Chair DYNAMICS, CONTROL AND NUMERICS FAU

Department of DATA SCIENCE

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

NATURWISSENSCHAFTLICHE FAKULTÄT

Practical Course: Modeling, Simulation, Optimization

Week 2

Daniël Veldman

Chair in Dynamics, Control, and Numerics, Friedrich-Alexander-University Erlangen-Nürnberg

Contents

2.A Review Exercise Week 1

Exercise 1

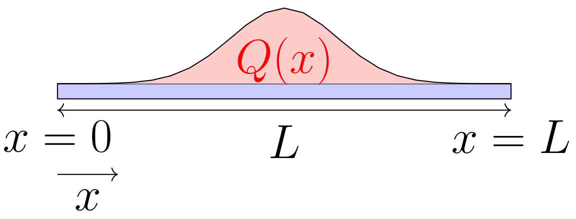


Figure: The considered aluminum rod

Consider the steady-state temperature distribution in the aluminum rod in Figure 1 with a length of $L = 0.3$ [m], a cross sectional area of $A_{\text{cs}} = 0.01$ [m^2], and a thermal conductivity of $k = 237$ [W/m/K]. Along the length of the rod, a constant heat load $Q(x) = Q_0 \exp(-(x - \frac{1}{2}L)^2/a^2)$ [W/m] is applied. The parameters for the heat load are $Q_0 = 100$ [W/m] and $a = 0.1$ [m].
The temperature increase w.r.t. a reference temperature of $T_0 = 293$ [K] is $T(x)$. At the left end of the rod, the temperature is fixed at the reference temperature T_0, i.e. $T(0) = 0$. At the right end of the rod, the (outgoing) heat flow is proportional to the temperature increase, i.e. $A_{\text{cs}} q(L) = hT(L)$ [W], where $h = 3$ [W/K] is the cooling coefficient and the outgoing heat flux is $q(L) = -k\frac{\mathrm{d}T}{\mathrm{d}x}(L)$.

Another hint for problem a.

$$\frac{\partial \rho_u}{\partial t}(t,x) = -A\frac{\partial q}{\partial x}(t,x) + Q(t,x).$$

We again need constitutive relations to complete the model.

Fourier's law of heat conduction

$$q(t,x) = -k\frac{\partial T}{\partial x}(t,x).$$

The coefficient k [W/m/K] is the thermal conductivity and $T(t,x)$ [K] is the temperature.
'Heat flows from locations with high temperatures to locations with low temperatures'

Internal energy

$$\rho_u(t,x) = cAT(t,x).$$

The coefficient c [J/K/m^3] heat capacity per unit length.

Part a: the BVP

Write down the boundary value problem for the temperature increase in the rod $T(x)$.

Solution: 1-D conservation law and Fourier's law of heat conduction:

$$\frac{\partial \rho_u}{\partial t}(t,x) = -A\frac{\partial q}{\partial x}(t,x) + Q(x), \qquad q(t,x) = -k\frac{\partial T}{\partial x}(t,x).$$

Steady-state: $\frac{\partial \rho_u}{\partial t} = 0$.

The resulting BVP:

$$A_{\mathrm{cs}}k\frac{\mathrm{d}^2T}{\mathrm{d}x^2}(x) + Q(x) = 0, \qquad Q(x) = Q_0 \exp\left(\frac{-(x-\frac{1}{2}L)^2}{a^2}\right).$$

$$T(0) = 0, \qquad A_{\mathrm{cs}}k\frac{\mathrm{d}T}{\mathrm{d}x}(L) = -hT(L).$$

Part b: finite difference discretization

$$A_{\mathrm{cs}}k\frac{\mathrm{d}^2T}{\mathrm{d}x^2}+Q(x)=0, \qquad Q(x)=Q_0\exp\left(\frac{-(x-\frac{1}{2}L)^2}{a^2}\right).$$

$$T(0)=0, \qquad A_{\mathrm{cs}}k\frac{\mathrm{d}T}{\mathrm{d}x}(L)=-hT(L).$$

$$\frac{A_{\mathrm{cs}}k}{\Delta x^2}\begin{bmatrix} 0 & \frac{\Delta x^2}{A_{\mathrm{cs}}k} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & & 0 & 0 & 0 & 0 \\ \vdots & \vdots & & & \ddots & & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{\Delta x}{2A_{\mathrm{cs}}k} & {\color{red}h\frac{\Delta x^2}{A_{\mathrm{cs}}k}} & \frac{-\Delta x}{2A_{\mathrm{cs}}k} \end{bmatrix}\begin{bmatrix} T_0 \\ T_1 \\ T_2 \\ T_3 \\ \vdots \\ T_{N-2} \\ T_{N-1} \\ T_N \\ T_{N+1} \end{bmatrix}+\begin{bmatrix} 0 \\ Q_1 \\ Q_2 \\ Q_3 \\ \vdots \\ Q_{N-2} \\ Q_{N-1} \\ Q_N \\ 0 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Part c: the analytic solution

Find a particular solution of the ODE:

$$T_{\text{part}}(x) = \frac{Q_0 a^2}{2A_{\text{cs}}k} \exp\left(\frac{-(x-\frac{1}{2}L)^2}{a^2}\right) + \frac{Q_0 a}{2A_{\text{cs}}k}\sqrt{\pi}(x-\tfrac{1}{2}L)\operatorname{erf}\left(\frac{x-\frac{1}{2}L}{a}\right),$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-y^2)\,\mathrm{d}y$.

