Spectral Simulation Methods

Numerical solution of partial differential equations

Andreas Greiner, Martin Ladecký, Lars Pastewka

July 4, 2024

Department of Microsystems Engineering
University of Freiburg

Many thanks to Jan Grießer, Anna Hoppe, Johannes Hörmann, Indre Jödicke, Maxim Kümmerle, Martin Ladecky, Antoine Sanner and the participants of the "Simulationstechniken" at the University of Freiburg in the winter terms of 2020/21 und 2021/22 for comments and edits.

Contents

Chapter 1

Introduction

Context: The term *simulation* refers to the numerical (computer-aided) solution of *models*. In this introductory chapter, we discuss how models of physical reality are build and present different classes of models. These models are usually described mathematically by means of *differential equations*, i.e. "simulation" is often (but not always) the numerical solution of a set of ordinary or partial differential equations.

1.1 Models

Models are approximations for the behavior of the physical world at certain length scales. For example, a model that explicitly describes atoms "lives" on length scales on the order of nm and may be appropriate to describe the growth of thin films in semiconductor manufacturing. We would not want to describe a macroscopic system or phenomenon that lives on scales of $\sim$ mm or beyond, such as how water flows out of a tap or how an airplane wing bends during takeoff, with such a model. Key to carrying out simulations is therefore the ability to match the physical phenomenon we want to describe with the appropriate model and the mathematical method required for its solution.

Note: While we *could* describe even macroscopic systems with atomic-scale models, this is typically prohibited by the computer resources available to us. Macroscopic systems consist of more than 10^{23} (Avogadro's number) atoms, whose positions we would not be able to fit into present day computers. In addition, the gist of the question we want to answer

may be hidden in such a fine-grained atomic-scale model like the legendary needle in a haystack.

Figure 1.1 shows on the vertical axis *length scales* and classes of models that live on these scales. On the shortest length scale, a quantum mechanical description is usually necessary. This means that if we want to resolve the world with Å resolution, we find ourselves at the level of quantum mechanics and all underlying models are of a quantum mechanical nature. Underlying quantum mechanics is the *Schrödinger equation*, whose (approximate) solution is implemented in various methods, such as *density functional theory* (Martin, 2004), a many-body description of the quantum mechanical electronic system. If we get rid of modeling the electron explicitly, we arrive at a class of simulation methods often referred to as *molecular dynamics* (Allen and Tildesley, 1989). The key mathematical object in molecular dynamics is the set of positions and velocities of all atoms, which means we have to introduce three position and three velocity variables for each of the n interacting particles. In contrast, in a quantum mechanical many-body description we are dealing with a field with three n position variables each, namely $\Psi(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n; t)$. This illustrates that formulating models on larger length scales requires some form of *coarse-graining*, i.e. removing information from a smaller scale model.

Note:

- $1\,\mathring{A} = 10^{-10}\,\mathrm{m}$
- The atoms that constitute our physical world are held together by quantum mechanics. Models based on quantum mechanical principles are also called *ab-initio* ("from the beginning") or *first principles* models. The fundamental equation that describes quantum mechanical objects is the *Schrödinger equation*. It is itself is in fact already an approximation, despite the fact that models derived from it are called *first principles* models!
- The single-particle Schrödinger equation is $i\hbar\frac{\partial}{\partial t}\Psi(\vec{r}, t) = \hat{H}\Psi(\vec{r}, t)$. This is a partial differential equation for the location- and time-dependent scalar matter field $\Psi(\vec{r}, t)$, with Planck's constant $\hbar$ and the Hamilton operator $\hat{H}$, which contains the details of the model. The solution of an equation of motion for many interacting particles, as described by a wavefunction with mathematical structure $\Psi(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n; t)$, is incomparably more complicated.

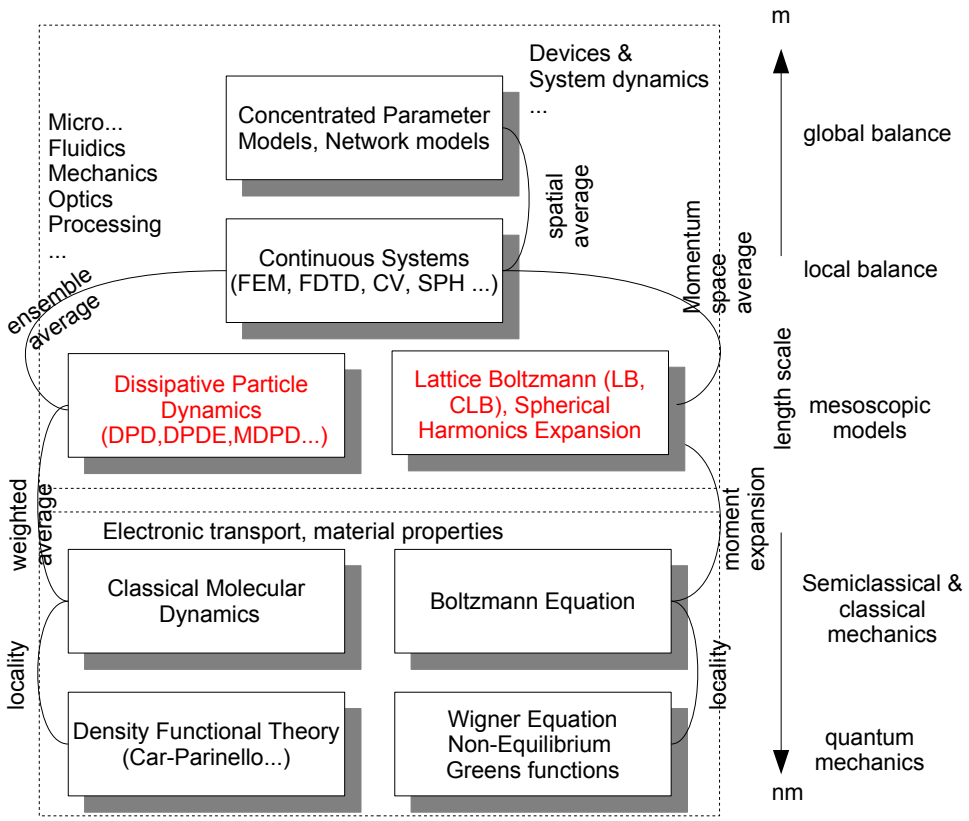

Figure 1.1: The vertical arrangement of the boxes corresponds to a length scale, with the shortest scales shown on the bottom. The boxes themselves show categories of models or simulation methods that are used on these scales. In this class we deal with the discretization of fields and choose a specific use case that falls into the *local balance* category.

- "Semiclassical" means that the motion of the particles is calculated according to classical mechanics, but the interactions between the particles are derived from quantum mechanical laws. This is of course an approximation that needs to be justified.
- "Mesoscopic" means that the model has an internal length scale and/or thermal fluctuations are important. These models usually operate on length scales above the atomic scale ($\sim$ nm) but below the scales of our perception of the environment ($\sim$ mm).
- "Balance" means that the core of the description is a *conserved quantity*. Conserved are e.g. particle numbers or mass (that is typically automatically conserved in models that have particles as the core mathematical object). The *balance equation* or balancing then simply counts the particles that flow into or out of a volume element over a certain time interval. Other conserved variables that can be balanced are momentum and energy. The balance equation is also called the *continuity equation*.

At the level of semiclassical and classical mechanics, also referred to as the kinetic level, models are either described by molecular dynamics or by the equation of motion of the single-particle probability density in phase space $f(\vec{r},\vec{p})$ - with location $\vec{r}$ and momentum $\vec{p}$ as independent variables. In the second case, we have a function $f(\vec{r}(t),\vec{p}(t),t)$ which depends on time both explicitly and implicitly via $\vec{r}(t)$ and $\vec{p}(t)$. Let us assume that we need to discretize $f(\vec{r}(t),\vec{p}(t),t)$ on regular grid of discrete sampling points. At a low resolution of 10 points per variable, this corresponds to already $10,000,000$ interpolation points. This may be manageable, but the resolution of such a model would not particularly good. This undertaking is therefore rather useless. We do not want to conceal the fact that there are methods for the numerical solution to the two problems described above, but these will not be discussed in detail in this class.

1.2 Particles

We can therefore roughly distinguish between two types of models: Models that have individual discrete elements, for example particles (atoms, molecules, grains, etc.), as their central mathematical objects and models that have continuous fields (electrostatic potential, ion concentrations, mechanical stresses

and strains) as the central objects. In the first type of model, evolution equations are formulated for discrete properties defined on the particles, such as their positions $\vec{r}_i$ and velocities $\vec{v}_i$.

For example, to describe the kinetics of these particles, we could solve Newton's equations of motion. This means that for each of the n particles we have to formulate 6 *ordinary differential equations (ODEs)*, which are coupled to each other, namely:

$$\dot{\vec{r}}_i(t) = \vec{v}_i(t) = \frac{\vec{p}_i(t)}{m_i} \tag{1.1}$$

This is the equation for the trajectory of the particle i in space. Since $\vec{r}_i$ is a vector, Eq. (1.1) is a system of 3 ordinary differential equations. differential equations. The velocity $\vec{v}_i$ of the particle i at time t is also subject to a system of differential equations, expressed most simply using the momentum $\vec{p}_i$,

$$\dot{\vec{p}}_i(t) = \vec{F}_i(t), \tag{1.2}$$

where $\vec{F}_i(t)$ is the force acting of particle i at time t. Equation (1.2) describes the temporal evolution of the momentum of the particle i. Equation (1.1) and (1.2) are each $3 \times n$ coupled ordinary differential equations. If, for example, we want to describe the movement of all molecules in a liter of water by a simulation, this is impossible due to the large number of equations and we must switch to a description using balance equations and fields.

Newton's equations of motion (1.1) and (1.2) are by their nature *basic physical principles*. They apply to atoms or planets. The nature of the force itself, $\vec{F}_{ij}$ in the equations above, depends on the nature of the physical system that we study. It is not necessarily a fundamental interaction, such as gravity, but may emerge from a complex interplay of multiple physical mechanisms. A simple example is the Lennard-Jones interaction with interaction energy

$$V_{ij} = 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{1.3}$$

and force

$$\vec{F}_{ij} = -4\varepsilon \left[12 \left(\frac{\sigma^{12}}{r_{ij}^{13}} \right) - 6 \left(\frac{\sigma^{6}}{r_{ij}^{7}} \right) \right] \hat{r}_{ij}. \tag{1.4}$$

We have written this in terms of a pair interaction and assumes that forces are pair-wise additive, meaning the total force on particle i is given by $\vec{F}_i = \sum_j \vec{F}_{ij}$. The quantity r_{ij} is the distance between the particles (here atoms or molecules)

i and j, and $\hat{r}_{ij}$ is the normal vector pointing from one to the other. The term $\propto r^{-13}$ describes the repulsion of the atoms due to the Pauli exclusion principle and the term $\propto r^{-7}$ describes the attraction of the atoms due to the London dispersion interaction (Müser et al., 2023). Both interactions are based on fundamental physical principles, but the formulation Eq. (1.4) reduces these complex phenomena to a simple constituating law. Such laws are often called *constitutive laws*. The numerical solution of Newton's equations of motion for atoms or molecules is called *molecular dynamics simulation*.

Note: The term constitutive law often appears in the context of field theories. For the Lennard-Jones potential, this term is rather unusual, but this law is nevertheless of a constitutive nature.

Another example of models with discrete elements are network models for electrical circuits. Here, an element links an electrostatic potential difference (energy difference) with a current, for example

$$i = u/R \tag{1.5}$$

describes the current i that flows through a resistor R across which the voltage drops by u. Such models are often referred to as "lumped-element models". Equation (1.5) naturally also has the quality of a *constitutive law*, as complex electronic processes are behind the individual parameter R. For a fully formulated model of an electric circuits we also need Kirchhoff's rules, that have the quality of *balance equations*. In Fig. 1.1, these models are therefore referred to as *global balance* models. "Lumped-element models" also lead to systems of ordinary differential equations, which are often solved numerically by explicit time propagation. Well-known representatives of this type of simulation software are, for example *SPICE* or *MATLAB Simulink*.

Such a global balance description is characterized by a lack of interest in local resolution. We are not interested in densities, but only in total masses, not in current densities but only in currents. This is best illustrated by the above-mentioned resistor whose contacts are at different potentials, which results in a current flow. We do not ask ourselves how the current is distributed in the resistor. We do not even ask whether the resistor is homogeneous or inhomogeneous. The model only requires the overall resistance R, essentially modeling the resistor as a black box to which we assign the value of a single parameter. This approach is discussed in detail in electrical engineering and systems theory.

1.3 Fields

However, if we now realize that our black box is only insufficiently described with one parameter, then we need to replace it with a more complex models, for example and equivalent circuit with details that resolve the internal state of the component. This in turn can be taken so far, that a continuum is created at the end - we have arrived at a *local balance* descriptions. Staying with the example of flow, we need parameters such as conductivity (or for fluids viscosity or diffusivity), which now describe the resistance to flow locally. These parameters can be obtained from experiments or *ab-initio* simulations but are required as input to (or the "parameterization" of) the local balance description.

Local balance means that we can assign density, concentration, temperature or similar quantity to each point in space. However, this means that the temporal changes in the local degrees of freedom - i.e. the momentum or velocity - are constrained by a *local, thermodynamic-equilibrium* condition. (In thermodynamic equilibrium, the momentum satisfies a Maxwell-Boltzmann distribution.) This local equilibrium does not mean that we no longer have dynamics: If we think of a swarm of gas or liquid molecules, then their individual velocities follow an equilibrium distribution function, but their mean follows the balance equation. The dynamics are therefore averaged over a huge number of these particles. Local balance also does not mean that different temperatures or densities cannot exist at different locations. The differences in these parameters are then the driving forces of the dynamics – temperature gradients, density gradients, etc.

