_0)$ is constant and with V as defined in (2.12)*. This is called the* translation equation*. Show that if $\eta_0(x) := \eta(x,0)$ is the initial perturbation at $t = 0$, the density at later times is given by the translated profile:*

$$\eta(x,t) = \eta_0(x - V_0\, t).$$

The translation is to the right or to the left, depending on the sign of V_0. Explain this result.

Example 2.3b. Car flow for anticipating drivers.
In Example 2.3a, the car velocity was supposed to depend only on the local density. It is known, however, that most drivers (fortunately) not only take the local density into consideration but also adapt their speed when they observe density *variations* ahead of them: When the density ahead is larger (resp., smaller) they reduce (resp., increase) speed. A simple model to take this effect into account is to let the car velocity also depend on the spatial derivative $\partial_x \rho$. So, to include this aspect, the car velocity could be expressed as

$$v(x,t) = \beta\,(\rho_m - \rho) - \alpha\,\partial_x \rho(x,t) \tag{2.14}$$

with α and β constants. □

Exercise 2.3b.

a. *Discuss why the coefficient α in* (2.14) *should be taken as positive to model the effect of anticipating driving. Argue that the constants in constitutive relation* (2.14) *must be chosen with some care, since otherwise cars may move backward if there is heavy traffic ahead of them.*

b. *Show that the balance equation is now given by*

$$\partial_t \rho + V(\rho)\,\partial_x \rho = \tfrac{1}{2}\alpha\,\partial_x^2 \rho^2 \tag{2.15}$$

with $V(\rho)$ as in (2.12)*. In* §4.6.1 *this equation is analyzed.*

Example 2.3c. Porous media: Water table.
A porous medium can consist of densely packed grains with fluid between the grains that fills the pores. Instead of describing flow through such a medium on a granular scale, one usually considers a macroscopic description. As a specific example, consider a (shallow)

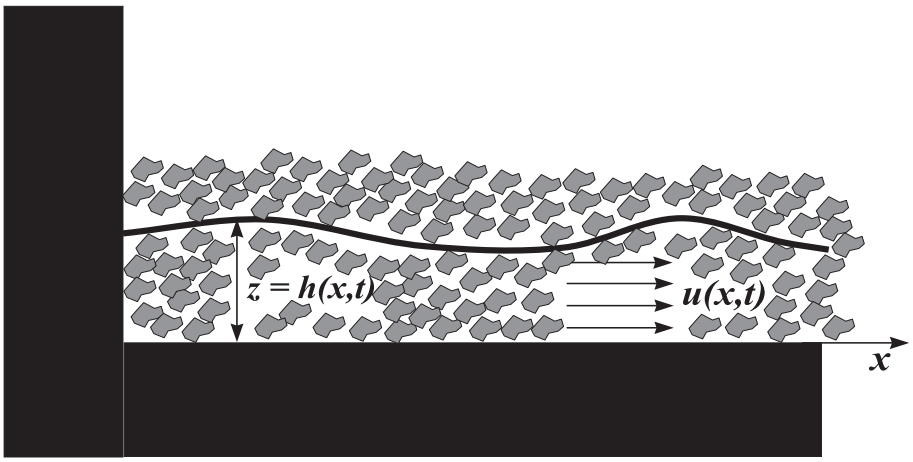

Figure 2.6. Flow through a porous medium which is isolated at one end.

aquifer (one horizontal direction, the x-axis) that is saturated with water above a horizontal bottom at $z = 0$. The water table is positioned at $z = h(x, t)$. See Fig. **2.6**. Let $u(x, t)$ denote the horizontal fluid velocity; since we assume the aquifer to be shallow, it is assumed that this velocity is independent of the height z. A standard constitutive relation is given by

$$u = -\kappa\, \partial_x\, h(x, t),$$

where (in the simplest cases) κ is a constant. This is a special case of the so-called Darcy law, which states that the local velocity is proportional to the gradient of the local pressure. □

Exercise 2.3c.

a. *Derive for this system the general equation expressing mass balance.*

b. *Substitute Darcy's law and obtain an equation for h above.*

c. *Assume that the aquifer is enclosed at the left endpoint at $x = a$ by a solid, impermeable rock, and at the right endpoint $x = b$ is connected to a lake. Derive the governing boundary conditions at both endpoints.*

Example 2.3d. Fourier's law for heat conduction.
A constitutive relation for heat conduction couples the heat flux Q and the temperature u. This depends very much on the material properties of the solid. From everyday experience we know that heat flows from places of high temperature to places of lower temperature. This daily-life principle is formalized in thermodynamics and is referred to as the *second law of thermodynamics*. These observations inspired Fourier to introduce as an empirical relation

$$Q = -\,D\, \partial_x u, \tag{2.16}$$

where D is the *heat conductivity*. In Fig. **2.3** Q is sketched for a Gaussian-like u-profile. The minus sign is included to make D positive. This coefficient may depend on time, position, and/or temperature, but in the simplest cases it is just a constant. This *Fourier's law* is much more general than the present application suggests. In fact, all kinds of diffusion processes have the same characteristic: there is a flow of some substance from positions with high concentrations to positions with low concentrations of that substance. Therefore, the resulting balance equation below also describes many diffusion processes.

Inserting Fourier's law in the balance law, the so-called heat or diffusion equation results:

$$\partial_t(\rho c u) = \partial_x\,(D\,\partial_x u) + S. \tag{2.17}$$

In the special case that ρ, c, and D are constant, the equation reduces to

$$\partial_t\, u = \kappa\,\partial_x^2 u + s \tag{2.18}$$

with $\kappa := D/\rho\, c$ the *thermal conductivity* and $s := S/\rho\, c$. In the absence of sources this equation reads as

$$\partial_t\, u = \kappa\,\partial_x^2 u. \tag{2.19}$$

See also §2.7 for its extension to more dimensions. The diffusion equation should be completed by specifying the boundary conditions, which describe the flow behavior at the endpoints of the rod, where the material is in contact with the surroundings. Frequently met boundary conditions are as follows:

- **Prescribed heat flux**. In this case, heat (and thus energy) is added or extracted via the boundary. A special case is that of an *insulated endpoint*. Then, $Q = 0$ at that boundary, and Fourier's law implies that, locally, it holds that

  $$\partial_x u = 0.$$

  So, the u-profile becomes flat when it approaches the boundary. Another often-applied condition is to take Q proportional to the temperature difference between rod and surroundings. This leads to *Newton's law of cooling*

  $$D\,\partial_x u = -k\,(u - u_0)$$

  with u_0 the temperature of the surroundings and $k > 0$ some coefficient measuring the speed of this process.

- **Prescribed temperature**. Experimentally this is arranged by keeping the endpoint in contact with a medium of which the heat capacity is practically infinite. Then the temperature of that medium is constant, no matter how much energy is exchanged. So then we have

  $$u = u_0$$

  at the boundary. □

Exercise 2.3d.

Consider a rod of length L of which both endpoints are kept at a fixed temperature u_0. Assume that the conduction is described by (2.19) *with uniform thermal conductivity κ.*

a. *Suppose that initially the temperature distribution is given by*

$$u(x,0) = u_0 + A\sin(m\pi x/L)$$

for some integer m and amplitude A. Find the temperature distribution for increasing time by searching for a solution of the form

$$u(x,t) = u_0 + f(t)\sin(m\pi x/L).$$

What is the effect of the value of m? Can you explain why the temperature approaches the equilibrium solution faster for larger values of m? To answer this question investigate the value of the flux.

b. *Now, we consider a rod of which the left endpoint is kept at a fixed temperature u_0 and the right endpoint is insulated. A constant local heat production s is present along the whole rod. See Fig.* **2.7**. *At some moment the temperature distribution inside the rod will become time independent, in a state independent of the initial temperature profile. Determine the stationary temperature distribution. Check that in the stationary situation the heat flux through the left endpoint is equal to the total heat supply. In the stationary situation, the left endpoint thus serves as a heating (or cooling) device for the surroundings.*

c. *Next, consider the same system as under* b *but now for a rod that consists of two materials. For $0 \le x \le L/2$ it has thermal conductivity κ_1, and for $L/2 \le x \le L$ it has thermal conductivity κ_2. Determine the stationary temperature gradient in the rod and show that its derivative is not continuous across $x = L/2$.*

Exercise 2.3dd.

Consider again heat conduction in a rod, described by (2.19) *with constant thermal conductivity κ.*

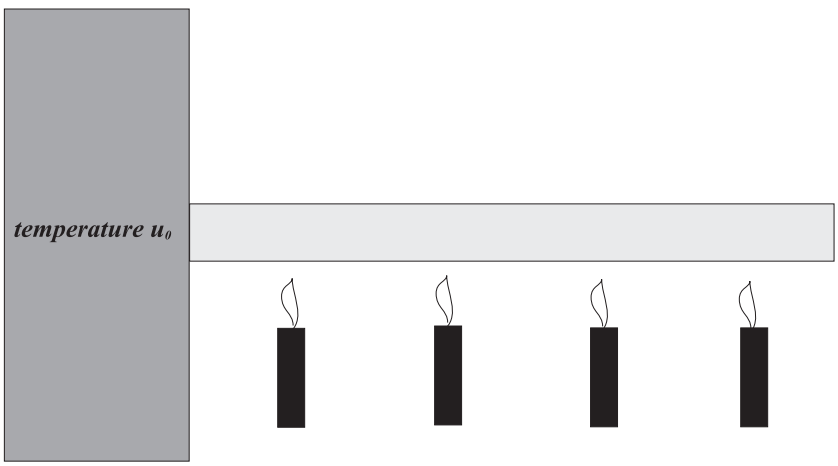

Figure 2.7. Heat-conducting rod, heated from outside, with one endpoint attached to a heat reservoir at constant temperature and with the other endpoint isolated.

a. *Show that if u satisfies this heat equation, u^2 satisfies*

$$\partial_t(u^2) = 2\kappa\left[\partial_x(u\,\partial_x u) - (\partial_x u)^2\right].$$

Compare this expression to (2.18)*. Which flux and which source term would you associate with u^2?*

b. *Consider* (2.19) *for a rod of length L with fixed temperatures at the endpoints:*

$$u(0,t) = u(L,t) = u_0 \quad \forall t.$$

Show that the norm

$$N_1(t) = \int_0^L (u(x,t) - u_0)^2\,dx$$

is monotonically decreasing in time for any solution that is nonuniform in x. Conclude from this that the temperature u decays to a uniform density if $t \to \infty$.

c. *Now consider the case that the rod is insulated at the endpoints:*

$$u_x(0,t) = u_x(L,t) = 0 \quad \forall t.$$

Show that the norm

$$N_2(t) = \int_0^L u^2(x,t)\,dx$$

is monotonically decreasing in time for any solution that is nonuniform in x. Again, conclude from this that the temperature u decays to a uniform density if $t \to \infty$.

d. *If for the system under* c *the initial density $u_0(x)$ is prescribed, what will be the final state in terms of u_0?*

Example 2.3e. Viscous and elastic behavior.
Above we have met with constitutive relations that coupled flux and mass densities. Here, we deal with constitutive relations in which forces play a role. The internal forces, i.e., the stresses, depend on the local deformation, the *strain*, and/or on the local velocities. In case of a dependence on the strain only, as is found in deformable bodies, such a system is said to respond *elastically*. If the local stress depends on the local velocity only, as is the case in simple fluids, such behavior is called *viscous*.

The preeminent example of elastic behavior is the spring, as sketched in Fig. **1.1**. For small deviations from the equilibrium length, the spring force is linearly proportional to the displacement. The external force $F^{elastic}$, needed to keep the spring stretched, satisfies Hooke's law

$$F^{elastic} = +kx \tag{2.20}$$

with x the displacement and k the spring constant measuring the spring stiffness.

Viscous behavior is found in fluids and gases. If an object moves through such a medium, it is slowed down due to frictional forces. These forces are proportional to the

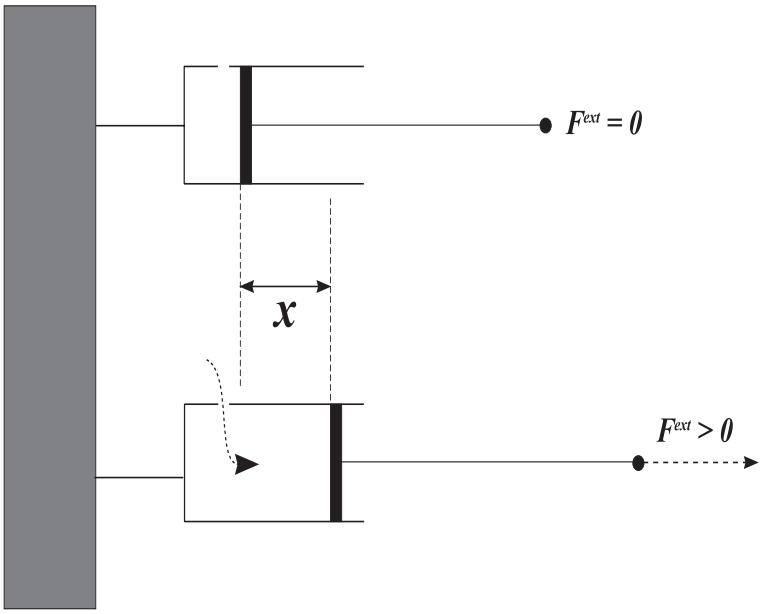

Figure 2.8. The dashpot as prototype of viscous behavior in one dimension.

velocity of the object. Such a behavior is appropriately modeled by a dashpot, as sketched in Fig. **2.8**. The external force $F^{viscous}$, needed to keep the dashpot moving, is proportional to the speed of the piston:

$$F^{viscous} = \eta\, v := \eta\, \partial_t x \tag{2.21}$$

with the coefficient η called the *viscosity* and v the piston speed.

If both local deformation and local velocity determine the local stress, as is the case for fluids with a complex internal structure, such a system is said to show *viscoelastic behavior*. Most materials behave in such a mixed way. We shall deal with two models that describe this behavior, the *Maxwell* model and the *Kelvin* model. Both have their own merits, and which describes the dynamics best depends on the system. □

Viscoelastic behavior: Maxwell model

This model describes the viscoelastic behavior of a one-dimensional system, say, a rod, via a spring and a dashpot in series, as sketched in Fig. **2.9**. The external force $F(t)$ is equally felt by both the spring and the dashpot. Let us denote the displacements of spring and dashpot by x_1 and x_2, respectively. The spring will respond with a displacement $x_1 = F/k$, according to (2.20). The dashpot will start to move at speed $\dot{x}_2 = F/\eta$, according to (2.21). The total displacement $x = x_1 + x_2$ of the series will vary with velocity

$$v(t) = \dot{x}_1 + \dot{x}_2 = \frac{1}{k}\,\partial_t F + \frac{1}{\eta}\,F. \tag{2.22}$$

With the definition of *relaxation time* as $\lambda := \eta/k$, the Maxwell model reads as

$$F + \lambda\,\partial_t F = \eta\, v. \tag{2.23}$$

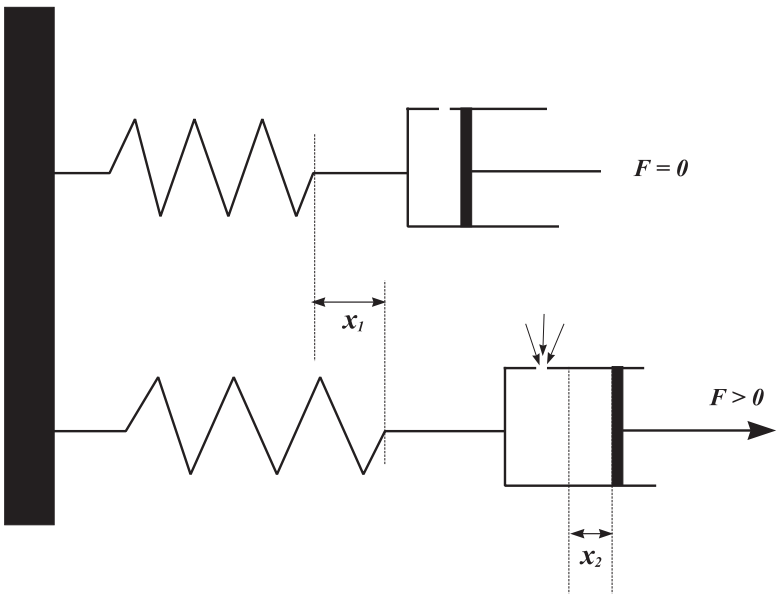

Figure 2.9. In one dimension the Maxwell model assumes the response of a rod to an external force as being described by the behavior of a spring and a dashpot in series.

So, if F is given, we can directly calculate v, and if v is given, we find F from solving a first-order ordinary differential equation (ODE).

Exercise 2.3e.

Let us study the case in which not the external force $F(t)$ exerted to the series is prescribed but the initial displacement x_0 and the velocity $v(t)$ of the endpoint. Equation (2.23) *is then read as a differential equation for F:*

$$\partial_t F = -\frac{1}{\lambda} F + k\, v. \tag{2.24}$$

a. *Check that for $t > 0$ the stress is given by*

$$F(t) = e^{-\frac{t}{\lambda}}\, F_0 + k \int_0^t e^{-\frac{t-t'}{\lambda}}\, v(t')\, dt'.$$

b. *Interpret this expression in terms of memory of the system.*

c. *At $t = 0$ we start with a displacement x_0 from equilibrium and fix the system in this position. Thus, the velocity is vanishing all the time. Determine the initial stress F_0 in the system and also $F(t)$, its response in time.*

d. *Next, we start in the equilibrium position and bring the system in oscillation by enforcing that the velocity is given by $v(t) = \sin(t)$. Find the solution by evaluating*

the integral under a. *Determine the phase difference which, after some transient time, is found between the force and the amplitude.*

Viscoelastic behavior: Kelvin model

This model describes one-dimensional viscoelastic behavior via a spring and a dashpot in parallel, as sketched in Fig. **2.10**. Now the spring and the dashpot have the same displacement x. The stress in the spring is, according to (2.20), given by $F^{elastic} = kx$, and the stress in the dashpot satisfies $F^{viscous} = \eta\dot{x}$, according to (2.21). The total force is then given by $F = F^{elastic} + F^{viscous}$ and satisfies

$$F = k\,x + \eta\,\partial_t x \tag{2.25}$$

or

$$x + \lambda\,\partial_t x = \frac{1}{k}F. \tag{2.26}$$

So, given $x(t)$ we can directly solve $F(t)$, and if $F(t)$ is prescribed, we find $x(t)$ from solving the differential equation. □

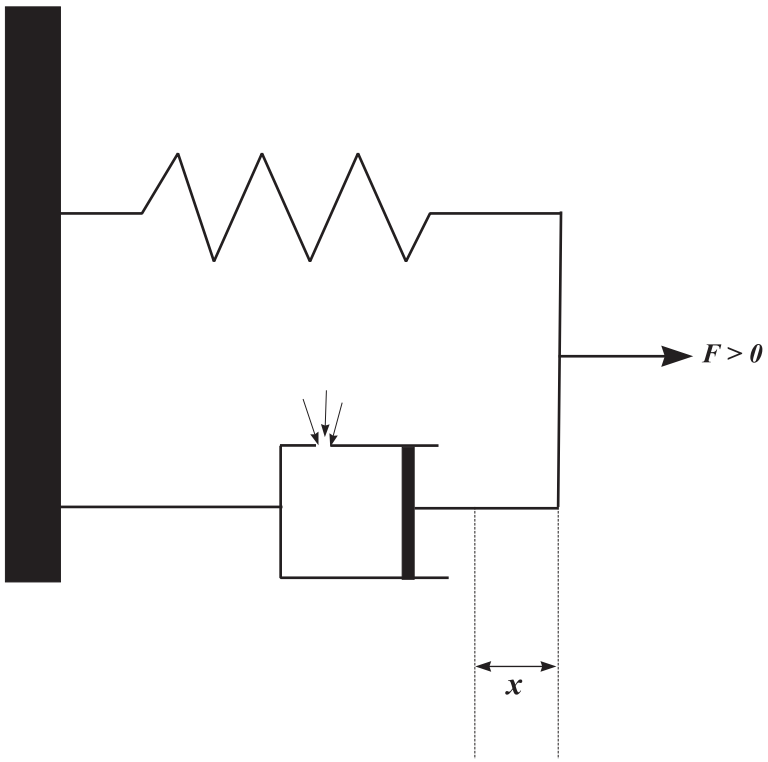

Figure 2.10. In one dimension the Kelvin model assumes the response of a rod to an external force as described by the behavior of a spring and a dashpot in parallel.

Exercise 2.3ee.

In the Kelvin model let the external force $F(t)$ be prescribed. Equation (2.25) *can then be solved for $x(t)$ if the initial position x_0 is given.*

a. *Check that if we apply a nonvanishing external force, the position of the system is given by*

$$x(t) = e^{-\frac{t}{\lambda}} x_0 + k \int_0^t e^{-\left(\frac{t-t'}{\lambda}\right)} F(t')\, dt'.$$

b. *Interpret this expression in terms of memory of the system.*

c. *Calculate the behavior of the system if we start at $x_0 > 0$, and if no external force is applied, so that $F(t) = 0$. What is the final position of the system?*

d *Next, we start in the equilibrium position and bring the system in oscillation by driving it with a force $F(t) = \sin(t)$. Evaluate the integral under* a. *Determine the phase difference which, after some transient time, is found between the force and the amplitude.*

Example 2.3f. Maxwell model in continuous media.
Let us apply the Maxwell model to a continuous medium that behaves in a viscoelastic way. Here, we restrict ourselves to the one-dimensional case, so the result applies to, e.g., a flexible rod or thread. It is somewhat unnatural to do this analysis in one dimension, but the present formulation is highly appropriate to be generalized to more dimensions, as will be done in §2.8. So, we consider a long flexible thread extending along the x-axis. At position x and time t the thread moves at speed $v(x,t)$. Local stretching of the thread leads to a local stress $\sigma(x,t)$. The constitutive relation describes how the stress is related to the deformations or to the velocities. In the Maxwell model, σ and v are related. To find this relation, consider a filament of length δx, positioned along the interval $[x, x+\delta x]$. See Fig. **2.11**. If the velocity difference $v(x+\delta x,t) - v(x,t)$ between the endpoint velocities is not vanishing, the filament will be stretched (or may shrink), and this

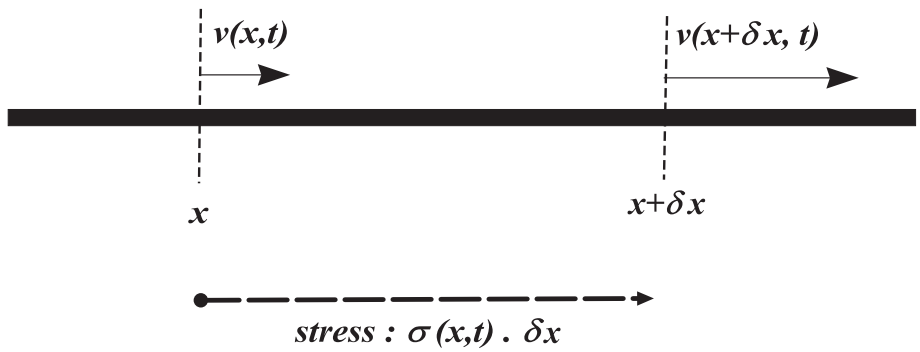

Figure 2.11. A filament of length δx will vary its length if the velocities at the endpoints are not the same.

influences its internal stress. We define this stress σ by $F = \sigma\,\delta x$, where F is the force needed to keep the filament in its stretched shape. The stress is thus defined per unit of length. Next, we apply (2.22), taking into account that now both endpoints may move. This leads to

$$v(x+\delta x,t) - v(x,t) = +\frac{1}{k}\,\partial_t\sigma\,\delta x + \frac{1}{\eta}\,\sigma\,\delta x. \tag{2.27}$$

Dividing by δx and taking the limit $\delta x \to 0$ we arrive at the continuous Maxwell model in one dimension:

$$\sigma + \lambda\,\partial_t\sigma = \eta\,\partial_x\,v. \tag{2.28}$$

The three-dimensional equivalent is given in Example 2.7c. □

Exercise 2.3f.

Check that the solution of the Maxwell model (2.28) *reads, in integral form,*

$$\sigma(x,t) = \frac{\eta}{\lambda}\int_{-\infty}^{t} e^{-\left(\frac{t-t'}{\lambda}\right)}\partial_x v(x,t')\,dt'$$

if at $t = -\infty$ *the material was stressless. Interpret this in terms of memory of the system.*

2.4 Transport theorem

We have shown how balance laws can be derived in one dimension for scalar quantities. We started our presentation with these examples for illustrative purposes, since in one dimension the concepts are quite clear. Here, we show how all conservation laws in flowing media can be derived from one general principle, the so-called transport theorem.

Let us consider a property with density $f(\mathbf{x},t)$. One may think of f as a scalar such as mass, heat, or energy, but it may also be identified with a vector quantity like momentum. For the moment we do not specify it. In fact, the power of the transport theorem is that it is so generally applicable. Property f is transported through the system, and the corresponding flux density is denoted by $\mathbf{Q}(\mathbf{x},t)$. Property f may also be produced and/or annihilated. The rate at which this happens is given by the source density $S(\mathbf{x},t)$. See Fig. **2.12**. Let us take an arbitrary volume V in the medium. We follow the so-called Eulerian approach, in which the control volume does not move with the flow but has a fixed position with respect to an external coordinate frame. The amount of f in V, $F(t)$, say, is given by the integral

$$F(t) = \int_V f(\mathbf{x},t)\,dV.$$

Conservation of the property under consideration implies that, per unit of time, the change in F equals the flux through the surface A of V plus the production of f in V. On the one hand, we have

$$\frac{dF}{dt} = \int_V \frac{\partial f}{\partial t}\,dV,$$

since the volume V is fixed in time. On the other hand, this change must be equal to

$$\frac{dF}{dt} = -\int_A \mathbf{Q}\cdot\mathbf{n}\,dA + \int_V S\,dV.$$

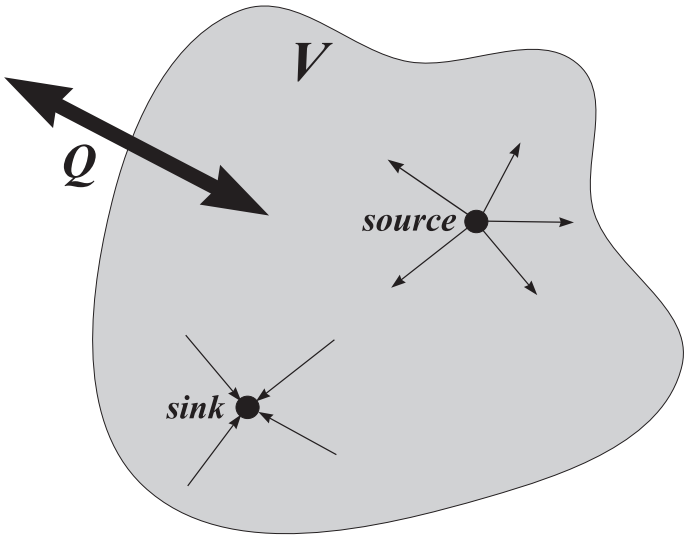

Figure 2.12. Fluxes, sources, and sinks in a fixed volume V.

Here we use the $\cdot$ (center dot) notation for the usual inner product of vectors. The first term on the right-hand side represents the total outward flux through the surface A with $\mathbf{n}$ the outward normal to A with unit length. The second term is the source term. We now invoke the Gauss theorem, which states that for a vector density such as $\mathbf{Q}$ we have that

$$\int_A \mathbf{Q} \cdot \mathbf{n}\, dA = \int_V \nabla \cdot \mathbf{Q}\, dV. \tag{2.29}$$

The so-called nabla operator ∇ is defined as

$$\nabla := (\partial_x, \partial_y, \partial_z)$$

with respect to Cartesian coordinates (x, y, z). The divergence of some vector field is easily expressed in terms of the nabla operator. For example,

$$\operatorname{div}(\mathbf{Q}) := \nabla \cdot \mathbf{Q} = \partial_x Q_1 + \partial_y Q_2 + \partial_z Q_3.$$

This allows us to rewrite the surface integral as a volume integral. Collecting all terms together, we conclude that conservation of f in V at any time t is expressed by the global or integral form of the *transport theorem*:

$$\int_V \{\partial_t f + \nabla \cdot \mathbf{Q} - S\}\, dV = 0. \tag{2.30}$$

The transport theorem relates f to the flux $\mathbf{Q}$ and the source S. If the integrand is continuous, we may apply the lemma of Lagrange (see Exercise 2.2a), which yields the local or differential form

$$\partial_t f + \nabla \cdot \mathbf{Q} = S. \tag{2.31}$$

In the following sections we apply the transport theorem to mass, heat, momentum, and energy, respectively.

2.5 Mass balance in three dimensions

If we interpret the property f as mass with density $\rho(\mathbf{x}, t)$, the flux is given by $\mathbf{Q} = \rho\,\mathbf{v}$. Substitution into (2.31) leads to the *continuity equation*

$$\partial_t \rho + \nabla \cdot (\rho\,\mathbf{v}) = S \tag{2.32}$$

with S the mass source density. Using the identity

$$\nabla \cdot (\rho\,\mathbf{v}) = \rho\,\nabla \cdot \mathbf{v} + \mathbf{v} \cdot \nabla\,\rho,$$

we may rewrite this as

$$\partial_t\,\rho + \mathbf{v} \cdot \nabla\,\rho + \rho\,\nabla \cdot \mathbf{v} = S.$$

The first two terms can be taken together by introducing the notion of *material time derivative*, also called *total time derivative*:

$$D_t := \partial_t + \mathbf{v} \cdot \nabla. \tag{2.33}$$

This time derivative measures the change in the property under consideration while the observer moves with the flow and thus follows a *streamline*, i.e., the path of a material point in the flow. The operator $\mathbf{v} \cdot \nabla$ accounts for the movement of the observer. So, an alternative form for (2.32) is

$$D_t \rho + \rho \nabla \cdot \mathbf{v} = S. \tag{2.34}$$

A special case is that of *incompressible* flow. In this case, the amount of mass in an arbitrary volume is conserved when the volume is transported with the flow, although the volume may change its shape when flowing. In an incompressible flow we have that $D_t \rho = 0$. In the absence of sinks and sources, the continuity equation then reduces to the *incompressibility condition*

$$\nabla \cdot \mathbf{v} = 0. \tag{2.35}$$

Example 2.5a. Pipe line transport.
Consider a pipe through which batches of immiscible fluids, e.g., oil and water, are pumped. See Fig. **2.13**. Since the batches are incompressible, they all travel at the same speed. Let the batches have densities $\rho_1, \rho_2, \ldots$, etc. The mass density $\rho(x, t)$ along the pipe is then

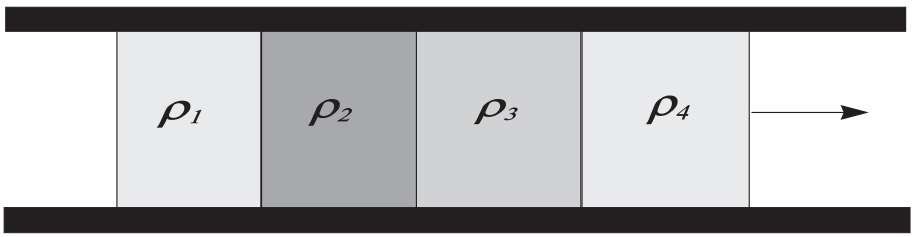

Figure 2.13. Pipe line transport of immiscible fluid batches.

constant in each batch and jumps across batch interfaces. For fixed position x, the derivative $\partial_t \rho$ vanishes as long as a batch passes x but sharply peaks when an interface passes. For an observer traveling at the same speed as the batches, ρ does not change. This is expressed by $D_t \rho = 0$. □

Exercise 2.5a.
Check that for this pipe transport condition (2.35) *holds.*

2.6 Heat balance in three dimensions

The derivation of a balance equation for heat is very similar to the derivation above for mass. As in Example 2.2a, f is now identified with the heat density $\rho\, c\, u$. The local heat balance then reads as

$$\partial_t (\rho\, c\, u) + \nabla \cdot \mathbf{Q} = S.$$

This is the generalization to more dimensions of (2.6). *Fourier's law*, which relates the heat flux $\mathbf{Q}$ to the gradient of the temperature, takes in more dimensions the form

$$\mathbf{Q} = -\, D\, \nabla u, \tag{2.36}$$

where D is the (scalar) *heat conductivity coefficient*. Insertion of this model for the flux in the local balance law leads to

$$\partial_t (\rho\, c\, u) - \nabla \cdot (D\, \nabla u) = S. \tag{2.37}$$

In the special case that ρ, c, and D are constant, the equation becomes

$$\partial_t u = \kappa\, \Delta u + s, \tag{2.38}$$

where $\kappa := D\,/\,(\rho\, c)$ is the thermal conductivity, $s := S\,/\,(\rho\, c)$, and Δ is the *Laplace operator*, which in Cartesian coordinates is given by

$$\Delta := \nabla^2 = \nabla \cdot \nabla = \partial_x^2 + \partial_y^2 + \partial_z^2\,.$$

2.7 Momentum

To apply the transport theorem to momentum, we first have to introduce the concept of stress in continuous media. Stress can be considered as a generalization of the concept of pressure. Pressure is a scalar quantity and represents the force exerted on the mass in a point due to the presence of surrounding mass. In isotropic media, this force is the same in all directions, depending on position only. However, in general this force does depend on the direction and thus cannot be represented by a scalar density. Instead, the stress density assigns a tensor to each point. Its precise definition is given in the following subsection.

2.7.1 Stress in continuous media

The definition of the stress tensor in a point **x** of a medium is illustrated in Fig. **2.14**. In **x** we position an imaginary surface element of area A. The details of the shape of this surface element are not relevant; one could think of a disk or a square. The area A should be small compared to the macroscopic dimensions of the medium but large compared to the molecular dimensions. The position in space of the surface element is determined by its normal **n**, which is assumed to have unit length. The material on the side of the surface element opposed to **n** exerts forces on the material at the same side as **n**. Compared to the macroscopic size of the flow as a whole, the microscopic forces have short ranges. Thus, only forces between molecules very close to the surface are relevant. Let us denote by $\mathbf{F_n}$ the sum of all these forces through the surface element. If A is small enough, $\mathbf{F_n}$ will be proportional to the area A. The *stress* $\boldsymbol{\sigma}_\mathbf{n}$ is defined as the force $\mathbf{F_n}$ normalized to a unit area, so

$$\boldsymbol{\sigma}_\mathbf{n} := \frac{\mathbf{F_n}}{A} .$$

The stress $\boldsymbol{\sigma}_\mathbf{n}$ can be decomposed into a component in the plane of the surface element and a component parallel to **n**. The first is called the *shear stress* and the second the *normal stress* in **x**. The shear stress tends to drag the surface element parallel to itself, whereas the normal stress tends to move it in the direction of its normal **n**.

If the direction of the normal **n** is varied, the stress vector $\boldsymbol{\sigma}_\mathbf{n}$ will vary. By definition, the linear mapping that relates the vector **n** to the vector $\boldsymbol{\sigma}_\mathbf{n}$ is the *stress tensor* $\boldsymbol{\sigma}$. In formula,

$$\boldsymbol{\sigma}_\mathbf{n} = \boldsymbol{\sigma} \cdot \mathbf{n}. \tag{2.39}$$

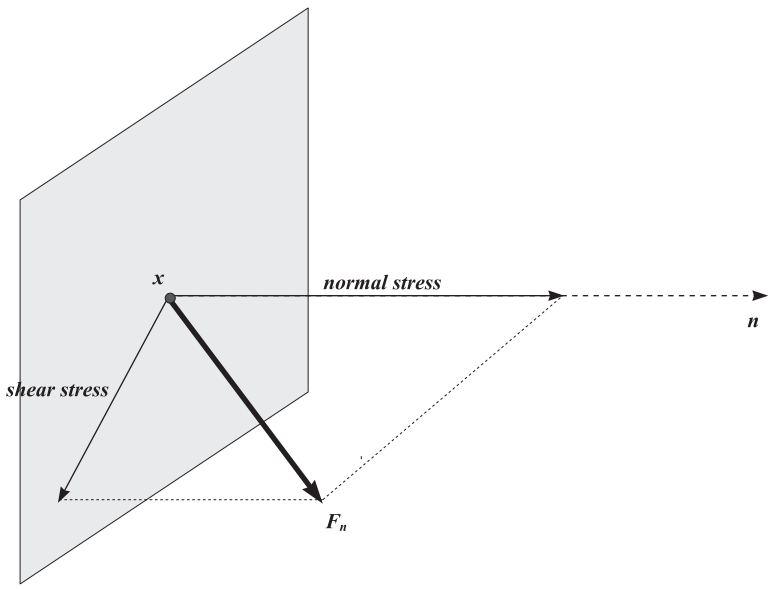

Figure 2.14. The stress felt by a surface element can be decomposed into the shear stress and the normal stress.

Here we use the · (center dot) notation for the usual matrix-vector product. Note that throughout this book the · notation is used for both the inner product of vectors and the matrix-vector product. This convention stems from tensor theory notation, where · stands for *contraction* of two tensors of any type.

If the stress tensor is known, the stress can be calculated in any direction. With respect to a fixed coordinate frame, σ is represented by a 3×3 matrix with elements depending on $(\mathbf{x}, t)$. The functional form of the matrix elements will vary from coordinate system to coordinate system.

Exercise 2.7a. *Hydrostatic pressure.*
In isotropic media such as gases and water, the stress tensor is simply given by

$$\boldsymbol{\sigma} = -p\,\mathbf{I}$$

with p the hydrostatic pressure and $\mathbf{I}$ the unit tensor.

a. *Can you interpret why the minus sign is introduced by convention?*

b. *How does p depend on position in the case of water at rest?*

c. *In Fig.* **2.15** *three objects A, B, and C are sketched, all with density much smaller than the density of water. Determine how the direction and magnitude of the stress varies over the sides of these objects. Realize that the total force exerted on an object is the integral of the stress over its surface. Conclude from this that object A will rise, object B will be pressed against the bottom, and cylinder C will stand still but that this stationary state is easily perturbed.*

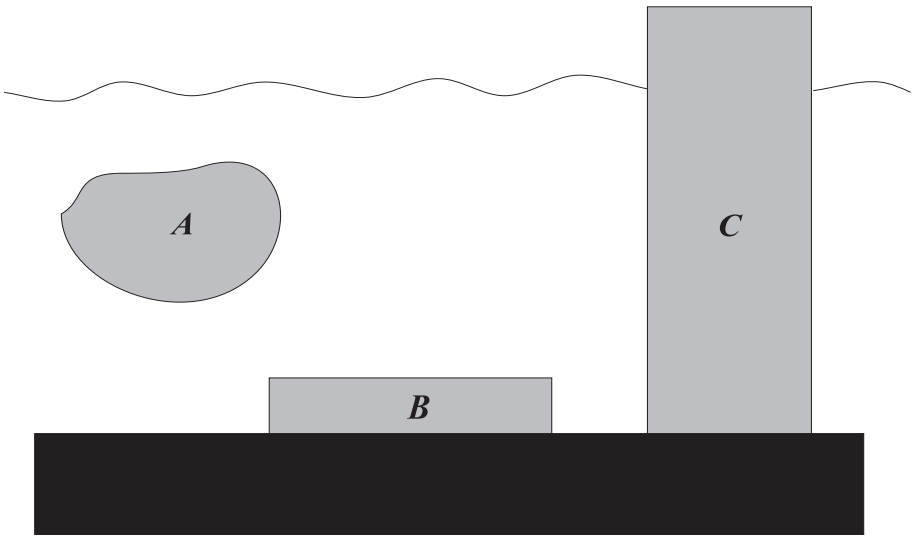

Figure 2.15. Three emerged bodies with mass densities smaller than the density of the fluid experience totally different buoyancy forces.

For nonisotropic materials like polymeric melts, the stress tensor is usually written in the form

$$\boldsymbol{\sigma} = -p\,\mathbf{I} + \boldsymbol{\sigma}^e \tag{2.40}$$

with the extra stress tensor $\boldsymbol{\sigma}^e$ measuring the deviation from an isotropic material.

2.7.2 Momentum balance

Here, we identify property f in (2.31) with momentum $\rho\,\mathbf{v}$. Its flux $\mathbf{Q}$ is the rate at which momentum is transported:

$$\mathbf{Q} := \rho\,\mathbf{v}\mathbf{v}.$$

Here, the so-called dyadic product $\mathbf{v}\,\mathbf{v}$ of the velocity vector $\mathbf{v}$ with itself is introduced. This stems from the fact that two directions are involved: momentum itself has a direction, and it is transported in a direction. The resulting flux is a tensor with two indices. If the components of $\mathbf{v}$ are indicated by v_i, $i = 1, 2, 3$, then the definition of the dyadic product is

$$(\mathbf{v}\,\mathbf{v})_{i,j} = v_i\,v_j. \tag{2.41}$$

Next, we want to apply the transport theorem (2.31). To that end we need to think about the interpretation of S in (2.31), the source/sink of momentum. If $\mathbf{p}(t)$, defined as the integral

$$\mathbf{p}(t) := \int_V \rho\,(\mathbf{x}, t)\,\mathbf{v}\,(\mathbf{x}, t)d\mathbf{x},$$

is the momentum of the flow within a control volume V, then Newton's second law states that the change of $\mathbf{p}$ in time equals the sum $\mathbf{F}_{total}$ of the forces exerted on the material within V. So,

$$\frac{d\,\mathbf{p}}{d\,t} = \mathbf{F}_{total}.$$

From this we conclude that S has to be identified with $\mathbf{F}_{total}$. It is the sum of two forces. First, so-called body forces apply. These represent the effect of external force fields on the material within V, due to, e.g., electromagnetic fields and gravitation. For example, in case of gravity the body force is given by $\rho\mathbf{g}$, with $\mathbf{g}$ the (constant) gravitational acceleration in the downward direction. Although gravity is always present, it may often be neglected, since the dynamics of highly viscous flows is dominated by the internal stresses. Body forces act on each point in V. We represent all body forces together by the density $\mathbf{f}_b(\mathbf{x}, t)$.

Second, the contact force $\mathbf{F}_c$ applies, which is due to the interaction with the surroundings of V. This contact force is nothing but the effect of the stress acting on the surface A of V. Its total contribution is given by an integral of the stress over the surface. As explained in §2.7.1, the stress in a point $\mathbf{x}$ on the surface is obtained through multiplying the stress tensor $\boldsymbol{\sigma}$ and the outward normal $\mathbf{n}$ at $\mathbf{x}$ to the surface. The total contact force is then given by the surface integral

$$\mathbf{F}_c(t) = \int_A \boldsymbol{\sigma}\,(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x})\,dA. \tag{2.42}$$

It is convenient to transform the surface integral into a volume integral by means of the Gauss divergence theorem:

$$\mathbf{F}_c(t) = \int_V \nabla \cdot \boldsymbol{\sigma}(\mathbf{x}, t)\, dV.$$

Application of the integral form (2.30) of the transport theorem then leads to

$$\int_V (\partial_t(\rho\,\mathbf{v}) + \nabla \cdot (\rho\,\mathbf{v}\,\mathbf{v})\, dV = \int_V (\nabla \cdot \boldsymbol{\sigma} + \mathbf{f}_b)\, dV.$$

Assuming the integrands to be continuous, we obtain the differential form

$$\partial_t(\rho\,\mathbf{v}) + \nabla \cdot (\rho\,\mathbf{v}\mathbf{v}) = \nabla \cdot \boldsymbol{\sigma} + \mathbf{f}_b. \tag{2.43}$$

Note that this vector expression stands for three equations. If the mass is conserved, we have $S = 0$ in the continuity equation (2.32). Combination of the momentum equation with the continuity equation (2.32) then yields

$$\rho\, D_t\mathbf{v} = \nabla \cdot \boldsymbol{\sigma} + \mathbf{f}_b. \tag{2.44}$$

If the system is at rest and we thus have $\mathbf{v} = 0$, this reduces to

$$\nabla \cdot \boldsymbol{\sigma} = -\,\mathbf{f}_b.$$

This simply expresses that in a free, deformable body the internal stresses must be balanced by external forces to keep the body in a fixed shape. Another way to establish a fixed shape is to apply contact forces to its surface. Then, the geometry is governed by the differential equations

$$\nabla \cdot \boldsymbol{\sigma} = 0,$$

and the contact forces are accounted for via boundary conditions.

2.7.3 Constitutive relations in three dimensions

To apply the momentum equation (2.44) to a specific flow, the character of the fluid has to be specified via a constitutive relation. In the present context this implies that the stress tensor has to be coupled to the strain, i.e., the deformations, and/or to the velocities in the flow. Here, we present some examples of such constitutive relations that are widely used.

Example 2.7a. Newtonian flow, Navier–Stokes equations.
Newton realized that many fluids show viscous behavior and that this implies that the local stress is proportional to the local velocity gradient. In Example 2.3f we introduced viscous and viscoelastic behaviors and showed how these are described in one dimension. For viscous fluids we have to generalize (2.28). In doing that, we must take into account that any stress tensor must be symmetric. This directly follows from the fact that an arbitrary subvolume in a medium at rest is also at rest and does not spontaneously start to rotate. The absence of rotational forces in this situation implies that the stress tensor is symmetric. The extra stress tensor, introduced in (2.40), reads for incompressible *Newtonian flow* as

$$\boldsymbol{\sigma}^e = 2\,\eta\,\mathbf{D}. \tag{2.45}$$

The factor of 2 is by convention, η is the viscosity, and the *rate of deformation tensor* $\mathbf{D}$ is defined as

$$\mathbf{D} := \tfrac{1}{2}(\nabla\mathbf{v} + (\nabla\mathbf{v})^T).$$

The upper index T denotes $transposed$, so $\mathbf{D}$ is symmetric by construction. This Newtonian constitutive relation applies for all fluids consisting of small, more or less spherical particles, for example, water. Substituting the Newtonian constitutive relation into the momentum equation (2.44), we get the famous *Navier–Stokes equation*,

$$\rho\, D_t\mathbf{v} = -\nabla p + 2\,\eta\,\nabla\cdot\mathbf{D} + \mathbf{f_b}, \tag{2.46}$$

which in combination with the continuity equation (2.32) describes numerous flow phenomena. Here, a relation between pressure p and density ρ still must be specified. This so-called equation of state is just another constitutive relation, in addition to the Newtonian stress model, needed to complete the balance equations. □

Exercise 2.7b.

a. *Show that for incompressible flow, for which it holds that* $\nabla\cdot\mathbf{v} = 0$*, the Navier–Stokes equations read as*

$$\rho\, D_t\mathbf{v} = -\nabla p + \eta\,\Delta\mathbf{v} + \mathbf{f_b}. \tag{2.47}$$

b. *In practice,* (2.47) *has to be used in dimensionless form. Let us consider a system with typical length scale* l_0*, typical mass density* ρ_0*, and typical velocity* v_0*. One could think of a flow in a pipe of diameter* l_0*, mean density* ρ_0*, and mean flow velocity* v_0*. Show, by scaling in a natural way via* $\mathbf{v} = \mathbf{v}^* v_0$*,* $\rho = \rho^*\rho_0$*,* $t = t^* l_0/v_0$*, etc., that* (2.47) *may be written in dimensionless form as*

$$\rho^* D_t^*\mathbf{v}^* = -\nabla^* p^* + \frac{1}{Re}\Delta^*\mathbf{v}^* + \mathbf{f_b}^* \tag{2.48}$$

with the Reynolds number Re *defined as*

$$Re := \frac{\rho_0\, l_0\, v_0}{\eta}\,. \tag{2.49}$$

The value of the dimensionless Reynolds number provides much information about the type of flow. If $Re \ll 1$, the viscous term in (2.48) with $\Delta^*\mathbf{v}^*$ is dominant. The system is then reasonably well described by a much simpler equation, the so-called Euler equation:

$$\Delta^*\mathbf{v}^* = 0\,. \tag{2.50}$$

The Laplace operator has a diffusive character: it tends to smooth out velocity differences via internal friction. The dynamics of systems described by the Euler equation is slow, since these systems are dominated by viscosity. If, on the contrary, $Re \gg 1$, the internal friction is relatively unimportant and the dynamics is governed by

$$\rho^* D_t^*\mathbf{v}^* = -\nabla^* p^*\,. \tag{2.51}$$

Then, velocity differences are not damped and the nonlinearity in this equation may lead to turbulent flow.

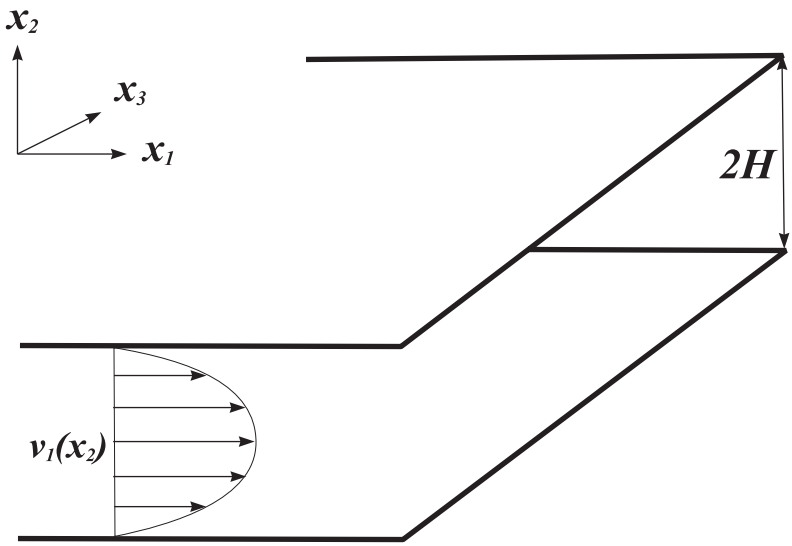

Figure 2.16. Laminar flow between two horizontal plates.

Exercise 2.7c. *Poiseuille flow.*

We consider the steady flow of a Newtonian fluid between two horizontal plates, as sketched in Fig. **2.16**. *We take Cartesian coordinates as indicated in the figure, with the fluid flowing in the x_1-direction, the x_2-direction in an upward vertical position, and the x_3-axis orthogonal to both x_1 and x_2. Thanks to symmetry, the velocity profile will not depend on x_3. The velocity profile is laminar, i.e., the velocity has a component in the x_1-direction only, and this component depends on x_2 only:*

$$\mathbf{v} = (v_1(x_2), 0, 0).$$

This implies that the fluid flows in horizontal layers with each layer having its own velocity. The fluid is pushed through the slit by a constant pressure difference Δp over the slit. This implies that the local pressure depends on x_1 as $p(x_1) = -\Delta p\, x_1/L$, with L the length of the slit in the x_1-direction. Furthermore, we assume that the fluid is incompressible.

a. *Show that if no body forces are relevant, the Navier–Stokes equation for this flow reduces to*

$$\rho\, \partial_t v_1 = -\frac{1}{L}\Delta p + \eta\, \partial_{x_2}^2 v_1. \tag{2.52}$$

b. *The slit height is $2H$ with x_2 ranging over $-H \le x_2 \le H$. At the boundary the no-slip boundary condition applies, so*

$$v_1(H) = v_1(-H) = 0.$$

Show that in the stationary situation this results in a parabolic-shaped profile, as sketched in Fig. **2.16**. *Such a solution is called a Poiseuille velocity profile is and given by*

$$v_1(x_2) = \frac{\Delta p}{2\eta L}\left(H^2 - x_2^2\right). \tag{2.53}$$

c. *Calculate the flux Q through the slit, i.e., the amount of fluid per unit of time through a slit of unit width. Then, evaluate the integral*

$$Q = \int_{-H}^{H} v_1(x_2)\, dx_2\ . \tag{2.54}$$

Example 2.7b. Generalized Newtonian flow.
If the fluid contains long molecules, as is the case in polymer solutions and polymer melts, the flow behavior becomes elastic in character, since the long molecular chains may entangle and form a (temporary) network. The viscoelasticity of such materials requires more complicated constitutive equations. One way to model this is to take η depending on the local deformation, thus on $\mathbf{D}$. The resulting model is called *generalized Newtonian flow*. It is mainly applied to so-called laminar flows, as sketched in Fig. **2.16**, in which the velocity has only one component $v_1(x_2)$. Laminar flow consists of horizontal layers which move at different speeds. Then, we can introduce the *shear rate* by

$$\dot\gamma = \partial_{x_2} v_1. \tag{2.55}$$

This notation is by convention. The shear rate measures the velocity differences between neighboring layers in the flow. The generalized Newtonian model reads as

$$\eta(\dot\gamma) = m\,\dot\gamma^n. \tag{2.56}$$

The constants m, n depend on the specific flow. This model may predict that the internal friction becomes smaller, if the velocity profile becomes steeper. This phenomenon is indeed observed in many flowing polymer melts and is referred to as *shear thinning*. □

Exercise 2.7d. *Generalized Poiseuille flow.*
Perform again the steps a–d *in Exercise* 2.7c *with* η *not constant but given by the model* (2.56). *For convenience, make use of the fact that the velocity profile will be symmetric with respect to* $x_2 = 0$ *so that only the interval* $0 \le x_2 H$ *needs to be considered. How does the symmetry condition that you must apply at* $x_2 = 0$ *read? How does the expression for the flux Q now depend on the extra parameters m and n?*

Example 2.7c. Maxwell model.
Another approach to describing viscoelastic behavior is to generalize the model introduced in (2.28) to more dimensions. In the three dimensional *Maxwell model*, the stress tensor satisfies the equation

$$\boldsymbol{\sigma}^e + \lambda \partial_t\, \boldsymbol{\sigma}^e = 2\,\eta\,\mathbf{D}.$$

This differential equation has to be read elementwise. In view of the symmetry of the tensors, we meet here with six separate differential equations.

□

2.8 Energy balance

With the experience of the balance equations for mass, heat, and momentum as presented above, it is quite easy to derive the balance relation for energy in continuous media. We hardly need to introduce new concepts. To apply the transport theorem (2.31), we identify the property f in the transport theorem with energy density e. For many systems the local energy is the sum of kinetic and internal energy:

$$e = \frac{1}{2}\,\rho\, v^2 + \rho\, U \tag{2.57}$$

with e, ρ, v, and U all depending on ($\mathbf{x}$,t). Here, v is the length of the vector $\mathbf{v}$. The kinetic energy is related to the average of the molecular motions, while the internal energy U, defined per unit of mass, represents all other types of energy stored in the molecules, for example, via random thermal motions, mutual interactions, or interactions with external fields. In Example 2.3c, we met with the internal energy. There, we focused on that part of the internal energy referred to as heat, i.e., the molecular vibrations that determine the temperature. The internal energy U introduced here may also include other types of energy.

The total energy E in a control volume V is given by

$$E(t) = \int_V e(\mathbf{x}, t)\, dV.$$

Changes in E are due to convection of energy through the surface of V, heat flux through the surface, and work done by the contact force $\mathbf{F}_c$ and/or body force $\mathbf{f}_b$, introduced in §2.7.2. It is important to discern carefully between the two types of energy flux through the surface. The energy flux $\mathbf{Q}_e = e\,\mathbf{v}$ represents the energy transport due to convection by moving molecules, carrying energy with them. The heat flux $\mathbf{Q}_h$ is due to transfer of energy by interactions of vibrating molecules, for example, via collisions. If there is no convection and thus $\mathbf{v} = 0$, we have $\mathbf{Q}_e = 0$, but $\mathbf{Q}_h$ in general will not be vanishing. The energy flux $\mathbf{Q}_e$ is accounted for automatically by the transport theorem, whereas the heat flux must be included separately. The total of heat H passing through the boundary of V per unit of time is given by

$$H = \int_A \mathbf{Q}_h \cdot \mathbf{n}\, dA = \int_V \nabla \cdot \mathbf{Q}_h\, dV. \tag{2.58}$$

The work W_c, done by the contact force per unit of time, is obtained by multiplying $\mathbf{F}_c$, given by (2.39), with the velocity $\mathbf{v}$ and integrating over the surface of V. Using Gauss's theorem (2.29) and the fact that the stress tensor is symmetric, we obtain the expression

$$W_c = \int_A \mathbf{v} \cdot \boldsymbol{\sigma} \cdot \mathbf{n}\, dA = \int_V \nabla \cdot (\boldsymbol{\sigma} \cdot \mathbf{v})\, dV. \tag{2.59}$$

For the work W_b done by the body force we have to multiply $\mathbf{f}_b$ with the velocity $\mathbf{v}$ and integrate over V. So,

$$W_b = \int_V \mathbf{v} \cdot \mathbf{f}_b\, dV. \tag{2.60}$$

The energy balance now reads

$$\frac{dE}{dt} = W_c + W_b - H. \tag{2.61}$$

The minus signs stem from taking the normal to the surface pointing outward. From the transport theorem (2.31) we obtain the differential form of the balance equation:

$$\partial_t\, e + \nabla \cdot (e\,\mathbf{v}) = \nabla \cdot (\boldsymbol{\sigma} \cdot \mathbf{v}) + \mathbf{v} \cdot \mathbf{f}_b - \nabla \cdot \mathbf{Q}_h. \tag{2.62}$$

This equation concerns the total energy, i.e., the sum of kinetic and internal energies. Equations for the two types of energy separately can also be derived. The balance equation for the kinetic energy density is obtained by taking the inner product of the velocity $\mathbf{v}$ with both sides of the momentum equation (2.43). Rewriting the result with the help of the continuity equation (2.32) we find

$$\partial_t \left(\tfrac{1}{2}\,\rho\, v^2\right) = -\nabla \cdot \left(\tfrac{1}{2}\,\rho\, v^2\,\mathbf{v}\right) - \mathbf{v} \cdot (\nabla \cdot \boldsymbol{\sigma}) + \mathbf{v} \cdot \mathbf{f}_b \tag{2.63}$$

with $v^2 := \mathbf{v} \cdot \mathbf{v}$.

If we subtract the balance of kinetic energy (2.63) from the balance of the total energy (2.62), we obtain the balance of internal energy

$$\partial_t(\rho U) = -\nabla \cdot (\rho\, U\mathbf{v}) - \nabla \cdot \mathbf{Q}_h - \boldsymbol{\sigma} : (\nabla \mathbf{v}). \tag{2.64}$$

In the last term, : (the colon) indicates full contraction of two matrices. For (square) matrices A and B with elements a_{ij} and b_{ij} this is defined as

$$A : B = \sum_{i,j} a_{ij}\, b_{j\,i}.$$

The last term in (2.64) represents energy dissipation. Due to internal friction between the layers of the laminar flow, heat is produced locally. The production rate is proportional to the viscosity η.

Exercise 2.8.

a. *Derive the balance of kinetic energy* (2.63) *from the momentum balance* (2.44) *using the continuity equation* (2.32).

b. *Derive the balance of internal energy* (2.64) *by subtracting the balance of kinetic energy* (2.63) *from the balance of total energy* (2.62).

