Numerics of Partial Differential Equations

Victor A. Kovtunenko

Dr. Habil. Docent

Institute for Mathematics and Scientific Computing, Karl-Franzens University of Graz, NAWI Graz;

Austrian Academy of Sciences (ÖAW);

Lavrent'ev Institute of Hydrodynamics, Siberian Division of the Russian Academy of Sciences

Email address: `https://homepage.uni-graz.at/victor.kovtunenko/`

The script is based on the lectures MATH.453 (SS 2017): *Numerics of Partial Differential Equations* held at the University of Graz.

The author thanks Anna V. Zubkova for the help preparing LaTex files and figures for the script.

Contents

Introduction

The script mainly follows the sources by (Clason 2013, Ciarlet and Lions 1991, Grossmann, Roos and Stynes 2007, Hackbusch 1992, Keeling 2016), other references will be given when used.

1. Classification of PDE problems

1.1. Classification of PDEs. Let the following data be prescribed:

(i) a domain (open set) $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, which can be bounded with the boundary $\partial\Omega =: \Gamma$, as well as unbounded;

(ii) symmetric coefficients $a_{ij}(x) = a_{ji}(x)$, $b_i(x)$, $c(x) \in C(\Omega)$ for $i, j = 1, \ldots, d$ and $x = (x_1, \ldots, x_d)$, which imply continuous functions.

For a twice differentiable function $u(x) : C^2(\Omega) \mapsto \mathbb{R}$ (or $\mathbb{C}$), we consider a *differential operator* $L : C^2(\Omega) \mapsto C(\Omega)$ given by

$$L\Big(x, \frac{\partial}{\partial x}\Big)u := -\sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i} + c(x). \tag{1.1}$$

The total symbol is defined as the corresponding *polynomial operator*

$$p(x, \xi) := L(x, \xi) = -\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j + \sum_{i=1}^{d} b_i(x)\xi_i + c(x),$$

the *principle symbol* is set to

$$\sigma(x, \xi) := \sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j = \xi^\top A(x)\xi, \quad \xi = \begin{pmatrix}\xi_1\\ \vdots \\ \xi_d\end{pmatrix}, \xi^\top = (\xi_1, \ldots, \xi_d),$$

where the symmetric d-by-d matrix $A(x) = \begin{bmatrix} a_{11}(x) \ldots a_{1d}(x) \\ \vdots \\ a_{d1}(x) \ldots a_{dd}(x)\end{bmatrix} \in \mathrm{Sym}(\mathbb{R}^{d\times d})$.

Definition 1.1. *For $d = 2$, the second-order differential operator L given in* (1.1) *is called*

- ***elliptic*** *in $x \in \Omega$, if $\det A(x) > 0$, that is $a_{11}(x)a_{22}(x) - a_{12}^2(x) > 0$, then two non-zero eigenvalues λ_1, λ_2 have the same sign;*
- ***parabolic*** *in $x \in \Omega$, if $\det A(x) = 0$, that is $a_{11}(x)a_{22}(x) - a_{12}^2(x) = 0$, then $\lambda_1 = 0$;*
- ***hyperbolic*** *in $x \in \Omega$, if $\det A(x) < 0$, that is $a_{11}(x)a_{22}(x) - a_{12}^2(x) < 0$, then λ_1, λ_2 have different signs.*

Exercise 1.1. Extend Definition 1.1 for $d > 2$.

Example 1.1.

- *Laplace equation*: $u_{x_1x_1} + u_{x_2x_2} = 0$,
  $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\det A = 1 > 0$, $\lambda_1 = \lambda_2 = 1$, hence elliptic;
- *heat/diffusion equation*: $u_t - u_{xx} = 0$,
  $A = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$, $\det A = 0$, $\lambda_1 = 0$, $\lambda_2 = -1$, hence parabolic;
- *wave equation*: $u_{tt} - u_{xx} = 0$,
  $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, $\det A = -1 < 0$, $\lambda_1 = 1$, $\lambda_2 = -1$, hence hyperbolic.

Consider the system of $d \geq 2$ first-order PDEs for vectors $u = (u_1, \dots, u_d) : C^1(\Omega) \mapsto \mathbb{R}^d$ such that

$$L\Big(x, \frac{\partial}{\partial x}\Big)u = \Big\{ \sum_{j,k=1}^{d} a_{ij}^k(x) \frac{\partial u_j}{\partial x_k} + c_i(x), \quad i = 1, \dots, d \Big\}. \tag{1.2}$$

Denoting the matrices $A^k = \begin{bmatrix} a_{11}^k \dots a_{1d}^k \\ \vdots \\ a_{d1}^k \dots a_{dd}^k \end{bmatrix}$ for $k = 1, \dots, d$, the principle symbol

$$\sigma(x, \xi) := \det\Big(\sum_{k=1}^{d} A^k(x)\xi_k\Big) = \xi^\top A(x)\xi$$

composes a matrix A, which determines type of the corresponding differential operator (1.2) for systems according to Definition 1.1.

Other examples of typical PDE are presented in (Kovtunenko 2010a), and their appearance in applications see in (Kovtunenko 2010b) and the references therein.

For a given right-hand side $f(x) \in C(\Omega)$, we set the *initial/boundary value problem (I/BVP)*: Find a solution $u(x) \in C^2(\Omega)$ satisfying the equation:

$$Lu(x) = f(x), \quad x \in \Omega,$$

and

- boundary conditions (BC) for elliptic operator L;
- BC with respect to x and initial conditions (IC) with respect to t for parabolic L;
- BC with respect to x and two IC with respect to t for hyperbolic L.

Definition 1.2 (Hadamar). *A problem is called* ***well-posed****, if*

(i) *a solution exists (existence);*
(ii) *the solution is unique (uniqueness);*
(ii) *the solution depends continuously on data (stability).*

1.2. Elliptic boundary-value problem (BVP). We assume that

(i) $A(x) \in \mathrm{Spd}(\mathbb{R}^{d\times d})$ for all $x \in \Omega$ and there exists $\underline{a} > 0$ such that

$$\xi^\top A(x)\xi \geqslant \underline{a}\|\xi\|^2, \quad \xi \in \mathbb{R}^d, \quad \|\xi\|^2 := \sum_{i=1}^{d} \xi_i^2;$$

(ii) the coefficients $a_{ij}(x)$, $b_i(x)$, $c(x)$ are uniformly bounded for $x \in \Omega$.

For the elliptic differential equation

$$L\Big(x, \frac{\partial}{\partial x}\Big)u = f \quad \text{in } \Omega$$

stated in the bounded domain with the unit normal vector $n = (n_1, \dots, n_d)^\top$ at the boundary $\partial\Omega$ and outward to Ω, the usual *boundary conditions* are:

- ***Dirichlet (1st kind)***: $u = \phi$ on $\partial\Omega$, $\phi \in C(\partial\Omega)$;
- ***Neumann (2nd kind)***: $\displaystyle\sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial u}{\partial x_i}n_j = \psi$ on $\partial\Omega$, $\psi \in C(\partial\Omega)$;
- ***Robin (3rd kind)***: $\displaystyle\sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial u}{\partial x_i}n_j + \alpha(x)u = \psi$ on $\partial\Omega$, $\alpha \in C(\partial\Omega)$;
- ***mixed BC***.