Homogeneous solution of the ODE $\frac{\mathrm{d}^2T}{\mathrm{d}x^2} = 0$ is given by $T_{\text{hom}}(x) = Ax + B$.
Insert $T_{\text{part}}(x) + T_{\text{hom}}(x)$ into the BCs.

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} T(0) \\ A_{\text{cs}}k\frac{\mathrm{d}T}{\mathrm{d}x}(L) + hT(L) \end{bmatrix} = \begin{bmatrix} T_{\text{part}}(0) \\ A_{\text{cs}}k\frac{\mathrm{d}T_{\text{part}}}{\mathrm{d}x}(L) + hT_{\text{part}}(L) \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ A_{\text{cs}}k + hL & h \end{bmatrix}\begin{bmatrix} A \\ B \end{bmatrix}$$

$$\frac{\mathrm{d}T_{\text{part}}}{\mathrm{d}x}(x) = \frac{Q_0 a}{A_{\text{cs}}k}\frac{\sqrt{\pi}}{2}\operatorname{erf}\left(\frac{x-\frac{1}{2}L}{a}\right).$$

This is a linear system from which A and B can be determined.

Resulting solution is thus $T(x) = T_{\text{part}}(x) + Ax + B$.

2.B Time-dependent problems

Motivating example: Diffusion of mass

$$\frac{\partial \rho}{\partial t}(t,x) = -\frac{\partial \phi}{\partial x}(t,x).$$

To complete the model, we need a *constitutive relation*
that relates the mass flux $\phi(t,x)$ to the mass density $\rho(t,x)$.

We could for example use.

Fick's law

$$\phi(t,x) = -D\frac{\partial \rho}{\partial x}.$$

The coefficient D [m^2/s] is called the diffusivity.
'Mass flows from locations with high concentrations to locations with low concentrations'

We then obtain

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

Motivating example: Heat conduction

$$\frac{\partial \rho_u}{\partial t}(t,x) = -A_{\mathrm{cs}}\frac{\partial q}{\partial x}(t,x) + Q(t,x).$$

We again need constitutive relations to complete the model.

Fourier's law of heat conduction

$$q(t,x) = -k\frac{\partial T}{\partial x}(t,x).$$

The coefficient k^* [W/m/K] is the thermal conductivity and $T(t,x)$ [K] is the temperature.
'Heat flows from locations with high temperatures to locations with low temperatures'

Internal energy

$$\rho_u(t,x) = cA_{\mathrm{cs}}T(t,x).$$

The coefficient c [J/K/m^3] heat capacity per unit volume.

We thus obtain

$$cA_{\mathrm{cs}}\frac{\partial T}{\partial t}(t,x) = kA_{\mathrm{cs}}\frac{\partial^2 T}{\partial x^2}(t,x) + Q(t,x). \tag{1}$$

2.C Spatial discretization

Spatial discretization / Method of Lines (MOL) / Semi-discretization

Suppose we want to approximate the solution $u(t,x)$ of the initial value problem

$$\frac{\partial u}{\partial t}(t,x) = \kappa \frac{\partial^2 u}{\partial x^2}(t,x) + f(t,x), \qquad (t,x) \in (0,T) \times (0,L),$$

$$u(t,0) = 0, \qquad \frac{\partial u}{\partial x}(t,L) = 0, \qquad u(0,x) = u_0(x).$$

Introduce an M-point grid in the interval $[0,L]$ with a grid spacing $\Delta x = L/(M-1)$

Also introduce $f_m(t) = f(t,x_m)$ and the approximations $u_m(t) \approx u(t,x_m)$.