Such models fall into the category of *field theories*, and their mathematical description is based on *partial* differential equations. (This is in contrast to the ordinary differential equations of discrete models.) A *transport theory* is a specific class of field theory that is based on the balancing mass, momentum or energy and requiring constitutive laws for the description of the material behavior. These constitutive laws contain *transport parameters* such as the viscosity or diffusion constant. There are also field theories that have the character of a basic physical principle. This is, for example, the Schrödinger equation mentioned above or the Maxwell equations of electrodynamics.

1.4 Which model is the right one?

Choosing and formulating the right model is a form of art. Just because a theory is called "quantum mechanics" (and leaves one or the other in awe at its complexity), it does not necessarily offer the solution to the problem

that we are trying to solve. Too much detail can even be a hindrance and we must constantly ask ourselves how much detail is necessary in model and simulation. We always need ask ourselves before we start a simulation: "Is a simulation of this complexity really necessary, or can I simplify the problem?" The simulation should be seen as a tool and not as an end in itself, according to the American mathematician Richard Wesley Hamming (*1915, †1998): "*The purpose of computing is insight, not numbers*".

Chapter 2

Transport theory

Context: We introduce the foundations of transport theory, in particular how to balance conserved quantity. This leads to the *continuity equation*. We start from an illustrating example, the diffusion of suspended particles.

2.1 Diffusion and drift

Diffusive transport is easily accessible via the image of the "random walk", a random stochastic movement of particles. Random motion was first described by the botanist *Robert Brown* (1773-1858) who observed random motion of grains of pollen suspended in water. He lend his name for the now common term *Brownian motion* or *Brownian molecular motion*. Robert Brown did not know about molecules at his time. He initially believed that the movement was due to active processes (the "force of life" of the pollen), but could then show that inactive matter also exhibits this random motion. Today we know that this movement is caused by thermal fluctuations, i.e. molecules that randomly hit suspended particles and push them in random directions. This explanation requires the existence of atoms and was popularized in 1905 by Albert Einstein (Einstein, 1905).

Brownian molecular motion leads to diffusive transport. Figure 2.1 shows a simple qualitative thought experiment. The configuration in Fig. 2.1a shows a localization of the "pollen" in the left half of the domain shown. Due to their random movement (shown as an example by the red line in Fig. 2.1a), some of the pollen will cross the dashed boundary line into the right half and also come back again. After a certain time, the initial state can no longer be identified and the pollen are distributed throughout the domain (Fig. 2.1b). The concentration is now constant. The pollen continue to move, but on average the same number of pollen move to the left as to the right. In the

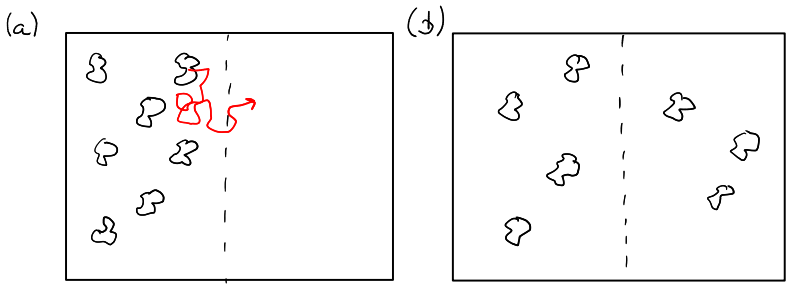

Figure 2.1: Illustration of diffusion. The "pollen" in (a) move randomly in the domain. After a certain time (b), the initial concentration difference between the left and right parts of the domain is equalized.

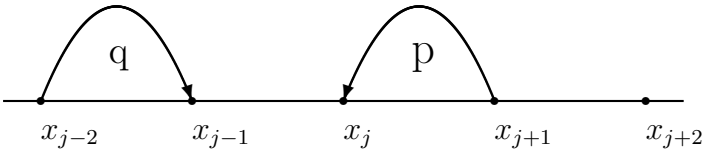

Figure 2.2: Random movement in one dimension is given by transition probabilities p (for a movement to the left) and q for a movement to the right.

case shown in Fig. 2.1a, this left/right symmetry is broken which leads to a finite flux to the right.

This thought experiment can be easily formalized mathematically. We consider a particle that performs a random movement in one dimension is performed. We start with a particle that randomly jumps back and forth on a straight line. The straight line lies along the x-direction. The particle can only move to predetermined positions on the x-axis, which we denote by x_j and which are equidistant, $x_j - x_{j-1} = \Delta x$ for $j \in \mathbb{Z}$ (see Fig. 2.2).

A particle jumps to the left with a probability p and to the right with a probability probability q to the right. In addition, we have the probability of finding a particle at time t at position x. This is given on the 1D grid by the function $P(x_j, t)$.

2.1.1 Diffusion

We first consider the case $p = q = 1/2$, i.e. that the probabilities for the jumps to the left and right are identical. We assume that the particles jump from one place to the neighboring one in a discrete, finite and constant time

step τ. Then the probability of finding a particle at time $t+\tau$ at location x is

$$P(x, t+\tau) = \frac{1}{2}P(x+\Delta x, t) + \frac{1}{2}P(x-\Delta x, t), \tag{2.1}$$

where $P(x-\Delta x, t)$ is the probability of finding a particle at position $x-\Delta x$ and $P(x+\Delta x, t)$ the probability of finding a particle at $x+\Delta x$, both at time t.

By subtracting $P(x,t)$ on both sides and dividing by τ, we obtain the following equivalent form:

$$\frac{P(x, t+\tau) - P(x,t)}{\tau} = \frac{\Delta x^2}{2\tau}\frac{P(x+\Delta x, t) - 2P(x,t) + P(x-\Delta x, t)}{\Delta x^2} \tag{2.2}$$

We can now make the limit transition to the "continuum". Taking $\tau \to 0$ and at the same time $\Delta x \to 0$ while maintaining

$$\lim_{\Delta x \to 0, \tau \to 0} \frac{\Delta x^2}{2\tau} = D \tag{2.3}$$

yields

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2}. \tag{2.4}$$

This is the well-known diffusion equation. In multiple dimensions, the second derivative becomes the Laplace operator ∇^2,

$$\frac{\partial P(x,t)}{\partial t} = D\nabla^2 P(x,t). \tag{2.5}$$

This equation is only correct if the diffusion constant is actually constant and does not vary spatially.

Note: The operator ∇ is a vector of the partial derivatives in the Cartesian direction, i.e.

$$\nabla = \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix}. \tag{2.6}$$

Applying it to a scalar function $f(x,y,z)$ yields the gradient,

$$\nabla f = \operatorname{grad} f = \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{pmatrix}. \tag{2.7}$$

The Laplacian is sometimes denoted by ∇^2 (often in the anglosaxon literature) or Δ (e.g. in the German literature). It is explicitly given by

$$\Delta = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.8}$$

We will use ∇^2 for the Laplacian throuhout this text.

2.1.2 Drift

What happens if the probabilities for the jumps to the right or left are not equal, $p \neq q$ (but of course $p + q = 1$)? We still assume discrete, uniform time steps and equidistant sampling points.

In this case, we have

$$P(x, t + \tau) = pP(x + \Delta x, t) + qP(x - \Delta x, t) \tag{2.9}$$

which yields

$$\frac{P(x, t + \tau) - P(x, t)}{\tau} = \frac{\Delta x^2}{\tau} \frac{pP(x + \Delta x, t) - P(x, t) + qP(x - \Delta x, t)}{\Delta x^2}. \tag{2.10}$$

This can be simplified by writing

$$p = \frac{1}{2} - \varepsilon \quad \text{and} \quad q = \frac{1}{2} + \varepsilon \quad \text{with} \quad 0 \leq |\varepsilon| \leq \frac{1}{2} \quad \text{or} \quad 2\varepsilon = q - p, \tag{2.11}$$

where ε now indicates how much more likely a jump to *right* is than to the left. A positive ε therefore means that the particles will move to the right on average – this is called *drift*. We can now write Eq. (2.10) using ε, giving

$$\begin{aligned} \frac{P(x, t + \tau) - P(x, t)}{\tau} =& \frac{\Delta x^2}{2\tau} \frac{P(x + \Delta x, t) - 2P(x, t) + P(x - \Delta x, t)}{\Delta x^2} \\ & - \frac{2\varepsilon \Delta x}{\tau} \frac{P(x + \Delta x, t) - P(x - \Delta x, t)}{2\Delta x}. \end{aligned} \tag{2.12}$$

In the limit $\tau \to 0$ and $\Delta x \to 0$ we require

$$\lim_{\Delta x \to 0, \tau \to 0} \frac{\Delta x^2}{2\tau} = D \quad \text{and} \quad \lim_{\Delta x \to 0, \tau \to 0} \frac{2\varepsilon \Delta x}{\tau} = v \tag{2.13}$$

and thus obtain the drift-diffusion equation

$$\frac{\partial P(x, t)}{\partial t} = \left(D \frac{\partial^2}{\partial x^2} - v \frac{\partial}{\partial x} \right) P(x, t). \tag{2.14}$$

Here, the first summand on the right-hand side again describes the diffusion process. The second summand is a drift process and v is a constant *drift* velocity. (From Eq. (2.13) and (2.14) it can be seen that the unit of v corresponds exactly to a velocity). It is the speed at which the particle moves (on average) along the x-axis.

Note: The motion of our particle was modeled using a *probability density* P. In the thermodynamic limit, i.e. for many particles (usually of the order of Avogadro's number $N_\mathrm{A} \sim 10^{23}$), this probability becomes the (mass) density ρ or the concentration (number density) c. We can therefore simply replace the probability P in the above equations with a concentration c. The reason for this is that we can write the concentration as an *ensemble* mean,

$$c(x,t) = \langle 1 \rangle (x,t), \tag{2.15}$$

where the mean value is defined as

$$\langle f(x) \rangle (x,t) = f(x) P(x,t). \tag{2.16}$$

2.2 Continuity

The equations (2.5) and (2.14) mix two concepts that we want to treat separately now: The conservation of the number of particles (continuity) and the process that leads to a flow of particles (diffusion or drift). The number of particles is conserved simply because we cannot create atoms out of nothing or destroy them into nothing. If we have a certain number of particles N_tot in our overall system, we know that this number

$$N_\mathrm{tot}(t) = \int \mathrm{d}^3 r \, c(\vec{r}, t) \tag{2.17}$$

cannot change over time: $\mathrm{d}N_\mathrm{tot}/\,\mathrm{d}t = 0$. The integral in Eq. (2.17) is carried out over the total volume of our system, essentially the physical world of the model.

For a small section of our physical world with volume V, the number of particles can change because they can flow over the walls of the sample volume (see Fig. 2.3). The change in the number of particles within V is given by

$$\dot{N}_\mathrm{V} = \frac{\partial}{\partial t} \int_V \mathrm{d}^3 r \, c(\vec{r}, t) = \int_V \mathrm{d}^3 r \, \frac{\partial c}{\partial t}. \tag{2.18}$$

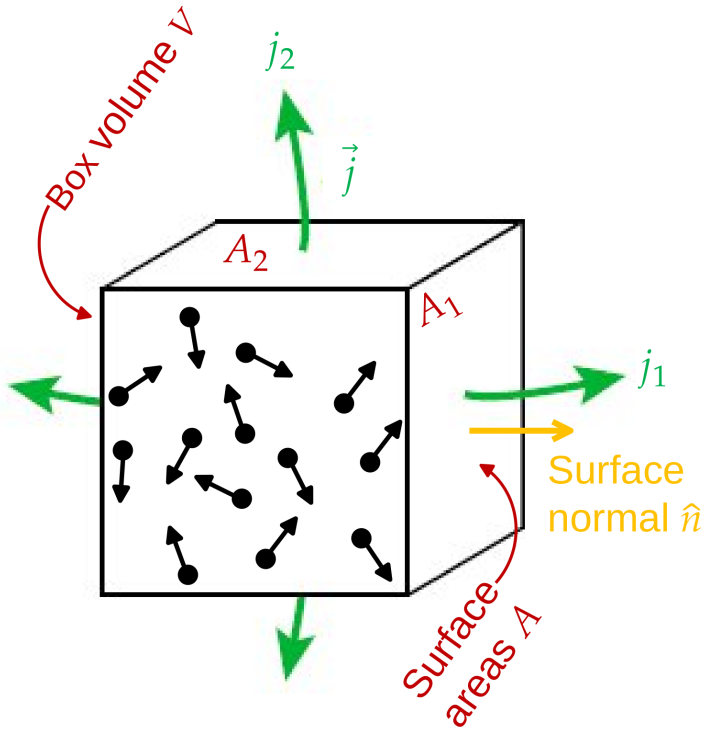

Figure 2.3: Particles can only leave the volume V through the side walls. The change in the number of particles N over a time interval τ is therefore given by the number of particles flowing through the walls. For this we need the particle flows j. The number of particles flowing through a surface is then given by $j\,A\tau$, where A is the area of the side wall.

However, the change $\dot{N}_{\mathrm{V}}$ must also be given by the number of particles flowing through the side walls. For a cube (Fig. 2.3) with six walls, it is approximately given

$$\begin{aligned}\dot{N}_{\mathrm{V}} = &- j_{\mathrm{right}} A_{\mathrm{right}} - j_{\mathrm{left}} A_{\mathrm{left}} \\ &- j_{\mathrm{above}} A_{\mathrm{above}} - j_{\mathrm{bottom}} A_{\mathrm{bottom}} \\ &- j_{\mathrm{front}} A_{\mathrm{front}} - j_{\mathrm{back}} A_{\mathrm{back}}\end{aligned} \tag{2.19}$$

if the walls are small enough so that j is almost constant over A. We have, in passing, introduced the current density j has the unit number of particles/time/area. The quantities jA are hence the number of particles flowing per unit time through one of the walls with are A.

The scalar current density j describes the current flowing out of the surface. For a general vectorial current density $\vec{j}$, which indicates the strength and direction of the particle current, the total current density flowing out of the

volume through wall i is given by $j_i = \vec{j}_i \cdot \hat{n}_i$, where $\hat{n}_i$ is the normal vector pointing outwards on wall i. The current through the wall is therefore only the component of $\vec{j}$ that is parallel to the surface normal (or perpendicular to the wall). With this argument, we can generalize the expression for the change in number of particles to

$$\dot{N}_\mathrm{V} = -\int_{\partial V} \mathrm{d}^2 r\, \vec{j}(\vec{r}) \cdot \hat{n}(\vec{r}) \tag{2.20}$$

where ∂V denotes the surface area of the volume V. This equation explicitly indicates that both the flux $\vec{j}$ and the surface normal $\hat{n}$ depend on the position $\vec{r}$ on the surface.