Lemma 1.1 (Weak maximum principle). *(Kuttler 2003, Th.6.2, p.137) Let $c \equiv 0$ and $u \in C^2(\Omega) \cup C(\overline{\Omega})$, then $Lu \leqslant 0$ in Ω follows that $\max\limits_{x\in\overline{\Omega}} u(x) \leqslant \max\limits_{x\in\partial\Omega} u(x)$, where the closure $\overline{\Omega} = \Omega \cup \partial\Omega$.*

Lemma 1.2 (Comparison principle). *Let $c \geqslant 0$ and $u, v \in C^2(\Omega) \cup C(\overline{\Omega})$, then $Lu \leqslant Lv$ in Ω and $u \leqslant v$ on $\partial\Omega$ follows that $u \leqslant v$ in $\overline{\Omega}$.*

Exercise 1.2. Prove that the weak maximum principle follows the comparison principle. Hint: consider $\Omega_+ = \{x \in \Omega : (u - v)(x) \geqslant 0\}$ and $\Omega_- = \{x \in \Omega : (u - v)(x) \leqslant 0\}$.

We specify a *homogeneous Dirichlet problem*: Find $u \in C^2(\Omega) \cup C(\overline{\Omega})$ such that

$$\begin{cases} Lu = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

Lemma 1.3. *The comparison principle follows uniqueness of the solution.*

For existence of a solution, ***regularity of the domain*** is needed:

Definition 1.3. *A boundary of domain Ω is said to belong to the class $C^{m,\alpha}$ (write $\partial\Omega \in C^{m,\alpha}$) with $m \in \mathbb{N}_0$, $\alpha \in (0, 1]$, if*

(i) *a finite number of open balls $(B_i)_{i\in I}$ exists such that $\partial\Omega \subset \bigcup\limits_{i\in I} B_i$;*

(ii) *a Hölder continuous function $\phi_i \in C^{m,\alpha}(\overline{B}_i)$ exists, that is*

$$\|\phi_i\|_{C^{m,\alpha}(B_i)} = \|\phi_i\|_{C^m(B_i)} + \max_{|\beta|=m} \sup_{x\neq y\in B_i} \frac{|D^\beta\phi_i(x) - D^\beta\phi_i(y)|}{\|x - y\|^\alpha} < \infty,$$

where $\|\phi_i\|_{C^m(B_i)} = \displaystyle\sum_{|\beta|=0}^{m} \sup_{x\in B_i} |D^\beta\phi_i|$ with the multi-index $\beta = (\beta_1, \dots, \beta_d)$ of length $|\beta| = \beta_1 + \ldots + \beta_d$, $D^\beta = \frac{\partial^{\beta_1}\ldots\partial^{\beta_d}}{\partial x_1^{\beta_1}\ldots\partial x_d^{\beta_d}}$, which maps one-to-one $\partial\Omega \cap B_i$ to the image in the hyper-space $x_d = 0$, and $\Omega \cap B_i$ to the image in the half-space $x_d > 0$.

In particular, $\partial\Omega \subset C^{0,1}$ is called the Lipschitz-continuous boundary.

Theorem 1.1 (Existence). *(Wu, Yin and Wang 2006, Th.8.2.7, p.249) Let $c \geqslant 0$, $\partial\Omega \in C^{2,\alpha}$ and $a_{ij}, b_i, c \in C^\alpha(\Omega)$, then the Dirichlet problem has a unique classical solution.*

1.3. Parabolic initial-boundary value problem (IBVP). In the time-space cylinder $Q_T := (0,T) \times \Omega$ with a final time $T > 0$, for a given $f \in C(Q_T)$, initial data $u^0 \in C_0(\overline{\Omega})$ (continuous function with the trace $u^0 = 0$ on $\partial\Omega$), find a solution $u(t,x) \in C^{2,1}(Q_T) \cap C(\overline{Q_T})$ (i.e. $x \mapsto u : C^2(\Omega) \cap C(\overline{\Omega})$, $t \mapsto u : C^1(0,T) \cap C([0,T])$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t} + L\Big(t, x, \dfrac{\partial}{\partial x}\Big) u = f & \text{in } Q_T, \\ u = 0 \quad \text{on } (0,T) \times \partial\Omega, \\ u = u^0 \quad \text{in } \Omega \text{ as } t = 0. \end{cases}$$

Lemma 1.4 (Weak parabolic maximum principle). *(Kuttler 2003, Th.6.7, p.142) Let $c \equiv 0$ and $u \in C^{2,1}(Q_T) \cap C(\overline{Q_T})$, then $\dfrac{\partial u}{\partial t} + Lu \leqslant 0$ in Q_T follows that*

$$\max_{(t,x) \in \overline{Q_T}} u(x) \leqslant \max_{(t,x) \in \overline{Q_T} :\, x \in \partial\Omega \ \text{or} \ t = 0} u(x).$$

Lemma 1.5 (Parabolic comparison principle). *Let $c \geqslant 0$ and $u, v \in C^{2,1}(Q_T) \cap C(\overline{Q_T})$, then*

$$\begin{cases} \dfrac{\partial u}{\partial t} + Lu \leqslant \dfrac{\partial v}{\partial t} + Lv \quad \text{in } Q_T, \\ u \leqslant v \quad \text{on } (0,T) \times \partial\Omega, \\ u \leqslant v \quad \text{in } \Omega \text{ as } t = 0 \end{cases}$$

follows that $u \leqslant v$ in Q_T.

Theorem 1.2 (Existence). *(Wu et al. 2006, Th.8.3.7, p.251) Let $\partial\Omega \in C^{2,\alpha}$ and $a_{ij}, b_i, c, f \in C^\alpha(Q_T)$, $\alpha > 0$, then there exists a unique classical solution of the parabolic IBVP.*

Note that $c \geqslant 0$ is not needed here.

1.4. Hyperbolic IBVP. For given $f \in C(Q_T)$, $v^0 \in C(\overline{\Omega})$ and $u^0 \in C_0(\overline{\Omega})$, find $u \in C^2(Q_T) \cap C^{1,0}(\overline{Q_T})$ (i.e. $x \mapsto u : C^2(\Omega) \cap C^1(\overline{\Omega})$, $t \mapsto u : C^2(0,T) \cap C([0,T])$ such that

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} + L\Big(t, x, \dfrac{\partial}{\partial x}\Big) u = f & \text{in } Q_T, \\ u = 0 \quad \text{on } (0,T) \times \partial\Omega, \\ u = u^0,\ u_t = v^0 \quad \text{in } \Omega \text{ as } t = 0. \end{cases}$$

Existence needs compatibility conditions.

CHAPTER 1

Numerics of Initial Boundary Value Problems

2. Discretization

For a given right-hand side $f \in C(\Omega)$ we start with a model *Dirichlet problem for the Poisson equation* in the square $\Omega = (0,1)^2$: Find a solution $u(x) \in C^2(\Omega) \cap C(\overline{\Omega})$ such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega =: \Gamma. \end{cases} \tag{2.1}$$

We call $\overline{\Omega_h} = \{x_{ij} \in \overline{\Omega},\ i,j = 0,\dots,N\}$ the uniform *grid* consisted of equidistant *grid points* $x_{ij} = x_{(i,j)} = (ih, jh)$ with the *mesh size* $h > 0$. The *grid functions* are composed by values $u_{ij} = u_{(i,j)} \approx u(x_{ij})$, $f_{ij} = f_{(i,j)} \approx f(x_{ij})$ associated to the grid points.

