Numerics of partial differential equations: instationary problems

J.M. Melenk, M. Faustmann

Institute of analysis and scientific computing
TU Wien
SS 2024

Contents

1 Eigenvalue problems **1**
- 1.1 Solvability of the abstract eigenvalue problem . . . 1
  - 1.1.1 Characterization of eigenvalues . . . 5
- 1.2 FEM discretization . . . 7
  - 1.2.1 Convergence of the eigenvalues . . . 8
  - 1.2.2 Convergence of eigenfunctions . . . 11
  - 1.2.3 Remarks . . . 13
  - 1.2.4 A more general approach using operator theory . . . 13
- 1.3 Numerical solution of the algebraic eigenvalue problem . . . 15
  - 1.3.1 Classical techniques . . . 15
  - 1.3.2 Preconditioned gradient method . . . 16
  - 1.3.3 Numerical example . . . 18
- 1.4 Weyl's asymptotic formula . . . 18
  - 1.4.1 Example: Rectangles . . . 19
  - 1.4.2 Monotonicity properties . . . 19
  - 1.4.3 Proof of Weyl's formula . . . 20
  - 1.4.4 Can you hear the shape of a drum? . . . 22

2 Parabolic problems **24**
- 2.1 Variational formulation in space . . . 25
- 2.2 Semi-discretization (method of lines) . . . 27
- 2.3 Fully discrete methods . . . 30
  - 2.3.1 Time discretization . . . 30
  - 2.3.2 The implicit Euler method . . . 32
  - 2.3.3 The θ-scheme . . . 35
  - 2.3.4 Stability of θ-scheme—the CFL-condition . . . 36
  - 2.3.5 Numerical examples . . . 37
- 2.4 Weakening the notion of solution . . . 39
  - 2.4.1 A short introduction to Bochner spaces . . . 40
  - 2.4.2 Weak derivatives in Bochner spaces . . . 41
  - 2.4.3 Weak formulation for the heat equation . . . 42
- 2.5 Numerical approximation for non-smooth initial data . . . 44
  - 2.5.1 Smoothing property . . . 44
  - 2.5.2 Reduction to the analysis of the Ritz-projection error . . . 45
  - 2.5.3 Convergence of semi-discretization for incompatible data . . . 47
  - 2.5.4 Time-discretization error for incompatible data . . . 48
- 2.6 Discontinuous Galerkin . . . 51
- 2.7 Space-time formulations . . . 55
  - 2.7.1 Variational formulation and solvability . . . 55
  - 2.7.2 A space-time least squares formulation . . . 57
- 2.8 Approximation for the Navier Stokes equations . . . 61
  - 2.8.1 Splitting methods . . . 62
  - 2.8.2 Splitting-schemes for Navier-Stokes . . . 64
  - 2.8.3 Numerical example . . . 66

3 Hyperbolic equations **67**
3.1 Popular examples . . . 67
3.2 The wave equation . . . 69
3.2.1 Semi-discretization . . . 69
3.2.2 Two formulations as first order system . . . 70
3.2.3 Time-stepping methods for wave equations . . . 71
3.2.4 Numerical example . . . 75
3.3 The advection equation – finite differences . . . 76
3.3.1 FD for the advection equation . . . 76
3.3.2 Upwinding . . . 80
3.3.3 Splitting methods . . . 81
3.4 von Neumann-analysis . . . 82
3.4.1 The leap frog method . . . 84
3.5 Dissipative methods . . . 86
3.5.1 Preliminary view . . . 86
3.5.2 Dissipative methods . . . 87
3.6 Space-time-DG . . . 91
3.6.1 Numerical example . . . 96
3.7 DG and time-stepping . . . 96
3.7.1 Numerical example . . . 98

Literature **98**

Chapter 1

Eigenvalue problems

Motivation: many problems in natural science and engineering lead to problems of finding eigenvalues, e.g. in structural mechanics (eigenfrequencies of elastic bodies) or in quantum mechanics, ...

Goal: numerical approximation of eigenvalues and eigenfunctions.
model problem: Given a domain $\Omega \subset \mathbb{R}^d$, find a function u and $\lambda \in \mathbb{R}$, such that

$$-\Delta u = \lambda u \quad \text{on } \Omega, \qquad u|_{\partial\Omega} = 0. \tag{1.1}$$

Remark 1.1 The eigenvalue problem (EVP) (1.1) has physical meaning: The eigenvalues in λ in (1.1) are indeed positive, and the values $\sqrt{\lambda}$ correspond to the eigenfrequencies of a clamped membrane. ∎

We consider problem (1.1) in a weak, variational sense

$$\text{Find } (u,\lambda) \in H_0^1(\Omega) \setminus \{0\} \times \mathbb{R} \text{ s.t.} \quad \int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega uv \qquad \forall v \in H_0^1(\Omega). \tag{1.2}$$

Example 1.2 Let $\Omega = (0,\pi)$ and consider

$$-u'' = \lambda u \quad \text{on } \Omega, \qquad u(0) = u(\pi) = 0.$$

The general solution to the differential equation $-u'' - \lambda u = 0$ is $u(x) = C_1 \sin(\sqrt{\lambda}x) + C_2 \cos(\sqrt{\lambda}x)$. In order to obtain non-trivial solutions, there must hold $\sqrt{\lambda} = n$, $n \in \mathbb{N}$. Consequently, we obtained the eigenpairs (u_n, λ_n)

$$u_n(x) = \sin(\sqrt{\lambda_n}x), \qquad \lambda_n = n^2, \qquad n = 1, 2, \ldots.$$

Observation: the eigenfunctions satisfy two orthogonalities

$$\begin{aligned} (u_n, u_m)_{L^2(\Omega)} &= 0 \qquad n \neq m, \\ (u_n, u_m)_{H_0^1(\Omega)} := \int_\Omega u_n' u_m' &= 0 \qquad n \neq m. \end{aligned}$$

Moreover, $(u_n)_{n\in\mathbb{N}}$ is an orthogonal basis of $L^2(\Omega)$ (and, as we will see later on, also in $H_0^1(\Omega)$). ∎

1.1 Solvability of the abstract eigenvalue problem

In the present chapter, we consider a more general setting:

- V, H Hilbert spaces (over $\mathbb{R}$)
- $V \subset H$ dense with continuous, compact embedding
- $a : V \times V \to \mathbb{R}$ symmetric, continuous, bilinear, coercive
- $(\cdot,\cdot)_H \ldots$ scalar product on H

- $a(\cdot,\cdot)$ induces an inner product on V, which is equivalent to the inner product on V.

The abstract eigenvalue problem is then given as

$$\text{Find } (u,\lambda) \in V \setminus \{0\} \times \mathbb{R} \quad \text{s.t.} \qquad a(u,v) = \lambda (u,v)_H \qquad \forall v \in V. \tag{1.3}$$

To obtain solvability of the problem, we use the theory of compact operators. Let $T : H \to V$ be the solution operator for the problem

$$\text{Find } u \in V \text{ s.t.} \qquad a(u,v) = (f,v)_H \qquad \forall v \in V, \tag{1.4}$$

meaning that $Tf \in V$ is characterized by

$$a(Tf,v) = (f,v)_H \qquad \forall v \in V. \tag{1.5}$$

As $V \subset H$, we can restrict T to V and consider it as an operator in V. We have:

Theorem 1.3 *(i) $T : V \to V$ is compact.*

(ii) $T : V \to V$ is self-adjoint w.r.t. $a(\cdot,\cdot)$.

(iii) T is positive on V in the sense $a(Tf,f) > 0$ for all $f \in V \setminus \{0\}$.

Proof: *ad (i):* The operator $T : V \to V$ is a composition of the compact embedding $V \subset H$ and the continuous operator $T : H \to V$ (Lax-Milgram!):

$$T : V \hookrightarrow H \to V$$

Therefore, $T : V \to V$ is compact as composition of a continuous and compact operator.
ad (ii): Let f, $v \in V$. Then,

$$a(Tf,v) = (f,v)_H = (v,f)_H = a(Tv,f) \overset{a \text{ sym.}}{=} a(f,Tv).$$

ad (iii):

- For $f \in V \setminus \{0\}$, there holds $a(Tf,f) = (f,f)_H \geq 0$.
- As the embedding $V \subset H$ is injective, $\|f\|_H = 0$ would imply that $f = 0$ (as element in V).

□

The variational eigenvalue problem (1.3) is equivalent to an EVP for T:

Lemma 1.4 $(u,\lambda) \in V \setminus \{0\} \times \mathbb{R} \setminus \{0\}$ *solves (1.3)* $\Longleftrightarrow$ $(u,\lambda) \in V \setminus \{0\} \times \mathbb{R} \setminus \{0\}$ *solves*

$$Tu = \frac{1}{\lambda} u. \tag{1.6}$$

Proof: "$\Longrightarrow$": Let (u,λ) be a solution to (1.3). Then,

$$a(u,v) = \lambda (u,v)_H = \lambda a(Tu,v) = a(\lambda Tu, v) \qquad \forall v \in V.$$

As $a(\cdot,\cdot)$ is an inner product, we have

$$u = \lambda Tu.$$

"$\Longleftarrow$": Let (1.6) hold. Then, for all $v \in V$

$$a(\lambda^{-1} u, v) = a(Tu,v) = (u,v)_H,$$

which is (1.3). □

Lemma 1.4 provides that the sought eigenpairs (u,λ) of (1.3) are given as eigenpairs $(u, 1/\lambda)$ of T. As T is compact and self-adjoint (w.r.t. $a(\cdot,\cdot)$), we can apply spectral theory for compact operators.

Theorem 1.5 (Spectral theorem for compact, self-adjoint operators) *Let X be a Hilbert space over $\mathbb{C}$ and $A : X \to X$ be compact and self-adjoint. Then, there holds:*

(i) The spectrum $\sigma(A) = \{\mu \in \mathbb{C} \,|\, \mu - A : X \to X$ is not continuously invertible$\}$ is a countable set with the only possible accumulation point being 0. The elements $\mu \in \sigma(A) \setminus \{0\}$ are called eigenvalues.

(ii) $\sigma(A) \subset \mathbb{R}$. If A is non-negative, i.e. $(Ax, x)_X \geq 0$ for all x, then $\sigma(A) \subset [0, \|A\|_X] \subset [0, \infty)$.

(iii) For all $\mu \in \sigma(A) \setminus \{0\}$ the dimension of $\operatorname{Ker}(\mu - A)$ is finite. The number $\dim \operatorname{Ker}(\mu - A) \in \mathbb{N}$ is called multiplicity[1] *of the eigenvalue μ. The space $\operatorname{Ker}(\mu - A)$ is called* eigenspace *to the eigenvalue μ.*

(iv) For μ_1, $\mu_2 \in \sigma(A) \setminus \{0\}$ with $\mu_1 \neq \mu_2$ and $u_1 \in \operatorname{Ker}(\mu_1 - A)$ and $u_2 \in \operatorname{Ker}(\mu_2 - A)$ there holds $(u_1, u_2)_X = 0$.

(v) X has an ONB consisting of eigenvectors of A. More precisely: Let the countable set $\sigma(A) \setminus \{0\}$ be written as sequence $\mu_1, \mu_2, \ldots,$, where every eigenvalue is written possibly multiple times according to its multiplicity. Then, there exists a sequence $(e_n)_n \subset X$ with

$$(e_n, e_m)_X = \delta_{n,m}$$

and $X = \operatorname{Ker} A \oplus_X \operatorname{span}\{e_1\} \oplus_X \operatorname{span}\{e_2\} \oplus_X \cdots$.

(vi) Every $x \in X$ can be written as Fourier series

$$x = a + \sum_{n=1}^{\infty} (x, e_n)_X e_n,$$

where $a \in \operatorname{Ker} A$. Moreover,

$$Ax = \sum_{n=1}^{\infty} \mu_n (x, e_n)_X e_n.$$

Proof: A real version of this theorem will be shown in the exercise. □

Corollary 1.6 *There exists a sequence $(u_n, \lambda_n)_{n \in \mathbb{N}} \subset V \setminus \{0\} \times \mathbb{R}$ with the following properties:*

(i) $a(u_n, v) = \lambda_n (u_n, v)_H \quad \forall v \in V$.

(ii) The sequence $(\lambda_n)_{n \in \mathbb{N}}$ satisfies $0 < \lambda_1 \leq \lambda_2 \leq \cdots$ and $\lim_{n \to \infty} \lambda_n = \infty$. Moreover, every eigenvalues of (1.3) has finite multiplicity.

(iii) $(u_n)_{n \in \mathbb{N}}$ is an ONB in H. In particular, every $u \in H$ can be written as

$$u = \sum_{n=1}^{\infty} (u, u_n)_H u_n.$$

(iv) $(\lambda_n^{-1/2} u_n)_{n \in \mathbb{N}}$ is an ONB in $(V, a(\cdot, \cdot))$ In particular, every $u \in V$ can be written as

$$u = \sum_{n=1}^{\infty} a(u, \lambda_n^{-1/2} u_n) \lambda_n^{-1/2} u_n.$$

[1] more precisely: $r_g := \dim \operatorname{Ker}(\mu - A)$ is the *geometric* multiplicity of μ. The *smallest* number $\alpha \in \mathbb{N}$ with $\operatorname{Ker}(\mu - A)^\alpha = \operatorname{Ker}(\mu - A)^{\alpha+1}$ is called the *ascent* of μ. The *algebraic* multiplicity of the eigenvalue μ is then defined as $r_a := \dim \operatorname{Ker}(\mu - A)^\alpha$. Obviously, $r_a \geq r_g$. In the case of self-adjoint operators, there holds $\alpha = 1$, such that $r_a = r_g$. This follows from: If $\alpha \geq 2$, then there exists $x \neq 0$ with $x \in \operatorname{Ker}(\mu - A)^\alpha$ and $x \notin \operatorname{Ker}(\mu - A)^{\alpha-1}$. As $\alpha \geq 2$, we can consider $(\mu - A)^{\alpha-2} x$ and using the self-adjointness of $\mu - A$, there holds $0 = ((\mu - A)^\alpha x, (\mu - A)^{\alpha-2} x)_X = ((\mu - A)^{\alpha-1} x, (\mu - A)^{\alpha-1} x)_X = \|(\mu - A)^{\alpha-1} x\|_X^2 > 0$, a contradiction.

Proof: We apply Theorem 1.5 for the compact operator T and $X = (V, a(\cdot,\cdot))$.
Step 1: We show $\operatorname{Ker} T = \{0\}$. To see this, let $u \in \operatorname{Ker} T$. Then,

$$0 = a(0, v) = a(Tu, v) = (u, v)_H \qquad \forall v \in V.$$

As V is dense in H, there follows $u = 0$ (as element of H and due to the injectivity of the embedding $V \subset H$ also as element of V).
Step 2: Let $(e_n, \mu_n)_{n\in\mathbb{N}}$ be the (normalized) eigenpairs of the operator T provided by Theorem 1.5, i.e.,

a) $Te_n = \mu_n e_n$ for all $n \in \mathbb{N}$.

b) $(e_n)_{n\in\mathbb{N}}$ is an ONB of $(V, a(\cdot,\cdot))$.

With $\lambda_n := 1/\mu_n$, there holds (compare Lemma 1.4):

$$a(e_n, v) = \lambda_n (e_n, v)_H \qquad \forall v \in V.$$

Step 3: We claim that the $(e_n)_{n\in\mathbb{N}}$ are also pairwise orthogonal w.r.t. $(\cdot,\cdot)_H$:

$$(e_n, e_m)_H = \lambda_n (e_n, \lambda_n^{-1} e_m)_H = a(e_n, \lambda_n^{-1} e_m) = \lambda_n^{-1} \delta_{n,m}.$$

Step 4: Now, we define the functions u_n in the statement of the theorem as

$$u_n := \sqrt{\lambda_n}\, e_n.$$

Then,

- $(u_n, u_m)_H = \delta_{n,m}$.
- $a(u_n, u_m) = \lambda_n \delta_{n,m}$.
- $(\lambda_n^{-1/2} u_n)_{n\in\mathbb{N}}$ is an ONB of $(V, a(\cdot,\cdot))$.

It remains to show that the span of $(u_n)_{n\in\mathbb{N}}$ is dense in H. Let $\Pi_N : H \to \operatorname{span}\{u_1, \dots, u_N\}$ be the orthogonal projection (in H). We claim

$$\lim_{N\to\infty} \|u - \Pi_N u\|_H = 0.$$

For every $u \in V$ there holds

$$u = \sum_{n=1}^{\infty} a(u, \lambda_n^{-1/2} u_n) \lambda_n^{-1/2} u_n = \sum_{n=1}^{\infty} (u, u_n)_H u_n$$
$$\Pi_N u = \sum_{n=1}^{N} (u, u_n)_H u_n = \sum_{n=1}^{N} a(u, \lambda_n^{-1/2} u_n) \lambda_n^{-1/2} u_n.$$

This means that the projection $\Pi_N u$ coincides with the truncated (orthogonal-)basis expansion of u in the space $(V, a(\cdot,\cdot))$! This directly implies

$$\lim_{N\to\infty} \|u - \Pi_N u\|_V = 0 \qquad \forall u \in V.$$

A density argument now also gives the result for $u \in H$: Let $u \in H$ and $\varepsilon > 0$. Choose $u_\varepsilon \in V$ with $\|u - u_\varepsilon\|_H \le \varepsilon$. Then,

$$\|u - \Pi_N u\|_H \le \underbrace{\|u - u_\varepsilon\|_H}_{\le \varepsilon} + \underbrace{\|\Pi_N (u - u_\varepsilon)\|_H}_{\le \|(u-u_\varepsilon)\|_H \le \varepsilon} + \underbrace{\|u_\varepsilon - \Pi_N u_\varepsilon\|_H}_{\le C\|u_\varepsilon - \Pi_N u_\varepsilon\|_V \to 0 \text{ For } N\to\infty}$$

where the constant $C > 0$ comes from the continuous embedding $V \subset H$. □

1.1.1 Characterization of eigenvalues

For self-adjoint, compact operators A, the *Rayleigh-quotient*

$$R_A(x) := \frac{(Ax, x)_X}{(x, x)_X}, \qquad x \in X \setminus \{0\}$$

is an important tool to characterize eigenvalues[2].

Lemma 1.7 *Let A be self-adjoint on the Hilbert space X. Let R_A be the Rayleigh-quotient. Then,*

$$\|A\|_X = \sup_x |R_A(x)| .$$

For compact operators A the supremum is indeed attained and the maximizer is an eigenvector. Moreover, $\|A\|_X$ is then an eigenvalue.

Proof: The last statement is shown in the exercises.
The statement

$$s := \sup_x |R_A(x)| \leq \|A\|_X$$

is easily directly shown. To prove $\|A\|_X \leq s$ we use the self-adjointness of A. Let x, $y \in X$ be arbitrary. Then,

$$\begin{aligned} 2\,|(Ax, y)_X + (Ay, x)_X| &= |(A(x+y), (x+y))_X - (A(x-y), (x-y))_X| \\ &\leq s\left(\|x+y\|_X^2 + \|x-y\|_X^2\right) = 2s\left(\|x\|_X^2 + \|y\|_X^2\right). \end{aligned} \tag{1.7}$$

We now set $y = tAx$ with $t > 0$ to be chosen later. Then, by self-adjointness of A, we have

$$4t\|Ax\|_X^2 \leq 2s\left(\|x\|_X^2 + t^2\|Ax\|_X^2\right)$$

i.e.,

$$\left(2t - st^2\right)\|Ax\|_X^2 \leq s\|x\|_X^2.$$

In order to obtain a sharp estimate, we now choose $t > 0$ s.t. the left-hand side is as large as possible. This is achieved by the choice $t = 1/s > 0$. Consequently, we obtain

$$\frac{1}{s}\|Ax\|_X^2 \leq s\|x\|_X^2.$$

As x is arbitrary, we have obtained $\|A\|_X \leq s$.
Strictly speaking, we have assumed $s > 0$ in the proof. In the case $s = 0$ we can use the same argument by replacing s by $s + \varepsilon$ with arbitrary $\varepsilon > 0$ in (1.7). Then, we obtain $\|A\|_X \leq s + \varepsilon$ for arbitrary $\varepsilon > 0$.
□

For $A = T$ and $X = (V, a(\cdot,\cdot))$, the Rayleigh-quotient R_A reads as

$$R_A(x) = \frac{(x, x)_H}{a(x, x)}.$$

We consider the reciprocal

$$R(x) := \frac{a(x, x)}{(x, x)_H} \qquad x \in V \setminus \{0\}. \tag{1.8}$$

There holds:

[2]For self-adjoint operators, the Rayleigh-quotient is also important in numerical methods such as the Ritz method.

Theorem 1.8 (Minimum principle) *Let the eigenvalues λ_n be sorted by size: $\lambda_1 \le \lambda_2 \le \cdots$ (also according for their multiplicity). Let u_n be the corresponding eigenfunctions. Then,*

(i) $R(u_n) = \lambda_n$ *for all* n.

(ii) $\lambda_1 = \min_{u \in V} R(u)$.

(iii) With

$$V_m := \operatorname{span}\{u_1, \ldots, u_m\}, \tag{1.9}$$

$$V_m^\perp := \{v \in V \,|\, a(v,w) = 0 \quad \forall w \in V_m\} = \{v \in V \,|\, (v,w)_H = 0 \quad \forall w \in V_m\} \tag{1.10}$$

there holds

$$\lambda_m = \min_{v \in V_{m-1}^\perp} R(v), \qquad m = 2, 3, \ldots. \tag{1.11}$$

Proof: *ad (i):* is clear.
ad (ii): Let $v \in V$. Then, due to $a(u_n, v) = \lambda_n (u_n, v)_H$ and the orthogonalities satisfied by the functions u_n, there holds

$$v = \sum_{n=1}^{\infty} (v, u_n)_H u_n = \sum_{n=1}^{\infty} a(v, \lambda_n^{-1/2} u_n) \lambda_n^{-1/2} u_n$$

$$\|v\|_H^2 = \sum_{n=1}^{\infty} |(v, u_n)_H|^2$$

$$a(v,v) = \sum_{n=1}^{\infty} |a(v, \lambda_n^{-1/2} u_n)|^2 = \sum_{n=1}^{\infty} \lambda_n |(v, u_n)_H|^2$$

This gives

$$R(v) = \frac{a(v,v)}{\|v\|_H^2} = \frac{\sum_{n=1}^{\infty} \lambda_n |(v, u_n)_H|^2}{\sum_{n=1}^{\infty} |(v, u_n)_H|^2}.$$

As the λ_n are sorted ascending, there follows

$$\min_{v \in V} R(v) = \lambda_1.$$

ad (iii): Let $v \in V_{m-1}^\perp$. Then, v has the representation

$$v = \sum_{n=1}^{\infty} (v, u_n)_H u_n = \sum_{n=m}^{\infty} (v, u_n)_H u_n.$$

This implies

$$R(v) = \frac{a(v,v)}{\|v\|_H^2} = \frac{\sum_{n=m}^{\infty} \lambda_n |(v, u_n)_H|^2}{\sum_{n=m}^{\infty} |(v, u_n)_H|^2}$$

and by the monotonicity of $\lambda_n \uparrow$, there follows

$$\min_{v \in V_{m-1}^\perp} R(v) = \lambda_m.$$

□

The explicit use of the eigenvectors u_n to describe the eigenvalues λ_m, $m \ge 2$ in Theorem 1.8 is oftentimes cumbersome, as the eigenvectors may not be explicitly known. The following theorem avoids that.

Theorem 1.9 (Minimax-principle) *There holds*

$$\lambda_m = \min_{\substack{E_m \subset V \\ \dim E_m = m}} \max_{v \in E_m} R(v), \qquad m = 1, 2, \ldots.$$

Proof: "$\geq$": Let $E_m := V_m = \operatorname{span}\{u_1, \dots, u_m\}$. Then, every $v \in V_m$ has the representation $v = \sum_{n=1}^m \alpha_n u_n$ with $\alpha_n = (v, u_n)_H$ and

$$R(v) = \frac{\sum_{n=1}^m \lambda_n \alpha_n^2}{\sum_{n=1}^m \alpha_n^2}.$$

Maximizing over all $(\alpha_n)_{n=1}^m \in \mathbb{R}^m$ gives, using the monotonicity of the $\lambda_n \uparrow$, that

$$\max_{v \in V_m} R(v) = \lambda_m. \tag{1.12}$$

"$\leq$": Let $E_m \subset V$ with $\dim E_m = m$. We show $\lambda_m \leq \max_{v \in E_m} R(v)$. Choose $v \in E_m \setminus \{0\}$ such that

$$(v, u_n)_H = 0, \qquad n = 1, \dots, m-1.$$

(this is possible due to the dimension theorem of linear algebra as only $m-1$ *linear* conditions are posed). Then, $0 \neq v \in V_{m-1}^\perp \cap E_m$ and by Theorem 1.8

$$\lambda_m = \min_{w \in V_{m-1}^\perp} R(w) \leq R(v) \leq \max_{w \in E_m} R(w).$$

□

1.2 FEM discretization

In order to discretize the variational eigenvalue problem (1.3), we choose a finite dimensional subspace $V_h \subset V$ and consider the problem

$$\text{Find } (u_h, \lambda_h) \in V_h \setminus \{0\} \times \mathbb{R} \text{ s.t.} \qquad a(u_h, v) = \lambda_h (u_h, v)_H \qquad \forall v \in V_h. \tag{1.13}$$

Choosing a basis $\{\varphi_1, \dots, \varphi_N\}$ of V_h, the discrete eigenvalue problem (1.13) is then equivalent to an algebraic (generalized) eigenvalue problem given by

$$\text{Find } (\mathbf{u}_h, \lambda_h) \in \mathbb{R}^N \setminus \{0\} \times \mathbb{R} \text{ s.t.} \qquad \mathbf{A}\mathbf{u}_h = \lambda_h \mathbf{M}\mathbf{u}_h, \qquad \mathbf{A}_{ij} = a(\varphi_j, \varphi_i), \qquad \mathbf{M}_{ij} = (\varphi_j, \varphi_i)_H. \tag{1.14}$$

Remark 1.10 For moderate problem sizes N, one can solve the matrix eigenvalue problem (1.14) with standard methods of numerical linear algebra, e.g. the QR-algorithm (if $\mathbf{M} = \mathrm{Id}$ or if you use $\mathbf{M}^{-1}\mathbf{A}$) or variants such as the QZ algorithm (For general $\mathbf{M}$). This returns all eigenvalues and eigenvectors with effort $O(N^3)$. For large N, iterative methods are used that only return a small part of the spectrum. This is for two reasons: 1) the costs are prohibitive and QR-type algorithms can make very poor use of the occupation structure of $\mathbf{A}$, $\mathbf{M}$ (typically these are *sparsely* populated matrices); 2) one is only interested in a small part of the spectrum anyway, because the numerical approximations of a large part of the spectrum are very poor (more on that later). ∎

Exercise 1.11 Show that the minimax-principle also holds for the discretized problem (1.13), i.e., there holds

$$\lambda_{h,m} = \min_{\substack{E_m \subset V_h \\ \dim E_m = m}} \max_{v \in E_m} R(v). \tag{1.15}$$

∎

With the help of Exercise 1.11, we can show that the the discrete eigenvalues converge from above.

Theorem 1.12 *There holds*

$$\lambda_m \leq \lambda_{h,m}, \qquad m = 1, \dots, N.$$

Proof: Using both the discrete minimax-principle (1.15) and the continuous minimax-principle, we obtain due to $V_h \subset V$:

$$\lambda_{h,m} = \min_{\substack{E_m \subset V_h \\ \dim E_m = m}} \max_{v \in E_m} R(v) \geq \min_{\substack{E_m \subset V \\ \dim E_m = m}} \max_{v \in E_m} R(v) = \lambda_m.$$

□

Exercise 1.13 If the approximation spaces V_h are nested, then the convergence is even monotone: Let $V_h \subset V_{h'} \subset V$, then

$$\lambda_m \leq \lambda_{h',m} \leq \lambda_{h,m}, \qquad m = 1, \ldots, N.$$

■

1.2.1 Convergence of the eigenvalues

We define the *Ritz-projector* $P_h : V \to V_h$ as the orthogonal projection onto V_h in the $a(\cdot,\cdot)$-inner product, i.e., $P_h u \in V_h$ is characterized by

$$a(u - P_h u, v) = 0 \qquad \forall v \in V_h. \tag{1.16}$$

The Ritz-projector satisfies (see FEM lecture)

- P_h is linear.
- Best-approximation property: $\|u - P_h u\|_V \leq C \inf_{v \in V_h} \|u - v\|_V$.
- Orthogonal projection in $a(\cdot,\cdot)$: There holds $\|P_h\|_E \leq 1$, where $\|\cdot\|_E := a(\cdot,\cdot)^{1/2}$ denotes the so called *energy norm*.

Now, the key question for the convergence of eigenvalues and eigenfunctions is how the space $V_m \subset V$ (spanned by the first m eigenfunctions) and its projection $P_h V_m \subset V_h$ are connected.
For that we define the quantity

$$\sigma_{h,m} := \inf_{v \in V_m} \frac{\|P_h v\|_H}{\|v\|_H}, \qquad 1 \leq m \leq N. \tag{1.17}$$

Lemma 1.14 *If, for some $m \in \{1, \ldots, N\}$, we have $\sigma_{h,m} > 0$, then, for this m, there holds*

$$\lambda_m \leq \lambda_{h,m} \leq \sigma_{h,m}^{-2} \lambda_m. \tag{1.18}$$

Proof: (*Note:* $1/\sigma_{h,m}$ is the norm of the inverse of the operator $P_h : V_m \to P_h V_m$ (exercise!). If $V_m = P_h V_m$, then $P_h|_{V_m} = \mathrm{Id}$, which follows from the projection property of P_h: We have $P_h v = v$ for $v \in V_h$ and the assumption $P_h V_m = V_m$ leads to $V_m \subset V_h$, so that $P_h v = v$ for $v \in V_m$. Therefore, in the case $P_h V_m = V_m$, we have $\sigma_{h,m}^{-1} = 1$ and thus we expect that $1 - \sigma_{h,m}^{-1}$ is a (rough) measure, how close the spaces V_m and $P_h V_m$ are.)
We use the minimax-principle, this time with the space $E_m = P_h V_m$. For that, we need to show that the assumption $\sigma_{h,m} > 0$ leads to $\dim E_m = m$: provided $\dim E_m = \dim(P_h V_m) < m$, then there exists $v \in V_m \setminus \{0\}$ with $P_h v = 0$, which is a contradiction to $\sigma_{h,m} > 0$.
The discrete minimax-principle (1.15) now gives with $E_m = P_h V_m$

$$\begin{aligned}\lambda_{h,m} &\leq \max_{v \in E_m} R(v) = \max_{v \in E_m} \frac{a(v,v)}{\|v\|_H^2} = \max_{v \in V_m} \frac{a(P_h v, P_h v)}{\|P_h v\|_H^2} \leq \max_{v \in V_m} \frac{a(v,v)}{\|P_h v\|_H^2} \\ &= \max_{v \in V_m} \frac{a(v,v)}{\|v\|_H^2} \frac{\|v\|_H^2}{\|P_h v\|_H^2} \leq \lambda_m \sigma_{h,m}^{-2},\end{aligned}$$

where, in the last step, we used (1.12). □

Lemma 1.14 indicates that, for fixed m, we want to show:[3]

$$\lim_{h\to 0} \sigma_{h,m} = 1.$$

The following lemma shows that this is possible.

Lemma 1.15 *There holds*

$$\sigma_{h,m}^2 \geq 1 - \frac{2\|a\|\sqrt{m}}{\lambda_1} \sup_{v\in V_m} \frac{\|v - P_h v\|_V^2}{\|v\|_H^2}.$$

Proof: Let $v \in V_m$ with $\|v\|_H = 1$ being written as

$$v = \sum_{n=1}^m \alpha_n u_n, \qquad \sum_{n=1}^m |\alpha_n|^2 = 1.$$

Due to $\|v\|_H = 1$, there holds

$$\begin{aligned} 1 - \|P_h v\|_H^2 &= (v,v)_H - (P_h v, P_h v)_H = (v - P_h v, v + P_h v)_H = (v - P_h v, 2v + P_h v - v)_H \\ &= -\|v - P_h v\|_H^2 + 2(v - P_h v, v)_H \end{aligned}$$

and consequently

$$\|P_h v\|_H^2 = 1 - 2(v - P_h v, v)_H + \|v - P_h v\|_H^2 \geq 1 - 2(v - P_h v, v)_H.$$

With $a(z, u_n) = \lambda_n (z, u_n)_H$ for all $z \in V$ and $a(v - P_h v, w) = 0$ for all $w \in V_h$, we calculate

$$\begin{aligned} (v - P_h v, v)_H &= \sum_{n=1}^m \alpha_n (v - P_h v, u_n)_H \\ &= \sum_{n=1}^m \frac{\alpha_n}{\lambda_n} a(v - P_h v, u_n) \\ &= \sum_{n=1}^m \frac{\alpha_n}{\lambda_n} a(v - P_h v, u_n - P_h u_n) \\ &\leq \sum_{n=1}^m \frac{|\alpha_n|}{\lambda_n} \|a\| \|v - P_h v\|_V \|u_n - P_h u_n\|_V \\ &\leq \|v - P_h v\|_V \frac{\|a\|}{\lambda_1} \underbrace{\sqrt{\sum_{n=1}^m |\alpha_n|^2}}_{=1} \underbrace{\sqrt{\sum_{n=1}^m \|u_n - P_h u_n\|_V^2}}_{\leq \sqrt{m} \sup\limits_{\substack{w\in V_m \\ \|w\|_H = 1}} \|w - P_h w\|_V} \\ &\leq \frac{\|a\|}{\lambda_1} \sqrt{m} \sup_{\substack{w\in V_m \\ \|w\|_H = 1}} \|w - P_h w\|_V^2. \end{aligned}$$

This implies

$$\|P_h v\|_H^2 \geq 1 - 2\frac{\|a\|}{\lambda_1} \sqrt{m} \sup_{\substack{w\in V_m \\ \|w\|_H = 1}} \|w - P_h w\|_V^2.$$

From $\|v\|_H = 1$ the claim follows. □

Lemma 1.15 shows that we obtain error estimates, if the family of discrete space $(V_h)_{h>0}$ has approximation properties. We thus assume that

$$\forall v \in V\colon \lim_{h\to 0} \inf_{w\in V_h} \|v - w\|_V = 0. \tag{1.19}$$

[3] until now the spaces $V_h \subset V$ were arbitrary, as we want to apply FEM, we think about them as piecewise polynomial spaces, e.g. $S^{p,1}(\mathcal{T}_h)$, on a mesh with maximal mesh-size h.

Theorem 1.16 *Assume that there holds (1.19). Then, for every $m \in \mathbb{N}$, there exists a $h_0 > 0$ and a constant $C_m > 0$, such that, for every $h < h_0$, we have*

$$0 \leq \lambda_{h,m} - \lambda_m \leq C_m \sup_{v \in V_m} \inf_{w \in V_h} \frac{\|v - w\|_V^2}{\|v\|_H^2}.$$

Proof: Again, we consider a $v \in V_m$ with $\|v\|_H = 1$, written as $v = \sum_{n=1}^m \alpha_n u_n$ with $\sum_{n=1}^m |\alpha_n|^2 = 1$. Then, there holds

$$\|v - P_h v\|_V \leq \sum_{n=1}^m |\alpha_n| \|u_n - P_h u_n\|_V \leq \underbrace{\sqrt{\sum_{n=1}^m |\alpha_n|^2}}_{=1} \sqrt{\sum_{n=1}^m \|u_n - P_h u_n\|_V^2}.$$

By assumption (1.19), we can approximate each of the m functions u_n, $n = 1, \ldots, m$, arbitrarily well, if h is small enough, i.e., we have

$$\forall \varepsilon > 0 \quad \exists h_0 > 0 \quad \forall h < h_0 \colon \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \|v - P_h v\|_V \leq \varepsilon.$$

Employing the elementary inequality

$$\frac{1}{1 - \varepsilon} \leq 1 + 2\varepsilon \qquad \text{for } \varepsilon > 0 \text{ sufficiently small}$$

together with Lemma 1.15, for sufficiently small h, there follows

$$\sigma_{h,m}^{-2} \leq 1 + C \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \|v - P_h v\|_V^2.$$

Due to the best-approximation property of the Ritz-projector, we obtain (with a different constant $C > 0$ as above)

$$\sigma_{h,m}^{-2} \leq 1 + C \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \inf_{w \in V_h} \|v - w\|_V^2.$$

Employing Lemma 1.14 this shows the theorem. □

Remark 1.17

- Theorem 1.16 shows that the convergence of the eigenvalues is *double as fast* as one can expect for the convergence of the eigenfunctions.
- Theorem 1.16 shows that, for fixed m, one can expect convergence $\lambda_m = \lim_{h \to 0} \lambda_{h,m}$, if one works with FEM spaces with maximal mesh-width h. ∎

Example 1.18 We consider our 1D model problem

$$\begin{aligned} -u'' &= \lambda u \quad \text{in} \quad (0, \pi) \\ u(0) &= u(\pi) = 0. \end{aligned}$$

Then, the continuous eigenpairs are given by $(u_m, \lambda_m) = (\sin(mx), m^2)$, $m \in \mathbb{N}$. Discretizing the model problem with lowest order finite elements on a uniform mesh on $[0, \pi]$ with mesh size h leads to the discrete eigenvalues (programming exercise!)

$$\lambda_{h,m} = \frac{6}{h^2} \frac{1 - \cos(mh)}{2 + \cos(mh)}$$

and the discrete eigenfunctions are the piecewise linear interpolants of the continuous eigenfunctions $\sin(m\cdot)$ in the grid points. By Taylor expansion, we observe

$$\lambda_{h,m} = m^2 + \frac{m^4}{12} h^2 + \mathcal{O}(m^6 h^4).$$

Consequently, we see that the error satisfies $|\lambda_m - \lambda_{h,m}| = \mathcal{O}(h^2)$. However, one should note that the hidden constant depends on m. This is expected: Larger m lead to more oscillating eigenfunction and thus a more fine mesh is needed to maintain the same accuracy of the approximation. ∎

1.2.2 Convergence of eigenfunctions

We mainly consider the case of a *simple* eigenvalue λ_m. Convergence theory for eigenvalues with higher multiplicity $k \geq 2$ is possible but more tedious.

We have already defined the eigenfunctions $(u_n)_{n\in\mathbb{N}}$ in Corollary 1.6. In the same way, one can define the *discrete eigenfunctions* $(u_{h,n})_{n=1}^N$. They form an ONB of the discrete space $(V_h, (\cdot,\cdot)_H)$ and the functions $(\lambda_{h,n}^{-1/2} u_n)_{n=1}^N$ form an ONB of $(V_h, a(\cdot,\cdot))$.

An important question in the convergence theory of eigenvalue problems is how well the eigenvalue λ_m is separated from the rest of the spectrum: $\min\{|\lambda_m - \lambda_i| \,|\, i \in \mathbb{N} \setminus \{m\}\} > 0$ (note: λ_m is required to be a simple eigenvalue). Correspondingly, we need a quantity that measures the extent to which the discrete spectrum $\{\lambda_{h,i} \,|\, 1 \leq i \leq N\}$ is separated from λ_m. More precisely: The number

$$\rho_{h,m} := \max_{\substack{1\leq i\leq N \\ i\neq m}} \frac{\lambda_m}{|\lambda_m - \lambda_{h,i}|} \tag{1.20}$$

is a measure of how far the discrete eigenvalues that *not* converge to λ_m, are separated from λ_m.

Exercise 1.19 Show that: Under the conditions of Theorem 1.16 and the assumption that λ_m is a simple eigenvalue, there is an $h_0 > 0$ and a $c_m > 0$ such that for all $h < h_0$ there holds $\rho_{h,m} \leq c_m$. ∎

The following lemma shows that (except for scaling) the discrete eigenfunction $u_{h,m}$ is actually closely related to the eigenfunction u_m:

Lemma 1.20 *Let λ_m be a simple eigenvalue and let the approximation property (1.19) hold. Then, there exists $h_0 > 0$ such that for all $h < h_0$ there holds: There is a sign $\sigma \in \{\pm 1\}$ such that*

$$\|u_m - \sigma u_{h,m}\|_H \leq 2(1 + \rho_{h,m})\|u_m - P_h u_m\|_H. \tag{1.21}$$

Proof: The continuous eigenfunctions $(u_n)_{n\in\mathbb{N}}$ and the discrete eigenfunctions $(u_{h,n})_{n=1}^N$ are unique except for the sign as we assumed that they are ONBs. We now choose the sign σ of $u_{h,m}$ by the normalization condition

$$(P_h u_m, u_{h,m})_H \geq 0. \tag{1.22}$$

With this choice, we now want to show (1.21) with $\sigma = +1$.

The functions $(u_{h,n})_{n=1}^N$ form an ONB of $(V_h, (\cdot,\cdot)_H)$. Let $v_{h,m} \in \operatorname{span}\{u_{h,m}\}$ be the H-orthogonal projection of $P_h u_m$, i.e.

$$v_{h,m} := (P_h u_m, u_{h,m})_H u_{h,m}.$$

We note

$$\begin{aligned} P_h u_m - v_{h,m} &= \sum_{\substack{n=1 \\ n\neq m}}^N (P_h u_m, u_{h,n})_H u_{h,n} \\ \|P_h u_m - v_{h,m}\|_H^2 &= \sum_{\substack{n=1 \\ n\neq m}}^N |(P_h u_m, u_{h,n})_H|^2. \end{aligned}$$

Using the triangle inequality, we obtain

$$\|u_m - u_{h,m}\|_H \leq \|u_m - P_h u_m\|_H + \|P_h u_m - v_{h,m}\|_H + \|v_{h,m} - u_{h,m}\|_H \tag{1.23}$$

and consequently, we have to estimate these three terms separately.

Using the property that the functions $(u_n)_{n\in\mathbb{N}}$ and $(u_{h,n})_{n=1}^N$ are eigenfunctions, we can transfer the $(\cdot,\cdot)_H$-inner product in an $a(\cdot,\cdot)$-inner product, in order to use the Galerkin-orthogonality of the Ritz-projector:

$$(P_h u_m, u_{h,n})_H = \lambda_{h,n}^{-1} a(P_h u_m, u_{h,n}) = \lambda_{h,n}^{-1} a(u_m, u_{h,n}) = \frac{\lambda_m}{\lambda_{h,n}} (u_m, u_{h,n})_H.$$

This implies that

$$\begin{aligned}(\lambda_{h,n}-\lambda_m)(P_h u_m, u_{h,n})_H &= \lambda_m (u_m - P_h u_m, u_{h,n})_H \\ \Longrightarrow \|P_h u_m - v_{h,m}\|_H^2 &= \sum_{\substack{n=1\\ n\neq m}}^{N} |(P_h u_m, u_{h,n})_H|^2 \\ &\leq \rho_{h,m}^2 \sum_{\substack{n=1\\ n\neq m}}^{N} |(u_m - P_h u_m, u_{h,n})_H|^2 \\ &\leq \rho_{h,m}^2 \sum_{n=1}^{N} |(u_m - P_h u_m, u_{h,n})_H|^2 \leq \rho_{h,m}^2 \|u_m - P_h u_m\|_H^2, \end{aligned} \tag{1.24}$$

where the last estimate follows from Bessel's inequality. We have thus bounded the second term in (1.23) in the desired way. It remains to estimate $v_{h,m} - u_{h,m}$. There holds

$$\begin{aligned} v_{h,m} - u_{h,m} &= ((P_h u_m, u_{h,m})_H - 1)\, u_{h,m} \\ \Longrightarrow \|v_{h,m} - u_{h,m}\|_H &= |1 - (P_h u_m, u_{h,m})_H|\,. \end{aligned}$$

The reverse triangle inequality together with the normalization assumption $(P_h u_m, u_{h,m})_H \geq 0$ implies

$$\underbrace{\|u_m\|_H}_{=1} - \|u_m - v_{h,m}\|_H \leq \underbrace{\|v_{h,m}\|_H}_{|(P_h u_m, u_{h,m})_H| = (P_h u_m, u_{h,m})_H} \leq \underbrace{\|u_m\|_H}_{=1} + \|u_m - v_{h,m}\|_H. \tag{1.25}$$

Consequently, we obtain

$$|1 - (P_h u_m, u_{h,m})_H| \leq \|u_m - v_{h,m}\|_H$$

and further

$$\|v_{h,m} - u_{h,m}\|_H = |1 - (P_h u_m, u_{h,m})_H| \leq \|u_m - v_{h,m}\|_H. \tag{1.26}$$

Applying the triangle inequality in (1.26) together with (1.24) leads to

$$\begin{aligned} \|v_{h,m} - u_{h,m}\|_H &\leq \|u_m - v_{h,m}\|_H \leq \|u_m - P_h u_m\|_H + \|P_h u_m - v_{h,m}\|_H \\ &\leq \|u_m - P_h u_m\|_H + \rho_{h,m} \|u_m - P_h u_m\|_H. \end{aligned}$$

This gives together with (1.24)

$$\begin{aligned} \|u_m - u_{h,m}\|_H &\leq \|u_m - P_h u_m\|_H + \|P_h u_m - v_{h,m}\|_H + \|v_{h,m} - u_{h,m}\|_H \\ &\leq \|u_m - P_h u_m\|_H + \rho_{h,m} \|u_m - P_h u_m\|_H + (1+\rho_{h,m}) \|u_m - P_h u_m\|_H, \end{aligned}$$

which is the claimed statement. □

Lemma 1.25 provides the convergence of the eigenvalues in the (rather weak) $\|\cdot\|_H$-norm. For statements in the $\|\cdot\|_V$-norm, the following result can be used, which also applies to error estimators for eigenvalue problems (see exercise).

Lemma 1.21 *Let $(u, \lambda) \in V \setminus \{0\} \times \mathbb{R}$ be an eigenpair. Then, for all $v \in V \setminus \{0\}$ there holds*

$$a(v,v) - \lambda \|v\|_H^2 = a(u-v, u-v) - \lambda (u-v, u-v)_H.$$

Proof: Follows from direct calculation and using $a(u,z) = \lambda (u,z)_H$ for all $z \in V$. □

If we insert $u = u_m$ and $v = u_{h,m}$ into Lemma 1.21, we obtain a convergence statement for $u_m - u_{h,m}$ from Lemma 1.20 and Theorem 1.16.

Theorem 1.22 *Let λ_m be a simple eigenvalue and let (1.19) hold. Then, there exists an $h_0 > 0$ and a $C > 0$ such that for all $h < h_0$, there holds (here the sign of $u_{h,m}$ is always like in the proof of Lemma 1.20)*

$$\|u_m - u_{h,m}\|_V \le C \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \inf_{v_h \in V_h} \|v - v_h\|_V$$
$$\|u_m - u_{h,m}\|_H \le C \|u_m - P_h u_m\|_H.$$

Proof: Lemma 1.21 applied with $u = u_m$ and $v = u_{h,m}$ and the normalizations $\|u_m\|_H = 1 = \|u_{h,m}\|_H$ provides

$$\begin{aligned}\lambda_{h,m} - \lambda_m &= \lambda_{h,m}\|u_{h,m}\|_H^2 - \lambda_m \|u_{h,m}\|_H^2 = a(u_{h,m}, u_{h,m}) - \lambda_m \|u_{h,m}\|_H^2 \\ &= a(u_m - u_{h,m}, u_m - u_{h,m}) - \lambda_m \|u_m - u_{h,m}\|_H^2.\end{aligned}$$

This gives

$$a(u_m - u_{h,m}, u_m - u_{h,m}) = \lambda_{h,m} - \lambda_m + \lambda_m \|u_m - u_{h,m}\|_H^2,$$

which, together with Lemma 1.20 and Theorem 1.16 gives the claimed estimate. □

1.2.3 Remarks

- The convergence statement in Theorem 1.16 requires the approximability of $V_m = \text{span}\{u_1, \dots, u_m\}$ and not simply $\text{span}\{u_m\}$ (in the case of simple eigenvalues). In fact, the statements can also be sharpened in such a way that for simple eigenvalues there holds

$$\lambda_{h,m} - \lambda_m \le C \inf_{v \in V_h} \|u_m - v\|_V^2. \tag{1.27}$$

- We have only considered the symmetric case here, where the operator T is self-adjoint. The convergence theory for non-symmetric bilinear forms $a(\cdot, \cdot)$ is possible. However, not all convergence statements from the symmetric case are transferable, mainly because geometric and algebraic multiplicities of eigenvalues no longer coincide. For example, the doubling of the convergence rate as in (1.27) is no longer guaranteed. ∎

1.2.4 A more general approach using operator theory

Goal: Illustration of a more general approach using functional analytical techniques, which can also be generalized to non-symmetric problems and can also provide quantifiable convergence statements.

