Aalto-yliopisto

AALTO UNIVERSITY

FALL 2013

Numerics Related to Partial Differential Equations

NUUTTI HYVÖNEN

November 28, 2013

Numerics Related to Partial Differential Equations

Nuutti Hyvönen

Department of Mathematics, Aalto University, P.O.Box 11100, FI-00076 Aalto, Finland
E-mail address: nuutti.hyvonen@aalto.fi
URL: http://users.tkk.fi/nhyvonen/

Contents

Chapter 1. Interpolation and discrete Fourier transform 1
1.1. Polynomial interpolation 1
1.2. Least squares polynomial fitting 3
1.3. Fourier series (a recap) 5
1.4. Discrete Fourier transform (DFT) 7

Chapter 2. Numerical solution of ordinary differential equations 13
2.1. Existence and uniqueness (counterexamples) 13
2.2. Picard–Lindelöf iteration and unique solvability 14
2.3. Continuous dependence on the initial value 18
2.4. Explicit solution methods 19
2.5. Stability and stiff systems 22
2.6. Implicit solution methods 24

Chapter 3. Numerical solution of the Laplace equation (finite difference method) 27
3.1. One-dimensional case 27
3.2. Two-dimensional case 32

Chapter 4. Numerical solution of the heat equation (finite difference method) 37

Chapter 5. Numerical solution of the wave equation (finite difference method) 43
5.1. Representation as a first order system 44
5.2. Direct discretization of the second time derivative 47
5.3. Generalizations 49

Chapter 6. Galerkin approximation and finite element method (in a nutshell) 51
6.1. Galerkin approximation 52
6.2. Multidimensional case 55
6.3. Finite element method 56

CHAPTER 1

Interpolation and discrete Fourier transform

The only available information about an examined function is often its values at some (sparse) set of grid points, which is not enough for many applications like differentiation. Under such circumstances one may resort to some form of interpolation, e.g., with the help of polynomials.

1.1. Polynomial interpolation

Suppose that the values of a function

$$f : \mathbb{C} \to \mathbb{C}$$

are known at $n+1 \in \mathbb{N}$ distinct points $x_0, x_1, \dots, x_n \in \mathbb{C}$, and the task is to find a polynomial of order n,

$$p_n(x) = \sum_{k=0}^{n} a_k x^k = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + a_n x^n, \qquad x \in \mathbb{C}, \tag{1.1}$$

such that

$$p_n(x_j) = f(x_j), \qquad j = 0, 1, \dots, n. \tag{1.2}$$

In other words, one is asked to find such coefficients $a := [a_0, a_1, \dots, a_n]^{\mathrm{T}} \in \mathbb{C}^{n+1}$ that (1.2) is satisfied.

Lemma 1.1. *The above interpolation problem has a unique solution, that is, the equations* (1.2) *define a unique polynomial of the form* (1.1).

PROOF. Let us define a linear mapping $V : \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$ via

$$V : \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \mapsto \begin{bmatrix} p_n(x_0) \\ p_n(x_1) \\ \vdots \\ p_n(x_n) \end{bmatrix},$$

where p_n is the polynomial defined by (1.1). It is easy to see that V can be represented in a matrix form as

$$Va = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix} a. \tag{1.3}$$

In the following, we identify V with this so-called *Vandermonde matrix*, i.e., we interpret $V \in \mathbb{C}^{(n+1)\times(n+1)}$. With the help of V the interpolation task can be

given in the following form: Find $a \in \mathbb{C}^{n+1}$ so that

$$Va = y, \tag{1.4}$$

where $y := [f(x_0), f(x_1), \ldots, f(x_n)]^{\mathrm{T}} \in \mathbb{C}^{n+1}$.

From the course *Mat-1.1020 L2* we recall that (1.4) has a unique solution if and only if V is invertible, which is equivalent to V being injective, that is,

$$Va = 0 \iff a = 0.$$

Assume that $Va = 0$ for some $a \in \mathbb{C}^{n+1}$. According to the definition of V this is equivalent to

$$p_n(x_j) = 0, \qquad j = 0, 1, \ldots, n,$$

meaning that the nth degree polynomial p_n has $n+1$ distinct roots. It is well known that the only polynomial p_n with this property is the trivial polynomial $p_n \equiv 0$,[1] proving that $a_0 = a_1 = \cdots = a_n = 0$. In consequence, V is injective and the proof is complete. □

The following theorem gives information about the error that is made if a function is replaced by its polynomial interpolant. For simplicity, we assume that $f : \mathbb{R} \to \mathbb{R}$.

Theorem 1.2. *Assume that the points $x_0, x_1, \ldots, x_n$ are in ascending order on the interval $[x_0, x_n]$ and $f \in C^{n+1}$, i.e., f is $n+1$ times continuously differentiable. Let p_n be the (nth degree) polynomial interpolant of f with respect to the aforementioned set of points. For any $x \in [x_0, x_n]$, it holds that*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=0}^{n} (x - x_j) \tag{1.5}$$

with some $\xi = \xi(x) \in [x_0, x_n]$.

PROOF. If $x = x_j$ for some $j = 0, 1, \ldots, n$, both sides of (1.4) vanish and the claim holds trivially.

We may thus assume that $x \neq x_j$ for all $j = 0, 1, \ldots, n$. Let us define two auxiliary functions:

$$w(s) = \prod_{j=0}^{n} (s - x_j)$$

and

$$g(s) = f(s) - p_n(s) - \frac{f(x) - p_n(x)}{w(x)} w(s),$$

which is well defined since $w(x) \neq 0$. Clearly, $g(x) = 0$, as well as $g(x_j) = 0$, $j = 0, 1, \ldots, n$, since

$$p_n(x_j) = f(x_j) \quad \text{and} \quad w(x_j) = 0, \qquad j = 0, 1, \ldots, n.$$

Hence, g has $n+2$ distinct roots on the interval $[x_0, x_n]$.

According to the *Rolle's theorem* (draw a picture or consult, e.g., Wikipedia), between any two roots of g there lies (at least) one root of g', resulting in at least $n+1$ roots of g' on the interval $[x_0, x_n]$. Continuing analogously, it is easy to deduce that $g^{(n+1)}$ has at least one root, say ξ, between x_0 and x_n. Because p_n is by definition a polynomial of degree n, $p^{(n+1)}$ is identically zero. On the other

[1] Fundamental theorem of algebra.

hand, as w is a polynomial of degree $n+1$, whose leading order coefficient is 1, it is straightforward to convince oneself that $w^{(n+1)}(s) \equiv (n+1)!$. To sum up,

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(x) - p_n(x)}{w(x)}(n+1)!\,,$$

or equivalently,

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} w(x),$$

which completes the proof. □

In particular, Theorem 1.2 induces the estimate

$$|f(x) - p_n(x)| \le \frac{C}{(n+1)!} \prod_{j=0}^{n} |x - x_j|, \qquad x \in [x_0, x_n],$$

where

$$C = \max_{\xi \in [x_0, x_n]} \left| f^{(n+1)}(\xi) \right|.$$

In other words, if the $(n+1)$th derivative of f does not attain very large values on the interval $[x_0, x_n]$ and the set of interpolation points is relatively dense, the discrepancy between f and p_n is not huge.

Remark 1.3. The interpolation polynomial p_n of $f : \mathbb{C} \to \mathbb{C}$ with respect to the distinct points $x_0, x_1, \ldots, x_n \in \mathbb{C}$ can be written explicitly with the help of the so-called *Lagrange polynomial*s,

$$l_m(x) = \prod_{k \ne m} \frac{x - x_k}{x_m - x_k}, \qquad m = 0, 1, \ldots, n.$$

Indeed, l_m is obviously an nth order polynomial and

$$l_m(x_j) = \delta_{m,j} := \begin{cases} 1, & j = m, \\ 0, & j \ne m, \end{cases} \qquad m, j = 0, 1, \ldots, n.$$

Consequently, it must hold that

$$p_n(x) = \sum_{m=0}^{n} f(x_m)\, l_m(x)$$

since the polynomial interpolant is unique due to Theorem 1.2.

Be that as it may, the above introduced linear algebraic interpretation of the polynomial interpolation problem enables a straightforward extension to the case where the polynomial order is less than the number of mesh points and the 'interpolant' is sought for in the sense of *least squares*.

1.2. Least squares polynomial fitting

We will next consider the practically relevant situation where (noisy) values of a function $f : \mathbb{R} \to \mathbb{R}$ are known at the points $x_0, x_1, \ldots, x_n \in \mathbb{R}$ and the task is to find a polynomial of degree $m \le n$,

$$p_m(x) = \sum_{k=0}^{m} a_k x^k = a_0 + a_1 x + \cdots + a_{m-1} x^{m-1} + a_m x^m, \qquad x \in \mathbb{R}, \tag{1.6}$$

which approximates f in the sense of least squares. To be more precise, the aim is to determine the coefficients $a := [a_0, a_1, \dots, a_m]^{\mathrm{T}} \in \mathbb{R}^{m+1}$ so that the square sum

$$\sum_{j=0}^{n} \Big(f(x_j) - p_m(x_j) \Big)^2$$

is minimized. (Re)defining the linear map $V : \mathbb{R}^{m+1} \to \mathbb{R}^{n+1}$ by

$$V : \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \mapsto \begin{bmatrix} p_m(x_0) \\ p_m(x_1) \\ \vdots \\ p_m(x_n) \end{bmatrix}$$

and noting that V may once again be identified with a Vandermonde matrix,

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ 1 & x_1 & x_1^2 & \dots & x_1^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix}, \tag{1.7}$$

the least squares problem can be given in an equivalent form as

$$\arg\min_{a \in \mathbb{R}^{m+1}} |y - Va|^2, \tag{1.8}$$

where $|\cdot|$ denotes the Euclidean norm and $y := [f(x_0), f(x_1), \dots, f(x_n)]^{\mathrm{T}} \in \mathbb{R}^{n+1}$.

Theorem 1.4. *Assume that $x_0 < x_1 < x_2 < \cdots < x_n$ and $m \le n$. Then, the least squares polynomial fitting problem* (1.8) *has a unique solution, which is determined by the unique solution of the* normal equation

$$V^{\mathrm{T}} V a = V^{\mathrm{T}} y, \tag{1.9}$$

where $V \in \mathbb{R}^{(n+1)\times(m+1)}$ is the Vandermonde matrix defined in (1.7).

PROOF. To begin with, we prove that (1.9) has a unique solution, which is equivalent to showing that the square matrix $V^{\mathrm{T}}V \in \mathbb{R}^{(m+1)\times(m+1)}$ injective. In fact, we will do slightly better and show that $V^{\mathrm{T}}V$ is positive definite:

$$a^{\mathrm{T}} V^{\mathrm{T}} V a = (Va)^{\mathrm{T}} (Va) = |Va|^2 \ge 0,$$

where the equality holds if and only if $Va = 0 \in \mathbb{R}^{n+1}$. By looking at (1.7), it is obvious that $Va = 0$ means the mth order polynomial p_m defined by the coefficients $a \in \mathbb{R}^{m+1}$ has $n+1$ distinct roots. Since $n \ge m$, it must hold $p_m \equiv 0$ and, in particular, $a = 0 \in \mathbb{R}^{m+1}$. Hence,

$$a^{\mathrm{T}} V^{\mathrm{T}} V a > 0 \qquad \text{for all } \mathbb{R}^{m+1} \ni a \neq 0,$$

which also demonstrates injectivity.

It remains to prove that the unique solution of (1.9) indeed uniquely solves (1.8), i.e., is the minimizer of

$$g(a) := |y - Va|^2 = y^{\mathrm{T}} y - 2a^{\mathrm{T}} V^{\mathrm{T}} y + a^{\mathrm{T}} V^{\mathrm{T}} V a, \qquad a \in \mathbb{R}^{m+1}.$$

This could be established by means of linear algebra, but we take here a more straightforward approach and compute the gradient of g (an exercise):

$$\nabla g(a) = -2V^{\mathrm{T}} y + 2V^{\mathrm{T}} V a.$$

It follows that the only point where the gradient of g vanishes is the one satisfying (1.9). As g is smooth and its minimum cannot obviously be attained at 'infinity', the critical point defined by (1.9) must coincide with the unique solution of (1.8). (Actually, one could easily show that the *Hessian* of g is the constant, positive definite matrix $2V^{\mathrm{T}}V$, and so the fact that the solution of (1.8) is the unique minimizer of g follows from basic optimization theory.) □

1.3. Fourier series (a recap)

Let us recall some basic concepts related to Fourier series. The Fourier coefficients of a square integrable function f are defined by the formula

$$\hat{f}(j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\, e^{-ijt} dt, \qquad j \in \mathbb{Z}. \tag{1.10}$$

They may be interpreted as (scaled) scalar projections of f on the orthonormal basis

$$\frac{1}{\sqrt{2\pi}}\, e^{ijt}, \qquad j \in \mathbb{Z},$$

of the space of square integrable functions

$$L^2([-\pi,\pi]) := \left\{ g : [-\pi,\pi] \to \mathbb{C} \;\middle|\; \int_{-\pi}^{\pi} |f(t)|^2 < \infty \right\}.$$

The Fourier series

$$\sum_{j=-\infty}^{\infty} \hat{f}(j)\, e^{ijt}$$

converges towards $f \in L^2([-\pi,\pi])$ in the sense that

$$\Big\| f - \sum_{j=-n}^{n} \hat{f}(j)\, e^{ij\,\cdot} \Big\|_{L^2([-\pi,\pi])} = \Big(\int_{-\pi}^{\pi} \Big| f(t) - \sum_{j=-n}^{n} \hat{f}(j)\, e^{ijt} \Big|^2 dt \Big)^{1/2} \longrightarrow 0$$

as n goes to infinity. In addition, it is known that

$$f(t) = \lim_{n\to\infty} \sum_{j=-n}^{n} \hat{f}(j)\, e^{ijt} =: \sum_{j=-\infty}^{\infty} \hat{f}(j)\, e^{ijt}$$

if f is differentiable at t.

The real form of the Fourier series is

$$\sum_{j=-\infty}^{\infty} \hat{f}(j)\, e^{ijt} = \frac{a_0}{2} + \sum_{j=1}^{\infty} \Big(a_j \cos(jt) + b_j \sin(jt) \Big),$$

where

$$a_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\, \cos(jt)\, dt \qquad \text{and} \qquad b_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\, \sin(jt)\, dt.$$

It holds that

$$\begin{cases} a_j = \hat{f}(j) + \hat{f}(-j), & j \geq 0, \\ b_j = i\big(\hat{f}(j) - \hat{f}(-j)\big), & j \geq 1, \end{cases} \qquad \text{and} \qquad \hat{f}(j) = \begin{cases} \frac{1}{2}(a_j - ib_j), & j > 0, \\ \frac{1}{2}a_0, & j = 0, \\ \frac{1}{2}(a_j + ib_j), & j < 0. \end{cases}$$

The so-called *Fourier cosine and sine series* are needed in many applications. To define these concepts, we consider a square integrable function $\tilde{f} \in L^2([0,\pi])$ and introduce the coefficients

$$\tilde{a}_j = \frac{2}{\pi}\int_0^\pi \tilde{f}(t)\,\cos(jt)\,dt \qquad \text{and} \qquad b_j = \frac{2}{\pi}\int_0^\pi \tilde{f}(t)\,\sin(jt)\,dt.$$

The Fourier cosine and sine series of $\tilde{f}$ are

$$\frac{\tilde{a}_0}{2} + \sum_{j=1}^{\infty} \tilde{a}_j \cos(jt) \qquad \text{and} \qquad \sum_{j=1}^{\infty} \tilde{b}_j \sin(jt),$$

respectively. It can be shown that these converge towards $\tilde{f}$ in the norm of $L^2([0,\pi])$.

Theorem 1.5. *The Fourier cosine and sine series of $\tilde{f} \in L^2([0,\pi])$ converge to $\tilde{f}$ in the topology of $L^2([0,\pi])$, that is,*

$$\left\| \tilde{f} - \Big(\frac{\tilde{a}_0}{2} + \sum_{j=1}^{n} \tilde{a}_j \cos(j\,\cdot)\Big) \right\|_{L^2([0,\pi])} \longrightarrow 0, \tag{1.11}$$

$$\left\| \tilde{f} - \sum_{j=1}^{n} \tilde{b}_j \sin(j\,\cdot) \right\|_{L^2([0,\pi])} \longrightarrow 0$$

as n goes to infinity.

PROOF. We start by considering (1.11) and continuing $\tilde{f}$ as an even function to the interval $[-\pi,\pi]$,

$$f(t) = \begin{cases} \tilde{f}(t), & t \in [0,\pi], \\ \tilde{f}(-t), & t \in [-\pi,0). \end{cases}$$

Clearly, $f \in L^2([-\pi,\pi])$, and so

$$\left\| f - \Big(\frac{a_0}{2} + \sum_{j=1}^{n} \Big(a_j \cos(j\,\cdot) + b_j \sin(j\,\cdot)\Big)\Big) \right\|_{L^2([-\pi,\pi])} \longrightarrow 0, \tag{1.12}$$

as $n \to \infty$. Since f and $\cos(j\,\cdot)$ are both even functions, their product is also even, and we have

$$a_j = \frac{1}{\pi}\int_{\pi}^{\pi} f(t)\,\cos(jt)\,dt = \frac{2}{\pi}\int_0^\pi \tilde{f}(t)\,\cos(jt)\,dt = \tilde{a}_j, \qquad j = 0,1,2,\dots\ .$$

On the other hand, as $\sin(j\,\cdot)$ is odd, its product with f is odd, leading to

$$b_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\,\sin(jt)\,dt = 0, \qquad j = 1,2,\dots\ .$$

In consequence, it follows from (1.12) that

$$\left\| f - \Big(\frac{\tilde{a}_0}{2} + \sum_{j=1}^{n} \tilde{a}_j \cos(j\,\cdot)\Big) \right\|_{L^2([-\pi,\pi])} \longrightarrow 0$$

when $n \to \infty$. By restricting the attention to the subinterval $[0,\pi]$, we have proved (1.11).

The claim for the sine series can be proved in a similar manner by continuing $\tilde{f}$ as an odd function, i.e., introducing

$$f(t) = \begin{cases} \tilde{f}(t), & t \in [0, \pi], \\ -\tilde{f}(-t), & t \in [-\pi, 0), \end{cases}$$

for which $a_j = 0$, $j = 0, 1, \ldots$, and $b_j = \tilde{b}_j$, $j = 1, 2, \ldots$. □

1.4. Discrete Fourier transform (DFT)

In many applications (e.g., signal or image processing), one wants to compute the Fourier coefficients for some underlying function f. This can rarely be achieved by directly evaluating the integrals (1.10). There are at least two reasons:

(i) The explicit form of f is often unknown; the available information is only a set of pointwise values $f(t_0), f(t_1), \ldots, f(t_{n-1})$.
(ii) Even if f is known explicitly, it is typically impossible to represent its Fourier coefficients with the help of elementary functions.

As a consequence, it is essential to consider how Fourier coefficients can be approximated numerically.

A finite number of Fourier coefficients can be estimated by means of numerical integration, e.g., with the help of quadrature rules such as the *trapezoidal rule*. An alternative approach is to employ interpolation: Let us assume that $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and differentiable, in which case it is known that

$$f(t) = \sum_{j=-\infty}^{\infty} \hat{f}(j)\, e^{ijt} \tag{1.13}$$

for all $t \in \mathbb{R}$. Moreover, suppose that the values of f are known on the equidistant grid

$$t_k = k\frac{2\pi}{n}, \qquad k = 0, 1, \ldots, n-1, \quad n \in \mathbb{N}, \tag{1.14}$$

over one period of f.

Remark 1.6. The points (1.14) form a uniform mesh over the interval $[0, 2\pi]$ — excluding the right end point 2π, which would not provide any extra information as $f(0) = f(2\pi)$. This is a traditional choice for the *discrete Fourier transform* (DFT).