Alternatively, we can also group the change in the number of particles Eq. (2.19) as follows:

$$\begin{aligned} \dot{N}_\mathrm{V} = &-(j_\mathrm{right} + j_\mathrm{left}) A_\mathrm{right/left} \\ &-(j_\mathrm{top} + j_\mathrm{bottom}) A_\mathrm{top/bottom} \\ &-(j_\mathrm{front} + j_\mathrm{rear}) A_\mathrm{front/back} \end{aligned} \tag{2.21}$$

Here we have used the fact that $A_\mathrm{right} = A_\mathrm{left} \equiv A_\mathrm{right/left}$. But now

$$\begin{aligned} j_\mathrm{right} &= \hat{x} \cdot \vec{j}(x + \Delta x/2, y, z) = j_x(x + \Delta x/2, y, z) \quad \text{and} \\ j_\mathrm{left} &= -\hat{x} \cdot \vec{j}(x - \delta x/2, y, z) = -j_x(x - \delta x/2, y, z) \end{aligned} \tag{2.22}$$

since $\hat{n} = \hat{x}$ for the right wall but $\hat{n} = -\hat{x}$ for the left wall. Here, $\hat{x}$ is the normal vector along the x-axis of the coordinate system. The sign of the surface normal is therefore reversed between the right and left surfaces. The same applies to the top/bottom and front/back walls. We can further rewrite this equation as

$$\begin{aligned} \dot{N} = &-\frac{j_x(x + \Delta x/2, y, z) - j_x(x - \Delta x/2, y, z)}{\Delta x} V \\ &-\frac{j_y(x, y + \Delta y/2, z) - j_y(x, y - \Delta y/2, z)}{\Delta y} V \\ &-\frac{j_z(x, y, z + \Delta z/2) - j_z(x, y, z - \Delta z/2)}{\Delta z} V, \end{aligned} \tag{2.23}$$

since $V = A_\mathrm{right/left} \Delta x = A_\mathrm{top/bottom} \Delta y = A_\mathrm{front/back} \Delta z$. However, the factors in front of the volume V in Eq. (2.23) are now exactly the difference quotients of the flows j_i, in the x, y and z directions respectively. For small volumes (and small Δx, etc.) this becomes

$$\dot{N} = -\int_V \mathrm{d}^3 r\, \nabla \cdot \vec{j}(\vec{r}). \tag{2.24}$$

We have just heuristically derived the divergence theorem (see also Eq. (2.26)) to express Eq. (2.20) as a volume integral.

Note: We have expressed the *divergence* of a vectorial field $\vec{f}(\vec{r})$ through the nabla operator,

$$\nabla \cdot \vec{f} = \operatorname{div} \vec{f} = \frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_z}{\partial z} \tag{2.25}$$

The *divergence theorem* is an important result of vector calculus. It converts an integral over a volume V into an integral over the surface ∂V of this volume. For a vector field $\vec{f}(\vec{r})$ applies:

$$\int_V \mathrm{d}^3 r \, \nabla \cdot \vec{f}(\vec{r}) = \int_{\partial V} \mathrm{d}^2 r \, \vec{f}(\vec{r}) \cdot \hat{n}(\vec{r}) \tag{2.26}$$

Here $\hat{n}(\vec{r})$ is the normal vector which points outwards on the edge ∂V of the volume V. Note that in one dimension this reduces to

$$\int_a^b \mathrm{d}x \, \frac{\partial f}{\partial x} = f(b) - f(a) \tag{2.27}$$

while is the integration rule we all know from high school. The divergence theorem is hence a generalization of this integration rule to functions of many variables.

Equation (2.18) and (2.24) together result in

$$\int_V \mathrm{d}^3 r \, \left\{ \frac{\partial c}{\partial t} + \nabla \cdot \vec{j} \right\} = 0. \tag{2.28}$$

Since this applies to any volume V, the equation

$$\frac{\partial c}{\partial t} + \nabla \cdot \vec{j} = 0 \tag{2.29}$$

must also hold. This equation is called *continuity equation*. It describes the conservation of the number of particles or the mass of the system.

Note: In the derivation presented here, we have already implicitly used the *strong* formulation and a *weak* formulation of a differential equation. Equation (2.29) is the strong formulation of the continuity equation. This

requires that the differential equation is satisfied for every spatial point $\vec{r}$. A corresponding weak formulation is Eq. (2.28). Here it is only required that the equation is fulfilled in a kind of mean value, here as an integral over a sample volume V. Within the volume, the strong form need not be satisfied, but the integral over these deviations (which we will later call "residuum") must vanish. The weak formulation is thus an approximation for finite sample volumes V. In many numerical approaches, a weak equation is solved exactly for a certain (approximate) initial function.

We can still require that "particles" are produced within our sample volume. In the current interpretation of the equation, this could be, for example, chemical reactions that convert one type of particle into another. An identical equation applies to heat transport, because just like particle numbers, also the energy is a conserved quantity. Here, a source term would be the production of heat, e.g. by a heating element. Given a flow Q (with unit number of particles/time/volume), the particle or heat source, the continuity equation can be extended to

$$\frac{\partial c}{\partial t} + \nabla \cdot \vec{j} = Q. \tag{2.30}$$

The continuity equation with source term is also sometimes referred to as the *balance equation*.

Note: Equation (2.30) describes the change in concentration c over time. A related question is how to solve this equation after a very long time - when a dynamic equilibrium has been reached and the concentration no longer varies but is *stationary*. This equilibrium is then characterized by the fact that $\partial c/\partial t = 0$. The equation

$$\nabla \cdot \vec{j} = Q \tag{2.31}$$

is the *stationary* variant of the continuity equation.

2.2.1 Drift

Let us come back to transport processes, first to drift. If all particles in our sample volume move with the speed $\vec{v}$, this leads to a particle flow

$$\vec{j}_{\text{drift}} = c\vec{v}. \tag{2.32}$$

When inserted into the continuity equation (2.29), this results in the drift contribution to the drift-diffusion equation (2.14).

2.2.2 Diffusion

From our thought experiment above, it is clear that the diffusion current must always go in the direction of the low concentration, i.e. in the opposite direction to the gradient ∇c of the concentration. The corresponding current is given by

$$\vec{j}_{\text{Diffusion}} = -D\nabla c. \tag{2.33}$$

When inserted into the continuity equation (2.29), this results in the diffusion equation (2.5).

The entire drift-diffusion equation therefore has the form

$$\frac{\partial c}{\partial t} + \nabla \cdot \{-D\nabla c + c\vec{v}\} = 0. \tag{2.34}$$

In contrast to equations (2.5) and (2.14), this equation also applies if the diffusion constant D or drift velocity $\vec{v}$ varies spatially.

Note: We have introduced transport theory here in terms of a particle concentration c. However, the continuity equation generally describes the *conservation* of a certain quantity, in our case the number of particles (or equivalently the mass). Other physically conserved quantities are momentum and energy. The continuity equation for the momentum leads to the Navier-Stokes equation. The continuity equation for the energy leads to the heat conduction equation.

Chapter 3

Numerical solution

Context: We will now put the transportation problem aside for a while and devote ourselves to the *numerical* solution of differential equations. This chapter shows the basics of the numerical analysis of differential equations and introduces a few important concepts, in particular the series expansion and the residual. The presentation here follows chapter 1 from Boyd (2000).

3.1 Series expansion

In abstract notation, we are looking for unknown functions $u(x, y, z, ...)$ that solve a set of differential equations

$$\mathcal{L}u(x, y, z, \ldots) = f(x, y, z, \ldots) \tag{3.1}$$

must be fulfilled. Here, $\mathcal{L}$ is a (not necessarily linear) operator that contains the differential (or integral) operations. We now introduce an important concept for the (numerical) solution of the differential equation: We approximate the function u by a truncated *series expansion* of N terms. We write

$$u_N(x, y, z, \ldots) = \sum_{n=1}^{N} a_n \varphi_n(x, y, z, \ldots) \tag{3.2}$$

where the φ_n are called "basis functions". We will discuss the properties of these basis functions in more detail in the next chapter.

We can now write the differential equation as,

$$\mathcal{L}u_N(x, y, z, \ldots) = f(x, y, z, \ldots). \tag{3.3}$$

This representation means that we have now replaced the question of the unknown function u with the question of the unknown coefficients a_n. We only have to let the differential operator $\mathcal{L}$ act on the (known) basis functions φ_n and we can calculate this analytically.

What remains is to determine the coefficients a_n. These coefficients are numbers, and these numbers can be calculated by a computer. Equation (3.2) is of course an approximation. For certain basis functions, it can be shown that these are "complete" and can therefore represent certain classes of functions exactly. However, this is only true under the condition that the series Eq. (3.2) is extended to $N \to \infty$. For all practical applications (such as implementations in computer code), however, this series expansion must be aborted. A "good" series expansion approximates the exact solution already at low N with a small error. With this statement, we would of course have to specify how we want to quantify errors. Numerically, we then search for the exact coefficients a_n that minimize the error. The choice of a good basis function is non-trivial.

3.2 Residual

An important concept is that of the *residual*. Our goal is to solve Eq. (3.1). The exact solution would be $\mathcal{L}u - f \equiv 0$. However, since we can only construct an approximate solution, this condition will not be fulfilled exactly. We define the residual as exactly this deviation from the exact solution, namely

$$R(x, y, z, \ldots; a_0, a_1, \ldots, a_N) = \mathcal{L}u_N(x, y, z, \ldots) - f(x, y, z, \ldots). \tag{3.4}$$

The residual is therefore a kind of measure for the error we make. The strategy for numerically solving the differential equation Eq. (3.1) is now to determine the coefficients a_n in such a way that the residual Eq. (3.4) is minimal. We have thus mapped the solution of the differential equation to an optimization problem. The different numerical methods, which we will discuss in the next chapters, are mainly determined by the specific optimization strategy.

Note: Numerical methods for *optimization* are a central core of the numerical solution of differential equations and thus of simulation techniques. There are countless optimization methods that work better or worse in different situations. We will first treat such optimizers as "black boxes". At the end of the course, we will return to the question of optimization and discuss some well-known optimization methods. The term *minimization method* is often used synonymously with optimization methods. A good

overview of optimization methods can be found in the book by Nocedal and Wright (2006).

3.3 A first example

We now want to concretize these abstract ideas using an example and introduce a few important terms. Let's look at the one-dimensional boundary value problem,

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} - (x^6 + 3x^2)u = 0, \tag{3.5}$$

with the boundary conditions $u(-1) = u(1) = 1$. (I.e. $x \in [-1, 1]$ is the domain on which we are looking for the solution.) In this case, the abstract differential operator $\mathcal{L}$ takes the concrete form

$$\mathcal{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - (x^6 + 3x^2) \tag{3.6}$$

is given. The exact solution to this problem is given by

$$u(x) = \exp\left[(x^4 - 1)/4\right]. \tag{3.7}$$

We now guess an approximate solution as a series expansion for this equation. This approximate solution should already fulfill the boundary conditions. The equation

$$u_2(x) = 1 + (1 - x^2)(a_0 + a_1 x + a_2 x^2) \tag{3.8}$$

is constructed in such a way that the boundary conditions are fulfilled. We can express these as

$$u_2(x) = 1 + a_0(1 - x^2) + a_1 x(1 - x^2) + a_2 x^2(1 - x^2) \tag{3.9}$$

to exponentiate the basis functions $\varphi_i(x)$. Here $\varphi_0(x) = 1 - x^2$, $\varphi_1(x) = x(1 - x^2)$ and $\varphi_2(x) = x^2(1 - x^2)$. Since these basis functions are non-zero on the entire domain $[-1, 1]$, this basis is called a *spectral* basis. (Mathematically: The carrier of the function corresponds to the domain.)

In the next step, we must find the residual

$$R(x; a_0, a_1, a_2) = \frac{\mathrm{d}^2 u_2}{\mathrm{d}x^2} - (x^6 + 3x^2)u_2 \tag{3.10}$$

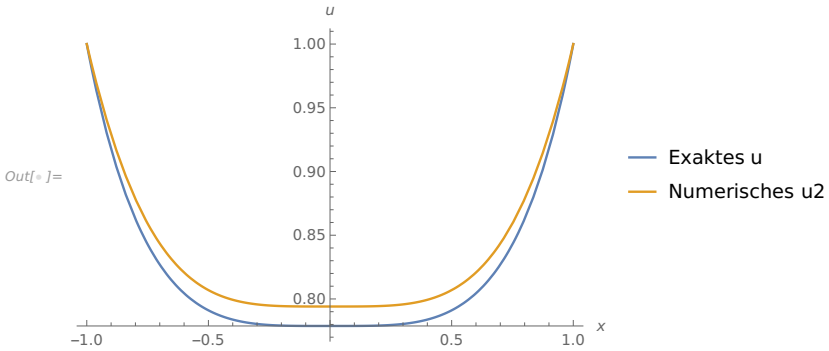

Figure 3.1: Analytical solution $u(x)$ and "numerical" approximate solution $u_2(x)$ of the GDGL (3.5).

minimize. For this we choose a strategy called *collocation*: We require that the residual vanishes exactly at three selected points:

$$R(x_i; a_0, a_1, a_2) = 0 \quad \text{for} \quad x_0 = -1/2, x_1 = 0 \text{ und } x_2 = 1/2. \tag{3.11}$$

Note: The disappearance of the residual at x_i does not mean that $u_2(x_i) \equiv u(x_i)$, i.e. that at x_i our approximate solution corresponds to the exact solution. We are still restricted to a limited set of functions, namely the functions covered by Eq. (3.8).