A basic assumption will be the approximation property (1.19), which we formulate slightly differently here as

$$\lim_{h \to 0} P_h u = u \qquad \forall u \in V. \tag{1.28}$$

Theorem 1.23 *(i) Let $(u_h, \lambda_h)_{h>0}$ be a bounded sequence of discrete eigenpairs (i.e. $\|u_h\|_V \le 1$ and $\lambda_h \le C$ with a constant $C > 0$ independent of h). Then, there exists a subsequence $(u_{h'}, \lambda_{h'})_{h'>0}$ and an eigenpair $(u, \lambda) \in V \setminus \{0\} \times \mathbb{R}$ of (1.2) with $u_{h'} \to u$ (in V) and $\lambda_{h'} \to \lambda$.*

(ii) Let λ be an eigenvalue of (1.2). Then, there exists a sequence $(u_h, \lambda_h)_{h>0}$ of discrete eigenpairs with $\lambda_h \to \lambda$.

For the proof we need an operator representation of the discrete eigenvalue problem:

Lemma 1.24 *(i) Let (u_h, λ_h) be a discrete eigenpair. Then,*

$$\lambda_h P_h T u_h = u_h \tag{1.29}$$
$$\lambda_h P_h T P_h u_h = P_h u_h = u_h. \tag{1.30}$$

(ii) Every eigenpair $(u,\lambda) \in V \setminus \{0\} \times \mathbb{R} \setminus \{0\}$ *of*

$$P_h T P_h u = \frac{1}{\lambda} u \tag{1.31}$$

is a discrete *eigenpair, i.e.,* $u = P_h u \in V_h$ *and* $(u,\lambda) \in V_h \times \mathbb{R}$ *solves (1.13).*

Proof: *ad (i):* We employ the self-adjointness of T and P_h w.r.t. $a(\cdot,\cdot)$: Let $(u_h,\lambda_h) \in V_h \times \mathbb{R}$. Then,

$$\begin{aligned} & a(u_h, v) = \lambda_h (u_h, v)_H \qquad \forall v \in V_h \\ \iff & \underbrace{a(u_h, P_h v)}_{=a(P_h u_h, v) = a(u_h, v)} = \lambda_h \underbrace{(u_h, P_h v)_H}_{a(TP_h v, u_h) = a(P_h v, T u_h) = a(v, P_h T u_h)} \qquad \forall v \in V \\ \iff & u_h = \lambda_h P_h T u_h. \end{aligned}$$

This shows (1.29). (1.30) follows from (1.29) by applying P_h.
ad (ii): Let (u,λ) satisfy (1.31). Then, $u = \lambda P_h T P_h u \in V_h$. In particular, all calculations in the proof of (i) can be done as well and show that (u,λ) is a discrete eigenpair. □

Proof of Theorem 1.23: The key tool is the *norm convergence* $P_h T \to T$, which follows from

$$P_h T - T = \underbrace{(P_h - I)}_{\to 0 \text{ pointwise}} \underbrace{T}_{\text{compact}}.$$

ad (i): Let $(u_h,\lambda_h)_{h>0}$ be a bounded sequence of discrete eigenpairs with

$$\|u_h\|_E = 1 \qquad \text{and} \qquad |\lambda_h| \le C \qquad \forall h > 0.$$

After taking a subsequence (which we again denote by $(u_h,\lambda_h)_{h>0}$) we can assume that (as the embedding $V \subset H$ is compact)

$$u_h \overset{V}{\rightharpoonup} u \in V, \tag{1.32}$$

$$u_h \overset{H}{\to} u, \tag{1.33}$$

$$T u_h \overset{V}{\to} T u, \tag{1.34}$$

$$\lambda_h \to \lambda \in \mathbb{R}. \tag{1.35}$$

Now, we may argue that

$$\underbrace{u_h}_{\to u} = \lambda_h P_h T u_h = \underbrace{\lambda_h}_{\to \lambda} \left(\underbrace{(P_h T - T)}_{\to 0 \text{ in norm}} \underbrace{u_h}_{\text{bounded}} + \underbrace{T(u_h - u)}_{\to 0 \; T \text{ compact}} + T u \right).$$

This means that the right-hand side converges in the norm (i.e. strongly) against $\lambda T u$. This also implies weak convergence to $\lambda T u$. As the right-hand side converges to u weakly and weak limits are unique, we have

$$u = \lambda T u$$

as well as strong convergence $u_h \to u$.
It remains to show that $u \neq 0$ and $\lambda \neq 0$. This follows from

$$1 = \|u_h\|_E^2 = a(u_h,u_h) = \lambda_h (u_h,u_h)_H = \lambda_h \|u_h\|_H^2 \overset{(1.33),(1.35)}{\to} \lambda \|u\|_H^2.$$

This implies $\lambda \neq 0$ and $\|u\|_H \neq 0$.
ad (ii): We argue by means of a perturbation argument (compare, e.g. the proof of the Bauer-Fike theorem). We write

$$T = P_h T P_h + \delta_h,$$

where $\|\delta_h\|_V \to 0$, since

$$\delta_h = T - P_h T P_h = (I - P_h) T + P_h T (I - P_h)$$

can be written as composition of compact operators and operators converging pointwise to zero. The operator P_hTP_h is compact and self-adjoint (w.r.t. $a(\cdot,\cdot)$) and therefore has an ONB $(e_n, \mu_{h,n})$ of $(V, a(\cdot,\cdot))$ consisting of eigenvectors

$$P_hTP_he_n = \mu_{h,n}e_n, \qquad n = 1, \dots.$$

By the spectral theorem, the eigenvalues $\mu_{h,n} = 0$ span $\operatorname{Ker} P_hTP_h$. According to Lemma 1.24, (ii), the pairs $(e_n, 1/\mu_{h,n})$ with $\mu_{h,n} \neq 0$ are discrete eigenpairs. With the discrete eigenvalues with $\lambda_{h,i} > 0$, $i = 1, \dots, N$, we thus have

$$\mu_{h,n} = \lambda_{h,n}^{-1}, \qquad n = 1, \dots, N, \qquad \mu_{h,n} = 0, \qquad n = N+1, \dots.$$

Now, let λ be an eigenvalue of the continuous problem. Set $\mu := \lambda^{-1}$. We claim:

$$\inf\{|\mu - \mu_{h,n}| \,:\, n = 1, \dots, N\} \to 0 \quad \text{for } h \to 0. \tag{1.36}$$

It is sufficient to consider $\inf\{|\mu - \mu_{h,n}| \,:\, n = 1, \dots, N\} > 0$. Then, λ is not a discrete eigenvalue and $\mu - P_hTP_h$ is invertible:

$$(\mu - P_hTP_h)v = \sum_{n=1}^{\infty}(\mu - \mu_{h,n})a(v, e_n)e_n \quad \text{and} \quad (\mu - P_hTP_h)^{-1}v = \sum_{n=1}^{\infty}\frac{1}{\mu - \mu_{h,n}}a(v, e_n)e_n.$$

In particular, we obtain

$$\|(\mu - P_hTP_h)^{-1}\|_E = \max_n \frac{1}{|\mu - \mu_{h,n}|} = \frac{1}{\min_n |\mu - \mu_{h,n}|}.$$

Now, let $0 \neq u$ be an eigenvector corresponding to the eigenvalue λ. Then, $\mu u = Tu = P_hTP_hu + \delta_h u$ and consequently

$$\begin{aligned} &u = (\mu - P_hTP_h)^{-1}\delta_h u \qquad \Longrightarrow \qquad \|u\|_E \le \|(\mu - P_hTP_h)^{-1}\|_E \|\delta_h\|_E \|u\|_E \\ &\Longrightarrow 1 \le \|(\mu - P_hTP_h)^{-1}\|_E \|\delta_h\|_E \Longrightarrow \min_n |\mu - \mu_{h,n}| \le \|\delta_h\|_E \to 0. \end{aligned}$$

We have thus obtained that the discrete eigenvalues approximate the continuous ones. □

1.3 Numerical solution of the algebraic eigenvalue problem

Goal: given symmetric, positive definite matrices $\mathbf{A}$, $\mathbf{M}$, obtain a good approximation to one eigenvalue λ of $\mathbf{A}x = \lambda \mathbf{M}x$.

1.3.1 Classical techniques

1. *Inverse iteration*

   Given starting vector $x^{(0)}$, loop over (until accurate enough):

   - compute $x^{(n+1)} := \mathbf{A}^{-1}\mathbf{M}x^{(n)}$
   - normalize $x^{(n+1)} := x^{(n+1)}/\|x^{(n+1)}\|$ ($\|\cdot\|$ suitable norm)
   - $\lambda^{(n+1)} := R(x^{(n+1)})$

   The method converges (asymptotically) linear with a rate of convergence depending on the separation of the smallest eigenvalue λ_1 and second smallest eigenvalue λ_2. In the considered *symmetric* case, the error in the m-th step is of order $O((\lambda_1/\lambda_2)^{2m})$. In particular, the convergence *does not depend* on the whole spectrum or the ratio of the largest and smallest eigenvalue ($\to$ condition number).

As the first step of the algorithm is a solution of a linear system, it is (for not too large problem sizes) advisable to compute an LU-factorization (or rather Cholesky factorization) of $\mathbf{A}$ once, which can be reused for forward-/backward-substitution in each step of the interation.

Modification: *inverse iteration with shift*: For given $\mu \in \mathbb{R}$ compute $x^{(n+1)} := (\mathbf{A} - \mu\mathbf{M})^{-1}\mathbf{M}x^{(n)}$ in the first step.

Converges to eigenvalue closest to μ (denoted by λ_1) with rate $O((|\lambda_1 - \mu|/|\lambda_2 - \mu|)^{2m})$ with λ_2 being the second closest eigenvalue to μ.

2. *Rayleigh-quotient iteration*

   Given starting vector $x^{(0)}$ (normalized), $\lambda^{(0)} := R(x^{(0)})$ loop over (until accurate enough):

   - compute $x^{(n+1)} := (\mathbf{A} - \lambda^{(n)}\mathbf{M})^{-1}\mathbf{M}x^{(n)}$
   - normalize $x^{(n+1)} := x^{(n+1)}/\|x^{(n+1)}\|$
   - $\lambda^{(n+1)} := R(x^{(n+1)})$

   In the considered *symmetric* case, the method has cubic convergence to the eigenvalue closest to the initial guess. However, in each step of the method a linear system with a different system matrix has to be solved, therefore a computed LU-factorization can *not* be reused.

Remark 1.25 Block variants exist to compute more than one eigenvalues. Krylov methods (such as the Lanczos method) can be employed as well to compute extremal eigenvalues. However, these methods may converge slowly for large FEM matrices as the convergence depends on $\lambda_{max} - \lambda_{min}$. ∎

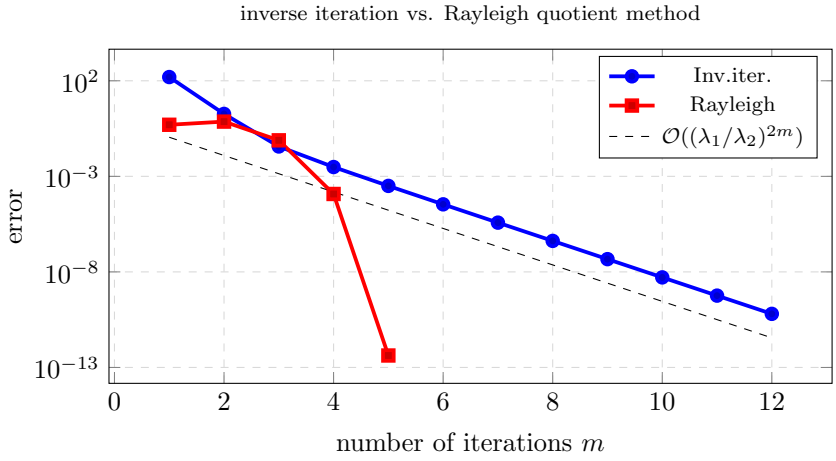

Figure 1.1: Convergence of inverse iteration and Rayleigh-quotient iteration for $\mathbf{A} = \texttt{diag}(1, 3, 5, \ldots, 315)$, $\mathbf{M} = \mathrm{Id}$. Thereby, $\lambda_1 = 1$, $\lambda_2 = 3$ and the initial guess for the Rayleigh-quotient iteration is chosen such that $\lambda^{(0)} := 1.5$.

Note that these classical algorithms assume that a linear system with matrix $\mathbf{A}$ or $(\mathbf{A} - \mu\mathbf{M})$ can be solved with reasonable effort and exactness.

1.3.2 Preconditioned gradient method

Question: what to do, if exact solution of the linear systems (i.e. application of $\mathbf{A}^{-1}$ or $(\mathbf{A} - \lambda^{(n)}\mathbf{M})^{-1}$) is not possible?
Idea: smallest eigenvalue is minimizer of Rayleigh-quotient $\implies$ minimize R using a gradient descent method. This gives the iteration

$$x^{(n+1)} := x^{(n)} - \omega^{(n)}\nabla R(x^{(n)}),$$
$$\nabla R(x) = 2\frac{\mathbf{A}x - R(x)\mathbf{M}x}{(x, \mathbf{M}x)_2}$$

with descent direction $\nabla R(x^{(n)})$ and step length $\omega^{(n)}$ to be chosen accordingly (below!).

Problem: The iteration is very slow, if the condition number $\mathrm{cond}(\mathbf{A})$ is large.

Solution: Use a *preconditioner* $\mathbf{T}$ ($\mathbf{T}$ SPD) and employ the preconditioned iteration

$$x^{(n+1)} := x^{(n)} - \omega^{(n)}\mathbf{T}\nabla R(x^{(n)}). \tag{1.37}$$

The method (1.37) is called *preconditioned inverse iteration (PINVIT)*, see also the exercises below. The step size $\omega^{(n)}$ is chosen such that $R(x^{(n+1)})$ is minimal, i.e., one solves the optimization problem

$$\min\left\{R(x)\,|\,x \in \mathrm{span}\{x^{(n)}, \mathbf{T}\nabla R(x^{(n)})\}\right\}. \tag{1.38}$$

By construction, this gives $R(x^{(n+1)}) \leq R(x^{(n)})$.

Exercise 1.26 Show that the minimization problem (1.38) is equivalent to a 2×2-EVP.

Exercise 1.27 Show that, if the step size $\omega^{(n)}$ is chosen by (1.38), then the scaling of $\mathbf{T}$ is irrelevant, i.e., one obtains the same iterate employing $\mathbf{T}$ or $\alpha\mathbf{T}$ for arbitrary $\alpha > 0$. Consequently, the iteration process is oftentimes written as

$$x^{(n+1)} = x^{(n)} - \widetilde{\omega}^{(n)}\mathbf{T}(\mathbf{A}x^{(n)} - R(x^{(n)})\mathbf{M}x^{(n)})$$

and $\widetilde{\omega}^{(n)}$ is computed from (1.38).

Remark 1.28 One can improve the algorithm dscribed above by minimizing over large spaces. The LO(B)PCG method (locally optimal (block) preconditioned conjugate gradient method) actually solves the 3×3-EVP

$$\min\left\{R(x)\,|\,x \in \mathrm{span}\{x^{(n)}, x^{(n-1)}, \mathbf{T}\nabla R(x^{(n)})\}\right\}. \tag{1.39}$$

Note that this iteration can be very cheaply realized, if the application of $\mathbf{T}$, $\mathbf{A}$ and $\mathbf{M}$ is cheap. ∎

The choice of preconditioner $\mathbf{T}$ is crucial for the performance of the algorithm. The optimal choice $\mathbf{T} = \mathbf{A}^{-1}$, however, might be to expensive to employ. Thus, we are considering preconditioners which are related to $\mathbf{A}$ (or actually $\mathbf{A}^{-1}$) by parameters γ_1, γ_0 such that

$$\gamma_1(x, \mathbf{T}^{-1}x) \leq (x, \mathbf{A}x) \leq \gamma_2(x, \mathbf{T}^{-1}x) \qquad \forall x. \tag{1.40}$$

In literature, this is also called spectral equivalency and implies that the condition number of $\mathbf{TA}$ is bounded by γ_2/γ_1.

Exercise 1.29 Consider the optimal preconditioner, i.e., $\mathbf{T} = \mathbf{A}^{-1}$. Show that the step size $\omega^{(n)}$ in (1.37) can be chosen such that one effectively does one step of inverse iteration. ∎

Concerning convergence of PINVIT with step size choice (1.38) (and therefore also of LOPCG) we have the following theorem.

Theorem 1.30 *With γ_1, $\gamma_2 > 0$ from (1.40) define $\gamma := (\gamma_2 - \gamma_1)/(\gamma_1 + \gamma_2)$. Assume $\lambda_i < R(x^{(n)}) < \lambda_{i+1}$.*
Then, for $x^{(n+1)}$ there holds: Either $R(x^{(n+1)}) \leq \lambda_i$ or $\lambda_i < R(x^{(n+1)}) < \lambda_{i+1}$ with

$$0 < \frac{R(x^{(n+1)}) - \lambda_i}{\lambda_{i+1} - R(x^{(n+1)})} \leq \sigma^2 \frac{R(x^{(n)}) - \lambda_i}{\lambda_{i+1} - R(x^{(n)})}, \qquad \sigma = \gamma + (1-\gamma)\frac{\lambda_i}{\lambda_{i+1}} < 1$$

Proof: See literature, [1, 10, 14]. □

Remark 1.31 Typically, PINVIT converges (asymptotically) linearly to the smallest eigenvalue. Note that the quality of the preconditioner $\mathbf{T}$ can be explicitly seen in the error bounds in the previous theorem as γ and in consequence the constraction factor σ depends on the parameters γ_1, γ_2, which are directly related to $\mathbf{T}$. In particular, choosing an optimal preconditioner, the convergence does not depend on $\mathrm{cond}\,\mathbf{A}$, but only how far the two smallest eigenvalues are separated.
However, the task of finding a good preconditioner is not easy, as we have the conflicting goals: we aim for (1.40) with moderate constants γ_1, γ_2, but the application $y \mapsto \mathbf{T}y$ should be cheap (much cheaper than computing $\mathbf{A}^{-1}$!). ∎

1.3.3 Numerical example

Example 1.18 shows the expected convergence of order h^2 for the 1d eigenvalue problem $-u'' = \lambda u$. Note that in the 1d case, the discrete eigenvalues are known explicitly, thus no eigenvalue solver needs to be employed.
We therefore focus on the 2d case.

Example 1.32 Let $\Omega = [0, \pi]^2$ be a square. Then, by separation of variables, the eigenvalues and eigenfunctions are given by

$$\lambda_{m,n} = m^2 + n^2, \qquad m, n \in \mathbb{N}$$
$$u_{m,n} = \sin(mx)\sin(ny), \qquad m, n \in \mathbb{N}$$

In the following, we employ PINVIT (using a multigrid preconditioner as approximate inverse) in NgSolve (see jupyter-notebook on TUWEL) to approximate the 4 smallest eigenvalues using lowest order FEM.

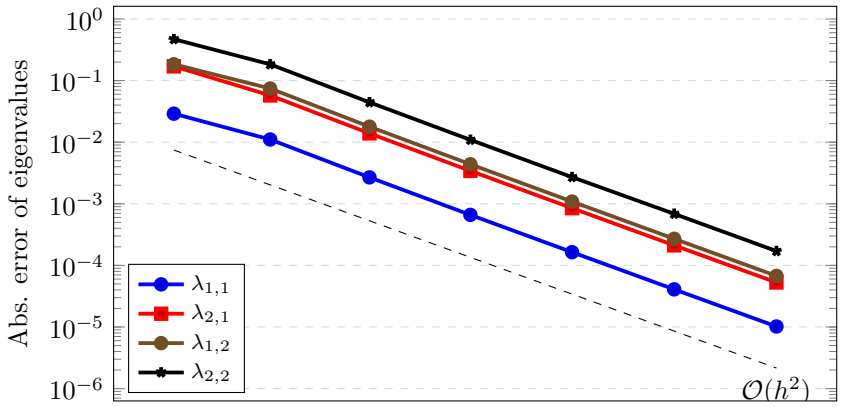

Figure 1.2: Errors $|\lambda_{m,n} - \lambda_{h,m,n}|$ for $m, n = 1, 2$ on the square.

As one can see, all eigenvalues are approximated with rate $\mathcal{O}(h^2)$ (note: $\dim V_h = N \sim h^{-2}$). Moreover, we have a double eigenvalue at $\lambda = 5$. While the error curves are close, they are not identical, thus, we have two *distinct* discrete eigenvalues approximating λ with one-dimensional eigenspaces! An approximation to an eigenfunction (which lies in a two dimensional eigenspace) for λ can thus only be good, if it is given by a linear combination in the one dimensional discrete eigenspaces. For more details on the treatement of multiple eigenvalues, we refer to literature, e.g., [4].

1.4 Weyl's asymptotic formula

Goal: Show how eigenvalues of the Laplacian (with Dirichlet boundary conditions) on a given domain $\Omega \subset \mathbb{R}^2$ are behaving asymptotically. More precisely, we define the counting function

$$N_{\Omega,D}(\lambda) := \operatorname{card}\{n \,|\, \lambda_n \leq \lambda\} \tag{1.41}$$

and aim to show the asymptotic behaviour

$$N_{\Omega,D}(\lambda) \sim \frac{|\Omega|}{4\pi}\lambda, \qquad \lambda \to \infty. \tag{1.42}$$

Remark 1.33 Similar formulas also exists for $\Omega \subset \mathbb{R}^d$, $d > 2$ and other types of boundary conditions. ∎

1.4.1 Example: Rectangles

Lemma 1.34 *Let $R = (0,a) \times (0,b)$. Then, the eigenvalues of the Dirichlet model problem on R are given by*

$$\lambda_{\ell,m} = \frac{\pi^2 \ell^2}{a^2} + \frac{\pi^2 m^2}{b^2}, \qquad \ell, m = 1, 2, \ldots, \tag{1.43}$$

$$N_{R,D}(\lambda) := \mathrm{card}\{(\ell,m) \in \mathbb{N}^2 \,|\, \lambda_{\ell,m} \leq \lambda\} = \lambda \frac{ab}{4\pi} + \mathcal{O}(\sqrt{\lambda}), \qquad \lambda \to \infty. \tag{1.44}$$

Proof: *Proof of (1.43):* Using a separation of variables, one directly obtains the eigenfunctions and eigenvalues as $u_{\ell,m}(x,y) := \sin(\pi \ell x/a)\sin(\pi m y/b)$ and $\lambda_{\ell,m} = \pi^2\ell^2/a^2 + \pi^2 m^2/b^2$.

Proof of (1.44): $N_{R,D}(\lambda)$ represents the number of grid points in $\mathbb{Z}^+ \times \mathbb{Z}^+$ that lie in the (closed) ellipse $\mathcal{E} = \{(x,y) \in \mathbb{R}^2 \,|\, (x/A)^2 + (y/B)^2 \leq 1\}$ with semi-axis $A = \sqrt{\lambda}a/\pi$ and $B = \sqrt{\lambda}b/\pi$. We note that the number of points inside the ellipse is bounded by the area of the first quarter of the ellipse. Thus, $N_{R,D}(\lambda) \leq \frac{1}{4}\mathrm{area}(\mathcal{E})$. For a lower bound, one may subtract one layer of grid points around the boundary of the ellipse, which gives $N_{R,D}(\lambda) \geq \frac{1}{4}\mathrm{area}(\mathcal{E}) - \mathcal{O}(\mathrm{boundarylength}(\mathcal{E}))$. For large λ, we therefore have

$$N_{R,D}(\lambda) = \frac{1}{4}\mathrm{area}(\mathcal{E}) + \mathcal{O}(\mathrm{boundarylength}(\mathcal{E})) = \frac{\lambda ab}{4\pi} + \mathcal{O}(\sqrt{\lambda}).$$

□

In the same way, one otains a formula for the Neumann Problem (boundary condition $\partial_n u = 0$).

Lemma 1.35 *Let $R = (0,a) \times (0,b)$. Then, the eigenvalues of the Neumann model problem on R are given by*

$$\lambda_{\ell,m} = \frac{\pi^2 \ell^2}{a^2} + \frac{\pi^2 m^2}{b^2}, \qquad \ell, m = 0, 1, 2, \ldots, \tag{1.45}$$

$$N_{R,N}(\lambda) := \mathrm{card}\{(\ell,m) \,|\, \lambda_{\ell,m} \leq \lambda\} = \lambda \frac{ab}{4\pi} + \mathcal{O}(\sqrt{\lambda}), \qquad \lambda \to \infty. \tag{1.46}$$

Proof: Here, the eigenfunctions are given by $u_{\ell,m}(x,y) = \cos(\pi \ell x/a)\cos(\pi m y/b)$ and the second statement follows as in the Dirichlet case. □

1.4.2 Monotonicity properties

We denote by $\lambda_{n,D}(\Omega)$ (Dirichlet) and $\lambda_{n,N}(\Omega)$ (Neumann) the (sorted) eigenvalues of

$$\begin{aligned} -\Delta u &= \lambda u \quad \text{in } \Omega, & u &= 0 \quad \text{on } \partial\Omega, \\ -\Delta u &= \lambda u \quad \text{in } \Omega, & \partial_n u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

By the minimax- and maximum-minimum principle (exercise), we obtain with the Rayleigh-quotient $R(v) = \|\nabla v\|^2_{L^2(\Omega)}/\|v\|^2_{L^2(\Omega)}$ that

$$\lambda_{n,D}(\Omega) = \min_{\substack{E_n \subset H^1_0(\Omega) \\ \dim E_n = n}} \max_{v \in E_n} R(v) = \max_{z_1,\ldots,z_{n-1} \in L^2(\Omega)} \min_{\substack{v \in H^1_0(\Omega) \\ (v,z_i)_{L^2(\Omega)} = 0, \\ i=1,\ldots,n-1}} R(v), \tag{1.47}$$

$$\lambda_{n,N}(\Omega) = \min_{\substack{E_n \subset H^1(\Omega) \\ \dim E_n = n}} \max_{v \in E_n} R(v) = \max_{z_1,\ldots,z_{n-1} \in L^2(\Omega)} \min_{\substack{v \in H^1(\Omega) \\ (v,z_i)_{L^2(\Omega)} = 0, \\ i=1,\ldots,n-1}} R(v). \tag{1.48}$$

The Dirichlet eigenvalues satsify a simple monotonicity principle.

Lemma 1.36 *Let domains $\Omega \subset \Omega'$ be given. Then,*

$$\lambda_{n,D}(\Omega) \geq \lambda_{n,D}(\Omega') \qquad \forall n \in \mathbb{N}. \tag{1.49}$$

Proof: Exercise, use that $H_0^1(\Omega) \subset H_0^1(\Omega')$. *note:* an analogous principle does *not* hold for the Neumann eigenvalues. □

The Dirichlet eigenvalues $\lambda_{n,D}(\Omega)$ and the Neumann eigenvalues $\lambda_{n,N}(\Omega)$ satisfy the following relation.

Lemma 1.37 *There holds*

$$\lambda_{n,D}(\Omega) \geq \lambda_{n,N}(\Omega) \qquad \forall n \in \mathbb{N}.$$

Proof: Follows from the minimax-principle and $H_0^1(\Omega) \subset H^1(\Omega)$. □

Exercise 1.38 Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots$ and $0 \leq \widetilde{\lambda}_1 \leq \widetilde{\lambda}_2 \leq \cdots$ with $\widetilde{\lambda}_n \leq \lambda_n$ for all $n \in \mathbb{N}$. Then,

$$\operatorname{card}\{n \,|\, \lambda_n \leq \lambda\} =: N(\lambda) \leq \widetilde{N}(\lambda) := \operatorname{card}\{n \,|\, \widetilde{\lambda}_n \leq \lambda\} \qquad \forall \lambda \geq 0. \tag{1.50}$$

We now consider the relationship between the eigenvalues of subdomains and the eigenvalues on the entire domain.

Lemma 1.39 *Let $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2 \cup \cdots \cup \overline{\Omega}_m$, with pairwise disjoint subdomains Ω_i, $i = 1, \ldots, m$. Let $\widetilde{\lambda}_1 \leq \widetilde{\lambda}_2 \leq \cdots$ be the sorted (accounting for the multiplicity) eigenvalues in $\{\lambda_{n,D}(\Omega_i) \,|\, n \in \mathbb{N}, \quad i \in \{1, \ldots, m\}\}$. Then,*

$$\widetilde{\lambda}_n = \max_{z_1,\ldots,z_{n-1} \in L^2(\Omega)} \min_{\substack{v \in L^2(\Omega) \\ v|_{\Omega_i} \in H_0^1(\Omega_i), i=1,\ldots,m \\ (v,z_j)_{L^2(\Omega)} = 0, j=1,\ldots,n-1}} R(v). \tag{1.51}$$

Note that here the Rayleigh quotient is understood as

$$R(v) = \frac{\sum_{i=1}^m \|\nabla v\|^2_{L^2(\Omega_i)}}{\sum_{i=1}^m \|v\|^2_{L^2(\Omega_i)}}. \tag{1.52}$$

Proof: Exercise. The essential observation is that the right-hand side of (1.51) is the characterization of an eigenvalue problem that can be decomposed into the *independent* eigenvalue problems of the m subdomains Ω_i. □

A similar result does also hold for the Neumann problem.

Lemma 1.40 *Let $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2 \cup \cdots \cup \overline{\Omega}_m$, with pairwise disjoint subdomains Ω_i, $i = 1, \ldots, m$. Let $\widetilde{\lambda}_1 \leq \widetilde{\lambda}_2 \leq \cdots$ be the sorted (accounting for the multiplicity) eigenvalues in $\{\lambda_{n,N}(\Omega_i) \,|\, n \in \mathbb{N}, \quad i \in \{1, \ldots, m\}\}$. Then,*

$$\widetilde{\lambda}_n = \max_{z_1,\ldots,z_{n-1} \in L^2(\Omega)} \min_{\substack{v \in L^2(\Omega) \\ v|_{\Omega_i} \in H^1(\Omega_i), i=1,\ldots,m \\ (v,z_j)_{L^2(\Omega)} = 0, j=1,\ldots,n-1}} R(v). \tag{1.53}$$

Again, R is to be understood as in (1.52).

1.4.3 Proof of Weyl's formula

Theorem 1.41 (Weyl's asymptotic formula) *Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain with piecewise smooth boundary. Then, the eigenvalues of the Laplace Dirichlet problem satisfy*

$$\lim_{\lambda \to \infty} \frac{N_{\Omega,D}(\lambda)}{\lambda} = \frac{|\Omega|}{4\pi}. \tag{1.54}$$

Proof: For $h > 0$, define the (infinite) regular, rectangular meshes

$$\mathcal{T}_h = \{(ih, (i+1)h) \times (jh, (j+1)h) \,|\, i, j \in \mathbb{Z}\}.$$

We consider the union of rectangles R_h and $\widetilde{R}_h$ that approximate Ω from inside and outside: Let $R_1, \dots, R_m$ be pairwise disjoint rectangles from $\mathcal{T}_h$ that lie *entirely* in Ω and $\widetilde{R}_1, \dots, \widetilde{R}_{\widetilde{m}}$ be rectangles from $\mathcal{T}_h$ that have non-empty intersections with $\overline{\Omega}$. Define

$$R_h := \bigcup_{i=1}^{m} R_i \subseteq \Omega, \qquad \widetilde{R}_h := \bigcup_{i=1}^{\widetilde{m}} \widetilde{R}_i \supseteq \Omega.$$

Step 1: Lemma 1.36 shows

$$\lambda_{n,D}(\Omega) \leq \lambda_{n,D}(R_h) \qquad \forall n \in \mathbb{N}.$$

With Exercise 1.38, we obtain

$$N_{\Omega,D}(\lambda) \geq N_{R_h,D}(\lambda) \qquad \forall \lambda > 0.$$

As R_h is a union of rectangles, we can estimate $\lambda_{n,D}(R_h)$ with explicit formulas

$$\begin{aligned}
\lambda_{n,D}(R_h) &\overset{(1.47)}{=} \max_{z_1,\dots,z_{n-1} \in L^2(R_h)} \min_{\substack{v \in H_0^1(R_h) \\ (v,z_i)_{L^2(R_h)}=0, i=1,\dots,n-1}} R(v) \\
&\leq \max_{z_1,\dots,z_{n-1} \in L^2(R_h)} \min_{\substack{v \in H_0^1(R_j), j=1,\dots,m \\ (v,z_i)_{L^2(R_h)}=0, i=1,\dots,n-1}} R(v) \\
&\overset{\text{Lemma 1.39}}{=} \widetilde{\lambda}_{n,D},
\end{aligned}$$

where $\widetilde{\lambda}_{1,D} \leq \widetilde{\lambda}_{2,D} \leq \cdots$ denotes the sorted sequence of eigenvalues $\{\lambda_{\ell,D}(R_j) \,|\, j = 1, \dots, m, \ell \in \mathbb{N}\}$. For every $j \in \{1, \dots, m\}$, Lemma 1.34 implies

$$N_{R_j,D}(\lambda) = \frac{\lambda |R_j|}{4\pi} + \mathcal{O}(\sqrt{\lambda}).$$

As the spectra are the same for all R_j and the R_j are pairwise disjoint, we obtain

$$\operatorname{card}\{n \,|\, \widetilde{\lambda}_{n,D} \leq \lambda\} = \sum_{j=1}^{m} N_{R_j,D}(\lambda) = \frac{\lambda |R_h|}{4\pi} + \mathcal{O}(\sqrt{\lambda}).$$

(Note: the hidden constant in the $\mathcal{O}(\cdot)$-notation does not depend on λ, but does depend on h and m.) This gives

$$N_{\Omega,D}(\lambda) \geq N_{R_h,D}(\lambda) \geq \frac{\lambda |R_h|}{4\pi} + \mathcal{O}(\sqrt{\lambda}).$$

Step 2: From

$$\lambda_{n,D}(\Omega) \overset{\text{Lemma 1.36}}{\geq} \lambda_{n,D}(\widetilde{R}_h) \overset{\text{Lemma 1.37}}{\geq} \lambda_{n,N}(\widetilde{R}_h),$$

we obtain together with Exercise 1.38

$$N_{\Omega,D}(\lambda) \leq N_{\widetilde{R}_h,N}(\lambda).$$

In order to estimate $N_{\widetilde{R}_h,N}(\lambda)$, we use

$$\begin{aligned}
\lambda_{n,N}(\widetilde{R}_h) &\overset{(1.48)}{=} \max_{z_1,\dots,z_{n-1} \in L^2(\widetilde{R}_h)} \min_{v \in H^1(\widetilde{R}_h)} R(v) \\
&\geq \max_{z_1,\dots,z_{n-1} \in L^2(\widetilde{R}_h)} \min_{\substack{v \in L^2(\widetilde{R}_h) \\ v|_{\widetilde{R}_j} \in H^1(\widetilde{R}_j)}} R(v) \\
&\overset{\text{Lemma 1.40}}{=} \widetilde{\lambda}_{n,N},
\end{aligned}$$

where $\widetilde{\lambda}_{1,N} \leq \widetilde{\lambda}_{2,N} \leq \cdots$ denotes the sorted sequence of elements in $\{\lambda_{\ell,N}(\widetilde{R}_j) \,|\, j = 1, \ldots, \widetilde{m}, \ell \in \mathbb{N}\}$. For every j, there holds by Lemma 1.35

$$N_{\widetilde{R}_j,N}(\lambda) = \lambda \frac{|\widetilde{R}_j|}{4\pi} + \mathcal{O}(\sqrt{\lambda}).$$

As the spectra of all rectangles $\widetilde{R}_j$ are the same, we arrive at

$$\text{card}\{n \,|\, \widetilde{\lambda}_{n,N} \leq \lambda\} = \sum_{j=1}^{\widetilde{m}} N_{\widetilde{R}_j,N}(\lambda) = \lambda \frac{|\widetilde{R}_h|}{4\pi} + \mathcal{O}(\sqrt{\lambda}),$$

which is

$$N_{\Omega,D}(\lambda) \leq N_{\widetilde{R}_h,N}(\lambda) \leq \lambda \frac{|\widetilde{R}_h|}{4\pi} + \mathcal{O}(\sqrt{\lambda}).$$

Step 3: Steps 1 and 2 provide

$$\frac{|R_h|}{4\pi} + \mathcal{O}(\lambda^{-1/2}) \leq \frac{N_{\Omega,D}(\lambda)}{\lambda} \leq \frac{|\widetilde{R}_h|}{4\pi} + \mathcal{O}(\lambda^{-1/2}).$$

This implies

$$\frac{|\widetilde{R}_h|}{4\pi} \geq \limsup_{\lambda\to\infty} \frac{N_{\Omega,D}(\lambda)}{\lambda} \geq \liminf_{\lambda\to\infty} \frac{N_{\Omega,D}(\lambda)}{\lambda} \geq \frac{|R_h|}{4\pi}.$$

Taking the limit $h \to 0$, there holds $\lim_{h\to} |R_h| = \lim_{h\to} |\widetilde{R}_h| = |\Omega|$ and consequently

$$\lim_{\lambda\to\infty} \frac{N_{\Omega,D}(\lambda)}{\lambda} = \frac{|\Omega|}{4\pi}.$$

□

We finish this section with a statement on the multiplicity of the smallest eigenvalue of the Laplace Dirichlet problem.

Lemma 1.42 *The smallest eigenvalue λ_1 of the Laplace Dirichlet problem is simple, i.e., we have $0 < \lambda_1 < \lambda_2$.*

Proof: Literature. □

1.4.4 Can you hear the shape of a drum?

The eigenvalues of the Laplace Dirichlet problem are completely determined by the domain Ω. This gives rise to the question, whether the converse statement holds as well: Can one, knowing all eigenvalues $\lambda_{n,D}(\Omega)$, $n \in \mathbb{N}$, uniquely (up to trivial congruence transformations) reconstruct the domain Ω?

The answer to this question posed in [9] under the titel "can you hear the shape of a drum?" was answered in [6] as negative as there two non-congruent domains were constructed that have the same spectrum.

We verify this observation numerically using NgSolve (see Jupyter Notebook on TUWEL). Discretizing the eigenvalue problem in a FEM space $S^{p,1}(\mathcal{T}_h)$ on meshes $\mathcal{T}_h$) with maximal mesh sizes $h = \frac{1}{4}$ and different polynomial degrees $p = 1, 2, 3$ produces for both in Figure 1.3 depicted domains the eigenvalues in the following table. Hereby, 10 steps of PINVIT were employed to solve the matrix eigenvalue problem.

	hen			arrow		
$h = \frac{1}{4},\ p:$	1	2	3	1	2	3
λ_1	10.93	10.20	10.17	10.87	10.20	10.17
λ_2	15.92	14.68	14.64	15.87	14.69	14.64
λ_3	23.43	20.83	20.74	23.26	20.83	20.74
$h = \frac{1}{8},\ p:$	1	2	3	1	2	3
λ_1	10.36	10.17	10.16	10.36	10.17	10.16
λ_2	14.95	14.64	14.63	14.93	14.66	14.63
λ_3	21.39	20.74	20.72	21.34	20.74	20.72

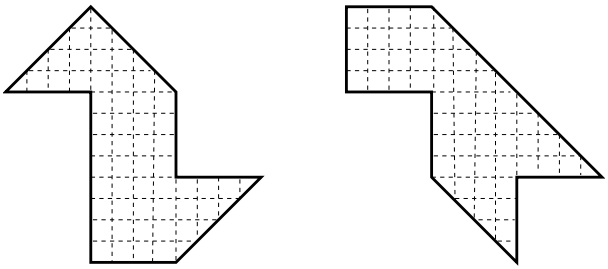

Figure 1.3: two non-congruent, isospectral domains

As theoretically expected, we also observe numerically that the domains produce the same eigenvalues. Figure 1.4 plots the discrete eigenfunctions, which, however, are different.

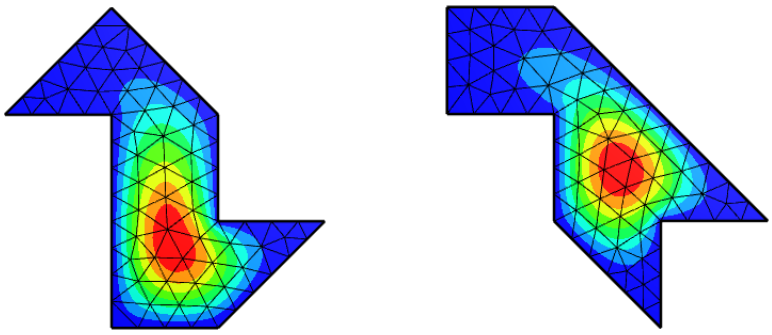

Figure 1.4: first eigenfunction on the domains

Chapter 2

Parabolic problems

A popular class of second order time dependent problems are parabolic equations, which are first order in time and second order in space. Prime examples are

- the heat equation $u_t - \Delta u = f$,
- the (incompressible) Navier-Stokes equations

$$\begin{aligned} \mathbf{u}_t - \mu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{f}, \\ \operatorname{div}\mathbf{u} &= 0. \end{aligned}$$

Let $\Omega \subset \mathbb{R}^d$, $T > 0$. As a model problem, we consider the heat equation on the space-time cylinder $\Omega_T := \Omega \times (0,T)$, i.e., with given data f, u_0, we want to solve

$$u_t - \Delta u = f \quad \text{in } \Omega_T, \tag{2.1a}$$

$$u = 0 \quad \text{on } \Gamma_T := \partial\Omega \times (0,T), \quad \text{"parabolic boundary"} \tag{2.1b}$$

$$u(\cdot,0) = u_0 \quad \text{in } \Omega. \tag{2.1c}$$

Remark 2.1 Problem (2.1) describes the physical model of a temperature distribution in an object Ω at time t. Hereby, u_0 is the initial temperature distribution and f describes an external heat source. The parabolic boundary condition means that there is isolation on the boundary, i.e., the temperature is fixed there.

Some possible numerical methods

1. *Method of lines*: this is classical and probably the most common method. One *fixes* a discretization in space and then solves a resulting ODE system with a numerical method for ODEs (Euler, Runge-Kutta, ...).

   Advantages:

   - "time marching" is memory efficient, no restrictions on end time T.
   - Combination of classical discretizations both in space and time $\implies$ well-understood tools for analysis and implementations available.

   Disadvantage: fixed discretization in space $\rightarrow$ adaptivity not possible!

2. *Rothe-method:* fix the time discretization and solve a space problem only in each time step (allows adaptivity in space, but not in time).

3. *Space-time method:* Do not treat space and time differently, discretize problem (2.1) in $\mathbb{R}^{d+1}$.

   Advantage: adaptivity in space and time possible.

   Disadvantage: more expensive than method of lines.

The roadmap for this chapter is as follows:

- We at first only provide a weak formulation for the heat equation in space to provide a framework for the method of lines, treating the time derivatives in a classical way.
- Then, we introduce the method of lines (semi-discretization) and provide error bounds for the semi-discretization. This is then combined with an *implicit* time stepping method to obtain a fully discrete scheme.
- Afterwards, we weaken the requirements made for the time derivative by deriving a variational formulation in certain Bochner spaces.
- For the fully discrete method we present an error analysis.
- Rather than doing a finite difference method in time, one could also do a Galerkin method in time, which leads to a DG method in Section 2.6.
- Then, we consider full space-time discretizations, at first discussing inf-sup stability of the continuous formulation and then present a different formulation based on least squares FEM that is actually coercive.
- Finally, we introduce numerical methods for the Navier-Stokes equations.

2.1 Variational formulation in space

Goal: framework for method of lines using FEM in space.

Notation: for $\Omega \subset \mathbb{R}^d$, we consider the Hilbert space $L^2(\Omega)$ and denote by $\langle \cdot, \cdot \rangle_{L^2}$ the $L^2(\Omega)$-inner product. Moreover, we consider the bilinear form $a(u,v) := \int_\Omega \nabla u \cdot \nabla v$.

Multiplying (2.1) with a test-function $v = v(x)$ (not depending on t!) and integration over Ω leads to the variational formulation for any classical solution u to (2.1)

$$\langle u_t(\cdot,t), v\rangle_{L^2} + a(u(\cdot,t), v) = \langle f(\cdot,t), v\rangle_{L^2} \qquad \forall v \in H_0^1(\Omega). \tag{2.2}$$

This is well-defined, provided for any fixed t there holds $u(\cdot,t) \in H_0^1(\Omega)$.

Concerning the derivative in time, different formulations can be chosen. Let X be a Banach space.

- Classical derivative as function $u : (0,T) \to X$ defined in the sense: an element $u'(t) \in X$ is called derivative of u at $t \in (0,T)$, if

$$\lim_{h \to 0} \left\| \frac{u(t+h) - u(t)}{h} - u'(t) \right\|_X = 0. \tag{2.3}$$

- Weak derivative: multiplying (2.2) with a test function (depending only on t), integration over $(0,T)$ and integration by parts in time (more on that later).

At first, we consider the case of classical derivatives in time, i.e., we consider function spaces

$$C^1((0,T);X) := \{u : (0,T) \to X : u \text{ is continuously differentiable}\},$$

where the derivative is understood as in (2.3). Inductively, we can define in the same way the spaces $C^m((0,T);X)$ for $m \in \mathbb{N} \cup \{\infty\}$.

We now consider the problem: Find $u \in C^1((0,T); H_0^1(\Omega)) \cap C([0,T]; H_0^1(\Omega))$ such that for all $t \in (0,T)$

$$\langle u'(t), v\rangle_{L^2} + a(u(t), v) = \langle f(t), v\rangle_{L^2} \qquad \forall v \in H_0^1(\Omega) \tag{2.4a}$$

$$u(0) = u_0 \qquad \text{(as functions in } H_0^1(\Omega)). \tag{2.4b}$$

For the data, this formulation requires

- $f \in C([0,T]; L^2(\Omega))$,
- $u_0 \in H_0^1(\Omega)$.

Remark 2.2 We note that formulation (2.4) is not a weak formulation in the idea of the FEM, as time derivatives are not treated in a weak sense[1]. However, this formulation is convenient to employ ODE theory. Other notions of solution can, e.g., be found in lectures on PDE theory.
In particular, we note that the notion of solution can be weakened to also consider initial functions $u_0 \in L^2(\Omega)$.

Exercise 2.3 For $u \in C^1((0,T); H_0^1(\Omega))$, there holds

- $t \mapsto \|u(t)\|^2_{L^2(\Omega)}$ is continuously differentiable,
- $\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} = 2\langle u'(t), u(t)\rangle_{L^2}$,
- $u \in C^1((0,T); L^2(\Omega))$.

Theorem 2.4 *(Energy inequality) Let $\gamma > 0$ be the coercivity constant of $a(\cdot,\cdot)$, i.e. $\gamma\|v\|^2_{H^1(\Omega)} \leq a(v,v)\ \forall v \in H_0^1(\Omega)$ and let u solve (2.4). Then, there holds*

$$\|u\|_{L^2(\Omega)} \leq e^{-\gamma t}\|u_0\|_{L^2(\Omega)} + \int_0^t e^{-\gamma(t-s)}\|f(s)\|_{L^2(\Omega)} ds.$$

Proof: *Step 1:* We at first assume that $\|u(s)\|_{L^2(\Omega)} > 0$ for all $0 < s < t$. Then, there holds

$$\frac{d}{dt}\|u(t)\|_{L^2(\Omega)} = \frac{\langle u'(t), u(t)\rangle_{L^2}}{\|u(t)\|_{L^2(\Omega)}}.$$

Using, for fixed s, the test function $v = u(s)$ in (2.4) gives

$$\langle f(s), u(s)\rangle_{L^2} \overset{(2.4)}{=} \langle u'(s), u(s)\rangle_{L^2} + a(u(s), u(s)) = \|u(s)\|_{L^2(\Omega)} \frac{d}{dt}\|u(t)\|_{L^2(\Omega)}\Big|_{t=s} + a(u(s), u(s))$$

and by Cauchy-Schwarz this implies

$$a(u(s), u(s)) + \|u(s)\|_{L^2(\Omega)} \frac{d}{dt}\|u(t)\|_{L^2(\Omega)}\Big|_{t=s} \leq \|f(s)\|_{L^2(\Omega)}\|u(s)\|_{L^2(\Omega)}. \tag{2.5}$$

Since $a(u(s), u(s)) \geq \gamma\|u(s)\|^2_{H^1(\Omega)} \geq \gamma\|u(s)\|^2_{L^2(\Omega)}$, we obtain

$$\gamma\|u(s)\|_{L^2(\Omega)} + \frac{d}{dt}\|u(t)\|_{L^2(\Omega)}\Big|_{t=s} \leq \|f(s)\|_{L^2(\Omega)} \qquad \forall\ 0 \leq s \leq t.$$

An integrating factor for the left-hand side of this differential inequality is given by $e^{\gamma t}$, which leads to

$$\frac{d}{dt}\left(e^{\gamma t}\|u(t)\|_{L^2(\Omega)}\right)\Big|_{t=s} = e^{\gamma s}\left(\gamma\|u(s)\|_{L^2(\Omega)} + \frac{d}{dt}\|u(t)\|_{L^2(\Omega)}\Big|_{t=s}\right) \leq e^{\gamma s}\|f(s)\|_{L^2(\Omega)}.$$

Now, integration over $(0,t)$ provides

$$e^{\gamma t}\|u(t)\|_{L^2(\Omega)} - e^{\gamma 0}\|u(0)\|_{L^2(\Omega)} \leq \int_0^t e^{\gamma s}\|f(s)\|_{L^2(\Omega)} ds$$

or

$$\|u(t)\|_{L^2(\Omega)} \leq e^{-\gamma t}\|u_0\|_{L^2(\Omega)} + \int_0^t e^{-\gamma(t-s)}\|f(s)\|_{L^2(\Omega)} ds.$$

[1] it is also possible to differentiate $u \in L^1_{loc}(0,T; H_0^1(\Omega))$ in a distributional sense.