Passing $h \searrow 0$ provides the ***Taylor expansion*** in the direction x_1:

$$u((i+1)h, jh) = u(ih, jh) + h\frac{\partial u}{\partial x_1}(ih, jh) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x_1^2}(ih, jh) + \mathrm{O}(h^3),$$
$$u((i-1)h, jh) = u(ih, jh) - h\frac{\partial u}{\partial x_1}(ih, jh) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x_1^2}(ih, jh) + \mathrm{O}(h^3);$$

analogously, in the direction x_2:

$$u(ih, (j \pm 1)h) = u(ih, jh) \pm h\frac{\partial u}{\partial x_2}(ih, jh) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x_2^2}(ih, jh) + \mathrm{O}(h^3),$$

where the short notation ± stands, respectively, for two options "+" and "-". After summation of these four asymptotic relations we get:

$$\begin{aligned} &u((i+1)h, jh) + u((i-1)h, jh) + u(ih, (j+1)h) + u(ih, (j-1)h) \\ &\qquad = 4u(ih, jh) + h^2\Big(\frac{\partial^2 u}{\partial x_1^2}(ih, jh) + \frac{\partial^2 u}{\partial x_2^2}(ih, jh)\Big) + \mathrm{O}(h^3). \end{aligned}$$

We note that $\dfrac{\partial^2 u}{\partial x_1^2} + \dfrac{\partial^2 u}{\partial x_2^2} = \Delta u$. Therefore, based on the above expansion, the problem (2.1) is approximated by a *discrete problem* on the grid $\overline{\Omega_h}$ such that

$$\begin{cases} 4u_{(i,j)} - u_{(i+1,j)} - u_{(i-1,j)} - u_{(i,j+1)} - u_{(i,j-1)} = h^2 f_{(i,j)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad i,j = 1,\dots,N-1, \\ u_{(0,j)} = u_{(N,j)} = u_{(i,0)} = u_{(i,N)} = 0, \quad i,j = 0,\dots,N. \end{cases} \tag{2.2}$$

From this consideration we arrive at generalization in $\mathbb{R}^d$ for

- ***grid functions*** $u_h : \overline{\Omega_h} \mapsto \mathbb{R}$;

- ***discrete function spaces:***

$U_h := \{u_h : \overline{\Omega_h} \mapsto \mathbb{R}\}, \quad U_h^0 := \{u_h \in U_h : \ u_h = 0 \text{ on } \Gamma_h := \overline{\Omega_h} \cap \Gamma\};$

- ***difference operators:***

$$(D_i^+ u)(x) = \frac{1}{h}[u(x + he^i) - u(x)] \text{ called } \textit{forward difference quotient},$$
$$(D_i^- u)(x) = \frac{1}{h}[u(x) - u(x - he^i)] \text{ called } \textit{backward difference quotient},$$
$$(D_i^0 u)(x) = \frac{1}{2}[D_i^+ + D_i^-]u(x) \text{ called } \textit{central difference quotient}$$

with respect to the unit vector: $e^i = (0, \dots, 0, \underset{i\text{th}}{1}, 0, \dots, 0)$ in $\mathbb{R}^d$.

2.1. Expansion of difference operators.

Lemma 2.1. *(Hackbusch 1992, Lemma 4.1.1., p.39) The following expansions of the difference operators hold:*

(i) $(D_i^\pm u)(x) = \dfrac{\partial u}{\partial x_i}(x) + hR^\pm(x)$ *with the reminders* $R^\pm$ *such that*

$$|R^+(x)| \leqslant \frac{1}{2}\|u\|_{C^2([x,x+he^i])}, \quad |R^-(x)| \leqslant \frac{1}{2}\|u\|_{C^2([x-he^i,x])};$$

(ii) $(D_i^0 u)(x) = \dfrac{\partial u}{\partial x_i}(x) + h^2 R^0(x), \quad |R^0(x)| \leqslant \dfrac{1}{6}\|u\|_{C^3([x-he^i,x+he^i])};$

(iii) $(D_i^- D_i^+ u)(x) = \dfrac{\partial^2 u}{\partial x_i^2}(x) + h^2 R(x), \quad |R(x)| \leqslant \dfrac{1}{12}\|u\|_{C^4([x-he^i,x+he^i])}.$

2.2. Discrete norms. For grid functions $u_h \in U_h^0$ given over equidistant points of a uniform grid $\overline{\Omega_h}$, in $\Omega_h := \overline{\Omega_h} \cap \Omega$ consisted of interior points x_h we set

- ***discrete maximum norm***: $\quad \|u_h\|_{\infty,h} := \max\limits_{x_h \in \Omega_h} |u_h(x_h)|,$
- ***discrete*** L^p***-norm*** (where $p \in [1, \infty)$): $\quad \|u_h\|_{L^p,h}^p := h^d \sum\limits_{x_h \in \Omega_h} |u_h(x_h)|^p,$
- ***discrete*** L^2***-norm***: $\quad \|u_h\|_{0,h}^2 := h^d \sum\limits_{x_h \in \Omega_h} |u_h(x_h)|^2 = (u_h, u_h)_h,$
- ***discrete*** H^1***-norm***: $\|u_h\|_{1,h}^2 := \|u_h\|_{0,h}^2 + |u_h|_{1,h}^2$ and the ***seminorm***:

$$|u_h|_{1,h}^2 := h^d \sum_{x_h \in \Omega_h} \sum_{i=1}^{d} |(D_i^+ u_h)(x_h)|^2 = \sum_{i=1}^{d} (D_i^+ u_h, D_i^+ u_h)_h,$$

where the latter two norms are induced by the scalar product for $u_h, v_h \in U_h^0$:

$$(u_h, v_h)_h := h^d \sum_{x_h \in \Omega_h} u_h(x_h) v_h(x_h). \tag{2.3}$$

In *non-equidistant case*, replace h^d with the *weight* $|D_h(x_h)| := \displaystyle\int_{D_h(x_h)} dx$ over appropriate subdomains $D_h(x_h)$ (e.g. Voronoi boxes (5.1)), respectively, instead (2.3) define the scalar product:

$$(u_h, v_h)_h := \sum_{x_h \in \Omega_h} |D_h(x_h)| u_h(x_h) v_h(x_h).$$

In fact, in the equidistant case the Voronoi box $D_h(x_h) = \big\{x \in \Omega : \|x - x_h\| < \frac{h}{2}\big\}$ has the area $|D_h(x_h)| = h^d$.

We note that the following *norm equivalence* is h-dependent:

$$\min_{x_h \in \Omega_h} |D_h(x_h)|^{1/2} \|u_h\|_{\infty,h} \leqslant \|u_h\|_{0,h} \leqslant |\Omega|^{1/2} \|u_h\|_{\infty,h},$$

and in the equidistant case: $\min\limits_{x_h \in \Omega_h} |D_h(x_h)|^{1/2} = h^{d/2}$. As $h \searrow 0^+$, the following *norm limit* holds:

$$\lim_{h\to 0} \|u(x_h)\|_{0,h} = \|u\|_{L^2(\Omega)}, \quad \lim_{h\to 0} \|u(x_h)\|_{\infty,h} = \|u\|_{L^\infty(\Omega)}.$$

2.3. Consistency, convergence, stability of discretization. Consider the *continuous problem* given by an abstract operator equation: Find $u \in U$ such that

$$Lu = f \tag{2.4}$$

with $L : U \mapsto V$, $f \in V$ over normed vector spaces U, V. For a discrete counterpart $L_h : U_h \mapsto V_h$, $f_h \in V_h$, the related *discrete problem* reads: Find $u_h \in U_h$ such that

$$L_h u_h = f_h. \tag{2.5}$$

Let a *restriction operator* $r_h : U \mapsto U_h$, $(r_h u)(x_h) = u(x_h)$ for $x_h \in \overline{\Omega_h}$ be defined.

