Daniel Leitner and Otmar Scherzer

Numerical Methods for the Solution of Differential Equations

Lecture Notes

Winter 2015/16

Warning: The lecture notes are subject to change during the semester

Computational Science Center
University of Vienna
A-1090 Vienna, Austria

Contents

Chapter 1

Examples of ODEs

We present a few examples of ordinary differential equations first and indicate some ways to solve them analytically.

1.1 Falling Body

We model the movement of a vertically falling body, which position at time t is described by its height $h(t)$.

Newton's second law of motion implies that the acceleration of the body, that is, the change of its speed, is proportional to the forces acting on the body. That is,

$$F = ma\,,$$

where F denotes the forces, m denotes the mass of the body, and a is its acceleration. Note, that the units of F are Newton (N), of the mass is (kg) and of a is m/s^2 (meter over second squared). The acceleration on the other hand is the change of the speed, which is itself the change of the position of the body. Therefore

$$F = m\ddot{h}(t)\ . \tag{1.1}$$

The main force F is gravity, which, for small heights h, equals approximately mg, where $g \approx 9.81\mathrm{m/s}^2$ is the gravitational acceleration at the earth's surface. The gravitation is acting downwards such that from (1.1) it follows that

$$m\ddot{h}(t) = -mg\ .$$

If either the body is very light or it is falling fast, it is necessary to take into account air friction, which slows down the downwards movement of the body. One possibility is to model air friction as a force proportional to the square of the body's velocity. Because friction always works against the current movement, the sign of the corresponding force will be opposite to the sign of $\dot{h}$. Thus we obtain the refined model

$$m\ddot{h}(t) = -c\,\mathrm{sgn}(\dot{h}(t))\,\dot{h}(t)^2 - mg\,,$$

where c is some material constant describing the drag of the body.

In order to obtain a complete description of the movement of the body, we will need in addition a description of the state of the body at some initial time t_0, where we begin our considerations. More precisely, we will need its initial position h_0 and its initial velocity v_0. Then, assuming this model is correct, the movement of the body is completely described by the *differential equation*

$$\begin{aligned} m\ddot{h}(t) &= -c\,\mathrm{sgn}(\dot{h}(t))\,\dot{h}(t)^2 - mg\,,\\ h(t_0) &= h_0\,,\\ \dot{h}(t_0) &= v_0\;. \end{aligned}$$

1.2 Separation of Variables at the Example of Population Dynamics

Now consider a simple model that describes the evolution of a population over some period of time. That is, we know the population p_0 at some given time t_0, and we want to obtain an estimate $p(t)$ of the population at some future time $t > t_0$.

As a basic model, we assume that the *rate of change* of the population is given by some function $N(t,p)$ that depends only on the time and the size of the population. Then the function p that describes the population solves the differential equation

$$\dot{p}(t) = N\big(t,p(t)\big)\,, \qquad\qquad p(t_0) = p_0\;. \tag{1.2}$$

One very simple model assumes that the number of births and deaths within a certain amount of time is proportional to the size of the population that is,

$$N\big(t,p(t)\big) = (R-S)p(t)\,,$$

with R, S are the birth and death rates, respectively. Thus (1.2) becomes

$$\dot{p}(t) = (R - S)\, p(t)\ .$$

Using the initial state $p(t_0) = p_0$, we obtain with this model the population dynamics

$$p(t) = p_0 e^{(R-S)(t-t_0)}\ .$$

That is, depending on the sign of $R - S$, either the population increases or decreases exponentially.

Now we try to introduce the effects of overpopulation into the model by assuming that the death rate depends on the size of the population. That is, instead of assuming a constant death rate $S > 0$, we assume that S is a function of p. The simplest model is to assume the death rate being proportional to p, setting

$$S(p) = \sigma\, p$$

for some constant $\sigma > 0$. Then we obtain the equation (the *logistic differential equation*)

$$\dot{p}(t) = \big(R - \sigma p(t)\big)\, p(t)\ . \tag{1.3}$$

In the following, we will compute the analytic solution of this equation. We note first that the derivative of p is positive if $R > \sigma p$ (and the population p is positive, which we tacitly assume), while it is negative if $R < \sigma p$. In other words, the population increases as long as $p < R/\sigma$, while it decreases for $p > R/\sigma$. In particular, this implies that the long term behavior of the population will be stagnating at the value $p = R/\sigma$.

In order to solve the logistic differential equation, we define

$$\rho := R/\sigma\,,$$

and rewrite the equation as

$$\frac{1}{(\rho - p(t))\, p(t)} \frac{dp(t)}{dt} = \sigma\ .$$

Integrating both sides of this equation with respect to t, we obtain

$$\int_{t_0}^{\hat{t}} \frac{1}{(\rho - p(t))\, p(t)} \frac{dp(t)}{dt}\, dt = \int_{t_0}^{\hat{t}} \sigma\, dt\ .$$

Now, we make a change of variables $t \to p := p(t)$, such that formally

$$\frac{dp(t)}{dt} dt = dp \, .$$

Using this identity we obtain the equation

$$\int_{p(t_0)}^{p(\hat{t})} \frac{1}{(\rho - p)p} \, dp = \sigma \hat{t} \, , \quad \forall \hat{t} \geq t_0 \, . \tag{1.4}$$

Now note that

$$\begin{aligned}
&\int_{p(t_0)}^{p(\hat{t})} \frac{1}{(\rho - p)p} \, dp \\
=&\frac{1}{\rho} \int_{p(t_0)}^{p(\hat{t})} \frac{1}{\rho - p} + \frac{1}{p} \, dp \\
=&\frac{1}{\rho} \left(\big(- \ln \left|\rho - p(\hat{t})\right| + \ln \left|p(\hat{t})\right| \big) - \underbrace{\big(- \ln |\rho - p(t_0)| + \ln |p(t_0)| \big)}_{=:C} \right) \\
=&\frac{1}{\rho} \ln \left| \frac{p(\hat{t})}{\rho - p(\hat{t}))} \right| - C \, .
\end{aligned}$$

This, together with (1.4) shows that the function p satisfies:

$$\frac{1}{\rho} \ln \left| \frac{p(t)}{\rho - p(t)} \right| = \sigma t + C \, , \qquad \forall t \geq 0 \, .$$

Multiplying the equation with ρ and taking the exponential, it follows that

$$\left| \frac{p(t)}{\rho - p(t)} \right| = e^{\rho \sigma t + \rho C} = e^{\rho \sigma t} \, e^{\rho C} \, ,$$

which is equivalent to

$$\left| \frac{\rho - p(t)}{p(t)} \right| = e^{-\rho \sigma t} \, e^{-\rho C} \, , \tag{1.5}$$

Now we define a new constant

$$D := \pm e^{-\rho C} \, ,$$

where $D > 0$ if $\rho - p(t) > 0$ and else otherwise.

Then this last equation reads as

$$\frac{\rho}{p(t)} - 1 = De^{-\rho\sigma t}\,,$$

which in turn implies that

$$p(t) = \frac{\rho}{1 + De^{-\rho\sigma t}}\,.$$

This is the general form of a solution of the differential equation (1.3). The specific solution satisfying $p(t_0) = t_0$ is determined by determining D from

$$p_0 = p(t_0) = \frac{\rho}{1 + De^{-\rho\sigma t_0}}\,.$$

The method, which we applied in the last example is called *separation of variables*:

Definition 1.1. *An ODE that can be transformed into the form*

$$f(y(t))\dot{y}(t) = g(t)\,, \tag{1.6}$$

where the function $f\colon \mathbb{R} \to \mathbb{R}$ only depends on y and not on t, and the function $g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ only depends on t and not on y, is called ordinary differential equation (of first order) with separable variables*.*

The general strategy for solving such differential equations is to substitute $t \to y := y(t)$. Since

$$\dot{y} = \frac{dy}{dt}\,, \tag{1.7}$$

we can *formally* multiply (1.6) with dt and obtain the *formal* equation

$$f(y)\,dy = g(t)\,dt\,.$$

Now we can apply indefinite integrals to both sides and obtain the equation

$$\int f(y)\,dy = \int g(t)\,dt + C\,,$$

where $C \in \mathbb{R}$ is some constant that appears due to the indefinite integration. Note, that the first integration is with respect to y and the right hand side reveals an integration with respect to t.

If it is possible to compute the integrals of f and g analytically, we obtain an equation the solution necessarily has to satisfy. If, in addition, it is possible to solve this equation for y, we indeed obtain an analytic (general) solution of the differential equation.

Example 1.2. *Consider the ODE*

$$(T^2 - t^2)\,\dot{y} + ty = 0\,,$$

where $T > 0$ is some given constant. This equation is not in a separable form (1.6)*, but can be brought into such. We rewrite the equation to*

$$\frac{\dot{y}}{y} = -\frac{t}{T^2 - t^2}\,,$$

which is possible for $y \neq 0$ and $t \neq \pm T$. We rewrite this formally as

$$\frac{dy}{y} = -\frac{t}{T^2 - t^2}dt\,.$$

Now, integration of both sides of the equation leads to

$$\ln|y| = \frac{1}{2}\ln\left|T^2 - t^2\right| + C\,.$$

Taking the exponential of the equation, we obtain

$$|y| = e^C\sqrt{|T^2 - t^2|}\,.$$

Replacing the constant $e^C > 0$ by the constant $D \in \mathbb{R}$ also encoding the sign of y, we get

$$y(t) = D\sqrt{|T^2 - t^2|}\,. \tag{1.8}$$

The constant $D \in \mathbb{R}$ still has to be determined using the initial condition $y(t_0) = y_0$. Inserting this condition into the general solution, we see that

$$y_0 = y(t_0) = D\sqrt{|T^2 - t_0^2|}\,,$$

and therefore

$$D = \frac{y_0}{\sqrt{|T^2 - t_0^2|}}\,. \tag{1.9}$$

Note that we have assumed during the computation of the solution of the ODE that $y_0 \neq 0$ and $t \neq \pm T$. It can be easily seen, however, that the derivation above also covers the situation where $y_0 = 0$ and $t_0 \neq \pm T$. There, the constant function $y = 0$ is the unique solution of the ODE, at least until the time reaches one of the values $\pm T$.

The case $t_0 = \pm T$, however, is different. Then, if $y_0 = 0$, for every *constant $D \in \mathbb{R}$ the function* (1.8) *satisfies the ODE and therefore is a solution. If, however, $y_0 \neq 0$, then the ODE has no solution at all– the ODE and the initial conditions are inconsistent.*

Finally, note that all the solutions are valid only locally; that is, there exists at least a time interval $[t_0, t_0 + \varepsilon)$ for some $\varepsilon > 0$ on which the solution exists and can be written as (1.8) *with D given by* (1.9)*. For general ODEs, this is all that can be said about the solution. In this special case, one can specify the length of the interval on which the solution looks like* (1.8)*: If $t_0 > T$, then the formula* (1.8) *is valid on $[t_0, +\infty)$. If, however, $-T < t_0 < T$, then the solution is*

$$y(t) = \begin{cases} D_1\sqrt{|T^2 - t^2|} & \textit{if } t \in [t_0, T]\,, \\ D_2\sqrt{|T^2 - t^2|} & \textit{if } t \in [T, +\infty)\,, \end{cases} \qquad \textit{with} \begin{cases} D_1 = y_0/\sqrt{T^2 - t_0^2}\,, \\ D_2 \in \mathbb{R} \textit{ arbitrary.} \end{cases}$$

In particular, the solution is only unique up to time T. Similarly, if $t_0 < -T$, then

$$y(t) = \begin{cases} D_1\sqrt{|T^2 - t^2|} & \textit{if } t \in [t_0, -T]\,, \\ D_2\sqrt{|T^2 - t^2|} & \textit{if } t \in [-T, T]\,, \\ D_3\sqrt{|T^2 - t^2|} & \textit{if } t \in [T, +\infty)\,, \end{cases} \qquad \textit{with} \begin{cases} D_1 = y_0/\sqrt{T^2 - t_0^2}\,, \\ D_2 \in \mathbb{R} \textit{ arbitrary,} \\ D_3 \in \mathbb{R} \textit{ arbitrary.} \end{cases}$$

1.3 Homogeneous ODEs

Definition 1.3. *An ODE of the form*

$$\dot{y} = f\Big(\frac{y}{t}\Big)\,, \tag{1.10}$$

with $f\colon \mathbb{R} \to \mathbb{R}$, is called of homogeneous type*.*

If we are given an ODE of homogeneous type, we can solve it by starting with the substitution

$$z(t) = \frac{y(t)}{t}\,.$$

For the right hand side of (1.10) we are left with the term $f(z)$. For the left hand side of (1.10) we use the product rule and obtain

$$\dot{y} = \frac{dy}{dt} = \frac{d(tz)}{dt} = z + t\frac{dz}{dt} = z + t\dot{z} \ .$$

Thus we have for the variable z the differential equation

$$z + t\dot{z} = f(z) \ .$$

Now it is easy to see that this ODE is of separable type: We can bring it in the form

$$\frac{\dot{z}}{f(z) - z} = \frac{1}{t} \ .$$

This ODE can now be solved by separation of variables, and we obtain a solution $z(t)$. At the end, we obtain the solution y by $y(t) = tz(t)$.

Example 1.4. *Consider the ODE*

$$\dot{y} = \Big(\frac{y}{t}\Big)^2 \ .$$

It is easy to see that this ODE is homogeneous with $f(y/t) = (y/t)^2$. Using the substitution $y = tz$ we obtain

$$z + t\dot{z} = z^2$$

and therefore

$$\frac{\dot{z}}{z^2 - z} = \frac{1}{t} \ .$$

Separation of variables (1.7) *shows that*

$$\frac{1}{z^2 - z} dz = \frac{1}{t} dt \, .$$

Integrating this equation, we obtain the indefinite integral equation

$$\int \frac{1}{z^2 - z} dz = \int \frac{1}{t} dt + C$$

or by calculating the integrals

$$\ln\left|\frac{z-1}{z}\right| = \ln|t| + C \ .$$

Now, assuming that $t > 0$ and $z \geq 1$, which depends on the initial condition, we get

$$\frac{z-1}{z} = Dt$$

for some constant $D = \exp(C) \in \mathbb{R}^+$ depending on the initial value. Solving for z we obtain

$$z(t) = \frac{1}{1-Dt}$$

(note that z is greater than 1) and, after substitution of $y = tz$

$$y(t) = \frac{t}{1-Dt} \; .$$

1.4 Linear ODEs

Definition 1.5. *An ODE that can be written as*

$$\dot{y} + f(t)y = g(t)$$

for some functions $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ is called linear ODE of first order*.*

Here, first order means that the highest derivative of the unknown function y that appears in the equation is the first one. Linear means that all the expressions are linear in the unknown y and its derivatives.

As in the case of linear algebraic equations, the linearity of an equation has some implications on the structure of its solutions. To that end we consider the *homogeneous* equation[1]

$$\dot{y} + f(t)y = 0 \; .$$

If we are given two solutions y_1 and y_2 of this equation (with possibly different initial conditions), then

$$\begin{aligned} \dot{y}_1 + f(t)y_1 &= 0 \, , \\ \dot{y}_2 + f(t)y_2 &= 0 \; . \end{aligned}$$

Consequently also

$$\frac{d}{dt}(y_1 + y_2) + f(t)(y_1 + y_2) = \dot{y}_1 + f(t)y_1 + \dot{y}_2 + f(t)y_2 = 0 \, ,$$

[1] Homogeneous means that the right hand side of the equation is zero, that is, $g = 0$

which shows that also $y_1 + y_2$ is a solution of the ODE. More general, if y_1 and y_2 solve the ODE and c_1, $c_2 \in \mathbb{R}$, then the linear combination $c_1 y_1 + c_2 y_2$ is also a solution.

In order to solve the (inhomogeneous) equation

$$\dot{y} + f(t)y = g(t) \tag{1.11}$$

we first observe that (1.11) is equivalent to

$$h(t)\dot{y} + h(t)f(t)y = h(t)g(t)\,, \tag{1.12}$$

at least, if $h\colon \mathbb{R} \to \mathbb{R}$ is a function that is different from zero.

Now the idea is to choose the function h in such a way that the left hand side of (1.12) is itself a derivative. More precisely, we try to find $h\colon \mathbb{R} \to \mathbb{R}$ such that

$$h(t)\dot{y} + h(t)f(t)y = \frac{d}{dt}(hy) = \dot{h}y + h\dot{y}\;. \tag{1.13}$$

If (1.13) holds, then the equation (1.12) reads as follows,

$$\frac{d}{dt}(hy) = h(t)g(t)\,,$$

which after integration becomes:

$$h(t)y(t) = \int^t g(s)h(s)\,ds + C\;. \tag{1.14}$$

For this reason, a function h satisfying (1.13) is called an *integrating factor* for the ODE (1.11).

Thus, (1.13) is satisfied, if h satisfies

$$h(t)f(t)y = \dot{h}(t)y\;.$$

Dividing this equation by y, we see that h has to satisfy the ODE

$$\dot{h} = f(t)h\;.$$

This ODE can be solved by separation of the variables, and we obtain the integrating factor

$$h(t) = D\exp\Bigl(\int^t f(s)\,ds\Bigr)\;.$$

Inserting this integrating factor in (1.14), we obtain

$$y(t) = \frac{\int^t \left[g(s) D \exp\left(\int^s f(r)\, dr\right)\right] ds + C}{D \exp\left(\int^t f(s)\, ds\right)} ,$$

or, setting $\tilde{C} := C/D$,

$$y(t) = \frac{\int^t \left[g(s) \exp\left(\int^s f(r)\, dr\right)\right] ds + \tilde{C}}{\exp\left(\int^t f(s)\, ds\right)} .$$

The following example provides a relation between ordinary and partial differential equations.

Example 1.6. *Let $u(x,t)$, $-1 \leq x \leq 1$, be the temperature distribution at time t in a slab of length $l = 2$. Assuming constant conductivity $\sigma = 1$, u satisfies the* heat conduction *equation:*

$$u_t = \sigma u_{xx} = u_{xx} , \quad -1 < x < 1 , 0 < t < T . \tag{1.15}$$

This is now a partial differential equation *because it depends on derivatives of* **at least two** *variables x, t. By discretization of the x variable we can transform the partial differential equation in a system of ordinary differential equations.*

Let $v : [-1,1] \to \mathbb{R}$ be an arbitrary function satisfying $v(-1) = v(1) = 0$, then we get by integration by parts

$$\int_{-1}^{1} u_t(t,x) v(x)\, dx = \int_{-1}^{1} u_{xx}(t,x) v(x)\, dx = -\int_{-1}^{1} u_x(t,x) v_x(x)\, dx . \tag{1.16}$$

Assume that the temperatures $u(-1,t) := u_0(t)$ and $u(1,t) := u_1(t)$ are measured, then, for every $t > 0$, $u(t,x)$ can be approximated by a linear spline in space over the grid $\Delta = \{-1 = x_0 < x_1 < ... < x_\nu = 1\}$, that is

$$u(t,x) = \sum_{\hat{i}=0}^{\nu} y_{\hat{i}}(t) \Lambda_{\hat{i}}(x) , \tag{1.17}$$

where $\Lambda_{\hat{i}}$ is a linear hat function with peak at $x_{\hat{i}}$. Taking into account the boundary conditions we see that $y_0 = u_0(t)$ and $y_\nu = u_1(t)$. All other functions $y_{\hat{i}}$ are unknown.