Step 2: For the general case $\|u(s)\|_{L^2(\Omega)} \geq 0$ on $[0,T]$, (2.5) still holds, but we can not divide by $\|u(s)\|_{L^2(\Omega)}$ any more. However, using this inequality with $\sqrt{\|u(t)\|_{L^2}^2 + \varepsilon^2}$, essentially the same arguments can be employed and the statement follows from sending $\varepsilon \to 0$ in the last step. □

Exercise 2.5 Theorem 2.4 gives uniqueness of solutions to (2.4).

Remark 2.6 For $f \equiv 0$, the heat equation is *dissipative* (in $L^2(\Omega)$), i.e.,

$$\|u(t)\|_{L^2(\Omega)} \leq e^{-\gamma t}\|u(0)\|_{L^2(\Omega)}.$$

Then, for two different initial conditions $u_0,\ \tilde{u}_0$, there holds for the corresponding solutions $u(t),\ \tilde{u}(t)$ that $\|u(t) - \tilde{u}(t)\|_{L^2(\Omega)} \leq e^{-\gamma t}\|u_0 - \tilde{u}_0\|_{L^2(\Omega)}$. Reasonable numerical methods should also produce this qualitative behaviour.

Definition 2.7 (Evolution operator) *Let $f = 0$. We define the* evolution operator

$$E(t) : H_0^1(\Omega) \to H_0^1(\Omega),\ u_0 \mapsto u(t),$$

where $u(t)$ denotes the solution of (2.4) with initial data u_0.

By uniqueness of the solutions to the heat equation, the evolution operator has the *semi-group* property:

- $E(t+s) = E(t) \circ E(s), \qquad t,\ s \geq 0,$
- $E(0) = \mathrm{Id}.$

Using the eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$ and eigenfunctions $(\varphi_n)_{n\in\mathbb{N}}$ of the Dirichlet Laplacian, which form an ONB of $L^2(\Omega)$ and an orthogonal basis of $H_0^1(\Omega)$, we can explicitly express the evolution operator $E(t)$ as

$$E(t)u_0 = \sum_{n=1}^{\infty} e^{-\lambda_n t}\langle u_0, \varphi_n\rangle_{L^2}\varphi_n. \tag{2.6}$$

This formula actually suggests that the evolution operator can also be well defined for functions $u_0 \in L^2(\Omega)$ (more on that later).

Remark 2.8 (Duhamel-principle) Let f be sufficiently smooth. Then, the solution u to (2.4) can be written as

$$u(t) = E(t)u_0 + \int_0^t E(t-s)f(s)\,ds. \tag{2.7}$$

2.2 Semi-discretization (method of lines)

Goal: approximation of (2.4) by a (finite) system of ODEs.

Let $V_h \subset H_0^1(\Omega)$ with $\dim(V_h) = N < \infty$ and take a basis $\{\varphi_i \mid i = 1,\dots,N\}$. Let $u_{0,h} \in V_h$ be an approximation to u_0. Then, the *semi-discrete approximation* u_h to the solution u of (2.4) is given by the problem:
Find $u_h \in C^1((0,T);V_h) \cap C([0,T];V_h)$ such that

$$\langle u_h'(t), v_h\rangle_{L^2} + a(u_h(t), v_h) = \langle f(t), v_h\rangle_{L^2} \qquad \forall\ v_h \in V_h \tag{2.8a}$$

$$u_h(0) = u_{0,h} \tag{2.8b}$$

Now, we write $u_h(t)$ w.r.t. the given basis as $u_h(t) = \sum_{i=1}^{N} \mathbf{u}_i(t)\varphi_i$ and analogously $u_{0,h} = \sum_{i=1}^{N} \mathbf{u}_{0,i}\,\varphi_i$. Inserting this into (2.8) leads to the equivalent[2] system of ODEs

$$\mathbf{M}\mathbf{u}'(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{F}(t) \qquad t > 0 \tag{2.9a}$$
$$\mathbf{u}(0) = \mathbf{u}_0, \tag{2.9b}$$

where the *stiffness matrix* $\mathbf{A} \in \mathbb{R}^{N\times N}$ and the *mass matrix* $\mathbf{M} \in \mathbb{R}^{N\times N}$ are given by

$$\mathbf{A}_{ij} = a(\varphi_j, \varphi_i), \qquad \mathbf{M}_{ij} = \langle \varphi_j, \varphi_i \rangle_{L^2}, \quad i,j = 1,\dots,N \tag{2.10}$$

and the right-hand side $\mathbf{F}(t)$ is defined as

$$\mathbf{F}_i(t) = \langle f(t), \varphi_i \rangle_{L^2}. \tag{2.11}$$

Exercise 2.9 Show: The matrices $\mathbf{A}$ and $\mathbf{M}$ are SPD. Moreover, for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$ with $v = \sum_{i=1}^{N} \mathbf{v}_i\varphi_i$, $w = \sum_{i=1}^{N} \mathbf{w}_i\varphi_i$, there holds $\mathbf{v}^T\mathbf{M}\mathbf{w} = \langle w, v\rangle_{L^2}$ and $\mathbf{v}^T\mathbf{A}\mathbf{w} = a(w, v)$.

Remark 2.10 Exercise 2.9 implies that (2.9) is equivalent to

$$\mathbf{u}' = \mathbf{M}^{-1}\mathbf{F}(t) - \mathbf{M}^{-1}\mathbf{A}\mathbf{u}(t),$$
$$\mathbf{u}(0) = \mathbf{u}_0.$$

In particular, the existence and uniqueness of solutions to (2.8) holds.

For the semi-discrete case there also holds an energy inequality.

Lemma 2.11 *Let $r \in C([0,T]; L^2(\Omega))$ and $w_h \in C^1((0,T); V_h) \cap C([0,T]; V_h)$ satisfy*

$$\langle w_h'(t), v\rangle_{L^2} + a(w_h(t), v_h) = \langle r(t), v_h\rangle_{L^2} \text{ for all } v_h \in V_h,$$

and let $\gamma > 0$ denote the coercivity constant of $a(\cdot,\cdot)$. Then,

$$\|w_h(t)\|_{L^2(\Omega)} \le e^{-\gamma t}\|w_h(0)\|_{L^2(\Omega)} + \int_0^t e^{-\gamma(t-s)}\|r(s)\|_{L^2(\Omega)}ds.$$

Proof: Analogous to Theorem 2.4. □

Remark 2.12 In the same way as for the continuous problem, we can define a semi-discrete evolution operator E_h that maps, for fixed t,

$$E_h(t) : V_h \to V_h,\ u_{0,h} \mapsto u_h(t)$$

[2]The equivalence follows from the calculation

$$\begin{aligned}
&u_h \text{ solves (2.8a)}\\
&\Leftrightarrow \left\langle \sum_{j=1}^{N} \mathbf{u}_j'(t)\varphi_j, \sum_{i=1}^{N} \mathbf{v}_i\varphi_i \right\rangle_{L^2} + a\left(\sum_{j=1}^{N} \mathbf{u}_j(t)\varphi_j, \sum_{i=1}^{N} \mathbf{v}_i\varphi_i\right) = \left\langle f(t), \sum_{i=1}^{N} \mathbf{v}_i\varphi_i \right\rangle_{L^2} \qquad \forall\, \mathbf{v} \in \mathbb{R}^N\\
&\Leftrightarrow \sum_{i,j=1}^{N} \mathbf{u}_j'(t)\mathbf{v}_i\langle\varphi_j,\varphi_i\rangle_{L^2} + \sum_{i,j=1}^{N} \mathbf{u}_j(t)\mathbf{v}_i\, a(\varphi_j,\varphi_i) = \sum_{i=1}^{N} \mathbf{v}_i\langle f(t),\varphi_i\rangle_{L^2} \qquad \forall\, \mathbf{v} \in \mathbb{R}^N\\
&\Leftrightarrow \mathbf{v}^T\mathbf{M}\mathbf{u}'(t) + \mathbf{v}^T\mathbf{A}\mathbf{u}(t) = \mathbf{v}^T\mathbf{F}(t) \qquad \forall\, \mathbf{v} \in \mathbb{R}^N\\
&\Leftrightarrow \mathbf{M}\mathbf{u}'(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{F}(t)
\end{aligned}$$

with u_h denoting the solution to (2.9) with initial data $u_{0,h}$ and $f \equiv 0$. Lemma 2.11 shows that this is a bounded operator in $L^2(\Omega)$.
Using the discrete eigenfunctions and eigenvalue, an expansion like (2.6) holds (exercise). Moreover, the discrete Duhamel principle holds:

$$u_h(t) = E_h(t)u_{0,h} + \int_0^t E_h(t-s)\Pi^{L^2} f(s)\, ds$$

with the L^2-orthogonal projection $\Pi^{L^2} : L^2(\Omega) \to V_h$. ∎

In order to obtain a bound for the semi-discretization, we employ the Ritz projection $P_h : H_0^1(\Omega) \to V_h$ from (1.16).

Theorem 2.13 *Let u solve (2.4) and u_h solve (2.8). Then,*

$$\begin{aligned}\|u(t) - u_h(t)\|_{L^2(\Omega)} \le \|u(t) - P_h u(t)\|_{L^2(\Omega)} + e^{-\gamma t}\|u_h(0) - P_h u(0)\|_{L^2(\Omega)} \\ + \int_0^t e^{-\gamma(t-s)}\|u'(s) - P_h u'(s)\|_{L^2(\Omega)} ds.\end{aligned}$$

Proof: We decompose $u_h(t) - u(t) = \underbrace{u_h(t) - P_h u(t)}_{=:\psi(t)} + \underbrace{P_h u(t) - u(t)}_{=:\varrho(t)}$ and employ the triangle inequality to obtain

$$\|u_h(t) - u(t)\|_{L^2(\Omega)} \le \|\psi(t)\|_{L^2(\Omega)} + \|\varrho(t)\|_{L^2(\Omega)}.$$

Step 1: Linearity and boundedness of P_h together with $u \in C^1((0,T); H_0^1(\Omega))$ imply $(P_h u)' = P_h u'$, since

$$\begin{aligned}\lim_{k\to 0} \left\| \frac{1}{k}\left(P_h u(t+k) - P_h u(t)\right) - P_h u'(t)\right\|_{H^1(\Omega)} &= \lim_{k\to 0} \left\| P_h\left(\frac{u(t+k)-u(t)}{k} - u'(t)\right)\right\|_{H^1(\Omega)} \\ &\le \lim_{k\to 0} \|P_h\| \left\|\frac{u(t+k)-u(t)}{k} - u'(t)\right\|_{H^1(\Omega)} = 0.\end{aligned}$$

Step 2: Step 1 implies $\psi \in C^1((0,T); V_h) \cap C([0,T]; V_h)$. Moreover, for any $v \in V_h$, there holds

$$\begin{aligned}\langle \psi'(t), v\rangle_{L^2} + a(\psi(t), v) &= \langle u_h', v\rangle_{L^2} + a(u_h, v) - \langle P_h u', v\rangle_{L^2} - a(P_h u, v) \\ &\overset{(2.8)}{=} \langle f(t), v\rangle_{L^2} - \langle P_h u', v\rangle_{L^2} - a(P_h u, v) \\ &= \langle f(t), v\rangle_{L^2} - \langle P_h u', v\rangle_{L^2} - a(u, v) \\ &\overset{(2.4)}{=} \langle u', v\rangle_{L^2} - \langle P_h u', v\rangle_{L^2} = \langle u' - P_h u'(t), v\rangle_{L^2}.\end{aligned}$$

Step 3: Lemma 2.11 applied to ψ gives

$$\|\psi(t)\|_{L^2(\Omega)} \le e^{-\gamma t} \underbrace{\|\psi(0)\|_{L^2(\Omega)}}_{=\|u_h(0) - P_h u(0)\|_{L^2(\Omega)}} + \int_0^t e^{-\gamma(t-s)}\|u'(s) - P_h u'(s)\|_{L^2(\Omega)} ds,$$

which shows the claimed estimate. □

Remark 2.14 Theorem 2.13 shows that the semi-discretization error $\|u(t) - u_h(t)\|_{L^2(\Omega)}$ can be estimated by an approximation error $\|u(t) - P_h u(t)\|_{L^2(\Omega)}$ and two additional terms that describe the error accumulation for times $0 \le s < t$. This is also referred to as "memory" of parabolic equations. ∎

As is typical in FEM theory, assuming regularity for u, one can derive explicit error estimates in terms of h.

Corollary 2.15 *Let $V_h = S_0^{1,1}(\mathcal{T})$, with $\mathcal{T}$ being a shape-regular, regular triangulation. Assume $u \in C^3(\bar{\Omega} \times [0,T])$ for the solution u to (2.4). Let $u_{0,h} \in V_h$ be either $u_{0,h} = P_h u_0$ or the piecewise linear interpolation $u_{0,h} = I u_0$. Then,*

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch \max_{0 \leq s \leq t} \left(|u(s)|_{H^2(\Omega)} + |u_t(s)|_{H^2(\Omega)} \right).$$

If Ω is convex, then

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch^2 \max_{0 \leq s \leq t} \left(|u(s)|_{H^2(\Omega)} + |u_t(s)|_{H^2(\Omega)} \right).$$

Proof: For all $v \in H^2(\Omega)$, there holds using the nodal interpolant Iv

- $\|v - P_h v\|_{L^2(\Omega)} \leq \|v - P_h v\|_{H^1(\Omega)} \leq C\|v - Iv\|_{H^1(\Omega)} \leq Ch|v|_{H^2(\Omega)}$.
- If Ω is convex, the Aubin-Nitsche duality argument gives $\|v - P_h v\|_{L^2(\Omega)} \leq Ch^2|v|_{H^2(\Omega)}$.

Then, the statement follows from Theorem 2.13. □

Remark 2.16 Note that the regularity requirement $u \in C^3(\bar{\Omega} \times [0,T])$ is very generous. A more careful look at Theorem 2.13 shows that one only needs $u(t) \in H^2(\Omega)$ as well as $\int_0^t \|u'(s)\|^2_{H^2(\Omega)} ds < \infty$.

2.3 Fully discrete methods

2.3.1 Time discretization

The semi-discretization approach in the previous section leads to the system of ODEs

$$\mathbf{M}\mathbf{u}' + \mathbf{A}\mathbf{u} = \mathbf{F}, \qquad \mathbf{u}(0) = \mathbf{u}_0 \tag{2.12}$$

with $\mathbf{M}$ and $\mathbf{A}$ SPD.

Question: how to discretize this ODE in time?

Answer: employ methods for *stiff ODEs*, i.e., A-stable or even L-stable methods.

In order to motivate this choice, we need to understand the qualitative behaviour of solutions to (2.12). Therefore, we transform (2.12) into a system of decoupled ODEs.

Theorem 2.17 *Let $\mathbf{A}$ and $\mathbf{M} \in \mathbb{R}^{N \times N}$ be SPD. Then, for the generalized eigenvalue problem*

$$\text{find } (\mathbf{v}, \lambda) \in \mathbb{R}^N \setminus \{0\} \times \mathbb{C}, \text{ s.t. } \mathbf{A}\mathbf{v} = \lambda \mathbf{M}\mathbf{v} \tag{2.13}$$

there holds:

(i) The eigenvalues λ satisfy $\lambda > 0$.

(ii) There are N eigenpairs $(\mathbf{v}_i, \lambda_i)$, $i = 1, \ldots, N$, which are orthogonal w.r.t. $(\cdot,\cdot)_{\mathbf{A}}$ and $(\cdot,\cdot)_{\mathbf{M}}$, i.e.,

$$(\mathbf{v}_i, \mathbf{v}_j)_{\mathbf{M}} = \langle \mathbf{M}\mathbf{v}_i, \mathbf{v}_j \rangle_2 = 0 \qquad \forall i \neq j,$$
$$(\mathbf{v}_i, \mathbf{v}_j)_{\mathbf{A}} = \langle \mathbf{A}\mathbf{v}_i, \mathbf{v}_j \rangle_2 = 0 \qquad \forall i \neq j.$$

(iii) The matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N) \in \mathbb{R}^{N \times N}$ diagonalizes $\mathbf{M}$ and $\mathbf{A}$ simultaneously, i.e,

$$\mathbf{V}^T \mathbf{M} \mathbf{V} = \mathbf{D}_1$$
$$\mathbf{V}^T \mathbf{A} \mathbf{V} = \mathbf{D}_2$$

with $\mathbf{D}_1, \mathbf{D}_2$ being diagonal matrices.

(iv) If the $\mathbf{v}_i$ are normalized, such that $(\mathbf{v}_i, \mathbf{v}_j)_{\mathbf{M}} = \delta_{ij}$, then

$$\mathbf{V}^T\mathbf{M}\mathbf{V} = \mathbf{I}, \quad \mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{D} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}.$$

Proof: Exercise. *Hint:* Consider the EVP $\mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{M}^{-\frac{1}{2}}\mathbf{x} = \lambda\mathbf{x}$. □

Now, define $\widetilde{\mathbf{u}} = \mathbf{V}^{-1}\mathbf{u}$, $\widetilde{\mathbf{F}} = \mathbf{V}^T\mathbf{F}$, $\widetilde{\mathbf{u}}_0 = \mathbf{V}^{-1}\mathbf{u}_0$, then (2.12) is equivalent to

$$\widetilde{\mathbf{u}}' + \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix} \widetilde{\mathbf{u}} = \widetilde{\mathbf{F}}, \quad \widetilde{\mathbf{u}}(0) = \widetilde{\mathbf{u}}_0. \tag{2.14}$$

(2.14) is a stiff system of ODEs, if some of the eigenvalues $\lambda_i > 0$ are large (see lecture "numerics of ODEs" for more details on that). For parabolic problems, this is indeed the case.

Theorem 2.18 *Let $\mathcal{T}$ be a regular, shape-regular triangulation and set $h_{min} = \min_{K\in\mathcal{T}} h_K$. Take a basis of $V_h = S_0^{1,1}(\mathcal{T})$ consisting of hat functions and let $\mathbf{A}$ and $\mathbf{M}$ be the corresponding stiffness and mass matrix. Then, there exists a constant $C > 0$, depending only on the shape-regularity of $\mathcal{T}$, such that*

$$C^{-1}\|u\|^2_{L^2(\Omega)} \le |u|^2_{H^1(\Omega)} \le \frac{C}{h^2_{min}}\|u\|^2_{L^2(\Omega)} \qquad \forall\ u \in S_0^{1,1}(\mathcal{T}).$$

In particular, the eigenvalues λ_i of (2.13) satisfy

$$C^{-1} \le \min_{i=1,\dots,N} \lambda_i \le \max_{i=1,\dots,N} \lambda_i \le \frac{C}{h^2_{min}}. \tag{2.15}$$

Proof: Exercise. In particular, the upper bound $O(h^{-2}_{min})$ is sharp. □

Single step methods

In the following, we consider single step methods on an uniform temporal mesh (theory for non-uniform meshes is also possible) and define the time steps as $t_n = nk$, where $k > 0$ is a fixed time step length. We briefly introduce the notions of convergence and stability for such methods, details can be found in the lecture "numerics of ODEs".

We consider single step methods with the properties:

1. (Stability function R) Applied to the scalar ODE $y' = \lambda y$, one step of the method can be written as
$$y_1 = R(k\lambda)y_0$$
with some function R.
2. (Coordinate invariance) Let $\mathbf{B} \in \mathbb{R}^{N\times N} = \mathbf{V}^{-1}\mathbf{D}\mathbf{V}$ be diagonalizable with diagonal matrix $\mathbf{D} = \mathrm{diag}(d_1, \dots, d_N)$. Then, the change of basis $\mathbf{x} \mapsto \mathbf{V}\mathbf{x}$ should commute with the application of the method to $\mathbf{y}' = \mathbf{B}\mathbf{y}$, i.e., for $\widetilde{\mathbf{y}}_1 = \mathbf{V}\mathbf{y}_1$ and $\widetilde{\mathbf{y}}_0 = \mathbf{V}\mathbf{y}_0$, there holds
$$\widetilde{\mathbf{y}}_1 = \mathrm{diag}(R(kd_1), \dots, R(kd_N))\widetilde{\mathbf{y}}_0.$$

Possible choices (written for a general ODE (system) $y' = f(t, y)$) are

- *explicit Euler:* $y_1 = y_0 + kf(t_0, y_0)$ with the stability function $R(z) = 1 + z$;
- *implicit Euler:* $y_1 = y_0 + kf(t_1, y_1)$ with the stability function $R(z) = \frac{1}{1-z}$;

- *θ-scheme:* $y_1 = y_0 + kf(t_0 + \theta(t_1 - t_0), y_0 + \theta(y_1 - y_0))$ as generalization:
  - $\theta = 0$: explicit Euler;
  - $\theta = 1$: implicit Euler;
  - $\theta = \frac{1}{2}$: Crank-Nicolson method with the stability function $R(z) = \frac{1+z/2}{1-z/2}$.
- more general: Runge-Kutta methods (see lecture numerics of ODEs). They always satisfy the coordinate invariance property (exercise)!

We call a method *consistent of order p*, if one step of the method allows for an error bound $|y(k) - y_1| \leq Ck^{p+1}$. This gives that the function $R(\cdot)$ describing the method should have the asymptotics

$$R(z) = e^z + \mathcal{O}(|z|^{p+1}) \qquad \text{as } z \to 0. \tag{2.16}$$

In the lecture numerics of ODEs, one sees that consistency of order p together with stability (more on that later) gives convergence of order p.
The Euler methods are of order $p = 1$, the Crank-Nicolson method is second order $p = 2$.

Applying the ODE method to $y' = \lambda y$ for $\lambda < 0$ (stiff ODE!) with exact solution $y(t) = e^{\lambda t} y_0$, shows that numerical solutions should stay bounded. Thus, due to

$$|y_n| = |R(k\lambda)|^n |y_0|$$

we want to have a method satisfying $|R(z)| \leq 1$ for all $z < 0$ (more on that later). Such methods are called *A-stable*.

For $\lambda < 0$ the exact solution even has the property that $y(t) \to 0$ for $t \to \infty$. Methods that reproduce this qualitative behaviour, i.e., satisfy $R(z) \to 0$ for $z \to -\infty$, are called *L-stable*.

The simplest A-stable method is the implicit Euler method considered in the following.

Example 2.19 For the implicit Euler method, the stability function satisfies

$$|R(z)| = \frac{1}{|1 - z|} = \frac{1}{1 - z} < 1 \qquad \forall z < 0.$$

Consequently, the method is A-stable. Moreover, there holds $R(z) \to 0$ for $z \to -\infty$, i.e., the method is even L-stable. The necessity of stable methods can be see in the numerical examples in Section 2.3.5.

2.3.2 The implicit Euler method

Applying the implicit Euler method to the variational formulation of our semi-discrete model problem gives

$$\left\langle \frac{u_h^{n+1} - u_h^n}{k}, v \right\rangle_{L^2} + a(u_h^{n+1}, v) = \langle f(t_{n+1}), v \rangle_{L^2} \qquad \forall\, v \in V_h, \tag{2.17}$$

where $k > 0$ denotes the time step, $t_n = nk$ and $u_h^n \approx u_h(t_n)$. In matrix notation (2.17) translates to

$$\frac{1}{k}\mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + \mathbf{A}\mathbf{u}^{n+1} = \mathbf{F}^{n+1}, \tag{2.18}$$

i.e., in each step, a linear system

$$(\mathbf{M} + k\mathbf{A})\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n + k\mathbf{F}^{n+1}$$

has to be solved.

Remark 2.20 In terms of computational efficiency, it is advisable to compute a Cholesky-decomposition of $\mathbf{M} + k\mathbf{A}$ once and then do forward/backward substitution only in each time step. ∎

Assuming sufficient regularity for the exact solution, we can now obtain an error estimate for the fully discrete method.

Theorem 2.21 *Let u solve (2.4) and additionally assume $u \in C^2([0,T];L^2(\Omega))$. Let $k_0 > 0$ be fixed and denote by λ_{min} the smallest eigenvalue[3] of*

$$\lambda \mathbf{M} x = \mathbf{A} x,$$

where $\mathbf{M}$ and $\mathbf{A}$ are mass matrix and stiffness matrix of the spatial discretization. Then, there exists $b > 0$, depending only on k_0 and λ_{min}, such that for all $k \in (0, k_0]$ there holds

$$\begin{aligned}\|u_h^n - u(t_n)\|_{L^2(\Omega)} &\le \|u(t_n) - P_h u(t_n)\|_{L^2(\Omega)} + e^{-bt_n}\|u_{0,h} - P_h u_0\|_{L^2(\Omega)} \\ &\quad + \int_0^{t_n} e^{-b(t_n - t)} \Big(\|u'(t) - P_h u'(t)\|_{L^2(\Omega)} + k \, \|u''(t)\|_{L^2(\Omega)} \Big) \, dt, \end{aligned} \tag{2.19}$$

as well as

$$\begin{aligned}&\|u_h^n - u(t_n)\|_{L^2(\Omega)}^2 + \sum_{j=1}^n \|u_h^j - u_h^{j-1}\|_{L^2(\Omega)}^2 + k|u_h^j - u(t_j)|_{H^1(\Omega)}^2 \\ &\qquad \le C\Big(\|u_{0,h} - P_h u_0\|_{L^2(\Omega)}^2 + \int_0^{t_n} \|u' - P_h u'\|_{L^2(\Omega)}^2 + k^2 \|u''\|_{L^2(\Omega)}^2 \, dt \\ &\qquad\quad + \|u(t_n) - P_h u(t_n)\|_{L^2(\Omega)}^2 + k \sum_{j=1}^n |u(t_{j-1}) - P_h u(t_{j-1})|_{H^1(\Omega)}^2 \Big). \end{aligned} \tag{2.20}$$

Proof: *Proof of (2.19):* We split the error into

$$u_h^n - u(t_n) = \underbrace{u_h^n - P_h u(t_n)}_{=:\psi_h^n} + \underbrace{P_h u(t_n) - u(t_n)}_{=:\varrho^n}$$

and obtain recurrence formulas for the error that will then be solved explicitly. In the following, we abbreviate $D^k \psi_h^n := \frac{1}{k}(\psi_h^n - \psi_h^{n-1})$.
Step 1: (recurrence relation for ψ^n) By definition of u_h^n and u, we have

$$\begin{aligned} \frac{1}{k}\langle u_h^n - u_h^{n-1}, v\rangle_{L^2} + a(u_h^n, v) &= \langle f(t_n), v\rangle_{L^2} \qquad \forall\ v \in V_h \\ \langle u'(t_n), v\rangle_{L^2} + a(u(t_n), v) &= \langle f(t_n), v\rangle_{L^2} \qquad \forall\ v \in H_0^1(\Omega), \end{aligned} \tag{2.21}$$

and applying both equations gives for all $v \in V_h$

$$\begin{aligned}\langle D^k \psi_h^n, v\rangle_{L^2} + a(\psi_h^n, v) &= \langle D^k u_h^n, v\rangle_{L^2} + a(u_h^n, v) - \langle D^k P_h u(t_n), v\rangle_{L^2} - a(P_h u(t_n), v) \\ &= \langle f(t_n), v\rangle_{L^2} - \langle D^k P_h u(t_n), v\rangle_{L^2} - a(u(t_n), v) \\ &= \langle u'(t_n) - D^k P_h u(t_n), v\rangle_{L^2} =: \langle w^n, v\rangle_{L^2}. \end{aligned}$$

Using $v = \psi_h^n \in V_h$ as test function, we obtain

$$\boxed{\|\psi_h^n\|_{L^2(\Omega)}^2 - \langle \psi_h^{n-1}, \psi_h^n\rangle_{L^2} = -k \, a(\psi_h^n, \psi_h^n) + k\langle w^n, \psi_h^n\rangle_{L^2}.} \tag{2.22}$$

Now, the Poincaré inequality $|\psi|_{H^1(\Omega)}^2 \ge \lambda_{min}\|\psi\|_{L^2(\Omega)}^2$ together with $a(\psi,\psi) = |\psi|_{H^1(\Omega)}^2$ and the Cauchy-Schwarz inequality imply

$$\|\psi_h^n\|_{L^2(\Omega)}^2 \le \|\psi_h^{n-1}\|_{L^2(\Omega)}\|\psi_h^n\|_{L^2(\Omega)} - k\lambda_{min}\|\psi_h^n\|_{L^2(\Omega)}^2 + k\|w^n\|_{L^2(\Omega)}\|\psi_h^n\|_{L^2(\Omega)}.$$

Dividing by the L^2-norm of ψ_h^n and $(1 + k\lambda_{min})$ this implies

$$\|\psi_h^n\|_{L^2(\Omega)} \le \frac{1}{1 + k\lambda_{min}}\|\psi_h^{n-1}\|_{L^2(\Omega)} + \frac{k}{1 + k\lambda_{min}}\|w^n\|_{L^2(\Omega)}.$$

[3] for the heat equation, we have $\lambda_{min} = \mathcal{O}(1)$, i.e., this is harmless.

Step 2: (solving the recurrence) Iteratively applying step 1 gives

$$\|\psi_h^n\|_{L^2(\Omega)} \le (1+k\lambda_{min})^{-n}\|\psi_h^0\|_{L^2(\Omega)} + \frac{k}{1+k\lambda_{min}}\sum_{j=1}^{n}(1+k\lambda_{min})^{-(n-j)}\|w^j\|_{L^2(\Omega)}.$$

As the function $x \mapsto (1+x)^{-1/x}$ is monotonously increasing, with $\lambda_{min}k \le \lambda_{min}k_0$ and $t_n = nk$ there follows

$$\begin{aligned}(1+\lambda_{min}k)^{-n} &= (1+\lambda_{min}k)^{-t_n\lambda_{min}/(\lambda_{min}k)} \le (1+\lambda_{min}k_0)^{-t_n\lambda_{min}/(\lambda_{min}k_0)} = e^{-bt_n}\\ (1+\lambda_{min}k)^{-(n-(j-1))} &\le (1+\lambda_{min}k_0)^{-t_{n-j+1}\lambda_{min}/(\lambda_{min}k_0)} = e^{-b(t_n-t_{j-1})},\end{aligned}$$

where $b>0$ is defined by $(1+\lambda_{min}k_0)^{-\lambda_{min}/(\lambda_{min}k_0)} = e^{-b}$. This leads to

$$\|\psi_h^n\|_{L^2(\Omega)} \le e^{-bt_n}\|\psi_h^0\|_{L^2(\Omega)} + k\sum_{j=1}^{n} e^{-b(t_n-t_{j-1})}\|w^j\|_{L^2(\Omega)}.$$

Step 3: (estimate of w^j) It remains to estimate $w^j = u'(t_j) - D^k P_h u(t_j)$. Using Taylor expansion, we may write

$$\begin{aligned}u'(t_j) &= \frac{u(t_j)-u(t_{j-1})}{k} + \frac{1}{k}\int_{t_{j-1}}^{t_j}(t-t_{j-1})u''(t)dt\\ &= \frac{P_hu(t_j)-P_hu(t_{j-1})}{k} + \frac{u(t_j)-u(t_{j-1})}{k} - \frac{P_hu(t_j)-P_hu(t_{j-1})}{k} + \frac{1}{k}\int_{t_{j-1}}^{t_j}(t-t_{j-1})u''(t)dt\\ &= D^kP_hu(t_j) + \frac{1}{k}\int_{t_{j-1}}^{t_j}u'(t)-P_hu'(t)dt + \frac{1}{k}\int_{t_{j-1}}^{t_j}(t-t_{j-1})u''(t)dt.\end{aligned}$$

Consequently, we obtain

$$\|w^j\|_{L^2(\Omega)} \le \frac{1}{k}\int_{t_{j-1}}^{t_j}\|u'(t)-P_hu'(t)\|_{L^2(\Omega)}dt + \int_{t_{j-1}}^{t_j}\|u''(t)\|_{L^2(\Omega)}dt. \tag{2.23}$$

Together with step 2 this implies

$$\begin{aligned}\|\psi_h^n\|_{L^2(\Omega)} &\le e^{-bt_n}\|\psi_h^0\|_{L^2(\Omega)} + \sum_{j=1}^{n}e^{-b(t_n-t_{j-1})}\int_{t_{j-1}}^{t_j}\|u'(t)-P_hu'(t)\|_{L^2(\Omega)} + k\|u''(t)\|_{L^2(\Omega)}dt\\ &\le e^{-bt_n}\|\psi_h^0\|_{L^2(\Omega)} + \int_0^{t_n}e^{-b(t_n-t)}\Big(\|u'(t)-P_hu'(t)\|_{L^2(\Omega)} + k\|u''(t)\|_{L^2(\Omega)}\Big)dt,\end{aligned}$$

which shows the first statement of the theorem.

Proof of (2.20): We start with (2.22) and estimate

$$\begin{aligned}\underbrace{\langle\psi_h^n-\psi_h^{n-1},\psi_h^n\rangle_{L^2}}_{=\frac{1}{2}\left(\|\psi_h^n\|^2-\|\psi_h^{n-1}\|^2+\|\psi_h^n-\psi_h^{n-1}\|^2\right)} &+k|\psi_h^n|^2_{H^1(\Omega)} = k\langle w^n,\psi_h^n\rangle_{L^2}\\ &\le k\|w^n\|_{L^2(\Omega)}\|\psi_h^n\|_{L^2(\Omega)} \le kC_P\|w^n\|_{L^2(\Omega)}|\psi_h^n|_{H^1(\Omega)}.\end{aligned}$$

Consequently,

$$\begin{aligned}\|\psi_h^n\|^2_{L^2(\Omega)} - \|\psi_h^{n-1}\|^2_{L^2(\Omega)} + \|\psi_h^n-\psi_h^{n-1}\|^2_{L^2(\Omega)} &+ 2k|\psi_h^n|^2_{H^1(\Omega)}\\ &\le 2kC_P|\psi_h^n|_{H^1(\Omega)}\|w^n\|_{L^2(\Omega)}\\ &\le k|\psi_h^n|^2_{H^1(\Omega)} + C_P^2k\|w^n\|^2_{L^2(\Omega)}.\end{aligned}$$

Subtracting $k|\psi_h^{n+1}|_{H^1(\Omega)}$ on both sides and employing this estimate for $j = 1, \dots, n$ leads to

$$\|\psi_h^n\|^2_{L^2(\Omega)} - \|\psi_h^0\|^2_{L^2(\Omega)} + \sum_{j=1}^n \|\psi_h^j - \psi_h^{j-1}\|^2_{L^2(\Omega)} + k|\psi_h^j|^2_{H^1(\Omega)} \leq C_P^2 k \sum_{j=1}^n \|w^j\|^2_{L^2(\Omega)}.$$

As with (2.23) we can estimate

$$k \sum_{j=1}^n \|w^j\|^2_{L^2(\Omega)} \leq \int_0^{t_n} \|u' - P_h u'\|^2_{L^2(\Omega)} + k^2 \|u''\|^2_{L^2(\Omega)}\, dt,$$

there holds

$$\|\psi_h^n\|^2_{L^2(\Omega)} + \sum_{j=1}^n \|\psi_h^j - \psi_h^{j-1}\|^2_{L^2(\Omega)} + k|\psi_h^j|^2_{H^1(\Omega)} \leq C_P^2 \left(\|\psi_h^0\|^2_{L^2(\Omega)} + \int_0^{t_n} \|u' - P_h u'\|^2_{L^2(\Omega)} + k^2 \|u''\|^2_{L^2(\Omega)}\, dt \right).$$

Defining $\rho^j := P_h u(t_j) - u(t_j)$, there follows

$$\|\rho^n\|^2_{L^2(\Omega)} + \sum_{j=0}^{n-1} \underbrace{\|\rho^{j+1} - \rho^j\|^2_{L^2(\Omega)}}_{\leq k \int_{t_j}^{t_{j+1}} \|u' - P_h u'\|^2_{L^2(\Omega)}} + k|\rho^{j+1}|^2_{H^1(\Omega)} \leq \|\rho^n\|^2_{L^2(\Omega)} + k \int_0^{t_n} \|u' - P_h u'\|^2_{L^2(\Omega)} + k \sum_{j=0}^{n-1} |u(t_j) - P_h u(t_j)|^2_{H^1(\Omega)}$$

and finally, with $u_h^{j+1} - u(t_{j+1}) = \psi_h^{j+1} + \rho^{j+1}$ and the triangle inequality, the claimed estimate. □

Corollary 2.22 *Let u solve (2.4) and assume $u \in C^3(\bar{\Omega} \times [0,T])$. Let u_h^n denote the fully discrete approximation employing $V_h = S_0^{1,1}(\mathcal{T})$ with $u_{0,h} = P_h u_0$ and the implicit Euler method in time. Then:*

(i) There exists a constant $C > 0$ independent of h, k such that

$$\|u_h^n - u(t_n)\|_{L^2(\Omega)} \leq C(h + k).$$

(ii) Additionally, if Ω is convex, then

$$\|u_h^n - u(t_n)\|_{L^2(\Omega)} \leq C(h^2 + k).$$

Proof: Theorem 2.21 gives

$$\|u_h^n - u(t_n)\|_{L^2(\Omega)} \leq \|u(t_n) - P_h u(t_n)\|_{L^2(\Omega)} + \int_0^{t_n} \|u'(t) - P_h u'(t)\|_{L^2(\Omega)} + k\, \|u''(t)\|_{L^2(\Omega)}\, dt.$$

Employing the approximation properties of P_h as in Corollary 2.15 shows that the first two terms on the right-hand side are of order $\mathcal{O}(h)$ or even of $\mathcal{O}(h^2)$ for convex domains Ω. □

2.3.3 The θ-scheme

As mentioned, the θ-scheme defined as

$$y^{i+1} = y^i + k f(t_i + \theta(t_{i+1} - t_i), y_i + \theta(y_{i+1} - y_i))$$

is a generalization of the implicit Euler method. Applied to the heat equation the most common cases $\theta = 0, 1$ and $\theta = 1/2$ lead to linear systems provided in Figure 2.1.

$$
\begin{aligned}
\theta = 0 \text{ (explicit Euler):} &\quad \mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + k\mathbf{A}\mathbf{u}^n = k\mathbf{F}(t_n) \\
\theta = 1 \text{ (implicit Euler):} &\quad \mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + k\mathbf{A}\mathbf{u}^{n+1} = k\mathbf{F}(t_{n+1}) \\
\theta = 1/2 \text{ (Crank-Nicolson):} &\quad \mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + \frac{k}{2}(\mathbf{A}\mathbf{u}^n + \mathbf{A}\mathbf{u}^{n+1}) = k\mathbf{F}(t_{n+\frac{1}{2}})
\end{aligned}
$$

Figure 2.1: Matrix form of θ-scheme for the heat equation.

Statements similar to Theorem 2.21 also holds for the θ-scheme, e.g., one can show the following stability estimate:

Exercise 2.23 Let u_h^n be given by the θ-scheme with $\theta \in (1/2, 1]$. Set $u_h^{n+\theta} := u_h^n + \theta(u_h^{n+1} - u_h^n)$ and $t_{j+\theta} := t_j + \theta(t_{j+1} - t_j)$. Then, there holds

$$
\|u_h^n\|^2_{L^2(\Omega)} + \sum_{j=0}^{n-1} k|u_h^{j+\theta}|^2_{H^1(\Omega)} + (2\theta - 1)\|u_h^{j+1} - u_h^j\|^2_{L^2(\Omega)} \leq \|u_h^0\|^2_{L^2(\Omega)} + C\sum_{j=0}^{n-1} k\|f(t_{j+\theta})\|^2_{L^2(\Omega)}.
$$

Hint: consider the test-function $v = \psi^{n+\theta} = u_h(t_{n+\theta}) - P_h u(t_{n+\theta})$.

For sufficient regular solutions, one obtains higher order convergence for the Crank-Nicolson method.

Theorem 2.24 (CN is second order) *Let u_h^n be defined by*

$$
\frac{1}{k}\langle u_h^{n+1} - u_h^n, v\rangle_{L^2} + a\left(\frac{u_h^{n+1} - u_h^n}{2}, v\right) = \langle f(t_n + k/2), v\rangle_{L^2} \qquad \forall v \in V_h.
$$

Then, there holds

$$
\begin{aligned}
\|u_h^n - u(t_n)\|_{L^2(\Omega)} \leq \|u_{0,h} - P_h u_0\|_{L^2(\Omega)} &+ \|u(t_n) - P_h u(t_n)\|_{L^2(\Omega)} \\
&+ \int_0^{t_n} \|u' - P_h u'\|_{L^2(\Omega)} + Ck^2\|u'''\|_{L^2(\Omega)}\, dt
\end{aligned}
$$

Proof: Analogous to Theorem 2.21, (2.19), use $v = (\psi_h^{n+1} + \psi_h^n)/2$ as test-function. □

2.3.4 Stability of θ-scheme—the CFL-condition

By definition of A-stability and L-stability, we see from the definitions of the stability functions R that the explicit Euler method is not A-stable, the Crank-Nicolson method is A-stable and the implicit Euler method is A-stable and L-stable. More general, the stability properties of the θ-scheme can be seen in Figure 2.2.

$0 \leq \theta < 1/2$	method *not* A-stable	order 1
$\theta = 1/2$	method A-stable	order 2
$1/2 < \theta \leq 1$	method A-stable (even L-stable)	order 1

Figure 2.2: Stability properties and convergence order of θ-scheme.

For $0 \leq \theta < 1/2$ the θ-scheme is not A-stable but only *conditionally stable*, i.e., there is a condition on the employed step sizes for the method to work, the so called *CFL condition*[4]. This is in practice very restrictive!

[4]Richard Courant, Kurt Friedrichs, Hans Lewy

In order to see this, we consider one step of the method, which can be written as

$$u_h^{n+1} = Ru_h^n + kF^n$$

with a linear operator $R : V_h \to V_h$. Therefore, we obtain (not specifying the norm on purpose)

$$\|u_h^{n+1}\| \le \|R\|\|u_h^n\| + k\|F^n\|.$$

The Gronwall lemma then gives

$$\|u_h^N\| \le \|R\|^N \|u_h^0\| + \cdots$$

With $t_N = Nk = T =$ final time, there holds

$$\|u_h^N\| \le \|R\|^{T/k} \|u_h^0\| + \cdots$$

This shows that there should hold $\|R\| \le 1 + Ck$, since then

$$\|R\|^{T/k} \le (1 + Ck)^{T/k} = (1 + Ck)^{TC/(Ck)} \le e^{CT}$$

holds *uniformly* in k (otherwise: $\|R\| \ge 1+\delta$ for a $\delta > 0$ independent of k gives $\|R\|^{T/k} \ge (1+\delta)^{T/k} \to \infty$ for $k \to 0$).

In order to keep the analysis of $\|R\|$ simple, we consider $\|R\|$ on the matrix level. The iteration is then written as $\mathbf{u}^{n+1} = \mathbf{R}\mathbf{u}^n + \cdots$, with the matrix $\mathbf{R}$ given in Fig. 2.3.
For any matrix norm, there holds $\|\mathbf{R}\| \le \rho(\mathbf{R})$ with the spectral radius $\rho(\mathbf{R}) = \max\{|\lambda| \,|\, \lambda \in \sigma(\mathbf{R})\}$. Therefore, the condition $\rho(\mathbf{R}) \le 1 + Ck$ is a reasonable condition for stability. In fact, the stronger condition

$$\rho(\mathbf{R}) \le 1 \tag{2.24}$$

is usually imposed, which gives a bound on $\|u_h^N\|$ independent of the end time T. Fig. 2.3 illustrates this condition for our fully discrete methods. Note that Theorem 2.18 gives $\lambda_{max} \sim C/h^2$. Thus, for the θ-scheme, the condition (2.24) can be written in terms of mesh size in space h and time step length k. For the explicit Euler method this gives the CFL condition

$$\frac{k}{h^2} = O(1) \qquad \text{i.e.} \qquad k \le Ch^2.$$

For the implicit Euler method and the Crank-Nicolson method no restrictions apply as seen in Fig. 2.3.

	$\mathbf{R}$	$\sigma(\mathbf{R})$	$\rho(\mathbf{R})$	stable?
$\mathbf{R}_{expl}$	$\mathbf{I} - k\mathbf{M}^{-1}\mathbf{A}$	$\{1 - k\lambda \mid \lambda \in \sigma\}$	$\lvert 1 - k\lambda_{max}\rvert$	if $k \le \frac{2}{\lambda_{max}}$
$\mathbf{R}_{impl}$	$(\mathbf{M} + k\mathbf{A})^{-1}\mathbf{M}$	$\left\{\frac{1}{1+k\lambda} \mid \lambda \in \sigma\right\}$	$\frac{1}{\lvert 1+k\lambda_{min}\rvert} \le 1$	for all $k > 0$
$\mathbf{R}_{CN}$	$\left(\mathbf{M} + \frac{k}{2}\mathbf{A}\right)^{-1}\left(\mathbf{M} - \frac{k}{2}\mathbf{A}\right)$	$\left\{\frac{1-k/2\lambda}{1+k/2\lambda} \mid \lambda \in \sigma\right\}$	≤ 1	for all $k > 0$

Figure 2.3: Analysis of $\mathbf{R}$ in $\mathbf{u}^{n+1} = \mathbf{R}\mathbf{u}^n + \cdots$. $\qquad \sigma = \{\lambda : \exists \mathbf{x} \ne 0 \text{ with } \mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}\}$

Exercise 2.25 For $\theta \in (0, 1/2)$, there also has to hold the CFL condition $k \le Ch^2$ for stability.

2.3.5 Numerical examples

Example 2.26 We consider the heat equation for $d = 1$ on $(0, 1) \times [0, 1]$ with data $u_0 = 1$ and $f \equiv 0$. For the time discretization, we consider the explicit Euler, implicit Euler and the Crank-Nicholson method. Figure 2.4 shows the computed numerical approximations and the exact solution. For the explicit Euler method we employ step sizes $k = \frac{2.001}{\lambda_{max}} \ge \frac{2}{\lambda_{max}}$ (which violates the stability bound) and $k = \frac{1.999}{\lambda_{max}} \le \frac{2}{\lambda_{max}}$ (which satisfies the stability bound).

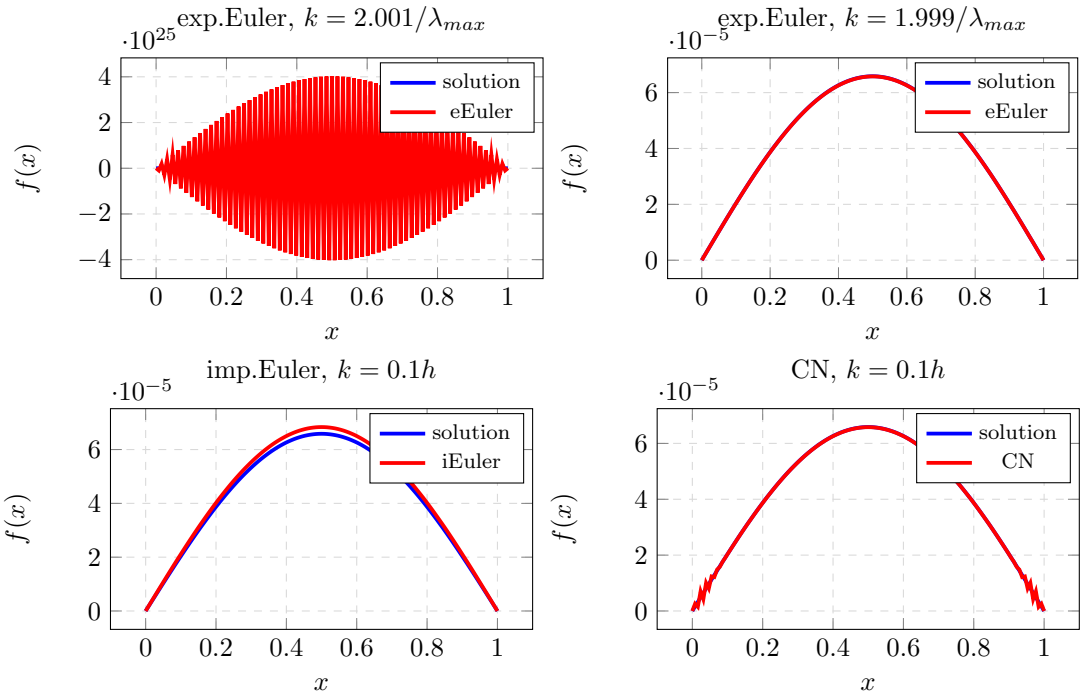

Figure 2.4: Comparison of exact and numerical solution for the 1D heat equation, $u_0 = 1, f = 0$.