Example 2.8a. Temperature profile in Poiseuille flow.
Let us apply (2.64) to calculate the temperature density in the Poiseuille flow dealt with in Exercise 2.7c. The flow geometry is described there and sketched in Fig. **2.16**. If we assume all material constants, such as ρ and η, are independent from the temperature, the momentum equation can be solved separately from the energy equation. So, for the calculation of the temperature density, we can take the Poiseuille velocity profile (2.53) for granted. In the steady state the temperature density depends on x_2 only, just as is the case for the velocity profile. For steady state calculations, the left-hand side of (2.64) vanishes. Since we are interested in heat, we take $U = c\,T$. The stress tensor $\boldsymbol{\sigma}$ can be calculated from the

Poiseuille profile (2.53). The pressure varies linearly along the slit, as discussed in part a of Exercise 2.7c. As constitutive relation for the heat flux we take the Fourier law (2.36). □

Exercise 2.8a.

a. *Derive from* (2.64) *the differential equation governing the steady state temperature profile.*

b. *As for the temperature boundary conditions, we take both plates at the same constant temperature* T_0. *Find an expression for the temperature profile by solving the equation under* a.

c. *Discuss the shape of the temperature profile. Take into consideration the local heat production due to internal friction. Discuss also the corresponding heat flux profile.*

2.9 Challenging problem: Shallow water waves

Here we apply the mass and momentum conservation laws to the description of water waves. The resulting model describes the desired evolution of surface waves in the approximation of shallow water. In this example many concepts introduced above merge into an integrated form. Consider a layer of fluid, for instance, water. See Fig. **2.17**. A deviation of the still water level will produce surface waves that evolve under the influence of gravity. Let us simplify the analysis a bit by assuming that the water movements are independent of one horizontal coordinate, for which we take the y-axis. So, we study the water movements as functions of the horizontal x-axis and the upward vertical z-axis. The water level in absence of waves is taken at $z = 0$ and the flat bottom (assumed to be impermeable) is at $z = -H$ so that H is the depth of the layer. We assume that the bottom is flat, so H is constant,

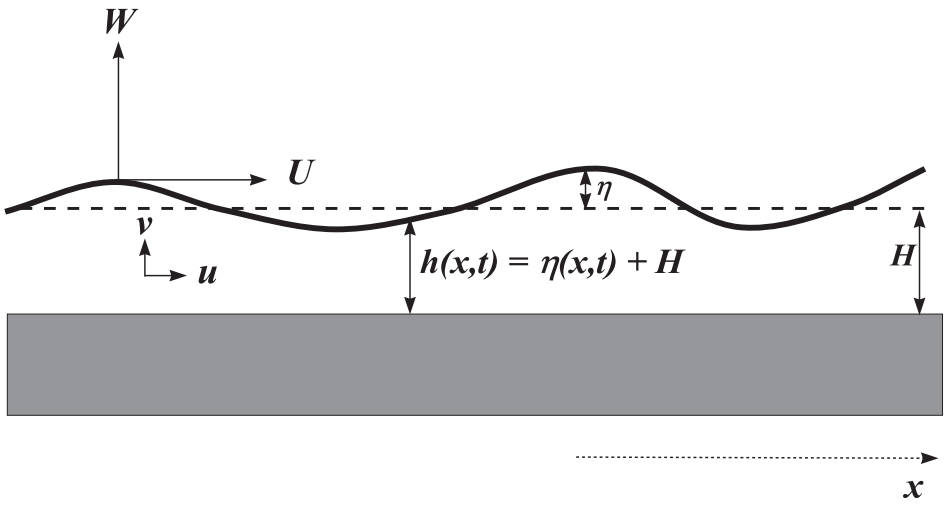

Figure 2.17. Sketch of shallow water waves.

and that the layer is shallow, so H is small. Furthermore, the water has constant density ρ. Effects of surface tension are not taken into account.

The free surface is represented by the function $z = \eta(x, t)$. The height h of the free surface above the bottom is then given by $h(x, t) = \eta(x, t) + H$. Inside the fluid, the horizontal and vertical components of the fluid velocity $\mathbf{v}$ are denoted by (u, w), respectively. At the free surface, these velocity components are referred to as (U, W).

Mass balance

First consider mass conservation. It will turn out that this leads to a dynamic equation for the free surface. In deriving the mass balance, we follow the same reasoning as in §2.2 but with one new element: The free surface changes its position in time. Let us consider the mass between two vertical lines at $x = a$ and $x = b$. See Fig. **2.17**. This mass is given by the area under the curve:

$$m_{a,b}(t) = \rho \int_a^b h(x, t)\, dx = \rho\, H\, (b - a) + \rho \int_a^b \eta(x, t)\, dx. \tag{2.65}$$

Changes in time of this mass are caused by water fluxes through the vertical lines and give rise to variations in the free surface. The water flux $Q(x, t)$ through a vertical line at x is given by

$$Q(x, t) = \rho \int_{-H}^{\eta(x,t)} u(x, z, t)\, dz. \tag{2.66}$$

Conservation of mass implies

$$\frac{dm_{a,b}}{dt}(t) = -[Q(b, t) - Q(a, t)]\,. \tag{2.67}$$

This leads to the balance equation

$$\partial_t\, \eta(x, t) = -\,\partial_x \int_{-H}^{\eta(x,t)} u(x, z, t)\, dz. \tag{2.68}$$

Exercise 2.9a.
Derive the mass balance (2.68).

The spatial derivative in the right-hand side of (2.68) has an effect on the integrand and on the upper boundary in the integral:

$$\partial_x \int_{-H}^{\eta} u(x, z, t)\, dz = u(x, \eta(x, t), t)\, \partial_x \eta(x, t) + \int_{-H}^{\eta} \partial_x\, u(x, z, t)\, dz. \tag{2.69}$$

The first term on the right-hand side is expressed in quantities at the free surface only. Also, the second term can be put into that form by applying the incompressibility assumption. From $\nabla \cdot \mathbf{v} = \partial_x u + \partial_z w = 0$, the second term on the right-hand side can be simplified:

$$\begin{aligned} \int_{-H}^{\eta} \partial_x\, u(x, z, t)\, dz &= -\int_{-H}^{\eta} \partial_z w(x, z, t)\, dz \\ &= -w(x, \eta(x, t), t) + w(x, -H, t) = -W(x, t). \end{aligned} \tag{2.70}$$

Here, we have used that the bottom is impermeable: $w(x, -H, t) = 0$. The dynamics of the free surface thus satisfy the *kinematic relation*

$$\partial_t \eta = -U \partial_x \eta + W. \tag{2.71}$$

This relation is valid without any approximation and can be seen as a variant of the continuity equation. It is convenient to read the right-hand side as the inner product of two vectors:

$$\partial_t \eta = (U, W) \cdot (-\partial_x \eta, 1). \tag{2.72}$$

This equation simply expresses that the evolution of the free surface is determined by the component of the free surface velocity (U, W) normal to the surface. To clarify this, we remark that the free surface is parameterized with parameter x and given by the vector $(x, \eta(x, t))$. The tangent vector of the free surface is then obtained by differentiating with respect to x. This yields the vector $(1, \partial_x \eta)$. The normal vector of the surface is thus given by $(-\partial_x \eta, 1)$. Note that both tangent and normal vectors are *not* normalized here.

In the case of shallow water, (2.71) simplifies. For shallow water, it is reasonable to neglect the z-dependence of the horizontal velocity component u. Note that this does not mean that we take the vertical velocity w equal to zero, which would violate the incompressibility condition. We thus take $u = u(x, t)$. Direct integration of the integral in (2.68) is now possible. This results in

$$\partial_t \eta + \partial_x((\eta + H)\, u) = 0, \tag{2.73}$$

or, equivalently,

$$\partial_t h + \partial_x (h\, u) = 0. \tag{2.74}$$

The conservation of mass thus leads to this kinematic relation for shallow water, which describes the dynamics of the water surface.

Momentum balance

In the shallow water approximation, the layer is treated as a one-dimensional object and the variations in the z-direction are averaged out by integrating over z. The momentum at x is given by

$$M(x, t) = \int_{-H}^{\eta} \rho\, u(x, t)\, dz = \rho\, h\, u(x, t) \tag{2.75}$$

and the momentum flux by

$$Q(x, t) = \int_{-H}^{\eta} \rho\, u^2(x, t)\, dz = \rho\, h\, u^2(x, t). \tag{2.76}$$

The stress tensor has the isotropic form mentioned in Exercise 2.7a. The local pressure is denoted by $p(x, z, t)$, and the total pressure along a vertical line at x is given by

$$P(x, t) = \int_{-H}^{\eta} p(x, z, t)\, dz. \tag{2.77}$$

The transport theorem (2.31) and/or the momentum balance (2.44) then lead to

$$\partial_t M + \partial_x Q = -\partial_x P. \tag{2.78}$$

This can be written as

$$\partial_t (\rho\, h\, u) = -\partial_x \left\{ \rho\, h\, u^2 + \int_{-H}^{\eta} p(x,z,t)\, dz \right\}.$$

In a shallow layer, p is well approximated by the *hydrostatic pressure*

$$p = \rho\, g\, (\eta - z) + p_{atm},$$

where g is the gravitational acceleration and p_{atm} the atmospheric pressure. If we take for convenience $p_{atm} = 0$, the governing equation becomes

$$\partial_t (hu) = -\partial_x \left\{ h\, u^2 + \frac{1}{2}\, g\, h^2 \right\}.$$

By using the continuity equation (2.74), this can be simplified to

$$\partial_t\, u + u\, \partial_x u = -g\, \partial_x h \tag{2.79}$$

or

$$\partial_t\, u = -\partial_x \left\{ \frac{1}{2} u^2 + g\, h \right\}. \tag{2.80}$$

Comparison with the transport theorem (2.31) suggests we read this as a conservation law for u and interpret $\frac{1}{2}u^2 + g\, h$ as the flux of u.

The two equations (2.74), (2.79) form a closed set of equations and describe the desired evolution of the surface waves in the approximation for shallow layers.

Exercise 2.9b.
Consider the nonlinear surface wave equations above for small wave heights.

a. *Show that the full equations have as trivial solution*

$$u(x,t) := 0, \quad \eta(x,t) := 0.$$

b. *The linearized equations can be written and lead to a second-order wave equation for η:*

$$\partial_t^2\, \eta = c^2\, \partial_x^2 \eta \;\; with \; c = \sqrt{g\, H}.$$

This is the standard wave equation. Check that the general solution is given by

$$\eta(x,t) = f_1(x - c\, t) + f_2(x + c\, t)$$

and that it consists of waves $f_1(x - ct)$ traveling to the right and waves $f_2(x + ct)$ traveling to the left at velocity c.

c. *Show that the full equations have also special solutions of the form*

$$u(x,t) := U, \quad \eta(x,t) := 0$$

corresponding to a uniform flow in the x-direction with constant velocity U in a layer with a flat surface.

Chapter 3
Basic Concepts

In this chapter we present and discuss a number of basic concepts that should be part of the standard luggage of any modeler. As shown in Chapter 2, many models in science are formulated in terms of differential equations, since we are nearly always interested in the time evolution of systems. In this chapter we present an overview of the basic concepts, both for ordinary differential equations (ODEs) and partial differential equations (PDEs), emphasizing the similarities between both cases. The emphasis is on linear models. Among all models, the linear ones form a special class. Linearity allows for superposition: any linear combination of solutions is itself a solution. This backbone of linear theory makes it possible to develop standard methods for a wide class of linear models. This implies that these models are fairly well understood, and unexpected dynamics will not show up in these systems. Well-known methods based on linear concepts are Fourier and Laplace transformations. It is not our purpose to discuss these powerful techniques here. We prefer to outline the essentials of the general procedure underlying these methods: expansion of the solution in terms of an appropriately chosen basis set. This expansion idea is the most widely applied procedure to determine the (approximated) solution of both linear and nonlinear systems. We introduce this concept via examples. In Chapter 4 we use it to determine stability of systems, and in Chapter 5 we discuss this approach in a more abstract setting.

For nonlinear models the situation is different from the linear case. Not one well-established scheme but a great variety of techniques is available. Still, much current research is devoted to finding new ways to encounter these models. Even a totally new breakthrough is possible, as shown by the rapid progress in chaos theory, which has a history that goes back no more than about 40 years. In analyzing nonlinear equations, one often falls back onto the insights gained in the context of linear models. That is why one often tries to reduce a nonlinear model to a linear one, which in general is easier to solve. The price one usually has to pay for this advantage is that the reduced model is reliable in only a limited interval in space and/or time.

3.1 State and state space

In the preceding chapters we met with many models. As introduced in §1.1, any mathematical model consists of equations representing the relations between the relevant properties

of the system under consideration. In these models we meet with *dependent* and *independent variables*, together with *parameters*. For systems described by ODEs, the dependent variables are functions of one variable t, which we like to interpret as time. For systems described by PDEs, they also depend on position $\mathbf{x}$. In practice we are nearly always interested in the evolution in time of systems. To cope with the evolutionary character of most models we introduce the concept of *state*.

3.1.1 State

To specify a unique solution of an ODE, we have to specify initial conditions. In the case of a PDE these are supplemented with boundary conditions. The initial conditions contain the information that completely characterizes the system at one moment in time. We call this amount of information a *state* of the system. For systems that are invariant under translation in time, the so-called autonomous systems, the moment of specification of initial data is not important. However, for nonautonomous systems, the moment of specification is itself part of the information.

From the theory of differential equations it is well known that for autonomous systems that involve only first-order time derivatives, it suffices to specify as initial conditions the values of all dependent variables. From that information one can deduce their values at other times via integration in time. So, for autonomous first-order systems, a state of the system is nothing but a set of values for the dependent variables.

If second-order time derivatives are involved, one has to include the first-order time derivatives of the dependent variables in the initial conditions. Then the state has to include these time derivatives, too, and this doubles its number of elements. (And so on, if higher-order time derivatives are involved.)

Since the dynamics of most mechanical systems are governed by the second law of Newton ($F = m\,a$), which essentially contains accelerations, i.e., second-order time derivatives, the state of these systems includes both the positions and the velocities of the constituting parts. This is illustrated in the following simple example. We remark that if constraints are present, the values of the variables cannot be specified independently, and this then leads to a reduction of the state.

Example 3.1a. State of a driven harmonic oscillator.
Let us consider the driven, damped, harmonic oscillator (see also Example 1.2a): a particle with mass m moves in one direction under influence of a spring, a friction, and a driving force. The dependent variable is its position u, which is a function of the independent variable time t. The equation of motion reads as

$$m\ddot{u} + c\dot{u} + ku = F_0 \sin \omega t. \tag{3.1}$$

The state of this oscillator is a two-dimensional vector $\mathbf{u}$ having as components the position $u_1 := u$ and the velocity $u_2 := \dot{u}$. It is convenient to rewrite the second-order equation (3.1) as two first-order equations:

$$\begin{cases} \dot{u}_1 = u_2, \\ \\ \dot{u}_2 = -\dfrac{c}{m}u_2 - \dfrac{k}{m}u_1 + \dfrac{F_0}{m}\sin \omega t. \end{cases} \tag{3.2}$$

This can be concisely written in the standard form

$$\dot{\mathbf{u}} = \mathbf{A} \cdot \mathbf{u} + \mathbf{b} \tag{3.3}$$

with the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ given by

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{pmatrix}, \quad \mathbf{b}(t) = \begin{pmatrix} 0 \\ \frac{F_0}{m}\sin(\omega t) \end{pmatrix}. \tag{3.4}$$

□

Note that throughout this book the · (center dot) notation is used for both the inner product of vectors and the matrix-vector product. This convention stems from tensor theory notation, where · stands for *contraction* of two tensors of any type.

3.1.2 State space

ODEs

If an ODE system has n dependent variables and is of order k with respect to time derivation, its state $\mathbf{u}$ is a vector of length $N = nk$: it contains all dependent variables and their time derivatives of order $1, 2, \ldots, k-1$. The set of all possible state vectors is denoted by U, the *state space*, and we have $U \subset \mathbb{R}^N$. In practice, we nearly always meet with conditions on the variables and their derivatives. For example, if we model the dynamics of a racing car, we know beforehand that its trajectory is bounded to follow a prescribed racing course and its speed will not exceed a maximum because of power limitations of the engine. So, the state vectors attain only values that satisfy the constraints. This means that in most models U is a proper subset of $\mathbb{R}^N$. The vectors in U are called *feasible* or *admissible* states.

In the two-dimensional case the state space is often referred to as the *phase plane*. One axis in this plane is used for the position, and the other, orthogonal to the first one, is used for the velocity (or the momentum) of the object. Note that the motion of a body along a line is one-dimensional from the physical point of view but two-dimensional from the mathematical point of view since in the state space formulation it corresponds to a motion in a virtual plane, the phase plane.

PDEs

Similar remarks apply to PDEs. The dependent variables now depend on position $\mathbf{x}$, too. If a PDE system has n dependent variables and is of order k with respect to time derivation, its state $\mathbf{u}$ is a set of $N = nk$ functions: all dependent variables and their time derivatives of order $1, 2, \ldots, k-1$. Several restrictions are in force. First, the elements of $\mathbf{u}$ must satisfy certain conditions on continuity and differentiability. These follow from the properties of the system to be modeled. Second, they must satisfy the boundary conditions. Third, additional constraints may hold. For example, the model for a vibrating string in Example 3.1b in this chapter holds only if the amplitude of the string is much smaller than the length of the string. The states that satisfy all conditions, and are thus *feasible*, together form the state space U. In the case of an ODE, the state space is finite dimensional. In the case of a PDE, this space is a function space and thus infinitely dimensional.

Boundary conditions
In models described by PDEs, the variable $\mathbf{x}$ runs over a set $\Omega \in \mathbb{R}^n$ with boundary $\partial\Omega$. The solution $\mathbf{u}(\mathbf{x}, t)$ of a PDE has to satisfy certain boundary conditions on $\partial\Omega$ in order to be uniquely defined. In the modeling practice one mostly; meets a few types of boundary conditions. Let u be one of the components of the solution; then the most common boundary conditions for this component are as follows:

Dirichlet conditions:

$$u(\mathbf{x}, t) = f(\mathbf{x}, t), \quad \mathbf{x} \in \partial\Omega\ , \quad \forall t\ . \tag{3.5}$$

Neumann conditions:

$$\mathbf{n} \cdot \nabla u = g(\mathbf{x}, t), \quad \mathbf{x} \in \partial\Omega\ , \quad \forall t \tag{3.6}$$

with $\mathbf{n}(\mathbf{x}, t)$ the (outward) normal at the boundary $\partial\Omega$.

Mixed or Robin conditions:

$$u(\mathbf{x}, t) + c\,\mathbf{n} \cdot \nabla u = h(\mathbf{x}, t), \quad \mathbf{x} \in \partial\Omega\ , \quad \forall t. \tag{3.7}$$

Specification of the functions f, g, and h and the constant c is part of the modeling process. The situation may be quite complicated, e.g., when on different parts of the boundary $\partial\Omega$ different types of conditions are in force or when different types of conditions hold for different dependent variables.

If the functions f, g, and h in the boundary conditions above are vanishing, the conditions are *homogeneous*. Then a linearity property holds: if two functions satisfy the boundary conditions, then any linear combination of them also satisfies these conditions.

3.1.3 Evolution equations

By definition, each point $\mathbf{u}_0 \in U$ corresponds to a unique *state* of the system and may act as an initial condition. For convenience, we always take $t_0 = 0$ for the initial time here. If time evolves, the state vector $\mathbf{u}(t)$ with $\mathbf{u}(0) = \mathbf{u_0}$ follows a trajectory through the state space U, governed by the ODE or PDE under consideration. These differential equations act as equations of motion. ODE and PDE systems can be written in completely similar forms. For ODEs we write

$$\begin{cases} \dot{\mathbf{u}} := \dfrac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}, t)\ , \\ \mathbf{u}(0) = \mathbf{u}_0\ , \end{cases} \tag{3.8}$$

and for PDEs we write

$$\begin{cases} \partial_t \mathbf{u} := \dfrac{\partial \mathbf{u}}{\partial t} = \mathbf{F}(\mathbf{u}, \mathbf{x}, t)\ , \\ \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x})\ . \end{cases} \tag{3.9}$$

In (3.9) it is not necessary to mention the boundary conditions explicitly, since they are satisfied by all states in U. Before discussing these evolution equations in more detail, we first make some general remarks:

- The evolution equations (3.8) and (3.9) have a relatively simple form, since they are first-order in time.
- The right-hand sides $\mathbf{F}$ could be interpreted as the velocity of the system when it follows a path in U. However, this velocity in the state space should not be confused with the velocity of the system in real space, not even in the case of mechanical systems.
- As mentioned above, if $\mathbf{F}$ does not explicitly depend on t, but only implicitly via $\mathbf{u}$, we call the system *autonomous*.
- The model is called *linear* if for any two solutions $\mathbf{u}_1(t)$ and $\mathbf{u}_2(t)$ it holds that an arbitrary linear combination $c_1\,\mathbf{u}_1(t) + c_2\,\mathbf{u}_1(t)$ is also a solution. Note that in the case of a PDE, linearity is a property not only of $\mathbf{F}$ but also of the boundary conditions, since these must be *homogeneous*.
- Special solutions of (3.8) and (3.9) are points $\mathbf{u}_0$ for which $\mathbf{F} = 0$ for all times. If the system starts in such a state, it will never leave it, since the evolution equation then prescribes that its position in U will not change. We call these special points *stationary states*. In the literature they are referred to with many different names, such as *critical states, singular states, equilibrium states, rest states*, and *steady states*. It may happen that a nonautonomous vector field has a stationary state. An ODE example is $\mathbf{F}(\mathbf{u}, t) = \mathbf{u}\,t$, which clearly has $\mathbf{u} = \mathbf{0}$ as a stationary state. However, it is the exception rather than the rule that a nonautonomous system has a stationary state.
- It has to be realized that in practice it is not possible to start a system exactly in a prescribed stationary state, since perturbing influences can never be avoided. This makes the stability analysis of stationary states of great practical interest. Do small perturbations cause the trajectory to leave the stationary state? These questions will be addressed in Chapter 4.

In case of an ODE, $\mathbf{F}$ in (3.8) is called the *vector field*. It is a vector-valued function with the same dimension as the state. If $\mathbf{F}$ has the special form

$$\mathbf{F}(\mathbf{u}, t) = \mathbf{A}(t) \cdot \mathbf{u} + \mathbf{b}(t), \tag{3.10}$$

the vector field is called *affine*. An affine vector field with $\mathbf{b}(t) := \mathbf{0}$ is called *linear*. Sometimes an affine system is called *linear* with the extra indication *inhomogeneous*. Then, the case $\mathbf{b}(t) := \mathbf{0}$ is denoted as *linear and homogeneous*.

Exercise 3.1a.

Classify the vector field in (3.3) *with* (3.4) *with respect to the linear, autonomous, and homogeneous properties.*

Exercise 3.1b.

In Exercise 1.4a *we introduced the simple system of a projectile, catapulted in the upward direction. Its height $z(t)$ is governed by the equation of motion*

$$m\ddot{z} = -mg.$$

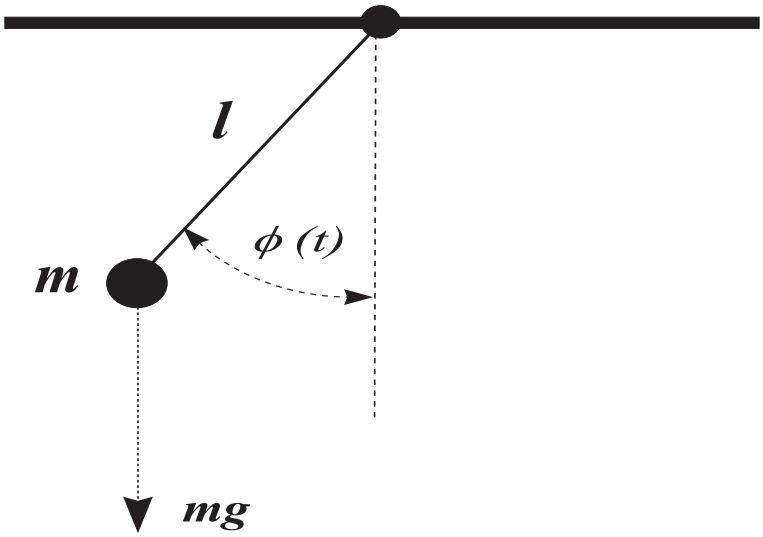

Figure 3.1. Pendulum of mass m and length ℓ under influence of a gravity force mg and friction.

Write this model in state space form and classify the vector field with respect to the linear, autonomous, and homogeneous properties.

Exercise 3.1c.
The equation of motion of the mathematical pendulum (see Fig. **3.1***) reads as*

$$m\,\ell\,\ddot{\varphi} + c\,l\,\dot{\varphi} + m\,g\,\sin\varphi = 0, \tag{3.11}$$

where we include the effect of friction. Write this vector field in state space form and classify it with respect to the linear, autonomous, and homogeneous properties. Next, use the approximation $\sin(\varphi) \sim \varphi$*, which holds for small amplitude oscillations around the downward position. How does this change the classification?*

In the case of a PDE, **F** in (3.9) is an operator acting on **u**. In most cases, its action involves taking derivatives of **u** with respect to **x**, but integral operators also can be involved. If **F** does not contain a derivation or integration, (3.19) reduces to a set of coupled ODEs: for each **x**-value, one ODE.

In the next example we present a well-known PDE, namely, the wave equation. This equation of motion will be derived in §6.2, in the context of the modeling of a vibrating polymer chain. See also (1.22) in §1.5.3 for a simpler wave equation.

Example 3.1b. Vibrating string.
Consider a string of length L vibrating in a plane. See Fig. **3.2**. Let us take the x-axis along the rest position of the string and denote its amplitude, measured orthogonal to this axis, by

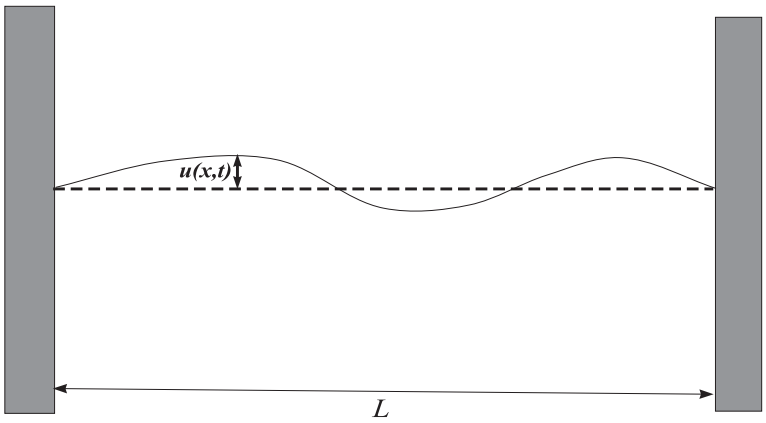

Figure 3.2. A vibrating string, clamped at both ends.

$u(x,t)$. For small amplitudes the string dynamics then satisfies the *wave equation*

$$\partial_t^2 u = c^2\, \partial_x^2 u \tag{3.12}$$

with the speed c depending on the material properties and the stress in the string. Since the string is clamped at both ends, the boundary conditions are

$$u(0,t) = u(L,t) = 0\,, \quad t \geq 0\,. \tag{3.13}$$

By introducing the state $\mathbf{u}(x,t) = (u(x,t), \dot{u}(x,t))$, we may write (3.12) in the state space form

$$\partial_t \mathbf{u} = \mathbf{F}(\mathbf{u})\,, \;\; \mathbf{F} = \begin{pmatrix} 1 \\ \\ c^2\, \partial_x^2 \end{pmatrix}. \qquad \square$$

In the next example we model the dynamics of an algae population. The resulting form of $\mathbf{F}$ is quite general and can also be used to model many other systems.

Example 3.1c. Algae dynamics.
Let us model the dynamics of an algae population in a long tube filled with water. See Fig. **3.3**. The water in the tube flows at constant speed v. Since the tube is much longer than the radius of its cross section, we may treat this system as being one-dimensional, taking the centerline of the tube as the x-axis. Let $u(x,t)$ represent the density of the population. We consider only first-order derivatives with respect to time so that the density forms the state of the system. The water entering the tube at the end $x = 0$ is clean and does not contain algae. At the other end, where $x = L$, a filter has been built in which the algae are captured so that there is no flux of algae leaving the tube. This is expressed by the boundary conditions

$$u(0,t) = 0\,, \quad \partial_x u(L,t) = 0 \;\; \forall t. \tag{3.14}$$

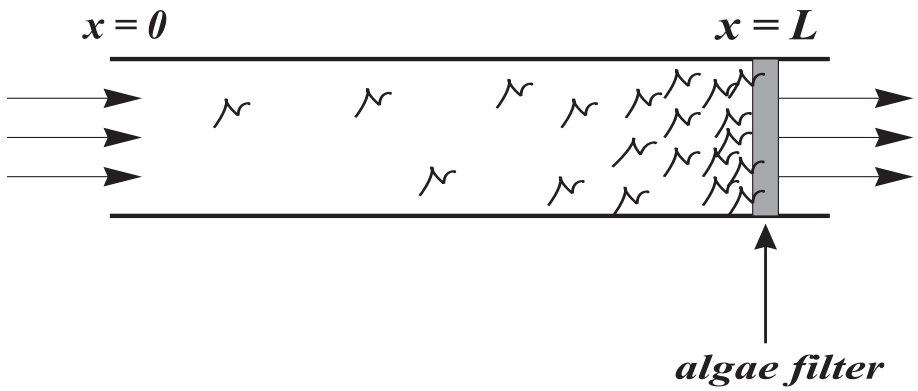

Figure 3.3. The time evolution of an algae population in a tube could be described by a PDE (3.15) in which the various terms represent the effects of consumption, growth, convection, and diffusion of algae.

Whether these boundary conditions lead to a well-posed problem depends on the form of **F**. If we have only these two boundary conditions, the operator **F** should be second order with respect to spatial derivatives. Several mechanisms apply by which the population density $u(x, t)$ may change. We assume the algae dynamics to be governed by

$$\partial_t u = c_0 + c_1\, u + c_2\, \partial_x u + c_3\, \partial_x^2 u. \tag{3.15}$$

In general the coefficients in such an expression may depend on $u(x, t)$, x, and t. The system may then be nonlinear, inhomogeneous, and/or nonautonomous. However, for illustrative purposes we take them to be constant.

The different terms in (3.15) represent different physical phenomena. This is most easily explained by focusing on each term separately, ignoring for now the boundary conditions. The respective terms represent the following effects:

a. $\partial_t u = c_0$.
This term represents that the density u increases ($c_0 > 0$) or decreases ($c_0 < 0$) linearly at each point x, so the solution is simply given by $u(x, t) = u(x, 0) + c_0 t$. In the present model the algae are consumed, e.g., by fish. For negative c_0 this term describes that a fixed number of algae is eaten per unit of time. Thus the c_0 term acts as a sink.

b. $\partial_t u = c_1 u$.
This term represents exponential growth ($c_1 > 0$) or decay ($c_1 < 0$) of the population, and the solution is $u(x, t) = u(x, 0)e^{c_1 t}$. If enough food is available in the water, the algae tend to multiply very fast with the density growing exponentially.

c. $\partial_t u = c_2\, \partial_x u$.
This is the *convective term*, since it represents the effect of flow of the water in the tube. It is easily concluded that c_2 must be equal to the water velocity v. Given an initial algae density $u_0(x)$, we have that the density at later times is given by $u(x, t) = u_0(u + v\, t)$: the initial profile is propagated along the tube at speed v with preserved shape.

d. $\partial_t u = c_3\, \partial_x^2 u$.
This is a *diffusive* term, met earlier in Example 2.3d. It describes algae that tend to diffuse through the water to places where the concentration is lower. Diffusion processes are usually much slower than convection processes.

A model like the one in (3.15) can be conveniently solved using the methods dealt with in §3.3.2. In Exercise 3.3e this will be done for a special case. □

3.2 ODEs

The simplest ODE system, which is still often met in practice, is given by the linear, autonomous system

$$\begin{cases} \dot{\mathbf{u}} = \mathbf{A} \cdot \mathbf{u}, \\ \\ \mathbf{u}(0) = \mathbf{u}_0 \end{cases} \tag{3.16}$$

with the matrix $\mathbf{A}$ being time independent. It has the explicit solution

$$\mathbf{u}(t) = e^{\mathbf{A}t} \cdot \mathbf{u}_0 \tag{3.17}$$

with the matrix exponential $\exp(\mathbf{A}t)$ defined via its Taylor series

$$e^{\mathbf{A}t} = \mathbf{I} + t\mathbf{A} + \tfrac{1}{2}t^2\mathbf{A}^2 + \tfrac{1}{3!}t^3\mathbf{A}^3 + \cdots, \tag{3.18}$$

where $\mathbf{I}$ is the unit matrix. A simple expression like (3.17) in general does not hold if $\mathbf{A}$ is time dependent. Only under a severe condition is the solution of

$$\dot{\mathbf{u}} = \mathbf{A}(t) \cdot \mathbf{u}$$

given by

$$\mathbf{u}(t) = e^{\bar{\mathbf{A}}(t)} \cdot \mathbf{u}_0 \tag{3.19}$$

with

$$\bar{\mathbf{A}}(t) = \int_0^t \mathbf{A}(t')\, dt'.$$

This condition is that $\mathbf{A}(t)$ and $\bar{\mathbf{A}}(t)$ commute, i.e.,

$$\mathbf{A}(t) \cdot \bar{\mathbf{A}}(t) = \bar{\mathbf{A}}(t) \cdot \mathbf{A}(t)\ , \quad t \geq 0\ . \tag{3.20}$$

Then, also $\exp(\bar{\mathbf{A}}(t))$ and $\mathbf{A}(t)$ commute, as directly follows from expansion (3.18).

Exercise 3.2a.
Consider the matrix $\mathbf{A}(t)$ *given by*

$$\mathbf{A}(t) = \begin{pmatrix} 1 & 2t \\ \\ 0 & 1 \end{pmatrix}.$$

a. *Calculate* $\bar{\mathbf{A}}(t)$.

b. *Show that* $\mathbf{A}(t)$ *and* $\bar{\mathbf{A}}(t)$ *commute.*

c. *Find explicit expressions in terms of* t *for the matrix elements of the operator* $\exp(\bar{\mathbf{A}}(t))$ *by using the Taylor expansion* (3.18) *and the fact that the* n*th powers of the matrix* $\bar{\mathbf{A}}(t)$ *have a simple structure,* $n = 1, 2, 3, \ldots$.

The solution of the inhomogeneous linear system

$$\begin{cases} \dot{\mathbf{u}} = \mathbf{A} \cdot \mathbf{u} + \mathbf{b}(t), \\ \\ \mathbf{u}(0) = \mathbf{u}_0, \end{cases} \tag{3.21}$$

with $\mathbf{A}$ a constant matrix, is simply found by introducing

$$\mathbf{v} := e^{-\mathbf{A}t} \cdot \mathbf{u}, \tag{3.22}$$

which satisfies

$$\begin{cases} \dot{\mathbf{v}} = \mathbf{b}(t), \\ \\ \mathbf{v}(0) = \mathbf{u}_0. \end{cases} \tag{3.23}$$

Integrating this ODE leads to an extension of (3.17):

$$\mathbf{u}(t) = e^{\mathbf{A}t} \cdot \mathbf{u}_0 + \int_0^t e^{\mathbf{A}(t-t')} \cdot \mathbf{b}(t')\, dt'. \tag{3.24}$$

This famous expression is usually referred to as the *variation of constants formula*.

Exercise 3.2b.
Check by substitution that (3.24) *is the solution of* (3.21).

In (3.24) we recognize a superposition principle. The solution of (3.21) turns out to be the sum of the solution (3.17) of its homogeneous part (3.16) and a so-called particular solution, given by the integral in (3.24). The latter represents the effect of the inhomogeneous term $\mathbf{b}(t)$. The explicit form (3.24) reveals that the particular solution is the solution of (3.21) if we take vanishing initial value; thus $\mathbf{u}_0 = 0$.

Exercise 3.2c.
Consider the system $\dot{\mathbf{u}} = \mathbf{A}(t) \cdot \mathbf{u}$ *with* $\mathbf{A}(t)$ *as in Exercise* 3.2a*. Write this system as two separate ODEs for the components of* $\mathbf{u} = (u_1, u_2)$*. The equation for* u_2 *can be directly solved. Substitute the solution for* u_2 *in the equation for* u_1*. This leads to a linear, inhomogeneous equation for* u_1*. Solve this equation using the variation of constants formula* (3.24) *and check your answer to that of Exercise* 3.2a*, where the same problem has been solved in an alternative way.*

3.2.1 Linearizing ODEs

For nonlinear ODEs we often want to make use of the insights gained for linear ones. This is especially the case if we study stability properties of solutions, which is the subject of Chapter 4. Here we prepare this by introducing the *linearization* of a nonlinear ODE. Let $\mathbf{u}_0$ be a stationary point of a nonlinear, autonomous vector field $\mathbf{F}$ so that

$$\mathbf{F}(\mathbf{u}_0) = \mathbf{0}.$$

If we take as the initial state $\mathbf{w}_0 = \mathbf{u}_0 + \varepsilon \mathbf{v}_0$, with $||\mathbf{v}_0|| = 1,\ \varepsilon \ll 1$, and thus $\mathbf{w}_0$ close to $\mathbf{u}_0$, we know that the solution $\mathbf{w}(t)$ of

$$\begin{cases} \dot{\mathbf{w}} = \mathbf{F}(\mathbf{w}), \\ \mathbf{w}(0) = \mathbf{w}_0 := \mathbf{u}_0 + \varepsilon \mathbf{v}_0 \end{cases} \tag{3.25}$$

will remain in the vicinity of $\mathbf{u}_0$, at least for some period after the initial time. The difference $\mathbf{v}(t) = \mathbf{w}(t) - \mathbf{u}_0$ satisfies

$$\begin{cases} \dot{\mathbf{v}} = \mathbf{F}(\mathbf{u}_0 + \mathbf{v}), \\ \mathbf{v}(0) = \varepsilon \mathbf{v}_0. \end{cases} \tag{3.26}$$

For small times we may apply a Taylor expansion of $\mathbf{F}(\mathbf{u}_0 + \mathbf{v})$ around $\mathbf{u}_0$, assuming that $\mathbf{F}$ is differentiable at $\mathbf{u}_0$. So we use

$$\mathbf{F}(\mathbf{u}_0 + \mathbf{v}) \approx \mathbf{F}(\mathbf{u}_0) + \mathbf{J}(\mathbf{u}_0) \cdot \mathbf{v} = \mathbf{J}(\mathbf{u}_0) \cdot \mathbf{v}, \tag{3.27}$$

leaving out terms of the order of $\|\mathbf{v}\|^2$. The Jacobi matrix $\mathbf{J}(\mathbf{u})$ of $\mathbf{F}(\mathbf{u})$ is defined as

$$\mathbf{J}_{ij} = \frac{\partial F_i}{\partial u_j}, \quad i, j = 1, \ldots, n, \tag{3.28}$$

where F_i is the ith component of $\mathbf{F}$ and u_j the jth component of $\mathbf{u}$.

So, $\mathbf{v}(t)$ initially satisfies

$$\dot{\mathbf{v}} = \mathbf{J}(\mathbf{u_0}) \cdot \mathbf{v}. \tag{3.29}$$

This is called the *linearization* of the vector field $\mathbf{F}$ around $\mathbf{u}_0$. Since the linearized equation (3.29) has a constant matrix, it is of type (3.16) and thus is easy to solve.

Exercise 3.2d.

a. *Linearize the vector field of the pendulum in Exercise* 3.1c *(equation* (3.11)*), around the downward position at rest, and thus around* $(\varphi, \dot{\varphi}) = (0, 0)$.

b. *Linearize this system also around the upward position at rest and thus around* $(\varphi, \dot{\varphi}) = (\pi, 0)$.

3.2.2 Expansions in basis vectors

Solutions (3.17) and (3.24) of the linear systems (3.16) and (3.21), respectively, considerably simplify if we make use of expansions in a convenient basis set. To that end we may make use

of the eigenvectors of the matrix $\mathbf{A}$. If the constant $n \times n$ matrix $\mathbf{A}$ has n different eigenvectors $\phi_i,\ i = 1, \ldots, n$, corresponding to eigenvalues $\lambda_i,\ i = 1, \ldots, n$, these eigenvectors form a basis in $\mathbb{R}^n$ and thus also in the state space $U \subset \mathbb{R}^n$. This implies that the initial perturbation can then be uniquely written as a linear combination of the ϕ_i:

$$\mathbf{u}_0 = \sum_{i=1}^{n} c_i\, \phi_i. \tag{3.30}$$

Similarly, for each time t the inhomogeneous term $\mathbf{b}(t)$ can be uniquely expanded:

$$\mathbf{b}(t) = \sum_{i=1}^{n} b_i(t)\, \phi_i. \tag{3.31}$$

Note that the expansion coefficients are now time dependent. Substituting this expansion in (3.24), we find that the solution $\mathbf{u}(t)$ of an inhomogeneous linear system can be written as

$$\mathbf{u}(t) = \sum_{i=1}^{n} \left[c_i\, e^{\lambda_i t} + \int_0^t e^{\lambda_i (t-t')}\, b_i(t')\, dt' \right] \phi_i. \tag{3.32}$$

From (3.32) it is easy to conclude how $\mathbf{u}(t)$ behaves for $t \to \infty$. If $Re\ (\lambda_i) < 0$ for all $i = 1, \ldots, n$, the first term, which stems from the initial condition, will converge to zero if $t \to \infty$. So, the system "forgets" its initial value, if time proceeds. This is sometimes called fading memory: after some transient period the effect of the initial state $\mathbf{u}_0$ is forgotten, and the dynamics of the system are completely determined by the effect of $\mathbf{b}(t)$, the driving force.

Exercise 3.2e.

Let us apply this concept to the damped, driven harmonic oscillator dealt with in Example 3.1a.

a. *Calculate the eigenvalues and eigenfunctions of the matrix* $\mathbf{A}$ *of the oscillator.*

b. *Find the expansion coefficients in* (3.31) *for the driving force* $\mathbf{b}(t)$.

c. *Calculate the behavior of the position* $u(t)$ *of the oscillator after the transient period, i.e., when the initial condition is forgotten, by evaluating the integrals in* (3.32).

If $\mathbf{A}$ has less than n eigenvectors, the situation is a bit more complicated. Then, the set of eigenvectors does not form a complete basis and we have to extend it. For the general case we refer to the standard literature.

Exercise 3.2f. *Symmetric matrices.*

A special case is met if the matrix $\mathbf{A}$ *is symmetric, i.e., if* $A_{ij} = A_{ji}$. *Then it holds that*

$$\mathbf{x} \cdot \mathbf{A} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{A} \cdot \mathbf{x} \quad \forall\, \mathbf{x}, \mathbf{y}.$$

a. *Show that an* $n \times n$ *symmetric matrix* $\mathbf{A}$ *has* n *eigenvalues if they are counted with the correct multiplicities.*

b. *Show that the eigenvalues of a symmetric matrix are real.*

c. *Show that eigenvectors corresponding to different eigenvalues of a symmetric matrix are orthogonal.*

From this exercise we may draw an important conclusion for symmetric matrices. Since in the subspace of all eigenvectors corresponding to the same eigenvalue we can always find an orthogonal basis, the eigenvectors of a symmetric matrix may be assumed to form an orthogonal and complete (i.e., spanning $\mathbb{R}^n$) set.

3.2.3 WKB approximation

The WKB method, named for Wentzel, Kramer, and Brillouin, is a powerful method for dealing with certain classes of linear inhomogeneous equations, i.e., for which the coefficients are not constant but vary slowly in time or space. Note that "slowly" is a relative indication. Here, it is important that the model equations are cast in dimensionless form. Then, the natural time or length scale is normally used to make the relevant time or length scale of the dimensionless variable of unit order. If the interval over which the inhomogeneous coefficient varies in unit order is long compared to the unit interval, then the inhomogeneity is referred to as being slow. The idea is to start from a quasi-homogeneous approximation, i.e., assuming at each time/position the coefficients to be constant, determining the homogeneous solutions at the current instant/position, and then gluing these solutions together such that the inhomogeneous equation is satisfied on relatively long time/space scales.

To make this more specific, consider the simple linear wave equation in one spatial dimension z with propagation velocity c:

$$\partial_t^2 u = c^2 \partial_z^2 u.$$

First, suppose that c depends only on space, so $c = c\,(z)$. Then we look for solutions that are harmonic in time: at given frequency ω solutions of the form $u(z,t) = v(z)e^{i\omega t}$ are sought. This leads to the equation

$$\partial_z^2\, v + k^2(z)v = 0 \text{ with } k(z) = \omega/c(z).$$

A specific physical situation described by this equation is light propagation through an inhomogeneous one-dimensional medium that is slowly varying. Then $v = v\,(z)$ represents the amplitude of the electric field, and $k = k\,(z)$ is related to the optical refractive index of the medium, the inverse of the propagation speed. Such a system will also be studied in Exercise 3.2i and in §6.4.3.

If, on the other hand, c depends only on time $c = c\,(t)$, we look for solutions that are periodic in space: for each k, solutions of the form $u(z,t) = w(t)e^{ikz}$ are sought. This leads to the equation

$$\partial_t^2 w + \omega(t)^2 w = 0 \text{ with } \omega(t) = k\,c(t).$$

This equation describes a harmonic oscillator with frequency that depends on time. A simple example of this is a pendulum of changing length.

Example 3.2a. Pendulum of varying length.
Consider a planar pendulum (cf., Exercise 1.3b) of which the length $\ell(t)$ varies in time. Such a pendulum has no definite period, since in general the oscillations are not periodic. However, if we choose a fixed time interval, we can determine an average period and an average length over that interval. If the length variation over an average period is small with respect to the average length of the pendulum, we call the variations slow. Let us derive the correct governing equation[1] for the angle ϕ with the vertical for a pendulum of mass m. The velocity of the pendulum mass is given by $\ell(t)\,\partial_t\,\phi$. The angular momentum p of the pendulum around its rotation center is given by

$$p = m\,\ell^2(t)\,\partial_t\phi.$$

The second law of Newton states that the temporal derivative of p equals the moment of force exerted on the pendulum mass. The latter force is the component of the vertical gravitational force $m\,g\,\sin\phi$ tangential to the velocity of the mass, and thus orthogonal to the pendulum rod, times the pendulum length. So, the equation of motion reads as

$$\partial_t\left[m\,\ell^2(t)\,\partial_t\,\phi\right] + m\,g\,\ell(t)\sin\phi = 0.$$

To cast this equation in a standard form, we introduce the new independent variable z such that $\ell^2(t)\,\partial_t := \partial_z$, i.e., $\partial z/\partial t = 1/\ell^2(t)$. Since $\partial z/\partial t > 0$, this transformation is invertible. Then, we may rewrite the equation of motion in the form

$$\partial_z^2 u + \omega^2\,(z)\sin u = 0 \;\text{ with }\; \omega^2(z) := g\ell^3(t(z)),$$

where $u(z) := \phi(t(z))$. Its linearized version

$$\partial_z^2\,u + \omega^2(z)\,u = 0 \tag{3.33}$$

describes a linear spring without damping and with varying spring constant. If the variations of ℓ are slow in time, the variations of ω as a function of z are also slow. □

Example 3.2b. Water waves above a nonflat bottom.
Another example is given by surface waves above a slowly varying bottom. See Fig. **3.4**. Here the natural length scale is the average water depth $\bar{H}$. Variations in the water depth are slow if $\Delta H/\bar{H} \ll 1$, where ΔH is the change in depth over a horizontal distance $\bar{H}$. With $c = \sqrt{gH(x)}$, where $H(x)$ is the local depth of the layer, the equation for the surface elevation reads as

$$\partial_t^2\eta - \partial_x\left[c^2\partial_x\eta\right] = 0.$$

For a derivation, see Exercise 2.9b. Looking for time-harmonic waves $\eta(x,t) = u(x)\cos(\omega t)$ and applying a transformation from x to y according to $c^2\,(x)\,\partial_x = \partial_y$, we arrive at an equation similar to (3.33). □

[1] A note and a warning about modeling: in some treatments of this problem one starts with the formulation for a pendulum with constant length in the form $\partial_t^2\phi + g/\,(m\ell)\sin\phi = 0$ and then lets ℓ depend on time; this is a different, and *wrong*, model.

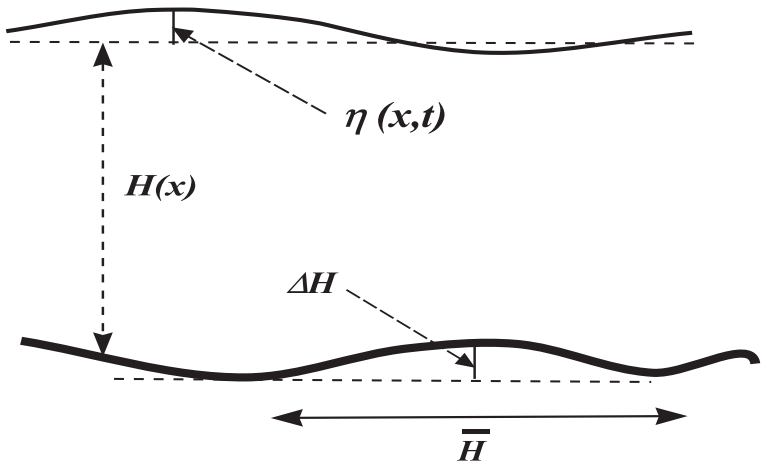

Figure 3.4. Shallow water waves above a nonflat bottom, with η denoting the water elevation above the rest level and $H(x)$ the slowly varying water depth.

Intuitive derivation of the WKB approximation

To provide the reader with a feeling of the essence of the WKB approach, we study an equation already met above:

$$\partial_z^2 u + k^2(z)u = 0.$$

Despite the simple appearance of this equation, exact solutions in closed form usually cannot be found, except for specific functions $k(z)$. However, when we assume that the variations in $k(z)$ are slow, it is possible to obtain the famous WKB approximation. As mentioned above, "slow" has meaning only if the equation is in dimensionless form. Variations are measured with respect to some natural length scale for the variable z.

To make more explicit that k depends on z in a slowly varying way, we introduce a small parameter to measure the change. That is, we write $k(z) = K(\varepsilon z)$ with $\varepsilon > 0$ small. We will use the notation $\mathcal{E}$ to denote the operator defined by

$$\mathcal{E}(u) := \partial_z^2 u + K^2(\varepsilon z)u = 0. \tag{3.34}$$

The slow variation assumption implies that on intervals of unit length the change in k is small. But note that on longer intervals, e.g., of length $\mathcal{O}(1/\varepsilon)$, the change in k will be of order one, and thus not small. The essence of the WKB method is that the error in the approximation uniformly vanishes for $\varepsilon \to 0$ (usually proportional to ε) on very long intervals, namely, of length $\mathcal{O}(1/\varepsilon)$.

A first suggestion for an approximate solution is found by looking for an approximation of the form

$$v = A\, e^{iK(\varepsilon z)z}. \tag{3.35}$$

Substituting this Ansatz in (3.35), we obtain that the error, also called the *residue*, is given by

$$\mathcal{E}(v) = \varepsilon[2iAK' - 2AK'zK]e^{iKz} + \mathcal{O}(\varepsilon^2)\,. \tag{3.36}$$

Here and in the following a prime denotes differentiation with respect to the argument of the function under consideration. Result (3.36) is reliable but not very surprising. It expresses that on a bounded and fixed interval the approximate solution continuously depends on the parameter ε. This was to be expected, since this continuous dependence result generally holds for ODEs [22]. However, we are interested in a better result, namely, in an approximation that is of the order ε on an interval of length $1/\varepsilon$. The residue (3.36) of the proposed approximation (3.35) does not have this property in view of the presence of the factor εz in the residue.

An improvement is obtained by looking for an approximation of the form

$$w = Ae^{i\theta(z)} \tag{3.37}$$

with θ a phase factor to be determined. This Ansatz leads to the residue

$$\mathcal{E}(w) = Ae^{i\theta}[i\theta'' - \theta'^2 + K^2].$$

If one chooses

$$\theta' := \partial_z\theta = \pm K(\varepsilon z)\ , \tag{3.38}$$

then $\theta'' = \pm\varepsilon\, K'' = \mathcal{O}(\varepsilon)$ and the residue is of order ε on the long z-interval. Integrating this simple condition we find that

$$\theta - \theta_0 = \int_0^z K(\varepsilon z')dz' = \frac{1}{\varepsilon}\int_0^{\varepsilon z} K(\zeta)\, d\zeta.$$

To compare the two different choices (3.35) and (3.37) for the phases, observe that the difference can be written as

$$\theta(z) - K(\varepsilon z)\, z = \frac{1}{\varepsilon}\int_0^{\varepsilon z} [K(\zeta) - K(\varepsilon z)]\, d\zeta.$$

This difference is small if $z \in [0, 1]$, since then $\theta(z) - K(\varepsilon z)\, z = \mathcal{O}(\varepsilon)$. However, for longer intervals the difference may become of order one.

The WKB approximation is a further improvement by allowing also the amplitude A to change slowly. Now the Ansatz

$$U(z) = A(\varepsilon z)\, e^{i\theta(z)}$$

is used. It yields the residue

$$\mathcal{E}(U) = e^{i\theta}[i\,A\theta'' + 2\varepsilon i\,A'\theta' - A\,\theta'^2 + A\,K^2 + \mathcal{O}(\varepsilon^2)]\ . \tag{3.39}$$

With the above choice for θ, the order ε term in the right-hand side can be made to vanish by letting A satisfy

$$A' = -\frac{-\theta''}{2\varepsilon\theta'}A = -\frac{K'}{2K}A\ ,$$

where we use (3.38).

This has as solution

$$A = \frac{A_0}{\sqrt{K}} .$$

Resuming, we have the following result.

Proposition 3.2a. WKB approximation.
An asymptotically valid solution of the equation

$$\partial_z^2 u + K^2(\varepsilon z)u = 0 \tag{3.40}$$

with slowly varying coefficient $K = K(\varepsilon z)$ is given by the WKB approximation

$$\tilde{u}(z) = \frac{1}{\sqrt{K(\varepsilon z)}} e^{i\theta(z)} \text{ with } \partial_z \theta(z) = K(\varepsilon z). \tag{3.41}$$

This approximation satisfies the equation to order $\mathcal{O}(\varepsilon^2)$, since

$$\mathcal{E}(\tilde{u}(z)) = \varepsilon^2 e^{i\theta(z)} \frac{d^2}{d\zeta^2} \left(\frac{1}{\sqrt{K(\zeta)}} \right), \quad \zeta = \varepsilon z, \tag{3.42}$$

uniformly on z-intervals of length of order $\mathcal{O}(1/\varepsilon)$.

Exercise 3.2g.
Check that the ε^2-term in (3.39) *is indeed given by expression* (3.42)*.*

Higher-order WKB approximations
The result above can be obtained in a different way from a coordinate transformation. At the same time we will find that it leads to a complete hierarchy of expressions that approximate the solution in increasing order of accuracy [12], provided that the inhomogeneous coefficient is sufficiently smooth.

Consider the original equation $\mathcal{E}(u) := \partial_z^2 u + K^2(\varepsilon z)\, u = 0$. Now define, motivated by the WKB approximation found above, a new variable

$$\theta := \int_0^z K(\varepsilon z')\, dz',$$

and introduce the transformation

$$B(\theta) := \sqrt{K(\varepsilon z)}\, u(z).$$

Then the following equation for B is found:

$$K^{3/2} \left\{ B'' + \left[1 - \varepsilon^2 \left\{ \frac{K''}{2K^3} - \frac{3K'^2}{4K^4} \right\} \right] B(\theta) \right\} = 0, \tag{3.43}$$

and so

$$B'' + \left[1 - \varepsilon^2 \left\{ \frac{K''}{2K^3} - \frac{3K'^2}{4K^4} \right\} \right] B(\theta) = 0. \tag{3.44}$$

Exercise 3.2h.
Check the result in (3.43).

Result (3.44) leads to some important observations and results:

1. If we neglect the terms of order ε^2, we find the simple equation

$$B'' + B(\theta) = 0,$$

which has the obvious solution

$$B = A_0\, e^{i\theta}.$$

So, the first-order WKB approximation is recovered.

2. Equation (3.44) with $\varepsilon \neq 0$ resembles the original equation (3.34), in which z is replaced with θ and the inhomogeneous function K^2 is replaced with

$$1 - \varepsilon^2 \left\{ \frac{K''}{2K^3} - \frac{3K'^2}{4K^4} \right\}.$$

This term is slowly varying in θ. Hence one can apply the WKB approximation to (3.44). This then leads to a residue of order $\mathcal{O}(\varepsilon^4)$.

3. It is also possible to apply the WKB transformation once more and to repeat the process to deal with inhomogeneities that are slowly varying and have smaller and smaller deviations. In this way a whole hierarchy of WKB approximations can be obtained, leading to increasingly better results. Of course, a condition for the possibility of repeating this process is the sufficient differentiability of the function K. For infinitely often differentiable K, the residue can be made arbitrarily small.

Exercise 3.2i. *Transmission-reflection in slowly varying media.*
Let us consider the optical problem of a slowly varying medium. We write the governing equation for the electric field as

$$\partial_z^2 E + k^2(z)\, E = 0 \;\; \textit{with} \;\; k = \frac{\omega\, n}{c},$$

where n is the index of refraction, a slowly varying function. More specifically, we will consider a function k that has the form of a ramp: it connects two constant levels, say, a smooth transition from level k_1 for $z \leq 0$, and a level k_2 for $z \geq L$ with $L \gg 1$. See Fig. **3.5**. *As a small parameter we choose $\varepsilon = 1/L$. Note that at $z = 0$ and $z = L$, discontinuities in the first derivative of k exist.*

a. *Write the WKB approximation that is correct up to $O(\varepsilon)$ on the whole real line and check the continuity.*

b. *Calculate the first derivative of this function with respect to z and show that it is continuous only up to $\mathcal{O}(1)$.*

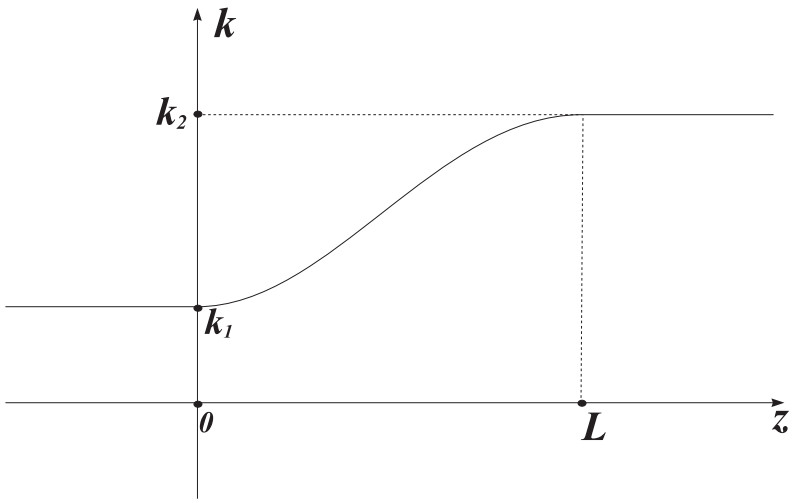

Figure 3.5. Smoothly varying behavior of the coefficient function $k(z)$.

