Mat-1.3652 Finite difference methods

0. Practical issues

The course will be lectured during Period I. The preliminary timetable is as follows:

Week 1: Introduction and motivation

Week 2-3: Linear multi-step methods

Week 4: Runge-Kutta methods

Week 5: Parabolic PDEs

Week 6: Hyperbolic PDEs.

These hand-written lecture notes should be almost all-inclusive. Other useful reading includes

- D.F. Griffiths and D.J. Higham, Numerical methods for ordinary differential equations, Springer. (weeks 1-4)
- T. Eirola and O. Nevanlinna, Discretizing differential equations, (available in Noppa, weeks 5-6).

Two thirds of the grade is based on a stream-lined exam (exact timing to-be-decided) and one third on weekly exercises: about 4 per week, graded 0-3, returned to the assistant (at the latest) at the beginning of the exercise session. (First exercise session on September 19.)

1. Introduction: Initial value problems

For most of the course, we consider the initial value problem

$$x'(t) = f(t, x(t)), \quad t > 0, \qquad (1.1)$$
$$x(0) = x_0.$$

Here t can be interpreted as time, $x(t) \in \mathbb{R}^n$ describes the state of the investigated system at time t, the initial value $x_0 \in \mathbb{R}^n$ determines the initial state of the system, and $f : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ gives the dependence of the change in the system on the time and the current state.

1.1 Existence and uniqueness

Although the general approach of this course is numerical, it is good to know under which conditions (1.1) has a unique solution (otherwise we would not know if the numerical solution makes any sense).

Example. Let $n=1$,

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

and $x_0 = 0$. Then, (1.1) does not have a solution.

"Hand-waving proof": The solution curve $x(t)$ cannot move away from the initial value $x(0) = x_0 = 0$ due to the sign of its derivative, but obviously $x(t) \equiv 0$ is not a solution.

1/3

This example motivates the following assumption, which will be upheld throughout this course:

Assumption 1. The function on the right-hand side of (1.1),

$$f : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

is assumed to be *continuous*. (Sometimes the domain of definition for f may vary, though.)

Under this assumption, we have local existence.

Theorem 1.1 (Peano theorem)
The initial value problem (1.1) has *a* solution for $t \in [0, T)$ for some $T > 0$.

The proof is omitted.

Notice that Theorem 1.1 is not totally satisfactory as it does not state anything about uniqueness. This problem cannot be circumvented without further assumptions.

Example. Choose $n=1$, $f(t,x) = \mathrm{sgn}(x)|x|^{1/2}$ and $x_0 = 0$, meaning that $f : \mathbb{R} \to \mathbb{R}$ is continuous but not differentiable at the origin (the slope is vertical). It is "easy" to check that (1.1) has in this case a family of solutions

$$x(t) = \begin{cases} 0, & t \leq t_0 \\ \frac{1}{4}(t-t_0)^2, & t \geq t_0, \end{cases}$$

parametrized by $t_0 \geq 0$.

It turns out that "restricting the slope of f with respect to x" is the natural way to achieve uniqueness for (1.1). This does not require differentiability with respect to x, but a slightly weaker *Lipschitz condition* is sufficient.

Assumption 2. The function on the right-hand side of (1.1),

$$f : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

satisfies

Euclidean norm

$$|f(t,x) - f(t,y)| \leqslant L|x-y|$$

for some $L > 0$ and all $t \geqslant 0$, $x, y \in \mathbb{R}^n$.

(Once again, the requirement "for all $t \geqslant 0$ and $x, y \in \mathbb{R}^n$" could be weakened.)

Theorem 1.2. Under Assumptions 1 and 2, the initial value problem (1.1) has a unique solution $x : \mathbb{R}_+ \to \mathbb{R}^n$.

The proof is divided into three parts.

(i) By integrating (1.1), it follows that

$$x(t) = x_0 + \int_0^t f(s, x(s))\, ds \qquad (1.2)$$

is an equivalent formulation for (1.1) (assuming that $x : \mathbb{R}_+ \to \mathbb{R}^n$ is continuosly differentiable).

(ii) Gronwall inequality:

Lemma 1.3. Assume that $u : [0,T] \to \mathbb{R}$ is a continuous nonnegative function such that

$$u(t) \leq \grave{C} + K\int_0^t u(s)\,ds$$

for some $\grave{C}, K \geq 0$ and all $t \in [0,T]$. Then,

$$u(t) \leq \grave{C} e^{Kt}$$

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

Proof. Let us first assume that $\grave{C} > 0$ and denote $v(t) = \grave{C} + K\int_0^t u(s)\,ds$; in particular,

$$u(t) \leq v(t) > 0 \qquad \text{for all } t \in [0,T].$$

Thus,

$$\frac{d}{dt}\ln(v(t)) = \frac{v'(t)}{v(t)} = \frac{Ku(t)}{v(t)} \leq K, \quad t \in [0,T].$$

Integrating this inequality, we get

$$\ln(v(t)) \leq \ln(\overbrace{v(0)}^{=\grave{C}}) + Kt$$

$$\Rightarrow u(t) \leq v(t) \leq \grave{C} e^{Kt}, \qquad \forall\, t \in [0,T].$$

The case $\grave{C} = 0$ can be handled by noting that then the previous inequality holds with any $\varepsilon > 0$ replacing $\grave{C}$. ☺

(iii) Since Theorem 1.1 establishes existence, the remaining unsettled issue is uniqueness (actually, this is not quite precise, see Remark 1.4 below). Suppose that $x, \tilde{x} : \mathbb{R}_+ \longrightarrow \mathbb{R}^n$ are both solutions to (1.1), and thus also satisfy (1.2). Hence,

$$\begin{aligned} u(t) &:= |x(t) - \tilde{x}(t)| \\ &= \left| \int_0^t \big(f(s,x(s)) - f(s,\tilde{x}(s)) \big)\, ds \right| \\ &\leq \int_0^t |f(s,x(s)) - f(s,\tilde{x}(s))|\, ds \\ \text{(Assumption 2)} \quad &\leq L \int_0^t |x(s) - \tilde{x}(s)|\, ds = L \int_0^t u(s)\, ds, \end{aligned}$$

for all $t \geq 0$, meaning that

$$u(t) = |x(t) - \tilde{x}(t)| = 0$$

by the Gronwall inequality (with $C = 0$). This completes the proof of Theorem 1.2.

☺

Remark 1.4. To be precise, Theorem 1.1 only establishes the local existence, i.e. existence on some finite time interval, while in the proof of Theorem 1.2 we used the existence of a solution for all $t > 0$. This stronger result can be proved, e.g., by considering the Picard-Lindelöf iteration

$$\begin{cases} x^{k+1}(t) = x_0 + \int_0^t f(s, x^{(k)}(s))\,ds, & t > 0, \\ x^0(t) \equiv x_0, & t > 0, \end{cases}$$

which can be shown to converge to a global solution of (1.1) under Assumptions 1 and 2. (See, e.g., pages 7/6 - 7/12 of the lecture notes of L4.)

1/10

1.2 Continuous dependence on the initial data

When analyzing the convergence of numerical methods for solving (1.1), it is essential to know how inaccuracies in the initial value $x_0 \in \mathbb{R}^n$ propagate as time proceeds. To simplify the notation, we make a new definition.

Definition 1.5. The solution map $\psi : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ associated to the system

$$x'(t) = f(t, x(t)) \qquad (1.3)$$

is defined via

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

where $x : [\tau, \infty) \to \mathbb{R}^n$ is the solution of (1.3) with the initial condition $x(\tau) = u \in \mathbb{R}^n$.

In other words, ψ characterizes the dependence of the solution to (1.3) on (i) the time, (ii) the initial time, and (iii) the initial value.

Assuming that ψ is well defined, i.e., (1.3) has a unique solution for the considered initial conditions, it holds that (an exercise)

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

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

$$\psi(t,s,\psi(s,\tau,u)) = \psi(t,\tau,u).$$

The main result of this section is as follows:

<u>Theorem 1.6.</u> Under Assumptions 1 and 2, it holds that

$$|\psi(t,t_0,x_0)-\psi(t,t_0,\tilde{x}_0)| \leq e^{L(t-t_0)}|x_0-\tilde{x}_0|,$$

for all $t \geq t_0$.

<u>Proof.</u> For simplicity we only consider the case $t_0 = 0$; the general result can be deduced, e.g., with the help of the change of variables $\tau = t - t_0$.

Let $x := \psi(\cdot, 0, x_0)$ and $\tilde{x} := \psi(\cdot, 0, \tilde{x}_0)$ be the solutions of (1.1) with the initial values x_0 and $\tilde{x}_0$, respectively. Define

$$u(t) = |\psi(t,0,x_0) - \psi(t,0,\tilde{x}_0)|^2 = |x(t)-\tilde{x}(t)|^2$$

for $t \geq 0$. It follows that

$$\begin{aligned} u'(t) &= 2\left(x'(t)-\tilde{x}'(t)\right)\cdot\left(x(t)-\tilde{x}(t)\right) \\ &= 2\left(f(t,x(t))-f(t,\tilde{x}(t))\right)\cdot\left(x(t)-\tilde{x}(t)\right). \end{aligned}$$

Due to Assumption 2, we get

$$u'(t) \leqslant 2L\,|x(t)-\tilde{x}(t)|^2 = 2L\,u(t),$$

which can be integrated to obtain

$$u(t) \leqslant \overbrace{u(0)}^{=|x_0-\tilde{x}_0|^2} + 2L\int_0^t u(s)\,ds.$$

Now, the Gronwall inequality gives

$$\underbrace{|\psi(t,0,x_0)-\psi(t,0,\tilde{x}_0)|^2}_{u(t)} \leqslant |x_0-\tilde{x}_0|^2\, e^{2Lt},$$

which completes the proof. ☺

1.3 Higher order systems

During this course we will only/mainly consider initial value problems of the first order.

This is because an mth order system

$$\begin{cases} x^{(m)}(t) = f(t, \overset{x(t)}{\overset{\shortparallel}{x^{(0)}(t)}}, \dots, x^{(m-1)}(t)), & t > 0, \\ x^{(0)}(0) = x_0^{(0)}, \dots, x^{(m-1)}(0) = x_0^{(m-1)}, \end{cases}$$

can always be represented in the form of a first order system

$$\begin{cases} y'(t) = g(t, y(t)), & t > 0, \\ y(0) = y_0, \end{cases}$$

where

$$y(t) = \begin{bmatrix} x^{(0)}(t) \\ \vdots \\ x^{(m-1)}(t) \end{bmatrix}, \quad g(t, y(t)) = \begin{bmatrix} x^{(1)}(t) \\ \vdots \\ x^{(m-1)}(t) \\ f(t, x^{(0)}(t), \dots, x^{(m-1)}(t)) \end{bmatrix};$$

and y_0 defined accordingly.

2. Euler's method

Let us continue to consider the initial value problem

$$\begin{aligned} x'(t) &= f(t, x(t)), \qquad t > 0, \\ x(0) &= x_0. \end{aligned} \tag{2.1}$$

In this section, we will introduce the simplest numerical scheme for solving (2.1), the Euler's method. Analysis of its properties will be used as a motivotion for introducing more complicated numerical methods over the rest of this course.

[For the rest of this section (and the rest of this course if not stated otherwise), we assume that Assumptions 1 and 2 hold.]

Assume that the unique solution of (2.1) is twice continuously differentiable*, and consider the Taylor expansion of x around $t \geq 0$,

$$x(t+h) = x(t) + h \, \overset{f(t,x(t))}{\overset{\|}{x'(t)}} + R_1(t),$$

where the remainder, or the local truncation error (LTE) can be given in the Lagrange form

$$R_1(t) = \frac{1}{2} h^2 x''(\xi),$$

for some $\xi = \xi(t,h) \in [t, t+h]$.

In consequence,

$$x(t+h) = x(t) + h f(t, x(t)) + O(h^2), \quad (2.2)$$

uniformly for all $t \in [0,T]$, with $T < \infty$ fixed.

* As f is continuous, we already know that x is continuously differentiable. Higher regularity of x follows by assuming more smoothness from f.

The idea of the Euler's method is to drop $O(h^2)$ out of (2.2) and proceed iteratively; this corresponds to taking a (short) step in the direction of the tangent of the solution curve (passing through the current iterate).

After defining the grid points

$$t_j = jh, \quad j = 0, 1, 2, \dots,$$

where $h > 0$ is the time step, *the Euler's method* reads

$$x_{j+1} = x_j + h f(t_j, x_j), \quad j = 0, 1, 2, \dots .$$

The leading idea (or hope) is that x_j gives a good estimate for $x(t_j)$, the exact solution at the time $t = t_j$.

Remark 2.1. To simplify the notation, we will occasionally write $f_j := f(t_j, x_j)$ and $y_j := x(t_j)$.

Based on the above material, we know that the LTE of the Euler's method is of the order $O(h^2)$. In other words, if the iterate x_j is known to be exact, i.e. $x_j = y_j$ $(= x(t_j))$, then

$$x_{j+1} = y_{j+1} + O(h^2). \qquad (2.3)$$

This sounds of course good, but *does not*, unfortunately, result in a *global error* (GE)

$$e_m = x(t_m) - x_m$$

of the order $O(h^2)$ if $t_m = T$ is fixed (and m and $h > 0$ chosen so that $t_m = mh = T$).

The intuitive explanation for this "loss of global order" is the following: In order to decrease the local truncation error, the step size $h > 0$ must be decreased, which means that more steps must be taken to reach the fixed time $T > 0$ ($m \simeq T/h$). Thus, it seems intuitive that

$$(2.4) \qquad e_m = \underset{\substack{\uparrow \\ \text{number of needed steps}}}{\frac{T}{h}} \underbrace{O(h^2)}_{\text{LTE}} = O(h), \qquad t_m = T.$$

Unfortunately, the situation is not quite this simple: The formula (2.3) was derived assuming that the previous iterate x_j is exact; when the Euler's method is applied to (2.1), this obviously holds only for $j = 0$.

For one-step methods*, such as the Euler's method, the "logic" (2.4) turns anyway out to be valid. This is due to the continuous dependence on the initial value, i.e. Theorem 1.6. Anticipating future developments, we prove a slightly more general result than necessary. [For Euler $p=1$.]

Theorem 2.2. Assume that the considered one-step method satisfies

LTE with the help of the solution map from Definition 1.5.

$$|x_{j+1} - \psi(t_{j+1}, t_j, x_j)| \le C h^{p+1}, \quad C > 0.$$

Then

GE at t_j

$$|x_j - \psi(t_j, 0, x_0)| \le \frac{C}{L} h^p (e^{LT} - 1) \qquad (2.5)$$

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

Lipschitz constant, Assumpt. 2.

* "One-step method" means that x_{j+1} is defined with the help of x_j only. Multi-step methods use also earlier iterates.