Finite difference discretization (implicit BCs):

$$\frac{\mathrm{d}u_m}{\mathrm{d}t}(t) = \kappa \frac{u_{m+1}(t) - 2u_m(t) + u_{m-1}(t)}{\Delta x^2} + f_m(t), \qquad m = 1,2,\ldots,M,$$

$$u_1(t) = 0, \qquad \frac{u_{M+1}(t) - u_{M-1}(t)}{2\Delta x} = 0, \qquad u_m(0) = u_0(x_m).$$

Implicit or explicit implementation of the boundary conditions

Finite difference discretization (implicit BCs):

$$\frac{\mathrm{d}u_m}{\mathrm{d}t}(t) = \kappa \frac{u_{m+1}(t) - 2u_m(t) + u_{m-1}(t)}{\Delta x^2} + f_m(t), \qquad m = 1, 2, \dots, M,$$

$$u_1(t) = 0, \qquad \frac{u_{M+1}(t) - u_{M-1}(t)}{2\Delta x} = 0, \qquad u_m(0) = u_0(x_m).$$

This is a system of Diffferential Algebraic Equations (DAEs)

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} \mathbf{u}_1(t) \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1(t) \\ \mathbf{u}_2(t) \end{bmatrix} + \begin{bmatrix} \mathbf{f}(t) \\ 0 \end{bmatrix}.$$

Implicit or explicit implementation of the boundary conditions

Finite difference discretization (implicit BCs):

$$\frac{\mathrm{d}u_m}{\mathrm{d}t}(t) = \kappa\frac{u_{m+1}(t) - 2u_m(t) + u_{m-1}(t)}{\Delta x^2} + f_m(t), \qquad m = 1, 2, \dots, M,$$

$$u_1(t) = 0, \qquad \frac{u_{M+1}(t) - u_{M-1}(t)}{2\Delta x} = 0, \qquad u_m(0) = u_0(x_m).$$

This is a system of Diffferential Algebraic Equations (DAEs)

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathbf{u}_1(t)\\ 0\end{bmatrix} = \begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1(t)\\ \mathbf{u}_2(t)\end{bmatrix} + \begin{bmatrix}\mathbf{f}(t)\\ 0\end{bmatrix}.$$

Finite difference discretization (explicit BCs):

$$\frac{\mathrm{d}u_m}{\mathrm{d}t}(t) = \kappa\frac{u_{m+1}(t) - 2u_m(t) + u_{m-1}(t)}{\Delta x^2} + f_m(t), \qquad m = 2, 3, \dots, M-1,$$

$$\frac{\mathrm{d}u_M}{\mathrm{d}t}(t) = \kappa\frac{-2u_M(t) + 2u_{M-1}(t)}{\Delta x^2} + f_M(t), \qquad u_m(0) = u_0(x_m),$$

where we should remember that $u_1(t) = 0$.
This is a system of Ordinary Differential Equations (ODEs) for the free DOFs $\mathbf{u}_{\mathrm{f}}(t)$

$$\dot{\mathbf{u}}_{\mathrm{f}}(t) = \mathbf{A}_{\mathrm{ff}}\mathbf{u}_{\mathrm{f}}(t) + \mathbf{f}_{\mathrm{f}}(t).$$

The explicit implementation of the BCs is preferred in time-dependent problems.

2.D Temporal discretization

Linear ODEs

Consider the following system of linear ODEs:

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t) = \mathbf{A}\mathbf{u}(t) + \mathbf{f}(t), \qquad \mathbf{u}(0) = \mathbf{u}_0.$$

Linear ODEs

Consider the following system of linear ODEs:

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t) = \mathbf{A}\mathbf{u}(t) + \mathbf{f}(t), \qquad \mathbf{u}(0) = \mathbf{u}_0.$$

- Choose a uniform time grid $t_0, t_1, t_2, \ldots$ with $t_k = k\Delta t$.
- Define $\mathbf{f}^k := \mathbf{f}(t_k)$ and introduce the approximations $\mathbf{u}^k \approx \mathbf{u}(t_k)$.

Linear ODEs

Consider the following system of linear ODEs:

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t) = \mathbf{A}\mathbf{u}(t) + \mathbf{f}(t), \qquad \mathbf{u}(0) = \mathbf{u}_0.$$

- Choose a uniform time grid $t_0, t_1, t_2, \ldots$ with $t_k = k\Delta t$.
- Define $\mathbf{f}^k := \mathbf{f}(t_k)$ and introduce the approximations $\mathbf{u}^k \approx \mathbf{u}(t_k)$.

By Taylor's theorem

$$\mathbf{u}(t_{k+1}) = \mathbf{u}(t_k + \Delta t) = \mathbf{u}(t_k) + \Delta t \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_k) + \frac{\Delta t^2}{2} \frac{\mathrm{d}^2\mathbf{u}}{\mathrm{d}t^2}(\tau),$$

for some $\tau \in [t_k, t_{k+1}]$. Rearranging, we find

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_k) + O(\Delta t).$$

Linear ODEs

Consider the following system of linear ODEs:

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t) = \mathbf{A}\mathbf{u}(t) + \mathbf{f}(t), \qquad \mathbf{u}(0) = \mathbf{u}_0.$$

- Choose a uniform time grid $t_0, t_1, t_2, \ldots$ with $t_k = k\Delta t$.
- Define $\mathbf{f}^k := \mathbf{f}(t_k)$ and introduce the approximations $\mathbf{u}^k \approx \mathbf{u}(t_k)$.