The leading idea of DFT is to find n coefficients d_j, $j = 0, 1, \ldots, n-1$, such that for $k = 0, 1, \ldots, n-1$,

$$f(t_k) = \sum_{j=0}^{n-1} d_j\, e^{ijt_k} = \sum_{j=0}^{n-1} d_j\, x_k^j = d_0 + d_1 x_k + d_2 x_k^2 + \cdots + d_{n-1} x_k^{n-1}, \tag{1.15}$$

where

$$x_k := e^{it_k} = e^{ik\frac{2\pi}{n}} = (e^{i\frac{2\pi}{n}})^k = w^k$$

with the definition $w := e^{i\frac{2\pi}{n}}$. Obviously, (1.15) defines a polynomial interpolation problem with respect to the points $\mathbb{C} \ni x_k = w^k$, $k = 0, 1, \ldots, n-1$. Since t_k, $k = 0, 1, \ldots, n-1$, are distributed uniformly over $[0, 2\pi]$, the points x_k, $k = 0, 1, \ldots, n-1$, form an equidistant grid on the unit circle in the complex plane. In

particular, x_k, $k = 0, 1, \dots, n-1$, are distinct and it follows from Lemma 1.1 that (1.15) uniquely determines the coefficients $d := [d_0, d_1, \dots, d_{n-1}]^{\mathrm{T}} \in \mathbb{C}^n$. Moreover, d can be solved from the equation

$$F\,d = y, \tag{1.16}$$

where $y := [f(t_0), f(t_1), \dots, f(t_{n-1})]^{\mathrm{T}} \in \mathbb{C}^n$ and F is a Vandermonde matrix,

$$F = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \dots & x_{n-1}^{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & w^1 & w^2 & \dots & w^{n-1} \\ 1 & w^2 & w^{2\cdot 2} & \dots & w^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{n-1} & w^{2(n-1)} & \dots & w^{(n-1)^2} \end{bmatrix}.$$

Elementwise, F is defined as

$$F_{jk} = w^{(j-1)(k-1)}, \qquad j, k = 1, 2, \dots, n.$$

The inverse of F can also be given explicitly.

Lemma 1.7. *It holds that*

$$F^{-1} = \frac{1}{n}\overline{F},$$

where $\bar{\cdot}$ denotes elementwise complex conjugation.

PROOF. According to the definition of matrix multiplication,

$$(F\overline{F})_{jk} = \sum_{l=1}^{n} F_{jl}\overline{F}_{lk} = \sum_{l=1}^{n} w^{(j-1)(l-1)}\overline{w}^{(l-1)(k-1)}.$$

Because $\overline{w} = e^{-i\frac{2\pi}{n}} = w^{-1}$, we get

$$(F\overline{F})_{jk} = \sum_{l=1}^{n} w^{(l-1)((j-1)-(k-1))} = \sum_{l=1}^{n} (w^{(j-k)})^{(l-1)}.$$

In particular,

$$(F\overline{F})_{jj} = \sum_{l=1}^{n} 1^{(l-1)} = \sum_{l=1}^{n} 1 = n, \qquad j = 1, 2, \dots, n.$$

On the other hand, the formula for a finite geometric series gives for $j \neq k$,

$$(F\overline{F})_{jk} = \sum_{l=0}^{n-1} (w^{(j-k)})^l = \frac{1 - (w^{(j-k)})^n}{1 - w^{(j-k)}} = \frac{1 - e^{i(j-k)2\pi}}{1 - e^{i\frac{(j-k)}{n}2\pi}} = \frac{1-1}{1 - e^{i\frac{(j-k)}{n}2\pi}} = 0,$$

where the denominator does not vanish as $j - k$ cannot be a multiple of n.

Altogether, we have proved that

$$F\overline{F} = nI \iff F\Big(\frac{1}{n}\overline{F}\Big) = I,$$

where $I \in \mathbb{C}^{n\times n}$ is the identity matrix. This completes the proof. □

Let us summarize our findings: If $y := [f(t_0), f(t_1), \dots, f(t_{n-1})]^{\mathrm{T}} \in \mathbb{C}^n$, then the DFT coefficients $d = [d_0, d_1, \dots, d_{n-1}]^{\mathrm{T}} \in \mathbb{C}^n$ can be computed via

$$d = \frac{1}{n}\overline{F}\, y = \frac{1}{n}\begin{bmatrix} 1 & \overline{x}_0 & \overline{x}_0^2 & \dots & \overline{x}_0^{n-1} \\ 1 & \overline{x}_1 & \overline{x}_1^2 & \dots & \overline{x}_1^{n-1} \\ 1 & \overline{x}_2 & \overline{x}_2^2 & \dots & \overline{x}_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \overline{x}_{n-1} & \overline{x}_{n-1}^2 & \dots & \overline{x}_{n-1}^{n-1} \end{bmatrix} y,$$

where $x_j = e^{it_j} = e^{ij\frac{2\pi}{n}}$. Taking into account the complex conjugations, this can be equivalently stated as

$$d_j = \frac{1}{n}\sum_{k=0}^{n-1} f(t_k)\,(e^{-it_j})^k = \frac{1}{n}\sum_{k=0}^{n-1} f\Big(\tfrac{2\pi k}{n}\Big)\, e^{-ijk\frac{2\pi}{n}}, \qquad j = 0, 1 \dots, n-1. \tag{1.17}$$

The vector $d \in \mathbb{C}^n$ is called the *discrete Fourier transform* (DFT) of the 'signal' $y \in \mathbb{C}^n$. Observe that the *inverse discrete Fourier transform*, i.e., the point values of f (or the signal y) given d, can be computed using the original interpolation formula (1.15) or equivalently the matrix identity (1.16).

Although (1.15) resembles the Fourier series expansion (1.13) of a smooth enough function, it is not yet quite obvious how the Fourier coefficients $\{\hat{f}_j\}_{j\in\mathbb{Z}} \subset \mathbb{C}$ and the DFT $d \in \mathbb{C}^n$ are related. This imperfection is fixed by the following theorem.

Theorem 1.8. *If f is differentiable and 2π-periodic, then*

$$d_j = \sum_{l=-\infty}^{\infty} \hat{f}(j + ln), \qquad j = 0, 1, \dots, n-1, \tag{1.18}$$

where $d \in \mathbb{C}^n$ is the discrete Fourier transform of f.

PROOF. As it is known that the Fourier series of a differentiable function converges pointwise, we have

$$f(t_k) = \sum_{m=-\infty}^{\infty} \hat{f}(m)\, e^{imt_k}, \qquad k = 0, 1, \dots, n-1.$$

Substituting this representation of $f(t_k)$ in (1.17) results in

$$d_j = \frac{1}{n}\sum_{k=0}^{n-1}\Big(\sum_{m=-\infty}^{\infty} \hat{f}(m)\, e^{imt_k}\Big) e^{-ikt_j} = \frac{1}{n}\sum_{m=-\infty}^{\infty} \hat{f}(m) \sum_{k=0}^{n-1} e^{ik(m-j)\frac{2\pi}{n}}, \tag{1.19}$$

where the second step is a consequence of the definition of t_k in (1.14).

Let us deduce the value of the inner sum in (1.19),

$$s_m := \sum_{k=0}^{n-1} e^{ik(m-j)\frac{2\pi}{n}},$$

by dividing the possible values of $m \in \mathbb{Z}$ into two subsets. Assume first that $m = j + ln$ for some $l \in \mathbb{Z}$, which leads to

$$s_m = \sum_{k=0}^{n-1} e^{ikl2\pi} = \sum_{k=0}^{n-1} 1 = n.$$

Next, suppose that $m \neq j + ln$ for all $l \in \mathbb{Z}$, which means that there exists an integer $0 < r < n$ such that $m = j + ln + r$ for some $l \in \mathbb{Z}$. In consequence,

$$s_m = \sum_{k=0}^{n-1} e^{ik(ln+r)\frac{2\pi}{n}} = \sum_{k=0}^{n-1} e^{ikl2\pi}\, e^{ikr\frac{2\pi}{n}} = \sum_{k=0}^{n-1} (e^{ir\frac{2\pi}{n}})^k = \frac{1-(e^{ir\frac{2\pi}{n}})^n}{1-e^{ir\frac{2\pi}{n}}} = \frac{1-e^{ir2\pi}}{1-e^{ir\frac{2\pi}{n}}} = 0,$$

where we once again used the formula for a finite geometric series.

Hence, one only needs to account for those $m \in \mathbb{Z}$ that equal "j modulo n", i.e., those m for which $m - j$ is divisible by n, in the exterior sum of (1.19):

$$d_j = \frac{1}{n} \sum_{m=-\infty}^{\infty} \hat{f}(m)\, s_m = \frac{1}{n} \sum_{l=-\infty}^{\infty} n\hat{f}(j+ln) = \sum_{l=-\infty}^{\infty} \hat{f}(j+ln),$$

which completes the proof. □

Let us try to interpret the result of the above theorem: If the examined function f is 'nice', its Fourier coefficients $\{\hat{f}_j\}_{j\in\mathbb{Z}}$ converge quickly to zero when the absolute value of the frequency parameter $j \in \mathbb{Z}$ increases. If $n \in \mathbb{N}$ is not very small, it is thus reasonable to assume that in the sum (1.18) the greatest contribution comes from the Fourier coefficient with the smallest index in the sense of absolute value.

Hence, the first 'half' of the DFT coefficients satisfy

$$d_j \approx \hat{f}(j), \qquad j = 0, 1, \ldots, \frac{n}{2} - 1,$$

where and in the following we assume for simplicity that $n \in \mathbb{N}$ is even, while for the latter ones it holds that

$$d_j \approx \hat{f}(j-n), \qquad j = \frac{n}{2} + 1, \frac{n}{2} + 2, \ldots, n-1.$$

Finally, the DFT coefficient 'in the middle' contains equal amount of information about two different Fourier coefficients:

$$d_{\frac{n}{2}} \approx \hat{f}\Big(\frac{n}{2}\Big) + \hat{f}\Big(-\frac{n}{2}\Big).$$

The other way around,

$$\hat{f}(j) \approx d_j, \qquad j = 0, 1, \ldots, \frac{n}{2} - 1,$$

and

$$\hat{f}(-j) \approx d_{n-j}, \qquad j = 1, 2, \ldots, \frac{n}{2} - 1.$$

To sum up, the (low) positive Fourier frequencies are in their 'correct places' in the DFT vector, whereas the (low) negative frequencies are in the tail end of d.

Remark 1.9. Usually, the Fourier coefficients $\hat{f}(j)$ with $|j| \ll \frac{n}{2}$ can be estimated more accurately by the DFT than those for which $|j|$ is only slightly smaller than $\frac{n}{2}$. As an example, according to (1.18), the second lowest Fourier frequency (in absolute value) contributing to d_1 is $-n+1$, i.e. pretty high, while

for $d_{\frac{n}{2}-1}$ it is $-\frac{n}{2}-1$, which approximately as high as the 'primary frequency' $\frac{n}{2}-1$.

In practice, the DFT is almost always computed by means of the *fast Fourier transform* (FFT). If the DFT coefficients are computed using the matrix identity $d = \frac{1}{n}\overline{F}y$, the number of needed floating point operations behaves like $O(n^2)$ as a function of n. The FFT algorithm takes advantage of the special structure of $\overline{F}$ and computes the DFT in $O(n\log(n))$ operations. If n is large, as it is often in practical applications, the speed-up is *considerable*.

An implementation of the FFT algorithm can be found in practically all computational software, such as MATLAB; the details of the algorithm are not discussed on this course.

CHAPTER 2

Numerical solution of ordinary differential equations

The solution of an *initial value problem* (IVP) for an *ordinary differential equation* (ODE) cannot usually be written down explicitly, meaning that one must often resort to numerical solution techniques in practice. It is anyway good to know the conditions under which a unique solution exists.

2.1. Existence and uniqueness (counterexamples)

Let us consider the IVP

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0. \tag{2.1}$$

Here, $t \in \mathbb{R}$ may be interpreted as the time, $x(t) \in \mathbb{R}^n$ describes the state of the examined system at the time $t \in \mathbb{R}$, the *initial value* x_0 gives the state of the system at the (fixed) initial time $t_0 \in \mathbb{R}$, and the function $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ gives the time derivative of the system as a function of the time and the state of the system. (The domain of definition for f is not always the whole of $\mathbb{R} \times \mathbb{R}^n$ but it depends on the specific application.)

It seems believable that knowing the initial value x_0 and the 'rate of change' f constitutes enough information for uniquely determining the state of the system, $x(t)$, for any $t \geq t_0$ (or $t \leq t_0$). However, the situation is not quite this simple: both the existence and the uniqueness of a solution to (2.1) require some regularity from $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$.

Example 2.1. Suppose $n = 1$, $t_0 = 0$, $x_0 = 0$ and consider

$$f(t, x) = f(x) = \begin{cases} 1, & x < 0, \\ -1, & x \geq 0. \end{cases}$$

Then the IVP (2.1) has no solution. The proof of this claim is left as an exercise. (Hint: Convince yourself that the 'solution curve' cannot move away from the initial value, i.e. from zero, but the trivial function $x(t) \equiv 0$ is not a solution.)

The source for the 'nonexistence' problem encountered in Example 2.1 is the discontinuity of f. In fact, the *Peano existence theorem* states that the IVP (2.1) has a (local) solution around the initial time $t = t_0$ if f is continuous. However, this solution does not need to be unique:

Example 2.2. Assume again that $n = 1$, $t_0 = 0$, $x_0 = 0$, but this time around set

$$f(t, x) = f(x) = 3|x|^{2/3},$$

which is continuous everywhere in $\mathbb{R}$. In this case, the IVP (2.1) has an infinite number of solutions. Indeed, let us introduce the family of functions

$$x_{a,b}(t) = \begin{cases} (t-a)^3, & t < a, \\ 0, & a \le t \le b, \\ (t-b)^3, & t > b, \end{cases} \tag{2.2}$$

where the parameters satisfy $a \le 0$ and $b \ge 0$. Obviously, $x_{a,b}(0) = 0$ and straightforward calculations give

$$\begin{aligned} x'_{a,b}(t) &= 3(t-a)^2 = 3|t-a|^2 = 3|x_{a,b}(t)|^{2/3}, && t < a, \\ x'_{a,b}(t) &= 0 = 3|x_{a,b}(t)|^{2/3}, && a < t < b, \\ x'_{a,b}(t) &= 3(t-b)^2 = 3|t-b|^2 = 3|x_{a,b}(t)|^{2/3}, && t > b. \end{aligned}$$

Finally, checking the points $t = a$ and $t = b$ separately (e.g., draw a picture) yields

$$x'_{a,b}(a) = x'_{a,b}(b) = 0 = 3|x_{a,b}(a)|^{2/3} = 3|x_{a,b}(b)|^{2/3}.$$

To sum up, the formula (2.2) defines an infinite family of solutions for the IVP parametrized by a and b.

2.2. Picard–Lindelöf iteration and unique solvability

In this section, it will be shown that a *sufficient* condition for the unique solvability of (2.1) over some time interval $I := [t_-, t_+] \ni t_0$ is that $f : [t_-, t_+] \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies the *Lipschitz condition*

$$|f(t,x) - f(t,y)| \le L|x-y| \tag{2.3}$$

for some $L > 0$ and for all $t \in I$, $x, y \in \mathbb{R}^n$. We assume that this is the case for the rest of this chapter.

Suppose that the IVP (2.1) has a (continuously differentiable) solution and integrate both sides of the differential equation over the interval $[t_0, t]$:

$$\int_{t_0}^{t} x'(s)\,ds = \int_{t_0}^{t} f(s, x(s))\,ds.$$

Taking into account the initial condition of (2.1), we get

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\,ds. \tag{2.4}$$

Hence, if $x : \mathbb{R} \to \mathbb{R}^n$ is a solution of the IVP (2.1), then it also satisfies (2.4). On the other hand, any solution of (2.4) obviously has the property

$$x(t_0) = x_0 + \int_{t_0}^{t_0} f(s, x(s))\,ds = x_0,$$

and differentiating the integral identity (2.4) — recalling the *fundamental theorem of calculus* — results in

$$x'(t) = f(t, x(t)).$$

As a consequence, the IVP (2.1) and the integral equation (2.4) are equivalent. This means, in particular, that if (2.4) has a unique solution, then the same applies to (2.1).

Our aim is to employ the *Picard–Lindelöf iteration* to show that (2.4) has a unique solution on the interval I under our assumptions on f. The Picard–Lindelöf iteration produces a sequence of continuous functions $x^j : I \to \mathbb{R}^n$, $j = 0, 1, \dots$, as follows:

(1) Choose the 'initial guess' $x^0 : I \to \mathbb{R}^n$ to be the constant function $x^0 \equiv x_0$, where $x_0 \in \mathbb{R}^n$ is the initial value.
(2) Given $x^k : I \to \mathbb{R}^n$, define the next iterate $x^{k+1} : I \to \mathbb{R}^n$ by the formula

$$x^{k+1}(t) = x_0 + \int_{t_0}^{t} f(s, x^k(s))\, ds, \qquad t \in I.$$

We will prove that the Picard–Lindelöf iteration converges to a solution of (2.4) as $k \to \infty$.

Lemma 2.3. *The Picard–Lindelöf iteration converges uniformly to a continuous function $x : I \to \mathbb{R}^n$, that is,*

$$\max_{t \in I} |x(t) - x^k(t)| \longrightarrow 0,$$

as $k \to \infty$.

PROOF. Let us write the Picard–Lindelöf iterates in the form

$$x^k(t) = x^0(t) + S^k(t), \qquad k = 1, 2, \dots, \tag{2.5}$$

where

$$S^k(t) = \sum_{j=1}^{k} \Big(x^j(t) - x^{j-1}(t) \Big).$$

The idea is to prove the uniform convergence of the iteration with the help of the *Weierstrass M-test*: If it is shown that $|x^j(t) - x^{j-1}(t)| \le M_j \in \mathbb{R}$ for all $t \in I$ and

$$\sum_{j=1}^{\infty} M_j < \infty, \tag{2.6}$$

then the function sequence S^k, $k = 1, 2, \dots$, converges uniformly over I to some limit function S. In this case, x^k converges uniformly to $x = x^0 + S$ due to (2.5).

Let us prove by induction that

$$|x^j(t) - x^{j-1}(t)| \le \frac{1}{j!} K L^{j-1} |t - t_0|^j, \qquad t \in I, \tag{2.7}$$

where

$$K = \max_{t \in I} |f(t, x^0(t))| = \max_{t \in I} |f(t, x_0)|.$$

First of all,

$$|x^1(t) - x^0(t)| = \left| \int_{t_0}^{t} f(s, x^0(s))\, ds \right| \le \int_{t_0}^{t} |f(s, x^0(s))|\, ds \le K |t - t_0|,$$

which proves (2.7) for $j = 1$.

Suppose then that (2.7) holds for some $j = l \in \mathbb{N}$. We have

$$\begin{aligned} |x^{l+1}(t) - x^l(t)| &= \Big| \int_{t_0}^{t} f(s, x^l(s)) - f(s, x^{l-1}(s))\, ds \Big| \\ &\le \int_{t_0}^{t} |f(s, x^l(s)) - f(s, x^{l-1}(s))|\, ds \\ &\le \int_{t_0}^{t} L\, |x^l(s) - x^{l-1}(s)|\, ds \\ &\le \frac{1}{l!} K L^l \int_{t_0}^{t} |s - t_0|^l\, ds, \qquad t \in I, \end{aligned}$$

where the last two steps follow from the Lipschitz condition (2.3) and the assumption that (2.7) holds for $j = l$, respectively. Considering the cases $t > t_0$ and $t < t_0$ separately and integrating the polynomials $(s - t_0)^l$ and $(t_0 - s)^l$ over the interval $[t_0, t]$, we easily get

$$|x^{l+1}(t) - x^l(t)| \le \frac{1}{(l+1)!} K L^l |t - t_0|^{l+1},$$

which completes the inductive step. In consequence, (2.7) hold for any $j \in \mathbb{N}$.