<u>Proof.</u> It follows from the properties of the solution map that

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

Hence,

$$\begin{aligned}
\varepsilon_j := |x_j - \psi(t_j, 0, x_0)| &\le |x_j - \psi(t_j, t_{j-1}, x_{j-1})| \\
&\quad + |\psi(t_j, t_{j-1}, x_{j-1}) - \psi(t_j, t_{j-1}, \psi(t_{j-1}, 0, x_0))| \\
&\le \underbrace{\hat{C} h^{p+1}}_{\text{LTE}} + \underbrace{e^{L(t_j - t_{j-1})} |x_{j-1} - \psi(t_{j-1}, 0, x_0)|}_{\substack{\text{continuous dependence on} \\ \text{the initial value, Theorem 1.6}}} \\
&= \hat{C} h^{p+1} + e^{Lh} \varepsilon_{j-1}.
\end{aligned}$$

Taking into account that $\varepsilon_0 = |x_0 - x_0| = 0$, it follows recursively that

$$\begin{aligned}
\varepsilon_j &\le \hat{C} h^{p+1} + e^{Lh}(\hat{C} h^{p+1} + e^{Lh} \varepsilon_{j-2}) \\
&= \hat{C} h^{p+1}(1 + e^{Lh}) + e^{2Lh} \varepsilon_{j-2} \\
&\le \hat{C} h^{p+1}(1 + e^{Lh} + e^{2Lh}) + e^{3Lh} \varepsilon_{j-3} \\
&\le \hat{C} h^{p+1} \sum_{k=0}^{j-1} e^{kLh} \qquad \| \text{ geometric series} \\
&= \hat{C} h^{p+1} \frac{1 - (e^{Lh})^j}{1 - e^{Lh}} = \hat{C} h^{p+1} \frac{e^{Lt_j} - 1}{e^{Lh} - 1}.
\end{aligned}$$

2/8

Because $t_j \leq T$ by assumption and

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

the claim follows. ☺

Although the Euler's method seems adequate in the sense that it converges to the exact solution of (2.1) in the sense of (2.5) with $p=1$ — assuming that the exact solution (cf. (2.2)) x is continuously differentiable — there are a number of reasons to study more sophisticated methods as well.

(i) The convergence rate (2.5) is relatively slow when $p=1$; to decrease the error by an order of magnitude, the time step must also be decreased by the same order, resulting in the need to take many steps.

(ii) In practice, one often encounters so-called stiff systems that can be studied (to a certain extent) by considering the model scalar problem

$$x'(t) = \lambda x(t), \quad x(0) = x_0 \neq 0.$$

with $\lambda < 0$. Obviously,

$$x(t) = x_0 e^{\lambda t} \xrightarrow{t \to \infty} 0,$$

and one would like the corresponding numerical solution to exhibit this same general behavior, i.e.,

$$x_j \xrightarrow{j \to \infty} 0, \qquad (2.6)$$

independently of the step size $h > 0$. It is, however, easy to check that (2.6) is satisfied for the Euler's method if and only if

$$h < -\frac{2}{\lambda},$$

which is small if $\lambda << 0$.

2/10

(iii) Many times the studied initial value problem has some invariant, such as the total energy or symplecticness for Hamiltonian systems. In these situations, it is of essence to use some numerical method that preserves the same invariants.

For these (and other) reasons, we study on this course two general classes of methods for solving (2.1): linear multistep methods (LMM) and Runge-Kutta (RK) methods. Curiously, the Euler's method is the simplest member of both of these families.

3. Linear multistep methods

As always, we will consider the initial value problem

$$x'(t) = f(t, x(t)), \quad t > 0, \qquad (3.1)$$
$$x(0) = x_0,$$

but for simplicity we assume that $x : R_+ \to \mathbb{R}$ and $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, i.e. we deal with a scalar problem.*

The leading idea of the linear multistep methods is to look for the iterate $x_{j+k} \approx x((j+k)h)$ as a "solution" of a "difference equation"

No α_k as we can divide by it.

$$x_{j+k} + \alpha_{k-1} x_{j+k-1} + \dots + \alpha_0 x_j = h(\beta_k f_{j+k} + \beta_{k-1} f_{j+k-1} + \dots + \beta_0 f_j), \qquad (3.2)$$

cf. Remark 2.1.

* This is by no means an indispensable assumption; it is just easier to write down higher order derivatives in this setting.

Some remarks are in order:

- If $k=1$, $\alpha_0=-1$, $\beta_0=1$ and $\beta_1=0$, we get the Euler's method.
- If $\beta_k \neq 0$, the method is *implicit*, that is, x_{j+k} appears on both sides of (3.2): as itself on the left and in $f_{j+k} = f(t_{j+k}, x_{j+k})$ on the right. If $\beta_k = 0$ the method is said to be explicit.
- The recursion (3.2) can be used to define x_ℓ only if $\ell \geq k$, which means that the iterates $x_1, \ldots, x_{k-1}$ must be obtained using some other method (e.g., Euler's).

Of course, the coefficients $\alpha_0, \ldots, \alpha_{k-1} \in \mathbb{R}$ and $\beta_0, \ldots, \beta_k \in \mathbb{R}$ should be chosen in some reasonable way in order to get a functional method. We start with two examples: an implicit one-step method and an explicit two-step method.

<u>Example.</u> (Trapedzoidal rule)

Let $z : \mathbb{R} \to \mathbb{R}$ be an arbitrary three times continnously differentiable function. In particular, we have

$$z(t+h) = z(t) + h z'(t) + \frac{1}{2} h^2 z''(t) + O(h^3), \tag{3.3}$$

where the "constant" hidden in the O-term depends on z'''. Naturally, we can also write a Taylor expansion for z', resulting in

$$\begin{aligned} z'(t+h) &= z'(t) + h z''(t) + O(h^2) \\ \Rightarrow \quad z''(t) &= \frac{z'(t+h) - z'(t)}{h} + O(h), \end{aligned} \tag{3.4}$$

where the exact form of $O(h^2)$ once again depends on z'''. Combining the above two formulas, we get

$$z(t+h) = z(t) + \frac{1}{2} h \big(z'(t) + z'(t+h) \big) + O(h^3).$$

Choosing $z = x$, the solution of (3.1), it follows that

$$x(t+h) = x(t) + \frac{1}{2} h \big(f(t, x(t)) + f(t+h, x(t+h)) \big) + O(h^3),$$

assuming that x is smooth enough. The previous formula suggest the numerical method

$\alpha_0 = -1, \beta_0 = \beta_1 = \frac{1}{2}.$

$$(3.5) \quad x_{j+1} - x_j = \frac{1}{2} h (f_{j+1} + f_j), \quad j = 0, 1, \dots .$$

This so-called trapedzoidal rule has LTE of order $O(h^3)^*$; as it is a one-step method — x_{j+1} depends only on the previous iterate x_j if (3.5) is solvable — Theorem 2.2 demonstrates that the associated global error is of the order $O(h^2)$.

Example. (Two-step Adams-Bashforth) Instead of using (3.4) as for the trapedzoidal rule, we can as well write

$$z'(t-h) = z'(t) + (-h) z''(t) + O(h^2)$$

$$\Rightarrow z''(t) = \frac{z'(t) - z'(t-h)}{h} + O(h).$$

* Assuming once again that the solution $x(t)$ is smooth enough.

Plugging this in (3.3), we have

$$z(t+h) = z(t) + \frac{1}{2}h\left(3z'(t) - z'(t-h)\right) + O(h^3).$$

Choosing again $z=x$ and using (3.1), results in the method

$$x_{j+1} - x_j = \frac{1}{2}h\left(3f_j - f_{j-1}\right),$$

or in the "fundamental form"

$$x_{j+2} - x_{j+1} = \frac{1}{2}h\left(3f_{j+1} - f_j\right),$$

for which $\alpha_1 = -1$, $\alpha_0 = 0 = \beta_2$, $\beta_1 = \frac{3}{2}$, $\beta_0 = -\frac{1}{2}$. Notice that, as for the trapedzoidal rule, the LTE of the two-step Adams-Bashforth is of order $O(h^3)$, but Theorem 2.2 cannot be used to deduce GE order $O(h^2)$ because Theorem 2.2 is for one-step methods only. [The global error turns anyway out to be $O(h^2)$; see Section 3.2.]

3.1 Construction and consistency

Let us define the linear difference operator $\mathcal{L}_h : C^1(\mathbb{R}) \longrightarrow C(\mathbb{R})$ associated to the LMM (3.2) by

$$(\mathcal{L}_h z)(t) = z(t+kh) + \alpha_{k-1} z(t+(k-1)h) + \dots + \alpha_0 z(t) - h\left(\beta_k z'(t+kh) + \beta_{k-1} z'(t+(k-1)h) + \dots + \beta_0 z'(t)\right). \tag{3.6}$$

[By linear we mean that

$$(\mathcal{L}_h(\overset{\overset{\mathbb{R}}{\Downarrow}}{a} z + \overset{\overset{\mathbb{R}}{\Downarrow}}{b} w))(t) = a(\mathcal{L}_h z)(t) + b(\mathcal{L}_h w)(t),$$

which is easily checked using the definition of $\mathcal{L}_h$.]

<u>Definition 3.1.</u> The LMM (3.2), or the linear difference operator $\mathcal{L}_h$, is said to be consistent of order p if

$$(\mathcal{L}_h z)(t) = O(h^{p+1}) \tag{3.7}$$

for any smooth function z. The term consistent is used when (3.7) holds for some $p \geq 1$.

Notice that in the previous two examples we demonstrated that the trapedzoidal rule and the two-step Adams-Bashforth method are both consistent of order $p=2$.

Let us figure out what are the conditions on the coefficients $\alpha_0, \dots, \alpha_{k-1}$ and $\beta_0, \dots, \beta_k$ that guarantee that the method (3.2) is consistent:

By linearizing (first order Taylor expansions) all terms on the right-hand side of (3.6)*, we get

$$(\mathcal{L}_h z)(t) = \left(1 + \alpha_{k-1} + \dots + \alpha_0\right) z(t) + \left(kh + \alpha_{k-1}(k-1)h + \dots + \alpha_1 h\right) z'(t) + O(h^2) - h\left((\beta_k + \beta_{k-1} + \dots + \beta_0) z'(t) + O(h)\right)$$

$$= \left(1 + \alpha_{k-1} + \dots + \alpha_0\right) z(t) + h\Big(\left(k + (k-1)\alpha_{k-1} + \dots + \alpha_1\right) - \left(\beta_k + \beta_{k-1} + \dots + \beta_0\right)\Big) + O(h^2). \quad (3.8)$$

* In fact, for the derivatives we only need the "zeroth order" expansions.

In consequence, the LMM (3.2) is consistent (of some order) iff the first two terms on the right-hand side of (3.8) vanish, that is

$$\begin{cases} 1 + \alpha_{k-1} + \ldots + \alpha_0 = 0, \\ k + (k-1)\alpha_{k-1} + \ldots + \alpha_1 = \beta_k + \beta_{k-1} + \ldots + \beta_0 . \end{cases} \tag{3.9}$$

These conditions can be neatly represented with the help of the so-called characteristic polynomials:

Definition 3.2. The first and second characteristic polynomials of the LMM (3.2) are defined to be

$$\rho(r) = r^k + \alpha_{k-1} r^{k-1} + \ldots + \alpha_1 r + \alpha_0 ,$$

$$\sigma(r) = \beta_k r^k + \beta_{k-1} r^{k-1} + \ldots + \beta_1 r + \beta_0 ,$$

respectively.

Lemma 3.3. The LMM (3.2) is consistent if and only if

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1) .$$

Proof. An easy exercise.

Naturally, one should not settle for mere consistence (of order 1): Even for $k=1$, (3.2) contains three free parameters, $\alpha_0, \beta_0, \beta_1$, and so it seems intuitive that, in addition to (3.9), a third condition could probably be satisfied, hopefully resulting in a method that is consistent of order 2. (We have, in fact, already demonstrated that this is possible: the choice $\alpha_0=-1$, $\beta_0=\beta_1=\frac{1}{2}$ results in the trapedzoidal rule). If $k>1$, one can aim at even higher orders of consistence.

A general technique for obtaining methods of higher order is to write higher order Taylor expansions for the terms on the right-hand side of (3.6).

3/10

To be more precise, using (p+1)th order expansion for the terms involving z and pth order expansion for those involving z', we obtain the formula

$$(\mathcal{L}_h z)(t) = C_0 z(t) + C_1 h z'(t) + \dots + C_p h^p z^{(p)}(t)$$
$$+ C_{p+1} h^{p+1} z^{(p+1)} + O(h^{p+2}),$$

where $C_0 = \rho(1)$, $C_1 = \rho'(1) - \sigma(1)$ and also $C_2, \dots, C_{p+1}$ are certain linear combinations of the coefficients $\alpha_0, \dots, \alpha_{k-1}$ and $\beta_0, \dots, \beta_k$. If these are chosen so that

$$C_0 = C_1 = \dots = C_p = 0, \qquad (3.10)$$

then the LMM (3.2) is consistent of order p. The first nonzero coefficient C_{p+1} is called the *error constant*.

Example. Consider the LMM of the form

$$x_{j+2} + \alpha_0 x_j = h(\beta_1 f_{j+1} + \beta_0 f_j); \quad (3.11)$$

note that this is an explicit method and, furthermore, it has been chosen a priori that $\alpha_1 = 0$. Let us choose α_0, β_0 and β_1 so that (3.11) is consistent of the highest possible order.

In this case,

$$(\mathcal{L}_h z)(t) = z(t+2h) + \alpha_0 z(t) - h(\beta_1 z'(t+h) + \beta_0 z'(t)), \quad (3.12)$$

which means that we need Taylor expansions for $z(t+2h)$ and $z'(t+h)$:

$$z(t+2h) = z(t) + 2h z'(t) + \frac{(2h)^2}{2!} z''(t) + \frac{(2h)^3}{3!} z'''(t) + O(h^4),$$

$$z'(t+h) = z'(t) + h z''(t) + \frac{h^2}{2} z'''(t) + O(h^3),$$

(where the number of terms was chosen "cleverly" by taking into account the number of free parameters in (3.11)).