By Taylor's theorem

$$\mathbf{u}(t_{k+1}) = \mathbf{u}(t_k + \Delta t) = \mathbf{u}(t_k) + \Delta t \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_k) + \frac{\Delta t^2}{2}\frac{\mathrm{d}^2\mathbf{u}}{\mathrm{d}t^2}(\tau),$$

for some $\tau \in [t_k, t_{k+1}]$. Rearranging, we find

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_k) + O(\Delta t).$$

We thus find the following scheme.

Forward Euler

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = \mathbf{A}\mathbf{u}^k + \mathbf{f}^k, \qquad \mathbf{u}^0 = \mathbf{u}_0.$$

Backward Euler

Instead of making a Taylor series expansion of $\mathbf{u}(t_{k+1})$ around $t = t_k$, we can also expand $\mathbf{u}(t_k)$ in a Taylor series around $t = t_{k+1}$:

$$\mathbf{u}(t_k) = \mathbf{u}(t_{k+1} - \Delta t) = \mathbf{u}(t_{k+1}) - \Delta t \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) + \frac{\Delta t^2}{2} \frac{\mathrm{d}^2\mathbf{u}}{\mathrm{d}t^2}(\tau),$$

for some $\tau \in [t_k, t_{k+1}]$. Rearranging, we find

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) = \frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} + O(\Delta t).$$

Backward Euler

Instead of making a Taylor series expansion of $\mathbf{u}(t_{k+1})$ around $t = t_k$, we can also expand $\mathbf{u}(t_k)$ in a Taylor series around $t = t_{k+1}$:

$$\mathbf{u}(t_k) = \mathbf{u}(t_{k+1} - \Delta t) = \mathbf{u}(t_{k+1}) - \Delta t \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) + \frac{\Delta t^2}{2} \frac{\mathrm{d}^2\mathbf{u}}{\mathrm{d}t^2}(\tau),$$

for some $\tau \in [t_k, t_{k+1}]$. Rearranging, we find

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) = \frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} + O(\Delta t).$$

We thus find the following scheme.

Backward Euler

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = \mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}, \qquad \mathbf{u}^0 = \mathbf{u}_0.$$

Backward Euler

Instead of making a Taylor series expansion of $\mathbf{u}(t_{k+1})$ around $t = t_k$, we can also expand $\mathbf{u}(t_k)$ in a Taylor series around $t = t_{k+1}$:

$$\mathbf{u}(t_k) = \mathbf{u}(t_{k+1} - \Delta t) = \mathbf{u}(t_{k+1}) - \Delta t \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) + \frac{\Delta t^2}{2}\frac{\mathrm{d}^2\mathbf{u}}{\mathrm{d}t^2}(\tau),$$

for some $\tau \in [t_k, t_{k+1}]$. Rearranging, we find

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) = \frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} + O(\Delta t).$$

We thus find the following scheme.

Backward Euler

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = \mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}, \qquad \mathbf{u}^0 = \mathbf{u}_0.$$

Updates with forward and backward Euler:

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \Delta t(\mathbf{A}\mathbf{u}^k + \mathbf{f}^k), \qquad \mathbf{u}^{k+1} = (\mathbf{I} - \Delta t\mathbf{A})^{-1}(\mathbf{u}^k + \Delta t\mathbf{f}^{k+1}).$$

In backward Euler we need to solve a system of linear equations in every time step. Forward Euler is an *explicit scheme*, backward Euler is an *implicit scheme*.

θ-schemes

From the previous two slides, we have

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_k) + O(\Delta t), \qquad \frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) + O(\Delta t).$$

Take a convex combination (with $\theta \in [0,1]$)

$$(1-\theta+\theta)\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = (1-\theta)\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_k) + \theta\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_{k+1}) + O(\Delta t).$$

θ-scheme

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}^k + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}\right), \qquad \mathbf{u}^0 = \mathbf{u}_0.$$

For $\theta = 1/2$, we find the Crank-Nicolson scheme.

Crank-Nicolson

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = \tfrac{1}{2}\left(\mathbf{A}\mathbf{u}^k + \mathbf{f}^k\right) + \tfrac{1}{2}\left(\mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}\right), \qquad \mathbf{u}^0 = \mathbf{u}_0.$$

Convergence analysis

Two ingredients:

1) ODE with continuous solution $u(t)$.

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

2) Discrete numerical scheme

$$\mathbf{F}_{\Delta t}((\mathbf{u}^k)_k) = 0.$$

Theorem (Lax)

The numerical scheme is convergent if it is both

- *consistent and*
- *stable.*

Definition (Consistent numerical scheme)

The numerical scheme is consistent iff $\mathbf{F}_{\Delta t}((\mathbf{u}(t_k))_k) = O((\Delta t)^p)$ for some $p > 0$.