In particular, since $t, t_0 \in I = [t_-, t_+]$,

$$|x^j(t) - x^{j-1}(t)| \le \frac{1}{j!} K L^{j-1} (t_+ - t_-)^j =: M_j \qquad \text{for all } t \in I,\ j \in \mathbb{N}. \tag{2.8}$$

Hence, the uniform convergence of the sequence S^k, $k = 1, 2, \ldots$, follows from the Weierstrass M-test if we show that (2.6) holds for M_j, $j = 1, 2, \ldots$, defined by (2.8):

$$\sum_{j=1}^{\infty} M_j = \frac{K}{L} \sum_{j=1}^{\infty} \frac{1}{j!} L^j (t_+ - t_-)^j = \frac{K}{L} \Big(\sum_{j=0}^{\infty} \frac{1}{j!} L^j (t_+ - t_-)^j - 1 \Big) = \frac{K}{L} (e^{L(t_+ - t_-)} - 1) < \infty.$$

To finalize the proof, we note that according to the *Uniform limit theorem*, if a sequence of continuous functions converges uniformly on I to some limit function, then this limit is also continuous. □

Let us return to the definition of the Picard–Lindelöf iteration,

$$x^{k+1}(t) = x_0 + \int_{t_0}^{t} f(s, x^k(s))\, ds, \qquad k = 0, 1, \ldots\ . \tag{2.9}$$

Because of Lemma 2.3, the left-hand side of (2.9) converges toward the continuous limit function $x(t)$ for all $t \in I$ as $k \to \infty$. On the other hand, for the right-hand side of (2.9) it holds that

$$\begin{aligned} \Big| \int_{t_0}^{t} f(s, x(s))\, ds - \int_{t_0}^{t} f(s, x^k(s))\, ds \Big| &\le \int_{t_0}^{t} |f(s, x(s)) - f(s, x^k(s))|\, ds \\ &\le L \int_{t_0}^{t} |x(s) - x^k(s)|\, ds \\ &\le L\, |t - t_0| \max_{s \in I} |x(s) - x^k(s)| \longrightarrow 0 \end{aligned}$$

as $k \to \infty$ due to Lemma 2.3. Consequently, taking the limit of (2.9) when k goes to infinity leads to the equation

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds, \qquad t \in I,$$

that is, the continuous limit function of the Picard–Lindelöf iteration satisfies (2.4), and thus also (2.1).

Summarizing, we know that under the assumption that $f : I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies the Lipschitz condition (2.3), the IVP (2.1) has at least one solution over the interval I. To be completely satisfied, we still need to show that this solution is unique. A tool suited for this task is the Gronwall inequality.

Lemma 2.4 (Gronwall inequality)**.** *Let $u : [0, T] \to \mathbb{R}$ be continuous, nonnegative and satisfy*

$$u(t) \le C + K \int_0^t u(s)\, ds$$

for some $C, K \ge 0$ and all $t \in [0, T]$. Then it holds that

$$u(t) \le Ce^{Kt}$$

for all $t \in [0, T]$.

PROOF. To begin with, assume that $C > 0$ and define

$$v(t) := C + K \int_0^t u(s)\, ds,$$

whence

$$u(t) \le v(t) > 0, \qquad t \in [0, T].$$

Due to the fundamental theorem of calculus, we have

$$\frac{d}{dt} \ln(v(t)) = \frac{v'(t)}{v(t)} = K \frac{u(t)}{v(t)} \le K, \qquad t \in [0, T]. \tag{2.10}$$

Integrating (2.10) over the interval $[0, t]$ results in

$$\ln(v(t)) - \ln(v(0)) \le Kt \iff \ln(v(t)) = \ln(C) + Kt.$$

Since the exponential function is monotonically increasing,

$$u(t) \le v(t) \le e^{\ln(C)+Kt} = Ce^{Kt}, \qquad t \in [0, T],$$

which proves the claim when $C > 0$.

If $C = 0$, the preceding line of reasoning can be applied with any $C = \epsilon > 0$, meaning that

$$u(t) \le \epsilon e^{Kt} \longrightarrow 0,$$

for all $t \in [0, T]$ as $\epsilon \to 0$. Hence, it must hold that $u \equiv 0$, which completes the proof. □

Now we are ready to complete the proof of unique solvability for (2.1).

Theorem 2.5. *Assume that $f : I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies the Lipschitz condition (2.3). Then, the IVP (2.1) has a unique solution $x : I \to \mathbb{R}^n$ on the interval I.*

PROOF. By virtue of the Picard–Lindelöf iteration, we have already established that there exists at least one solution for (2.1) on I.

Let us thus assume that $x : I \to \mathbb{R}^n$ and $y : I \to \mathbb{R}^n$ are both (continuous) solutions of (2.1), which is equivalent to

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds \qquad \text{and} \qquad y(t) = x_0 + \int_{t_0}^{t} f(s, y(s))\, ds, \qquad t \in I.$$

Subtracting these two identities gives

$$|x(t) - y(t)| \le \int_{t_0}^{t} |f(s, x(s)) - f(s, y(s))|\, ds \le L \int_{t_0}^{t} |x(s) - y(s)|\, ds, \qquad t \in I. \tag{2.11}$$

where we used (2.3).

Assume that $t \ge t_0$, set $r = t - t_0 \ge 0$ and denote

$$u(\tau) = |x(t_0 + \tau) - y(t_0 + \tau)|.$$

It follows from (2.11) that

$$u(r) \le L \int_{t_0}^{t} |x(s) - y(s)|\, ds = L \int_{0}^{t-t_0} |x(t_0 + \tau) - y(t_0 + \tau)|\, d\tau = L \int_{0}^{r} u(\tau)\, d\tau,$$

where we employed the change of variables $\tau = s - t_0$. Because $u : [0, t_+ - t_0] \to \mathbb{R}$ is clearly nonnegative and continuous, it follows from the Gronwall inequality that

$$u(r) = 0, \quad r \in [0, t_+ - t_0] \quad \Longleftrightarrow \quad x(t) = y(t), \quad t \in [t_0, t_+],$$

which shows the uniqueness of the solution for (2.1) over the right half of I, i.e. $[t_0, t_+]$. The claim for the left half $[t_-, t_0]$ can be proved by a similar argument. □

2.3. Continuous dependence on the initial value

When considering the accuracy of numerical methods for solving IVPs of the type (2.1), it is essential to have information about how the solution depends on the initial value x_0 (or error in this initial value). To simplify the notation, we start by introducing a new definition.

Definition 2.6. The solution map $\psi : I \times I \times \mathbb{R}^n \to \mathbb{R}^n$ associated to the ODE

$$x'(t) = f(t, x(t)) \tag{2.12}$$

is defined via

$$\psi(t, \tau, u) = x(t),$$

where $x(t)$ is the unique solution of (2.12) with the initial condition $x(\tau) = u \in \mathbb{R}^n$.

The solution map characterizes the dependence of the solution to (2.12) on (i) the time, (ii) the initial time, and (iii) the initial value. It is easy to see that the following identities are valid

$$\frac{\partial}{\partial t}\psi(t, \tau, u) = f(t, \psi(t, \tau, u)),$$
$$\psi(\tau, \tau, u) = u.$$

Moreover, it follows from the uniqueness of the solution that

$$\psi\Big(t, s, \psi(s, \tau, u)\Big) = \psi(t, \tau, u), \tag{2.13}$$

which can be interpreted as follows: If one starts from the state u at time τ, follows the solution curve until time s, gives the 'current state' $\psi(s,\tau,u)$ as a new initial value at time s, and finally continues along the solution curve up to time t, the end result is the same as the one produced by starting from u at time τ and following the solution curve directly until time t. In other words, the 'intermediate stop' at time s has no effect.

The following theorem proves the continuous dependence of the solution to (2.1) on the initial data.

Theorem 2.7. *Assume that $f : I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies the Lipschitz condition (2.3). Then, it holds that*

$$|\psi(t,t_0,x_0) - \psi(t,t_0,y_0)| \le e^{L|t-t_0|}|x_0 - y_0| \tag{2.14}$$

for all $t \in I$.

PROOF. Let $x : I \to \mathbb{R}^n$ and $y : I \to \mathbb{R}^n$ be the solutions of (2.12) with the initial conditions $x(t_0) = x_0$ and $y(t_0) = y_0$, that is,

$$x'(t) = f(t,x(t)), \quad x(t_0) = x_0,$$

and

$$y'(t) = f(t,y(t)), \quad y(t_0) = y_0.$$

In other words, $x(t) = \psi(t,t_0,x_0)$ and $y(t) = \psi(t,t_0,y_0)$.

We denote

$$u(\tau) = |x(t_0+\tau) - y(t_0+\tau)|^2, \qquad \tau \in [0, t_+ - t_0].$$

Straightforward differentiation yields

$$\begin{aligned} u'(\tau) &= 2\Big(x'(t_0+\tau) - y'(t_0+\tau)\Big) \cdot \Big(x(t_0+\tau) - y(t_0+\tau)\Big) \\ &= 2\Big(f(t_0+\tau, x(t_0+\tau)) - f(t_0+\tau, y(t_0+\tau))\Big) \cdot \Big(x(t_0+\tau) - y(t_0+\tau)\Big) \\ &\le 2\Big|f(t_0+\tau, x(t_0+\tau)) - f(t_0+\tau, y(t_0+\tau))\Big| \Big|x(t_0+\tau) - y(t_0+\tau)\Big| \\ &\le 2L\Big|x(t_0+\tau) - y(t_0+\tau)\Big|^2 \le 2L\,u(\tau). \end{aligned}$$

Integrating this inequality over the interval $[0,s]$ leads to

$$0 \le u(s) \le u(0) + 2L\int_0^s u(\tau)\,d\tau.$$

Taking into account that $u(0) = |x_0 - y_0|^2$, the Gronwall inequality guarantees that

$$u(s) \le |x_0 - y_0|^2 e^{2Ls}, \qquad s \in [0, t_+ - t_0].$$

Now, the claim for $t \in [t_0, t_+]$ follows by choosing $s = t - t_0$ and taking the square root.

The case $t \in [t_-, t_0]$ can be handled by a similar argument. □

2.4. Explicit solution methods

Suppose the solution of the ODE

$$x'(t) = f(t,x(t)) \tag{2.15}$$

is known at time $t = \tau \in \mathbb{R}$ and let $h > 0$ be a given time step. Our preliminary aim is to approximate the value $x(\tau + h)$.

Assume that the solution $x : \mathbb{R} \to \mathbb{R}^n$ of (2.15) is so regular[1] that we can write a first order Taylor's expansion around τ:

$$x(\tau + h) = x(\tau) + hx'(\tau) + O(h^2) = x(\tau) + hf(\tau, x(\tau)) + O(h^2), \tag{2.16}$$

where $O(h^2)$ is an error term that satisfies

$$|O(h^2)| \leq Ch^2,$$

with a fixed $C > 0$ for any small enough $h > 0$. (In fact, C depends on the size of x'' in the vicinity of τ.) By dropping out the error term from (2.16), one gets the simplest numerical method for solving ODEs.

For simplicity, we assume that the initial condition is given at time $t_0 = 0$, i.e. $x(0) = x_0$, and define a set of grid points by

$$t_j = jh, \qquad j = 0, 1, 2, \dots .$$

Definition 2.8 (Euler's method)**.** We define the approximations $x_j \approx x(t_j) \in \mathbb{R}^n$, $j = 1, 2, \dots$, via

$$x_{j+1} = x_j + hf(t_j, x_j), \qquad j = 0, 1, \dots,$$

which gives the Euler's method.[2]

Since in a single step of the Euler's method one ignores an error term of the size $O(h^2)$, the *local truncation error* of the method is proportional to the square of the step size. In what follows, it will be shown that this leads to a *global error* of the form $O(h)$ due to the accumulation of the local errors (and the continuous dependence on the initial data).

Obviously, it would be nice to find numerical methods for which the local truncation error is smaller than $O(h^2)$, say, $O(h^3)$. One such algorithm is the (explicit) *midpoint rule*, which is motivated by the following observation.

Lemma 2.9. *If a solution to* (2.15) *is smooth enough (three times continuously differentiable) around $t = \tau$, then*

$$x(\tau + h) = x(\tau) + hf\Big(\tau + \frac{h}{2}, x(\tau) + \frac{h}{2} f(\tau, x(\tau))\Big) + O(h^3) \tag{2.17}$$

for small enough $h > 0$.

PROOF. Let us write a second order Taylor's expansion for $x : \mathbb{R} \to \mathbb{R}^n$:

$$\begin{aligned} x(\tau + h) &= x(\tau) + hx'(\tau) + \frac{h^2}{2} x''(\tau) + O(h^3) \\ &= x(\tau) + hf(\tau, x(\tau)) + \frac{h^2}{2} \frac{d}{d\tau} f(\tau, x(\tau)) + O(h^3) \\ &= x(\tau) + hf(\tau, x(\tau)) + \frac{h^2}{2}\Big(f_t(\tau, x(\tau)) + f_x(\tau, x(\tau))\, x'(\tau)\Big) + O(h^3) \\ &= x(\tau) + hf(\tau, x(\tau)) + \frac{h^2}{2}\Big(f_t(\tau, x(\tau)) + f_x(\tau, x(\tau))\, f(\tau, x(\tau))\Big) + O(h^3), \end{aligned} \tag{2.18}$$

[1] Notice that the smoothness of f also induces smoothness on x.

[2] Observe that j indicates here the 'discrete time' not a component of a vector

where $f_t(\tau, x(\tau)) \in \mathbb{R}^n$ denotes the derivative of f with respect to its first argument and $f_x(\tau, x(\tau)) \in \mathbb{R}^{n\times n}$ the derivative with respect to the second argument (a Jacobian matrix).[3] This is the needed form for the left-hand side of (2.17).

In order to handle the right-hand side of (2.17), we denote

$$g(h) = f\Big(\tau + \frac{h}{2},\, x(\tau) + \frac{h}{2} f(\tau, x(\tau))\Big).$$

In particular,

$$g'(h) = \frac{1}{2} f_t\Big(\tau + \frac{h}{2},\, x(\tau) + \frac{h}{2} f(\tau, x(\tau))\Big) + \frac{1}{2} f_x\Big(\tau + \frac{h}{2},\, x(\tau) + \frac{h}{2} f(\tau, x(\tau))\Big) f(\tau, x(\tau)),$$

and a Taylor's expansion gives

$$g(h) = g(0) + hg'(0) + O(h^2) = f(\tau, x(\tau)) + \frac{h}{2}\Big(f_t(\tau, x(\tau)) + f_x(\tau, x(\tau))\, f(\tau, x(\tau))\Big) + O(h^2).$$

Hence, the right-hand side of (2.17) allows the representation

(2.19)

$$x(\tau) + hg(h) = x(\tau) + hf(\tau, x(\tau)) + \frac{h^2}{2}\Big(f_t(\tau, x(\tau)) + f_x(\tau, x(\tau))\, f(\tau, x(\tau))\Big) + O(h^3).$$

The claim now follows by comparing the expansions for the left- and right-hand sides of (2.17), i.e. (2.18) and (2.19). □

Definition 2.10 (Midpoint rule)**.** We define the approximations $x_j \approx x(t_j) \in \mathbb{R}^n$, $j = 1, 2, \ldots$, via

$$x_{j+1} = x_j + hf\Big(t_j + \frac{h}{2},\, x_j + \frac{h}{2} f(t_j, x_j)\Big), \qquad j = 0, 1, \ldots,$$

which gives the (explicit) midpoint rule.

If the local truncation error of a numerical method is of the form $O(h^{p+1})$, it is said to be of order p. The Euler's method is thus of the first order and the midpoint rule of the second order. In a similar manner, one can look for numerical methods of even higher orders. As an example, the *Classical Runge–Kutta method* has a local truncation error $O(h^5)$, and it is thus a fourth order method.

We complete this section by studying how the local errors accumulate into a global error.

Theorem 2.11. *Let $\psi : I \times I \times \mathbb{R}^n$ be the solution map of (2.15) for a continuous f that satisfies the Lipschitz condition (2.3) (with, for simplicity, $I = [0, T]$). Assume that the considered numerical method is of order $p \geq 1$, that is, the corresponding local truncation error satisfies*

$$|x_{j+1} - \psi(t_{j+1}, t_j, x_j)| \leq Ch^{p+1}, \tag{2.20}$$

for some $C > 0$ whenever f is smooth enough. Then, the global error satisfies

$$\varepsilon_j := |x_j - \psi(t_j, 0, x_0)| \leq \frac{C}{L} h^p (e^{LT} - 1),$$

for all $t_j = jh \in [0, T]$.

[3] If you are afraid of Jacobian matrices, you may assume that $n = 1$ without loosing the main idea of the proof.

PROOF. According to (2.13),

$$\psi(t_j, 0, x_0) = \psi\Big(t_j, t_{j-1}, \psi(t_{j-1}, 0, x_0)\Big).$$

In consequence, it follows from the triangle inequality, (2.20) and the continuous dependence on the initial data (2.14) that

$$\begin{aligned}
\varepsilon_j &= |x_j - \psi(t_j, 0, x_0)| \\
&\leq |x_j - \psi(t_j, t_{j-1}, x_{j-1})| + \Big|\psi(t_j, t_{j-1}, x_{j-1}) - \psi\Big(t_j, t_{j-1}, \psi(t_{j-1}, 0, x_0)\Big)\Big| \\
&\leq Ch^{p+1} + e^{L(t_j - t_{j-1})}|x_{j-1} - \psi(t_{j-1}, 0, x_0))| \\
&= Ch^{p+1} + e^{Lh}\varepsilon_{j-1}.
\end{aligned}$$

We continue recursively to obtain

$$\begin{aligned}
\varepsilon_j &\leq Ch^{p+1} + e^{Lh}\varepsilon_{j-1} \\
&\leq Ch^{p+1} + e^{Lh}(Ch^{p+1} + e^{Lh}\varepsilon_{j-2}) \\
&= Ch^{p+1}(1 + e^{Lh}) + e^{2Lh}\varepsilon_{j-2} \\
&\leq Ch^{p+1}(1 + e^{Lh} + e^{2Lh}) + e^{3Lh}\varepsilon_{j-3} \\
&\leq Ch^{p+1}(1 + e^{Lh} + e^{2Lh} + \cdots + e^{(j-1)Lh}) + e^{jLh}\varepsilon_0.
\end{aligned}$$

Since

$$\varepsilon_0 = |x_0 - \psi(t_0, 0, x_0)| = |x_0 - x_0| = 0,$$

using the formula for a finite geometric series, we get

$$\varepsilon_j = Ch^{p+1}\sum_{l=0}^{j-1} e^{lLh} = Ch^{p+1}\frac{1 - (e^{Lh})^j}{1 - e^{Lh}} = Ch^{p+1}\frac{e^{L(jh)} - 1}{e^{Lh} - 1}.$$

Because

$$e^{Lh} = \sum_{l=0}^{\infty}\frac{1}{l!}(Lh)^l \geq 1 + Lh$$

and $0 \leq t_j = jh \leq T$ by assumption, we finally deduce the claim:

$$\varepsilon_j \leq Ch^{p+1}\frac{e^{Lt_j} - 1}{e^{Lh} - 1} \leq \frac{C}{L}h^p(e^{LT} - 1).$$

□

2.5. Stability and stiff systems

Often the system modeled by the IVP (2.1) includes phenomena that die out quickly. As an example, the half-life of some radioactive isotope may be considerably shorter than of the others, or some subprocess of a chemical reaction reaches equilibrium faster than the whole system. The ODEs that model these kinds of systems are called *stiff*.

To model the quickly changing phenomena accurately, it is typically necessary to use extremely small time steps in the numerical solver — far smaller than required by the 'other subprocesses' of the examined system. This potentially enormous computational load can be avoided by noticing that for rapidly stabilizing processes it is more essential to correctly model the long-term behavior than to get all details right when the process still has high 'change rate'

(or derivative). In consequence, it is of interest to study how different numerical methods succeed in predicting the state of a quickly stabilizing system at the time $t = \infty$.

The suitability of a numerical method for solving a stiff system can be tested by studying the model problem

$$x'(t) = \lambda x(t), \quad x(0) = x_0 \neq 0 \tag{2.21}$$

where $\mathrm{Re}\,(\lambda) < 0$[4] and $t \geq 0$. Notice that the exact solution of (2.21) is $x(t) = x_0 e^{\lambda t}$. In particular, it holds that

$$\lim_{t\to\infty} x(t) = 0. \tag{2.22}$$

It would thus be desirable that the discrete 'solution sequence' $x_j \approx x(jh) = x(t_j)$, $j = 0, 1, \ldots$, produced by some numerical method would have this same property, that is,

$$\lim_{j\to\infty} x_j = 0 \tag{2.23}$$

for any step size $h > 0$. If (2.23) holds for a given numerical method and some $h > 0$, the numerical method in question is said to be *stable* when applied to (2.21) with the considered step size. Unfortunately, no explicit method, such as the Euler's method or the midpoint rule, is stable for all $h > 0$.