Inserting (1.17) in (1.16) we get a system of differential equations for $y_1, ..., y_{\nu-1}$*:*

$$\sum_{\hat{i}=0}^{\nu} (y_{\hat{i}})_t(t) \int_{-1}^{1} \Lambda_{\hat{i}}(x) v(x)\, dx = - \sum_{\hat{i}=0}^{\nu} y_{\hat{i}}(t) \int_{-1}^{1} (\Lambda_{\hat{i}})_x(x) v_x(x)\, dx \ ,$$

where we choose $v(x) \in \left\{ \Lambda_{\hat{j}}(x) : \hat{j} = 1, ..., \nu - 1 \right\}$ *- this means that* v *is a hat function, which satisfies homogenous boundary conditions.*

Denote by

$$G := [\langle \Lambda_{\hat{i}}, \Lambda_{\hat{j}} \rangle]_{1 \le \hat{i}, \hat{j} \le \nu - 1} = \frac{h}{6} \begin{bmatrix} 4 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & 4 & 1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & 1 & 4 & 1 \\ \vdots & . & . & 0 & 1 & 4 \end{bmatrix}$$

and

$$A := [\langle {\Lambda_{\hat{i}}}', {\Lambda_{\hat{j}}}' \rangle]_{1 \le \hat{i}, \hat{j} \le \nu - 1} = h \begin{bmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & -1 & 2 & -1 \\ \vdots & \ddots & \ddots & 0 & -1 & 2 \end{bmatrix}$$

we get a compact description of the system

$$G y'(t) + A y(t) = b(t) \,, \tag{1.18}$$

where b *is an appropriate vector, which depends on* u_0 *and* u_1*.*

To completely specify the system (1.18) *we need initial values for* $y_1, ..., y_{\nu-1}$*, which are typically determined from interpolation of the initial temperature* $u(0, x)$*. Note however, that this is a system of equations (the solution* y *is vector valued).*

Chapter 2

Ordinary Differential Equations

In this chapter we study the numerical solution of ordinary differential equations

$$y' = f(t, y), \quad t \in [0, T] \text{ with the initial condition } y(0) = y_0 \, . \tag{2.1}$$

Here y is a vector valued function with respect to time t. We call (2.1) a *system of first order*.

We use the following convention:

y	$\in \mathbb{R}^{\nu \pm 1}$
x	$\in \mathbb{R}^{\nu + 1}$
$\hat{i}, \hat{j}$	Index for x, y
i, j	Index of iterations of numerical method
y_i, t_i	iterate y_i approximating $y(t_i)$

2.1 The Euler Method

The *(explicit) Euler-method* approximates y on a uniform grid

$$\Delta = \{0 = t_0 < t_1 < t_2 < ... < t_n\} \subseteq I$$

with the recursive formula

$$\boxed{y_{i+1} = y_i + (t_{i+1} - t_i) f(t_i, y_i) \, .}$$

The *implicit Euler-method* approximates the solution by

$$\boxed{y_{i+1} = y_i + (t_{i+1} - t_i) f(t_{i+1}, y_{i+1}) \, .} \tag{2.2}$$

Thereby in each step an equation has to be solved. This method is stable in a sense which has to be specified afterward. However, the method is rather slow.

2.2 Taylor Method

Let for the sake of simplicity of presentation $\nu = 1$. That is, we do not consider vector valued ODEs here and $y : [0, T] \to \mathbb{R}$.

If the existence of all higher order partial derivatives is assumed for y at $t = t_0$, then by Taylor series the value of y at any neighboring point $t_0 + h$ can be written as

$$y(t_0 + h) = y(t_0) + hy'(t_0) + \frac{h^2}{2!} y''(t_0) + \frac{h^3}{3!} y'''(t_0) + ...$$

Similarly higher derivatives of y at t_0 also can be computed by making use of the relation $y' = f(t, y)$:

$$\begin{aligned} y'' &= f_t + f_y y' = f_t + f_y f \,, \\ y''' &= f_{tt} + 2 f_{ty} y' + f_{yy} (y')^2 + f_y y'' \\ &= f_{tt} + 2 f_{ty} f + f_{yy} f^2 + f_y (f_t + f_y f) \,, \end{aligned}$$

and so on. Hence:

$$y(t_0{+}h) = y(t_0){+}hf{+}\frac{h^2}{2!}(f_t{+}f_y y'){+}\frac{h^3}{3!}(f_{tt}{+}2f_{ty}y'{+}f_{yy}(y')^2{+}f_y y''){+}\mathcal{O}(h^4) \,.$$

The Taylor's method then reads as

$$\begin{aligned} y_{i+1} =& y_i + hf(t_i, y_i) + \frac{h^2}{2!}(f_t(t_i, y_i) + f_y(t_i, y_i) y'(t_i, y_i)) + \\ & \frac{h^3}{3!} \Bigg(f_{tt}(t_i, y_i) + 2 f_{ty}(t_i, y_i) f(t_i, y_i) + f_{yy}(t_i, y_i) \\ & \quad + f^2(t_i, y_i) + f_y(t_i, y_i)(f_t(t_i, y_i) + f_y(t_i, y_i) f(t_i, y_i)) \Bigg) \,. \end{aligned}$$

2.3 Runge-Kutta Method

The disadvantage of both Euler-methods is the slow convergence (in dependence of the time discretization). Faster convergence can be obtained with an ansatz

$$\boxed{y_{i+1} = y_i + h\sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j)\,, \quad \sum_{j=1}^{s} b_j = 1\,,} \tag{2.3}$$

where η_j is an approximation for $y(t_i + c_j h)$.

Such methods are called *Runge-Kutta-methods* of *degree s*. In particular:

- The explicit Euler method is with $s = 1$ and $c_1 = 0$, $\eta_1 = y_i$.
- For the implicit Euler method we have $s = 1$, $c_1 = 1$, $\eta_1 = y_{i+1}$.

Because in (2.3) one calculates and approximation $y_{i+1} \approx y(t_{i+1})$ starting from $y_i \approx y(t_i)$ the method is called *single step* method. If other previous approximations $y_{i-1}, ...$ are used to determine y_{i+1}, then the method is called *multi-step* method.

2.4 Single Step Runge-Kutta Methods

We assume that $y_i = y(t_i)$, then by the fundamental theorem of differential calculus

$$\begin{aligned} y(t_{i+1}) - y_{i+1} &\underbrace{=}_{y(t_i)=y_i} y(t_{i+1}) - y(t_i) - h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \\ &= \int_{t_i}^{t_{i+1}} y'(t)\,dt - h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \\ &\underbrace{=}_{ODE} \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt \; - h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j)\;. \end{aligned}$$

We see that the *local error* gets small if

$$h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \approx \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt\;.$$

This suggest to take quadrature formulas for choosing $\{b_j\}$, $\{c_j\}$ and $\{\eta_j\}$.

Example 2.1.

- *Using the midpoint rule we get*

$$y_{i+1} = y_i + hf(t_i + \frac{h}{2}, \eta_1) \ , \tag{2.4}$$

where ideally $\eta_1 = y(t_i + \frac{h}{2})$. *Because this value of the solution* y *is not known we are looking for an approximation: The method of* Runge *(1895) used the approximation*

$$\eta_1 = y(t_i) + \frac{h}{2} y'(t_i) \approx y_i + \frac{h}{2} f(t_i, y_i) \ .$$

- *With the trapezoidal rule we find*

$$y_{i+1} = y_i + \frac{h}{2} f(t_i, y_i) + \frac{h}{2} f(t_{i+1}, \tilde{\eta}_1) \, ,$$

where $\tilde{\eta}_i \approx y(t_i + h)$. *If we proceed as in the Runge method and if we use the approximation*

$$\tilde{\eta}_1 = y_i + hy'(t_i) \, ,$$

then we get the method of Heun.

The *Runge-Kutta methods* rely on the following choice of coefficients:

$$\boxed{\eta_j \approx y(t_i + c_j h) = y(t_i) + \int_{t_i}^{t_i + c_j h} y'(t) \, dt \ = y(t_i) + \int_{t_i}^{t_i + c_j h} f(t, y(t)) \, dt \ .} \tag{2.5}$$

For the approximate evaluation there are used again quadrature formulas which, for the evaluation of $f(t, y)$, use the same nodal values $f(t_i + c_j h, \eta_j)$, $j = 1, ..., s$, as they are used for calculating y_{i+1}. Thus we make the following ansatz:

$$\eta_j = y_i + h \sum_{k=1}^{s} a_{jk} f(t_i + c_k h, \eta_k) \, , \quad \sum_{k=1}^{s} a_{jk} = c_j \ . \tag{2.6}$$

The coefficients $\{a_{jk}, b_j, c_j\}$ are summarized in a quadratic tableau (*Runge-Kutta Abc* or *Butcher-tableau*):

$$\begin{array}{c|c} \mathrm{c} & \mathrm{A} \\ \hline & \mathrm{b}^T \end{array} \quad = \quad \begin{array}{c|ccccc} c_1 & a_{1,1} & \dots & \dots & \dots & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & \dots & \dots & \dots \\ c_3 & a_{3,1} & a_{3,2} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ c_s & a_{s,1} & \dots & \dots & a_{s,s-1} & a_{s,s} \\ \hline & b_1 & b_2 & \dots & b_{s-1} & b_s \end{array}$$

where $A = [a_{j,k}] \in \mathbb{R}^{s\times s}$, $b = [b_1, ..., b_s]^T \in \mathbb{R}^s$ and $c = [c_1, ..., c_s]^T \in \mathbb{R}^s$.

Example 2.2. *For the explicit and implicit Euler method we have the following tableau, respectively:*

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

The method of Runge requires to add a trivial equation to transform it into the general scheme:

$$\begin{array}{rcl} \eta_0 & = & y_i + h\sum_{k=0}^{1} 0 \cdot f(t_i + c_k h, \eta_k) \\ \eta_1 & = & y_i + \frac{h}{2} f(t_i + \frac{h}{2}, \eta_0) \\ y_{i+1} & = & y_i + hf(t_i + \frac{h}{2}, \eta_1) \; . \end{array}$$

The tableau then reads as follows:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

2.5 Ill-Conditioned ODE

There exist ODEs for which error and noise significantly influence the solution. Such problems are called *ill–conditioned*, and cannot be cured by a numerical approach. As an illustration we consider the system

$$u_1' = 2u_2 \text{ and } u_2' = 2u_1$$

for which the general solution is

$$u_1 = ae^{2t} + be^{-2t} \text{ and } u_2 = ae^{2t} - be^{-2t} \; .$$

Taking the initial conditions

$$u_1(0) = 3 \text{ and } u_2(0) = -3$$

we have

$$a + b = 3 \text{ and } a - b = -3 \, ,$$

and therefore $a = 0$ and $b = 3$, and the solution of the system is

$$u_1 = 3e^{-2t} \text{ and } u_2 = -3e^{-2t} \; .$$

However, if we put

$$u_1(0) = 3 + \varepsilon \text{ and } u_2(0) = -3\,,$$

(assume that ε is some noise), then we have

$$a + b = 3 + \varepsilon \text{ and } a - b = -3\,,$$

which gives $a = \frac{\varepsilon}{2}$ and $b = 3 + \frac{\varepsilon}{2}$, and therefore the solution is

$$u_1 = \frac{\varepsilon}{2}e^{2t} + (3 - \varepsilon)e^{-2t} \text{ and } u_2 = \frac{\varepsilon}{2}e^{2t} - \left(3 - \frac{\varepsilon}{2}\right)e^{-2t}\,.$$

For fixed $\varepsilon > 0$ the term $\frac{\varepsilon}{2}e^{2t}$ gets dominant for large t.

Ill-conditioning can also occur for a single first-order ODE: Consider for example

$$y' = 3y - t^2$$

for which the general solution is

$$y = Ce^{3t} + \frac{t^2}{3} + \frac{2t}{9} + \frac{2}{27}\,.$$

If we take as initial condition $y(0) = \frac{2}{27} + \varepsilon$, then $C = \varepsilon$. Again, the term Ce^{3t} gets dominant for large t. Thus for every small error ε the error term will dominate the exact solution.

Ill-conditionedness is a property of the equation and cannot be cured with numerical algorithms.

2.6 Stiff ODE's

Stiffness is a phenomenon rather than a definition in a rigorous mathematical setting. The terminology *stiff* probably originates from chemical reaction problems which exhibit tight coupling of various reactions of different scales.

Since there is no rigorous mathematical definition of stiffness, we can only describe it phenomenologically.

Example 2.3. *Consider the differential equation*

$$y'(t) = -15y(t)\,, \quad t \geq 0\,, \qquad y(0) = 1\,. \tag{2.7}$$

The exact solution is

$$y(t) = e^{-15t}\,,$$

which satisfies $y(t) \to 0$ for $t \to \infty$.

Numerically we see a completely different behavior for various methods.

1. *The Euler method with a step size of $h = 1/4$ oscillates and the solution blows up very rapidly. The iterates $y_i, i = 0, \ldots, 10$,*

   $$[1, -3, 8, -21, 57, -157, 433, -1189, 3271, -8995, 24736]^T\,.$$

   While the exact solution is

   $$[1, 0.0235, 0.0006, 0, 0, 0, 0, 0, 0, 0, 0]^T\,.$$

2. *The iterates with Euler's method with step size $h = 1/8$ are bounded:*

   $$\begin{aligned}&[1, -0.8750, 0.7656, -0.6699, 0.5862, -0.5129,\\ &\quad 0.4488, -0.3927, 0.3436, -0.3007, 0.2631, -0.2302,\\ &\quad 0.2014, -0.1762, 0.1542, -0.1349, 0.1181, -0.1033,\\ &\quad 0.0904, -0.0791, 0.0692]^T\,.\end{aligned}$$

   The exact solution is

   $$[1, 0.1534, 0.0235, 0.0036, 0.0006, 0.0001, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]\,.$$

3. *The* trapezoidal method, *defined by,*

   $$\begin{aligned}y_{i+1} &= y_i + \frac{h}{2}(f(t_i, y_i) + f(t_{i+1}, y_{i+1}))\\ &= \frac{2-15h}{2+15h}y_i\end{aligned}$$

   gives with step size $h = 1/8$

   $$[1, 0.0323, 0.0010, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]\,,$$

   which is monotonically decreasing.

Example 2.4. *One of the most prominent examples of a stiff ODEs is a system that describes the chemical reaction of Robertson:*

$$\begin{aligned} y_1' &= -4.10^{-2}y_1 + 10^4 y_2 y_3 \,, \\ y_2' &= 4.10^{-2}y_1 - 10^4 y_2 y_3 - 3.10^7 y_2^2 \,, \\ y_3' &= 3.10^7 y_2^2 \,. \end{aligned} \tag{2.8}$$

On a short time interval the numerical solution of the system does not make problems, however for large t (let us say 10^{11}) it does.

2.6.1 Stiffness Ratio

Consider the linear inhomogeneous system

$$y'(t) = \mathcal{A}y(t) + f(t)\,, \tag{2.9}$$

where $y = y(t), f = f(t) \in \mathbb{R}^\nu$ and $\mathcal{A} \in \mathbb{R}^{\nu\times\nu}$ is symmetric with eigenvalues $\lambda_{\hat{i}} \in \mathbb{C}$ and eigenvectors $y_{\hat{i}}$, $\hat{i} = 1, \ldots, \nu$. We assume that the matrix $\mathcal{A}$ can be diagonalized: That is, there exists a matrix $Y = (y_1, \ldots, y_\nu)$, consisting of the columns of $y_{\hat{i}}$, such that

$$\mathcal{A} = Y\Lambda Y^{-1}\,,$$

where Λ is the diagonal-matrix consisting of eigenvalues of $\mathcal{A}$, and $y_{\hat{i}}$, $\hat{i} = 1, \ldots, \nu$ forms an orthonormal basis of $\mathbb{R}^\nu$ [2]. So, if $\mathcal{A}$ can be diagonalized, then with

$$\boxed{z(t) = Y^{-1}y(t) \text{ and } g(t) = Y^{-1}f(t)} \tag{2.10}$$

we get

$$z'(t) = Y^{-1}y'(t) = \Lambda Y^{-1}y(t) + Y^{-1}f(t) = \Lambda z(t) + g(t)\,, \tag{2.11}$$

or in other words

$$z_{\hat{i}}'(t) = \lambda_{\hat{i}} z_{\hat{i}}(t) + g_{\hat{i}}(t)\,. \tag{2.12}$$

The solution of this system is determined by the method of *variations of constants*: This procedure makes use of the ansatz

$$\boxed{z_{\hat{i}}(t) = c_{\hat{i}}(t)e^{\lambda_{\hat{i}}t}\,.} \tag{2.13}$$

Then

$$z_{\hat{i}}'(t) = c_{\hat{i}}'(t)e^{\lambda_{\hat{i}}t} + c_{\hat{i}}(t)\lambda_{\hat{i}}e^{\lambda_{\hat{i}}t}\,.$$

To satisfy the differential equation (2.12) we have to satisfy

$$c'_{\hat{i}}(t)e^{\lambda_{\hat{i}}t} + \underline{c_{\hat{i}}(t)\lambda_{\hat{i}}e^{\lambda_{\hat{i}}t}} = \underline{c_{\hat{i}}(t)\lambda_{\hat{i}}e^{\lambda_{\hat{i}}t}} + g_{\hat{i}}(t)\,,$$

or in other words

$$c'_{\hat{i}}(t) = e^{-\lambda_{\hat{i}}t}g_{\hat{i}}(t)\ .$$

Thus we get

$$c_{\hat{i}}(t) - c^{(0)}_{\hat{i}} = \int_0^t c'_{\hat{i}}(\tau)\,d\tau = \int_0^t g_{\hat{i}}(\tau)e^{-\lambda_{\hat{i}}\tau}\,d\tau\ .$$

And thus

$$\begin{aligned} z_{\hat{i}}(t) &= c_{\hat{i}}(t)e^{\lambda_{\hat{i}}t} \\ &= \left(\int_0^t g_{\hat{i}}(\tau)e^{-\lambda_{\hat{i}}\tau}\,d\tau + c^{(0)}_{\hat{i}}\right)e^{\lambda_{\hat{i}}t} \\ &= \int_0^t g_{\hat{i}}(\tau)e^{\lambda_{\hat{i}}(t-\tau)}\,d\tau + c^{(0)}e^{\lambda_{\hat{i}}t}\,, \end{aligned}$$

or in compact vector notation

$$z(t) = \int_0^t g(\tau)e^{\Lambda(t-\tau)}\,d\tau + c^{(0)}_{\hat{i}}e^{\Lambda t}\,,$$

where here $e^{\Lambda(t-\tau)} = [e^{\lambda_{\hat{i}}(t-\tau)}]_{1\le\hat{i}\le\nu}$.