3/12

Plugging these expansions in (3.12), we get

$$(\mathcal{L}_h z)(t) = \overbrace{(1+\alpha_0)}^{C_0} z(t) + \overbrace{(2-\beta_1-\beta_0)}^{C_1} h z'(t) + \underbrace{(2-\beta_1)}_{C_2} h^2 z''(t) + \underbrace{\left(\tfrac{4}{3}-\tfrac{1}{2}\beta_1\right)}_{C_3} h^3 z'''(t) + O(h^4).$$

For the first three terms C_0, C_1 and C_2 to vanish, we require

$$\left\{\begin{array}{l} 1+\alpha_0 = 0 \\ 2-\beta_1-\beta_0 = 0 \\ 2-\beta_1 = 0 \end{array}\right\} \iff \left\{\begin{array}{l} \alpha_0 = -1, \\ \beta_0 = 0, \\ \beta_1 = 2. \end{array}\right.$$

Notice that the first two equations of the above linear system correspond to the consitency conditions (3.9) — as they should. There are not enough parameters to make C_3 vanish, and so it is the error constant $C_3 = \frac{4}{3} - 1 = \frac{1}{3}$.

The resulting scheme (consisten of order 2)

$$x_{j+2} - x_j = h f_{j+1}$$

is known as the "leap-frog" method.

3/13

3.2 Convergence and zero stabilty

Since the k-step LMM (3.2) has $2k+1$ free parameters, it seems plausible that one can choose them so that (3.10) is satisfied with $p = 2k$, which would lead to a method that is consistent of order $2k$ (or $2k-1$ for explicit methods). However, unlike for one-step methods such as the Euler's method, for multistep methods consistency *does not* imply convergence (cf. Theorem 2.2), which ends up restricting the order of *reasonable* LMMs quite heavily. To make sense of this claim, we naturally need to first define what exactly convergence means in the multistep context.

Recall the general k-step LMM from (3.2), i.e.

$$x_{j+k} + \sum_{l=0}^{k-1} \alpha_l x_{j+l} = h \sum_{l=0}^{k} \beta_l f_{j+l}, \quad (3.13)$$

for numerically solving

$$x'(t) = f(t, x(t)), \quad t > 0, \qquad (3.14)$$
$$x(0) = x_0.$$

Since (3.13) cannot be used to define x_l for $l<k$, it must be complemented with the starting values

$$x_0 = \eta_0, \quad x_1 = \eta_1, \quad \ldots, \quad x_{k-1} = \eta_{k-1}, \qquad (3.15)$$

which can be obtained, e.g., by using some one-step method such as the Euler's method. [The first equation of (3.15) is just for notational convenience as x_0 is already defined by the to-be-solved problem (3.14).]

Naturally, one cannot assume that $\eta_l = x(t_l)$, $l = 1, \ldots, k-1$, i.e., that the starting values are exact, but the least that should be expected is that

$$\lim_{h \to 0} \eta_l = x_0 , \qquad l = 1, \ldots, k-1 . \quad (3.16)$$

In other words, as the time step gets infinitely small, all starting values should approximate the initial value of (3.14) infinitely well.

<u>Definition 3.4.</u> The LMM (3.13) with starting values satisfying (3.16) is said to be convergent on $[0,T]$ if

$$\lim_{\substack{h \to 0 \\ jh = t}} x_j = x(t)$$

for all $t \in [0,T]$, with x being the unique solution of (3.14).

The following example demonstrates that consistence does not imply convergence for LMMs.

<u>Example.</u> The two-step method

$$x_{j+2} + 4x_{j+1} - 5x_j = h(4f_{j+1} + 2f_j) \quad (3.17)$$

is consistent of order 3 (, which is the highest possible order for an explicit two-step method; see page 3/13). Unfortunately, it does not converge, rendering it useless in practice:

Consider the trivial problem

$$x'(t) = 0, \qquad x(0) = 1$$

and apply (3.17) to it with the starting values $\eta_0 = 1$ and $\eta_1 = 1 + h$, which obviously satisfy (3.16).

In this case (3.17) becomes

$$x_{j+2} + 4x_{j+1} - 5x_j = 0; \quad x_0 = 1, \; x_1 = 1+h. \tag{3.18}$$

This is a homogeneous difference equation of the second order, and can be solved via the trial $x_j = r^j$:

$$r^{j+2} + 4r^{j+1} - 5r^j = 0$$

$$\Leftrightarrow r^j \underbrace{(r^2 + 4r - 5)}_{= \rho(r)} = 0 \tag{3.19}$$

$$\Leftrightarrow r = 0 \quad \text{or} \quad r = 1 \quad \text{or} \quad r = -5.$$

[Notice that $r=1$ must be a solution of (3.18) as (3.17) is known to be consistent; see Lemma 3.2.]

From this it follows that the general solution of (3.18) is (without initial conditions)

$$x_j = A\,1^j + B(-5)^j, \qquad j = 0, 1, \ldots,$$

and taking the initial conditions into account

$$x_j = 1 + \frac{1}{6}h\left(1 - (-5)^j\right).$$

3/18

If we, e.g., consider the fixed time $t = 1 = jh$, i.e. $h = 1/j$, it follows that

$$x_j = 1 + \frac{1}{6j}\left(1 - (-5)^j\right),$$

which diverges as $j \to \infty$ (and $h \to 0$). This means that (3.17) is <u>not</u> a convergent method.

[<u>Remark 3.5.</u> Linear difference equations can be solved analogously to linear (constant-coefficient) differential equations: the trial $x_j = r^j$ gives the linearly independent solutions for the homogeneous equation*, and the general solution for a nonhomogeneous equation is obtained by adding any particular solution of the nonhomogeneous equation to the linear combination of the "homogeneous" solutions.

* If there are multiple roots, things get slightly more complicated.]

Observe that the problems of the previous example originate from the fact that (3.19), i.e. the first characteristic polynomial of the method (3.17), has a root that has modulus larger than one; in fact a multiple root of modulus exactly one would have also meant trouble (see, e.g., Example 5.3 of G&H). This observation motivates the following definitions.

Definition 3.6. A polynomial is said to satisfy the *root condition* if all of its roots are within the closed unit disk in the complex plain, and the roots on the unit circle are simple.

3/20

Definition 3.7. The LMM (3.13) is called *zero-stable* if its first characteristic polynomial ρ (cf. Definition 3.2) satisfies the root condition.

It turns out that consistency (Definition 3.1) and zero-stability induce convergence.

Theorem 3.8. The LMM (3.13) is convergent if and only if it is both consistent and zero-stable.

Proof. Dahlquist 1956.

One can think that consistency guarantees that the next iterate is "accurate enough" if the previous k iterates are *exact*. On the other hand, zero-stability guarantees that (small) errors in the previous iterates do not *accumulate* undesirably.

Unfortunately, zero-stability restricts the best possible order of consistency for LMMs considerably.

Theorem 5.7. (First Dahlquist Barrier)
The order of a zero-stable k-step LMM (3.13) satisfies

(i) $p \leq k+2$ if k is even,

(ii) $p \leq k+1$ if k is odd

(iii) $p \leq k$ if the LMM is explicit (or more generally $\beta_k \leq 0$).

Proof. Dahlquist 1959.

According to the rationale on page 3/13, a k-step method can be of order $2k$ (or $2k-1$ for explicit methods). Hence, the requirement of zero-stabilty, or convergence, approximately halves the attainable order of consistency.

3/22

We complete the discussion on the convergence of LMMs with a (slightly vague) theorem on the speed of global convergence.

Theorem 3.9. If the LMM (3.13) is zero-stable, consistent of order p, and the starting values $\eta_0, \eta_1, \dots, \eta_{k-1}$ of (3.15) are "accurate enough", then the global error of the method behaves like $O(h^p)$, that is,

$$|x_j - x(t_j)| \leq C h^p$$

for all $t_j = jh \in [0, T]$ with $T > 0$ fixed, assuming that the solution of (3.1), $x : \mathbb{R}_+ \to \mathbb{R}^n$ is smooth enough.

3.3 Absolute stability and stiff systems

Often the system modelled by an initial value problem includes phenomena that die out quickly. As an example, the half-life of some radioactive isotope may be considerably shorter than of the others, or some subprocess of a chemical reaction may reach its equilibrium faster than the whole system. To model such quickly changing phenomena accurately, it is typically necessary to use extremely small time steps in the numerical solver — far smaller than required by the "other parts" of the examined system.

However, this potentially enormous computational load can be avoided by noticing that for rapidly stabilizing processes it is more essential to correctly model the long-term behavior than to get all details right when the process still has large "change rate" or derivative. In consequence, it is of interest to study how different numerical methods succeed in predicting the state of a quickly stabilizing system at time "$t=\infty$".

3/25

It turns out that the suitability of a numerical scheme for solving stiff problems (i.e. problems of the type described on the previous two pages) can be studied by applying them to the simple test problem

$$x'(t) = \underset{\substack{\cap \\ \mathbb{C}}}{\lambda} x(t), \qquad x(0) = \overset{\substack{0 \\ \#}}{x_0}, \qquad (3.20)$$

which is known to have the solution $x(t) = x_0 e^{\lambda t}$. In particular,

$$\lim_{t \to \infty} x(t) = 0 \qquad (3.21)$$

if $\operatorname{Re} \lambda < 0$. Naturally, it would be desirable that when a numerical method is applied to (3.20), the resulting sequence x_j, $j = 0, 1, 2, \dots$,

3/26

would have this same general behavior, that is,

$$\lim_{j\to\infty} x_j = 0 \quad \text{if } \operatorname{Re}\Lambda < 0, \tag{3.22}$$

independently of the step size $h>0$.

It turns out that for most methods — in particular, for all explicit methods — it is too much to ask that (3.22) holds for all $h>0$.

We start deeper analysis with two definitions and a remark.

Definition 3.10. An LMM is said to be *absolutely stable* at $\hat{h} := h\Lambda \in \mathbb{C}$ if, when applied to (3.20) with step size $h>0$, it produces a numerical solution * satisfying $x_j \longrightarrow 0$ when $j \longrightarrow \infty$.

* for any choice of starting values.

Remark 3.11. As we shall see in what follows, in analysis of absolute stability the parameters $h>0$ and $\lambda \in \mathbb{C}$ always occur as the product $h\lambda$, which motivates the introduction of $\hat{h} \in \mathbb{C}$ in Definition 3.10.

Definition 3.12. The *region of absolute stability* for an LMM is defined as the set $\mathcal{R} \subset \mathbb{C}$ of those $\hat{h} = h\lambda$ for which the LMM is absolutely stable.

As the exact solution of (3.20) satisfies $\lim\limits_{t\to\infty} x(t) = 0$ *if and only if* $\operatorname{Re}\lambda < 0$, it would seem optimal that $\mathcal{R}$ is the (open) left half-plane of $\mathbb{C}$.

3/28

The following example demonstrates that $\mathcal{R}$ can be both larger or smaller than this "optimal" set.

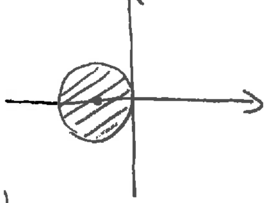
Example. If the Euler's method is applied to (3.20), one arrives at the recursion

$$x_{j+1} = x_j + \overbrace{h\lambda}^{\hat{h}} x_j, \quad j = 0, 1, \ldots,$$

or equivalently,

$$x_j = (1 + \hat{h})^j x_0, \quad j = 0, 1, 2, \ldots .$$

Hence, $\lim_{j \to \infty} x_j = 0$ for $x_0 \neq 0$ if and only if



$$|1 + \hat{h}| < 1 \iff |\hat{h} - (-1)| < 1,$$

and so the region of absolute stability for the Euler's method is the open disk of radius one around the point $-1 \in \mathbb{C}$.

3/29

On the other hand, the implicit (or backward) Euler's method

$$x_{j+1} = x_j + h f(t_{j+1}, x_{j+1}), \quad j = 0,1,2,\dots,$$

produces the recursion

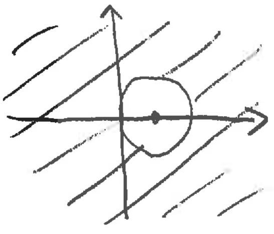
$$x_{j+1} = x_j + h\lambda x_{j+1}$$

$$\Rightarrow \quad x_{j+1} = \frac{x_j}{1 - \underbrace{h\lambda}_{\hat{h}}}, \quad j = 0,1,2,,$$

resulting in

$$x_j = \frac{1}{(1-\hat{h})^j} x_0, \quad j = 0,1,2,\dots.$$

Therefore, the region of absolute stability for the implicit Euler's method is



$$\mathcal{R} = \{\hat{h} \in \mathbb{C} \mid |\hat{h}-1| > 1\},$$

i.e. the open exterior of the disk of radius one around $1 \in \mathbb{C}$.

3/30

To sum up, if $\lim_{t\to\infty} x(t) = 0$ then $\lim_{j\to\infty} x_j = 0$ for the implicit Euler's method, but the same holds for the (explicit) Euler's method only if the step size $h > 0$ is small enough (relative to λ). This actually reflects a more general trend: the regions of absolute stability are typically larger for implicit than for explicit methods.

Let us then approach the question whether a given $\hat{h}$ belongs to $\mathcal{R}$ for a general (consistent and zero-stable) k-step method

$$
\begin{aligned}
x_{j+k} + \alpha_{k-1} x_{j+k-1} + \dots + \alpha_0 x_j \qquad\qquad (3.23) \\
= h(\beta_k f_{j+k} + \dots + \beta_0 f_j), \quad j = 0, 1, \dots
\end{aligned}
$$

Applying this to (3.20), we get

$$x_{j+k} + \alpha_{k-1} x_{j+k-1} + \ldots + \alpha_0 x_k \qquad (3.24)$$
$$= h(\lambda \beta_k x_{j+k} + \ldots + \lambda \beta_0 x_k)$$

$$\Rightarrow (1-\hat{h}\beta_k)x_{j+k} + (\alpha_{k-1} - \hat{h}\beta_{k-1})x_{j-k-1} + \ldots + (\alpha_0 - \hat{h}\beta_0)x_j$$
$$= 0, \quad j = 0, 1, 2, \ldots$$

This is a homogeneous linear difference equation that can be solved by the trial $x_j = r^j$, leading to

$$(1-\hat{h}\beta_k) r^k + (\alpha_{k-1} - \hat{h}\beta_{k-1}) r^{k-1} + \ldots + (\alpha_0 - \hat{h}\beta_0) = 0$$

$$\Leftrightarrow \rho(r) - \hat{h}\sigma(r) = 0, \qquad (3.25)$$

where ρ and σ are the first and second characteristic polynomials of the LMM (3.23) introduced in Definition 3.2.

A couple of definitions are in order.

Definition 3.13. The polynomial

$$p(r) = p_{\hat{h}}(r) = \rho(r) - \hat{h}\,\sigma(r)$$

is called the *stability polynomial* of the LMM (3.23).

Definition 3.14. A polynomial is said to satisfy the *strict root condition* if all of its roots lie within the *open* unit disk in the complex plane.

The following theorem (implicitly) characterizes the region of absolute stability for a given LMM.

Theorem 3.15. An LMM is absolutely stable at $\hat{h} = h\lambda \in \mathbb{C}$ if and only if its stability polynomial $p_{\hat{h}}$ satisfies the strict root condition.