In the case of even a small violation of the stability bound, the numerical solution oscillates very strongly and is completely inaccurate.

In the next example, we take the exact solution $u(x,t) = e^{-t}x(1-x)$ (and correspondingly obtain u_0, f by inserting the solution into the PDE), and study the convergence behaviour of the fully discrete methods in Figure 2.5. As expected, we obtain first order convergence $\mathcal{O}(h)$ (as we have $k = 0.1h$ and expect $\mathcal{O}(k + h^2)$) for the implicit Euler method (unconditionally) and for the explicit Euler method provided the stability bound is satisfied. The Crank-Nicolson method even gives second order convergence $\mathcal{O}(h^2)$ (as we have $k = 0.1h$ and expect $\mathcal{O}(k^2 + h^2)$).

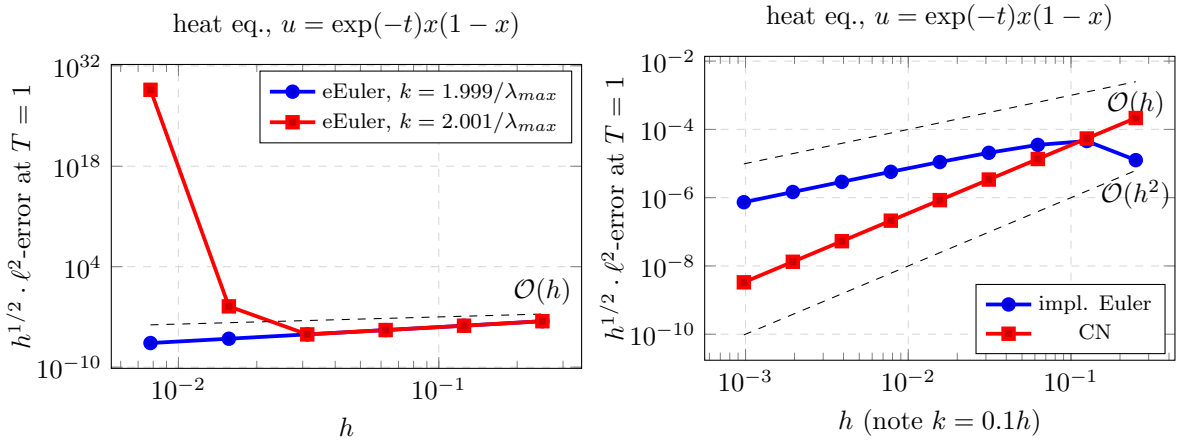

Figure 2.5: Convergence of the fully discrete schemes.

Example 2.27 We consider the 2D-heat equation on the unit square $\Omega = [0,1]^2$

$$u_t - \Delta u = 0 \quad \text{in } \Omega_T, \tag{2.25}$$

$$u = 0 \quad \text{on } \partial\Omega \times (0,T), \qquad u(x,0) = \sin(\pi x)\sin(\pi y). \tag{2.26}$$

Then, the exact solution is given by $u(x,y,t) = e^{-2\pi^2 t}\sin(\pi x)\sin(\pi y)$. We use explicit and implicit Euler for this problem implemented in ngsolve. As seen in Figure 2.6 there is energy dissipation as expected. However, for the explicit Euler, we again observe instability (Figure 2.7) if the time step is chosen too large.

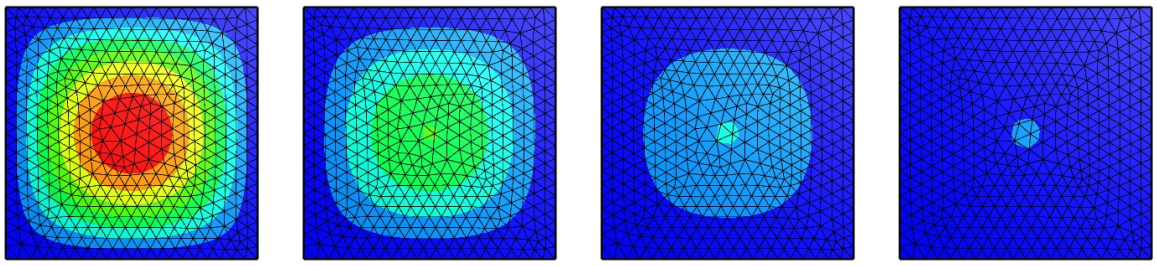

Figure 2.6: solution implicit Euler, first four time steps, $k = 0.05$, colour scale fixed between 0 (blue) and 1 (red).

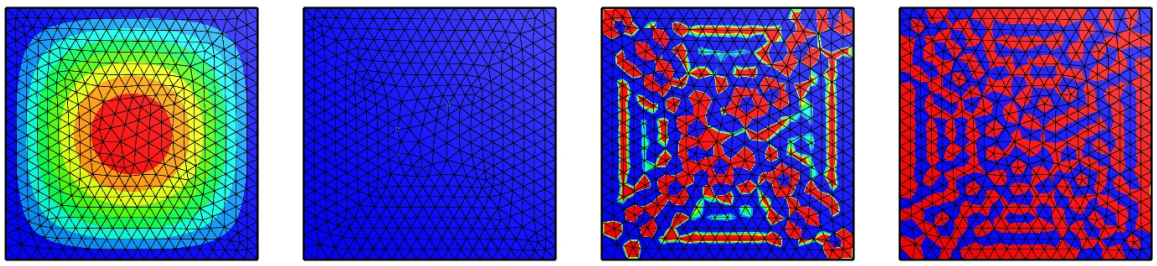

Figure 2.7: solution explicit Euler, first four time steps, $k = 0.05$, colour scale fixed between 0 (blue) and 1 (red).

Remark 2.28 The previous examples show that the explicit Euler method should be avoided! In order to keep the computational cost small, it is beneficial to make large time steps (in each time step you have to solve a linear system!). For the explicit Euler, this means that you have to know $\lambda_{\max}$ in advance and by no means should violate the stability bound. ∎

2.4 Weakening the notion of solution

Until now: classical solution in time, i.e., regularity $u \in C^1((0,T); H_0^1(\Omega)) \cap C^0([0,T], H_0^1(\Omega))$.

Observation:

- this requires $u_0 \in H_0^1(\Omega)$ ("compatible initial data").
- $u_0 \in H_0^1(\Omega)$ is not natural in applications as well as in the mathematical structure of the problem (compare the solution formula in (2.6)).

2.4.1 A short introduction to Bochner spaces

Let X be a Banach space and $(a,b) \subset \mathbb{R}$.
Goal: reasonable definition of $\int_a^b f(t)\,dt$ for a function $f : (a,b) \to X$.

Idea: work similarly to definition of Lebesgue integral by starting with step functions and then taking limits.

For elementary functions (note χ_A is the characteristic function for the set A), defined as $g : t \mapsto \sum_{i=1}^n f_i \chi_{A_i}(t)$ with $A_1, \dots, A_n \subset (a,b)$ (Lebesgue-)measurable and $g_1, \dots, g_n \in X$, the integral is canonically defined as

$$\int_a^b g(t)dt := \sum_{i=1}^n g_i \int_{A_i} dt = \sum_{i=1}^n g_i |A_i| \in X.$$

Then, for every function $f : (a,b) \to X$, which can be approximated pointwise almost everywhere by elementary functions (this property is called Bochner-measurable), i.e., there is a sequence $(f_n)_{n\in\mathbb{N}}$ of elementary functions with

$$\lim_{n\to\infty} f_n(t) = f(t) \qquad \text{a.e. as limit in X,} \tag{2.27}$$

the Bochner integral can be defined as limit of the integrals of the f_n.

Definition 2.29 (Bochner-Integral) *Let $f : (a,b) \to X$ and $(f_n)_n$ a sequence of elementary functions with (2.27). Assume additionally*

$$\lim_{n\to\infty} \int_a^b \|f(t) - f_n(t)\|_X \, dt = 0. \tag{2.28}$$

Then, f is called Bochner-integrable *and*

$$\int_a^b f(t)\,dt := \lim_{n\to\infty} \int_a^b f_n(t)\,dt$$

is called the Bochner integral of f.

Exercise 2.30 Show that: The Bochner integral is well defined, i.e., the limit is independent of the choice of sequence $(f_n)_n$, which has the pointwise limit f. ∎

An important characterization of Bochner integrability is given by the following theorem.

Theorem 2.31 *$f : (a,b) \to X$ is Bochner integrable $\iff$ f is Bochner-measurable and $t \mapsto \|f(t)\|_X$ is integrable. Then, there also holds*

$$\left\| \int_a^b f(t)\,dt \right\|_X \le \int_a^b \|f(t)\|_X \, dt. \tag{2.29}$$

Proof: Exercise. "$\Longrightarrow$": easy.
"$\Longleftarrow$": Define

$$\widetilde{f}_n(t) := \begin{cases} f_n(t) & \text{if } \|f_n(t)\|_X \le (1+\varepsilon)\|f(t)\|_X \\ 0 & \text{otherwise} \end{cases}$$

and employ the Lebesgue dominated convergence theorem. □

The previous theorem motivates that, analogous to the classical Lebesgue spaces L^p for $p \in [1,\infty)$, one can define L^p-spaces of Bochner integrable functions denoted by $L^p(a,b;X)$. The norm on these spaces is given by

$$\|f\|_{L^p(a,b;X)} := \left(\int_a^b \|f(t)\|_X^p \, dt \right)^{1/p}, \quad p \in [1,\infty), \qquad \|f\|_{L^\infty(a,b;X)} := \operatorname*{ess\,sup}_{t\in(a,b)} \|f(t)\|_X. \tag{2.30}$$

By definition, the space $L^1(a,b;X)$ coincides with the space of Bochner integrable functions.

Theorem 2.32 *The spaces $L^p(a,b;X)$ are Banach spaces. Is X a Hilbert space, then $L^2(a,b;X)$ is also a Hilbert space with scalar product*

$$(u,v)_{L^2(a,b;X)} = \int_a^b (u(t),v(t))_X \, dt.$$

In the following sections, it will be useful to interchange the Bochner integral with a linear functional on X.

Theorem 2.33 *Let $f \in L^1(a,b;X)$. Then,*

$$\langle g, \int_a^b f(t)\,dt\rangle_{X'\times X} = \int_a^b \langle g, f(t)\rangle_{X'\times X} dt \qquad \forall g \in X'. \tag{2.31}$$

Proof: We employ a density argument.
Step 1: Let f be a elementary function, then (2.31) holds by direct calculation.
Step 2: Let $g \in X'$. Then both the left-hand and right-hand side in (2.31) define a continuous linear functional on $L^1(a,b;X)$ (here, we do not show the measurability of the appearing functions):

$$\begin{aligned}
\left|\int_a^b \langle g, f(t)\rangle_{X'\times X}\, dt\right| &\leq \int_a^b |\langle g, f(t)\rangle_{X'\times X}|\, dt \leq \int_a^b \|g\|_{X'}\|f(t)\|_X\, dt = \|g\|_{X'}\|f\|_{L^1(a,b;X)} \\
\left|\langle g, \int_a^b f(t)\,dt\rangle_{X'\times X}\right| &\leq \|g\|_{X'}\left\|\int_a^b f(t)\,dt\right\|_X \leq \|g\|_{X'}\int_a^b \|f(t)\|_X\, dt = \|g\|_{X'}\|f\|_{L^1(a,b;X)}.
\end{aligned}$$

As the linear functionals are equal on a dense subset by step 1, they, by continuity, also are equal on $L^1(a,b;X)$. □

2.4.2 Weak derivatives in Bochner spaces

The space $H^1(a,b;X)$

We denote by $C_0^\infty((a,b);\mathbb{R})$ the space of infinitely continuously differentiable real-valued functions with compact support.

Definition 2.34 *Let $u \in L^1(a,b;X)$. Then, the* distributional derivative *of u is the linear mapping $C_0^\infty((a,b);\mathbb{R}) \to X$ given by*[5]

$$C_0^\infty((a,b);\mathbb{R}) \ni \varphi \mapsto -\int_a^b u(t)\varphi'(t)\,dt. \tag{2.32}$$

If this linear mapping has a representation by a function $v \in L^1_{loc}(a,b;X)$ (meaning L^1 on each compact subset of (a,b)), i.e.,

$$-\int_a^b u(t)\varphi'(t)\,dt = \int_a^b v(t)\varphi(t)\,dt \qquad \forall \varphi \in C_0^\infty((a,b);\mathbb{R})$$

we call v the weak derivative *of u.*

If X is separable, then the weak derivative is indeed unique. Now, we can define a Sobolev-type Bochner space.

Definition 2.35 *Let X be a separable Hilbert space. Then, we define the space $H^1(a,b;X) := \{u \in L^2(a,b;X) \,|\, u' \in L^2(a,b;X)\}$, which is a Hilbert space with scalar product*

$$(u,v) \mapsto \int_a^b \langle u(t), v(t)\rangle_X + \langle u'(t), v'(t)\rangle_X \, dt.$$

[5] the appearing integrals are Bochner-integrals!

The space $W(a,b;X;Y)$

Not every function $u \in L^2(a,b;X)$ has a distributional derivative, which can be represented by an X-valued Bochner integral. For our purpose it is relevant that the distributional derivative of u has a representation as a Bochner integral in a larger space $Y \supset X$ (i.e. in a weaker topology).
We assume

$$X,\ Y \text{ separable Hilbert spaces, } X \subset Y \text{ dense.} \tag{2.33}$$

Exercise 2.36 Let X, Y satisfy (2.33). Let $u \in L^1(a,b;X)$. Then, $u \in L^1(a,b;Y)$ and with the embedding $\iota_{X\to Y} : X \to Y$

$$\iota_{X\to Y} \int_a^b u(t)\,dt = \int_a^b \iota_{X\to Y} u(t)\,dt.$$

Here, the left integral is a Bochner integral in X and the right one in Y. For sake of readability we will not write the embedding in the following any more. ∎

Definition 2.37 *Let X, Y satisfy (2.33). Let $u \in L^1(a,b;X)$. A function $v \in L^1(a,b;Y)$ is called* weak derivative *of u, if*

$$-\int_a^b u(t)\varphi'(t)\,dt = \int_a^b v(t)\varphi(t)\,dt \qquad \forall \varphi \in C_0^\infty(a,b),$$

which is to be understood as equality in Y (also employing exercise 2.36). We also write u' for the weak derivative.

Definition 2.38 *Let X, Y satisfy (2.33). We then define $W(a,b;X;Y) = \{u \in L^2(a,b;X) \,|\, u' \in L^2(a,b;Y)\}$, which is a Hilbert space with scalar product*

$$(u,v) \mapsto \int_a^b \langle u(t), v(t)\rangle_X + \langle u'(t), v'(t)\rangle_Y\,dt.$$

For the spaces $H^1(a,b;X)$ and the generalization $W(a,b;X;Y)$ there hold many statements that also hold for classical Sobolev spaces, e.g., density of smooth (X valued) functions or, similarly to the 1D Sobolev embedding $H^1(0,1) \subset C([0,1])$, that

$$W(a,b;X;X') \subset C([a,b];Y) \tag{2.34}$$

with X' being the dual space of X.
Moreover, for $X \hookrightarrow Y \hookrightarrow X'$ (dense) with X, Y being separable Hilbert spaces and for $u, v \in W(a,b;X;X')$ integration by parts formulas of the form

$$\langle u(t), v(t)\rangle_Y - \langle u(s), v(s)\rangle_Y = \int_s^t \langle u'(\tau), v(\tau)\rangle_{X'\times X} + \langle u(\tau), v'(\tau)\rangle_{X\times X'}\,d\tau$$

hold.

2.4.3 Weak formulation for the heat equation

Recap: with the eigenpairs of the Dirichlet Laplacian $(\varphi_n, \lambda_n)_{n\in\mathbb{N}} \subset H_0^1(\Omega) \times \mathbb{R}$ (which are an ONB of $L^2(\Omega)$), we defined the evolution operator for the heat equation as

$$E(t)u_0 = \sum_{n\in\mathbb{N}} e^{-\lambda_n t}\langle u_0, \varphi_n\rangle_{L^2}\varphi_n.$$

The right-hand side is even well defined for $u_0 \in L^2(\Omega)$, thus we can extend the evolution operator to an operator mapping

$$E(t) : L^2(\Omega) \to H_0^1(\Omega).$$

Note that for all $v \in L^2(\Omega)$ with basis expansion $v = \sum_{n\in\mathbb{N}} v_n\varphi_n$, Sobolev norms can be expressed as

$$\begin{aligned}\|v\|^2_{L^2(\Omega)} &= \sum_{n\in\mathbb{N}} |v_n|^2,\\ |v|^2_{H^1(\Omega)} &= \sum_{n\in\mathbb{N}} \lambda_n |v_n|^2,\\ \|v\|_{H^{-1}(\Omega)} &:= \sup_{w\in H^1_0(\Omega)} \frac{\langle v,w\rangle_{L^2}}{|w|_{H^1(\Omega)}} = \sup_{(w_n)_{n\in\mathbb{N}}} \frac{\sum_{n\in\mathbb{N}} v_n w_n}{\sqrt{\sum_{n\in\mathbb{N}} \lambda_n |w_n|^2}} = \sqrt{\sum_{n\in\mathbb{N}} |v_n|^2 \lambda_n^{-1}}.\end{aligned}$$

The last equality follows from Cauchy-Schwarz for sums and the particular choice $w_n = \frac{1}{\lambda_n} v_n$.
This observation then gives the following mapping properties for the evolution operator.

Lemma 2.39 *Let $u_0 \in L^2(\Omega)$ and denote by $E(t)$ the extended evolution operator. Then, there holds*

$$t \mapsto E(t)u_0 \in L^2(0,T;H^1_0(\Omega)), \tag{2.35}$$

$$t \mapsto \partial_t(E(t)u_0) \in L^2(0,T;H^{-1}(\Omega)), \tag{2.36}$$

$$t \mapsto E(t)u_0 \in C([0,T];L^2(\Omega)). \tag{2.37}$$

The norms in (2.35)–(2.37) can be estimated by $C\|u_0\|_{L^2(\Omega)}$.

Proof: We only illustrate (2.36). As $\partial_t E(t)u_0 = \sum_n -\lambda_n e^{-\lambda_n t}\langle u_0,\varphi_n\rangle_{L^2}\varphi_n$ (the interchange between differentiation and summation can be argued with standard arguments), we have

$$\int_0^T \|\partial_t E(t)u_0\|^2_{H^{-1}(\Omega)}\,dt = \int_0^T \sum_n \lambda_n^2 e^{-2\lambda_n t} \frac{|\langle u_0,\varphi_n\rangle_{L^2}|^2}{\lambda_n}\,dt \le \frac{1}{2}\|u_0\|^2_{L^2(\Omega)},$$

where we used that summation and integration can be interchanged and $\int_0^T \lambda_n e^{-2\lambda_n t} \le \frac{1}{2}$. □

Motivation: Lemma 2.39 *suggests*:

- $u_0 \in L^2(\Omega)$ is sensible as initial data;
- one can expect $u \in L^2(0,T;H^1_0(\Omega))$ and $u' \in L^2(0,T;H^{-1}(\Omega))$ for a reasonable weak formulation of the heat equation;
- by (2.34), then there holds $u \in C([0,T];L^2(\Omega))$, i.e., initial values can be well defined.

In order to derive a suitable weak formulation, we again start at the variational formulation in space

$$\langle u'(t),v\rangle_{L^2(\Omega)} + a(u(t),v) = \langle f(t),v\rangle_{L^2(\Omega)} \qquad \forall v \in H^1_0(\Omega). \tag{2.38}$$

Rather than interpreting the derivative in a classical sense, we now aim for weak derivatives in time. More precisely, we set $V = H^1_0(\Omega)$ and $H = L^2(\Omega)$ and seek a solution $u \in L^2(0,T;V)$. The term including the time derivative in (2.38) is actually well defined even for $u'(t) \in V'$: Using a distributional interpretation of the derivative (see (2.32)) gives the formulation

$$-\int_0^T \langle u(t),v\rangle_{L^2(\Omega)}\varphi'(t)\,dt + \int_0^T a(u(t),v)\varphi(t)\,dt = \int_0^T \langle f(t),v\rangle_{L^2(\Omega)}\varphi(t)\,dt \qquad \forall v \in H^1_0(\Omega)\ \ \forall \varphi \in C^\infty_0(0,T).$$

If we require now $u \in L^2(0,T;V)$ with $u' \in L^2(0,T;V')$, we calculate using the definition of weak derivative[6]

$$\begin{aligned}-\int_0^T \langle u(t),v\rangle_{L^2(\Omega)}\varphi'(t)\,dt &= -\int_0^T \langle u(t),v\rangle_H \varphi'(t)\,dt = -\int_0^T \langle u(t),v\rangle_{V\times V'}\varphi'(t)\,dt\\ &\overset{\text{Thm. 2.33}}{=} -\langle v, \int_0^T u(t)\varphi'(t)\,dt\rangle_{V'\times V} = -\langle v, \int_0^T u(t)\varphi'(t)\,dt\rangle_{V\times V'} \overset{u'\in L^2(0,T;V')}{=} \langle v, \int_0^T u'(t)\varphi(t)\,dt\rangle_{V\times V'}.\end{aligned}$$

[6] note that more precisely, one actually would write $\langle u(t),v\rangle_{L^2(\Omega)} = \langle \iota u(t), \iota v\rangle_H$ with the embedding $\iota : V \to H$ and obtain the second equality as $\langle \iota u(t),v\rangle_H = \langle u(t),\iota' v\rangle_{V\times V'}$

Thus, we obtained the formulation: Find $u \in L^2(0,T;V)$ with $u' \in L^2(0,T;V')$, s.t.

$$\int_0^T \langle u'(t), v\rangle_{V'\times V}\varphi(t)\,dt + \int_0^T a(u(t),v)\varphi(t)\,dt = \int_0^T \langle f(t), v\rangle_{L^2(\Omega)}\varphi(t)\,dt \qquad \forall v \in H_0^1(\Omega) \qquad \forall \varphi \in C_0^\infty(0,T).$$

Variation over φ, we obtain that this can also be (equivalently) understood as pointwise almost everywhere formulation in time:

$$\langle u'(t), v\rangle_{V'\times V} + a(u(t),v) = \langle f(t), v\rangle_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega) \quad \text{a.e.}$$

Up until now, we tested with test functions vanishing at $t = 0$. Including the initial data, we observe that $u \in W(0,T;V;V')$ implies that $u \in C([0,T];H)$, i.e., $u_0 \in L^2(\Omega)$ is allowed. This leads to the standard weak formulation for the heat equation: Find $u \in L^2(0,T;V)$ with $u' \in L^2(0,T;V')$ such that

$$\begin{aligned} \langle u'(t), v\rangle_{V'\times V} + a(u(t),v) &= \langle f(t), v\rangle_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega) \quad \text{a.e.}, && \text{(2.39a)} \\ u(0) &= u_0 \in H. && \text{(2.39b)} \end{aligned}$$

Existence and uniqueness of solutions can e.g. be shown by employing a Galerkin method (see lecture "partial differential equations") or by employing inf-sup theory (see Section 2.7.1).

Having derived the weak formulation with less regularity requirements on the data and the solution, we can now proceed as in Section 2.2, replace the space $H_0^1(\Omega)$ by a FEM space and obtain a corresponding (weakly in time) ODE. The error and stability analysis of the method, however, needs to be done with different tools, which is the topic of the following section.

2.5 Numerical approximation for non-smooth initial data

Goal: convergence theory for $u_0 \in L^2(\Omega)$ (until now: only for compatible data $u_0 \in H_0^1(\Omega)$).

Again, we aim to split the error into

$$\|u(t_n) - u_h^n\| \le \|u(t_n) - u_h(t_n)\| + \|u_h(t_n) - u_h^k\|$$

and treat the semi-discrete error in Section 2.5.3 and the time discretization error in Section 2.5.4.

Difficulty with $\|u(t_n) - u_h(t_n)\|$: with $V = H_0^1(\Omega), H = L^2(\Omega)$, we can not expect that $u' \in L^2(0,T;V)$ (only $u' \in L^2(0,T;V')$ by solvability theory). The analysis of the semi-discretization error in Theorem 2.13 leads to

$$\|u(t) - u_h(t)\|_{L^2} \le e^{-\gamma t}\|u_0 - P_h u_0\|_{L^2} + \|u(t) - P_h u(t)\|_{L^2} + \int_0^t e^{-\gamma(t-s)}\|u'(s) - P_h u'(s)\|_{L^2}\,ds;$$

and therefore needed $u_0 \in V$ and $u' \in L^1(0,T;L^2(\Omega))$.

Simplification: in the following, we will only consider the case $f \equiv 0$, which already shows all the typical phenomena for the heat equation, similar statements also hold for $f \neq 0$ (assuming sufficient regularity).

2.5.1 Smoothing property

A typical property of parabolic equations is that solutions are very smooth for any $t > 0$, thus the described problems only appear at $t = 0$.
In order to capture this behaviour, we again look at the evolution operator $E(t)$ and observe that it satisfies mapping properties in certain t-weighted spaces.

Lemma 2.40 *Let $u_0 \in L^2(\Omega)$ and $E(t)$ be the evolution operator. Then,*

(i) The mapping $t \mapsto E(t)u_0$ is in $C^\infty((0,\infty);H_0^1(\Omega))$ and $C^\infty((0,\infty);L^2(\Omega))$.

(ii) For every $m \in \mathbb{N}_0$, there holds $\|\frac{d^m}{dt^m}E(t)u_0\|_{H^1(\Omega)} \le C_m t^{-1/2-m}\|u_0\|_{L^2(\Omega)}$.

(iii) For every $m \in \mathbb{N}_0$, there holds $\|\frac{d^m}{dt^m} E(t)u_0\|_{L^2(\Omega)} \le C_m t^{-m} \|u_0\|_{L^2(\Omega)}$.

(iv) $\int_0^t \|E(s)u_0\|^2_{H^1(\Omega)}\, ds \le C\|u_0\|^2_{L^2(\Omega)}$.

(v) $\int_0^t s^2 \|\frac{d}{dt} E(s)u_0\|^2_{H^1(\Omega)}\, ds \le C\|u_0\|^2_{L^2(\Omega)}$.

Proof: Exercise. Similar to Lemma 2.39. For (iv) one e.g. calculates

$$\begin{aligned}\int_0^t |E(t)u_0|^2_{H^1(\Omega)}\, ds &= \int_0^t \sum_n \lambda_n |\langle E(s)u_0, \varphi_n\rangle_{L^2}|^2\, ds = \int_0^t \sum_n \lambda_n e^{-2\lambda_n s} |\langle u_0, \varphi_n\rangle_{L^2}|^2\, ds \\ &\le C \sum_n |\langle u_0, \varphi_n\rangle_{L^2}|^2.\end{aligned}$$

For (ii) and (iii) one uses that $\sup_{x>0} xe^{-x} < \infty$. □

Remark 2.41 If $\partial\Omega \in C^\infty$, then the eigenfunctions of the Dirichlet Laplacian φ_n are smooth as well (on $\overline{\Omega}$), and one can show that $E(t)u_0 \in C^\infty((0,\infty); H^k(\Omega))$ for every $k \in \mathbb{N}$. However, the singularity at $t = 0$ still remains. ∎

2.5.2 Reduction to the analysis of the Ritz-projection error

As in the proof of Theorem 2.13, we split the error into two contributions

$$e_h(t) := u_h(t) - u(t) = \underbrace{u_h(t) - P_h u(t)}_{=\psi(t)} + \underbrace{P_h u(t) - u(t)}_{=\rho(t)} \tag{2.40}$$

and recall the *error equation* (where the derivatives are understood as elements in $L^2(0,T;V')$)

$$\langle \psi'(t), v\rangle_{L^2} + a(\psi(t), v) = -\langle \rho'(t), v\rangle_{L^2} \qquad \forall v \in V_h. \tag{2.41}$$

The key difference to the previously derived error analysis is that, in the present setting, $\psi(0)$ is not well defined. A key observation of Lemma 2.40 is that one can not expect u' to be in $L^2(0,T;H^1_0(\Omega))$ or $L^2(0,T;L^2(\Omega))$, but difficulties do only appear at $t = 0$. Thus, we can work with t-weighted spaces and refine the estimates from Theorem 2.13 in these norms.

Lemma 2.42 *For all $t > 0$, there hold the stability bounds*

$$\int_0^t \|\psi(s)\|^2_{L^2(\Omega)}\, ds \le Ct\|\Pi^{L^2} e_h(0)\|^2_{L^2(\Omega)} + C\int_0^t \|\rho(s)\|^2_{L^2(\Omega)}\, ds, \tag{2.42}$$

and

$$\begin{aligned}t\|\psi(t)\|^2_{L^2(\Omega)} + \int_0^t s|\psi(s)|^2_{H^1(\Omega)}\, ds &\le C\int_0^t s^2\|\rho'(s)\|^2_{L^2(\Omega)} + \|\rho(s)\|^2_{L^2(\Omega)}\, ds \\ &\quad + C\Big(\sup_{s\in(0,t)} s\|\rho(s)\|^2_{L^2(\Omega)} + t\|\Pi^{L^2} e_h(0)\|^2_{L^2(\Omega)}\Big).\end{aligned} \tag{2.43}$$

Proof: *ad (2.42):* The idea is to employ a so called parabolic duality argument (i.e. to cleverly choose a test-function). Let $t = t_0$ be fixed and consider the backwards equation

$$\begin{aligned}-z_t - \Delta z &= \psi \quad \text{in } \Omega \times (0,t_0), \\ z &= 0 \quad \text{on } \partial\Omega \times (0,t_0), \\ z(t_0) &= 0.\end{aligned}$$

Now, consider a semi-discrete approximation, i.e., $z_h \in C^1((0,t_0); V_h) \cap C^0([0,t_0]; V_h)$ satisfying

$$\begin{aligned}-\langle z_h'(s), v\rangle_{L^2} + a(z_h(s), v) &= \langle \psi(s), v\rangle_{L^2} \qquad \forall v \in V_h, \quad s \in (0,t_0) \\ z_h(t_0) &= 0.\end{aligned}$$

For this problem, we have the stability estimate (proof in exercises!)

$$\int_0^{t_0} \|z_h'(s)\|_{L^2(\Omega)}^2 \, ds + t_0^{-1} \|z_h(0)\|_{L^2(\Omega)}^2 \leq C \int_0^{t_0} \|\psi(s)\|_{L^2(\Omega)}^2 \, ds. \tag{2.44}$$

Now, for $s \in (0, t_0)$, we obtain

$$\begin{aligned}
\|\psi(s)\|_{L^2(\Omega)}^2 &= -\langle z_h'(s), \psi(s) \rangle_{L^2} + a(z_h(s), \psi(s)) \\
&= -\frac{d}{dt} \langle z_h(s), \psi(s) \rangle_{L^2} + \langle z_h(s), \psi'(s) \rangle_{L^2} + a(z_h(s), \psi(s)) \\
&= -\frac{d}{dt} \langle z_h(s), \psi(s) \rangle_{L^2} - \langle \rho'(s), z_h(s) \rangle_{L^2} \\
&= -\frac{d}{dt} \langle z_h(s), \psi(s) \rangle_{L^2} + \langle \rho(s), z_h'(s) \rangle_{L^2} - \frac{d}{dt} \langle z_h(s), \rho(s) \rangle_{L^2} \\
&= -\frac{d}{dt} \langle z_h(s), e_h(s) \rangle_{L^2} + \langle \rho(s), z_h'(s) \rangle_{L^2}.
\end{aligned}$$

Integration over (ε, t_0) for $0 < \varepsilon < t_0$ together with $z_h(t_0) = 0$ gives

$$\int_\varepsilon^{t_0} \|\psi(s)\|_{L^2(\Omega)}^2 \leq \underbrace{\langle z_h(\varepsilon), e_h(\varepsilon) \rangle_{L^2}}_{\to \langle z_h(0), e_h(0) \rangle_{L^2} \text{since } e_h \in C([0,T]; L^2)} + \underbrace{\int_\varepsilon^{t_0} \|\rho(s)\|_{L^2(\Omega)} \|z_h'(s)\|_{L^2(\Omega)} \, ds}_{\leq \sqrt{\int_0^{t_0} \|\rho\|_{L^2(\Omega)}^2} \sqrt{\int_0^{t_0} \|z_h'\|_{L^2(\Omega)}^2}}.$$

As $z_h(0) \in V_h$, we have $\langle z_h(0), e_h(0) \rangle_{L^2} = \langle z_h(0), \Pi^{L^2} e_h(0) \rangle_{L^2}$. Using (2.44) this leads to

$$\int_0^{t_0} \|\psi(s)\|_{L^2(\Omega)}^2 \, ds \leq C \Big(\sqrt{\int_0^{t_0} \|\rho(s)\|_{L^2(\Omega)}^2 \, ds} + \sqrt{t_0} \|\Pi^{L^2} e_h(0)\|_{L^2(\Omega)} \Big) \sqrt{\int_0^{t_0} \|\psi(s)\|_{L^2(\Omega)}^2 \, ds}.$$

ad (2.43): The choice $v = t\psi(t)$ in (2.41) gives

$$\frac{1}{2} \frac{d}{dt} (t \|\psi(t)\|_{L^2(\Omega)}^2) + t a(\psi(t), \psi(t)) = -t \langle \rho'(t), \psi(t) \rangle_{L^2} + \frac{1}{2} \|\psi(t)\|_{L^2(\Omega)}^2.$$

Integration over (ε, t) leads to

$$\frac{1}{2} t \|\psi(t)\|_{L^2(\Omega)}^2 + \int_\varepsilon^t s |\psi(s)|_{H^1(\Omega)}^2 \, ds = \frac{1}{2} \varepsilon \|\psi(\varepsilon)\|_{L^2(\Omega)}^2 - \underbrace{\int_\varepsilon^t \langle s\rho'(s), \psi(s) \rangle_{L^2} \, ds}_{\leq \sqrt{\int_0^t s^2 \|\rho'(s)\|_{L^2}^2 \, ds} \sqrt{\int_0^t \|\psi(s)\|_{L^2}^2 \, ds}} + \frac{1}{2} \int_\varepsilon^t \|\psi(s)\|_{L^2(\Omega)}^2 \, ds.$$

It remains to provide an estimate for $\limsup_{\varepsilon \to 0} \varepsilon \|\psi(\varepsilon)\|_{L^2(\Omega)}^2$. We write $\psi = e_h - \rho$ and obtain due to $e_h \in C([0,T]; L^2(\Omega))$

$$\begin{aligned}
\limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|\psi(\varepsilon)\|_{L^2(\Omega)} &\leq \limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|e_h(\varepsilon)\|_{L^2(\Omega)} + \limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|\rho(\varepsilon)\|_{L^2(\Omega)} \\
&= \limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|\rho(\varepsilon)\|_{L^2(\Omega)} \leq \sup_{s \in (0,t)} \sqrt{s} \|\rho(s)\|_{L^2(\Omega)}.
\end{aligned}$$

Together with Young's-inequality and (2.42), we finally arrive at

$$t \|\psi(t)\|_{L^2(\Omega)}^2 + \int_0^t s |\psi(s)|_{H^1(\Omega)}^2 \, ds \leq C \Big(\int_0^t s^2 \|\rho'(s)\|_{L^2(\Omega)}^2 + \|\rho\|_{L^2(\Omega)}^2 \, ds + \sup_{s \in (0,t)} s \|\rho(s)\|_{L^2(\Omega)}^2 + t \|\Pi^{L^2} e_h(0)\|_{L^2(\Omega)}^2 \Big).$$

□

2.5.3 Convergence of semi-discretization for incompatible data

In Corollary 2.15, we employed piecewise linear finite elements to obtain rates of convergence in terms of h for the semi-discretization. More generally, one only needs certain approximation properties of the Ritz-projection. We formulate this as a notation fixing a parameter r describing the rate.

Notation 2.43 *Let $r \in (0,1]$ be defined such that for all $v \in H_0^1(\Omega)$ and $P_h : H_0^1(\Omega) \to V_h$*

$$\|v - P_h v\|_{L^2(\Omega)} \leq Ch^r \|v - P_h v\|_{H^1(\Omega)} \leq Ch^r \|v\|_{H^1(\Omega)}.$$

For convex or smooth domains, by the Aubin-Nitsche duality argument, there holds $r = 1$ provided V_h is a standard FEM space. On non-convex polygons, the parameter r depends on the interior angles of the polygon (see elliptic shift theorems).
We formulate the case $r = 1$ together with additional approximation properties for $H^2(\Omega)$-functions as assumption:

Assumption 2.44 *There holds Notation 2.43 with $r = 1$ and for all $v \in H^2(\Omega) \cap H_0^1(\Omega)$*

$$\|v - P_h v\|_{L^2(\Omega)} \leq Ch^2 \|v\|_{H^2(\Omega)}.$$

For *compatible* initial data $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$, one obtains the optimal rate (as in Corollary 2.15).

Theorem 2.45 *Let Assumption 2.44 hold. Let $f \equiv 0$ and $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$ and u solve (2.39). Assume either $u_{h,0} = \Pi^{L^2} u_0$ or $u_{h,0} = P_h u_0$. Then,*

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch^2 \|u_0\|_{H^2(\Omega)}.$$

We now consider the case of *incompatible* initial data.

Theorem 2.46 *Let r be given as in Notation 2.43. Let $u_0 \in L^2(\Omega)$ and $f \equiv 0$. Let u_h be the semi-discrete approximation with initial data $u_{h,0} = \Pi^{L^2} u_0$ and u solve (2.39). Then,*

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch^r t^{-1/2} \|u_0\|_{L^2(\Omega)}. \tag{2.45}$$

If additionally Assumption 2.44 holds, then

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch^2 t^{-1} \|u_0\|_{L^2(\Omega)}. \tag{2.46}$$

Proof: *ad (2.45):* Due to $u_h(t) - u(t) = \psi(t) + \rho(t)$ and Lemma 2.42, we have to estimate $\|\rho(s)\|_{L^2(\Omega)}$ and $\|\rho'(s)\|_{L^2(\Omega)}$. The additional term $\Pi^{L^2} e_h(0)$ vanishes due to the choice $u_{h,0} = \Pi^{L^2} u_0$.
For $s > 0$, there holds $u(s) = E(s)u_0 \in H_0^1(\Omega)$ and by assumption that

$$\|\rho(s)\|_{L^2(\Omega)} \leq Ch^r \|u(s)\|_{H^1(\Omega)}, \qquad \|\rho'(s)\|_{L^2(\Omega)} \leq Ch^r \|u'(s)\|_{H^1(\Omega)}.$$

Using Lemma 2.40, we arrive at

$$t\|\psi(t)\|_{L^2(\Omega)}^2 \leq \int_0^t \|\rho(s)\|_{L^2(\Omega)}^2 \, ds + \int_0^t s^2 \|\rho'(s)\|_{L^2(\Omega)}^2 \, ds + \sup_{0<s<t} s\|\rho(s)\|_{L^2(\Omega)}^2 \leq Ch^{2r} \|u_0\|_{L^2(\Omega)}^2.$$

ad (2.46): The proof is rather technical and therefore not done here. The main idea is to define, for $v \in L^2(\Omega)$, an error operator $F_h(t)v := E_h(t)\Pi^{L^2} v - E(t)v$ using the discrete and continuous evolution operators.
Using the semi-group properties of the evolution operators the operator $F_h(t)$ can be written as compositions of operators $F_h(t/2)$ and $E(t/2)$. This allows to use the smoothing property of $E(t/2)$ and in turn Theorem 2.45 and a variant of Lemma 2.40 in $H^2(\Omega)$ to obtain the scaling in h and t. □

2.5.4 Time-discretization error for incompatible data

Here, we now aim to estimate the time stepping error $u_h(t_n) - u_h^n$ under reduced regularity assumptions (compared to e.g. Theorem 2.21 before).
We aim for estimates of the form

$$\|u_h(t_n) - u_h^n\|_{L^2(\Omega)} \leq Ck^p t_n^{-p} \|u_{h,0}\|_{L^2(\Omega)}, \qquad n = 1, 2, \ldots,$$

where p denotes the order of the time-stepping scheme. Note that similar estimates can also be derived for the case $f \neq 0$.

As in Section 2.3.1, we consider single step methods described by their stability functions $R(\cdot)$.

Error analysis

With the eigenvalues $\lambda_{h,n}$ and eigenfunctions $\varphi_{h,n}$, $n = 1, \ldots, N$ of the discretized Dirichlet Laplacian (posed on V_h), we define the discrete spectrum $\sigma_h := \{\lambda_{h,n} \,|, n = 1, \ldots, N\}$ and can write the evolution operator as

$$E_h(t)u_{h,0} = \sum_{m=1}^{N} e^{-\lambda_{h,m}t} \langle u_{h,0}, \varphi_{h,m} \rangle_{L^2} \varphi_{h,m}.$$

Setting $u_h^0 := u_{h,0}$, one step of the time-stepping method then, by assumptions on $R(\cdot)$, leads to

$$u_h^1 = \sum_{m=1}^{N} R(-\lambda_{h,m}k) \langle u_h^0, \varphi_{h,m} \rangle_{L^2} \varphi_{h,m}.$$

Correspondingly, we obtain u_h^n as

$$u_h^n = \sum_{m=1}^{N} R(-\lambda_{h,m}k)^n \langle u_h^0, \varphi_{h,m} \rangle_{L^2} \varphi_{h,m}.$$

It is convenient to introduce the function

$$F_n(z) := e^{-zn} - (R(-z))^n.$$

This allows to write the error at t_n as

$$\|u_h(t_n) - u_h^n\|_{L^2(\Omega)}^2 = \sum_{m=1}^{N} |F_n(k\lambda_{h,m})|^2 |\langle u_h^0, \varphi_{h,m} \rangle_{L^2}|^2 \leq \sup_{\lambda \in \sigma_h} |F_n(k\lambda)|^2 \|u_h^0\|_{L^2(\Omega)}^2.$$

Consequently, we have to control $F_n(k\lambda)$ *uniformly in* $\lambda_{h,m}$ and explicit in n. The consistency condition (2.16) $R(z) = e^z + \mathcal{O}(|z|^{p+1})$ gives good control for $F_n(k\lambda_{h,m})$ (fixed n) for $\lambda_{h,m}$ where $k\lambda_{h,m}$ is small. The remaining eigenvalues $\lambda_{h,m}$ are treated with additional stability assumptions for $R(\cdot)$.

As motivated in Section 2.3.1, it is reasonable to assume:

(I) R is defined on $(-\infty, 0]$ and satisfies $|R(z)| \leq 1$ for all $z \in (-\infty, 0]$.

(II) R is defined on $(-\infty, 0]$ and $\forall z_0 > 0 \quad \exists q(z_0) < 1$ with $|R(z)| \leq q(z_0) < 1$ for all $z < -z_0$.

Any A-stable Runge-Kutta method, such as the Crank-Nicolson method (or more general Gauß-methods), satisfies (I). The stronger assumption (II) is satisfied by L-stable Runge-Kutta methods such as the implicit Euler method (or more general Radau IIA-methods).

Lemma 2.47 *Assume that the consistency condition (2.16) holds.*

(i) Let $c \in (0, 1)$. Then, there is $z_0 > 0$ and $C > 0$, such that

$$|F_n(z)| \leq Cnz^{p+1} e^{-cnz} \qquad 0 < z < z_0.$$

(ii) Additionally assume the stability condition (I). Then,

$$|F_n(z)| \le Cz^p \qquad \forall z \in (0,\infty).$$

(iii) Additionally assume the stability condition (II). Then, for every $z_0 > 0$, there exist constants $c, C > 0$ such that

$$|F_n(z)| \le Ce^{-cn} \le Cz^p \qquad \forall z \ge z_0.$$

Proof: *ad (i):* The consistency condition (2.16) gives for sufficiently small $\lambda_0 > 0$

$$|e^{-\lambda} - R(-\lambda)| \le C\lambda^{p+1} \qquad 0 \le \lambda \le \lambda_0. \tag{2.47}$$

As $R(-z) = 1 - z + \mathcal{O}(z^2)$, for every $c \in (0,1)$, there exists a $z_0 > 0$ such that

$$|R(-z)| \le e^{-cz} \qquad 0 \le z \le z_0. \tag{2.48}$$

Together with (2.47) this gives

$$\begin{aligned} |F_n(z)| = |e^{-zn} - (R(-z))^n| &= \left|e^{-z} - R(-z)\right| \left|\sum_{j=0}^{n-1} (R(-z))^{n-1-j} e^{-jz}\right| \\ &\le Cz^{p+1} n e^{-c(n-1)z} \le Cz^p, \end{aligned} \tag{2.49}$$

where we used $\max_{j\in\{0,\dots,n-1\}} -jz - cz(n-1-j) = -cz(n-1)$ in the first inequality and $\sup_{x>0} xe^{-x} < \infty$ in the last inequality.
ad (ii): Let z_0 be given as in the statement of (i). As (2.49) is the sought estimate for $z \in (0, z_0)$, it is enough to consider $z \ge z_0$. The stability assumption $|R(\zeta)| \le 1$ for $\zeta \in (-\infty, 0]$ implies for $z \ge z_0$

$$|F_n(z)| \le |e^{-zn}| + |(R(-z))^n| \le 2 \le Cz_0^p \le Cz^p$$

with a suitable constant $C > 0$.
ad (iii): There holds

$$|F_n(z)| = |e^{-nz} - (R(-z))^n| \le e^{-nz} + |R(-z)|^n.$$

For $z \ge z_0$, there holds the assumption $|R(-z)| \le q < 1$, which gives

$$|F_n(z)| \le e^{-nz} + |R(-z)|^n \le e^{-nz_0} + q^n.$$

This finishes the proof. □

Theorem 2.48 *Assume the consistency condition (2.16) and the stability assumption (II). Then,*

$$\|u_h(t_n) - u_h^n\|_{L^2(\Omega)} \le Ck^p t_n^{-p} \|u_{h,0}\|_{L^2(\Omega)}.$$

Proof: We have to show that

$$\sup_{\lambda\in\sigma_h} |F_n(k\lambda)| \le Ck^p t_n^{-p} = Ck^p (kn)^{-p} = Cn^{-p}. \tag{2.50}$$

Let z_0 be given as in the statement of Lemma 2.47, (i). For $\lambda \in \sigma_h$ with $k\lambda \le z_0$, there holds by Lemma 2.47, (i)

$$|F_n(k\lambda)| \le Cn(k\lambda)^{p+1} e^{-cnk\lambda} = Cn^{-p}(t_n\lambda)^{p+1} e^{-c\lambda t_n} \le Cn^{-p},$$

since $\sup_{x>0} x^{p+1}e^{-x} < \infty$. For $\lambda \in \sigma_h$ with $k\lambda \ge z_0$, we employ Lemma 2.47, (iii) to estimate

$$|F_n(k\lambda)| \le Ce^{-cn} \le Cn^{-p}.$$

This shows (2.50). □

Theorem 2.48 assumes $|R(\infty)| < 1$, which does not hold for the Crank-Nicolson method. However, this can be fixed by doing two steps of the implicit Euler method at the beginning.