Definition (Stable numerical scheme)

The numerical scheme is stable iff there exists a constant K independent of Δt such that $\|\mathbf{u}^k - \mathbf{u}(t_k)\| \leq K \|\mathbf{F}_{\Delta t}((\mathbf{u}(t_k))_k)\|$

Consistency

The computations on the previous slide already show that

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}(t_k) + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}(t_{k+1}) + \mathbf{f}^{k+1}\right) + O(\Delta t).$$

But for the Crank-Nicolson scheme ($\theta = \frac{1}{2}$) we can do better

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = \frac{1}{2}\left(\mathbf{A}\mathbf{u}(t_k) + \mathbf{f}^k\right) + \frac{1}{2}\left(\mathbf{A}\mathbf{u}(t_{k+1}) + \mathbf{f}^{k+1}\right) + O((\Delta t)^2).$$

(Exercise: check this using Taylor series expansions)

Proving stability (1/2)

We have

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}(t_k) + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}(t_{k+1}) + \mathbf{f}^{k+1}\right) + \mathbf{r}_k.$$

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}^k + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}\right), \qquad \mathbf{u}(t_0) = \mathbf{u}^0 = \mathbf{u}_0,$$

where the residues $\mathbf{r}_k$ are $O(\Delta t)$ (or $O((\Delta t)^2)$ if $\theta = \frac{1}{2}$).

Proving stability (1/2)

We have

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}(t_k) + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}(t_{k+1}) + \mathbf{f}^{k+1}\right) + \mathbf{r}_k.$$

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}^k + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}\right), \qquad \mathbf{u}(t_0) = \mathbf{u}^0 = \mathbf{u}_0,$$

where the residues $\mathbf{r}_k$ are $O(\Delta t)$ (or $O((\Delta t)^2)$ if $\theta = \frac{1}{2}$).

Introduce $\mathbf{e}^k := \mathbf{u}^k - \mathbf{u}(t_k)$ and subtract the first equation from the second:

$$\frac{\mathbf{e}^{k+1} - \mathbf{e}^k}{\Delta t} = (1-\theta)\mathbf{A}\mathbf{e}^k + \theta\mathbf{A}\mathbf{e}^{k+1} - \mathbf{r}_k, \qquad \mathbf{e}^0 = 0.$$

Proving stability (1/2)

We have

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k)}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}(t_k) + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}(t_{k+1}) + \mathbf{f}^{k+1}\right) + \mathbf{r}_k.$$

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}^k + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}\right), \qquad \mathbf{u}(t_0) = \mathbf{u}^0 = \mathbf{u}_0,$$

where the residues $\mathbf{r}_k$ are $O(\Delta t)$ (or $O((\Delta t)^2)$ if $\theta = \frac{1}{2}$).

Introduce $\mathbf{e}^k := \mathbf{u}^k - \mathbf{u}(t_k)$ and subtract the first equation from the second:

$$\frac{\mathbf{e}^{k+1} - \mathbf{e}^k}{\Delta t} = (1-\theta)\mathbf{A}\mathbf{e}^k + \theta\mathbf{A}\mathbf{e}^{k+1} - \mathbf{r}_k, \qquad \mathbf{e}^0 = 0.$$

Rearranging shows that

$$(\mathbf{I} - \theta\Delta t\mathbf{A})\mathbf{e}^{k+1} = (1-\theta)\Delta t\mathbf{A}\mathbf{e}^k - \mathbf{r}_k$$

$$\mathbf{e}^{k+1} = \mathbf{B}\mathbf{e}^k - \Delta t\mathbf{b}_k, \qquad \mathbf{e}^0 = 0,$$

where

$$\mathbf{B} = (\mathbf{I} - \theta\Delta t\mathbf{A})^{-1}\left(\mathbf{I} + (1-\theta)\Delta t\mathbf{A}\right), \qquad \mathbf{b}_k = (\mathbf{I} - \theta\Delta t\mathbf{A})^{-1}\mathbf{r}_k.$$

Note that $\mathbf{b}_k = O(\Delta t)$ (or $O((\Delta t)^2)$ if $\theta = 1/2$).

Proving stability (2/2)

$$\mathbf{e}^{k+1} = \mathbf{B}\mathbf{e}^k - \Delta t \mathbf{b}_k, \qquad \mathbf{e}^0 = 0,$$

When $\|\mathbf{B}\| > 1$, the scheme is clearly unstable.