As will be discussed extensively in §3.3.4*, a solution of the form* $e^{i(k\,z-\omega\,t)}$ *can be interpreted as a mode traveling to the right. In a nonuniform part of the medium, such a wave will be partly reflected, partly transmitted. The scattering problem is the problem of finding for* $z \leq 0$ *a solution consisting of an incoming wave (traveling to the right) of a given amplitude, plus a reflected wave (traveling to the left) with unknown amplitude* r *(reflected) and for* $z \geq L$ *a transmitted wave (traveling to the right) of an unknown amplitude* t*. This means that the solution sought is a function* E *that satisfies for some values of* r *and* t

$$E(z) = e^{ik_1 z} + re^{-ik_1 z} \quad \textit{for} \quad z \leq 0,$$
$$E(z) = t\, e^{ik_2 z} \qquad\quad \textit{for} \quad z \geq L,$$

while in the interval $(0, L)$ *the WKB approximation is used for a mode that travels to the right and one that travels to the left so that the total solution is continuous and once differentiable, also in* $z = 0$ *and* $z = L$*.*

c. *Write the WKB approximation in the interval* $(0, L)$ *and determine the amplitudes of the modes.*

d. *If one calculates the reflection coefficient, one finds that it depends on the discontinuities:*

$$r = \frac{1}{4}\frac{1}{L}\left\{\frac{k'(0^+)}{k(0)^2} - e^{2i\theta(L)}\,\frac{k'(L^-)}{k(L)^2}\right\} + \mathcal{O}\left(\frac{1}{L^2}\right).$$

Here, $k'(0^+)$ *denotes the right-hand side derivative of* k *with respect to its argument, and similarly for* $k'(L^-)$*. Investigate this formula by dimensional analysis. Observe that a smooth differentiable index function* k *has higher-order reflection, namely,* $r = \mathcal{O}(\frac{1}{L^2})$*.*

e. *To find the solution correct up to* $\mathcal{O}(\varepsilon^3)$*, and so with* $\partial_z E$ *correct up to* $\mathcal{O}(\varepsilon^2)$*, a higher-order WKB-approximation is required, e.g., by taking the WKB approximation of the*

transformed WKB equation. This improved result looks like

$$r = \frac{1}{4}\frac{1}{L^2}\left\{\frac{k''(0)}{k(0)^\alpha} - e^{i\varphi}\,\frac{k''(L)}{k(L)^\alpha}\right\} + \mathcal{O}\left(\frac{1}{L^3}\right)$$

for some exponent α and phase factor φ. Determine α from dimensional analysis. Derive the formula.

3.3 PDEs

The simplest PDE systems are linear and autonomous. They have the form

$$\partial_t \mathbf{u} = \mathbf{F}(\mathbf{u}) \tag{3.45}$$

with $\mathbf{F}$ a linear operator, which implies that the boundary conditions are homogeneous. An example of such a linear $\mathbf{F}$ is given in (3.15). To solve linear PDEs, standard methods are available. That is why one often tries to reduce a nonlinear system by an approximating linear one. We shall deal with that procedure in the next subsection. Just as for ODEs, expansions in terms of appropriate basis functions are also highly useful for PDEs. Of the existing methods, Fourier series and integrals are probably most well known. One should be aware that the Fourier approach is only one member of a whole class of methods. In §3.3.2 we outline the main principles of such expansions in some detail.

3.3.1 Linearizing PDEs

Analogous to linearization of ODEs, we may linearize a PDE around a stationary state. However, this requires the introduction of a new concept. For ODEs, the vector field $\mathbf{F}$ is a function of the state vector and its derivative is given by the Jacobi matrix defined in (3.28). In the case of PDEs, $\mathbf{F}$ is an operator acting on states $\mathbf{u}$ which are functions themselves. We meet with the question of how $\mathbf{F}(\mathbf{u})$ varies if $\mathbf{u}$ varies. (The analogue of "derivative" of a function is in case of an operator the "directional derivative" or "variational derivative" $\mathbf{F}'$ of $\mathbf{F}$.) For given functions $\mathbf{u}(\mathbf{x})$ and $\mathbf{v}(\mathbf{x})$, the definition of $\mathbf{F}'$ reads as

$$\mathbf{F}'(\mathbf{u};\mathbf{v}) = \lim_{\varepsilon\to 0}\frac{\mathbf{F}(\mathbf{u}+\varepsilon\mathbf{v}) - \mathbf{F}(\mathbf{u})}{\varepsilon}\,. \tag{3.46}$$

This derivative itself is an operator. It depends not only on $\mathbf{u}$, the state at which the derivative is taken, but also on $\mathbf{v}$, the direction in which the state $\mathbf{u}$ is perturbed. To get acquainted with this concept, let us look at some simple examples.

Exercise 3.3a.

For convenience, we consider the one-dimensional case, when $\mathbf{F}$ and $\mathbf{u}$ are scalar. Calculate the derivative $F'(u;v)$ of F by applying definition (3.46) *in the following cases:*

a. $F(u) = u$,

b. $F(u) = u^n$, $n = 2, 3, \ldots$,

c. $F(u) = \partial_x u$,

d. $F(u) = \partial_x^n u$, $n = 2, 3, \ldots$,

e. $F(u) = \sin u$,

f. $F(u) = \exp(u)$,

g. $F(u) = \exp(\partial_x u)$.

This exercise shows that in many cases $\mathbf{F}'(\mathbf{u}; \mathbf{v})$ is linear in $\mathbf{v}$. Then, we may write

$$\mathbf{F}'(\mathbf{u}; \mathbf{v}) = \mathbf{F}'(\mathbf{u}) \cdot \mathbf{v}. \tag{3.47}$$

If this property holds (and if this operator is bounded), we call $\mathbf{F}'$ the *Frechet derivative* of $\mathbf{F}$. Using this concept, we may linearize a general nonlinear PDE $\partial_t \mathbf{u} = \mathbf{F}(\mathbf{u})$ around a given state. So, let us consider a stationary profile $\mathbf{u}_0(\mathbf{x})$, for which

$$\mathbf{F}(\mathbf{u}_0) = \mathbf{0}.$$

We are interested in what happens with the solution if we choose an initial profile close to $\mathbf{u}_0$. To that end we start at $\mathbf{w}_0 = \mathbf{u}_0 + \varepsilon \mathbf{v}_0$ with $\varepsilon \ll 1$ and $||\mathbf{v}_0|| = 1$. So, $\mathbf{w}_0$ is close to $\mathbf{u}_0$. The sum $\mathbf{u}_0(\mathbf{x}) + \varepsilon \mathbf{v}_0(\mathbf{x})$ is a function in the state space U, so it satisfies the boundary conditions. Note that $\mathbf{u}_0(\mathbf{x})$ alone already satisfies these conditions. This implies that the perturbation $\mathbf{v}_0(\mathbf{x})$ satisfies homogeneous boundary conditions.

We know that the solution $\mathbf{w}(t)$ starting at $\mathbf{w}_0$ will remain in the vicinity of $\mathbf{u}_0$, at least for some period after the initial time. So, initially the difference $\mathbf{v}(t) = \mathbf{w}(t) - \mathbf{u}_0$ satisfies the linear approximation

$$\partial_t \mathbf{v} = \mathbf{F}'(\mathbf{u}_0) \cdot \mathbf{v}, \tag{3.48}$$

provided that $\mathbf{F}$ has property (3.47). This is the *linearization* of the nonlinear PDE $\partial_t \mathbf{u} = \mathbf{F}(\mathbf{u})$ around $\mathbf{u}_0$. Note the similarity of (3.29) for ODEs and (3.48) for PDEs.

Exercise 3.3b.

We consider a nonlinear version of the model (3.15) *in Example* 3.1c *for the modeling of algae dynamics. To that end we take only two effects into account and let the coefficients depend on u itself:*

$$\partial_t u = u^2 \, \partial_x u + u \, \partial_x^2 u \; . \tag{3.49}$$

The constant (in time) and uniform (in space) state $u := 1$ *is a stationary state of this system. Linearize the system around this stationary state.*

3.3.2 Expansions in basis functions

Just as for ODEs, expanding the solution of a PDE in terms of basis functions can be of great help in solving the PDE. In the case of a linear PDE with constant coefficients, a particularly convenient choice is to expand in terms of the eigenfunctions of the spatial operator. For convenience we take here for the spatial domain Ω a finite, one-dimensional interval and consider the scalar case, so the solution u and the operator F are scalar. Let us consider the

space $\mathcal{L}^2(\Omega)$ of (possibly complex valued) functions $u(x)$ that are square integrable on the interval Ω, thus for which it holds that

$$\int_\Omega |u|^2(x)\,dx < \infty. \tag{3.50}$$

We assume these functions satisfy homogeneous boundary conditions. Note that this space is quite wide. Its elements may be discontinuous and even singular (diverging to ∞) in isolated points, provided that these singularities are such that they do not destroy the integrability condition (3.50). In $\mathcal{L}^2(\Omega)$ a natural inner product of two functions u and v is defined as

$$(u,\ v) = \int_\Omega u(x)\,v^*(x)\,dx, \tag{3.51}$$

where v^* is the complex conjugate of v.

Orthonormal basis functions

In Exercise 3.2f we pointed out that for ODEs the case of a symmetric matrix is quite advantageous, since the eigenvectors then constitute a natural orthogonal basis set. For the PDE a similar situation is met if the spatial operator F in (3.45) is *self-adjoint*, since then the corresponding eigenfunctions are orthogonal, as we shall show. The operator F is self-adjoint if it has the property that

$$(F\,u,\ v) = (u,\ F\,v) \tag{3.52}$$

for any pair $u, v \in \mathcal{L}^2(\Omega)$.

Exercise 3.3c.

Consider the operator

$$F = -\partial_x^2 \tag{3.53}$$

acting on scalar functions $u(x)$, $x \in [0, 2\pi]$ that satisfy homogeneous boundary conditions $u(0) = u(2\pi) = 0$ and are two times differentiable.

a. *Show that F is self-adjoint on this function space. Apply partial integration and make use of the homogeneous boundary conditions.*

b. *Is F still self-adjoint if we take as boundary conditions $u(0) = 0$ and $\partial_x u(2\pi) = 0$? Consider also the case that $\partial_x u(0) = \partial_x u(2\pi) = 0$.*

The eigenfunctions $\varphi_\lambda(x)$ and eigenvalues λ of the operator F satisfy the eigenvalue equation

$$F\varphi_\lambda = \lambda\,\varphi_\lambda. \tag{3.54}$$

Furthermore, they have to satisfy the boundary conditions under consideration. If F is self-adjoint, we have that

$$(F\,\varphi_\lambda, \varphi_\lambda) = \lambda(\varphi_\lambda, \varphi_\lambda) = (\varphi_\lambda, F\,\varphi_\lambda) = \lambda^*(\varphi_\lambda, \varphi_\lambda). \tag{3.55}$$

So, these eigenvalues are real: $\lambda \in \mathbb{R}$. Also the eigenfunctions $\varphi_\lambda(x)$ have nice properties. Since

$$(F\,\varphi_\lambda, \varphi_{\lambda'}) - (\varphi_\lambda, F\,\varphi_{\lambda'}) = (\lambda - \lambda')\,(\varphi_\lambda, \varphi_{\lambda'}) = 0, \tag{3.56}$$

we conclude that eigenfunctions corresponding to different eigenvalues are orthogonal:

$$(\varphi_\lambda, \varphi_{\lambda'}) = 0, \quad \lambda \neq \lambda'. \tag{3.57}$$

If to one eigenvalue several eigenfunctions correspond, we can always find an orthogonal basis spanning the subspace of these eigenfunctions. So, self-adjoint operators give rise to a set of eigenfunctions that forms an orthogonal basis in $\mathcal{L}^2(\Omega)$. So, any function in this space can be written as a linear combination of the φ_λ. This property forms the basis of, among others, Fourier theory, which is the subject of the next example.

Example 3.3a. Fourier series.
Fourier series are nothing more than expansions in terms of the eigenfunctions of the operator $F = -\partial_x^2$. Which eigenfunctions are used depends on the boundary conditions. If we take for Ω the interval $[0, 2\pi]$ and homogeneous Dirichlet boundary conditions, the eigenfunctions are solutions of

$$\begin{cases} -\partial_x^2 \varphi = \lambda \varphi, \\ \\ \varphi_\lambda(0) = \varphi_\lambda(2\pi) = 0. \end{cases} \tag{3.58}$$

This yields a set of discrete eigenvalues

$$\lambda_n = \left(\frac{n}{2}\right)^2, \quad n = 1, 2, 3, \ldots,$$

with corresponding eigenfunctions

$$\varphi_n(x) = \sin\left(\frac{1}{2} n x\right). \qquad \square$$

Exercise 3.3d.
Derive the eigenvalues and eigenfunctions of the operator $-\partial_x^2$ if we take as boundary conditions

a. $\varphi_\lambda(0) = \partial_x \varphi_\lambda(2\pi) = 0$,
b. $\partial_x \varphi_\lambda(0) = \partial_x \varphi_\lambda(2\pi) = 0$.

The set of eigenvalues of an operator is called its *spectrum*. As seen above, the spectrum and corresponding eigenfunctions $\varphi_n(x)$ depend not only on the form of the operator F but also on the boundary conditions. As shown in Example 3.3a, a finite interval Ω leads to a discrete spectrum λ_n with n running over integers. It is convenient to normalize the eigenfunctions $\varphi_n(x)$ such that $(\varphi_n, \varphi_n) = 1$. Their orthonormality is then expressed by

$$(\varphi_n, \varphi_m) = \delta_{n,m} \tag{3.59}$$

with $\delta_{n,m}$ the Kronecker delta defined as $\delta_{n,m} = 1$ if $n = m$ and $\delta_{n,m} = 0$ if $n \neq m$. After normalization, the φ_n's form an orthonormal basis in $\mathcal{L}^2(\Omega)$. Every function in this space can be written as

$$u(x) = \sum_{n \in N} c_n \varphi_n(x). \tag{3.60}$$

An advantage of working with an orthonormal basis is that the coefficients c_n are simply given by the inner products:

$$c_n = (u, \varphi_n). \tag{3.61}$$

In the following heat diffusion problem we show how powerful the application of expansions in basis functions is. □

Example 3.3b. Heat diffusion in a rod.
Let us consider heat diffusion in a long, thin rod of length L. See Fig. **3.6**. The governing equation was derived in §2.3. The temperature $u(x, t)$ in the rod satisfies the linear PDE

$$\partial_t u = \kappa\, \partial_x^2 u. \tag{3.62}$$

At one end, where $x = 0$, we keep the temperature at a fixed level u_0 by connecting the rod to a heat reservoir which has a very large thermal conductivity and a huge heat capacity. At the other end, where $x = L$, we prescribe the heat flux. So, the boundary conditions are

$$u(0, t) = u_0, \quad -\kappa\, \partial_x u(L, t) = Q_0, \quad t \geq 0. \tag{3.63}$$

The stationary state $u^s(x)$ is easily found from the condition $\partial_x^2 u^s = 0$. It is a linear function of x, given by

$$u^s(x) = -\frac{Q_0}{\kappa} x + u_0.$$

This problem has nonhomogeneous boundary conditions. However, if we consider the difference

$$v(x, t) := u(x, t) - u^s(x), \tag{3.64}$$

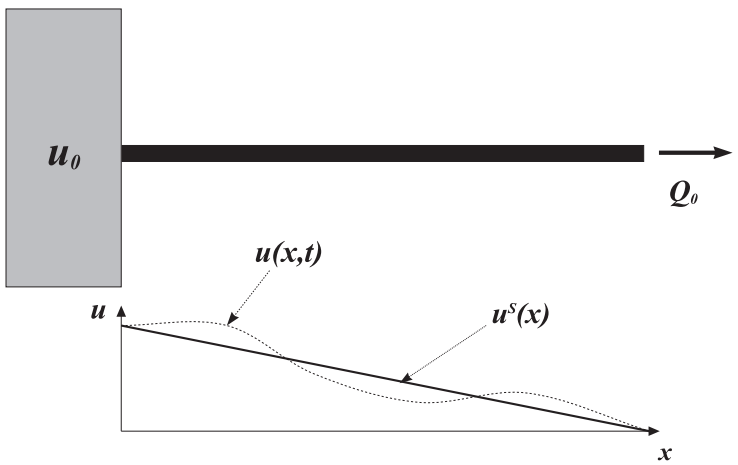

Figure 3.6. Heat diffusion in a long rod, with a fixed temperature at one end and a fixed outgoing flux at the other end.

we observe that v satisfies the original PDE (3.62) but now with homogeneous boundary conditions. So,

$$\begin{cases} \partial_t v = \kappa\, \partial_x^2 v, \\ \\ v(0,t) = 0, \quad -\kappa\, \partial_x v(L,t) = 0, \quad t \geq 0. \end{cases} \tag{3.65}$$

Now we may apply the procedure of expansion in basis functions outlined above. The eigenvalue problem

$$\begin{cases} -\partial_x^2 \varphi = \lambda\, \varphi, \\ \\ \varphi(0) = \partial_x\, \varphi(L) = 0 \end{cases} \tag{3.66}$$

leads to a discrete set of eigenvalues

$$\lambda_n = \left[\frac{(n+\frac{1}{2})\pi}{L} \right]^2, \quad n = 0, 1, 2, \ldots$$

with corresponding eigenfunctions

$$\varphi_n(x) = \sqrt{\frac{2}{L}}\, \sin\left[\frac{(n+\frac{1}{2})\pi x}{L} \right]. \tag{3.67}$$

The prefactor is chosen such that the φ_n are normalized. Since the operator in (3.66) is self-adjoint, the φ_n are orthogonal. By the way, this orthonormality also can be directly checked from the explicit representation (3.67).

We now use that the functions $\varphi_n(x)$ form a basis set in the space $\mathcal{L}^2[0, L]$. So, assuming $v(x, t)$ defined in (3.64), to be in this space, we may write the expansion

$$v(x,t) = \sum_{n=0}^{\infty} c_n(t)\, \varphi_n(x). \tag{3.68}$$

Note that the time dependence of v is contained in the coefficients $c_n(t)$, whereas the spatial dependence is represented by the $\varphi_n(x)$. This is why this technique is also referred to as *separation of variables*. The $c_n(t)$'s follow from the requirement that this expansion must satisfy (3.62):

$$\sum_{n=0}^{\infty} \partial_t\, c_n(t)\, \varphi_n(x) = \kappa \sum_{n=0}^{\infty} c_n(t)\, \partial_x^2\, \varphi_n(x) = -\kappa \sum_{n=0}^{\infty} c_n(t)\, \lambda_n\, \varphi_n(x). \tag{3.69}$$

Taking on both sides the inner product with one specific basis function, we find that each $c_n(t)$ satisfies the ODE

$$\partial_t c_n(t) = -\kappa\, \lambda_n\, c_n(t). \tag{3.70}$$

The step from (3.69) to (3.70) is referred to as *projecting out* the nth basis function and relies on the orthogonality property (3.59). We conclude that the $c_n(t)$, $n = 0, 1, 2, \ldots$, are given by

$$c_n(t) = c_n(0)\, e^{-\kappa \lambda_n t}.$$

The solution of (3.62) is thus

$$u(x,t) = u^s(x) + \sum_{n=0}^{\infty} c_n(0)\, e^{-\kappa\lambda_n t}\, \varphi_n(x). \tag{3.71}$$

Note the similarity and the difference between this expansion in eigenfunctions for a linear PDE and the expansion in eigenvectors (3.32) for a linear ODE. Since in (3.71) infinitely many basis functions are involved, the PDE case is sometimes referred to as being infinitely dimensional.

For completeness, we mention that the coefficients $c_n(0)$ follow from the initial temperature profile $v_0(x) := u(x,0) - u^s(x)$. Setting $t = 0$ we find

$$v_0(x) = \sum_{n=0}^{\infty} c_n(0)\, \varphi_n(x).$$

Again, the projection procedure can be applied. It leads to

$$c_n(0) = (v_0, \varphi_n). \qquad \square$$

Example 3.3b makes clear that the solution of a linear PDE with constant coefficients can be conveniently expressed in terms of the spectrum and eigenfunctions of the spatial operator. This requires the solution of a linear eigenvalue problem. For many operators these solutions are explicitly known so that analytical expressions are available in terms of infinite series. In practice these series expansions are useful mainly if the convergence is reasonably fast so that in the evaluation one can restrict oneself to a limited number of terms. The convergence is dominated by the behavior of the exponentials in (3.71). If, for fixed value of t, the eigenvalues λ_n increase fast with increasing n, the exponentials $\exp(-\kappa\lambda_n t)$ decay very fast as functions of n. Then, only the lower n-values need to be taken into account.

Exercise 3.3e.

We again consider the algae dynamics in a tube of length L modeled in Example 3.1c. *Let us assume that there is no convection in the tube. Then, the algae dynamics is governed by the PDE*

$$\partial_t u = c_0 + c_1\, u + c_2\, \partial_x u \tag{3.72}$$

with c_0, c_1, and c_2 constants. The boundary conditions are

$$u(0,t) = 0, \quad \partial_x u(L,t) = 0 \quad \forall t. \tag{3.73}$$

Find an explicit expression for the solution $u(x,t)$ by applying an expansion in basis functions in a similar way as done in Example 3.3b. *Investigate the behavior of the population for $t \to \infty$ as a function of the coefficients c_0, c_1, and c_2. Interpret these findings.*

A much more extended application of this technique is given in §6.1, where the dynamical behavior of dissolved polymer molecules is modeled.

3.3.3 Infinite spatial intervals

If the spatial interval Ω extends to infinity, the approach followed above for finite Ω no longer applies. In the past it took some time and effort to find out how both cases can be treated in a way that preserves the analogies. In Fourier theory, it leads to Fourier integrals instead of Fourier series. In this subsection we can give the reader only a flavor of the ideas involved. First we study what happens if the length of Ω becomes longer and longer.

Let us consider the spectrum of the operator $-\partial_x^2$ on the interval $\Omega = [-L, L]$. We know that the eigenvalue problem

$$-\partial_x^2 \varphi_\lambda = \lambda \varphi_\lambda \tag{3.74}$$

has sines and cosines as solutions. It is convenient to deal with both types of solutions at once by introducing complex notation. For convenience, we write $\lambda = k^2$. The eigenfunctions are then given by $\varphi_k(x) = \exp(ikx)$. For these complex functions we cannot simultaneously impose boundary conditions on both the real and the imaginary parts. To handle these parts on the same footing, it is common use to prescribe so-called periodic boundary conditions: $\varphi_k(-L) = \varphi_k(L)$. This directly leads to the eigenfunctions

$$\varphi_n(x) = e^{ik_n x}\,, \quad k_n = \frac{n\pi}{L}\,, \quad n = 0, \pm 1, \pm 2, \ldots\,. \tag{3.75}$$

They form a basis set of the square integrable, complex-valued functions on the interval $[-L, L]$. The inner product in this space is given by (3.51). For the $\varphi_n(x)$ we have that

$$(\varphi_n, \varphi_{n'})_L = \int_{-L}^{L} e^{i(k_n - k_{n'})x} dx.$$

If $n = n'$, this yields $(\varphi_n, \varphi_n) = 2L$, so the φ_n can be normalized. For $n \neq n'$ we find

$$(\varphi_n, \varphi_{n'})_L = 2\frac{\sin[(k_n - k_{n'})L]}{k_n - k_{n'}} = 2L\frac{\sin[(n - n')\pi]}{(n - n')\pi} = 0. \tag{3.76}$$

So, the φ_n are also orthogonal, as expected. The interesting point is what happens if we take the limit $L \to \infty$. Then, the discrete spectrum in (3.75) converges to the continuous spectrum $\{k \in \mathbb{R}\}$, and the eigenfunctions then are

$$\varphi_k(x) = \frac{1}{\sqrt{2\pi}}\, e^{ikx}\,, \quad k \in \mathbb{R}. \tag{3.77}$$

Here, we introduce a normalization factor for later convenience. However, in the limit $L \to \infty$ these basis functions cannot be normalized in the usual sense.

The delta function

To settle this normalization problem Dirac introduced the so-called delta function $\delta(x)$. This is not a function but an operator, defined by specifying its action when applied to a

well-defined set of test functions. Such an operator is sometimes referred to as a *generalized function*. The action of the delta function $\delta(x)$ is defined by the following:

a. *When $\delta(x)$ is applied to a continuous function $u(x)$, and the result is integrated over an arbitrary interval I containing the origin, the resulting value is equal to $u(0)$. We call this the* filter property *of the delta function. In formula,*

$$(\delta, u) := \int_I \delta(x)\, u(x) dx = u(0). \tag{3.78}$$

b. *In particular, when the delta function is applied to the function $u(x) = 1$, and when we take $I := \mathbb{R}$, then it holds that*

$$\int_{\mathbb{R}} \delta(x)\, dx = 1. \tag{3.79}$$

The delta function has many representations; see, e.g., [28]. A convenient approach consists of a limiting procedure, in which one constructs a so-called delta sequence, a sequence of functions $d(x, L)$, depending on a parameter L. The procedure is then as follows:

- Multiply the test function with $d(x, L)$.
- Integrate the product over an arbitrary interval containing the origin. The result will depend on L.
- Take the limit $L \to \infty$.

An example of such a delta sequence is

$$d(x, L) = \frac{1}{\pi} \frac{\sin(x\, L)}{x}. \tag{3.80}$$

For large values of L, these functions show intensely oscillating behavior, and this causes the contribution to the integral from any interval not including the origin to vanish. One often meets the statement that "a delta function vanishes outside the origin." Note that its definition is much more subtle. Moreover, the elements of a delta sequence generally do not vanish outside the origin.

Another representation is the fundamental solution of the heat equation, which we already encountered in §1.5.2. It is given by

$$d(x, L) = \frac{1}{\sqrt{\pi L}}\, e^{-\frac{x^2}{L}}\ . \tag{3.81}$$

Exercise 3.3f.

a. *Prove that* (3.80) *indeed forms a delta sequence. To that end, substitute $y := x\, L$ in the integrals*

$$\frac{1}{2\pi} \int_{-a}^{b} \frac{\sin(x\, L)}{x} u(x)\, dx$$

with $a, b > 0$, and take the limit $L \to \infty$.

b. *Show that the "tent" functions*

$$d(x, L) = \begin{cases} L\,|1 - x\,L|\,, & |x| \le \dfrac{1}{L}\,, \\ 0 & \textit{otherwise}, \end{cases}$$

form a delta sequence.

c. *Show that the functions defined in* (3.81) *form a delta sequence.*

From comparing (3.76) and (3.80) we find that the inner product $(\varphi_n, \varphi_{n'})_L$ thus acts as a delta sequence with the variable not x but $k_n - k_{n'}$. Therefore, in the limit $L \to \infty$ we may write

$$(\varphi_k, \varphi_{k'}) = \delta(k - k')\,. \tag{3.82}$$

So, in the limit $L \to \infty$, normalization (3.59) in terms of the Kronecker delta is replaced with normalization (3.82) in terms of the delta function.

3.3.4 Plane waves

From the preceding section we learn that the complex exponentials $\varphi_k(x)$ in (3.77) are appropriate basis functions to expand square integrable functions. They usually are referred to as *plane waves*. For any function $u(x) \in \mathcal{L}^2(\mathbb{R})$ we may write the Fourier integral

$$u(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{u}(k)\, e^{ikx}\, dk. \tag{3.83}$$

The coefficient function $\hat{u}(k)$ is the *Fourier transform* of $u(x)$ and given by the inner product

$$\hat{u}(k) = (u, \varphi_k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{u}(k) e^{-ikx}\, dk. \tag{3.84}$$

Let us show how a linear PDE can be solved by use of the Fourier representation (3.83). We assume F to have a simple form, namely,

$$F = c_0 + c_1 \partial_x + c_2 \partial_x^2 + \cdots + c_N \partial_x^N := p(\partial_x).$$

So, p—sometimes called the *symbol* of the differential operator F—is a polynomial in ∂_x of degree N with constant coefficients $c_0, \ldots, c_N$. The action of F on a basis function is then given by

$$F\, e^{ikx} = p(ik)\, e^{ikx}\,,$$

since taking the derivative of a plane wave with respect to x is just multiplication by the factor ik.

If we assume that the solutions $u(x, t)$ of $\partial_t u = F(u)$ are square integrable with respect to their spatial dependence, they can be written as a Fourier integral

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{u}(k, t)\, e^{ikx} dk. \tag{3.85}$$

The time dependence is incorporated in the transform $\hat{u}(k,t)$, since the basis functions are time independent. Substitution of (3.85) into the PDE then yields

$$\int_{-\infty}^{+\infty} \partial_t \hat{u}(k,t)\, e^{ikx} dk = \int_{-\infty}^{+\infty} \hat{u}(k,t)\, p(ik)\, e^{ikx} dk.$$

By multiplying both sides with $e^{ik'x}$ and integrating over x we project out one basis function, using the orthonormality of the basis functions. Eventually we find that the time evolution of $\hat{u}(k,t)$ is governed by the ODE

$$\partial_t \hat{u}(k,t) = p(ik)\, \hat{u}(k,t), \tag{3.86}$$

which has the solution

$$\hat{u}(k,t) = \hat{u}(k,0)\, e^{p(ik)t}. \tag{3.87}$$

It is convenient to introduce the notation

$$\omega(k) = i\ p(ik). \tag{3.88}$$

The dependence of ω on k is the so-called dispersion relation. This term is easily understood from substituting (3.87) in (3.85). This yields

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{u}(k,0)\, e^{i(kx-\omega(k)t)}\, dk. \tag{3.89}$$

Note that according to (3.84), the function $\hat{u}(k,0)$ is the Fourier transform of the initial profile $u(x,0)$.

From the derivation of (3.89) we conclude that any linear PDE in one dimension can be solved by writing its solution $u(x,t)$ as a linear combination of the functions

$$\varphi_k(x,t) := \frac{1}{\sqrt{2\pi}} e^{i(kx-\omega(k)t)}, \quad k \in \mathbb{R}, \tag{3.90}$$

where $\omega(k)$ is characteristic for the operator F under consideration. These functions are referred to as *monochromatic modes* or just *modes*. They are periodic in space, since k is real. The spatial period or *wave length* is given by

$$\lambda = \frac{2\pi}{k}.$$

We call k the *wave number* and $\omega(k)$ the *frequency* of $\varphi_k(x,t)$. The frequency is in general complex, and we may split it into its real and imaginary parts:

$$\omega(k) = \omega^{re}(k) + i\, \omega^{im}(k).$$

From (3.90) we see that a mode with wave number k has a so-called phase velocity v_{ph} given by

$$v_{ph} = \frac{\omega^{re}(k)}{k}. \tag{3.91}$$

Note that for $k > 0$, the direction of velocity v_{ph} is to the right if $\omega^{re}(k) > 0$ and to the left if $\omega^{re}(k) < 0$. With this notation the modes may be written as

$$\varphi(x,t) = \frac{1}{2\sqrt{2\pi}} e^{\omega^{im}(k)t}\ e^{ik(x-v_{ph}t)}. \tag{3.92}$$

From (3.92) we immediately conclude the following:

- If $\omega^{im}(k) = 0$, the kth mode preserves its amplitude; we call such a mode *conservative*.
- If $\omega^{im}(k) < 0$, the amplitude of the kth mode decreases exponentially in amplitude; such a mode is called *dissipative*.
- If $\omega^{im}(k) > 0$, the kth mode shows exponential increase in amplitude; such a mode will not remain bounded if time proceeds.

Exercise 3.3g.
Consider the PDE

$$\partial_t u = \partial_x^2 u + \partial_x^3 u + \partial_x^4 u\ .$$

a. *For which values of the wave number k are the modes of this system dissipative?*
b. *Calculate the dispersion relation for the following PDE:*
   i. $\partial_t^2 u = c^2 \partial_x^2 u$ *(wave equation),*
   ii. $\partial_t^2 u = c^2 \partial_x^2 u + \alpha\, u$,
   iii. $\partial_t^2 u = c^2 \partial_x^2 u + \alpha\, \partial_x^4 u$,
   iv. $\partial_t^2 u + \alpha\, \partial_t u + \beta\, u + \gamma\, \partial_x u + \delta\, \partial_x^2 u = 0$ *(telegraph equation).*

   For which values of the wave number k (in terms of the constants α, β, γ, δ, and c) are the modes of these systems dissipative? Calculate the phase velocities, too.

Example 3.3c. Heat diffusion in an infinitely long rod.
Let us consider heat diffusion in an infinitely long rod. The temperature $u(x,t)$ satisfies

$$\partial_t u = \kappa\, \partial_x^2 u\ . \tag{3.93}$$

As an initial profile we take a strongly peaked distribution

$$u(x,0) = q\ \delta(x) \tag{3.94}$$

for some constant q. According to (3.89), we may write $u(x,t)$ in the form

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{u}(k)\, e^{i(kx-\omega(k)t)} dk. \tag{3.95}$$

□

Exercise 3.3h.

a. *Show that the dispersion relation corresponding to* (3.93) *is given by*

$$\omega(k) = -i\,k^2.$$

b. *Show that the Fourier transform* $\hat{u}(k)$ *of the initial profile* (3.94) *is given by*

$$\hat{u}(k) = q.$$

c. *Conclude that the system is dissipative and then give the physical interpretation of this conclusion.*

d. *Find* $u(x,t)$ *by evaluating the integral in* (3.95)*. You might make use of the equality*

$$\int_{-\infty}^{+\infty} e^{(-a\,x^2+b\,x)}\,dx = \sqrt{\frac{\pi}{a}}\,e^{\frac{b^2}{4a}}, \tag{3.96}$$

where a *is a positive real constant and* b *an arbitrary complex number.*

e. *Compare the result found in* d *with the analysis and results in section* 1.5.2.

3.3.5 Group velocity

The phase velocity v_{ph} defined in (3.91) has a clear interpretation: it is the speed at which a mode propagates. However, if we take a superposition of modes, as is effectively done in the Fourier integral (3.89), the notion of speed is not that simple. Only if the modes involved have wave numbers that are not very different does it make sense to introduce a measure for the speed of the superposition as a whole. We show this first for the superposition of two modes.

We consider a system with dispersion relation $w(k)$. A superposition of two modes with wave numbers k_1 and k_2 reads as

$$u(x,t) = e^{i(k_1 x-\omega(k_1)t)} + e^{i(k_2 x-\omega(k_2)t)}\ .$$

Introducing the differences

$$\Delta k = \tfrac{1}{2}(k_2 - k_1), \quad \Delta\omega = \tfrac{1}{2}(\omega(k_2) - \omega(k_1))$$

and the averages

$$k_0 = \tfrac{1}{2}(k_1 + k_2), \quad \omega_0 = \tfrac{1}{2}(\omega(k_2) + \omega(k_1)),$$

we may write $u(x,t)$ in the form

$$u(x,t) = 2\cos(\Delta k\,x - \Delta\omega\,t)\,e^{i(k_0 x-\omega_0 t)}.$$

So, the superposition $u(x,t)$ has the character of the mode $\exp(i\,(k_0 x - \omega_0 t))$, the so-called carrier wave, but its amplitude is affected by the presence of the cosinus term. Note that the

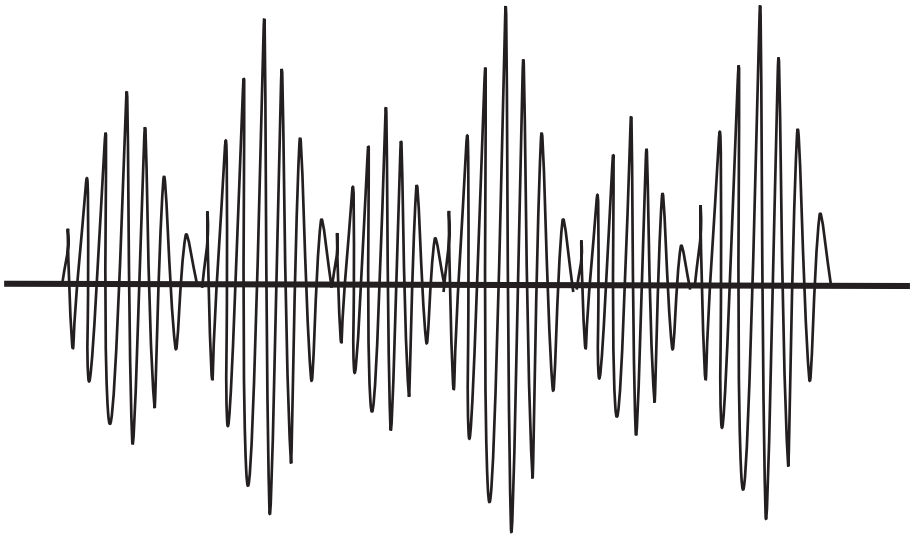

Figure 3.7. A train of beats.

latter term, the so-called modulation, varies in both space and time. It is the real part of a mode with velocity $\Delta\omega/\Delta k$. This motivates us to introduce the *group velocity* v_{gr} by

$$v_{gr}(k) = \frac{d\omega}{dk}.$$

For $\Delta k \ll k_0$, the modulation travels at approximately the group velocity $v_{gr}(k_0)$. Then the optical effect on observing $Re\ (u(x,t))$ is that of a train of beats as plotted in Fig. **3.7**. The envelope propagates at speed $v_{gr}(k_0)$, while the carrier wave propagates at the phase speed ω_0/k_0. Hence, the carrier wave moves with respect to the envelope, since in general $d\omega/dk \neq \omega/k$.

If we superimpose infinitely many modes, nearly all kinds of behavior in space and time can be represented, as follows from Fourier theory. A special structure is obtained if one adds modes with wave numbers k that are centered around a given value k_0. Such a superposition is called a *wave group* or *wave package*. Let the distribution of the k-values in the package at $t = 0$ be given by the spectrum $c(k)$ and let the system have dispersion relation $\omega(k)$. Then, the time evolution of the package is given by

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} c(k)\, e^{i(kx-\omega(k)t)}\, dk. \tag{3.97}$$

If $c(k)$ is strongly peaked around k_0 and vanishes outside a small neighborhood of k_0, the resulting effect in observing $u(x,t)$ is that of an envelope that travels with approximately the group velocity $v_{gr}(k_0)$ in much the same way as the beat pattern above. This can be concluded from a Taylor expansion of $\omega(k)$ around k_0:

$$\omega(k) = \omega(k_0) + (k-k_0)v_{gr}(k_0) + \mathcal{O}((k-k_0)^2).$$

If the quadratic term is neglected, we get the approximation

$$u(x,t) \approx \frac{1}{\sqrt{2\pi}}\, e^{-i(\omega(k_0)-k_0 v_{gr}(k_0))t} \int_{-\infty}^{+\infty} c(k)\, e^{ik(x-v_{gr}(k_0)t)}\, dk \tag{3.98}$$
$$= \frac{1}{\sqrt{2\pi}}\, e^{-i(\omega(k_0)-k_0 v_{gr}(k_0))t}\, u_0(x - v_{gr}(k_0)t),$$

where $u_0(x) := u(x,0)$. This shows that the initial profile $u_0(x)$ is translated with the group velocity $v_{gr}(k_0)$. At the same time, it is subject to a time modulation with a period depending on the difference between the phase and the group velocity.

A more precise approach to define the velocity of a wave package is to consider the position of its center of mass, which is given by

$$X(t) = \frac{\displaystyle\int_{-\infty}^{+\infty} x\, |u(x,t)|^2\, dx}{\displaystyle\int_{-\infty}^{+\infty} |u(x,t)|^2 dx}.$$

According to standard mechanics, the speed of the package can be defined as the time derivative of its center of mass $X(t)$. This is referred to as the *centrovelocity* v_c of the package:

$$v_c(t) = \frac{dX}{dt}(t). \tag{3.99}$$

The centrovelocity is also obtained by averaging the group velocity $v_{gr}(k)$ over all k-values, using as weighting function $|\hat{u}(k,t)|^2$ with $\hat{u}(k,t)$ the Fourier transform of $u(x,t)$:

$$v_c(t) = \frac{\displaystyle\int_{-\infty}^{+\infty} v_{gr}(k)\, |\hat{u}(k,t)|^2 dk}{\displaystyle\int_{-\infty}^{+\infty} |\hat{u}(k,t)|^2 dk}, \tag{3.100}$$

where

$$\hat{u}(k,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u(x,t) e^{-ikx} dx.$$

Exercise 3.3i.

Prove that (3.99) *and* (3.100) *are indeed equivalent.*

Exercise 3.3j.

Derive the dispersion relation for the so-called dispersive wave equation

$$(1 - \partial_x^2)\, \partial_t u = -\partial_x u.$$

Show that the phase velocity is always positive. Determine for which wavelengths the group velocity is positive. Investigate the limiting cases of very long and very short wavelengths.

Exercise 3.3k.

Consider the system described by

$$\partial_t^2 u = c^2 \partial_x^2 u + \partial_x^4 u.$$

Write the solution $u(x,t)$ in the form of a Fourier integral. Evaluate this integral if

$$u(x,0) = \begin{cases} 1, & 0 \le x \le 2\pi, \\ 0, & x < 0, x > 2\pi, \end{cases}$$

and

$$\partial_t u(x,0) = 0.$$

Chapter 4
Stability and Robustness

Any model is only a limited description of the phenomenon under consideration, since all effects could never be included. For instance, in writing the equation of motion for a simple pendulum, we usually ignore the effects of air friction, friction in the joint, the extensibility of the cord, etc. Often there are good reasons to ignore effects beforehand. However, such a choice is always based on intuition, and intuition varies from person to person, giving the art of modeling a personal flavor. This is not what we like in science. If we develop a model, we should clearly realize the purpose for which it is meant and justify that the effects we intend to ignore are really of minor importance. For instance, neglecting air resistance in the pendulum motion may be acceptable for small time intervals, but it is certainly not reliable if one wants to know how long it takes a real pendulum to slow down.

In this chapter we address issues concerning stability and robustness. The central question is

Do small perturbations have small consequences?

The *stability* analysis of models concerns the question of how sensitive the solution is for $t \to \infty$ to inaccuracies in the initial state. In practice, one can never specify the initial state perfectly. In some situations this inherent inaccuracy has dramatic consequences; in others, not. For example, if we were able to position a pendulum at rest in the perfect upright position, it would remain there forever, provided that all perturbations could be avoided. However, it is common experience that we are not able to manage this, and the pendulum will start to swing, thus leaving its initial position over a distance that is not related to the inaccuracy in the initial state. So, a small perturbation has big consequences. We therefore tend to call the upright position "unstable." This is in contrast to the downward position. Although we also are not able to prepare the pendulum in exactly this position, the inaccuracy has few consequences, since the pendulum stays very close to it forever. We tend to call the downward position "stable," since a small error in the initial condition has only small consequences.

As for *robustness*, the same questions as for stability are relevant, but now with respect to the model equations themselves. Most physical effects can be included in a model by adding an extra term. The coefficient in front of this term measures the strength of the effect

relative to the contributions of the other effects. To estimate how sensitive the solutions of a model are to some effect, one could play with the value of this coefficient and observe how the solution, for given initial values, depends on this coefficient. If the influence of the extra term appears to be small in all cases, this may justify not including the effect in the model. We then call the model robust with respect to this effect. The notions of stability and robustness are of great importance in weighing the relevance of a mathematical model.

In the following we illuminate these concepts with examples and then make things more precise in mathematical terms. We deal with stability first and later with robustness. We try to treat ordinary differential equation (ODE) and partial differential equation (PDE) systems as much as possible on an equal footing, since this stresses that stability and robustness are quite general concepts.

4.1 Stability

In §3.2.3 we showed that both ODE and PDE systems can be written in forms that clearly express the evolutionary character:

$$\begin{cases} \dot{\mathbf{u}} = \mathbf{F}(\mathbf{u}), \\ \mathbf{u}(0) = \mathbf{u}_0, \end{cases} \tag{4.1}$$

and

$$\begin{cases} \partial_t \mathbf{u} = \mathbf{F}(\mathbf{u}), \\ \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \end{cases} \tag{4.2}$$

respectively. The dependence of the solution $\mathbf{u}(t)$ for $t \to \infty$ on the initial state $\mathbf{u}_0$ is the subject we want to discuss first, i.e., stability.

When studying (in)stability, it is natural to focus on two types of special orbits: stationary solutions and periodic solutions. For many real-life systems these types of behavior are preferable. One usually wants the systems to operate in or close to one of these states. Due to perturbations, it is in practice never possible to choose the initial conditions such that the system is stationary or periodic right from the start. Therefore, an important question is whether the system converges to such a state if $t \to \infty$ or at least remains in its vicinity. We shall first illustrate these notions for linear systems using very common systems, namely, the harmonic oscillator and heat diffusion. The behaviors of these systems are explicitly known, so we can easily deduce what will happen if $t \to \infty$. Their properties are characteristic for all linear systems with constant coefficients. However, the stability notions we meet in these systems are so generic that in the next section we can readily extend them to the general cases.

Example 4.1a. Harmonic motion and stability.
Let us consider the driven, damped harmonic oscillator, described by

$$m\ddot{u} + c\dot{u} + ku = F_0 \cos(\omega t)\ , \quad u(0) = u_0\ ,\ \dot{u}(0) = v_0, \tag{4.3}$$

with m, k, c, and F_0 the mass, spring constant, friction coefficient, and amplitude of the driving force, respectively. We may rewrite the second-order equation (4.3) into two first-order equations by introducing the state vector $\mathbf{u} := (u, \dot{u})$. This can be concisely written

in the form

$$\dot{\mathbf{u}} = \mathbf{A} \cdot \mathbf{u} + \mathbf{b} \ , \ \ \mathbf{u}(0) = \mathbf{u}_0 := (u_0, v_0), \tag{4.4}$$

with the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ given by (cf. Example 3.1a)

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -\dfrac{k}{m} & -\dfrac{c}{m} \end{pmatrix} \ , \quad \mathbf{b}(t) = \begin{pmatrix} 0 \\ \dfrac{F_0}{m}\cos(\omega t) \end{pmatrix} \ . \tag{4.5}$$

Let us first consider the homogeneous case by taking $F_0 = 0$ and thus $\mathbf{b}(t) := \mathbf{0}$. The only stationary state of this system is then the zero solution $\mathbf{u}_0 = (0, 0)$, thus with the oscillator in its rest position and vanishing velocity. How does the system behave if we perturb this rest state, either by stretching the spring a bit or by forcing it to move at some velocity (or by doing both)? Let us take an initial condition such that $\mathbf{u}_0 \neq (0, 0)$. Then, the solution is explicitly known, as already pointed out in §3.2:

$$\mathbf{u}(t) = e^{\mathbf{A}t} \cdot \mathbf{u}_0 = c_1 \, e^{\lambda_1 t} \phi_1 + c_2 \, e^{\lambda_2 t} \phi_2 \ . \tag{4.6}$$

Here, λ_1, λ_2 are the two (different) eigenvalues of the matrix $\mathbf{A}$,

$$\lambda_1 = -\frac{c}{2m} + \sqrt{\Big(\frac{c}{2m}\Big)^2 - \frac{k}{m}} \ , \quad \lambda_2 = -\frac{c}{2m} - \sqrt{\Big(\frac{c}{2m}\Big)^2 - \frac{k}{m}} \ , \tag{4.7}$$

with ϕ_1, ϕ_2 the corresponding eigenvectors. To ensure that these are independent, we exclude the case of critical damping by requiring $c^2 - 4km \neq 0$. The coefficients c_1, c_2 are determined by the expansion of the initial condition $\mathbf{u_0}$ in terms of the basis vectors ϕ_1 and ϕ_2:

$$\mathbf{u}_0 = c_1 \, \phi_1 + c_2 \, \phi_2 \ . \tag{4.8}$$

From (4.6) we directly conclude that for $t \to \infty$ only three different types of behavior can occur. For convenience we label the eigenvalues such that $Re\ (\lambda_1) \geq Re\ (\lambda_2)$, with Re denoting the real part. Let us take an initial value with $c_1 \neq 0$ and see what happens with the corresponding solution if $t \to \infty$.

1. If Re $(\lambda_1) < 0$, $\mathbf{u}(t)$ converges exponentially fast to the zero solution. We then call the zero solution *asymptotically stable*.

2. If Re $(\lambda_1) = 0$, the first term in (4.6) will not diverge but also will not damp out for $t \to \infty$. So, in this case the perturbed solution will remain in the vicinity of the zero solution. We then call the zero solution *(Lyapunov) stable*.

3. If Re $(\lambda_1) > 0$, then $||\mathbf{u}(t)|| \to \infty$ if $t \to \infty$. We then call the zero solution unstable.

The three possibilities mentioned are sketched in Fig. **4.1**.

From (4.7) we may conclude that for the undriven harmonic oscillator, Re $(\lambda_1) \leq 0$ for all parameter values. As soon as friction is present ($c > 0$), we have that Re $(\lambda_1) < 0$, so then the system is asymptotically stable. In the absence of friction ($c = 0$), we have that Re $(\lambda_1) = 0$, so then stability occurs. Since the case $c < 0$ is not physically acceptable, instability does not occur in this system.

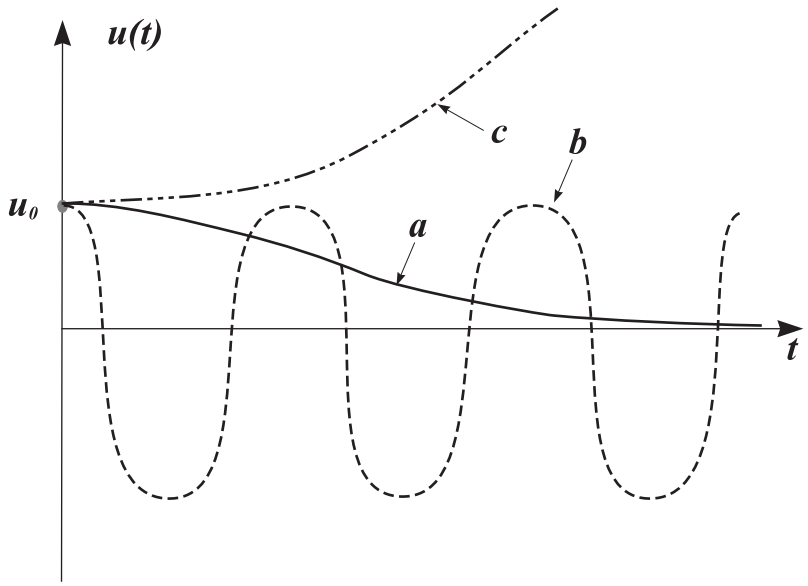

Figure 4.1. Sketch of (a) an asymptotically stable, (b) a stable, and (c) an unstable solution of the harmonic oscillator.

Next, we take into consideration the driving force by taking $F_0 \neq 0$. Let us first consider the case $\omega = 0$, so that the driving force is constant. As an example we could think of a harmonic oscillator in vertical position with the gravitational force acting on the mass, so $F_0 = m\,g$. See Fig. **4.2**. Then the system has as a stationary state

$$\mathbf{u}_0 = \Big(\frac{mg}{k}, 0\Big), \tag{4.9}$$

a rest position shifted in the direction of the force. Introducing a shifted coordinate system via $\bar{\mathbf{u}} := (u - \frac{mg}{k}, v)$, we directly see that $\bar{\mathbf{u}}$ satisfies the homogeneous equation

$$\dot{\bar{\mathbf{u}}} = \mathbf{A} \cdot \bar{\mathbf{u}}. \tag{4.10}$$

So, this case reduces to the homogeneous case dealt with above, and the long-term behavior of the difference $\bar{\mathbf{u}}(t)$ is fully determined by the real parts of the eigenvalues of $\mathbf{A}$. We conclude that the stability properties of the shifted rest position are the same as those of the zero solution of the homogeneous system.

Next, we consider a periodic driving force by requiring $\omega \neq 0$. According to the variation of constants formula (3.24), the solution of (4.3) now reads as

$$\mathbf{u}(t) = e^{\mathbf{A}t} \cdot \mathbf{u}_0 + \int_0^t e^{\mathbf{A}(t-t')} \cdot \mathbf{b}(t')\,dt' =: \mathbf{u}^t(t) + \mathbf{u}^p(t) \tag{4.11}$$

with $\mathbf{u}^t(t)$ the so-called transient solution and $\mathbf{u}^p(t)$ given by an integral. The transient solution is independent from the driving force and, as explained above, dies out if all

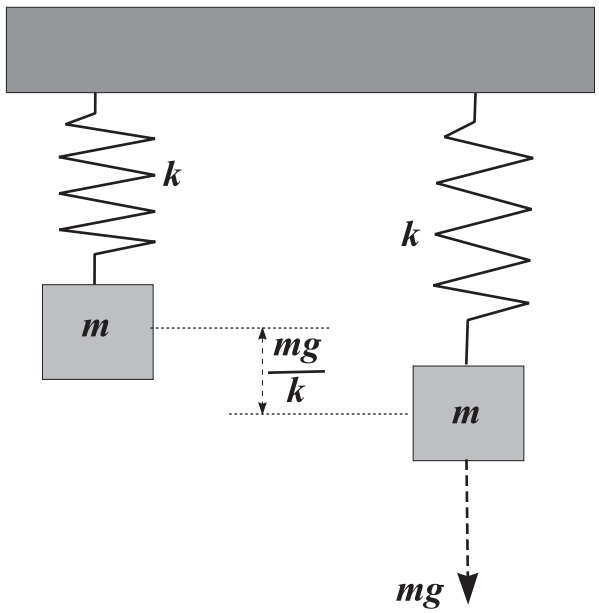

Figure 4.2. Vertical harmonic oscillator under influence of the (constant) gravity force.

eigenvalues of the matrix $\mathbf{A}$ have negative real parts. So, in that case the system "forgets" its initial value and only $\mathbf{u}^p(t)$ persists. For a special choice of the parameters the phenomenon of *resonance* occurs and $||\mathbf{u}^p(t)|| \to \infty$ if $t \to \infty$. See for this case Exercise 4.1a. For all other cases, $\mathbf{u}^p(t)$ will converge to a periodic solution, as can be checked by evaluating the integral in (4.11) (cf. Exercise 4.1a). It has the form

$$u^{per}(t) = a\,cos(\omega t - \phi_0). \tag{4.12}$$

The corresponding orbit in state space is $\mathbf{u}^{per}(t) = (u^{per}(t), \dot{u}^{per}(t))$. The amplitude a and phase ϕ_0 are most easily obtained from substitution of this expression in (4.1). An important observation is that the difference

$$\bar{\mathbf{u}}(t) := \mathbf{u}(t) - \mathbf{u}^{per}(t) \tag{4.13}$$

again satisfies the homogeneous equation (4.10). So, we conclude that the stability of a periodic solution of the periodically driven oscillator is equal to the stability of the zero solution of the undriven oscillator. □

From this example we deduce that the stability of any stationary or periodic solution of a linear or affine system with constant coefficients is fully determined by the real parts of the eigenvalues of the coefficient matrix $\mathbf{A}$. For these special systems one may refer to the stability of the entire system instead of to the stability of a specific solution.

As a warning we emphasize that the stability properties of linear systems with a time-dependent matrix $\mathbf{A}(t)$ are not determined by the eigenvalues of $\mathbf{A}(t)$. These time-dependent eigenvalues do not provide any information about the stability of the system.

Exercise 4.1a. *Resonance.*

a. *Calculate the eigenvectors of* $\mathbf{A}$ *in* (4.5) *as functions of the parameters* m, k, *and* c.

b. *Determine the stability properties of the rest state for* $c > 0$.

c. *How does this change if* $c \downarrow 0$*?*

d. *For what values of the parameters does the rest state become unstable? Interpret the corresponding (un)physical properties of the oscillator.*

e. *Calculate the amplitude* a *in* (4.12) *and deduce for which parameter values* resonance *occurs, i.e.,* $a \to \infty$. *Check that in this case* $\mathbf{u}^p(t)$ *diverges if* $t \to \infty$ *by analyzing the integral in* (4.11).

In the next exercise the same stability ideas as in Example 4.1a are illustrated. The model is simpler but contains the same essentials.

Exercise 4.1b. *Climate control in a cathedral.*

Let us study the dynamics of the temperature $T(t)$ *in a big building, say, a cathedral. The building is not internally heated, so its temperature is fully governed by the heat exchange with the environment. For convenience, we assume* $T(t)$ *to be uniform over the building. Let the temperature of the environment be given by* $T_{ext}(t)$. *See* Fig. **4.3**. *According to Fourier's law (cf. Example* 2.3d*), the heat exchange is proportional to the temperature difference. This leads to the ODE*

$$\dot{T} = -\kappa(T - T_{ext}) \tag{4.14}$$

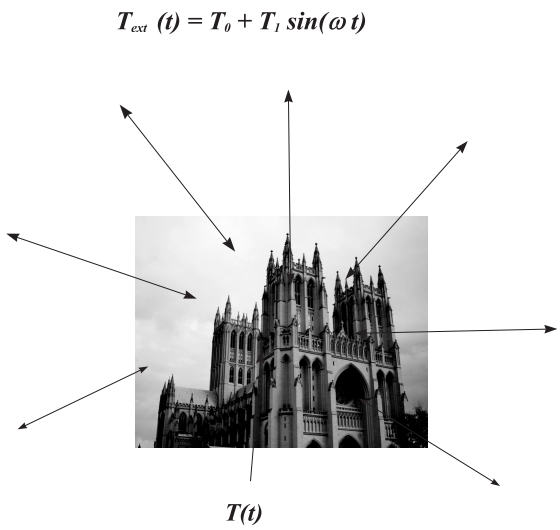

Figure 4.3. Climate control in a cathedral is determined mainly by the exchange of heat through walls, windows, roof, and floor.

with κ the thermal conductivity. *For simplicity we approach the day and night behavior of $T_{ext}(t)$ by a sinusoidal function*

$$T_{ext}(t) = T_0 + T_1 \sin(\omega t) \tag{4.15}$$

with $\omega = (2\pi)/24\ h^{-1}$. This ODE is brought into dimensionless form by introducing $u := (T - T_0)/T_0$ and $\tau := \omega t$:

$$\dot{u} + c_1 u = c_2 \sin\tau \tag{4.16}$$

with dimensionless constants $c_1 = \kappa/\omega$ and $c_2 = (\kappa T_1)/(\omega T_0)$.

a. *Write* (4.16) *in state space form and identify the matrix* $\mathbf{A}$ *and the vector* $\mathbf{b}(t)$. *What is, in this scalar case, the eigenvalue λ of* $\mathbf{A}$*?*

b. *Evaluate $u^p(t)$, defined in* (4.11)*, which in this scalar case is given by*

$$u^p(t) = \int_0^t e^{\lambda(t-t')} b(t')\, dt'. \tag{4.17}$$

c. *Check that $u^p(t)$ approaches the periodic solution*

$$u^{per}(t) = c_2\,(c_1\ \sin t - \cos t)/(1 + c_1^2) \tag{4.18}$$

if $t \to \infty$.

d. *Determine the stability properties of $u^{per}(t)$.*

Until now we have dealt with stability properties of linear ODEs. In the next example we study the same topics for linear PDEs and show that the essential features are identical for ODEs and PDEs.