Definition 2.1. *A discretization* (2.5) *of the problem* (2.4) *is called*

- ***consistent****, when the consistency error decays to zero:*

$$\|L_h(r_h u) - f_h\|_{V_h} \to 0 \quad \text{as } h \to 0,$$

and consistent of order p, if $\|L_h(r_h u) - f_h\|_{V_h} = \mathrm{O}(h^p)$*;*

- ***convergent****, when the discretization error decays to zero:*

$$\|r_h u - u_h\|_{U_h} \to 0 \quad \text{as } h \to 0,$$

and convergent of order q, if $\|r_h u - u_h\|_{U_h} = \mathrm{O}(h^q)$*;*

- ***stable****, when there exists $C > 0$ such that*

$$\|u_h - v_h\|_{U_h} \leqslant C \|L_h u_h - L_h v_h\|_{V_h} \quad \text{for } u_h, v_h \in U_h.$$

Note that for linear operators L_h, stability is equivalent to boundedness of the inverse operator $\|L_h^{-1}\| \leqslant C$.

Lemma 2.2. *Let the problems* (2.4) *and* (2.5) *have solutions. If discretization is consistent and stable, then it is convergent.*

3. Finite difference method (FDM) for elliptic BVP

For twice differentiable functions $u(x) : \Omega \mapsto \mathbb{R}$, consider a second-order differential operator given in the *divergence-form*:

$$Lu := -\mathrm{div}(A(x)\nabla u) = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i} \left[\sum_{j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_j} \right]. \tag{3.1}$$

We assume that $A = (a_{ij})_{i,j=1}^d \in C^1(\Omega)$ is uniformly *symmetric positive-definite* matrix for $x \in \Omega$ and there exists $\underline{a} > 0$ such that

$$\xi^\top A(x) \xi = \sum_{i,j=1}^{d} a_{ij}(x) \xi_i \xi_j \geqslant \underline{a} \sum_{i=1}^{d} \xi_i^2 = \underline{a} \|\xi\|^2 \quad \text{for all } \xi \in \mathbb{R}^d.$$

In particular, if $A(x) \equiv I$ is the identity matrix, then $Lu = -\mathrm{div}(\nabla u) = -\Delta u$ is the *Laplace operator*.

For a given right-hand side $f \in C(\Omega)$, consider a ***Dirichlet problem*** in the cube: Find $u(x) \in C^2(\Omega) \cap C(\overline{\Omega})$ such that

$$\begin{cases} Lu = f & \text{in } \Omega = (0,1)^d, \\ u = 0 & \text{on } \partial\Omega =: \Gamma. \end{cases} \tag{3.2}$$

In the closure $\overline{\Omega}$ we introduce the *uniform grid* of mesh-size $h = 1/N > 0$ for a fixed natural number $N \in \mathbb{N}$:

$$\overline{\Omega_h} = \{x_h = (x_1, \dots, x_d)^\top \in \mathbb{R}^d : \quad x_i = hn_i,\ n_i \in \{0, 1, \dots, N\},\ i = 1, \dots, d\}.$$

Recall the discrete spaces $U_h = \{u_h : \overline{\Omega_h} \mapsto \mathbb{R}\}$, $U_h^0 = \{u_h \in U_h : u_h = 0 \text{ on } \Gamma_h\}$, and set $V_h = \{u_h : \Omega_h \mapsto \mathbb{R}\}$.

According to (3.1) we define the *discrete difference operator* $L_h : U_h \mapsto V_h$,

$$L_h u_h = -\sum_{i=1}^d D_i^- \left[\sum_{j=1}^d a_{ij}(x) D_j^+ u_h \right],$$

and set the *finite-dimensional (discrete) problem* for $f_h := f(x_h)$: Find $u_h \in U_h^0$ such that

$$L_h u_h = f_h. \tag{3.3}$$

Similarly to Lemma 2.1, the *first-order consistency estimate* holds:

$$\|L_h(r_h u) - f_h\|_{V_h} = \max_{x_h \in \Omega_h} |([L_h - L]u)(x_h)| \leqslant C \|A\|_{C^1(\overline{\Omega})} \|u\|_{C^3(\overline{\Omega})} h, \quad C > 0,$$

since $L_h(r_h u) = (L_h u)(x_h)$ and $f_h = f(x_h) = (Lu)(x_h)$. Moreover, for the Laplace operator (when $A(x) \equiv I$), it follows the *second-order consistency estimate*:

$$\max_{x_h \in \Omega_h} |([\Delta_h - \Delta]u)(x_h)| \leqslant \frac{d}{12} \|u\|_{C^4(\overline{\Omega})} h^2. \tag{3.4}$$

Example 3.1. In 2d, for vectors $u_h = (u_{(i,j)})_{i,j=1}^N \in U_h^0$ under the Dirichlet BC: $u_{(0,j)} = u_{(N,j)} = u_{(i,0)} = u_{(i,N)} = 0$, after reordering

$$v_h := (u_{(1,1)}, \dots, u_{(N-1,1)}, \dots, u_{(N-1,1)}, \dots, u_{(N-1,N-1)}) \in \mathbb{R}^{(N-1)^2},$$

the discrete Laplace operator $-\Delta_h$ implies multiplication with the *block tridiagonal matrix*, which is *sparse*:

$$-\Delta_h u_h = \frac{1}{h^2} \begin{pmatrix} T & -I & & 0 \\ -I & \ddots & \ddots & \\ & \ddots & \ddots & -I \\ 0 & & -I & T \end{pmatrix} v_h, \quad T = \begin{pmatrix} 4 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 4 \end{pmatrix}.$$

Our aim is to estimate the convergence error $e_h := u_h - r_h u$ based on the consistency error d_h:

$$f_h - L_h(r_h u) = L_h u_h - L_h(r_h u) = L_h(u_h - r_h u) =: d_h,$$

where

$$\|d_h\|_{\infty,h} \leqslant C \begin{cases} \|u\|_{C^3(\overline{\Omega})} h, \\ \|u\|_{C^4(\overline{\Omega})} h^2 & \text{for } L = -\Delta. \end{cases}$$

3.1. Solvability. The following technical tools are employed later on.

Lemma 3.1 (*Discrete Green's formula*). *(Grossmann et al. 2007, Lemma 2.33, p.65) The operator $-D_i^+$ is adjoint to D_i^- with respect to the inner product $(\,\cdot\,,\,\cdot\,)_h$:*

$$(D_i^- u_h, v_h)_h = -(u_h, D_i^+ v_h)_h \quad \textit{for } u_h, v_h \in U_h^0.$$

Lemma 3.2 (*Discrete Poincaré inequality*). *(Grossmann et al. 2007, Lemma 2.35, p.66) There exists $K_{\mathrm{P}}(\Omega) > 0$ such that*

$$\|u\|_{0,h}^2 \leqslant K_{\mathrm{P}}(\Omega)|u_h|_{1,h}^2 \quad \textit{for } u_h \in U_h^0.$$

In particular, $K_{\mathrm{P}}(\Omega) = 1$ for $\Omega = (0,1)^d$.

Theorem 3.1 (Solvability). *For each $f_h \in V_h$, the discrete problem* (3.3) *has a unique solution.*

3.2. H^1-convergence error.

Theorem 3.2 (H^1-convergence error). *(Grossmann et al. 2007, Th 2.36, p.67) Let the solution of* (3.2) *be smooth $u \in C^3(\overline{\Omega})$. Then there exists $C > 0$ such that*

$$\|u_h - r_h u\|_{0,h} \leqslant \sqrt{K_{\mathrm{P}}(\Omega)}\|u_h - r_h u\|_{1,h} \leqslant C\|u\|_{C^3(\overline{\Omega})} h.$$

For $L = -\Delta$, if $u \in C^4(\overline{\Omega})$, then

$$\|u_h - r_h u\|_{0,h} \leqslant \sqrt{K_{\mathrm{P}}(\Omega)}\|u_h - r_h u\|_{1,h} \leqslant C\|u\|_{C^4(\overline{\Omega})} h^2.$$

Remark 3.1. Applying the norm equivalence $h^{d/2}\|e_h\|_{\infty,h} \leqslant \|e_h\|_{0,h} \leqslant ch$ we can estimate the maximum norm $\|e_h\|_{\infty,h} \leqslant ch^{1-d/2}$ for $d = 1$ only.