Exercise 2.49 Consider the method: Do *two* steps of the implicit Euler method and afterwards only apply the Crank-Nicolson time-stepping. Show that

$$\|u_h(t_n) - u_h^n\|_{L^2(\Omega)} \leq Ck^2 t_n^{-2} \|u_h^0\|_{L^2(\Omega)}.$$

Remark 2.50 In contrast to Section 2.3.1, the estimates in Theorem 2.48 and Exercise 2.49 are not uniform in t_n: the constant gets worse for $t_n \to 0$. In order to have good estimates up to $t = 0$, one has to employ non-uniform meshes in time (and possibly in space). ∎

2.6 Discontinuous Galerkin

Goal: up until now, we considered single-step methods for time discretization (i.e. finite difference techniques), here we aim for a *Galerkin*-discretization in time. While the formulation in itself is a bit more complicated, one obtains a method that inherits advantages of Galerkin methods. In particular, one obtains a space-time method with good stability properties that also allows for easy parallelization in time.

Setting: for simplicity, we employ the method for the semidiscrete equation. We assume that V_h is finite dimensional, with slight modifications the choice $V_h = H_0^1(\Omega)$ would also be possible.

Let $u_h \in C^1((0,T);V_h) \cap C^0([0,T];V_h)$ satisfy

$$\begin{aligned} \langle u_h'(t), v\rangle_{L^2} + a(u_h(t), v) &= \langle f(t), v\rangle_{L^2} \qquad \forall v \in V_h, \quad t \in (0,T), && \text{(2.51a)} \\ u_h(0) &= u_{h,0} \in V_h. && \text{(2.51b)} \end{aligned}$$

Let $\mathcal{T}_k$ be a mesh on $(0,T)$ given by the knots $0 = t_0 < t_1 < \ldots < t_N = T$ and elements $K_n = (t_n, t_{n+1})$ and define $k_n := t_{n+1} - t_n$. For piecewise continuous (w.r.t. the mesh $\mathcal{T}_k$) functions v with values in V_h, we define the *jump* at the knot t_n, $n = 1, \ldots, N-1$ by

$$[\![v]\!]_n := v(t_n+) - v(t_n-).$$

The numerical method seeks a function U in the space

$$X_{k,h} := S^{p,0}(\mathcal{T}; V_h) = \{u \,:\, u|_K \in \mathcal{P}_p(K; V_h)\} \qquad \text{with} \qquad \mathcal{P}_p(K; V_h) = \Big\{ \sum_{i=0}^{p} u_{K,i} t^i \,:\, u_{K,i} \in V_h \Big\},$$

i.e., the space of piecewise (discontinuous) polynomials of degree p (in t) with values in V_h.
In order to motivate the method, multiply (2.51) with a test function $w \in X_{k,h}$, integrate over $[0,T]$ and then use integration by parts on each element of $\mathcal{T}_k$, which gives

$$\sum_{K \in \mathcal{T}} - \int_K \langle u_h(t), w'(t)\rangle_{L^2}\, dt + \langle u_h, w\rangle_{L^2}|_{\partial K} + \int_K a(u_h(t), w(t))\, dt = \int_0^T \langle f(t), w(t)\rangle_{L^2}\, dt.$$

Now, replacing the exact solution u_h by the approximation $U \in X_{k,h}$, we obtain

$$\sum_{K \in \mathcal{T}} - \int_K \langle U(t), w'(t)\rangle_{L^2}\, dt + \langle U, w\rangle_{L^2}|_{\partial K} + \int_K a(U(t), w(t))\, dt = \int_0^T \langle f(t), w(t)\rangle_{L^2}\, dt \qquad \forall w \in X_{k,h}. \tag{2.52}$$

Since both the approximation U and the test function w are possibly discontinuous, equation (2.52) is no complete system of equations for U: equation (2.52) rather decomposes into independent equations on the subintervals $K \in \mathcal{T}_k$. In other words: there is a coupling between neighbouring elements missing (in each knot t_n there are two approximations $U(t_n-)$ and $U(t_n+)$ that do not influence each other at all). In order to couple neighbouring elements, we choose for each knot t_n a value $\widehat{U}(t_n)$ and set this to be the function value at t_n, i.e., we consider

$$\sum_{K \in \mathcal{T}} - \int_K \langle U(t), w'(t)\rangle_{L^2}\, dt + \langle \widehat{U}, w\rangle_{L^2}|_{\partial K} + \int_K a(U(t), w(t))\, dt = \int_0^T \langle f(t), w(t)\rangle_{L^2}\, dt \qquad \forall w \in X_{k,h}. \tag{2.53}$$

Integration by parts back gives

$$\sum_{K \in \mathcal{T}} \int_K \langle U'(t), w(t)\rangle_{L^2}\, dt + \langle -U + \widehat{U}, w\rangle_{L^2}|_{\partial K} + \int_K a(U(t), w(t))\, dt = \int_0^T \langle f(t), w(t)\rangle_{L^2}\, dt \qquad \forall w \in X_{k,h}. \tag{2.54}$$

We now choose

$$\widehat{U}(t_n) := U(t_n-).$$

This choice can be motivated that the heat equation is only well-posed *forward in time*. With this choice and additionally setting $\widehat{U}(0) := u_{h,0}$, we obtain

$$\int_0^T \langle U'(t), w(t)\rangle_{L^2} + a(U(t), w(t))\,dt + \langle U(0+), w(0+)\rangle_{L^2} + \sum_{n=1}^{N-1} \langle [\![U]\!]_n, w(t_n+)\rangle_{L^2}$$
$$= \int_0^T \langle f(t), w(t)\rangle_{L^2}\,dt + \langle u_{h,0}, w(0+)\rangle_{L^2} \qquad \forall w \in X_{k,h}.$$

This now gives the *dG(p)-method*:

$$\text{Find } U \in X_{k,h} \text{ s.t. for all } w \in X_{k,h} \text{ there holds } B(U,w) = l(w), \text{ with} \tag{2.55}$$
$$B(U,w) := \sum_{K\in\mathcal{T}} \int_K \langle U'(t), w(t)\rangle_{L^2} + a(U(t), w(t))\,dt + \sum_{n=1}^{N-1} \langle [\![U]\!]_n, w(t_n+)\rangle_{L^2} + \langle U(0+), w(0+)\rangle_{L^2},$$
$$l(w) := \int_0^T \langle f(t), w(t)\rangle_{L^2}\,dt + \langle u_{h,0}, w(0+)\rangle_{L^2}.$$

Unique solvability of (2.55) follows from coercivity of $B(\cdot,\cdot)$.

Lemma 2.51 *There holds*

$$B(U,U) \geq \int_0^T a(U(t), U(t))\,dt + \frac{1}{2}\|U(T-)\|^2_{L^2(\Omega)} \qquad \forall U \in X_{k,h}.$$

Proof: In order to remove the special treatment of the first element K_0, we extend $U \in X_{k,h}$ by 0 for $t < 0$ fort. Consequently, $[\![U]\!]_0 = U(0+)$, and we obtain

$$B(U,w) = \sum_{K\in\mathcal{T}} \int_K \langle U'(t), w(t)\rangle_{L^2} + a(U(t), w(t))\,dt + \sum_{n=0}^{N-1} \langle [\![U]\!]_n, w(t_n+)\rangle_{L^2}.$$

Elementary calculations show

$$\begin{aligned} 2\int_{t_n}^{t_{n+1}} \langle U', U\rangle_{L^2}\,dt + 2\langle [\![U]\!]_n, U(t_n+)\rangle_{L^2} &= \|U(t_{n+1}-)\|^2_{L^2(\Omega)} - \|U(t_n+)\|^2_{L^2(\Omega)} + 2\|U(t_n+)\|^2_{L^2(\Omega)} \\ &\quad - 2\langle U(t_n-), U(t_n+)\rangle_{L^2} \\ &\geq \|U(t_{n+1}-)\|^2_{L^2(\Omega)} - \|U(t_n+)\|^2_{L^2(\Omega)} + \|U(t_n+)\|^2_{L^2(\Omega)} - \|U(t_n-)\|^2_{L^2(\Omega)} \\ &= \|U(t_{n+1}-)\|^2_{L^2(\Omega)} - \|U(t_n-)\|^2_{L^2(\Omega)}. \end{aligned} \tag{2.56}$$

Since $U(0-) = 0$, this directly implies

$$B(U,U) \geq \int_0^T a(U(t), U(t))\,dt + \frac{1}{2}\|U(T-)\|^2_{L^2(\Omega)},$$

which is the claimed coercivity. □

Exercise 2.52 (i) The dG(0)-method (i.e. $p = 0$) is a (modified) implicit Euler method of the form

$$\frac{1}{k_n}\langle U^{n+1} - U^n, v\rangle_{L^2} + a(U^{n+1}, v) = \frac{1}{k_n}\int_{t_n}^{t_{n+1}} \langle f(t), v\rangle_{L^2}\,dt \qquad \forall v \in V_h, \qquad U^0 := u_{h,0},$$

where we employed the notation $U^n = U|_{(t_{n-1},t_n)}$.

(ii) The dG-method (2.55) is a time stepping method, i.e., one can compute $U|_{K_0}$, $U|_{K_1}, \ldots$, in succession. In order to see that, introduce the jump at $t_0 = 0$ by

$$[\![U]\!]_0 := U(t_0+) - u_{h,0}. \tag{2.57}$$

Show that the dG-method can be written as: Given $U_{K_{n-1}}$, find $U|_{K_n} \in \mathcal{P}_p(K_n; V_h)$ such that

$$\int_{t_n}^{t_{n+1}} \langle U'(t), w(t)\rangle_{L^2} + a(U(t), w(t))\,dt + \langle [\![U]\!]_n, w(t_n+)\rangle_{L^2} = \int_{t_n}^{t_{n+1}} \langle f(t), w(t)\rangle_{L^2}\,dt \qquad \forall w \in \mathcal{P}_p(K_n; V_h). \tag{2.58}$$

■

In order to analyze the error $U - u_h$, we employ a suitable interpolation operator.

Lemma 2.53 *Let $u \in C([0,T]; V_h)$. Let the interpolant $Iu \in X_{k,h}$ be elementwise defined on $K_n = (t_n, t_{n+1})$ by*

$$u(t_{n+1}) = u(t_{n+1}-) = (Iu)(t_{n+1}-), \tag{2.59}$$

$$\int_{t_n}^{t_{n+1}} t^\ell (u(t) - (Iu)(t))\,dt = 0, \qquad \ell = 0, \dots, p-1. \tag{2.60}$$

Let $\|\cdot\|_*$ be a norm on V_h. Then, for all $u \in C^{p+1}([0,T]; V_h)$, there holds

$$\max_{t\in[t_n,t_{n+1}]} \|u(t) - (Iu)(t)\|_*^2 \le C k_n^{2p+1} \int_{t_n}^{t_{n+1}} \|u^{(p+1)}(s)\|_*^2\,ds.$$

Proof: The interpolation operator is well-defined: choosing a basis $(e_i)_{i=1}^N$ of V_h, one can write $u \in C([0,T]; V_h)$ in the form $u(t) = \sum_{i=1}^N u_i(t) e_i$. In the same way, one can write $(Iu)(t) = \sum_{i=1}^N \widetilde{u}_i(t) e_i$ with polynomials $\widetilde{u}_i$, thus, one obtains systems of equations for $\widetilde{u}_i$, which are uniquely solvable.

The error estimate follows from a scaling argument. On the reference element $(0,1)$ a corresponding, scaled operator $\widehat{I}$ satisfies

$$\|\widehat{I}\widehat{u}\|_{C([0,1];(V_h,\|\cdot\|_*))} \le \Lambda \|\widehat{u}\|_{C([0,1];(V_h,\|\cdot\|_*))}$$

with a constant $\Lambda > 0$ independent of $\widehat{u}$. Moreover, it still reproduces polynomials of degree p, i.e,

$$\widehat{I}\pi = \pi \qquad \forall \pi \in \mathcal{P}_p([0,1]; V_h).$$

Let $T_p\widehat{u}$ be the Taylor polynomial of degree p of $\widehat{u}$ around an arbitrary point in $[0,1]$. Then,

$$\|\widehat{u} - \widehat{I}\widehat{u}\|_C = \|\widehat{u} - T_p\widehat{u} - \widehat{I}(\widehat{u} - T_p\widehat{u})\|_C \le (1+\Lambda)\|\widehat{u} - T_p\widehat{u}\|_C \le C\sqrt{\int_0^1 \|\widehat{u}^{(p+1)}(t)\|_*^2\,dt},$$

where the last step used Taylor expansion with remainder in integral form together with Cauchy-Schwarz. Now, scaling of the inequality, using $k_n = t_{n+1} - t_n$, gives

$$\|u - Iu\|_{C([t_n,t_{n+1}];(V_h,\|\cdot\|_*))} \le C k_n^{p+1/2} \sqrt{\int_{K_n} \|u^{(p+1)}(t)\|_*^2\,dt},$$

which shows the lemma. □

We now analyze the dG(p)-method. For that, we use the interpretation of the method as time-stepping scheme to derive a recurrence for the error. We start with the error equation

$$B(u_h - U, w) = 0 \qquad \forall w \in X_{k,h},$$

which follows by construction of the approximation (consistency of the method). As in (2.58), we obtain the elementwise equation

$$\int_{t_n}^{t_{n+1}} \langle (U-u_h)', w\rangle_{L^2} + a(U-u_h, w)\,dt + \langle [\![U-u_h]\!]_n, w(t_n+)\rangle_{L^2} = 0 \qquad \forall w \in \mathcal{P}_p(K_n; V_h). \tag{2.61}$$

Here, we used that the exact solution u_h is continuous in the knots t_n, $n = 1, \dots, N-1$ such that $[\![u_h]\!]_n = 0$ for all n. Moreover, for the special case $n = 0$, we have extended the exact solution u_h by

a constant $u_{h,0}$ for $t<0$ (in the same way as for U such that the jump is well defined for $n=0$). We decompose the error $U-u_h$ into

$$U-u_h=(U-Iu_h)+(Iu_h-u_h)=:\psi+\rho.$$

In order to treat the jump at $t_0=0$ correctly, we also extend Iu_h by $u_{h,0}$ for $t<0$. [7] This now gives

$$\psi(t_0-)=0. \tag{2.62}$$

By construction of I, we have

$$\int_{t_n}^{t_{n+1}} t^\ell \rho(t)\,dt=0, \qquad \ell=0,\dots,p-1, \qquad \rho(t_{n+1}-)=0.$$

Together with the (elementwise) error equation (2.61) this implies

$$\begin{aligned}
&\int_{t_n}^{t_{n+1}} \langle\psi',w\rangle_{L^2}+a(\psi,w)\,dt+\langle[\![\psi]\!]_n,w(t_n+)\rangle_{L^2}=-\int_{t_n}^{t_{n+1}}\langle\rho',w\rangle_{L^2}+a(\rho,w)\,dt-\langle[\![\rho]\!]_n,w(t_n+)\rangle_{L^2}\\
&=\int_{t_n}^{t_{n+1}}\langle\rho,w'\rangle_{L^2}-a(\rho,w)\,dt-\langle\rho(t_{n+1}-),w(t_{n+1}-)\rangle_{L^2}+\langle\rho(t_n+),w(t_n+)\rangle_{L^2}-\langle[\![\rho]\!]_n,w(t_n+)\rangle_{L^2}\\
&=-\int_{t_n}^{t_{n+1}} a(\rho,w)\,dt.
\end{aligned} \tag{2.63}$$

We now derive a recursion for ψ. The elementary calculations in (2.56) hold in the same way for ψ as well, i.e.,

$$2\int_{t_n}^{t_{n+1}}\langle\psi',\psi\rangle_{L^2}\,dt+2\langle[\![\psi]\!]_n,\psi(t_n+)\rangle_{L^2}\geq\|\psi(t_{n+1}-)\|^2_{L^2(\Omega)}-\|\psi(t_n-)\|^2_{L^2(\Omega)}.$$

Taking the test function $w=\psi$ in (2.63) and employing the Cauchy-Schwarz inequality this leads to

$$\begin{aligned}
\|\psi(t_{n+1}-)\|^2_{L^2(\Omega)}+2\int_{t_n}^{t_{n+1}}a(\psi,\psi)\,dt&\leq\|\psi(t_n-)\|^2_{L^2(\Omega)}-2\int_{t_n}^{t_{n+1}}a(\rho,\psi)\,dt\\
&\leq\|\psi(t_n-)\|^2_{L^2(\Omega)}+\int_{t_n}^{t_{n+1}}|\rho(t)|^2_{H^1(\Omega)}\,dt+\int_{t_n}^{t_{n+1}}|\psi(t)|^2_{H^1(\Omega)}\,dt.
\end{aligned}$$

Subtracting and afterwards dropping the positive term $\int_{t_n}^{t_{n+1}}a(\psi,\psi)=\int_{t_n}^{t_{n+1}}|\psi(t)|^2_{H^1(\Omega)}$, the recursion can be solved employing the Gronwall lemma, which gives

$$\|\psi(t_{n+1}-)\|^2_{L^2(\Omega)}\leq\|\psi(0-)\|^2_{L^2(\Omega)}+\int_0^{t_{n+1}}|\rho(t)|^2_{H^1(\Omega)}\,dt.$$

Since Lemma 2.53 provides the estimate $\|\rho\|^2_{C^0([t_n,t_{n+1}];(V_h;|\cdot|_{H^1(\Omega)}))}\leq Ck_n^{2p+1}\int_{K_n}|u_h^{(p+1)}(t)|^2_{H^1(\Omega)}\,dt$, and we have set $\psi(0-)=0$, we have actually show the following theorem.

Theorem 2.54 *For $n=0,1,\dots,$ there holds*

$$\|U(t_{n+1}-)-u_h(t_{n+1})\|^2_{L^2(\Omega)}\leq\int_0^{t_{n+1}}|\rho(t)|^2_{H^1(\Omega)}\,dt\leq C\sum_{m=0}^{n}k_m^{2p+2}\int_{t_m}^{t_{m+1}}|u_h^{(p+1)}(t)|^2_{H^1(\Omega)}\,dt.$$

Remark 2.55 Theorem 2.54 gives convergence of approximations at the knots t_n. In fact, from these estimates one can also obtain estimates for $\sup_{t\in(t_n,t_{n+1})}\|U(t)-u_h(t)\|_{L^2}$ for all n, [15, Thm. 12.2]. An indication that this is possible comes from Exercise 2.56, as there time derivatives of U can be controlled. ∎

Exercise 2.56 Show that for all n there holds

$$\int_{t_n}^{t_{n+1}}(t-t_n)\|U'(t)\|^2_{L^2}\,dt+(t_{n+1}-t_n)|U(t_{n+1})|^2_{H^1}\leq C\int_{t_n}^{t_{n+1}}\|f(t)\|^2_{L^2}\,dt+|U(t)|^2_{H^1}\,dt.$$

Hint: consider the test function $\widetilde{U}\in X_{k,h}$ defined by $\widetilde{U}|_{(t_n,t_{n+1})}(t):=(t-t_n)U'(t)$.

[7] All these extensions only have the purpose to correctly identify the jump $[\![\cdot]\!]_0$ and thus avoid case distinction of $n=0$ and $n>0$.

2.7 Space-time formulations

Goal: formulation that treats time just as an additional variable.

2.7.1 Variational formulation and solvability

The framework of the Section 2.4.3 also allows for a space-time formulation. For simplicity, we consider $u_0 = 0$ in the following. Recall that V, H are Hilbert spaces with $V \hookrightarrow H \hookrightarrow V'$ (dense).
We start with the derived distributional formulation: Find $u \in L^2(0,T;V)$ with $u' \in L^2(0,T;V')$ such that

$$-\int_0^T \langle u(t), v\rangle_H \varphi'(t)\,dt + \int_0^T a(u(t),v)\varphi(t)\,dt = \int_0^T \langle f(t), v\rangle_H \varphi(t)\,dt \quad \forall v \in V \quad \forall \varphi \in C_0^\infty(0,T), \tag{2.64a}$$

$$u(0) = 0 \in H. \tag{2.64b}$$

We define the spaces

$$X := \{u \in L^2(0,T;V) \,|\, u' \in L^2(0,T;V') \quad \text{and } u(0) = 0\}, \qquad Y := L^2(0,T;V). \tag{2.65}$$

By the following theorem, there holds that (2.64) is equivalent to the problem: Find $u \in X$ such that

$$\int_0^T \langle u'(t), v(t)\rangle_{V'\times V} + a(u(t),v(t))\,dt = \int_0^T \langle f(t), v(t)\rangle_{V'\times V}\,dt \quad \forall v \in Y; \tag{2.66}$$

here, we already considered the more general case of $f \in L^2(0,T;V')$.

Theorem 2.57 *Formulations (2.39) and (2.64) are equivalent. For $u_0 = 0$, these formulations are equivalent to (2.66).*

Note that formulation (2.66) has a different test and trial space, and thus leads to a so called Petrov-Galerkin method. For solvability, we employ inf-sup theory.

Theorem 2.58 *The bilinear form $B : X \times Y \to \mathbb{R}$ given by*

$$(u,v) \mapsto B(u,v) := \int_0^T \langle u'(t), v(t)\rangle_{V'\times V} + a(u(t),v(t))\,dt$$

satisfies

$$\inf_{u\in X} \sup_{v\in Y} \frac{B(u,v)}{\|u\|_X \|v\|_Y} \geq \gamma > 0 \tag{2.67}$$

$$\forall 0 \neq v \in Y \quad \exists u \in X: \quad B(u,v) \neq 0. \tag{2.68}$$

Proof: *ad (2.67):* Let $\mathbf{A} : V \to V'$ be the Gram-operator to the bilinear form $a(\cdot,\cdot)$, i.e.,

$$\langle \mathbf{A}u, v\rangle_{V'\times V} = a(u,v) \qquad \forall u, v \in V.$$

Due to Lax-Milgram, $\mathbf{A}$ is continuously invertible. In particular,

$$\|z\|_{V'} \leq C\|\mathbf{A}^{-1}z\|_V \qquad \forall z \in V'.$$

Moreover, coercivity of $a(\cdot,\cdot)$ implies positive definiteness of $\mathbf{A}^{-1}$, since

$$\langle \mathbf{A}^{-1}z, z\rangle_{V\times V'} = \langle \mathbf{A}^{-1}z, \mathbf{A}\mathbf{A}^{-1}z\rangle_{V\times V'} = a(\mathbf{A}^{-1}z, \mathbf{A}^{-1}z) \geq \alpha\|\mathbf{A}^{-1}z\|_V^2 \geq \alpha'\|z\|_{V'}^2 \qquad \forall z \in V'.$$

For given $u \in X$, we make the ansatz $v(t) = \mathbf{A}^{-1}u'(t) + u(t)$Inserting this in the bilinear form gives

$$B(u,v) \quad = \quad \int_0^T \langle u'(t), \mathbf{A}^{-1}u'(t) + u(t)\rangle_{V'\times V} + a(u(t), \mathbf{A}^{-1}u'(t) + u(t))\,dt.$$

Integration by parts gives

$$\int_0^T \langle u'(t), u(t)\rangle_{V'\times V}\,dt = \frac{1}{2}\int_0^T \frac{d}{dt}\|u(t)\|_H^2\,dt = \frac{1}{2}\|u(T)\|_H^2.$$

Consequently, as $a(u(t), \mathbf{A}^{-1}u'(t)) = \langle u'(t), u(t)\rangle_{V'\times V}$,

$$B(u,v) \;=\; \underbrace{\int_0^T \langle u'(t), \mathbf{A}^{-1}u'(t)\rangle_{V'\times V}\,dt}_{\geq C\|u'\|^2_{L^2(0,T;V')}} + \underbrace{\frac{1}{2}\|u(T)\|^2_{L^2(\Omega)}}_{\geq 0} + \underbrace{\int_0^T a(u(t),u(t))\,dt}_{\geq C\|u\|^2_{L^2(0,T;V)}} + \underbrace{\int_0^T a(u(t),\mathbf{A}^{-1}u'(t))\,dt}_{=\frac{1}{2}\|u(T)\|^2_H \geq 0}.$$

Thus, we have obtained

$$\begin{aligned} B(u,v) &\geq C\left[\|u\|^2_{L^2(0,T;V)} + \|u'\|^2_{L^2(0,T;V')}\right] = C\|u\|_X^2, \\ \|v\|_Y &\leq \|u\|_{L^2(0,T;V)} + \|\mathbf{A}^{-1}\|\|u'\|_{L^2(0,T;V')} \leq (1+\|\mathbf{A}^{-1}\|)\|u\|_X. \end{aligned}$$

ad (2.68): Let $v \in Y$ with

$$B(u,v) = 0 \qquad \forall u \in X. \tag{2.69}$$

We have to show that $v = 0$.
Step 1: we claim that the distributional derivative of $v \in L^2(0,T;V)$ is in $L^2(0,T;V')$. We explicitly give v' by defining the function

$$z(t) := \mathbf{A}v(t) \qquad t \in (0,T).$$

Then, $z \in L^2(0,T;V')$, due to $v \in L^2(0,T;V)$. By the following calculation, we indeed obtain that $z = v'$ as for arbitrary $w \in V$ and $\varphi \in C_0^\infty(0,T)$, we have

$$\begin{aligned} \langle w, \int_0^T z(t)\varphi(t)\,dt\rangle_{V\times V'} &\overset{\text{Thm. 2.33}}{=} \int_0^T \langle w, z(t)\rangle_{V\times V'}\varphi(t)\,dt = \int_0^T \langle w, \mathbf{A}v(t)\rangle_{V\times V'}\varphi(t)\,dt \\ &= \int_0^T a(v(t), w\varphi(t))\,dt \overset{B(\cdot,v)=0}{=} -\int_0^T \langle w\varphi'(t), v(t)\rangle_{V'\times V}\,dt \\ &= -\langle w, \int_0^T v(t)\varphi'(t)\,dt\rangle_{V\times V'} \end{aligned}$$

and z is the weak derivative of v.
Step 2: since $v \in L^2(0,T;V)$ and $v' \in L^2(0,T;V')$ integration by parts is possible and, for arbitrary $u \in X$, we obtain

$$\begin{aligned} \langle u(T), v(T)\rangle_H - \langle u(0), v(0)\rangle_H &= \int_0^T \langle u'(t), v(t)\rangle_{V'\times V} + \langle v'(t), u(t)\rangle_{V'\times V}\,dt \\ &\overset{B(\cdot,v)=0}{=} \int_0^T -a(u(t),v(t)) + \langle v'(t), u(t)\rangle_{V'\times V}\,dt. \end{aligned} \tag{2.70}$$

Taking a function u of the form $u(t) = w\varphi(t)$ with $w \in V$ and $\varphi \in C_0^\infty(0,T)$ in (2.70), there follows

$$\int_0^T \Big(-\langle v'(t), w\rangle_{V'\times V} + a(v(t),w)\Big)\varphi(t)\,dt = 0 \qquad \forall \varphi \in C_0^\infty(0,T).$$

Variation over $\varphi \in C_0^\infty(0,T)$ gives

$$-\langle v'(t), w\rangle_{V'\times V} + a(v(t),w) = 0 \quad \text{ a.e.,} \tag{2.71}$$

or in other words

$$v'(t) = \mathbf{A}v(t) \quad \text{ a.e. .} \tag{2.72}$$

Choosing $u(t) = tw$ with $w \in V$ arbitrary in (2.70) leads to

$$T\langle w, v(T)\rangle_H = \int_0^T t\underbrace{\left(-a(w,v(t)) + \langle v'(t), w\rangle_{V'\times V}\right)}_{\overset{(2.71)}{=}0}\,dt = 0 \qquad \forall w \in V. \tag{2.73}$$

Consequently, we have in total

$$v'(t) = \mathbf{A}v(t) \quad \text{a.e.} \qquad \text{and} \qquad v(T) = 0 \quad \text{in } L^2(\Omega). \tag{2.74}$$

Step 3: (2.74) implies $v \equiv 0$: by (2.72), we obtain

$$0 = \int_0^T -a(u(t), v(t)) + \langle v'(t), u(t)\rangle_{V'\times V}\, dt \qquad \forall u \in L^2(0,T;V).$$

Choosing $u = v$ and exploiting $v(T) = 0$ together with integration by parts, there holds

$$-\frac{1}{2}\|v(0)\|_H^2 = \int_0^T a(v(t), v(t))\, dt \geq 0,$$

which now implies $v = 0$. □

Corollary 2.59 *The variational formulation (2.66) is uniquely solvable for all $f \in L^2(0,T;V')$ and there holds*

$$\|u\|_X \leq \frac{1}{\gamma}\|f\|_{L^2(0,T;V')}.$$

Remark 2.60 The case of inhomogeneous initial conditions is possible as well. Hereby, one chooses the test-space

$$\widetilde{Y} := Y \times H$$

and the bilinear form $\widetilde{B}$ and right-hand side $\widetilde{l}$

$$\widetilde{B}(u,(v,w)) := B(u,v) + (u(0),w)_H, \qquad \widetilde{l}((v,w)) := l(v) + (u_0,w)_H.$$

Now, going to the discrete setting, inf-sup stable problems come with the additional difficulty that the inf-sup condition has to be verified for the choice of discrete spaces as well (compare FEM for Stokes!). This is in contrast to coercive problems, where coercivity is transferred to any closed subspace.
For the heat equation this is indeed a major drawback and in literature discrete inf-sup stability is usually only shown in weaker, mesh-dependent norms and not the natural norms of the problem.

2.7.2 A space-time least squares formulation

Problem: uniform inf-sup stability for the discretization of space-time methods can not be expected!

Goal: different kind of space-time method that is uniformly inf-sup stable for *any* choice of discrete subspace.

Idea: reformulate the PDE as minimization of a quadratic functional ("least squares functional"). Such methods, called *least squares finite element methods*, can also be applied to other model problems, we refer to [3] for an overview. The presented approach in this subsection is based on [5].

We again consider the heat equation

$$u_t - \Delta u = f \qquad \text{in } \Omega_T := \Omega \times (0,T), \tag{2.75a}$$
$$u(x,t) = 0 \qquad \text{on } \partial\Omega \times (0,T) \tag{2.75b}$$
$$u(\cdot,0) = u_0 \qquad \text{in } \Omega \tag{2.75c}$$

and assume for the data $f \in L^2(0,T;L^2(\Omega))$ (note: crucial, does not work in $L^2(0,T;H^{-1}(\Omega))$) and $u_0 \in L^2(\Omega)$.

Introducing the variable $\sigma = \nabla u$, we obtain (as for the Laplacian) a mixed formulation (in strong form)

$$\begin{aligned} u_t - \operatorname{div}\sigma &= f \\ \sigma - \nabla u &= 0. \end{aligned}$$

Squaring the equations (including the initial condition) and integrating gives the *least squares functional*

$$J(v,\psi) := \|\psi - \nabla v\|^2_{L^2(\Omega_T)} + \|v_t - \operatorname{div}\psi - f\|^2_{L^2(\Omega_T)} + \|v(0) - u_0\|^2_{L^2(\Omega)}. \tag{2.76}$$

As the solution to (2.75) satisfies $J(u,\nabla u) = 0$, a reasonable idea is to minimize (2.76). An appropriate Hilbert space for this problem is given by

$$U := \{(v,\psi) \ : \ v \in W(0,T;H^1_0(\Omega);H^{-1}(\Omega)), \psi \in L^2(\Omega_T)^d, v_t - \operatorname{div}\psi \in L^2(\Omega_T)\}$$

endowed with the natural norm

$$\|(v,\psi)\|^2_U := \|v\|^2_{L^2(0,T;H^1_0(\Omega))} + \|v_t\|^2_{L^2(0,T;H^{-1}(\Omega))} + \|\psi\|^2_{L^2(\Omega_T)} + \|v_t - \operatorname{div}\psi\|^2_{L^2(\Omega_T)}.$$

The Euler-Lagrange equations for the functional J lead to the variational problem: Find $(u,\sigma) \in U$ such that

$$b((u,\sigma),(v,\psi)) = \ell(v,\psi) \qquad \forall (v,\psi) \in U, \tag{2.77}$$

where

$$\begin{aligned} b((u,\sigma),(v,\psi)) &:= \langle \nabla u - \sigma, \nabla v - \psi\rangle_{L^2(\Omega_T)} + \langle u_t - \operatorname{div}\sigma, v_t - \operatorname{div}\psi\rangle_{L^2(\Omega_T)} + \langle u(0), v(0)\rangle_{L^2(\Omega)}, \\ \ell(v,\psi) &:= \langle f, v_t - \operatorname{div}\psi\rangle_{L^2(\Omega)} + \langle u_0, v(0)\rangle_{L^2(\Omega)}. \end{aligned}$$

Note that this problem – in contrast to the previous space-time formulation – gives a symmetric Galerkin formulation. Thus, the Lax-Milgram setting applies and we obtain unique solvability for *any* closed subspace $U_h \subset U$. Note that $U_h = U$ in the following theorem is allowed and gives well-posedness of the formulation (2.77).

Theorem 2.61 *Let $U_h \subseteq U$ be a closed subspace. Then, the bilinear form $b(\cdot,\cdot)$ is continuous and coercive on U, and the problem: Find $(u_h,\sigma_h) \in U_h$ such that*

$$b((u_h,\sigma_h),(v_h,\psi_h)) = \ell(v_h,\psi_h) \qquad \forall (v_h,\psi_h) \in U_h \tag{2.78}$$

is uniquely solvable and there holds the quasi best-approximation property

$$\|(u-u_h,\sigma-\sigma_h)\|_U \le C \min_{(v_h,\psi_h)\in U_h} \|(u-v_h,\sigma-\psi_h)\|_U$$

with a constant independent of U_h.

Proof: Boundedness of the bilinear form $b(\cdot,\cdot)$ and the linear form $\ell(\cdot)$ follows from Cauchy-Schwarz and the embedding $C([0,T];L^2(\Omega)) \subset W(0,T;H^1_0(\Omega),H^{-1}(\Omega))$ (to treat the term with the initial condition). We show coercivity of $b(\cdot,\cdot)$ on U, then the remaining statements follow from the Lax-Milgram lemma. Let $(v,\psi) \in U$ be arbitrary. Then, we write

$$v_t - \Delta v = v_t - \operatorname{div}(\psi) + \operatorname{div}(\psi - \nabla v)$$

and apply the well-posedness of Corollary 2.59 with $f = v_t - \operatorname{div}\psi + \operatorname{div}(\psi - \nabla v)$ to obtain

$$\begin{aligned} \|v\|_{W(0,T;H^1_0(\Omega),H^{-1}(\Omega))} &\lesssim \|v_t - \operatorname{div}\psi\|_{L^2(0,T;H^{-1}(\Omega))} + \|\operatorname{div}(\psi - \nabla v)\|_{L^2(0,T;H^{-1}(\Omega))} + \|v(0)\|_{L^2(\Omega)} \\ &\lesssim \|v_t - \operatorname{div}\psi\|_{L^2(\Omega_T)} + \|\psi - \nabla v\|_{L^2(\Omega_T)} + \|v(0)\|_{L^2(\Omega)}. \end{aligned}$$

The triangle inequality then implies

$$\begin{aligned} \|\psi\|_{L^2(\Omega_T)} &\le \|\psi - \nabla v\|_{L^2(\Omega_T)} + \|\nabla v\|_{L^2(\Omega_T)} \le \|\psi - \nabla v\|_{L^2(\Omega_T)} + \|v\|_{L^2(0,T;H^1_0(\Omega))} \\ &\lesssim \|v_t - \operatorname{div}\psi\|_{L^2(\Omega_T)} + \|\psi - \nabla v\|_{L^2(\Omega_T)} + \|v(0)\|_{L^2(\Omega)}. \end{aligned}$$

Consequently, we obtain

$$\begin{aligned} \|(v,\psi)\|^2_U &= \|v\|^2_{L^2(0,T;H^1_0(\Omega))} + \|v_t\|^2_{L^2(0,T;H^{-1}(\Omega))} + \|\psi\|^2_{L^2(\Omega_T)} + \|v_t - \operatorname{div}\psi\|^2_{L^2(\Omega_T)} \\ &\lesssim \|v_t - \operatorname{div}\psi\|_{L^2(\Omega_T)} + \|\operatorname{div}(\psi - \nabla v)\|_{L^2(\Omega_T)} + \|v(0)\|_{L^2(\Omega)} \\ &= b((v,\psi),(v,\psi)), \end{aligned}$$

which shows coercivity. □

We now discretize the problem with piecewise linear finite elements. Let $\mathcal{T}_h$ be a shape-regular, regular mesh of the space time cylinder $\Omega \times (0,T)$. For simplicity, we assume that each element in $T \in \mathcal{T}_h$ has tensor product structure, i.e., $T = K_h \times \omega_h \in \mathcal{K}_h \times \Omega_h$, where $\mathcal{K}_h$ is a mesh on $(0,T)$ and Ω_h is a shape-regular, regular mesh on Ω. As discrete subspace, we take

$$U_h := S_0^{1,1}(\mathcal{T}_h) \times (S^{1,1}(\mathcal{T}_h))^d \subset U.$$

Note that by this choice, we have $\partial_t v_h - \operatorname{div} \psi_h \in L^2(\Omega \times (0,T))$ for all $(v_h, \psi_h) \in U_h$.
By the quasi-optimality result of Theorem 2.61, error estimates can be obtained by constructing an approximation (here using interpolation) in U_h.

As we have $u \in C([0,T]; L^2(\Omega))$, we may work with a piecewise linear interpolant $\mathcal{I}_h$ in time, which maps $\mathcal{I}_h : C([0,T]; L^2(\Omega)) \to S^{1,1}(\mathcal{K}_h; L^2(\Omega))$, where $S^{1,1}(\mathcal{K}_h; L^2(\Omega))$ denotes the space of piecewise linear mappings (in time) with values in $L^2(\Omega)$. Note that this is a semi-discrete operator. For some Hilbert space X (we will use $X = H_0^1(\Omega), H^{-1}(\Omega)$), this operator has the approximation properties (the proof is essentially identical to the scalar valued linear interpolation operator)

$$\begin{aligned} \|u - \mathcal{I}_h u\|_{L^2(0,T;X)} &\le Ch \|u\|_{H^1(0,T;X)} \qquad \forall u \in H^1(0,T;X), \\ \|(u - \mathcal{I}_h u)'\|_{L^2(0,T;X)} &\le Ch \|u\|_{H^2(0,T;X)} \qquad \forall u \in H^2(0,T;X). \end{aligned}$$

In space, we employ the classical $L^2(\Omega)$-orthogonal projection $\Pi_{0,h} : L^2(\Omega) \to S_0^{1,1}(\Omega_h)$ with the approximation properties for $u \in H^2(\Omega)$

$$\begin{aligned} \|u - \Pi_{0,h} u\|_{H^1(\Omega)} &\le Ch \|u\|_{H^2(\Omega)}, \\ \|u - \Pi_{0,h} u\|_{H^{-1}(\Omega)} &\le Ch^2 \|u\|_{H^1(\Omega)}. \end{aligned}$$

Combining both operators gives an operator mapping into (the first component in) U_h by

$$\mathcal{J}_{0,h} := \mathcal{I}_h \circ (Id \otimes \Pi_{0,h}) : C([0,T]; L^2(\Omega)) \to S_0^{1,1}(\mathcal{T}_h). \tag{2.79}$$

Note that this definition means that, at first, the operator $Id \otimes \Pi_{0,h}$ is applied, which does nothing in the time variable and applies the projection $\Pi_{0,h}$ in space, and afterwards the (semi-discrete in time) operator $\mathcal{I}_h$ is employed, which reproduces polynomials in the spatial variables and thus maps into the correct discrete function space.
Regarding approximation properties, we have the following lemma.

Lemma 2.62 *Let $\mathcal{J}_{0,h}$ be defined by (2.79). Then, for $u \in H^1(0,T; H_0^1(\Omega)) \cap L^\infty(0,T; H^2(\Omega))$ there holds*

$$\|u - \mathcal{J}_{0,h} u\|_{L^2(0,T;H_0^1(\Omega))} \le Ch \left(\|u\|_{H^1(0,T;H_0^1(\Omega))} + \|u\|_{L^\infty(0,T;H^2(\Omega))} \right). \tag{2.80}$$

If there holds $u \in H^2(0,T; H^{-1}(\Omega)) \cap L^\infty(0,T; H_0^1(\Omega))$, then

$$\|(u - \mathcal{J}_{0,h} u)'\|_{L^2(0,T;H^{-1}(\Omega))} \le Ch \left(\|u\|_{H^2(0,T;H^{-1}(\Omega))} + \|u\|_{L^\infty(0,T;H_0^1(\Omega))} \right).$$

Proof: By definition of $\mathcal{J}_{0,h}$, for $t \in (t_j, t_{j+1})$, we may write

$$\mathcal{J}_{0,h} u(t) = \mathcal{I}_h u(t) + \frac{(\Pi_{0,h} u(t_{j+1}) - u(t_{j+1}))(t - t_j) + (\Pi_{0,h} u(t_j) - u(t_j))(t_{j+1} - t)}{h}.$$

With the approximation properties of $\Pi_{0,h}$, we estimate for $t \in (t_j, t_{j+1}) \in \mathcal{K}_h$

$$\begin{aligned} \|u(t) - \mathcal{J}_{0,h} u(t)\|_{H^1(\Omega)} &\le \|u(t) - \mathcal{I}_h u(t)\|_{H^1(\Omega)} + \sum_{k=0,1} \|\Pi_{0,h} u(t_{j+k}) - u(t_{j+k})\|_{H^1(\Omega)} \\ &\le \|u(t) - \mathcal{I}_h u(t)\|_{H^1(\Omega)} + Ch \sum_{k=0,1} \|u(t_{j+k})\|_{H^2(\Omega)} \\ &\le \|u(t) - \mathcal{I}_h u(t)\|_{H^1(\Omega)} + 2Ch \sup_{t \in (t_j, t_{j+1})} \|u(t)\|_{H^2(\Omega)}. \end{aligned}$$

Now, integration over $(0,T)$ and application of the approximation properties of $\mathcal{I}_h$ applied with $X = H_0^1(\Omega)$ gives

$$\|u - \mathcal{J}_{0,h}u\|_{L^2(0,T;H_0^1(\Omega))} \leq Ch \left(\|u\|_{H^1(0,T;H_0^1(\Omega))} + \sup_{t\in(0,T)} \|u(t)\|_{H^2(\Omega)} \right),$$

which is the first estimate.
Similarly, we write

$$(\mathcal{J}_{0,h}u)'(t) = (\mathcal{I}_h u)'(t) + \frac{(\Pi_{0,h}u(t_{j+1}) - u(t_{j+1})) - (\Pi_{0,h}u(t_j) - u(t_j))}{h},$$

and application of the approximation properties of $\Pi_{0,h}$ in the $H^{-1}(\Omega)$-norm as well as the approximation properties of $\mathcal{I}_h$ applied with $X = H^{-1}(\Omega)$ give the second estimate in the same way as above. □

Note that the boundary condition does only enter through the application of the $L^2(\Omega)$-orthogonal projection mapping into $S_0^{1,1}(\Omega_h)$ and is not explicitly used in the proof. Thus, estimate (2.80) holds in the same way for an operator $\mathcal{J}_h := \mathcal{I}_h \circ (Id \otimes \Pi_h) : C([0,T]; L^2(\Omega)) \to S^{1,1}(\mathcal{T}_h)$, where $\Pi_h : L^2(\Omega) \to S^{1,1}(\Omega_h)$ is the $L^2(\Omega)$-orthogonal projection without the zero boundary condition.

The previous lemma gives an estimate in the energy norm for the heat equation, an estimate in the least-squares norm $\|\cdot\|_U$ can also be deduced.

Corollary 2.63 *Assume $u \in L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^2(\Omega)) \cap H^2(0,T;L^2(\Omega)) \cap L^\infty(0,T;H^3(\Omega))$. Then,*

$$\|(u - \mathcal{J}_{0,h}u, \nabla u - \mathcal{J}_h \nabla u)\|_U \leq Ch$$

with a constant $C > 0$ independent of h (depending on some norms of u).

Proof: Exercise. Follows from definition of the U-norm and a slight modification of the previous lemma (to give an estimate for $(u - \mathcal{J}_{0,h}u)'$ in the $L^2(0,T;L^2(\Omega))$-norm). □

Remark 2.64 The regularity requirement of the previous holds for solutions u to the heat equation, e.g., for smooth domains Ω and $u_0 \in H_0^1(\Omega) \cap H^3(\Omega)$, $f \in H^1(0,T;H^2(\Omega))$ and $f(0) + \Delta u_0 \in H_0^1(\Omega)$.

Remark 2.65 The assumption that elements in $\mathcal{T}_h$ can be weakened such that simplical meshes that are refinements of tensor product meshes are also allowed (note that the tensor product elements are prisms). In [5] a construction and error analysis for that can be found.

A drawback of the presented approach is, however, that the considered first order system has more unknowns than the previously considered discretizations and is thus more expensive to solve.

A nice feature of the least-squares formulation is that it has a *built in a posteriori error estimator*.
For $(v_h, \psi_h) \in U_h$ and $K \in \mathcal{T}_h$, we define the local error indicators

$$\eta_K^2(v_h, \psi_h) := \|\psi_h - \nabla v_h\|_{L^2(K)} + \|\partial_t v_h - \operatorname{div} \psi_h - f\|_{L^2(K)} + \|v_h(0) - u_0\|_{L^2(\partial K \cap \Omega \times \{0\})}$$

and the error estimator

$$\eta(v_h, \psi_h) := \left(\sum_{K\in\mathcal{T}_h} \eta_K^2(v_h, \psi_h) \right)^{1/2}.$$

The error estimator indeed is a good measure for the error by the following theorem.

Theorem 2.66 *Let $(u, \psi) \in U$ denote the solution to (2.77). The estimator η is reliable and efficient, i.e., there exist constants $C_1, C_2 > 0$ such that*

$$C_1\|(u - v_h, \sigma - \psi_h)\|_U \leq \eta(v_h, \psi_h) \leq C_2\|(u - v_h, \sigma - \psi_h)\|_U \qquad \forall (v_h, \psi_h) \in U. \tag{2.81}$$

Proof: The statement follows from coercivity and boundedness of $b(\cdot,\cdot)$ and

$$b((u - v_h, \sigma - \psi_h), (u - v_h, \sigma - \psi_h)) = \eta^2(v_h, \psi_h).$$

□

2.8 Approximation for the Navier Stokes equations

In this section we consider the incompressible Navier-Stokes equations. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, $T > 0$ and u_0 be given. We seek functions $\mathbf{u}$, p (bold symbols denote vector valued quantities) solving

$$\begin{aligned} \partial_t \mathbf{u} - \mu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega \times (0,T) \\ \operatorname{div} \mathbf{u} &= 0 && \text{in } \Omega \times (0,T) \\ \mathbf{u}(0,\cdot) &= \mathbf{u}_0 && \text{in } \Omega. \end{aligned}$$

Regarding boundary conditions, there are various different choices depending on the considered physical model. Oftentimes (think of flow through a pipe), the boundary $\Gamma := \partial\Omega$ is decomposed into an inflow boundary Γ_{in} (with inhomogeneous Dirichlet boundary condition there), wall boundary Γ_w (with homogeneous Dirichlet conditions) and an outflow boundary Γ_{out} (with homogeneous Neumann conditions). For simplicity, we assume homogeneous boundary conditions everywhere, i.e.,

$$\mathbf{u} = 0 \qquad \text{in } \partial\Omega \times (0,T).$$

In the following, we consider a weak formulation of the Navier-Stokes equations and apply the semi-discretization approach, i.e., after multiplication with a test-function $\mathbf{v}$ and integration by parts, we seek $\mathbf{u} : [0,T] \to H_0^1(\Omega)^d$ with $\mathbf{u}(0) = \mathbf{u}_0$ and $p : [0,T] \to L^2(\Omega)$ such that

$$\begin{aligned} \int_\Omega \partial_t \mathbf{u} \cdot \mathbf{v} + \int_\Omega \mu \nabla \mathbf{u} : \nabla \mathbf{v} + (\mathbf{u} \cdot \nabla \mathbf{u}) \cdot \mathbf{v} - \int_\Omega \operatorname{div} \mathbf{v}\, p &= \int_\Omega \mathbf{f} \cdot \mathbf{v} \qquad \forall \mathbf{v} \in H_0^1(\Omega)^d, \\ \int_\Omega \operatorname{div} \mathbf{u}\, q &= 0 \qquad \forall q \in L^2(\Omega). \end{aligned}$$

Regarding existence and uniqueness (open problem!) and other formulations, we refer to the lecture "nonlinear partial differential equations". Note that we take a saddle-point approach here. Alternatively, one could also add the side-constraint $\operatorname{div} \mathbf{u} = 0$ into the space for $\mathbf{u}$. However, then, in order to have a conforming FEM method, a discrete subspace with exact divergence free discrete functions (and accordingly such finite elements) has to be constructed.