Thus, in total we have:

$$\boxed{y(t) = Yz(t) = \int_0^t e^{\Lambda(t-\tau)}f(\tau)\,d\tau + e^{\Lambda t}Yc^{(0)}\ .} \tag{2.14}$$

Let us assume that

$$\Re(\lambda_{\hat{i}}) < 0\,,\quad \forall\hat{i} = 1,2,\ldots,\nu\,. \tag{2.15}$$

Let $\overline{\lambda},\underline{\lambda} \in \{\lambda_{\hat{i}}, i = 1,2,\ldots,n\}$ be the maximal absolute eigenvalues:

$$-\Re\overline{\lambda} = \left|\Re(\overline{\lambda})\right| \ge |\Re(\lambda_{\hat{i}})| \ge |\Re(\underline{\lambda})| = -\Re(\underline{\lambda}), \qquad i = 1,2,\ldots,n\ .$$

We now define the *stiffness ratio* as $\frac{\Re(\overline{\lambda})}{\Re(\underline{\lambda})}$. The crux with the siffness ratio is that it is severely affected by the smallest negative eigenvalue (equivalently the one with highest absolute value). This one however, is the best behaving analytically. Interestingly, it affects the numerics.

Remark 2.5. *The solution of the homogenous equation according to* (2.12) *(that is with $g_{\hat{\imath}} \equiv 0$) is given by*

$$z_{\hat{\imath}}(t) = ce^{\lambda_{\hat{\imath}} t} \, .$$

The name variations of constant for the ansatz (2.13) *is due to the fact that the constant c of the solution of the homogenuous system is replaced by the function $c_{\hat{\imath}}(t)$. That means that the constant is replaced by a function, that is it is varied now.*

Example 2.6. *Also the Example 1.6 results into a system of stiff ODEs' if n is large.*

2.6.2 A-Stability

The behavior of numerical methods on stiff problems can be analyzed by applying these methods to the test equation

$$y'(t) = \lambda y(t) \text{ with } y(0) = 1 \tag{2.16}$$

for some $\lambda \in \mathbb{C}^- := \{\lambda \in \mathbb{C} : \Re(\lambda) < 0\}$. The solution of this equation is $y(t) = e^{\lambda t}$. This solution is monotonically decreasing and approaches zero for $t \to \infty$ when $\Re(\lambda) < 0$.

Definition 2.7. *If the numerical method also exhibits the monotonicity behavior, then the method is said to be A-stable.*

A-Stability is a property of the numerical method and not of the equation. The test-equation (2.16) is behaving completely stable, but the outcome of the algorithm might not.

Now, we return to Runge-Kutta methods and define the *stability function*:

Definition 2.8.

$$R : \mathbb{C} \backslash \left\{ \frac{1}{\sigma} : 0 \neq \sigma \in \sigma(A) \right\} \to \mathbb{C} \, ,$$

$$\zeta \to 1 + \zeta b^T (\mathbb{1} - \zeta A)^{-1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

The stability domain of a Runge-Kutta method is defined by

$$\mathcal{S} := \{\zeta : |R(\zeta)| \leq 1\} \, .$$

We mention that for the test-equation (2.16)

- $y_i = (R(h\lambda))^i$, which explodes if $R(h\lambda) > 1$. Thus one needs to choose the step-size h in such a way that $h\lambda \in \mathcal{S}$.
- Note also, that the stability function and stability domain only depend on A and b, but **not** on c.

Theorem 2.9. *A Runge-Kutta method is A-stable if for given*

$$\zeta \in \mathbb{C}^- := \left\{\zeta = \zeta^r + \mathrm{i}\zeta^i \in \mathbb{C} : \zeta^r \leq 0\right\}$$

we have $|R(\zeta)| \leq 1$.

Example 2.10. *For the explicit Euler method we have*

$$R(\zeta) = 1 + \zeta\,,$$

and thus it is not A-stable.

Because the set of ζ *where*

$$|R(\zeta)|^2 = |1 + \zeta|^2 = (1 + \zeta^r)^2 + (\zeta^i)^2 \leq 1$$

is a circle with center $(-1, 0)$ *and radius* 1 *in the complex plane, the explicit Euler-method is only A-stable if the step-size* h *is chosen such that*

$$\boxed{|h\lambda - (-1,0)| \leq 1\,.}$$

- *For Example 2.3 we can guarantee A-stability if* $h \leq 1/15$*, which supports the numerical results of Example 2.3.*
- *If we are considering again a system of ODEs* (2.9)*, where the matrix* $\mathcal{A}$ *has eigenvalues*

  $$\left\{\lambda_i = \underbrace{\lambda_i^r}_{\leq 0} + \mathrm{i}\lambda_i^i : i = 1, \ldots, \nu\right\}\,,$$

  then for A-stability it is required that

  $$(h\lambda_i^r + 1)^2 + (h\lambda_i^i)^2 \leq 1\,.$$

Example 2.11. *The implicit Euler method is A-stable, because here* $R(\zeta) = (1-\zeta)^{-1}$ *and*

$$|1-\zeta|^2 = (1-\zeta^r)^2 + \zeta^{i2} = 1 - 2\zeta^r + |\zeta|^2 \geq 1 \text{ for } \zeta^r \leq 0 \; .$$

The step size is **not** *essential for A-stability.*

Example 2.12. *The tableau of the mid-point rule is defined by*

$$\begin{array}{c|c} c = 1/2 & A = 1/2 \\ \hline & b^T = 1 \end{array}$$

and the Runge-Kutta method has the form

$$y_{i+1} = y_i + hf(t_i + \frac{h}{2}, \eta_1) \, , \quad \eta_1 = y_i + \frac{h}{2} f(t_i + \frac{h}{2}, \eta_1) \; . \tag{2.17}$$

By combination of the two equations we get:

$$y_{i+1} = y_i + hf(t_i + \frac{h}{2}, (y_i + y_{i+1})/2) \; .$$

According to the definition of the stability function we have:

$$\begin{aligned} R(\zeta) &= 1 + \zeta b(1 - \zeta A)^{-1} \\ &= 1 + \zeta \left(1 - \frac{\zeta}{2}\right)^{-1} \\ &= \frac{1 + \frac{\zeta}{2}}{1 - \frac{\zeta}{2}} \\ &= 1 + \zeta + \zeta^2/2 + \zeta^3/4 + \ldots \, , \end{aligned}$$

which is a Möbius-function. This function satisfies

$$|R(\zeta)|^2 = \frac{\left(1 + \frac{\zeta^r}{2}\right)^2 + \left(\frac{\zeta^i}{2}\right)^2}{\left(1 - \frac{\zeta^r}{2}\right)^2 + \left(\frac{\zeta^i}{2}\right)^2} \leq 1 \, , \qquad \forall \zeta = \zeta^r + \mathrm{i}\zeta^i \in \mathbb{C}^- \; .$$

That is, the implicit midpoint rule is A-stable. Every choice of the step size is feasible.

2.7 Multi-Step Methods

The basic idea consists in approximating the integrand on the right hand side of

$$y(t_{i+l}) = y(t_{i-k}) + \int_{t_{i-k}}^{t_{i+l}} y'(\tau)\, d\tau = y(t_{i-k}) + \int_{t_{i-k}}^{t_{i+l}} f(\tau, y(\tau))\, d\tau$$

over an intervall $[t_{i-k}, t_{i+l}]$. Given some $s \in \mathbb{N}$ let

$$(t_j, f_j) := (t_j, f(t_j, y_j)) \text{ for } j = i-s, i-s+1, \ldots, i\,,$$

where $y_j \approx y(t_j)$, which we assume to be calculated already. The polynomial of degree s interpolating these values is given by

$$P_s(\tau) = \sum_{j=i-s}^{i} f_j L_j(\tau) \text{ with } L_j(\tau) = \prod_{j \neq \hat{j} = i-s}^{i} \frac{\tau - t_{\hat{j}}}{t_j - t_{\hat{j}}}\ .$$

The functions L_j are the basic Lagrange polynomials.

The s-th order multi-step method is defined by

$$y_{i+l} = y_{i-k} + \int_{t_{i-k}}^{t_{i+l}} P_s(\tau)\, d\tau\ .$$

The different methods depend on the choice of s, k and l.

- The s-th order *Adams-Bashford* methods is explicit and $l = 1$ and $k = 0$.
- The s-th order *Adams-Moulton* method is implicit and $l = 0$ and $k = 1$.

We are only studying the first five members of the Adams-Bashford for constant step-size:

Order s	Interpolant	Interpolation points
0	constant	(t_i, f_i)
1	linear	$(t_i, f_i), (t_{i-1}, f_{i-1})$
2	quadratic	$(t_i, f_i), (t_{i-1}, f_{i-1}), (t_{i-2}, f_{i-2})$
2	cubic	$(t_i, f_i), (t_{i-1}, f_{i-1}), (t_{i-2}, f_{i-2}), (t_{i-3}, f_{i-3})$
4	quartic	$(t_i, f_i), (t_{i-1}, f_{i-1}), (t_{i-2}, f_{i-2}), (t_{i-3}, f_{i-3}), (t_{i-4}, f_{i-4})$

- If $s = 0$, $l = 1$ and $k = 0$ then the Adams-Bashfort method satisfies $P_0 = f(t_i, y_i)$ and thus

  $$y_{i+1} = y_i + h_i f(t_i, y_i) \text{ with } h_i = t_i - t_{i-1}$$

  is exactly the Euler method.

- If $s = 1$, $l = 1$ and $k = 0$ we have

  $$P_1(\tau) = f_{i-1} + \frac{f_i - f_{i-1}}{h_{i-1}}(\tau - t_{i-1}) \ .$$

  Thus we obtain

  $$\begin{aligned}\int_{t_i}^{t_{i+1}} f(\tau, y(\tau))\, d\tau &\approx \int_{t_i}^{t_{i+1}} \left(f_{i-1} + \frac{f_i - f_{i-1}}{h_{i-1}}(\tau - t_{i-1}) \right) d\tau \\ &= \frac{h_i}{2} \left(\frac{h_i + 2h_{i-1}}{h_{i-1}} f_i - \frac{h_i}{h_{i-1}} f_{i-1} \right) \ .\end{aligned}$$

  In particular if $h = h_i = h_{i-1}$, then

  $$y_{i+1} = y_i + \frac{h}{2}(3f_i - f_{i-1}) \ .$$

The derivation of higher order methods is analogous. Here only the results for constant step-size are summarized:

Order s	Adam-Bashfort
0	$y_{i+1} = y_i + hf_i$
1	$y_{i+1} = y_i + \frac{h}{2}(3f_i - f_{i-1})$
2	$y_{i+1} = y_i + \frac{h}{12}(23f_i - 16f_{i-1} + 5f_{i-2})$
3	$y_{i+1} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3})$
4	$y_{i+1} = y_i + \frac{h}{720}(1901f_i - 2774f_{i-1} + 2616f_{i-2} - 1274f_{i-3} + 251f_{i-4})$

2.8 Step-Size Control for Runge-Kutta Methods

For the implementation of a single-step Runge-Kutta method it requires to choose an optimal step-length h. The larger h can be chosen, the less computational effort is required to calculate y at a given time T.

Runge-Kutta methods of order q with a step-size h have a local error (error in each iteration step) $\mathcal{O}(h^{q+1})$ and a global error $\mathcal{O}(h^q)$.

A common way of step-length control is by using a *control method*, which is a method of higher order than q and the same stability properties. Let $\hat{y}_i$ denote the iterates of the control method. Because it is of higher order we can expect that

$$\|\hat{y}_i - y(t_i)\|_2 << \|y_i - y(t_i)\|_2 \ ,$$

such that

$$\delta_i := \|y_i - \hat{y}_i\|_2 \approx \|y_i - y(t_i)\|_2 \ .$$

This means that δ_i provides a quantitative figure for $\mathcal{O}(h^{q+1})$.

Assuming that the Runge-Kutta method is of order q and the control method is of order $q+1$, the modification of the step-size is determined based on the error estimates:

$$y_{i+1} - y(t_{i+1}) = h^{q+1}\omega_i + \mathcal{O}(h^{q+2}) \text{ and } \hat{y}_{i+1} - y(t_{i+1}) = \mathcal{O}(h^{q+2}) \, ,$$

with an, in general, unknown ω_i. Thus we have

$$\delta_{i+1} = \delta_{i+1}(h) = \|y_{i+1} - \hat{y}_{i+1}\|_2 = \left\|h^{q+1}\omega_i + \mathcal{O}(h^{q+2})\right\|_2 \approx \left\|h^{q+1}\omega_i\right\|_2 \ .$$

With a different step-size $\tilde{h}$ we get

$$\delta_{i+1}(\tilde{h}) \approx \left\|\tilde{h}^{q+1}\omega_i\right\|_2 = \left(\frac{\tilde{h}}{h}\right)^{q+1} h^{q+1}\|\omega_i\|_2 = \left(\frac{\tilde{h}}{h}\right)^{q+1} \delta_{i+1}(h) \ .$$

As a consequence

$$\tilde{h} = \tau \left(\frac{\varepsilon}{\delta_i(h)}\right)^{1/(q+1)} h \tag{2.18}$$

Guozhi: *maybe*

$$\tilde{h} = \tau \left(\frac{\varepsilon}{\delta_{i+1}(h)}\right)^{1/(q+1)} h \tag{2.19}$$

is the largest step-size (typically used in the next iteration based on the step-size h of the current iteration) such that, with $\tau = 1$, $\delta_{i+1}(\tilde{h}) \leq \varepsilon$. In practice one chooses a relaxation parameter $\tau \approx 0.8$ to maintain a trade-off between accuracy and computational effort.

In practical realization one uses a control Runge-Kutta with the same slope $f(t_i + c_j h, \eta_j)$. Therefore, the Runge-Kutta method and its control method read as follows:

$$\boxed{\begin{aligned} y_{i+1} &= y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j)\,, \\ \hat{y}_{i+1} &= y_i + h \sum_{j=1}^{s} \hat{b}_j f(t_i + c_j h, \eta_j)\,. \end{aligned}} \tag{2.20}$$

where

$$\eta_j = y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu)\,, \quad j = 1, \dots, s\,.$$

Most Runge-Kutta methods are constructed in such a ways that they are optimal with respect to convergence order. This means that there is no freedom to construct the control method with the same slopes $\{\eta_j\}$. The Fehlberg-trick deals with this problem and is explained for an example.

Example 2.13. *We choose as the control-method the classical Runge-Kutta method with $s = q = 4$, which has tableau*

$$\begin{array}{c|c} c & A \\ \hline & \hat{b}^T \end{array} \quad = \quad \begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

In concrete terms the classical Runge-Kutta method reads as follows:

$$\begin{aligned}
\eta_1 &= y_i\,,\\
\eta_2 &= y_i + \frac{h}{2} f(t_i, \eta_1)\\
&= y_i + \frac{h}{2} f(t_i, y_1)\,,\\
\eta_3 &= y_i + \frac{h}{2} f\left(t_i + \frac{h}{2}, \eta_2\right)\\
&= y_i + \frac{h}{2} f\left(t_i + \frac{h}{2}, y_i + \frac{h}{2} f(t_i, y_1)\right)\,,\\
\eta_4 &= y_i + hf\left(t_i + \frac{h}{2}, \eta_3\right)\\
&= y_i + hf\left(t_i + \frac{h}{2}, y_i + \frac{h}{2} f\left(t_i + \frac{h}{2}, y_i + \frac{h}{2} f(t_i, y_1)\right)\right)\,,
\end{aligned}$$

and

$$y_{i+1} = y_i + h\left(\frac{1}{6} f(t_i, \eta_1) + \frac{1}{3} f\left(t_i + \frac{h}{2}, \eta_2\right) + \frac{1}{3} f\left(t_i + \frac{h}{2}, \eta_3\right) + \frac{1}{6} f\left(t_i + h, \eta_4\right)\right)\,.$$

Now, we are looking for an embedded Runge-Kutta method with weights b_j, $j = 1, \ldots, 4$, which satisfy

$$\sum_{j=1}^{4} b_j = 1\,, \quad \sum_{j=1}^{4} b_j c_j = \frac{1}{2}\,, \quad \sum_{j=1}^{4} b_j c_j^2 = \frac{1}{3} \text{ and } \sum_{j=1}^{4} b_j \sum_{\nu=1}^{4} a_{j\nu} c_\nu = \frac{1}{6}\,.$$

This conditions actually guarantee that the method is of third order.
Guozhi: why it is third order? I still can not understand, maybe a reference?

Inserting the coefficients from the control-method (because the slopes should be identical) we get the equation

$$\begin{bmatrix} 1 & 1 & 1 & 1\\ 0 & 1/2 & 1/2 & 1\\ 0 & 1/4 & 1/4 & 1\\ 0 & 0 & 1/4 & 1/2 \end{bmatrix} \begin{bmatrix} b_1\\ b_2\\ b_3\\ b_4 \end{bmatrix} = \begin{bmatrix} 1\\ 1/2\\ 1/3\\ 1/6 \end{bmatrix}\,.$$

This matrix is regular and thus $b = \hat{b}$. That means that there does not exist any Runge-Kutta methods of order 3 with the same slopes.

The Fehlberg-trick adds a fifth column to the Runge-Kutta tableau: Let

$$\hat{b}_5 = 0 \text{ and } \eta_5 = \hat{y}_{i+1} = y_i + h \sum_{j=1}^{4} \hat{b}_j f(t_i + c_j h, \eta_j) \ .$$

The order conditions result in the equation

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1/2 & 1/2 & 1 & 1 \\ 0 & 1/4 & 1/4 & 1 & 1 \\ 0 & 0 & 1/4 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/6 \end{bmatrix} .$$

The rank of the matrix is again 4 *and the null-space consists of vectors spanned by* $[0, 0, 0, 1, -1]^t$*. A possible parameter choice for for* (2.20) *is therefore*

$$b = [1/6, 1/3, 1/3, 0, 1/6]^t \text{ and } \hat{b} = [1/6, 1/3, 1/3, 1/6, 0]^t \ .$$

It can actually be shown that the embedded method is in fact of order $q = 3$*.* ***Guozhi: again, why it is order 3 here?***

Chapter 3

Boundary Value Problems

For motivating purposes we study first boundary value problems for ordinary differential equations at hand of a simple test example:

$$\begin{aligned} L[u] = -u'' + bu' + cu = f \text{ in } (0,1)\,, \\ u(0) = u(1) = 0\;. \end{aligned} \tag{3.1}$$

b, c, and f can be functions on $(0,1)$.

It can be shown that this differential equation has a unique solution provided that

$$c(x) \geq 0\,, \qquad \forall x \in (0,1)\;.$$

This will always be assumed in the following.