3.3. L^∞-convergence error.

Lemma 3.3 (*Discrete maximum principle*). *(Grossmann et al. 2007, Lemma 2.39, p.68) For grid functions $u_h \in U_h$, if $-(\Delta_h u_h)(x_h) \leqslant 0$ for $x_h \in \Omega_h$, then it follows*

$$\max_{x_h \in \overline{\Omega_h}} u_h(x_h) \leqslant \max_{x_h \in \Gamma_h} u_h(x_h).$$

Lemma 3.4 (*Discrete comparison principle*). *For $u_h, v_h \in U_h$, if*

$$\begin{cases} -(\Delta_h u_h)(x_h) \leqslant -(\Delta_h v_h)(x_h) & \textit{for } x_h \in \Omega_h, \\ u_h(x_h) \leqslant v_h(x_h) \quad \textit{for } x_h \in \Gamma_h, \end{cases}$$

then $u_h(x_h) \leqslant v_h(x_h)$ for all $x_h \in \overline{\Omega_h}$.

Exercise 3.1. Derive Lemma 3.4 from the maximum principle.

Theorem 3.3 (L^∞-convergence error). *(Grossmann et al. 2007, Th 2.41, p.69) Let the solution of* (3.2) *be smooth $u \in C^4(\overline{\Omega})$ for $L = -\Delta$, then*

$$\|u_h - r_h u\|_{\infty,h} \leqslant \frac{d}{96}\|u\|_{C^4(\overline{\Omega})} h^2.$$

4. Properties of discrete operators

4.1. M-matrices. We start with the following

Definition 4.1 (M-matrix). *(Hackbusch 1992, Def.4.3.1, p.45)*

(i) $A = (a_{ij})_{i,j=1}^N \in \mathbb{R}^{N\times N}$ *is called L-matrix, if* $a_{ii} > 0$ *and* $a_{ij} \leqslant 0$ *for* $i \neq j$.

(ii) *L-matrix which is inverse monotone (exists* $A^{-1} \geqslant 0$*) is called M-matrix.*

For example, $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ is M-matrix with $A^{-1} = \frac{1}{4}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$.

Lemma 4.1 (M-criterion). *Let A be L-matrix. A is inverse monotone if and only if one of the following conditions (i)–(iv) holds.*

(i) *There exists a* ***majoring element*** $v \in \mathbb{R}^N$ *such that* $v > 0$ *and* $Av > 0$, *in this case, the stability estimate holds*

$$\sup_{\|x\|=1} \|A^{-1}x\| =: \|A^{-1}\| \leqslant \frac{\|v\|}{\min\limits_{i\in\{1,\dots,N\}} (Av)_i}. \tag{4.1}$$

(ii) *A is* ***strongly diagonally dominant****:* $a_{ii} > \sum\limits_{j\neq i} |a_{ij}|$ *for all* $i = 1, \dots, N$, *in this case,* $\|A^{-1}\| \leqslant 1 \big/ \min\limits_{i\in\{1,\dots,N\}} \big(a_{ii} - \sum\limits_{j\neq i} |a_{ij}|\big)$.

(iii) *A is* ***diagonally dominant****:* $a_{ii} \geqslant \sum\limits_{j\neq i} |a_{ij}|$ *for all* $i = 1, \dots, N$ *with strict inequality at least for one index i, and* ***irreversible****: there is no permutation matrix* $B \in \mathbb{R}^{N\times N}$ *(all entries in all rows and columns are 0 except exactly one entry 1) such that* $BAB^\top = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$ *is block upper triangular matrix.*

(iv) *The* ***spectral radius*** *(the maximal eigenvalue) of the following matrix*

$$P := I - (\mathrm{diag}(A))^{-1} A$$

is less than 1.

For (i) see (Grossmann et al. 2007, Th.2.73, p.70), for other criteria see (Hackbusch 1992, Criterion 4.3.7 and Lemma 4.3.9, p.48).

Example 4.1. Consider the ***discrete Laplace operator*** $-\Delta_h$.

- It is evidently *L-matrix*, criterion (iii) argues *M-matrix.*
- The *majoring element* $v(x) = x_1(1 - x_1) + \ldots + x_d(1 - x_d)$ is such that $v > 0$ in $\Omega = (0, 1)^d$, with the maximum norm $\|v\| = 1/4$ and $-\Delta_h v = \begin{pmatrix} 2 \\ \vdots \\ 2 \end{pmatrix}$, provides the criterion (i) and the stability estimate: $\| -\Delta_h^{-1}\| \leqslant \frac{\|v\|}{\min\limits_{i\in\{1,\dots,N\}} (-\Delta_h v)_i} = \frac{1/4}{2} = \frac{1}{8}$.
- From the consistency estimate (3.4) implying

$$|(\Delta_h(u_h - r_h u))(x_h)| = |([\Delta - \Delta_h]u)(x_h)| \leqslant \tfrac{d}{12} \|u\|_{C^4(\overline{\Omega})} h^2 =: 2\alpha$$

we evaluate the *convergence error* by applying the Cauchy–Schwarz inequality:

$$\|u_h - r_h u\|_{\infty,h} := \sup_{x_h\in\Omega_h} |(u_h - r_h u)(x_h)| = \sup_{x_h\in\Omega_h} |(-\Delta_h^{-1})(-\Delta_h)(u_h - r_h u)(x_h)|$$
$$\leqslant \| - \Delta_h^{-1}\| \sup_{x_h\in\Omega_h} |(-\Delta_h)(u_h - r_h u)(x_h)| \leqslant \frac{1}{8} 2\alpha = \frac{\alpha}{4}$$

that agrees the assertion of Theorem 3.3.

4.2. Difference stencil. Consider the discrete Laplace operator $-\Delta_h$ in 2d. For inner points x_h of a uniform grid we construct a 9-point stencil with unknown weights $c_{(i,j)}$ (see Fig. 4.2) by means of the approximation

$$L_h u_h(x_h) = \frac{1}{h^2} \sum_{i,j\in\{-1,0,1\}} c_{(i,j)} u_h(x_1 + ih, x_2 + jh).$$

After expansion in Taylor's series it has the asymptotic form:

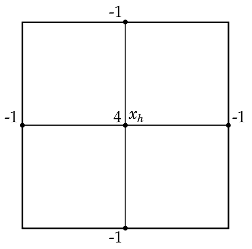


FIGURE 4.1. 5-point stencil

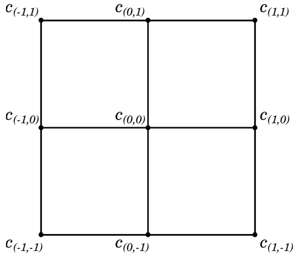


FIGURE 4.2. 9-point stencil

$$\begin{aligned} L_h u(x_h) = \frac{1}{h^2}\Big\{ & c_{(0,0)} u(x_h) + \sum_{i,j=\pm1} c_{(i,j)} \big[u(x_h) + ih\frac{\partial u}{\partial x_1}(x_h) + jh\frac{\partial u}{\partial x_2}(x_h) \\ & + ijh^2 \frac{\partial^2 u}{\partial x_1 \partial x_2}(x_h) + \frac{h^2}{2}\Delta u(x_h) + \mathrm{O}(h^3) \big] \\ & + \sum_{i=\pm1} c_{(i,0)} \big[u(x_h) + ih\frac{\partial u}{\partial x_1}(x_h) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x_1^2}(x_h) + \mathrm{O}(h^3) \big] \\ & + \sum_{j=\pm1} c_{(0,j)} \big[u(x_h) + jh\frac{\partial u}{\partial x_2}(x_h) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x_2^2}(x_h) + \mathrm{O}(h^3) \big] \Big\} \end{aligned}$$