Example 4.1b. Heat diffusion and stability.
Here we revisit heat diffusion in a long, thin rod of length L. A similar system has already been worked out in Example 3.3b, and now we focus on its stability properties. The time evolution of its temperature $u(x, t)$ satisfies the linear PDE

$$\partial_t u = \kappa\ \partial_x^2 u. \tag{4.19}$$

At both ends, where $x = 0$ and $x = L$, we keep the temperature at a fixed level u_0. So, the boundary conditions are

$$u(0, t) = u(L, t) = u_0, \quad t \geq 0. \tag{4.20}$$

The stationary state clearly is $u^s := u_0$. We are interested in the stability of u^s. How does the system behave for $t \to \infty$ if we start with an initial profile different from u^s? This question is easy to answer since the difference

$$v(x, t) := u(x, t) - u^s \tag{4.21}$$

satisfies (4.19) with homogeneous boundary conditions. According to Example 3.3b an explicit expression for this difference can be derived:

$$v(x,t) = \sum_{n=0}^{\infty} c_n \, e^{-\kappa \lambda_n t} \varphi_n(x) \tag{4.22}$$

with c_n some constants determined by the initial condition, and λ_n and $\varphi_n(x)$ the eigenvalues and eigenfunctions of the eigenvalue problem corresponding to the spatial operator in (4.19) with homogeneous boundary conditions:

$$\begin{cases} -\partial_x^2 \varphi = \lambda \varphi, \\ \\ \varphi(0) = \varphi(L) = 0. \end{cases} \tag{4.23}$$

Similar to Example 3.3b, the λ_n are real in this case. From (4.22) we may directly deduce what will happen: the eigenvalues λ_n fully determine whether the solution $u(x,t)$ will converge to the steady state u^s, will start to oscillate around it, or will blow up. The important observation is that the conclusions about stability of a stationary solution of a linear PDE are very similar to the conclusions drawn in the case of a linear ODE.

One might wonder whether this also holds for the stability of a periodic solution of a PDE system. To investigate this we extend the above heat diffusion problem a bit by adding a source term:

$$\partial_t u = \kappa \; \partial_x^2 u + s(x,t). \tag{4.24}$$

If we take the source periodic, e.g., by setting $s(x,t) := \bar{s}(x)\cos(t)$, this equation may have a periodic solution $u^{per}(x,t)$. As for its stability, we observe that the difference

$$v(x,t) := u(x,t) - u^{per}(x,t) \tag{4.25}$$

with both $u(x,t)$ and u^{per} solutions of (4.24) also satisfies (4.19), but with homogeneous boundary conditions. So, the stability properties of a periodic solution of the inhomogeneous linear system are equal to the stability properties of the stationary solution of the corresponding homogeneous system. □

From this example we see that the situations for linear and affine ODEs and PDEs with constant coefficients are very similar: in the case of an ODE the stability of stationary and periodic solutions is fully determined by the real parts of the eigenvalues of the coefficient matrix **A**, and in the case of a PDE the eigenvalues of the spatial operator play the same role. We conclude that for linear ODE or PDE systems we may refer to the stability of the system as a whole, instead of referring to the stability of a particular solution.

4.2 Stability definitions

Although the systems studied in the examples in §4.1 deal with a very special type of equations, we want to emphasize that the three types of behavior for $t \to \infty$ observed in these systems are generic. This inspires us to formalize them by introducing general stability definitions. We do this in a general setting, simultaneously for ODEs and PDEs.

Let $\bar{\mathbf{u}}(t)$ be a stationary or periodic solution of (4.1) or (4.2). To determine the stability of $\bar{\mathbf{u}}$, we take an initial value $\mathbf{v}_0$ in the vicinity of $\bar{\mathbf{u}}$ and study the solution of the perturbed problem

$$\begin{cases} \dot{\mathbf{v}} = \mathbf{F}(\mathbf{v}), \\ \mathbf{v}(0) = \mathbf{v}_0, \end{cases} \tag{4.26}$$

or in the PDE case

$$\begin{cases} \partial_t \mathbf{v} = \mathbf{F}(\mathbf{v}), \\ \mathbf{v}(\mathbf{x}, 0) = \mathbf{v}_0(\mathbf{x}). \end{cases} \tag{4.27}$$

Stability of stationary solutions
If $\bar{\mathbf{u}}$ is a stationary solution, we have in the ODE case that $\bar{\mathbf{u}} = \mathbf{u}_0$ for some constant vector $\mathbf{u}_0$. The distance between $\mathbf{v}(t)$ and $\mathbf{u}_0$ is measured by the Euclidean norm

$$d(\mathbf{v}(t), \mathbf{u}_0) = ||\mathbf{v}(t) - \mathbf{u}_0|| \ .$$

For the PDE, the state space U is a space of vector-valued functions, and a stationary solution is given by some function $\bar{\mathbf{u}} = \mathbf{u}_0(x)$. In this case there is not one natural candidate for the distance function. A possible choice could be

$$d(\mathbf{v}(t), \mathbf{u}_0) := ||\mathbf{v}(t) - \mathbf{u}_0||_2 \ ,$$

where we use the standard norm for quadratically integrable functions:

$$||\mathbf{u}||_2 = \left\{ \int_\Omega ||\mathbf{u}(\mathbf{x})||^2 \, d\mathbf{x} \right\}^{1/2} . \tag{4.28}$$

If in the distance function $d(.,.)$ the Euclidean norm $||.||$ is used in the ODE case and $||.||_2$ in the PDE case, both of these cases can be dealt with on an equal footing.

We call $\mathbf{u}_0$ *stable* if for all t the distance $d(\mathbf{v}(t), \mathbf{u}_0)$ can be made arbitrarily small by taking the distance $d(\mathbf{v}_0, \mathbf{u}_0)$, which measures the strength of the initial perturbation, smaller and smaller. This is made precise in the following definition.

Stability: *A stationary solution* $\mathbf{u}_0$ *of* (4.1) *or* (4.2) *is (Lyapunov) stable if for any solution of the perturbed equation* (4.26) *or* (4.27) *it holds that for any* $\varepsilon > 0$ *we can find a* $\delta(\varepsilon) > 0$ *such that for all* $t \geq 0$

$$d(\mathbf{v}(t), \mathbf{u}_0) < \varepsilon \quad \textit{if} \quad d(\mathbf{v}_0, \mathbf{u}_0) < \delta \ .$$

In Fig. **4.4**(a) the situation is sketched for a scalar system.

A stronger form of stability is met not only if $\mathbf{u}_0$ is Lyapunov stable but also if any perturbation of $\mathbf{u}_0$ damps out for $t \to \infty$. This leads to the following definition.

Asymptotic stability: *A stationary state* $\mathbf{u}_0$ *of* (4.1) *or* (4.2) *is asymptotically stable if* $\mathbf{u}_0$ *is (Lyapunov) stable and there exists a neighborhood of* $\mathbf{u}_0$ *such that any solution* $\mathbf{v}(t)$ *starting in this neighborhood converges to* $\mathbf{u}_0$ *for* $t \to \infty$*, i.e.,*

$$\lim_{t \to \infty} d(\mathbf{v}(t), \mathbf{u}_0) = 0. \tag{4.29}$$

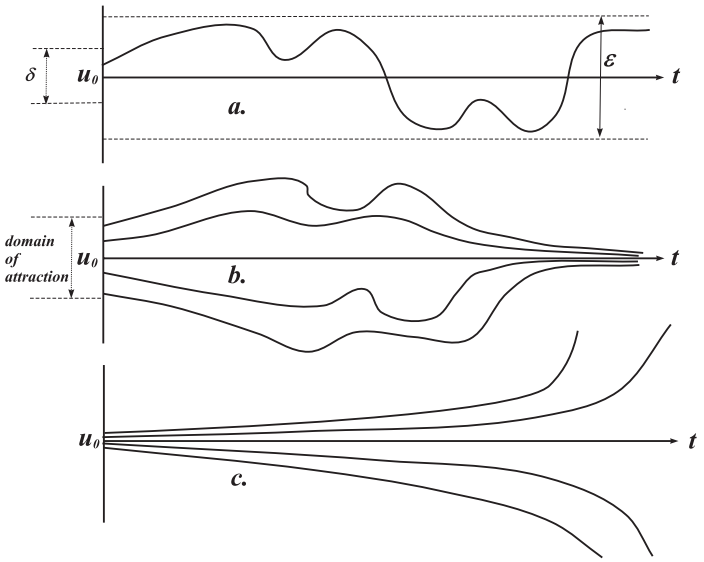

Figure 4.4. Sketch of the behavior of scalar solutions as functions of time in the vicinity of a stable stationary point (a), an asymptotically stable point (b), and an unstable stationary point (c).

An asymptotically stable solution thus acts as an *attractor*. In Fig. **4.4**(b) the situation is sketched for a scalar system. The set of all initial conditions that are starting points of solutions that converge to $\mathbf{u}_0$ is called the *domain of attraction* of $\mathbf{u}_0$. In general, this domain is a restricted neighborhood of $\mathbf{u}_0$. For linear systems the domain of attraction always coincides with the whole space of feasible initial conditions, as illustrated in Example 4.1a.

Stability of periodic solutions

If $\bar{\mathbf{u}}$ is periodic, we have that $\bar{\mathbf{u}}(t+nT) = \bar{\mathbf{u}}(t),\ n = 0, 1, 2, \ldots$, for some period T. The stability of a periodic solution as a whole is called *orbital stability*. This kind of stability is very similarly defined as the stability of a stationary point. We follow the time evolution of a solution $\mathbf{v}(t)$ starting in the vicinity of $\bar{\mathbf{u}}$ and measure the distance to the set $\mathbf{C}$ of all points on the closed orbit defined as

$$\mathbf{C} = (\bar{\mathbf{u}}(t)|\ 0 \leq t < T). \tag{4.30}$$

Such a perturbed solution $\mathbf{v}(t)$ may for $t \to \infty$ remain close to $\mathbf{C}$, converge to it, or leave it. To specify such behavior, we need a measure for the distance between an arbitrary point in the state space, $\mathbf{w}$, say, and the closed curve $\mathbf{C}$. We define it as

$$d(\mathbf{w}, \mathbf{C}) = \min_{\mathbf{z}\in\mathbf{C}} ||\mathbf{w} - \mathbf{z}|| \ . \tag{4.31}$$

See Fig. **4.5**. Note that this definition also applies if $\bar{\mathbf{u}}$ is a stationary solution, since that is nothing but a periodic solution with period $T = 0$ and $\mathbf{C}$ contracted on one point.

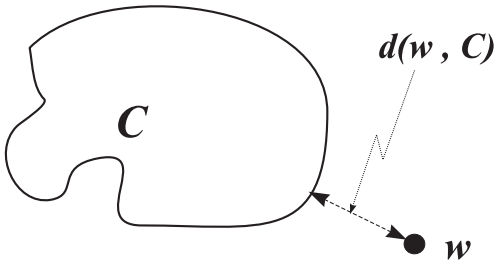

Figure 4.5. *Distance between a point* **w** *and a periodic orbit corresponding to a closed curved* **C**.

We call $\bar{\mathbf{u}}$ *orbitally stable* if for all t the distance $d(\mathbf{v}(t), \mathbf{C})$ can be made arbitrarily small by taking the initial distance $d(\mathbf{v}_0, \mathbf{C})$, which measures the strength of the initial perturbation, smaller and smaller. This is made precise in the following definition.

Orbital stability: *A periodic solution* $\bar{\mathbf{u}}$ *of* (4.1) *or* (4.2) *is orbitally stable if for any solution of the perturbed equation* (4.26) *or* (4.27) *it holds that for any* $\varepsilon > 0$ *we can find a* $\delta(\varepsilon) > 0$ *such that for all* $t > 0$

$$d(\mathbf{v}(t), \mathbf{C}) < \varepsilon \quad \textit{if} \quad d(\mathbf{v}_0, \mathbf{C}) < \delta(\varepsilon).$$

Just as above, we may also introduce a stronger form of orbital stability not only if $\bar{\mathbf{u}}$ is orbitally stable but also if the perturbations damp out for $t \to \infty$. This leads to the next definition.

Asymptotic orbital stability: *A periodic state* $\bar{\mathbf{u}}$ *of* (4.1) *or* (4.2) *is asymptotically orbitally stable if* $\bar{\mathbf{u}}$ *is orbitally stable and there exists a neighborhood of the closed curve* **C** *such that any solution* $\mathbf{v}(t)$ *starting in this neighborhood converges to* **C** *for* $t \to \infty$*, i.e.,*

$$\lim_{t\to\infty} d(\mathbf{v}(t), \mathbf{C}) = 0.$$

Phase shift

An important phenomenon of asymptotic orbital stability is as follows. Consider an asymptotically orbitally stable solution with closed curve **C**. A solution $\mathbf{v}(t)$ starting in a point $\mathbf{v}_0$ on **C** will stay on **C** forever. Another solution, $\mathbf{w}(t)$, say, starting in a point $\mathbf{w}_0$ close to $\mathbf{v}_0$ but not on **C**, will converge to **C** if time proceeds. So, in the limit $t \to \infty$ both $\mathbf{v}(t)$ and $\mathbf{w}(t)$ follow the same curve given by **C**, but maybe with some distance between them, as sketched in Fig. **4.6**. This implies that both solutions follow the same trajectory, but with a *phase shift*. The phase shift ϕ is defined as the number for which

$$\lim_{t\to\infty} ||\mathbf{v}(t-\phi) - \mathbf{w}(t)|| = 0\ .$$

In general, ϕ may be a complicated function of the vector field and the initial values $\mathbf{v}_0$ and $\mathbf{w}_0$. The concept of a phase shift is illustrated in the following exercise.

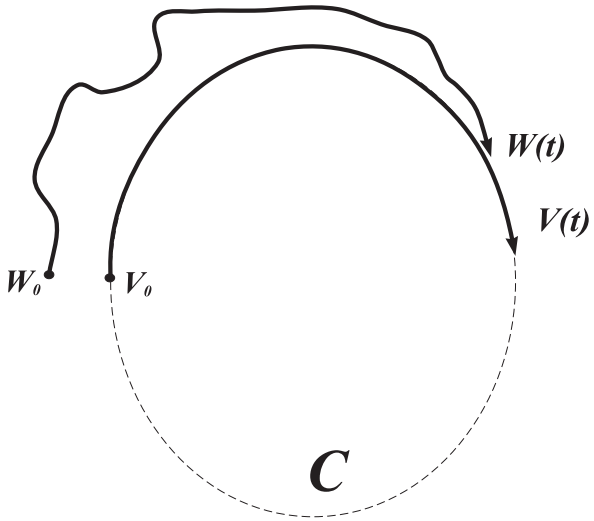

Figure 4.6. All solutions starting on and in the vicinity of an asymptotically stable periodic orbit **C** converge to **C**, but possibly with a phase shift.

Exercise 4.2a.

We consider a system in the plane. In terms of polar coordinates (r, φ) *the equations read as*

$$\begin{cases} \dot{r} = 1 - r, \\ \dot{\varphi} = r. \end{cases}$$

This system has the periodic solution $r = 1$, $\varphi = \varphi_0 + (t - t_0)$, *corresponding to the unit circle. For convenience we choose* $\varphi_0 = 0$ *and* $t_0 = 0$ *in the following. The first equation can be solved by separation of variables. If we write the initial value* r_0 *as* $r_0 = 1 + \delta$, *with* $\delta > -1$, *the solution* $r(t)$ *is given by*

$$r(t) = 1 + \delta\, e^{-t}.$$

From this it is clear that every solution starting off the unit circle will converge to it as $t \to \infty$. *We can find* $\varphi(t)$ *from integrating the second equation:*

$$\varphi(t) = t + \delta(1 - e^{-t}).$$

From this we observe that the two solutions starting at $(r_0, \varphi_0) := (1, 0)$ *and* $(r_0, \varphi_0) := (1 + \delta, 0)$ *both follow the unit circle as* $t \to \infty$, *but with a constant phase shift* δ.

Instability

Eventually, instability is simply defined as follows:

Instability: *A stationary state* $\bar{\mathbf{u}}$ *of* (4.1) *or* (4.2) *is* unstable *if it is not Lyapunov stable. Such a solution is thus certainly not asymptotically stable.*

The instability of a periodic solution is completely similarly defined.

Instability implies that in every neighborhood of an unstable solution $\bar{\mathbf{u}}$, however small this neighborhood may be, at least one solution that leaves the vicinity of $\bar{\mathbf{u}}$ starts. In Fig. **4.4**(c) the situation is sketched for an unstable stationary point.

4.3 Linearization

To establish the stability of stationary solutions of nonlinear systems we can readily make use of the insights gained for linear systems. In the nonlinear case we cannot speak of the stability of the system as a whole but can speak only of the stability of a specific solution. According to the stability definitions above, we have to determine the time evolution of solutions starting close to the solution under consideration. For example, if all solutions in a certain neighborhood around this stationary point, however small it may be, converge to it, this point is asymptotically stable. The idea now is to exploit the fact that the time evolution of solutions starting very close to a stationary point will initially be governed by the *linearization* of the system around this point. As explained in §3.3.1, the linearization of the ODE

$$\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u}) \tag{4.32}$$

around a stationary point $\bar{\mathbf{u}}$ is given by

$$\dot{\mathbf{v}} = \mathbf{J}(\bar{\mathbf{u}}) \cdot \mathbf{v} \tag{4.33}$$

with $\mathbf{J}$ the Jacobi matrix of $\mathbf{F}$. Similarly, the linearization of a PDE

$$\partial_t \mathbf{u} = \mathbf{F}(\mathbf{u}) \tag{4.34}$$

is given by

$$\partial_t \mathbf{v} = \mathbf{F}'(\bar{\mathbf{u}})\, \mathbf{v} \tag{4.35}$$

with $\mathbf{F}'$ the Frechet derivative of $\mathbf{F}$. The spectrum of the linearized systems (4.33) and (4.35) is formed by the eigenvalues of $\mathbf{J}$ and $\mathbf{F}'$, respectively. The signs of the real parts of these eigenvalues determine to a great extent the stability of $\bar{\mathbf{u}}$, since the following statement holds. For more discussions on this topic we refer to, e.g., [22] for ODEs and [20, 33] for PDEs.

> **Linearization and Stability:** *If a stationary solution $\bar{\mathbf{u}}$ of the linearized system is asymptotically stable, respectively, unstable, then it is also an asymptotically stable, respectively, unstable, solution of the nonlinear system.*

Note that this statement does not cover the case of a stable (but not asymptotically stable) solution. It implies that the spectrum of the linearized system provides all the information needed to determine the stability properties of the stationary state in the cases of asymptotic stability and instability. In the first case all eigenvalues have negative real parts, and in the latter case at least one eigenvalue has a positive real part. If one or more eigenvalues have vanishing real parts and the others negative real parts, the linearized system is stable but

not asymptotically stable. In this case linearization does not provide any information about the stability of the nonlinear system.

Example 4.3a.
To illustrate the study of stability with the help of linearization, the damped pendulum is ideal. Its equation of motion reads as

$$m\,\ell\,\ddot{\varphi} + c\,l\,\dot{\varphi} + m\,g\,\sin\varphi = 0. \tag{4.36}$$

This system has already been treated in Example 1.3b, via dimensional analysis, and in Exercise 3.1c. Its state vector is $(\varphi, \dot{\varphi})$ and its vector field reads as

$$\mathbf{F}(\varphi, \dot{\varphi}) = \begin{pmatrix} f_1 \\ \\ f_2 \end{pmatrix} = \begin{pmatrix} \dot{\varphi} \\ \\ -\dfrac{c}{m}\,\dot{\varphi} - \dfrac{g}{l}\sin\varphi \end{pmatrix}. \tag{4.37}$$

The Jacobi matrix of the pendulum is given by

$$\mathbf{J}(\varphi, \dot{\varphi}) = \begin{pmatrix} \dfrac{\partial f_1}{\partial \varphi} & \dfrac{\partial f_1}{\partial \dot{\varphi}} \\ \\ \dfrac{\partial f_2}{\partial \varphi} & \dfrac{\partial f_2}{\partial \dot{\varphi}} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \\ -\dfrac{g}{l}\cos\varphi & -\dfrac{c}{m} \end{pmatrix}.$$

So, $\mathbf{J}$ depends only on the amplitude φ and not on the velocity $\dot{\varphi}$. The stationary states follow from the conditions $f_1 = f_2 = 0$. They are $(0, 0)$, corresponding to the downward position, and $(\pi, 0)$, corresponding to the upward position. The eigenvalues of $\mathbf{J}(0, 0)$ are

$$\lambda_{\pm} = \frac{1}{2m}\left(-c \pm \sqrt{c^2 - \frac{4gm^2}{l}}\right).$$

These eigenvalues are real if $c \geq 2m\sqrt{g\ell}$ and complex otherwise. Furthermore, Re $(\lambda_{\pm}) < 0$ as long as $c > 0$. This implies that the downward position is asymptotically stable, however small the friction may be. The stability of the upward rest state is investigated in Exercise 4.3a. The domain of attraction of the downward rest position does not follow from linearization. From physical practice we know that this domain contains a great deal of the state space, but it does not coincide with it. For example, the upward stationary position does not belong to it, not do all points in which a trajectory starts that converges to the upward rest position for $t \to \infty$ under the influence of friction.

Also the imaginary parts of the $\lambda_{\pm}$ provide information about the behavior of solutions for $t \to \infty$. From the real parts we know that the bigger the friction coefficient c is, the faster the solutions will converge to the rest position. In the case of critical damping, when Im $(\lambda_{\pm}) = 0$, this convergence occurs without oscillations: the pendulum approaches the downward position without passing it. It will approach it in a creeping way. If Im $(\lambda_{\pm}) \neq 0$, the pendulum will approach the rest position in an oscillating way, and the bigger the imaginary parts, the higher the frequency of the oscillations will be. □

Exercise 4.3a.

Let us first study the stability properties of the pendulum if friction is ignored, so by setting $c = 0$, perform the following steps:

a. *Calculate the Jacobi matrix and its eigenvalues for the upward position $(\pi, 0)$ and draw conclusions about its stability.*

b. *Do the same for the downward position $(0, 0)$.*

Next, we switch on the friction by taking $c > 0$.

c. *Describe for both $c = 0$ and $c > 0$ the behavior of the system after being perturbed from a stationary point; perturb both its position and its velocity. Conclude from these insights the consequences of switching on the friction for the stability of the stationary points.*

The next example shows the application of linearization in the case of a PDE system.

Example 4.3b. Nonlinear heat diffusion.
We consider heat diffusion in a rod of length L with a thermal conductivity coefficient D, which depends on the temperature itself. This diffusion behavior is often found in practice. The dynamics of the temperature distribution $u(x, t)$ is then governed by the diffusion equation

$$\rho\, c\, \partial_t u = \partial_x (D(u)\, \partial_x u). \tag{4.38}$$

For convenience we take both the density ρ and the heat capacity c constant and equal to unity, so $\rho = 1$, $c = 1$, and similarly for the length of the rod: $L = 1$. Let the conductivity D be a differentiable function. Then we may rewrite the heat equation in the form

$$\partial_t u = D'\, (\partial_x u)^2 + D\, \partial_x^2 u := F(u) \tag{4.39}$$

with $D' := \partial_u D$. Both D and D' thus depend on u. The Frechet derivative $F'(u)$ of $F(u)$ can be calculated as pointed out in §3.3.1. This operator is in this case given by

$$F'(u) = 2\, D'\, (\partial_x u)\, \partial_x + D''\, (\partial_x u)^2 + D'\, (\partial_x^2 u) + D\, \partial_x^2\, . \tag{4.40}$$

If we take as boundary conditions equal temperatures at both ends,

$$u(0, t) = u(1, t) = 1, \quad t \geq 0, \tag{4.41}$$

we have as a stationary solution the uniform distribution $\bar{u} := 1$. Linearizing around this stationary state, we arrive at

$$\partial_t\, v = F'(\bar{u})\, v = D_1\, \partial_x^2 v\, , \tag{4.42}$$

where $D_1 := D(1)$. To determine the stability properties of $\bar{u}$ we thus need the spectrum of the spatial operator $D_1\, \partial_x^2$. This follows from the eigenvalue problem

$$D_1\, \partial_x^2\, \varphi = \lambda\, \varphi$$

with φ satisfying the homogeneous boundary conditions $\varphi(0) = \varphi(1) = 0$.

The corresponding eigenfunctions and eigenvalues are

$$\varphi_n(x) = \sin n\pi x \ , \quad \lambda_n = -(n\pi)^2 D_1 \ , \quad n = 1, 2, \ldots \ .$$

Note that $D_1 > 0$, since D is the conductivity. The spectrum $\{\lambda_n\}$, $n = 1, 2$, is thus completely real and negative. From this we may conclude that $\bar{u} := 1$ is asymptotically stable. □

Exercise 4.3b.

In Example 4.3b *we introduced the conductivity* $D(u)$. *Here, we specify a functional form for* $D(u)$ *and take*

$$D(u) = e^{\alpha u} \tag{4.43}$$

with α *some constant controlling the strength of the dependence of* D *on* u. *This dependence vanishes for* $\alpha = 0$ *and increases with increasing* α *values. As boundary conditions we prescribe*

$$u(0, t) = u_0 \ , \quad u(1, t) = u_1 \quad \forall t \geq 0 \ . \tag{4.44}$$

a. *Show, using separation of variables, that the steady temperature distribution in the rod is now given by*

$$\bar{u}^s(x) = \frac{1}{\alpha} ln[(e^{u_1} - e^{u_0})x + e^{u_0}] \ , \quad 0 \leq x \leq 1 \ . \tag{4.45}$$

b. *Determine the stability of* $\bar{u}^s$. *Follow the same steps as in Example* 4.1b, *so start with deriving the PDE governing the dynamics of the difference* $u(x, t) - \bar{u}^s(x)$.

4.4 Robustness

In the preceding sections we studied the sensitivity of systems with respect to perturbations of the initial condition and defined the stability of stationary and periodic solutions. Here, we focus on the sensitivity of solutions to changes in the parameters. In other words, we deal with the question of how robust solutions are under parameter perturbations. This question is highly relevant in practice. When modeling real-life systems, one always meets with the question of which effects have to be included and which effects can be left out. To keep models simple and manageable, one wants to include only those effects that essentially determine the behavior of the system. However, it is not always clear in advance which effects are dominant. A useful strategy then is to include an effect of which the relevance is doubted still in the model. At the same time one includes a parameter that controls the strength of the effect. By varying the parameter we can switch on and switch off the effect. Studying robustness then boils down to determining the sensitivity of the solutions to this parameter. We emphasize that comparison of the magnitudes of different terms in an equation makes sense only if the model is in dimensionless form. So, it is always worthwhile to apply the techniques in §1.3 first, before performing a robustness analysis.

For example, the modeling of a simple system like a pendulum requires a number of decisions. Air resistance, friction in the joint, and extensibility of the cord are all effects

that in practice may play a role but are discarded most of the time. Intuition and/or experiments may lead to such a decision. However, this can be rather misleading. Neglect of air resistance does not influence the pendulum motion on short time scales, but in the long run this effect will force the pendulum to come to rest. In the next example we discuss this a bit further.

Example 4.4a. Effect of friction on a pendulum.
The swinging pendulum, under influence of gravitation and air friction, is governed by the ODE

$$m\,\ell\,\ddot{\varphi} + c\,\ell\,\dot{\varphi} + m\,g\,\sin\varphi = 0$$

with $\varphi(t)$ the angle with the vertical, m the mass, ℓ the length of the pendulum, and g the gravitational constant. In this model air friction is taken to be linearly proportional to the velocity $\ell\dot{\varphi}$ with friction coefficient c. One expects the effect of air friction to be very small. To compare the orders of magnitude of the different terms, we turn the model into a dimensionless form. It is convenient to introduce the dimensionless time $\tau = t\sqrt{g/\ell}$. This leads to

$$\ddot{\varphi} + \varepsilon\,\dot{\varphi} + \sin\varphi = 0$$

with the derivatives taken with respect to τ and the dimensionless parameter ε given by

$$\varepsilon = \frac{c}{m}\sqrt{\frac{\ell}{g}}\,.$$

Substituting realistic values for c, m, and ℓ, we find that, indeed, $\varepsilon \ll 1$. This implies that initially the friction term has only a minor effect. However, in the long run the friction term is always felt and even becomes so dominant that it determines the behavior for $t \to \infty$. For $\varepsilon = 0$, the system will continue to move when it starts outside its stationary points, whereas for $\varepsilon > 0$ it always converges to the rest state, however small ε may be. We conclude that for $\varepsilon = 0$ any solution, except for the stationary solutions, is not robust with respect to perturbations of ε.

A related aspect in which the presence of friction is decisive is that for $\varepsilon = 0$ the system is independent from the mass m, whereas for $\varepsilon > 0$ the mass is important. Since $\varepsilon \sim 1/m$, the larger the mass of the pendulum, the longer it takes to come to rest. □

This example illustrates that even inclusion of a very small term may have dramatic consequences if $t \to \infty$. Let us formalize these ideas here. As for stability we showed that it can be introduced for ODEs and PDEs at the same time by using a general concept of "distance" between solutions. Here, we follow the same approach. We study a system described by an ODE,

$$\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u};\mathbf{p}); \quad \mathbf{u}(0) = \mathbf{u}_0, \tag{4.46}$$

or by its PDE analogue. The parameter vector $\mathbf{p} \in \mathbb{R}^m$ measures the strength of one or more effects. We can always scale the parameters such that for $\mathbf{p} = \mathbf{0}$ the terms representing the effects vanish. For clarity we shall focus on robustness with respect to one particular effect. So, we vary only one parameter and indicate it by p. The solution of (4.46) will then depend on p and we denote if as $\mathbf{u}^p(t)$. If the effect is switched off, and thus $p = 0$, the solution

is denoted as $\mathbf{u}^0(t)$. The robustness of $\mathbf{u}^0(t)$ with respect to p is determined by the behavior of the distance between $\mathbf{u}^p(t)$ and $\mathbf{u}^0(t)$ for $t \to \infty$.

Asymptotic robustness: *A stationary or periodic solution* $\mathbf{u}^0$ *of* (4.46) *is* asymptotically robust *(or* asymptotically structurally stable*) with respect to perturbations in the parameter p, if there exists an interval of p values around* $p = 0$ *for which* $d(\mathbf{u}^p(t), \mathbf{C}^0) \to 0$ *if* $t \to \infty$.

Here, the set $\mathbf{C}$ is defined in (4.30) and $d(.,.)$ is the distance function we defined earlier in (4.31). If $\mathbf{u}^0$ is a stationary solution, then $\mathbf{C}$ consists of only that point.

A weaker form of robustness is found if not all perturbations of $p = 0$ damp out for $t \to \infty$ but if the difference can be made arbitrarily small for all times by taking p small enough. Its formula reads as follows.

Robustness: *The stationary or periodic solution* $\mathbf{u}^0$ *of* (4.46) *is* robust *(also called* structurally stable*) if for any* $\varepsilon > 0$ *we can find a* $\delta(\varepsilon) > 0$ *such that the difference* $d(\mathbf{u}^p(t), \mathbf{C}) < \varepsilon$ *for all* $t \geq 0$ *if* $|p| < \delta$.

In all other cases we call $\mathbf{u}^0$ *not robust* or *structurally unstable* with respect to the effect under consideration.

As for stationary states, all information about the robustness of $\mathbf{u}^0$ with respect to some p is thus given by the time evolution of the difference $\mathbf{z}(t) := \mathbf{u}^p(t) - \mathbf{u}^0$. Since both $\mathbf{u}^p$ and $\mathbf{u}^0$ satisfy (4.46), we may write

$$\dot{\mathbf{z}} = \mathbf{F}(\mathbf{u}^0 + \mathbf{z};\, p) - \mathbf{F}(\mathbf{u}^0;\, 0)\ , \quad \mathbf{z}(0) = \mathbf{0},$$

or its PDE analogue. Let us assume that the vector field $\mathbf{F}(\mathbf{u};\, p)$ is differentiable with respect to both $\mathbf{u}$ and p. For most vector fields used in practice this assumption holds. We are interested in the behavior of $\mathbf{z}$ for small values of p. Since $\mathbf{z}(0) = \mathbf{0}$, $\|\mathbf{z}\|$ will initially be close to zero. In that (beforehand unknown) time window we apply a Taylor expansion of $\mathbf{F}$ with respect to both variables $\mathbf{u}$ and p. Neglecting terms of quadratic and higher order in $||\mathbf{u}||$ and p, we obtain in the ODE case

$$\dot{\mathbf{z}} = \mathbf{J}(\mathbf{u}^0) \cdot \mathbf{z} + p\, \partial_p \mathbf{F}(\mathbf{u}^0;\, 0) \tag{4.47}$$

and in the PDE case

$$\partial_t \mathbf{z} = \mathbf{F}'(\mathbf{u}^0)\, \mathbf{z} + p\, \partial_p \mathbf{F}(\mathbf{u}^0;\, 0). \tag{4.48}$$

If we compare (4.33) and (4.34), derived to investigate stability via linearization, with (4.47) and (4.48), we observe that in the last cases we have an extra term, which acts as an inhomogeneous source term.

In the case of ODE, the perturbation equation (4.47) has the standard form

$$\dot{\mathbf{z}} = \mathbf{A} \cdot \mathbf{z} + \mathbf{b} \tag{4.49}$$

with $\mathbf{A} := \mathbf{J}(\mathbf{u}^0)$ and $\mathbf{b} := p\ \partial_p \mathbf{F}(\mathbf{u}^0;\, 0)$. If $\mathbf{u}^0$ is periodic, this equation is not simple to solve. If $\mathbf{u}^0$ is a stationary state, both $\mathbf{A}$ and $\mathbf{b}$ are constant, and then we can say much more.

For example, if $\mathbf{b} \neq 0$, the origin $\mathbf{z} = 0$ apparently is not a stationary state of this equation. This implies that the presence of the effect represented by $p \neq 0$ is always felt for $t \to \infty$. The system is thus not asymptotically robust. If $\mathbf{A}$ is a nonsingular matrix, i.e., the inverse $\mathbf{A}^{-1}$ exists, (4.49) has the stationary state $-\mathbf{A}^{-1} \cdot \mathbf{b}$. From §4.3 we know that the difference $\mathbf{z}(t)$ will converge to it if all eigenvalues of $\mathbf{A}$ have negative real parts. So, if for $p = 0$ we start in the stationary state $\mathbf{u}_0$ and switch on the effect by setting $p \neq 0$, the system will then converge to the shifted stationary state $\mathbf{u}_0 - \mathbf{A}^{-1} \cdot \mathbf{b}$. Since the shift is proportional to p, we can get it as small as we like by decreasing the value of p. The system is thus robust (or structurally stable).

Example 4.4b. Switching on gravity.
If we study a harmonic oscillator that moves horizontally, gravity will not be relevant. By gradually rotating the oscillator toward a vertical position, the influence of gravity will be increasingly felt. This effect is included in the model by adding a term representing gravitation:

$$m\,\ddot{u} + c\,\dot{u} + ku = -\varepsilon\, m\, g.$$

Then, if ε increases from 0 to 1, the effect of gravity is gradually switched on. In state space form with $\mathbf{u} := (u, \dot{u})$ the equation of motion reads as (cf. Example 3.1a)

$$\dot{\mathbf{u}} = \mathbf{A} \cdot \mathbf{u} + \varepsilon\, \mathbf{b} \tag{4.50}$$

with

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -\dfrac{k}{m} & -\dfrac{c}{m} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ -g \end{pmatrix}.$$

The stationary state of the system depends on ε and is given by

$$u^{\varepsilon} = -\varepsilon\, \frac{mg}{k}.$$

We now start in the unperturbed system with $\varepsilon = 0$ and stationary state $u^0 = 0$ and switch on gravity by setting ε at some positive value. Equation (4.47) then describes what will happen with the difference $z(t) = u(t) - u^0$. Since the vector field in (4.50) is linear in both $\mathbf{u}$ and ε, we find that (4.47) for $\mathbf{z}$ is identical to (4.50) for $\mathbf{u}$. Since all eigenvalues of $\mathbf{A}$ have negative real parts, the perturbation $z(t)$ will converge to the value u^{ε} if $t \to \infty$. Since the difference $u^{\varepsilon} - u^0$ can be made arbitrarily small by taking ε smaller and smaller, the stationary state u^0 is robust with respect to gravity. □

The robustness of periodic orbits is not that simple to establish. In the next examples we show several typical possibilities. For illustrative purposes we take simple systems for which the solutions are explicitly known so that it is easy to check what will happen.

Example 4.4c. Perturbing the spring constant.
Let us study whether the harmonic oscillator is robust under perturbations of the spring constant. We look at the equation

$$m\ddot{u} + (k + \varepsilon)\, u = 0\,, \quad u(0) = 0\,, \ \dot{u}(0) = 1.$$

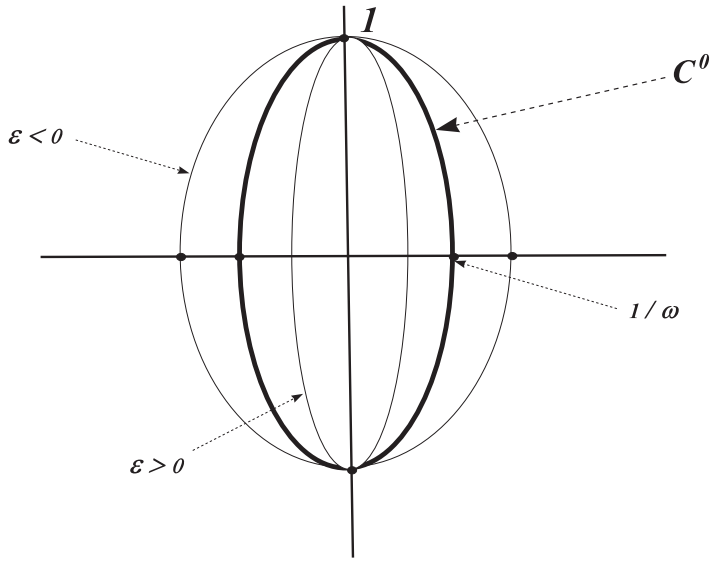

Figure 4.7. *Ellipses in the phase plane for $\varepsilon = 0$, $\varepsilon > 0$, and $\varepsilon < 0$.*

It has the solution

$$u^\varepsilon(t) = \frac{1}{\omega}\sin(\omega t)$$

with $\omega^2 = (k+\varepsilon)/m$. In the (two-dimensional) state space this solution reads as $\mathbf{u}^\varepsilon(t) = (\frac{1}{\omega}\sin(\omega t), \cos(\omega t))$, which corresponds to an ellipse, as sketched in Fig. **4.7**. If we denote the closed orbit corresponding to $\mathbf{u}^\varepsilon(t)$ by $\mathbf{C}^\varepsilon$, then it is clear that these closed curves smoothly deform under variations of ε. Since $\mathbf{C}^\varepsilon \to \mathbf{C}^0$ if $\varepsilon \to 0$, the system is robust under variations of the spring constant, as long as $k > 0$. However, if we vary around $k = 0$, and thus if we study the equation

$$m\ddot{u} + \varepsilon\, u = 0\,, \quad u(0) = 0\,,\ \dot{u}(0) = 1\,,$$

with ε varying around 0, the situation is different. The solution for $\varepsilon = 0$ is $u(t) = t$ and thus is not periodic. However, for $\varepsilon > 0$ the solution is periodic, whereas for $\varepsilon < 0$ the solution will grow exponentially. This implies that switching on the effect of the spring essentially changes the system behavior. So, in this case the system is not robust under that action. □

Example 4.4d. Perturbing the friction coefficient.
A similar conclusion as in the preceding example holds if we perturb the friction coefficient. The solutions of

$$m\ddot{u} + (c+\varepsilon)\,\dot{u} + ku = 0$$

are robust under small variations of ε as long as $c \neq 0$. However, if we vary the friction coefficient around zero, we have to consider the equation

$$m\ddot{u} + \varepsilon\,\dot{u} + ku = 0.$$

Solutions $u^0(t)$ without friction are periodic. A solution $u^\varepsilon(t)$ with $\varepsilon > 0$ converges to the stationary rest state, whereas for $\varepsilon < 0$ solutions not starting in a stationary position will continue to move faster and faster. So, in this case the system is not robust with respect to the introduction of friction. □

4.5 Singular perturbations

In Examples 4.4c and 4.4d we observed that the cases $k = 0$ and $c = 0$, respectively, were of a special character. This phenomenon is the subject of this section. This kind of behavior is often met in modeling situations and deserves thorough attention. In a very general setting one could formulate the effect under consideration as follows. Let $\mathbf{u}$ be the solution of

$$K(\mathbf{u}, \varepsilon) = 0,$$

where K may represent a relation of any kind, e.g., an algebraic expression or a differential equation. Now we focus on the dependence of the solution on the parameter ε. If the solution for $\varepsilon = 0$ is essentially different from the solutions for $\varepsilon \neq 0$, we call the system *singular* with respect to variations of ε around 0. Or in other words, we deal with *singular perturbations*. On purpose this formulation is vague, since it is difficult to specify what is meant by "essentially different." Instead of trying to introduce a definition we shall demonstrate the effect of singular perturbations via examples.

In the first example K represents an algebraic equation.

Example 4.5a. A singularly perturbed algebraic equation.
Let us consider the algebraic equation.

$$K(x, \varepsilon) := \varepsilon x^2 - x + 1 = 0 \tag{4.51}$$

and vary ε around 0. For $\varepsilon \neq 0$ we have two roots

$$x_\pm = \frac{1}{2\varepsilon}(1 \pm \sqrt{1 - 4\varepsilon}),$$

whereas for $\varepsilon = 0$ we have only one root. In the limit $\varepsilon \to \infty$ the character of the solution set typically changes. What happens in this limit? For small ε this expression can be expanded. The leading orders in ε are

$$\begin{aligned} x_+ &= \frac{1}{\varepsilon} - 1 + \mathcal{O}(\varepsilon), \\ x_- &= 1 + \mathcal{O}(\varepsilon). \end{aligned}$$

So for $\varepsilon \to 0$ we have that $x_+ \to \infty$ and $x_- \to 1$. Thus one of the roots converges to the single root of the so-called reduced equation, which is obtained by setting $\varepsilon = 0$,

$$-x + 1 = 0,$$

whereas the other root shifts to $+\infty$. □

Exercise 4.5a.

Plot the solutions x_+ and x_- as functions of ε and observe what happens if $\varepsilon \to 0$.

Our second example concerns the inertia term $m\,\ddot{u}$ in the equation of motion of the harmonic oscillator:

$$m\,\ddot{u} + c\,\dot{u} + k\,u = 0,$$

In a naive approach one might expect that the effect of taking the mass m smaller and smaller would lead to a smooth transition. However, this is not the case. It is a general notion in ODE and PDE cases that the parameter in front of the highest-order derivative plays a special role. An indication for this is that for $m \neq 0$ we need two initial conditions to specify a unique solution, whereas for $m = 0$ we need only one.

Before dealing with our second example we want to make some general remarks on singular perturbations in cases of ODEs or PDEs. In general they are hard to handle. However, many tools are available to treat these systems if it happens that the term with the highest derivative is relevant only in a small window. In general we call such a window a *transition region*. It should be noted that transition regions may occur in both time and space. In the latter case such a region is usually referred to as a *boundary layer*. Boundary layers, for example, are often met in fluid mechanics. As explained in Chapter 2, the dynamics of fluids and gases is described by equations expressing conservation of mass, momentum, and energy. Famous examples are the Navier–Stokes equations for Newtonian flow. If one tries to find a stationary solution of these equations, one usually ends up with an ODE or a PDE in which the independent variable is position. If a stationary solution of a fluid or gas is sought in the vicinity of an object, it may happen that some terms in the equation of motion are relevant only in a thin layer near the object. The corresponding equation in the layer may then be singularly perturbed, giving rise to one or more boundary layers. These phenomena have inspired the development of a sophisticated treatment of this kind of singular perturbations, the so-called matched asymptotic expansions technique. These techniques go far beyond the scope of this book. But, apart from the technical details, the concept of a transition region is of great importance from a modeler's point of view. This is because the existence of a transition region is the consequence of the fact that the dominant forces in the layer are essentially different from the dominant forces outside the layer. For an experienced modeler the physical interpretation and the presence of a small parameter in the model may never be separate phenomena. When a model is made dimensionless and a small parameter appears, the modeler should immediately associate this with properties of the system under consideration and be aware of the possible presence of transition regions. This insight will guide the modeler in how to solve the equations, both in an analytical and a numerical approach.

In the following simple example the flavor of the idea of a transition region and the power of the matched asymptotic expansions technique are nicely demonstrated.

Example 4.5b. An ultralight oscillator.
We consider a damped harmonic oscillator with nearly vanishing mass ε, described by

$$K(u, \varepsilon) := \varepsilon\,\ddot{u} + c\,\dot{u} + k\,u = 0\,, \quad u(0) = 1,\ \dot{u}(0) = 1. \tag{4.52}$$

This well-known equation of motion expresses the balance of three forces: the inertial force $\varepsilon\ddot{u}$, the friction force $c\dot{u}$, and the spring force ku. The mass ε is assumed to be very small.

This implies that the inertial force is relevant only if the acceleration $\ddot{u}$ is big. Since the oscillator is damped, this will not be the case in the long run. So, we expect that the inertia term can be ignored most of the time, perhaps with the exception of a transition window $[0, \delta]$ directly after the start. The necessity of the presence of such a transition region also follows from the fact that if we ignore the inertia term for all times, the two initial conditions can not be satisfied, since the reduced equation

$$c\,\dot{u}^o + k\,u^o = 0 \tag{4.53}$$

is of first order. The single initial condition for this reduced equation will be specified at $t = \delta$. As an approximation we assume that (4.53) describes the dynamics outside the transition region, and thus for $t \geq \delta$. The superscript in u^o refers to "outer solution." This equation expresses that outside the transition region the dynamics is determined by the balance of friction and spring force only. To find the relevant balance of forces in the transition region, we *stretch* this region by applying the transformation $t = \delta\tau$. The new time variable τ then varies over $[0, 1]$ throughout the region. After the transformation, (4.52) reads as

$$u'' + c\frac{\delta}{\varepsilon}u' + k\frac{\delta^2}{\varepsilon}u = 0 \tag{4.54}$$

with the derivatives taken with respect to τ. The key idea is that this equation must also hold in the limit $\varepsilon \to 0$, at least on the boundary where $\tau = 1$. To avoid divergence of some terms in that limit, we must assume that the thickness δ of the region is a function of ε. Intuition suggests that the smaller ε is, the thinner the region will be. Let us try the scaling relation $\delta \sim \varepsilon^p$, i.e., δ scales with some power of ε. The value of p cannot be smaller than one, since then the friction term in (4.54) would diverge if $\varepsilon \to 0$. A general guideline for the choice of p is to take the smallest value for which none of the coefficients in the "stretched" version of the equation of motion diverges in this limit. In (4.54) this implies the choice $\delta = \varepsilon$. The dynamics in the transition region are then characterized by a balance of the inertial and friction forces:

$$(u^i)'' + c\,(u^i)' = 0, \quad u^i(0) = 1,\ (u^i)'(0) = \varepsilon. \tag{4.55}$$

The superscript in u^i refers to "inner solution." Solving (4.55) we obtain

$$u^i(\tau) = \frac{\varepsilon}{c}(1 - e^{-c\tau}) + 1.$$

Having found the approximation u^i in the transition region, we have to match it to the approximative solution $u^o(t)$ outside the region. The transition between the two regions takes place at $t = \varepsilon$, i.e., $\tau = 1$. The outer solution u^o is found from solving (4.53) with as initial condition

$$u^o(t = \varepsilon) = u^i(\tau = 1) = \frac{\varepsilon}{c}(1 - e^{-c}) + 1 =: q\ .$$

This yields

$$u^o(t) = q\,e^{-\frac{k}{c}(t-\varepsilon)}.$$

So, eventually we have constructed the approximation

$$u(t) = \begin{cases} u^i(t/\varepsilon), & 0 < t \leq \varepsilon, \\ \\ u^0(t), & t \geq \varepsilon. \end{cases} \tag{4.56}$$

Note that instead of solving the complete model equation, we have solved two reduced equations. In the present example it is straightforward to solve the equation of motion directly, so there is no need to apply this technique. However, in general the reduced equations are easier to solve, and then this technique really exploits the presence of a small parameter. □

Exercise 4.5b.

The preceding example is somewhat artificial since the equation of motion (4.52) *is so simple that it can be solved directly. This allows for a comparison of the exact solution with the approximating solution* (4.56)*. Check that the exact solution of* (4.52) *is given by*

$$u(t) = a\, e^{q_+ t} - b\, e^{q_- t} \tag{4.57}$$

with

$$q_\pm = \frac{-c \pm \sqrt{c^2 - 4k\varepsilon}}{2\varepsilon},$$

$$a = \frac{1 - q_-}{q_+ - q_-},$$

$$b = \frac{1 - q_+}{q_+ - q_-}.$$

Plot expressions (4.56) *and* (4.57) *for parameter values* $c = k = 1$ *taking for* ε *a series of decreasing values:* $\varepsilon = 0.1$*,* $\varepsilon = 0.05$*,* $\varepsilon = 0.01$*, etc. What do you observe?*

In Example 4.5b we concluded with the presence of a transition region from physical reasoning: the presence of friction tempers the acceleration. To check this intuitive reasoning, in the following exercise we study the same system without friction. The important message is that a small parameter in front of the highest derivative does not necessarily always imply a transition region. So, trying to construct a transition region solution without insight into the phenomena represented by the model is a waste of time.

Exercise 4.5c.

Consider the undamped oscillator with nearly vanishing mass ε*:*

$$\varepsilon\, \ddot{u} + k\, u = 0, \quad u(0) = 1,\ \dot{u}(0) = 1.$$

a. *Check that the solution is given by*

$$u(t) = \cos\left(\sqrt{\tfrac{k}{\varepsilon}}\, t\right) + \sqrt{\tfrac{\varepsilon}{k}} \sin\left(\sqrt{\tfrac{k}{\varepsilon}}\, t\right).$$

b. *Check that the inertia term* $\varepsilon\ddot{u}$ *is never negligible for* $t > 0$.

In the following example we deal with the occurrence of a boundary layer, i.e., a transition region in space. The model is a boundary value problem, so boundary conditions are specified at both ends of a spatial interval.

Example 4.5c. Spread of a pollutant.
We consider a factory that pollutes a neighboring canal by draining off some material into the canal. See Fig. **4.8**. The pollutant diffuses through the canal and is at the same time destructed by some biological process. So, near the factory the water will be polluted, but far from it the pollution will no longer be detectable. We are interested in the question of what the typical distance is after which the concentration $u(x, t)$ of the pollutant has become negligibly small. Of course, this depends on the diffusion coefficient, the rate of drain-off from the factory, and the rate of destruction. If we model the diffusion of the pollutant with the help of the well-known diffusion equation derived in Example 2.3d, and model the destruction of the pollutant as exponentially decaying in time, we arrive at the following model:

$$\partial_t u = D\,\partial_x^2 u - k\,u \tag{4.58}$$

with boundary conditions (the factory is located at $x = 0$)

$$\partial_x u(0) = c, \quad u(\infty) = 0,$$

where c is the amount of material that per unit of time is drained off into the canal. We are looking for the stationary state of the model, which satisfies the equation

$$\varepsilon\,\partial_x^2 u - u = 0, \quad \partial_x u(0) = c, \quad u(\infty) = 0, \tag{4.59}$$

with $\varepsilon = D/k$. We assume that the diffusion is relatively slow compared to the destruction process, so we have $\varepsilon \ll 1$. We thus meet with a singularly perturbed equation.

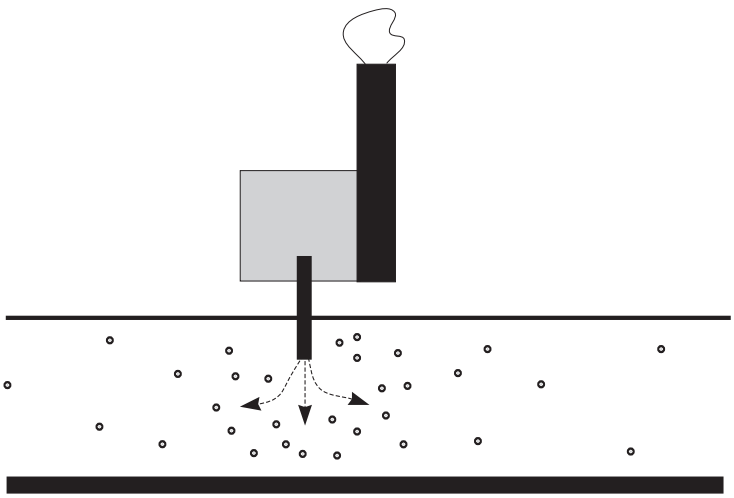

Figure 4.8. Sketch of a factory polluting a canal.

This simple equation of motion expresses the balance of the inertial term $\varepsilon\,\partial_x^2 u$ and the destruction term u. Since $\varepsilon \ll 1$, the inertial term is relevant only if the second derivative $\partial_x^2 u$ is big. The necessity of the presence of a boundary layer also follows from the fact that if we ignore the inertia term, the two initial conditions cannot be satisfied at the same time. Since the pollution has its origin at $x = 0$, we expect that there exists transition region $[0, \delta]$. We stretch this region by applying the transformation $x = \delta\, y$. The new variable y varies over $[0, 1]$ throughout the region. After the transformation, (4.59) reads as

$$\frac{\varepsilon}{\delta^2}\,\partial_y^2 u - u = 0, \quad \partial_y u(0) = \delta\, c, \quad u(\infty) = 0. \tag{4.60}$$

This equation must also hold in the limit $\varepsilon \to 0$. This suggests we try the scaling relation $\delta = \sqrt{\varepsilon}$. The inner solution u^i in the transition region is then

$$u^i(y) = c_1\, e^y + c_2\, e^{-y}.$$

The boundary condition at $x = 0$ yields that $c_1 - c_2 = \delta\, c$. Having found this approximation in the transition region, we have to match it to the outer solution $u^o(t)$ outside the boundary layer. This outer solution is directly found by setting $\varepsilon = 0$ and using the boundary condition at ∞. This yields that $u^o(t) := 0$. Matching then yields

$$u^o(x = \delta) = u^i(y = 1) = c_1\, e + c_2\, e^{-1} = 0.$$

So, we find that $c_1 = \delta\, c/(1+e^2)$ and $c_2 = -e^2\delta\, c/(1+e^2)$. Eventually we have constructed the approximation

$$u(x) = \begin{cases} u^i(x/\sqrt{\varepsilon}), & 0 < x \le \sqrt{\varepsilon}, \\ 0, & x \ge \sqrt{\varepsilon}. \end{cases} \tag{4.61}$$

Now we are able to answer the question about the maximal pollution distance: this information is given by the thickness of the boundary layer and is thus equal to $\sqrt{\varepsilon} = \sqrt{D/k}$. □

Exercise 4.5d.

Equation (4.59) *can also be solved directly. This allows for a comparison of the exact solution with the approximating solution. Write down the exact solution of* (4.59)*. Plot the exact solution and the approximative solution* (4.61)*, taking for* ε *a series of decreasing values. What do you observe?*

4.6 Challenging problems

4.6.1 Traffic flow

In Example 2.3b and Exercise 2.3b we met with a model for traffic flow. If we take a constitutive relation that models the response of anticipating drivers,

$$v(x, t) = \beta(\rho_m - \rho(x, t)) - \alpha\partial_x \rho(x, t), \tag{4.62}$$

the traffic density $\rho(x,t)$ is governed by the PDE

$$\partial_t \rho = -V(\rho)\,\partial_x \rho + \tfrac{1}{2}\,\alpha\,\partial_x^2(\rho^2) \tag{4.63}$$

with $V(\rho) = \beta(\rho_m - 2\rho)$ and β, ρ_m, and α given (positive) parameters. Their interpretation is given in Example 2.3b.

Exercise 4.6a.

a. *Show that any concentration ρ_0 that does not depend on x is a stationary state ρ_0. So, it also does not depend on t. We call such a state uniform in space and constant in time. What are the natural lower and upper bounds for ρ_0?*

b. *Linearize* (4.63) *around the uniform and constant state ρ_0.*

c. *Apply a coordinate transformation $(x,t) \to (y,t)$ with y defined as*

$$y = x - V(\rho_0)t.$$

Write the linearized equation, obtained under b *in terms of y and t. Interpret the linearized equation.*

d. *Consider an interval $0 \le y \le L$. Note that this window moves with velocity $V(\rho_0)$ along the road. At some moment, let a perturbation of ρ_0 be given by*

$$\rho(y) = \rho_0 + \sin\Big(\frac{my}{L}\Big).$$

How does this perturbation initially evolve in time?

4.6.2 Population models

We study the population dynamics of some species. For simplicity we assume that this species lives in one spatial coordinate (taken as the x-axis) and can migrate along this axis. One can think of an algae population in a pipe line.

Exercise 4.6b.

a. *Let us first ignore migration. The normalized population density $u(x,t)$ is assumed to satisfy the ODE ($\alpha > 0$)*

$$\partial_t u = \alpha u(1-u) := K(u). \tag{4.64}$$

This equation of motion expresses that u will grow if the density is small, apparently because enough food is available. However, if the population density becomes high, the restricted availability of food limits the growth of u. Sketch the vector field $K(u)$ as a function of u and determine the solution of (4.64) *for any $x \in [0,1]$ starting from some given initial profile $u_0(x) := u(x,0)$. Equation* (4.64) *can be solved via separation of variables, since for any x we have that the solution $u(x,t)$ satisfies the relation*

$$\int_{u_0(x)}^{u(x,t)} \frac{1}{u'(1-u')}\,du' = \alpha \int_0^t dt'$$

and the integral on the left-hand side can be easily evaluated (since $(u(1-u))^{-1} = u^{-1} + (1-u)^{-1}$).

b. *Find the equilibrium population densities, and determine their stability.*

c. *Now suppose that a spatially nonuniform population $u(x,t)$ will migrate in a way as modeled by Fourier's law for diffusion. See* §2.4. *Show that the following equation is then a simple model for the migration:*

$$\partial_t u = \kappa\, \partial_x^2 u,$$

where κ is a constant. What name would you assign to κ?

d. *Now combine both effects by simply adding the two terms in the vector field:*

$$\partial_t u = \kappa\, \partial_x^2 u + \alpha\, u(1-u).$$

What are the stationary densities of this model? Determine the linearization of the equation around each equilibrium solution.

f. *Investigate the linearized equations in* d *and derive the dispersion relations. Conclude which perturbations will grow in time. Investigate what happens with an arbitrary small, localized initial population.*

g. *Next, assume that the species lives on a bounded interval and that on the boundaries the concentration is kept at a zero level by some external influence. Investigate again what happens with an arbitrary small, localized initial density profile.*

4.6.3 Lindstedt's method

We consider the oscillations of an oscillator with nonlinear spring constant. The nonlinearity is represented by an extra term in the equation of motion:

$$\ddot{x} + x - \varepsilon x^3 = 0.$$

Exercise 4.6c.

a. *Sketch the vector field in the phase plane for $\varepsilon = 0$ for the case of spring softening $\varepsilon > 0$ and the case of spring hardening $\varepsilon < 0$. Conclude from the vector field that all bounded solutions are periodic.*

b. *The solution $x_\varepsilon(t)$ with initial conditions*

$$x^\varepsilon(0) = A,\ \dot{x}^\varepsilon(0) = 0$$

is bounded and periodic if A is small enough. Determine the solution $x^0(t)$, so with $\varepsilon = 0$, and its frequency.

c. *If $\varepsilon \neq 0$, the frequency of the vibration will depend on the amplitude A. Substitute a regular expansion*

$$x^\varepsilon(t) = x^0(t) + \varepsilon\, x_1(t) + \varepsilon^2\, x_2(t) + \cdots$$

and show that already in first order in ε resonance is observed (which means that the series does not approximate the periodic $x^\varepsilon(t)$ uniformly on $t \geq 0$).

d. *Anticipating that $x^\varepsilon(t)$ is periodic, scale time t with the (still unknown) frequency ω, $\tau = \omega t$ and write $\omega = 1 + \varepsilon\,\omega_1 + \varepsilon^2\,\omega_2 + \cdots$. Derive the scaled equation using τ and substitute the series for x^ε and ω.*

e. *Show that* $\cos^3 t = \frac{3}{4}\cos t + \frac{1}{4}\cos 3t$. *Now solve the equations in* d *up to first order in ε. Return to the variable t to find the first-order approximation of the periodic solution x_ε and show that*

$$\omega = 1 - \tfrac{3}{8}\,\varepsilon\,A^2 + \mathcal{O}(\varepsilon^2)$$

gives the dependence of the frequency on the amplitude.