For the simplicity of presentation we consider an equidistant grid

$$\Delta_h = \{x_i = ih : i = 1, \ldots, n-1, h = 1/n\} \subseteq (0,1)\;. \tag{3.2}$$

We denote by

$$\vec{u} = (u(x_1), \ldots, u(x_{n-1}))^t \in \mathbb{R}^{n-1} \tag{3.3}$$

the vector of the exact solution u of (3.1) on the grid Δ_h (3.2). In addition, we assume Dirichlet boundary conditions

$$0 = u(x_0) = u(x_n) = 0\;.$$

For the numerical solution we look for an approximating vector

$$\vec{u}_h = (u_1, \ldots, u_{n-1})^t \in \mathbb{R}^{n-1}\;. \tag{3.4}$$

For this purpose we discretize L from (3.1) by approximating the derivatives of u at the positions $x = x_i$ via *difference quotients*. Thereby we have several alternatives:

- *One-sided forward-difference operator:*

$$D_h^+[u](x) = \frac{u(x+h) - u(x)}{h} \sim u'(x) \; .$$

- *One-sided backward-difference operator:*

$$D_h^-[u](x) = \frac{u(x) - u(x-h)}{h} \sim u'(x) \; .$$

- *Central difference quotient:*

$$D_h[u](x) = \frac{u(x+h) - u(x-h)}{2h} \sim u'(x) \; . \tag{3.5}$$

Moreover, the second derivative can be approximated by a central difference quotient

$$D_h^2[u](x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \sim u''(x) \; . \tag{3.6}$$

Example 3.1. *We study a simple situation of* (3.1) *with* $b, c \equiv 0$*, that is* $-u'' = f$*. We approximate* u'' *by* $D_h^2[u]$ *at the nodal points* Δ_h*. Taking into account the Dirichlet boundary conditions* $u(x_0) = u(x_n) = 0$ *we get the discretized equation:*

$$\underbrace{\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \end{bmatrix}}_{=:\vec{f}} = - \begin{bmatrix} u''(x_1) \\ u''(x_2) \\ \vdots \\ u''(x_{n-1}) \end{bmatrix} \sim \underbrace{h^{-2} \begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}}_{=:L_h} \underbrace{\begin{bmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_{n-1}) \end{bmatrix}}_{\vec{u}} .$$

Because $\vec{u}$ *should be approximated by* $\vec{u}_h$*, we will use the following linear equation to determine* $\vec{u}_h$*:*

$$L_h \vec{u}_h = \vec{f} \; . \tag{3.7}$$

The Eigenvalues of L_h *are* $4h^{-2}\sin^2(kh\pi/2)4$*,* $k = 1, \dots, n-1$*. The function sinc*$(x) := \frac{\sin(x)}{x}$ *is monotonically decreasing in* $[0, \pi/2]$ *such that*

$$sinc(x) \geq sinc\left(\frac{\pi}{2}\right) = \frac{2}{\pi} \, , \quad \forall x \in [0, \pi/2] \, ,$$

which implies that:

$$\left\|L_h^{-1}\right\|_2 = \frac{1}{\lambda_{\min}(L_h)} = \max_{1\leq k\leq n-1} \frac{h^2}{4\sin^2(kh\pi/2)} \leq \frac{1}{4} \,.$$

Consequently,

$$\begin{aligned} \|\vec{u}-\vec{u}_h\|_2 &= \left\|L_h^{-1}(L_h\vec{u}-\vec{f})\right\|_2 \\ &\leq \left\|L_h^{-1}\right\|_2 \left\|L_h\vec{u}-\vec{f}\right\|_2 \\ &\leq \frac{1}{4}\left\|L_h\vec{u}-\vec{f}\right\|_2 \,. \end{aligned} \tag{3.8}$$

If $\left\|L_h\vec{u}-\vec{f}\right\|_2$ *converges to* 0 *for* $h\to 0$*, then* L_h *is called* consistent*. If there exists an estimate of the form* (3.8)*, then* consistency *implies* stability*.*

In the following we determine error estimates for difference quotients:

Lemma 3.2. *Let* $u \in C^2[0,1]$ *and* $x \in [h, 1-h]$*. Then, for one-sided difference quotients we have the estimate*

$$\left|D_h^{\pm}[u](x) - u'(x)\right| \leq \frac{1}{2}\left\|u''\right\|_\infty h \,.$$

For a central difference quotient and $u \in C^3[0,1]$ *we even have:*

$$|D_h[u](x) - u'(x)| \leq \frac{1}{6}\left\|u'''\right\|_\infty h^2 \,.$$

For D_h^2 *we have: Let* $u \in C^4[0,1]$ *and* $x \in [h,1-h]$*, then:*

$$\left|D_h^2[u](x) - u''(x)\right| \leq \frac{1}{12}\left\|u''''\right\|_\infty h^2 \,. \tag{3.9}$$

Proof. We prove exemplary the assertion for the central difference quotient. Let $u \in C^3[0,1]$, then it follows from Taylor expansion around $x \in (0,1)$:

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{1}{2}h^2u''(x) + \frac{1}{6}h^3u'''(\zeta_+)\,, \\ u(x-h) &= u(x) - hu'(x) + \frac{1}{2}h^2u''(x) - \frac{1}{6}h^3u'''(\zeta_-)\,, \end{aligned}$$

for some $\zeta_\pm$ satisfying $x-h<\zeta_-<x<\zeta_+<x+h$. Therefore

$$u(x+h) - u(x-h) = 2hu'(x) + \frac{1}{6}h^3(u'''(\zeta_+) + u'''(\zeta_-))\,,$$

and thus

$$\left|\frac{u(x+h)-u(x-h)}{2h}-u'(x)\right| \leq \frac{1}{6}h^2 \sup\left\{|u'''(\zeta)| : \zeta \in [0,1]\right\} ,$$

which gives the assertion. □

Example 3.3. *Applied to the differential equation* (3.1) *we find that, provided the solution of the differential equation is 4× continuously differentiable, that*

$$\left\|\underbrace{L_h}_{=-D_h^2}\vec{u}-\vec{f}\right\|_\infty \leq \frac{1}{12}\left\|u''''\right\|_\infty h^2 = \frac{1}{12}\left\|f''\right\|_\infty h^2 .$$

In the following we discretize the operator L defined in (3.1)*. We use the discretization $D_h^2[u]$ for approximating u''. Moreover, the first derivative is approximated by either one of the difference quotients $D_h^+[u]$, $D_h^-[u]$, $D_h[u]$. Using different difference quotients gives different diagonal matrices:*

$$L_h = h^{-2}\begin{bmatrix} d_1 & s_1 & & 0 \\ r_2 & d_2 & \ddots & \\ & \ddots & \ddots & s_{n-2} \\ 0 & & r_{n-1} & d_{n-1} \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)} , \qquad (3.10)$$

where for

- D_h^+*:*

$$\begin{aligned} d_i &= 2 - hb(x_i) + h^2c(x_i) , \\ r_i &= -1 , \\ s_i &= -1 + hb(x_i) , \end{aligned} \qquad (3.11)$$

- D_h^-*:*

$$\begin{aligned} d_i &= 2 + hb(x_i) + h^2c(x_i) , \\ r_i &= -1 - hb(x_i) , \\ s_i &= -1 , \end{aligned} \qquad (3.12)$$

- D_h*:*

$$\begin{aligned} d_i &= 2 + h^2c(x_i) , \\ r_i &= -1 - hb(x_i)/2 , \\ s_i &= -1 + hb(x_i)/2 . \end{aligned} \qquad (3.13)$$

The approximate solution is determined as the solution of the linear system (3.7)*.*

Definition 3.4. *A difference method has* order of consistence q *if*

$$\left\|L_h\vec{u}-\vec{f}\right\|_\infty = \max|(L_h\vec{u})_i - f_i| \leq Ch^q \, .$$

Note, that in this definition $\vec{u}$ is the vector of the solution of the infinite dimensional problem at the nodal points.

Theorem 3.5. *Let the solution of the boundary value problem* (3.1) *be* 4× *continuously differentiable (which is for instance the case if b, c, f are* 2× *continuously differentiable). Then the difference method* (3.7) *has the order of consistency q:*

- *$q = 2$, if the central difference quotient D_h is used for approximating u';*
- *$q = 1$, if forward or backward difference quotients $D_h^\pm$ are used for approximating u'.*

3.1 Singularly Perturbed Problems

We start with an example:

Example 3.6. *Let $\varepsilon > 0$ be small. We investigate the solution of*

$$-\varepsilon u'' + u' = 1 \text{ in } (0,1)\,, \qquad u(0) = u(1) = 0\, . \tag{3.14}$$

The exact solution is

$$u_\varepsilon(x) = x - v_\varepsilon(x)\,, \qquad v_\varepsilon(x) = \frac{e^{x/\varepsilon}-1}{e^{1/\varepsilon}-1}\, . \tag{3.15}$$

Using the difference method with central difference quotient for u' results in a linear equation

$$L_h\vec{u}_h = \frac{\varepsilon}{h^2}\begin{bmatrix} 2 & -1+\frac{h}{2\varepsilon} & & 0 \\ -1-\frac{h}{2\varepsilon} & 2 & \ddots & \\ & \ddots & \ddots & -1+\frac{h}{2\varepsilon} \\ 0 & & -1-\frac{h}{2\varepsilon} & 2 \end{bmatrix} = \vec{f}\, .$$

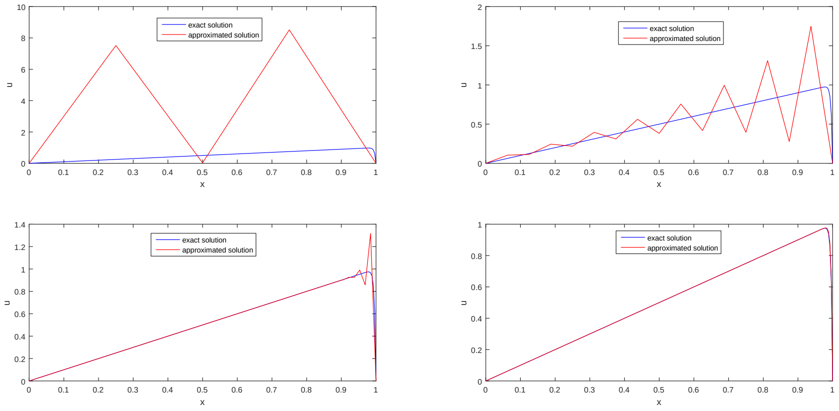


Figure 3.1: Plot of the exact solution and numerical approximation with different step size. We choose $\varepsilon = 1/2^8$, and the step sizes are $1/2^2, 1/2^4, /1/2^6, 1/2^8$ from left to right and from up to down respectively.

For $h > 2\varepsilon$ there occur significant oscillations. Intuitively this is the result of overdetermined boundaries. If $\varepsilon = 0$, then it is a ordinary differential equation, which is fully determined by one initial condition. The solution of

$$u' = 1 \textit{ in } (0,1)\,, \qquad u(0) = 0 \tag{3.16}$$

is given by $u(x) = x$.

Now, we consider the more general equations

$$L[u] = -\varepsilon u'' + bu' + cu = f \text{ in } (0,1)\,, \qquad u(0) = u(1) = 0\,. \tag{3.17}$$

To avoid oscillations one can uses *up-wind* scheme, where one uses a forward difference scheme if $b(x_i) > 0$ and a backward scheme if $b(x_i) < 0$. This results in the system

$$L_h \vec{u}_h = \frac{1}{h^2}\begin{bmatrix} d_1 & s_1 & \cdots & 0 \\ r_2 & d_2 & \ddots & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & \cdots & r_n & d_n \end{bmatrix} = \vec{f}\,.$$

with

$$\begin{aligned} d_i &= 2\varepsilon + h\,|b(x_i)| + h^2 c(x_i)\,, \\ r_i &= -\varepsilon - h b_i^+(x_i)\,, \\ s_i &= -\varepsilon + h b_i^-(x_i)\ . \end{aligned}$$

3.2 Shooting Methods

We are concerned with the boundary value problem

$$u'' = f(x,u,u') \text{ in } (0,1)\,, \quad u(0) = u(1) = 0\ . \tag{3.18}$$

Let us assume that we know $\alpha = u'(0)$, then the solution of (3.18) is also a solution of the initial value problem

$$v'' = f(x,v,v') \text{ in } (0,1)\,, \quad v(0) = 0 \text{ and } v'(0) = \alpha\ . \tag{3.19}$$

Shooting methods are rather intuitive: There is selected an initial guess α and (3.19) is solved. Then α is optimized as long as $v(1) \approx u(1) = 0$.

This problem is then formulated as a nonlinear equation

$$F(\alpha) = 0\,, \tag{3.20}$$

where $F : \mathbb{R} \to \mathbb{R}$ with $F(\alpha) = v(1)$. The solution of (3.19) then solves (3.18).

The nonlinear equation (3.20) can be determined with a gradient of Newton's method for instance. For this purpose we need the derivative of F.

Lemma 3.7. *$F'(\alpha) = w_\alpha(1)$, where w_α solves the initial value problem*

$$\begin{aligned} w'' &= f_u(x, v_\alpha, v_\alpha')w + f_{u'}(x, v_\alpha, v_\alpha')w'\,, \\ w(0) &= 0\,, \qquad w'(0) = 1\ . \end{aligned} \tag{3.21}$$

Note that this linear equation can be solved with Runge-Kutta method.

Example 3.8. *We study the example*

$$u'' + uu' = -1\,, \qquad u(0) = u(1) = 0\ .$$

Newton's method reads as follows:

$$\alpha_{k+1} = \alpha_k - F(\alpha_k)/F'(\alpha_k)\,, \quad k = 0, 1, \dots\ .$$

In every iteration step we have to evaluate the initial value problems (3.19) *and* (3.21)*:*

$$\begin{aligned} v'' + vv' &= -1\,, \qquad & v(0) = 0\,, v'(0) = \alpha_k\,, \\ w'' + vw' + v'w &= 0\,, \qquad & w(0) = 0\,, w'(0) = 1\,, \end{aligned} \tag{3.22}$$

which gives

$$F(\alpha_k) = v(1) \text{ and } F'(\alpha_k) = w(1)\ .$$

The two differential equations are coupled and therefore we are required to solve a system of four differential equations of first order for four functions $y_1 = v$*,* $y_2 = v'$*,* $y_3 = w$ *and* $y_4 = w'$*:*

$$\begin{aligned} y_1' &= y_2\,, & y_1(0) &= 0\,, \\ y_2' &= -1 - y_1 y_2\,, & y_2(0) &= \alpha_k\,, \\ y_3' &= y_4\,, & y_3(0) &= 0\,, \\ y_4' &= -y_1 y_4 - y_2 y_3\,, & y_4(0) &= 1\ . \end{aligned} \tag{3.23}$$

Chapter 4

Interpolation

We study the problem of interpolation of function samples $y_0 = y(x_0), \ldots, y_l = y(x_l)$ from a function $y : [a, b] \to \mathbb{R}$ on the grid

$$\Delta = \{a = x_0 < x_1 < \ldots < x_l = b\} \ . \tag{4.1}$$

The *grid size* is defined by

$$h := \max_{i=1,\ldots,l} h_i \, , \quad h_i = x_i - x_{i-1} \ .$$

Notation: m is the degree of the polynomial, $l + 1$ is the number of interpolation points.

4.1 Lagrange Interpolation

Historically, the first interpolation methods are based on polynomials:

Definition 4.1. Π_m *denotes the space of polynomials of degree* $\leq m$*.*

Polynomial interpolation consists in determining a polynomial $p \in \Pi_m$ such that

$$p(x_i) = y_i \, , \quad i = 0, \ldots, m \ . \tag{4.2}$$

Definition 4.2. *We denote by*

$$w(x) := \prod_{i=0}^{m} (x - x_i) \in \Pi_{m+1}$$

the nodal *polynomial at* Δ*. The polynomial*

$$\boxed{l_i(x) := \frac{w(x)}{(x-x_i)w'(x_i)} = \prod_{j=0,j\neq i}^{m} \frac{x-x_j}{x_i-x_j} \in \Pi_m, \quad x \neq x_i} \tag{4.3}$$

is called Lagrange-polynomial.

The Lagrange-polynomial satisfies

$$l_i(x_j) = \delta_{ij} \ . \tag{4.4}$$

The polynomial

$$p(x) = \sum_{i=0}^{m} y_i l_i(x) \ .$$

satisfies $p(x_j) = \sum_{i=0}^{m} y_i l_i(x_j) = y_j$, that is the interpolation exercise. The polynomial is unique in the space of functions in Π_m.

Example 4.3. *Every function* $f(x)$ *is interpolated at nodal points* a *and* b *by the linear polynomial*

$$\boxed{p(x) = f(a) - \frac{f(b)-f(a)}{b-a}(x-a) \ .}$$

4.2 Trigonometric Interpolation

Here we consider the grid

$$\Delta = \left\{ t_0 = 0 < t_1 = \frac{2\pi}{l} < \ldots < t_{l-1} = (l-1)\frac{2\pi}{l} \right\} , \tag{4.5}$$

which equally subdivides the interval $[0, 2\pi)$ into l subintervals.

The goal is to interpolate sample values $\{y_0, y_1, \ldots, y_{l-1}\}$ at Δ with a function of the form

$$y(t) = \frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos(jt) + \sum_{j=1}^{m} b_j \sin(jt) \ .$$

Such functions are called *trigonometric polynomial* of degree m.