Assume that $\|\mathbf{B}\| \leq 1$, then

$$|\mathbf{e}^{k+1}| \leq |\mathbf{e}^k| + \Delta t |\mathbf{b}_k|, \qquad \Rightarrow \qquad |\mathbf{e}^k| \leq \Delta t \sum_{k=0}^{k-1} |\mathbf{b}_k| \leq Ck(\Delta t)^2,$$

where it was used that $\mathbf{b}_k$ is $O(\Delta t)$, i.e. there exists a C such that $|\mathbf{b}_k| \leq C\Delta t$.

Proving stability (2/2)

$$\mathbf{e}^{k+1} = \mathbf{B}\mathbf{e}^k - \Delta t \mathbf{b}_k, \qquad \mathbf{e}^0 = 0,$$

When $\|\mathbf{B}\| > 1$, the scheme is clearly unstable.

Assume that $\|\mathbf{B}\| \leq 1$, then

$$|\mathbf{e}^{k+1}| \leq |\mathbf{e}^k| + \Delta t |\mathbf{b}_k|, \qquad \Rightarrow \qquad |\mathbf{e}^k| \leq \Delta t \sum_{k=0}^{k-1} |\mathbf{b}_k| \leq Ck(\Delta t)^2,$$

where it was used that $\mathbf{b}_k$ is $O(\Delta t)$, i.e. there exists a C such that $|\mathbf{b}_k| \leq C\Delta t$.

So the error after a *fixed number of k time-steps* is of $O((\Delta t)^2)$.
However, the error at a fixed time-instant T, i.e. the error after $K = T/\Delta t$ is

$$|\mathbf{e}^K| = CK(\Delta t)^2 = CT\Delta t = O(\Delta t).$$

Stability regions

Recall that

$$\mathbf{B} = (\mathbf{I} - \theta \Delta t \mathbf{A})^{-1} (\mathbf{I} + (1-\theta)\Delta t \mathbf{A}) \, .$$

Suppose that $\mathbf{v}$ is an eigenvalue of $\mathbf{A}$, i.e. that $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$. Then also

$$\mathbf{B}\mathbf{v} = \frac{1 + (1-\theta)\lambda\Delta t}{1 - \theta\lambda\Delta t}\mathbf{v}.$$

Stability regions

Recall that

$$\mathbf{B} = (\mathbf{I} - \theta \Delta t \mathbf{A})^{-1} (\mathbf{I} + (1-\theta)\Delta t \mathbf{A}) .$$

Suppose that $\mathbf{v}$ is an eigenvalue of $\mathbf{A}$, i.e. that $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$. Then also

$$\mathbf{B}\mathbf{v} = \frac{1 + (1-\theta)\lambda\Delta t}{1 - \theta\lambda\Delta t}\mathbf{v}.$$

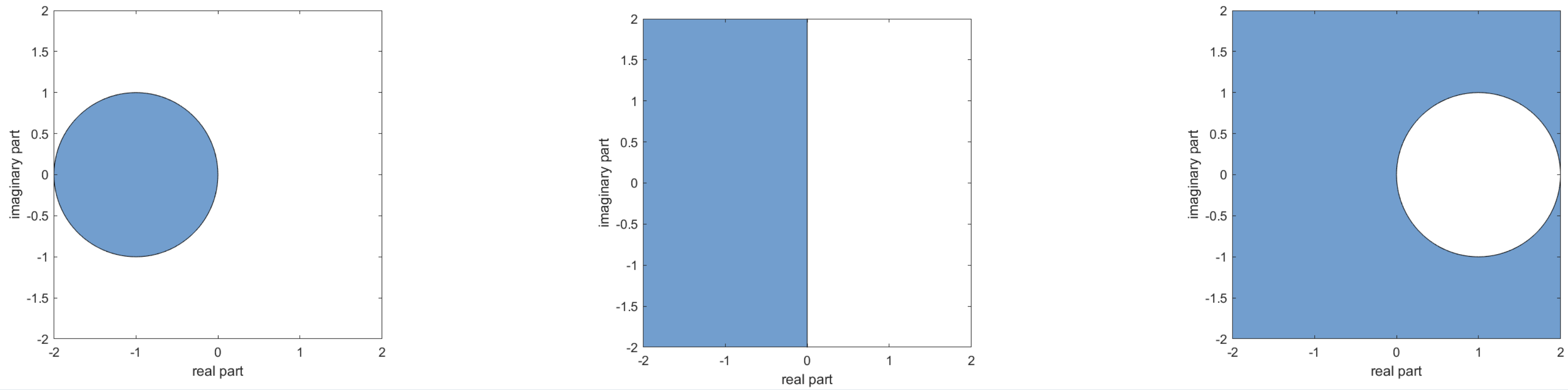
The scheme is thus stable when

$$\left| \frac{1 + (1-\theta)\lambda\Delta t}{1 - \theta\lambda\Delta t} \right| \leq 1, \qquad \text{for all } \lambda \in \sigma(\mathbf{A}).$$