The consistency needs the following conditions:

- by $\frac{\partial u}{\partial x_1}$: $\sum_{i=\pm1} i\big(c_{(i,0)} + \sum_{j=\pm1} c_{(i,j)}\big) = 0$,
- by $\frac{\partial u}{\partial x_2}$: $\sum_{j=\pm1} j\big(c_{(0,j)} + \sum_{i=\pm1} c_{(i,j)}\big) = 0$,
- by $\frac{\partial^2 u}{\partial x_1 \partial x_2}$: $\sum_{i,j=\pm1} ijc_{(i,j)} = 0$,

and, due to the symmetry $c_{(\pm1,0)} = c_{(0,\pm1)} =: c_1$ and $c_{(-1,\pm1)} = c_{(1,\pm1)} =: c_2$ also:

- by u: $\sum_{i,j=\pm1,0} c_{(i,j)} = 0$,
- by $\frac{\partial^2 u}{\partial x_1^2}$: $\sum_{i=\pm1}\big(c_{(i,0)} + \sum_{j=\pm1} c_{(i,j)}\big) = 2c_1 + 4c_2$,
- by $\frac{\partial^2 u}{\partial x_2^2}$: $\sum_{j=\pm1}\big(c_{(0,j)} + \sum_{i=\pm1} c_{(i,j)}\big) = 2c_1 + 4c_2$.

Then $c_0 + 4c_1 + 4c_2 = 0$, where $c_0 := c_{(0,0)}$, and we calculate

$$L_h u(x_h) = \frac{1}{h^2}\frac{h^2}{2}(2c_1 + 4c_2)\Delta u(x_h) + \mathrm{O}(h)$$

which for $c_1 + 2c_2 = -1$ describes the *consistent 9-point stencil* (see Fig. 4.3). To

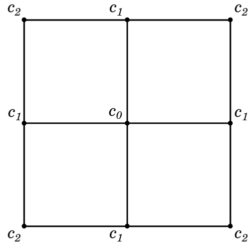


FIGURE 4.3. A consistent 9-point stencil

compare, $c_2 = 0$, $c_1 = -1$, $c_0 = 4$ implies the standard 5-point stencil (see Fig. 4.1).

EXAMPLE 4.2. Consider $c_2 = -\frac{1}{6}$, $c_1 = -\frac{2}{3}$, $c_0 = \frac{10}{3}$.

In this case, the expansion refined with high-order asymptotic terms reads

$$\begin{aligned} L_h u(x_h) = -\Delta u(x_h) - \frac{1}{h^2}\Big\{\frac{2}{3}\Big[\frac{h^4}{4!}\Big(2\frac{\partial^4 u}{\partial x_1^4}(x_h) + 2\frac{\partial^4 u}{\partial x_2^4}(x_h)\Big) + \mathrm{O}(h^6)\Big] \\ + \frac{1}{6}\Big[\frac{h^4}{4!}\Big(4\frac{\partial^4 u}{\partial x_1^4}(x_h) + 4\frac{\partial^4 u}{\partial x_2^4}(x_h)\Big) + \frac{h^4}{2!2!}4\frac{\partial^4 u}{\partial x_1^2 \partial x_2^2}(x_h) + \mathrm{O}(h^6)\Big]\Big\} \\ = -\Delta u(x_h) - \frac{1}{h^2}\Big\{\frac{h^4}{12}\Big(\frac{\partial^4 u}{\partial x_1^4} + \frac{\partial^4 u}{\partial x_2^4} + 2\frac{\partial^4 u}{\partial x_1^2 \partial x_2^2}\Big)(x_h) + \mathrm{O}(h^6)\Big\} \\ = -\Delta u(x_h) - \frac{h^2}{12}\Delta^2 u(x_h) + \mathrm{O}(h^4). \end{aligned}$$

Moreover, we obtain the *4-order consistency* when we approximate f as

$$\begin{aligned} f_h = \frac{1}{12}\Big[8f(x_h) + \sum_{i=\pm 1} f(x_h + ihe^1) + \sum_{j=\pm 1} f(x_h + jhe^2)\Big] \\ = \frac{1}{12}\Big[8f(x_h) + 2f(x_h) + h^2\frac{\partial^2 u}{\partial x_1^2}(x_h) + \mathrm{O}(h^4) + 2f(x_h) + h^2\frac{\partial^2 u}{\partial x_2^2}(x_h) + \mathrm{O}(h^4)\Big] \\ = f(x_h) + \frac{h^2}{12}\Delta f(x_h) + \mathrm{O}(h^4). \end{aligned}$$

4.3. Shortley–Weller difference operator. For boundary points x_h given by parameters $s_N, s_W, s_S, s_E \in (0,1]$, see Fig. 4.4, the Shortley–Weller operator:

$$\begin{aligned} -\Delta_h^{SW} u_h(x_h) = \frac{2}{h^2}\Big[\Big(\frac{1}{s_E s_W} + \frac{1}{s_S s_N}\Big)u_h(x_h) \\ - \frac{1}{s_W(s_E + s_W)}u_h(x_h + s_W h e^1) - \frac{1}{s_E(s_E + s_W)}u_h(x_h - s_E h e^1) \\ - \frac{1}{s_N(s_S + s_N)}u_h(x_h + s_N h e^2) - \frac{1}{s_S(s_S + s_N)}u_h(x_h - s_S h e^2)\Big]. \end{aligned}$$

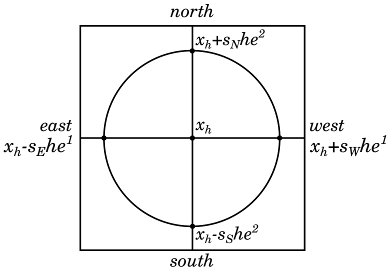


FIGURE 4.4. Shortley–Weller difference operator

LEMMA 4.2. *(Hackbusch 1992, Th.4.8.4, p.81) The Shortley–Weller difference operator has the following properties:*

(i) *if $s_N = s_W = s_S = s_E = 1$, then $-\Delta_h^{SW} = -\Delta_h$;*
(ii) *consistency error is of order one: $\|\big([\Delta - \Delta_h^{SW}]u\big)(x_h)\| = \mathrm{O}(h)$;*
(iii) *it is M-matrix;*
(iv) *for the domain $\Omega \subset (x_1^0, x_1^0 + X) \times \mathbb{R}$ inside a strip, the majoring element $v = \frac{1}{2}(x_1 - x_1^0)(x_1^0 + X - x_1)$ provides estimate of the inverse operator: $\|(-\Delta_h^{SW})^{-1}\| \leqslant \frac{X^2}{8}$.*

4.4. Discretization of mixed derivative. The 9-point stencil with weights $c_{(i,j)}$ from Fig. 4.2 is consistent with the mixed derivative, when

$$\frac{1}{h^2} \sum_{i,j \in \{-1,0,1\}} c_{(i,j)} u_h(x_1 + ih, x_2 + jh) = \frac{\partial^2 u}{\partial x_1 \partial x_2}(x_h) + \mathrm{O}(h).$$

The form yields any consistent stencil, see (Grossmann et al. 2007, p.84):

$$S_h^{12} := \frac{1}{4h^2} \begin{bmatrix} -1 + a - b + c & 2(-a - c) & 1 + a + b + c \\ 2(-a + b) & 4a & 2(-a - b) \\ 1 + a - b - c & 2(-a + c) & -1 + a + b - c \end{bmatrix}$$

of the 1st-order, and 2nd-order if $b = c = 0$. The common choice is $a = b = c = 0$.