Chapter 5
Variational Modeling

In the previous chapters we have seen various models that describe a large variety of natural, physical, and technical problems. The models, of different complexity, were typically formulated in terms of differential equations for state variables that were functions of space and time. To design the models, we used physical laws and balance and conservation arguments, as described in Chapter 2.

This chapter deals with different important modeling tools and arguments. In the first section we describe how many problems from the natural sciences are governed by a variational principle. This means that the physical state of such a system is completely determined by the property that a specific functional achieves a lowest value (or a critical value) on some constraint set. For most applications the domain on which the functional has to be investigated is a function space and thus is infinite dimensional. For illustrational purposes we introduce the formulation of the problem of the hanging chain in some detail. For this system, the set of all admissible shapes consists of curves that have given length and connect given supports at its end. The actual shape of the hanging chain, called catenary, is such that its potential energy is as small as possible. This corresponds to a shape with the center of gravity as low as possible.

In §5.2 we discuss that there is a natural, variationally consistent way to simplify a given variational problem by approximating the functional, or the set on which the functional has to be investigated. Using expansions in terms of basis functions, we may reduce high-dimensional models to low-dimensional ones. The same idea can lead to efficient numerical algorithms. The best-known examples are Fourier theory and finite elements, which are discussed in more detail in §5.5.

Section 5.3 links the variational formulation to equations that have to be satisfied by the minimizer of the functional. We will restrict our attention to essentials and treat the so-called direct problem: we show that Fermat's algorithm (i.e., the derivative of a function vanishes in a point where it achieves a minimal, or critical, value) can be generalized to infinite-dimensional variational problems. For the functionals that are most often encountered, the so-called density functionals, this equation is an ordinary differential equation (ODE) or partial differential equation (PDE), called the Euler–Lagrange equation.

In §5.4 we reconsider the idea of variational restriction and explain the consequences for the equation. In §5.5 we briefly treat some well-known numerical methods in the

variational perspective and design the algorithms in a constructive way by variational restriction.

5.1 Variational principles

We start this section with some historical remarks. After that, we give the general formulation of variational principles and illustrate the ideas by discussing the principle of minimal potential energy.

5.1.1 Optimality in the natural sciences

> *Je suis convaincu que par tout la nature agit selon quelque principe d'un maximum ou minimum.* (Euler, 1746)[1]

This quotation of one of the greatest scientists who shaped the modern mathematical description and investigation of the natural sciences clearly expresses the variational way of thinking. The belief that optimization was important to describe natural phenomena was verified and exploited by Euler for various problems, in order to present a more thorough investigation of the problems. Farther-reaching conclusions were drawn by another scientist:

> *Des loix du mouvement ou l'action est toujours employee avec la plus grande economie, demontreront l'existence de l'Etre supreme.* (Maupertuis, 1757)[2]

But this point of view belongs to metaphysics and as such is not very fruitful for a deeper investigation of natural phenomena.

Optimization problems have been known since ancient times. Well known is *Dido's problem*. According to the legend, some inhabitants of North Africa offered queen Dido all the land she could surround with the hide of a bull. She allegedly maximized her real estate holdings and founded the city of Carthago by slitting the hide into a very long thin strip which she then formed into a circle. Many other problems can be formulated as *geodetic problems*, where one investigates those curves (or surfaces) with the property that a functional measuring the length (or the area) is as small or large as possible. A major example is the following:

> **Fermat's principle** (1662): *The actual trajectory of a light ray between two points in an inhomogeneous medium has the property that the time (or optical length) required to transverse the curve is as small as possible when compared to the time required for any other curve between the points.*

In fact, the investigation of this principle led Fermat to the mathematical result that will be stated below as Fermat's algorithm. From this principle, *Snell's law* about the refraction of light at the interface between two media can be derived. An alternative point of view (looking for the evolution of light fronts, the surfaces that can be reached by the light from a point source in a give time) was investigated by Huygens in 1695. *Huygens' principle*,

[1] I believe that everywhere in nature some maximum or minimum principle is in force.

[2] The laws of motion or action are always employed in the most economical sense, showing the existence of the Supreme Being.

of vital importance for the basic understanding of light propagation, can be considered as a major example of what has since become known as a duality method.

For dynamical systems without dissipation, there is a dynamic variational formulation known as the *principle of stationary action*. This is studied for finite dimensional systems (described by ODEs) in classical mechanics. We will encounter these so-called Hamiltonian and Poisson systems in §6.3. All these notions can be generalized to infinite dimensions; the governing equations are then PDEs. For instance, basic laws such as the Maxwell equations of electromagnetism and the Euler equations of fluid dynamics can be shown to have similar variational principles and a Hamiltonian or Poisson structure. We will see examples of wave equations describing surface water waves in §6.3. For more general dynamic principles see [13].

The historical remarks above reveal that although the analytical methods of the classical calculus of variations were developed in the eighteenth century by scientists like Newton, Euler, and Lagrange, some basic ideas could already be found in the seventeenth century. It is clear that practical problems from physics provided the initial motivation for the beautiful mathematical theory that has been developed since then. The interested reader may consult, e.g., [11, 2, 14].

5.1.2 General formulation of optimization problems

Most variational principles are based on an optimization problem. Such an optimization problem has the following basic ingredients:

- A set of so-called admissible elements $\mathcal{M}$, usually some subset of a finite or infinite-dimensional state space $\mathcal{U}$. These are the feasible states that are taken into account to find the actual optimal state.
- A functional $\mathcal{L}$, defined on $\mathcal{U}$ (or only on $\mathcal{M}$).

The optimization problem of $\mathcal{L}$ on $\mathcal{M}$ can then generally be described as the problem to find, or characterize, the elements $\hat{u}$ that minimize the functional on $\mathcal{M}$. We may denote the optimization problem symbolically as

$$\hat{u} \in \min \ \{ \mathcal{L}(u) \mid u \in \mathcal{M} \},$$

and the actual minimizers $\hat{u}$ are the elements for which, by definition,

$$\mathcal{L}(\hat{u}) \leq \mathcal{L}(u) \ \text{ for all } \ u \in \mathcal{M}.$$

Example 5.1a. Least squares method.
This mathematically motivated problem is a prime example of an often-used optimization problem. The method is often motivated by the desire to bring order into a set of measurements of some system. Bringing order in the data means to find a (simple) relation between the variables, and hence implicitly to make a model from the data. For example, let there be given a sequence of two variables, say, (x_k, y_k), $k = 1, \ldots, N$. The variables x_k might denote successive times at which the temperatures y_k are measured. When plotted in the (x, y) plane, these measurements may suggest a certain relationship, such as an exponential

or linear temperature increase. In most cases the measurements will have experimental errors. To get more quantitative information, it is then desired to find an exponential or linear relation (the model) that fits the data best. It is essential to specify what "best" means. For instance, when errors are expected to occur only in the y-quantity, a natural approach is to formulate the following minimization problem:

$$\min_u \{\mathcal{L}(u) | u \in \mathcal{M}\}$$

with

$$\mathcal{L}(u) = \sum_k |y_k - u(x_k)|^2 ,$$

where $\mathcal{M}$ is the set of model functions. The choice of the distance function measuring the distance between the data points and the line is in some sense free. The present choice of the sum of squares is motivated by the fact that it allows for an elegant, explicit expression for the parameters of the model functions in terms of the data.

A general strategy is to choose $\mathcal{M}$ such that its elements are parameterized functions with the number of parameters as small as possible. Often such parameters are called *collective coordinates*: they specify the particular element from the restricted class of shape functions. In §5.2 we will come back to this approach. The advantage is that the resulting optimization problem is finite dimensional, since the corresponding model manifold is finite dimensional.

For instance, for fitting a straight line through data points, $\mathcal{M}$ is the two dimensional set $\mathcal{M} = \{u(x) = a\,x + b \mid a, b \in \mathbb{R}\}$. This leads to the minimization problem

$$\min_{a,b} \left\{ \sum_k |y_k - (a\,x_k + b)|^2 \mid a, b \in \mathbb{R} \right\} .$$

When fitting an exponential curve, one takes for $\mathcal{M}$ the three-dimensional set

$$\mathcal{M} = \left\{ u(x) = a\,e^{b\,x} + c \mid a, b, c \in \mathbb{R} \right\} .$$

A similar but usually high-dimensional case is obtained if we want to approximate a given function on an interval, $[0, L]$, say, by a linear combination of basis functions. For instance, Fourier theory aims to approximate a given function $f(x)$ by a truncated series of harmonic functions:

$$u(x) = \sum_{k=0}^{N} \left[a_k \cos(2\pi x/L) + b_k \sin(2\pi x/L) \right] .$$

The complete Fourier representation requires the limit $N \to \infty$, so that $\mathcal{M}$ is infinite dimensional, but in practice one usually cuts off the series at some $N < \infty$ so that $\mathcal{M}$ is finite dimensional. □

5.1.3 Principle of minimal potential energy

By way of illustration we will now deal with the principle of minimal potential energy (PMPE). It is a special case of a variational principle (also called "action principle") and

it does not hold in general. For steady states, the kinetic energy of a system vanishes and the total energy is the potential energy. The actual physical state is then described by the PMPE. This means that on a set of admissible states $\mathcal{M}$, which are the physically acceptable states that do not violate given constraints, the potential energy functional $\mathcal{V}$ assigns a value $\mathcal{V}(u)$ to each state $u \in \mathcal{M}$, and the physical state attained in practice is the state $\hat{u}$ that minimizes $\mathcal{V}$ on $\mathcal{M}$. We introduce this idea by means of the example of a hanging chain. In the following this system will be used several times to illustrate variational principles. Its explicit solution will be given in Example 5.3c.

Example 5.1b. Catenary: The hanging chain.
Consider a single mass particle under the influence of gravity. It will fall down (dynamic state) until it reaches the floor (static state). Its potential energy is proportional to the height, and in the steady state the height of the particle is as low as possible.

If many mass particles are connected to each other to form a chain, the principle describes that the actual form of the chain, when fixed at the endpoints, will be such that its total potential energy is as low as possible. Finding the actual form is the famous problem of the hanging chain. See Fig. **5.1**. We will describe this in some detail by looking at a simple model for the chain.

The first assumption is to consider the chain to be inextensible. Its length is thus fixed at $2L$, say. The other assumption is that the chain can bend without any effort, i.e., is completely flexible. To simplify matters, we assume that the chain has uniform mass density ρ and is very thin, i.e., we may neglect its thickness with respect to its length. If we

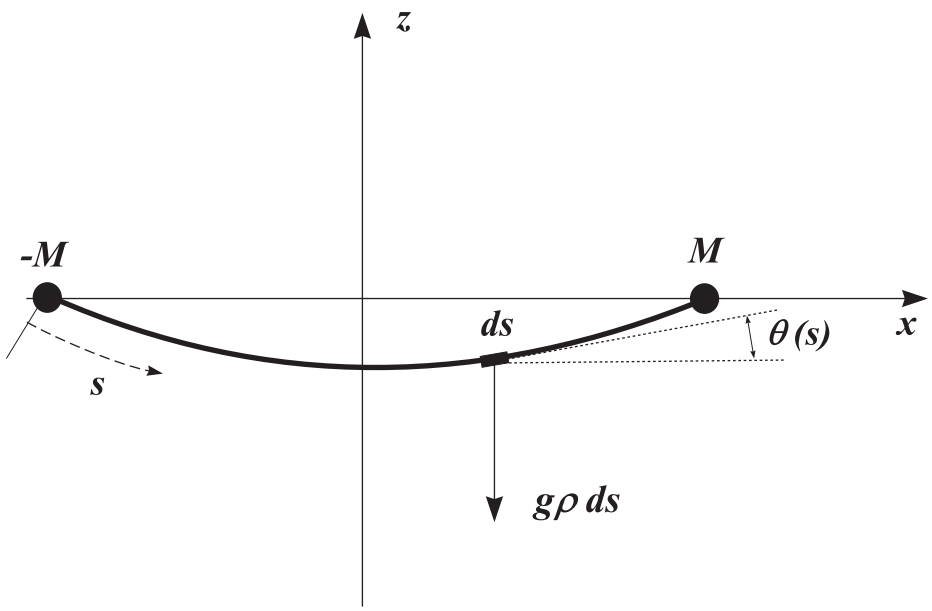

Figure 5.1. Under influence of gravity, a hanging chain attains a shape referred to as "catenary."

fix its two endpoints, the actual shape of the chain under gravitation is such that its potential energy is as low as possible. This is equivalent to saying that its center of gravity is as low as possible. Under the assumptions stated above, the shape is called a *catenary*.

We now describe the optimization problem in explicit formulas. Given the positions of the two fixed endpoints, it is natural to take the vertical plane (i.e., parallel to gravity) through the supports as domain of the chain. The vertical axis is taken as the z-axis and the horizontal as the x-axis. For simplicity we take a symmetric situation, with the supports at coordinates $(\pm M, 0)$ with, of course, $0 < M < L$. The shape of the chain curve can be described as a parameter curve $s \to (x(s), z(s))$, where s is the arc length, running over the interval $[-L, L]$. For a mass element of infinitesimal length ds at height z under the influence of gravity, with acceleration g, the potential energy is $\rho\, g\, z\, ds$, and so the total potential energy is given by

$$\mathcal{V} = \int_{-L}^{L} \rho\, g\, z(s)\, ds.$$

The set of admissible elements here are all curves of given length $2L$ that connect two given points, so

$$\mathcal{M} = \Big\{ (x(s), z(s)) | s \in [-L, L],\ x(\pm L) = \pm M,\ z(\pm L) = 0, \\ \sqrt{(\partial_s x)^2 + (\partial_s z)^2} = 1 \Big\}, \tag{5.1}$$

where the condition $\sqrt{(\partial_s x)^2 + (\partial_s z)^2} = 1$ guarantees that the curve is inextensible and thus has fixed length. The variational principle reads as

$$\min_{x,z} \ \{ \mathcal{V}(x, z) \mid (x, z) \in \mathcal{M} \}.$$

Other, equivalent, formulations are possible. For instance, we could write the shape as the graph of a function $z(x)$. This requires the reasonable assumption that the shape is smooth and can be described as a function of x. Then, the set of admissible functions consists of functions $z(x)$ defined on the interval $[-M, M]$ satisfying the conditions

$$\mathcal{M} = \left\{ z(x) | \ \int_{-M}^{M} \sqrt{1 + (\partial_x z)^2}\, dx = 2L,\ z(\pm M) = 0 \right\}, \tag{5.2}$$

and the variational formulation then reads as

$$\min_{z} \ \left\{ \ \int_{-M}^{M} \rho\, g\, z(x) \sqrt{1 + (\partial_x z)^2}\, dx \mid z \in \mathcal{M} \right\}. \tag{5.3}$$

□

Exercise 5.1a.

a. *Exploit the symmetry by introducing the angle $\theta(s)$ of the tangent with the horizontal axis as a new variable. See Fig.* **5.1**. *Then it holds that $\partial_s x = \cos\theta$, $\partial_s z = \sin\theta$, and $\theta(0) = 0$. Show that the potential energy can be written, using $z(M) = 0$ and applying partial integration, as*

$$\mathcal{V} = -2\rho\, g \int_0^L s \sin\theta\, ds,$$

and that the condition $x\,(L) = M$ transforms into

$$\int_0^L \cos\theta \, ds = M.$$

Hence, with $\theta = \theta\,(s)$ as a dependent variable, the problem can be formulated as

$$\min_{\theta} \left\{ -2\,\rho\,g \int_0^L s \sin\theta \, ds \mid \int_0^L \cos\theta \, ds = M,\ \theta(0) = 0 \right\}.$$

b. *Take any two points of a hanging chain as new points of support. Argue that the catenary shape remains the shape that is found from a PMPE formulation with the new support points. Conclude that each part of an optimal curve for a PMPE is also an optimal curve for a local (partial) PMPE. Is the reverse also true, i.e., is a curve consisting of a combination of local optimal curves also globally optimal?*

Example 5.1c. Gaudi's engineering reflection principle.
Antonio Gaudi (1852–1926) was the architect of the Sagrada Familia in Barcelona, Spain, a huge cathedral of which the construction is not yet finished. Gaudi used many mathematical ideas in the construction, one of which is the PMPE. See [7]. He realized that in equilibrium the hanging chain has a shape in which only tension forces are present. As a thought experiment he reversed the direction of gravity and concluded that then only compression forces will be present in the chain. So, wanting to construct lightweight arches, he decided that the optimal arches have the shape of a catenary but now reflected in the x-axis.

At the time of Gaudi computers were not available, and no extensive calculations or simulations were possible. Yet he simulated the construction in an experimental way, using directly the idea of PMPE with the addition of reversion. In Fig. **5.2**. the accompanying picture of a (reversed) model of the cathedral is given, which was used in practice. In this hanging model a system of threads represent columns, arches, walls, and vaults. Sachets with lead shot resemble the weight of small or larger building parts. From these experiments it is possible to deduce how the whole structure should hang to ensure that a global dynamic equilibrium is obtained. In the actual construction phase, parts of the construction are being built at different times. Note that each part of the construction should be optimally constructed, since only then will the whole construction be optimal. □

5.2 Variational approximation and restriction

5.2.1 General methodology

Consider a problem described by a variational principle

$$\text{Crit } \{\, \mathcal{L}(u) \mid u \in \mathcal{M} \,\}.$$

Here, Crit means that all stationary (or critical) points of $\mathcal{L}$ are sought. In such points the partial derivatives of $\mathcal{L}$ vanish. In §5.3.2 we will describe how this has to be interpreted if $\mathcal{L}$ is a functional. Suppose that this exact formulation is too complicated to be studied and

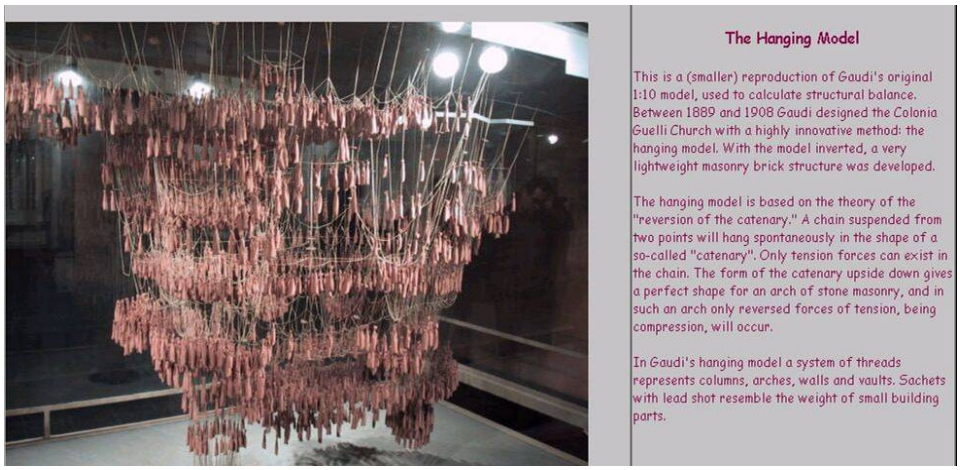

Figure 5.2. Reverse model used by Gaudi to design the shape of the cathedral in Barcelona.

we aim to simplify it. This can be done by either simplifying the functional $\mathcal{L}$ or changing the model manifold $\mathcal{M}$, or both. In doing both we replace the exact variational principle by another, simplified one:

$$\text{Crit } \{ \mathcal{L}_a(u) \mid u \in \mathcal{M}_a \}.$$

The quality of this approximate variational principle as a model for the original problem depends of course on the properties of the chosen approximate functional $\mathcal{L}_a$ and manifold $\mathcal{M}_a$. No general statements can be made without restriction to special cases, and even for a specific problem and approximation, the relation between the solutions of both variational problems may be difficult to investigate.

When we are dealing with an explicit *minimization* problem, there is a direct monotonicity result for the optimal value under severe restrictions. Namely, if $\mathcal{M}_a \subset \mathcal{M}$ and $\mathcal{L}(u) \leq \mathcal{L}_a(u)$ for all $u \in \mathcal{M}$, then it is obvious that

$$\min \{ \mathcal{L}(u) \mid u \in \mathcal{M} \} \leq \min \{ \mathcal{L}_a(u) \mid u \in \mathcal{M}_a \}.$$

In particular, when only $\mathcal{M}$ is changed but $\mathcal{L} = \mathcal{L}_a$, the difference in the optimal values may be used as a measure for the accuracy of the approximation. See also §5.4.2. Intuitively speaking, the smaller the difference, the better the approximate solution approximates the exact optimal solution. With some care the approximate model may lead to good approximations.

The approximation of the functional may be motivated by physical assumptions. For instance, in the example of PMPE, the potential energy may be approximated based on some simplification of the physical description. We illustrate the idea of simplifying the functional for the problem of the hanging chain.

Example 5.2a. Small deflections of the hanging chain.
We consider the catenary problem in the case in which the chain is just a little bit longer than the length between the supports. Then the chain will deviate only a little from the

horizontal axis. This will imply that we essentially may linearize the governing equation. The start is the original formulation (5.3). The special case we want to consider is characterized by

$$\frac{L}{M} = 1 + \varepsilon^2$$

with $0 < \varepsilon << 1$. Giving the chain a triangular shape and applying Pythagoras's theorem, it is easily deduced that the maximum possible distance between chain and horizontal axis is about $\varepsilon\, M$. In the stationary state, the chain will have a smooth shape without wrinkles. This implies that necessarily $|\partial_x z|$ is of order ε and we may write

$$\sqrt{1 + (\partial_x z)^2} \approx 1 + \frac{1}{2}\,(\partial_x z)^2 = 1 + \mathcal{O}(\varepsilon^2)\,.$$

This motivates us to approximate the potential energy given in (5.3) as

$$\int_{-M}^{M} \rho\, g\, z(x) \sqrt{1 + (\partial_x z)^2}\, dx \approx \int_{-M}^{M} \rho\, g\, z(x)\, dx$$

and the length constraint given in (5.2) as

$$\int_{-M}^{M} \sqrt{1 + (\partial_x z)^2}\, dx \approx 2M + \int_{-M}^{M} \frac{1}{2}\,(\partial_x z)^2\, dx,$$

so

$$\int_{-M}^{M} \frac{1}{2}\,(\partial_x z)^2\, dx = 2\varepsilon^2.$$

We thus arrive at the approximated problem

$$\min_z \left\{ \int_{-M}^{M} \rho\, g\, z(x)\, dx \;\middle|\; \int_{-M}^{M} \frac{1}{2}\,(\partial_x z)^2\, dx = 2\varepsilon^2 \right\} \tag{5.4}$$

for a fixed value of ε. □

5.2.2 Low- and high-dimensional restrictions

Instead of simplifying the functional $\mathcal{L}$, much more often we will deal with the case that we want to change the set $\mathcal{M}$ of admissible elements. The choice of the subset depends on the purpose we have in mind.

If we want to design a numerical algorithm for accurate calculations, the subspace is chosen to be high-dimensional. It will consist of finite-dimensional representations of the original state variables, for instance, finite Fourier truncations or representation with finite elements. Usually the dimension of the space is taken as a parameter, expecting that for increasing dimension the approximation will improve.

When the aim is to obtain a simple description of the qualitative behavior, a very low dimensional manifold is preferable, and the choice of the manifold will already reflect important aspects of the phenomena one is interested in. If, for instance, one is interested in describing soliton-like phenomena (i.e., localized solutions that propagate nearly undeformed), a low-dimensional Fourier truncation would be useless, but a description with

Gaussian profiles may be appropriate. Then, the height and width of the Gaussians act as parameters, thus defining a two- (parameter) dimensional manifold.

Both cases described here, low- and high-dimensional modeling, can be seen as specifying the manifold by a set of parameterized functions. When we use the notation $U(\mathbf{p})$ for the basis functions, where $\mathbf{p}$ stands for a collection of parameters, the manifold is given by

$$\mathcal{M} = \{\, U(\mathbf{p}) \mid \mathbf{p} \,\}.$$

High-dimensional Fourier or finite element representations can be written in this form. Then, the parameters are the coefficients in a linear representation of the functions with respect to certain basis functions ϕ_k:

$$U(\mathbf{p}) = \sum_k p_k \, \phi_k$$

with $\mathbf{p} = (p_1, p_2, \ldots)$. Any choice of sets that form a complete set of basis functions will lead to the Ritz–Galerkin type of discretizations that will be described in §5.4.1 of this chapter.

Low-dimensional models use specific functions depending only on a few parameters. For instance, the manifold of Gaussian-shaped functions, with height and width as collective coordinates, is given by

$$\mathcal{M} = \left\{\, A \exp\left[-x^2/\sigma^2\right] \;\middle|\; A, \sigma \in \mathbb{R} \right\}. \tag{5.5}$$

The parameters can in fact also include functions. For example, using such Gaussians to approximate a heat conducting process with a localized heat source in an infinitely long iron rod, the parameters A, σ will have to be functions of time.

Example 5.2b. The catenary and global trial functions.
From experience we guess that the shape of the catenary will be roughly something like a cosine shape or some suitable polynomial. We specify this guess by finding subsets of the original infinite-dimensional space of shape functions that are described as suitable few-parameter families of functions. Taking into account the constraints and the expected symmetry, we could take trial functions of the form $z(x) = -a\cos(bx) - c$, where a, b, c are the parameters to be estimated. The boundary conditions fix the value of c, but a and b are free parameters:

$$z(x; a, b) = a\,[\cos(b\,M) - \cos(b\,x)]. \tag{5.6}$$

In view of (5.2), the length constraint of the catenary is given by the integral

$$I(a, b) := \int_{-M}^{M} \sqrt{1 + a^2 b^2 \sin^2(bx)}\, dx.$$

Substituting (5.6) in (5.3), we find for the functional to be minimized the expression

$$\begin{aligned} \mathcal{L}(a, b) &:= \int_{-M}^{M} \rho\, g\, z(x) \sqrt{1 + (\partial_x z)^2}\, dx \\ &= \rho\, g \int_{-M}^{M} a\,[\cos(b\,M) - \cos(bx)] \sqrt{1 + a^2 b^2 \sin^2(bx)}\, dx\,. \end{aligned} \tag{5.7}$$

So, we arrive at the following two-parameter optimization problem, which is the restricted version of (5.3):

$$\min_{a,b}\{\ \mathcal{L}(a,b)\mid I(a,b)=2L\ \}\ . \tag{5.8}$$

This illustrates that by restricting the manifold by means of trial functions, the infinite-dimensional problem becomes finite dimensional. It is an example of the use of *global basis functions*. □

Exercise 5.2a.

a. *Formulate with the same set of trial functions the restriction of the simplified formulation* (5.4) *in Example* 5.2a*. Calculate all integrals by hand, and solve the optimization problem.*

b. *Also solve problem* (5.8) *using these trial functions, and show that the solution under* a *is a good approximation of the solution under* b *for small deflections.*

Example 5.2c. Finite element method.
In this example we present the basic idea of the finite element method. Instead of the (global) cosine profiles used in Example 5.2b, let us approximate the shape by a simple piecewise linear function with three parameters a_0, a_+, and a_-:

$$u = a_0\,T_0\,(x) + a_-\,T_-\,(x) + a_+\,T_+(x), \tag{5.9}$$

where T_0, T_+, and T_- are the tent functions of unit height defined on intervals of length M:

$$T_0\,(x) = \begin{cases} (2x+M)\,/M & \text{for } x\in[-M/2,\,0],\\ (M-2x)/M & \text{for } x\in[0,\,M/2],\\ 0 & \text{for } x\notin[-M/2,\,M/2], \end{cases}$$

and $T_\pm(x) = T_0\,(x\pm M/2)$. See Fig. **5.3**. These basis functions are not differentiable on the whole interval. Yet, their piecewise constant derivative is defined everywhere except at a few points, and we can give a well-defined meaning to the value of the functionals restricted to this set. All relevant expressions are then dependent on the coefficients (a_0, a_+, a_-). This thus leads to a constrained three-dimensional optimization problem.

The choice of the special basis functions here is an example of the use of finite elements. The terminology "finite element" refers to the intervals of support of the chosen basis functions. The basis functions are now local, vanishing outside the single element on which they are defined. Taking smaller elements than done here, this turns into one of the most powerful numerical methods currently available. For an accurate scheme, many elements are being used (a finer grid), which corresponds to increasing the dimension of the restricted manifold, leading to a better approximation of the exact solution. See also §5.5. □

Exercise 5.2b.

a. *Substitute the basis functions* (5.9) *in* (5.2) *and* (5.3) *and calculate the restricted functions for the catenary problem explicitly and solve the optimization problem.*

b. *Using symmetry, define the problem now on the half interval* $[0, M]$ *with a suitable boundary condition at* $x = 0$*. Treat the problem in a similar fashion as under* a *but now with slightly adjusted basis functions.*

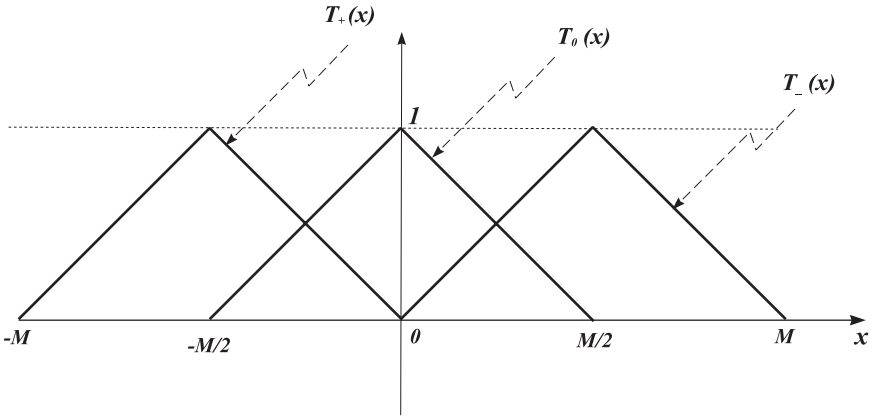

Figure 5.3. Tent functions are examples of localized basis functions used in the finite element method.

5.3 Variational calculus

The general structure of variational problems given above is completely determined by the functional $\mathcal{L}$ and the set $\mathcal{M}$ on which the minimizing (or critical) points of the functional are to be sought. This formulation is very different from problems that are described by explicit relations between variables, such as algebraic or differential relations. Yet we know from elementary calculus that at a point where a function of one variable is minimal, the derivative of the function vanishes. Hence we get an equation for the optimal solution. For infinite-dimensional optimization problems with density functionals that are differentiable a similar result holds: the optimal element will satisfy some (ordinary or partial) differential equation, the so-called Euler-Lagrange equation.

5.3.1 Finite-dimensional problems

The basic idea of finding stationary points of finite-dimensional problems is treated in any real analysis course. Here, we summarize the techniques and main results for convenience.

Unconstrained problems
A standard result in one dimension for any local investigation is the following.

> **Algorithm of Fermat:** *If a differentiable function $f : \mathbb{R} \to \mathbb{R}$ attains an extreme value at the point $\hat{x}$, then its derivative vanishes at that point:*
>
> $$f'(\hat{x}) = 0.$$

Fermat did not write this equation. He reasoned that a small variation Δx in x near a minimizer produces a variation in the function value that is quadratic in orders of Δx. This fundamental insight justifies attaching his name to this mathematical algorithm. Fermat did not know the concept of derivative of functions other than polynomials. It was Leibniz who

introduced in 1684 the concept of derivative of an arbitrary function. In modern notation, this algorithm results from the observation that

$$f(x + \Delta x) - f\,(x) = f'\,(x)\;\Delta x + O\left(\Delta x^2\right).$$

The condition $f'(\hat{x}) = 0$ characterizes a stationary (or critical) point. For a minimization problem this condition is necessary but not sufficient, since maximizers and saddle points satisfy this condition just as well.

Knowing the above result for functions of one variable, the generalization to functions of more variables, n-dimensional problems, is simple. For $F : \mathbb{R}^n \to \mathbb{R}$, the gradient ∇F is defined with the help of the *nabla* operator (see also §2.5) as the column vector

$$\nabla F(\mathbf{x}) := \begin{pmatrix} \partial_{x_1} F(\mathbf{x}) \\ \ldots \\ \partial_{x_n} F(\mathbf{x}) \end{pmatrix}.$$

The (Frechet) derivative of functions defined on $\mathbb{R}^n$ has already been defined in §3.3.1 as

$$F'(\mathbf{x};\,\mathbf{y}) := F'(\mathbf{x}) \cdot \mathbf{y} \equiv (\nabla F(\mathbf{x}),\;\mathbf{y}),$$

which is the directional derivative of F at the point $\mathbf{x}$ in the direction of $\mathbf{y}$. Note that this implies that in $\mathbb{R}^n$ an inner product $(.,.)$ has been chosen.

For $\hat{\mathbf{x}}$ to minimize F on $\mathbb{R}^n$, a necessary condition is that the directional derivative vanishes in any direction $\mathbf{y}$:

$$\nabla F(\hat{\mathbf{x}}),\;\mathbf{y} = 0.$$

Since $\mathbf{y}$ is arbitrary here, we arrive at the condition

$$\nabla F(\hat{\mathbf{x}}) = \mathbf{0}.$$

This is the direct generalization of Fermat's algorithm to n dimensions.

Constrained problems: Lagrange multipliers

When the constraint set is not the whole space but a lower-dimensional manifold, variations can be taken only in directions tangent to the manifold: the admissible variations are restricted. At an optimal solution, the directional derivative of the functional is only known to vanish in these tangent directions. Information about the other directions cannot be obtained. When the manifold is explicitly given by a level set of a differentiable function, say,

$$\mathcal{M} = \{\mathbf{x} \mid G(\mathbf{x}) = 0\,\},$$

the tangent directions ($\mathbf{y}$) are determined by $\nabla G\,(\mathbf{x}) \cdot \mathbf{y} = 0$. Hence there are no admissible variations perpendicular to the constraint manifold, i.e., in the direction $\nabla G\,(\mathbf{x})$. For an optimal solution of F on this set, we then have $\nabla F\left(\hat{\mathbf{x}}\right) \cdot \mathbf{y} = 0$ for all variations for which $\nabla G\left(\hat{\mathbf{x}}\right) \cdot \mathbf{y} = 0$. This leads to the conclusion that $\nabla F\left(\hat{\mathbf{x}}\right)$ can have a component only in the normal direction $\nabla G\left(\hat{\mathbf{x}}\right)$. Stated differently, for some λ it should hold that

$$\nabla F\left(\hat{\mathbf{x}}\right) = \lambda\,\nabla G\left(\hat{\mathbf{x}}\right).$$

Here, λ is the so-called Lagrange multiplier. This result is often formulated as the following.

> **Lagrange multiplier rule:** *A solution of a constrained problem with functional F and constraint function G is also a solution of a corresponding unconstrained problem with functional $\bar{F}(\mathbf{x}, \lambda) := F(\mathbf{x}) - \lambda\, G(\mathbf{x})$, since this leads to the same system of equations $\nabla F(\hat{\mathbf{x}}) = \lambda\, \nabla G(\hat{\mathbf{x}})$ and $G(\hat{\mathbf{x}}) = 0$.*

Extension to the case that the constrained set is an intersection of various level sets is possible, and we get just as many Lagrange multipliers as the number of level sets involved. This result directly generalizes to infinite dimensions, which we formulate below.

5.3.2 Basic notions of variational calculus

Now we generalize the variational calculus from finite- to infinite-dimensional problems. All these notions are part of the classical calculus of variations.

We refer to a mapping from a function space, thus an infinite-dimensional space, to the real numbers as a *functional*. The basic properties for mappings defined on finite-dimensional spaces can be generalized to functionals with the following guidelines:

- As already defined in §3.3.1, the directional derivative can be defined just as easily as for the finite-dimensional case by restricting the functional to one-dimensional lines. The directional derivative is also called the *first variation*.
- Generalization of the gradient concept leads to the notion of *variational derivative*, to be introduced below. The specific expression depends on the choice of the L_2-inner product for functions under consideration.

Just as in finite dimensions, the second derivative provides insight into the character of a critical point and thus reflects minimization properties. In the calculus of variations these aspects are dealt with in the *theory of first and second variations*. Now we will show some details of the generalization to infinite-dimensional problems.

Variational derivative

The directional derivative of functions defined on function spaces was introduced in §3.3.1. Here we discuss this concept a bit more extensively. We write the functional depending on (scalar or vector) functions belonging to some space $\mathcal{U}$ as $\mathcal{L} : u \to \mathcal{L}(u)$, for $u \in \mathcal{U}$. An *admissible variation* is an element, classically denoted by δu, such that $u + \varepsilon\, \delta u$ belongs again to $\mathcal{U}$ (or its tangent space at the point u, in case the set $\mathcal{U}$ is a manifold). Then the directional derivative is called the *first variation* of the functional at the point u in the direction δu. It is defined in the usual way as

$$\delta\mathcal{L}(u; \delta u) := \lim_{\varepsilon \to 0} \frac{\mathcal{L}(u + \varepsilon\, \delta u) - \mathcal{L}(u)}{\varepsilon}.$$

Note that the only technical action needed to calculate this expression is differentiation with respect to the scalar variable ε. In most cases the resulting functional can be written as the inner product of an expression with the variation

$$\delta\mathcal{L}(u; \delta u) = (\delta\mathcal{L}(u), \delta u).$$

Then, the function $\delta\mathcal{L}(u)$ is called the *variational derivative*. Its form depends on the inner product used. When dealing with the variational derivative, we will always use the L_2-inner product, defined in §3.3.2. For $\hat{u}$ to be a stationary point it must hold for any (admissible) variation δu that

$$(\delta\mathcal{L}(u), \delta u) = 0. \tag{5.10}$$

This condition is called the *weak* or *variational formulation* of the problem. When $\delta\mathcal{L}(u)$ is continuous, this is equivalent to the *pointwise* or *strong* formulation

$$\delta\mathcal{L}(u) = 0,$$

which is referred to as the *Euler–Lagrange equation*. It is in most cases a differential equation for u. The weak formulation is more general than the strong one.

Lagrange multipliers in infinite dimensions

The Lagrange multiplier rule can be generalized to infinite dimensions, too. The simplest case, to which we will restrict ourselves, is when the constraint set is given by the intersection of a finite number of level sets of explicit functionals. So, we consider the constrained problem

$$\hat{u} \in \text{Crit } \{\, \mathcal{L}(u) \mid u \in \mathcal{M},\ \mathcal{K}_1(u) = \kappa_1, \ldots, \mathcal{K}_n(u) = \kappa_n \} \,,$$

where $\mathcal{L}, \mathcal{K}_1, \ldots, \mathcal{K}_n$ are given functionals, with variational derivatives $\delta\mathcal{L}, \delta\mathcal{K}_1, \ldots, \delta\mathcal{K}_n$. In the following we assume that these derivatives are linearly independent at $\hat{u}$, in the sense that $\delta\mathcal{K}_1(\hat{u}), \ldots, \delta\mathcal{K}_n(\hat{u})$ are linearly independent functions. Then, in a critical point $\hat{u}$ the *Lagrange multiplier rule* holds: there exist constants $\lambda_1, \ldots, \lambda_n$ such that

$$\delta\mathcal{L}\left(\hat{u}\right) = \lambda_1 \delta\mathcal{K}_1\left(\hat{u}\right) + \cdots + \lambda_n \delta\mathcal{K}_n\left(\hat{u}\right).$$

Equivalently, we can say that for certain values $\lambda_1, \ldots, \lambda_n$ a critical point $\hat{u}$ of the constrained problem is also a critical point of the unconstrained problem with functional $\mathcal{L} - \lambda_1\mathcal{K}_1 \cdots - \lambda_n\mathcal{K}_n$:

$$\hat{u} \in \text{Crit } \{\, \mathcal{L}(u) - \lambda_1\mathcal{K}_1(u) \ldots - \lambda_n\mathcal{K}_n(u) \mid u \in \mathcal{M} \,\}.$$

The critical point $\hat{u}$, and the Lagrange multipliers $\lambda_1, \ldots, \lambda_n$, depend on the values of $\kappa_1, \ldots, \kappa_n$. When certain smoothness assumptions are satisfied, these Lagrange multipliers depend on the function $V(\kappa_1, \ldots, \kappa_n) := \mathcal{L}\left(\hat{u}(\kappa_1, \ldots, \kappa_n)\right)$ according to

$$\lambda_i = \frac{\partial V}{\partial \kappa_m}, \quad i = 1, \ldots, n.$$

Variational principles

When a given system is described by some equation, say,

$$\mathcal{E}(u) = 0,$$

and a functional $\mathcal{L}$ can be found such that the solutions of $\mathcal{E}(u) = 0$ are in one-to-one correspondence with the critical points of $\mathcal{L}$, i.e., the solutions of

$$\delta\mathcal{L}(u) = 0\,,$$

the equation is called a *variational equation*, and the problem is said to be described by a *variational principle* or to possess a *variational structure*. The problem of finding which infinite-dimensional problems are variational is called the *inverse problem of variational calculus*. We will not deal with this problem in general but restrict ourselves to cases for which either the problem is directly formulated as a variational problem or a given equation (together with its boundary conditions) can be directly seen to be related to a variational formulation.

One of the most intriguing aspects of variational problems in functions spaces is the correct treatment of boundary conditions. In finite dimensions this aspect is not relevant. In the following we will pay attention to this point via examples. We will use at various places the notation $u_x \equiv \partial_x u$, $L_u \equiv \partial_u L$, etc.

Example 5.3a. Single integral problems.
Let us consider the standard problem in the calculus of variations. On the set of functions defined on an interval, $[0, \ell]$, say, we take as a density functional

$$\mathcal{L}(u) = \int_0^\ell L(x, u, u_x)\, dx. \tag{5.11}$$

Here L, a function depending on three variables, is called the Lagrangian density of the Lagrange functional $\mathcal{L}$. If L is differentiable in its second and third variables, its directional derivative is given by

$$\delta\mathcal{L}(u; v) = \int_0^\ell \left[L_u\, v + L_{u_x}\, v_x\right] dx.$$

Partial integration gives

$$\delta\mathcal{L}(u; v) = \int_0^\ell \left[L_u - \frac{d}{dx} L_{u_x}\right] v\, dx + [L_{u_x}\, v]_0^\ell. \tag{5.12}$$

If we admit only variations with $v(0) = v(\ell) = 0$, the last term on the right-hand side vanishes. The condition that (5.12) must hold for any v that satisfies vanishing boundary conditions leads to the so-called Euler-Lagrange equations

$$L_u - \frac{d}{dx} L_{u_x} \equiv \frac{\partial L}{\partial u} - \frac{d}{dx}\left(\frac{\partial L}{\partial u_x}\right) = 0. \tag{5.13}$$

In general this is a second-order ODE for the function $u(x)$. Note that these equations play a highly important role in classical mechanics. The variable x used here is then identified with time, and the fact that the solution of some dynamical system is characterized by the equation of motion following from the critical points of (5.11) is then called the *action principle*. See, e.g., [11] or Chapter 11 of [22].

If the system under consideration is *autonomous*, i.e., L does not explicitly depend on x, this implies that the system has translation symmetry. Then, the second-order ODE reduces to a first-order ODE, and the following expression, a so-called integral of the system, is constant in x:

$$u_x \frac{\partial L}{\partial u_x} - L = E. \tag{5.14}$$

The expression on the left-hand side is called the *Hamiltonian* of the system. For the relation between the Lagrangian and the Hamiltonian of a system we refer to Examples 5.3c and 5.3d. For autonomous systems the Hamiltonian is apparently constant in x. This means that in the phase plane spanned by (u, u_x) the solutions are (parts of) level sets of the Hamiltonian. □

Exercise 5.3a.
Check that for autonomous systems, (5.14) *follows from* (5.13) *by differentiating both sides of* (5.14) *with respect to* x.

Example 5.3b. Sturm–Liouville boundary value problems.
For given functions $p(x)$, $q(x)$, and $f(x)$ the ODE

$$-\partial_x[p(x)\,\partial_x u] + q(x)\,u = f(x) \tag{5.15}$$

with boundary conditions

$$u\,(0) = u\,(\ell) = 0$$

constitutes a boundary value problem that is often met in mathematical physics; this is named for Sturm and Liouville. This type of problem has as variational principle

$$\text{Crit}\ \left\{ \frac{1}{2}\int_0^\ell \left[p\,(\partial_x u)^2 + qu^2 - 2fu\right] dx \;\middle|\; u\,(0) = u\,(\ell) = 0 \right\}.$$

If we release the boundary condition at $x = 0$, we get the slightly modified variational principle

$$\text{Crit}\ \left\{ \frac{1}{2}\int_0^\ell \left[p\,(\partial_x u)^2 + qu^2 - 2fu\right] dx \;\middle|\; u\,(\ell) = 0 \right\}.$$

The first variation in direction v is now given by

$$\int_0^\ell (p\,(\partial_x u)\,(\partial_x v) + q\,u\,v - f\,v)\,dx,$$

which can be rewritten in the form

$$\int_0^\ell (-\partial_x(p\,\partial_x u) + qu - f)\,v\,dx + [p\,(\partial_x u)\,v]_{x=0}^{x=\ell}.$$

By first taking variations that vanish at $x = 0$, we get the same differential equation in the interval $(0, \ell)$ as before. For variations v that do not vanish at $x = 0$ we get in addition the condition $p\,\partial_x u = 0$ at $x = 0$. Hence, we obtain the boundary condition at the origin from the fact that we do not restrict the admissible variations in this endpoint. A boundary condition originating from such an argument is called a *natural boundary condition*. □

Example 5.3c. Catenary.
Now we will determine the precise form of the catenary using the formulation in (5.2) and (5.3). According to the Lagrange multiplier rule applied to these two equations, the shape $z(x)$ of the catenary minimizes for some multiplier λ the functional

$$\rho\, g \int_{-M}^{M} (z-\lambda)\sqrt{1+z_x^2}\, dx.$$

The system is autonomous, so the Hamiltonian is thus conserved and the solution satisfies for some E the ODE

$$\frac{z-\lambda}{\sqrt{1+z_x^2}} = E.$$

With the identity $\cosh^2\xi - \sinh^2\xi = 1$ it is easy to check that the solution of this ODE is explicitly given by

$$z = \lambda + E\cosh(x/E)\,.$$

The two parameters E, λ in this solution must be chosen such that the support conditions $z(\pm M) = 0$ and the restriction on the total length of the chain are satisfied. □

Exercise 5.3b.
Calculate explicitly the parameters E and λ from the support and length conditions on the catenary and express the shape of the catenary in terms of the original parameters L and M. See also Example 5.1.b.

Exercise 5.3c. *Solitons.*
We have seen in §1.5.3 *that the KdV equation describes surface waves above a flat bottom. In normalized form and in a coordinate system moving with the speed of infinitely long waves, the governing equation for the wave amplitude $u(x,t)$ reads as*

$$\partial_t u + \partial_x^3 u + u\,\partial_x u = 0\,. \tag{5.16}$$

Let us consider the special family of solitary wave solutions called solitons. *These are waves that travel with undisturbed shape and at constant speed. Their profile decays at infinity and consists of a single positive hump.*

a. *Use for the approximate model manifold "waves" that have a Gaussian function as a profile, as used in* (5.5)*, and that translate at a certain speed. So, as basis functions we take*

$$U(x,t;A,\sigma,\lambda) := A\ \exp -(x-\lambda t)^2/\sigma^2.$$

Derive from dimensional analysis a relation between the parameters A, σ, and λ.

b. *Denote the soliton solution by $S(x-\lambda t)$, where the function S represents the shape of the soliton and λ its speed. Write the equation for S by substituting in* (5.16)*. Using the fact that S has to decay at infinity, integrate the equation once and show that the constant of integration has to vanish. Check that this equation can also be interpreted as the equation for the critical points of the constrained variational principle*

$$\min_u \{\, \mathcal{E}(u) \mid I(u) = \gamma \,\},$$

where

$$\mathcal{E} = \int \left[-\left(\frac{1}{2}\partial_x u\right)^2 + \frac{1}{6}u^3 \right] dx$$

is related to the wave energy and

$$I = \frac{1}{2}\int u^2\, dx$$

to its momentum. Hence, the shape of the soliton is such that the energy is critical for a given value of the momentum. Changing the value of the momentum, a whole family is obtained. Note that the translation speed of the soliton acts as a Lagrange multiplier.

c. *Now use again the Gaussian functions as an approximation for the soliton profile, and restrict the constrained minimization problem to a restricted problem for functions of only two parameters, namely, amplitude A and width σ. Use this restricted optimization problem to find relations between width, amplitude, and speed (multiplier) of the solitons. What is the advantage of this approach compared to the result under* a*?*

d. *Viewing the resulting multiplier equation as a critical point of the functional $\mathcal{E}(u) - \lambda\, I\,(u)$, use energy conservation to obtain a first-order ODE for S. Perform a phase plane analysis, and conclude that homoclinic orbits starting from and ending in the origin exist only for positive values of λ. These homoclinic orbits correspond to the soliton profiles sought for. An explicit expression for the soliton profiles turns out to be of the form*

$$S\,(x) = \frac{a}{\cosh^2\,(bx)}\,.$$

Find the coefficients a and b explicitly, and relate these to the value of λ. Show that these results agree with the ones found above.

e. *Instead of the KdV equation in normalized form* (5.16)*, we now want to use real physical variables. With the surface elevation denoted by η, the full KdV equation for nonlinear surface waves above a layer of depth H is given by*

$$\partial_t \eta + c\left(\partial_x \eta + \frac{H^2}{6}\partial_x^3 \eta\right) + \eta\,\partial_x \eta = 0,$$

where $c = \sqrt{gH}$ with g the constant acceleration of gravity. Now repeat the above arguments and show in addition the dependence of the results on the depth H.

Example 5.3d. Lagrangian formulation of classical mechanics.
Special cases of the systems mentioned above are mechanical systems without friction, frequently studied in classical mechanics. The typical situation, and notation, concerns a system that is described with the position vector $\mathbf{q} = \mathbf{q}\,(t)$, a $3N$-vector, including the positions of N mass points in space. The velocities are represented by the $3N$-vector $\dot{\mathbf{q}}$. The Lagrangian function is the density $L(\mathbf{q}, \dot{\mathbf{q}}, t)$ with which the so-called action functional

$$\int L(\mathbf{q}, \dot{\mathbf{q}}, t)\, dt \tag{5.17}$$

is associated. As described in Example 5.3a, the dynamics of such a Lagrangian system are governed by the *action-principle*, i.e., the physical solution is found as that trajectory in the $(\mathbf{q}, \dot{\mathbf{q}})$ space that makes the action functional between the initial and the final time stationary. This principle leads to the insight that the equations of motion of the system are given by the Euler–Lagrange equations applied to L:

$$\nabla_{\mathbf{q}} L - \frac{d}{dt}(\nabla_{\dot{\mathbf{q}}} L) \equiv \frac{\partial L}{\partial \mathbf{q}} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\mathbf{q}}}\right) = \mathbf{0}\,.$$

This is a system of $3N$ second-order ODEs. Note that boundary conditions are not considered, which means that the variations are taken to vanish at the endpoints of the time interval under consideration.

Let us restrict ourselves to so-called conservative systems for which the forces are given by the gradient of some time and velocity independent potential function $V(\mathbf{q})$. Then, L is given by the difference between kinetic and potential energy:

$$L = \frac{1}{2}\,\dot{\mathbf{q}} \cdot \mathbf{M} \cdot \dot{\mathbf{q}} - V(\mathbf{q}), \tag{5.18}$$

where $\mathbf{M}$ is the so-called mass-matrix. This matrix is symmetric and in many cases diagonal with on the diagonal the masses of the particles. The Euler–Lagrange equations for such a system read as

$$\frac{d}{dt}\,(\mathbf{M} \cdot \dot{\mathbf{q}}) = -\nabla_{\mathbf{q}} V.$$

In this simple case, this system of $3N$ second-order equations is usually referred to as the *Newton equations*. For a system with only one particle these equations reduce to the well-known second law of Newton $\mathbf{F} = m\,\mathbf{a}$. The total energy of such systems is the sum of potential and kinetic energy, and it is easy to check that the total energy is conserved and thus given by a constant, E, say:

$$\frac{1}{2}\,\dot{\mathbf{q}} \cdot \mathbf{M} \cdot \dot{\mathbf{q}} + V(\mathbf{q}) = E.$$

For instance, for a harmonic oscillator of mass m and spring constant k we have

$$L = \frac{1}{2}\,m\,\dot{\mathbf{q}}^2 - \frac{1}{2}\,k\,\mathbf{q}^2\,.$$

The corresponding Newton equation reads as

$$m\,\ddot{\mathbf{q}} + k\,\mathbf{q} = 0,$$

and the total energy is given by

$$E = \frac{1}{2}\,m\,\dot{\mathbf{q}}^2 + \frac{1}{2}\,k\,\mathbf{q}^2\,. \qquad \square$$

Example 5.3e. Hamiltonian formulation of classical mechanics.
For many Lagrangian systems an alternative formulation is possible. Roughly speaking, the idea is to replace each second-order equation by two first-order equations. Instead of using positions and velocities $(\mathbf{q}, \dot{\mathbf{q}})$ as variables, we switch to positions and momenta $(\mathbf{q}, \mathbf{p})$, with the momentum vector defined as

$$\mathbf{p} := \nabla_{\dot{\mathbf{q}}} L(\mathbf{q}, \dot{\mathbf{q}}, t).$$

This transformation of variables is possible if the mapping $\dot{\mathbf{q}} \to \nabla_{\dot{\mathbf{q}}} L(\mathbf{q}, \dot{\mathbf{q}})$ is invertible for each $\mathbf{q}$ and t. Just as in Example 5.3d we consider conservative systems with the Lagrangian given by (5.18). The momentum vector then has the simple form

$$\mathbf{p} := \mathbf{M} \cdot \dot{\mathbf{q}} \,,$$

and the second law of Newton states that

$$\frac{d}{dt}\mathbf{p} = \mathbf{F} = -\nabla_{\mathbf{q}} V \,.$$

This system of $3N$ second-order ODEs for $\mathbf{q}(t)$

$$\mathbf{M} \cdot \ddot{\mathbf{q}} = -\nabla_{\mathbf{q}} V$$

can be split up into a system of $6N$ first-order ODEs for $\mathbf{q}(t)$ and $\mathbf{p}(t)$:

$$\dot{\mathbf{q}} = \mathbf{M}^{-1} \cdot \mathbf{p} \,, \tag{5.19}$$

$$\dot{\mathbf{p}} \;\; = -\nabla_{\mathbf{q}} V \,. \tag{5.20}$$

These are the *Hamilton equations*. These equations can also be obtained as the Euler–Lagrange equations applied to the so-called canonical action functional

$$\mathcal{A}(\mathbf{q}, \mathbf{p}) = \int \left[\dot{\mathbf{q}} \cdot \mathbf{p} - H(\mathbf{q}, \mathbf{p})\right] dt \,, \tag{5.21}$$

which is nothing more than the action functional in (5.17), rewritten in terms of $(\mathbf{q}, \mathbf{p})$ instead of $(\mathbf{q}, \dot{\mathbf{q}})$. The Hamiltonian function is generally defined as

$$H(\mathbf{q}, \mathbf{p}, t) := \dot{\mathbf{q}} \cdot \mathbf{p} - L.$$

Note that initially L is expressed in terms of $(\mathbf{q}, \dot{\mathbf{q}}, t)$ so that the right-hand side has to be rewritten in terms of $(\mathbf{q}, \mathbf{p}, t)$. The Hamiltonian corresponding to the Lagrangian (5.18) is

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\, \mathbf{p} \cdot \mathbf{M}^{-1} \cdot \mathbf{p} + V(\mathbf{q}) \,.$$

Taking together the two $3N$-vectors $\mathbf{q}$ and $\mathbf{p}$ as one $6N$-vector $\mathbf{u} := (\mathbf{q}, \mathbf{p})$, we obtain the Hamilton equations in the characteristic form

$$\partial_t \mathbf{u} = \mathbf{J} \cdot \nabla_{\mathbf{u}} H(\mathbf{u}) \quad \text{with } \mathbf{J} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix}, \tag{5.22}$$

where $\mathbf{I}$ is the $3N \times 3N$ unit matrix. The matrix J is thus skew-symmetric. We will encounter a generalized form for infinite-dimensional systems in §6.3 to describe surface water waves. □

Exercise 5.3d.

Determine the Euler–Lagrange equations associated with the functional $\mathcal{A}(\mathbf{q}, \mathbf{p})$ *in* (5.21). *Note that now variations with respect to* $\mathbf{q}$ *as well as with respect to* $\mathbf{p}$ *are required. Check that the resulting equations can be written in the form of the Hamiltonian equations.*

5.4 Variational restriction

Let us again consider a variational problem, specified by a functional $\mathcal{L}$ and a set $\mathcal{M}$ of admissible elements

$$\mathcal{L} : \mathcal{M} \to \mathbb{R}.$$

The governing variational equation is given by $\delta\mathcal{L}(u) = 0$. Consider the restriction to a model manifold $\mathcal{M}_a$ of parameterized functions $U(\mathbf{p})$ as introduced in §5.3.2:

$$\mathcal{M}_a = \{\, U(\mathbf{p}) \mid \mathbf{p} \,\} \subset \mathcal{M},$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_N)$ is the parameter vector of length N. The case $N \to \infty$ is included in the following. The restriction of the functional to this manifold

$$\mathcal{L}_a : \mathcal{M}_a \to \mathbb{R}$$

leads to a function(al) defined on the parameter space:

$$\mathcal{L}_a(\mathbf{p}) := \mathcal{L}(U(\mathbf{p})).$$

Instead of looking for critical points of $\mathcal{L}$ on $\mathcal{M}$, we now look for the critical points of the restricted functional $\mathcal{L}_a(\mathbf{p})$. In the minimization process we may take only variations that are tangent to the manifold $\mathcal{M}_a$. This leads to the conditions

$$\frac{\partial \mathcal{L}_a}{\partial p_k} = \left(\delta\mathcal{L}(U(\mathbf{p})), \frac{\partial U}{\partial p_k} \right) = 0, \qquad k = 1, \ldots, N. \tag{5.23}$$

So, the governing equation for critical points of the restricted functional is found by differentiation with respect to the parameters $\mathbf{p}$:

$$\nabla_{\mathbf{p}} \mathcal{L}_a = \mathbf{0}.$$

This last equation is the weaker requirement that the projection of the original equation vanishes on the tangent manifold, i.e., in all directions $(\partial U/\partial p_k)$ determined by the restricted set. We thus find that for a solution $\bar{U}$ of the restricted model it holds that

$$\delta\mathcal{L}(\bar{U}) \perp T\mathcal{M}_a\left(\bar{U}\right),$$

where $T\mathcal{M}_a$ is the tangent manifold of $\mathcal{M}_a$ at $\bar{U}$. Here, orthogonality is meant with respect to the L_2-inner product.