To concretize this claim, let us study how the Euler's method fairs when applied to (2.21). According to Definition 2.8,

$$x_{j+1} = x_j + hf(t_j, x_j) = x_j + h\lambda x_j = (1 + h\lambda)x_j$$

since for (2.21) we have $f(t, x) = \lambda x$. Using this identity recursively, we get

$$x_j = (1 + h\lambda)^j x_0 \implies |x_j| = |1 + h\lambda|^j |x_0|, \quad j = 0, 1, \ldots .$$

Denoting $R(z) = 1 + z$, one may thus give an explicit condition for the stability of the Euler's method:

$$|R(h\lambda)| = |1 + z| < 1. \tag{2.24}$$

The function R is called the *stability function* (of the Euler's method) and the set

$$S = \{z \in \mathbb{C} \mid |R(z)| < 1\} \tag{2.25}$$

is the corresponding *stability region*. To sum up, the Euler's method provides a stable solution for (2.21) if and only if $h\lambda$ belongs to the stability region (2.25). Observe that (2.25) is the open disk of radius one around the point -1 in the complex plane.

In particular, if $\mathbb{R} \ni \lambda < 0$, it is easily deduced that

$$\lim_{j\to\infty} x_j = 0 \iff 0 < h < -\frac{2}{\lambda},$$

where x_j, $j = 0, 1, \ldots$, is the numerical solution of (2.21) produced by the Euler's method. In other words, if $\lambda \ll 0$ — when the exact solution of (2.21) goes to zero very fast as $t \to \infty$ — the stability of the Euler's method requires an extremely small step size $h > 0$.

[4] For certain reasons, stability is usually studied with $\lambda \in \mathbb{C}$, but $\lambda \in \mathbb{R}$ is the most important case on this course

It would be optimal if the stability region of a numerical method contained the open left half of the complex plane, since then (2.22) for the exact solution of (2.21) would imply (2.23) for the numerical method. This property is called *A-stability*. As the stability function of any explicit numerical method[5] is a polynomial, the corresponding stability region is bounded; cf. (2.24) and the homework set of week 9. Thus, no explicit numerical method can be A-stable — which is unfortunate.

2.6. Implicit solution methods

The fundamental remedy for instability issues is the introduction of *implicit solution methods*. We continue to study the numerical solution of the ODE

$$x'(t) = f(t, x(t)). \tag{2.26}$$

Assume once again that the exact solution to (2.26) is smooth enough close to the point $\tau \in \mathbb{R}$ and let us write a Taylor's expansion at $t = \tau + h$:

$$x(\tau) = x(\tau + h) + (-h)x'(\tau + h) + O(h^2) = x(\tau + h) - hf\Big(\tau + h, x(\tau + h)\Big) + O(h^2),$$

which is equivalent to

$$x(\tau + h) = x(\tau) + hf\Big(\tau + h, x(\tau + h)\Big) + O(h^2).$$

By dropping out the local truncation error $O(h^2)$, we get the *implicit Euler's method*.

Definition 2.12 (Implicit Euler's method)**.** We define the estimates $x_j \approx x(t_j) \in \mathbb{R}^n$, $j = 1, 2, \ldots$, via

$$x_{j+1} = x_j + hf(t_{j+1}, x_{j+1}), \qquad j = 0, 1, \ldots, \tag{2.27}$$

which gives the implicit Euler's method.

Notice that x_{j+1} appears on both sides of (2.27), so it must be solved from (2.27)[6]; in practice, this usually happens by some numerical (Newton-type) method. In consequence, the implementation of the implicit Euler's method is not as straightforward as that of the (explicit) Euler's method. The same applies to other implicit methods, as well.

However, the extra workload required by implicit methods is often justifiable due to their superior stability properties. Indeed, let us apply the implicit Euler's method to the test problem (2.21), which leads to the recursion

$$x_{j+1} = x_j + h\lambda x_{j+1} \iff x_{j+1} = \frac{1}{1 - h\lambda} x_j, \qquad j = 0, 1, \ldots,$$

if $h\lambda \neq 1$. In other words,

$$x_j = R(h\lambda)^j x_0,$$

where

$$R(z) = \frac{1}{1 - z}, \qquad z \in \mathbb{C},$$

[5] More precisely, of any *explicit Runge–Kutta method*.

[6] Hence, the word "implicit".

is the stability function of the implicit Euler's method. In consequence, the stability region of the implicit Euler's method is

$$S = \{z \in \mathbb{C} \mid |R(z)| < 1\} = \left\{z \in \mathbb{C} \;\middle|\; \left|(1-z)^{-1}\right| < 1\right\} = \{z \in \mathbb{C} \mid |1-z| > 1\}, \tag{2.28}$$

i.e., the open exterior of the disk with radius one around the point 1 in the complex plane. In particular, the open left half of the complex plane is contained in S, and thus the implicit Euler's method is A-stable. Recall this means that (2.22) for the exact solution of (2.21) implies (2.23) for the implicit Euler's method independently of the step size $h > 0$. However, since the stability region (2.28) is not exactly the open left half of the complex plane, it may happen that (2.23) is valid even if (2.22) is not: The implicit Euler's method satisfies (2.23) for many λ with $\mathrm{Re}(\lambda) \geq 0$, for which (2.22) obviously does not hold.

As the local truncation error of the implicit Euler's method is $O(h^2)$, Theorem 2.11 tells us that the corresponding global error is of the form $O(h)$ and the implicit Euler's method is of order $p = 1$. As for the explicit methods, it is also possible to build implicit methods of arbitrarily high order.

Definition 2.13 (Implicit midpoint rule)**.** We define the estimates $x_j \approx x(t_j) \in \mathbb{R}^n$, $j = 1, 2, \ldots$, via

$$x_{j+1} = x_j + hf\Big(t_j + \frac{h}{2}, \frac{1}{2}(x_j + x_{j+1})\Big), \qquad j = 0, 1, \ldots,$$

which gives the implicit midpoint rule.

The implicit midpoint rule is a second order method — local error $O(h^3)$, global error $O(h^2)$ — like its explicit counterpart. It also has good stability properties: If the implicit midpoint rule is applied to (2.21), one ends up with

$$x_{j+1} = x_j + h\lambda\Big(\frac{1}{2}(x_j + x_{j+1})\Big) \iff \Big(1 - \frac{1}{2}h\lambda\Big)x_{j+1} = \Big(1 + \frac{1}{2}h\lambda\Big)x_j.$$

Solving for x_{j+1} yields

$$x_{j+1} = R(h\lambda)x_j \implies R(h\lambda)^j x_0, \qquad j = 0, 1, \ldots,$$

where

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}$$

is the stability function of the implicit midpoint rule.

It can be shown (Homework 2) that the stability region of the implicit midpoint rule is exactly the open left half of the complex plane. Hence,

$$\lim_{t\to\infty} x(t) = 0 \iff \lim_{j\to\infty} x_j = 0,$$

where $x(t) = x_0 e^{\lambda t}$ is the exact solution of

$$x'(t) = \lambda x(t), \quad x(0) = x_0 \neq 0, \qquad \lambda \in \mathbb{C},$$

and x_j, $j = 0, 1, \ldots$, is the corresponding numerical solution produced by the implicit midpoint rule.

CHAPTER 3

Numerical solution of the Laplace equation (finite difference method)

The basic idea of *finite difference method*s is to approximate the considered (partial) differential operator by a finite-dimensional linear operator, i.e. by a matrix, that takes grid values of a function as its input. In this chapter, we study how a finite difference method can be employed for numerical solution of the Laplace (or Poisson) equation in simple geometries.

3.1. One-dimensional case

Consider a simple *Dirichlet* boundary value problem: Find $u : [0,1] \to \mathbb{R}$ such that

$$\begin{cases} u''(x) = f(x), & x \in (0,1), \\ u(0) = \alpha, \quad u(1) = \beta, \end{cases} \tag{3.1}$$

where $f : [0,1] \to \mathbb{R}$ is continuous[1] and $\alpha, \beta \in \mathbb{R}$. It is known that (3.1) has a unique solution; this could be proved, e.g., by using techniques studied during the first half of the course.

Let us figure out how the solution of (3.1) can be approximated numerically if the values of f are known on the spatial grid

$$x_j = jh, \qquad j = 1, 2, \ldots, m, \tag{3.2}$$

where $h = 1/(m+1)$. Our aim is to estimate the values of the solution u to (3.1) on this same grid; observe that the points $x_0 = 0$ and $x_{m+1} = 1$ are not interesting because $u(0)$ and $u(1)$ are defined by the *boundary conditions* of (3.1).

Let $v : \mathbb{R} \to \mathbb{R}$ be an arbitrary four times continuously differentiable function. We approximate its second derivative at $x \in \mathbb{R}$ with the help of suitable Taylor's expansions with $h > 0$:

$$v(x+h) = v(x) + hv'(x) + \frac{h^2}{2}v''(x) + \frac{h^3}{6}v'''(x) + O(h^4),$$
$$v(x-h) = v(x) + (-h)v'(x) + \frac{(-h)^2}{2}v''(x) + \frac{(-h)^3}{6}v'''(x) + O(h^4).$$

Adding these formulas, we get

$$v(x-h) + v(x+h) = 2v(x) + h^2 v''(x) + O(h^4),$$

or equivalently,

$$v''(x) = \frac{1}{h^2}\Big(v(x-h) - 2v(x) + v(x+h)\Big) + O(h^2). \tag{3.3}$$

[1] This assumption could be relaxed considerably.

By dropping out the error term $O(h^2)$ from (3.3), one obtains the standard second order central difference approximation.

For the solution of (3.1) and the grid points (3.2), the difference approximation (3.3) takes the form

$$\begin{aligned} u''(x_1) &\approx \frac{1}{h^2}\Big(\alpha - 2u(x_1) + u(x_2)\Big), \\ u''(x_j) &\approx \frac{1}{h^2}\Big(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\Big), \qquad j = 2, \ldots, m-1, \\ u''(x_m) &\approx \frac{1}{h^2}\Big(u(x_{m-1}) - 2u(x_m) + \beta\Big), \end{aligned} \tag{3.4}$$

where we used the boundary conditions $u(0) = u(x_0) = \alpha$ and $u(1) = u(x_{m+1}) = \beta$. Equating the right-hand sides of these equations with the corresponding grid values of f leads to a system of linear equations:

$$\begin{aligned} \frac{1}{h^2}(-2u_1^h + u_2^h) &= f(x_1) - \frac{\alpha}{h^2}, \\ \frac{1}{h^2}(u_{j-1}^h - 2u_j^h + u_{j+1}^h) &= f(x_j), \qquad j = 2, \ldots, m-1, \\ \frac{1}{h^2}(u_{m-1}^h - 2u_m^h) &= f(x_m) - \frac{\beta}{h^2}, \end{aligned} \tag{3.5}$$

where (hopefully) $u_j^h \approx u(x_j)$, $j = 1, \ldots, m$. The linear system (3.5) can be written as a matrix equation,

$$\Delta_{D-D}^h u^h = b^h, \tag{3.6}$$

where $u^h = [u_1^h, u_2^h, \ldots, u_m^h]^{\mathrm{T}}$, $b^h = [f(x_1) - \alpha/h^2, f(x_2), \ldots, f(x_{m-1}), f(x_m) - \beta/h^2]^{\mathrm{T}} \in \mathbb{R}^m$ and

$$\Delta_{D-D}^h = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{m\times m} \tag{3.7}$$

is a tridiagonal matrix.

The Dirichlet boundary value problem (3.1) can be solved approximately by determining the vector $u^h \in \mathbb{R}^m$ that satisfies (3.6); the unique solvability of (3.6) is a by-product of Lemma 3.1 presented below.[2] To be more precise, assuming that u and its first four derivatives are continuous on $[0,1]$[3], it holds that

$$\max_{1\le j\le m} |u(x_j) - u_j^h| \le Ch^2 \|u\|_{C^4([0,1])}, \qquad C > 0, \tag{3.8}$$

where

$$\|u\|_{C^4([0,1])} = \max_{0\le\eta\le 4} \max_{x\in[0,1]} |u^{(\eta)}(x)|.$$

[2] A square matrix is injective, which is equivalent to invertibility, if and only if zero is not its eigenvalue.

[3] The smoothness of u depends on the regularity of the source f in (3.1).

The proof of (3.8) is omitted, but it should, however, be noted that similar estimates can also be proved for more general problems than (3.1): If the exact solution of the considered *partial differential equation* (PDE) is smooth enough, the numerical solution produced by a (reasonable) finite difference method is typically as accurate as the discretization of differential operators in question — in our case, the accuracy is $O(h^2)$, as indicated by (3.3).

For reasons that will become more apparent in the following chapters, we next determine the eigenvalues and eigenvectors of Δ^h_{D-D}.

Lemma 3.1. *The eigenvalues of the matrix* $\Delta^h_{D-D} \in \mathbb{R}^{m\times m}$ *are*

$$\lambda_l = -\frac{4}{h^2}\sin^2\Big(\frac{\pi l h}{2}\Big), \qquad l = 1, 2, \ldots, m,$$

with the corresponding orthonormal eigenvectors

$$v^l = \sqrt{2h}\begin{bmatrix} \sin(\pi l x_1) \\ \sin(\pi l x_2) \\ \sin(\pi l x_3) \\ \vdots \\ \sin(\pi l x_m) \end{bmatrix} \in \mathbb{R}^m, \qquad l = 1, 2, \ldots, m,$$

where $\sqrt{2h}$ *is a normalization constant. In particular, all eigenvalues of* Δ^h_{D-D} *lie in the open interval* $(-4/h^2, 0)$.

PROOF. We denote $a^l = \Delta^h_{D-D} v^l$ and aim at proving that $a^l = \lambda_l v^l$, $l = 1, 2, \ldots, m$. It is easy to see that the jth component of a^l satisfies

$$a^l_j = \frac{\sqrt{2h}}{h^2}\Big(\sin(\pi l x_{j-1}) - 2\sin(\pi l x_j) + \sin(\pi l x_{j+1})\Big), \tag{3.9}$$

which holds for all $j = 1, 2, \ldots, m$ — including $j = 1$ and $j = m$ since $\sin(\pi l x_0) = \sin(0) = 0$ and $\sin(\pi l x_{m+1}) = \sin(l\pi) = 0$. According to the *sum and difference identities of trigonometric functions*,

$$\begin{aligned}\sin(\pi l x_{j\pm1}) = \sin(\pi l(j \pm 1)h) &= \sin(\pi l j h)\cos(\pi l h) \pm \cos(\pi l j h)\sin(\pi l h) \\ &= \sin(\pi l x_j)\cos(\pi l h) \pm \cos(\pi l x_j)\sin(\pi l h).\end{aligned}$$

Substituting these in (3.9) yields

$$\begin{aligned} a^l_j &= \frac{\sqrt{2h}}{h^2}\Big(2\sin(\pi l x_j)\cos(\pi l h) - 2\sin(\pi l x_j)\Big) \\ &= \frac{2}{h^2}\Big(\cos(\pi l h) - 1\Big)\sqrt{2h}\,\sin(\pi l x_j) = \frac{2}{h^2}\Big(\cos(\pi l h) - 1\Big)v^l_j, \end{aligned}$$

for all $j = 1, 2, \ldots, m$. Due to the definition of a^l, the fact that λ_l is an eigenvalue of Δ^h_{D-D}, with the associated eigenvector $v^l \in \mathbb{R}^m$, follows now from the trigonometric identity

$$\cos(\pi l h) - 1 = -2\sin^2\Big(\frac{\pi l h}{2}\Big).$$

Since sine is monotonically increasing on the open interval $(0, \pi/2)$, the eigenvalues

$$\lambda_l = -\frac{4}{h^2}\sin^2\Big(\frac{\pi l h}{2}\Big) = -\frac{4}{h^2}\sin^2\Big(\frac{\pi l}{2(m+1)}\Big), \qquad l = 1, 2, \ldots, m,$$

form a monotonically decreasing sequence, that is,

$$0 > \lambda_1 > \lambda_2 > \cdots > \lambda_{m-1} > \lambda_m.$$

In particular, we have found m distinct eigenvalues for $\Delta^h_{D-D} \in \mathbb{R}^{m\times m}$. Since it is well known that for a symmetric matrix, such as Δ^h_{D-D}, the eigenvectors corresponding to different eigenvalues are mutually orthogonal,[4] the only remaining thing to prove is that $|v^l| = 1$, $l = 1, \ldots, m$.

It holds that

$$\begin{aligned}
\frac{1}{2h}|v^l|^2 &= \sum_{j=1}^{m} \sin^2(\pi l x_j) = \frac{1}{2}\sum_{j=1}^{m}\Big(1 - \cos(2\pi l x_j)\Big) \\
&= \frac{m}{2} - \frac{1}{2}\sum_{j=1}^{m} \mathrm{Re}\left(e^{i2\pi l x_j}\right) = \frac{m}{2} - \frac{1}{2}\mathrm{Re}\Big(\sum_{j=1}^{m} e^{i2\pi l x_j}\Big) \\
&= \frac{m+1}{2} - \frac{1}{2}\mathrm{Re}\Big(\sum_{j=0}^{m}\left(e^{i2\pi l h}\right)^j\Big) \\
&= \frac{m+1}{2} - \frac{1}{2}\mathrm{Re}\left(\frac{1-\left(e^{i2\pi l h}\right)^{m+1}}{1-e^{i2\pi l x_j}}\right).
\end{aligned}$$

Recalling that $h = 1/(m+1)$, this simplifies to

$$\frac{1}{2h}|v^l|^2 = \frac{1}{2h} - \frac{1}{2}\mathrm{Re}\left(\frac{1-e^{i2\pi l}}{1-e^{i2\pi l x_j}}\right) = \frac{1}{2h}, \qquad l = 1, 2, \ldots, m,$$

since $e^{i2\pi l} = 1$ for all $l \in \mathbb{Z}$. A multiplication by $2h$ completes the proof. □

Let us next examine what happens if one of the two *Dirichlet boundary conditions* in (3.1) is changed into a *Neumann boundary condition*, that is, we consider the mixed boundary value problem

$$\begin{cases} u''(x) = f(x), & x \in (0,1), \\ -u'(0) = \alpha, \quad u(1) = \beta. \end{cases} \tag{3.10}$$

We continue to assume that f is known at the grid points (3.2) and approximate the second derivative of u by the formulas

$$\begin{aligned}
u''(x_1) &\approx \frac{1}{h^2}\Big(u(x_0) - 2u(x_1) + u(x_2)\Big), \\
u''(x_j) &\approx \frac{1}{h^2}\Big(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\Big), \qquad j = 2, \ldots, m-1, \\
u''(x_m) &\approx \frac{1}{h^2}\Big(u(x_{m-1}) - 2u(x_m) + \beta\Big).
\end{aligned} \tag{3.11}$$

Notice that the difference between these equations and (3.4) is that now we cannot directly use the left boundary condition of (3.10) to replace $u(x_0)$ in the first equation of (3.11). As (3.11) constitutes of m equations, the $m+1$ point values $u(x_0), u(x_1), \ldots, u(x_m)$ appearing in (3.11) cannot be directly estimated by equating the right-hand sides of (3.11) with the corresponding grid values of f. Hence, we need to get rid of one grid value of u in (3.11); this is established

[4] $\lambda_k (v^l)^{\mathrm{T}} v^k = (v^l)^{\mathrm{T}}(\lambda_k v^k) = (v^l)^{\mathrm{T}} A v^k = ((v^l)^{\mathrm{T}} A v^k)^{\mathrm{T}} = (v^k)^{\mathrm{T}} A v^l = \lambda_l (v^k)^{\mathrm{T}} v^l = \lambda_l (v^l)^{\mathrm{T}} v^k \;\Rightarrow\; (v^l)^{\mathrm{T}} v^k = 0$

by using the Neumann condition $-u'(0) = \alpha$ to give $u(x_0) = u(0)$ approximately with the help of $u(x_1)$, $u(x_2)$ and α.

Lemma 3.2. *If $v : [0,1] \to \mathbb{R}$ is smooth enough close to the left end point $x = 0$, then*

$$v'(0) = \frac{1}{2h}\Big(-3v(0) + 4v(h) - v(2h)\Big) + O(h^2)$$

for all small enough $h > 0$.