We restrict attention to the case that $l = 2m+1$, in which case we can expect that we can solve the trigonometric interpolation exercise uniquely ($2m+1$ nodal values and $2m+1$ interpolation values): The interpolation exercise reads as follows:

Given $\{y_0, y_1, \ldots, y_{2m}\}$ determine $a_0, \{a_1, \ldots, a_m\}, \{b_1, \ldots, b_m\}$ such that

$$\boxed{y_k = \frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos(jt_k) + \sum_{j=1}^{m} b_j \sin(jt_k)\,, \quad \forall k = 0, 1, \ldots, 2m\,.} \tag{4.6}$$

The coefficients $\{a_i, b_i\}$ can be determined analytically: For this purpose we use the following expression of the sums of cos and sin:

$$\begin{aligned}
\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) &= \begin{cases} 0 & \text{for} \quad j \neq \hat{j} \text{ and } j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\}\,, \\ \frac{l}{2} & \text{for} \quad j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}\,, \\ l & \text{for} \quad j = \hat{j} = 0\,, \end{cases} \\
\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\sin(jt_k) &= 0 \text{ for } j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\}\,, \\
\sum_{k=0}^{l-1} \sin(\hat{j}t_k)\sin(jt_k) &= \begin{cases} 0 & \text{for} \quad j \neq \hat{j} \text{ and } j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\}\,, \\ \frac{l}{2} & \text{for} \quad j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}\,, \\ 0 & \text{for} \quad j = \hat{j} = 0\,. \end{cases}
\end{aligned} \tag{4.7}$$

These equalities are determined from the summation formulas

$$\begin{aligned}
\cos(jt_k)\cos(\hat{j}t_k) &= \frac{1}{2}\Big(\cos((j-\hat{j})t_k) + \cos((j+\hat{j})t_k)\Big) \\
&= \frac{1}{2}\mathrm{Re}\left(e^{\mathrm{i}(j-\hat{j})t_k} + e^{\mathrm{i}(j+\hat{j})t_k}\right)\,, \\
\cos(jt_k)\sin(\hat{j}t_k) &= \frac{1}{2}\Big(\sin((j+\hat{j})t_k) - \sin((j-\hat{j})t_k)\Big) \\
&= \frac{1}{2}\mathrm{Im}\left(e^{\mathrm{i}(j+\hat{j})t_k} - e^{\mathrm{i}(j-\hat{j})t_k}\right)\,, \\
\sin(jt_k)\sin(\hat{j}t_k) &= \frac{1}{2}\Big(\cos((j-\hat{j})t_k) - \cos((j+\hat{j})t_k)\Big) \\
&= \frac{1}{2}\mathrm{Re}\left(e^{\mathrm{i}(j-\hat{j})t_k} - e^{\mathrm{i}(j+\hat{j})t_k}\right)\,.
\end{aligned}$$

We only show the first identity of (4.7), the others are left as exercises:

Theorem 4.4.

$$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = \begin{cases} 0 & \textit{for} \quad j \neq \hat{j} \textit{ and } j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\} , \\ \frac{l}{2} & \textit{for} \quad j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\} , \\ l & \textit{for} \quad j = \hat{j} = 0 , \end{cases}$$

Proof: Denoting

$$q_+ = e^{\mathrm{i}(j+\hat{j})\frac{2\pi}{l}} \text{ and } q_- = e^{\mathrm{i}(j-\hat{j})\frac{2\pi}{l}} ,$$

it follows that

$$\begin{aligned} \sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) &= \frac{1}{2}\mathrm{Re}\sum_{k=0}^{l-1}\left(e^{\mathrm{i}(j-\hat{j})t_k} + e^{\mathrm{i}(j+\hat{j})t_k}\right) \\ &= \frac{1}{2}\mathrm{Re}\left(\sum_{k=0}^{l-1} q_-^k + \sum_{k=0}^{l-1} q_+^k\right) . \end{aligned}$$

Let us denote by

$$\sum := \sum_{k=0}^{l-1} q_-^k + \sum_{k=0}^{l-1} q_+^k .$$

- If $j = \hat{j} = 0$, then $q_+ = q_- = 1$. therefore

  $$\sum := 2l .$$

  Thus in turn

  $$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = l .$$

- If $j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}$, then $q_- = 1$, and $\sum_{k=0}^{l-1} q_-^k = \sum_{k=0}^{l-1} 1 = l$.

  Because $j + \hat{j} = 2j \in \{2, \ldots, l-1\}$ we see that $q_+ \neq 1$.

  The second term of $\sum$ is a geometric sum, that is

  $$\sum_{k=0}^{l-1} q_+^k = \frac{1 - q_+^l}{1 - q_+} = \frac{1 - e^{\mathrm{i}(j+\hat{j})2\pi}}{1 - q_+} = 0 .$$

  Therefore

  $$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = \frac{1}{2}\mathrm{Re}(0 + l) = \frac{l}{2} .$$

- if $j \neq \hat{j}$, then $0 \neq j - \hat{j}$ and $j + \hat{j} \in \{1, \ldots, l-2\}$. Therefore, both $l(j \pm \hat{j})\frac{2\pi}{l}$ are multipliers of 2π, and thus $q_{\pm}^{l} = 1$, which means that $\sum = 0$. Therefore

$$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = 0 \ .$$

□

Now, we continue with determining the coefficient $\{a_j, b_j\}$. Thereby we use three types of equalities (4.7) above :

- From (4.6) it follows that by taking into account that $\cos(0t_k) = \cos(0) = 1$,

$$\sum_{k=0}^{l-1} y_k = l\frac{a_0}{2} + \sum_{j=1}^{m} a_j \underbrace{\sum_{k=0}^{l-1} \cos(0t_k)\cos(jt_k)}_{(1.\text{in}\,(4.7)\ \text{with}\ \hat{j}=0)=0}$$
$$+ \sum_{j=1}^{m} b_k \underbrace{\sum_{k=0}^{l-1} \cos(0t_k)\sin(jt_k)}_{(2.\text{in}\,(4.7)\ \text{with}\ \hat{j}=0)=0} \ .$$

That is

$$a_0 = \frac{2}{l}\sum_{k=0}^{l-1} y_k \ .$$

- Let $\hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}$. Then, by multiplication of (4.6) with cosine functions and summation gives

$$\begin{aligned}
&\sum_{k=0}^{l-1} y_k \cos(\hat{j} t_k)\\
=&\frac{a_0}{2} \underbrace{\sum_{k=0}^{l-1} \cos(\hat{j} t_k)}_{(1.\text{in }(4.7)\text{ with } j=0)=0}\\
&+\sum_{j=1}^{m} a_j \underbrace{\sum_{k=0}^{l-1} \cos(j t_k)\cos(\hat{j} t_k)}_{(1.\text{in }(4.7))=\frac{l}{2}\delta_{j,\hat{j}}}\\
&+\sum_{j=1}^{m} b_j \underbrace{\sum_{k=0}^{l-1} \sin(j t_k)\cos(\hat{j} t_k)}_{(2.\text{in }(4.7))=0}\,,\\
=&\frac{l}{2} a_{\hat{j}}\;. \qquad \forall 1 \le \hat{j} < \frac{l}{2}\;.
\end{aligned}$$

That is

$$a_{\hat{j}} = \frac{2}{l}\sum_{k=0}^{l-1} y_k \cos(\hat{j} t_k)\;. \tag{4.8}$$

- Let $\hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}$. Multiplication of (4.6) with sine functions and summation gives

$$\begin{aligned}
&\sum_{k=0}^{l-1} y_k \sin(\hat{j} t_k)\\
=&\sum_{k=0}^{l-1}\left(\frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos(j t_k) + \sum_{j=1}^{m} b_j \sin(j t_k)\right)\sin(\hat{j} t_k)\,,\\
=&\frac{l}{2} b_{\hat{j}}\;.
\end{aligned}$$

Or in other words:

$$b_{\hat{j}} = \frac{2}{l}\sum_{k=0}^{l-1} y_k \sin(\hat{j} t_k)\;. \tag{4.9}$$

It is common to change to a complex number notation:

$$\boxed{c_{\hat{\jmath}} = a_{\hat{\jmath}} + \mathrm{i} b_{\hat{\jmath}} = \frac{2}{l}\sum_{k=0}^{l-1} y_k(\cos(\hat{\jmath} t_k) + \mathrm{i}\sin(\hat{\jmath} t_k)) = \frac{2}{l}\sum_{k=0}^{l-1} y_k \exp(\mathrm{i}\hat{\jmath} t_k)\ .} \tag{4.10}$$

Definition 4.5. *The discrete Fourier transform (DFT) of a set of* n *complex data values* $\{y_k : k = 0, \dots, l-1\}$*, which are evenly spaced in* $[0, 2\pi)$ *is the set*

$$\left\{ c_{\hat{\jmath}} = \sum_{k=0}^{l-1} y_k \exp(\mathrm{i}\hat{\jmath} t_k) : \hat{\jmath} = 0, 1, \dots, l-1 \right\}\ .$$

Note, that in comparison with (4.10) *the prefactor* $\frac{2}{l}$ *is left out.*

4.3 Fast Fourier Transform (FFT)

Is an algorithm for fast evaluation of the DFT.

Let

$$\omega = \omega_l = \exp\left(\mathrm{i}\frac{2\pi}{l}\right)\ .$$

With this notation the DFT becomes

$$\left\{ c_{\hat{\jmath}} = \sum_{k=0}^{l-1} y_k \omega^{\hat{\jmath} k} : \hat{\jmath} = 0, 1, \dots, l-1 \right\}\ .$$

We explain the FFT for a 4×4 system, that is for $l = 4$. In this case $\omega = \exp\left(\mathrm{i}\frac{2\pi}{4}\right) = \mathrm{i}$. The linear relation of the DFT is as follows:

$$\begin{aligned}
\omega^0 y_0 + \omega^0 y_1 + \omega^0 y_2 + \omega^0 y_3 &= c_0\ ,\\
\omega^0 y_0 + \omega^1 y_1 + \omega^2 y_2 + \omega^3 y_3 &= c_1\ ,\\
\omega^0 y_0 + \omega^2 y_1 + \omega^4 y_2 + \omega^6 y_3 &= c_2\ ,\\
\omega^0 y_0 + \omega^3 y_1 + \omega^6 y_2 + \omega^9 y_3 &= c_3\ .
\end{aligned}$$

Let

$$F_4 := \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^3 & \omega^6 & \omega^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & 1 & \omega^2 \\ 1 & \omega^3 & \omega^2 & \omega \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{bmatrix}\ .$$

be the *Fourier*-matrix.

Thus the system in matrix vector notation reads as follows:

$$F_4 \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} c_0 \\ c_2 \\ c_1 \\ c_3 \end{bmatrix} .$$

The matrix F_4 can be factorized as follows:

$$F_4 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & \mathrm{i} \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -\mathrm{i} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .$$

The last matrix puts the odd indices in front of the even ones. The middle matrix consists of two Fourier matrices of half size. In general we have

$$\boxed{F_{2n} = \begin{bmatrix} I_n & D_n \\ I_n & -D_n \end{bmatrix} \begin{bmatrix} F_n & 0 \\ 0 & F_n \end{bmatrix} \begin{bmatrix} 1 & 0 & & & & & & \\ & 0 & 1 & 0 & & & & \\ & & & 0 & 1 & 0 & & \\ & & & & & & & \\ 0 & 1 & 0 & & & & & \\ & & 0 & 1 & 0 & & & \\ & & & & 0 & 1 & 0 & \\ & & & & & & & \end{bmatrix} ,}$$

where I_n is the n-dimensional unitary matrix and $D_n = \mathrm{diag}(1, \omega, \ldots, \omega^{n-1})$ with $\omega = \omega_n = \exp(2\pi\mathrm{i}/n)$. The last matrix puts the odd lines on the top of the matrix and shuffles the even to the end. See Strang [4].

We calculate matrix vector multiplications after the factorization:

- We have to perform and index renumbering. Since there are no multiplications needed it does not count to the complexity.

- we need to perform two times the Fourier matrix multiplication of size $n/2$.

- We need n multiplications, when multiplying with the diagonal matrices $D_n F_n$.

Thus we have the recursive complexity

$$e(2n) = 2e(n) + \mathcal{O}(n) \ .$$

$\mathcal{O}(n)$ here refers to the fact that at most order n operations, such as multiplications with ω are performed. Let $n = 2^p$ (in our example $n = 4$ and $p = 2$). From the master theorem it follows that $e(n) = \mathcal{O}(n \log_2(n))$.

4.4 Spline Interpolation

Let $\Delta = \{a = x_0 < x_1 < \ldots < x_l = b\}$ be a grid on the interval $[a, b]$. A *step function* is a function, which satisfies

$$s(x) = s_i \, , \quad x_{i-1} \le x < x_i \, , \quad i = 1, \ldots, l \ .$$

The set of all step functions is denoted by $S_{0,\Delta}$. It is a vector space of dimension l. As basis functions we use the characteristic functions $\chi_i = \chi_{[x_{i-1},x_i)}$, $i = 1, \ldots, l$. Thus

$$s(x) = \sum_{i=1}^{l} s_i \chi_i(x) \ .$$

Remark 4.6. *Let $f : [a, b] \to \mathbb{R}$. The step function $s(x) = \sum_{i=1}^{l} s_i \chi_i(x)$ with*

$$s_i = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} f(x) \, dx \, , \quad i = 1, \ldots, l \tag{4.11}$$

is the best approximating step function with respect to the norm $\rho \to \sqrt{\int_a^b \rho^2(x) \, dx}$. That is the functional

$$\rho \in S_{0,\Delta} \to \int_a^b (f(x) - \rho(x))^2 \, dx$$

is minimal for s.

4.5 Linear Splines

Definition 4.7. *A* spline of degree n *is a function s, which is $(n-1)$–times differentiable in (a, b) and on every interval $[x_{i-1}, x_i)$ a polynomial of degree n. The space of splines of order n is denoted by $S_{n,\Delta}$.*

Of particular importance are the linear splines ($n = 1$) and cubic splines ($n = 3$).

Remark 4.8.

- *$S_{n,\Delta}$ is an $(n + l)$-dimensional space. Thus, in order to determines a spline of degree n, we have to specify the values at l nodal points and n additional conditions.*
- *A basis for $S_{1,\Delta}$ are formed by the* hat functions *Λ_i, $i = 0, 1, \ldots, l$, which are continuous, piecewise linear, and satisfy*

$$\Lambda_i(x_j) = \delta_{ij}\,, \quad i, j = 0, \ldots, l\,. \tag{4.12}$$

- (4.12) *allows for an easy computation of the interpolating spline: Let $y_0, \ldots, y_l$ sample values. Then, $s = \sum_{i=0}^{l} y_i\Lambda_i \in S_{1,\Lambda}$ is the unique spline, which satisfies*

$$s(x_i) = y_i\,, \quad i = 0, \ldots, l\,.$$

4.6 Cubic Splines

Cubic splines, that are the elements of $S_{3,\Delta}$, are used frequently in computer graphics.

We summarize some basic facts:

1. a cubic spline is two times differentiable.
2. A cubic spline is determined from $(l+3)$ measurements and conditions.
3. If $s \in S_{3,\Delta}$, then $s'' \in S_{1,\Delta}$. Thus

$$s'' = \sum_{i=0}^{l} \gamma_i\Lambda_i\,, \tag{4.13}$$

   where $\gamma_i = s''(x_i)\,, i = 0, \ldots, l\,,$. γ_i are called *moments* of s.

In the following we derive the conditions for determining a cubic spline. First, we need some auxiliary result: For an arbitrary function ρ, which is twice differentiable in $[x_{i-1}, x_i]$, we have:

$$\begin{aligned}
&\rho(x) - \rho(x_i) \\
&= \int_{x_i}^{x} \rho'(t) \cdot 1 \, dt \\
\underbrace{=}_{\text{Integration by parts}} \quad &\rho'(t)(t-x)\big|_{t=x_i}^{x} - \int_{x_i}^{x} \rho''(t)(t-x)\, dt \\
&= -\rho'(x_i)(x_i - x) - \int_{x_i}^{x} \rho''(t)(t-x)\, dt \; .
\end{aligned} \tag{4.14}$$

Moreover, for an arbitrary $t \in [x_{i-1}, x_i]$ we have

$$\begin{aligned}
s''(t) &= \gamma_{i-1}\Lambda_{i-1}(t) + \gamma_i \Lambda_i(t) \\
&= \gamma_{i-1}\frac{x_i - t}{x_i - x_{i-1}} + \gamma_i \frac{t - x_{i-1}}{x_i - x_{i-1}} \\
&= -\frac{\gamma_{i-1}}{h_i}(t - x_i) + \frac{\gamma_i}{h_i}(t - x_{i-1}) \\
&= \frac{\gamma_i - \gamma_{i-1}}{h_i}(t - x_i) + \gamma_i \, ,
\end{aligned} \tag{4.15}$$

which implies that for every $x \in [x_{i-1}, x_i)$

$$\begin{aligned}
&s(x) - s(x_i) + s'(x_i)(x_i - x) \\
\underbrace{=}_{(4.14)} &- \int_{x_i}^{x} s''(t)(t-x)\, dt \\
\underbrace{=}_{(4.15)} &- \frac{\gamma_i - \gamma_{i-1}}{h_i} \int_{x_i}^{x} (t - x_i)(t-x)\, dt - \gamma_i \int_{x_i}^{x} t - x \, dt \\
\underbrace{=}_{\text{Integration by parts}} \quad &\frac{\gamma_i - \gamma_{i-1}}{2h_i} \int_{x_i}^{x} (t - x_i)^2 \, dt \\
&+ \frac{\gamma_i - \gamma_{i-1}}{2h_i}(t - x_i)^2 (t - x)\bigg|_{t=x_i}^{x} \\
&- \frac{\gamma_i}{2}(t-x)^2\bigg|_{t=x_i}^{x} \\
=&\frac{\gamma_i - \gamma_{i-1}}{h_i}\frac{(x - x_i)^3}{6} + \gamma_i \frac{(x - x_i)^2}{2} \; .
\end{aligned} \tag{4.16}$$

Using the abbreviations we get

$$s_i = s(x_i) \text{ and } s'_i = s'(x_i) \text{ for } i = 0, \ldots, l \, .$$

Thus from (4.16) it follows that for every $x \in [x_{i-1}, x_i)$ and $i = 1, \ldots, l$

$$s(x) = s_i + s'_i(x - x_i) + \gamma_i \frac{(x-x_i)^2}{2} + \frac{\gamma_i - \gamma_{i-1}}{h_i} \frac{(x-x_i)^3}{6} \, . \tag{4.17}$$

In particular for $i = 1, \ldots, l$ we have

$$\begin{aligned} s_{i-1} &= s_i - s'_i h_i + \frac{\gamma_i h_i^2}{2} - \frac{(\gamma_i - \gamma_{i-1}) h_i^2}{6} \\ &= s_i - s'_i h_i + \frac{h_i^2}{6}(2\gamma_i + \gamma_{i-1}) \, , \\ s'_{i-1} &= s'_i - \frac{h_i}{2}(\gamma_{i-1} + \gamma_i) \, . \end{aligned} \tag{4.18}$$

Combinations of these equations shows

$$\begin{aligned} &\frac{s_{i+1} - s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i} \\ =&s'_{i+1} - \gamma_i \frac{h_{i+1}}{6} - \gamma_{i+1} \frac{h_{i+1}}{3} - s'_i + \gamma_{i-1} \frac{h_i}{6} + \gamma_i \frac{h_i}{3} \\ =&(\gamma_i + \gamma_{i+1}) \frac{h_{i+1}}{2} - \gamma_i \frac{h_{i+1}}{6} - \gamma_{i+1} \frac{h_{i+1}}{3} + \gamma_{i-1} \frac{h_i}{6} + \gamma_i \frac{h_i}{3} \\ =&\frac{1}{6}(h_{i+1}\gamma_{i+1} + 2\gamma_i(h_i + h_{i+1}) + \gamma_{i-1} h_i) \, . \end{aligned}$$

Writing this system in matrix notation we get

$$\underbrace{\frac{1}{6}\begin{bmatrix} h_1 & 2(h_1+h_2) & h_2 & & & 0 \\ & h_2 & 2(h_2+h_3) & \ddots & & \\ & & \ddots & \ddots & h_{l-1} & \\ & & & h_{l-1} & 2(h_{l-1}+h_l) & h_l \end{bmatrix}}_{\in\mathbb{R}^{(l-1)\times(l+1)}} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{l-1} \\ \gamma_l \end{bmatrix}$$

$$= - \underbrace{\begin{bmatrix} -h_1^{-1} & h_1^{-1}+h_2^{-1} & -h_2^{-1} & & & 0 \\ & -h_2^{-1} & h_2^{-1}+h_3^{-1} & \ddots & & \\ & & \ddots & \ddots & -h_{l-1}^{-1} & \\ & & & -h_{l-1}^{-1} & h_{l-1}^{-1}+h_l^{-1} & -h_l^{-1} \end{bmatrix}}_{\in\mathbb{R}^{(l-1)\times(l+1)}} \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{l-1} \\ s_l \end{bmatrix} . \tag{4.19}$$

The matrices in (4.19) have dimension $(l-1)\times(l+1)$, and thus are underdetermined. Thus, additional conditions are required: For a *natural cubic spline* we request in addition that

$$s''(a) = s''(b) = 0 \, . \tag{4.20}$$

However, (4.13) then shows, that

$$\gamma_0 = s''(a) = 0 \text{ and } \gamma_l = s''(b) = 0 \, . \tag{4.21}$$

Thus the system (4.19) simplifies to

$$\underbrace{\frac{1}{6}\begin{bmatrix} 2(h_1+h_2) & h_2 & & 0 \\ h_2 & 2(h_2+h_3) & \ddots & \\ & \ddots & \ddots & h_{l-1} \\ & & h_{l-1} & 2(h_{l-1}+h_l) \end{bmatrix}}_{:=\mathcal{G}\in\mathbb{R}^{(l-1)\times(l-1)}} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{l-1} \end{bmatrix} \tag{4.22}$$

$$= \begin{bmatrix} d_1 \\ \vdots \\ d_{l-1} \end{bmatrix} ,$$

where

$$d_i = \frac{s_{i+1} - s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i} = \frac{s_{i+1}}{h_{i+1}} - s_i \left(\frac{1}{h_i} + \frac{1}{h_{i+1}} \right) + \frac{s_{i-1}}{h_{i-1}} \,, \qquad i = 1, \dots, l-1 \,. \tag{4.23}$$

Example 4.9. *We consider an equidistant grid with step size h. For given $j = 1, \dots, l-1$ we determine the natural cubic spline s which satisfies*

$$s(x_i) = s_i = \delta_{ij} \,, \quad i = 0, \dots, l \,. \tag{4.24}$$

The system (4.22), (4.23) *reads as follows:*

$$\begin{aligned}
\frac{1}{6} \begin{bmatrix} 4 & 1 & & 0 \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 4 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{l-1} \end{bmatrix}
&= -\frac{1}{h^2} \begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_{l-1} \end{bmatrix} \\
&= -\frac{1}{h^2} \begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= \frac{1}{h^2} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -2 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \,.
\end{aligned} \tag{4.25}$$

Chapter 5

Mathematical Modelling using Partial Differential Equations

Many processes that are studied in the natural sciences are described by the change of a variable over time in dependence of the change of the variable in space. Such an equation is called a partial differential equation (PDE). In the following two examples are presented for motivating the methods for numerical solution of PDEs.