For the elliptic operator in non-divergence form compared to (3.1):

$$Lu = -a_{11}\frac{\partial^2 u}{\partial x_1^2} - 2a_{12}\frac{\partial^2 u}{\partial x_1 \partial x_2}, -a_{22}\frac{\partial^2 u}{\partial x_2^2},$$

by summation over points in the patch with the following stencil

$$\frac{1}{h^2} \begin{bmatrix} -a_{12}^- & -a_{22} + |a_{12}| & -a_{12}^+ \\ -a_{11} + |a_{12}| & 2(a_{11} + a_{22} - |a_{12}|) & -a_{11} + |a_{12}| \\ -a_{12}^+ & -a_{22} + |a_{12}| & -a_{12}^- \end{bmatrix},$$

where $a_{12}^+ = \max(0, a_{12}) \geq 0$ and $a_{12}^- = -\min(0, a_{12}) \geq 0$, constitutes an M-matrix L_h if the coefficients satisfy $a_{11} > |a_{12}|$ and $a_{22} > |a_{12}|$. Here we use $b = c = 0$, and either $a = -1$ as $a_{12} < 0$, or $a = 1$ as $a_{12} > 0$ such that

$$\text{either } S_h^{12} = \frac{1}{4h^2} \begin{bmatrix} -2 & 2 & 0 \\ 2 & -4 & 2 \\ 0 & 2 & -2 \end{bmatrix}, \quad \text{or } S_h^{12} = \frac{1}{4h^2} \begin{bmatrix} 0 & -2 & 2 \\ -2 & 4 & -2 \\ 2 & -2 & 0 \end{bmatrix}.$$

5. Finite Volume Method (FVM) on unstructured grids

Let $\Omega \subset \mathbb{R}^d$ be a convex polyhedron obeying flat faces. We distinguish

$$\begin{cases} \textit{interior grid points } x_i \in \Omega,\ i = 1, \dots, N \text{ and} \\ \textit{boundary grid points } x_i \in \Gamma,\ i = N+1, \dots, \bar{N}, \end{cases}$$

by the mean of index sets $J := \{1, \dots, N\}$ and $\bar{J} := \{1, \dots, \bar{N}\}$.

Definition 5.1. *For $i \in J$, a* ***Voronoi box*** *is called the geometric set*

$$\Omega_i = \bigcap_{j \in \bar{J},\, j \neq i} B_{ij} := \{x \in \Omega : \quad \|x - x_i\| < \|x - x_j\|\}. \tag{5.1}$$

Assume that $\overline{\Omega_i} \cap \Gamma = \emptyset$. We can define

- *midpoint* $x_{ij} := \frac{x_i + x_j}{2}$ such that $\|x_i - x_{ij}\| = \|x_j - x_{ij}\| = \frac{\|x_i - x_j\|}{2}$,
- *hyper-plane* $\partial B_{ij} \supset \overline{\Omega_i} \cap \overline{\Omega_j} =: \Gamma_{ij}$ with the *unit normal vector* $n_{ij} = \frac{x_i - x_j}{\|x_i - x_j\|}$,
- *boundary* $\Gamma_i := \bigcup_{j \in N_i} \Gamma_{ij}$ of the cell $\Omega_i = \bigcap_{j \in N_i} B_{ij}$ for *essential neighbours*

$$N_i = \{j \in \bar{J},\ j \neq i : \ (d-1)\text{-dimensional measure } |\overline{\Omega_i} \cap \overline{\Omega_j}| \neq 0\}.$$

5.1. Discetization of Poisson equation. We consider the *Dirichlet problem*: For $f(x) \in C(\Omega)$ and $g(x) \in C(\Gamma)$ find $u(x) \in C(\overline{\Omega}) \cap C^2(\Omega)$ such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma. \end{cases} \tag{5.2}$$

Integrate (5.2) by parts over the Voronoi box Ω_i with the unit normal vector $n_i = \bigcup_{j \in N_i} n_{ij}$ at $\partial\Omega_i$ which is outward to Ω_i, to derive:

$$\int_{\Omega_i} f\, dx = -\int_{\Omega_i} \Delta u\, dx = -\int_{\Gamma_i} \frac{\partial u}{\partial n_i}\, dS_x = -\sum_{j \in N_i} \int_{\Gamma_{ij}} \frac{\partial u}{\partial n_{ij}}\, dS_x.$$

Denoting by $|\cdot|$ the Hausdorff measure of a set and by $\operatorname{diam}(\cdot)$ its diameter, the following expansion holds

$$\int_{\Omega_i} f\, dx = |\Omega_i| f(x_i) + \mathrm{O}(|\Omega_i| \operatorname{diam}(\Omega_i)),$$

and similarly

$$\int_{\Gamma_{ij}} \frac{\partial u}{\partial n_{ij}}\, dS_x = |\Gamma_{ij}| \frac{\partial u}{\partial n_{ij}}(x_{ij}) + \mathrm{O}(|\Gamma_{ij}| \|x_i - x_j\|).$$

We apply the Taylor series:

$$u(x_j) = u(x_{ij}) + \|x_j - x_{ij}\| \frac{\partial u}{\partial n_{ij}}(x_{ij}) + \frac{1}{2}\|x_j - x_{ij}\|^2 \frac{\partial^2 u}{\partial n_{ij}^2}(x_{ij}) + \mathrm{O}(\|x_j - x_{ij}\|^3),$$

$$u(x_i) = u(x_{ij}) - \|x_i - x_{ij}\| \frac{\partial u}{\partial n_{ij}}(x_{ij}) + \frac{1}{2}\|x_i - x_{ij}\|^2 \frac{\partial^2 u}{\partial n_{ij}^2}(x_{ij}) + \mathrm{O}(\|x_i - x_{ij}\|^3)$$

and subtract them recalling $\|x_i - x_{ij}\| = \|x_j - x_{ij}\| = \frac{\|x_i - x_j\|}{2}$, such that

$$u(x_j) - u(x_i) = \|x_j - x_i\| \frac{\partial u}{\partial n_{ij}}(x_{ij}) + \mathrm{O}(\|x_j - x_i\|^3),$$

hence the normal derivative admits the expansion:

$$\frac{\partial u}{\partial n_{ij}}(x_{ij}) = \frac{u(x_j) - u(x_i)}{\|x_j - x_i\|} + \mathrm{O}(\|x_j - x_i\|^2).$$

Based on the above representations, we discretize the problem (5.2) as follows

$$\begin{cases} -\sum\limits_{j\in N_i} \frac{|\Gamma_{ij}|}{\|x_j - x_i\|}(u_j - u_i) = |\Omega_i| f(x_i), & i \in J, \\ u_i = g(x_i), & i \in \bar{J} \setminus J, \end{cases} \tag{5.3}$$

or, equivalently, in the matrix form $\sum_{j\in J} L_{ij} u_j = f_i, \quad i \in J$, with the coefficients

$$L_{ij} = \begin{cases} \sum\limits_{k\in N_i} \frac{|\Gamma_{ik}|}{\|x_k - x_i\|} & \text{if } j = i \in J, \\ -\frac{|\Gamma_{ij}|}{\|x_j - x_i\|} & \text{if } j \in N_i \cap J, \\ 0 & \text{otherwise,} \end{cases} \qquad f_i = |\Omega_i| f(x_i) + \sum_{k\in N_i \text{ at } \Gamma} \frac{|\Gamma_{ik}|}{\|x_k - x_i\|} g(x_k),$$

where the latter term is according to the inhomogeneous BC at Γ. Gathering the components $u_h = (u_i)_{i\in J}$, $f_h = (f_i)_{i\in J} \in \mathbb{R}^N$, and $L_h = (L_{ij})_{i,j\in J} \in \mathbb{R}^{N\times N}$ implies the *linear system*: $L_h u_h = f_h$.