From the collocation condition we now get a linear system of equations with three unknowns:

$$R(x_0; a_0, a_1, a_2) \equiv -\frac{659}{256}a_0 + \frac{1683}{512}a_1 - \frac{1171}{1024}a_2 - \frac{49}{64} = 0 \tag{3.12}$$

$$R(x_1; a_0, a_1, a_2) \equiv -2(a_0 - a_2) = 0 \tag{3.13}$$

$$R(x_2; a_0, a_1, a_2) \equiv -\frac{659}{256}a_0 - \frac{1683}{512}a_1 - \frac{1171}{1024}a_2 - \frac{49}{64} = 0 \tag{3.14}$$

$$\tag{3.15}$$

The solution of these equations results in

$$a_0 = -\frac{784}{3807}, \quad a_1 = 0 \quad \text{and} \quad a_2 = a_0. \tag{3.16}$$

Figure 3.1 shows the "numerical" solution $u_2(x)$ in comparison with the exact solution $u(x)$.

In the numerical example shown here, both the basis functions and the strategy for minimizing the residual can be varied. In the course of this

lecture, we will establish the finite elements as basis functions and use the Galerkin method as minimization strategy. To do this, we must first discuss properties of possible basis functions.

Note: The example shown here is a simple case of *discretization*. We have gone from a continuous function to the discrete coefficients a_0, a_1, a_2.

Chapter 4

Function spaces

Context: Before we dive deeper into the numerical solution of partial differential equations, we need to introduce a slightly abstract concept here: The concept of *function spaces*, or more concretely *Hilbert spaces*. Function spaces are useful because they formalize the series expansion and provide easy access to the coefficients of a series expansion through the concept of basis functions.

4.1 Vectors

As an introduction, let us recall the usual Cartesian vectors. We can represent a vector $\vec{a} = (a_1, a_2, a_3)$ as a linear combination of basis vectors $\hat{e}_1$, $\hat{e}_2$ and $\hat{e}_3$,

$$\vec{a} = a_1\hat{e}_1 + a_2\hat{e}_2 + a_3\hat{e}_3. \tag{4.1}$$

The unit vectors $\hat{e}_1$, $\hat{e}_2$ and $\hat{e}_3$ are of course the vectors that span the Cartesian coordinate system. (In previous chapters, we also use the notation $\hat{x} \equiv \hat{e}_1$, $\hat{y} \equiv \hat{e}_2$ and $\hat{z} \equiv \hat{e}_3$.) The numbers a_1, a_2 and a_3 are the components or *coordinates* of the vector, but also the coefficients multiplying the unit vectors in Eq. (4.1). In this sense, they are identical to the coefficients of the series expansion, with the difference that the $\hat{e}_i$s are orthogonal, i.e.

$$\hat{e}_i \cdot \hat{e}_j = \delta_{ij} \tag{4.2}$$

where δ_{ij} is the Kronecker-δ. Two Cartesian vectors $\vec{a}$ and $\vec{b}$ are orthogonal if the scalar product between them vanishes:

$$\vec{a} \cdot \vec{b} = \sum_i a_i^* b_i = 0 \tag{4.3}$$

Using the scalar product, we can obtain the components as $a_i = \vec{a} \cdot \hat{e}_i$. This is a direct consequence of the orthogonality of the basis vectors $\hat{e}_i$.

4.2 Functions

In the previous section, we claimed that the basis functions from Chapter 5 are not orthogonal. For this we need an idea for orthogonality of functions. With a definition of a scalar product between two *functions*, we can then define orthogonality as the vanishing of this scalar product.

We now introduce a scalar product on functions (pr function spaces). Given two functions $g(x)$ and $f(x)$ on the interval $x \in [a, b]$, *define* the scalar product as

$$(f, g) = \int_a^b \mathrm{d}x\, f^*(x) g(x), \tag{4.4}$$

where $f^*(x)$ is the complex conjugate of $f(x)$. This scalar product pr *inner* product is a map to a real number with the properties

- Positive definite: $(f, f) \geq 0$ and $(f, f) = 0 \Leftrightarrow f = 0$
- Sesquilinear: $(\alpha f + \beta g, h) = \alpha^*(f, h) + \beta^*(g, h)$ and $(f, \alpha g + \beta h) = \alpha(f, g) + \beta(f, h)$
- Hermitian: $(f, g) = (g, f)^*$

The scalar products Eq. (4.3) and (4.4) both fulfill these properties.

Note: The scalar product between two functions can be defined more generally with a weight function $w(x)$,

$$(f, g) = \int_a^b \mathrm{d}x\, f^*(x) g(x) w(x). \tag{4.5}$$

The question of orthogonality between functions can thus only be answered with respect to a certain definition of the scalar product. For example, *Chebyshev polynomials* are othogonal with respect to a scale product with weight function $w(x) = (1 - x^2)^{-1/2}$. Within this class, we will only use the case $w(x) = 1$.

4.3 Basis functions

Let us now return to the series expansion,

$$f_N(x) = \sum_{n=1}^{N} a_n \varphi_n(x). \tag{4.6}$$

The functions $\varphi_i(x)$ are called *basis functions*. A necessary property of the basis functions is their linear independence. The functions are linearly independent if none of the basis functions themselves can be written as a linear combination, i.e. in the form of the series expansion Eq. (4.6), of the other basis functions. This means that it must be fulfilled that

$$\sum_{n=1}^{N} a_n \varphi_n(x) = 0 \tag{4.7}$$

if and only if all $a_n = 0$. Linearly independent elements form a basis.

This basis is called complete if all relevant functions (= elements of the underlying vector space) can be mapped by the series expansion (4.6). (Proofs of the completeness of basis functions are complex and outside the focus of this class.) The coefficients a_n are called coordinates or coefficients. The number of basis functions or coordinates N is called the *dimension* of the vector space.

Note: A *vector space* is a set on which the operations of addition and scalar multiplication are defined with the usual properties, such as the existence of neutral and inverse elements and associative, commutative and distributive laws. If this space is defined on functions, it is also referred to as a *function space*. If there is also a scalar product such as Eq. (4.4), then we speak of a *Hilbert space*.

4.3.1 Orthogonality

Particularly useful basis functions are orthogonal. Using the scalar product, we can now define orthogonality for these functions. Two functions f and g are orthogonal if the scalar product vanishes, $(f, g) = 0$. A set of mutually orthogonal basis functions satisfies

$$(\varphi_n, \varphi_m) = \nu_n \delta_{nm}, \tag{4.8}$$

where δ_{nm} is the Kronecker-δ. For $\nu_n \equiv (\varphi_n, \varphi_n) = 1$ the basis is called *orthonormal*.

Orthogonality is useful because it shows us a way to obtain the coefficients of the series expansion (4.6):

$$(\varphi_n, f_N) = \sum_{i=1}^{N} a_i (\varphi_n, \varphi_i) = \sum_{i=1}^{N} a_i \nu_i \delta_{ni} = a_n \nu_n \tag{4.9}$$

which yields the coefficients as

$$a_n = \frac{(\varphi_n, f_N)}{(\varphi_n, \varphi_n)}. \tag{4.10}$$

The coefficients are given by the projection (the scalar product) of the function onto the basis vectors. Remember that the following also applies to Cartesian vectors: $a_n = \vec{a} \cdot \hat{e}_n$. (The normalization factor is omitted here because $\hat{e}_n \cdot \hat{e}_n = 1$, i.e. the Cartesian basis vectors are orthonormal.) The coefficient given by Eq. (4.10) can be thought of as coordinates of the function, similar to the coordinates in Cartesian space.

Note: A useful identity for expansion into orthogonal bases is *Parseval's theorem*. Because scalar products between different basis functions vanish, the square (or power) of a series expansion is given by

$$(f_N, f_N) = \sum_{n=1}^{N} |a_n|^2 \nu_n, \tag{4.11}$$

or for orthogonal basis functions

$$(f_N, f_N) = \sum_{n=1}^{N} |a_n|^2. \tag{4.12}$$

4.3.2 Fourier basis

A famous and important set of basis functions is the *Fourier basis*,

$$\varphi_n(x) = \exp\left(i q_n x\right), \tag{4.13}$$

on the interval $x \in [0, L]$ with $q_n = 2\pi n/L$ and $n \in \mathbb{Z}$. The Fourier basis is periodic on this interval and is shown in Fig. 4.1. It can easily be shown that

$$(\varphi_n, \varphi_m) = L\delta_{nm}, \tag{4.14}$$

so that the Fourier basis is orthogonal. The coefficients a_n of the Fourier series,

$$f_\infty(x) = \sum_{n=-\infty}^{\infty} a_n \varphi_n(x), \tag{4.15}$$

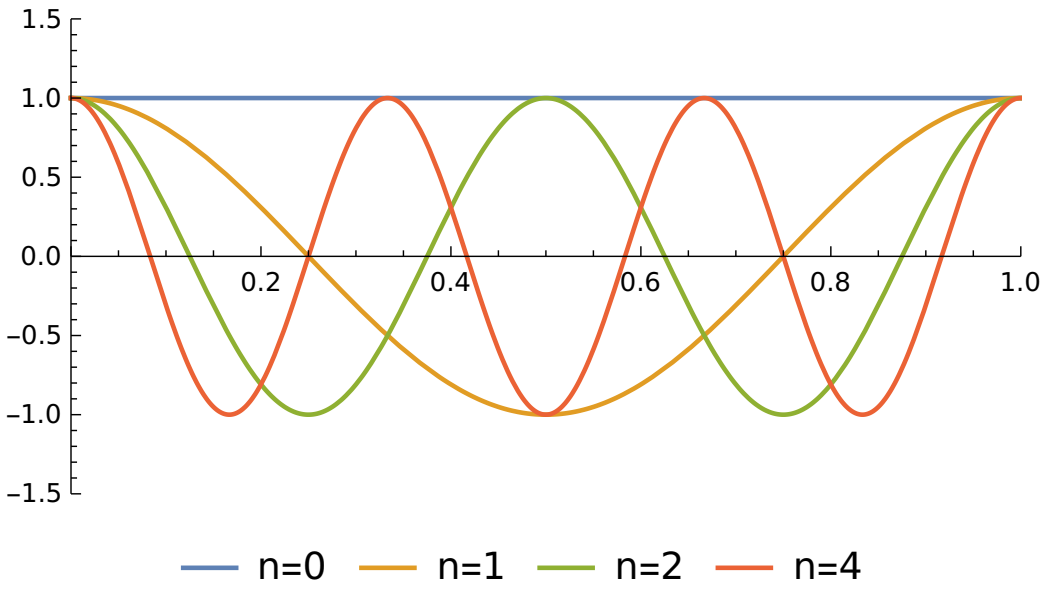

Figure 4.1: Real part of the Fourier basis functions, Eq. (4.13), for $n = 1, 2, 3, 4$. The higher order basis functions oscillate with a smaller period and represent higher frequencies

.

can thus be obtained as

$$a_n = \frac{1}{L}(\varphi_n, f_\infty) = \frac{1}{L}\int_0^L \mathrm{d}x\, f_\infty(x) \exp\left(-iq_n x\right). \tag{4.16}$$

This is the well-known formula for the coefficients of the Fourier series.

Note: Conceptually, the Fourier basis describes different frequency components, while the basis of the finite elements described in the next section describes spatial localization.

For real-valued function with $f_\infty(x) \equiv f_\infty^*(x)$, we get

$$\sum_{n=-\infty}^{\infty} a_n \varphi_n(x) \equiv \sum_{n=-\infty}^{\infty} a_n^* \varphi_{-n}(x) \tag{4.17}$$

because $\varphi_n^*(x) = \varphi_{-n}(x)$. This means $a_n = a_{-n}^*$ is a necessary condition to obtain a real-valued $f(x)$. This has implications for truncating the Fourier

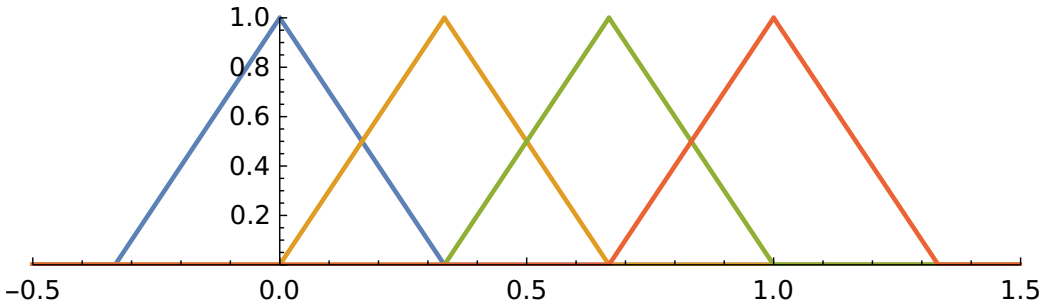

Figure 4.2: The base of the finite elements in its simplest, linear incarnation. Each basis function is a "marquee" that runs over a certain interval between 0 and 1 and back again, see also Eq. (4.19).

series to a finite number of terms N. In particular, we need to truncate symmetrically, i.e.

$$f_N(x) = \sum_{n=-(N-1)/2}^{(N-1)/2} a_n \exp(iq_n x) \tag{4.18}$$

with odd N.

4.3.3 Finite elements

Another basis set that is important for numerical analysis is the finite-element basis. In contrast to the Fourier basis, which only becomes zero at isolated points in the entire domain, the finite element basis is localized in space and is zero for large areas of the domain. It thus divides the domain into spatial sections.

In its simplest form, the basis consists of localized piece-wise linear functions, the "tent" functions,

$$\varphi_n(x) = \begin{cases} \frac{x-x_{n-1}}{x_n-x_{n-1}} & \text{for } x \in [x_{n-1}, x_n] \\ \frac{x_{n+1}-x}{x_{n+1}-x_n} & \text{for } x \in [x_n, x_{n+1}] \\ 0 & \text{else} \end{cases} \tag{4.19}$$

Here, the x_n are the *nodes* (also known as grid points) between which the tents are spanned. The functions are constructed in such a way that the maximum value is 1 and $\int_0^L \mathrm{d}x\, \varphi_n(x) = (x_{n+1} - x_{n-1})/2$. This basis is the simplest form of the finite element basis and is shown in Fig. 4.2. Higher order polynomials can be used for greater accuracy.