5.1 Heat equation

We consider the distribution of heat over time in a thin conductor that is fully isolated. The internal energy per unit volume Q in a small spatial region is given by

$$Q = c\rho u(x,t),$$

where c is the specific heat capacity, ρ is the mass density of the material and $u(x,t)$ is the temperature at position x at time t. The change in internal energy over time is

$$(2Adx)\frac{\partial}{\partial t}c\rho u(x,t) = AJ(x-dx,t) - AJ(x+dx,t),$$

where A is the cross section of the conductor, $2dx$ is the length of the control element, i.e. $2Adx$ is the volume of the element. The change in internal energy over time equals the flux into and out of the control element, where

$J(x,t)$ is the heat flux (heat energy surface flux) at position x and at time t. Assuming c, and ρ are time independent for $dt \to 0$ this yields

$$\begin{aligned} c\rho \frac{\partial}{\partial t} u(x,t) &= \lim_{dx \to 0} \frac{J(x-dx,t) - J(x+dx,t)}{2dx} \\ &= -\frac{\partial}{\partial x} J(x,t). \end{aligned} \tag{5.1}$$

Fouriers law states that the heat flux $J(x,t)$ is proportional to the change in temperature, i.e.

$$J(x,t) = -k \frac{\partial}{\partial x} u(x,t).$$

Inserting into Eqn (5.1) and assuming constant k and $c\rho > 0$, the heat equation in one dimension is given by

$$\boxed{\frac{\partial}{\partial t} u(x,t) = \frac{k}{c\rho} \frac{\partial^2}{\partial x^2} u(x,t).} \tag{5.2}$$

If we want to solve a PDE over time we need to know

1. the PDE including the spatial domain Ω
2. the initial conditions $u(x,0) = u_0(x)$, $x \in \Omega$
3. what happens at the boundaries of the domain $x \in \delta\Omega$ for $t > 0$

In the example of the 1-dimensional heat equation the domain Ω is any interval $[L,R]$. The initial heat distribution must be known in this interval, $u(x,0) = u_0(x)$, $x \in [L,R]$. The boundary conditions specify the behaviour at the left and right end of the conductor. The most important boundary conditions are

1. to set the temperature at the boundaries, i.e. $u(L,t) = u_L(t)$ at the left, and (or) $u(R,t) = u_R(t)$ at the right boundary. This is called a *Dirichlet* boundary condition.
2. to set the heat flux into or out of the domain, i.e. $c\rho \frac{\partial u}{\partial x}(L,t) = J_L(t)$ at the left, and (or) $c\rho \frac{\partial u}{\partial x}(R,t) = J_R(t)$ at the right boundary. This is called a *Neumann* boundary condition.

The heat equation in two or three dimension can be modelled in a similar way

$$\begin{aligned} c\rho\frac{\partial}{\partial t}u &= \nabla\cdot(k\nabla u) && \text{for } t>0, x\in\Omega \\ u(x,0) &= u_0(x) && \text{for } t=0, x\in\Omega \\ c\rho\nabla u\cdot n &= J(x,t) && \text{for } t>0, x\in\delta\Omega \end{aligned} \tag{5.3}$$

where $n := n(x)$ is the outward normal of the boundary surface.

Remark 5.1. *The Nabla operator is defined as $\nabla = (\partial/\partial x, \partial/\partial y\,[,\partial/\partial z])^T$ in two and three dimensions, and ∇u denotes the gradient of u, $\nabla\cdot u$ the divergence of u. The Laplace operator is written as $\Delta = \nabla\cdot\nabla = \partial^2/\partial x^2 + \partial^2/\partial y^2\,[+\partial^2/\partial z^2]$ and for constant k, c, and ρ the heat equation is often written as*

$$\frac{\partial}{\partial t}u = \frac{k}{c\rho}\Delta u.$$

5.2 Wave equation

We describe an elastic string that is fixed at the same vertical position at the right and left end ($\Omega = [L,R]$). Starting point is Newton's law of motion stating force F is mass m times acceleration a, i.e.

$$F = ma.$$

The mass of a control volume from x to $x+dx$ is given by $m = \rho A dx$, where A is the cross section of the elastic string, and ρ the material density.

If $u(x,t)$ describes the vertical displacement of the string from its equilibrium position, the acceleration is given by $a = \frac{\partial^2}{\partial t^2}u(x,t)$.

The force per cross section is the net vertical component of the tension:

$$\begin{aligned} \frac{F}{A} &= T\sin\theta_2 - T\sin\theta_1 \\ &\approx T(\theta_2-\theta_1) \\ &= T\left(\frac{\partial u}{\partial x}(x+dx,t) - \frac{\partial u}{\partial x}(x,t)\right), \end{aligned}$$

where the tension T is constant within the string, and θ_1 and θ_2 are the angles of the tangents in points $u(x,t)$ and $u(x+dx,t)$ respectively for a fixed t.

Putting these all together yields

$$\rho A dx \frac{\partial^2}{\partial t^2} u(x,t) = AT \left(\frac{\partial u}{\partial x}(x+dx,t) - \frac{\partial u}{\partial x}(x,t) \right)$$

and if make the control volume small, i.e. $dx \to 0$, we obtain the wave equation

$$\boxed{\frac{\partial^2}{\partial t^2} u(x,t) = c^2 \frac{\partial^2}{\partial x^2} u(x,t)} \tag{5.4}$$

where $c^2 = \frac{T}{\rho}$.

Dirichlet boundary conditions are used stating that there is no displacement at the suspension, i.e. $u(L,t) = 0$ and $u(R,t) = 0$ for all t.

As the PDE is second order in time one equation of describing the initial displacements, and one the initial velocities is needed:

$$\begin{aligned} u(x,0) &= u_0(x) \quad x \in [L,R] \\ \frac{\partial}{\partial t} u(x,0) &= u_{t,0}(x) \quad x \in [L,R] \end{aligned}$$

In two or three dimension the wave equation is given by

$$\frac{\partial^2}{\partial t^2} u = c^2 \Delta u.$$

5.3 Nondimensionalisation

In contrast to mathematical equations, in the modelling of physical processes variables and coefficients have units, since their values describe physical quantities. If two expressions are compared by some relation it is essential that they have the same units (e.g. it makes no sense to compare a meter to a second). Note that the physical units are completely independent of the spatial dimensions of the model.

While units can help to understand model variables and parameters, we need a dimensionless equation for a mathematical analysis. This is achieved by scaling the variables

$$u = [u]u^*,$$

where $[u]$ is the scale (with units) and u^* is the dimensionless variable. As an simple example the following ODE is nondimensionalised:

$$\frac{dN}{dt} = -\lambda N, \text{ with } N(0) = N_0,$$

with $t = [t]t^*$ and $N = [N]N^*$ this becomes

$$\frac{dN^*}{dt^*} = -\{[t]\lambda\}N^*, \text{ with } N^*(0) = \{N_0/[N]\},$$

where all terms (including the expressions in curly brackets) are dimensionless. Choosing the scales $[N] = N_0$ and $[t] = 1/\lambda$ the equation simplifies to

$$\frac{dN^*}{dt^*} = -N^*, \text{ with } N^*(0) = 1.$$

Remark 5.2. *By picking the right scales the heat equation without dimensions is given by*

$$\frac{\partial}{\partial t}u = \Delta u$$

and the wave equation by

$$\frac{\partial^2}{\partial t^2}u = \Delta u.$$

In more complicated examples some parameters might not vanish by scaling, but turn into dimensionless parameters, that determine the behaviour of the PDE. Examples of such dimensionless numbers are Reynolds number, Peclet number, or Darcy number.

Chapter 6

Classification of Second Order Linear Partial Differential Equations

In this chapter we present how to classify linear PDEs, and restrict our attention to partial differential equations of second order in two variables. Generally, such an equation for a function $u = u(x, y)$ reads as :

$$Au_{xx} + 2B_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0\ . \tag{6.1}$$

Here $A = A(x, y), \ldots, G = G(x, y)$ are again functions.

Definition 6.1. *A PDE* (6.1) *is called*

- elliptic *if* $AC - B^2 > 0$ *,*
- parabolic *if* $AC - B^2 = 0$ *, and*
- hyperbolic *if* $AC - B^2 < 0$ *.*

Example 6.2.
The wave equation

$$\frac{1}{c^2}u_{xx} - u_{yy} = 0$$

(note we changed the notation from t to y) is of the form (6.1) *with*

$$A = c^{-2}, B = 0, C = -1\ .$$

Because

$$AC - B^2 = -c^{-2} < 0\,,$$

the equation is hyperbolic.

The Laplace equation

$$u_{xx} + u_{yy} = 0$$

is of the form (6.1) *with*

$$A = 1, B = 0, C = 1\ .$$

Because

$$AC - B^2 = 1 \geq 0\,,$$

the equation is elliptic.

The heat equation

$$u_x - u_{yy} = 0$$

(note we changed the notation from t to x and x to y) is of the form (6.1) *with*

$$A = 0, B = 0, C = -1\ .$$

Because

$$AC - B^2 = 0\,,$$

the equation is parabolic.

These three PDE are the archetypical equations presenting each problem class.

Remark 6.3.

1. *For the classification on the* main symbol

   $$Au_{xx} + 2Bu_{xy} + Cu_{yy}$$

   is relevant. These are the terms of the differential equation of highest order (in our case this is 2*).*

2. $AC - B^2$ *is the determinant of the symmetric matrix*

   $$M = \begin{pmatrix} A & B \\ B & C \end{pmatrix}\ .$$

Denoting by $m_{11} = A$, $m_{12} = m_{21} = B$, $m_{22} = C$ and $x_1 = x$, $x_2 = y$ the main symbol reads as follows

$$\sum_{i,j=1}^{2} m_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} \ .$$

This acts as a basis for a classification in higher dimensions.

3. *If the coefficients A, B, C are not constant, but functions which depend on x and y in a non-trivial manner, then the type of the partial differential equation can be different at various points (x, y). For instance the differential equation*

$$x u_{xx} + u_{yy} = 0$$

is elliptic for $x > 0$, parabolic for $x = 0$, and hyperbolic for $x < 0$.

4. *The terminology elliptic, parabolic and hyperbolic is motivated from conic sections. A curve $(X, Y(X))$, which satisfies the equation*

$$AX^2 + 2BXY + CY^2 + DX + EY + F = 0 \qquad \textit{(with constant coefficients)}$$

is either an ellipsis, parabola, or an hyperbola, depending on the sign of $AC - B^2$. For instance

   (a) *$A = C = 1$, $B = D = E = 0$, $F = -1$ leads to $X^2 + Y^2 = 1$, a circle.*

   (b) *More general $A = \frac{1}{a^2}$, $C = \frac{1}{b^2}$, $B = D = E = 0$, $F = -1$ leads to $\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1$, an ellipsis.*

   (c) *$A = F = 1$, $C = -1$ and $B = D = E = 0$ lead to, $y^2 = x^2 + 1$, which is an hyperbola.*

   (d) *$A = 1$, $B = C = D = F = 0$, $E = -1$ leads to $Y = X^2$, which is a parabola.*

The most important property of this classification is that it is invariant under coordinate transformations, i.e. a coordinate transformation does not change the type of the differential equation.

We consider the change of coordinates:

$$\begin{aligned} L_1(x,y) &= ax + by\,, \\ L_2(x,y) &= cx + dy\,, \text{ and therefore,} \\ \frac{\partial (u \circ L)}{\partial x}(x,y) &= a\partial_1 u(L(x,y)) + c\partial_2 u(L(x,y))\,, \\ \frac{\partial (u \circ L)}{\partial y}(x,y) &= b\partial_1 u(L(x,y)) + d\partial_2 u(L(x,y))\,. \end{aligned}$$

Above notation is ugly, however, it should emphasize that on the right hand side we differentiate with respect to the first component and not the x variable.

The second order derivatives are given by

$$\begin{aligned} \frac{\partial^2 (u \circ L)}{\partial^2 x}(x,y) &= a^2\partial_1^2 u(L(x,y)) + 2ac\partial_{12}^2 u(L(x,y)) + c^2\partial_2^2 u(L(x,y))\,, \\ \frac{\partial^2 (u \circ L)}{\partial x \partial y}(x,y) &= ab\partial_1^2 u(L(x,y)) + (ad+bc)\partial_{12}^2 u(L(x,y)) + cd\partial_2^2 u(L(x,y))\,, \\ \frac{\partial^2 (u \circ L)}{\partial^2 y}(x,y) &= b^2\partial_1^2 u(L(x,y)) + 2bd\partial_{12}^2 u(L(x,y)) + d^2\partial_2^2 u(L(x,y))\,. \end{aligned}$$

This can be written now into compact matrix form:

$$\begin{aligned} &\underbrace{\begin{pmatrix} \frac{\partial^2 (u\circ L)}{\partial^2 x}(x,y) & \frac{\partial^2 (u\circ L)}{\partial x\partial y}(x,y) \\ \frac{\partial^2 (u\circ L)}{\partial x\partial y}(x,y) & \frac{\partial^2 (u\circ L)}{\partial^2 y}(x,y) \end{pmatrix}}_{=:C} \\ =&A \underbrace{\begin{pmatrix} \partial_1^2 u(L(x,y)) & \partial_{12}^2 u(L(x,y)) \\ \partial_{12}^2 u(L(x,y)) & \partial_2^2 u(L(x,y)) \end{pmatrix}}_{=:U''} A^T \end{aligned}$$

with

$$A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}\,.$$

Thus, we have

$$(\det(A))^2 \det(U'') = \det(C)\,,$$

and thus the determinants of C and A have equal signs. This shows that the type does not change by linear transformations.

Remark 6.4. *The characterisation can be loosely summarized as:*

- elliptic - *time independent,*
- parabolic - *time dependent and diffusive,*
- hyperbolic - *time dependent and wavelike with finite propagation speed.*

Chapter 7

Finite Difference Method

7.1 Elliptic Differential Equations

An one-dimensional elliptic PDE coincides with the BVP described in Chapter 3.

In the following we study the Poisson equation in two dimensions with Dirichlet boundary conditions:

$$\begin{aligned} -\Delta u(x,y) = -\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) u(x,y) &= f(x,y) \text{ for } (x,y) \in \Omega \\ u(x,y) &= \phi(x,y) \text{ for } (x,y) \in \delta\Omega \end{aligned} \tag{7.1}$$

For simplicity the domain Ω under investigation is the unit square $[0,1] \times [0,1]$.

We discretise the domain Ω by a equidistant rectangular grid with mesh size $h = 1/N$, and use the notation

$$\begin{aligned} x_{i,j} = (x_i, y_j) &:= (ih, jh) && \text{with } i,j \in 0 \dots N, \text{ and} \\ u_{i,j} = u(x_i, y_j) &:= u(ih, jh) && \text{with } i,j \in 0 \dots N. \end{aligned}$$

In this notation the one-sided forward and backwards difference operator, as well as the second order central difference operator for inner grid points are given by

$$\begin{aligned} D_x^-[u_h] &= \frac{1}{h}(u_{i,j} - u_{i-1,j}), & D_y^-[u_h] &= \frac{1}{h}(u_{i,j} - u_{i,j-1}), \\ D_x^+[u_h] &= \frac{1}{h}(u_{i+1,j} - u_{i,j}), & D_y^+[u_h] &= \frac{1}{h}(u_{i,j+1} - u_{i,j}), \\ D_x[u_h] &= \frac{1}{2h}(u_{i+1,j} - u_{i-1,j}), & D_y[u_h] &= \frac{1}{2h}(u_{i,j+1} - u_{i,j+1}), \\ D_x^2[u_h] &= \frac{1}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}), \end{aligned} \tag{7.2}$$

wbhere u_h is the restriction of u to the grid points, i.e. $u_h(x_{i,j}) = u(x_{i,j})$.