PROOF. Let us write two origin-centered Taylor's expansions for v:

$$v(h) = v(0) + hv'(0) + \frac{h^2}{2}v''(0) + O(h^3),$$
$$v(2h) = v(0) + (2h)v'(0) + \frac{(2h)^2}{2}v''(0) + O(h^3).$$

We multiply the first equation by four and subtract the second one:

$$4v(h) - v(2h) = 3v(0) + 2hv'(0) + O(h^3) \iff v'(0) = \frac{1}{2h}\Big(-3v(0) + 4v(h) - v(2h)\Big) + O(h^2),$$

which completes the proof. □

If the exact solution of (3.10) is smooth enough, Lemma 3.2 tells us that

$$-\alpha = u'(x_0) = u'(0) \approx \frac{1}{2h}\Big(-3u(x_0) + 4u(x_1) - u(x_2)\Big),$$

where we utilized the definition of the grid points in (3.2). Solving for $u(x_0)$ yields

$$u(x_0) \approx \frac{1}{3}\Big(4u(x_1) - u(x_2) + 2h\alpha\Big),$$

which may be substituted in the first equation of (3.11) to obtain

$$u''(x_1) \approx \frac{2}{3h^2}\Big(-u(x_1) + u(x_2) + h\alpha\Big). \tag{3.12}$$

Equating the right-hand sides of (3.12) and the other equations of (3.11) with the grid values $f(x_1), f(x_2), \ldots, f(x_m)$, finally results in the matrix equation

$$\Delta^h_{N-D} u^h = b^h,$$

where $u^h = [u^h_1, u^h_2, \ldots, u^h_m]^{\mathrm{T}} \in \mathbb{R}^m$ contains the estimates for $u(x_1), u(x_2), \ldots, u(x_m)$, $b^h = [f(x_1) - (2\alpha)/(3h), f(x_2), \ldots, f(x_{m-1}), f(x_m) - \beta/h^2]^{\mathrm{T}} \in \mathbb{R}^m$ and

$$\Delta^h_{N-D} = \frac{1}{h^2}\begin{bmatrix} -2/3 & 2/3 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

is a tridiagonal difference matrix.

Remark 3.3. The subscripts of the matrices Δ^h_{D-D} and Δ^h_{N-D} indicate the types of the considered boundary conditions: Δ^h_{D-D} is a discretization of the second spatial derivative with Dirichlet boundary conditions at both ends of $[0,1]$, whereas Δ^h_{N-D} has a Neumann boundary condition on the left and a Dirichlet condition on the right. In the same manner, one can also introduce Δ_{D-N}

and Δ_{N-N}. Like $\Delta^h_{D-D} \in \mathbb{R}^{m\times m}$, also $\Delta^h_{N-D}, \Delta^h_{D-N} \in \mathbb{R}^{m\times m}$ are invertible (an exercise). Furthermore, it can be shown (non-trivially) that the eigenvalues of Δ^h_{N-D} and Δ^h_{D-N} lie in the open interval $(-4/h^2, 0)$, which was deduced for Δ^h_{D-D} already in Lemma 3.1. However, $\Delta^h_{N-N} \in \mathbb{R}^{m\times m}$ is singular, i.e. noninvertible, as explained below.

If both boundary conditions are of the Neumann type, that is, we consider the Neumann boundary value problem

$$\begin{cases} u''(x) = f(x), & x \in (0,1), \\ -u'(0) = \alpha, \quad u'(1) = \beta, \end{cases} \tag{3.13}$$

the situation gets slightly more complicated. Indeed, the problem (3.13) has a solution only if

$$\int_0^1 f(x)\,dx = \int_0^1 u''(x)\,dx = u'(1) - u'(0) = \alpha + \beta. \tag{3.14}$$

In electrostatics, this condition can be explained as follows: If u models the electromagnetic potential inside the homogeneous 'body' $(0,1)$, then f may be interpreted as an internal current source and α, β as the currents flowing out from the endpoints of the body. Thus, (3.14) corresponds to the law of charge conservation. On the other hand, if (3.13) has a solution, it is only unique up to an additive constant: If u is a solution of (3.13), then so is $u + c$ because the differentiations in (3.13) make the constant $c \in \mathbb{R}$ disappear. This nonuniqueness corresponds to the freedom in the choice of the ground level of potential in electrostatics.

These fundamental properties of (3.13) are inherited by the associated discretized system that has

$$\Delta^h_{N-N} = \frac{1}{h^2}\begin{bmatrix} -2/3 & 2/3 & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 2/3 & -2/3 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

as its coefficient matrix. In particular, Δ^h_{N-N} is not invertible (an exercise).

3.2. Two-dimensional case

Next, we consider how the Laplace operator can be discretized in the open unit square $\Omega = (0,1)\times(0,1)$. Assume $v : \overline{\Omega} \to \mathbb{R}$ is smooth enough and, for simplicity, that

$$v(x) = 0 \quad \text{for all } x \in \partial\Omega, \tag{3.15}$$

which has the short-hand notation $v|_{\partial\Omega} = 0$. Moreover, suppose that the values of v are known at the two-dimensional grid points

$$(jh, kh) \in \Omega, \qquad j,k = 1,2,\ldots,m,$$

where $h = 1/(m+1)$. We denote

$$v_{j,k} = v(jh, kh), \qquad j,k = 1,2,\ldots,m,$$

and introduce the matrix of grid values

$$V = \begin{bmatrix} v_{1,1} & \dots & v_{1,m} \\ \vdots & & \vdots \\ v_{m,1} & \dots & v_{m,m} \end{bmatrix} \in \mathbb{R}^{m\times m}. \tag{3.16}$$

In the same spirit, $V_{xx} \in \mathbb{R}^{m\times m}$ and $V_{yy} \in \mathbb{R}^{m\times m}$ denote the matrices containing the grid values of v_{xx} and v_{yy}, respectively.

As in the one-dimensional case, the second (partial) derivatives of v can be approximated via

$$\begin{aligned} v_{xx}(jh, kh) &\approx \frac{1}{h^2}\big(v_{j-1,k} - 2v_{j,k} + v_{j+1,k}\big) \\ v_{yy}(jh, kh) &\approx \frac{1}{h^2}\big(v_{j,k-1} - 2v_{j,k} + v_{j,k+1}\big), \end{aligned}$$

where $j, k = 1, 2, \dots, m$. Notice that

$$v_{0,k} = v_{j,0} = v_{m+1,k} = v_{j,m+1} = 0, \qquad j, k = 0, 1, \dots, m, m+1,$$

by virtue of (3.15). In consequence, it is straightforward to deduce that

$$V_{xx} \approx \Delta^h_{D-D} V \qquad \text{and} \qquad V_{yy} \approx V(\Delta^h_{D-D})^{\mathrm{T}}. \tag{3.17}$$

Notice that in the first equation of (3.17), the 'one-dimensional' discretized second derivative with Dirichlet boundary conditions, i.e. $\Delta^h_{D-D} \in \mathbb{R}^m$, operates on the columns of V, whereas in the second equation it operates on the rows of V.[5] Summing the to formulas in (3.17), we get an approximate representation for the grid values of $\Delta v = v_{xx} + v_{yy}$, namely

$$\Delta^{h,2} V := \Delta^h_{D-D} V + V(\Delta^h_{D-D})^{\mathrm{T}}.$$

Obviously, the mapping $\Delta^{h,2} : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$ is linear, and thus it can be represented as an element of $\mathbb{R}^{m^2\times m^2}$ with respect to a suitable basis of $\mathbb{R}^{m\times m}$. To this end, let us introduce an auxiliary linear map $T : \mathbb{R}^{m\times m} \to \mathbb{R}^{m^2}$ which piles the columns of a given matrix to form a 'long vector'. To be precise, if $A = [a_1, a_2, \dots, a_m] \in \mathbb{R}^{m\times m}$, then

$$T(A) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}. \tag{3.18}$$

Notice that the inverse map $T^{-1} = \mathbb{R}^{m^2} \to \mathbb{R}^{m\times m}$ builds a square matrix out of a m^2-dimensional vector. In addition, we define the Kronecker product of $A, B \in \mathbb{R}^{m\times m}$ by

$$B \otimes A = \begin{bmatrix} B_{11}A & B_{12}A & \dots & B_{1m}A \\ B_{21}A & B_{22}A & \dots & B_{2m}A \\ \vdots & \vdots & \dots & \vdots \\ B_{m1}A & B_{m2}A & \dots & B_{mm}A \end{bmatrix} \in \mathbb{R}^{m^2\times m^2}.$$

[5] Since Δ^h_{D-D} is symmetric, the transposition in the second formula of (3.17) is needless. However, since difference matrices corresponding to other boundary conditions may not be symmetric, the transposition is included here for the sake of generality.

Lemma 3.4. *For any $A \in \mathbb{R}^{m\times m}$, it holds that*

$$\Delta^{h,2}A = \Delta^h_{D-D}A + A(\Delta^h_{D-D})^{\mathrm{T}} = T^{-1}\Big((I \otimes \Delta^h_{D-D} + \Delta^h_{D-D} \otimes I)\, T(A)\Big),$$

where $I \in \mathbb{R}^{m\times m}$ is the identity matrix.

PROOF. Although the proof is not very difficult — only technical and lengthy — it is omitted. □

Let us finally consider the Dirichlet boundary value problem

$$\begin{cases} \Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.19}$$

where Ω still is the unit square. If the grid values of the source $f : \Omega \to \mathbb{R}$ are stuffed into the matrix $F \in \mathbb{R}^{m\times m}$, that is,

$$F_{j,k} = f(jh, kh), \qquad j,k = 1,2,\dots,m,$$

and the to-be-solved approximate grid values of u into the matrix U,

$$U_{j,k} \approx u(jh, kh), \qquad j,k = 1,2,\dots,m,$$

then the discretized version of (3.19) reads

$$\Delta^h_{D-D}U + U(\Delta^h_{D-D})^{\mathrm{T}} = F. \tag{3.20}$$

In practice, this is solved by considering the corresponding 'vector formulation'

$$(I \otimes \Delta^h_{D-D} + \Delta^h_{D-D} \otimes I)u^h = f^h, \tag{3.21}$$

where $u^h = T(U) \in \mathbb{R}^{m^2}$ and $f^h = T(F) \in \mathbb{R}^{m^2}$. The equivalence of the formulations (3.20) and (3.21) is guaranteed by Lemma 3.4.

Remark 3.5. Different (homogeneous) boundary conditions on different edges of the unit square can be accounted for by using appropriate one-dimensional discretizations of the second derivative. As an example, if the considered problem is

$$\begin{cases} \Delta u = f & \text{in } \Omega, \\ u(0,\cdot) = 0 & \text{on } (0,1), \\ u_x(1,\cdot) = 0 & \text{on } (0,1), \\ -u_y(\cdot,0) = 0 & \text{on } (0,1), \\ u(\cdot,1) = 0 & \text{on } (0,1), \end{cases}$$

the equation (3.20) transforms into

$$\Delta^h_{D-N}U + U(\Delta^h_{N-D})^{\mathrm{T}} = F$$

and (3.20) into

$$(I \otimes \Delta^h_{D-N} + \Delta^h_{N-D} \otimes I)u^h = f^h.$$

Remark 3.6. Handling nonhomogeneous boundary conditions is naturally also possible: The boundary values contribute to the source vector $f^h \in \mathbb{R}^m$ in a certain way (a nontrivial exercise).

Remark 3.7. In a similar way, one could also discretize more general *elliptic* PDEs of the form

$$-\nabla \cdot (a\nabla u) + b \cdot \nabla u + cu = f \quad \text{in } \Omega$$

accompanied by suitable boundary conditions. Here, $a : \Omega \to \mathbb{R}$, $b : \Omega \to \mathbb{R}^2$ and $c : \Omega \to \mathbb{R}$ are regular enough coefficient functions, with case-dependent positivity properties. However, in more complicated geometries, the finite difference way of thinking becomes complicated, and it is often more sensible to resort to *finite element methods* (cf. Chapter 6).

CHAPTER 4

Numerical solution of the heat equation (finite difference method)

When a partial differential equation includes both time and a spatial variable, their different natures must be taken into account in the discretization: For the spatial variable, one usually has boundary conditions (or conditions for the behavior at infinity), while the behavior in time is restricted by an initial condition, that is, the state of the examined system is predefined, say, at $t = 0$. If the studied phenomenon is first discretized with respect to the spatial variable, one typically ends up with an IVP for a system of ODEs, which can then be numerically solved by resorting to the techniques in Chapter 2.

Consider the one-(spatial)-dimensional Dirichlet initial and boundary value problem

$$\begin{cases} u_t(x,t) = cu_{xx}(x,t) + f(x,t), & t > 0,\ x \in (0,1), \\ u(0,t) = u(1,t) = 0, & t > 0, \\ u(x,0) = g(x), & x \in (0,1). \end{cases} \tag{4.1}$$

If $c > 0$ and $f : (0,1) \times (0,\infty) \to \mathbb{R}$, $g : (0,1) \to \mathbb{R}$ are regular enough, it can be shown that (4.1) has a unique solution $u : [0,1] \times [0,\infty) \to \mathbb{R}$, whose smoothness (and behavior as $t \to \infty$) depends on the initial condition g and, in particular, on the source f.

We start by discretizing (4.1) with respect to the spatial variable; we continue to employ the one-dimensional spatial grid (3.2). Applying the standard central difference approximation (3.3) to u_{xx} gives

$$u_t(x_j,t) = \frac{c}{h^2}\Big(u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)\Big) + f(x_j,t) + O(h^2), \quad j = 1,2,\dots,m,\ t > 0, \tag{4.2}$$

assuming that u is smooth enough. Accounting for the Dirichlet boundary conditions ($u(x_0,t) = u(x_{m+1},t) = 0$) and dropping out the error term $O(h^2)$ in (4.2), the spatially discretized version of (4.1) becomes

$$\begin{cases} (u^h)'(t) = c\Delta^h_{D-D}u^h(t) + f^h(t), & t > 0, \\ u^h(0) = g^h, \end{cases} \tag{4.3}$$

where $\Delta^h_{D-D} \in \mathbb{R}^{m\times m}$ is the standard discretization of the second spatial derivative with Dirichlet boundary conditions and

$$u^h(t) = \begin{bmatrix} u^h_1(t) \\ u^h_2(t) \\ \vdots \\ u^h_m(t) \end{bmatrix}, \qquad f^h(t) = \begin{bmatrix} f(x_1,t) \\ f(x_2,t) \\ \vdots \\ f(x_m,t) \end{bmatrix}, \qquad g^h = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_m) \end{bmatrix}.$$

The idea is that $u_j^h(t) \approx u(x_j, t)$ for $j = 1, 2, \ldots, m$ and $t > 0$.

Remark 4.1. There are no boundary conditions in (4.3) because those of (4.1) have already been included in the structure of the difference matrix Δ_{D-D}^h.

As the problem (4.3) includes no spatial variable, it is an IVP for a system of ordinary differential equation. In particular, (4.3) can be discretized by resorting to the methodology of Chapter 2. As an example, the (explicit) Euler's method results in the recursion

$$\begin{cases} u_{k+1}^h = u_k^h + \delta\big(c\Delta_{D-D}^h u_k^h + f^h(t_k)\big), & k = 0, 1, \ldots, \\ u_0^h = g^h, \end{cases} \tag{4.4}$$

where the time grid is defined by $t_k = k\delta$, $k = 0, 1, 2, \ldots$, with $\delta > 0$ being the time step size, and $u_k^h \approx u^h(t_k)$. Putting together both the spatial and the time discretization, the hope is that $(u_k^h)_j \approx u(x_j, t_k)$, for $j = 1, 2, \ldots, m$ and $k = 0, 1, 2, \ldots$.

Unfortunately, the most simple discretization of (4.1) provided by (4.4) has no practical use due to the weak stability properties of the Euler's method. To concretize this claim, we assume that $f \equiv 0$, whence it is known that for any initial data g,

$$\lim_{t\to\infty} u(x, t) = 0$$

uniformly with respect to $x \in (0, 1)$. Although we omit the proof of this claim, its physical interpretation is simple: If the end points of the 'rod' $[0, 1]$ are at the fixed temperature 0 and the rod itself is not heated externally, it is obvious that at '$t = \infty$' the whole rod is at the zero temperature independently of the initial temperature distribution g. As for the stiff IVPs studied in Chapter 2, any reasonable numerical solution of (4.1) should also exhibit this same general behavior; in particular, the solution sequence produced by the iteration (4.4) should satisfy

$$\lim_{k\to\infty} u_k^h = 0 \in \mathbb{R}^n,$$

if $f^h(t) = 0$ for all $t \geq 0$. Notice that in this case (4.4) may be given compactly as

$$u_k^h = (I + \delta c\Delta_{D-D}^h)^k g^h, \qquad k = 0, 1, 2, \ldots, \tag{4.5}$$

where $I \in \mathbb{R}^{m\times m}$ is the identity matrix.

Lemma 4.2. *It holds that*

$$\lim_{k\to\infty} (I + \delta c\Delta_{D-D}^h)^k = 0 \in \mathbb{R}^{m\times m},$$

or equivalently

$$\lim_{k\to\infty} u_k^h = 0 \in \mathbb{R}^m$$

for u_k^h of (4.5) with all $g^h \in \mathbb{R}^m$, if and only if the time step size $\delta > 0$ satisfies

$$\delta < \frac{h^2}{2c}\left(\sin^2\left(\frac{m\pi}{2(m+1)}\right)\right)^{-1}.$$

PROOF. According to Lemma 3.1, $\Delta_{D-D}^h \in \mathbb{R}^{m\times m}$ has the eigenvalues

$$\lambda_l = -\frac{4}{h^2}\sin^2\Big(\frac{\pi l h}{2}\Big) = -\frac{4}{h^2}\sin^2\Big(\frac{\pi l}{2(m+1)}\Big), \qquad l = 1, 2, \ldots, m.$$

It follows that the matrix $I + \delta c \Delta^h_{D-D} \in \mathbb{R}^{m\times m}$ has the eigenvalues

$$\mu_l = 1 + \delta c \lambda_l, \qquad l = 1, 2, \dots, m.$$

Indeed, if v^l is the (normalized) eigenvector of Δ^h_{D-D} corresponding to the eigenvalue λ_l, then

$$(I + \delta c \Delta^h_{D-D}) v^l = v^l + \delta c \lambda_l v^l = (1 + \delta c \lambda_l) v^l,$$

for all $l = 1, 2, \dots, m$.

Due to the strict monotonicity of sine on the interval $(0, \pi/2)$, we know that

$$0 > \lambda_1 > \lambda_2 > \cdots > \lambda_{m-1} > \lambda_m,$$

and thus also

$$\mu_1 > \mu_2 > \cdots > \mu_{m-1} > \mu_m \tag{4.6}$$

since $c, \delta > 0$ by assumption. In particular, as $I + \delta c \Delta^h_{D-D} \in \mathbb{R}^{m\times m}$ has m distinct eigenvalues, it is diagonalizable. Hence, it holds that (Homework 11)

$$\lim_{k\to\infty} (I + \delta c \Delta^h_{D-D})^k = 0 \in \mathbb{R}^{m\times m}$$

if and only if $|\mu_l| < 1$ for all $l = 1, 2, \dots, m$. By virtue of (4.6), we thus have

$$\lim_{k\to\infty} (I + \delta c \Delta^h_{D-D})^k = 0 \iff \mu_1 < 1 \text{ and } \mu_m > -1.$$

First of all, since $\lambda_1 < 0$ and $c, \delta > 0$,

$$\mu_1 = 1 + \delta c \lambda_1 < 1$$

holds trivially. On the other hand,

$$\mu_m = 1 + \delta c \lambda_m = 1 - \frac{4\delta c}{h^2} \sin^2\Big(\frac{m\pi}{2(m+1)}\Big) > -1$$

if and only if

$$\delta < \frac{h^2}{2c} \left(\sin^2\Big(\frac{m\pi}{2(m+1)}\Big) \right)^{-1},$$

which completes the proof. □

Remark 4.3. Because for large $m \in \mathbb{N}$,

$$\frac{m\pi}{2(m+1)} \approx \frac{\pi}{2},$$

and thus

$$\sin^2\Big(\frac{m\pi}{2(m+1)}\Big) \approx 1,$$

the assertion of Lemma 4.2 should usually be read as

$$0 < \delta \le \frac{h^2}{2c} \implies \lim_{k\to\infty} (I + \delta c \Delta^h_{D-D})^k = 0.$$

The above result on the stability of the discretization (4.4) can be interpreted as follows: If the spatial discretization is fine, i.e., $h = 1/(m+1)$ is small, then the time step size $\delta > 0$ of the Euler's method must be chosen extremely small in order to get stable numerical solutions. As an example if $c = 1$ and $m = 99$, i.e. $h = 10^{-2}$, then the stability condition for the discretization (4.4) reads

$$\delta \le \frac{(10^{-2})^2}{2} = 5 \cdot 10^{-5},$$

which is rather restrictive. Due to these kinds of stability problems, the time discretization of the heat equation is usually done by some implicit method.