An important note at this point is that the basis of the finite elements *not* is orthogonal. In our one-dimensional case, the scalar product does not vanish for the nearest neighbors. This is the case because two neighbors each have an overlapping rising and falling edge. One obtains

$$M_{nn} \equiv (\varphi_n, \varphi_n) = \frac{1}{3}(x_{n+1} - x_{n-1}) \tag{4.20}$$

$$M_{n,n+1} \equiv (\varphi_n, \varphi_{n+1}) = \frac{1}{6}(x_{n+1} - x_n) \tag{4.21}$$

$$M_{nm} \equiv (\varphi_n, \varphi_m) = 0 \quad \text{for } |n - m| > 1 \tag{4.22}$$

for the scalar products.

Nevertheless, we can use these relations to determine the coefficients of a series expansion,

$$f_N(x) = \sum_{n=0}^{N-1} a_n \varphi_n(x), \tag{4.23}$$

which yields

$$\begin{aligned}(\varphi_n, f_N(x)) &= a_{n-1}(\varphi_n, \varphi_{n-1}) + a_n(\varphi_n, \varphi_n) + a_{n+1}(\varphi_n, \varphi_{n+1}) \\ &= M_{n,n-1}a_{n-1} + M_{nn}a_n + M_{n,n+1}a_{n+1}.\end{aligned} \tag{4.24}$$

We can express this as

$$(\varphi_n, f_N(x)) = [\underline{M} \cdot \vec{a}]_n \tag{4.25}$$

where $[\vec{v}]_n = v_n$ denotes the nth component of the vector enclosed by the two square brackets $[\cdot]_n$. The matrix $\underline{M}$ is *sparse*. For an orthogonal basis, such as the Fourier basis of section 4.3.2, this matrix is diagonal. For a basis with identical distances $x_{n+1} - x_n = 1$ of the grid points x_n, the matrix has the following form

$$\underline{M} = \begin{pmatrix} 2/3 & 1/6 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1/6 & 2/3 & 1/6 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1/6 & 2/3 & 1/6 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1/6 & 2/3 & 1/6 & 0 & \cdots & \\ 0 & 0 & 0 & 1/6 & 2/3 & 1/6 & \cdots & \\ 0 & 0 & 0 & 0 & 1/6 & 2/3 & \cdots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{4.26}$$

To find the coefficients a_n, we must solve a (sparse) linear system of equations. The matrix $\underline{M}$ is also called the *mass matrix*.

Note: Basis sets that are different from zero only at individual points are called *spectral* basis sets. In particular, the Fourier basis is a spectral basis set for periodic functions. The *orthogonal polynomials* are important spectral basis sets that are also used in numerical analysis. For example, Chebyshev polynomials are good basis sets for non-periodic functions defined on closed intervals. The finite-element basis is not a spectral basis.

Chapter 5

Approximation and interpolation

Context: We now apply the idea of basis functions to approximate functions. To do this, we return to the concept of the residual. The goal of function approximation is that the approximated function minimizes the residual. Building on these ideas, we will then discuss the approximation of differential equations in the next chapter.

5.1 Residual

In the previous section, we described how a series expansion can be constructed using basis functions. A typical series expansion contains a finite number of elements N and has the form

$$f_N(x) = \sum_{n=1}^{N} a_n \varphi_n(x), \tag{5.1}$$

where the $\varphi_n(x)$ are the basis functions introduced in the previous chapter.

We now want to approach the question of how we can approximate an arbitrary function $f(x)$ via such a basis function expansion. We define the residual

$$R(x) = f_N(x) - f(x), \tag{5.2}$$

which vanishes at every point x if $f_N(x) \equiv f(x)$. For an approximation we want to "minimize" this residual. (Minimizing in this context means to bring it as close to zero as possible.) We are looking for the coefficients a_n of the series, which approximate the function $f(x)$ in the sense of minimizing the residual.

At this point, it should be noted that the basis functions must be defined on the same support as the target function $f(x)$. For the approximation of a periodic function $f(x)$ we need a periodic basis.

5.2 Collocation

The first minimization strategy introduced here is *collocation*. This method requires that the residual disappears at selected collocation points y_n,

$$R(y_n) = 0 \quad \text{or} \quad f_N(y_n) = f(y_n). \tag{5.3}$$

The number of collocation points must correspond to the number of coefficients in the series expansion. The choice of ideal collocation points y_n itself is non-trivial, and we will only discuss specific cases here.

As a first example, we discuss an expansion into N finite elements. As collocation points we choose the interpolation points of the basis, $y_n = x_n$. At these sampling points, only one of the basis functions is non-zero, $\varphi_n(y_n) = 1$ and $\varphi_n(y_k) = 0$ if $n \neq k$. This means that the condition

$$R(y_n) = 0 \tag{5.4}$$

trivially leads to

$$a_n = f(y_n). \tag{5.5}$$

The coefficients a_n are therefore the function values at the collocation points. The approximation is a piece-wise linear function between the function values of $f(x)$.

As a second example, we discuss a Fourier series with corresponding N Fourier basis functions,

$$\varphi_n(x) = \exp\left(iq_n x\right). \tag{5.6}$$

In the context of a collocation method, we require that the residual vanishes on N equidistant points, $R(y_n) = 0$ with

$$y_n = nL/N, \tag{5.7}$$

where L/N is the grid spacing. The collocation condition is

$$\sum_{k=-(N-1)/2}^{(N-1)/2} a_k \exp\left(iq_k y_n\right) = \sum_{k=-(N-1)/2}^{(N-1)/2} a_k \exp\left(i2\pi \frac{kn}{N}\right) = f(y_n). \tag{5.8}$$

Equations (5.8) can now be solved for a_k. We use the fact that for equidistant collocation points the Fourier matrix

$$W_{kn} = \exp(i2\pi kn/N) = \left[\exp(i2\pi/N)\right]^{kn} \tag{5.9}$$

is unitary (except for a constant factor), i.e. its inverse is given by the adjoint:

$$\sum_{n=0}^{N-1} W_{kn} W_{nl}^* = \sum_{n=0}^{N-1} \left[\exp(i2\pi/N)\right]^{n(k-l)} = N\delta_{kl} \tag{5.10}$$

We can therefore multiply Eq. (5.8) by W_{nl}^* and sum over n. This results in

$$\sum_n \sum_k W_{kn} W_{nl}^* a_k = \sum_k N a_k \delta_{kl} = N a_l. \tag{5.11}$$

This means that the coefficients can be expressed as

$$a_l = \frac{1}{N} \sum_{n=0}^{N} f\left(\frac{nL}{N}\right) \exp\left(-i2\pi \frac{ln}{N}\right) = \frac{1}{N} \sum_{n=0}^{N} f(y_n) \exp\left(-iq_l y_n\right) \tag{5.12}$$

for $-(N-1) \leq l \leq N-1$. This is the *discrete Fourier transform (DFT)* of the function $f(y_n)$ discretized on the collocation points.

As a simple example, we show the approximation of the example function $f(x) = \sin(2\pi x)^3 + \cos(6\pi(x^2 - 1/2))$ using the Fourier basis and finite elements. Figure 5.2 shows this approximation for $2N+1 = 5$ and $2N+1 = 11$ basis functions with equidistant collocation points.

The figure shows that all approximations run exactly through the collocation points, as required by the collocation condition. The two approaches interpolate differently between the collocation points. The finite elements lead to a linear interpolation between the points. The Fourier basis is more complicated. The curve between the collocation points is called *Fourier interpolation*.

5.3 Weighted residuals

We would now like to generalize the collocation method. To do this, we introduce the concept of a *test function*. Instead of requiring that the residual vanishes at individual points, we require that the scalar product

$$(v, R) = 0 \tag{5.13}$$

with some function $v(x)$ disappears. If Eq. (5.13) vanishes for any test function $v(x)$, then the "weak" formulation Eq. (5.13) is identical to the strong formulation $R(x) = 0$. Equation (5.13) is called a "weak" formulation because the condition is only fulfilled in the integral sense. In particular, it is shown later that this weak formulation leads to a weak *solution*, which cannot

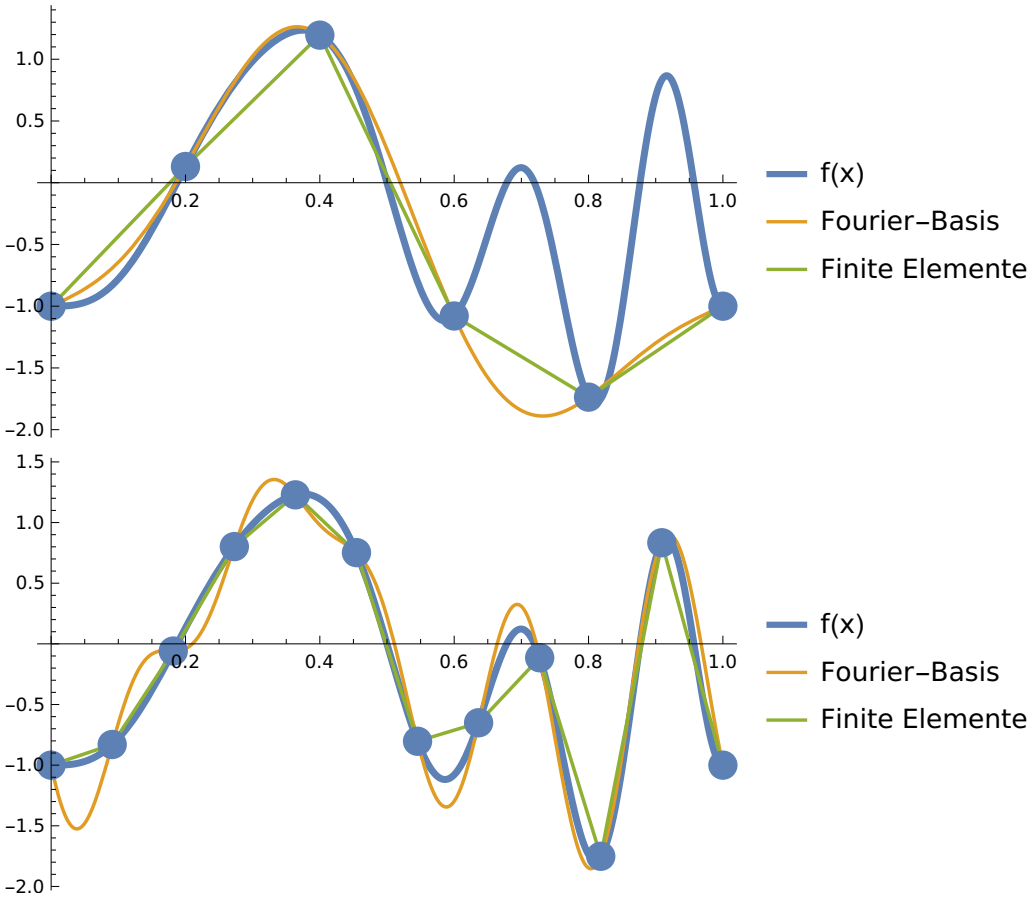

Figure 5.1: Approximation of the periodic function $f(x) = \sin(2\pi x)^3 + \cos(6\pi(x^2 - 1/2))$ on the interval $[0, 1]$ with a Fourier basis and finite elements. The function was approximated with 5 (top) and 11 (bottom) basis functions using the collocation method. The round dots show the collocation points. Both approximations run exactly through these collocation points. (The right collocation point is identical to the left one due to the periodicity). The approximation with $N = 5$ basis functions does not capture the two right oscillations of the target function $f(x)$ in both cases

.

satisfy the original (strong) PDGL at every point. The condition (5.13) is often called a *weighted residual*.

A special set of test functions leads directly to the collocation method. We choose the set of N test functions

$$v_n(x) = \delta(x - y_n) \tag{5.14}$$

where $\delta(x)$ is the Dirac δ function and y_n the collocation points. The condition $(v_n, R) = 0$ for all $n \in [0, N-1]$ leads directly to the collocation condition $R(y_x) = 0$.

Note: The Dirac δ function should be familiar from lectures on signal processing. The most important property of this function is the filter property,

$$\int_{-\infty}^{\infty} \mathrm{d}x\, f(x)\delta(x - x_0) = f(x_0), \tag{5.15}$$

i.e. the integral over the product of the δ function gives the function value at which the argument of the δ function disappears. All other properties follow from this, e.g.

$$\int \mathrm{d}x\, \delta(x) = \Theta(x), \tag{5.16}$$

where $\theta(x)$ is the (Heaviside) step function.

5.4 Galerkin method

The Galerkin method is based on the idea of using the basis functions φ_n of the series expansion as test functions. This leads to the N conditions

$$(\varphi_n, R) = 0, \tag{5.17}$$

which can be written as

$$(\varphi_n, f_N) = (\varphi_n, f). \tag{5.18}$$

For an orthogonal set of basis functions, this yields

$$a_n = \frac{(\varphi_n, f)}{(\varphi_n, \varphi_n)}. \tag{5.19}$$

This equation has already been discussed in section 4.3. For a non-orthogonal basis set, e.g. the basis of the finite elements, the Galerkin condition yields a

system of linear equations,

$$\sum_{m=1}^{N}(\varphi_n, \varphi_m)a_m = (\varphi_n, f), \tag{5.20}$$

where the matrix $A_{nm} = (\varphi_n, \varphi_m)$ is sparse for the finite elements.

Let us now return to our example function $f(x) = \sin(2\pi x)^3 + \cos(6\pi(x^2 - 1/2))$. Figure 5.2 shows the approximation of this function with Fourier and finite element basis sets and the Galerkin method. There are no collocation points and the approximation using finite elements does not exactly match the function to be approximated at the interpolation points. The function is only approximated in the integral sense.

Note: The Galerkin condition (see also Eq. (5.17))

$$(\varphi_n, R) = 0, \tag{5.21}$$

means that the residual is *orthogonal* to all basis functions. In other words, the residual can only contain contributions to the function that cannot be mapped with the given basis set. This implies that we can systematically improve our solution by extending the basis set.