Forward Euler ($\theta = 0$) — $|1 + \lambda\Delta t| \leq 1$

Crank-Nicolson ($\theta = \frac{1}{2}$) — $|1 + \frac{1}{2}\lambda\Delta t| \leq |1 - \frac{1}{2}\lambda\Delta t|$

Backward Euler ($\theta = 1$) — $|1 - \lambda\Delta t| \geq 1$



Summary

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t) = \mathbf{A}\mathbf{u}(t) + \mathbf{f}(t), \qquad \mathbf{u}(0) = \mathbf{u}_0.$$

θ-scheme

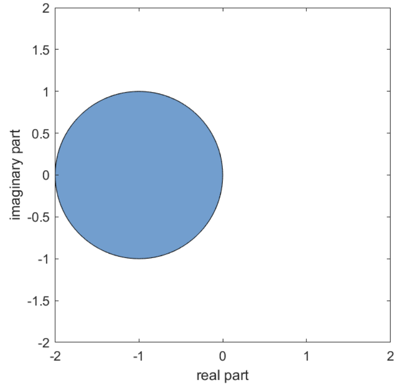
$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} = (1-\theta)\left(\mathbf{A}\mathbf{u}^k + \mathbf{f}^k\right) + \theta\left(\mathbf{A}\mathbf{u}^{k+1} + \mathbf{f}^{k+1}\right), \qquad \mathbf{u}^0 = \mathbf{u}_0.$$

The scheme is stable iff

$$|1 + (1-\theta)\lambda\Delta t| \leq |1 - \theta\lambda\Delta t|, \qquad \text{for all } \lambda \in \sigma(\mathbf{A}).$$

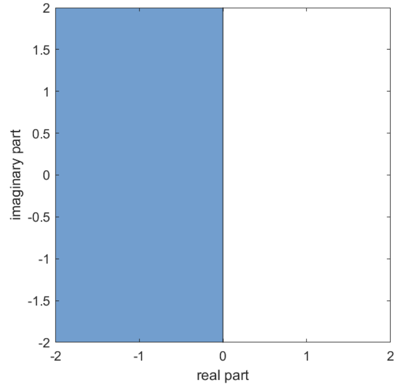
Forward Euler ($\theta = 0$)

$$|1 + \lambda\Delta t| \leq 1$$



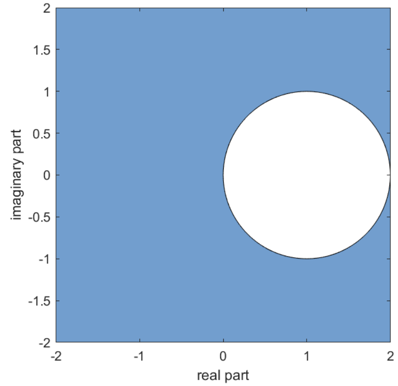
Crank-Nicolson ($\theta = \frac{1}{2}$)

$$|1 + \tfrac{1}{2}\lambda\Delta t| \leq |1 - \tfrac{1}{2}\lambda\Delta t|$$



Backward Euler ($\theta = 1$)

$$|1 - \lambda\Delta t| \geq 1$$



Chair
DYNAMICS, CONTROL
AND NUMERICS
FAU

Department of
DATA SCIENCE

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
NATURWISSENSCHAFTLICHE
FAKULTÄT

2.E Back to the spatial discretization

Returning to our original problem

Suppose we want to approximate the solution $u(t,x)$ of the initial value problem

$$
\begin{aligned}
&\frac{\partial u}{\partial t}(t,x) = \kappa \frac{\partial^2 u}{\partial x^2}(t,x) + f(t,x), && (t,x) \in (0,T) \times (0,L), \\
&u(t,0) = 0, \qquad \frac{\partial u}{\partial x}(t,L) = 0, && u(0,x) = u_0(x).
\end{aligned}
$$

Introduce an M-point grid in the interval $[0,L]$ with a grid spacing $\Delta x = L/(M-1)$

Also introduce $f_m(t) = f(t,x_m)$ and the approximation $u_m(t) \approx u(t,x_m)$.
Finite difference discretization (explicit BCs):

$$
\begin{aligned}
&\frac{\mathrm{d}u_m}{\mathrm{d}t}(t) = \kappa \frac{u_{m+1}(t) - 2u_m(t) + u_{m-1}(t)}{\Delta x^2} + f_m(t), && m = 2, 3, \ldots, M-1, \\
&\frac{\mathrm{d}u_M}{\mathrm{d}t}(t) = \kappa \frac{-2u_M(t) + 2u_{M-1}(t)}{\Delta x^2} + f_M(t), && u_m(0) = u_0(x_m),
\end{aligned}
$$

where we should remember that $u_1(t) = 0$.