Proof. From problem 1 of exercise session 2, we know that the general form of the solution to (3.24) is

$$x_j = A_1 g_1(r_1, j) + A_2 g_2(r_2, j) + \dots + A_k g_k(r_k, j),$$

where $A_1, \dots, A_k$ are arbitrary constants, $r_1, \dots, r_k$ are the roots of (3.25), i.e. $p_{\hat{h}}(r) = 0$, counted according to their multiplicity, and

$$g_\ell(r, j) = j^m r^j$$

for all $\ell = 1, \dots, k$ and some $m = m(\ell) \in \mathbb{N}_0$. Thus, asking that $\lim_{j\to\infty} x_j = 0$ for arbitrary $A_1, \dots, A_k$ is equivalent to $|r_1|, \dots, |r_k| < 1$.

Determining the region of absolute stability analytically for a given LMM is in general a tedious task: one needs to figure out how the moduli of the roots for a polynomial behave as functions of the coefficients of that polynomial. For 2-step methods this can be done, but for more general LMMs it is more natural to study the properties of the region of absolute stability numerically (see, e.g., G&H, Section 6.3).

The property that $\mathcal{R}$ contains the left half of the complex plain is so important that it merits a definition.

Definition 3.16. An LMM is said to be A-stable if its region of absolute stability includes the open left half plane

$$\mathbb{C}_- := \{ w \in \mathbb{C} \mid \operatorname{Re} w < 0 \}.$$

According to previous examples, the implicit Euler's method is A-stable, whereas the Euler's method is not. For the trapezoidal rule $\mathcal{R} = \mathbb{C}_-$ (an exercise), meaning that it is 'exact' in the sense that

$$\lim_{t\to\infty} x(t) = 0 \quad \overset{\text{for any } h}{\iff} \quad \lim_{j\to\infty} x_j = 0$$

where x is the exact solution of (3.20) and $\{x_j\}_{j=0}^{\infty}$ the corresponding numerical approximation by the trapezoidal rule. In fact, the trapezoidal rule is somewhat special as indicated by the following theorem (the proof is omitted for obvious reasons):

<u>Theorem 3.17</u>. (Second Dahlquist barrier)

1. There is no explicit A-stable LMM.
2. An A-stable (implicit) LMM cannot have consistency order $p > 2$.

As the trapezoidal rule is A-stable and consistent of order $p=2$, it is in a certain sense "as good as it gets". Bear in mind, however, that (unlike zero-stability) A-stability is not an indispensable requirement for the feasibility of an LMM.

To complete this section, we consider how the absolute stability considerations change if (3.20) is replaced by the n-dimensional system

$$u'(t) = A\,u(t), \qquad u(0) = \overset{\neq 0}{u_0}, \qquad (3.26)$$

where $u : \mathbb{R}_+ \to \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ and $u_0 \in \mathbb{R}^n$.

It is well known that the unique solution of (3.26) can be given as

$$u(t) = e^{tA} u_0 , \qquad t \geq 0, \qquad (3.27).$$

where the matrix exponent function $e^{\cdot} : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}$ is defined by its "Taylor series"

$$(3.28) \qquad e^{B} = \sum_{l=0}^{\infty} \frac{1}{l!} B^{l} = I + B + \frac{1}{2!} B^{2} + \frac{1}{3!} B^{3} + \dots .$$

[To check that (3.27) really solves (3.26), it is enough to differentiate (3.27) "term by term" using the definition (3.28).]

In order to have a meaningful discussion on absolute stability for (3.26), we need to figure out when $u(t)$ of (3.27) satisfies $\lim_{t\to\infty} u(t) = \underset{\overset{\uparrow}{\mathbb{R}^{n}}}{0}$.

<u>Theorem 3.18.</u> The solution of (3.26) satisfies $\lim_{t\to\infty} u(t) = 0 \in \mathbb{R}^n$ ** if and only if

$$\operatorname{Re}\lambda_j < 0, \qquad j=1,2,\dots,n,$$

where λ_j, $j=1,2,\dots,n$, are the eigenvalues of $A\in\mathbb{R}^{n\times n}$ counted according to their (algebraic) multiplicity.

** for all $u_0 \in \mathbb{R}^n$

<u>Proof.</u> We prove the claim assuming that there exist linearly independent eigenvectors $v_1, v_2, \dots, v_n \in \mathbb{R}^n$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ * (in other words, the geometric and algebraic multiplicity of λ_j coincide for all $j=1,\dots,n$). In this case A is diagonalizable,

* $Av_j = \lambda_j v_j$

(3.29) $A = V\Lambda V^{-1} \iff V^{-1}AV = \Lambda$,

where $V = [v_1, v_2, \dots, v_n] \in \mathbb{R}^{n\times n}$ and $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$.

Obviously, it holds that

$$
\begin{aligned}
V^{-1}e^{tA}V &= V^{-1}\left(\sum_{l=0}^{\infty} \frac{1}{l!}(tA)^l\right)V \\
&= \sum_{l=0}^{\infty} \frac{t^l}{l!} V^{-1}A^lV \\
&= \sum_{l=0}^{\infty} \frac{t^l}{l!} (V^{-1}AV)^l && \| \; VV^{-1} = I \\
&= \sum_{l=0}^{\infty} \frac{t^l}{l!} \Lambda^l && \| \; \text{Note that } \Lambda^l \text{ is the same as } \mathrm{diag}(\lambda_1^l, \dots, \lambda_n^l). \\
(3.29\tfrac{1}{2}) \qquad &= \mathrm{diag}(e^{t\lambda_1}, e^{t\lambda_2}, \dots, e^{t\lambda_n}).
\end{aligned}
$$

Thus, $\lim\limits_{t\to\infty} V^{-1}e^{tA}V = 0 \in \mathbb{R}^{n\times n}$ if and only if $\mathrm{Re}\,\lambda_j < 0$, $j = 0, 1, \dots, n$, and the same applies to

$$\lim_{t\to 0} e^{tA} = V\left(\lim_{t\to 0} V^{-1}e^{tA}V\right)V^{-1}$$

as $VBV^{-1} = 0$ if and only if $B = 0$ for any matrix $B \in \mathbb{R}^{n\times n}$. The claim follows from the representation (3.27) since obviously

$$B = 0 \in \mathbb{R}^{n\times n} \iff Bu_0 = 0 \in \mathbb{R}^n \quad \text{for all } u_0 \in \mathbb{R}^n.$$

☺

<u>Remark 3.19.</u> If the coefficient matrix $A \in \mathbb{R}^{n\times n}$ is not diagonalizable, the above proof can be carried out by replacing (3.29) by the so-called Jordan normal form.

Consequently, if the eigenvalues of A have negative real parts, the solution of (3.26) satisfies $\lim_{t\to\infty} u(t) = 0$, and one would hope that the numerical solution $u_j \in \mathbb{R}^n$, $j = 0, 1, 2, \ldots$, produced by an LMM* would have this same property, i.e. $\lim_{j\to\infty} u_j = 0 \in \mathbb{R}^n$.

We redefine absolute stability for (3.26) in the natural way:

<u>Definition 3.20.</u> When an LMM is applied to (3.26) with a fixed $h > 0$, it is said to be absolutely stable if $\lim_{j\to\infty} u_j = 0 \in \mathbb{R}^n$ for all choices of starting values.

* Note that all introduced LMMs can also be used for systems ($n > 1$), although they have been analyzed only for $n = 1$.

The following theorem demonstrates that analysis of absolute stability for (3.26) can be reduced to the case of the scalar model problem (3.20).

Theorem 3.21. The k-step LMM (3.23), with time step $h>0$, is absolutely stable for (3.26) if and only if $h\lambda_j \in \mathcal{R}$ (cf. Definition 3.12) for all eigenvalues λ_j, $j=1,\dots,n$, of A.

Proof. As in the proof of Theorem 3.18, we only consider the case when $A = V\Lambda V^{-1}$ is diagonalizable, where we use the same notation as in Theorem 3.18.

When the LMM (3.23) is applied to (3.26), we get

$$u_{j+k} + \alpha_{k-1} u_{j+k-1} + \dots + \alpha_0 u_j \qquad (3.29)$$
$$= hA(\beta_k u_{j+k} + \beta_{k-1} u_{j+k-1} + \dots + \beta_0 u_j).$$

$j=0,1,2,\dots$

Let us define $x_j = V^{-1}u_j \in \mathbb{R}^n$, and multiply (3.29) from left by V^{-1}, resulting in

$$x_{j+k} + \alpha_{k-1}x_{j+k-1} + \dots + \alpha_0 x_j \qquad (3.30)$$
$$= h\underbrace{V^{-1}AV}_{=\Lambda}\Big(\beta_k \underbrace{x_{j+k}}_{V^{-1}u_{j+k}} + \dots + \beta_0 \underbrace{x_j}_{V^{-1}u_j}\Big), \quad j=0,1,\dots$$

Because $\Lambda \in \mathbb{R}^{n\times n}$ is diagonal, the equations of (3.30) decouple:

$$x^\ell_{j+k} + \alpha_{k-1}x^\ell_{j+k-1} + \dots + \alpha_0 x^\ell_j \qquad (3.31)$$
$$= h\Lambda_\ell\big(\beta_k x^\ell_{j+k} + \dots + \beta_0 x^\ell_j\big), \quad j=0,1,\dots,$$

for all $\ell = 1,2,\dots,n$ (where the upper index denotes the component of $x_j \in \mathbb{R}^n$).

Since (3.31) corresponds to applying the k-step method (3.23) to (3.20) with $\Lambda = \Lambda_\ell$, it follows from Definition 3.12 that $\lim_{j\to\infty} x^\ell_j = 0$ (for all starting values) if and only if $h\Lambda_\ell \in \mathcal{R}$.

Thus,

$$\lim_{j\to\infty} x_j = 0 \in \mathbb{R}^n \iff h\Lambda_\ell \in \mathcal{R} \quad \text{for all } \ell = 1,\dots,n.$$

Since $u_j = Vx_j$ (and V is invertible), the claim follows.

[To be quite precise, we should have been a bit more careful with the starting values $u_0,\dots,u_{k-1}$. However, when these go through all possible combinations in $\mathbb{R}^n$, so do $x_0 = V^{-1}u_0, \dots, x_{k-1} = V^{-1}u_{k-1}$, and we are on the safe side.]

☺

We complete Chapter 3 by revisiting the concept of a *stiff system*. To this end, Consider the initial value problem (3.26), assume that $\operatorname{Re}\lambda_j < 0$, $j=1,\dots,n$, for the eigenvalues of $A \in \mathbb{R}^{n\times n}$, and consider the ratio

$$\frac{\max\limits_{1\le j\le n} -\operatorname{Re}\lambda_j}{\min\limits_{1\le j\le n} -\operatorname{Re}\lambda_j} \qquad (3.32)$$

e.g. 10^6

If this ratio is large, (3.26) is called a stiff system. If we apply a method for which $\mathcal{R} \not\supseteq \mathbb{C}_-$ to such (3.26), the step length $h>0$ (from the point of view of stability) is typically restricted by the eigenvalue realizing the numerator of (3.32) (cf. Theorem 3.21), while the long-term behavior of the solution to (3.26) is affected the most by the one realizing the denominator (cf. $(3.29\frac{1}{2})$).

As a consequence, a lot of computational power must be spent to control the numerical instability caused by the quickly stabilizing part of the solution, although much larger time step would be enough to model the long-term behavior accurately enough. For this reason A-stable methods are recommendable for stiff systems (although they are always implicit).

As a final comment, we note that A-stable methods often outperform other techniques also when the right-hand side of (3.26) is nonlinear in $u(t)$, but analysis of such a situation is well outside the scope of this course.

4. Runge-Kutta methods

We return to the general initial value problem

$$\begin{aligned} x'(t) &= f(t, x(t)), \quad t > 0 \\ x(0) &= x_0, \end{aligned} \qquad (4.1)$$

but, as in the case of LMMs, simplify the analysis by assuming that $x: \mathbb{R} \to \mathbb{R}$ and $f: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. (Note that we could — and should — have sometimes assumed that $x: \mathbb{R} \to \mathbb{C}$ and $f: \mathbb{R} \times \mathbb{C} \to \mathbb{C}$ in Chapter 3, and also in the following.)

Compared to the Euler's method, the leading idea of LMMs was to get a more accurate value for x_{j+1} by using not only x_j but also other earlier iterates.

The Runge-Kutta (RK) methods are based on a slightly different principle: x_{j+1} is computed employing only* x_j, but update is formed in a more complicated, nonlinear fashion.

Let us start (in a pedagogically indefensible manner) by writing down (the arguably intimidating) general s-stage RK method

$$x_{j+1} = x_j + h \sum_{l=1}^{s} b_l k_l \,, \qquad (4.2)$$

where

$$k_l = f\Big(t_j + c_l h,\ x_j + h\sum_{m=1}^{s} a_{lm} k_m\Big), \qquad (4.3)$$
$$l = 1, 2, \dots, s.$$

The basic idea is the same as for the Euler's method: update x_j by taking a step in a "direction" defined by the right-hand side of (4.1).

* a one-step method

However, the "direction" in (4.2) is formed as a weighted mean* of k_l, $l=1,\dots,s$, which are (in general, implicitly) defined evaluations of f.

The coefficients in (4.2) and (4.3) are traditionally given as a Butcher array

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \qquad (4.4)$$

.

In particular, notice that if $a_{lm}=0$ for all $m \geq l$, the "updates" k_l, $l=1,\dots,s$, can be computed recursively (first k_1, then k_2 and so on) without solving any equations.

* $\sum_{l=1}^{s} b_l = 1$ turns out to be an consitency condition.

In this case the RK method is explicit; notice that for explicit methods the "coefficient matrix" $\mathcal{A} := (a_{\ell m})_{\ell,m=1}^{s}$ is strictly lower triangular. On the other hand, if $a_{\ell m} \neq 0$ for some $m \geq \ell$ computing k_ℓ, $\ell = 1, \dots, s$, requires solving at least one equation involving f, meaning that the method is implicit.

A few clarifying examples are in order:

Example. Choosing $s=1$, $c_{11} = a_{11} = 0$ and $b_1 = 1$ leads to

$$\begin{aligned} x_{j+1} &= x_j + h k_1, \\ k_1 &= f(t_j, x_j), \end{aligned}$$

which is just the Euler's method.

On the other hand, $s=1$, $c_{11} = a_{11} = 1$ and $b_1 = 1$ results in

$$\begin{aligned} x_{j+1} &= x_j + hk_1 \\ k_1 &= f(t_j + h,\ x_j + hk_1) \\ &= f(t_{j+1},\ x_{j+1}), \end{aligned}$$

which is the implicit Euler's method. The corresponding Butcher tables are

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \quad \text{and} \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}.$$

Take note that both Euler's methods are one-stage RK methods, but they are also one-step LMMs.