5.5 Least squares

An alternative approach to approximation is to minimize the square of the residual, (R, R), also knows as a *least squares* approach. For a general series expansion with N basis functions, we obtain

$$\begin{aligned}(R, R) &= (f, f) + (f_N, f_N) - (f_N, f) - (f, f_N) \\ &= (f, f) + \sum_{n=1}^{N}\sum_{m=1}^{N} a_n^* a_m(\varphi_n, \varphi_m) - \sum_{n=1}^{N} a_n^*(\varphi_n, f) - \sum_{n=1}^{N} a_n(f, \varphi_n).\end{aligned} \tag{5.22}$$

This error square is minimized if

$$\frac{\partial(R, R)}{\partial a_k} = \sum_{n=1}^{N} a_n^*(\varphi_n, \varphi_k) - (f, \varphi_k) = 0 \tag{5.23}$$

and

$$\frac{\partial(R, R)}{\partial a_k^*} = \sum_{n=1}^{N} a_n(\varphi_k, \varphi_n) - (\varphi_k, f) = 0. \tag{5.24}$$

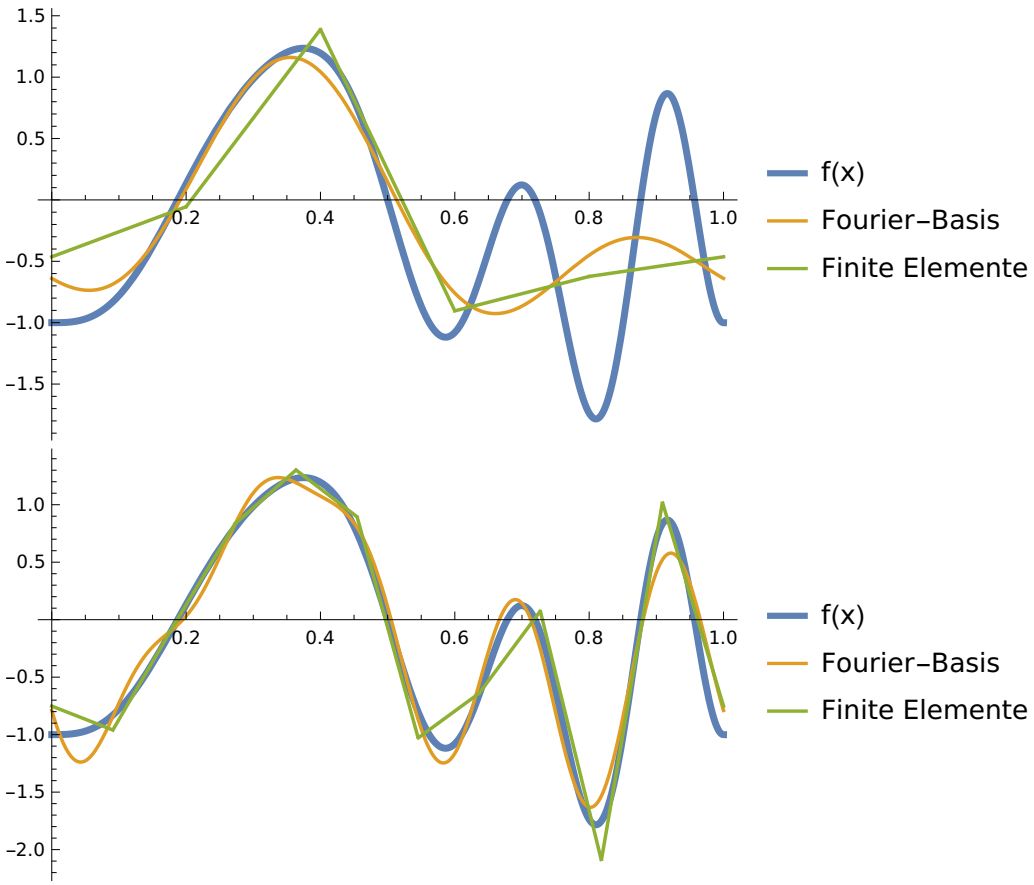

Figure 5.2: Approximation of the periodic function $f(x) = \sin(2\pi x)^3 + \cos(6\pi(x^2 - 1/2))$ on the interval $[0, 1]$ with a Fourier basis and finite elements. The figure shows an approximation 5 (top) and 11 (bottom) basis functions. The coefficients were determined using the Galerkin method. The approximation with 5 basis functions does not capture the two right oscillations of the target function $f(x)$ in both cases.

This expression is identical to Eq. (5.20) of the Galerkin method.

Chapter 6

Fourier spectral methods

Context: We now develop the ideas for solving partial differential equations outlined in the previous chapters. In this chapter, we specifically describe solution strategies using the Fourier basis. This leads to a solution method that belongs to the class of *spectral methods*. The chapter also describes how to extend the solution approaches to multidimensional spaces and how to treat nonlinear terms.

6.1 Differential operators

To solve differential equations, we now use exactly the same methods that we developed in the previous chapter: Minimizing the residual using the Galerkin method. Our residual now has the general form

$$R(x, y, z, \ldots; a_0, a_1, \ldots, a_N) = \mathcal{L}u_N(x, y, z, \ldots) - f(x, y, z, \ldots), \tag{6.1}$$

where the unknown function u_N is represented here as a series expansion into a certain basis $\varphi_n(x, y, z)$. The Galerkin method requires

$$(\varphi_n, R) = 0 \tag{6.2}$$

for each n.

We now discuss the Fourier basis for periodic functions on $x \in [0, L]$ in one dimension,

$$\varphi_n(x) = \exp(iq_n x) \tag{6.3}$$

with $q_n = 2\pi n/L$. The operator $\mathcal{L}$ can contain any differential operations

that act on the basis functions, for example

$$\frac{d}{dx}\varphi_n(x) = iq_n\varphi_n(x) \tag{6.4}$$

$$\frac{d^2}{dx^2}\varphi_n(x) = -q_n^2\varphi_n(x). \tag{6.5}$$

The derivatives of the (Fourier) basis functions result in the *same* basis function and an algebraic factor. It is also said that the basis functions *diagonalize* the differential operator. (This will be different for the finite elements discussed in the next chapter).

This property is particularly useful because, at least for linear differential equations, the residual becomes a trivial series expansion again and we can easily determine the coefficients using the orthogonality of the basis.

6.2 Poisson equation in one dimension

We use the (one-dimensional) Poisson equation,

$$\nabla^2\Phi \equiv \frac{\mathrm{d}^2\Phi}{\mathrm{d}x^2} = -\frac{\rho}{\varepsilon}, \tag{6.6}$$

as a demonstrator. Here ρ is a charge density and Φ is the electrostatic potential. The residual is therefore

$$R(x) = \frac{\mathrm{d}^2\Phi}{\mathrm{d}x^2} + \frac{\rho}{\varepsilon}, \tag{6.7}$$

and the solution of Eq. (6.6) is given by $R(x) = 0$.

Formally, we now write the potential as the series expansion

$$\Phi(x) \approx \Phi_N(x) = \sum_{n=-(N-1)/2}^{(N-1)/2} a_n\varphi_n(x), \tag{6.8}$$

whereby we will not explicitly specify the summation limits in the following. We also expand the right-hand side of Eq. (6.6) into a series with the same basis functions,

$$\rho_N(x) = \sum_{n=-(N-1)/2}^{(N-1)/2} b_n\varphi_n(x). \tag{6.9}$$

Substituting this into Eq. (6.7) we obtain

$$R_N(x) = -\sum_n a_n q_n^2\varphi_n(x) + \frac{1}{\varepsilon}\sum_n b_n\varphi_n(x). \tag{6.10}$$

We now multiply this from the left by the basis functions, (φ_k, R_N) (Galerkin method) and, due to the orthogonality of the basis functions, we obtain the equations

$$(\varphi_k, R_N) = -Lq_k^2 a_k + Lb_k/\varepsilon. \tag{6.11}$$

The factor L appears because the basis functions are not normalized. The condition $(\varphi_k, R_N) = 0$ leads to $a_k = b_k/(q_k^2\varepsilon)$. The approximate solution of the Poisson equation is thus given by

$$\Phi_N(x) = \sum_n \frac{b_n}{q_n^2\varepsilon}\varphi_n(x). \tag{6.12}$$

This is the Fourier series of the solution.

6.3 Transition to the Fourier transform

The Fourier basis Eq. (6.3) is periodic on a finite domain of length L. If we let the length L go to infinity, we get a formulation for non-periodic functions. This leads directly to the *Fourier transform*.

We write the series expansion as

$$\Phi_N(x) = \sum_{n=-N}^{N} a_n\varphi_n(x) = \sum_{n=-N}^{N} a_n \exp(iq_n x) = \sum_{n=-N}^{N} \frac{\Delta q}{2\pi}\,\tilde{\Phi}(q_n)\exp(iq_n x) \tag{6.13}$$

with $\Delta q = q_{n+1} - q_n = 2\pi/L$ and rescaled coefficients $\tilde{\Phi}(q_n) = La_n$. Here, only the factor $1 = L\Delta q/2\pi$ was inserted on the right-hand side of Eq. (6.13). This now helps to form the limits $L \to \infty$ and $N \to \infty$. In this case, $\Delta q \to dq$ and the sum becomes the integral. This yields

$$\Phi(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}q}{2\pi}\,\tilde{\Phi}(q)\exp(iqx)\,, \tag{6.14}$$

the *inverse Fourier transform*.

The (forward) transform is obtained via a similar argument. We now know that

$$\tilde{\Phi}(q_n) = La_n = L\frac{(\varphi_n, \Phi_N)}{(\varphi_n, \varphi_n)} = (\varphi_n, \Phi_N) = \int_0^L \mathrm{d}x\,\Phi_N(x)\exp(-iq_n x)\,. \tag{6.15}$$

In the limiting case $L \to \infty$ and $N \to \infty$ this becomes

$$\tilde{\Phi}(q) = \int_{-\infty}^{\infty} \mathrm{d}x\,\Phi(x)\exp(-iqx)\,, \tag{6.16}$$

of the Fourier transform. The Fourier transform is useful to obtain analytical solutions for partial differential equations on infinite domains.

Note: A tilde $\tilde{f}(q)$ denotes the Fourier transform of a function $f(x)$. The Fourier transform is a function of the wave vector q. In contrast, from the Fourier series we obtain countable coefficients a_n. The reason for this is the periodicity of the function.

6.4 Poisson equation in multiple dimensions

Similar to how we constructed an approximate solution for a differential equation using a series expansion, we can use the approach Eq. (6.14) to obtain analytical solutions. In this section, this is demonstrated using the Poisson equation in three dimensions.

In three dimensions, the Poisson equation is

$$\nabla^2\Phi \equiv \frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} + \frac{\partial^2\Phi}{\partial z^2} = -\frac{\rho}{\varepsilon}. \tag{6.17}$$

In contrast to Eq. (6.6), the partial derivative ∂ now appears here because $\Phi(x, y, z)$ depends on three variables (the Cartesian coordinates).

The generalization of the Fourier basis and thus also of the Fourier transform to three dimensions is trivial. A basis is obtained by multiplying basis functions in the Cartesian directions (x, y and z). Usually, you now need three indices for the coefficients, which denote the basis in x, y and z respectively. The result is a series expansion

$$\begin{aligned}\Phi_{NMO}(x, y, z) &= \sum_{n=-(N-1)/2}^{(N-1)/2} \sum_{m=-(M-1)/2}^{(M-1)/2} \sum_{o=-(O-1)/2}^{(O-1)/2} a_{nmo}\varphi_n(x)\varphi_m(y)\varphi_o(z) \\ &\equiv \sum_{n=-(N-1)/2}^{(N-1)/2} \sum_{m=-(M-1)/2}^{(M-1)/2} \sum_{o=-(O-1)/2}^{(O-1)/2} a_{nmo}\varphi_{nmo}(x, y, z)\end{aligned} \tag{6.18}$$

with (possibly different) truncation orders N, M and O. The basis set is given here by the functions $\varphi_{nmo}(x, y, z) = \varphi_n(x)\varphi_m(y)\varphi_o(z)$. Orthogonality of this basis set is trivially derived from the orthogonality of the one-dimensional basis functions $\varphi_n(x)$. The generalization of the Fourier transform follows

directly from this. The Fourier inverse transform is written as

$$\Phi(x,y,z) = \int_{-\infty}^{\infty} \frac{\mathrm{d}^3 q}{(2\pi)^3} \tilde{\Phi}(q_x, q_y, q_z) \exp\left(iq_x x + iq_y y + iq_z z\right), \tag{6.19}$$

where the Fourier transform $\tilde{\Phi}$ now naturally depends on three wave vectors q_x, q_y and q_z. The differential operator $\mathrm{d}^3 q = \mathrm{d}q_x\,\mathrm{d}q_y\,\mathrm{d}q_z$ is a shorthand notation for three-dimensional integration.

We can now insert Eq. (6.19) into the PDGL Eq. (6.17) and obtain

$$R(\vec{r}) = \int_{-\infty}^{\infty} \frac{\mathrm{d}^3 q}{(2\pi)^3} \left[\left(-q_x^2 - q_y^2 - q_z^2\right)\tilde{\Phi}(\vec{q}) + \frac{\tilde{\rho}(\vec{q})}{\varepsilon}\right] \exp\left(i\vec{q}\cdot\vec{r}\right) = 0 \tag{6.20}$$

with $\vec{r} = (x,y,z)$ and $\vec{q} = (q_x, q_y, q_z)$. This equation must be fulfilled for every x, y, z and therefore the argument of the integration must disappear, i.e.

$$-q^2\tilde{\Phi}(\vec{q}) + \frac{\tilde{\rho}(\vec{q})}{\varepsilon} = 0. \tag{6.21}$$

Note: An alternative argument is obtained by writing the Fourier transform of $R(x,y,z)$:

$$R(q_x', q_y', q_z') = \int \mathrm{d}^3 r\, R(\vec{r}) \exp\left(-iq_x x - iq_y y - iq_z z\right). \tag{6.22}$$

This contains terms of the form

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \exp\left(i(q_x - q_x')x\right) = 2\pi\delta(q_x - q_x'), \tag{6.23}$$

which are expressions of the orthogonality of the basis functions. Since the basis functions are now "parameterized" with a continuous q_x (instead of a discrete n), a Dirac δ function is obtained instead of the Kronecker δ in the orthogonality relation.