Applying the implicit Euler's method to the spatially discretized IVP (4.3) leads to the algorithm

$$\begin{cases} u_{k+1}^h = u_k^h + \delta\big(c\Delta_{D-D}^h u_{k+1}^h + f^h(t_{k+1})\big), & k = 0, 1, \dots, \\ u_0^h = g^h, \end{cases} \tag{4.7}$$

where the time grid t_k, $k = 0, 1, 2, \dots$, $f : (0, \infty) \to \mathbb{R}^m$ and $g^h \in \mathbb{R}^m$ are as in (4.4). Once again, the aim is that $(u_k^h)_j \approx u(x_j, t_k)$ for $j = 1, 2, \dots, m$ and $k = 0, 1, 2, \dots$. The 'next iterate' u_{k+1}^h can be solved from the first equation of (4.7) to obtain an 'explicit form' for the considered numerical method:

$$\begin{cases} u_{k+1}^h = (I - \delta c\Delta_{D-D}^h)^{-1}\big(u_k^h + \delta f^h(t_{k+1})\big), & k = 0, 1, \dots, \\ u_0^h = g^h. \end{cases} \tag{4.8}$$

Notice that this recursion is well defined: The matrix $I - \delta c\Delta_{D-D}^h \in \mathbb{R}^{m\times m}$ is invertible because it has m distinct eigenvalues that are strictly larger than zero[1]; see the proof of Lemma 4.2.

Analogously, an application of the implicit midpoint rule to (4.3) yields

$$\begin{cases} u_{k+1}^h = u_k^h + \delta\Big(c\Delta_{D-D}^h\Big(\frac{1}{2}(u_k^h + u_{k+1}^h)\Big) + f^h(t_k + \delta/2)\Big), & k = 0, 1, \dots, \\ u_0^h = g^h, \end{cases}$$

and solving for u_{k+1} gives

$$\begin{cases} u_{k+1}^h = \big(I - \frac{1}{2}\delta c\Delta_{D-D}^h\big)^{-1}\Big[\big(I + \frac{1}{2}\delta c\Delta_{D-D}^h\big)u_k^h + \delta f^h(t_k + \delta/2)\Big], & k = 0, 1, \dots, \\ u_0^h = g^h. \end{cases} \tag{4.9}$$

The algorithm (4.9) is the so-called the *Crank–Nicolson method*, which is a popular technique for solving *parabolic* PDEs. As in the case of (4.8), it is easy to see that all eigenvalues of $I - \frac{1}{2}\delta c\Delta_{D-D}^h$ are positive, and thus the iteration (4.9) is well defined for all $\delta > 0$.

Remark 4.4. Although the implementation of (4.8) or (4.9) requires inverting a sparse $m \times m$ matrix at each step of the time iteration,[2] this is usually worthwhile as both (4.8) and (4.9) produce stable solutions for *any* $\delta > 0$ (Homework 11), which makes it possible to use far larger time step sizes than in the case of (4.4).

If the boundary conditions of (4.1) are not homogeneous, this must be taken into account in the spatially discretized problem (4.3). As an example, let us consider the initial and boundary value problem

$$\begin{cases} u_t(x,t) = cu_{xx}(x,t) + f(x,t), & t > 0,\ x \in (0,1), \\ u(0,t) = \alpha,\ \ u_x(1,t) = \beta, & t > 0, \\ u(x,0) = g(x), & x \in (0,1). \end{cases} \tag{4.10}$$

[1] A square matrix is invertible if and only if zero is not its eigenvalue.

[2] It is actually debatable how the multiplications by the matrix inverse in (4.8) or (4.9) are performed most efficiently; on this course, you can always resort to the 'backslash' operation of MATLAB.

For the central grid points $x_2, x_3, \ldots, x_{m-1}$, we can directly use the spatially discretized equations (4.2), but for the endpoints x_1 and x_m we need to account for the difference in boundary conditions between (4.1) and (4.10). As when handling nonhomogeneous boundary conditions in Chapter 3, at x_1 we get the approximate equation

$$u_t(x_1,t) \approx \frac{c}{h^2}\Big(\alpha - 2u(x_1,t) + u(x_2,t)\Big) + f(x_1,t), \tag{4.11}$$

where the left boundary condition of (4.10), $u(0,t) = u(x_0,t) = \alpha$, was used. On the other hand, employing the 'right endpoint version' of Lemma 3.2,

$$\beta = u_x(1,t) = u_x(x_{m+1},t) \approx \frac{1}{2h}\Big(u(x_{m-1},t) - 4u(x_m,t) + 3u(x_{m+1},t)\Big),$$

i.e.,

$$u(x_{m+1},t) = \frac{1}{3}\Big(- u(x_{m-1},t) + 4u(x_m,t) + 2h\beta\Big), \tag{4.12}$$

we may approximate at x_m as follows:

$$\begin{aligned} u_t(x_m,t) &\approx \frac{c}{h^2}\Big(u(x_{m-1},t) - 2u(x_m,t) + u(x_{m+1},t)\Big) + f(x_m,t) \\ &\approx \frac{c}{h^2}\Big(\frac{2}{3}u(x_{m-1}) - \frac{2}{3}u(x_m,t) + \frac{2h}{3}\beta\Big) + f(x_m,t). \end{aligned} \tag{4.13}$$

Combining (4.11) and (4.13) with (4.2) for $j = 2, \ldots, m-1$, we obtain the spatially discretized problem corresponding to (4.10):

$$\begin{cases} (u^h)'(t) = c\Delta^h_{D-N}u^h(t) + b^h(t), & t > 0, \\ u^h(0) = g^h, \end{cases} \tag{4.14}$$

where Δ^h_{D-N} is the standard approximation of the second derivative with a Dirichlet boundary condition on the left and a Neumann condition on the right, $u^h : (0,\infty) \to \mathbb{R}^m$ gives the approximations for the grid values of the solution to (4.10), $g^h \in \mathbb{R}^m$ is as in (4.3) and

$$b^h(t) = \begin{bmatrix} f(x_1,t) + \dfrac{c\alpha}{h^2} \\ f(x_2,t) \\ \vdots \\ f(x_{m-1},t) \\ f(x_m,t) + \dfrac{2c\beta}{3h} \end{bmatrix}.$$

The boundary values appearing in (4.10) are thus visible in the source vector $b^h : (0,\infty) \to \mathbb{R}^m$ of (4.14); notice that the case when α and/or β are time-dependent can be handled analogously. The problem (4.14) can be solved by some suitable (implicit) method for IVPs, such as the implicit midpoint rule.

We complete the discussion on solving parabolic initial/boundary value problems using finite difference methods by considering a simple setting in the unit square $\Omega = (0,1) \times (0,1)$:

$$\begin{cases} u_t(x,t) = c\Delta u(x,t) + f(x,t), & t > 0,\ x = (x_1,x_2) \in \Omega, \\ -u_{x_1}(0,x_2,t) = u_{x_1}(1,x_2,t) = 0, & x_2 \in (0,1),\ t > 0, \\ u(x_1,0,t) = u_{x_2}(x_1,1,t) = 0, & x_1 \in (0,1),\ t > 0, \\ u(x,0) = g(x), & x \in \Omega. \end{cases} \tag{4.15}$$

Following the discretization of the two-dimensional Poisson equation (3.19), we denote by $U : (0,\infty) \to \mathbb{R}^{m\times m}$ the approximations for the grid values of the solution to (4.15), i.e.,

$$U_{j,k}(t) \approx u(jh, kh, t), \qquad j,k = 1,2,\dots,m, \quad t > 0,$$

where $h = 1/(m+1)$. Similarly, $F : (0,\infty) \to \mathbb{R}^{m\times m}$ and $G \in \mathbb{R}^{m\times m}$ are defined via

$$F_{j,k}(t) = f(jh, kh, t), \qquad G_{j,k} = g(jh, kh), \qquad j,k = 1,2,\dots,m.$$

Finally, let us denote

$$u^h(t) = T(U(t)), \qquad f^h(t) = T(F(t)), \qquad g^h = T(G),$$

where $T : \mathbb{R}^{m\times m} \to \mathbb{R}^{m^2}$, defined by (3.18), is the linear operator that piles the columns of a matrix into a vector.

Taking into account the boundary conditions, a spatially discretized version of (4.15) reads (cf. (3.21) and Remark 3.5)

$$\begin{cases} (u^h)'(t) = c(I \otimes \Delta^h_{N-N} + \Delta^h_{D-N} \otimes I)u^h(t) + f^h(t), & t > 0, \\ u^h(0) = g^h, \end{cases} \tag{4.16}$$

where $I \in \mathbb{R}^{m\times m}$ is the identity matrix. This IVP can subsequently be solved, e.g., by the implicit midpoint rule, resulting in

$$\begin{cases} u^h_{l+1} = (I - B)^{-1}\Big[(I + B)u^h_l + \delta f^h(t_l + \delta/2)\Big], & l = 0,1,\dots, \\ u^h_0 = g^h, \end{cases} \tag{4.17}$$

where $\delta > 0$ is the time step size and $B = \frac{1}{2}\delta c(I \otimes \Delta^h_{N-N} + \Delta^h_{D-N} \otimes I) \in \mathbb{R}^{m^2\times m^2}$. In such a discretization,

$$\left(T^{-1}(u^h_l)\right)_{j,k} \approx u(jh, kh, l\delta), \qquad j,k = 1,\dots,m, \quad l = 0,1,\dots,$$

with $u : \Omega \times (0,\infty) \to \mathbb{R}$ being the exact solution of (4.15).

Remark 4.5. Different homogeneous boundary conditions on the edges of the unit square can be handled by choosing the appropriate difference matrices (Δ_{D-D}, Δ_{N-D}, Δ_{D-N}, Δ_{N-N}). Nonhomogeneous boundary conditions affect also the source vector f^h.

CHAPTER 5

Numerical solution of the wave equation (finite difference method)

Let us start by considering a one-(spatial)-dimensional initial and boundary value problem for the wave equation:

$$\begin{cases} u_{tt}(x,t) = c^2 u_{xx}(x,t), & t > 0,\ x \in (0,1), \\ u(0,t) = u(1,t) = 0, & t > 0, \\ u(x,0) = f(x),\ \ u_t(x,0) = g(x), & x \in (0,1), \end{cases} \tag{5.1}$$

where $c > 0$ is the wave 'speed'. If the initial data $f : (0,1) \to \mathbb{R}$ and $g : (0,1) \to \mathbb{R}$ are regular enough, it can be demonstrated that the problem (5.1) has a unique solution $u : [0,1] \times [0,\infty) \to \mathbb{R}$. One possible physical interpretation of (5.1) is as follows: An 'ideal string' with fixed endpoints has initial shape f and initial velocity g. The solution of (5.1) at time $t > 0$, i.e. $u(\cdot\,, t)$, represents the shape of the string at that time. Unlike for the heat equation, the solution of (5.1) is composed of infinitely oscillating components. In particular, the 'energy' of the solution to (5.1) is constant over time:

Lemma 5.1. *Any smooth enough solution $u : [0,1] \times [0,\infty) \to \mathbb{R}$ of (5.1) satisfies*

$$\|u_t(\cdot\,,t)\|^2_{L^2([0,1])} + c^2 \|u_x(\cdot\,,t)\|^2_{L^2([0,1])} = \|g\|^2_{L^2([0,1])} + c^2 \|f'\|^2_{L^2([0,1])}$$

for all $t \geq 0$. The first term on the left can be interpreted as the kinetic energy and the second term as the potential energy. In particular, the total energy of the system is constant in time.

PROOF. To begin with, recall the definition of the norm $\|\cdot\|_{L^2([0,1])}$:

$$\|v\|_{L^2([0,1])} = \left(\int_0^1 |v(x)|^2 \, dx \right)^{1/2}.$$

By bravely changing the order of differentiation and integration, we deduce that

$$\frac{d}{dt} \|u_t(\cdot\,,t)\|^2_{L^2([0,1])} = \frac{d}{dt} \int_0^1 u_t(x,t)^2 \, dx = 2 \int_0^1 u_t(x,t) u_{tt}(x,t) \, dx. \tag{5.2}$$

Similarly,

$$\begin{aligned} \frac{d}{dt} \|u_x(\cdot\,,t)\|^2_{L^2([0,1])} &= 2 \int_0^1 u_{tx}(x,t) u_x(x,t) \, dx = 2 \int_0^1 u_{xt}(x,t) u_x(x,t) \, dx \\ &= 2 \Big[u_t(x,t) u_x(x,t) \Big]_{x=0}^{x=1} - 2 \int_0^1 u_t(x,t) u_{xx}(x,t) \, dx, \end{aligned} \tag{5.3}$$

where the last step follows from integration by parts. Since $u(0,\cdot)$ and $u(1,\cdot)$ are identically zero due to the boundary conditions of (5.1), so are their (time) derivatives, which means that the first term on the right-hand side of (5.3) vanishes. Thus, combining (5.2) and (5.3) yields

$$\frac{d}{dt}\left(\|u_t(\cdot,t)\|^2_{L^2([0,1])} + c^2\|u_x(\cdot,t)\|^2_{L^2([0,1])}\right) = 2\int_0^1 u_t(x,t)\Big(u_{tt}(x,t) - c^2u_{xx}(x,t)\Big)\,dx = 0$$

because u is a solution of (5.1). In consequence, the energy is constant in time. In particular, for any $t \geq 0$,

$$\begin{aligned}\|u_t(\cdot,t)\|_{L^2([0,1])} + c^2\|u_x(\cdot,t)\|_{L^2([0,1])} &= \|u_t(\cdot,0)\|_{L^2([0,1])} + c^2\|u_x(\cdot,0)\|_{L^2([0,1])}\\ &= \|g\|_{L^2([0,1])} + c^2\|f'\|_{L^2([0,1])},\end{aligned}$$

which completes the proof. □

As in the case of the heat equation, it is advisable to first discretize (5.1) with respect to the spatial variable. We continue to employ the grid points (3.2) and use the standard difference formula for the second (spatial) derivative to obtain

$$u_{tt}(x_j,t) = \frac{c^2}{h^2}\Big(u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)\Big) + O(h^2), \qquad j = 1,2,\ldots,m,$$

assuming that the exact solution of (5.1) is smooth enough. Taking the homogeneous Dirichlet boundary conditions of (5.1) into account and loosing the error term $O(h^2)$, this can be written in the form

$$\begin{cases} (u^h)''(t) = c^2\Delta^h_{D-D}u^h(t), & t > 0,\\ u^h(0) = f^h, \quad (u^h)'(0) = g^h, \end{cases} \tag{5.4}$$

where Δ^h_{D-D} is the 'one-dimensional' difference matrix with Dirichlet boundary conditions,

$$f^h(t) = \begin{bmatrix} f(x_1)\\ f(x_2)\\ \vdots\\ f(x_m)\end{bmatrix}, \qquad g^h = \begin{bmatrix} g(x_1)\\ g(x_2)\\ \vdots\\ g(x_m)\end{bmatrix} \tag{5.5}$$

and $u^h : [0,\infty) \to \mathbb{R}^m$ hopefully satisfies $u^h_j(t) \approx u(x_j,t)$, $j = 1,2,\ldots,m$.

We will consider two alternative ways of discretizing (4.3) with respect to time.

5.1. Representation as a first order system

Let us introduce an auxiliary function[1]

$$w^h = \begin{bmatrix} u^h\\ (u^h)'\end{bmatrix} : [0,\infty) \to \mathbb{R}^{2m}.$$

Based on (5.4), it is straightforward to verify that

$$(w^h)'(t) = \begin{bmatrix} (u^h)'(t)\\ (u^h)''(t)\end{bmatrix} = \begin{bmatrix} 0 & I\\ c^2\Delta^h_{D-D} & 0\end{bmatrix}\begin{bmatrix} u^h(t)\\ (u^h)'(t)\end{bmatrix}, \qquad t > 0,$$

[1]This is actually a standard trick for reducing higher order IVPs into first order systems.

where $0 \in \mathbb{R}^{m\times m}$ and $I \in \mathbb{R}^{m\times m}$ are the zero and the identity matrix, respectively. Denoting

$$M^h_{D-D} = \begin{bmatrix} 0 & I \\ c^2\Delta^h_{D-D} & 0 \end{bmatrix} \in \mathbb{R}^{2m\times 2m},$$

the spatially discretized problem (5.4) can be written in the alternative form

$$\begin{cases} (w^h)'(t) = M^h_{D-D} w^h(t), & t > 0, \\ w^h(0) = w^h_0, \end{cases} \tag{5.6}$$

where $w^h_0 = [f^h, g^h]^\mathrm{T} \in \mathbb{R}^{2m\times 2m}$.

The IVP (5.6) can now be solved numerically using one of the methods introduced in Chapter 2. More precisely, the Euler's method gives

$$w^h_{k+1} = w^h_k + \delta M^h_{D-D} w^h_k, \qquad k = 0, 1, 2, \ldots, \tag{5.7}$$

the implicit Euler's method results in

$$w^h_{k+1} = w^h_k + \delta M^h_{D-D} w^h_{k+1} \iff w^h_{k+1} = (I - \delta M^h_{D-D})^{-1} w^h_k, \qquad k = 0, 1, 2, \ldots, \tag{5.8}$$

and the implicit midpoint rule leads to

$$w^h_{k+1} = w^h_k + \delta M^h_{D-D}\Big(\frac{1}{2}(w^h_k + w^h_{k+1})\Big) \iff w^h_{k+1} = \Big(I - \frac{1}{2}\delta M^h_{D-D}\Big)^{-1}\Big(I + \frac{1}{2}\delta M^h_{D-D}\Big) w^h_k, \tag{5.9}$$

where $k = 0, 1, 2, \ldots$. Notice, in particular, that the matrices $(I - \delta M^h_{D-D})^{-1}$ and $(I - \frac{1}{2}\delta M^h_{D-D})^{-1}$ needed in (5.8) and (5.9), respectively, are well defined for any $\delta > 0$: The to-be-inverted matrices have full sets of distinct nonzero eigenvalues, cf. Homework 11. In all of the methods (5.7)–(5.9), the main idea is that

$$\begin{cases} (w^h_k)_j \approx u(x_j, t_k), & j = 1, 2, \ldots, m, \\ (w^h_k)_j \approx u_t(x_j, t_k), & j = m+1, m+2, \ldots, 2m, \end{cases}$$

meaning that one simultaneously obtains estimates for both the solution of (5.1) and its time derivative.

As implied by Lemma 5.1, the solution of (5.1) is composed of infinitely lasting oscillations; in particular, it does not include exponentially growing or decaying components. Any reasonable method for solving (5.1) should have this same property: The produced numerical solution should retain its 'energy', without 'exploding' or 'vanishing' as time moves forward. It can be argued that from this view point the method (5.9) is superior to (5.7) and (5.8) — in fact, there is no good reason for using (5.7) in practice.

Let us try to be a bit more explicit. Each of the methods (5.7), (5.8) and (5.9) can be given in the form

$$w^h_k = B^k w^h_0, \qquad k = 0, 1, 2, \ldots, \tag{5.10}$$

where B is either $I + \delta M^h_{D-D}$, $(I - \delta M^h_{D-D})^{-1}$ or $(I - \frac{1}{2}\delta M^h_{D-D})^{-1}(I + \frac{1}{2}\delta M^h_{D-D})$. All of these matrices have the same linearly independent eigenvectors $v^j \in \mathbb{C}^{2m}$, $j = 1, 2, \ldots, 2m$ (cf. Homework 11 and 12); however, the corresponding eigenvalues vary depending on the version of B. Let us denote

$$V = [v^1, v^2, \ldots, v^{2m}] \in \mathbb{C}^{2m\times 2m}.$$

Since the eigenvectors are linearly independent, any initial condition $w_0^h \in \mathbb{R}^{2m}$ for (5.6) can be written as their linear combination, that is, there exist $a = [a_1, a_2, \dots a_{2m}]^{\mathrm{T}} \in \mathbb{C}^{2m}$ such that

$$w_0^h = \sum_{l=1}^{2m} a_l v^l = Va.$$

As B has a full set of linearly independent eigenvectors, it is known to be diagonalizable, i.e.,

$$B = V\Lambda V^{-1}, \tag{5.11}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_{2m}) \in \mathbb{C}^{2m\times 2m}$ is a diagonal matrix carrying the eigenvalues of the considered $B \in \mathbb{C}^{2m\times 2m}$.[2] In consequence, the recursion (5.10) can be written as

$$w_k^h = (V\Lambda V^{-1})^k Va = V\Lambda^k a = V \begin{bmatrix} \lambda_1^k & & & \\ & \lambda_2^k & & \\ & & \ddots & \\ & & & \lambda_{2m}^k \end{bmatrix} a = \sum_{j=1}^{2m} a_j \lambda_j^k v^j, \tag{5.12}$$

where $k = 0, 1, 2, \dots$. Due to (5.12), it is obvious that the general behavior of the numerical solution produced by (5.10) as $k \to \infty$ depends on the magnitudes of the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_{2m} \in \mathbb{C}$.