Semi-discretization

For discretization, we employ the method of lines. We take an inf-sub stable finite dimensional pair $V_h \times Q_h \subset H_0^1(\Omega)^d \times L^2(\Omega)$, e.g., the Taylor-Hood pair. Then, the semi-discretization leads to the problem: Find $\mathbf{u}_h : [0,T] \to V_h$ with $\mathbf{u}_h(0) = \mathbf{u}_{0,h} \in V_h$ and $p_h : [0,T] \to Q_h$ such that

$$\begin{aligned} \langle \partial_t \mathbf{u}_h, \mathbf{v}_h \rangle_{L^2} + a(\mathbf{u}_h, \mathbf{v}_h) + c(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) &= \langle \mathbf{f}, \mathbf{v}_h \rangle \qquad \forall \mathbf{v}_h \in V_h \\ b(\mathbf{u}_h, q_h) &= 0 \qquad \forall q_h \in Q_h \end{aligned}$$

with the bilinear forms $a(\mathbf{u},\mathbf{v}) = \mu \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v}$ and $b(\mathbf{v},p) = -\int_\Omega \operatorname{div} \mathbf{v} p$ and the trilinear form $c(\mathbf{u},\mathbf{v},\mathbf{w}) = \int_\Omega (\mathbf{u} \cdot \nabla \mathbf{v}) \cdot \mathbf{w}$.

Choosing a basis $\{\boldsymbol{\phi}_j^u : j = 1,\dots,N_u\} \subset V_h$ and $\{\varphi_j^p : j = 1,\dots,N_p\} \subset Q_h$, we may write

$$\mathbf{u}_h(x,t) = \sum_{j=1}^{N_u} \mathbf{u}_{h,j}(t) \boldsymbol{\phi}_j^u(x) \qquad p_h(x,t) = \sum_{j=1}^{N_p} \mathbf{p}_{h,j}(t) \varphi_j^p(x)$$

and inserting this into the semi-discrete formulation gives the system of ODEs for the coefficient vectors $\mathbf{u}_h$ and $\mathbf{p}_h$ as

$$\mathbf{M}\mathbf{u}_h'(t) + \mathbf{A}\mathbf{u}_h(t) + \mathbf{C}(\mathbf{u}_h(t))\mathbf{u}_h(t) + \mathbf{B}^T \mathbf{p}_h(t) = \mathbf{F}, \tag{2.82}$$

$$\mathbf{B}\mathbf{u}_h(t) = 0, \tag{2.83}$$

where $\mathbf{A}$ and $\mathbf{M}$ denote the usual stiffness and mass matrices and

$$\mathbf{C} : \mathbb{R}^{N_u} \to \mathbb{R}^{N_u \times N_u}, \quad (\mathbf{C}(\mathbf{v}_h))_{i,j} = c(\mathbf{v}_h, \boldsymbol{\phi}_i^u, \boldsymbol{\phi}_j^u).$$

Time integration

For the time integration of the *nonlinear* ODE (2.82), one can employ a single step method as for the heat equation. However, as is the same there, explicit time stepping schemes should be avoided since the appearing ODE is stiff and thus restrictions on the time step length would have to be applied (the matrix $\mathbf{A}$ appears in both semi-discretizations). Thus, one could employ the implicit Euler-method (or more general the θ-scheme with $\theta \in [1/2, 1]$). Writing $t_n = kn$ for the time steps and $\mathbf{u}_h^n = \mathbf{u}_h(t_n)$, $\mathbf{p}_h^n = \mathbf{p}_h(t_n)$, then a fully discrete scheme with the implicit Euler reads as

$$\begin{aligned} \mathbf{M}\frac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^n}{k} + (\mathbf{A} + \mathbf{C}(\mathbf{u}_h^{n+1}))\mathbf{u}_h^{n+1} + \mathbf{B}^T\mathbf{p}_h^{n+1} &= \mathbf{F}^{n+1}, \\ \mathbf{B}\mathbf{u}_h^{n+1} &= 0. \end{aligned}$$

A big drawback hereby is that a *nonlinear* system of equations has to be solved in each step. This is usually done with variants of the Newton method. Also note that the scheme results in two *coupled* equations for velocity and pressure (the incompressibility constraint results in a saddle-point type system)

$$\begin{pmatrix} \mathbf{M} + k\mathbf{A} + k\mathbf{C}(\mathbf{u}_h^{n+1}) & k\mathbf{B}^T \\ k\mathbf{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_h^{n+1} \\ \mathbf{p}_h^{n+1} \end{pmatrix} = \begin{pmatrix} \mathbf{M}\mathbf{u}_h^n + k\mathbf{F}^{n+1} \\ 0 \end{pmatrix}.$$

2.8.1 Splitting methods

Splitting methods are a very popular choice for the Navier-Stokes equations and they can be applied to many other problems (e.g. hyperbolic equations in the next chapter or problems in quantum mechanics). We motivate the idea of splitting methods by applying easy, classical techniques to a simple ODE example and afterwards present more involved fractional splitting methods that are popular especially for the Navier-Stokes equations.

Idea of splitting methods

We illustrate the main idea on a simple ODE-System

$$\mathbf{u}' = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{u}, \qquad \mathbf{u}(0) = \mathbf{u}_0, \tag{2.84}$$

with matrices $\mathbf{A}, \mathbf{B}$ (generalizations follow later).
Using the matrix exponential function, we obtain the solution as

$$\mathbf{u}(t) = e^{t(\mathbf{A}+\mathbf{B})}\mathbf{u}_0.$$

Goal: a good approximation to $\mathbf{u}(k)$, where k denotes the time step size.

Rules of the game: $\mathbf{u}_0 \mapsto e^{k(\mathbf{A}+\mathbf{B})}\mathbf{u}_0$ is hard to realize/expensive, but the mappings $\mathbf{v} \mapsto e^{k\mathbf{A}}\mathbf{v}$ and $\mathbf{v} \mapsto e^{k\mathbf{B}}\mathbf{v}$ are easy/cheap.

Classical methods to approximate $e^{k(\mathbf{A}+\mathbf{B})}$ are:

- Lie-splitting: $e^{t(\mathbf{A}+\mathbf{B})} \approx e^{t\mathbf{B}}e^{t\mathbf{A}}$. Algorithmically one write for a step of length k

$$\mathbf{u}_{1/2} := e^{k\mathbf{A}}\mathbf{u}_0, \qquad \mathbf{u}_1 := e^{k\mathbf{B}}\mathbf{u}_{1/2}.$$

- Strang-splitting: $e^{t(\mathbf{A}+\mathbf{B})} \approx e^{1/2t\mathbf{A}}e^{t\mathbf{B}}e^{1/2t\mathbf{A}}$. Algorithmically one write for a step of length k

$$\mathbf{u}_{1/3} := e^{k/2\mathbf{A}}\mathbf{u}_0, \qquad \mathbf{u}_{2/3} := e^{k\mathbf{B}}\mathbf{u}_{1/3}, \qquad \mathbf{u}_1 := e^{k/2\mathbf{A}}\mathbf{u}_{2/3}.$$

  Note: the Strang-splitting seems more expensive than the Lie-splitting. However, if one does multiple steps after each other the computational costs of both methods is roughly the same. This follows from

$$e^{k/2\mathbf{A}}e^{k\mathbf{B}}e^{k/2\mathbf{A}} \quad e^{k/2\mathbf{A}}e^{k\mathbf{B}}e^{k/2\mathbf{A}} \quad \dots \quad e^{k/2\mathbf{A}}e^{k\mathbf{B}}e^{k/2\mathbf{A}} = e^{k/2\mathbf{A}}e^{k\mathbf{B}}e^{k\mathbf{A}}e^{k\mathbf{B}} \dots e^{k\mathbf{B}}e^{k/2\mathbf{A}}.$$

- The SWSS ("symmetrically weighted sequential splitting") method: $e^{t(\mathbf{A}+\mathbf{B})} \approx \frac{1}{2}\left(e^{t\mathbf{A}}e^{t\mathbf{B}} + e^{t\mathbf{B}}e^{t\mathbf{A}}\right)$. Algorithmically one write for a step of length k

$$\mathbf{u}_{1/2} := e^{k\mathbf{A}}e^{k\mathbf{B}}\mathbf{u}_0, \qquad \widetilde{\mathbf{u}}_{1/2} := e^{k\mathbf{B}}e^{k\mathbf{A}}\mathbf{u}_0, \qquad \mathbf{u}_1 := \frac{1}{2}\left(\mathbf{u}_{1/2} + \widetilde{\mathbf{u}}_{1/2}\right).$$

Note that, in general, matrices $\mathbf{A}, \mathbf{B}$ do not commute, i.e., $\mathbf{AB} \neq \mathbf{BA}$, and thus there *does not hold* $e^{t(\mathbf{A}+\mathbf{B})} = e^{t\mathbf{A}}e^{t\mathbf{B}}$. This means that the methods described above really are different approximations to the exact solution.

The Lie-splitting is a first order method and the Strang-splitting and the SWSS-method are second order (proof: Taylor!). In principle, it is possible to construct methods of higher order. The most popular methods, however, are the Lie- and Strang-splitting.

Application of splitting methods

Splitting methods are not limited to linear ODEs, possibilities are

- A, B can be (differential-)operators or their (spatial-)discretizations,
- A, B can be non-linear.

Splitting methods are a popular choice, if the operators A and B have different characteristics and thus need different (numerical) treatment/discretizations.

For operators A, B the matrix exponentials in the previous chapter are replaced by the evolution operators E_A and E_B of the problems

$$\begin{aligned} u' &= Au, \qquad u(0) = u_0 \\ u' &= Bu, \qquad u(0) = u_0. \end{aligned}$$

Then, one step (of length k) of the Lie-Splitting reads as

$$u_1 := E_B(k)E_A(k)u_0$$

and one step of the Strang-Splitting reads as

$$u_1 := E_A(k/2)E_B(k)E_A(k/2)u_0.$$

In practice, this evolution operators are not computable and thus replaced by discrete evolution operators, e.g., one step of explicit or implicit Euler.

Example 2.67 Adding a lower order terms to the heat equation gives the equation

$$u_t = \Delta u + \mathbf{b} \cdot \nabla u.$$

Then, the splitting $Au = \Delta u$ and $Bu = \mathbf{b} \cdot \nabla u$ means that one has to solve a heat equation for E_A and an advection equation (see next chapter!) for E_B. ∎

Splitting methods can also significantly reduce the computational effort on the discrete level.

Example 2.68 (ADI–alternating directions implicit variant) Consider the 2D heat equation with the domain $\Omega = (0,1)^2$

$$\begin{aligned} u_t &= u_{xx} + u_{yy} && \text{in } \Omega \times (0,T) \\ u(x,t) &= 0 && \text{on } \partial\Omega \times (0,T). \end{aligned}$$

We employ semi-discretization with piecewise linear finite elements on a uniform mesh with $n \times n$ knots (i.e. n knots per coordinate direction). As we should use implicit methods in time (i.e. the implicit Euler), we have to solve a 2D-space problem in each time step. With typical lexicographic numbering of the unknowns the (sparse) matrices have bandwidth $b = \mathcal{O}(n)$. Consequently, the LU-decomposition

for solution of the linear system needs $\mathcal{O}(bn^2) = \mathcal{O}(n^3)$ (note that the size of the matrices is $N = n^2$) memory.

A possible splitting method is given by $Au = u_{xx}$ and $Bu = u_{yy}$. Then, the discrete evolutions E_A and E_B consists of solutions of *decoupled* 1D equations, which produce linear systems of size n (which are even tridiagonal). This implies that the cost for one step of the Lie-splitting is only $2n\mathcal{O}(n) = \mathcal{O}(n^2) = \mathcal{O}(N)$, which is optimal. ∎

2.8.2 Splitting-schemes for Navier-Stokes

For the Navier-Stokes equations splitting methods are usually applied together with fractional time steps, i.e., one decomposes the interval (t_i, t_{i+1}) into finitely many sub-intervals and apply a (possibly different) splitting scheme on each subinterval.
We write the Navier-Stokes equations (or its semi-discretization) as operator evolution equation in the general form

$$\partial_t u + A(u) = f$$

and aim to split the operator $A = A_1 + A_2$.

The Peaceman-Rachford scheme

An operator splitting method with a similar flavour to the ADI variant of the previous example is the rather old and popular *Peaceman-Rachford scheme*, which makes a half-step of an implicit Euler for A_1 and explicitly treats A_2 and then makes a half step with flipped roles:

$$\begin{aligned} \frac{u^{n+1/2} - u^n}{k/2} + A_1 u^{n+1/2} + A_2 u^n &= f^{n+1/2}, \\ \frac{u^{n+1} - u^{n+1/2}}{k/2} + A_1 u^{n+1/2} + A_2 u^{n+1} &= f^{n+1}, \end{aligned}$$

where $u^{n+\xi} = u(t_{n+\xi})$ and $t_{n+\xi} = t_n + \xi k$.

In order to analyze stability of the scheme, we apply it to the ODE system $\mathbf{u}' = \mathbf{A}\mathbf{u}$ with the operator splitting $\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A} =: A_1 + A_2$ with $\alpha + \beta = 1$. Then, one step of the scheme reads as

$$\mathbf{u}^{n+1} = \left(1 + \frac{k}{2}\beta\mathbf{A}\right)^{-1} \left(1 - \frac{k}{2}\alpha\mathbf{A}\right) \left(1 + \frac{k}{2}\alpha\mathbf{A}\right)^{-1} \left(1 - \frac{k}{2}\beta\mathbf{A}\right) \mathbf{u}^n.$$

Thus, the stability function of the Peaceman-Rachford scheme reads as

$$R(z) = \frac{(1 - \alpha z/2)(1 - \beta z/2)}{(1 + \alpha z/2)(1 + \beta z/2)} = 1 - z + \frac{z^2}{2} - (\alpha^2 + \beta^2 + \alpha\beta)\frac{z^3}{4} + \mathcal{O}(|z|^4).$$

Consequently, the method is A-stable and of second order. However, as $R(z) \to 1$ for $z \to -\infty$, we do not have L-stability.

Remark 2.69 Note that the Peaceman-Rachford scheme can be seen as two steps of Lie-splittings, where the first step has as (discrete) evolution operators the explicit Euler for E_{A_2} followed by the implicit Euler for E_{A_1} and the roles are reversed in the second step.

Applied to the semi-discrete Navier Stokes equations, we choose the splitting

$$\begin{aligned} A_1 &= \frac{1}{2}\mathbf{A} + \mathbf{B}^T \qquad \text{together with the incompressibility constraint,} \\ A_2 &= \frac{1}{2}\mathbf{A} + \mathbf{C}(\cdot). \end{aligned}$$

Note that parts of $\mathbf{A}$ should always be treated implicitly as it is a stiff part of the ODE.

Thus, the scheme reads as: Given $\mathbf{u}_h^n$ compute $\mathbf{u}_h^{n+1/2}, \mathbf{p}_h^{n+1/2}, \mathbf{u}_h^{n+1}$ by

$$\textit{Step } 1: \begin{cases} \mathbf{M}\frac{\mathbf{u}_h^{n+1/2}-\mathbf{u}_h^n}{k/2} + \frac{1}{2}\mathbf{A}\mathbf{u}_h^{n+1/2} + \mathbf{B}^T\mathbf{p}_h^{n+1/2} = \mathbf{F}^{n+1/2} - \frac{1}{2}\mathbf{A}\mathbf{u}_h^n - \mathbf{C}(\mathbf{u}_h^n)\mathbf{u}_h^n \\ \mathbf{B}\mathbf{u}_h^{n+1/2} = 0 \end{cases}$$

$$\textit{Step } 2: \left\{ \mathbf{M}\frac{\mathbf{u}_h^{n+1}-\mathbf{u}_h^{n+1/2}}{k/2} + \frac{1}{2}\mathbf{A}\mathbf{u}_h^{n+1} + \mathbf{C}(\mathbf{u}_h^{n+1})\mathbf{u}_h^{n+1} = \mathbf{F}^{n+1} - \frac{1}{2}\mathbf{A}\mathbf{u}_h^{n+1/2} - \mathbf{B}^T\mathbf{p}_h^{n+1/2}. \right.$$

Note that in the first step the nonlinear term is treated explicitly (thus no solution of a nonlinear system is needed) and in the second step the incompressibility constraint is treated explicitly (which just is the second line of step 1 and thus already computed) and therefore no saddle-point system is needed.

The fractional-step-θ-scheme

In this scheme, for some $\theta < 1/2$, three steps are made and it reads as

$$\begin{aligned} \frac{u^{n+\theta} - u^n}{\theta k} + A_1 u^{n+\theta} + A_2 u^n &= 0 \\ \frac{u^{n+1-\theta} - u^{n+\theta}}{(1-2\theta)k} + A_1 u^{n+\theta} + A_2 u^{n+1-\theta} &= 0 \\ \frac{u^{n+1} - u^{n+1-\theta}}{\theta k} + A_1 u^{n+1} + A_2 u^{n+1-\theta} &= 0 \end{aligned}$$

In order to analyze stability of the scheme, we again apply it to the ODE $\mathbf{u}' = \mathbf{A}\mathbf{u}$ with the operator splitting $\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$ with $\alpha + \beta = 1$. Then, with $\theta' = (1-2\theta)$, one step of the method reads as[8]

$$\mathbf{u}^{n+1} = (1+\alpha\theta k\mathbf{A})^{-2}\,(1-\beta\theta k\mathbf{A})^2\,(1+\beta\theta' k\mathbf{A})^{-1}\,(1-\alpha\theta' k\mathbf{A})\,\mathbf{u}^n.$$

Thus, the stability function of the scheme reads as

$$R(z) = \frac{(1-\beta\theta z)^2(1-\alpha\theta' z)}{(1+\alpha\theta z)^2(1+\beta\theta' z)} = 1 - z + \Big(1 + (\beta-\alpha)(2\theta^2 - 4\theta + 1)\Big)\frac{z^2}{2} + \mathcal{O}(|z|^3).$$

Now, the choice of θ, α, β influences both the accuracy and stability of the scheme. As

$$\lim_{z\to-\infty} |R(z)| = \frac{\beta}{\alpha},$$

there holds A-stability, if $\alpha > \beta$. However, the method is second order, if $\alpha = \beta$ or $\theta = 1 - 1/\sqrt{2}$. Taking $\theta = 1 - 1/\sqrt{2}$, the choices

$$\alpha = \frac{1-2\theta}{1-\theta} \qquad \beta = \frac{\theta}{1-\theta}$$

give $\alpha + \beta = 1$ as well as $\beta/\alpha = \theta/(1-2\theta) < 1$ and consequently, a second order, and a so called *strongly A-stable* method[9].

Applied to the semi-discrete Navier Stokes equations this reads as: Given $\mathbf{u}_h^n$, compute $\mathbf{u}_h^{n+\theta}, \mathbf{u}_h^{n+1-\theta}, \mathbf{u}_h^{n+1}$ as well as $p^{n+\theta}, p^{n+1}$ from

$$\textit{Step } 1: \begin{cases} \mathbf{M}\frac{\mathbf{u}_h^{n+\theta}-\mathbf{u}_h^n}{\theta k} + \alpha\mathbf{A}\mathbf{u}_h^{n+\theta} + \mathbf{B}^T\mathbf{p}_h^{n+\theta} = \mathbf{F}^{n+\theta} - \beta\mathbf{A}\mathbf{u}_h^n - \mathbf{C}(\mathbf{u}_h^n)\mathbf{u}_h^n \\ \mathbf{B}\mathbf{u}_h^{n+\theta} = 0 \end{cases}$$

$$\textit{Step } 2: \left\{ \mathbf{M}\frac{\mathbf{u}_h^{n+1-\theta}-\mathbf{u}_h^{n+\theta}}{(1-2\theta)k} + \beta\mathbf{A}\mathbf{u}_h^{n+1-\theta} + \mathbf{C}(\mathbf{u}_h^{n+1-\theta})\mathbf{u}_h^{n+1-\theta} = \mathbf{F}^{n+\theta} - \alpha\mathbf{A}\mathbf{u}_h^{n+\theta} - \mathbf{B}^T\mathbf{p}_h^{n+\theta} \right.$$

$$\textit{Step } 3: \begin{cases} \mathbf{M}\frac{\mathbf{u}_h^{n+1}-\mathbf{u}_h^{n+1-\theta}}{\theta k} + \alpha\mathbf{A}\mathbf{u}_h^{n+1} + \mathbf{B}^T\mathbf{p}_h^{n+1} = \mathbf{F}^{n+1} - \beta\mathbf{A}\mathbf{u}_h^{n+1-\theta} - \mathbf{C}(\mathbf{u}_h^{n+1-\theta})\mathbf{u}_h^{n+1-\theta} \\ \mathbf{B}\mathbf{u}_h^{n+1} = 0 \end{cases}$$

Note that, again, we have decoupled the incompressibility constraint and the nonlinear convection term. Step 1 and step 3 lead to solution of *linear systems* of equations, while step 2 leads to a non-linear system, but without the need of the saddle point structure.

[8] the operators in the brackets commute, since $(1+\sigma\mathbf{A})^{-1}(1+\mu\mathbf{A}) = (1+\sigma\mathbf{A})^{-1}(1+\mu\mathbf{A})(1+\sigma\mathbf{A})(1+\sigma\mathbf{A})^{-1} = (1+\sigma\mathbf{A})^{-1}(1+\sigma\mathbf{A})(1+\mu\mathbf{A})(1+\sigma\mathbf{A})^{-1} = (1+\mu\mathbf{A})(1+\sigma\mathbf{A})^{-1}$ for all $\sigma, \mu \in \mathbb{R}$.

[9] This means that $\lim_{z\to-\infty} |R(z)| < 1$, which also induces convergence to 0 of the discrete solution for $\mathrm{Re}\,\lambda < 0$ as in the (stronger!) case of L-stability.

The IMEX method

A popular simplification of the fractional θ-step method is the first order IMEX method

$$\mathbf{M}\frac{\mathbf{u}_h^{n+1}-\mathbf{u}_h^n}{k}+\mathbf{A}\mathbf{u}_h^{n+1}+\mathbf{B}^T\mathbf{p}_h^{n+1}=\mathbf{F}^{n+1}-\mathbf{C}(\mathbf{u}_h^n)\mathbf{u}_h^n$$
$$\mathbf{B}\mathbf{u}_h^{n+1}=0,$$

i.e., the side constraint is treated implicitly, but the nonlinearity only explicit.
This corresponds to the choice $\alpha=1$, $\beta=0$, $\theta=1$ in the fractional θ-step method and therefore is of first order.
IMEX methods of higher order can also be constructed, which is typically done with *partitioned Runge-Kutta methods* (see literature or lecture "numerics of ODEs").

For error analysis of the fully discrete methods, we refer to literature, e.g., [8]. In principle, the employed techniques are similar to the analysis done for the heat equation, however, the non-linearity and saddle-point structure of the problem induce a lot more effort into deriving the estimates.

2.8.3 Numerical example

We consider the incompressible Navier-Stokes equations with viscosity $\nu=0.001$ in two dimensions.
As showcase, we take the Schäfer-Turek benchmark (see NgSolve tutorials), which models flow in a rectangular pipe around a circular obstacle. Here, the top and bottom part of the rectangle have zero boundary conditions, while on the lateral boundaries there is an inflow (left) and outflow (right) boundary condition.

For the initial condition, a parabolic flow profile is chosen, with which a (stationary!) Stokes problem is solved to obtain an initial velocity profile (depicted in the top left picture in Figure 2.8). For time discretization the IMEX method of the previous subsection is chosen and the resulting velocity profiles at some time-steps are depicted in Figure 2.8.

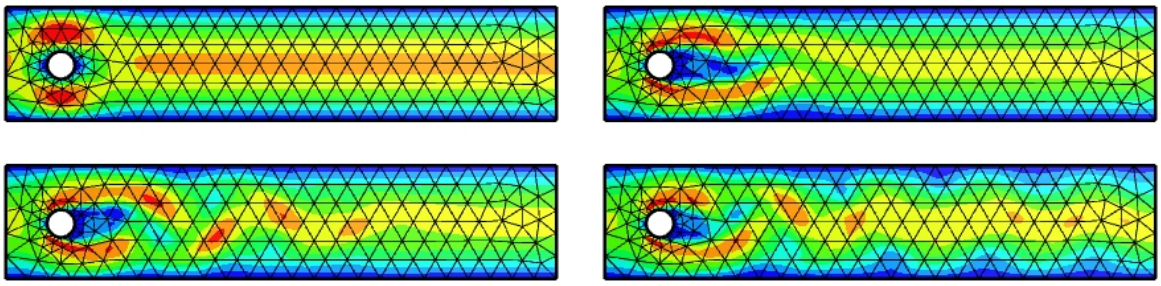

Figure 2.8: solution to incompressible Navier Stokes at $T=0$ (left,top), $T=0.5$ (right,top), and $T=1$ (left,bottom) $T=5$ (right,bottom).

Chapter 3

Hyperbolic equations

We now consider a different type of time-dependent partial differential equations, so called *hyperbolic* PDEs. In principle, similar methods as for parabolic equations can be employed and as such, we consider approaches based on *semi-discretization* and *space time formulations*. Additionally, here, we discuss finite difference methods both in space and time, which are rather popular with engineers.

The roadmap for this chapter is as follows:

- We start with the wave equation and employ methods derived in the previous chapter based on semi-discretization and time-stepping. A key difference to the parabolic case is, that the continous equation satisfies energy conservation (rather than dissipation) and thus numerical methods that preserve this have to be constructed.
- Then, we consider the advection equation and introduce finite difference schemes. The main focus hereby is a stability analysis of the methods.
- Afterwards, we consider a space-time discontinuous Galerkin formulation, which generalizes the DG approach of Section 2.6. Finally, we show that this approach can also be formulated as DG in space combined with a time-stepping method.

3.1 Popular examples

Wave equations

The prime example of a second order hyperbolic PDE is the (scalar, acoustic) wave equation

$$u_{tt} - c\Delta u = f,$$

where $c \in \mathbb{R}$ is the so called wave speed.

Hyperbolic conservation laws

A more general definition of hyperbolicity can also be given for first order (systems of) equations. Such equations share the same characteristic behaviour with the wave equation.

A general form of (systems of) *conservation laws* is given by

$$\partial_t \mathbf{u} + \sum_{j=1}^{d} \partial_{x_j}(\mathbf{f}^j(\mathbf{u})) = 0 \qquad (x,t) \in \Omega \times (0,\infty). \tag{3.1}$$

- The components $\mathbf{u}_j : \Omega \to \mathbb{R}$, $j = 1, \dots, s$ of the function $\mathbf{u} : \Omega \to \mathbb{R}^s$ are called *states*, the vector $\mathbf{u}$ *state vector*.
- The functions $\mathbf{f}^j : \mathbb{R}^s \to \mathbb{R}^s$, $j = 1, \dots, d$, are called *flux functions*.

A conservation law is called hyperbolic, if it additionally satisfies a diagonalization property for its linearization.

Definition 3.1 (hyperbolicity) *The conservation law (3.1) is called* hyperbolic*, if for every state vector* $\mathbf{u}$ *and every direction* $\underline{\omega} \in \mathbb{R}^d \setminus \{0\}$*, the matrix*

$$\mathbf{A}(\mathbf{u}, \underline{\omega}) := \sum_{j=1}^{d} \omega_j D_{\mathbf{u}} \mathbf{f}^j(\mathbf{u})$$

is diagonalizable (in $\mathbb{R}$*).*
If, for every state $\mathbf{u}$ *and every direction* $\underline{\omega}$*, the eigenvalues are pairwise distinct, then the system is called* strictly hyperbolic*.*

For the historically important 1D-case, this can be written more easily.

Exercise 3.2 Let $d = 1$ and consider the conservation law

$$\partial_t \mathbf{u} + \partial_x(\mathbf{F}(\mathbf{u})) = 0. \tag{3.2}$$

Show that: this conservation law is hyperbolic, if $D\mathbf{F}(\mathbf{u})$ is real diagonalisable for all state vectors $\mathbf{u}$. ∎

An important special case are scalar conservation laws.

Example 3.3 Let $s = 1$, then the functions $f^1, \ldots, f^d$ are real valued. Writing $\mathbf{F}(u) := (f^1, \ldots, f^d)^\top$, one obtains a conservation law of the form

$$\partial_t u + \nabla \cdot (\mathbf{F}(u)) = 0. \tag{3.3}$$

The easiest example of that form is given by the *advection equation*

$$u_t + \mathbf{b} \cdot \nabla u = 0. \tag{3.4}$$

∎

Remark 3.4 Note that the name conservation law can be motivated by taking an arbitrary "control volume" $D \subset \mathbb{R}^d$ and integration, which gives, e.g., for a scalar conservation law that

$$\frac{d}{dt} \int_D u \, dx + \int_{\partial D} \mathbf{F}(u) \cdot n \, ds = 0,$$

which can be interpreted as mass conservation.

Conservation laws are oftentimes obtained by laws of physics (e.g., mass, energy of (angular) momentum conservation). Very popular examples come from fluid or gas dynamics such as the Euler equations.

Example 3.5 (Euler equations) The *Euler equations* describe the movement of a gas or fluid. Hereby, one has conservation of mass, energy and momentum. Denote by $\mathbf{v}(x,t) \in \mathbb{R}^3$ the speed of a particle at position x and time t, by $\rho(x,t)$ the density, by $p(x,t)$ the pressure and by $e(x,t)$ the specific intrinsic energy (temperature), then with the total energy $E = \rho e + \frac{1}{2}|\mathbf{v}|^2$ one obtains the equations for the conservation law

$$\begin{aligned}
\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) &= 0 && \text{mass conservation} \\
\partial_t(\rho \mathbf{v}_i) + \nabla \cdot (\rho \mathbf{v} \mathbf{v}_i) + \partial_{x_i} p &= 0, \quad i = 1,2,3, && \text{momentum conservation} \\
\partial_t E + \nabla \cdot (\mathbf{v}(E+p)) &= 0 && \text{energy conservation.}
\end{aligned}$$

Note that (in $\mathbb{R}^3$) this gives 5 equations for 6 unknowns. The missing equation can be obtained from a constitutive law, e.g., an ideal gas law of the form $p = \rho(\gamma - 1)e$ with a constant γ.
The Euler equations can be brought to form (3.1):

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho \mathbf{v}_1 \\ \rho \mathbf{v}_2 \\ \rho \mathbf{v}_3 \\ E \end{pmatrix}, \qquad \mathbf{f}^1 = \begin{pmatrix} \rho \mathbf{v}_1 \\ p + \rho \mathbf{v}_1^2 \\ \rho \mathbf{v}_1 \mathbf{v}_2 \\ \rho \mathbf{v}_1 \mathbf{v}_3 \\ \mathbf{v}_1(E+p) \end{pmatrix}, \qquad \mathbf{f}^2 = \begin{pmatrix} \rho \mathbf{v}_2 \\ \rho \mathbf{v}_1 \mathbf{v}_2 \\ p + \rho \mathbf{v}_2^2 \\ \rho \mathbf{v}_2 \mathbf{v}_3 \\ \mathbf{v}_2(E+p) \end{pmatrix}, \qquad \mathbf{f}^3 = \begin{pmatrix} \rho \mathbf{v}_3 \\ \rho \mathbf{v}_1 \mathbf{v}_3 \\ \rho \mathbf{v}_2 \mathbf{v}_3 \\ p + \rho \mathbf{v}_3^2 \\ \mathbf{v}_3(E+p) \end{pmatrix}.$$

∎

3.2 The wave equation

Let $\Omega \subset \mathbb{R}^d$ and $T \in (0, \infty]$. As a model problem, we consider the wave equation on the (infinite) space-time cylinder $\Omega_T := \Omega \times (0, T)$, i.e., with given data f, u_0, v_0, we want to solve

$$u_{tt} - \Delta u = f \qquad \text{in } \Omega_T, \tag{3.5a}$$
$$u(x,t) = 0 \qquad \text{on } \partial\Omega \times (0,T), \tag{3.5b}$$
$$u(\cdot,0) = u_0,\ u_t(\cdot,0) = v_0 \qquad \text{in } \Omega \tag{3.5c}$$

Note that for $\Omega = \mathbb{R}^d$ and $f = 0$, a solution is given by the *d'Alambert formula*

$$u(x,t) = \frac{u_0(x+t) + u_0(x-t)}{2} + \frac{1}{2}\int_{x-t}^{x+t} v_0(y)\, dy.$$

Exercise 3.6 Let $d = 1$. Then, a solution to the inhomogeneous problem is given by

$$u(x,t) = \frac{u_0(x+t) + u_0(x-t)}{2} + \frac{1}{2}\int_{x-t}^{x+t} v_0 + \frac{1}{2}\int_0^t \int_{x-(t-s)}^{x+(t-s)} f(y,s)\, dyds.$$

3.2.1 Semi-discretization

In the following, we only consider the method of lines with compatible initial data.
As for the heat equation, we multiply with a test function $w = w(x) \in H_0^1(\Omega)$ and integrate over Ω to obtain the formulation

$$\langle u_{tt}(t), w\rangle_{L^2} + a(u(t), w) = \langle f(t), w\rangle_{L^2}. \tag{3.6}$$

An appropriate function space for this problem to be well-defined requires

- $u \in L^2(0,T; H_0^1(\Omega))$, $\partial_t u \in L^2(0,T; L^2(\Omega))$, $\partial_{tt} u \in L^2(0,T; H^{-1}(\Omega))$
- $u(0) = u_0$ in $H_0^1(\Omega)$, $\partial_t u(0) = v_0$ in $L^2(\Omega)$.

Note that embeddings similarly to (2.34) hold and the regularities $u \in L^2(0,T; H_0^1(\Omega))$, $\partial_t u \in L^2(0,T; L^2(\Omega))$, $\partial_{tt} u \in L^2(0,T; H^{-1}(\Omega))$ imply also $u \in C([0,T]; H_0^1(\Omega))$ as well as $\partial_t u \in C([0,T]; L^2(\Omega))$, which means that the initial values can be well-defined.

Regarding unique solvability, we have (see [13, Thm. 8.1]) the following result.

Theorem 3.7 *Let $f \in L^2(0,T; L^2(\Omega))$, $u_0 \in H_0^1(\Omega)$ and $v_0 \in L^2(\Omega)$. Then, there exists a unique solution to (3.6), which satisfies $u \in L^2(0,T; H_0^1(\Omega))$, $\partial_t u \in L^2(0,T; L^2(\Omega))$, $\partial_{tt} u \in L^2(0,T; H^{-1}(\Omega))$ with*

$$\|u\|^2_{L^2(0,T;H_0^1(\Omega))} + \|\partial_t u\|^2_{L^2(0,T;L^2(\Omega))} \le C\left(\|u_0\|^2_{H^1(\Omega)} + \|v_0\|^2_{L^2(\Omega)} + \|f\|^2_{L^2(0,T;L^2(\Omega))}\right).$$

For the heat equation, we obtained an energy inequality, which showed a dissipative behaviour. For the wave equation, this is different, as we actually have energy conservation.

Lemma 3.8 *Let $f = 0$ and define $\mathcal{E}(t) := \frac{1}{2}\|\nabla u(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|u_t\|^2_{L^2(\Omega)}$. Let u solve (3.6) and additionally assume $\partial_t u \in L^2(0,T; H_0^1(\Omega))$. Then, there holds*

$$\mathcal{E}(t) = \mathcal{E}(0) = \frac{1}{2}\|\nabla u_0\|^2_{L^2(\Omega)} + \frac{1}{2}\|v_0\|^2_{L^2(\Omega)}. \tag{3.7}$$

Proof: For fixed $t > 0$, we take $w = u_t(t)$ as test function and obtain

$$\langle u_{tt}(t), u_t(t)\rangle_{L^2} + a(u(t), u_t(t)) = 0,$$

which implies

$$\frac{1}{2}\frac{d}{dt}\|u_t\|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}\|\nabla u\|^2_{L^2(\Omega)}$$

or $\frac{d}{dt}\mathcal{E}(t) = 0$. □

Now, we take a finite dimensional space $V_h \subset H^1_0(\Omega)$ with $\dim(V_h) = N < \infty$ with a basis $\{\phi_i \ : \ i = 1, \dots, N\}$. Replacing $H^1_0(\Omega)$ by V_h in (3.6) and u_0 by some $u_{0,h} \in V_h$ and v_0 by some $v_{0,h} \in V_h$ gives a semi-discrete solution $u_h \in L^2(0,T;V_h)$. Inserting a basis expansion of u_h and v_h into the equation gives in the same way as for the heat equation an equivalent system of ODEs

$$\mathbf{M}\mathbf{u}''(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{F}(t), \tag{3.8}$$

with mass matrix $\mathbf{M}$ and stiffness matrix $\mathbf{A}$, as well as the initial conditions

$$\mathbf{u}(0) = \mathbf{u}_0 \qquad \mathbf{u}'(0) = \mathbf{v}_0.$$

Note that for the semi-discrete problem with $\mathbf{F} = 0$, there still holds energy conservation

$$\mathcal{E}_h(t) := \frac{1}{2}\mathbf{u}'\mathbf{M}\mathbf{u}' + \frac{1}{2}\mathbf{u}\mathbf{A}\mathbf{u} = \text{const.}$$

Regarding a-priori estimates for the semi-discrete problem, a similar result to Theorem 2.13 holds (see [12] for a proof).

Theorem 3.9 *Let u solve (3.6) and u_h be its semi-discrete approximation. Let Assumption 2.44 hold. Then, assuming sufficient regularity of u and choosing $u_{0,h} = P_h u_0 \in V_h$ and $v_{0,h} = P_h v_0 \in V_h$, we have*

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \le Ch^2\Big(\|u(t)\|_{H^2(\Omega)} + \int_0^t \|u_{tt}(s)\|_{H^2(\Omega)} ds\Big)$$

$$|u(t) - u_h(t)|_{H^1(\Omega)} \le Ch\Big(\|u(t)\|_{H^2(\Omega)} + \int_0^t \|u_{tt}(s)\|_{H^2(\Omega)} ds\Big)$$

as well as

$$\|u'(t) - u_h'(t)\|_{L^2(\Omega)} \le Ch^2\Big(\|u_t(t)\|_{H^2(\Omega)} + \int_0^t \|u_{tt}(s)\|_{H^2(\Omega)} ds\Big).$$

3.2.2 Two formulations as first order system

Formulation 1

Introducing (on the PDE level) the function $v = \partial_t u$, the variational formulation can also be written as: Find $(u, v) \in L^2(0,T;H^1_0(\Omega)) \times L^2(0,T;L^2(\Omega))$ such that

$$\langle v_t(t), \psi\rangle_{L^2} + a(u(t), \psi) = \langle f(t), \psi\rangle_{L^2} \qquad \forall \psi \in H^1_0(\Omega) \tag{3.9a}$$

$$\langle u_t(t), \phi\rangle_{L^2} - \langle v(t), \phi\rangle_{L^2} = 0 \qquad \forall \phi \in L^2(\Omega) \tag{3.9b}$$

On the ODE-level this corresponds to the usual rewriting of the second order ODE (3.8) to a first order system by defining $\mathbf{v} = \mathbf{u}'$, which gives

$$\begin{aligned} \mathbf{u}' &= \mathbf{v} \\ \mathbf{M}\mathbf{v}' &= \mathbf{F} - \mathbf{A}\mathbf{u}, \end{aligned}$$

with initial conditions

$$\mathbf{u}(0) = \mathbf{u}_0 \qquad \mathbf{v}(0) = \mathbf{v}_0.$$

Formulation 2

We also rewrite the wave equation in a different way as a first order system. Introducing the (vector-valued) unknown $\boldsymbol{\sigma} = \int_0^t \nabla u$, the wave equation $u'' - \Delta u = f$ can be written

$$\begin{aligned} \boldsymbol{\sigma}' &= \nabla u \\ u' - \operatorname{div} \boldsymbol{\sigma} &= \tilde{f} \end{aligned}$$

with the integrated source $\tilde{f} = \int_0^t f$. Now, as for the Stokes problem, one can make a mixed variational formulation (for simplicity we treat the time derivative classically): Find $(\boldsymbol{\sigma}, u) \in C^1((0,T); H(\operatorname{div}, \Omega)) \times C^1((0,T); L^2(\Omega))$ such that

$$\begin{aligned} \langle \boldsymbol{\sigma}', \boldsymbol{\tau} \rangle_{L^2} + \langle u, \operatorname{div} \boldsymbol{\tau} \rangle_{L^2} &= 0 \qquad && \forall\, \boldsymbol{\tau} \in H(\operatorname{div}, \Omega) \\ \langle u', w \rangle_{L^2} - \langle w, \operatorname{div} \boldsymbol{\sigma} \rangle_{L^2} &= \langle \tilde{f}, w \rangle_{L^2} && \forall\, w \in L^2(\Omega). \end{aligned}$$

With mass matrices $\mathbf{M}_\sigma$ (corresponding to evaluation of $\langle \boldsymbol{\sigma}, \boldsymbol{\tau} \rangle_{L^2}$ with basis functions of $H(\operatorname{div}, \Omega)$-finite elements, e.g. Raviart-Thomas elements) and $\mathbf{M}_u$ (corresponding to evaluation of $\langle u, v \rangle_{L^2}$ with basis functions in $L^2(\Omega)$, e.g., piecewise constant functions) as well as the stiffness matrix $\mathbf{B}$ (corresponding to evaluation of the bilinear form $\langle u, \operatorname{div} \boldsymbol{\tau} \rangle_{L^2}$ at the (already) chosen FEM-bases in $L^2(\Omega) \times H(\operatorname{div}, \Omega)$), after space discretization we obtain the system of ODEs

$$\begin{pmatrix} \mathbf{M}_\sigma & 0 \\ 0 & \mathbf{M}_u \end{pmatrix} \begin{pmatrix} \boldsymbol{\sigma}' \\ \mathbf{u}' \end{pmatrix} = \begin{pmatrix} 0 & -\mathbf{B}^T \\ \mathbf{B} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\sigma} \\ \mathbf{u} \end{pmatrix} + \begin{pmatrix} 0 \\ \widetilde{\mathbf{F}} \end{pmatrix}.$$

In the following, we set $f = 0$. A nice advantage of formulation 2 is that the ODE system has the structure of a Hamiltonian system and thus conservation of energy is easily seen from

$$\frac{d}{dt} \left(\tfrac{1}{2} \boldsymbol{\sigma}^T \mathbf{M}_\sigma \boldsymbol{\sigma} + \tfrac{1}{2} \mathbf{u}^T \mathbf{M}_u \mathbf{u} \right) = \boldsymbol{\sigma}^T \mathbf{M}_\sigma \boldsymbol{\sigma}' + \mathbf{u}^T \mathbf{M}_u \mathbf{u}' = -\boldsymbol{\sigma}^T \mathbf{B}^T \mathbf{u} + \mathbf{u}^T \mathbf{B} \boldsymbol{\sigma} = 0.$$

3.2.3 Time-stepping methods for wave equations

In principal, one can reduce the second order ODE to a first order system (as done above), and apply some standard time-stepping method (Euler) for it. This will in general require the solution of linear systems of twice the size. In addition, the structure (symmetric and positive definite) may be lost, which makes it difficult to solve. Moreover, we want a numerical method that also matches the qualitative behaviour of energy conservation of the exact solution.

Exercise 3.10 Numerically verify that the Euler methods do not conserve energy for the scalar ODE $u'' + u = 0$.

In the following, we consider methods that allow for energy conservation and are based on either the second order ODE formulation or the formulations as first order systems.

The Crank-Nicolson method

We consider the first-order system (3.9) and apply the Crank-Nicolson scheme to both equations (note that in practice, one would do a semi-discretization in space first, for simplicity we formulate everything in H_0^1 and L^2 here, but a semi-discrete version holds as well). This gives

$$\langle v^n - v^{n-1}, \phi \rangle_{L^2} + \frac{k}{2} a(u^n + u^{n-1}, \phi) = \frac{1}{2} \langle f^n + f^{n-1}, \phi \rangle_{L^2} \qquad \forall \phi \in H_0^1(\Omega) \tag{3.10}$$

$$\langle u^n - u^{n-1}, \psi \rangle_{L^2} - \frac{k}{2} \langle v^n + v^{n-1}, \psi \rangle_{L^2} = 0 \qquad \forall \psi \in L^2(\Omega) \tag{3.11}$$

The following theorem shows that the *total energy*

$$E^n = \frac{1}{2} \|\nabla u^n\|^2_{L^2(\Omega)} + \frac{1}{2} \|v^n\|^2_{L^2(\Omega)}$$

is preserved.

Theorem 3.11 *Let $f = 0$. Then, there holds for all n*

$$E^n = \frac{1}{2}\|\nabla u^n\|^2_{L^2(\Omega)} + \frac{1}{2}\|v^n\|^2_{L^2(\Omega)} = \frac{1}{2}\|\nabla u^{n-1}\|^2_{L^2(\Omega)} + \frac{1}{2}\|v^{n-1}\|^2_{L^2(\Omega)} = E^{n-1},$$

where u^n, v^n are defined by (3.10).

Proof: We use the test-functions $\phi = u^n - u^{n-1}$ and $\psi = v^n - v^{n-1}$ in (3.10). This gives

$$\begin{aligned}
\langle v^n - v^{n-1}, u^n - u^{n-1}\rangle_{L^2} + \frac{k}{2}a(u^n + u^{n-1}, u^n - u^{n-1}) = 0,\\
\langle u^n - u^{n-1}, v^n - v^{n-1}\rangle_{L^2} - \frac{k}{2}\langle v^n + v^{n-1}, v^n - v^{n-1}\rangle_{L^2} = 0.
\end{aligned}$$

Subtracting both equations and multiplication with $\frac{2}{k}$ leads to

$$\begin{aligned}
0 &= a(u^n + u^{n-1}, u^n - u^{n-1}) + \langle v^n + v^{n-1}, v^n - v^{n-1}\rangle_{L^2}\\
&= a(u^n, u^n) - a(u^{n-1}, u^{n-1}) + \langle v^n, v^n\rangle_{L^2} - \langle v^{n-1}, v^{n-1}\rangle_{L^2},
\end{aligned}$$

which implies the energy conservation. □

Exercise 3.12 Show that for the implicit Euler method, there holds energy dissipation, i.e., $E^n \le E^{n-1}$.

The Newmark time-stepping method

We now consider the semi-discretization as second order ODE system

$$\mathbf{M}\mathbf{u}'' + \mathbf{A}\mathbf{u} = \mathbf{f}$$

and derive a very popular single-step method for this formulation, the so called *Newmark method*.

The Newmark method is based on second order Taylor expansion, thus we need the state $\mathbf{u}_n \approx \mathbf{u}(t_n)$ (for better readability we now write the time stepping indices as subscripts instead of superscripts), the velocity $\mathbf{u}'_n \approx \mathbf{u}'(t_n)$ and acceleration $\mathbf{u}''_n \approx \mathbf{u}''(t_n)$. For given state $\mathbf{u}_n$, the acceleration can be computed from the given ODE as $\mathbf{u}''_n = \mathbf{M}^{-1}(\mathbf{f}_n - \mathbf{A}\mathbf{u}_n)$ with $\mathbf{f}_n = \mathbf{f}(t_n)$.

Now, the next iterate for the state $\mathbf{u}_{n+1}$ is given by second order Taylor expansion, the next iterate for the velocity is given by first order Taylor expansion and the appearing second derivatives are weighted averages of old and new accelerations. Consequently, for parameters $\beta, \gamma \in \mathbb{R}$, the method reads as

$$\mathbf{u}_{n+1} = \mathbf{u}_n + k\mathbf{u}'_n + k^2\Big((\tfrac{1}{2} - \beta)\mathbf{u}''_n + \beta\mathbf{u}''_{n+1}\Big) \tag{3.12}$$

$$\mathbf{u}'_{n+1} = \mathbf{u}'_n + k\Big((1-\gamma)\mathbf{u}''_n + \gamma\mathbf{u}''_{n+1}\Big). \tag{3.13}$$

Inserting the formula for $\mathbf{u}_{n+1}$ into $\mathbf{M}\mathbf{u}'' + \mathbf{A}\mathbf{u} = \mathbf{f}$ at time t_{n+1} gives

$$\mathbf{M}\mathbf{u}''_{n+1} + \mathbf{A}\Big(\mathbf{u}_n + k\mathbf{u}'_n + k^2((\tfrac{1}{2} - \beta)\mathbf{u}''_n + \beta\mathbf{u}''_{n+1})\Big) = \mathbf{f}_{n+1}$$

or after rearrangement

$$\left(\mathbf{M} + \beta k^2\mathbf{A}\right)\mathbf{u}''_{n+1} = \mathbf{f}_{n+1} - \mathbf{A}\left(\mathbf{u}_n + k\mathbf{u}'_n + k^2(\tfrac{1}{2} - \beta)\mathbf{u}''_n\right). \tag{3.14}$$

Thus, the Newmark method requires in (3.14) the solution of a SPD linear system with system matrix $\mathbf{M} + k^2\beta\mathbf{A}$ and then has two explicit update formulas (3.12) and (3.13).