<u>Example</u> (two-stage methods).
The choices $S=2$, $a_{11} = a_{12} = 0$, $a_{21} = a_{22} = \frac{1}{2}$, $c_1 = 0$, $c_2 = 1$ and $b_1 = b_2 = 1/2$ correspond to

$$\begin{aligned} x_{j+1} &= x_j + \tfrac{1}{2}h(k_1+k_2), \\ k_1 &= f(t_j, x_j), \\ k_2 &= f(t_j+h, x_j+\tfrac{1}{2}h(k_1+k_2)) \\ &= f(t_{j+1}, x_{j+1}). \end{aligned}$$

It is easy to see that this is the trapezoidal rule, which thus is a one-step LMM but also a two-stage RK method with

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

The "explicit counterpart" of the trapezoidal rule is the Heun's method

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

which reads

$$\begin{aligned} x_{j+1} &= x_j + \tfrac{1}{2}h(k_1+k_2),\\ k_1 &= f(t_j, x_j),\\ k_2 &= f(t_j+h,\ x_j+hk_1)\\ &= f(t_j+h,\ x_j+h\,f(t_j,x_j)). \end{aligned}$$

The Heun's method has a local truncation error (LTE) of the order $O(h^3)$, as does the trapezoidal rule, and since it is a one-step method (but a two-stage RK method), it follows from Theorem 2.2 that the corresponding global error is $O(h^2)$. Finally, notice that the Heun's method is not an LMM.

4/8

4.1 Order conditions

As RK methods are one-step methods, their GEs are induced directly by the corresponding LTEs; see Thm. 2.2.

[Recal that the LTE of a given one-step method is defined as

$$\mathrm{LTE} = x(t_{j+1}) - x_{j+1},$$

where x is the (smooth enough) exact solution of (4.1) and x_{j+1} is produced assuming that the previous grid value x_j is exact, i.e. $x_j = x(t_j)$.]

As for LMMs, the determination of the order of RK methods requires tedious Taylor expansions.

Here, we determine all one-stage methods of order $p \geq 1$ $(\text{LTE} = O(h^{p+1}),\ \text{GE} = O(h^p))$ and explicit two-stage methods of order $p \geq 2$. Consult, e.g., G&H, Chapter 9 for more information on three- and four-stage methods.

<u>Lemma 4.1.</u> A general one-stage RK-method defined by

$$\begin{array}{c|c} c_1 & a_{11} \\ \hline & b_1 \end{array}$$

is of order $p=1$ if and only if $b_1 = 1$. Moreover, it is of order $p=2$ if and only if $b_1 = 1$ and $a_{11} = c_1 = \frac{1}{2}$.

Proof. A smooth enough solution of (4.1) satisfies

$$x(t_{j+1}) = x(t_j) + h x'(t_j) + \frac{h^2}{2} x''(t_j) + O(h^3)$$

$$= x(t_j) + h f(t_j, x(t_j))$$

$$(4.4) \qquad + \frac{h^2}{2} \big(f_t(t_j, x(t_j)) + f(t_j, x(t_j))\, f_x(t_j, x(t_j)) \big)$$

$$+ O(h^3),$$

where the second equality follows as in Problem 4 of Exercise 1.

On the other hand, the examined RK method is

$$x_{j+1} = x_j + h b_1 k_1, \qquad (4.5)$$

$$k_1 = f(t_j + c_1 h, x_j + h a_{11} k_1).$$

Let us treat $k_1 = k_1(h)$ as a function of the time step, and write a Taylor expansion:

$$k_1(h) = k_1(0) + h \Big[\underset{\substack{\uparrow \\ \text{total} \\ \text{derivative}}}{\frac{d}{dh}} k_1(h) \Big]_{h=0} + O(h^2).$$

With the help of the chain and product rules, we get

derivative w.r.t. the first argument

$$\frac{d}{dh} k(h) = f_t(t_j + c_1 h,\ x_j + a_{11} h k(h))\, c_1 + f_x(t_j + c_1 h,\ x_j + a_{11} h k(h)) \left(a_{11} k(h) + a_{11} h k'(h)\right),$$

derivative w.r.t. the second variable

and subsequently by setting $h = 0$,

$$k_1(h) = \underbrace{f(t_j, x_j)}_{k(0)} + h\Big(f_t(t_j, x_j)\, c_1 + f_x(t_j, x_j)\, a_{11} \underbrace{f(t_j, x_j)}_{k(0)} \Big) + O(h^2).$$

Substituting this in (4.5), it follows that

$$x_{j+1} = x_j + b_1 h f(t_j, x_j) + b_1 \frac{h^2}{2} \left(2 c_1 f(t_j, x_j) + 2 a_{11} f(t_j, x_j) f_x(t_j, x_j) \right) + O(h^3). \tag{4.6}$$

The assertion follows by comparing (4.4) and (4.6) under the assumption $x_j = x(t_j)$.

4/12

The highest order one-stage method defined by $\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$ is the implicit midpoint rule

$$x_{j+1} = x_j + hk_1,$$
$$k_1 = f(t_j + \tfrac{1}{2}h, x_j + \tfrac{1}{2}hk_1),$$

which can be written neatly as

$$x_{j+1} = x_j + f\left(\tfrac{1}{2}(t_j + t_{j+1}), \tfrac{1}{2}(x_j + x_{j+1})\right).$$

Lemma 4.2. An explicit two-stage RK method has order $p=2$ if and only if

$$(4.6\tfrac{1}{2}) \quad c_1 b_1 + c_2 b_2 = \tfrac{1}{2}, \quad a_{21} b_2 = \tfrac{1}{2}, \quad b_1 + b_2 = 1$$

[Recall that $a_{11} = a_{12} = a_{22} = 0$ as the considered methods are explicit.]

Proof. The Butcher table of the considered methods is $\begin{array}{c|cc} c_1 & 0 & 0 \\ c_2 & a_{21} & 0 \\ \hline & b_1 & b_2 \end{array}$.

4/13

In practice, this means

$$x_{j+1} = x_j + h(b_1 k_1 + b_2 k_2),$$
$$k_1 = f(t_j + c_1 h, x_j),$$
$$k_2 = f(t_j + c_2 h, x_j + a_{21} k_1).$$

As in the proof of Lemma 4.1, it is useful to write down Taylor expansions for $k_1 = k_1(h)$ and $k_2 = k_2(h)$:

$$k_1(h) = k_1(0) + \left[\frac{d}{dh} k_1(h)\right]_{h=0} + O(h^2)$$
$$= f(t_j, x_j) + h c_1 f_t(t_j, x_j) + O(h^2)$$

$$k_2(h) = k_2(0) + \left[\frac{d}{dh} k_2(h)\right]_{h=0} + O(h^2)$$
$$= k_2(0) + h\Big(f_t(t_j, x_j) c_2$$
$$+ f_x(t_j, x_j) a_{21} k_1(0)\Big) + O(h^2)$$
$$= f(t_j, x_j) + h c_2 f_t(t_j, x_j)$$
$$+ h a_{21} f(t_j, x_j) f_x(t_j, x_j) + O(h^2),$$

where we used exactly the same line of reasoning as in the proof of Lemma 4.1.

$4/14$

Altogether, this means that

$$x_{j+1} = x_j + h(b_1+b_2) f(t_j,x_j) + \frac{h^2}{2}\Big(2(c_1b_1+c_2b_2) f_t(t_j,x_j) + 2a_{21}b_2 f(t_j,x_j) f_x(t_j,x_j)\Big) + O(h^3).$$

Comparing this to (4.4) and using the "localizing" assumption $x_j = x(t_j)$, it follows that the method is of order $p=2$ (i.e. $\text{LTE} = O(h^3)$) if

$$b_1 + b_2 = 1,$$

$$c_1b_1 + c_2b_2 = \frac{1}{2},$$

$$a_{21}b_2 = \frac{1}{2}.$$

☺

Remark. Almost always the extra condition

$$\sum_{m=1}^{\ell} a_{\ell m} = c_m \qquad (4.7)$$

$4/15$

is imposed on the coefficients of a RK method. The motivation for this is the following: By introducing a "new unknown" $\tilde{x}(t) = [x(t), t]^T$, (4.1) can be written in an autonomous form

$$(4.8) \qquad \tilde{x}'(t) = \tilde{f}(\tilde{x}(t)) := \begin{bmatrix} f(t, x(t)) \\ 1 \end{bmatrix}, \qquad \tilde{x}(0) = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}.$$

The condition (4.7) ensures that the RK method gives the same numerical solution independently of whether it is applied to (4.1) or (4.8) (an exercise).

For Lemma 4.2 this means that $c_1 = 0$, $c_2 = a_{21}$, and so $(4.6\tfrac{1}{2})$ consists only of the latter two conditions.

4/16

Notice that the "order conditions" (4.6$\frac{1}{2}$) are nonlinear in the coefficients defining the explicit two-stage RK method; for LMMs the order conditions resulted in linear equations for the free coefficients (cf. (3.10)). This nonlinearity also occurs for higher number of stages than $s=2$. For example, an explicit three-stage method

$$(4.9) \quad \begin{array}{c|ccc} 0=c_1 & 0 & 0 & 0 \\ c_2 & a_{21} & 0 & 0 \\ c_3 & a_{31} & a_{32} & 0 \\ \hline & b_1 & b_2 & b_3 \end{array}, \qquad \sum_{m=1}^{3} a_{\ell m} = c_\ell, \quad \ell=1,2,3,$$

is of order $p=3$ if and only if

$$b_1+b_2+b_3=1, \quad b_2c_2+b_3c_3=\frac{1}{2}$$
$$b_2c_2^2+b_3c_3^2=\frac{1}{3}, \quad c_2a_{32}b_3=\frac{1}{6}. \qquad (4.10)$$

Although (4.9) has six free parameters — note that $a_{21} = c_2$ and $a_{31} = c_3 - a_{32}$ —, and the order $p=3$ is achieved by only satisfying four equations (4.10), it can be shown that an explicit three-stage method cannot have an order higher than $p=3$ (even if the extra conditions $\sum_{m=1}^{3} a_{\ell m} = c_\ell^{*}$, $\ell = 1,2,3$, were removed).

The same holds for two-stage methods as well: The conditions ($4.6\frac{1}{2}$) <u>cannot</u> be amended so that one would have a <u>solvable</u> system guaranteeing the order $p=3$ for a two-stage method.

* This condition does not affect the attainable order of a RK method.

The so-called classic RK method

$$\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}$$

is an explicit four-stage RK method of order $p=4$.

As the order conditions for RK methods are nonlinear, they are nontrivial to analyze: For example, it is not known what is the minimum number of stages required for the order $p=9$ (somewhere between $s=12$ and $s=17$).

4/19

4.2 Absolute stability

The concept of absolute stability is as relevant for RK methods as it is for LMMs.

We once again examine the simple test problem

(4.10) $$x'(t) = \lambda x(t), \quad t>0, \qquad x(0) = x_0,$$

where $\lambda \in \mathbb{C}$. Recall that $x(t) = x_0 e^{\lambda t} \xrightarrow{t\to\infty} 0$, if and only if if $\operatorname{Re}\lambda < 0$, and the definitions of absolute stability and the region of absolute stability:

Definition 4.3. A RK method is absolutely stable at $\hat{h} = h\lambda \in \mathbb{C}$ if, when applied to (4.10) with step size $h>0$ and any $x_0 \in \mathbb{C}$, it holds that $\lim_{j\to\infty} x_j = 0$.

Definition 4.4. The region of absolute stability $\mathcal{R} \subset \mathbb{C}$ for a RK method is the set of those $\hat{h} = h\lambda$ for which the RK method is absolutely stable.

The absolute stability of LMMs was studied in Section 3.3 by looking at the roots of the so-called stability polynomial. For RK methods the concept of stability function turns out to be more fruitful. Let us illustrate it via two examples.

4/21

<u>Example.</u> Consider the family of explicit two-stage RK methods

$$\begin{array}{c|cc} 0 & 0 & 0 \\ (2\theta)^{-1} & (2\theta)^{-1} & 0 \\ \hline & 1-\theta & \theta \end{array} \qquad (4.11)$$

parametrized by $\theta \in \mathbb{R}\setminus\{0\}$. It is easy to check that such a method satisfies the conditions $(4.6\frac{1}{2})$, and thus we are dealing with a second order method for any $\theta \in \mathbb{R}\setminus\{0\}$. Let us apply (4.11) to (4.10):

$$\begin{aligned} x_{j+1} &= x_j + h((1-\theta)k_1 + \theta k_2), \\ k_1 &= f(t_j, x_j) = \lambda x_j, \\ k_2 &= f(t_j + (2\theta)^{-1}h,\ x_j + (2\theta)^{-1}hk_1) \\ &= \lambda(x_j + (2\theta)^{-1}hk_1) \\ &= \lambda x_j + (2\theta)^{-1}h\lambda^2 x_j. \end{aligned}$$

4/22

Altogether we thus get

$$x_{j+1} = x_j + h\Big(\big((1-\theta)\Lambda x_j + \theta\big(\Lambda x_j + (2\theta)^{-1} h \Lambda^2 x_j\big)\Big)$$

$$= x_j + \underbrace{h\Lambda}_{\hat{h}} x_j + \frac{1}{2} \underbrace{h^2\Lambda^2}_{\hat{h}^2} x_j$$

$$= \underbrace{\left(1 + \hat{h} + \frac{1}{2}\hat{h}^2\right)}_{=:\, R(\hat{h})} x_j \,, \qquad j = 0,1,2,\dots .$$

This is a trivial difference equation with the solution

$$x_j = \left(R(\hat{h})\right)^j x_0 \,, \qquad j = 0,1,2,\dots .$$

$R(\hat{h})$ is called the *stability function* of the method(s) (4.11), and it obviously holds that

$$\hat{h} \in \mathcal{R} \iff |R(\hat{h})| < 1 .$$

Notice, in particular, that

$$R(\hat{h}) = e^{\hat{h}} + O(\hat{h}^3).$$

4/23

Example. Let us next examine the implicit midpoint rule

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}.$$

Applying this to the test problem (4.10) yields

$$\begin{aligned} x_{j+1} &= x_j + h k_1 , \\ k_1 &= f(t_j + \tfrac{1}{2}h , x_j + \tfrac{1}{2}h k_1) \\ &= \lambda (x_j + \tfrac{1}{2} h k_1) . \end{aligned}$$

Solving the latter equation for k_1 results in

$$k_1 = \frac{\lambda}{1 - \frac{1}{2}h\lambda} x_j ,$$

and thus

$$\begin{aligned} x_{j+1} &= x_j + \frac{\overbrace{h\lambda}^{\hat{h}}}{1 - \frac{1}{2}\underbrace{h\lambda}_{\hat{h}}} x_j \\ &= \frac{1 + \frac{1}{2}\hat{h}}{1 - \frac{1}{2}\hat{h}} x_j , \qquad j = 0, 1, 2, \dots . \end{aligned}$$

In consequence,

$$x_j = (R(\hat{h}))^j x_0, \quad j = 0, 1, 2, \ldots,$$

where

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

is the stability function of the implicit midpoint rule (cf. Problem 1, Exercise 3). As in the previous example,

$$|R(\hat{h})| < 1 \iff \hat{h} \in \mathcal{R} \ (= \mathbb{C}_-).$$

Notice that for $|\hat{h}| < 2$, we have

$$\frac{1}{1 - \frac{1}{2}\hat{h}} = \sum_{l=0}^{\infty} \left(\frac{1}{2}\hat{h}\right)^l = 1 + \frac{1}{2}\hat{h} + \frac{1}{4}\hat{h}^2 + O(\hat{h}^3)$$

due to the sum formula for a geometric series. Hence

$$R(\hat{h}) = \left(1 + \frac{1}{2}\hat{h}\right)\left(1 + \frac{1}{2}\hat{h} + \frac{1}{4}\hat{h}^2\right) + O(\hat{h}^3)$$
$$= 1 + \hat{h} + \frac{1}{2}\hat{h}^2 + O(\hat{h}^3) = e^{\hat{h}} + O(\hat{h}^3).$$

4/25

The findings of these two examples can be generalized for any RK method.