Returning to our original problem

Finite difference discretization (explicit BCs):

$$\frac{\mathrm{d}u_m}{\mathrm{d}t}(t) = \kappa \frac{u_{m+1}(t) - 2u_m(t) + u_{m-1}(t)}{\Delta x^2} + f_m(t), \qquad m = 2, 3, \ldots, M-1,$$

$$\frac{\mathrm{d}u_M}{\mathrm{d}t}(t) = \kappa \frac{-2u_M(t) + 2u_{M-1}(t)}{\Delta x^2} + f_M(t), \qquad u_m(0) = u_0(x_m),$$

where we should remember that $u_1(t) = 0$.
This is a system of Ordinary Diffferential Equations (ODEs) for the free DOFs $\mathbf{u}_\mathrm{f}(t)$

$$\dot{\mathbf{u}}_\mathrm{f}(t) = \mathbf{A}_\mathrm{ff}\mathbf{u}_\mathrm{f}(t) + \mathbf{f}_\mathrm{f}(t).$$

$$\mathbf{A}_\mathrm{ff} = \frac{\kappa}{\Delta x^2}\begin{bmatrix} -2 & 1 & 0 & & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & & -2 & 1 & 0 \\ 0 & 0 & 0 & & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 2 & -2 \end{bmatrix}.$$

Note: $\mathbf{A}_\mathrm{ff}$ depends on Δx,

The stability of the numerical scheme may thus depend on Δt and Δx!

A first observation

Claim: All eigenvalues of $\mathbf{A}_{\mathrm{ff}}$ are nonpositive.

Conclusion:
The Crank-Nicolson and Backward Euler scheme are stable (for all Δx and Δt).

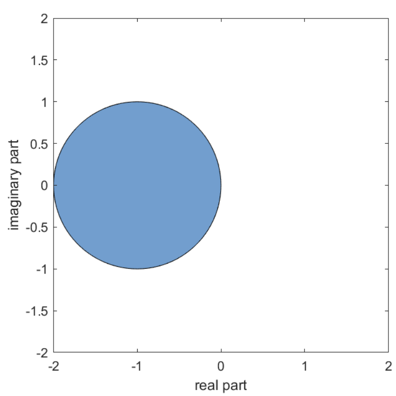
What about Forward Euler?

In the lecture next week, we will see how we can prove that

$$\sigma(\mathbf{A}_{\mathrm{ff}}) \subset [\tfrac{-4\kappa}{(\Delta x)^2}, 0]$$

Forward Euler ($\theta = 0$)

$$|1 + \lambda \Delta t| \leq 1$$



The Forward Euler scheme is stable when

$$|1 + \tfrac{-4\kappa}{(\Delta x)^2}\Delta t| \leq 1$$

$$1 + \tfrac{-4\kappa}{(\Delta x)^2}\Delta t \leq 1, \qquad \text{and} \qquad -\left(1 + \tfrac{-4\kappa}{(\Delta x)^2}\Delta t\right) \leq 1$$

$$\tfrac{-4\kappa}{(\Delta x)^2}\Delta t \leq 0, \qquad \text{and} \qquad \tfrac{4\kappa}{(\Delta x)^2}\Delta t \leq 2$$

Conclusion:
The Forward Euler scheme is stable when

$$\Delta t \leq \tfrac{1}{2\kappa}(\Delta x)^2$$

A nice trick for Finite Differences with Forward Euler

We consider

$$\frac{\partial u}{\partial t}(t,x) = \kappa \frac{\partial u^2}{\partial x^2}(t,x).$$

Finite differences+Forward Euler:

$$\frac{u_m^{k+1} - u_m^k}{\Delta t} = \kappa \frac{u_{m+1}^k - 2u_m^k + u_{m-1}^k}{(\Delta x)^2}$$

This scheme is of $O(\Delta t + (\Delta x)^2)$.

However, when we check the consistency error we see that

$$\frac{u(t_{k+1},x_m) - u(t_k,x_m)}{\Delta t} = \frac{\partial u}{\partial t}(t_k,x_m) + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(t_k,x_m) + O((\Delta t)^2)$$

$$\kappa \frac{u(t_k,x_{m+1}) - 2u(t_k,x_m) + u(t_k,x_{m-1})}{(\Delta x)^2} = \kappa \frac{\partial^2 u}{\partial x^2}(t_k,x_m) + \kappa \frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(t_k,x_m) + O((\Delta x)^4)$$

Note that $\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial t^2}(t_k,x_m) = \kappa^2 \frac{\partial^4 u}{\partial x^4}(t_k,x_m)$.
When $\Delta t = \frac{1}{6\kappa}(\Delta x)^2$ we get $O((\Delta t)^2 + (\Delta x)^4)$!
(But you need to discretize the BCs with the same rates...)