Therefore, Equation (7.1) is approximated by

$$\underbrace{-\Delta u(x,y) = f}_{L[u]=f} \approx \underbrace{-(D_x^2[u_h] + D_y^2[u_h]) = f(x_i, y_j)}_{L_h[u_h]=f_h},$$

where L is the differential operator, and L_h is the corresponding discrete differential operator. Therefore, the discrete Poisson equation is given by

$$\frac{1}{h^2}(4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}) = f_{i,j}\,, \tag{7.3}$$

for all inner points (x_i, y_j) with $i, j = 1 \dots N-1$. For Dirichlet boundary conditions the values at the boundaries are predetermined by

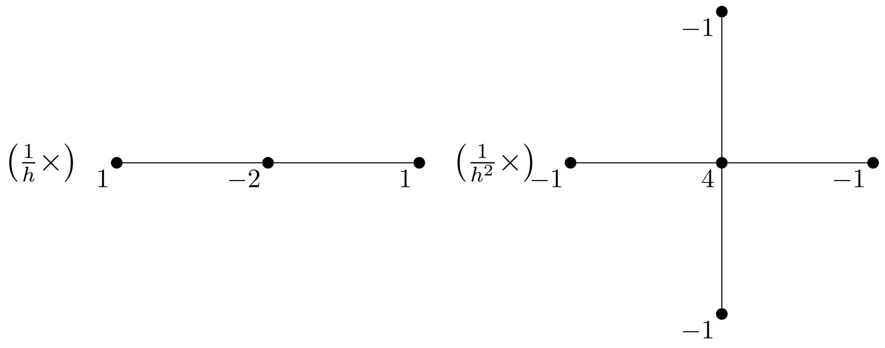
$$\begin{aligned} u_{i,j} = \phi(x_i, x_j) \text{ for } & i = \{0, N\}, j = 0 \dots N, \text{ top and bottom,} \\ \text{or } & i = 0 \dots N, j = \{0, N\}, \text{ left and right.} \end{aligned}$$

Definition 7.1. *The* local truncation error ξ_h *is the residuum of the exact solution* u^* *of* $L[u] = f$ *plugged into the discrete operator, i.e.* $\xi_h := L_h[u_h^*] - f_h$*. The asymptotic convergence order of the local truncation error when* $h \to 0$ *is called the* consistency order *of the method.*

Lemma 7.2. *The discrete Poisson equation (Eqn 7.3) has a consistency order of* $O(h^2)$*.*

Proof. by Taylor expansion. □

Note that discrete differential operators are often visualized by their stencils, e.g.:



$D_x^2[u]$ $-(D_x^2[u] + D_y^2[u])$

We solve Eqn (7.3) for all inner values u_{ij} i.e. there are $(N-1)^2$ unknowns and the same amount of equations. By sorting $u_{i,j}$ into a vector v_k (e.g. $v_k := u_{i,j}$ with $k := (i-1)(N-1)+j$) we can put it as a linear equation

$$Av = b, \tag{7.4}$$

where each row of the matrix describes Eqn (7.3) for a specific inner point $x_{i,j}$ for $i, j = 1 \ldots N-1$. This linear equation is reformulation of the discrete equation $L_h[u_h] = f_h$.

More generally, the number of entries per row of A is exactly the number of neighbours considered in the stencil plus the node itself. The vector b consists of the values $f_{i,j}$ together with the contribution of the Dirichlet boundary conditions. Note that the matrix A is sparse and the linear system must be solved using iterative methods.

In the case of Neumann boundary conditions the situation is slightly more complicated. To achieve a consistency order of $O(h^2)$ we have to approximate the boundary flux by central differences. Therefore, at the boundary the point is calculated from the boundary condition. Generally, the Neumann boundary condition for a flux ϕ is given by

$$\nabla u \cdot n = \phi.$$

In the case of the domain $\Omega = [0,1] \times [0,1]$ the outward normals are given by $n_{top} = (0,-1)^T, n_{bot} = (0,1)^T, n_{left} = (-1,0)^T, n_{right} = (0,1)^T$. We exemplify the approach for x_{ij} at the right boundary, i.e. $j = N - 1$,

$$\begin{aligned}\nabla u \cdot n_{right} = \phi \Rightarrow & \frac{\partial}{\partial x} u = \phi \\ & \approx \frac{1}{2h}(u_{i,N-2} - u_{i,N}) = \phi_{i,N-1} \\ & \Rightarrow u_{i,N} = u_{N-2} - 2h\phi_{i,N-1}\end{aligned}$$

Eqn (7.3) for the point $x_{i,N-1}$ is given by

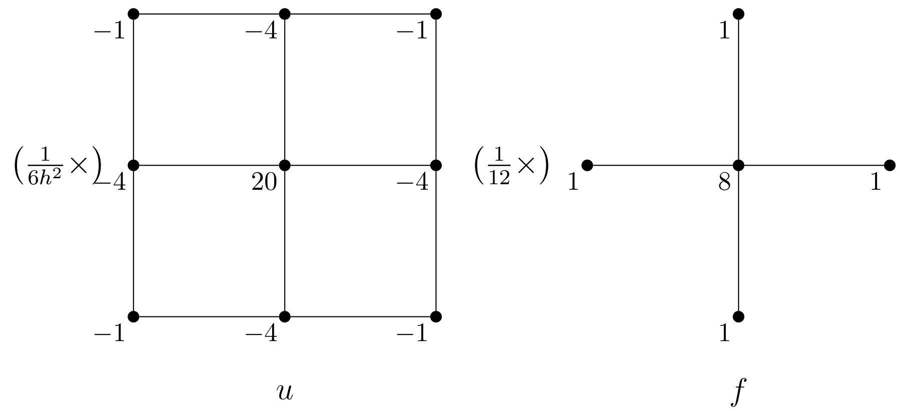
$$\frac{1}{h^2}(4u_{i,N-1} - u_{i+1,N-1} - u_{i,N-1} - u_{i,N} - u_{i,N-2}) = f_{i,N-1}\,,$$

and inserting $u_{i,N}$ from above yields

$$\frac{1}{h^2}(4u_{i,N-1} - u_{i+1,N-1} - u_{i,N-1} - 2u_{i,N-2}) = f_{i,N-1}\frac{1}{2h}\phi_{i,N-1}.$$

Therefore, the Neumann boundary condition changes the matrix A and (if the flow is not equal zero) the right hand side b.

The proposed method has consistency order $O(h^2)$. Higher order methods can be derived by using bigger stencils (i.e. using a bigger neighbourhood), and or by additionally using a stencil for f. A 9-point finite difference scheme with a consistency order of $O(h^4)$ is given by



Again, this can be shown by Taylor expansion.

Definition 7.3. *The* global error *is defined as $\epsilon_h := u_h - u_h^*$, where u_h is the solution of discrete equation $L_h u_h = f_h$ and u_h^* is the exact solution at the grid points. The asymptotic convergence order of the global error when $h \to 0$ is called the* convergence order *of the method*

For a linear discrete operator local and global error are related by

$$L_h \epsilon_h = -\xi_h,$$

therefore, the global error can be calculated from the local error

$$\epsilon_h = -L_h^{-1} \xi_h,$$

and the convergence order of ϵ_h is obtained by estimating

$$\|\epsilon_h\| \leq \left\|L_h^{-1}\right\| \|\xi_h\|$$

in a suitable norm.

For L_h of the model problem (see Eqn 7.3) the eigenfunctions $u_h^{k,l}$ and corresponding eigenvalues $\lambda^{k,l} > 0$ are known

$$\begin{aligned} u_h^{k,l}(x,y) &= \sin(k\pi x)\sin(l\pi y) \\ \lambda^{k,l} &= \frac{4}{h^2}\left(\sin^2(k\pi\frac{h}{2}) + \sin^2(l\pi\frac{h}{2})\right), \end{aligned}$$

for $k, l = 1 \dots N-1$.

Lemma 7.4. *For the inverse discrete operator L_h^{-1} of the discrete Poisson problem (Eqn 7.3)*

$$\left\|L_h^{-1}\right\|_2 \leq \frac{1}{2\pi^2} + O(h^2)$$

Proof. The largest eigenvalue of L_h^{-1} is $\frac{1}{\lambda_{min}}$, where λ_{min} is the smallest eigenvalue of L_h

$$\lambda_{min} = \lambda^{1,1} = \frac{4}{h^2}\left(2\sin^2(\pi\frac{h}{2})\right) \geq \frac{8}{h^2}\left(\frac{\pi^2 h^2}{4} + O(h^4)\right).$$

□

7.2 Parabolic Differential Equations

In this section we study the numerical solution of parabolic initial value problems exemplary for the model problem

$$\begin{aligned} u_t &= L[u] + f && \text{for } t \geq 0, x \in \Omega\,, \\ u(x,0) &= u_0(x) && \text{for } x \in \Omega\,, \\ u(x,t) &= \phi(x,t) && \text{for } t \geq 0, x \in \delta\Omega\,, \end{aligned} \tag{7.5}$$

where t is time, x is the spatial variable, $u(x,t)$ the unknown function, $u_t := \partial u \partial t$, and $f := f(x,t)$ a source or sink term.

Generally, an elliptic differential operator of second order has following structure

$$L[u](x,t) = \underbrace{\nabla \cdot (c(x)\nabla u(x,t))}_{diffusion} - \underbrace{p(x) \cdot \nabla u(x,t)}_{convection} - \underbrace{q(x)u(x,t)}_{reaction}. \tag{7.6}$$

It is possible to use the same methods as in the elliptic case and leave time continuous, yielding

$$\frac{du_h}{dt}(t) = L_h[u_h(t)] + f_h(t), \tag{7.7}$$

which is an stiff ODE and can be solved with the methods presented in the previous chapters. This approach is called *(vertical) methods of lines*, because it reduces the PDE to a system of ordinary differential equations with respect time.

Example 7.5. *Consider the discrete heat equation in one dimension using $D_x^2[u_h]$ (see Eqn 7.2), then*

$$u_i'(t) = \frac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1}).$$

The simplest way of to discretise time is to use the explicit or implicit Euler method applied to Eqn (7.7). The explicit Euler method is given by

$$\begin{aligned} \frac{u_h^{k+1} - u_h^k}{\tau} &= A_h u_h^k + f_h(t_k) \\ u_h^{k+1} &= (\tau A_h + I)u_h^k + \tau f_h(t_k)\,, \end{aligned} \tag{7.8}$$

where τ is the time step, the superscript k of u_h denotes the time index, $t_k = k\tau$, and A_h is the matrix describing the discrete operator L_h. Implicit Euler is given by

$$\begin{aligned} \frac{u_h^{k+1} - u_h^k}{\tau} &= A_h u_h^{k+1} + f_h(t_{k+1}) \\ (I - \tau A_h)u_h^{k+1} &= \tau f_h(t_k) + u_h^k\,, \end{aligned} \tag{7.9}$$

where a linear system must be solved in each time step.

Definition 7.6. *Similar to the elliptic case the* spatial local truncation error ξ_h *of a parabolic PDE is the residuum of the exact solution* u^* *of* $u_t = L[u]+f$ *plugged into the discrete operator at a fixed time* t_k*, i.e.*

$$\xi_h^k := u^{*\prime}(t_k) - L_h[u^*(t_k)] - f_h(t_k).$$

The local truncation error of the full scheme is calculated plugging the exact solution in the discrete operators regarding space and time

$$\zeta_h^k := B_h[u^*](t_k) - L_h[u^*](t_k) - f_h(t_k).$$

Again, the asymptotic convergence order of the local truncation error when $h \to 0$ *is called the* consistency order *of the method.*

Example 7.7. *Consider the heat equation using* $D_x^2[u_h]$ *and the explicit Euler, then*

$$\frac{u_i^{k+1} - u_i^k}{\tau} = \frac{1}{h^2}(u_{i-1}^k - 2u_i^k + u_{i+1}^k),$$

and more generally,

$$\frac{u_h^{k+1} - u_h^k}{\tau} = A_h u_h^k + f_h(t_k).$$

The local truncation error is given by

$$\begin{aligned} \zeta_h^k &= \frac{u_h^*(t_{k+1}) - u_h^*(t_k)}{\tau} - A_h u_h^*(t_k) - f_h(t_k) \\ &= \underbrace{\frac{u_h^*(t_{k+1}) - u_h^*(t_k)}{\tau} - u_h^{*\prime}(t_k)}_{=O(\tau)\ (Taylor)} + \underbrace{u_h^{*'}(t_k) - A_h u_h^*(t_k) - f_h(t_k)}_{=\left\|\xi_h^k\right\|=O(h^2)\ for\ t_k\ fixed} \end{aligned}$$

Therefore, the consistency order is of $O(\tau) + O(h^2)$*. For implicit Euler this can be shown in the same way.*

Definition 7.8. *The* global error *is defined (as in Def 7.3) for each time* t_k *as*

$$\epsilon_h^k := u_h^k - u_h^*(t_k).$$

Calculation of the global error from the local truncation error is more complicated than in the elliptic case. For explicit and implicit Euler the global error can be calculated by

$$\frac{\epsilon_h^{k+1} - \epsilon_h^k}{\tau} = A_h \epsilon_h^k - \xi_h(t_k) \tag{7.10}$$

and

$$\frac{\epsilon_h^{k+1} - \epsilon_h^k}{\tau} = A_h \epsilon_h^{k+1} - \xi_h(t_k)\,, \tag{7.11}$$

where $\xi_h(t_k) = O(\tau) + O(h^2)$ for central differences.

Definition 7.9. Stability *of a finite difference equation: we call the finite difference equation stable when two solution (u_h and v_h) that are close in the beginning stay close, i.e. there is an estimation*

$$\left\|u_h^k - v_h^k\right\| \leq c_1(t_k)\left\|u_h^0 - v_h^0\right\| + c_2(t_k) \sup_{0\leq l\leq k} \left\|\delta_h^l\right\|$$

where $c_1(t_k)$ and $c_2(t_k)$ are functions that are independent of τ and h, δ_h^k is the perturbation at time t_k.

Theorem 7.10. *Consistency and stability implicate convergence: If the finite difference equation is stable and consistent with order $\left\|\xi_h^k\right\| = O(h^p) + O(\tau^q)$, then the method is convergent with the same order, thus $\left\|\epsilon_h^k\right\| = O(h^p) + O(\tau^q)$.*

Proof. Take u_h and $v_h := u_h^*$ in the stability estimation. The perturbation is exactly the local truncation error, and with $u_h^0 = u_h^*(0)$ convergence is shown. □

In the following we analyse the stability of the implicit and explicit Euler method. For the explicit Euler we obtain a recursion for the global error from Eqn (7.10)

$$\epsilon_h^{k+1} = (I + \tau A_h)\epsilon_h^k + \tau \xi_h(t_k).$$

Stability depends on $\|I + \tau A_h\|_2$. The matrix for 1D Poisson with $D_x^2[u_h]$ has eigenvalues

$$1 + \tau\lambda^l = 1 - \frac{4\tau}{h^2}\sin^2(l\pi\frac{h}{2}) \in [1 - \frac{4\tau}{h^2}, 1],$$

thus,

$$\|I + \tau A_h\|_2 \approx |1 - \frac{4\tau}{h^2}| \leq 1 \text{ for } \tau \leq \frac{h^2}{2},$$

otherwise $\|I + \tau A_h\|_2 \geq 1$ and the method becomes unstable.

For the implicit Euler we follow the same approach, from Eqn (7.11) we obtain

$$\epsilon_h^{k+1} = (I - \tau A_h)^{-1}(\epsilon_h^k + \tau\xi_h(t_k)).$$

This method is unconditionally stable, since

$$\left\|(I-\tau A_h)^{-1}\right\|_2 \leq \frac{1}{1-\tau\lambda_1} \approx \frac{1}{1+\tau\pi^2},$$

with $\lambda_1 = \frac{4}{h^2}\sin^2(\pi\frac{h}{2}) = \pi^2 + O(h^2)$.

New methods for time integration can be derived from linearly combination of implicit and explicit Euler. This method is called the θ-method:

$$\frac{u_h^{k+1}-u_h^k}{\tau} = \theta A_h u_h^{k+1} + (1-\theta)A_h u_h^k + f_h(t_k). \tag{7.12}$$

The most common method is the *Crank-Nicolson method*, where $\theta = \frac{1}{2}$. The Crank-Nicolson method is unconditionally stable and has a convergence rate of $O(k^2+h^2)$ using central differences in space. For $\theta = 0$ the methods give the explicit Euler method, for $\theta = 1$ the implicit Euler method.

An example of an explicit scheme second order accurate in time is the *Dufort Frankel* method. In 1-dimension this is given by

$$\begin{aligned} \frac{u_i^{k+1}-u_i^{k-1}}{2\tau} &= \frac{1}{h^2}\left(u_{i-1}^k - 2\overline{u}_i + u_{i+1}^k\right) \quad \text{with} \\ \overline{u}_i &:= \frac{1}{2}\left(u_i^{k-1} + u_i^{k+1}\right). \end{aligned} \tag{7.13}$$

This method is stable as long as $\frac{\tau}{h} \to 0$, and second order in space and time (as long as $\tau = ch^2$) by using two grids at time k and $k-1$ to calculate the new time step $k+1$.

Summary of finite difference methods for the 1-dimensional heat equation:

Explicit Euler	$u_i^{k+1} = u_i^k + \frac{\tau}{h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right)$	$O(\tau)+O(h^2)$
Implicit Euler	$u_i^{k+1} = u_i^k + \frac{\tau}{h^2}\left(u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)$	$O(\tau)+O(h^2)$
Crank/Nicolson	$u_i^{k+1} = u_i^k + \frac{\tau}{2h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right)$ $+\frac{\tau}{2h^2}\left(u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)$	$O(\tau^2)+O(h^2)$
Dufort Frankel	$u_i^{k+1} = \frac{2\tau}{h^2+2\tau}\left(u_{i-1}^k + u_{i+1}^k\right) + \frac{h^2-2\tau}{h^2+2\tau}u_i^{k-1}$	$O(\tau^2)+O(h^2)+O(\frac{\tau^2}{h^2})$

7.3 Hyperbolic Differential Equations

The simplest example of a hyperbolic equation is the *linear transport equation* in one dimension:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0. \tag{7.14}$$

Its analytic solution is given by

$$u^*(x,t) = u_0(x - at), \tag{7.15}$$

where $u_0(x)$ is the initial data at time $t = 0$. Therefore, initial data is transported with velocity a from left to right. The equation can be used for stability analysis of numerical schemes.

Other examples of hyperbolic equations are:

- *Linear systems of scalar equations* are given by

  $$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0,$$

  where $u := u(x,t) \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n\times n}$ diagonalizable. Note that such systems of equations can be solved analytically by decoupling: Insert $A = X\Lambda X^{-1}$, where Λ is a diagonal matrix. Multiplication with X^{-1} from the left yields the diagonalized system of equations that can be solved for $v := A^{-1}u$.

- The *wave equation* in one dimension is

  $$\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = 0,$$

  where c is the finite speed of wave propagation. The wave equation can be transformed into a system of linear equations by introducing $v := \frac{\partial u}{\partial t}$, yielding

  $$\begin{aligned}\frac{\partial v}{\partial t} + c\frac{\partial u}{\partial x} &= 0.\\ \frac{\partial u}{\partial t} + c\frac{\partial v}{\partial x} &= 0.\end{aligned}$$

- The *Euler equations* are an example of a system of non-linear hyperbolic equations
$$\frac{\partial u}{\partial t} + u \cdot \nabla u = -\frac{1}{\rho_0}\nabla p + g$$
$$\nabla \cdot u = 0,$$
where $u \in R^n$ represents the flow velocity, $g \in R^n$ the gravitational acceleration, p the pressure, and ρ_0 the density.