For the method (5.9), which is based on the implicit midpoint rule, all eigenvalues of the system matrix $B = (I - \frac{1}{2}\delta M_{D-D}^h)^{-1}(I + \frac{1}{2}\delta M_{D-D}^h)$ are known to be of magnitude one (Homework 12), that is, in the formula (5.12) it holds that $|\lambda_j| = 1$, $j = 1, 2, \dots, 2m$, independently of $\delta > 0$. As a consequence, the numerical solution produced by (5.9) does not approach zero or blow up when $k \to \infty$, but the corresponding numerical solution w_k^h oscillates infinitely as the coefficients λ_j^k in (5.12) rotate around the unit circle in the complex plane. Consequently, the method (5.9) has the desired *stability* behavior.

In case of (5.7), i.e., when the time discretization is performed by the (explicit) Euler's method, *all* eigenvalues of $B = I + \delta M_{D-D}^h$ are known to have magnitude grater than one: $|\lambda_j| > 1$, $j = 1, 2, \dots 2m$. Hence,

$$\lim_{k\to\infty} |w_k^h| = \infty$$

for any nonzero $a \in \mathbb{C}^m$, i.e. for any nonzero $w_0^h = [f^h, g^h]^{\mathrm{T}}$. This blow-up is particularly severe if $\delta > 0$ is large or the initial data f, g for (5.1) are irregular, meaning that they contain high spatial frequencies.

Finally, the eigenvalues of the system matrix $B = (I - \delta M_{D-D}^h)^{-1}$ for (5.8) satisfy $|\lambda_j| < 1$ for all $j = 1, 2, \dots, 2m$, and it follows trivially from (5.12) that

$$\lim_{k\to\infty} w_k^h = 0 \in \mathbb{R}^{2m}$$

independently of $a \in \mathbb{C}^{2m}$, i.e. independently of the initial data $w_0^h \in \mathbb{R}^{2m}$. The high spatial frequencies decay particularly fast in the scheme (5.8): the corners in an angular wave disappear almost instantly.

[2] Since V is composed of the eigenvectors of B, we have $BV = [Bv^1, \dots, Bv^{2m}] = [\lambda_1 v^1, \dots, \lambda_{2m} v^{2m}] = V\Lambda$, and thus (5.11) follows by multiplying from the right by V^{-1}, which exists as the columns of V are linearly independent.

To sum up, it is often a good idea to use the implicit midpoint rule for solving the IVP (5.4) that results from the spatial discretization of (5.1). On the other hand, utilization of the (explicit) Euler's method is (almost) never an appropriate approach.

5.2. Direct discretization of the second time derivative

We continue to consider the spatially discretized wave equation (5.4), but this time around we will directly discretize the second time derivative. As usual, we choose a time step size $\delta > 0$ and denote by $u_k^h \in \mathbb{R}^m$ the estimate for the solution $u^h : [0, \infty) \to \mathbb{R}^m$ of (5.4) at time $t_k = k\delta$. By using the standard difference formula for the second (time!) derivative at t_k, we get an approximate version of (5.4):

$$\frac{1}{\delta^2}\left(u_{k+1}^h - 2u_k^h + u_{k-1}^h\right) = c^2 \Delta_{D-D}^h u_k^h, \qquad k = 1, 2, \dots .$$

Solving for u_{k+1}^h leads to the recursion

$$u_{k+1}^h = (2I + \delta^2 c^2 \Delta_{D-D}^h) u_k^h - u_{k-1}^h, \qquad k = 1, 2, \dots . \tag{5.13}$$

This is a so-called two-step method: Initializing (5.13) requires the first two iterates u_0^h and u_1^h — they are needed for computing u_2^h via (5.13) after which the recursion proceeds naturally.

The first initial condition of (5.4) is an obvious candidate for u_0^h, that is, we set $u_0^h = f^h \in \mathbb{R}^m$. On the other hand, assuming that the exact solution u^h of the spatially discretized problem (5.4) is smooth enough, we may write a second order Taylor's expansion at the origin:

$$\begin{aligned} u^h(t_1) = u^h(\delta) &= u^h(0) + \delta (u^h)'(0) + \frac{\delta^2}{2}(u^h)''(0) + O(\delta^3) \\ &= u^h(0) + \delta (u^h)'(0) + \frac{\delta^2 c^2}{2}\Delta_{D-D}^h u^h(0) + O(\delta^3) \\ &= \left(I + \frac{\delta^2 c^2}{2}\Delta_{D-D}^h\right) f^h + \delta g^h + O(\delta^3), \end{aligned} \tag{5.14}$$

where the penultimate step follows from the first equation of (5.4) and the last one from the corresponding initial conditions. Dropping the error term $O(\delta^3)$ from (5.14) gives our choice of u_1^h. The complete form of the numerical algorithm thus reads

$$\begin{cases} u_{k+1}^h = (2I + \delta^2 c^2 \Delta_{D-D}^h) u_k^h - u_{k-1}^h, \qquad k = 1, 2, \dots, \\ u_0^h = f^h, \quad u_1^h = \left(I + \frac{\delta^2 c^2}{2}\Delta_{D-D}^h\right) f^h + \delta g^h, \end{cases} \tag{5.15}$$

where f^h, $g^h \in \mathbb{R}^m$ are as in (5.5) and, hopefully, $(u_k^h)_j \approx u(x_j, t_k)$, with $j = 1, \dots, m$ and $k = 0, 1, \dots$.

As for the solution techniques considered in Section 5.1, it is desirable that the numerical method (5.15) does not produce solutions that contain exponentially increasing or decaying components. We will next investigate what kind of a condition needs to be imposed on the parameters $h = 1/(m+1)$ and $\delta > 0$ in order to reach this objective. For simplicity, we assume that $g^h = 0 \in \mathbb{R}^m$ and $f^h = v$ is one of the eigenvectors of the difference matrix Δ_{D-D}^h (cf. Lemma 3.1);

the resulting condition can straightforwardly be extended for more general initial value vectors.

Lemma 5.2. *Let $v \in \mathbb{R}^m$ be an eigenvector of Δ^h_{D-D} with the corresponding eigenvalue $\lambda \in (-\frac{4}{h^2}, 0)$. For the initial conditions $f^h = v$ and $g^h = 0$, the numerical solution produced by (5.15) satisfies $u^h_k = \beta_k v$, where the coefficients $\beta_k \in \mathbb{R}$, $k = 0, 1, 2, \ldots$, satisfy the recursion*

$$\beta_{k+1} = (2 + \delta^2 c^2 \lambda)\beta_k - \beta_{k-1}, \qquad k = 1, 2, \ldots . \tag{5.16}$$

PROOF. Due to the choice of the initial vectors, (5.15) yields

$$u^h_0 = v, \qquad u^h_1 = \Big(I + \frac{1}{2}\delta^2 c^2 \Delta^h_{D-D}\Big)v + \delta 0 = \Big(1 + \frac{1}{2}\delta^2 c^2 \lambda\Big)v, \tag{5.17}$$

which shows that $\beta_0 = 1$ and $\beta_1 = 1 + \frac{1}{2}\delta^2 c^2 \lambda$.

Let us then make an induction assumption that $u^h_{k-1} = \beta_{k-1} v$ and $u^h_k = \beta_k v$ for some $\beta_{k-1}, \beta_k \in \mathbb{R}$ and an arbitrary but fixed $k \in \mathbb{N}$; we already know that this holds for $k = 1$, and thus the claim follows by induction if we can show that $u^h_{k+1} = \beta_{k+1} v$ with β_{k+1} given by (5.16). Indeed,

$$\begin{aligned} u^h_{k+1} &= (2I + \delta^2 c^2 \Delta^h_{D-D})u^h_k - u^h_{k-1} = \beta_k (2I + \delta^2 c^2 \Delta^h_{D-D})v - \beta_{k-1} v \\ &= \beta_k (2 + \delta^2 c^2 \lambda)v - \beta_{k-1} v = \Big((2 + \delta^2 c^2 \lambda)\beta_k - \beta_{k-1}\Big)v, \end{aligned}$$

which completes the proof. □

We continue to assume the initial conditions $f^h = v$ and $g^h = 0$ and note that the recursion (5.16) can be presented in the 'matrix form'

$$\eta^{k+1} = A\eta^k \iff \eta^k = A^{k-1}\eta_1, \qquad k = 1, 2, \ldots,$$

where

$$\eta^k = \begin{bmatrix} \beta_k \\ \beta_{k-1} \end{bmatrix} \in \mathbb{R}^2 \qquad \text{and} \qquad A = \begin{bmatrix} 2 + \delta^2 c^2 \lambda & -1 \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

The corresponding 'initialization' is $\eta_1 = [1 + \frac{1}{2}\delta^2 c^2 \lambda, 1]^{\mathrm{T}} \in \mathbb{R}^2$ given by (5.17). In a similar manner as in Section 5.1, it can be reasoned that the numerical solution produced by (5.15) does not decay to zero or blow up — for the considered, simple initial data — if and only if both eigenvalues of A have magnitude one (Recall that $u^h_k = \beta_k v = \eta^k_1 v$.)

The eigenvalues of A can be solved from the polynomial equation

$$\det(A - \mu I) = \begin{vmatrix} 2 + \delta^2 c^2 \lambda - \mu & -1 \\ 1 & -\mu \end{vmatrix} = 0,$$

which is equivalent to

$$\mu^2 - \alpha\mu + 1 = 0 \iff \mu_\pm = \frac{1}{2}\Big(\alpha \pm \sqrt{\alpha^2 - 4}\Big),$$

where $\alpha = 2 + \delta^2 c^2 \lambda$. If $|\alpha| > 2$, the discriminant $\alpha^2 - 4$ is positive and $\mathbb{R} \ni \mu_+ > \mu_- \in \mathbb{R}$, and thus the absolute value of (at least) one of the two eigenvalues must be larger than one.[3] In consequence, we must have $|\alpha| \leq 2$, whence $\sqrt{4 - \alpha^2} \in \mathbb{R}$ and

$$\mu_\pm = \frac{1}{2}\Big(\alpha \pm i\sqrt{4 - \alpha^2}\Big).$$

[3] The case $\mu_+ = 1$ and $\mu_- = -1$ can be easily excluded.

In particular,

$$|\mu_\pm|^2 = \frac{1}{4}\Big(\alpha^2 + (4-\alpha^2)\Big) = 1,$$

which is good.

The *stability condition* of the method (5.15) is thus

$$-2 \le \alpha = 2 + \delta^2 c^2 \lambda \le 2 \tag{5.18}$$

where $\lambda \in (-\frac{4}{h^2}, 0)$ is still an (arbitrary) eigenvalue of the difference matrix Δ^h_{D-D}. By virtue of the negativity of λ, only the left-hand inequality of (5.18) imposes a nontrivial condition on $\delta > 0$:

$$\delta^2 c^2 \lambda > -4 \iff \delta \le \frac{2}{c\sqrt{-\lambda}}\,.$$

In consequence, the 'worst case', i.e., the most restrictive condition for $\delta > 0$, is encountered at the left endpoint of the interval $(-\frac{4}{h^2}, 0)$ for λ:

$$\delta \le \frac{h}{c},$$

which guarantees that the numerical solution produces by (5.15) continues to oscillate for eternity.

5.3. Generalizations

If the boundary conditions of (5.1) are replaced by some other (homogeneous) ones, the only thing that changes in the spatially discretized problem (5.4) is the difference matrix. As an example, the spatially discretized IVP corresponding to

$$\begin{cases} u_{tt}(x,t) = c^2 u_{xx}(x,t), & t > 0,\ x \in (0,1), \\ -u_x(0,t) = u(1,t) = 0, & t > 0, \\ u(x,0) = f(x),\ \ u_t(x,0) = g(x), & x \in (0,1), \end{cases}$$

is

$$\begin{cases} (u^h)''(t) = c^2 \Delta^h_{N-D} u^h(t), & t > 0, \\ u^h(0) = f^h,\ \ (u^h)'(0) = g^h, \end{cases}$$

which can be solved — depending on the prevailing mood — following either the methodology of Section 5.1 or that Section 5.2 (and substituting Δ^h_{N-D} for Δ^h_{D-D} in all formulas).

The discretization of the wave equation in the two-dimensional unit square is performed in the same way as for the heat equation in Chapter 4. Indeed, applying the standard discretization of the two-dimensional Laplace operator to, e.g.,

$$\begin{cases} u_{tt}(x,t) = c^2 \Delta u(x,t), & t > 0,\ x \in \Omega, \\ \dfrac{\partial}{\partial \nu} u(x,t) = 0, & t > 0,\ x \in \partial\Omega, \\ u(x,0) = f(x),\ \ u_t(x,0) = g(x), & x \in \Omega, \end{cases} \tag{5.19}$$

where $\Omega = (0,1) \times (0,1)$ and $\frac{\partial}{\partial \nu} v = \nu \cdot \nabla v$ denotes the normal derivative of v, i.e. the derivative of v in the direction of the unit normal vector $\nu : \partial\Omega \to \mathbb{R}^2$ of

the boundary $\partial\Omega$, leads to the spatially discretized IVP

$$\begin{cases} (u^h)''(t) = c(I \otimes \Delta^h_{N-N} + \Delta^h_{N-N} \otimes I)u^h(t), & t > 0, \\ u^h(0) = f^h, \quad (u^h)'(0) = g^h, \end{cases} \tag{5.20}$$

which can once again be numerically solved by resorting to the ideas of Sections 5.1 and 5.2. Here, $u^h : [0, \infty) \to \mathbb{R}^{m^2}$ and $f^h, g^h \in \mathbb{R}^{m^2}$ have the interpretations

$$\left(T^{-1}(u^h(t))\right)_{j,k} \approx u(jh, kh, t), \quad \left(T^{-1}(f^h)\right)_{j,k} \approx f(jh, kh), \quad \left(T^{-1}(g^h)\right)_{j,k} \approx g(jh, kh),$$

where $j, k = 1, 2, \ldots, m$ and $T : \mathbb{R}^{m \times m} \to \mathbb{R}^{m^2}$ is the linear operator defined by (3.18), with $T^{-1} : \mathbb{R}^{m^2} \to \mathbb{R}^{m \times m}$ being its inverse map. As in the one-dimensional case, a change in the boundary conditions of (5.19) can be handled by choosing the appropriate difference matrices for (5.20).

The considerations of this chapter could also be extended to the case of nonhomogeneous boundary conditions, but the details are omitted.

CHAPTER 6

Galerkin approximation and finite element method (in a nutshell)

We start by considering the one-dimensional elliptic boundary value problem

$$\begin{cases} -\dfrac{d}{dx}\Big(\kappa(x)\dfrac{d}{dx}u(x)\Big) + q(x)u(x) = f(x), & x \in (0,1), \\ u(0) = 0, \quad \kappa(1)u_x(1) + \gamma u(1) = \beta, \end{cases} \tag{6.1}$$

where $\beta, \gamma \in \mathbb{R}$, $\kappa : [0,1] \to \mathbb{R}$ is continuously differentiable and $q, f : [0,1] \to \mathbb{R}$ are continuous. In particular, choosing $\kappa \equiv 1$ and $q \equiv 0$ gives the Laplace equation with a homogeneous Dirichlet condition on the left and a Robin condition on the right. We assume that (6.1) has a unique solution and define a (test) function space V via[1]

$$V = \Big\{ v \in L^2([0,1]) \;\Big|\; v(0) = 0,\ v' \in L^2([0,1]) \Big\}. \tag{6.2}$$

In particular, $u \in V$.

Our aim is to write (6.1) in an alternative, integral form. To this end, we multiply the first equation of (6.1) by an arbitrary $v \in V$ and integrate over the interval $[0,1]$:

$$-\int_0^1 \frac{d}{dx}\Big(\kappa(x)\frac{d}{dx}u(x)\Big)v(x)\,dx + \int_0^1 q(x)u(x)v(x)\,dx \;=\; \int_0^1 f(x)v(x)\,dx.$$

Via integration by parts this can be transformed into

$$-\left(\Big[\kappa(x)u'(x)v(x)\Big]_{x=0}^{x=1} - \int_0^1 \kappa(x)u'(x)v'(x)\,dx\right) + \int_0^1 q(x)u(x)v(x)\,dx \;=\; \int_0^1 f(x)v(x)\,dx,$$

and accounting for the boundary conditions of (6.1) in the substitution term finally yields

$$\int_0^1 \kappa(x)u'(x)v'(x)\,dx + \int_0^1 q(x)u(x)v(x)\,dx \;+ \gamma u(1)v(1) \;=\; \int_0^1 f(x)v(x)\,dx + \beta v(1), \tag{6.3}$$

for all $v \in V$.

Definition 6.1. The equation (6.3) is called the variational formulation (or version) of (6.1).

Several remarks are in order:

[1] In fact, V is a so-called *Sobolev space*.

- To be quite precise, the derivative v' appearing in the definition of the variational space V is the so-called weak (or distributional) derivative of v, which exists for any $v \in L^2([0,1])$, but does not necessarily belong to $L^2([0,1])$ itself. (Be that as it may, you are allowed to just think about standard derivatives on this course.)
- It follows from the Cauchy–Schwarz inequality for integrals that the variational equation (6.3) is well defined if it is just assumed that $\kappa : [0,1] \to \mathbb{R}$ and $q : [0,1] \to \mathbb{R}$ are bounded (assuming that the solution itself belongs to V).
- The homogeneous Dirichlet boundary condition of (6.1) is visible in the definition of the variational space V, but the Robin condition (or the Neumann condition if $\gamma = 0$) is included implicitly in (6.3). (It is also possible to treat nonhomogeneous Dirichlet boundary conditions, but the situation gets *slightly* more complicated, and thus the details are omitted.)
- If the derivatives and boundary conditions of (6.1) are interpreted in a suitable *weak* manner, the original boundary value problem (6.1) is actually *equivalent* to its variational version (6.3).

Summary: *The variational problem* (6.3) *is well defined under more general assumptions than the (classical) boundary value problem* (6.1)*. However, the two problems are equivalent if* (6.1) *is interpreted in an appropriate (weak) sense*.

6.1. Galerkin approximation

Naturally, it is impossible to numerically solve (6.3) as such: The function space V is infinite-dimensional, which means that the solution $u \in V$ does not usually allow a 'finite-dimensional parametrization'. Moreover, there are infinite number of test functions $v \in V$ for which (6.3) is required to hold. A natural way to circumvent these problems related to the infinite-dimensionality of (6.3) is to search for an approximate solution u_h in some finite-dimensional subspace $V_h \subset V$ and to only require that (6.3) holds for all $v \in V_h$. Such $u_h \in V_h$ is called the *Galerkin approximation* of $u \in V$ in the subspace V_h.

For simplicity, let us assume that $\gamma = 0$ in (6.1), which means that the variational problem (6.3) takes the form

$$\int_0^1 \kappa(x)u'(x)v'(x)\,dx + \int_0^1 q(x)u(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx + \beta v(1) \tag{6.4}$$

for all $v \in V$. Given a finite-dimensional subspace $V_h \subset V$, our aim is to find $u_h \in V_h$ that satisfies the same equation as u, that is,

$$\int_0^1 \kappa(x)u_h'(x)v'(x)\,dx + \int_0^1 q(x)u_h(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx + \beta v(1) \tag{6.5}$$

but only for all $v \in V_h$. Since V_h is by assumption finite-dimensional, there exists a set of linearly independent basis functions for V_h,

$$v_1, v_2, \ldots, v_m \in V_h \subset V,$$

where $m \in \mathbb{N}$ is the dimension of V_h. This means that any $v \in V_h$ can be given as a (unique) linear combination ("basis")

$$v = \sum_{k=1}^{m} a_k v_k \tag{6.6}$$

and ("linear independence")

$$\sum_{k=1}^{m} a_k v_k \equiv 0 \quad \Longleftrightarrow \quad a_k = 0, \quad k = 1, 2, \ldots, m. \tag{6.7}$$

Without loss of generality, the solution $u_h \in V_h$ of (6.5) can thus be searched in the form

$$u_h = \sum_{k=1}^{m} a_k v_k. \tag{6.8}$$

As both sides of (6.5) are linear with respect to v, it is also easy to check that the condition "for all $v \in V_h$" can be replaced by the equivalent condition "for all v_j, $j = 1, 2, \ldots, m$".