Proposition 4.5. When a RK method of order $p \geq 1$ is applied to the test problem (4.10), the numerical solution is of the form

$$(4.12) \qquad x_j = (R(\hat{h}))^j x_0, \quad j = 0, 1, 2, \ldots$$

The stability function $R(\hat{h})$ satisfies

$$R(\hat{h}) = e^{\hat{h}} + O(\hat{h}^{p+1}). \qquad (4.13)$$

Moreover, for a s-stage *explicit* method $R(\hat{h})$ is a sth order polynomial, and for an implicit s-stage method

$$R(\hat{h}) = \frac{q_1(\hat{h})}{q_2(\hat{h})}, \qquad (4.14)$$

where q_1 and q_2 are polynomials of order s.

4/26

<u>Hand-waving proof</u>. It is "easy to see" that, when a general s-stage RK method is applied to (4.10), the relationship between consecutive iterates can be given in the form

$$q_2(\hat{h})\, x_{j+1} = q_1(\hat{h})\, x_j, \quad j = 0, 1, 2, \ldots,$$

where q_1 and q_2 are polynomials of order s. Moreover, if the method is explicit, x_{j+1} does not "appear" on the right-hand side of (4.2), meaning that $q_2(\hat{h}) \equiv 1$. This "proves" (4.12) and (4.14).

Finally, because $x_0 = x(0)$ is exact, for the first iterate of a pth order method it must hold that

$$x_1 = R(\hat{h})\, x_0 = x(h) + \overset{\text{LTE}}{O(h^{p+1})}$$
$$= e^{\lambda h} x_0 + O(h^{p+1}) = e^{\hat{h}} x_0 + O(\hat{h}^{p+1}),$$

which proves (4.13).

4/27

An important consequence of Proposition 4.5 is that no (reasonable, i.e. of order $p \geq 1$) explicit RK method can be A-stable. [Recall that a RK method (or an LMM) is A-stable if the open left half of the complex plane is contained in $\mathcal{R}$.]

Indeed, if $R(\hat{h})$ is a polynomial of order $s \geq 1$ — as it is for any reasonable explicit RK method — it holds that

$$\lim_{\mathbb{R} \ni \hat{h} \to -\infty} R(\hat{h}) = \infty \quad \text{or} \quad \lim_{\mathbb{R} \ni \hat{h} \to -\infty} R(\hat{h}) = -\infty ,$$

and thus it follows from (4.12) that the corresponding RK method is not A-stable (nor A_0-stable; see Problem 2, Exercise 3).

4/28

We complete the discussion on absolute stability of RK methods by checking what happens if the model scalar problem (4.10) is replaced by a linear system

$$u'(t) = Au(t), \qquad u(0) = u_0 \in \mathbb{R}^n, \qquad (4.15)$$

where $u : \mathbb{R} \to \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$.

It is "easy" to verify that if an explicit RK method is applied to (4.15), then

$$(4.16) \qquad u_j = (R(hA))^j u_0, \qquad j = 0, 1, 2, \dots,$$

where R is the stability function (a polynomial) of the considered explicit RK method and $\{u_j\}_{j=0}^{\infty} \subset \mathbb{R}^n$ is the corresponding numerical solution with the step size $h > 0$.

Example. When the Heun's method $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$ is applied to (4.15) it follows that

$$\begin{aligned} u_{j+1} &= u_j + \tfrac{1}{2}h(k_1 + k_2), \\ k_1 &= f(t_j, u_j) = Au_j, \\ k_2 &= f(t_j + h, u_j + hk_1) = Au_j + hA^2u_j, \end{aligned}$$

which leads to

$$\begin{aligned} u_{j+1} &= u_j + hAu_j + \tfrac{1}{2}h^2A^2u_j \\ &= R(hA)u_j, \quad j = 0,1,2, \end{aligned}$$

as claimed (cf. p. 4/22)

The following theorem characterize the case when $\lim_{j\to\infty} u_j = 0 \in \mathbb{R}^n$.

Theorem 4.6. An (explicit) RK method is absolutely stable when applied to (4.15), that is, $\lim_{j\to\infty} u_j = 0^{*}$, if and only if $h\Lambda_\ell \in \mathcal{R}$ for all eigenvalues $\Lambda_1, \dots, \Lambda_n \in \mathbb{C}$ of $A \in \mathbb{R}^{n\times n}$.

$*$ for any $u_0 \in \mathbb{R}^n$

Proof. As in the case of LMMs (cf. Theorem 3.21), we prove the assertion only for a diagonalizable A,

$Av_\ell = \Lambda_\ell v_\ell$

$$A = V \Lambda V^{-1}, \tag{4.17}$$

where $V = [v_1, v_2, \dots, v_n] \in \mathbb{R}^{n\times n}$ is built of the linearly independent eigenvectors of A and $\Lambda = \operatorname{diag}(\Lambda_1, \dots, \Lambda_n)$.

[If the eigendecomposition did not exist, one could resort to the Jordan normal form of A.]

According to (4.16), $\lim_{j\to\infty} u_j = 0$ for all $u_0 \in \mathbb{R}^n$ if and only if $\lim_{j\to\infty} (R(hA))^j = 0 \in \mathbb{R}^{n\times n}$.

We will show that the latter holds if and only if $h\Lambda_\ell \in \mathcal{R}$ for all $\ell = 1, \dots, n$.

To begin with, notice that for any $m \in \mathbb{N}_0$,

$$(hA)^m = (V(h\Lambda)V^{-1})^m = V\underbrace{(h\Lambda)^m}_{=\mathrm{diag}(h^m\Lambda_1^m, \dots, h^m\Lambda_n^m)}V^{-1}.$$

Hence, by linearity,

$$\underbrace{R(hA)}_{\text{polynomial}} = V\underbrace{R(h\Lambda)}_{=\mathrm{diag}(R(h\Lambda_1), \dots, R(h\Lambda_n))}V^{-1}.$$

Similarly,

$$\begin{aligned}(R(hA))^j &= V(R(h\Lambda))^j V^{-1} \\ &= V\,\mathrm{diag}(R(h\Lambda_1)^j, \dots, R(h\Lambda_n)^j)\, V^{-1} \\ &\xrightarrow{j\to\infty} 0 \in \mathbb{R}^{n\times n}\end{aligned}$$

if and only if

$$\lim_{j\to\infty} R(h\Lambda_\ell)^j = 0, \qquad \ell = 1, 2, \dots, n.$$

$$\Longleftrightarrow \ |R(h\Lambda_\ell)| < 1 \ \Longleftrightarrow \ h\Lambda_\ell \in \mathcal{R}, \quad \ell = 1, 2, \dots, n,$$

which completes the proof. ☺

Recall from Theorem 3.18 that the exact solution of (4.15) satisfies

$$(4.18) \qquad \lim_{t\to\infty} u(t) = 0 \in \mathbb{R}^n \quad \text{for all } u_0 \in \mathbb{R}^n$$

if and only if $\operatorname{Re}\Lambda_\ell < 0$, $\ell = 1, \dots, n$. As the stability region of an explicit RK method is bounded, the condition $h\Lambda_\ell \in \mathcal{R}$ of Theorem 4.6 is very restrictive on the time step if $\operatorname{Re}\Lambda_\ell \ll 0$ for some ℓ, and so $\lim_{j\to\infty} u_j \neq 0$ may hold even if (4.18) is valid, unless $h > 0$ is <u>small</u> enough.

Luckily Theorem 4.6 holds also for implicit RK methods*, though the proof is slightly more involved (the stability function is no longer a polynomial). We omit the proof and only consider an important example.

Example. Let us apply the implicit midpoint rule $\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$ to (4.15):

$$u_{j+1} = u_j + h k_1$$

$$\mathbb{R}^n \ni k_1 = f(t_j + \tfrac{1}{2}h, u_j + \tfrac{1}{2}hk_1)$$

$$= Au_j + \tfrac{1}{2}hAk_1 .$$

Consequently,

$$k_1 = (I - \tfrac{1}{2}hA)^{-1} A u_j ,$$

* which may be, e.g., A-stable

and altogether

$$
\begin{aligned}
u_{j+1} &= u_j + (I - \tfrac{1}{2}hA)^{-1}(hA)u_j \\
&= (I - \tfrac{1}{2}hA)^{-1}\big((I - \tfrac{1}{2}hA) + hA\big)u_j \\
&= (I - \tfrac{1}{2}hA)^{-1}(I + \tfrac{1}{2}hA)u_j, \quad j = 0, 1, \ldots . \qquad (4.18)
\end{aligned}
$$

Notice that the inverse matrix $(I - \frac{1}{2}hA)^{-1}$ exists for small enough $h > 0$ due to, e.g., Neumann series representation.

Once again, for simplicity, let us assume that $A = V\Lambda V^{-1}$ is diagonalizable and abuse the notation slightly by writing $R(hA) = (I - \frac{1}{2}hA)^{-1}(I + \frac{1}{2}hA)$. Let us figure out the eigenvalues of $R(hA)$.

Obviously,

$$\left(I - \tfrac{1}{2}hA\right) v_\ell = \left(1 - \tfrac{1}{2}h\Lambda_\ell\right) v_\ell$$

$$\Leftrightarrow \quad \frac{1}{1 - \frac{1}{2}h\Lambda_\ell} v_\ell = \left(I - \tfrac{1}{2}hA\right)^{-1} v_\ell$$

and

$$\left(I + \tfrac{1}{2}hA\right) v_\ell = \left(1 + \tfrac{1}{2}h\Lambda_\ell\right) v_\ell$$

for any eigenpair $\Lambda_\ell \in \mathbb{C}$, $v_\ell \in \mathbb{C}^n$ of A.

Thus,

$$R(hA)v_\ell = \underbrace{\frac{1 + \frac{1}{2}h\Lambda_\ell}{1 - \frac{1}{2}h\Lambda_\ell}}_{R(h\Lambda_\ell)} v_\ell, \quad \ell = 1, \dots, n.$$

As A was assumed to be diagonalizable, $v_1, \dots, v_n \in \mathbb{R}^n$ are linearly independent, and we have thus found n linearly independent eigenvectors for $\overset{\mathbb{R}^{n\times n}}{\overset{\in}{R(hA)}}$ with the corresponding eigenvalues $R(h\Lambda_1), \dots, R(h\Lambda_n)$.

In particular, $R(hA)$ is also diagonalizable,

$$R(hA) = V \underbrace{R(h\Lambda)}_{\mathrm{diag}(R(h\lambda_1), \dots, R(h\lambda_n))} V^{-1}.$$

Since by (4.18)

$$u_j = (R(hA))^j, \quad j = 0, 1, \dots,$$

as in the proof of Theorem 4.6, we deduce that

$$\lim_{j \to \infty} u_j = 0 \in \mathbb{R}^n \iff h\lambda_\ell \in \mathcal{R}, \quad \ell = 1, \dots, n.$$

Because $\mathcal{R} = \mathbb{C}_-$ for the implicit midpoint rule, it follows that

$$\lim_{t \to \infty} u(t) = 0 \iff \lim_{j \to \infty} u_j = 0$$

independently of the time step $h > 0$.

Remark 4.7. In stability analysis, we have already many times used the following general result: For an arbitrary matrix $B \in \mathbb{C}^{n \times n}$

$$\lim_{j \to \infty} B^j = 0 \in \mathbb{C}^{n \times n}$$

if and only if all eigenvalues of B, say, $\mu_1, \mu_2, \dots, \mu_n$, satisfy

$$|\mu_\ell| < 1, \quad \ell = 1, \dots, n.$$

The proof has also been implicitly given in case B is diagonalizable, but it has also been noted that the general case can be handled with the help of the Jordan normal form of B.

What was left out?

Before moving on to (briefly) consider parabolic and hyperbolic PDEs, let us list some important topics* related to LMMs and RKMs that were left out:

Adaptive step size selection:

Often the requirements on the time step size $h > 0$ change as a function of time: For example, the modelled system may exhibit rapid changes — calling for a small $h > 0$ — at early time instants, but in the

* Here only two are listed, although there are many more.

long run the alterations are relative slow permitting a longer time step. In such a case, it is reasonable to treat $h = h_j$ as a function of the "discrete time" and try to minimize the computational cost by choosing it *adaptively*. To put it *really* short, the leading idea is often to compute numerical solutions by methods of two different orders: If they give significantly different results, the step size should be decreased; if there is not much difference between the two methods, the step size may be increased.

Implementation of implicit methods:

At each step of an implicit LMM or RKM, one must in general solve a (nonlinear) system of (algebraic) equations.

Usually, this must be done numerically, which considerably increases the computational load; in fact, it is not even completely clear a priori, whether the equations defining the next iterate have a solution (they do have for a small enough $h>0$). In consequence, attention must be paid to efficient implementation of implicit methods.