This "generalized multiplier rule" is understandable, since the constraint that is a consequence of the restriction to the restricted manifold that is a subset of the original manifold leaves uncertainty about $\delta\mathcal{L}(\bar{U})$ in directions in which no variations can be considered, i.e., in directions perpendicular to the tangent space of the restricted manifold. In the next section we work this out further by specifying $\mathcal{M}_a$ more explicitly.

5.4.1 Ritz–Galerkin projections

Let us construct the restricted manifold $\mathcal{M}_a$ using a finite number of basis functions ϕ_k , $k = 1, \ldots, N$. We take for the $U(\mathbf{p})$ linear combinations of the basis functions:

$$U(\mathbf{p}) = \sum_{k=1}^{N} p_k \, \phi_k.$$

In this case we have that

$$\nabla_{\mathbf{p}} U(\mathbf{p}) = (\phi_1, \phi_2, \ldots, \phi_N),$$

and (5.23) becomes

$$(\delta\mathcal{L}(\bar{U}), \phi_k) = 0, \quad k = 1, .., N. \tag{5.24}$$

This is known as the *Ritz–Galerkin method*: approximating the solution with a finite number of N basis functions, the infinite-dimensional equation is replaced with a set of N equations by taking the projection of the variational derivative on the N basis functions. This last formulation is also often referred to as (the truncation of) the weak formulation of the problem.

Defining the restriction of the functional by

$$\tilde{\mathcal{L}}(\mathbf{p}) := \mathcal{L}\left(\sum_{k=1}^{N} p_k \, \phi_k\right), \tag{5.25}$$

we find that the Ritz–Galerkin method replaces the variational problem $\delta\mathcal{L}(u) = 0$ with the condition of vanishing of the gradient of the function $\tilde{\mathcal{L}}(\mathbf{p})$:

$$\nabla_{\mathbf{p}} \tilde{\mathcal{L}}(\mathbf{p}) = \mathbf{0}. \tag{5.26}$$

Example 5.4a. Truncation of Fourier series.
A well-known example of this method is the Fourier representation, in which the functions from the original space are written as a Fourier series and the model manifold is the restriction to a finite number of Fourier modes. The Ritz–Galerkin approach requires that the original equation is satisfied along all directions of the restricted manifold: the residue $\delta\mathcal{L}\left(\bar{U}\right)$ will only have Fourier modes different from those in the model manifold. □

5.4.2 Variational accuracy

If for the solution $\bar{u}$ of an equation $\mathcal{E}(u) = 0$ an approximate solution $\bar{u}_{app}$ is obtained, in whatever way, we can consider the *residual*, given by the value $|\mathcal{E}(\bar{u}_{app})|$. In most cases it is not hard to estimate the value of the residual if the (in)accuracy $||\bar{u} - \bar{u}_{app}||$ of the approximation is known. The other way around, i.e., to find from knowledge of the residual the accuracy of the approximate solution, is much more difficult. This is also true for variational problems. However, in various problems, the *value* of the functional at a critical point may give some indication of the accuracy of the approximation. This especially holds for minimization problems: the closer the approximate value is to the exact value, the better

the approximation will be. Results in this direction are obtained in the theory of second variation, which studies the quadratic terms in a Taylor expansion. Generally speaking, at an optimal solution $\bar{u}$ we have that $\delta L(\bar{u}) = 0$. Then, in a Taylor approximation of the functional, we can write for a variation δu around the optimal solution

$$\mathcal{L}(\bar{u} + \varepsilon\delta u) = \mathcal{L}(\bar{u}) + \varepsilon^2 \mathcal{Q}(\delta u) + O\left(\varepsilon^3\right).$$

Here $\mathcal{Q}$ is the so-called second variation. When this second variation is nondegenerate, for instance, when it is a positive definite quadratic form, the value of the residue leads to an estimate for the error of the approximation.

The second-order effect at a critical point, $\mathcal{L}(\bar{u} + \varepsilon\delta u) \approx \mathcal{L}(\bar{u}) + O\left(\varepsilon^2\right)$, does lead to an interesting result. It implies that the value of the functional can be obtained more accurately than the solution itself: from a first-order approximation of the solution, the value of the functional is obtained in second-order accuracy. This is called *variational accuracy.*

Example 5.4b. Eigenvalue problem of Sturm–Liouville type.
An eigenvalue problem of Sturm–Liouville type on an interval $[0, \ell]$ is given by

$$-\partial_x [p(x)\, \partial_x u] + q(x)\, u(x) = \lambda\, u(x),$$

together with homogeneous boundary conditions (either Dirichlet, Neumann, or mixed conditions). See also Example 5.3b. If we take Dirichlet conditions, the eigenfunctions u are the critical points of the so-called *Rayleigh quotient* and the value of that quotient is the value of the corresponding eigenvalue:

$$\lambda = \text{Crit}\,\{\, \mathbb{R}(u) \mid u(0) = u(\ell) = 0 \,\},$$

where the *Rayleigh quotient* is defined as

$$\mathbb{R}(u) = \int \left[p\,(\partial_x u)^2 + q u^2\right] dx \Big/ \int u^2\, dx.$$

In particular, when p is a strictly positive function, the lowest eigenvalue is finite and is given by the minimization of the Rayleigh quotient. Then, the variational accuracy result implies that if an approximation of the eigenfunction is of order ε, the eigenvalue is found with higher accuracy, namely, of order ε^2. □

Exercise 5.4a.

a. *Verify that for $p = 1, q = 0$, and $\ell = \pi$, the eigenvalues and eigenfunctions of the above Sturm–Liouville problem in Example* 5.4b *are*

$$u_k = \sin(kx), \quad \lambda_k = k^2.$$

b. *Take a set of appropriate (e.g., polynomial) functions and determine an approximation for the lowest eigenvalue by variational restriction.*

c. *Using these trial functions, show that the variational accuracy result holds in this specific case.*

5.5 Scientific computing

As indicated in §5.2.2, the difference between low-dimensional models and high-dimensional models for numerical calculations is not a principle difference. For low-dimensional models the choice of the model manifold is essential and directly determines the quality and validity of the model. Usually, it will be a nonlinear manifold. However, when looking for numerical algorithms, the model manifold is mostly taken to be a linear subspace, of finite but high dimension N. It is characteristic then to consider N as a parameter that can be varied according to the desired accuracy: not a single manifold but a sequence of linear manifolds which will span the whole space in the limit $N \to \infty$. Hence, roughly speaking, the solution will always be captured (with desired accuracy) by taking N sufficiently large. Of course, the specific choice of the N-dimensional manifolds is determined by the basis functions used, and one choice may be more efficient or more tractable than another, but for increasing N the differences will become smaller and smaller.

For variational problems, the variational restriction method guarantees that the variational property of the basic equation is reflected in the restricted problem, since the so-called discretized system will be obtained as the equation for critical points of the restricted functional. Hence, the discretization procedure is variationally consistent. It should be noted that this consistency in itself is not directly related to accuracy or efficiency. The latter aspects will depend on the specific application.

The basic ingredients of any model are the state variables and the relations between them. For numerical discretizations of complicated sets of PDEs, the approximation of the state variables belongs to approximation theory and is formally not related to the way the equations are approximated. However, in the variational restriction method, the discretization of the equations follows directly from the choice of the model manifold, i.e., from the approximation of the state variables. So, the discretization of the state variables and the discretization of the equations are closely connected in the process of variational restriction. In fact, we have seen in §5.4.1 that in a general Ritz–Galerkin way, the restriction of the equation leads to a discretized system obtained by projecting the original equation in the directions of the basis functions. For restricted manifolds that are nonlinearly dependent on the parameters, the projection of the original equation is along the tangent directions of the manifold.

In the following we will briefly consider some simple ways to approximate state variables, and then we show the resulting discretizations for characteristic problems.

5.5.1 Approximation of the state variables

A simple but illustrative example that we will use in the rest of this section to illuminate the ideas is to study systems for which the state space has as elements real or complex-valued functions defined on an interval, say,

$$u : [0, \ell] \to \mathbb{R} \text{ (or } C) .$$

For example, one could think of temperature profiles in a rod of length ℓ. We will not specify the smoothness of the functions higher than continuity, but when required, we will assume sufficient regularity to perform certain operations later on. In the first instance we will also disregard possible boundary conditions.

Approximation theory deals with approximating the infinite-dimensional space of functions on $[0, \ell]$ by a finite-dimensional one of N dimensions. We list and comment on a few of the most well-known methods.

Collocation
We choose in the interval $I \equiv [0, \ell]$ a set of N points (not necessarily equidistant) with $x_1 < x_2 < \cdots < x_N$. In the collocation procedure, the function $u(x)$ is discretized by an N-dimensional vector $\mathbf{u} := (u_1, \ldots, u_N)$ with

$$u_k := u(x_k), \quad k = 1, \ldots, N.$$

The reverse way, interpreting the vector as an approximation of the function, is in fact not defined, since the vector does not contain any information about the function in the intervals between the gridpoints. Intuitively one often thinks of linear or spline interpolation in between the points. In fact, many modern-day plotting programs on computers will easily transform such a sequence of points by a smooth curved line. Without additional information, the collocation method does not define a model manifold that can be considered as a subspace of the original state space. Related to this is the fact that the map $u(x) \to \mathbf{u}$ could be described with functionals like $u_k = \int u(x)\, \delta(x - x_k)\, dx$, but since function values at a specific point are not continuously defined in the L_2-sense (as is related to the fact that the delta function is not square integrable), this is not a continuous functional in the usual L_2-sense. The collocation method also has its consequence when looking for approximations of the derivatives of such a function. As said above, no information in between the grid points is available, so one has to define the derivatives without this information. The usual approach is to replace the differential quotients with difference quotients, i.e., taking *finite differences*, which is also the common name for such methods. Euler forward, Euler backward, central differences, etc., are then some of the many choices.

Finite elements
In the finite element method the spatial domain is discretized, too. The resulting small spatial domains are referred to as elements. In one dimension the elements are just the intervals between the grid points. In two dimensions one often uses triangles. In the finite element approach, local basis functions are used, which are nonzero only on a few adjacent elements. The smoother the basis functions are required to be, the more extensive the support of the basis functions will have to be. The simplest example in one dimension is the use of tent functions, also used in Example 5.2c. These are piecewise linear functions with support on two adjacent elements, the intervals left and right of a mesh point x_k. At x_k itself the tent function takes the value 1. See Fig. **5.3**. For simplicity we assume a uniform grid of mesh size h. In formula the tent functions are then given by

$$T_k(x) = \left(1 - \frac{|x - x_k|}{h}\right) H(x - x_{k-1})\, H(x_{k+1} - x), \tag{5.27}$$

where H denotes the Heaviside function with $H(x) = 0$, if $x < 0$, and $H(x) = 1$, if $x \geq 0$. The following approximation then coincides on the grid points with the function $u(x)$ and

linearly interpolates between the grid points:

$$\hat{u}(x) := \sum_{k=1}^{N} u_k \, T_k(x). \tag{5.28}$$

When basis functions of quadratic or higher order are used, the interpolation between the grid points becomes smoother.

A slightly different way to define the coefficients u_k in (5.28) is to use a continuous functional to determine the coefficients so that they depend continuously on the function. This can be done, for instance, by taking

$$\hat{u}(x) := \sum_{k=1}^{N} \hat{u}_k \, T_k(x) \tag{5.29}$$

with the coefficients $\hat{u}_k$ defined by

$$\hat{u}_k = \int_I u(x) \, T_k^*(x) \, dx.$$

Here, the dual basis functions T_k^* are introduced. They satisfy the orthonormality condition (for all k and j)

$$\int_I T_k(x) \, T_j^*(x) \, dx = \delta_{kj}.$$

Fourier truncation

If the functions to be considered are periodic, with period ℓ, say, standard Fourier representation may be used:

$$u(x) = \sum_{k=-\infty}^{\infty} c_k \, e^{2\pi i k x/\ell}$$

with the coefficients given by

$$c_k := \frac{1}{\ell} \int_0^{\ell} u(x) \, e^{-2\pi i k x/\ell} \, dx.$$

Contrary to the basis functions used in the finite element approach, the Fourier basis functions form a set of *global* basis functions. They are global, since the support covers the whole interval, and, as a consequence, local changes in $u(x)$ affect all Fourier coefficients. The representation is infinite-dimensional and we thus need infinitely many coefficients. Truncation leads to finite approximations. A truncation to n modes leads to an $N = (2n+1)$-dimensional manifold and is described by

$$\hat{u}(x) = \sum_{k=-n}^{n} c_k \, e^{2\pi i k x/L} \, .$$

Note that for each value of n the approximations form subspaces of the original space.

5.5.2 Variational treatment of Sturm–Liouville problems

We will now show a simple example of variational discretization. In §6.3 and §6.4 we will meet more complicated examples.

Consider a variational problem with functional $\mathcal{L}(u)$ on a function space $U = \{u : [0, \ell] \to \mathbb{R}\}$. As specific example we will take the Sturm–Liouville type of functional

$$\mathcal{L}(u) = \int_0^\ell \left[\frac{1}{2} p(x) (\partial_x u)^2 + \frac{1}{2} q(x) u^2 - f(x) u \right] dx \tag{5.30}$$

with p, q, and f given functions. See also Example 5.3b. We take homogeneous Dirichlet conditions: $u(0) = u(\ell) = 0$. The governing Euler–Lagrange equation leads to the standard Sturm–Liouville boundary value problem

$$\begin{aligned} -\partial_x(p\, \partial_x u) + q\, u &= f, \\ u(0) = u(\ell) &= 0. \end{aligned} \tag{5.31}$$

We will use the finite element approach with, as basis functions, the piecewise linear tent functions (5.27) on a uniform grid $0 = x_0 \le x_1 \le x_2 \le \cdots \le x_{N+1} = \ell$ with grid size $\ell/(N+1)$. The tent functions centered around the endpoints are only partly within the interval $[0, \ell]$. The homogeneous boundary conditions cause the first and last coefficients to vanish: $c_0 = c_{N+1} = 0$. Each function is then approximated by an element from the model manifold:

$$u^{(N)}(x) = \sum_{k=1}^{N} c_k\, T_k(x). \tag{5.32}$$

The derivative of $u^N(x)$ is piecewise continuous, and the functional $\mathcal{L}(u)$ can be evaluated for such functions. Inserting the superposition into the functional $\mathcal{L}$, we obtain a function of the N-vector $\mathbf{c} := (c_1, \ldots, c_N)$ defined by

$$\hat{\mathcal{L}}(\mathbf{c}) := \mathcal{L}\left(u^{(N)}\right).$$

The finite element approximation of the Sturm–Liouville problem is then given by the variational problem for the critical points of $\hat{\mathcal{L}}$, i.e., by the N algebraic equations

$$\nabla_{\mathbf{c}} \hat{\mathcal{L}}(\mathbf{c}) = \mathbf{0}, \text{ i.e., } \partial_{c_1} \hat{\mathcal{L}}(\mathbf{c}) = \cdots = \partial_{c_N} \hat{\mathcal{L}}(\mathbf{c}) = 0\,.$$

An explicit expression for the function $\hat{\mathcal{L}}(\mathbf{c})$ is easily obtained from substitution of (5.32) in (5.30). It results in a quadratic expression in the c_k, which in matrix notation reads as

$$\hat{\mathcal{L}}(\mathbf{c}) = \frac{1}{2}\, \mathbf{c} \cdot \mathbf{P} \cdot \mathbf{c} + \frac{1}{2}\, \mathbf{c} \cdot \mathbf{Q} \cdot \mathbf{c} - \mathbf{f} \cdot \mathbf{c},$$

where the elements of the matrices $\mathbf{P}$, $\mathbf{Q}$ and the vector $\mathbf{f}$ are given by

$$P_{kj} = \int_0^\ell p\ (\partial_x T_k)\left(\partial_x T_j\right)\ dx, \tag{5.33}$$

$$Q_{kj} = \int_0^\ell q\ T_k\ T_j\ dx, \tag{5.34}$$

$$f_k = \int_0^\ell f\ T_k\ dx. \tag{5.35}$$

The governing algebraic equation is then

$$\nabla_{\mathbf{c}} \hat{\mathcal{L}}\left(\mathbf{c}\right) = \mathbf{P}\cdot\mathbf{c} + \mathbf{Q}\cdot\mathbf{c} - \mathbf{f} = \mathbf{0}.$$

This is a system of N linear equations for the N components of the vector $\mathbf{c}$. Comparing this with the original Sturm–Liouville equation (5.30), it is seen that the differential operator $-\partial_x(p\,(x)\,\partial_x)$ is transformed into the matrix $\mathbf{P}$, the multiplication operator $q\,(x)$ into the matrix $\mathbf{Q}$, and the function f into the vector $\mathbf{f}$. The matrices $\mathbf{P}$ and $\mathbf{Q}$ are symmetric. This is an immediate consequence of the consistent variational approach: the restriction of the quadratic parts of the functional $\mathcal{L}$ of u have become symmetric bilinear functions of $\mathbf{c}$.

With usual conditions that guarantee that this problem has a solution (for instance, positivity of $p\,(x)$ and $q\,(x)$ suffices), this algebraic system has a solution, which has to be found numerically.

Actually, in the formulas above we have not yet used the specific definition of the basis functions T_k. Their specific form comes in when the matrices are calculated explicitly. Then also the practical advantage of the finite element approach may be appreciated: the finite element matrices obtained with the local tent functions are very sparse. Since only the product of two overlapping, thus neighboring, tent functions does not vanish, the matrices are in this case tridiagonal. On the other hand, matrices obtained with global basis functions will usually be full. From a computational point of view, the finite element method is highly preferable for solving the linear algebra problem.

Example 5.5a. Sparsity of finite element matrices.
As for the sparsity of the finite element method matrices, observe that although the integrals in (5.33) and (5.34) are over the entire interval $(0,\,\ell)$, the confinement of the tent functions has as a consequence that the integral of the product $T_k\ T_j$, and likewise the product of their derivatives, does not vanish only if $k = j$, or $k = j \pm 1$. Hence, both $\mathbf{P}$ and $\mathbf{Q}$ are tridiagonal matrices. Their structure and character can best be interpreted for the special case that the functions p and q are constant. For $p = q = 1$ we have

$$P = \frac{1}{h}\text{tridiag}\,(-1,\,2,\,-1)\,, \tag{5.36}$$

$$Q = \frac{1}{6h}\text{tridiag}\,(1,\,4,\,1)\,. \tag{5.37}$$

We find that the above procedure has discretized the differential operator $-\partial_x^2$ and the multiplication operator $q\,(x)$ in the following way:

$$\partial_x^2 u(x_k) \to \frac{c_{k+1} - 2c_k + c_{k-1}}{h^2}, \tag{5.38}$$

$$u(x_k) \to \frac{1}{6}c_{k+1} + \frac{2}{3}c_k + \frac{1}{6}c_{k-1}. \tag{5.39}$$

The expression for the differential operator is the same expression as obtained when the central difference method of the second derivative is used. However, the discretization of the multiplication operator—here the identity $q = 1$—shows that the present discretization is not simply the value at the point of consideration, as would be the case when using a simple Taylor expansion around x_k. Instead, it is a weighted average, in which the values at two adjacent grid points are involved. □

Exercise 5.5a.

a. *Numerically implement the discretization scheme above.*

b. *Check your implementation by taking special cases for which the solution is known (at least qualitatively), for instance, for the case* $p \equiv 1, q \equiv 0, f \equiv 1$.

c. *Determine the order of the numerical scheme: the error of the solution as a function of the grid size.*

d. *Take* $p \equiv 1, q \equiv -1$ *and different loading functions* f. *Explain by plotting the resulting deflections that the shapes are what one expects from the interpretation of the equation as a loaded string with mass density and spring constant equal to unity and* $f(x)$ *representing the force exerted on the string at position* x.

Chapter 6
Advanced Models

This chapter contains a number of sophisticated models that may serve several purposes. First, the models are chosen such that they illustrate the application of most modeling tools introduced in the preceding sections. Although most of these features are also explained through minor examples of a classroom character in the foregoing chapters, it is highly useful to meet them also in the context of models that describe complicated real-life situations from different fields. The mathematical tools to be used are mentioned in the introduction of each of these case studies.

6.1 Polymer dynamics and vibrating strings

The following case study forms an illustration of the concepts dealt with in
§1.3 (dimensional analysis),
§2.1 (discrete versus continuous models),
§2.3 (constitutive relations),
§3.1.3 (evolution equations), and
§3.3.2 (expansion in basis functions).

In this case study we model the dynamical behavior of long flexible objects. These strings may stretch and possess little bending stiffness. Examples are long molecules in a polymer melt and DNA molecules in a dilute solution, but a vibrating string in a musical instrument also fits the present description. In the first instance we focus on modeling the dynamics of polymer chains. It will turn out that a vibrating string is described by an alternative version of the general model, the well-known wave equation.

Polymer molecules consist of a series of thousands of connected monomers. We restrict ourselves to very long, unbranched molecules, of which polyethylene is a common example. See Fig. **6.1**. In the case of DNA, the chain of constituting molecules has no periodicity, but for the dynamics of the molecule these differences are of minor importance. We assume the polymer chains or DNA molecules to be dissolved in a solvent and take the concentration of chains to be low. Then, the chains have little interaction with each other and are completely surrounded by solvent molecules. The solvent consists of small particles, so it behaves in a Newtonian manner, as described in Example 2.7a. The viscosity

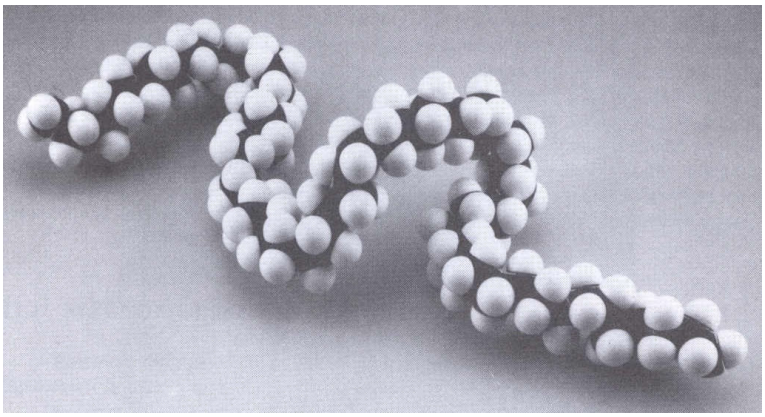

Figure 6.1. Geometrical model of a polymer chain.

of the solvent tends to slow down any motion in the long run. The situation is sketched in Fig. **6.2**. The motion of an individual chain can be observed experimentally. Here, we aim at a description of its dynamics. The first question that arises is how to model a flexible chain in three dimensions. Readers who are interested in reading further about polymer dynamics are referred to [5, 8, 19].

6.1.1 Bead-spring representation

A useful and therefore popular representation of a flexible chain without bending stiffness is a bead-spring model. In this so-called Rouse model a chain of mass M is discretized by cutting it into N identical pieces, each of mass $m = M/N$. These beads are treated as point masses. In the unstretched situation, the chain has equilibrium length L_0. The equilibrium

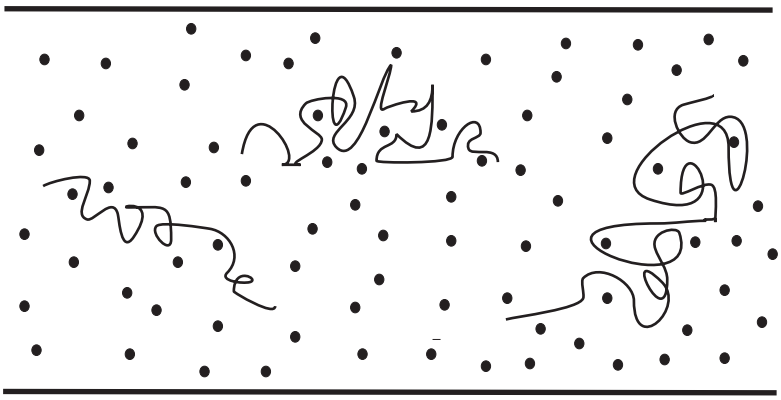

Figure 6.2. Long, flexible, dissolved polymer molecules are hindered in their motions by the surrounding solvent molecules, but they still move under influence of Brownian forces.

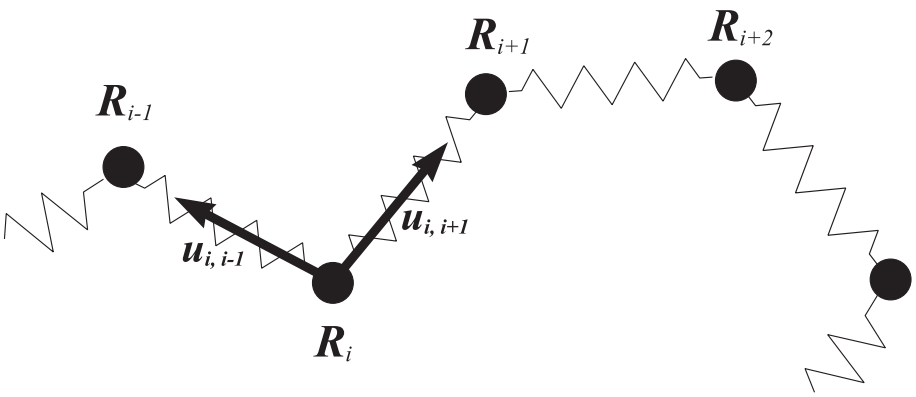

Figure 6.3. A long flexible object is readily represented by a mass-spring chain.

distance between neighboring beads is then given by $b = L_0/(N-1)$. It should be noted that the beads do not necessarily coincide with individual monomers but may represent a much longer part of the chain. The N beads are connected by $N-1$ identical springs, as sketched in Fig. **6.3**. The bigger N is, the more details of the dynamical behavior of the chain are represented, of course. We denote the position of bead i at time t by $\mathbf{R}_i(t)$. Per bead the second law of Newton, which expresses the balance of forces, reads as

$$\underbrace{m\,\partial_t^2 \mathbf{R}_i}_{\text{inertia term}} \quad = \quad \underbrace{-\zeta\,\partial_t \mathbf{R}_i}_{\text{friction}} \quad + \quad \underbrace{\mathbf{F}_{i,i+1} + \mathbf{F}_{i,i-1}}_{\text{spring forces}} . \tag{6.1}$$

Since the concentration of chains is assumed to be low (dilute solution), each chain interacts only with solvent particles. This interaction leads to friction. As usual, we model the friction by assuming it to be linearly proportional to the velocity of the bead under consideration. Further, $\mathbf{F}_{i,i+1}$ is the elastic force exerted by bead $i+1$ on bead i. The strength of this force is taken to be proportional to the stretching of the interval between beads i and $i+1$ relative to the equilibrium distance b. In formula,

$$\mathbf{F}_{i,i+1} = k\,\frac{|\mathbf{R}_{i+1} - \mathbf{R}_i| - b}{b}\,\mathbf{u}_{i,i+1} \tag{6.2}$$

with the unit vector $\mathbf{u}_{i,i+1}$ parallel to the vector connecting beads i and $i+1$:

$$\mathbf{u}_{i,i+1} = \frac{\mathbf{R}_{i+1} - \mathbf{R}_i}{|\mathbf{R}_{i+1} - \mathbf{R}_i|} \tag{6.3}$$

and with k the spring constant, which characterizes the "stretchability" of the chain. Since the end beads $i = 1$ and $i = N$ have only one neighboring bead, (6.1) does only hold for $i \neq 1, N$. However, if we define $\mathbf{F}_{1,0} = \mathbf{F}_{N,N+1} = 0$, this equation holds for all $i = 1, \ldots, N$. Since the solution has a certain temperature, the solvent particles exert random forces on the beads. The dynamics of the chain thus has a stochastic component, the so-called Brownian motion. To keep the analysis transparent, the Brownian force will be left out first and be incorporated at a later stage.

6.1.2 Continuous limit

The equations of motions for the Rouse chain could be numerically integrated. In that approach it would be advantageous to take the number of beads N as small as possible to save computing time. However, in reality polymer chain dynamics are better described by taking the chain as a continuum. That is why we study here the limit $N \to \infty$. In this so-called continuous limit we replace the bead-spring model with a continuous elastic. In the limit $N \to \infty$ the mass $M = Nm$ is kept constant. Then, instead of treating a discrete mass system, we smear the mass out along the chain and work with a continuous mass distribution. In this limit the set of discrete points $\mathbf{R}_i,\ i = 1, \ldots N$, is replaced with the continuous curve $\mathbf{R}(x, t)$, where x labels the physical points on the chain. The parameter x varies over the interval $[0,\ L_0]$ with L_0 the length of the unstretched chain. We emphasize that x labels the mass points along the chain but is *not* the arc length along the chain. The chain is flexible and may adjust its length, whereas x varies over a fixed interval. The curve $\mathbf{R}(x, t)$ is a mapping of this fixed interval onto the real chain conformation at time t. See Fig. **6.4**.

The continuous limit is often applied in modeling and it is therefore worthwhile to spend some time on it here. In this limit we distribute the mass M over increasingly more but smaller beads. Although this transition is sometimes thought to be trivial, much care is needed. We consider the equilibrium situation with all springs at their equilibrium length $b = L_0/(N-1) \approx L_0/N$ and the mass uniformly distributed along the chain. First, we multiply (6.1) at both sides with $N/L_0 \approx 1/b$:

$$\frac{Nm}{L_0}\partial_t^2 \mathbf{R}_i = -\frac{\zeta N}{L_0}\partial_t \mathbf{R}_i + \frac{\mathbf{F}_{i,i+1} + \mathbf{F}_{i,i-1}}{b}\ . \tag{6.4}$$

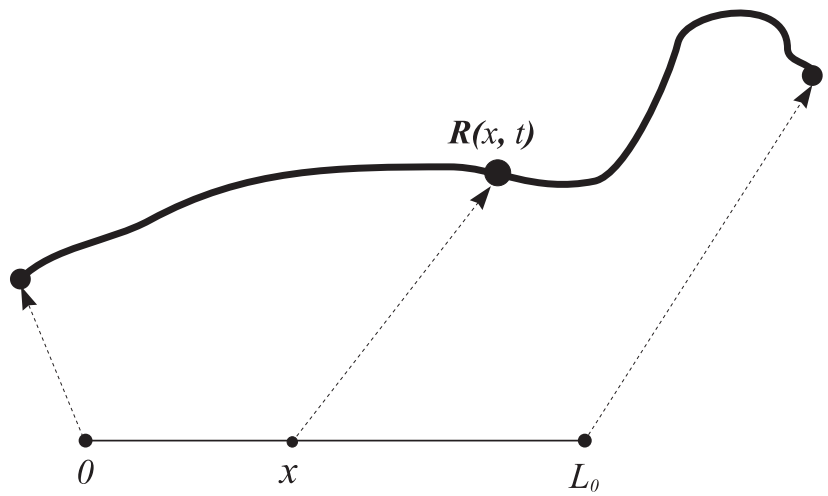

Figure 6.4. In the continuous limit a polymer chain is represented as a parameter curve $\mathbf{R}(x, t)$, with the parameter x running over the fixed parameter interval $[0, L_0]$.

In the continuous limit the following insights are important:

- $\frac{Nm}{L_0} = \frac{M}{L_0} \to \rho$, the mass per unit of length of the unstretched string.
- The friction force must scale with the mass, so $\zeta \sim m = \frac{M}{N}$. This implies that $\frac{\zeta N}{L_0} \to$ constant. In the continuous case the coefficient ζ has to be read as the friction coefficient per unit of length.
- The spring constant k is going to play the role of tension in the chain. Its value is determined by both the material properties and the boundary conditions.

In the continuous limit the unit vectors $\mathbf{u}_{i,i+1}$ and $\mathbf{u}_{i,i-1}$ both converge to the unit tangent vector of the chain at the position of bead i, but with opposite signs. So, for small b we have

$$\mathbf{u}_{i,i-1} \sim -\,\mathbf{u}_{i,i+1}\ . \tag{6.5}$$

From (6.2), (6.3), and (6.5) we find for small b

$$\mathbf{F}_{i,i+1} + \mathbf{F}_{i,i-1} \approx k\frac{\mathbf{R}_{i+1} + \mathbf{R}_{i-1} - 2\mathbf{R}_i}{b}. \tag{6.6}$$

In the limit $b \to 0$ we have that

$$\frac{\mathbf{R}_{i+1} + \mathbf{R}_{i-1} - 2\,\mathbf{R}_i}{b^2} \to \partial_x^2\,\mathbf{R}(x,t). \tag{6.7}$$

Eventually, we find that the continuous elastic chain is described by the equation of motion

$$\rho\,\partial_t^2\mathbf{R} = -\zeta\,\partial_t\mathbf{R} + k\,\partial_x^2\mathbf{R}. \tag{6.8}$$

This partial differential equation (PDE) is second order in both the time and the spatial derivative. So, we need two conditions in time, the initial profile $\mathbf{R}(x,0)$ and its time derivative $\partial_t\mathbf{R}(x,0)$, and two spatial boundary conditions.

So far, so good. The model equation (6.8) has been obtained after a number of tedious steps. It is a general necessity in modeling to check at such a stage whether the resulting model indeed describes the phenomena one is interested in. So, here we have to put the question: does (6.8) describe the dynamical behavior of a long flexible object. Without constructing the solution of this model explicitly, we can already draw some conclusions from analyzing the steady states of the model. To that end we set the time derivatives in (6.8) at zero and find that a steady state has to satisfy

$$\partial_x^2\mathbf{R}\ = 0. \tag{6.9}$$

So, all components of $\mathbf{R}$ must be linear functions of x. This implies that in the stationary state the conformation of the molecule or string is a straight line. We conclude that our model describes chains that resist bending. This is not what we intended. Apparently, in our modeling process an unwanted property crept in. It is not hard to detect when this happened: it might stem from the steps taken in (6.5) and (6.6). Repairing this shortcoming requires a more subtle treatment of the continuous limit and this will certainly complicate

the model. In practice, molecular chains never reach a stationary state because of Brownian motion, unless they are extremely cooled. Since we are mainly interested in the dynamics of such chains, we will proceed with (6.8) as our model equation.

Our conclusions concerning the steady states have implications for the boundary conditions we apply. For example, we could prescribe position and direction of the chain at one endpoint, so

$$\mathbf{R}(0,t)\ ,\quad \partial_x\mathbf{R}(0,t).$$

An alternative is to prescribe the positions at both ends:

$$\mathbf{R}(0,t)\ ,\quad \mathbf{R}(L_0,t).$$

Note that in this case the stationary state has a built-in internal stress if the distance between these two points does not equal L_0. In §6.1.4 we discuss boundary conditions that do not lead to built-in internal stress.

Exercise 6.1a. *Dimensionless equations of motion.*
It is highly useful to make (6.8) *dimensionless. In the next two subsections we will use the results of this exercise to point out how reduced versions of the equation of motion describe a string vibrating in air and a polymer molecule in a highly viscous solvent. The continuous chain has as dependent variables the components of* $\mathbf{R}$*, and as independent variables* x *and* t*. The components of* $\mathbf{R}$ *and* x *have the dimension of length and* t *of time. The parameters are* ρ*,* ζ*,* k*, and* L_0*. The parameter* ρ *has the dimension of mass divided by length.*

a. *Determine the dimensions of the parameters* ρ*,* ζ*, and* k *in* (6.8)*.*

b. *Introduce the dimensionless quantities*

$$\mathbf{R}^* = \frac{1}{L_0}\mathbf{R}\ ,\quad x^* = \frac{1}{L_0}x\ ,\quad t^* = \frac{\zeta L_0^2}{k}t. \tag{6.10}$$

Show that with these scalings (6.8) *gets the dimensionless form (omitting the * index for convenience)*

$$\alpha\,\partial_t^2\mathbf{R} = -\partial_t\mathbf{R} + \partial_x^2\mathbf{R}\ , \tag{6.11}$$

where α *is given by*

$$\alpha = \frac{\rho k}{\zeta^2 L_0^2}\ . \tag{6.12}$$

c. *Introduce the dimensionless quantities*

$$\mathbf{R}^* = \frac{1}{L_0}\mathbf{R}\ ,\quad x^* = \frac{1}{L_0}x\ ,\quad t^* = L_0\sqrt{\frac{\rho}{k}}t. \tag{6.13}$$

Show that with these scalings (6.8) *gets the dimensionless form (omitting the * index for convenience)*

$$\partial_t^2\mathbf{R} = -\beta\,\partial_t\mathbf{R} + \partial_x^2\mathbf{R}, \tag{6.14}$$

where β *is given by*

$$\beta = \frac{\zeta L_0}{\sqrt{\rho k}}. \tag{6.15}$$

6.1.3 Vibrating string

A great advantage of the dimensionless forms (6.11) and (6.14) is that the relative importance of the different terms can be weighted. This heavily relies on the system under consideration. For example, a vibrating string in a musical instrument oscillates in free air, and the friction will thus be very small. Substituting realistic values for the parameters, we find that for such a system, $\beta \ll 1$. Neglecting the friction term with β in (6.14), we find that $\mathbf{R}$ satisfies the *wave equation*

$$\partial_t^2 \mathbf{R} = \partial_x^2 \mathbf{R}. \tag{6.16}$$

A vibrating string is clamped at both sides and performs small-amplitude oscillations around a rest position. Let us take the rest position along the x-axis and let us assume that the motion is in the (x, z) plane. Writing $u := R_3$, we arrive at the vibrating string equation in the usual, scalar form

$$\partial_t^2 u = \partial_x^2 u. \tag{6.17}$$

Since a string is clamped at both sides, we have Dirichlet boundary conditions:

$$u(0, t) = 0\,, \quad u(1, t) = 0 \ \ \forall t. \tag{6.18}$$

6.1.4 Polymer dynamics

Contrary to the vibrating string, a polymer molecule in a solvent experiences a lot of friction. Substituting realistic values for the parameters, we find that $\alpha \ll 1$, whereas β is not necessarily small. For realistic polymer chains the inertia term with α in (6.11) is thus negligible compared to the other forces. The physical reason for this is that the viscous environment of the chain hinders fast accelerations of the molecules. This agrees with the observation that under Brownian motion the shape of such a chain changes rather slowly. So, this leads to the conclusion that we may take as an equation of motion

$$\partial_t \mathbf{R} = \partial_x^2 \mathbf{R}\,. \tag{6.19}$$

Comparing this to (2.19) in Chapter 2 we conclude that the motion of the chain is described by a similar equation as heat diffusion in a solid. This is a nice example of the power of mathematical modeling: one and the same model may apply to very different systems.

In the following we study the dimensionless PDE (6.19). Since it is first order in time, we only need to specify the initial configuration $\mathbf{R}_0(x) := \mathbf{R}(x, 0)$. The boundary conditions follow from taking the continuous limit of the equation of motion (6.1) for the cases $i = 1$ and $i = N$. Then, we get as boundary conditions

$$\partial_x \mathbf{R}(0, t) = \mathbf{0}\,, \quad \partial_x \mathbf{R}(1, t) = \mathbf{0} \ \ \forall t. \tag{6.20}$$

These conditions express that the endpoints are stress-free, i.e., the chain stretching vanishes at the endpoints. Note that these boundary conditions do not fix the position of the ends, so the chain may move freely.

To solve (6.19) analytically, we may apply the method outlined in Example 3.5b. To expand the solution, we take as basis functions the eigenfunctions of the operator ∂_x^2 that

satisfy the boundary conditions (6.20). So, we solve the eigenvalue problem

$$\partial_x^2 \phi = \lambda \phi \tag{6.21}$$

with boundary conditions $\partial_x \phi(0) = \partial_x \phi(1) = 0$. From this we find a discrete set of eigenvalues:

$$\lambda_n = -(n\pi)^2 , \quad n = 0, 1, 2, \ldots , \tag{6.22}$$

with corresponding eigenfunctions

$$\phi_n = A_n \cos(n\pi x). \tag{6.23}$$

The constants A_n follow from normalization, as done in the following exercise.

Exercise 6.1b.

a. *Check, by evaluating the normalization condition* $(\phi_n, \phi_n) := \int_0^1 \phi_n^2(x)\, dx = 1$, *that the normalized eigenfunctions, to be used as a basis set, are given by*

$$\phi_n(x) = \begin{cases} 1 & \text{if } n = 0, \\ \sqrt{2}\cos(n\pi x) & \text{if } n \geq 1. \end{cases}$$

b. *Check that*

$$(\phi_n, \phi_{n'}) = \delta_{n,n'} . \tag{6.24}$$

So, according to (6.24), we have an orthonormal set of basis functions, the so-called normal modes. The first few are sketched in Fig. **6.5**. The $\{\phi_n\}$ form a basis in $\mathcal{L}^2([0, 1])$.

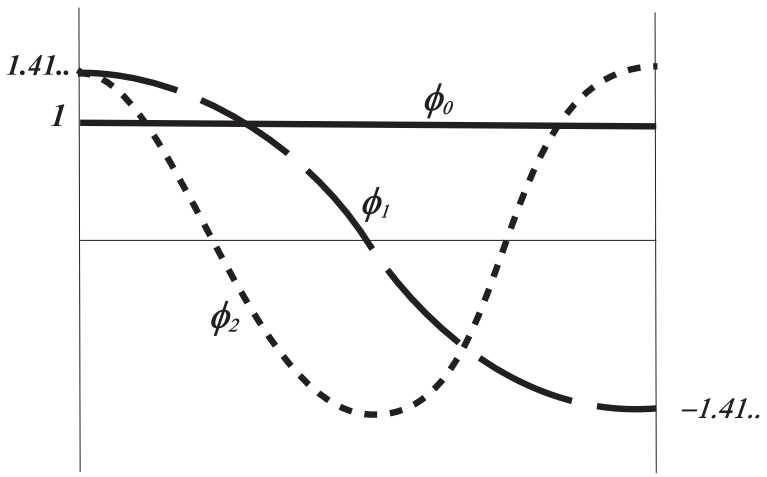

Figure 6.5. The basis functions or normal modes for n = 0, 1, and 2, used to describe the motion of a molecular chain in a dilute solution.

For any value of time t we assume the components of $\mathbf{R}(x,t)$ to belong to this space. So, we obtain the expansion

$$\mathbf{R}(x,t) = \sum_{n=0}^{\infty} \mathbf{r}_n(t)\,\phi_n(x) \tag{6.25}$$

with vector coefficients $\mathbf{r}_n \in \mathbb{R}^3$. Note that this representation yields a nice separation of variables. One can read $\phi_n(x)$ as a scalar standing wave with the amplitude $\mathbf{r}_n(t)$ giving both its orientation in space and its time dependence.

Having obtained a general expression for $\mathbf{R}(x,t)$, it remains to determine the $\mathbf{r}_n(t)$ from substituting the series into the PDE (6.19). This leads to

$$\sum_{n=0}^{\infty} \partial_t\,\mathbf{r}_n(t)\,\phi_n = \sum_{n=0}^{\infty} \mathbf{r}_n(t)\,\lambda_n\,\phi_n(x). \tag{6.26}$$

Taking the inner product of both sides with a particular basis function and using the orthonormality of the ϕ_n, we find that

$$\partial_t\,\mathbf{r}_n = \lambda_n\,\mathbf{r}_n\ ,\quad n = 0, 1, 2, \ldots, \tag{6.27}$$

with solution

$$\mathbf{r}_n(t) = e^{\lambda_n t}\,\mathbf{r}_n(0)\ . \tag{6.28}$$

The initial values $\mathbf{r}_n(0)$ follow from the inner product of each component of the initial profile $\mathbf{R}_0(x)$ with the nth basis function. We summarize this as

$$\mathbf{r}_n(0) = (\phi_n, \mathbf{R}_0). \tag{6.29}$$

Eventually, the solution of (6.19) reads as

$$\mathbf{R}(x,t) = \sum_{n=0}^{\infty} \mathbf{r}_n(0)\,e^{\lambda_n t}\,\phi_n(x)\ . \tag{6.30}$$

From this expression with decaying exponentials it is seen that the effect of the initial condition is "forgotten" if $t \to \infty$.

Exercise 6.1c.

a. *Conclude from* (6.30) *the chain behavior if* $t \to \infty$. *To which length does the chain converge?*

b. *Find the typical dimensionless decay time and from this the dimensionful decay time. Use in your argument that* λ_n *is given by* (6.22).

Exercise 6.1d.

a. *We change the boundary conditions in* (6.20) *such that*

$$\partial_x\mathbf{R}(0,t) = \partial_x\mathbf{R}(1,t) = (1,0,0) \quad \forall t. \tag{6.31}$$

The physical meaning of these conditions is twofold: first, the internal stress in the string at the endpoints is prescribed, and, second, the orientation of the string in space is fixed at the endpoints. Derive the general solution of (6.19) *in the case of these boundary conditions along the same lines as in* (6.21)–(6.30).

b. *Answer the questions in Exercise* 6.1c *but now for the solution derived under* a *of the present exercise.*

6.1.5 Brownian motion

In polymer melt dynamics the Brownian motion is a dominating effect. The next step is therefore to include this stochastic force.

The solvent molecules not only hinder the chain in its motion but they also exert forces on the chain molecules via collisions, since they themselves move in a random fashion due to their thermal energy. This *Brownian motion* has the effect that the chain will never attain a fixed, equilibrium configuration but will continuously change its shape. We include the *Brownian motion* by incorporating a stochastic force $\mathbf{f}$ in the dimensionless equation of motion (6.19):

$$\partial_t \mathbf{R} = \partial_x^2 \mathbf{R} + \mathbf{f} \; . \tag{6.32}$$

For a specific chain the action of $\mathbf{f}$ is unpredictable. We can specify only certain average properties. To that end we must realize that in the solution many chains are present, which all experience Brownian forces. Each chain in the ensemble follows its own trajectory and attains its own configurations, depending on the random forces exerted on that chain. Any property of these chains, e.g., the stretching at position x and time t, is thus a stochastic variable. To calculate properties of an "average chain," we just average the variable under consideration over the ensemble. Such an ensemble average is indicated by $\langle \ldots \rangle$.

The stochastic Brownian force $\mathbf{f}(x,t)$ at position x and time t is characterized by specification of its first two moments:

a. $$\langle \mathbf{f}(x,t) \rangle \; = \; 0 \;\; \forall x \in [0,1] \, , \;\; \forall t. \tag{6.33}$$

b. $$\langle \mathbf{f}(x,t) \mathbf{f}(x',t') \rangle \; = \; c_B \, \delta(x-x') \, \delta(t-t') \, \mathbf{I}. \tag{6.34}$$

Property a expresses that the direction of the Brownian force is random so that the mean value of this force vanishes on average. The delta functions $\delta(x-x')$ and $\delta(t-t')$ in property b express that the Brownian forces at different positions along a chain and at different times are completely uncorrelated. The unit matrix $\mathbf{I}$ indicates that also the different components of $\mathbf{f}(x,t)$ at position x and time t are uncorrelated. The coefficient c_B is a measure for the strength of the fluctuations $\mathbf{f}$. From the so-called fluctuation-dissipation theorem—a theory that goes beyond the scope of this book—it is known that c_B is linearly proportional to both the temperature of the solution and the friction coefficient ζ.

The extra force $\mathbf{f}$ hardly changes the derivation in the steps (6.19)–(6.27). After inclusion of the Brownian force, (6.27) now reads as

$$\partial_t \mathbf{r}_n = \lambda_n \mathbf{r}_n + \mathbf{f}_n \tag{6.35}$$

with $\mathbf{f}_n$ given by

$$\mathbf{f}_n = (\phi_n, \mathbf{f}). \tag{6.36}$$

Just as in (6.29), this inner product has to be read per component of $\mathbf{f}$. Equation (6.35) is an inhomogeneous linear equation as introduced in (3.11). Its solution is thus given by the variation of constants formula (3.16):

$$\mathbf{r}_n(t) = e^{\lambda_n t}\,\mathbf{r}_n(0) + \int_0^t e^{\lambda_n (t-t')}\,\mathbf{f}_n(t')\,dt'. \tag{6.37}$$

Substituting this expression in (6.25), we can calculate any average property of the ensemble of chains we like. For that we need the statistical properties of $\mathbf{f}_n$. They directly follow from the defining properties (6.33) and (6.34) of $\mathbf{f}$ itself:

$$\begin{aligned} \langle \mathbf{f}_n(t) \rangle &= \left\langle \int_0^1 \phi_n(x) \mathbf{f}(x,t)\,dx \right\rangle \\ &= \int_0^1 \phi_n(x) \langle \mathbf{f}(x,t) \rangle\,dx \\ &= 0 \ \forall t, \end{aligned} \tag{6.38}$$

$$\begin{aligned} \langle \mathbf{f}_n(t)\,\mathbf{f}_{n'}(t') \rangle &= \left\langle \int_0^1 \phi_n(x)\mathbf{f}(x,t)\,dx \int_0^1 \phi_{n'}(x')\mathbf{f}(x',t')\,dx' \right\rangle \\ &= \int_0^1 \int_0^1 \phi_n(x)\phi_{n'}(x') \langle \mathbf{f}(x,t)\,\mathbf{f}(x',t') \rangle\,dx\,dx' \\ &= c_B\,\delta(t-t')\,\mathbf{I} \int_0^1 \phi_n(x)\phi_{n'}(x)\,dx \\ &= c_B\,\delta(t-t')\,\mathbf{I}\,(\phi_n, \phi_{n'}) \\ &= c_B\,\delta(t-t')\,\delta_{n,n'}\,\mathbf{I}. \end{aligned} \tag{6.39}$$

Center of mass

For the diffusion of a chain as a whole through the solvent, only the motion of its center of mass is relevant. It is given by the average of all positions along the chain:

$$\mathbf{R}_{cm}(t) = \int_0^1 \mathbf{R}(x,t)\,dx. \tag{6.40}$$

Exercise 6.1e.

a. *Show that* $\int_0^1 \phi_n(x)\,dx = 0$ *if* $n \geq 1$ *and* $\int_0^1 \phi_0(x)\,dx = 1$.

b. *Use these properties to show that*

$$\mathbf{R}_{cm}(t) = \mathbf{r}_0(t) = \mathbf{R}_{cm}(0) + \int_0^t \mathbf{f}_0(t')\,dt' \tag{6.41}$$

by substituting (6.25) *with* (6.37) *in* (6.40). *So,* $\mathbf{R}_{cm}(t)$ *follows a stochastic path.*

c. *Show that for the average center of mass position it holds that*

$$\langle \mathbf{R}_{cm}(t) \rangle = \mathbf{R}_{cm}(0). \tag{6.42}$$

So, although the chain may travel a lot, its center of mass stays on average in the initial position, since the direction of the Brownian motion is random.

From (6.41) we may conclude that the longer we wait, the farther $\mathbf{R}_{cm}(t)$ may temporarily travel from its starting position $\mathbf{R}_{cm}(0)$. The average distance of $\mathbf{R}_{cm}(t)$ to $\mathbf{R}_{cm}(0)$ is measured by the standard deviation of this difference. It is given by the square root of the variance

$$\langle |\mathbf{R}_{cm}(t) - \mathbf{R}_{cm}(0)|^2 \rangle \; = \int_0^t \int_0^t \langle (\mathbf{f}_0(t'), \mathbf{f}_0(t'')) \rangle \, dt' \, dt'', \tag{6.43}$$

where we used (6.41).

Exercise 6.1f.
Show that this variance is given by

$$\langle |\mathbf{R}_{cm}(t) - \mathbf{R}_{cm}(0)|^2 \rangle \; = \; \text{Trace}(\mathbf{I}) \; c_B \; t = 3 \; c_B \; t. \tag{6.44}$$

So, the standard deviation of $\mathbf{R}_{cm}(t) - \mathbf{R}_{cm}(0)$ scales with $\sqrt{t}$. This is typical for a random walk process in three dimensions. Brownian motion is a standard example of such a process.

End-to-end vector
The end-to-end vector $\widehat{\mathbf{R}}(t)$, sketched in Fig. **6.6**, is defined as

$$\widehat{\mathbf{R}}(t) = \mathbf{R}(1, t) - \mathbf{R}(0, t). \tag{6.45}$$

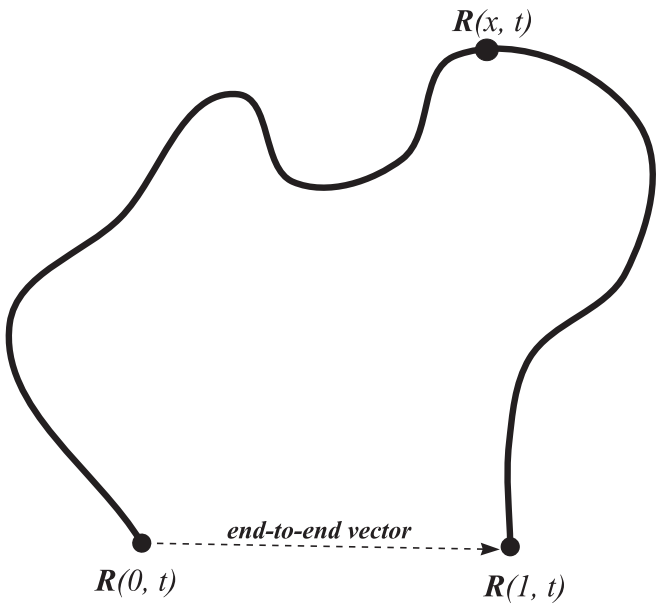

Figure 6.6. The end-to-end vector of a long polymer or DNA chain.

Exercise 6.1g.
Show that

$$\widehat{\mathbf{R}}(t) = -\sqrt{8} \sum_{n=1,3,5,\ldots} \mathbf{r}_n(t). \tag{6.46}$$

Memory effects
The correlation $\langle \widehat{\mathbf{R}}(t)\widehat{\mathbf{R}}(0)\rangle$, with the dyadic product defined as $(\widehat{\mathbf{R}}\widehat{\mathbf{R}})_{ij} := \widehat{\mathbf{R}}_i\widehat{\mathbf{R}}_j$, provides information about the memory of the system with respect to its initial conformation. If time proceeds, $\widehat{\mathbf{R}}(t)$ will shrink, expand, and rotate in a stochastic way. From (6.46) we have that

$$\left\langle \widehat{\mathbf{R}}(t)\widehat{\mathbf{R}}(0)\right\rangle \ = 8 \sum_{n,m=1,3,5,\ldots} \langle \mathbf{r}_n(t)\, \mathbf{r}_m(0)\rangle. \tag{6.47}$$

Exercise 6.1h.
a. *Show that*

$$\left\langle \widehat{\mathbf{R}}(t)\widehat{\mathbf{R}}(0)\right\rangle \ = 8\, \mathbf{I} \sum_{n=1,3,\ldots} \frac{1}{\lambda_n} e^{-\lambda_n t} \ . \tag{6.48}$$

b. *Show that*

$$\left(\widehat{\mathbf{R}}(t), \widehat{\mathbf{R}}(0)\right) = \text{Trace } \left(\widehat{\mathbf{R}}(t)\widehat{\mathbf{R}}(0)\right) \ , \tag{6.49}$$

and conclude from this that

$$\left\langle \left(\widehat{\mathbf{R}}(t), \widehat{\mathbf{R}}(0)\right)\right\rangle \ = 24 \sum_{n=1,3,\ldots} \frac{1}{\lambda_n} e^{-\lambda_n t}. \tag{6.50}$$

The spectrum $\{\lambda_n\}$ of relaxation rates determines the "fading" memory of the system. It is clear that the $n = 1$ term strongly determines the sum. So,

$$\tau := \lambda_1^{-1} \tag{6.51}$$

is the characteristic decay time (in dimensionless units) of the system.

Exercise 6.1i.
Let us introduce the position vector $\mathbf{P}(t)$ *by*

$$\mathbf{P}(t) = \tfrac{1}{2}[\mathbf{R}(0,t) + \mathbf{R}(1,t)] \ . \tag{6.52}$$

a. *Give the geometrical interpretation of* $\mathbf{P}(t)$.
b. *Expand* $\mathbf{P}(t)$ *in terms of the basis functions* $\phi_n(x)$*, using* (6.25) *with* (6.37).
c. *Calculate the average distance vector* $\langle \mathbf{R}_{cm}(t) - \mathbf{P}(t)\rangle$ *and interpret the result.*
d. *Calculate also the variance* $\langle |\mathbf{R}_{cm}(t) - \mathbf{P}(t)|^2\rangle$ *and interpret the result.*

6.2 Fiber spinning

The following case study forms an illustration of the concepts dealt with in
§1.3 (dimensional analysis),
§2.3 (constitutive relations),
§2.5 (conservation of mass),
§2.7 (conservation of momentum),
§4.1 (stability), and
§4.3 (linearization).

In this case study we model the dynamical behavior of some aspects of the fiber spinning process. During the spinning of fibers a molten polymer is extruded through a small hole, the die, in the wall of some barrel. The fiber is stretched in the air and then wound around a chill roll. The geometry of the process is sketched in Fig. **6.7**. The stretching of the fiber is forced via the chill roll either by maintaining a fixed rotational speed or by maintaining a fixed stress level in the longitudinal direction of the fiber. In the air and, mostly, on the roll the fiber cools and solidifies. When the fiber runs through the air, its radial cross section decreases. The consequence is that its axial speed increases, since the total volume flux in the axial direction must be conserved. In the spinning practice it has been found that this system can become unstable. Then, the shape of the fiber is no longer constant, but oscillating. This in particular may happen if the spinning speed exceeds a critical value, i.e., if the chill roll rotates too fast. But also other parameters appear to be of influence, such as the diameter of the die and the elastic properties of the fiber material. The phenomenon of unstable dynamics limits production rates. So, it is important to understand

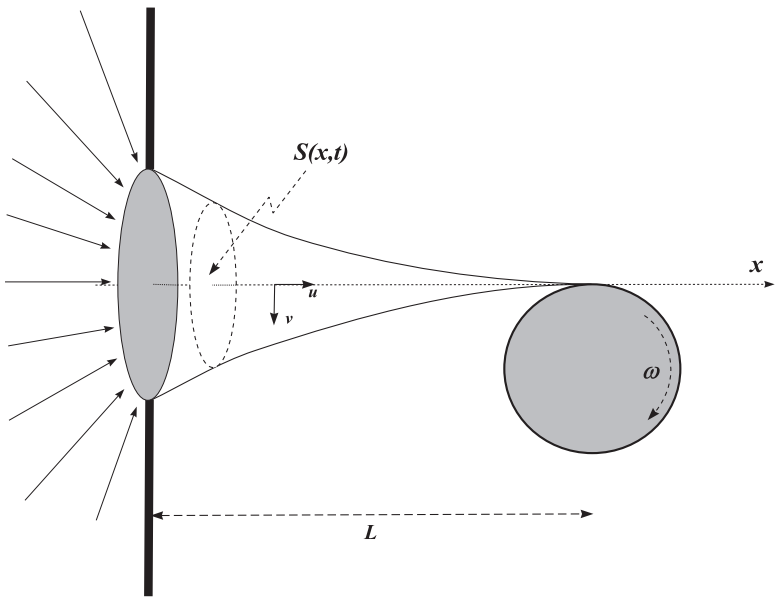

Figure 6.7. Sketch of the fiber spinning process. The scales shown are not realistic. The fiber is long and very thin, so its length L is much larger than its radius.

the mechanisms leading to it. If reliable models would become available to predict the occurrence of instabilities, it might be possible to avoid them by choosing the parameters in a "safe" range. Here we want to develop a relatively simple model that includes enough essential characteristics of the system to describe the phenomena. The aim is to determine whether this model describes the transition from stable to unstable behavior of the fiber and, in particular, for which parameter values the model predicts that this will happen. In practice, such a model could be used to improve the efficiency of the spinning process by indicating how the parameters should be chosen to postpone unstable dynamics.