Three frequently used explicit finite difference methods for the 1-dimensional linear transport (Eqn 7.14) are given by

$$\begin{array}{lll}
\text{Lax Friedrich} & u_i^{k+1} = \frac{1}{2}(u_{i+1}^k + u_{i-1}^k) - a\frac{\tau}{2h}(u_{i+1}^k - u_{i-1}^k) & O(\tau) + O(h^2) \\
\\
\text{Euler Upwind} & u_i^{k+1} = u_i^k - a\frac{\tau}{h}(u_i^k - u_{i-1}^k) \text{ for } a > 0 & O(\tau) + O(h) \\
 & u_i^{k+1} = u_i^k - a\frac{\tau}{h}(u_{i+1}^k - u_i^k) \text{ for } a < 0 & \\
\\
\text{Lax Wendroff} & u_{i+1/2}^{n+1/2} = \frac{1}{2}(u_{i+1}^n + u_i^n) - a\frac{\tau}{2h}(u_{i+1}^n - u_i^n) & O(\tau^2) + O(h^2) \\
 & u_{i-1/2}^{n+1/2} = \frac{1}{2}(u_i^n + u_{i-1}^n) - a\frac{\tau}{2h}(u_i^n - u_{i-1}^n) & \\
 & u_i^{n+1} = u_i^n - a\frac{\tau}{h}\left(u_{i+1/2}^{n+1/2} - u_{i-1/2}^{n+1/2}\right) &
\end{array}$$

In the following we will analyse the stability of the Lax Friedrich and Euler upwind scheme. In order for any explicit scheme to work, the Courant-FriedrichsLewy (CFL) condition must be satisfied. The CFL condition in one dimension is given by
$$\left|\frac{a\tau}{h}\right| < 1, \tag{7.16}$$
where a is the velocity from Eqn (7.14). The CFL condition is a necessary condition, ensuring that the range of dependence is large enough in the numerical scheme.

Theorem 7.11. *Lax Friedrich method is stable if the CFL condition is satisfied.*

Proof. We proof that $\|u^{k+1}\|_1 \leq \|u^k\|_1$, where $\|\cdot\|_1$ is the discrete L_1 norm, i.e. $\|u^k\|_1 = h\sum_i |u_i|$:
The Lax Friedrich scheme is given by
$$u_i^{k+1} = \frac{1}{2}(u_{i+1}^k + u_{i-1}^k) - a\frac{\tau}{2h}(u_{i+1}^k - u_{i-1}^k)$$

Taking the absolute value, multiplying by h, and summation over all i yields

$$\begin{aligned}\|u^{k+1}\|_1 &= h\sum_i |\frac{1}{2}(u^k_{i+1}+u^k_{i-1}) - a\frac{\tau}{2h}(u^k_{i+1}-u^k_{i-1})| \\ &\leq \frac{h}{2}\left(\sum_i |(1-\frac{a\tau}{h})u^k_{i+1}| + \sum_i |(1+\frac{a\tau}{h})u^k_{i-1}|\right)\end{aligned}$$

If the CFL condition holds $1-\frac{a\tau}{h}\geq 0$ and $1+\frac{a\tau}{h}\geq 0$, therefore

$$\begin{aligned}\|u^{k+1}\|_1 &\leq \frac{h}{2}\left((1-\frac{a\tau}{h})\sum_i |u^k_{i+1}| + (1+\frac{a\tau}{h})\sum_i |u^k_{i-1}|\right) \\ &= \frac{1}{2}\left((1-\frac{a\tau}{h})\|u^k\|_1 + (1+\frac{a\tau}{h})\|u^k\|_1\right) = \|u^k\|_1\end{aligned}$$

□

Theorem 7.12. *The Euler upwind scheme is stable if the CFL condition is satisfied.*

Proof. Again, we proof that $\|u^{k+1}\|_1 \leq \|u^k\|_1$:
For $a > 0$ we obtain

$$\begin{aligned}\|u^{k+1}\|_1 &= h\sum_i |u^k_i - a\frac{\tau}{h}(u^k_i - u^k_{i-1})| \\ &= h\sum_i |(1-a\frac{\tau}{h})u^k_i + a\frac{\tau}{h}u^k_{i-1}|\end{aligned}$$

and with the CFL condition

$$\begin{aligned}\|u^{k+1}\|_1 &\leq h\left((1-a\frac{\tau}{h})\sum_i |u^k_i| + a\frac{\tau}{h}\sum_i |u^k_{i-1}|\right) \\ &= (1-a\frac{\tau}{h})\|u^k\|_1 + a\frac{\tau}{h}\|u^k\|_1 = \|u^k\|_1\end{aligned}$$

The proof for $a < 0$ follows analogously. □

Chapter 8

Finite Element Method

In this chapter we are considering *finite element methods* (FEM) for the solution of elliptic differential equations. We restrict attention to space dimension two for simplicity. The domain Ω, on which we solve the differential equation, has piecewise linear boundary Γ, is bounded and connected.

8.1 Weak solutions

The basis of finite element methods are *weak* solutions. We just give a short sketch of the basics of this theory:

Definition 8.1. *$H^1(\Omega)$ denotes the space of square integrable functions with square integrable derivatives and the inner product*

$$\langle u, v\rangle_{H^1(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega uv \, dx \; .$$

The associated norm is denoted by $\|\cdot\|_{H^1(\Omega)}$ and the semi-norm is denoted by

$$|u|_{H^1(\Omega)} = \int_\Omega |\nabla u|^2 \; dx \; .$$

Functions $u \in H^1(\Omega)$ are not necessarily continuous in Ω, but it is still possible to define boundary values. For us it is not really important how they can be defined rigorously, we just do as if they point evaluations can be done. Of particular importance are the set of zero-Dirichlet data:

$$H_0^1(\Omega) = \left\{ u \in H^1(\Omega) : u|_\Gamma = 0 \right\} \; ,$$

which is a closed linear subspace of $H^1(\Omega)$. For functions in $H_0^1(\Omega)$ the Poincare-Friedrich inequality is valid:

$$\gamma_\Omega \left\|u\right\|_{H^1(\Omega)} \leq \left|u\right|_{H^1(\Omega)} , \qquad \forall u \in H_0^1(\Omega) . \tag{8.1}$$

After this clarification of notation we are investigating now elliptic differential equations first:

$$L[u] := -\nabla \cdot (\sigma \nabla u) + cu = f \text{ in } \Omega \tag{8.2}$$

with Dirichlet boundary conditions

$$u = 0 \text{ on } \Gamma . \tag{8.3}$$

Aside from some smoothness conditions (which we do not discuss in detail) essential conditions are the following:

$$0 < \sigma_0 \leq \sigma(x) \leq \sigma_\infty \text{ and } 0 \leq c(x) \leq c_\infty .$$

This conditions are essential for ellipticity. A classical solution (referring to standard theory) is one, where the second derivative is continuous.

The basics of weak solutions are partial integration:

$$\begin{aligned}\int_\Omega fv\,dx \underbrace{=}_{(8.2)} -\int_\Omega \nabla \cdot (\sigma \nabla u) v\,dx + \int_\Omega cuv\,dx \\ = \int_\Omega \sigma \nabla u \nabla v\,dx + \int_\Omega cuv\,dx - \int_\Gamma v\sigma \frac{\partial u}{\partial n}\,ds .\end{aligned}$$

Definition 8.2. *A weak solution of the* homogenous Dirichlet-problem*, that is of* (8.2) *and* (8.3)*, is a solution of*

$$\int_\Omega fv\,dx = \int_\Omega \sigma \nabla u \nabla v\,dx + \int_\Omega cuv\,dx , \qquad \forall v \in H_0^1(\Omega) . \tag{8.4}$$

Remark 8.3. *The weak solution is unique.*

The *inhomogenous Dirichlet problem* consists in solving (8.2) together with boundary conditions:

$$u = g \text{ on } \Gamma . \tag{8.5}$$

We extend the function g from Γ to Ω and denote such an extension by u_0. With u_0 we reduce (8.2), (8.5) to a homogenous Dirichlet problem. In fact $w := u - u_0$ solves the homogenous Dirichlet problem

$$L[w] = f - L[u_0] , \qquad w|_\Gamma = 0 .$$

The (in-)homogenous *Dirichlet problem* has a unique solution:

The *Neumann problem* consists in the solution of (8.2) with boundary conditions:

$$\sigma \frac{\partial u}{\partial n} = g \text{ on } \Gamma \, . \tag{8.6}$$

The weak form of the equation is again derived by partial integration:

$$\begin{aligned} \int_\Omega f v \, dx \underbrace{=}_{(8.2)} & \int_\Omega \nabla \cdot (\sigma \nabla u) v \, dx + \int_\Omega cuv \, dx \\ = & \int_\Omega \sigma \nabla u \nabla v \, dx + \int_\Omega cuv \, dx - \int_\Gamma v \sigma \frac{\partial u}{\partial n} \, ds \\ \underbrace{=}_{(8.6)} & \int_\Omega \sigma \nabla u \nabla v \, dx + \int_\Omega cuv \, dx - \int_\Gamma vg \, ds \, . \end{aligned}$$

Now, we consider a general strategy for solving elliptic differential equations: Let

$$\begin{aligned} a(u,v) &:= \int_\Omega \sigma \nabla u \cdot \nabla v \, dx + \int_\Omega cuv \, dx \, , \\ l(v) &:= \int_\Omega f v \, dx \, . \end{aligned} \tag{8.7}$$

a is called *bilinear* form because it is linear in every component on $V = H^1(\Omega)$. Moreover, l is a linear operator on V. With this notation we have a compact formulation of the differential equation (8.2), (8.3):

$$a(u,v) = l(v) \, , \qquad \forall v \in H_0^1(\Omega) \, . \tag{8.8}$$

Note, that the space $H_0^1(\Omega)$ is designed such that the solution satisfies homogenous Dirichlet conditions.

8.2 Galerkin approach

To determine an approximate solution we use a *Galerkin*-approach: We select a finite dimensional subspace $V_h \subseteq H_0^1(\Omega)$ and determine $u_h \in V_h$ satisfying

$$a(u_h, v_h) = l(v_h) \, , \qquad \forall v_h \in V_h \, . \tag{8.9}$$

If $\{\phi_1, \dots, \phi_n\}$ is a basis of V_h, then the ansatz

$$u_h = \sum_{i=1}^n u_i \phi_i$$

results in the linear equation

$$\begin{aligned} A\vec{u}_h = \vec{b} \text{ with } \vec{u}_h = (u_1, \dots, u_n)^T, \\ A = [a(\phi_i, \phi_j)]_{ij} \in \mathbb{R}^{n\times n}, \vec{b} = [l(\phi_j)]_j \in \mathbb{R}^n \ . \end{aligned} \tag{8.10}$$

The matrix A is called *stiffness matrix*.

Example 8.4. *We solve the following Dirichlet-problem by a Galerkin approach:*

$$-u'' = f \ \textit{in} \ (0,1) \ \textit{with} \ u(0) = u(1) = 0 \ .$$

Let V_h be the space of linear splines on a equidistant grid:

$$\Delta_h = \{x_i = ih : 0 \le i \le n, h = 1/n\}$$

satisfying homogenous boundary data. The space of linear splines consists of linear combinations of hat-functions Λ_i, $i = 1, \dots, n-1$, which are equal to 1 at the nodal points i/n. That gives the stiffness matrix:

$$a(\Lambda_i, \Lambda_j) = \int_0^1 \Lambda_i'(x)\Lambda_j'(x)\,dx = \begin{cases} 2/h & i = j \\ -1/h & |i-j| = 1 \\ 0 & \textit{else} \end{cases}$$

In this case the Galerkin method requires to solve the equation:

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \vdots & \\ & \vdots & \vdots & -1 \\ & & -1 & 2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{pmatrix},$$

where

$$b_i = \int_0^1 f(x)\Lambda_i(x)\,dx\,, \qquad i = 1, \dots, n-1 \ .$$

8.3 Triangulations

For realizing a Galerkin method we need an appropriate *ansatz space* $V_h \subseteq H^1(\Omega)$. Typically finite Element methods are based on *triangulations* of the domain Ω.

Definition 8.5. *A set of open triangles $\Gamma = \{T_1, \ldots, T_m\}$ is called* regular triangulation *of Ω, if*

1. $T_i \cap T_j = \emptyset \quad \forall i \neq j$,
2. $\bigcup_{i=1}^m \overline{T_i} = \overline{\Omega}$,
3. *for $i \neq j$ we have either*
   - *(a)* $\overline{T_i} \cap \overline{T_j} = \emptyset$,
   - *(b)* $\overline{T_i} \cap \overline{T_j}$ *is a joint corner of T_i and T_j, or*
   - *(c) a common edge.*

The corners of the triangle are called corners.

On a triangulation we define linear ansatz functions (in analogy to linear splines).

Theorem 8.6. *Let Γ be a regular triangulation of a polygonal domainΩ with nodal points x_i, $i = 1, \ldots, n$. Then there exist continuous functions $\Lambda_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$, satisfying:*

1. $\Lambda_i(x_j) = \delta_{ij}\,, \quad i, j = 1, \ldots, n\,,$
2. $\Lambda_i(x) = \beta_{ik} + \alpha_{ik} \cdot x$ *for $x \in T_k$ with $\alpha_{ik} \in \mathbb{R}^2$, $\beta_{ik} \in \mathbb{R}$.*

$V^\Gamma = span\{\Lambda_1, \ldots, \Lambda_n\}$ consists of piecewise linear functions with respect to Γ.

The gradient of an element V^Γ is piecewise constant and we have $V^\Gamma \subseteq H^1(\Omega)$.

Definition 8.7. *The tupel (Γ, V^Γ) is called* finite elements.

The analog of Lagrange interpolation for finite elements reads as follows:

Theorem 8.8. *Let Γ be a regular triangulation of $\Omega \subseteq \mathbb{R}^2$ with nodal points $\{x_i : i = 1, \ldots, n\}$. Let be given $\{y_i : i = 1, \ldots, n\}$. Then $\psi = \sum_{i=1}^n y_i \Lambda_i \in V^\Gamma$ and*

$$\psi(x_i) = y_i\,, \quad i = 1, \ldots, n\,.$$

8.4 Stiffness Matrix

A finite element method is reduced to the solution of the linear matrix equation

$$A\vec{u}_h = b\,, \tag{8.11}$$

with *stiffness matrix* A, where

$$a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \sigma\nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx\ . \tag{8.12}$$

We empghasize that the matrix is sparse.

For every triangle $T_k \in \Gamma$ the matrix

$$S_k = \left[\int_{T_k} \sigma\nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx\right]_{ij} \in \mathbb{R}^{n\times n} \tag{8.13}$$

consists of all integrals over the triangle T_k. These matrices S_k are called *element stiffness matrices.* Because of

$$\begin{aligned} a(\phi_i, \phi_j) &= \int_\Omega \sigma\nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx \\ &= \sum_{k=1}^{m} \int_{T_k} \sigma\nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx \end{aligned}$$

we have

$$A = \sum_{k=1}^{m} S_k\ . \tag{8.14}$$

To determine the element stiffness matrices one uses the transformation

$$\Phi(s,t) = x_1 + s(x_2 - x_1) + t(x_3 - x_1)\,, \tag{8.15}$$

which maps the *reference triangle*

$$D = \left\{z = (s,t)^t : s > 0, t > 0, s + t < 1\right\} \tag{8.16}$$

onto the triangle $T \in \Gamma$ with corners $x_i = (\zeta_i, \eta_i)^t$, $i = 1, 2, 3$. Therefore, we have

$$\Phi'(s,t) = [x_2 - x_1 \quad x_3 - x_1] = \left[\begin{array}{cc} \zeta_2 - \zeta_1 & \zeta_3 - \zeta_1 \\ \eta_2 - \eta_1 & \eta_3 - \eta_1 \end{array}\right]\ .$$

These two vectors are linear independent in a triangle which is not degenerate. Thus

$$d = \det\Phi' = (\zeta_2 - \zeta_1)(\eta_3 - \eta_1) - (\zeta_3 - \zeta_1)(\eta_2 - \eta_1) \neq 0 \, .$$

Thus

$$\Phi'^{-1}(x) = \frac{1}{d} \left[\begin{array}{cc} \eta_3 - \eta_1 & \zeta_1 - \zeta_3 \\ \eta_1 - \eta_2 & \zeta_2 - \zeta_1 \end{array} \right] (x) \, .$$

Example 8.9. *We are calculating the element stiffness matrix $S = [s_{ij}]$ for $L[u] = -\Delta u$ and a triangle $T \in \Gamma$ with corners x_1, x_2 and x_3. We denote by Λ_i, $i = 1, 2, 3$ the hut functions, with nodal value 1 at x_i and 0 else, respectively. Therefore,*

$$\begin{aligned} s_{ij} &= \int_T \nabla_x \Lambda_i(x) \cdot \nabla_x \Lambda_j(x) \, dx \\ &= \int_D \nabla_x \Lambda_i(\Phi(z)) \cdot \nabla_x \Lambda_j(\Phi(z)) \, |det\Phi'| \, dz \\ &= |d| \int_D \Phi'^{-t} \nabla_z \Lambda_i(\Phi(z))) \cdot (\Phi'^{-t} \nabla_z \Lambda_j(\Phi(z)) \, dz \, . \end{aligned}$$

The function $\Lambda_i(\Phi(\cdot))$ is again an hat function over D with nodal value 1 at z_i and 0 else. Therefore,

$$G := \left[\begin{array}{c} \nabla_z(\Lambda_1(\Phi(\cdot))^t \\ \nabla_z(\Lambda_2(\Phi(\cdot))^t \\ \nabla_z(\Lambda_3(\Phi(\cdot))^t \end{array} \right] = \left[\begin{array}{cc} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{array} \right] .$$

Therefeore, the integrands s_{ij} are constant on D. Moreover, the area of D is 0.5. Consequently,

$$\begin{aligned} [s_{ij}]_{i,j} &= \frac{|d|}{2} G \Phi'^{-1} \Phi'^{-t} G^t \\ &= \frac{1}{2\,|d|} \left[\begin{array}{cc} \eta_2 - \eta_3 & \zeta_3 - \zeta_2 \\ \eta_3 - \eta_1 & \zeta_1 - \zeta_3 \\ \eta_1 - \eta_2 & \zeta_2 - \zeta_1 \end{array} \right] \left[\begin{array}{ccc} \eta_2 - \eta_3 & \eta_3 - \eta_1 & \eta_1 - \eta_2 \\ \zeta_3 - \zeta_2 & \zeta_1 - \zeta_3 & \zeta_2 - \zeta_1 \end{array} \right] \in \mathbb{R}^{3\times 3} \end{aligned}$$

References which were used to prepare this notes are [3, 1, 2].

Bibliography

[1] P. Deuflhard and A. Hohmann. *Numerische Mathematik I. Eine algorithmisch orientierte Einführung*. De Gruyter, Berlin, 1993. 2., überarb. Aufl.

[2] G. H. Golub and Ch. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1996. 3.

[3] M. Hanke. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Teubner, Stuttgart, Leipzig, Wiesbaden, 2002.

[4] G. Strang. Wavelet transforms versus Fourier transforms. *Bull. Amer. Math. Soc.*, 28(2):288–305, 1993.