We use the short notation $[U]^m_n := U^m - U^n$ for the time steps in the following. For some parameters, the Newmark method satisfies a discrete energy conservation.

Theorem 3.13 *Let* $\mathbf{f} = 0$. *Then,*

$$\left[\tfrac{1}{2}(\mathbf{u}')^T\mathbf{M}\mathbf{u}' + \tfrac{1}{2}\mathbf{u}^T\mathbf{A}_{eq}\mathbf{u}\right]_n^{n+1} = -(\gamma - \tfrac{1}{2})(\mathbf{u}_{n+1} - \mathbf{u}_n)\mathbf{A}_{eq}(\mathbf{u}_{n+1} - \mathbf{u}_n),$$

with the so called equivalent stiffness matrix $\mathbf{A}_{eq}$ *given by*

$$\mathbf{A}_{eq} = \mathbf{A} + (\beta - \tfrac{1}{2}\gamma)k^2\mathbf{A}\mathbf{M}^{-1}\mathbf{A}.$$

Proof: Essentially direct calculations, write

$$\left[\tfrac{1}{2}(\mathbf{u}')^T\mathbf{M}\mathbf{u}' + \tfrac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u}\right]_n^{n+1} = \tfrac{1}{2}(\mathbf{u}'_{n+1} + \mathbf{u}'_n)^T\mathbf{M}(\mathbf{u}'_{n+1} - \mathbf{u}'_n)^T + \tfrac{1}{2}(\mathbf{u}_{n+1} + \mathbf{u}_n)^T\mathbf{A}(\mathbf{u}_{n+1} - \mathbf{u}_n)$$

and then use the defining equations (3.12), (3.13) and (3.14) multiple times. The lengthy calculation can be found in [11]. □

From this, we get the conservation of the modified energy depending on the parameter γ:

- $\gamma = \frac{1}{2}$: conservation;
- $\gamma > \frac{1}{2}$: damping;
- $\gamma < \frac{1}{2}$: growth of energy (unstable).

If $\mathbf{A}_{eq}$ is positive definite, then this conservation proves stability. This is unconditionally true if $\beta \geq \frac{1}{2}\gamma$ (the method is called unconditionally stable). If $\beta < \frac{1}{2}\gamma$, the allowed time step is limited by

$$k^2 \leq \lambda_{max}(\mathbf{M}^{-1}\mathbf{A})^{-1}\frac{1}{\frac{1}{2}\gamma - \beta}.$$

For second order problems we have $\lambda_{max}(\mathbf{M}^{-1}\mathbf{A}) \sim h^{-2}$, and thus $k \leq Ch$ which is a CFL type condition. Choices for β and γ of particular interests are:

- $\gamma = \frac{1}{2}, \beta = \frac{1}{4}$: unconditionally stable, conservation of original energy ($\mathbf{A}_{eq} = \mathbf{A}$), method is equivalent to Crank-Nicolson.
- $\gamma = \frac{1}{2}, \beta = 0$: conditionally stable. We have to solve

  $$\mathbf{M}\mathbf{u}''_{n+1} = \mathbf{f}_{n+1} - \mathbf{A}(\mathbf{u}_n + k\mathbf{u}'_n + \tfrac{k^2}{2}\mathbf{u}''_n),$$

  which is explicit, if $\mathbf{M}$ is cheaply invertible.

For $\gamma > 1/2$, the Newark method is of first order, for $\gamma = 1/2$, one even has quadratic convergence in time.

Methods for the Hamiltonian first order system

Methods tailored for the skew-symmetric (Hamiltonian) structure are so called *symplectic methods* (again, for details see lecture numerics of ODEs). The easiest symplectic method is the *symplectic Euler* method, which reads as

$$\begin{aligned}\mathbf{M}_\sigma\frac{\boldsymbol{\sigma}^{n+1} - \boldsymbol{\sigma}^n}{k} &= -\mathbf{B}^T\mathbf{u}^n,\\ \mathbf{M}_u\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{k} &= \mathbf{B}\boldsymbol{\sigma}^{n+1}.\end{aligned}$$

For updating the second variable, the new value of the first variable is used. For the analysis, we can reduce the large system to 2×2 systems, where β are singular values of $\widetilde{\mathbf{B}} := \mathbf{M}_\sigma^{-1/2}\mathbf{B}\mathbf{M}_u^{-1/2}$ (square-roots of eigenvalues of $\widetilde{\mathbf{B}}^T\widetilde{\mathbf{B}}$):

$$\boldsymbol{\sigma}' = -\beta\mathbf{u} \qquad \mathbf{u}' = \beta\boldsymbol{\sigma}.$$

The symplectic Euler method can be written as

$$\begin{pmatrix} \sigma_{n+1} \\ u_{n+1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ k\beta & 1 \end{pmatrix}\begin{pmatrix} 1 & -k\beta \\ 0 & 1 \end{pmatrix}}_{\mathbf{T}=\begin{pmatrix} 1 & -k\beta \\ k\beta & 1-(k\beta)^2 \end{pmatrix}} \begin{pmatrix} \sigma_n \\ u_n \end{pmatrix}$$

The eigenvalues of $\mathbf{T}$ satisfy $\lambda_1\lambda_2 = \det(\mathbf{T}) = 1$, and, if $k\beta \leq \sqrt{2}$, they are conjugate complex, and thus $|\lambda_1| = |\lambda_2| = 1$. Thus, the discrete solution is oscillating without damping or growth.

The symplectic Euler method is of first order.

Mass lumping

A drawback of some methods presented here is that, even for the explicit type methods, one has to solve a linear system with system matrix being some mass matrix $\mathbf{M}$.
A technique to circumvent this is so called *mass lumping*. Hereby the mass matrix $\mathbf{M}$ is replaced by the diagonal matrix $\overline{\mathbf{M}}$ with entries $\overline{\mathbf{M}}_{jj} = \sum_k \mathbf{M}_{jk}$, i.e., the row sums of $\mathbf{M}$.

This procedure can be understood as calculation of the time derivative term using numerical quadrature. Let $d = 2$, $\mathcal{T}_h$ be a shape-regular, regular triangulation and $V_h = S_0^{1,1}(\mathcal{T}_h)$ with a basis $\{\phi_i \ : \ i = 1, \dots, N\}$ of hat-functions. For $T \in \mathcal{T}_h$, the nodal quadrature formula corresponds to evaluation in the vertices $x_{T,\ell}$, $\ell = 1, 2, 3$ of T, i.e.,

$$Q_{T,h}(g) = \frac{|T|}{3}\sum_{\ell=1}^{3} g(x_{T,\ell}) \approx \int_T g.$$

Applying this for all elements for the time derivative term in the semi-discretization defines the approximative bilinear form

$$\langle u_h'', v\rangle_h = \sum_{T\in\mathcal{T}_h} Q_{T,h}(u_h'' v).$$

Now the key observation is that the FEM-basis functions ϕ_i satisfy that $(\phi_i\phi_j)(x_T) = 0$ for $i \neq j$ for all nodes x_T of the mesh. Thus, as all quadrature points are nodes of the mesh this implies $\langle \phi_i, \phi_j\rangle_h = 0$ for $i \neq j$, i.e., the corresponding mass matrix is diagonal!
Fix ϕ_j and denote by x^j the corresponding node in the mesh $\mathcal{T}_h$. We now show that

$$\langle \phi_j, \phi_j\rangle_h = \sum_{i=1}^{n} \langle \phi_j, \phi_i\rangle_{L^2}.$$

Denoting by $\omega(x^j) := \bigcup\{\overline{T} \ : \ T \in \mathcal{T}_h, x^j \in \overline{T}\}$ the so called patch of the mesh node x^j, by definition of the quadrature rule, we have

$$\langle \phi_j, \phi_j\rangle_h = \sum_{T\in\mathcal{T}_h, T\subset\omega(x^j)} \frac{|T|}{3}\sum_{\ell=1}^{3}\phi_j^2(x_{T,\ell}) = \frac{|\omega(x^j)|}{3}.$$

The L^2-inner products $\langle \phi_j, \phi_i\rangle_{L^2}$ are only non-zero, if two basis functions (hat functions!) ϕ_i, ϕ_j correspond to neighboring nodes (i.e. nodes that are connected via an edge in the mesh) $x^i_{T_1}$ and $x^j_{T_2}$. Moreover, the integral $\int_T \phi_i\phi_j$ does vanish for all elements T with $x^i_{T_1} \vee x^j_{T_2} \notin \overline{T}$.
Transformation to the reference element shows that for any element T with nodes $x^i_{T_1}, x^j_{T_2} \in \overline{T}$, there holds

$$\int_T \phi_i\phi_j = \frac{|T|}{12}, \qquad \int_T \phi_j^2 = \frac{|T|}{6}.$$

Now, for every x^j and $T \in \mathcal{T}_h, T \subset \omega(x^j)$ there exist only two nodes with this property. Thus,

$$\sum_{i=1}^{n}\langle \phi_j, \phi_i\rangle_{L^2} = \sum_{i=1}^{n}\sum_{T\in\mathcal{T}_h, T\subset\omega(x^j)} \langle \phi_j, \phi_i\rangle_{L^2(T)} = \sum_{T\in\mathcal{T}_h, T\subset\omega(x^j)} \frac{|T|}{6} + 2\frac{|T|}{12} = \frac{|\omega(x^j)|}{3}$$

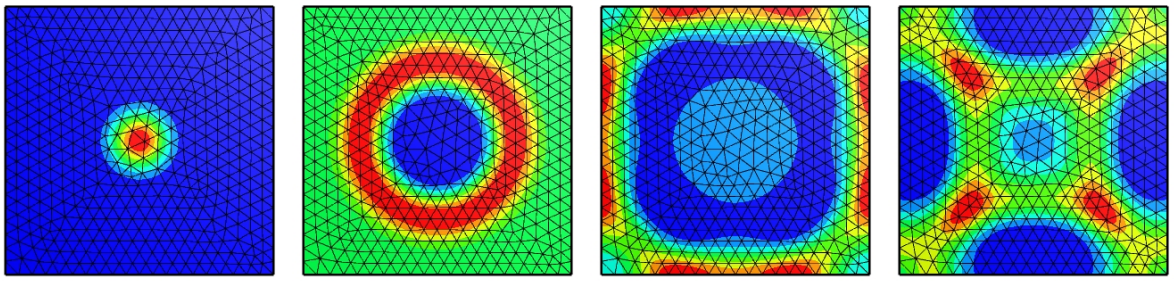

Figure 3.1: Numerical solution to the wave equation at $T = 0, 0.25, 0.5, 0.75$.

and we have shown the claimed indentity.
Now, inserting a basis expansion of u_h and v into the formulation

$$\langle u_h'', v\rangle_h + a(u_h, v) = \langle f, v\rangle_{L^2} \qquad \forall v \in V_h$$

gives the semi-discrete system

$$\overline{\mathbf{M}}\mathbf{u}'' + \mathbf{A}\mathbf{u} = \mathbf{F},$$

i.e., the mass lumped system.

Note that this quadrature rule is of second order, i.e., the quadrature error can be bounded by Ch^2, and thus error bounds for the semi-discretization error using mass lumping similar to Theorem (3.9) still hold.

3.2.4 Numerical example

We consider the 2D-wave equation on the unit square $\Omega = [0,1]^2$ with Neumann boundary conditions

$$u_t - \Delta u = 0 \quad \text{in } \Omega_T, \tag{3.15}$$
$$\partial_n u = 0 \quad \text{on } \partial\Omega \times (0,T), \tag{3.16}$$
$$u(x,0) = u_0 \qquad u_t(x,0) = v_0 \tag{3.17}$$

We use the method of lines with piecewise cubic finite elements ($p = 3$) in space and Newmark timestepping ($\gamma = 1/2$, $\beta = 1/4$).

In Figure 3.1, we choose $u_0 = \exp(-100 \cdot |x - 0.5|^2)$ and $v_0 = u_0$ and observe that the initial peak in the middle (think of it as a stone dropping into water) produces a circular wave travelling until it hits the boundary, where it is reflected. In contrast to the heat equation, there is no dissipative behaviour observed.

Moreover, in Figure 3.2, one can observe that the energy difference between the initial discrete energy $E^0 = \frac{1}{2}\mathbf{u}_0\mathbf{A}\mathbf{u}_0 + \frac{1}{2}\mathbf{v}_0\mathbf{A}\mathbf{v}_0$ is essentially preserved over time when the Newmark method is employed for time-stepping.

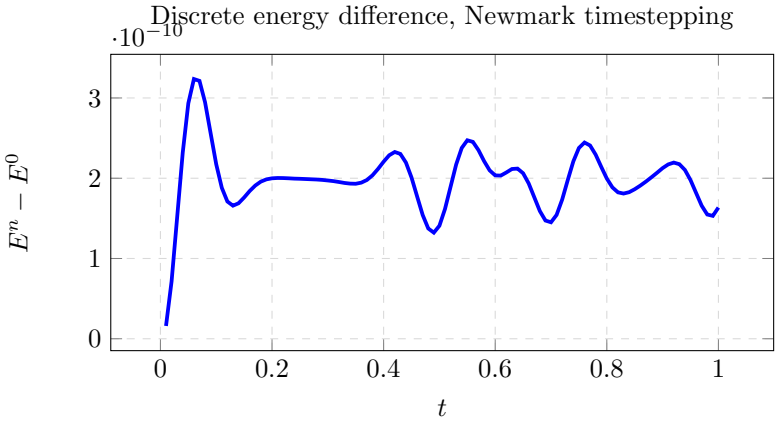

Figure 3.2: Difference between initial energy E^0 and energy E^n at time-step t_n.

3.3 The advection equation – finite differences

Idea: Obtain a numerical method by replace derivatives by evaluations of difference quotients. This is very popular for engineers, much less so for mathematicians.

Advantages:

- very easy to understand and implement;
- can be used for *non-linear* equations.

Disadvantages: hard to treat complicated geometries, boundary conditions.

Focus here: stability of methods in time ($\to$ do not care about boundary conditions by considering full-space problems). also for simplicity: 1D-space only here.

3.3.1 FD for the advection equation

As model problem, with $a \in \mathbb{R}$, we consider the advection equation

$$\begin{aligned} u_t + a u_x &= g && \text{in } \mathbb{R} \times (0, \infty) \\ u(x, 0) &= u_0(x) && \text{in } \mathbb{R}. \end{aligned}$$

For the data g and u_0 we assume compact support.

For the case $g \equiv 0$, the exact solution is given by

$$u(x, t) = u_0(x - at). \tag{3.18}$$

In order to approximate derivatives by difference quotients, we consider a uniform mesh $x_i = ih$, $i \in \mathbb{Z}$ in space and a uniform mesh in time $t_n = nk$, $n = 0, 1, \ldots$. We aim to compute $u_i^n \approx u(x_i, t_n)$.

Notation: We call a sequence $(U_i)_{i\in\mathbb{Z}}$ *grid function*. One can think of this as values at the knots $x_i = ih$. In order to denote the current time step n, we use super scripts. Thus, the grid function $U^n = (U_i^n)_{i\in\mathbb{Z}}$ can be seen as values in the points (x_i, t_n).

Moreover, for grid functions $U = (U_i)_{i\in\mathbb{Z}}$, we introduce the difference operators $D^+U_i := U_{i+1} - U_i$ (forward difference), $D^-U_i := U_i - U_{i-1}$ (backward difference) and $D_0U_i := U_{i+1} - U_{i-1}$ (symmetric difference).

We consider the following *explicit* methods:

1. *forward time/backward space*: take the right-difference quotient in time and the left difference quotient in space, i.e., approximate a 1D- derivative by $(u(x+h) - u(x))/h$, which leads to

$$\frac{u_i^{n+1} - u_i^n}{k} + a\frac{u_i^n - u_{i-1}^n}{h} = g(x_i, t_n). \tag{3.19}$$

2. *forward time/forward space*: take the right-difference quotient in time and space, which leads to

$$\frac{u_i^{n+1} - u_i^n}{k} + a\frac{u_{i+1}^n - u_i^n}{h} = g(x_i, t_n). \tag{3.20}$$

3. *Lax-Friedrichs*: take the central difference quotient in space, which leads to

$$\frac{1}{k}\left(u_i^{n+1} - \frac{1}{2}\left(u_{i+1}^n + u_{i-1}^n\right)\right) + \frac{a}{2h}\left(u_{i+1}^n - u_{i-1}^n\right) = g(x_i, t_n). \tag{3.21}$$

With the grid function $U^n := (u_i^n)_{i\in\mathbb{Z}}$, all schemes have the form

$$U^{n+1} = EU^n + kG^n, \tag{3.22}$$

with the so called *propagation operator* E, which is a linear operator on the space of grid functions, and G^n denoting the grid function $(g(x_i, t_n))_{i\in\mathbb{Z}}$.

As usual, the key properties are the notions of consistency and stability. Consistency is the error of the method in one step with exact initial data, i.e., let u be the exact solution and a grid function U_{kh}^n given by

$$U_{kh,i}^n := u(x_i, t_n),$$

then, the consistency error is

$$\tau_i^{n+1} := \frac{1}{k}\left[U_{kh,i}^{n+1} - ((EU_{kh}^n)_i + kG_i^n)\right].$$

Using the equation $u_t + au_x = g$, this can also be more conveniently written as

$$\tau_i^{n+1} = \frac{1}{k}\left(U_{kh,i}^{n+1} - (EU_{kh}^n)_i\right) - (u_t(x_i, t_n) + au_x(x_i, t_n)) \tag{3.23}$$

for sufficiently smooth functions u.

Exercise 3.14 Use Taylor expansion to show that for the methods introduced above there holds

$$|\tau_i^n| \le C\,(k + h)\,,$$

with a constant $C > 0$ depending on u, but not on h, k.

Stability measures the amplification of previous errors by the numerical method. Define the error

$$\varepsilon_i^n := u(x_i, t_n) - u_i^n = U_{kh,i}^n - u_i^n.$$

For the method and the consistency error, there holds

$$\begin{aligned} U^{n+1} &= EU^n + kG^n \\ U_{kh}^{n+1} &= EU_{kh}^n + kG^n + k\tau^{n+1}. \end{aligned}$$

Due to the linearity of E, we obtain the recursion

$$\varepsilon^{n+1} = E\varepsilon^n + k\tau^{n+1}.$$

We now fix a norm on the space of grid functions

$$\|(V_i)_{i\in\mathbb{Z}}\|_{\ell_h^1} := \sum_{i\in\mathbb{Z}} h|V_i|.$$

There holds

$$\|\varepsilon^n\|_{\ell_h^1} \le \|E^n\|_{\ell_h^1}\|\varepsilon^0\|_{\ell_h^1} + k\sum_{\ell=1}^{n}\|E^{n-\ell}\|_{\ell_h^1}\|\tau^\ell\|_{\ell_h^1}.$$

Thus, for fixed T and $0 \le nk \le T$, it is reasonable to require

$$\sup_{n:0\le nk\le T}\|E^n\|_{\ell_h^1} \le C_T, \tag{3.24}$$

with a constant $C_T > 0$ independent of k and h, since then

$$\|\varepsilon^n\|_{\ell_h^1} \le C_T\Big[\|\varepsilon^0\|_{\ell_h^1} + \sup_{\ell\le n}\|\tau^\ell\|_{\ell_h^1}\underbrace{\sum_{\ell=1}^{n}k}_{\le T}\Big].$$

For the considered methods, there follows

$$\sup_{n:0\le nk\le T}\|\varepsilon^n\|_{\ell_h^1} \le C_T\left[\|\varepsilon^0\|_{\ell_h^1} + C(k+h)\right].$$

If the starting error $\|\varepsilon^0\|_{\ell_h^1}$ is $\mathcal{O}(h)$, which holds for nodal evaluation or even for taking mean values over elements, the overall error will be of order $\mathcal{O}(k+h)$.
Nonetheless, the stability property (3.24) is key. Practically it holds, provided

- $\|E\|_{\ell_h^1} \le 1$
- or slightly weaker $\|E\|_{\ell_h^1} \le 1 + Ck$ with a constant $C > 0$ not depending on k or h.

As seen also previously for other model problems, the stability property will only hold for explicit methods, if the quantity k/h is sufficiently small, i.e., there holds a CFL condition of the form

$$\frac{|ak|}{h} \le c, \tag{3.25}$$

with some constant $c > 0$ independent of k, h, but dependent on the problem and chosen numerical method.

Example 3.15 (Advection equation on torus) We consider

$$\begin{aligned} &u_t + u_x = 0, \qquad x \in (0,1), \quad t > 0, \qquad u(0,t) = u(1,t) \quad \forall t > 0\\ &u(x,0) = \sin(\pi x) \end{aligned}$$

Figure 3.3–3.8 show the behaviour of different numerical methods. In particular, the CFL-condition $\lambda = \frac{k|a|}{h} \le 1$ is essential for the explicit methods. ∎

Exercise 3.16 Show that, for the Lax-Friedrichs scheme, under the CFL-condition $|ak/h| \le 1$, there holds

$$\|E\|_{\ell_h^1} \le 1.$$

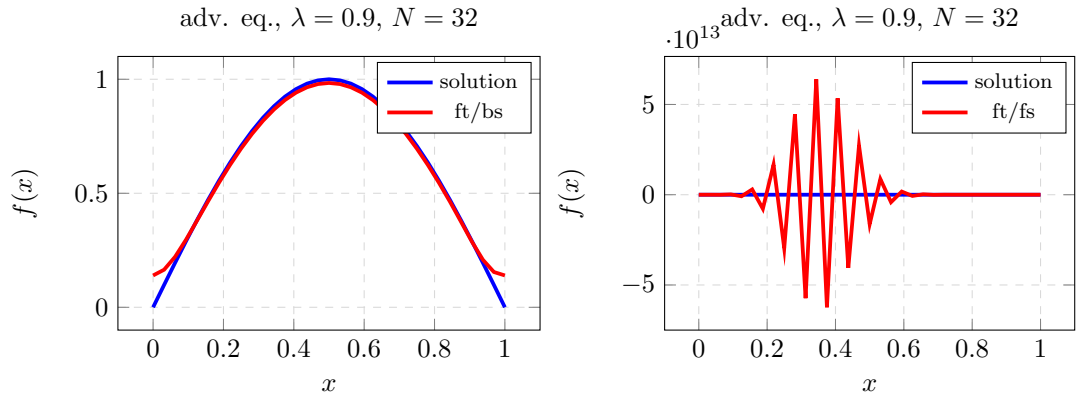

Figure 3.3: comparison of exact solution and approximation by ft/bs and ft/fs.

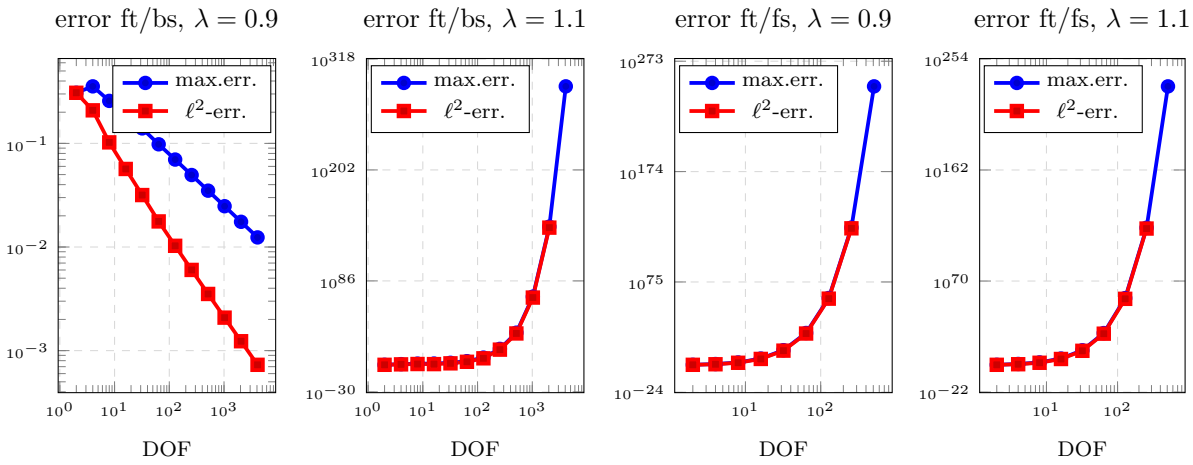

Figure 3.4: influence of CFL-condition for ft/bs and ft/fs.

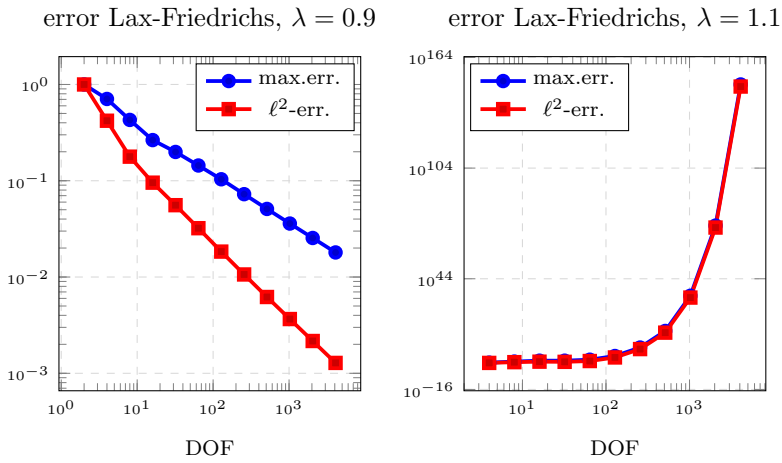

Figure 3.5: influence of CFL-condition for Lax-Friedrichs.

3.3.2 Upwinding

We now consider the forward time/forward space and forward time/backward space methods. The key for stability of the methods is the sign of a:

$$\begin{cases} \text{use ft/bs (3.19)} & \text{if } a > 0 \\ \text{use ft/fs (3.20)} & \text{if } a < 0\ . \end{cases} \tag{3.26}$$

Theorem 3.17 (stability of upwinding) *Under the CFL-condition $|ak/h| \leq 1$, the propagation operator E of the upwind method (3.26) satisfies $\|E\|_{\ell^1_h} \leq 1$. For sufficiently smooth solutions, the error is then of order $\mathcal{O}(k+h)$.*

Proof: Consider the case $a > 0$. For a grid function $(V_i)_{i\in\mathbb{Z}}$, there holds

$$(EV)_i = V_i - \frac{ak}{h}\,(V_i - V_{i-1})$$

and consequently

$$\begin{aligned} \|EV\|_{\ell^1_h} &\leq \sum_i h\left[\left|1-\frac{ka}{h}\right||V_i| + |V_{i-1}|\left|\frac{ka}{h}\right|\right] \\ &\overset{(3.25)}{=} \left(1-\frac{ak}{h}\right)\sum_i h|V_i| + \frac{ak}{h}\sum_i h|V_{i-1}| = \|V\|_{\ell^1_h}, \end{aligned}$$

which gives the statement. □

Exercise 3.18 Show that: Under the assumptions of Theorem 3.17 and the CFL-condition $|a|k/h \leq 1$, there holds $\|E\|_{\ell^\infty} \leq 1$ (operator norm), i.e. the method is stable in ℓ^∞. ∎

Physical plausibility of the CFL condition/upwinding

Considering the ft/fs method, we see that the numerical value at point (x_i, t_n) is only constructed from values at (x_i, t_{n-1}) and (x_{i+1}, t_{n-1}). Inductively, this can be traced back until $t = 0$ and thus the numerical method only uses the initial data at the points $(x_{i+j}, 0)$ for $j = 0, \ldots, n$. For the case $a > 0$, this means that only values right of x_i are used (see red dots in Figure 3.6). This is called *numerical domain of dependence*.

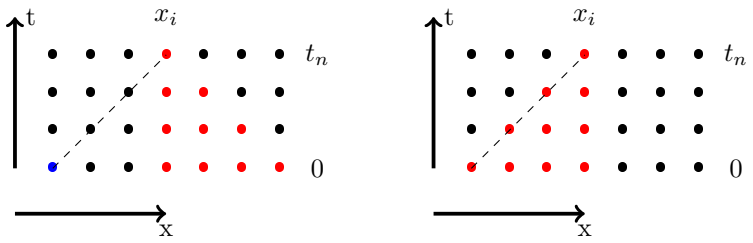

Figure 3.6: The ft/fs method (left) and the ft/bs (right) for $a = 1$.

However, by the solution formula for $f = 0$, the exact solution at point (x_i, t_n) is given by $u_0(x_i - at_n)$, i.e., it does only depend on the values of u_0 at a point left of x_i (*domain of dependence of the true solution*). Taking initial data that is e.g. 1 in the blue dot in Figure 3.6 and zero for all $x \geq x_i$, then $u(x_i, t_n) = 1$, but the numerical approximation will be $u_i^n = 0$, *irrespectively of the step sizes!*.

Thus, if the numerical domain of dependence is not containing the domain of dependence of the exact solution, the method can not work! We have thus obtained:

- For $a > 0$ the ft/fs-method can not work. In the same way, for $a < 0$, the ft/bs-method can not work.
- For $a < 0$ the ft/fs-method can only work, if $nh \geq at_n$, which is exactly the CFL condition.
- For $a > 0$ the ft/bs-method can only work, if $nh \geq at_n$, which is exactly the CFL condition.

Remark 3.19 Upwinding can also be formulated for systems of equations. Let $\mathbf{A}$ be a constant matrix and consider

$$\mathbf{U}_t + \mathbf{A}\mathbf{U}_x = \mathbf{0}.$$

If $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ is diagonalizable, then one does ft/bs in transformed variables $\widetilde{\mathbf{U}} = \mathbf{V}^{-1}\mathbf{U}$ with $\mathbf{D}_{ii} > 0$ and ft/fs for components with $\mathbf{D}_{ii} < 0$. Write

$$\mathbf{D} = \mathbf{D}^+ + \mathbf{D}^-,$$

where $\mathbf{D}^+$ is given by $\mathbf{D}^+_{ii} = \max\{\mathbf{D}_{ii}, 0\}$ and $\mathbf{D}^-$ by $\mathbf{D}^-_{ii} = \min\{\mathbf{D}_{ii}, 0\}$. The upwind method then reads as

$$\widetilde{\mathbf{U}}^{n+1}_j = \widetilde{\mathbf{U}}^n_j - \frac{k}{h}\mathbf{D}^+(\widetilde{\mathbf{U}}^n_j - \widetilde{\mathbf{U}}^n_{j-1}) - \frac{k}{h}\mathbf{D}^-(\widetilde{\mathbf{U}}^n_{j+1} - \widetilde{\mathbf{U}}^n_j).$$

Transforming back to $\mathbf{U}$-variables, gives with

$$\mathbf{A}^+ = \mathbf{V}\mathbf{D}^+\mathbf{V}^{-1}, \qquad \mathbf{A}^- = \mathbf{V}\mathbf{D}^-\mathbf{V}^{-1}$$

that

$$\mathbf{U}^{n+1}_j = \mathbf{U}^n_j - \frac{k}{h}\mathbf{A}^+(\mathbf{U}^n_j - \mathbf{U}^n_{j-1}) - \frac{k}{h}\mathbf{A}^-(\mathbf{U}^n_{j+1} - \mathbf{U}^n_j).$$

3.3.3 Splitting methods

In the following, we employ the splitting methods of Section 2.8.1 to obtain an approximation for the *2D*-advection equation.

Example 3.20 Consider the advection equation in two spatial dimensions

$$u_t + au_x + bu_y = 0 \qquad \text{in } \Omega = \mathbb{R}^2, \qquad u(\cdot, 0) = u_0(\cdot). \tag{3.27}$$

We use the splitting $Au = au_x$ and $Bu = bu_y$. For a function $v = v(x, y)$, the exact evolutions are given by

$$(E_A v)(x, y, t) = v(x - at, y), \qquad (E_B v)(x, y, t) = v(x, y - bt).$$

Consequently, one step of the Lie-splitting $E_B E_A$ is given by

$$v \mapsto v(x - at, y - bt).$$

This is even the exact solution, i.e., the Lie-splitting is exact! As the operators A, B commute, this could have been expected. ∎

We also apply an IMEX scheme to deal with an advection diffusion equation.

Example 3.21 Let $a > 0$ and consider the advection-diffusion equation

$$u_t = (u_{xx} + u_{yy}) - au_x.$$

Finite difference discretization (backward in space since $a > 0$) in space give the ODE system

$$\frac{d}{dt}u_{i,j} = \frac{1}{h^2}\left[D^+_x D^-_x + D^+_y D^-_y\right] u_{i,j} - \frac{a}{h}D^-_x u_{i,j}$$

The operator A corresponds to a spatial discretization of the Laplacian and the operator B to the transport term. Writing $u = \phi + \psi$ and employing implicit Euler for ϕ and explicit Euler for ψ leads to

$$\begin{aligned} \frac{1}{k}\left(\phi_{i,j}^{n+1} - \phi_{i,j}^{n}\right) &= \frac{1}{h^2}\left[D_x^+ D_x^- + D_y^+ D_y^-\right] u_{i,j}^{n+1}, \\ \frac{1}{k}\left(\psi_{i,j}^{n+1} - \psi_{i,j}^{n}\right) &= -\frac{a}{h} D_x^- u_{i,j}^{n}. \end{aligned}$$

Summing up, this gives

$$u_{i,j}^{n+1} = u_{ij}^{n} + \frac{k}{h^2}\left[D_x^+ D_x^- + D_y^+ D_y^-\right] u_{i,j}^{n+1} - \frac{ak}{h} D_x^- u_{i,j}^{n}. \tag{3.28}$$

As the transport term was treated explicitly, the CFL-condition $k|a|/h \leq 1$ needs to hold. A classical implicit Euler would lead to

$$u_{i,j}^{n+1} = u_{ij}^{n} + \frac{k}{h^2}\left[D_x^+ D_x^- + D_y^+ D_y^-\right] u_{i,j}^{n+1} - \frac{ak}{h} D_x^- u_{i,j}^{n+1} \tag{3.29}$$

without CFL condition. However, the linear system in (3.28) is SPD, while the linear system in (3.29) is non symmetric and thus more expensive to solve. ∎

3.4 von Neumann-analysis

The classical stability analysis is done in the norm

$$\|(V_i)\|_{\ell_h^2} := \left(\sum_i h|V_i|^2\right)^{1/2}.$$

Correspondingly, we denote by ℓ_h^2 the space of sequences with finite norm.
The reason for that lies in the fact that it is very convenient to employ Fourier analysis. We define

- The "Fourier-transform" of a grid function $(v_j)_{j\in\mathbb{Z}}$ (recall $x_j = jh$):

$$\widehat{v}(\xi) := (\mathcal{F}_h(v))(\xi) := h\sum_j e^{-\mathbf{i}\xi x_j} v_j, \qquad \xi \in [-\pi/h, \pi/h]$$

- the L_h^2-norm:

$$\|\widehat{v}\|_{L_h^2}^2 := \int_{-\pi/h}^{\pi/h} |\widehat{v}(\xi)|^2 \, d\xi$$

- The convolution of two sequences $u = (u_i)_i$, $v = (v_i)_i$

$$(u * v)_i := h\sum_j u_{i-j} v_j = h\sum_j u_j v_{i-j}$$

Theorem 3.22 *(i) (Parseval) $\mathcal{F}_h$ is an isomorphism $\ell_h^2 \to L_h^2$:*

$$\sqrt{2\pi}\|(v_i)_{i\in\mathbb{Z}}\|_{\ell_h^2} = \|\widehat{v}\|_{L_h^2}, \qquad \widehat{v} = \mathcal{F}_h((v_i)_{i\in\mathbb{Z}}).$$

Its inverse is given by

$$v_j = (\mathcal{F}_h^{-1}(\widehat{v}))_j = \frac{1}{2\pi}\int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x_j}\widehat{v}(\xi)\, d\xi.$$

(ii) For $(u_i)_{i\in\mathbb{Z}} \in \ell_h^2$ and $(v_i)_{i\in\mathbb{Z}} \in \ell_h^1$, there holds $u * v \in \ell_h^2$ and

$$\widehat{(u * v)}(\xi) = \widehat{u}(\xi)\widehat{v}(\xi)$$

(iii) (Translation) For $j_0 \in \mathbb{Z}$ there is $\mathcal{F}_h((v_{j+j_0})_{j\in\mathbb{Z}})(\xi) = e^{\mathbf{i}\xi x_{j_0}}\widehat{v}(\xi)$

(iv) (Modulation) For $\xi_0 \in \mathbb{R}$ there is $\mathcal{F}_h((e^{\mathbf{i}\xi_0 x_j}v_j)_{j\in\mathbb{Z}})(\xi) = \widehat{v}(\xi - \xi_0)$

(v) (Dilation) For $m \in \mathbb{Z}\setminus\{0\}$ there is $\mathcal{F}_h((v_{mj})_{j\in\mathbb{Z}})(\xi) = \widehat{v}(\xi/m)/|m|$

(vi) (Conjugation) $\mathcal{F}_h((\overline{v_j})_{j\in\mathbb{Z}})(\xi) = \overline{\widehat{v}(-\xi)}$

Proof: Exercise. □

Consider a single-step method with propagation operator E of the form

$$(Ev)_i = \sum_{\ell=-r}^{s} \alpha_\ell v_{i+\ell}$$

with coefficients α_ℓ. Then, the operator E is of convolution type

$$(Ev) = a * v, \qquad a_\ell = \frac{1}{h}\alpha_{-\ell}.$$

Correspondingly, we have

$$\widehat{(Ev)}(\xi) = \widehat{a}(\xi)\widehat{v}(\xi),$$

where $\widehat{a}$ is called *amplification factor*.

Exercise 3.23 There holds

$$\|E\|_{\ell_h^2} = \max_{\xi\in[-\pi/h,\pi/h]} |\widehat{a}(\xi)|$$

Consequently, the stability analysis is reduced to the calculation of $\widehat{a}$. A method satisfies the *von-Neumann-stability condition*, if the corresponding amplification factor $\widehat{a}$ satisfies

$$|\widehat{a}(\xi)| \le 1 + Ck \qquad \forall \xi \in [-\pi/h, \pi/h], \tag{3.30}$$

where $C > 0$ is a constant independent of k.

Example 3.24 (Upwind) The upwind method for the advection equation with $a = -1$ and $g \equiv 0$ is given by

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \lambda(v_{j+1}^n - v_j^n), \qquad \lambda = \frac{k}{h}.$$

This can be written as $Ev^n = a * v^n$ with

$$a_j = \frac{1}{h}\lambda\delta_{-1,j} + \frac{1}{h}(1-\lambda)\delta_{j,0}, \qquad \delta_{n,m} = \text{Kronecker } \delta.$$

Thus, the Fourier transform can be calculated as

$$\widehat{a}(\xi) = h(e^{-\mathbf{i}\xi x_{-1}}a_{-1} + e^{-\mathbf{i}\xi x_0}a_0) = \lambda e^{\mathbf{i}\xi h} + (1-\lambda),$$

and, for $0 < \lambda \le 1$, there holds

$$|\widehat{a}(\xi)| \le 1 \qquad \forall \xi \in [-\pi/h, \pi/h].$$

This means that the method satisfies the von-Neumann stability condition (3.30). ∎

In practice, the calculation of $g(\xi) := \widehat{a}(\xi)$ is oftentimes shortened by using a "calculation rule": make the ansatz $v_j^n = g^n e^{\mathbf{i}\xi jh}$ and insert it into the method, this gives a formula for $g(\xi)$.
The reason for including the term Ck in the stability condition (3.30) comes from the desire to also treat equations with lower order terms.

Operator	Symbol
Forward diff. D_+: $u_{i+1} - u_i$	$e^{\mathbf{i}\xi} - 1$
Backward diff. D_-: $u_i - u_{i-1}$	$1 - e^{-\mathbf{i}\xi}$
symmetric diff. D_0: $u_{i+1} - u_{i-1}$	$2\mathbf{i}\sin\xi$
δ defined by: $u_{i+1/2} - u_{i-1/2}$	$2\mathbf{i}\sin(\xi/2)$
$D_+ \circ D_-$	$-4\sin^2(\xi/2)$

Figure 3.7: Fourier symbols of some difference operators.

Exercise 3.25 Let $c : \mathbb{R} \to \mathbb{R}$ be continuous with $\|c\|_{L^\infty} < \infty$. Consider the equation

$$u_t - u_x + c(x)u = 0$$

and employ the upwind method

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \lambda(v_{j+1}^n - v_j^n) + kc(x_j)v_j, \qquad \lambda = \frac{k}{h}. \tag{3.31}$$

Let E_0 be the evolution corresponding to $c \equiv 0$. Example 3.24 shows that $\|E_0\|_{\ell_h^2} \leq 1$.
Show that $\|E\|_{\ell_h^2} \leq \|E_0\|_{\ell_h^2} + k\|c\|_{L^\infty}$ holds and deduce that E satisfies the von-Neumann condition. ∎

Exercise 3.26 Consider a discretization of the heat equation $u_t - u_{xx} = 0$ with explicit Euler in time and symmetric differences in space given by

$$u_i^{n+1} - u_i^n = \sigma\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right), \qquad \sigma := \frac{k}{h^2}$$

Show that the amplification factor is $g(\xi) = 1 - 4\sigma\sin^2(\xi h/2)$. What can you say about stability of the method in dependence on k and h? ∎

Remark 3.27

- The von-Neumann analysis can also be applied for systems (and thus also for multi step methods) – see exercises.
- In general, it only provides *necessary* conditions for stability, but not sufficient conditions. In practice, however, it does provide a fairly good idea of what the CFL condition is.
- For problems with non-constant coefficients, one typically proceeds as follows: In a first step one determines the range of values of the coefficients and then perform a von Neumann analysis for the equation with frozen coefficients ("freezing of coefficients"). In principle, this can also be done for nonlinear equations. In this way, of course, one does not obtain sharp estimates for the CFL condition, but often useful indicators.

3.4.1 The leap frog method

We now consider a very common *2-step* method, the so called leap frog method.

Note: Multi-step methods can be interpreted as vector single-step methods (exercise!) and thus be analyzed similarly to single-step methods. However, oftentimes a direct analysis is easier.

For the advection equation the *leap frog method* replaces the derivatives by central difference quotients, i.e., it reads as

$$\frac{u_i^{n+1} - u_i^{n-1}}{2k} + a\frac{u_{i+1}^n - u_{i-1}^n}{2h} = 0. \tag{3.32}$$

Note that no value u_i^n appears in the approximation of the time derivative, this has been "leaped over", which explains the name of the method. Leap frog also is an explicit method.

Remark 3.28

- As a two step method, leap frog needs two starting values u_0^0 and u_i^1, $i \in \mathbb{Z}$. The values u_i^1 can be computed with a single step method.

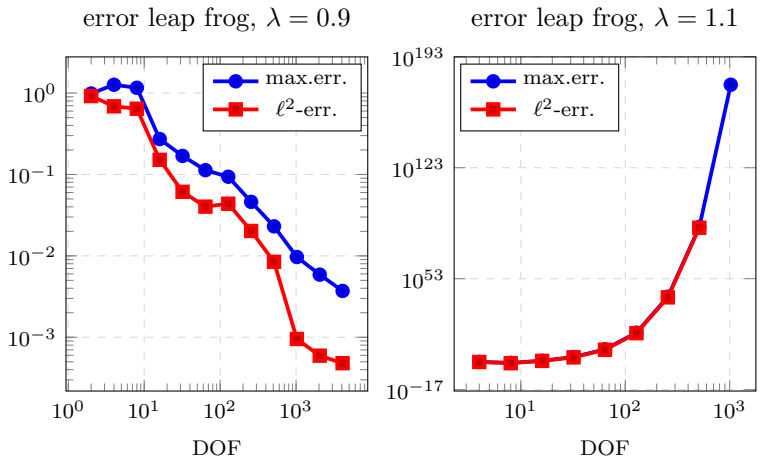

Figure 3.8: influence of CFL-condition for the leap frog method.

- Defining

$$U_i^n := \begin{pmatrix} u_i^n \\ u_i^{n-1} \end{pmatrix}$$

the method (3.32) can be formulated as a single-step method.

■

We now make a von-Neumann stability analysis: the Fourier transform $\widehat{u}^n(\xi) := \mathcal{F}_h((u_i^n)_i)(\xi)$ satisfies

$$\widehat{u}^{n+1}(\xi) + 2\mathbf{i}a\lambda \sin(\xi h)\widehat{u}^n(\xi) - \widehat{u}^{n-1}(\xi) = 0, \qquad \lambda = \frac{k}{h}. \tag{3.33}$$

For fixed ξ, h, this is a recurrence relation in n that can be solved.

Exercise 3.29 Let α, $\beta \in \mathbb{C}$. Let x_1, x_2 be solutions of $0 = x^2 + \alpha x + \beta$. Consider all sequences $(v_n)_n$ with

$$v_{n+1} + \alpha v_n + \beta v_{n-1} = 0, \qquad \forall n \geq 1.$$

Then, there holds

(a) The space of sequences with this property is a 2-dimensional vector space.
(b) If $x_1 \neq x_2$, then the sequences $(x_1^n)_n$ and $(x_2^n)_n$ are a basis for this space.
(c) If $x_1 = x_2$, then the sequences $(x_1^n)_n$ and $(nx_1^n)_n$ are a basis for this space.

Now, let $g_+(\xi)$ and $g_-(\xi)$ be solutions to

$$g^2 + 2\mathbf{i}a\lambda \sin(\xi h)g - 1 = 0,$$

i.e.,

$$g_\pm(\xi h) = -\mathbf{i}a\lambda \sin(\xi h) \pm \sqrt{1 - (a\lambda)^2 \sin^2(\xi h)}.$$

Consequently, we can write $\widehat{u}^n$ as

$$\widehat{u}^n(\xi) = \gamma(\xi)(g_+(\xi h))^n + \delta(\xi)(g_-(\xi))^n \qquad \text{falls } g_+(\xi h) \neq g_-(\xi h),$$
$$\widehat{u}^n(\xi) = \gamma(\xi)(g_+(\xi h))^n + \delta(\xi)n(g_+(\xi))^n \qquad \text{falls } g_+(\xi h) = g_-(\xi h).$$

The functions $\gamma(\xi)$, $\delta(\xi)$ are determined by the starting values $\widehat{u}^0(\xi)$, $\widehat{u}^1(\xi)$. Stability of the discretization requires

$$|g_+(\xi h)| \leq 1 \quad \text{and} \quad |g_-(\xi h)| \leq 1 \qquad \forall \xi \in (-\pi/h, \pi/h), \tag{3.34}$$
$$\text{if } g_+(\xi) = g_-(\xi) \text{ then} |g_+(\xi h)| < 1. \tag{3.35}$$

We now observe:

- $|a\lambda| > 1$ implies $|g_-(-\pi/(2h))| > 1$, which means that the CFL-condition $|a\lambda| \leq 1$ has to hold.
- $|a\lambda| \leq 1$ implies $|g_+|^2 = |g_-|^2 = 1 - (a\lambda)^2 \sin^2(\xi h) + (a\lambda \sin(\xi h))^2 = 1$, i.e., the method is stable apart from the case $g_+(\xi h) = g_-(\xi h)$. In this case $1 = |a\lambda \sin(\xi h)|$, i.e. $|a\lambda| = 1$ and $\xi = \pm\pi/(2h)$. Then, $g_+ = g_- = \pm\mathbf{i}$, i.e. condition (3.35) is violated.

Summary: If $|a\lambda| < 1$, then leap frog is stable, if $|a\lambda| = 1$, it is (weakly) unstable.

Leap frog for the wave equation

The leap frog method is also a popular time-stepping method for the wave equation. Applied to the semi-discrete system

$$\mathbf{M}\mathbf{u}_{tt} + \mathbf{A}\mathbf{u} = \mathbf{f}$$

this reads as

$$\mathbf{M}\frac{\mathbf{u}^{n+1} - 2\mathbf{u}^n + \mathbf{u}^{n-1}}{k^2} + \mathbf{A}\mathbf{u}^n = \mathbf{f}^n.$$

For the leap frog method there holds energy conservation, which is desirable for the wave equation.

Exercise 3.30 Define the discrete energy

$$E^{n+1/2} := \langle (\mathbf{I} - \frac{k^2}{4}\mathbf{K})\frac{\mathbf{z}^{n+1} - \mathbf{z}^n}{k}, \frac{\mathbf{z}^{n+1} - \mathbf{z}^n}{k}\rangle + \langle \mathbf{K}\frac{\mathbf{z}^{n+1} + \mathbf{z}^n}{2}, \frac{\mathbf{z}^{n+1} + \mathbf{z}^n}{2}\rangle$$

and show that $E^{n+1/2} = E^{n-1/2}$.