The modeling of the spinning system may serve as a general exercise, since it is typical for many flow problems. We first discuss the consequences of the conservation of mass and momentum. The special structure of the system—a long, thin fiber—allows for a reduction of the equations. From the reduced model the stationary state is easily deduced. The stability analysis of this state is not that simple. A fully analytical treatment is hard to present. Instead, we will formulate a promising numerical approach. For further reading about fiber spinning, see [27, 32, 34].

6.2.1 Modeling fiber spinning

The quantity we are interested in is the geometry of the fiber. If the process takes place under stable conditions, this geometry is expected to be stationary. As sketched in Fig. **6.8** we assume the fiber to run horizontally with the x-axis as the line connecting the center of the die and the top of the chill roll. One could ask whether gravity is of influence on the geometry. To investigate that systematically we could introduce gravity in the model and study the robustness of the model with respect to this effect. In the first instance we prefer to ignore gravity, to keep the model tractable. We expect this choice to be reliable in view of the high stresses dominating the dynamics of the fiber. Without gravity the form of the fiber in the stationary state will be symmetric around the x-axis. This suggests the use of cylindrical coordinates (x, r, ϕ). Let us denote the cross section perpendicular to the x-axis by $S(x,t)$ with area $A(x,t)$. The model should yield the profile $A(x,t),\ 0 \le x \le L,\ t \ge 0$, as a result. The other important quantity of the system is the velocity profile $\mathbf{v}$. In view of the cylindrical symmetry, it has two components, an axial one u and a radial one v, so $\mathbf{v} = (u, v)$. Since the cross section area $A(x,t)$ decreases as x increases, we can directly conclude that the radial velocity v must be in the direction of the center line of the fiber. The fiber is a long, thin object, and the radial speeds will be orders of magnitude smaller than the axial one. That is why in the following we shall average over the cross section and leave out the details of the dynamics in radial direction. This implies that v will not play a role.

Conservation of mass

Let ρ be the mass density in the fiber. Since the material is incompressible, ρ is constant. From §2.4 we know that conservation of mass is expressed by

$$\frac{d}{dt}\int_V \rho \, dV + \int_{\partial V} \rho \, \mathbf{v} \cdot \mathbf{n} \, dA = 0 \tag{6.53}$$

for an arbitrary volume V with surface ∂V within the fiber. Here, $\mathbf{n}$ is outward normal to the surface ∂V. Since ρ is constant, this quantity drops out. In §2.4 we took only for

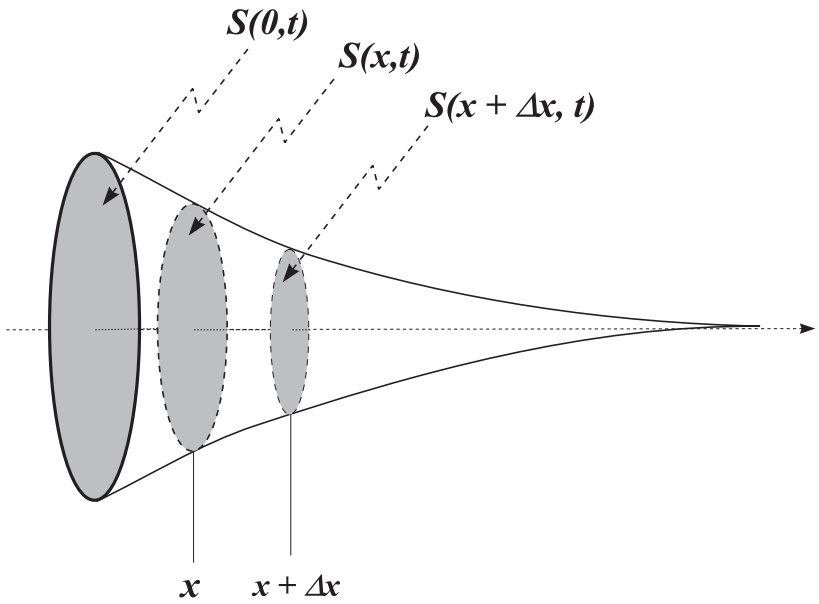

Figure 6.8. Volume V with surface ∂V between two radial cross sections at x and $x + \Delta x$.

V volumes that are fixed in time. Note that (6.53) is more general and also holds if the volume deforms when following the flow. However, then it is not allowed to interchange derivation and integration. Let us take for V the volume at time t between the two cross sections at x and $x + \Delta x$ with $\Delta x > 0$, as indicated in Fig. **6.8**. For small Δx, the volume V is approximately part of a cone. The material enters and leaves this volume only through the cross sections $S(x, t)$ and $S(x + \Delta x, t)$. From these insights (6.53) can be rewritten in the form

$$\frac{d}{dt}\int_x^{x+\Delta x} A(x', t)\, dx' + \left[\int_{S(x+\Delta x,t)} u\, dA - \int_{S(x,t)} u\, dA\right] = 0\,. \tag{6.54}$$

Introducing the axial velocity averaged over a cross section by

$$\bar{u}(x, t) := \frac{\int_{S(x,t)} u\ dA}{A(x, t)}\,, \tag{6.55}$$

we may rewrite (6.54) in the form

$$\frac{d}{dt}\int_x^{x+\Delta x} A(x', t)\, dx' = -[\bar{u}(x + \Delta x, t)A(x + \Delta x, t) - \bar{u}(x, t)A(x, t)]\,. \tag{6.56}$$

Dividing by Δx and taking the limit $\Delta x \to 0$, we obtain for this system the mass balance in differential form:

$$\partial_t A = -\partial_x[\bar{u}A]\,. \tag{6.57}$$

Conservation of momentum
To derive the momentum balance, we proceed as above for mass conservation. So, we start from the conservation equation in its most elementary form, as expressed in §2.7.2. The momentum balance then reads as

$$\frac{d}{dt}\int_V \rho\,\mathbf{v}\,dV = \int_{\partial V} \boldsymbol{\sigma}\cdot\mathbf{n}\,dA \tag{6.58}$$

with σ the stress tensor introduced in §2.8 and $\mathbf{n}$ the outward normal. It will appear soon that we only need the (x,x)-component of the stress tensor. For shortness in notation we shall denote this component by τ, so

$$\tau(\mathbf{x},t) := \sigma_{11}(\mathbf{x},t).$$

Taking the volume V as above, we note that at the free surface of the fiber the normal stress vanishes. The surface integral thus runs only over the cross sections $S(x,t)$ and $S(x+\Delta x,t)$. The corresponding normal vectors are given by $\mathbf{n}=(-1,0)$ and $\mathbf{n}=(1,0)$, respectively. In the following we will average over variations in the y-direction. So, only the equation for the x-component is relevant. For that we obtain

$$\rho\frac{d}{dt}\int_V u\,dV = \int_{S(x+\Delta x,t)} \tau\,dA - \int_{S(x,t)} \tau\,dA\,. \tag{6.59}$$

Using (6.55) we can write the integral on the left-hand side of (6.59) as

$$\int_V u\,dV = \int_x^{x+\Delta x}\left[\int_{S(x',t)} u\,dA\right]dx' = \int_x^{x+\Delta x} \bar{u}\,A\,dx'\,. \tag{6.60}$$

Just as for the velocity, it is convenient to average the axial stress over a cross section. We thus define

$$\bar{\tau}(x,t) := \frac{\int_{S(x,t)}\tau\,dA}{A(x,t)}\,. \tag{6.61}$$

This definition allows us to write (6.59) in the form

$$\rho\frac{d}{dt}\int_x^{x+\Delta x}\bar{u}A\,dx' = \bar{\tau}(x+\Delta x,t)A(x+\Delta x,t) - \bar{\tau}(x,t)A(x,t). \tag{6.62}$$

Dividing by Δx and taking the limit $\Delta x\to 0$ we obtain the momentum equation in the form

$$\rho\,\partial_t[\bar{u}\,A] = \partial_x[\bar{\tau}\,A]. \tag{6.63}$$

6.2.2 Newtonian flow

To charaterize the flow properties of the fiber we have to specify a constitutive relation for the dependence of the stress on the deformation. As explained in §2.7.3, purely viscous materials like water satisfy a constitutive relation in which the stress is directly proportional to the gradient of the velocity. In the present case this leads to

$$\tau = \eta\,\partial_x u. \tag{6.64}$$

Here, η is the viscosity. If we average both sides over the cross section, we get the relation

$$\bar{\tau} = \eta\, \partial_x \bar{u}. \tag{6.65}$$

Viscoelastic materials are described by more intricate constitutive relations. In the present case we restrict ourselves to Newtonian flow to avoid complications that may obscure the analysis. If this relatively simple description would already lead to unstable behavior of the fiber, we may expect that more complicated constitutive relations will do the same.

Mathematical model
Summarizing the achievements above, we have the model equations

$$\begin{aligned} \partial_t A &= -\partial_x[\bar{u}A], \\ \rho\, \partial_t[\bar{u}A] &= \eta\, \partial_x[A\, \partial_x \bar{u}] \end{aligned} \tag{6.66}$$

for the variables $A(x,t)$ and $\bar{u}(x,t)$. As for the boundary conditions, we need one condition for A and two for $\bar{u}(x,t)$, since the first derivative of A and the second derivative of $\bar{u}(x,t)$ are involved. The one for A reads as

$$A(0,t) = A_0\ ,\quad t \geq 0\ , \tag{6.67}$$

with A_0 the area of the die. Since the flow leaves the barrel at constant speed, we have the condition

$$\bar{u}(0,t) = u_0\ ,\quad t \geq 0\ . \tag{6.68}$$

At the side of the chill roll where $x = L$ we could prescribe the speed

$$\bar{u}(L,t) = u_L\ ,\quad t \geq 0\ . \tag{6.69}$$

An alternative is to keep the stress fixed:

$$\bar{\tau}(L,t)A(L,t) = \eta\, A(L,t)\, \partial_x \bar{u}(L,t) = F_c\ ,\quad t \geq 0\ . \tag{6.70}$$

Dimensionless model
As explained in Chapter 1, it is always convenient to nondimensionalize the model. Straightforward scalings are

$$x^* = \frac{x}{L},\quad t^* = \frac{u_0\, t}{L},\quad A^* = \frac{A}{A_0},\quad \bar{u}^* = \frac{\bar{u}}{u_0}. \tag{6.71}$$

If we apply (6.71) and omit the $*$ index for convenience, we get equations that are quite similar to the dimensionful model (6.66):

a. $$\partial_t A = -\partial_x[\bar{u}A], \tag{6.72}$$

b. $$\partial_t[\bar{u}A] = \alpha\, \partial_x[A\, \partial_x \bar{u}].$$

The only parameter in these equations is

$$\alpha := \frac{\eta}{\rho\, L\, u_0}. \tag{6.73}$$

This allows for an important conclusion. Since $\rho \approx 1$, the order of magnitude of α is determined by the ratio of the viscosity η and the product $L\,u_0$. If the viscosity dominates the dynamics, i.e., $\alpha \gg 1$, the left-hand side of (6.72b) could be omitted so that this equation would considerably reduce. On the other hand, if $\alpha \ll 1$, the right-hand side of (6.72b) could be omitted, which would imply that the flux $\bar{u}A$ is constant in time.

The dimensionless equivalents of (6.67) and (6.68) are

$$A(0,t) = 1\ ,\quad \bar{u}(0,t) = 1\ ,\quad t \geq 0\ . \tag{6.74}$$

Exercise 6.2a.

Make boundary conditions (6.69) *and* (6.70) *dimensionless, using* (6.71).

6.2.3 Stationary solution

To determine the stationary solution of our model, we assume all profiles to be time independent so that the derivatives with respect to time vanish.

Exercise 6.2b.

Show that the stationary solution of (6.72) *is given by*

$$\bar{u}^s(x) = a\,e^{b\,x}\ ,\quad A^s(x) = c\,e^{d\,x}, \tag{6.75}$$

and determine the constants a, b, c, and d from the boundary conditions. Treat the cases (6.69) *and* (6.70) *separately.*

6.2.4 Stability analysis

We are interested in the stability of the stationary profiles $\bar{u}^s(x)$ and $A^s(x)$ in (6.75), since in practice the fiber spinning process may become unstable, especially if the speed u_L of the chill roll is increased. In view of the observed instabilities we assume that the cylindrical symmetry is always conserved. A reasonable question is whether a critical speed u_{crit} of the chill roll exists, such that the process is stable for $u_L < u_{crit}$ and unstable for $u_L > u_{crit}$. Analogously, we could put forth the question of whether a critical value F_{crit} for the parameter F_c exists. If such a critical speed or force exists, the next question is to determine how it depends on the parameters ρ, η, L, and A_0. In Chapter 4 we showed how the stability of the stationary profiles can be investigated. The idea is to perturb them slightly and to observe the time evolution of the perturbations. If all possible perturbations damp out, we may conclude that the stationary state is asymptotically stable. If at least one perturbation can be found that grows in time, the system state is unstable.

To apply standard linear stability analysis, we linearize the model equations around the stationary solution, given by (6.75).

Exercise 6.2c.

Since in the model (6.72) *the product $A\bar{u}$ is present, which essentially is the material flux, it might be convenient to apply the transformation $p := A$ and $q := A\,\bar{u}$.*

a. *Show that in terms of* p, q *the model reads as*

$$\partial_t p = -\partial_x q, \tag{6.76}$$

$$\partial_t q = \alpha\, \partial_x \left[p\, \partial_x \left\{ \frac{q}{p} \right\} \right] \tag{6.77}$$

with at $x = 0$ *the boundary conditions*

$$p(0,t) = 1\ ,\quad q(0,t) = 1\ ,\quad t \geq 0, \tag{6.78}$$

and at $x = 1$ *either (in case of* (6.69)*)*

$$\frac{q}{p} = \frac{u_L}{u_0}\ ,\quad t \geq 0, \tag{6.79}$$

or (in case of (6.70)*)*

$$p\, \partial_x \left\{ \frac{q}{p} \right\} = \frac{L\, F_c}{\eta\, A_0\, u_0} := F_c^*\ ,\quad t \geq 0\ . \tag{6.80}$$

b. *Give the stationary solutions* p^s *and* q^s *using the results in* (6.75).

c. *Linearize* (6.76) *and* (6.77) *around the stationary state by writing* $p = p^s + p_1$ *and* $q = q^s + q_1$ *and show that the linearized equations for the perturbations* p_1 *and* q_1 *take the form*

$$\partial_t p_1 = -\partial_x q_1\ , \tag{6.81}$$

$$\partial_t q_1 = \alpha\, \partial_x [p^s\, \partial_x (f\, q_1 - g\, p_1) + h\, p_1],$$

where, for convenience, we introduce the notation

$$f(x) := \frac{1}{p^s}\ ,\quad g(x) := \frac{q^s}{(p^s)^2}\ ,\quad h(x) := \partial_x \left\{ \frac{q^s}{p^s} \right\}. \tag{6.82}$$

The perturbations p_1 and q_1 vanish at $x = 0$ for all times, since the unperturbed solutions p^s and q^s already satisfy the boundary conditions there. The boundary conditions at $x = 1$ are more intricate, as the following exercise shows.

Exercise 6.2d.

a. *Show by writing* $p = p^s + p_1$ *and* $q = q^s + q_1$ *in* (6.79) *that in that case the boundary condition for* p_1 *and* q_1 *at* $x = 1$ *is*

$$\frac{p_1}{q_1} = \frac{p^s}{q^s}\ ,\quad t \geq 0. \tag{6.83}$$

b. *Show by writing* $p = p^s + p_1$ *and* $q = q^s + q_1$ *in* (6.80) *that in that case the boundary condition for* p_1 *and* q_1 *at* $x = 1$ *is*

$$p^s\, \partial_x (f\, q_1 - g\, p_1) + h\, p_1 = 0\ ,\quad t \geq 0. \tag{6.84}$$

Crucial for stability or instability is the question of how initial perturbations $p_1(x,0)$ and $q_1(x,0)$ develop for $t > 0$. The next step is to investigate whether the linearized model has solutions of the form

$$p_1(x,t) = e^{\lambda t} r(x)\ , \quad q_1(x,t) = e^{\lambda t} s(x) \tag{6.85}$$

for some eigenvalue λ and functions $r(x)$ and $s(x)$ that satisfy the boundary conditions. Substitution of this Ansatz leads to an eigenvalue problem of the form

$$\lambda\, r = -\partial_x s, \tag{6.86}$$

$$\lambda\, s = \alpha\, \partial_x [p^s\, \partial_x (f\, s - g\, r) + h\, r]. \tag{6.87}$$

Note that the boundary conditions for r and s are the same as for p_1 and q_1 and thus similar to (6.83) or (6.84).

It is to be expected that solutions satisfying the boundary conditions exist only for special values of λ. These values together form the so-called spectrum of the linearized problem. This spectrum depends on the values of the parameters. If the spectrum contains at least one λ-value with positive real part, the process is unstable.

In general it is not simple to solve (6.86) and (6.87). Therefore, it is worth investigating the special cases $\alpha \ll 1$ and $\alpha \gg 1$. If $\alpha \ll 1$, the right-hand side of (6.87) can be omitted. Then, it is easily concluded that the only solutions are $r = s = 0$. So, the model does not admit perturbations if the viscosity is relatively small. In the practice of fiber spinning, the case $\alpha \gg 1$ is much more relevant. Then, the left-hand side of (6.87) can be omitted, which reduces the problem considerably. To interpret this reduction, it is convenient to go back to the model before it was put in dimensionless form. If we omit the left-hand side of (6.72b), we find that $A\, \partial_x \bar{u}$ must be uniform along the fiber, i.e., constant in space. Note that $A\, \partial_x \bar{u}$ is proportional to the axial stress. At this point a clear difference between boundary conditions (6.69) and (6.70) arises. If boundary condition (6.70) is in force, the axial stress is determined by the parameter F_c at the chill roll. In dimensionless units we then have

$$A^*\, \partial_x \bar{u}^* = F_c^* \tag{6.88}$$

everywhere along the fiber and constant in time. However, if boundary condition (6.69) applies, the axial stress is also uniform along the fiber, but not necessarily constant in time. The next exercise shows that it is this freedom that admits the possibility of unstable behavior.

Exercise 6.2e.

As discussed above, for $\alpha \gg 1$ we conclude that the quantity in (6.87) *between square brackets does not depend on x. So,*

$$p^s\, \partial_x (f\, s - g\, r) + h\, r = c \tag{6.89}$$

for some constant c. Note that $c = 0$, if boundary condition (6.70) *applies, since in that case fluctuations in the axial stress are not allowed, as explained above.*

a. *Rewrite the linearized problem* (6.86) *and* (6.87) *in the standard form*

$$\partial_x \mathbf{r} = \mathbf{M}(x,\lambda)\mathbf{r} + \mathbf{b} \tag{6.90}$$

with $\mathbf{r} := (r, s)$ *and* $\mathbf{M}$ *a* 2×2 *matrix. The system* (6.86), (6.87) *is of second order for the two variables* (r, s). *Rewriting this system in first-order form, one expects a state vector of length* 4, *namely,* $(r, s, \partial_x r, \partial_x s)$. *However, in view of* (6.86) $\partial_x s$ *can be expressed in terms of* r, *and in view of* (6.89) $\partial_x r$ *can be expressed in terms of* r *and* s. *So, it suffices to take as a state vector* (r, s).

Determine the elements of the matrix $\mathbf{M}$ *and the vector* $\mathbf{b}$.

b. *Conclude that in the case of boundary condition* (6.70), *when* $c = 0$ *and thus* $\mathbf{b} = \mathbf{0}$, *the linearized problem has no nontrivial solution. What does this imply for the stability of the stationary state?*

c. *Investigate in the case of* (6.69) *for which* λ*-values* (6.90) *with boundary conditions* (6.83) *has a nontrivial solution. Note that the value of* c *in* (6.89) *is not important as long as we have* $c \neq 0$. *That is because the perturbation functions* r *and* s *may also be scaled with an arbitrary nonvanishing factor. Use this freedom to choose a convenient value for* c *and thus* $\mathbf{b}$.

Exercise 6.2f.

To determine the spectrum of the linearized problem, one has to invoke numerical methods. One could use a shooting method, i.e., an iterative procedure to determine λ *such that* (6.83) *is satisfied. In such a method the system* (6.90) *is integrated over the interval* $[0 \leq x \leq 1]$ *for some trial value of* λ *with vanishing initial conditions. In general, the obtained solution will not satisfy boundary condition* (6.83) *at* $x = 1$. *In each iteration step the value of* λ *is adjusted such that the solution more accurately satisfies* (6.83).

a. *Conclude from the spectrum obtained this way whether the stationary solution is stable or not.*

b. *Investigate whether a critical value for* u_{crit} *exists such that the system is unstable if for the velocity of the chill roll it holds that* $u_L > u_{crit}$. *If this is the case, how then does* u_{crit} *depend on the other parameters?*

6.3 Surface water waves

The following case study forms an illustration of the concepts dealt with in
§2.7.3 (Navier–Stokes equations),
§3.3.1 (directional derivative),
§3.3.4 (plane waves and Fourier transforms),
§3.3.5 (group velocity), and
Chapter 5 (variational modeling).

In this section we mainly illustrate a nontrivial application of variational modeling as described in general terms in Chapter 5. To that end we extend the Challenging problem in §2.9 and formulate the water wave problem in a much more general form. It is worthwhile to read §2.9 before studying this section.

We consider the motion of waves on a layer of fluid and make some simplifying assumptions, only meant to make the presentation more transparent, restricting ourselves

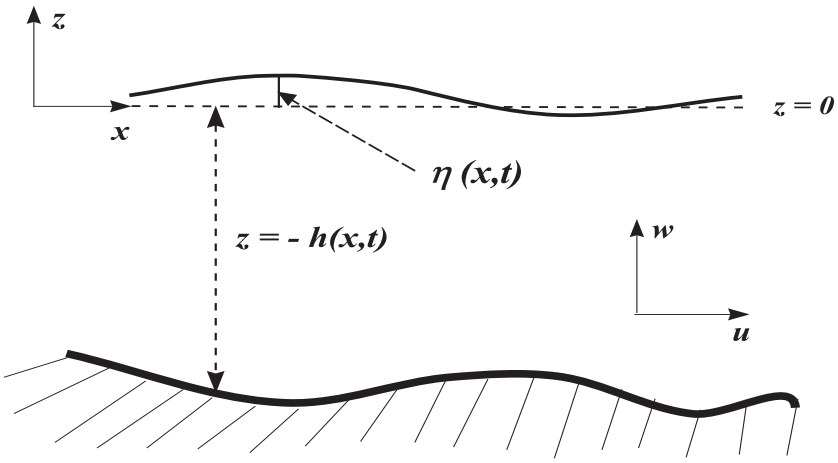

Figure 6.9. Sketch of the water waves above a nonflat bottom profile.

to essentials. We assume that there is no surface tension, although this could be included easily. Another simplifying assumption is that we will deal with straight wave fronts. Then, only one horizontal spatial coordinate x is involved, running from $-\infty$ to ∞. In the orthogonal horizontal direction the waves are assumed to be homogeneous. As for the boundary conditions, we assume all relevant quantities to vanish at infinity. The situation is sketched in Fig. **6.9**. The vertical direction (in the opposite direction of gravity) is taken as the z-axis. The still water level is at $z = 0$, and the bottom is at $z = -h(x, t)$, so we allow a nonflat bottom shape, and even time-dependent changes in the bottom. The latter are relevant, for example, in the study of the effects of tectonic plate motions or underwater landslides that may generate tsunamis.

We restrict ourselves to irrotational fluid. This is the main restriction, but it is quite common when modeling waves on a fluid surface. Specifically, if $\mathbf{v} = (u, w)$ denotes the Eulerian fluid velocity, irrotational motion has the property that

$$\text{rot } \mathbf{v} = \partial_z u - \partial_x w = 0.$$

It has the consequence that a fluid potential $\Phi(x, z, t)$ can be introduced such that

$$\mathbf{v} = \nabla\Phi = (\partial_x \Phi, \partial_z \Phi).$$

This potential is determined up to an additive constant (or function of time). The irrotationality assumption simplifies the analysis essentially. We also assume that the surface elevation can be described as the graph of a function given by $z = \eta(x, t)$, say, as sketched in Fig. **6.9**. This excludes the description of breaking, overtaking, waves.

We start the analysis with Luke's variational principle. This principle explicitly describes the fluid motion in the interior and gives the essential conditions at the free surface and the impermeable bottom. After having found and interpreted the resulting equations, we will "reduce" the description to quantities that are defined on the free surface only. In that

way we essentially reduce the spatial dimensions from two to one. After this reduction, a purely dynamic system for surface elevation and fluid potential at the surface results. Inherited from Luke's variational principle, these dynamic equations have a specific structure: it is a Hamiltonian system akin to systems in classical mechanics, now generalized to infinite dimensional systems described by partial differential equations.

As an advanced application of variational restriction, we will then simplify the equations in a consistent way by approximating the kinetic energy functional for two specific types of fluid motions. The two cases that we will consider are the linear theory of surface waves, which includes the full description of linear dispersion, and the shallow water equations, which include the full nonlinear terms of the problem. Dispersion and nonlinearity are the basic effects that are present in any model of gravity driven surface waves. As an important class of such models, we discuss Boussinesq-type equations, which take these effects into account in a certain approximation. Restricting our attention to waves running mainly in one direction, we encounter the Korteweg–de Vries (KdV) equation, and we will present its structure. When considering wave groups instead of individual waves, we show that the governing equation for the complex amplitude is the nonlinear-Schrödinger (NLS) equation, which also has a Hamiltonian structure.

Hence both the basic equations, as well as all the simplified models, have a Hamiltonian structure. We will show that the Hamiltonian structure of the equations may be retained when we discretize the spatial dependence of the functions by variational projection. This leads to numerical discretizations that are fully consistent and can be used for numerical simulations.

6.3.1 Luke's variational formulation

In the Challenging problem in §2.9 we derived the equations of motion for surface water waves from first principles under the restrictive condition that the layer is shallow. Here, we show that for irrotational flow the equations of motion can be obtained without any restriction from the variational approach named for Luke [21]. So, our starting point is the next result.

Luke's variational principle
Let $\Phi(x, z, t)$ be the fluid potential for two-dimensional irrotational, inviscid fluid flow, and let $\mathbf{v} = (u, w) = \nabla\Phi$ be the Eulerian velocity. Let $\eta(x, t)$ denote the surface elevation, and let the bottom shape be given by $z = -h(x, t)$. Then, the full equations for incompressible, irrotational fluid flow under a pressure-free atmosphere are given by the stationary points of the functional

$$P(\Phi, \eta) := \int \mathcal{P}(\Phi, \eta)\, dt,$$

where the functional $\mathcal{P}(\Phi, \eta)$ is given by

$$\mathcal{P}(\Phi, \eta) = \int \left[\int_{-h}^{\eta} \left\{ \partial_t \Phi + \frac{1}{2}|\nabla\Phi|^2 + g\, z \right\} dz \right] dx.$$

Here and in the following the x-integration is taken along the whole real axis; g denotes the gravitational constant of acceleration. □

To derive the equations of motion from the variational principle, we have to take the variational derivatives, introduced in §3.3.1 and §5.3.2, of the functional $\mathcal{P}(\Phi, \eta)$ with respect to the variables Φ and η, and to set the derivatives equal to zero. First, the vanishing of the first variation of the functional with respect to variations $\delta\Phi$ in Φ leads to

$$\int \left[\int \left[\int_{-h}^{\eta} \{\partial_t (\delta\Phi) + \nabla\Phi \cdot \nabla (\delta\Phi)\} \, dz \right] dx \right] dt = 0.$$

We write the first term in the form

$$\int_{-h}^{\eta} \partial_t (\delta\Phi) \, dz = \partial_t \left[\int_{-h}^{\eta} \delta\Phi \, dz \right] - (\delta\Phi)_{z=\eta} \, \partial_t \eta - (\delta\Phi)_{z=-h} \, \partial_t h.$$

For the second term we apply the Gauss theorem for partial integration and find

$$\int \int_{-h}^{\eta} \nabla\Phi \cdot \nabla (\delta\Phi) \, dz \, dx = \int \left[- \int_{-h}^{\eta} (\Delta\Phi) \, \delta\Phi \, dz + [(\partial_N \Phi) \, \delta\Phi]_{z=-h}^{z=\eta} \right] dx \, .$$

We conclude that vanishing for all variations $\delta\Phi$ leads to the Laplace equation in the interior fluid and to one bottom and one free surface boundary condition:

$$\begin{aligned} &\text{a.} && \Delta\Phi = 0, \quad h < z < \eta, \\ &\text{b.} && \partial_t h = \nabla\Phi \cdot N_b, \quad z = -h, \quad N_b = (-\partial_x h, -1), \\ &\text{c.} && \partial_t \eta = \nabla\Phi \cdot N_s, \quad z = \eta, \quad N_s = (-\partial_x \eta, 1) \, . \end{aligned} \tag{6.91}$$

The equation in the interior layer (6.91a) is the continuity equation. It expresses the incompressibility of the fluid: $\nabla \cdot \mathbf{v} = 0$, with $\mathbf{v} = \nabla\Phi$ being the Eulerian fluid velocity. Equation (6.91b) expresses the impermeability of the bottom (no flow through the bottom). Similarly, equation (6.91c) is the kinematic surface condition expressing that no water flows "through" the water surface.

Arbitrary variations with respect to η directly lead to

$$\partial_t \Phi + \frac{1}{2} |\nabla\Phi|^2 + g \, \eta \, (x, t) = 0 \tag{6.92}$$

at the water surface $z = \eta \, (x, t)$. This is the dynamic-free surface equation, or *Bernoulli's equation*, stating that the pressure at the surface of the water should vanish. This is the pressure condition when we assume that above the layer there is a pressure-free atmosphere (no wind or other forces acting on the free surface).

Exercise 6.3a.

Rederive Bernouilli's equation by starting from the Euler equations for a fluid with constant mass density $\rho = 1$:

$$\partial_t \mathbf{v} + (\mathbf{v} \cdot \nabla) \, \mathbf{v} = -\nabla p - g \, \mathbf{e}_z.$$

This momentum conservation equation has been derived in §2.7.3 *for inviscid flow. Here, the left-hand side is the total or material time derivative of the fluid velocity, p the pressure, and* $\mathbf{e}_z$ *the unit vector in the vertical direction. For irrotational flow, insert the relation* $\mathbf{v} = \nabla\Phi$ *and observe that the equations can be integrated, leading to Bernoulli's equation* (6.92) *given above, if the atmosphere is pressure-free.*

6.3.2 Reformulation and dimension reduction

The above problem is a formulation for the two-dimensional fluid layer in terms of the coordinates x and z and variables η and Φ. We can reduce it to a problem in one space variable x only. This reduction is possible since Φ is the solution of a Laplace problem on the whole domain, and this solution is fully determined by the prescribed values on the boundaries of the domain, and thus by one-dimensional information. We shall show that the reduction is an essential improvement; however, it is at the cost that a functional, the kinetic energy, has to be introduced, which is not easily explicitly expressed in the relevant variables.

To perform the reduction we take as basic variables the surface elevation $\eta(x,t)$ as before and the free surface potential

$$\phi(x,t) := \Phi(x, z = \eta(x,t), t).$$

Note that the latter is nothing more than the flow potential introduced earlier, but it is now restricted to the water surface. Both η and ϕ depend on x and t only, and not on z. For simplicity of exposition, we assume here that the bottom is time independent: $\partial_t h = 0$. Introducing the velocity type of quantity

$$u := \partial_x \phi$$

and the vertical velocity at the free surface

$$W(x,t) = w(x, \eta(x,t), t),$$

we can write the free surface equations (6.91c) and (6.92) as

$$\begin{aligned} \partial_t \eta &= -u\,\partial_x \eta + W(1+\eta_x^2), \\ \partial_t u &= -\partial_x \left[g\,\eta + \frac{1}{2}u^2 - \frac{1}{2}W^2(1+\eta_x^2) \right]. \end{aligned} \tag{6.93}$$

If we could express W explicitly in terms of η and ϕ (or in η and u), we would have a closed system. To that end, we introduce the kinetic energy $\mathcal{K}(\phi,\eta)$ as a functional of the basic quantities:

$$\mathcal{K}(\phi,\eta) = \min_{\Phi} \left[\int \left\{ \int_{-h}^{\eta} \frac{1}{2}|\nabla\Phi|^2\, dz \right\} dx \;\middle|\; \Phi = \phi \text{ at } z = \eta \right]. \tag{6.94}$$

The minimizing function Φ satisfies the linear problem for the Laplace equation with the correct boundary conditions:

$$\begin{aligned} \Delta\Phi &= 0, \quad -h < z < \eta, \\ \nabla\Phi \cdot N_b &= 0, \quad z = -h, \quad N_b = (-\partial_x h, -1), \\ \Phi &= \phi, \quad z = \eta. \end{aligned} \tag{6.95}$$

Since for a nonmoving bottom it holds that

$$\int_{-h}^{\eta} \partial_t \Phi \, dz = \partial_t \int_{-h}^{\eta} \Phi \, dz - \phi \, \partial_t \eta,$$

the functional $\mathcal{P}(\Phi, \eta)$ in Luke's variational principle can be rewritten as

$$\begin{aligned} \mathcal{P}(\Phi, \eta) &= \int \left[\int_{-h}^{\eta} \left\{ \partial_t \Phi + \frac{1}{2} |\nabla \Phi|^2 + g\, z \right\} dz \right] dx \\ &= - \int \phi \, \partial_t \eta \, dx + \mathcal{K}(\phi, \eta) + \int \frac{1}{2} g \left(\eta^2 - h^2 \right) dx \\ &\quad + \partial_t \int_{-h}^{\eta} \Phi \, dz. \end{aligned} \tag{6.96}$$

If we disregard an irrelevant term at the boundary of the time interval together with the term $\frac{1}{2} \int g\, h^2 \, dx$, since it is insensitive for variations with respect to η and ϕ, Luke's variational principle then comes down to finding the stationary points of

$$\int \mathcal{P}(\phi, \eta) \, dt = \int \left\{ \int \phi \, \partial_t \eta \, dx - \mathcal{H}(\phi, \eta, t) \right\} dt$$

with respect to ϕ and η. Here, we introduced the functional

$$\mathcal{H}(\phi, \eta) = \mathcal{K}(\phi, \eta) + \int \frac{1}{2} g\, \eta^2 \, dx \; .$$

Reflection on Hamiltonian structure

Recapitulating the results so far, we note that the variations in Φ in Luke's principle have consistently and equivalently been replaced with arbitrary variations in ϕ. The resulting variational principle

$$Crit_{\phi,\eta} \int \left\{ \int \phi \, \partial_t \eta \, dx - \mathcal{H}(\phi, \eta) \right\} dt \tag{6.97}$$

is known as a *canonical action principle*. It is a generalization of the similar principle in classical mechanics to infinite-dimensional systems. See Example 5.3e and textbooks on classical mechanics like [11]. The variables ϕ, η are called canonical variables, and $\mathcal{H}(\phi, \eta)$ is the *Hamiltonian*. Since no time-dependent boundary conditions are present, this Hamiltonian represents the total energy of the system, the sum of potential and kinetic energy. The governing equations, obtained by variations with respect to ϕ and η, are the *Hamiltonian equations*:

$$\begin{aligned} \partial_t \eta &= \delta_\phi \mathcal{H}(\phi, \eta), \\ \partial_t \phi &= -\delta_\eta \mathcal{H}(\phi, \eta). \end{aligned} \tag{6.98}$$

These equations involving variational derivatives directly generalize the well-known Hamiltonian equations for finite-dimensional (discrete) systems in terms of partial derivatives. The equations can be rewritten in the same way as in (5.22), that is, as

$$\partial_t \begin{pmatrix} \eta \\ \phi \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \delta_\eta \mathcal{H}(\phi, \eta) \\ \delta_\phi \mathcal{H}(\phi, \eta) \end{pmatrix}. \tag{6.99}$$

Any dynamical system with state variable U (possibly a vector function) is called a *Hamiltonian system* if it evolves according to an equation of the form

$$\partial_t U = \Gamma \, \delta_U \mathcal{H}(U,t), \tag{6.100}$$

where Γ is a skew-symmetric operator (the so-called structure map).[1] One of the most significant properties of a Hamiltonian system is that the Hamiltonian (the energy) is conserved during the evolution if the Hamiltonian is autonomous, i.e., does not explicitly depend on time. This easily follows from the skew-symmetry of the operator Γ. Using the notation $(.\,,\,.)$ for the inner product as defined in §5.3.2 for the variational derivative, we may write

$$\begin{aligned}\frac{d}{dt}\mathcal{H}(U,t) &= (\delta_U \mathcal{H}(U,t)\ ,\ \partial_t U) + \frac{\partial}{\partial t}\mathcal{H}(U,t)\\ &= (\delta_U \mathcal{H}(U,t)\ ,\ \Gamma\,\delta_U \mathcal{H}(U,t)) + \frac{\partial}{\partial t}\mathcal{H}(U,t)\\ &= \frac{\partial}{\partial t}\mathcal{H}(U,t),\end{aligned} \tag{6.101}$$

since $(\delta_U \mathcal{H}(U,t),\ \Gamma\,\delta_U \mathcal{H}(U,t)) = 0$ in view of the skew-symmetry of Γ. Hence,

$$\frac{d}{dt}\mathcal{H}(U,t) = 0 \quad \text{if} \quad \frac{\partial}{\partial t}\mathcal{H}(U,t) = 0. \tag{6.102}$$

For the surface wave problem under consideration, this property implies that if we take the bottom profile constant in time, i.e., $h(x,t) = h(x)$, the Hamiltonian is autonomous and the total energy is conserved during the evolution. When $\partial_t h \neq 0$, energy will be added or subtracted from the fluid.

Recapitulating the result about the modeling of surface waves so far, we observe that, at least formally, the problem is now well formulated as a problem in basic variables that depend only on x and t but not on z. The price to be paid is that we deal with the rather difficult functional $\mathcal{K}$. We will show the advantage of this reduction in the next subsections.

It may be observed that the functional $\mathcal{K}(\phi,\eta)$, and hence the total Hamiltonian, is invariant for adding a constant to ϕ. This is to be expected, since not the potential Φ but the velocity given by its spatial derivatives are the physical variables. This motivates us to consider u and η as the basic pair of variables, instead of ϕ and η, with $u(x,t) := \partial_x \phi$. Writing the Hamiltonian then as $\mathcal{H}(u,\eta)$, we have the relation

$$\delta_\phi \mathcal{H} = -\partial_x \, \delta_u \mathcal{H},$$

where the minus sign stems from partial integration. The equations of motion become

$$\partial_t \begin{pmatrix} \eta \\ u \end{pmatrix} = \begin{pmatrix} 0 & -\partial_x \\ -\partial_x & 0 \end{pmatrix} \begin{pmatrix} \delta_\eta \mathcal{H}(u,\eta) \\ \delta_u \mathcal{H}(u,\eta) \end{pmatrix}. \tag{6.103}$$

[1] More precisely, it is called a Hamiltonian system only if Γ is also invertible; if not, it is called a Poisson system. It is even possible to allow that Γ itself depends on the state variable, $\Gamma = \Gamma(U)$; then much of the basic properties still hold provided that Γ satisfies the so-called Jacobi conditions; see, e.g., [13].

Note that this equation has again a Hamiltonian structure. The equations are in this way in conservative form, i.e., the time derivatives are given by spatial derivatives of corresponding fluxes:

$$\begin{aligned} \partial_t \eta &= -\partial_x \left[\delta_u \mathcal{H}(u,\eta)\right] = -\partial_x \left[\delta_u \mathcal{K}(u,\eta)\right], \\ \partial_t u &= -\partial_x \left[\delta_\eta \mathcal{H}(u,\eta)\right] = -\partial_x \left[\delta_\eta \mathcal{K}(u,\eta) + g\,\eta\right]. \end{aligned} \tag{6.104}$$

Exercise 6.3b.

We have derived the Hamilton equations from manipulations in the variational formulation, which proves their validity in a direct way. Yet it is useful to write these equations in full, which should coincide with the original equations. To do this, recall the Hamiltonian equations in potential form:

$$\begin{aligned} \partial_t \eta &= \delta_\phi \mathcal{H}(\phi,\eta) = \delta_\phi \mathcal{K}(\phi,\eta), \\ \partial_t \phi &= -\left[\delta_\eta \mathcal{K}(\phi,\eta) + g\,\eta\right]. \end{aligned} \tag{6.105}$$

The variational derivative $\delta_\phi \mathcal{K}(\phi,\eta)$ is found from the relation for variations $\delta\phi$ and corresponding variations $\delta\Phi$ of the potential Φ in the interior,

$$\left(\delta_\phi \mathcal{K}(\phi,\eta)\ ,\ \delta\phi\right) = \int \left[\int \left[\nabla\Phi\cdot\nabla\delta\Phi\right] dz\right] dx.$$

Here, Φ is the solution of (6.95).

a. *Show that, using the properties of Φ, this reduces to*

$$\left(\delta_\phi \mathcal{K}(\phi,\eta)\ ,\ \delta\phi\right) = \int \left(\nabla\Phi\cdot N_s\right)\,\delta\phi\, dx$$

with N_s the nonnormalized normal on the fluid surface. Hence,

$$\delta_\phi \mathcal{K}(\phi,\eta) = \nabla\Phi\cdot N_s,$$

which reproduces the kinematic boundary condition.

b. *Finding the dynamic equation is a bit more complicated, since the restriction to the surface of Φ has consequences for its derivatives. For instance,*

$$\partial_t \phi = \frac{d}{dt}\Phi(x,\eta(x,t),t) = \left[\partial_t \Phi\right]_{z=\eta} + \left[\partial_z \Phi\right]_{z=\eta} \partial_t \eta.$$

When calculating $\delta_\eta \mathcal{K}(\phi,\eta)$, we have to realize that we then vary the boundary of the integration domain and that at this boundary the value of Φ has to remain equal to ϕ. Hence, we get a correction when differentiating with respect to η. The change in Φ at the surface resulting from a change $\delta\eta$ in η is given by $\delta\phi = \delta\eta\left[\partial_z\Phi\right]_{z=\eta}$. Show therefore that

$$\delta_\eta \mathcal{K}(\phi,\eta) = \left(\frac{1}{2}|\nabla\Phi|^2\right)_s - \delta_\phi \mathcal{K}\left[\partial_z\Phi\right]_{z=\eta}.$$

c. *Show that from this it indeed follows that* $\partial_t \phi = -\left[\delta_\eta \mathcal{K}(\phi, \eta) + g\,\eta\right]$ *leads to the correct dynamic surface condition.*

Exercise 6.3c.

In dynamical systems we are always interested in symmetries, which often correspond to conservation properties. This is also the case in the present surface wave problem. In particular, suppose that the bottom is flat: $h(x) = \bar{h}$. *Then, it is clear that the problem exhibits a translation symmetry: the physics does not change under a translation in the horizontal direction. The corresponding functional that is a "constant of motion" (invariant integral) is the horizontal momentum functional* I *given by*

$$I(u, \eta) := \int u\,\eta\,dx.$$

a. *Show in a direct way by using the Hamiltonian equations that*

$$\frac{d}{dt} I(u, \eta) = 0.$$

b. *To motivate that this functional is called a "momentum" functional, consider the Hamiltonian system with as Hamiltonian the functional* I *itself. Show then that the solutions can be easily found and are nothing but a translation in space at constant speed. Show that this is consistent with the translation symmetry described above, by observing that the original Hamiltonian is independent of a translation of the variables* (η, ϕ). *Argue that the flows of the* H *and* I *dynamics commute: starting with a translation over a distance* α *followed by dynamic evolution during time* T *is the same as starting with the dynamic evolution during time* T *after which a translation* α *is performed.*

6.3.3 Special cases of wave models by variational restriction

Above we obtained a reduction from two to one spatial dimension by expressing the interior fluid motion via the introduction of the kinetic energy in terms of the potential at the free surface. However, in this formulation the kinetic energy is crucial and it makes the full surface wave problem difficult. We will now consider various reduced models by simplifying the kinetic energy. All models will keep the Hamiltonian structure. We restrict ourselves to specific types of wave fields (such as small amplitude waves, or waves above shallow layers) for which the kinetic energy gets a simpler form.

Small amplitude waves: Linear dispersion

The kinetic energy in (6.94) can be written as

$$\mathcal{K}(\phi, \eta) = \frac{1}{2} \int \phi\,[\nabla\Phi \cdot N_s]_{z=\eta}\,dx$$

and contains the derivative normal to the free surface. If we were able to express this normal derivative $\nabla\Phi \cdot N_s$ in terms of (ϕ, η), we would have at hand an explicit functional of the kinetic energy. This is possible for small-amplitude waves, for which the surface elevation

and also the bottom variations are small. This will be worked out in detail, but first we give the basic argument.

For small-amplitude waves and small bottom variations, we replace the free surface and the bottom with a flat surface, i.e., we replace the fluid domain with a horizontal strip, $-\bar{h} < z < 0$, where $\bar{h}$ is a constant, effective depth. Then the normal derivative of the potential, the vertical fluid velocity component at the flat surface, depends linearly on the surface potential, say, through some linear operator D:

$$\left[\partial_z \Phi\right]_{z=0} = D\,\phi.$$

This operator is called the *Dirichlet-to-Neumann operator*: it assigns to the Dirichlet value ϕ the normal derivative of the solution of the Laplace problem in the strip. Hence, it incorporates (in a nontrivial way) effects of the water motion in the interior of the layer. The kinetic energy then becomes the quadratic form

$$\mathcal{K}(\phi, \eta) = \frac{1}{2}\int \phi\,\left[\partial_z \Phi\right]_{z=\eta}\,dx = \frac{1}{2}\int \phi\,D\,\phi\,dx. \tag{6.106}$$

We will now determine the operator D by solving the Laplace problem in the strip:

$$\begin{aligned} \Delta\Phi &= 0 \text{ for } -\bar{h} < z < 0, \\ \partial_z \Phi &= 0 \text{ at } z = -\bar{h}, \\ \Phi &= \phi \text{ at } z = 0. \end{aligned} \tag{6.107}$$

To proceed, we apply Fourier transformation in x, denoting by $\hat{\phi}$ the Fourier transform of ϕ. As to be derived in Exercise 6.3d, the solution is given by (suppressing the time dependence for the moment)

$$\Phi(x, z) = \int \hat{\phi}(k)\,\frac{\cosh\left(k\left(z+\bar{h}\right)\right)}{\cosh\left(k\bar{h}\right)} e^{ikx}\,dk.$$

Hence, from Parseval's identity we find that

$$\begin{aligned} \mathcal{K}(\phi, \eta) &= \frac{1}{2}\int \phi\left[\partial_z \Phi\right]_{z=0}\,dx \\ &= \pi \int \hat{\phi}\left[\partial_z \hat{\Phi}\right]^*_{z=0}\,dk \\ &= \pi \int \hat{\phi}(k)\,k\tanh\left(k\bar{h}\right)\,\hat{\phi}^*(k)\,dk\,. \end{aligned} \tag{6.108}$$

From this we see that D is a pseudodifferential operator, which means that after Fourier transformation the operator becomes a multiplication with a function $\hat{D}(k)$:

$$\widehat{D\phi}(k) = \hat{D}(k)\,\hat{\phi}(k)\,.$$

From the above expression (and Exercise 6.3d), $\hat{D}$ is explicitly given by

$$\hat{D}(k) = k\tanh\left(k\bar{h}\right).$$

The operator D is symmetric and positive, corresponding to the fact that $\hat{D}$ is real and nonnegative. The governing linear Hamiltonian equations are

$$\partial_t \eta = D\phi, \quad \partial_t \phi = -g\,\eta.$$

These can be written as a second-order equation for any of the variables. For instance,

$$\partial_t^2 \eta = -g\,D\,\eta.$$

As dealt with in §3.3.4, this equation has $\omega = \Omega(k)$ as dispersion relation with

$$\omega^2 = g\hat{D}(k) =: \Omega^2(k),$$

where Ω is explicitly given by

$$\Omega(k) = k\sqrt{\frac{g\tanh\left(k\bar{h}\right)}{k}}. \tag{6.109}$$

The second-order equation can then be written as

$$\partial_t^2 \eta + \Omega^2(-i\partial_x)\,\eta = 0\,.$$

The fact that Ω is a nonlinear function of k means that surface waves have dispersion: the monochromatic mode $\exp i\,(kx - \Omega(k)\,t)$ travels with a phase velocity—see §3.3.4—given by

$$c_{ph}(k) = \frac{\Omega(k)}{k} = \sqrt{\frac{g\tanh\left(k\bar{h}\right)}{k}}, \tag{6.110}$$

which thus depends on k. Since Ω is a concave function of k, waves of shorter (longer) wavelengths travel at slower (faster) speeds so that waves of different wavelengths starting at the same position become separated, i.e., they disperse. Note that they do not dissipate, since they do not vanish after some time. We remark that infinitely long waves have the largest phase speed, given by taking the limit for $k \to 0$:

$$c_\infty = \sqrt{g\bar{h}}.$$

If we denote by C_{ph} the pseudodifferential operator corresponding to the phase speed c_{ph}, the kinetic energy can in case of linear dispersion be rewritten as

$$\mathcal{K}_{lin}(\phi) = \frac{1}{2}\int \phi\,D\,\phi\,dx = \frac{1}{2}\int u\,\left(C_{ph}^2/g\right)\,u\,dx =: \mathcal{K}_{lin}(u),$$

or, in terms of Fourier variables,

$$\mathcal{K}_{lin} = \pi\int\left[\frac{\tanh\left(k\bar{h}\right)}{k}\,|\hat{u}|^2\right]dk.$$

Exercise 6.3d.

a. *Find the solution of the Laplace problem* $\Delta\Phi = 0$ *on the strip* $-\bar{h} < z < 0$ *by determining the coefficients* A, B *in the general solution at wavenumber* k

$$\Phi(x,z) = e^{ikx}\left[Ae^{k(z+\bar{h})} + Be^{-k(z+\bar{h})}\right]$$

from the surface condition $\Phi = \phi$ *at* $z = 0$ *and the bottom condition* $\partial_z\Phi = 0$ *at* $z = -\bar{h}$. *Verify that it is given by*

$$\Phi = \hat{\phi}(k)\,\frac{\cosh\left(k\left(z+\bar{h}\right)\right)}{\cosh\left(k\bar{h}\right)}e^{ikx},$$

where $\hat{\phi}(k)$ *is the Fourier transform of* ϕ.

b. *Show that the Fourier transform of the Dirichlet-to-Neumann operator* D *is given by*

$$\hat{D}(k) = k\tanh\left(k\bar{h}\right)$$

and derive from this the results given above.

Example 6.3a. Long wave approximations.
A Taylor expansion of the dispersion relation (6.109) $\Omega(k)$ near $k = 0$ leads to

$$\frac{\Omega^2\bar{h}}{g} = (k\bar{h})^2 - \frac{1}{3}(k\bar{h})^4 + \frac{2}{15}(k\bar{h})^6 - \frac{17}{315}(k\bar{h})^8 + \mathcal{O}\left((k\bar{h})^{10}\right).$$

In lowest order of $\left(k\bar{h}\right)$ we thus obtain

$$\Omega^2 = g\bar{h}k^2 = c_\infty^2 k^2$$

with $c_\infty = \sqrt{g\bar{h}}$ the phase speed of infinitely long waves. Observe that the next higher order expansion has no definite sign: for sufficiently small wavelengths, and thus large k-values, negative signs appear. This leads to unstable solutions, since the resulting frequencies are complex valued for sufficiently small wavelengths so that the corresponding waves grow or damp out. This flaw could be repaired in an artificial way, for instance, as follows:

$$\begin{aligned}\frac{\Omega^2\bar{h}}{g} &\approx (k\bar{h})^2 - \frac{1}{3}(k\bar{h})^4 = \bar{h}^2k^2\left(1 - \frac{\bar{h}^2k^2}{3}\right) \\ &\approx \bar{h}^2k^2\left(1 + \frac{\bar{h}^2k^2}{3}\right)^{-1},\end{aligned} \tag{6.111}$$

where the last approximation is positive and thus leads to stable behavior for all k. □

Shallow water equations
We will now derive an approximation for the kinetic energy of the so-called shallow water approximation. See also §2.9. This means that the wavelengths of the waves under

consideration are much larger than the depth of the layer. This specifies the meaning of "shallow," relating the depth to the type of waves under consideration: $k\,\bar{h} \ll 1$. In the present approximation we allow the bottom to vary in space: $h = h(x)$. Now, $\bar{h}$ is the average water depth.

This assumption leads to the idea of approximating the fluid potential at each depth by its value at the free surface: $\Phi(x, z) \approx \phi(x)$ independent of z. Taking this approximation in the kinetic energy integral (6.106) leads to the shallow water approximation

$$\mathcal{K}_{sw} = \frac{1}{2}\int (\eta + h)\, u^2\, dx - \int \phi\, \partial_t\, h\, dx.$$

The term $\eta\, u^2$ in the integrand is the only term that leads to nonlinearity in the Hamiltonian equations. In potential form they read as

$$\begin{aligned} \partial_t \phi &= -g\,\eta - \frac{1}{2}u^2, \\ \partial_t \eta &= -\partial_x\left[(h+\eta)\,\partial_x \phi\right] - \partial_t h. \end{aligned} \tag{6.112}$$

The last equation can be rewritten like

$$\partial_t\,(h+\eta) = -\partial_x\left[(h+\eta)\,u\right]\,,$$

which is clearly the continuity equation expressing mass conservation. The first equation is Bernoulli's equation at the surface with approximation for the squared velocity. Using $u = \partial_x \phi$ we can rewrite both equations, leading to the *shallow water equation* (SWE) in conservative form

$$\begin{aligned} \partial_t u &= -\partial_x\left[g\eta + \frac{1}{2}u^2\right], \\ \partial_t\,(h+\eta) &= -\partial_x\left[(h+\eta)\,u\right]\,. \end{aligned} \tag{6.113}$$

These equations are of a special mathematical structure. The two equations form a set of *first-order hyperbolic equations*. Equations with such a structure can be analyzed with the method of characteristics. We refer to textbooks such as [23, 33] on PDEs for further details.

Boussinesq-type equations

The linear theory treated above deals with small amplitude waves and incorporates interior fluid motion, leading to the dispersion property of waves. The SWEs, on the other hand, neglect variations in horizontal velocities in the interior layer, but take the nonlinear effects of finite bottom variations and surface elevations fully into account. Boussinesq-type models are obtained if both dispersive and nonlinear effects are approximated in a simplified and balanced way. In practice this means that the dispersion operator is approximated and the nonlinearity is taken as in the SWE. For instance, a nondefinite (unstable) Boussinesq equation is obtained by approximating the dispersion in (6.110) so that

$$c^2_{ph,approx} \approx g\,\bar{h}\left(1 - \frac{1}{3}\left(k\bar{h}\right)^2\right).$$

A positive definite approximation is obtained by

$$c^2_{ph,approx} \approx g\,\bar{h}/\left(1+\frac{1}{3}\left(k\bar{h}\right)^2\right).$$

There are many variants of Boussinesq-type equations, each with its own merits and approximative quality. In deriving such approximations, the Hamiltonian character of the basic equations is not always retained, and conservation properties like energy and momentum conservation (when applicable) have then to be investigated in an ad hoc way. Deriving approximate equations by approximating the kinetic energy functional in the Hamiltonian formulation, as above and as in [13], guarantees from the start that the Hamiltonian structure will be retained in the approximate equations, with successive consequences for the conservation properties. All these equations will have a Hamiltonian structure of the form (6.103). As a specific example of a Boussinesq equation derived along the lines as sketched here, the reader may consult, for instance, [16]. This last model includes bottom motions; the resulting numerical code, implemented with a pseudospectral method, is able to simulate generation of tsunamis from tectonic plate motions or underwater landslides.

KdV-type equations

The equations above describe waves in one spatial dimension that can run in both directions. Because of the nonlinearity, right-and left-traveling waves are coupled. Yet it is possible to consider waves that are mainly running in one direction, which occur most often in real-life situations. This can be viewed as another restriction, and from Boussinesq-type equations such unidirectional models can be derived. We will not go into details of the intricate restriction method but will only point out that then the two dependent variables u and η become coupled, and that one equation for only η can be obtained that is again of Hamiltonian form, namely,

$$\partial_t \eta = -\partial_x \delta_\eta \mathcal{H}_1(\eta), \tag{6.114}$$

where the Hamiltonian $\mathcal{H}_1$ is an approximation of the total energy, and (for $h=1$) given by

$$\mathcal{H}_1(\eta) = \int \left[\frac{1}{2}c_\infty \eta^2 - \frac{1}{12}\left(\partial_x \eta\right)^2 + \frac{1}{3}\eta^3\right] dx.$$

This functional leads to the KdV equation, considered earlier in §1.5.3. Indeed, for this functional, equation (6.114) reads as

$$\partial_t \eta + \partial_x \left[c_\infty \eta + \frac{1}{6}\partial_x^2 \eta + \eta^2\right] = 0.$$

Wave groups

It should be noted that the equations above describe individual waves. We will now consider so-called wave groups, a collection of waves with almost the same wavelength. Here, we will make use of the theory in §3.3.4. The collective behavior of a group can be described in the most natural way in terms of a complex amplitude multiplying the *carrier wave*, which is the monochromatic harmonic wave with averaged wavelength. We will briefly present here the main idea of wave groups, mainly emphasizing the Hamiltonian structure of the governing equation for the amplitude.

We start with a linear dispersive wave equation for unidirectional waves with dispersion relation $\omega = \Omega(k)$. The equation is of the form

$$\partial_t \eta + i\Omega(-i\partial_x)\eta = 0,$$

where we substitute the operator $-i\,\partial_x$ in the dispersion relation $\Omega(k)$. It has normal mode solutions of the form $\eta = \exp(i[kx - \Omega(k)t])$. As shown in §3.3.4, the linearity implies that the solution for the initial value problem can be written in terms of a Fourier integral:

$$\eta(x,t) = \int \widehat{\eta_0}(k)\, e^{(i[kx-\Omega(k)t])}\, dk,$$

where $\widehat{\eta_0}(k)$ is the Fourier transform of the initial wave profile at $t = 0$.