Equation (6.21) can easily be solved analytically. This yields

$$\tilde{\Phi}(\vec{q}) = \frac{\tilde{\rho}(\vec{q})}{\varepsilon q^2} \tag{6.24}$$

with $q = |\vec{q}|$. This is equivalent to solving Eq. (6.12) for the Poisson equation on a periodic domain. The difficulty now lies in evaluating the back and forth transformation for a given $\rho(x,y,z)$.

Example: As an example, we now consider the solution for a point charge Q at the origin,

$$\rho(x, y, z) = Q\delta(x)\delta(y)\delta(z). \tag{6.25}$$

The Fourier transform of the charge density ρ is obtained from Eq. (6.16),

$$\tilde{\rho}(q_x, q_y, q_z) = Q. \tag{6.26}$$

I.e. the Fourier transform of the electrostatic potential is given by (see Eq. (6.24))

$$\tilde{\Phi}(\vec{q}) = \frac{Q}{\varepsilon q^2}, \tag{6.27}$$

and thus the representation in real space is

$$\begin{aligned} \Phi(\vec{r}) =& \int_{-\infty}^{\infty} \frac{\mathrm{d}^3 q}{(2\pi)^3} \frac{Q}{\varepsilon q^2} \exp\left(i\vec{q}\cdot\vec{r}\right) \\ =& \frac{Q}{(2\pi)^3\varepsilon} \int_0^{\infty} \mathrm{d}q \int_0^{2\pi} \mathrm{d}\phi \int_{-1}^{1} \mathrm{d}(\cos\theta)\, \exp\left(iqr\cos\theta\right) \end{aligned} \tag{6.28}$$

where $\mathrm{d}^3 q = q^2\, \mathrm{d}q\, \mathrm{d}\phi\, \mathrm{d}(\cos\theta)$ with azimuth angle ϕ and elevation angle θ, was used (see also Fig. 6.1). We require here (without limiting the generality) that $\vec{r}$ points in the direction of the zenith.

One obtains

$$\begin{aligned} \Phi(\vec{r}) =& \frac{Q}{(2\pi)^2\varepsilon} \int_0^{\infty} \mathrm{d}q \int_{-1}^{1} \mathrm{d}(\cos\theta)\, \exp\left(iqr\cos\theta\right) \\ =& \frac{Q}{(2\pi)^2\varepsilon} \int_0^{\infty} \mathrm{d}q\, \frac{\exp(iqr) - \exp(-iqr)}{iqr} \\ =& \frac{Q}{(2\pi)^2\varepsilon} \int_{-\infty}^{\infty} \mathrm{d}q\, \frac{\sin qr}{qr} \\ =& \frac{Q}{4\pi\varepsilon r}, \end{aligned} \tag{6.29}$$

where $\int \mathrm{d}x\, \sin x/x = \pi$ was used. This is the known solution for the electrostatic potential of a point charge. It is also called the fundamental solution or *Green's function* of the (three-dimensional) Poisson equation.

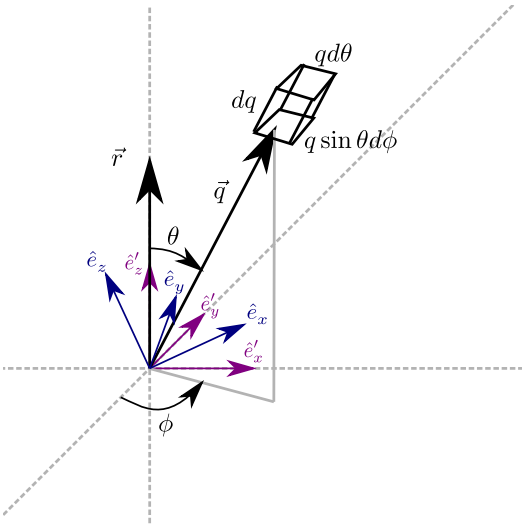

Figure 6.1: Volume element for integration in spherical coordinates.

Chapter 7

Discrete convolutions

Context: This chapter revisits the discrete Fourier transform (DFT) and illustrates how it can be used for the efficient computation of cyclic convolutions. Nonlinear terms in partial differential lead to noncyclic convolutions in spectral methods. Direct application of the DFT then yield aliasing errors, which can be corrected through dealiasing procedures.

7.1 Discrete Fourier transform

The Fourier series represents functions on finite domains of length L as

$$f(x) = \frac{1}{L} \sum_{n=-\infty}^{\infty} \tilde{f}(q_n) \exp(iq_n x) \tag{7.1}$$

with $q_n = 2\pi n/L$. The inverse transform is given by

$$\tilde{f}(q_n) = \int_0^L \mathrm{d}x\, f(x) \exp(-iq_n x)\,. \tag{7.2}$$

We now evaluate the Fourier series only on discrete, equidistant sampling points $x_k = kL/N$. The inverse transform Eq. (7.2) then becomes

$$\tilde{f}(q_n) = \frac{L}{N} \sum_{k=0}^{N-1} f(x_k) \exp(-iq_n x_k)\,, \tag{7.3}$$

where the sum is the discrete variant of $\int \mathrm{d}x \approx \sum \Delta x$ with grid spacing $\Delta x = L/N$. The phase-factor $q_n x_k = 2\pi kn/N$ is an integer multiple of 2π if n is an integer multiple of N, hence $\tilde{f}(q_{n+\alpha N}) = \tilde{f}(q_n)$ because $k(n+\alpha N)/N =$

$kn/N + k\alpha$ for integer $\alpha \in \mathbb{Z}$. Formally this means that Eq. (7.1) diverges. This is consistent with the interpretation that the discretely sampled $f(x_k)$ is a convolution of $f(x)$ with a Dirac comb. We can truncate the forward transform to

$$f(x_k) = \frac{1}{L} \sum_{n=M}^{M+N-1} \tilde{f}(q_n) \exp(iq_n x_k) = \sum_{n=M}^{M+N-1} W_{kn} \hat{f}_n, \tag{7.4}$$

where $\hat{f}_n = \tilde{f}(q_n)/L$ and we are using the DFT-matrix $W_{kn} = \exp(i2\pi kn/N)$, see also Eq. (5.9). This indeed yields the correct discretely sampled function, as can be seen by inserting the inverse transform Eq. (7.5) into Eq. (7.4). The choice of $M \in \mathbb{Z}$ in Eq. (7.4) remains completely arbitrary, but a typical choice is $M = 0$. Equation (7.4) is called the *discrete Fourier transform (DFT)* with inverse Eq. (7.5) which can be rewritten to

$$\hat{f}_n = \frac{1}{N} \sum_{k=0}^{N-1} W^*_{nk} f(x_k). \tag{7.5}$$

The DFT is typically computed using a fast Fourier transform algorithm (FFT), that reduces the computational complexity from $O(N^2)$ of a naive implementation of Eq. (7.4) to $O(N \log N)$.

7.2 Cyclic convolutions

The FFT is useful to reduce the complexity of computing cyclic convolutions from $O(N^2)$ to $O(N \log N)$. A cyclic convolution of the discrete series a_n and b_n is given by

$$c_k = \sum_{n=0}^{N-1} a_n b_{a(k-n)} \tag{7.6}$$

where $a(m) = m + \alpha N$ with $\alpha \in \mathbb{Z}$ such that $0 \leq a(m) < N$. We can insert the DFT expression, Eq. (7.4), which automatically fulfills the cyclic property

to obtain

$$\begin{aligned} c_k &= \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} W_{nm} \hat{a}_m \sum_{l=0}^{N-1} W_{k-n,l} \hat{b}_l \\ &= \frac{1}{N^2} \sum_{m=0}^{N-1} \sum_{l=0}^{N-1} \hat{a}_m W_{kl} \hat{b}_l \sum_{n=0}^{N-1} W_{n,m-l} \\ &= \frac{1}{N^2} \sum_{m=0}^{N-1} \sum_{l=0}^{N-1} \hat{a}_m W_{kl} \hat{b}_l N \delta_{m,l} \\ &= \frac{1}{N} \sum_{m=0}^{N-1} W_{km} \hat{a}_m \hat{b}_m, \end{aligned} \tag{7.7}$$

which is the discrete Fourier transform representation of the product, i.e.

$$\hat{c}_m = \hat{a}_m \hat{b}_m. \tag{7.8}$$

This means the convolution requires an element-wise product with complexity $O(N)$, plus three fast Fourier transforms, which all have complexity $O(N \log N)$.

7.3 Nonlinear terms and aliasing

We now discuss the treatment of nonlinear terms in numerical solution of partial differential equations with Fourier spectral methods. As an example, let us regard the simple multiplication of two functions,

$$h(x) = f(x)g(x). \tag{7.9}$$

We expand both $f(x)$, $g(x)$ and $h(x)$ into a truncated Fourier series, i.e. we write

$$f(x) \approx \sum_{n=-N}^{N} a_n \exp\left(iq_n x\right), \tag{7.10}$$

$$g(x) \approx \sum_{m=-N}^{N} b_m \exp\left(iq_m x\right), \tag{7.11}$$

$$h(x) \approx \sum_{k=-N}^{N} c_k \exp\left(iq_k x\right). \tag{7.12}$$

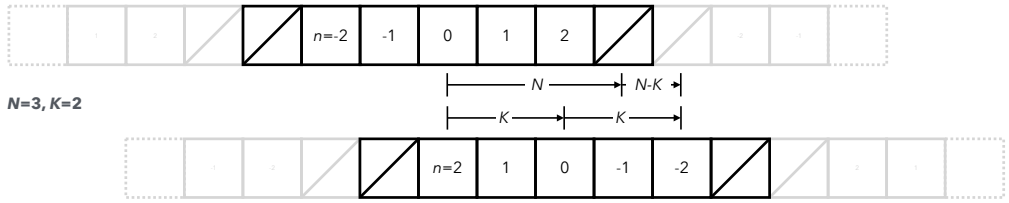

Figure 7.1: Example of a discrete convolution. The crossed-out boxes indicate the zero-padded region. The gray boxes indicate the cyclic (periodic) continuation implicit in the convolution computed with the FFT. Period interactions are eliminated if $K + K \leq N + (N - K)$, leading to the 2/3 rule.

To simplify notation, we use N to denote the truncation of the sums which now run over $2N + 1$ terms.

Inserting into Eq. (7.9) yields

$$h(x) \approx \sum_{n=-N}^{N} \sum_{m=-N}^{N} a_n b_m \exp\left(i q_{n+m} x\right). \tag{7.13}$$

The discrete form of this equation can be obtained from the Galerkin method, i.e. by multiplying with $\exp\left(i q_k x\right)$ from the left, which yields

$$c_k = \frac{1}{L}\left(\exp(i q_k x), h(x)\right) = \sum_{n=-N}^{N} \sum_{m=-N}^{N} a_n b_m \delta_{k,n+m} = \sum_{n=-N}^{N} a_n b_{k-n} \tag{7.14}$$

An important observation is that c_k is nonzero in the range $-2N \leq k \leq 2N$, but we truncated the sum at N. Since b_m in Eq. (7.8) is not cyclic, using the discrete Fourier transform to compute this convolution introduces an error. This type of numerical error is called *aliasing*.

Aliasing can be removed from the cyclic convolution by zero padding the two vectors. We set $a_n = 0$ and $b_n = 0$ for $n > K$. The value of c_k is then exact for $k \leq K$ if $2K \leq N + (N - K)$, yielding $K \leq 2N/3$ (see Fig. 7.1 for an illustration). This means we can compute the convolution, Eq. (7.14), using the FFT for the cyclic convolution at the expanse of discarding 1/3 of the wavevectors. This is called the 2/3 rule (Orszag, 1971).

Appendix A

Differential equations

Context: Most of the phenomena we encounter in engineering are very well described by differential equations. We remember the discrete network models from electrical engineering and systems theory. They are described by a system of linear ordinary differential equations with time as an independent variable. described. We also remember the diffusion process, such as the heat transport in a component on a heat sink that is exposed to a heat source. This phenomenon is best described using a partial differential equation (partial differential equation). In this chapter, we deal with an abstract classification of differential equations. The diffusion process will be repeated in more detail in the next chapter.

A.1 Ordinary differential equations

We recall the classification (properties) of *ordinary* differential equations (ODEs) and recognize the different types of differential equations. For all these differential equations, we are always interested in a solution for a certain initial value (or boundary value), e.g. $x(t = 0) = x_0$ etc. This initial value is always part of the definition of the differential equation.

A.1.1 Linearity

A linear differential equation is, for example

$$m\ddot{x}(t) + c\dot{x}(t) + kx = f(t) \tag{A.1}$$

which describes the damped and driven harmonic oscillator, while

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + \mu(x^2 - 1)\frac{\mathrm{d}x}{\mathrm{d}t} + x = 0 \tag{A.2}$$

is a non-linear equation of motion for x. It describes the so-called van der Pol oscillator. The non-linearity can be recognized here by the fact that x^2 multiplies the derivative $\mathrm{d}x/\,\mathrm{d}t$.

Note: The first or higher order derivative is a linear operation, since

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}\lambda f(x) = \lambda\frac{\mathrm{d}^n}{\mathrm{d}x^n}f(x) \tag{A.3}$$

for a constant λ and

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[f(x)+g(x)\right] = \frac{\mathrm{d}^n}{\mathrm{d}x^n}f(x) + \frac{\mathrm{d}^n}{\mathrm{d}x^n}g(x). \tag{A.4}$$

Time derivatives are displayed with a dot,

$$\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t). \tag{A.5}$$

For functions of a variable, the derivative is often displayed with a dash,

$$f'(x) = \frac{\mathrm{d}}{\mathrm{d}x}f(x). \tag{A.6}$$

This is no longer possible for functions with several variables. We will therefore always explicitly use the differential operator here.