5. Parabolic PDEs

We will demonstrate how the methods for solving initial value problems of the type

$$y'(t) = f(t, y(t)), \quad y(0) = y_0$$

can be employed in the framework of parabolic partial differential equations (PDEs) by considering the one-spatial-dimensional Dirichlet initial and boundary value problem

$$(5.1) \quad \begin{cases} u_t(x,t) = c\, u_{xx}(x,t), & t>0,\ x\in(0,1) \\ u(0,t) = u(1,t) = 0, & t>0 \\ u(x,0) = g(x), & x\in(0,1). \end{cases}$$

This is as simple parabolic

PDE as there exists. One possible physical interpretation for (5.1) is the following: $u: (0,1) \times \mathbb{R}_+ \to \mathbb{R}$ is the temperature of the rod $(0,1)$ over positive time instants, with the initial temperature distribution $g: (0,1) \to \mathbb{R}$. The end points of the rod are held at constant "zero temperature".*

<u>Theorem 5.1.</u> Assume that $g \in L^2(0,1)$ (i.e. g is square integrable) and $c > 0$. Then (5.1) has a unique solution

$$u(x,t) = \sum_{l=1}^{\infty} \beta_l e^{-c\pi^2 l^2 t} \sin(\pi l x), \tag{5.2}$$

where

* $c > 0$ is the constant thermal diffusivity, which is related to the thermal conductivity.

where

$$\beta_\ell = 2\int_0^1 g(x)\sin(\pi\ell x), \quad \ell=1,2,\dots,$$

are the Fourier sine coefficients of $g:(0,1)\to\mathbb{R}$.

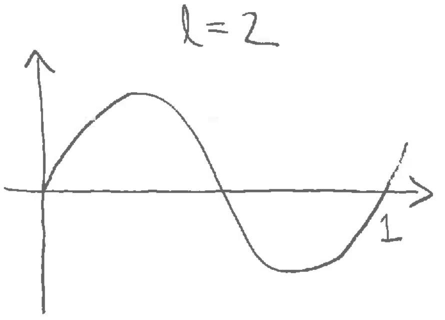
Proof. An exercise.

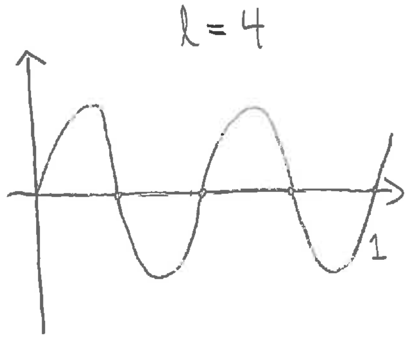
In particular, note that $u(x,t)\xrightarrow{t\to\infty}0$ for all $x\in(0,1)$.

uniformly

Interpretation of (5.2):

The initial temperature $g(x)$ is divided in spatial frequencies $\{\sin(\pi\ell\,\cdot)\}_{\ell=1}^{\infty}$, which form an orthonormal basis of $L^2(0,1)$, but also satisfy the boundary conditions of (5.1). The higher the spatial frequency, the faster the temperature differences disappear





because the strength of the heat flow is proportional to the derivative of the temperature gradient. For example, if $c=1$, at time $t=1$, the "component of g in the direction of $\sin(\pi\,\cdot)$" is multiplied by $e^{-\pi^2} \approx 5.17 \cdot 10^{-5}$ whereas the component in the direction of, say, $\sin(10\pi\,\cdot)$ by $e^{-100\pi^2} = 0$ *.

In other words, the representation (5.2) suggests that "different components" of the solution to (5.1) die out with *considerably* different speeds. [(5.1) has "stiff" written all over itself.]

* within working precision of Matlab

The standard way to numerically solve (parabolic problems like) (5.1) is to first discretize the spatial derivatives to obtain an initial value problem for a system of ordinary differential equations, and then use the techniques discussed in Chapters 3 and 4 to obtain an approximation for $u(x,t)$. Here we handle the second spatial derivative by a standard difference scheme, although in more general settings the finite element method is arguably the prefered choice (Mat-1.3650 Finite element method).

Let us assume that $g:(0,1)\to\mathbb{R}$ is smooth enough and its point values are known at spatial grid points

$$x_j = jh, \quad j=1,2,\ldots,n,$$

where the mesh parameter is $h=\frac{1}{n+1}>0$.

For any four times continuously differentiable function $v:(0,1)\to\mathbb{R}$ it holds that

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

Adding these and solving for v'', one gets the standard central second order difference approximation

$$v''(x) = \frac{1}{h^2}\big(v(x-h) - 2v(x) + v(x+h)\big) + O(h^2). \tag{5.3}$$

Taking into account the boundary conditions of (5.1), an application of (5.3) to the solution of (5.1) at the grid points x_j, $j=1,2,\dots,n$, gives

$$(5.4) \qquad \begin{bmatrix} u_{xx}(x_1,t) \\ u_{xx}(x_2,t) \\ \vdots \\ u_{xx}(x_n,t) \end{bmatrix} = \frac{1}{h^2} \underbrace{\begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}}_{=:\, A \,\in\, \mathbb{R}^{n\times n}} \begin{bmatrix} u(x_1,t) \\ u(x_2,t) \\ \vdots \\ u(x_n,t) \end{bmatrix},$$

which holds "modulo $O(h^2)$" for any (fixed) $t>0$ independently of the smoothness of $g \in L^2(0,1)$. [It is "easy" to see that (5.2) defines an infinitely smooth function for any $t>0$.]

Using (5.4) to approximate the right-hand side of (5.1), one ends up with the initial value problem

The boundary conditions of (5.1) are accounted for in the structure of A.

$$U'(t) = c\,A\,U(t), \qquad U(0) = G, \tag{5.5}$$

where $U : \mathbb{R}_+ \longrightarrow \mathbb{R}^n$ is a "semidiscrete" approximation of $u : (0,1) \times \mathbb{R}_+ \longrightarrow \mathbb{R}$, that is (hopefully),

$$U(t) \approx [u(x_1,t), u(x_2,t), \dots, u(x_n,t)]^T,$$

and G contains the grid values of $g : (0,1) \longrightarrow \mathbb{R}$,

$$G = [g(x_1), g(x_2), \dots, g(x_n)]^T \in \mathbb{R}^n.$$

Now we could proceed by solving (5.5) numerically with the help of one of the LMMs or RKMs studied in Chapters 3 and 4.

However, it is advisable to first have a look at the eigenvalues of the (stiff!) matrix $A \in \mathbb{R}^{n \times n}$.

<u>Lemma 5.2.</u> The eigenvalues of the matrix A (cf (5.4)) are

$$\Lambda_\ell = -\frac{4}{h^2} \sin^2\left(\frac{\ell \pi}{2n+2}\right), \quad \ell = 1, 2, \dots, n,$$

with the corresponding orthonormal eigenvectors

$$v^\ell = \underbrace{\sqrt{2h}}_{\text{normalization constant}} \begin{bmatrix} \sin\left(\frac{\ell\pi}{n+1}\right) \\ \sin\left(\frac{2\ell\pi}{n+1}\right) \\ \vdots \\ \sin\left(\frac{n\ell\pi}{n+1}\right) \end{bmatrix} \in \mathbb{R}^n,$$

$\ell = 1, 2, \dots, n$.

Proof. We denote $a^l = Av^l$.

It is easy to see that

$$(5.6)\quad a^l_j = \frac{\sqrt{2h}}{h^2}\left(\sin\left(\frac{(j-1)l\pi}{n+1}\right) - 2\sin\left(\frac{jl\pi}{n+1}\right) + \sin\left(\frac{(j+1)l\pi}{n+1}\right)\right),$$

the component $\rightarrow j$

which holds for all $j = 1, 2, \dots, n$ (also for $j=1$ and $j=n$ because $\sin 0 = \sin l\pi = 0$).

Since (MAOL)

$$\sin\left(\frac{(j\pm1)l\pi}{n+1}\right) = \sin\left(\frac{jl\pi}{n+1}\right)\cos\left(\frac{l\pi}{n+1}\right) \pm \cos\left(\frac{jl\pi}{n+1}\right)\sin\left(\frac{l\pi}{n+1}\right),$$

(5.6) transforms into

$$a^l_j = \frac{1}{h^2}\underbrace{\left(2\left(\cos\left(\frac{l\pi}{n+1}\right)-1\right)\right)}_{-4\sin^2\left(\frac{l\pi}{2(n+1)}\right)}\underbrace{\sqrt{2h}\sin\left(\frac{jl\pi}{n+1}\right)}_{v^l_j},$$

for $j = 1, 2, \dots, n$ (and $l = 1, 2, \dots, n$).

In other words, $Av^l = \lambda_l v^l$, $l = 1, \dots, n$, as claimed.

Since $\sin(\cdot)$ is a strictly increasing function on the interval $(0, \frac{\pi}{2})$, it follows that

$$0 > \Lambda_1 > \Lambda_2 > \dots > \Lambda_n > -\frac{4}{h^2},$$

and thus we have found all n eigenvalues of A.

Because the eigenvectors corresponding to distinct eigenvalues of a Hermitian (in particular, real and symmetric) matrix are orthogonal, the only remaining thing to prove is that $|v^\ell| = 1$, $\ell = 1, 2, \dots, n$. This is left as an exercise.

If one wants (5.5) to be an accurate approximation of the original model problem (5.1) the mesh parameter $h = \frac{1}{n+1}$ needs to be small enough. This means that the smallest eigenvalue of A, i.e.

$$\Lambda_n = -\frac{4}{h^2}\sin\left(\underbrace{\frac{n}{2n+2}}_{\approx \frac{1}{2}}\pi\right) \approx -\frac{4}{h^2} \ll 0$$

has typically huge absolute value, meaning that (5.5) is very stiff (the largest eigenvalue satisfies $\Lambda_1 \approx -\frac{4}{h^2}\Bigl(\underbrace{\frac{\pi}{2n+2}}_{\pi h/2}\Bigr)^2 = -\pi^2$ for $n \gg 1$).

Recalling that for the Euler's method $\mathcal{R} \cap \mathbb{R} = (-2, 0)$, it follows from Theorem 4.6 that the Euler's method is absolutely stable for (5.5) if the time step is less than $\frac{h^2}{2}$, i.e. very small.

The same conclusion holds also for other explicit methods: the time step, say $\delta > 0$, needs to be of the order $\delta \sim h^2$ to achieve stability (recall that the region of absolute stability for any explicit RKM or LMM is bounded). Hence, (5.5) should be solved by some (A-stable) implicit method.

Probably, the most common choice is the *implicit midpoint rule*, which leads to the scheme

$$(5.7) \quad \mathbb{R}^n \ni U_{k+1} = \left(I - \tfrac{1}{2}\delta c A\right)^{-1}\left(I + \tfrac{1}{2}\delta c A\right) U_k,$$

$k = 0, 1, 2, \dots$ (cf. (4.18)). $(U_0 = G)$ Here,

$$(U_k)_j \approx [U(\underbrace{k\delta}_{=:t_k})]_j \approx u(x_j, t_k), \qquad j = 1, \dots, n, \quad k = 0, 1, 2, \dots .$$

Observe that since the right-hand side of (5.5) is linear with respect to U, the trapezoidal rule

$$U_{k+1} - U_k = \frac{1}{2}\delta\left(cAU_{k+1} + cAU_k\right),$$

k = 0, 1, 2, …, also results in the scheme (5.7). [The equivalence between the implicit midpoint rule and the trapezoidal rule ceases to hold if the original problem (5.1) includes a source term; cf. Exercise 6.]

The numerical method (5.7) for solving parabolic problems is so popular that it has a special name: the *Crank-Nicolson method*. Because the spatial discretization (5.4) is of order two, as is the implicit midpoint rule (or the trapezoidal rule) employed in the time discretization of (5.5), and the iteration (5.7) is absolutely stable for any time step $\delta > 0$ (see 4/33 – 4/36 and Lemma 5.2), it could be proved that

$$(5.8) \qquad |u(x_j, t_k) - (U_k)_j| \leq \grave{C}(h^2 + \delta^2), \quad \grave{C} > 0,$$

if the intial data $g : (0,1) \to \mathbb{R}$ is smooth enough. In consequence, the errors originating from the spatial and time discretizations are of the same order; in a sense, there is no reason to implement a higher order scheme for solving (5.5), unless (5.4) is

$5/16$

replaced by a higher order approximation as well.

Remark 5.3. The multiplication by the inverse of $I - \frac{1}{2}\delta c A$ is not extremely expensive as this matrix is sparse. As the form of $I - \frac{1}{2}\delta c A$ does not change between different time steps — unless the step size $\delta > 0$ is chosen adaptively or the "thermal diffusivity" $c > 0$ is time-dependent — it may also be reasonable to compute a suitable matrix decomposition for $I - \frac{1}{2}\delta c A$ (QR or Cholesky), which makes multiplication with the inverse matrix cheap.

5/17

To conclude the discussion on discretization of parabolic problems, let us investigate how the iteration (5.7) affects different "spatial frequencies". To this end, we recall the eigenpairs $\{(\Lambda_\ell, v^\ell)\}_{\ell=1}^{n} \subset \mathbb{R}_- \times \mathbb{R}^n$ from Lemma 5.2,

$$\Lambda_\ell = -\frac{4}{h^2}\sin^2\left(\frac{\ell\pi h}{2}\right); \quad v^\ell_j = \sqrt{2h}\,\sin(j\ell\pi h),$$

$j = 1, \dots, n$, where $h = \frac{1}{n+1}$. Since the eigenvectors v^ℓ, $\ell = 1, \dots, n$, form an orthonormal basis of $\mathbb{R}^n$, the initial value $U_0 = \underset{\mathbb{R}^n}{\underset{\uparrow}{G}}$ can be given as

$$U_0 = \sum_{\ell=1}^{n} \gamma_\ell v^\ell,$$

where $\overset{\overset{\mathbb{R}}{\cup}}{\gamma_\ell} = U_0^T v^\ell$, $\ell = 1, \dots, n$, are the projections of U_0 on the "eigendirections". As a consequence, it follows from (5.7) that (cf. 4/35)

$$U_k = \sum_{\ell=1}^{n} \underbrace{\left(\frac{1+\frac{1}{2}\delta c \Lambda_\ell}{1-\frac{1}{2}\delta c \Lambda_\ell}\right)}_{R(\delta c \Lambda_\ell)}{}^{k} \gamma_\ell v^\ell \,, \qquad (5.9)$$

where R is the stability function of the implicit midpoint rule (or the trapezoidal rule). Since the implicit midpoint rule is A-stable and $\delta c \Lambda_\ell < 0$, we know that the multiplier $R(\delta c \Lambda_\ell)^k \gamma_\ell$ of each "spatial frequency" v^ℓ goes to zero as $k \to \infty$. However, the highest frequency v^ℓ does not typically experience the strongest damping (as is the case for the exact solution (5.2)):

If, motivated by (5.8), one has, e.g., chosen $\delta = h$, for the highest "spatial frequency" $l = n$ it holds that

$$R(\delta c \underset{-4/h^2}{\underset{\approx}{\Lambda_n}}) \approx \frac{1 - 2ch^{-1}}{1 + 2ch^{-1}} \approx -1 ,$$

if $h \ll 2c$. Hence, although $R(\delta c\Lambda_n)^k$ goes to zero as $k \longrightarrow \infty$, this does not happen very quickly if $(\delta =)\, h$ is small. In actual numerical computations this "slow damping of high frequencies" by the Crank-Nicolson method shows up if the initial data $g : (0,1) \longrightarrow \mathbb{R}$ has high-frequency components: for example corners in the graph of g are not smoothened out as quickly as they should be in light of (5.2).