Note that, in contrast to the Crank-Nicolson method, the leap frog method is explicit (thus needs CFL condition!), where only a linear system with matrix $\mathbf{M}$ has to be solved. Again, employing mass lumping this can be done very efficiently.

3.5 Dissipative methods

3.5.1 Preliminary view

Starting point:

1. For linear equations with constant coefficients, von-Neumann analysis provides a simple tool to derive a stability analysis.

2. For *systems*

   $$\mathbf{U}^{n+1} = E\mathbf{U}^m + k\mathbf{G}^n$$

   one can analogously look at the Fourier transform $\widehat{\mathbf{E}}(\xi)$. For simplicity, one analyzes the spectral radius $\rho(\widehat{\mathbf{E}})$ for $\xi \in (-\pi/h, \pi/h)$. Thus, the method satisfies the von-Neumann condition, if $\rho(\widehat{\mathbf{E}}(\xi)) \leq 1 + Ck$ for all $\xi \in (-\pi/h, \pi/h)$.

Problems: variable coefficients? Nonlinear equations?

Standard procedure: von-Neumann analysis for *frozen coefficients*, i.e., one does von-Neumann analysis for all (expected) values of the coefficients. Oftentimes, this gives good indicators for stability or step size restrictions. For so called dissipative methods and certain problem classes (e.g. parabolic problems), the method of freezing of coefficients is indeed good enough.

However, the following example shows that the method also may fail.

Example 3.31 (Zabusky-Kruskal-method for KdV equation) We consider the KdV equation

$$u_t + (\alpha u^2 + \nu u_{xx})_x = 0$$

with a leap-frog type discretization

$$u_j^{n+1} = u_j^{n-1} - \frac{2\alpha\mu}{3}\left(u_{j-1}^n + u_j^n + u_{j+1}^n\right)\left(u_{j+1}^n - u_{j-1}^n\right) - \frac{\nu\mu}{h^2}\left(u_{j+2}^n - 2u_{j+1}^n + 2u_{j-1}^n - u_{j-2}^n\right), \qquad \mu = \frac{k}{h}.$$

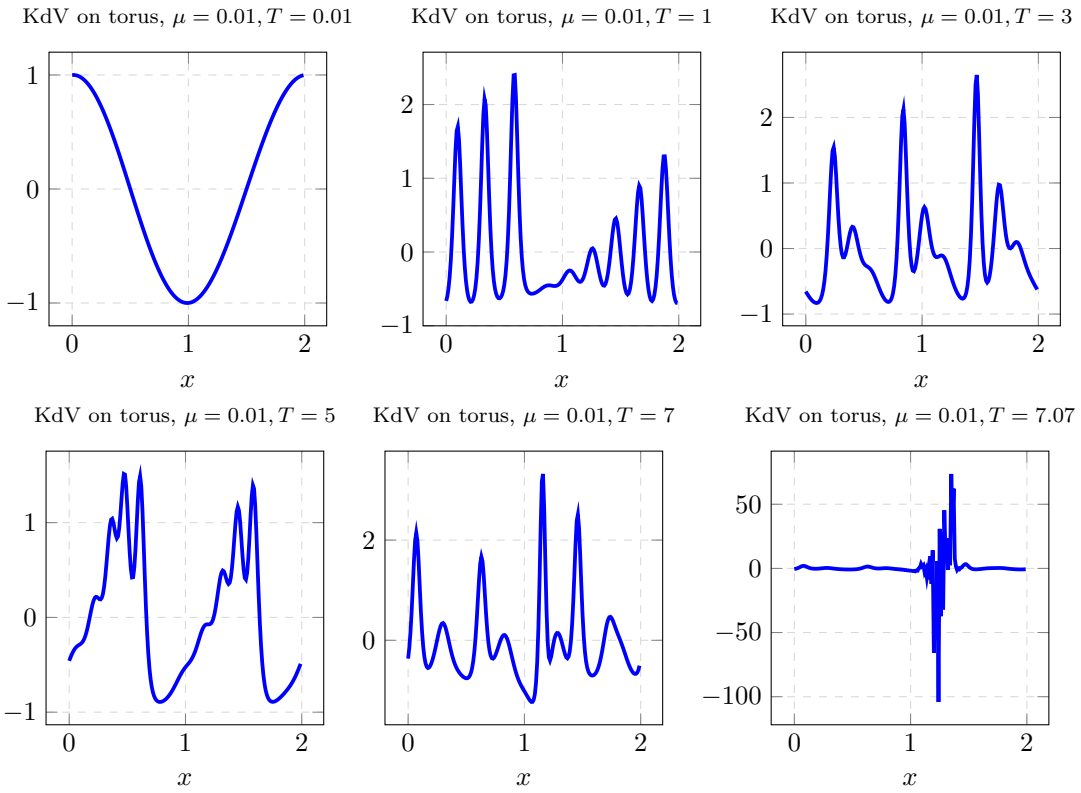

Figure 3.9: Zabusky-Kruskal-method, solution plot for the KdV equation.

Using von-Neumann analysis after freezing of coefficients gives

$$k < \frac{h}{4|\nu|/h^2 + 2|\alpha u_{max}|}.$$

We now consider the method with

- $\nu = 0.022^2, \qquad \alpha = 0.5, \qquad u_0(x) = \cos(\pi x), \qquad u(0,t) = u(2,t).$
- One can show that $|u| \leq 5$.
- $\Longrightarrow$ One expects that $h = 0.01$ and $k < 0.004$ produce a stable method.
- In the numerical example in Fig. 3.9, we thus take $h = 0.01$ and $k = 0.0001$.

The numerical simulation in Fig. 3.9 shows that this (conservative!) step size choice is not good enough! At time $T > 21/\pi$ a blow-up happens, even tough the exact solution exists for all times. ∎

The failure of the Zabusky-Kruskal-method [16][1] is caused by an instability in the sense that high-frequency components of the solution (i.e. with large ξ in Fourier image) are amplified by the non-linearity.
Thus, a common way is to use methods which induce (some) dissipation.

3.5.2 Dissipative methods

Definition 3.32 *A method with propagation operator* $\mathbf{E}$ *is called* dissipative *of order* $2r$*, if*

$$\rho(\widehat{\mathbf{E}}(\xi)) \leq (1 - \delta|\xi h|^{2r})(1 + Ck)$$

for constants C, $\delta > 0$ *independent of* k, h.

[1] in this paper the notion of solitons was introduced

Example 3.33 (Lax-Wendroff) For the advection equation with $a = -1$, the method is given by

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \frac{1}{2}\lambda(v_{j+1}^n - v_{j-1}^n) + \frac{1}{2}\lambda^2(v_{j+1}^n - 2v_j^n + v_{j-1}^n), \qquad \lambda = k/h.$$

Using the ansatz $v_j^n = g^n e^{\mathbf{i}\xi jh}$ and division by $g^n e^{\mathbf{i}\xi jh}$ leads to

$$g(\xi) = 1 + \frac{1}{2}\lambda(e^{\mathbf{i}\xi h} - e^{-\mathbf{i}\xi h}) + \frac{1}{2}\lambda^2(e^{\mathbf{i}\xi h} - 2 + e^{-\mathbf{i}\xi h}) = 1 + \mathbf{i}\lambda \sin \xi h - 2\lambda^2 \sin^2 \frac{\xi h}{2}.$$

An elementary calculation shows that for $\lambda \in (0, 1]$ the condition

$$\sup_{\xi\in[-\pi/h,\pi/h]} |g(\xi)|^2 = \sup_{\xi\in[-\pi/h,\pi/h]} \left(1 - 4\lambda^2(1-\lambda^2)\sin^4(\xi h/2)\right) \overset{\lambda\in(0,1]}{\leq} 1$$

is satisfied, i.e., the method satisfies the von-Neumann condition (3.30).
Moreover, the amplification factor also shows that the method is dissipative of order $2r = 4$, *if* $\lambda = k/h < 1$ (use $\sqrt{1-x} = 1 - 1/2x + O(x^2)$ for small x). There holds $g(\xi) \approx 1$ for $\xi \approx 0$, but for $|\xi| \approx \pi/h$ there even $|g(\xi)| \approx 1 - 4\lambda^2(1-\lambda^2) < 1$, if $\lambda < 1$. In other words: While the low frequency terms (due to consistency!) are neither amplified nor damped, the high frequency solution components (and therefore also the error contributions) are damped. ∎

For parabolic equations dissipative methods are very natural as the continuous problem has the same property.

Example 3.34 Consider the explicit Euler discretization for the heat equation from Exercise 3.26, which has the amplification factor $g(\xi)$ given by $g(\xi) = 1 - 4\sigma \sin^2(\xi h/2)$. Stability ($|g(\xi)| \leq 1$) requires $\sigma \leq 1/2$. For $\sigma \leq 1/2$ the method is dissipative of order 2.
Exercise: For the implicit Euler, the amplification factor is $g(\xi) = 1/(1 + 4\sigma \sin^2(\xi h/2))$. The method is therefore stable and dissipative of order 2. ∎

In the easiest case, dissipativity and consistency gives stability.

Theorem 3.35 (Kreiss) *Consider strictly hyperbolic systems*

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0,$$

i.e., the constant matrix $\mathbf{A}$ *has pairwise distinct real eigenvalues. Then, an (explicit) difference method is stable, if it is consistent and dissipative of order* $2r > 0$.

Proof: See [2, Thm. 5.2] and e.g. [7, Thm. 6.5.2]. □

Remark 3.36 The interplay of consistency and dissipativity has already been seen for parabolic equations in Section 2.5.4 in the analysis of the function F_n: for "small" $z = \lambda_n k$ we used the consistency, for large $z = \lambda_n k$ we used the L-stability of the one-step method. Similar ideas are the basis of Theorem 3.35: for consistent methods, one expects that the stability behaviour for small ξ is obtained from the continuous problem, while the stability behaviour for large ξ is an additional property.

As theorem 3.35 shows, dissipativity of a method can be a useful property. In fact many methods (also for hyperbolic problems) have some dissipation – the Lax-Wendroff method is an example. While, for parabolic problems such as the heat equation, dissipation is a property of the continuous problem, for hyperbolic problems, it is usually not. Therefore, one usually tries to keep the dissipation of the method as low as possible.

Derivation of Lax-Wendroff

As previously done, e.g. for parabolic problems, many methods are derived by doing a semi-discretization in space (FEM,FD) and then employing a time stepping method. The Lax-Wendroff method (here for the advection equation) is derived differently: With the difference operators D_+, D_-, D_0 (e.g. $D_+u(x) = u(x+h) - u(x)$, $D_0u(x) = u(x+h) - u(x-h)$) Taylor expansion gives

$$u(t+k,x) = u + ku_t + \frac{k^2}{2}u_{tt} + \mathcal{O}(k^3).$$

Employing the equation $u_t + au_x = 0$ gives formally

$$u_t = -au_x \qquad \Longrightarrow \qquad u_{tt} = a^2 u_{xx}. \tag{3.36}$$

Thus, we can replace time derivatives by spatial derivatives and afterwards approximate those by difference quotients (which leads to the additional term $\mathcal{O}(h^2)$):

$$u(t+k,x) = u - \frac{k}{2h}aD_0u + \frac{k^2}{2h^2}a^2D_+D_-u + k\mathcal{O}(k^2+h^2).$$

This shows that the method

$$u_i^{n+1} = u_i^n - \frac{\lambda}{2}\left(u_{i+1}^n - u_{i-1}^n\right) + \frac{\lambda^2}{2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^m\right) = 0, \qquad \lambda = \frac{ka}{h}$$

is consistent of order 2 (or more precisely: $(2,2)$).

Modified equation

Fourier analysis (i.e., von Neumann analysis) is a way to understand the behaviour of discretizations. Another way to understand the qualitative behaviour of methods is to interpret them as a "better" approximation to another *modified equation* and to stipulate that the qualitative behaviour of this continuous equation describes the numerical method.
We illustrate this with three classical examples:

Example 3.37 (Modified equation for updwind) We consider $u_t + au_x = 0$ with $a < 0$. The upwind method is given by

$$u_i^{n+1} = u_i^n - a\lambda(u_{i+1}^n - u_i^n), \qquad \lambda = \frac{k}{h}.$$

Taylor expansion around (t,x) gives

$$\begin{aligned}
u(t+k,x) &= u + ku_t + \frac{k^2}{2}u_{tt} + \frac{k^3}{6}u_{ttt} + \cdots, \\
u(t,x+h) &= u + hu_x + \frac{h^2}{2}u_{xx} + \frac{h^3}{6}u_{xxx} + \cdots, \\
\tau &= \frac{u(t+k,x)-u}{k} + a\frac{u(t,x+h)-u}{h} = \left[u_t + \frac{k}{2}u_{tt} + \frac{k^2}{6}u_{ttt} + \cdots\right] + a\left[u_x + \frac{h}{2}u_{xx} + \frac{h^2}{6}u_{xxx} + \cdots\right] \\
&= u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} + \mathcal{O}(k^2+h^2)
\end{aligned}$$

Exploiting $u_t + au_x = 0$, we obtain that the upwind method has consistency order $(1,1)$. However, if we assume that u solves the equation

$$u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} = 0, \tag{3.37}$$

then we obtain consistency order $(2,2)$. Equation (3.37) is called the modified equation. Typically, this equation is reformulated by replacing time derivatives by spatial derivatives: from (3.37), one obtains (formally) by differentiation w.r.t. t and x:

$$u_{tt} + au_{xt} = \mathcal{O}(k+h), \qquad u_{xt} + au_{xx} = \mathcal{O}(k+h), \tag{3.38}$$

such that we can write (3.37) (dropping the terms of order $\mathcal{O}(h^2+k^2)$ and using $kh \le k^2+h^2$)

$$0 = u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} \overset{(3.38)}{=} u_t + au_x + \frac{1}{2}\left[-kau_{xt} + hau_{xx} + k\mathcal{O}(k+h)\right]$$
$$\overset{(3.38)}{=} u_t + au_x + \frac{1}{2}\left[ka^2u_{xx} + hau_{xx} + k\mathcal{O}(k+h)\right] = u_t + au_x + \frac{1}{2}\left[ka^2 + ha\right]u_{xx} + \mathcal{O}(k^2+h^2).$$

Thus, the above stated upwind method approximates the equation

$$u_t + au_x - \nu u_{xx} = 0, \qquad \nu := -\frac{1}{2}\left(ka^2 + ah\right) \overset{a\le 0}{=} \frac{h}{2k/h}\left(\frac{k}{h}|a| - \frac{k^2}{h^2}a^2\right) \tag{3.39}$$

with consistency order $(2,2)$. We note that $\nu > 0$, if the CFL-condition holds. One can think that the upwind method approximates the advection equation, but also with higher accuracy the heat equation (3.39) with (small) diffusion. As the heat equation is dissipative, one expects that the upwind method is dissipative as well. ∎

Exercise 3.38 Show that for the Lax-Friedrichs method the modified equation is

$$u_t + au_x + \nu u_{xx} = 0, \qquad \nu = \frac{h}{2\lambda}\left(1 - \lambda^2 a^2\right), \qquad \lambda = \frac{k}{h}.$$

In particular, the diffusion constant ν is bigger for Lax-Friedrichs than for upwind.

Exercise 3.39 Show that for the Lax-Wendroff method the modified equation is

$$u_t + au_x + \frac{ah^2}{6}\left(1 - \frac{k^2}{h^2}a^2\right)u_{xxx} = 0.$$

Example 3.40 We consider the advection equation with periodic boundary conditions

$$u_t - u_x = 0 \quad \text{ on } (0,1)\times\mathbb{R}^+, \qquad u(0,t) = u(1,t) \quad \forall t > 0. \tag{3.40}$$

In order to determine the influence of the high-frequency components of the initial condition, we take $u_0(x) = \chi_{[0.25,0.75]}(x)$ (and consequently the exact solution remains a piecewise constant function).
In Fig. 3.10 we plot the results of different numerical methods at $T = 1$ and for $k/h = 0.5$.
The modified equation for the upwind and Lax-Friedrichs methods are parabolic equations with quite a lot dissipation. This can be seen in Fig. 3.10.
The modified equation of Lax-Wendroff is an equation of third order, for which dispersion (see remark below) is important, which explains the oscillations in the numerical solution.

∎

Remark 3.41 (Dispersion) The solution u of

$$u_t + au_x = 0, \qquad u(0,x) = u_0(x)$$

is

$$u(t,x) = u_0(x-at) = \frac{1}{\sqrt{2\pi}}\int_\xi \widehat{u}_0(\xi)e^{\mathrm{i}\xi(x-at)}\,d\xi.$$

This representation comes directly from the Fourier inversion formula (we scale the Fourier transform and its inverse with $\frac{1}{\sqrt{2\pi}}$) or after Fourier transformation of the PDE and solution afterwards.
We interpret this as follows: The Fourier component $\widehat{u}_0(\xi)e^{\mathrm{i}\xi x}$ corresponding to the frequency ξ expands with speed a (to the right). The propagation speed a is *independent* of ξ.
Let now $\xi \mapsto a(\xi)$ be a function of ξ. Then, $v(x,t) = \frac{1}{\sqrt{2\pi}}\int_\xi \widehat{u}_0(\xi)e^{\mathrm{i}\xi(x-a(\xi)t)}\,d\xi$ is a function with the following property: the Fourier component of $v(\cdot,0)$ corresponding to the frequency ξ moves with speed $a(\xi)$.
More general, one speaks of *dispersion*, if the Fourier components expand with ξ-dependent speed.

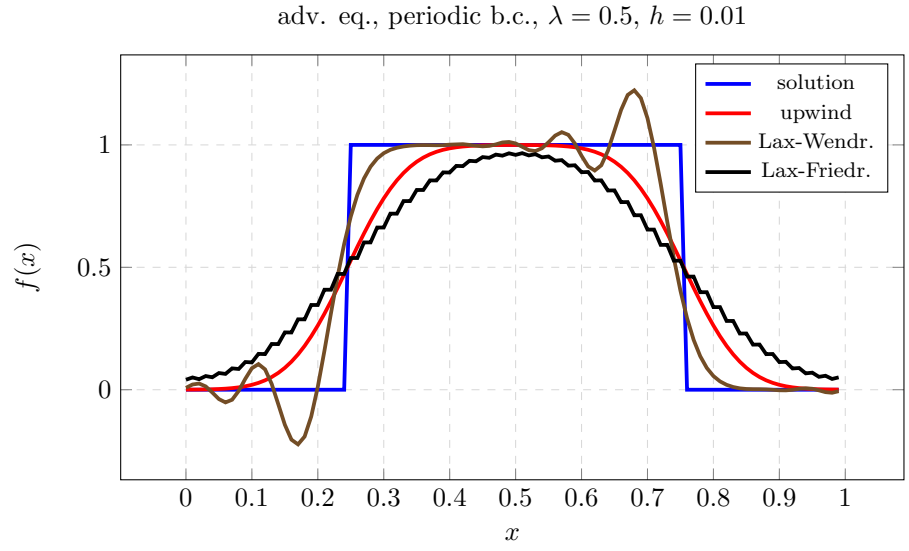

Figure 3.10: Upwind, LW, LF for advection equation with discontinuous initial data.

We solve the equation

$$u_t + au_x + \nu u_{xxx} = 0, \qquad u(0,x) = u_0(x)$$

Fourier transformation (in x) gives the ODE

$$\widehat{u}(t,\xi) + a\mathbf{i}\xi\widehat{u}(t,\xi) + \nu(\mathbf{i}\xi)^3\widehat{u}(t,\xi) = 0, \qquad \widehat{u}(0,\xi) = \widehat{u}_0(\xi)$$

with solution

$$\widehat{u}(t,\xi) = \exp(-(a\mathbf{i}\xi + \nu(\mathbf{i}\xi)^3)t)\widehat{u}_0(\xi).$$

Fourier inversion leads to

$$u(t,x) = \frac{1}{\sqrt{2\pi}}\int_\xi e^{\mathbf{i}\xi x}e^{-(a\mathbf{i}\xi+\nu(\mathbf{i}\xi)^3)t}\widehat{u}_0(\xi)d\xi.$$

In particular, we observe that the expansion of different Fourier components of the initial values of u_0 have different speed. This is observed in the Lax-Wendroff method in Fig. 3.10 and explains the oscillations. ∎

3.6 Space-time-DG

We consider a more general form of (3.4):

$$u_t + \mathbf{b}(x,t)\cdot\nabla u + c(x,t)u = g. \tag{3.41}$$

In contrast to parabolic equations, the t-variable in (3.41) does not have a special role. Rewriting it as $(t,x) = (x_0, x_1, \ldots, x_d)$, we can also write this problem as

$$\mathbf{b}(x)\cdot\nabla u + c(x)u = g \quad \text{ on } \Omega, \tag{3.42}$$

where now $\Omega \subset \mathbb{R}^{d+1}$.
As a first order equation, one can not impose boundary conditions on the entire boundary $\partial\Omega$. We define the inflow, outflow and characteristic boundary by

$$\Gamma^- := \{x \in \partial\Omega\colon \mathbf{b}(x)\cdot n(x) < 0\}, \tag{3.43}$$
$$\Gamma^+ := \{x \in \partial\Omega\colon \mathbf{b}(x)\cdot n(x) > 0\}, \tag{3.44}$$
$$\Gamma^= := \{x \in \partial\Omega\colon \mathbf{b}(x)\cdot n(x) = 0\}. \tag{3.45}$$

Here, $n(x)$ denotes the (outer) normal vector in the point $x \in \partial\Omega$.
For the equation (3.42), we can only impose a boundary condition on Γ^- or Γ^+. Therefore, we consider the boundary value problem

$$\begin{aligned} \mathbf{b}(x) \cdot \nabla u + c(x)u &= g \quad \text{on } \Omega, && \text{(3.46a)} \\ u &= 0 \quad \text{on } \Gamma^-. && \text{(3.46b)} \end{aligned}$$

Example 3.42 Consider the problem

$$-u' + u = g \quad \text{on } (0,1),$$

and argue that one can not require $u(0) = 0 = u(1)$ in a sensible way, only $u(0) = 0$ *or* $u(1) = 0$. ∎

Our goal is a numerical method for (3.46). Let $\mathcal{T}$ be a mesh that resolves Γ^-, i.e., every edge e of an element in $\mathcal{T}$ either satisfies $e \subset \Gamma^-$ or $e \subset \partial\Omega \setminus \Gamma^-$. Moreover, for $K \in \mathcal{T}$, we denote by n_K its outer normal vector.

For motivation, we assume that the solution u of (3.46) is sufficiently smooth. Let v be a piecewise smooth test function, i.e., $v|_K$ is smooth for all $K \in \mathcal{T}$. Multiplication of (3.46a) with v, integration over K and integration by parts give

$$\int_K gv = \int_K (cu + \mathbf{b} \cdot \nabla u)v = \int_K u(cv - u\nabla \cdot (\mathbf{b}v)) + \int_{\partial K} (\mathbf{b} \cdot n_K)uv.$$

Summation over all elements $K \in \mathcal{T}$ then leads to

$$\sum_{K \in \mathcal{T}} \int_K u\,(cv - \nabla \cdot (\mathbf{b}v)) + \int_{\partial K} (\mathbf{b} \cdot n_K)uv = \sum_{K \in \mathcal{T}} \int_K gv.$$

We now introduce the notion of *flux* on the boundary of the element K:

$$f_K = (\mathbf{b} \cdot n_K)u.$$

It will be convenient to denote the common edge/face (we will use the word edge for both in the following, even tough for $d > 2$ this means a hyperplane) between two elements K and K' by

$$K|K',$$

where the order of the elements fixes an orientation.
Moreover, we define the neighbours of an element by

$$\mathcal{N}(K) := \{K' \in \mathcal{T} \mid \ K \text{ and } K' \text{ share a manifold with co-dimension } 1\}.$$

For the discretization, we assume that u is only piecewise smooth, i.e., in the space

$$\mathcal{S}^{p,0}(\mathcal{T}) := \{u \in L^2(\Omega) \colon u|_K \in \mathcal{P}_p \qquad \forall K \in \mathcal{T}\}.$$

for given $p \in \mathbb{N}_0$. For a numerical realization, it is convenient to also take test functions v from this space. As for the DG for parabolic equations, this leads to the fact that there is no coupling between $u|_K$ and $u|_{K'}$ for neighbouring elements K and K'.
We realize this coupling by fixing a so called *numerical flux* on the common edge $K|K'$, where two approximations $u|_K$ and $u|_{K'}$ of the exact solution are available. Thus, we replace the flux f_K by some $\widehat{f}_{K|K'}$. A reasonable choice would be, e.g.,

$$\widehat{f}_{K|K'} = (\mathbf{b} \cdot n_K)\widehat{u}_{K|K'},$$

where the choice of $\widehat{u}$ has many possibilities. Plausible choices seem to be e.g.

- $\widehat{u}_{K|K'} = \frac{1}{2}(u|_K + u|_{K'})|_e$
- $\widehat{u}_{K|K'} = u|_K$

- $\widehat{u}_{K|K'} = u|_{K'}$

For boundary edges $e \subset \Gamma^-$, one would require[2]

$$\widehat{u}|_e = 0 \qquad \forall e \subset \Gamma^-.$$

In the present case, the numerical flux $\widehat{f}_{K|K'}$ is uniquely defined by the choice of $\widehat{u}$ on the edges. We have thus derived the following numerical method:

$$\begin{aligned} &\text{Find } u \in S^{p,0}(\mathcal{T}) \text{ s.t.} \qquad (3.47)\\ &B_{DG}^{Trans}(u,v) := \sum_{K\in\mathcal{T}} \int_K u\,(cv + \nabla\cdot(\mathbf{b}v)) + \int_{\partial K} (\mathbf{b}\cdot n_K)\widehat{u}v = \int_\Omega gv =: l(v) \qquad \forall v \in S^{p,0}(\mathcal{T}). \end{aligned}$$

Usually, this formulation is integrated by parts back and we obtain

$$\begin{aligned} &\text{Find } u \in S^{p,0}(\mathcal{T}) \text{ s.t.} \qquad (3.48)\\ &B_{DG}^{Trans}(u,v) = \sum_{K\in\mathcal{T}} \int_K cuv + \mathbf{b}\cdot\nabla uv + \int_{\partial K} (\mathbf{b}\cdot n_K)(\widehat{u}-u)v = \int_\Omega gv \qquad \forall v \in S^{p,0}(\mathcal{T}) \end{aligned}$$

The choice of numerical flux is essential for the quality of the numerical method. We illustrate this on the following example.

Example 3.43 Let $g(x) = 1 + x$ and consider

$$u' + u = g \quad \text{ on } (0,1), \qquad u(0) = 0.$$

Then, the exact solution is $u(x) = x$. Let $x_i = ih$, $i = 0,\dots,N$ and $K_i = (x_{i-1}, x_i)$. For the choice $p = 0$ (i.e. piecewise constants), we write for $u|_{K_i} = u_i$ and also $v|_{K_i} = v_i$ for the piecewise constant test functions. Then the DG-bilinear form reads as

$$\begin{aligned} B_{DG}^{Trans}(u,v) &= \sum_{i=1}^N \int_{K_i} (u'+u)v + (\widehat{u}(x_i) - u_i)v_i - (\widehat{u}(x_{i-1}) - u_i)v_i \\ &= \sum_{i=1}^N \int_{K_i} uv + (\widehat{u}(x_i) - \widehat{u}(x_{i-1}))v_i \\ &= \sum_{i=1}^N \left[hu_i + (\widehat{u}(x_i) - \widehat{u}(x_{i-1}))\right] v_i, \end{aligned}$$

which leads to the linear system

$$\int_{K_i} g = hu_i + \widehat{u}(x_i) - \widehat{u}(x_{i-1}), \qquad i = 1,\dots,N.$$

We consider two choices of $\widehat{u}(x_i)$:

- *upwind flux:* $\widehat{u}(x_j) = u_j$ for $1 \le j \le N$ and $\widehat{u}(x_0) = 0$ (and $\widehat{u}(x_N) = u_N$);
- *central flux:* $\widehat{u}(x_j) = \frac{1}{2}(u_j + u_{j+1})$ for $1 \le j \le N-1$ and $\widehat{u}(x_0) = 0$ and $\widehat{u}(x_N) = u_N$.

In Fig. 3.11, we observe that the choice of the *upwind flux* gives a good approximation to the true solution as well as (expected!) convergence $\mathcal{O}(h)$, while the choice of *central flux* does not lead to a convergent method. ∎

Key for the choice of numerical flux is that it is dependent on the sign of $\mathbf{b}\cdot n_K$ in (3.49). As we will see below, the right choice of numerical flux will lead to a numerical method with good stability properties.

The *upwind flux* $(\mathbf{b}\cdot n_K)\widehat{u}$ is defined by the following choice of $\widehat{u}$ on every edge:

[2] this choice seems obvious, but is not necessary!

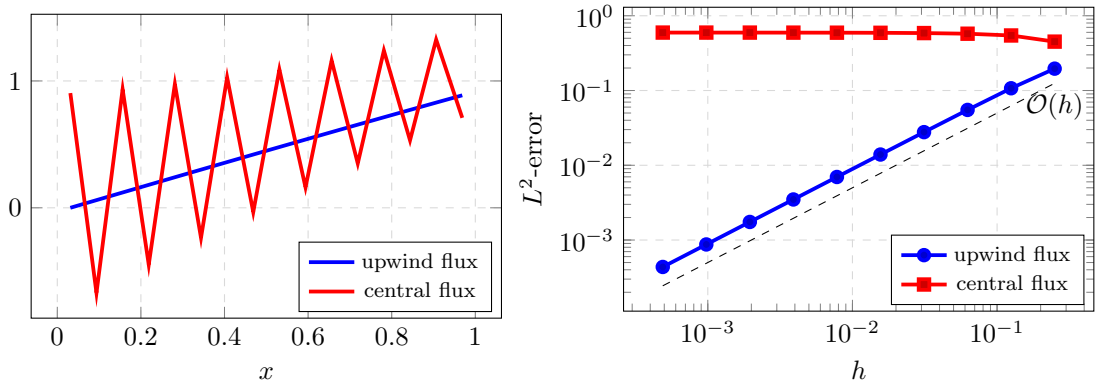

Figure 3.11: DG for the 1D example; left: solution plot, right: error plot.

- Let e be an *interior* edge of $\mathcal{T}$ that is shared by K and K'. For $x \in e$, we define

$$\begin{aligned} \widehat{u}(x) &= \text{whatever} \qquad \text{if } \mathbf{b}(x)\cdot n_K(x) = \mathbf{b}(x)\cdot n_{K'}(x) = 0 \\ \widehat{u}(x) &= u|_K(x) \qquad \text{if } \mathbf{b}(x)\cdot n_K(x) > 0 \\ \widehat{u}(x) &= u|_{K'}(x) \qquad \text{if } \mathbf{b}(x)\cdot n_{K'}(x) > 0. \end{aligned}$$

- Let e be an edge on Γ^-: then, we define $\widehat{u}|_e = 0$.
- Let e be an edge on $\partial\Omega \setminus \Gamma^-$: then, we define $\widehat{u}|_e$ as limit of u from the neighbouring element.

This method has good stability properties, which we show in the following. For that we need to define the *jump* $[\![u]\!]$ on an edge $e = K|K'$:

$$\begin{aligned} [\![u]\!]|_e &= u|_K n_K + u|_{K'} n_{K'} \qquad \text{if } e = K|K' \text{ is an interior edge,} \\ [\![u]\!]|_e &= u|_K n_K \qquad \text{if } e \text{ is an edge on the boundary with } e \subset \partial K. \end{aligned}$$

Theorem 3.44 *Assume*

$$c - \frac{1}{2}\nabla\cdot\mathbf{b} \geq c_0 > 0 \qquad \text{on } \overline{\Omega}.$$

With the jump $[\![\cdot]\!]$ and the choice of "upwind flux", there holds for smooth functions u that

$$B_{DG}^{Trans}(u,u) \geq \sum_{K\in\mathcal{T}} \|\sqrt{c_0}u\|^2_{L^2(K)} + \sum_{e\in\mathcal{E}} \frac{1}{2}\||\mathbf{b}\cdot n_K|^{1/2}[\![u]\!]\|^2_{L^2(e)},$$

where $\mathcal{E}$ denotes the set of all edges of $\mathcal{T}$. For boundary edges, the jump is just the value of the trace.

Proof: For smooth functions u, there holds $u\nabla u = \nabla(\frac{1}{2}u^2)$. Thus, for every element $K \in \mathcal{T}$, we have

$$\int_K u(cu + \mathbf{b}\cdot\nabla u) = \int_K u^2\left(c - \frac{1}{2}\nabla\cdot\mathbf{b}\right) + \int_{\partial K}\frac{1}{2}\mathbf{b}\cdot n_K u^2.$$

Consequently,

$$B_{DG}^{Trans}(u,u) = \sum_{K\in\mathcal{T}}\int_K u^2 \underbrace{\left(c - \frac{1}{2}\nabla\cdot\mathbf{b}\right)}_{\geq c_0} + \int_{\partial K}\mathbf{b}\cdot n_K\left[\widehat{u}u - \frac{1}{2}u^2\right].$$

We write the sum $\sum_{K\in\mathcal{T}}\int_{\partial K}$ as sum over all edges. Hereby, we consider:

- Let e be an *interior* edge, shared by the elements K and K'. Let $x \in e$. Let w.l.o.g. K the element with $\mathbf{b}(x) \cdot n_K(x) > 0$ (the case $\mathbf{b}(x) \cdot n_K(x) = 0$ is not interesting). We then calculate due to $n_K = -n_{K'}$ and the choice of $\widehat{u}(x)$:

$$\begin{aligned}
&\mathbf{b}(x) \cdot n_K(x) \left(\widehat{u}(x) u_K(x) - \frac{1}{2} u_K(x)^2 \right) + \mathbf{b}(x) \cdot n_{K'}(x) \left(\widehat{u}(x) u_{K'}(x) - \frac{1}{2} u_{K'}(x)^2 \right) \\
&= \mathbf{b}(x) \cdot n_K(x) \left(u_K(x)^2 - \frac{1}{2} u_K(x)^2 - u_K(x) u_{K'}(x) + \frac{1}{2} u_{K'}(x)^2 \right) \\
&= \mathbf{b}(x) \cdot n_K(x) \frac{1}{2} (u_K(x) - u_{K'}(x))^2.
\end{aligned}$$

  Note that this calculation also holds true for the case $\mathbf{b}(x) \cdot n_K(x) = 0$.

- If e is a boundary edge with $e \subset \Gamma^-$, then $\widehat{u} = 0$ on e and consequently

$$\mathbf{b} \cdot n_K \left(\widehat{u} - \frac{1}{2} u \right) u = -\mathbf{b} \cdot n_K \frac{1}{2} u^2 = \frac{1}{2} |\mathbf{b} \cdot n_K| u^2 = \frac{1}{2} |\mathbf{b} \cdot n_K| |[\![u]\!]|^2,$$

  where we exploited the definition of the jump on the boundary edges in the last step.

- If e is a boundary edge with $e \subset \partial\Omega \setminus \Gamma^-$, then $\mathbf{b} \cdot n_K \geq 0$ and $\widehat{u} = u$. Consequently,

$$\mathbf{b} \cdot n_K \left(\widehat{u} - \frac{1}{2} u \right) u = \frac{1}{2} \mathbf{b} \cdot n_K u^2 = \frac{1}{2} \mathbf{b} \cdot n_K |[\![u]\!]|^2,$$

  where we again exploited the definition of the jump on the boundary edges.

Combining all edge contributions, we obtain

$$\sum_{K \in \mathcal{T}} \int_{\partial K} \mathbf{b} \cdot n_K \left(\widehat{u} - \frac{1}{2} u \right) u = \frac{1}{2} \sum_{e \in \mathcal{E}} \| |\mathbf{b} \cdot n_K|^{1/2} [\![u]\!] \|^2_{L^2(e)},$$

where we (sloppily) write n_K for a normal vector on e. □

Theorem 3.44 shows that the bilinear form B_{DG}^{Trans} is coercive. In particular, this gives unique solvability of the discrete method.

The derivation of the variational formulation additionally shows that the method is consistent in the following sense:
If u is a solution to (3.46) that additionally satisfies the regularity requirement $u \in H^1(\Omega)$, then

$$B_{DG}^{Trans}(u, v) = l(v) \qquad \forall v \in S^{p,0}(\mathcal{T}). \tag{3.49}$$

This implies Galerkin orthogonality

$$B_{DG}^{Trans}(u - u_N, v) = 0 \qquad \forall v \in S^{p,0}(\mathcal{T}). \tag{3.50}$$

3.6.1 Numerical example

We consider the equation

$$\mathbf{b}(x) \cdot \nabla u = f \quad \text{on } (0,1)^2,$$

with given wind $\mathbf{b}(x,y) := (1, \sin(4\pi x)/2)$ and source $f = \exp(-400 \cdot (y-0.5)^2)$. The discretization is done with DG in ngsolve with $p = 2$ and upwinding numerical flux. Figure 3.12 depicts the wind as well as the numerical solution.

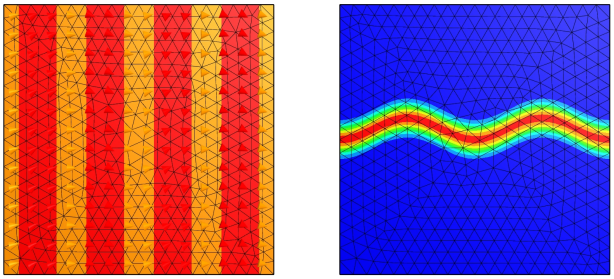

Figure 3.12: Numerical solution (right) to (3.42) with wind $\mathbf{b}$ depicted on the left.

3.7 DG and time-stepping

In the previous section, we used that the time variable does not warrant a special treatment and derived a space-time method. In practice, as for parabolic problems, time-stepping is the more common method and can be employed together with DG in space.

Setting: In order to not care about boundary conditions, we take the simplest case $\Omega = \mathbb{R}^d$ (treatment of boundary conditions is a topic on its own). Moreover, we assume that u_0 has compact support.
We consider the hyperbolic conservation law

$$u_t + \nabla \cdot \mathbf{f}(u) = 0 \quad \text{on } \Omega \times \mathbb{R}^+, \qquad u(\cdot, 0) = u_0 \tag{3.51}$$

Let $\mathcal{T}$ be a triangulation of Ω. For (discontinuous) test-functions $v \in S^{p,0}(\mathcal{T})$, we obtain after elementwise integration by parts

$$\sum_{K \in \mathcal{T}} \frac{d}{dt} \int_K uv - \int_K \nabla v \cdot \mathbf{f}(u) + \int_{\partial K} n_K \cdot \mathbf{f}(u) v = 0 \qquad \forall v \in S^{p,0}(\mathcal{T}).$$

As usual with DG methods, we have to choose a *numerical flux* $\widehat{f} = \widehat{f}(u, v, n)$ to enforce a coupling of neighboring elements. Denote by $\mathcal{N}(K) := \{K' \in \mathcal{T} \,:\, \overline{K} \cap \overline{K'} \neq \emptyset\}$ the neighboring elements of $K \in \mathcal{T}$. Thus, we have the numerical method: Find $u \in S^{p,0}(\mathcal{T})$, such that

$$\sum_K \frac{d}{dt} \int_K uv - \int_K \nabla v \cdot \mathbf{f}(u) + \sum_{K' \in \mathcal{N}(K)} \int_{K|K'} \widehat{f}(u_K, u_{K'}, n_K) v = 0 \qquad \forall v \in S^{p,0}(\mathcal{T}),$$

where we again used the notation $K|K'$ for the common edge between the elements K and K'. For the numerical flux, we define some desired properties.

Definition 3.45 (Numerical Flux) *Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \,|\, \|x\|_2 = 1\}$. A function $\widehat{f} : \mathbb{R} \times \mathbb{R} \times \mathbb{S}^{d-1} \to \mathbb{R}$ is called numerical flux, if it is locally Lipschitz-continuous. The numerical flux is called*

- consistent, *if $\widehat{f}(u, u, \mathbf{n}) = \mathbf{f}(u) \cdot \mathbf{n}$ for all u.*
- conservative, *if $\widehat{f}(u, v, \mathbf{n}) = -\widehat{f}(v, u, -\mathbf{n})$.*

- monotone, *if $\widehat{f}$ is monotone increasing in the first argument and monotone decreasing in the second argument:* $\widehat{f}(\uparrow, \downarrow, \mathbf{n})$.

Example 3.46 Consider the advection equation, i.e., $f(u) = \mathbf{b}u$ with constant $\mathbf{b}$. Upwinding means choosing on $K|K'$ the numerical flux

$$\widehat{f}(u_K, u_{K'}, \mathbf{n}_K) = \begin{cases} \mathbf{b} \cdot \mathbf{n}_K u_K & \text{if } \mathbf{b} \cdot \mathbf{n}_K > 0 \\ \mathbf{b} \cdot \mathbf{n}_K u_{K'} & \text{if } \mathbf{b} \cdot \mathbf{n}_K < 0. \end{cases}$$

This choice of flux is conservative (exercise!) and consistent. Note that this choice of flux can also be written (without case distinction) as

$$\widehat{f}(u_K, u_{K'}, \mathbf{n}_K) = \frac{1}{2}\mathbf{b} \cdot \mathbf{n}_K \left(u_{K'} + u_K\right) - \frac{1}{2} \left|\mathbf{b} \cdot \mathbf{n}_K\right| \left(u_{K'} - u_K\right).$$

Remark 3.47 The property $\widehat{f}(u, v, \mathbf{n}) = -\widehat{f}(v, u, -\mathbf{n})$ expresses a conservation property: With the test-function $v \equiv 1$, we obtain

$$\frac{d}{dt}\int_\Omega u = \sum_{K\in\mathcal{T}} \frac{d}{dt}\int_K u = -\sum_{K\in\mathcal{T}} \sum_{K'\in\mathcal{N}(K)} \int_{K|K'} \widehat{f}(u_K, u_{K'}, \mathbf{n}_K).$$

Writing this as a sum over edges (i.e. hyperplanes), together with the observation that each interior edge e in the mesh is shared by two elements K_e, K'_e, this leads to

$$\begin{aligned} \frac{d}{dt}\int_\Omega u &= -\sum_{K\in\mathcal{T}} \sum_{K'\in\mathcal{N}(K)} \int_{K|K'} \widehat{f}(u_K, u_{K'}, \mathbf{n}_K) = -\sum_e \int_e \widehat{f}(u_{K_e}, u_{K'_e}, \mathbf{n}_{K_e}) + \int_e \widehat{f}(u_{K'_e}, u_{K_e}, \mathbf{n}_{K'_e}) \\ &= -\sum_e \int_e \widehat{f}(u_{K_e}, u_{K'_e}, \mathbf{n}_{K_e}) - \widehat{f}(u_{K_e}, u_{K'_e}, \mathbf{n}_{K_e}) = 0, \end{aligned}$$

where we have used the conservative property in the last equation. ∎

Now, the DG-method can be combined with a time-stepping scheme, e.g. the Euler-methods. The easiest case is the explicit Euler method. Note that, as for previous semi-discrete methods, the term involving the time-derivative produces a mass matrix, which now is block-diagonal (as there is no coupling between the elements/basis functions there) and thus very efficiently inverted. Moreover, for the explicit Euler method, there still holds the conservation property from Remark 3.47:

Exercise 3.48 Formulate the time-stepping DG-method with explicit Euler method. Denote by u^n the numerical approximation at time t_n. Show that

$$\int_\Omega u^{n+1} = \int_\Omega u^n \qquad \forall n \in \mathbb{N}_0.$$

Remark 3.49 There are many choices for time-stepping. Oftentimes, a method is desirable that reproduces some (monotonicity) property of the exact solution. For scalar conservation laws there might hold something like $\|u(\cdot, t)\|_{L^\infty} \leq \|u_0(\cdot)\|_{L^\infty}$. Methods that reproduce such a property, i.e. produce approximations with $\|u^{n+1}\| \leq \|u^n\|$ (in some norm, on the ODE-level) are called *strong stability preserving methods.*

3.7.1 Numerical example

We consider the instationary transport equation

$$u_t + \operatorname{div}(\mathbf{b}u) = 0 \quad \text{on } (0,1)^2 \times \mathbb{R}^+, \qquad u(\cdot,0) = u_0 \tag{3.52}$$

with given circular wind $\mathbf{b}(x,y) = (y - 0.5, 0.5 - x)$ and initial condition $u_0(x,y) = \exp(-100 \cdot [(x - 0.5)^2 + (y - 0.75)^2])$.
The spatial discretization is done with DG in ngsolve with $p = 3$ and upwinding numerical flux. For time discretization we employ the explicit Euler method. Figure 3.13 depicts the wind as well as the numerical solution at different times.

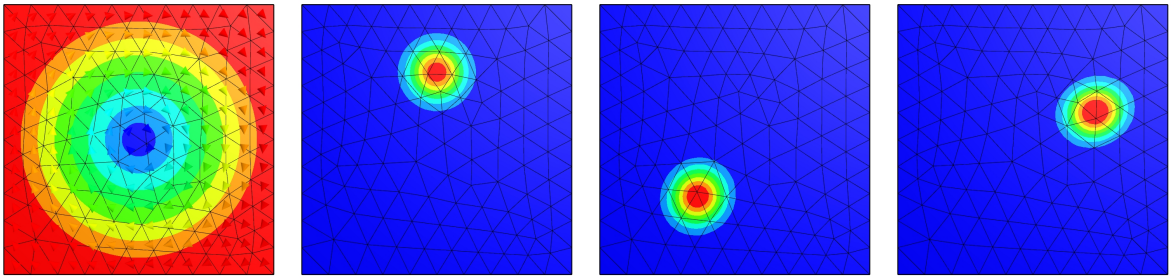

Figure 3.13: Numerical solution to (3.52). Wind $\mathbf{b}$ depicted left, then initial condition at $T = 0$ and numerical solutions at $T = 10, 20$.

Bibliography

[1] Merico E. Argentati, Andrew V. Knyazev, Klaus Neymeyr, Evgueni E. Ovtchinnikov, and Ming Zhou. Convergence theory for preconditioned eigenvalue solvers in a nutshell. *Found. Comput. Math.*, 17(3):713–727, 2017.

[2] Uri M. Ascher. *Numerical methods for evolutionary differential equations*, volume 5 of *Computational Science & Engineering*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[3] Pavel B. Bochev and Max D. Gunzburger. *Least-squares finite element methods*, volume 166 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[4] D. Boffi. Finite element approximation of eigenvalue problems. *Acta Numer.*, 19:1–120, 2010.

[5] Thomas Führer and Michael Karkulik. Space-time least-squares finite elements for parabolic equations. *Comput. Math. Appl.*, 92:27–36, 2021.

[6] Carolyn Gordon, David L. Webb, and Scott Wolpert. One cannot hear the shape of a drum. *Bull. Amer. Math. Soc. (N.S.)*, 27(1):134–138, 1992.

[7] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. *Time-dependent problems and difference methods*. Pure and Applied Mathematics (Hoboken). John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2013.

[8] Volker John. *Finite element methods for incompressible flow problems*, volume 51 of *Springer Series in Computational Mathematics*. Springer, Cham, 2016.

[9] Mark Kac. Can one hear the shape of a drum? *Amer. Math. Monthly*, 73(4):1–23, 1966.

[10] Andrew V. Knyazev and Klaus Neymeyr. A geometric theory for preconditioned inverse iteration. III. A short and sharp convergence estimate for generalized eigenvalue problems. volume 358, pages 95–114. 2003. Special issue on accurate solution of eigenvalue problems (Hagen, 2000).

[11] Steen Krenk. Energy conservation in Newmark based time integration algorithms. *Comput. Methods Appl. Mech. Engrg.*, 195(44-47):6110–6124, 2006.

[12] Stig Larsson and Vidar Thomée. *Partial differential equations with numerical methods*, volume 45 of *Texts in Applied Mathematics*. Springer-Verlag, Berlin, 2003.

[13] J.-L. Lions and E. Magenes. *Problèmes aux limites non homogènes et applications. Vol. 1*, volume No. 17 of *Travaux et Recherches Mathématiques*. Dunod, Paris, 1968.

[14] Klaus Neymeyr and Ming Zhou. The block preconditioned steepest descent iteration for elliptic operator eigenvalue problems. *Electron. Trans. Numer. Anal.*, 41:93–108, 2014.

[15] Vidar Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1997.

[16] N. J. Zabusky and M. D. Kruskal. Interaction of solitons in a collisionless plasma and the recurrence of initial states. *Phys. Rev. Lett.*, 15:240–243, Aug 1965.