A.1.2 Order

The order of a differential equation is given by the highest derivative that appears in the equation. Eq. (A.1) and Eq. (A.2) are examples of second-order differential equations.

A.1.3 Systems

A system of first-order differential equations is formed, for example, by the equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x(m-ny), \tag{A.7}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -y(\gamma-\delta x), \tag{A.8}$$

the well-known Räuber-Beute equations or Lotka-Volterra equations. Equations (A.7) and (A.8) are still non-linear.

Differential equations of higher order can be rewritten into a system of 1st order equations. In the example of the damped harmonic oscillator,

$$m\ddot{x}(t) + c\dot{x}(t) + kx = f(t), \tag{A.9}$$

we replace $\dot{x} = y$ and thus obtain two first-order equations instead of the original second-order equation, namely

$$\dot{x} = y \tag{A.10}$$
$$m\dot{y} = -cy - kx + f(t) \tag{A.11}$$

A.2 Partial differential equations

Partial differential equations (PDEs) are differential equations with more than one independent variable. As an example, we imagine a time-dependent heat transport problem in one dimension. This is represented by a diffusion equation for the local temperature of the system. The temperature is therefore represented as a function of two independent variables, the time t and the spatial position x, $T(x,t)$. The time evolution of the temperature is given by

$$\frac{\partial T(x,t)}{\partial t} = \kappa \frac{\partial^2 T(x,t)}{\partial x^2}, \tag{A.12}$$

where κ denotes the heat conduction coefficient. This equation was developed by Joseph Fourier (*1768, †1830), whom we will encounter again during this course.

Note: In Eq. (A.12) $\partial/\partial t$ denotes the *partial derivative*. This is the derivative with respect to one of the arguments (here t), i.e. the variation of the function if all other arguments are kept constant. With ODEs, in contrast to PDEs, only derivatives with respect to one variable (usually the time t) occur, which are then denoted by the differential operator $\mathrm{d}/\,\mathrm{d}t$.

A.2.1 First order

Quasilinear PDEs of the first order, i.e. equations of the form

$$P(x,t;u)\frac{\partial u(x,t)}{\partial x} + Q(x,t;u)\frac{\partial u(x,t)}{\partial t} = R(x,t;u), \tag{A.13}$$

for an (unknown) function $u(x,t)$ and the initial condition $u(x,t=0)=u_0(x)$ can be systematically traced back to a system of coupled first-order ODEs. We want to investigate this important property.

Note: In Eq. (A.13) a representation with two variables x and t was chosen for illustration. In general, we can write

$$\sum_i P_i(\{x_i\};u)\frac{\partial u(\{x_i\})}{\partial x_i} = R(\{x_i\};u) \tag{A.14}$$

The notation used here is $u(\{x_i\}) = u(x_0, x_1, x_2, \ldots)$, i.e. the curly brackets denote all degrees of freedom x_i.

Equation (A.13) can be transformed to a system of ODEs. This is called the method of characteristics. We can then apply the formalisms (analytical or numerical) for solving systems of ODEs that we learned about in the lecture "Differential Equations".

We proceed as follows:

1. First, we parameterize the independent variables in Eq. (A.13) with a parameter s according to $x(s)$ and $t(s)$.

2. We then form the *total derivative* of $u(x(s),t(s))$ to s

$$\frac{\mathrm{d}u(x(s),t(s))}{\mathrm{d}s} = \frac{\partial u(x(s),t(s))}{\partial x}\frac{\mathrm{d}x(s)}{\mathrm{d}s} + \frac{\partial u(x(s),t(s))}{\partial t}\frac{\mathrm{d}t(s)}{\mathrm{d}s}. \tag{A.15}$$

3. By comparing the coefficients of the total derivative (A.15) with the PDE (A.13), you can see that this DGL is solved exactly when

$$\frac{dx(s)}{ds} = P(x,t,u), \tag{A.16}$$
$$\frac{dt(s)}{ds} = Q(x,t,u) \quad \text{und} \tag{A.17}$$
$$\frac{du(s)}{ds} = R(u(s)). \tag{A.18}$$

   is fulfilled. This describes the solution along certain curves in the (x,t)-plane.

We have thus converted the PDE into a set of coupled first-order ODEs, Eq. (A.21)-(A.23).

Example: The transport equation

$$\frac{\partial u(x,t)}{\partial t} + c\frac{\partial u(x,t)}{\partial x} = 0 \tag{A.19}$$

with the initial condition $u(x, t = 0) = u_0(x)$ is to be solved. We proceed according to the recipe above:

1. We parameterize the variables x and t with the help of a new variable s, i.e. $x(s)$ and $t(s)$. We are now looking for an expression with which we can determine $x(s)$ and $t(s)$.

2. We now ask how the function $u(x(s), t(s))$ behaves. This function describes the change in an initial value $u(x(0), t(0))$ with the variable s. The total derivative becomes

$$\frac{\mathrm{d}u(x(s),t(s))}{\mathrm{d}s} = \frac{\partial u}{\partial t}\frac{\mathrm{d}t(s)}{\mathrm{d}s} + \frac{\partial u}{\partial x}\frac{\mathrm{d}x(s)}{\mathrm{d}s}. \tag{A.20}$$

3. The total derivative is identical to the partial differential equation that we want to solve if

$$\frac{\mathrm{d}x(s)}{\mathrm{d}s} = c \quad \text{and} \tag{A.21}$$

$$\frac{\mathrm{d}t(s)}{\mathrm{d}s} = 1. \tag{A.22}$$

   In this case, the following applies

$$\frac{\mathrm{d}u(s)}{\mathrm{d}s} = 0. \tag{A.23}$$

4. The general solutions for the three ordinary differential equations (A.21)-(A.23) are given by

$$x(s) = cs + \text{const.}, \tag{A.24}$$

$$t(s) = s + \text{const.} \quad \text{and} \tag{A.25}$$

$$u(s) = \text{const.} \tag{A.26}$$

5. With the initial conditions $t(0) = 0$, $x(0) = \xi$ and $u(x, t = 0) = f(\xi)$ you get $t = s$, $x = ct + \xi$ and $u = f(\xi) = f(x - ct)$,

The initial condition $f(\xi)$ is transported with the speed c in the positive x-direction. The solution for u remains constant, as the derivative of u is zero, so u retains the value given by the initial condition. The field $u(x, 0)$ is therefore shifted at a constant speed c: $u(x, t) = u(x - ct, 0)$.

A.2.2 Second order

Examples of second-order PDEs are the...

- ...wave equation:

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 \tag{A.27}$$

- ...diffusion equation (which we will look at in more detail here):

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \tag{A.28}$$

- ...Laplace equation (which we will also get to know better):

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{A.29}$$

The second order here refers to the second derivative. These examples are formulated for two variables, but these differential equations can also be written down for more degrees of freedom.

For two variables, the general form of second-order linear PDEs is

$$a(x,y)\frac{\partial^2 u}{\partial x^2} + b(x,y)\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\frac{\partial^2 u}{\partial y^2} = F\left(x, y; u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right), \tag{A.30}$$

where F itself must of course also be linear in the arguments if the entire equation is to be linear. We now classify 2nd order PDEs, but note that this classification is not exhaustive and that it only applies pointwise. The latter means that the PDE can fall into a different classification at different points in space.

We first assume that $F = 0$ and a, b, c are constant. Then we get:

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = 0. \tag{A.31}$$

We rewrite this equation as the quadratic form

$$\begin{pmatrix} \partial/\partial x \\ \partial/\partial y \end{pmatrix} \cdot \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \cdot \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \end{pmatrix} u = \nabla \cdot \underline{C} \cdot \nabla u = 0 \tag{A.32}$$

We can now diagonalize the coefficient matrix $\underline{C}$. This for to

$$\underline{C} = \underline{U} \cdot \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \cdot \underline{U}^T, \tag{A.33}$$

where $\underline{U}$ is unitary due to the symmetry of $\underline{C}$, $\underline{U}^T \cdot \underline{U} = \underline{1}$. The geometric interpretation of the operation $\underline{U}$ is a rotation. We now introduce transformed coordinates x' and y' so that

$$\nabla = \underline{U} \cdot \nabla' \tag{A.34}$$

with $\nabla' = (\partial/\partial x', \partial/\partial y')$. In other words, the transformation matrix is given as

$$\underline{U} = \begin{pmatrix} \partial x'/\partial x & \partial y'/\partial x \\ \partial x'/\partial y & \partial y'/\partial y \end{pmatrix}. \tag{A.35}$$

Equation (A.31) becomes

$$\lambda_1 \frac{\partial^2 u}{\partial x'^2} + \lambda_2 \frac{\partial^2 u}{\partial y'^2} = 0. \tag{A.36}$$

We have diagonalized the coefficients of the differential equation. For any twice differentiable function $f(z)$, is

$$u(x', y') = f\left(\sqrt{\lambda_2} x' + i\sqrt{\lambda_1} y'\right) \tag{A.37}$$

a solution of Eq. (A.36).

The analytical expression for the eigenvalues is:

We now distinguish three cases:

- The case $\det \underline{C} = \lambda_1 \lambda_2 = ac - b^4/4 = 0$ with $b \neq 0$ and $a \neq 0$ leads to a parabolic PDE. This PDE is called parabolic because the quadratic form Eq. (A.32) or (A.33) describes a parabola. (This is of course an analogy. You have to replace the differential operators with coordinates for this to work). Without restriction of generality, let $\lambda_2 = 0$. Then we get

  $$\frac{\partial^2 u}{\partial x'^2} = 0. \tag{A.38}$$

  This is the canonical form of a parabolic PDE.

- The case $\det \underline{C} = \lambda_1 \lambda_2 = ac - b^2/4 > 0$ leads to an elliptic PDE. This PDE is called elliptic because the quadratic form Eq. (A.32) or (A.33) describes an ellipse for a constant right-hand side. (For $\lambda_1 = \lambda_2$ it is a circle). We now convert the equation for the elliptical case to a standardized form and introduce the scaled coordinates $x' = \sqrt{\lambda_1} x''$ and $y' = \sqrt{\lambda_2} y''$. Eq. (A.36) then becomes the canonical elliptic PDE

  $$\frac{\partial^2 u}{\partial x''^2} + \frac{\partial^2 u}{\partial y''^2} = 0. \tag{A.39}$$

The canonical elliptic PDE is therefore the Laplace equation, Eq. (A.39) (here in two dimensions). Solutions of the Laplace equation are called *harmonic functions*.

- The case $\det \underline{C} = \lambda_1 \lambda_2 = ac - b^2/4 < 0$ results in the so-called hyperbolic PDE. This PDE is called hyperbolic because the quadratic form Eq. (A.32) or (A.33) for a constant right-hand side describes a hyperbola. Without restricting the generality, we now require $\lambda_1 > 0$ and $\lambda_2 < 0$. Then we can again introduce scaled coordinates $x' = \sqrt{\lambda_1} x''$ and $y' = \sqrt{-\lambda_2} y''$ so that

$$\frac{\partial^2 u}{\partial x''^2} - \frac{\partial^2 u}{\partial y''^2} = \begin{pmatrix} \partial u/\partial x'' \\ \partial u/\partial y'' \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} \partial u/\partial x'' \\ \partial u/\partial y'' \end{pmatrix} = 0. \tag{A.40}$$

  We can now use a further coordinate transformation, namely a rotation by 45°, to bring the coefficient matrix in Eq. (A.40) to a form in which the diagonal elements are 0 and the secondary diagonal elements are 1. This results in the differential equation

$$\frac{\partial^2 u}{\partial x''' \partial y'''} = 0, \tag{A.41}$$

  where x'''' and y'''' are the corresponding rotated coordinates. This equation is the canonical form of a hyperbolic PDE and is equivalent to Eq. (A.31) in the new variables x''' and y'''.

For higher dimensional problems, we need to look at the eigenvalues of the coefficient matrix $\underline{C}$. The PDE is called *parabolic* if there is an eigenvalue that vanishes, but all other eigenvalues are either greater or less than zero. The PDE is called *elliptic* if all eigenvalues are either greater than zero or less than zero. The PDE is called *hyperbolic* if there is exactly one negative eigenvalue and all others are positive or if there is exactly one positive eigenvalue and all others are negative. It is clear that for PDEs with more than two variables, these three classes of PDEs are not exhaustive and there are coefficient matrices that fall outside this classification scheme. For problems with exactly two variables, this classification leads to the conditions for the determinants of the coefficient matrix mentioned above.

These three types of 2nd-order linear PDEs can also be solved analytically for some problems. In the following, we give an example of this.

Example: We solve the one-dimensional wave equation.

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \tag{A.42}$$

by separating the variables. To do this, we take the approach $u(x,t) = X(x)T(t)$, which leads to

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} = \frac{1}{c^2}\frac{1}{T}\frac{\partial^2 T}{\partial t^2}. \tag{A.43}$$

In Eq. (A.43), the left-hand side depends only on the variable x, while the right-hand side depends only on t. For any x and t, this equation can only be fulfilled if both sides are equal to a constant and we thus obtain

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} = -k^2 = \frac{1}{c^2}\frac{1}{T}\frac{\partial^2 T}{\partial t^2}\,. \tag{A.44}$$

This results in the following two equations

$$\frac{\partial^2 X}{\partial x^2} + k^2 X = 0$$

with the solution $X(x) = e^{\pm ikx}$ and

$$\frac{\partial^2 T}{\partial t^2} + \omega^2 T = 0$$

with the solution $T(t) = e^{\pm i\omega t}$, where we have set $\omega^2 = c^2k^2$. have set. This example needs initial conditions to be completed so that we can can find a solution.

Bibliography

M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 1989.

J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Dover Publications, New York, 2000.

A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.*, 17:549, 1905.

R. M. Martin. *Electronic Structure*. Cambridge University Press, 2004.

M. H. Müser, S. V. Sukhomlinov, and L. Pastewka. Interatomic potentials: achievements and challenges. *Advances in Physics: X*, 8(1):2093129, Jan. 2023.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd ed edition, 2006.

S. Orszag. On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *Journal of the Atmospheric Sciences*, 28: 1074–1074, Sept. 1971.