Now suppose that the initial profile has a spectrum that is confined to a small neighborhood of some wave number k_0, say. So, the spectrum is strongly peaked around k_0. Then, with $k = k_0 + \kappa$ and $\omega_0 = \Omega(k_0)$, the solution can be written as

$$\eta(x,t) = e^{i[k_0 x - \omega_0 t]} \int \alpha\ (\kappa)\, e^{(i[\kappa(x-V_0 t)-\Omega_{res}(\kappa)t])}\, d\kappa + cc,$$

where cc denotes the complex conjugate of the preceding term. Here $\alpha(\kappa) = \widehat{\eta_0}(k_0+\kappa)$ has small support and the integral over κ is over this support. Furthermore, V_0 is the so-called group velocity at k_0 defined as $V_0 = d\,\Omega/dk|_{k=k_0}$. The residual frequency Ω_{res} is defined as

$$\Omega(k_0+\kappa) = \Omega(k_0) + V_0\,\kappa + \Omega_{res}(\kappa).$$

In a good approximation Ω_{res} is given by the quadratic expression

$$\Omega_{res}(\kappa) \approx -\beta\kappa^2 \qquad \text{with} \qquad \beta = -\frac{1}{2}\frac{d^2\Omega}{dk^2}|_{k=k_0}. \tag{6.115}$$

Here, the minus sign is chosen to get β positive for the linear surface wave problem. The solution can now be written as

$$\eta(x,t) = A\, e^{i[k_0 x - \omega_0 t]} + cc \tag{6.116}$$

with the complex valued amplitude A given by

$$A(\xi,\tau) = \int \alpha(\kappa)\, e^{(i[\kappa\xi - \Omega_{res}(\kappa)\tau])}\, d\kappa.$$

Here, a coordinate system moving with the group velocity has been introduced: $\xi := x - V_0 t$, $\tau := t$. From the form of this expression it follows that A is the solution of the initial value problem of another dispersive wave equation, namely,

$$\partial_\tau A + i\,\Omega_{res}\left(-i\,\partial_\xi\right)A = 0,$$

with the initial value $A(\xi,0)$ being equal to the Fourier transform of $\alpha(\kappa)$. With the quadratic approximation (6.115) of the residual spectrum, the equation reads as

$$\partial_\tau A + i\,\beta\,\partial_\xi^2 A = 0\,. \tag{6.117}$$

Description (6.116) represents the surface elevation as a carrier wave, the monochromatic mode (or plane wave) $e^{i\,[k_0 x-\omega_0 t]}$, multiplied by the complex amplitude A. Since α has small support, A will have large support, and η resembles a wave group, also called "wave package," a collection of waves with slowly changing amplitude (compared to the carrier wave length). An equation like (6.117) is called an *amplitude*, or *envelope equation*. A specific example was considered in §3.3.5, the bichromatic wave group.

Exercise 6.3e.

The so-called Gaussian spectrum is given by

$$\alpha\,(\kappa) = \frac{1}{\sqrt{\pi}} \exp\left(-\frac{\kappa^2}{2\sigma^2}\right).$$

Now, an explicit expression for A can be obtained if the residual dispersion relation is approximated by the quadratic expression (6.115).

a. *Verify that this solution is given by*

$$A = \frac{1}{\sqrt{2}\pi} \frac{\exp\left[-\frac{1}{2}\xi^2\sigma^2/\left(1-i\,\beta\,\sigma^2\tau\right)\right]}{\sqrt{1-i\,\beta\,\sigma^2\tau}}.$$

Note that the narrower the initial amplitude spectrum is, i.e., σ small, the broader the amplitude will be, and that, while broadening, the amplitude will decay slowly with $1/(\sigma\sqrt{\tau})$ for large times.

b. *Plot the evolution of the initial elevation at various times.*

c. *Show that this solution is also obtained from the linear heat equation*

$$\partial_\tau A - \partial_\xi^2 A = 0$$

via the transformation $\tau \to -i\,\tau$. Please realize the difference: the heat equation is clearly dissipative, whereas the envelope equation is conservative.

Until now we have neglected nonlinear effects. Nonlinear terms will generate new modes. It can be shown that, if such effects are taken into account, in lowest order the resulting equation will be of the form

$$\partial_\tau A + i\,\beta\,\partial_\xi^2 A + i\,\gamma\,|A|^2\,A = 0 \tag{6.118}$$

with real coefficient γ. This is the famous so-called non-linear Schrödinger (NLS) equation, which is the characteristic envelope equation to describe the evolution of wave groups in dispersive wave equations with second- or third-order nonlinearity. Also, this equation has a Hamiltonian structure, which is now complex. Indeed, the equation can be written as

$$\partial_\tau A = i\,\delta\mathcal{H}\,(A)\,,$$

where the complex unit i constitutes the structure map and the Hamiltonian is given by

$$\mathcal{H}(A) = \int \left[\frac{1}{2}\beta \left|\partial_\xi A\right|^2 - \frac{1}{4}\gamma |A|^4 \right] d\xi.$$

Exercise 6.3f.

Verify, using the inner product for complex quantities defined in §3.3.2*, that the complex unit is skew symmetric (and therefore a structure map) and that the variational derivative of $\mathcal{H}$ is indeed given by*

$$\delta\mathcal{H}(A) = -\beta\, \partial_\xi^2 A - \gamma |A|^2 A.$$

In conclusion, we can say that not only the wave equations that describe the surface waves but also the envelope equation (6.118) have a Hamiltonian structure. In the next section we will retain this structure when designing spatial discretizations for numerical calculations.

6.3.4 Spatial discretization of Hamiltonian systems

We recall the general formulation of a Hamiltonian system introduced above: a dynamical system with state u (possibly a vector function) that evolves according to an equation of the form

$$\partial_t u = \Gamma\, \delta_u \mathcal{H}(u) \tag{6.119}$$

with the structure map Γ being a skew-symmetric operator. As is the case for the surface wave problem, we assume in the following Γ to be independent of u and the system to be autonomous, i.e., the Hamiltonian $\mathcal{H}$ does not explicitly depend on time.

In the setting of surface waves above, the state variable u is a vector function with two components depending on the spatial variable x and the time t. In most cases the state has more than one component. However, for notational ease we will take u to have one component. Extension to the more component case is rather straightforward.

Basis functions

We will now derive a finite-dimensional discretization of a Hamiltonian system in a consistent way. We use a set of basis functions $\{\phi_k(x)\},\ k = 1, 2, \ldots$, in the function space, and approximate $u(x,t)$ by a finite truncation:

$$u(x,t) \approx \Sigma_{k=1}^{N} \tilde{u}_k(t)\, \phi_k(x). \tag{6.120}$$

In the following we denote the vector containing the coefficients by $\tilde{\mathbf{u}}(t) := (\tilde{u}_1(t), \ldots, \tilde{u}_N(t))$. So, instead of dealing with functions in an infinite-dimensional space, we now work with vectors in a finite-dimensional space.

Note that it is not necessary that the set of basis functions is complete, although this may be an advantage. The only requirement is that the basis functions to be used lead to an approximation of the state with reasonable accuracy. Completeness is an issue of importance if one wants to study convergence properties of the method taking the limit $N \to \infty$. In practice, the basis functions must be chosen such that for a relatively small value of N the approximation is accurate enough, since then the numerical effort is kept at a tractable level.

To get fast convergence of the expansions, one should choose basis functions that fit to the expected shape of the functions to be represented.

We will show that the equation of motion of the resulting finite-dimensional system for the restricted state vector $\tilde{\mathbf{u}}(t)$ is again of the form of a Hamiltonian system, and it reads as

$$\partial_t \tilde{\mathbf{u}} = \tilde{\Gamma} \nabla_{\tilde{\mathbf{u}}} \tilde{\mathcal{H}}(\tilde{\mathbf{u}}). \tag{6.121}$$

Note that the variational derivative in (6.119) is now replaced by the ∇ operator. $\tilde{\mathcal{H}}(\tilde{\mathbf{u}})$ is the restriction of the Hamiltonian:

$$\tilde{\mathcal{H}}(\tilde{\mathbf{u}}) := \mathcal{H}\left(\Sigma_{k=1}^{N} \tilde{u}_k(t)\, \phi_k(x)\right), \tag{6.122}$$

and $\tilde{\Gamma}$ is a skew-symmetric matrix that approximates the effect of the structure map Γ.

Result (6.121) is the most compelling reason to develop numerical algorithms in a variationally consistent way. Observe, for instance, that energy conservation for the original (autonomous) system directly translates to conservation of the (restricted) energy in the finite-dimensional system. To derive (6.121), substitute (6.120) in the left-hand side of the original equation (6.119) and take the inner product with any basis function ϕ_m. The result is the mth component of the vector

$$M\, \partial_t \tilde{\mathbf{u}}, \tag{6.123}$$

where M is the square, symmetric matrix

$$M_{k,m} = (\phi_k, \phi_m).$$

To discretize the right-hand side of (6.119), we first expand as

$$\delta_u \mathcal{H}(u) \approx \Sigma_{j=1}^{N} f_j \phi_j$$

for some coefficients f_j. Then, we note that

$$(\Sigma f_j \phi_j, \phi_k) \approx (\delta_u \mathcal{H}(u), \phi_k) = \partial_{\tilde{u}_k} \mathcal{H}\left(\Sigma\, \tilde{u}_j\, \phi_j\right) = \partial_{\tilde{\mathbf{u}}_k} \tilde{\mathcal{H}}(\tilde{\mathbf{u}}),$$

which in matrix notation reads as (using that M is a symmetric matrix)

$$M\, \mathbf{f} = \nabla_{\tilde{u}} \tilde{\mathcal{H}}(\tilde{\mathbf{u}}).$$

Then the right-hand side becomes $(\Gamma \delta_u \mathcal{H}(u), \phi_m) = \Sigma_k f_k\, (\Gamma \phi_k, \phi_m)$. Introducing the matrix $\tilde{\gamma}$

$$\tilde{\gamma}_{km} = (\Gamma\, \phi_k, \phi_m),$$

we find $(\Gamma\, \delta_u \mathcal{H}(u), \phi_m) = (\tilde{\gamma} \mathbf{f})_m$ and with $\mathbf{f} = M^{-1} \nabla_{\tilde{\mathbf{u}}} \tilde{\mathcal{H}}(\tilde{\mathbf{u}})$ this leads to

$$(\Gamma\, \delta_u \mathcal{H}(u), \phi_m) = \left(\tilde{\gamma} M^{-1} \nabla_{\tilde{\mathbf{u}}} \tilde{\mathcal{H}}(\tilde{\mathbf{u}})\right)_m.$$

Equating, in matrix notation, left- and right-hand-side results to $M\,\partial_t \tilde{\mathbf{u}} = \tilde{\gamma}\, M^{-1} \nabla_{\tilde{\mathbf{u}}} \tilde{\mathcal{H}}(\tilde{\mathbf{u}})$, and defining

$$\tilde{\Gamma} = M^{-1} \tilde{\gamma} M^{-1},$$

we obtain (6.121).

Note that $\tilde{\Gamma}$ is skew-symmetric from the fact that $\tilde{\gamma}$ is skew-symmetric.

Implementation aspects for the surface wave problem

The result above is in principle directly applicable to the surface water problem. We will not go into the details of the full implementation; instead we discuss some aspects related to the choice of basis functions and their consequences. As seen above, we only have to investigate the discretization of the structure map Γ and the restriction of the Hamiltonian. Referring to Chapter 5, we will discuss the spectral and the finite element method. For this method we will discuss only the use of piecewise linear basis functions, the so-called hat functions.

Taking an orthonormal basis of global spectral functions, the matrix M is diagonal. It is even the identity matrix if the functions are normalized. For the local basis functions in the finite element method, this matrix becomes three-diagonal, provided that the support of the basis functions is restricted to two adjacent elements. The structure operator is essentially the first derivative. Then, the same reasoning as for M shows that $\tilde{\gamma}$ is tridiagonal, with zeros on the main diagonal, since it is skew-symmetric. For a spectral discretization this leads to a tridiagonal dimensional skew-symmetric matrix $\tilde{\Gamma}$, while for the finite element method this becomes a seven-diagonal skew symmetric matrix.

For the restriction of the Hamiltonian, the dispersive and nonlinear terms deserve separate attention. The effect of the linear dispersion shows up in quadratic terms in the Hamiltonian, which can easily be transformed in the spectral discretization using Parseval's identity. This is even the case when working with pseudodifferential operators. For finite element method discretization, the pseudodifferential operator has to be truncated in some way to a differential operator, which then leads to quadratic forms with diagonal matrices. The nonlinearity of the problem results from cubic expressions in the Hamiltonian. Now the spectral discretization will lead to full matrices in the trilinear expressions, whereas the finite element method leads to matrices with band-diagonal structures.

6.4 Optics

The following case study forms an illustration of the concepts dealt with in
§3.2.3 (WKB approximation),
§3.3.2 (expansions in basis functions),
§3.3.4 (plane waves), and
Chapter 5 (variational modeling).

In this section we will deal with optics as an area of application. Optics is a part of electromagnetic theory. We will introduce some major topics about optical wave guiding, an important aspect of integrated optics. The variational ideas and tools introduced in Chapter 5 will be extensively used.

6.4.1 Introduction

Optical pulse transmission deals with the propagation of optical pulses, for instance, through glass fibers. This is highly important in present-day telecommunication applications. The glass through which the light travels has dispersive and nonlinear properties, which, when dissipation is neglected, lead to equations similar to the ones describing surface water waves, as presented in §6.3. However, in case of water waves, dispersion is an effect of the fluid motion in the interior of the underlying layer, and nonlinearity is a direct effect of gravity. On the other hand, in optics the dispersive and nonlinear effects depend on properties of the material, and different materials lead to a large variety of dispersive and nonlinear effects. We will not deal with nonlinear dispersive pulse propagation here. Instead, we will consider some topics from integrated optics, thereby restricting ourselves to linear materials.

Integrated optics deals with devices to handle and select information carried by the pulses. Such micrometer scale devices consist of dielectric materials through which light is guided and manipulated to achieve a certain functionality. Typically, these devices are building blocks for optical chips and optical sensors: waveguides, wavelength splitters, filters, and Y-junctions are among the simplest and most important devices. In practice, these devices have three-dimensional geometries, but for modeling and developing insight, often planar structures are considered to simplify the analysis. Here we will briefly show how a variety of methods, including variational restriction, can be used for modeling some of the problems from optical propagation in simple geometries.

Helmholtz functional for monochromatic light

Electromagnetic waves are described by *Maxwell equations*. Restricting to nonmagnetic dielectric materials in planar configurations, and to time-harmonic light, the governing equation is the *Helmholtz equation*. Let us consider the transmission of light that is polarized in the y-direction, so the field has only one component: $\mathbf{E} = (0, E(x, z), 0)$. The field amplitude E satisfies the Helmholtz equation

$$\Delta E + k^2(x, z)E = 0. \tag{6.124}$$

If the optical properties of the slab are not homogeneous, the coefficient k depends on the position in the slab. This coefficient is given by

$$k(x, z) = k_0\, n(x, z)$$

with $n(x, z)$ the so-called refractive index. Its value depends on the dielectric material through which the light travels and is inversely proportional to the propagation speed of the light. The quantity k_0 is the wavenumber of light in a vacuum, i.e., $k_0 = \omega/c$, with ω the frequency of the light and c the speed of light in a vacuum.

The Helmholtz equation is a prominent example of a variational equation. Indeed, define the so-called Helmholtz functional

$$\mathcal{H}(E) := \frac{1}{2}\int_\Omega \left[|\nabla E|^2 - k^2E^2\right]\, dA,$$

where the integration is taken over the domain of consideration Ω, a region in the (x, z)-plane. We will not specify boundary conditions, but we require variations to vanish at the

boundary. Later on we will be more precise in specific cases. The critical points of the Helmholtz equation precisely satisfy the Helmholtz equation, as can be easily verified by investigating the first variation for arbitrary variations, leading to

$$\begin{aligned}\delta\mathcal{H}(E;\delta E) &= \int \left[\nabla E\cdot\nabla(\delta E) - k^2\, E\,\delta E\right]\, dA \\ &= -\int \left[\Delta E + k^2(x,z)E\right]\delta E\, dA\,. \end{aligned} \tag{6.125}$$

The last term is obtained after partial integration and taking into account that the perturbations δE vanish on the boundary of Ω. Vanishing of (6.125) for all variations δE leads to the Helmholtz equation (6.124).

At material interfaces, the index may be discontinuous and the Helmholtz equation has to be interpreted in a weak sense. As we shall show below, this leads to the requirement that for a continuous field its normal derivative should be continuous, so that

$$E \text{ and } \partial_n E \text{ are continuous at material interfaces.}$$

We shall now show that a variational solution automatically satisfies the desired *interface conditions*, which is an essential advantage when using the variational formulation for numerical purposes. Indeed, let $\Omega = \Omega_1 \cup \Omega_2$, where the subdomains $\Omega_{1,2}$ share a common part Γ of their boundary. Then, a similar reasoning as above, but now for each subdomain, leads to

$$\begin{aligned}\delta\mathcal{H}(E;\delta E) &= \int_{\Omega_1\cup\Omega_2} \left[\nabla E\cdot\nabla\delta E - k^2 E\cdot\delta E\right]\, dA \\ &= -\int_{\Omega_1}\left[\Delta E + k^2\, E\right]\delta E\, dA - \int_{\Omega_2}\left[\Delta E + k^2\, E\right]\delta E\, dA \\ &\quad + \int_\Gamma \nabla E\cdot\mathbf{n}\,\delta E_1\, dA - \int_\Gamma \nabla E\cdot\mathbf{n}\,\delta E_2\, dA\,. \end{aligned} \tag{6.126}$$

Here, the normal, denoted by n, points outward to Ω_1. Vanishing of (6.126) for arbitrary variations in each of the subdomains that are first required to vanish at all boundaries, including Γ, lead to the Helmholtz equation in each of these domains. Next, the resulting expression

$$\delta\mathcal{H}(E;\delta E) = \int_\Gamma \nabla E\cdot\mathbf{n}\,\delta E_1\, dA - \int_\Gamma \nabla E\cdot\mathbf{n}\,\delta E_2\, dA \tag{6.127}$$

can be considered for nonvanishing variations on the common boundary. All physically realistic fields must be continuous on all of Ω. This continuity of the field in the interior implies that the variations in the subdomains must be equal at Γ, $\delta E_1 = \delta E_2$ at Γ. From this it then follows from (6.127) that $\nabla E\cdot\mathbf{n}$ is continuous along Γ. Hence, from the variational principle it follows that wherever defined in the classical sense, the Helmholtz equation is satisfied, and that at interfaces with jumps in the material properties, the continuity conditions are satisfied. Vanishing of the first variation for all continuous variations,

$$\delta\mathcal{H}(E;\delta E) = 0 \text{ for all continuous variations } \delta E,$$

is in fact the compact *variational formulation* of the classical Helmholtz equation and the interface conditions. Also for nonvariational problems such weak formulations can be written down. As said, a major advantage is that in numerical discretizations the interface conditions are a consequence of the formulation and do not have to be imposed additionally a priori.

In the next two subsections we illustrate these notions by dealing with two simple, yet basic, problems: wave guiding in two dimensions and transmittance/reflection of light through a dielectric slab.

6.4.2 Waveguide optics

We now consider in various ways the basic problem of waveguiding. In free space, or in any homogeneous dielectric material, light will diverge so that an initially sharp bundle will spread. Waveguides are then used to "confine" the light and prevent it from spreading.

In geometrical optics light is described in terms of light rays propagating along straight lines at a speed that is depending on the density of the material. This local speed is, of course, always lower than the light speed in vacuum. A light ray traveling through a medium that is optically more dense than its surroundings and for which the index of refraction is thus larger than that of the surroundings will be totally reflected at the boundary when the angle of incidence is large enough, according to Snell's law of refraction. Then, confinement in the optically dense medium can be achieved by successive total internal reflections. However, interference of successive rays will cause that only at discrete angles light is really guided. This propagation and interference is described by physical optics, in which separate rays are not followed but the continuous electric field at each point is calculated, which satisfies the Helmholtz equation.

Consider a strip infinitely extending in the z-direction, of width $2L$ in the x-direction, surrounded by free space, and with piecewise constant index

$$n(x) = \begin{cases} n_1 & \text{for } |x| < L \,, \\ 1 & \text{for } |x| > L \,. \end{cases}$$

We will look for "guided light" traveling in the z-direction. "Guided" here means that we require the field to vanish for $|x| \to \infty$. Separation of variables in the Helmholtz equation, i.e., assuming that

$$E(x, z) = \phi(x)\, u(z),$$

we obtain the relation

$$\left[\partial_x^2 \phi + k_0^2 n^2(x) \phi\right] u + \phi\, \partial_z^2 u = 0$$

or

$$\frac{\left[\partial_x^2 \phi + k_0^2 n^2(x) \phi\right]}{\phi} = -\frac{\partial_z^2 u}{u} \,.$$

So, for some separation constant β^2 it must hold that

$$\begin{aligned} &\text{a.} \quad \partial_z^2 u + \beta^2 u = 0 \,, \\ &\text{b.} \quad \partial_x^2 \phi + k_0^2\, n^2(x)\, \phi = \beta^2\, \phi \,. \end{aligned} \tag{6.128}$$

The solution for u corresponds to traveling waves in the z-direction provided that β^2 is real and positive, since then

$$E(x, z) = \phi(x) \exp(\pm i\, \beta\, z)\,.$$

For β^2 real and negative so-called evanescent fields are obtained, which will not be considered here any further.

The value of β has still to be found and should be such that there is a confined transversal field profile ϕ that satisfies (6.128b) with the conditions

$$\phi \text{ and } \partial_x \phi \text{ continuous at } |x| = L \tag{6.129}$$

and

$$\phi(x) \to 0 \text{ for } |x| \to \infty. \tag{6.130}$$

This is an eigenvalue problem, which will determine possible values of β. Solutions are called guided modes. Depending on the width L and the index n_1, there may be one or more solutions, as we will show in the following.

Explicit construction of guided modes
In the uniform exterior, a solution of (6.128b), also called a mode, is of the form

$$\phi(x) = A_+ \exp\left[-\sqrt{(\beta^2 - k_0^2)}\,(x - L)\right] \text{ for } x > L$$

for some amplitude A_+. For $x < -L$ a similar expression holds. These modes satisfy (6.130) if $\beta^2 > k_0^2$. In the interior of the waveguide we have index n_1 and with the notation $k_1 = k_0\, n_1$ the solution is given by

$$\phi(x) = a \cos\left(\sqrt{(k_1^2 - \beta^2)}\, x + \alpha\right)\,.$$

It remains to satisfy condition (6.129) at the interfaces $x = \pm L$. This leads to four equations for the four unknowns $A_\pm$, α, and a. However, since the problem is linear, there are only three essential unknowns, which indicate that only for specific choices of β will there be a solution. The conditions are easily found by writing the interface conditions:

$$a \cos\left(\sqrt{(k_1^2 - \beta^2)}\, L \pm \alpha\right) = A_\pm,$$
$$\mp\sqrt{k_1^2 - \beta^2}\, a \sin\left(\sqrt{(k_1^2 - \beta^2)}\, L \pm \alpha\right) = -\sqrt{(\beta^2 - k_0^2)} A_\pm.$$

From this it follows that β and α have to satisfy

$$\sqrt{(k_1^2 - \beta^2)} \tan\left(\sqrt{(k_1^2 - \beta^2)}\, L \pm \alpha\right) = \sqrt{(\beta^2 - k_0^2)}\,. \tag{6.131}$$

We find two possible families of solutions, symmetric and skew-symmetric modes, for $\alpha = 0$ and $\alpha = \pi/2$, respectively. The remaining condition for β is the trigonometric equation

$$\sqrt{(k_1^2 - \beta^2)} \tan\left(\sqrt{(k_1^2 - \beta^2)} L + \alpha\right) = \pm\sqrt{(\beta^2 - k_0^2)}\,. \tag{6.132}$$

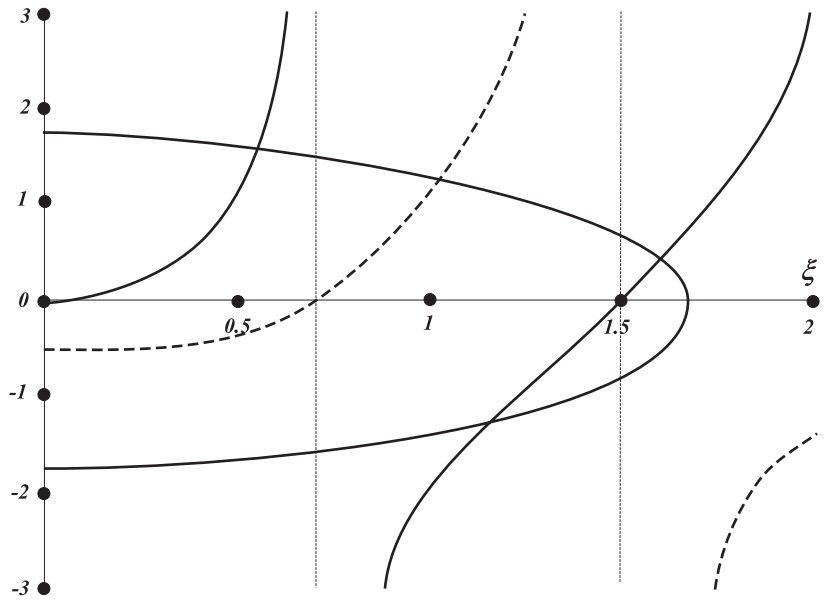

Figure 6.10. Graphical representation of the solutions of (6.131) for the parameter choices $L = 2$, $k_0 = 1$, and $k_1^2 - k_0^2 = 3$ and for the cases $\alpha = 0$ and $\alpha = \pi/2$ (dotted line).

The solution set can be viewed graphically as depicted in Fig. **6.10** with on the horizontal axis $\xi := \sqrt{k_1^2 - \beta^2}$, so $\beta^2 = k_1^2 - \xi^2$. Vertically are plotted the graphs of the right-hand side of (6.131), $\sqrt{\beta^2 - k_0^2} = \sqrt{\left(k_1^2 - k_0^2\right) - \xi^2}$ (a circle) and the left-hand side $\sqrt{(k_1^2 - \beta^2)}\tan\left(\sqrt{(k_1^2 - \beta^2)}L + \alpha\right) = \xi\tan(\xi L + \alpha)$ for $\alpha = 0$ and $\alpha = \pi/2$. Intersection points of these graphs correspond to the solutions sought for. In the plots, characteristic values for k_0, k_1, and L have been taken, but it should be observed that the number of solutions depends on the width L and the index contrast $k_1^2 - k_0^2$. In all cases there is at least one solution; the number of guided modes is always finite.

Effective boundary conditions

Instead of solving the eigenvalue problem on the whole real axis, as done above, it is possible to write down a formulation in the wave guide interior itself with suitable boundary conditions, leading to a complete formulation for the x-interval $[-L, L]$. This is achieved by observing that when the value of β would be known, we can write the expression in the exterior as done above, from which it follows that at an interface

$$\partial_x \phi = \mp\sqrt{(\beta^2 - k_0^2)}\,\phi \quad \text{at } x = \pm L\ .$$

But, since the solution sought for has to be continuous together with its derivative, the same expression should hold for the solution inside the wave guide. Hence a total formulation

would be

$$\partial_x^2\phi + k_1^2\,\phi = \beta^2\,\phi \quad \text{for} \;\; |x| < L, \tag{6.133}$$

$$\partial_x\phi = \mp\sqrt{(\beta^2 - k_0^2)}\,\phi \quad \text{at} \;\; x = \pm L. \tag{6.134}$$

This confined formulation is particularly useful when the problem has to be calculated numerically; this will be the case, for instance, when the index in the waveguide is not constant but varying. Then, instead of having to solve the problem on the whole real line, we can discretize the problem on the finite interval $[-L, L]$. The boundary conditions are therefore called *effective boundary conditions*: they replace at the boundary the effect of the solution in the exterior.

Of course, the price to be paid for the confinement is that now also the boundary condition contains the eigenvalue β to be determined. For this reason such a problem is called a nonlinear eigenvalue problem. In fact, in this case it is a quadratic eigenvalue problem, which can be seen by introducing $\mu := \sqrt{\beta^2 - k_0^2}$ as the eigenvalue. Then μ appears linearly in the boundary conditions and quadratically in the interior equation. Taking the solution inside the waveguide as above, we find directly the same conditions as in the previous subsection.

Effective index and dimension reduction

The eigenvalue β^2 found for guided modes is the square of the wave number of the mode for its propagation in the z-direction, and β is often called the *propagation constant*. It can be related to an effective index N_e by writing $\beta = k_0\,N_e$. This has as interpretation, and motivation for the name, that the propagation in the z-direction is as if the mode is propagating in a uniform medium with index N_{eff}. Since $N_{eff} < n_1$, the effect of the mode being exponentially confined in the transversal direction to the waveguide is that it reduces the actual index. In the ray picture of the light, this can be understood, since the light does not propagate parallel to the waveguide but makes an angle and remains within the guide by internal reflections at the boundaries. For this longer (than straight) trajectory, the effective index has to be smaller.

This may lead to the practical question of whether it would be possible to get an estimate or approximation for the effective index without solving completely the eigenvalue problem for ϕ that yields the value of β. If that is possible, we essentially reduce a two-dimensional problem to a one-dimensional problem, with the approximation for the effective index replacing the exact value of the propagation constant. This has as generalization the question how to reduce three-dimensional problems that have a strongly planar field distribution (either because the field outside the plane vanishes rapidly, or is nearly uniform) to "effective" two-dimensional problems.

Specifically, illustrating this for the waveguide problem, let us reconsider the method of separation of variables used above. From the Ansatz $E(x, z) = \phi(x)\,u(z)$ we found that

$$\left[\partial_x^2\phi + k_0^2 n^2(x)\phi\right] u(z) + \phi(x)\partial_z^2 u = 0.$$

To get a reduction to an equation in the z variable only, one has to get rid of the x-dependence, replacing this equation by some equation of the form

$$\partial_z^2 u + \tilde{\beta}^2\,(z)\,u = 0,$$

where we even allow the effective coefficient $\tilde{\beta}^2$ to depend on z. For the exact transversal profile $\phi(x)$ this was possible, but it is not clear what to do when that profile is only approximately known. For instance, one choice would be an integration over x, leading to

$$\partial_z^2 u + \tilde{\beta}^2 u = 0 \text{ with } \tilde{\beta}^2 = \frac{\int \left[\partial_x^2 \phi + k_0^2 n^2(x)\phi\right] dx}{\int \phi \, dx}, \tag{6.135}$$

provided that the denominator does not vanish. Another choice would be

$$\partial_z^2 u + \bar{\beta}^2 u = 0 \text{ with } \bar{\beta}^2 = \frac{\int \left[(\partial_x \phi)^2 - k_0^2 n^2(x)\phi^2\right] dx}{\int \phi^2 \, dx}. \tag{6.136}$$

Of these and other heuristic, or ad hoc, methods it is not clear in advance which one is the "best." In any case, when using the exact profile, both methods mentioned above produce the same and "correct" result $\tilde{\beta} = \bar{\beta} = \beta$.

The road to be described next will make it clear that the second choice (6.136) is the best, and variational accuracy will imply that a first-order error in the profile will produce only a second-order error in the value of the quotient that approximates β.

Variational aspects of waveguiding

We illustrate the application of variational methods using the waveguide problem treated above.

We consider the same simple waveguide geometry as above but allow the index to change in both the x- and the z-directions. Now we make the Ansatz that the field can be separated:

$$E(x, z) = \phi(x, z)\, u(z) \, ,$$

where ϕ is real and u may be complex. This thus defines the model manifold, which is a subset of all possible fields. Substitution in the Helmholtz functional

$$\mathcal{H}(E) = \int\int \left[|\partial_x E|^2 + |\partial_z E|^2 - k^2(x, z)|E|^2\right] dx\, dz$$

leads to the restricted functional

$$\begin{aligned}\mathcal{H}(u, \phi) = \int \Bigg[&\left(\int \phi^2 \, dx\right) |\partial_z u|^2 \\ &- \left\{\int k^2(x, z)\, \phi^2 \, dx - \int \left[(\partial_x \phi)^2 + (\partial_z \phi)^2\right] dx \right\} |u|^2 \Bigg] \, dz.\end{aligned}$$

Let us first consider the case that k is independent of z, so $k = k(x)$. Then, we also look for a function ϕ independent of z, and we may introduce a scaled version of $\mathcal{H}(u, \phi)$ by

$$\mathcal{H}_\phi(u) := \int \left[|\partial_z u|^2 - \lambda(\phi)\, |u|^2\right] dz \, ,$$

where we divided by $\int \phi^2 \, dx$, and where $\lambda(\phi)$ is the functional

$$\lambda(\phi) = \frac{\int \left[(\partial_x \phi)^2 - k^2(x)\, \phi^2\right] dx}{\int \phi^2 \, dx} \, .$$

For given ϕ, the equation for u follows from the critical points of the functional $u \to \mathcal{H}_\phi(u)$ and reads as

$$\partial_z^2 u + \lambda(\phi)\, u = 0\,,$$

where $\lambda(\phi)$ now plays the role of squared propagation constant. If it were known, and positive, we would find the traveling modes in the z-direction, $u(z) = \exp\left(\pm i\sqrt{\lambda(\phi)}\, z\right)$.

No variations in ϕ were considered so far. Taking into account variations in ϕ leads to the conclusion that an optimal electric field should be a critical point of the functional $\phi \to \lambda(\phi)$. This is the *Rayleigh functional*. Differentiating this functional we find that such a critical point satisfies the equation

$$\partial_x^2 \phi + k^2(x)\phi = \beta^2 \phi.$$

In fact, the value of β^2 is precisely the value of $\lambda(\phi)$ when evaluated at the optimal solution. Hence, we recover the same equation (and the same solutions) as in the above treatment of (6.131). From this we can conclude that of the two ad hoc ways (6.135) and (6.136) to formulate the differential problem into a one-dimensional problem, the second choice (6.136) is the most appropriate choice. We can now extract even more information obtained from variational accuracy. This tells us that a first-order error in the electric field leads to only a second error in the value: the eigenvalue can be approximated to a higher order than the eigenfunction.

The above method can be generalized to cover the case that the waveguide is (slowly) varying in the z-direction: $k = k(x, z)$. Again, we take as Ansatz the model manifold $E(x, z) = \phi(x, z)u(z)$. This implies that we look for a *quasi-homogeneous approximation*: at each position z_0 the waveguide with index $k(x, z_0)$ is considered, for which $\phi(x, z_0)$, the local mode profile, which depends on z_0, can be calculated as before. This quasi-homogeneous approximation couples the transversal fields and can be expected to be a good approximation if the variations in the propagation direction are small, i.e., for slow variations in the z-direction. A similar procedure as above then leads to

$$\mathcal{H}_\phi(u) = \int \left[N(\phi)\,|\partial_z u|^2 - \mu(\phi, z)\,|u|^2\right] dz,$$

where both $N(\phi)$ and $\mu(\phi)$ are functionals of ϕ, thus also depending on z:

$$N(\phi) = \int \phi^2 dx\,,$$
$$\mu(\phi) = \int k^2(x, z)\,\phi^2 dx - \int \left[(\partial_x\phi)^2 + (\partial_z\phi)^2\right] dx.$$

Since we assume slow variations, the dependence of ϕ on z is slowly varying, which motivates us to neglect the term $\int (\partial_z\phi)^2\, dx$. Then,

$$\mu(\phi) \approx -\int \left[(\partial_x\phi)^2 - k^2(x, z)\phi^2\right] dx\,.$$

For given ϕ, the equation for u follows again from the critical point equation for $u \to \mathcal{H}_\phi(u)$, and hence we have that $\partial_z^2 u + \mu(\phi)\, u = 0$. Since both functionals N and μ are

homogeneous of degree two, it is no restriction to assume the function ϕ to be normalized, $N(\phi) = 1$. Then, we find essentially the same eigenvalue problem as before. The main difference is that now z enters the eigenvalue equation through the dependence of k on z:

$$\partial_x^2\phi + k^2(x,z)\phi = \beta\phi \,.$$

Example 6.4a. Tapered waveguide.
We consider a waveguide with variable thickness in the x-direction. For simplicity we consider a symmetric configuration, $\{(x,z)||x| < L(z)\}$. To make formulas easier, we assume that instead of the exponential decay in the uniform exterior, we invoke hard mirror boundary conditions, which means that we require E to vanish at the boundary $|x| = L(z)$. Then, at each position z, the guided modes are easily found. We will restrict to the principal mode given by $\phi(x,z) = \left(1/\sqrt{L(z)}\right)\cos(\pi x/2L(z))$. For this mode it holds that

$$\int_{-L}^{L} \phi^2\,dx = 1, \quad \int_{-L}^{L} (\partial_x\phi)^2\,dx = \frac{\pi^2}{4L^2(z)} \,.$$

The resulting functional for u then is

$$\int \left[|\partial_z u|^2 - \left\{ k_1^2 - \frac{\pi^2}{4L^2(z)} \right\} |u|^2 \right] dz \,,$$

leading to the equation for the propagation part

$$\partial_z^2 u + \left\{ k_1^2 - \frac{\pi^2}{4L^2(z)} \right\} u = 0 \,.$$

This is the correct result for the special case that L is constant, and an approximation if $L(z)$ is slowly varying. The resulting inhomogeneous one-dimensional problem can analytically be tackled with the WKB method for slow variations and numerically with methods described in the next section. □

6.4.3 Variational approach to the WKB approximation

In §3.2.3 it was shown that the WKB method provides asymptotically correct approximations for the problem of light propagating in a one-dimensional medium with slowly varying transmission coefficient. Also in the approximations for two-dimensional waveguiding problems treated above, we end up with one-dimensional inhomogeneous problems. We intend to apply the WKB method here, too. Using the variational methods of Chapter 5, we first reconsider the WKB method.

For simplicity of exposition, we consider the one-dimensional problem of light propagation in an inhomogeneous medium. Recall that for such optical structures with index depending on z, so $k = k(z)$, the governing Helmholtz functional and equation are given by

$$\mathcal{H}(E) := \int \left[|\partial_z E|^2 - k^2(z)|E|^2 \right] dz,$$

which gives rise to the equation

$$\partial_z^2 E + k^2(z)E = 0.$$

The derivations in §3.2.3 of the WKB approximation can also be done within the variational framework, as we will show now. To that end we write the electrical field E in terms of a real amplitude and a phase,

$$E = A(z)\, e^{i\theta(z)},$$

and insert this into the Helmholtz functional:

$$\mathcal{H}(A,\theta) := \int \left[(\partial_z A)^2 + A^2\,(\partial_z\theta)^2 - k^2(z)A^2\right] dz\ .$$

The resulting Euler–Lagrange equations are obtained by variations with respect to A and θ, and are read as

$$\begin{aligned} &\text{a.} && -\partial_z^2 A + A\,(\partial_z\theta)^2 - k^2(z)A = 0\ , \\ &\text{b.} && -\partial_z\left[A^2\,\partial_z\theta\right] = 0\ . \end{aligned} \tag{6.137}$$

The last conservation law, also referred to as the Poynting identity, leads to the constraint

$$A^2\partial_z\theta = P$$

for some constant P. Substitution in (6.137a) yields the amplitude equation

$$-\partial_z^2 A + \frac{P^2}{A^3} - k^2(z)A = 0\ .$$

This equation is just the Euler–Lagrange equation of the functional

$$\int \left[(\partial_z A)^2 - \frac{P^2}{A^2} - k^2(z)A^2\right] dz.$$

Note the minus sign in front of the second term in the integrand, which would not appear when—wrongly—the constraint is simply put directly into the original Helmholtz functional.

The WKB approximation is obtained by taking in a quasi-homogeneous way the *equilibrium solution* of the amplitude equation:

$$A^2 = \frac{P}{k(z)}.$$

This leads to the WKB approximation

$$E(z) = \frac{A_0}{\sqrt{k(z)}} e^{i\theta(z)} \quad \text{with}\ \ \partial_z\theta(z) = k(z).$$

6.4.4 Transmittance through a finite slab

Let us again consider the one-dimensional transmittance problem. However, we now relax the restriction that the change in material properties is slow. Then, we will generally not be able to find good approximations via the WKB approach. The main reason is that the inhomogeneity will cause light to be reflected in a complicated way. Since the WKB approximation takes into account only waves running in one direction, neglecting any reflection, this method is not reliable here.

Since there is no systematic procedure with which to obtain approximations in the general case of inhomogeneous material, we approach the problem in a different way: we introduce (exact) artificial boundary conditions such that the problem can be tackled in a numerical way. The formulation of these boundary conditions is not trivial: these must describe the physics correctly. This in particular means that transmittance or reflection at a boundary must be modeled correctly, and this depends on the properties in the exterior, outside the slab. The profit of such a description is that it allows for a reliable implementation of numerical calculations.

Consider a slab of thickness L and infinite width of some dielectric material. See Fig. **6.11**. Note that in the following we take the propagation direction of the waves, which is orthogonal to the slab, as the z-axis. Outside the slab we assume the medium to be uniform, having constant index. The dielectric may have arbitrary index variations inside the interval $z \in [0, L]$. So, we do not restrict ourselves to slowly varying or stepwise changing indices. Outside the interval we have $k(z) = k_-$ for $z < 0$ and $k(z) = k_+$ for $z > L$. An incoming wave from the left is assumed to have amplitude A, and no incoming waves from the right are included. Reflected and transmitted waves with a priori unknown reflection coefficient r and transmittance coefficient t have to be taken into account, so we have

$$E(z) = \begin{cases} A\exp(ik_-z) + r\exp(-ik_-z) & \text{for } z < 0, \\ t\exp(ik_+(z-L)) & \text{for } z > L. \end{cases}$$

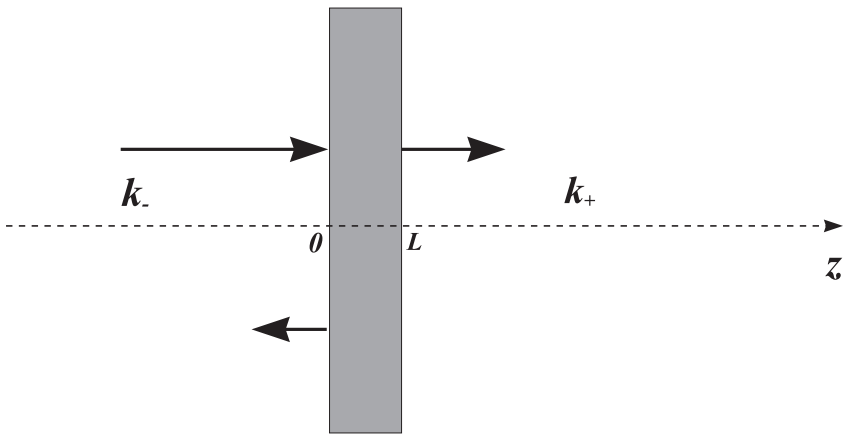

Figure 6.11. Transmittance of electromagnetic waves through a slab of thickness L.

The problem is to characterize the solution, in particular to find the reflection and transmission coefficients r and t. A priori a relation between the solutions on both sides of the slab can be found from the *Poynting quantity*, which is here defined as

$$P(z) := \text{Im}\left[E^* \partial_z E\right].$$

Here, E^* denotes the complex conjugate. It can readily be verified that for a solution of the Helmholtz equation P is conserved:

$$\partial_z P = 0.$$

P can easily be evaluated in a uniform medium and we find that

$$P = k_-\left[|A|^2 - |r|^2\right] \quad \text{for} \quad z < 0$$

and

$$P = k_+|t|^2 \qquad \text{for} \quad z > L.$$

The relation that expresses momentum conservation through the slab is then

$$k_-\left[|A|^2 - |r|^2\right] = k_+|t|^2 \,.$$

Transparent influx boundary conditions

We now transform the problem to a confined problem, a boundary value problem on the interval $[0, L]$, by deriving so-called transparent influx boundary conditions (TIBC).

At the left, $z < 0$, we have an incoming wave $A\, e^{ik_- z}$ traveling to the right. Let us take as TIBC at $z = 0$ the condition

$$\partial_z E + ik_- E = 2ik_- A.$$

Clearly, in the region $z < 0$, a superposition of the incoming wave with *any* reflected wave, $E = Ae^{ik_- z} + re^{-ik_- z}$ satisfies this relation at $z = 0$. Since both E and $\partial_z E$ have to be continuous at $z = 0$, this relation should therefore also be satisfied by the solution inside the slab. And conversely, when the interior solution satisfies this condition at $z = 0$, it can be smoothly matched to a solution in the exterior $z < 0$. In the same way, now using the condition that no waves come in from the right, we prescribe a transparent boundary condition at $z = L$:

$$\partial_z E - ik_+ E = 0 \,.$$

Indeed, *any* solution travelling to the right for $z > L$, i.e., $E = t\, e^{ik_+ z}$, satisfies this condition, and hence this condition should also hold at $z = L$ for the solution in the interior of the slab.

If the index is constant over the whole real line, so $k(z) = k_0$, the influx at $z = 0$ of the wave $Ae^{ik_0 z}$ should be equal to the outflux at $z = L$, which indeed satisfies the TIBCs:

$$\begin{aligned} &\partial_z E + ik_0 E = 2ik_0 A \quad \text{at } z = 0, \\ &\partial_z E - ik_0 E = 0 \qquad \text{at } z = L. \end{aligned} \tag{6.138}$$

When a mirror is placed at $z = L$, the boundary condition reads $E = 0$ at $z = L$. The mirror will reflect the total wave at $z = L$, and hence the solution is

$$E = Ae^{ik_0 z} - A\, e^{ik_0(2L - z)} \,.$$

This solution indeed satisfies the TIBC at $z = 0$,

$$\partial_z E + ik_0 E = 2ik_0 A \quad \text{at } z = 0,$$

so that the reflected wave is nicely transmitted through $z = 0$. The value of the solution at $z = 0$ is $E(0) = A\left(1 - e^{2Lik_0}\right)$. Trying to formulate Dirichlet boundary conditions will never be possible since the reflected wave is generated inside the slab and is unknown in advance. Therefore the trick in the TIBC is that neither Dirichlet nor Neumann conditions are used, but a combination of the two. In fact, in the T(I)BC one can recognize the Dirichlet-to-Neumann operator introduced in §6.3.3. This idea can be generalized to higher-dimensional problems, too.

Variational formulation
The full description of the confined problem derived above reads as

$$\begin{aligned} &\partial_z^2 E + k^2(z)E = 0\,, \\ &\partial_z E + ik_- E = 2ik_- A \quad \text{at } z = 0\,, \\ &\partial_z E - ik_+ E = 0 \qquad \text{at } z = L\,. \end{aligned} \tag{6.139}$$

This problem has a variational structure. Not only the equation in the interior has this structure, but also the boundary conditions. The total formulation uses the sum of the Helmholtz functional, and a specific boundary functional. Solutions of the boundary value problem (6.139) are in one-to-one correspondence with the critical points of the functional

$$\begin{aligned} \mathcal{L}(E) = &\int_0^L \frac{1}{2}\left[(\partial_z E)^2 - k^2 E^2\right] dz \\ &- \frac{1}{2} ik_- E(0)^2 - \frac{1}{2} ik_+ E(L)^2 + 2ik_- A E(0)\,. \end{aligned} \tag{6.140}$$

This follows with the standard reasoning from the vanishing of the first variation for arbitrary variations δE:

$$\begin{aligned} \delta\mathcal{L}(E;\delta E) = &\int_0^L \left[(\partial_z E)(\partial_z \delta E) - k^2\, E\, \delta E\right] dz \\ &- ik_-\, E(0)\, \delta E(0) - ik_+\, E(L)\, \delta E(L) + 2ik_-\, A\, \delta E(0) \\ = &\int_0^L \left[-\partial_z^2 E - k^2 E\right] \delta E\, dz + (\partial_z E)\, \delta E|_0^L \\ &-ik_-\, E(0)\, \delta E(0) - ik_+\, E(L)\, \delta E(L) \; + 2ik_-\, A\, \delta E(0)\,. \end{aligned}$$

Restricting first to variations that vanish at the boundaries, so $\delta E(0) = \delta E(L) = 0$, we find the correct Helmholtz equation. Variations that are arbitrary at the boundary lead to the condition

$$(\partial_z E)\, \delta E|_0^L - ik_-\, E(0)\, \delta E(0) - ik_+\, E(L)\, \delta E(L) + 2ik_-\, A\, \delta E(0) = 0\,.$$

From this it follows that

$$-(\partial_z E)|_0 - ik_- E(0) + 2ik_- A = 0$$

and

$$(\partial_z E)|_L - ik_+ E(L) = 0.$$

We conclude that not only the Helmholtz equation but also the TIBC automatically follows from the condition $\delta\mathcal{L}(E;\delta E) = 0$.

Note that we did not specify the interior index structure. Of course, only for simple structures can explicit solutions be found, for instance, when k is piecewise constant. Well-known examples are gratings, when the index changes periodically. Above, the emphasis has been on the procedure to derive the confined formulation, because this formulation can also be exploited for numerical discretizations in case of an arbitrary internal index profile.

Exercise 6.4a.

Consider the case with constant index k_1 in the interior interval.

a. *Calculate explicitly the solution by matching the general solutions in the three separate intervals using the interface conditions.*

b. *Show that the solution satisfies the TIBCs and that it indeed can be found only by solving the problem on the bounded interval with the TIBCs.*

c. *Considering the reflection and transmission, determine the values of the length L such that the transmittance is minimal and maximal.*

Exercise 6.4b.

Extend the problem in Exercise 6.4a *to the case that k is piecewise constant in the interval, leading to transfer matrix techniques. This is possible since for the piecewise constant index, the solution is simple in each interval of constant index, while the connection between different intervals is determined by interface conditions; an algebra software package may be helpful.*

Numerical implementation

To tackle the optical transmittance problem numerically, we follow the same approach as in §5.5.2. A discretization for the problem of optical transmittance through a dielectric medium is now easily obtained by using the functional (6.140):

$$\begin{aligned}\mathcal{L}(E) = \int_0^L \frac{1}{2}\left[(\partial_z E)^2 - k^2 E^2\right] dz \\ - \frac{1}{2} ik_- E(0)^2 - \frac{1}{2} ik_+ E(L)^2 + 2ik_- A E(0)\,.\end{aligned}$$

A complication is that we have to deal with complex valued functions and with boundary terms in the functional. In view of the boundary terms we have to include the endpoints of the interval into the grid. This is done by taking a uniform mesh-size of length L/N and grid points $z_0 = 0 < z_1 < \cdots < z_N = L$. Let the functional be approximated with tent functions as defined in §5.5.2:

$$E^{(N)}(z) = \sum_{k=0}^{N} c_k\, T_k(z)\,,$$

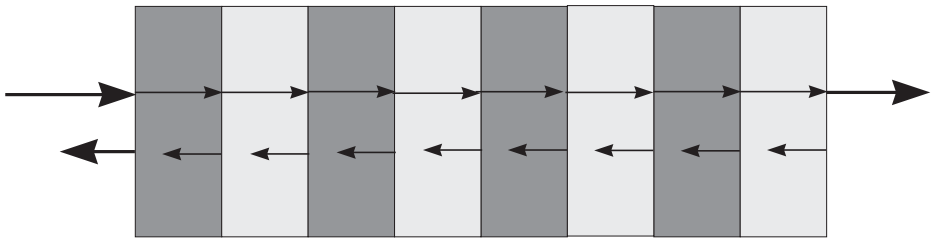

Figure 6.12. Transmittance of waves through a grating with four periods, each consisting of two layers with different indices.

where the coefficients may be complex valued. Inserting this expression into the functional leads to

$$\mathcal{L}(E^{(N)}) = \frac{1}{2}\left[\mathbf{c}\cdot\mathbf{P}\,\mathbf{c} - \mathbf{c}\cdot\mathbf{Q}\,\mathbf{c}\right] - \frac{1}{2}ik_-\,{c_0}^2 - \frac{1}{2}ik_+\,c_N^2 + 2ik_-\,A\,c_0\ ,$$

where the (real) matrices $\mathbf{P}$ and $\mathbf{Q}$ are defined in a similar way as in §5.5.2, and the vector $\mathbf{c}$ contains the coefficients: $\mathbf{c} := (c_0, c_2, \ldots, c_N)$. Requiring the derivative to vanish, we obtain an algebraic set of equations:

$$[\mathbf{P} - \mathbf{Q}]\,\mathbf{c} - \mathbf{B}\,\mathbf{c} + \mathbf{f} = \mathbf{0}\ .$$

Here, $\mathbf{B}$ is the matrix with only two nonzero elements: $B_{00} = ik_-$, $B_{NN} = ik_+$, and $\mathbf{f}$ is the vector with only one nonzero component: $f_0 = 2ik_-$. These last terms are a consequence of the boundary conditions.

Exercise 6.4c.

a. *Write a numerical implementation of the discretization scheme above.*

b. *Check your implementation by taking special cases for which the solution is known (at least qualitatively), for instance, for the case of a uniform index:* $k = k_0\,n_0 = \frac{\omega}{\mathrm{c}}\,n_0$ *with* n_0 *constant.*

c. *Then consider a grating consisting of four periods, in each of which are two layers of constant but different indices. See* Fig. **6.12**. *Calculate the transmittance plot of this grating, i.e., the absolute value of the amplitude of the transmitted wave, normalized by the amplitude of the incoming wave, as a function of* ω.

d. *Vary the values of the indices, and observe that the transmittance plot is not easily predicted.*

   We remark that this plot also can be obtained by solving the problem explicitly, as was suggested in Exercise 6.4b *with transfer matrix techniques; the periodicity can be exploited to simplify the calculations.*

Exercise 6.4d.

Construct a numerical scheme to calculate all the guided mode profiles and the corresponding propagation constant of a uniform waveguide. See §6.4.2.

Some numerical software packages use the variational (weak) formulation as explained here. The technical problems related to include the geometrical structure, and the corresponding mesh consisting of elements, are taken care of by the program itself, and extensive calculations can be performed quickly. Be aware, however, that the boundary conditions may not be as desired; in particular there are no facilities with which to implement TIBCs as we have described above for one-dimensional problems, and usually only approximate boundary conditions have been implemented.

Bibliography

[1] V.I. Arnold, *Ordinary Differential Equations*, MIT Press, Cambridge, MA, 1973.

[2] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.

[3] G.I. Barenblatt, *Dimensional Analysis*. Gordon and Breach Science Publishers, New York, 1987.

[4] G.I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge Texts in Applied Mathematics 14, Cambridge University Press, Cambridge, UK, 1996.

[5] R.B. Bird, R.C. Armstrong, O. Hassager, and C.F. Curtiss, *Dynamics of Polymeric Liquids*, *Vol.* 2*: Kinetic Theory*, John Wiley, New York, 1977.

[6] E. Buckingham, *On physically similar systems: Illustrations of the use of dimensional equations*, Physical Review, 1914, issue 4, pp. 345–376.

[7] M. Burry, *Expiatory Church of the Sagrada Familia, Antoni Gaudi*, Phaidon, San Francisco, 1993.

[8] M. Doi and S.F. Edwards, *The Theory of Polymer Dynamics*, Clarendon Press, Oxford, UK, 1986.

[9] N.D. Fowkes and J.J. Mahony, *An Introduction to Mathematical Modeling*, John Wiley, New York, 1994.

[10] A.C. Fowler, *Mathematical Models in the Applied Sciences*, Cambridge University Press, Cambridge, UK, 1997.

[11] H. Goldstein, *Classical Mechanics*, Addison-Wesley, Reading, MA, 1959.

[12] E. van Groesen, *WKB- and SVEA-approximative methods for slowly varying nonlinear dielectric media*, Optical and Quantum Electronics, 30, 1998, pp. 467–474.

[13] E. van Groesen and E.M. de Jager, *Mathematical Structures in Continuous Dynamical Systems*, Elsevier, Amsterdam, 1994.

[14] T.W.B. Kibble, *Classical Mechanics*, McGraw-Hill, London, 1966.

[15] R. Haberman, *Mathematical Models*: Mechanical Vibrations, Population Dynamics, and Traffic Flow, Classics in Applied Mathematics 21, SIAM, Philadelphia, 1998.

[16] G. Klopman, M. Dingemans, and E. van Groesen, *A variational model for fully nonlinear water waves of Boussinesq type*, in Proceedings of the 20th International Workshop on Water Waves and Floating Bodies, J. Grue, ed., Spitsbergen, Norway, 2005. Available online at http://www.math.uio.no/iwwwfb/include/proc1508.pdf

[17] D.J. Korteweg, and F. de Vries, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves*, Philosophical Magazine, 39, 1895, 422–443.

[18] L.D. Landau and E.M. Lifshitz, *Fluid Mechanics*, Pergamon Press, Oxford, UK, 1959.

[19] R.G. Larson, *Constitutive Equations for Polymer Melts and Solutions*, Butterworths, Boston, 1988.

[20] C.C. Lin, *Theory of Hydrodynamic Stability*, Cambridge University Press, London, New York, 1955.

[21] J.C. Luke, *A variational principle for a fluid with a free surface*, Journal of Fluid Mechanics, 27, 1967, pp. 395–397.

[22] R.M.M. Mattheij and J. Molenaar, *Ordinary Differential Equations in Theory and Practice*, Classics in Appl. Math. 43, SIAM, Philadelphia, 2002.

[23] R.M.M. Mattheij, S.W. Rienstra, and J.H.M. ten Thije Boonkkamp, *Partial Differential Equations: Modeling, Analysis, Computations*, Monographs in Mathematical Modeling and Computation 10, SIAM, Philadelphia, 2005.

[24] J.W. Miles, *The Korteweg-de Vries Equation: A Historical Essay*, Journal of Fluid Mechanics, 106, 1981, pp. 131–147.

[25] D.R. Shier and K.T. Wallenius, *Applied Mathematical Modeling: A Multidisciplinary Approach*, Chapman and Hall/CRC, Boca Raton, FL, 2000.

[26] D. Basmadjian, *The Art of Modeling in Science and Engineering*, Chapman and Hall/CRC, Boca Raton, FL, 1999.

[27] J.C. Slattery and S. Lee, *Analysis of melt spinning*, Journal of Non-Newtonian Fluid Mechanics, 89, 2000, pp. 273–286.

[28] I. Stakgold, *Green's Functions and Boundary Value Problems*, 2nd ed, John Wiley, New York, 1998.

[29] T. Svobodny, *Mathematical Modeling for Industry and Engineering*, Prentice–Hall, Upper Saddle River, NJ, 1998.

[30] A.B. Tayler *Mathematical Models in Applied Mechanics*, Clarendon Press, Oxford, UK, 1986.

[31] G.I. Taylor, *The Formation of a Blast Wave by a Very Intense Explosion*. II. *The atomic explosion of* 1945, Proceedings of the Royal Society London Series A, 201, 1950, pp. 175–186.

[32] A. Yarin, *Free Liquid Jets and Films: Hydrodynamics and Rheology*, John Wiley, New York, 1993.

[33] E. Zauderer, *Partial Differential Equations of Applied Mathematics*, 2nd ed., John Wiley, New York, 1989.

[34] A. Ziabacki and H. Kawai, ed., *High-Speed Fiber Spinning*, John Wiley, New York, 1985.

Index