Altogether we have thus reasoned that the problem of finding the Galerkin approximation for the exact solution $u \in V$ of (6.4) in the finite-dimensional subspace V_h can be (re)formulated as follows: Find a coefficient vector $a \in \mathbb{R}^m$ such that

$$\int_0^1 \kappa(x)\Big(\sum_{k=1}^{m} a_k v_k'(x)\Big)v_j'(x)\,dx + \int_0^1 q(x)\Big(\sum_{k=1}^{m} a_k v_k(x)\Big)v_j(x)\,dx = \int_0^1 f(x)v_j(x)\,dx + \beta v_j(1)$$

for all $j = 1, 2, \ldots, m$. This can be written equivalently as

$$\sum_{k=1}^{m}\left(\int_0^1 \kappa(x)v_k'(x)v_j'(x)\,dx + \int_0^1 q(x)v_k(x)v_j(x)\,dx\right)a_k = \int_0^1 f(x)v_j(x)\,dx + \beta v_j(1), \tag{6.9}$$

for $j = 1, 2, \ldots, m$, which defines a system of m linear equations for the coefficients $a = [a_1, a_2, \ldots, a_m]^{\mathrm{T}}$. Like any system of linear equations, also (6.9) can be given in a matrix form:

$$Aa = d, \tag{6.10}$$

where $A \in \mathbb{R}^{m\times m}$ and $d \in \mathbb{R}^m$ are defined by

$$A_{j,k} = \int_0^1 \kappa(x)v_j'(x)v_k'(x)\,dx + \int_0^1 q(x)v_j(x)v_k(x)\,dx, \qquad j, k = 1, 2, \ldots, m,$$

and

$$d_j = \int_0^1 f(x)v_j(x)\,dx + \beta v_j(1), \qquad j = 1, 2, \ldots, m.$$

Assuming that the basis functions $v_1, \ldots, v_m$ are known,[2] the matrix equation (6.10) can be formed by means of (numerical) integration, and the Galerkin approximation is subsequently obtained by solving (6.10) and substituting the solution coefficients in (6.8). Notice that this procedure produces an approximate

[2] In practice, it is actually typical to first choose the basis function and then define V_h to be the space of their linear combinations, that is, $V_h = \operatorname{span}\{v_1, \ldots, v_m\}$.

solution for (6.1) in a form of a function $u_h : [0,1] \to \mathbb{R}$, not just approximate grid values as does the finite difference method (cf. Chapter 3).

Usually the unique solvability of (6.10) is proved by referring to the unique solvability of the original infinite dimensional variational problem (6.4), which is true under suitable assumptions on the coefficients κ and q. However, because tackling the unique solvability of (6.4) requires certain tools of *functional analysis*[3], we prove the unique solvability of the matrix equation (6.10) in a nonstandard way.

Theorem 6.2. *Assume that the basis functions $v_1, \ldots, v_m : [0,1] \to \mathbb{R}$ are continuously differentiable. If $\kappa(x) \geq c > 0$ and $q(x) \geq 0$ for all $x \in (0,1)$, then (6.10) has a unique solution $a \in \mathbb{R}^m$, that is, the system matrix A is invertible.*

PROOF. We will prove that A is positive definite, which is a stronger property than invertibility.[4] It holds that

$$
\begin{aligned}
a^{\mathrm{T}}Aa &= \sum_{j=1}^{m}\sum_{k=1}^{m} a_j A_{j,k} a_k \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m} a_j \left(\int_0^1 \kappa(x) v_j'(x) v_k'(x)\,dx + \int_0^1 q(x) v_j(x) v_k(x)\,dx \right) a_k \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m} \left(\int_0^1 \kappa(x) a_j v_j'(x) a_k v_k'(x)\,dx + \int_0^1 q(x) a_j v_j(x) a_k v_k(x)\,dx \right) \\
&= \int_0^1 \kappa(x) \sum_{j=1}^{m}\sum_{k=1}^{m} \left(a_j v_j'(x) a_k v_k'(x) \right) dx + \int_0^1 q(x) \sum_{j=1}^{m}\sum_{k=1}^{m} \left(a_j v_j(x) a_k v_k(x) \right) dx \\
&= \int_0^1 \kappa(x) \sum_{j=1}^{m} \left(a_j v_j'(x) \right) \sum_{k=1}^{m} \left(a_k v_k'(x) \right) dx + \int_0^1 q(x) \sum_{j=1}^{m} \left(a_j v_j(x) \right) \sum_{k=1}^{m} \left(a_k v_k(x) \right) dx \\
&= \int_0^1 \kappa(x) \left(\sum_{j=1}^{m} a_j v_j'(x) \right)^2 dx + \int_0^1 q(x) \left(\sum_{j=1}^{m} a_j v_j(x) \right)^2 dx \;\geq\; 0
\end{aligned}
\tag{6.11}
$$

due to the positivity assumptions on κ and q. Furthermore, since κ is strictly positive, the equality at the last step of (6.11) can hold only if

$$
\frac{d}{dx} \left(\sum_{j=1}^{m} a_j v_j(x) \right) = \sum_{j=1}^{m} a_j v_j'(x) = 0 \qquad \text{for all } x \in [0,1],
$$

by virtue of the assumed regularity of the basis functions.[5] In other words, the equality can hold in (6.11) only if

$$
\sum_{j=1}^{m} a_j v_j(x) = c \qquad \text{for all } x \in [0,1], \tag{6.12}
$$

---

[3] Riesz representation theorem or Lax–Milgram lemma.

[4] $(a \neq 0 \Rightarrow a^{\mathrm{T}}Aa > 0) \;\Rightarrow (a \neq 0 \Rightarrow Aa \neq 0) \Leftrightarrow \exists A^{-1}$

[5] If a continuous function is nonzero at a single point, it must also be nonzero in some nonempty open neighborhood of that point.

with some constant $c \in \mathbb{R}$. However, as $v_j \in V_h \subset V$, we have $v_j(0) = 0$ for all $j = 1, 2, \dots, m$, and thus the only possible constant in (6.12) is $c = 0$. Due to the assumption that $v_1, \dots, v_m$ are linearly independent (6.7), this means that $a = 0 \in \mathbb{R}^m$; in particular, the equality holds in (6.11) if only if $a = 0 \in \mathbb{R}^m$.

To sum up, we have shown that

$$a^{\mathrm{T}} A a > 0 \qquad \text{for all } 0 \neq a \in \mathbb{R}^m,$$

which completes the proof. □

6.2. Multidimensional case

Let us demonstrate how to form the variational version of an elliptic boundary value problem in higher dimensions with the help of a simple example: Let $\Omega \subset \mathbb{R}^n$ be a regular enough domain and assume that the boundary of Ω is divided into two disjoint components: $\partial\Omega = \Gamma_D \cup \Gamma_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. Our aim is to deduce the variational formulation of

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma_D, \qquad \frac{\partial u}{\partial \nu} = g \quad \text{on } \Gamma_N, \tag{6.13}$$

where $\frac{\partial}{\partial \nu} u = \nu \cdot \nabla u$ denotes the normal derivative of u, i.e. the derivative of u in the direction of the exterior unit normal vector $\nu : \partial\Omega \to \mathbb{R}^n$ of the boundary $\partial\Omega$.

In this case, the variational space is

$$V = \Big\{ v \in L^2(\Omega) \;\Big|\; v|_{\Gamma_D} \equiv 0; \;\; v_{x_j} \in L^2(\Omega), \; j = 1, \dots, m \Big\}.$$

In particular, observe that the (homogeneous) Dirichlet boundary condition of (6.13) is once again included in the definition of V. We multiply the first equation of (6.13) by an arbitrary element of V and integrate over Ω:

$$-\int_\Omega \Delta u v \, dx = \int_\Omega f v \, dx,$$

According to a Green's formula (cf. Homework 12, Set 1, Problem 4), this can be rewritten in the form

$$-\left(\int_{\partial\Omega} \frac{\partial u}{\partial \nu} v \, ds - \int_\Omega \nabla u \cdot \nabla v \, dx \right) = \int_\Omega f v \, dx. \tag{6.14}$$

Employing the boundary conditions of u and v, the boundary integral term of (6.14) can manipulated as follows:

$$\int_{\partial\Omega} \frac{\partial u}{\partial \nu} v \, ds = \int_{\Gamma_D} \frac{\partial u}{\partial \nu} v \, ds + \int_{\Gamma_N} \frac{\partial u}{\partial \nu} v \, ds = \int_{\Gamma_D} \frac{\partial u}{\partial \nu} 0 \, ds + \int_{\Gamma_N} g v \, ds = \int_{\Gamma_N} g v.$$

Plugging this expression back in (6.14) and reorganizing the terms, we finally deduce that the solution of (6.13), i.e. $u \in V$, should also satisfy the variational equation

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\Gamma_N} g v \, ds \qquad \text{for all } v \in V. \tag{6.15}$$

This is called the variational formulation of the elliptic boundary value problem (6.13).

As in the one-dimensional case:

- The problems (6.13) and (6.15) are equivalent if the former is interpreted in the correct weak sense.
- The Galerkin approximation for $u \in V$ with respect to a given finite-dimensional subspace $V_h \subset V$ is obtained by (numerically) solving (6.15) when V is replaced by V_h and the approximate solution u_h is also sought from V_h.

6.3. Finite element method

The feasibility of the Galerkin approximation in practical computations is affected at least by the following things:

(i) The approximation properties of the chosen subspace $V_h \subset V$. (It can be shown that the Galerkin approximation $u_h \in V_h$ gives in a certain sense the best approximation for the exact solution $u \in V$ in V_h.)
(ii) The computational cost of forming the matrix equation (6.10).
(iii) The computational cost of solving the matrix equation (6.10).

Regarding item (i), it would be natural to choose V_h to be some subspace whose approximation properties are well understood, such as the space of polynomials of a certain order (cf. Section (1.1)). However, this is not necessarily reasonable from the view point of items (ii) and (iii): Since polynomials take nonzero values (almost) everywhere, the formation of the system matrix A and the load vector d requires (numerical) integration over the whole domain for all basis functions v_j, $j = 1, 2, \ldots, m$, which can be expensive. For this same reason, A typically becomes full, i.e., most of its elements are nonzero; the supports of the basis functions v_j and v_k intersect for all $j, k = 1, 2, \ldots, m$.

The leading idea of the finite element method is to choose V_h to be the space spanned by certain piecewise polynomial functions, each of which is supported only in a *small* subset of the examined domain. In such a case, the matrix in (6.10) becomes sparse: Most elements $A_{j,k}$ are zero since the supports of v_j and v_k do not intersect for most $j \neq k$. In consequence, the formation of A is computationally inexpensive (cf. (i)) as is solving the equation (6.10) (cf. (ii)).[6] On the other hand, the predictability of the accuracy of the Galerkin approximation remains good (cf. (iii)) because the task of approximating functions by piecewise polynomials is well studied.

We will demonstrate the basic ideas of the finite element method by considering a simple one-dimensional model problem:

$$\begin{cases} -u''(x) + u(x) = f(x), & x \in (0,1), \\ -u'(0) = 0, \quad u(1) = 0. \end{cases} \tag{6.16}$$

The corresponding (infinite-dimensional) variational space is

$$V = \left\{ v \in L^2([0,1]) \;\middle|\; v(1) = 0,\ v' \in L^2([0,1]) \right\}, \tag{6.17}$$

which takes into account the homogeneous boundary condition of (6.16) at the right endpoint. As for the problem (6.1), by multiplying the first equation of (6.16) by an arbitrary $v \in V$, integrating over $[0,1]$, resorting to partial integration and, finally, utilizing the boundary conditions satisfied by u and v, one

[6] In general, it is far cheaper to invert a sparse than a full matrix.

obtains the variational version of (6.16):

$$\int_0^1 u'(x)v'(x)\,dx + \int_0^1 u(x)v(x)\,dx \;=\; \int_0^1 f(x)v(x)\,dx \qquad \text{for all } v \in V. \tag{6.18}$$

Let us then define the simplest finite element basis functions. We start by choosing a set of (possible nonuniform) grid points

$$0 = x_0 < x_1 < x_2 < \cdots < x_m < x_{m+1} = 1,$$

and introduce the corresponding piecewise linear 'tent functions':

$$v_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \le x \le x_j, \\[2ex] \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \le x \le x_{j+1}, \\[2ex] 0, & x \notin [x_{j-1}, x_{j+1}], \end{cases} \tag{6.19}$$

where $j = 0, 1, \ldots, m+1$ (and we have set $x_{-1} = 1$, $x_{m+2} = 1$). Notice that $v_j(x_{j-1}) = 0 = v_j(x_{j+1})$ and the value $v(x_j) = 1$ is uniquely defined, meaning that the functions v_j, $j = 0, 1, \ldots, m+1$, are continuous. For $1 \le j \le m$, the function v_j is identically zero on $[0, x_{j-1}]$, on $[x_{j-1}, x_j]$ it increases linearly up to the value 1, on $[x_j, x_{j+1}]$ it decreases linearly back down to zero, and on $[x_{j+1}, 1]$ it is again identically zero. The 'outermost' basis function form only a 'half tent': v_0 takes the value 1 at $x = 0$, on $[0, x_1]$ it decreases linearly down to zero, and it is identically zero for the rest of the way; v_{m+1} forms a 'mirror image' of v_0 at the other end point of $[0, 1]$. The (weak) derivative of v_j, $j = 0, 1, \ldots, m+1$, can also be given explicitly as a piecewise constant function:[7]

$$v_j'(x) = \begin{cases} \dfrac{1}{x_j - x_{j-1}}, & x_{j-1} < x < x_j, \\[2ex] \dfrac{1}{x_j - x_{j+1}}, & x_j < x < x_{j+1}, \\[2ex] 0, & x \notin [x_{j-1}, x_{j+1}]. \end{cases} \tag{6.20}$$

Due to the Dirichlet boundary condition in the definition of V in (6.17), $v_{m+1} \notin V$ and thus it is not included as a basis function of V_h:

$$V_h = \operatorname{span}\{v_0, v_1, \ldots, v_m\} = \Big\{ v \in V \;\Big|\; v = \sum_{j=0}^{m} c_j v_j, \;\; c_j \in \mathbb{R} \Big\},$$

which is a space of piecewise linear functions with respect to the grid points $x_0, x_1, \ldots, x_m$. The Galerkin approximation $u^h \in V_h$ that we are interested in is defined as the solution of (cf. (6.18))

$$\int_0^1 u_h'(x)v'(x)\,dx + \int_0^1 u_h(x)v(x)\,dx \;=\; \int_0^1 f(x)v(x)\,dx \qquad \text{for all } v \in V_h \subset V. \tag{6.21}$$

[7] Try not to be discouraged by the fact that v_j is not differentiable in the classical sense at x_{j-1}, x_j and x_{j+1}.

As suggested by (6.8), we search for u^h in the form

$$u_h = \sum_{k=0}^{m} a_j v_j,$$

which results in a matrix equation for $a = [a_0, a_1, \dots, a_m]^{\mathrm{T}} \in \mathbb{R}^{m+1}$,

$$Aa = d,$$

where $A \in \mathbb{R}^{(m+1)\times(m+1)}$ and $d \in \mathbb{R}^{m+1}$ are defined via

$$A_{j,k} = \int_0^1 v_j'(x)v_k'(x)\,dx + \int_0^1 v_j(x)v_k(x)\,dx, \qquad j,k = 0,1,\dots,m, \tag{6.22}$$

and

$$d_j = \int_0^1 f(x)v_j(x)\,dx = \int_{x_{j-1}}^{x_{j+1}} f(x)v_j(x)\,dx, \qquad j = 0,1,\dots,m,$$

because v_j is supported in $[x_{j-1}, x_{j+1}]$. Since the basis functions v_j, $j = 0,1,\dots,m$, and their derivatives are known explicitly, the elements of the matrix A can be written down explicitly with the help of the grid points (a voluntary exercise). Be that as it may, the most important observation is that

$$A_{j,k} = 0 \quad \text{if } |j-k| > 1$$

because for $|j-k| > 1$ the integrands in (6.22) are identically zero due to the 'narrow' supports of the basis functions. In consequence, A is tridiagonal, which means that it is not computationally expensive to form or to invert.

The finite element method has a number of advantages compared to the finite difference method:

- The grid points need not be distributed uniformly, but they can be adjusted based on the properties of the considered problem.
- The handling/discretization of boundary conditions easier — especially, in the case of Neumann and Robin conditions.
- The treatment of complicated geometries is more straightforward in higher spatial dimensions.
- The accuracy of the numerical solution can be increased by increasing the degree of the piecewise polynomial basis functions. (This can even be done locally.)

To complete the discussion, let us investigate what happens if we apply the finite element method with the uniform grid points $x_j = jh = j/(m+1)$, $j = 0,1,\dots,m,m+1$ to the one-dimensional model problem of Section 3.1,

$$\begin{cases} u''(x) = f(x), & x \in (0,1), \\ u(0) = 0, \quad u(1) = 0. \end{cases} \tag{6.23}$$

In this case the variational space contains two Dirichlet conditions,

$$V = \left\{ v \in L^2([0,1]) \;\middle|\; v(0) = v(1) = 0, \;\; v' \in L^2([0,1]) \right\}, \tag{6.24}$$

and the variational formulation of (6.23) is to find $u \in V$ such that[8]

$$-\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx \qquad \text{for all } v \in V. \tag{6.25}$$

[8]Take note that changing one of the two Dirichlet condition into a homogeneous Neumann condition would not alter (6.25) but only the variational space V.

Because of the boundary conditions in (6.24), the first v_0 and last v_{m+1} of the basis functions (6.19) do not belong to V, and thus we define

$$V_h = \operatorname{span}\{v_1, v_2, \dots, v_m\} \subset V$$

and look for the solution $u_h \in V_h$ of the Galerkin problem

$$-\int_0^1 u_h'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx \qquad \text{for all } v \in V_h$$

in the form $u_h = \sum_{k=1}^m a_j v_j$. As above, we end up at the matrix equation

$$Aa = d, \tag{6.26}$$

where $A \in \mathbb{R}^{m\times m}$ and $d \in \mathbb{R}^m$ are defined via

$$A_{j,k} = -\int_0^1 v_j'(x)v_k'(x)\,dx, \qquad j,k = 1,2,\dots,m,$$

and

$$d_j = \int_0^1 f(x)v_j(x)\,dx = \int_{x_{j-1}}^{x_{j+1}} f(x)v_j(x)\,dx, \qquad j = 1,2,\dots,m. \tag{6.27}$$

Recalling (6.20) and that the grid points $x_1, x_2, \dots, x_m$ are now distributed uniformly over $[0,1]$, it is easy to figure out that

$$A_{j,k} = \begin{cases} -\int_{x_{j-1}}^{x_j} \frac{1}{h^2}\,dx - \int_{x_j}^{x_{j+1}} \frac{1}{h^2}\,dx = -\frac{2}{h}, & k = j, \\ -\int_{x_{j-1}}^{x_j} \frac{1}{h(-h)}\,dx = \frac{1}{h}, & k = j-1, \\ -\int_{x_j}^{x_{j+1}} \frac{1}{(-h)h}\,dx = \frac{1}{h}, & k = j+1, \\ 0, & |j-k| > 1. \end{cases}$$

In other words, $A = h\Delta^h_{D-D}$, where $\Delta^h_{D-D} \in \mathbb{R}^{m\times m}$ is the difference matrix with Dirichlet boundary conditions from Chapter 3, and the equation (6.26) can be rewritten as

$$\Delta^h_{D-D} a = \frac{1}{h} d,$$

where the components of d are given by (6.27).

The lesson: In the case of uniform/symmetric meshes, the finite difference method and the finite element method often result in similar matrix equations.

THE END