One possible remedy for the bad behavior of high spatial frequencies of the initial data g under the C-N method (5.7) is to initiate (5.7) by taking first one step of the implicit Euler's method. This changes (5.9) into the form

$$U_k = \sum_{\ell=1}^{n} \left(\frac{1 + \frac{1}{2}\delta c \Lambda_\ell}{1 - \frac{1}{2}\delta c \Lambda_\ell} \right)^{k-1} \frac{1}{1 - \delta c \Lambda_\ell} \gamma_\ell v^\ell .$$

Because, if $\delta = h << 4c$,

$$\frac{1}{1 - \delta c \Lambda_n} \approx \frac{1}{1 + c\frac{4}{h}} \approx \frac{h}{4c} = \text{"small"},$$

the highest frequencies are damped more efficiently than in the standard C-N method.

5/21

Remark 5.4. Although we have only studied the simple model problem (5.1), most of the observed phenomena *generalize* to the case of a more general elliptic equation

$$u_t(x,t) = \nabla\cdot\big(c(x)\nabla u(x,t)\big), \quad x\in\Omega,\ t>0, \tag{5.10}$$

where $\Omega\subset\mathbb{R}^m$ is the examined spatial domain. After combining (5.10) with relevant initial and boundary conditions, suitable spatial discretization (e.g. some finite element method or difference scheme). leads to a semidiscrete problem of the type (5.5), with A depending on $c:\Omega\to\mathbb{R}_+$, the boundary conditions, the shape of Ω etc.. Be that as it may, the "general properties" of A typically remain the same as for the special case (5.4).

6. Hyperbolic PDEs

To illuminate the requirements on the (time) discretization of hyperbolic PDEs, we consider the model problem

$$(6.1) \quad \begin{cases} u_{tt}(x,t) = c^2 u_{xx}(x,t), & x \in (0,1),\ t > 0, \\ u(0,t) = u(1,t) = 0, & t > 0, \\ u(x,0) = g_1(x), \quad u_t(x,0) = g_2(x), & x \in (0,1). \end{cases}$$

One possible physical interpretation for (6.1) is as follows: An "ideal string" with fixed end points has initial shape $g_1 : (0,1) \to \mathbb{R}$ and initial velocity $g_2 : (0,1) \to \mathbb{R}$. The solution of (6.1) at time $t > 0$, i.e. $u(\cdot, t)$, represents the shape of the string at that time. The constant $c > 0$ is the "speed of waves".

Unlike for the model parabolic problem (5.1), the solution of the simplistic wave equation (6.1) is composed of infinitely oscillating components. In particular, the "energy" of the solution to (6.1) is constant over time:

<u>Lemma 6.1.</u> Any smooth enough solution $u:(0,1)\times\mathbb{R}_+\to\mathbb{R}$ of (6.1) satisfies

$$\underbrace{\|u_t(\cdot,t)\|^2_{L^2(0,1)}}_{\text{kinetic energy}} + c^2\overbrace{\|u_x(\cdot,t)\|^2_{L^2(0,1)}}^{\text{potential energy}} = \|g_2\|^2_{L^2(0,1)} + c^2\|g_1'\|^2_{L^2(0,1)}, \quad t\geq 0.$$

In particular, this total energy is constant in time.

<u>Proof.</u> To begin with, recall the definition of the norm $\|\cdot\|_{L^2(0,1)}$:

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

Obviously,

$$\frac{d}{dt}\|u_t(\cdot,t)\|_{L^2}^2 = \frac{d}{dt}\int_0^1 u_t(x,t)^2\,dx$$

$$= 2\int_0^1 u_{tt}(x,t)\,u_t(x,t)\,dx$$

and

$$\frac{d}{dt}\|u_x(\cdot,t)\|_{L^2}^2 = 2\int_0^1 \underbrace{u_{tx}(x,t)}_{u_{xt}(x,t)}\,u_x(x,t)\,dx$$

$$\overset{\text{partial integration}}{=} 2\Big[\underbrace{\Big|_0^1 u_t(x,t)u_x(x,t)}_{\substack{=0 \text{ due to}\\ \text{the boundary}\\ \text{conditions}\\ \text{of (6.1)}}} - \int_0^1 u_{xx}(x,t)\,u_t(x,t)\,dx\Big].$$

Hence, altogether

$$\frac{d}{dt}\left(\|u_t(\cdot,t)\|^2 + c^2\|u_x(\cdot,t)\|^2\right)$$

$$= 2\int_0^1 \underbrace{\left(u_{tt}(x,t) - c^2u_{xx}(x,t)\right)}_{=0} u_t(x,t)\,dx$$

$$= 0,$$

which shows that the total energy is a constant (and proves the claim).

In particular, when (6.1) is discretized, the numerical solution should also conserve some kind of "discrete energy" and be composed of infinitely continuing oscillations.

With this in mind, let us begin the discretization procedure. As for the parabolic model problem (5.1), we start by replacing the second spatial derivative in (6.1) by the matrix $A \in \mathbb{R}^{n \times n}$ from (5.4):

$$\begin{aligned} U''(t) &= c^2 A U(t), \quad t > 0, \\ U(0) &= G_1, \quad U'(0) = G_2, \end{aligned} \tag{6.2}$$

where $G_i = [g_i(x_1), \dots, g_i(x_n)]^T$, $i = 1, 2$, and $[U(t)]_j \approx u(x_j, t)$, $j = 1, 2, \dots, n$, $t \geq 0$.

Furthermore, the spatial grid points $x_j = jh$, $j=1,2,\ldots,n$, are defined as in Chapter 5, i.e. with $h = \frac{1}{n+1}$. One could now proceed by directly discretizing the second time derivative by, e.g., the standard central difference approximation, which easily leads to a two-step recursion, for which the (other) starting value $U_1 \in \mathbb{R}^n$ can be given with the help of the initial conditions of (6.2) and an origin-centered Taylor's expansion. Be that as it may, we take here a different approach and present (6.2) as a twice as large first order initial value problem (cf. Section 1.3).

To this end, we introduce a new unknown function

$$W := \begin{bmatrix} U \\ U' \end{bmatrix} : \mathbb{R}_+ \longrightarrow \mathbb{R}^{2n}.$$

A trivial calculation based on (6.2) demonstrates that

$$W'(t) = \underbrace{\begin{bmatrix} \overset{\mathbb{R}^{n\times n}}{\overset{\rotatebox{90}{$\in$}}{O}} & \overset{\mathbb{R}^{n\times n}}{\overset{\rotatebox{90}{$\in$}}{I}} \\ c^2 A & \underbrace{O}_{\in \mathbb{R}^{n\times n}} \end{bmatrix}}_{=: M \in \mathbb{R}^{2n\times 2n}} W(t), \quad t > 0, \tag{6.3}$$

which should be combined with the initial condition

$$W(0) = W_0 := \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \in \mathbb{R}^{2n}. \tag{6.4}$$

The initial value problem (6.3)-(6.4) is something that we should, in principle, be able to handle numerically based on Chapters 3 and 4. However, it is once again advisable to have first a look at the eigenpairs of the coefficient matrix $M \in \mathbb{R}^{2n \times 2n}$ in (6.3).

<u>Lemma 6.2.</u> The matrix $M \in \mathbb{R}^{2n \times 2n}$ has $2n$ distinct eigenvalues

$$\mathbb{C} \ni \mu_\ell^\pm = \pm i c \sqrt{-\Lambda_\ell}, \qquad \ell = 1, 2, \dots, n,$$

where $\Lambda_\ell < 0$, $\ell = 1, \dots, n$, are the n (negative) eigenvalues of $A \in \mathbb{R}^{n \times n}$ from Lemma 5.2. The corresponding eigenvectors are

$$w^{\ell,\pm} = \begin{bmatrix} v^\ell \\ \mu_\ell^\pm v^\ell \end{bmatrix} \in \mathbb{C}^{2n},$$

where $v^\ell \in \mathbb{R}^n$, $\ell = 1, 2, \dots, n$, are the eigenvectors of A. (numbered in the same order as Λ_ℓ).

Proof. A straightforward calculation gives

$$Mw^{l,\pm} = \begin{bmatrix} \mu_l^\pm v^l \\ \\ c^2 \underbrace{Av^l}_{\Lambda_l v^l} \end{bmatrix} = \begin{bmatrix} \mu_l^\pm v^l \\ \\ (\underbrace{\pm i c\sqrt{-\Lambda_l}}_{\mu_l^\pm})^2 v^l \end{bmatrix}$$

$$= \mu_l^\pm w^{l,\pm}, \qquad l = 1, 2, \ldots, n,$$

which proves the assertion. □

In particular, Lemma 6.2 tells us that M is diagonalizable (it has $2n$ distinct eigenvalues, resulting in a full set of eigenvectors),

$$M = WDW^{-1},$$

$\mathbb{C}^{2n\times 2n}$ $\swarrow$

where $W = [w^{1,+} w^{1,-}, w^{2,+}, w^{2,-}, \ldots, w^{n,+}, w^{n,-}]$ and $D = \operatorname{diag}(\mu_1^+, \mu_1^-, \ldots, \mu_n^+, \mu_n^-) \in \mathbb{C}^{2n\times 2n}$.

By looking at Theorem 4.6 and its proof — together with the example on pages 4/33 – 4/36 — it is easy to convince oneself that if an (explicit) RK method is applied to the initial value problem (6.3) – (6.4), the resulting numerical solution can be written in the form

$$w_j = W R(\delta D)^j W^{-1} w_0 , \quad j = 0, 1, 2, \dots , \tag{6.5}$$

where $\delta > 0$ is the time step length and

stability function of the studied RK method

$$R(\delta D) = \operatorname{diag}\big(R(\delta\mu_1^+), R(\delta\mu_1^-), \dots, R(\delta\mu_n^+), R(\delta\mu_n^-)\big).$$

Now, recall that, according to Lemma 6.1, if at least one of the two initial values g_1 and g_2 of (6.1) is not identically zero (and g_1 satisfies the boundary conditions of (6.1) so that $g_1 \not\equiv 0 \Rightarrow g_1' \not\equiv 0$), then either $u(\cdot,t)$ or $u_t(\cdot,t)$ (or both) is not identically zero for any $t>0$ (or even at "$t=\infty$"). In the discrete framework of (6.5), this means that (at the very least) the condition

$$\lim_{j\to\infty} |w_j| \neq 0$$

should be satisfied for any $0 \neq w_0 \in \mathbb{R}^{2n}$. Hence, $|R(\delta\mu_l^\pm)| \geq 1$ should hold for all $l = 1,2,\dots,n$.* On the other hand, in the light of Lemma 6.1, the behavior

$$\lim_{j\to\infty} |w_j| = \infty$$

is as unwanted, meaning that also $|R(\delta\mu_l^\pm)| \leq 1$. Putting the above arguments together, it follows that a reasonable method for solving (6.2) should satisfy

* see (6.5)

6/11

$$(6.6)\qquad |R(\delta\mu_\ell^\pm)| = 1, \qquad \ell = 1,2,\dots,n.^*$$

Let us investigate how the simplest methods fare with (6.6).

<u>Example.</u> The stability function of the Euler's method is

$$R(\hat{h}) = 1 + \hat{h}.$$

Thus,

$$|R(\delta\mu_\ell^\pm)|^2 = |1 + \delta\overset{\text{imaginary!}}{\mu_\ell^\pm}|^2 = 1 + \delta^2|\mu_\ell^\pm|^2$$

$$= 1 - \delta^2 c^2 \underbrace{\Lambda_\ell}_{<0\ \nwarrow\ \text{eigenvalue of } A} > 1,$$

for <u>all</u> $\ell = 1,2,\dots,n$, which means that for any $0 \neq w_0 \in \mathbb{R}^{2n}$, i.e. for any pair $G_1, G_2 \in \mathbb{R}^n$ for (6.2) not both identically zero,

$$\lim_{j\to\infty} |w_j| = \infty$$

for the Euler's method.

* In some situations, it is beneficial to have $|R(\delta\mu_\ell^\pm)|$ (slightly) less than one for $\ell \gg 1$.

The behavior of the numerical solution produced by the implicit Euler's method is exactly the opposite: Now

$$R(\hat{h}) = \frac{1}{1-\hat{h}} ,$$

and so

$$|R(\delta\mu_l^{\pm})|^2 = \frac{1}{|1-\delta\mu_l^{\pm}|^2} = \frac{1}{1-\delta^2c^2\Lambda_l} < 0$$

for <u>all</u> $l = 1, 2, \dots, n$. In consequence,

$$\lim_{j\to\infty} w_j = 0 \in \mathbb{R}^{2n}$$

independently of $w_0 \in \mathbb{R}^{2n}$.

Finally, let us note that the above problems can be avoided by using the <u>implicit midpoint rule</u> (or the trapezoidal rule) for which

$$R(\hat{h}) = \frac{1+\frac{1}{2}\hat{h}}{1-\frac{1}{2}\hat{h}} .$$

Hence,

$$|R(\delta\mu_\ell^\pm)|^2 = \frac{|1+\frac{1}{2}\delta\mu_\ell^\pm|^2}{|1-\frac{1}{2}\delta\mu_\ell^\pm|^2} = \frac{1-\frac{1}{2}\delta^2c^2\Lambda_\ell}{1-\frac{1}{2}\delta^2c^2\Lambda_\ell} = 1$$

for all $\ell = 1, 2, \ldots, n$, and so the numerical solution produced by the implicit midpoint rule for (6.2) keeps oscillating for eternity (without loosing or gaining energy in a certain sense).

☺

The main lesson is that the implicit midpoint rule (or the trapezoidal rule) is a reasonable technique for solving the initial value problem (6.3)–(6.4) originating from the hyperbolic model problem (6.1) via spatial discretization (and order reduction). In particular, the implicit midpoint rule conserves energy in a certain sense and, furthermore, its order of consistency $p=2$ "matches" the error in the spatial discretization when deducing (6.2); see (5.4). [The implicit midpoint rule is by no means perfect though; it suffers from *numerical dispersion* like many other numerical schemes.]

To complete this chapter — and these lecture notes — the reader should be reminded that the hyperbolic problems encountered in practice are typically *more* complicated than (6.1). However, such problems, and the associated initial value problems for ordinary differential equations, carry anyway often similar qualitative properties as (6.1) and (6.2).

THE END