Lemma 5.1. *(Grossmann et al. 2007, Lemma 2.60, p.94) L_h is M-matrix.*

Example 5.1. For *uniform grid* of the size h in $\Omega = (0,1)^2$, the Voronoi box is $\Omega_i = \big(x_i - \frac{h}{2}e^1, x_i + \frac{h}{2}e^1\big) \times \big(x_i - \frac{h}{2}e^2, x_i + \frac{h}{2}e^2\big)$ with $|\Omega_i| = h^2$, $\|x_i - x_j\| = h$, $|\Gamma_{ij}| = h$, the cardinality $|N_i| = 4$ for all $i \in J$ and $j \in \bar{J}$, thus

$$L_{ij} = \begin{cases} 4 & \text{if } j = i \in J, \\ -1 & \text{if } j \in N_i \cap J, \\ 0 & \text{otherwise,} \end{cases} \qquad f_i = h^2 f(x_i) + \sum_{k\in N_i \text{ at } \Gamma} g(x_k).$$

In this case, $L_h = -\Delta_h$ agrees the standard 5-point stencil.

5.2. L^∞-convergence error. On a uniform grid we have $|\Gamma_{ij}| = h^{d-1}$ in $\mathbb{R}^d$. Therefore, define the ***discretization parameter*** (mesh size) h by

$$h := \max_{i\in J} h_i, \quad h_i := \big(\max_{j\in N_i} |\Gamma_{ij}|\big)^{\frac{1}{d-1}}, \quad d \geq 2.$$

Theorem 5.1 (L^∞-convergence error). *(Grossmann et al. 2007, Lemma 2.63, p.96; Th.2.65, p.98) Let the solution of* (5.2) *be smooth $u \in C^4(\overline{\Omega})$, the areas of cells $(\Omega_i)_{i\in J}$ be uniformly bounded from below:*

$$|\Omega_i| \geqslant c_1 h_i^d, \quad i \in J, \quad c_1 > 0, \tag{5.4}$$

and the consistency estimate holds:

$$\Big|\sum_{j\in N_i} \frac{|\Gamma_{ij}|}{\|x_j - x_i\|}(u(x_j) - u(x_i)) + |\Omega_i| f(x_i)\Big| \leqslant c_2 h_i^{d+1}, \quad i \in J, \quad c_2 > 0. \tag{5.5}$$

Then the convergence error is of order h:

$$\|e_h\|_{\infty,h} = \|r_h u - u_h\|_{\infty,h} \leqslant ch, \quad c > 0.$$

Note that the left-hand side of (5.5) is $|(-L_h u)(x_h) + f(x_h)| = |([L - L_h]u)(x_h)|$ in Ω_i, thus its maximum over $i \in J$ implies the L^∞-consistency $\|[L - L_h]u\|_{\infty,h}$.

Remark 5.1. For the Neumann BC: $\partial u/\partial n = g$ on Γ it holds:

$$\int_{\Omega_i} f\,dx = -\sum_{j\in N_i}\int_{\Gamma_{ij}}\frac{\partial u}{\partial n_{ij}}\,dS_x - \sum_{k\in N_i \text{ at } \Gamma}\int_{\Gamma_{ik}}\frac{\partial u}{\partial n_{ik}}\,dS_x$$
$$= -\sum_{j\in N_i}\int_{\Gamma_{ij}}\frac{\partial u}{\partial n_{ij}}\,dS_x - \sum_{k\in N_i \text{ at } \Gamma}\int_{\Gamma_{ik}} g\,dS_x.$$

Exercise 5.1. Generalize FVM (5.3) for the inhomogeneous medium:

$$-\mathrm{div}(a(x)\nabla u) = f \quad \text{in } \Omega.$$

5.3. FVM for low-order convection terms. Consider the stationary diffusion-convection equation given in the divergence form: find $u(x)\in C(\overline{\Omega})\cap C^2(\Omega)$ such that

$$Lu(x) := -\mathrm{div}\big(a(x)\nabla u(x) - b(x)\cdot u(x)\big) = f(x) \quad \text{in } \Omega\subset\mathbb{R}^d,$$

where the right-hand side $f(x)\in C(\Omega)$, coefficients $a(x), b_i(x)\in C^1(\Omega)$ with vector $b=(b_1\dots,b_d)$, and $a>0$.

Using Voronoi diagram on unstructured grids we generalize FVM. Denote by $\beta_{ij} := b\cdot n_{ij}$ on the box sides Γ_{ij} and decompose $\beta_{ij} = \beta^+_{ij} - \beta^-_{ij}$ into the positive $\beta^+_{ij}\ge 0$ and the negative parts $\beta^-_{ij}\ge 0$, the 1-st order consistent approximation

$$(L_h u(x_h))_i = -\sum_{j\in N_i}\Big\{\Big[\frac{|\Gamma_{ij}|\,a(x_{ij})}{\|x_j-x_i\|} + \beta^-_{ij}\Big]u(x_j) - \Big[\frac{|\Gamma_{ij}|a(x_{ij})}{\|x_j-x_i\|} + \beta^+_{ij}\Big]u(x_i)\Big\}$$

constitutes L-matrix L_h.

6. FDM for parabolic IBVP

In the 1d-spatial domain $\Omega=(0,1)$, we consider a *parabolic equation*

$$\big[\frac{\partial}{\partial t} + L\big]u = f.$$

Namely, in the cylinder $Q_T = (0,T)\times\Omega$ we look for $u(t,x)\in C^{2,1}(Q_T)\cap C(\overline{Q_T})$ satisfying

$$(6.1)\quad \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = f & \text{for } t\in(0,T),\ x\in\Omega \quad (\textbf{\textit{diffusion/heat equation}}),\\ u(t,x) = g(t,x) & \text{for } t\in(0,T),\ x\in\partial\Omega := \{0,1\} \quad (\textit{Dirichlet BC}),\\ u(0,x) = u^0(x) & \text{for } t=0,\ x\in\Omega=(0,1) \quad (\textit{IC}). \end{cases}$$

6.1. Full discretization. For equidistant grid points: $x_i = ih$, $i=0,\dots,N$, $h=1/N$ and $t^k = k\tau$, $k=0,\dots,M$, $\tau = T/M$, we discretize (6.1) by *6-point schemes* with an extra parameter $\sigma\in[0,1]$: Find a discrete function $u_{(\tau,h)} = (u^k_i)_{i=0,\dots,N,\,k=0,\dots,M}$ such that

$$[D^+_\tau + L^\sigma_h]u_{(\tau,h)} = f_{(\tau,h)}$$

by means of the relations

$$(6.2)\qquad \begin{cases} \dfrac{u_i^{k+1}-u_i^k}{\tau} - D_h^- D_h^+(\sigma u_i^{k+1} + (1-\sigma)u_i^k) = f_i^k, \\ \qquad\qquad\qquad\qquad i=0,\dots,N, \quad k=0,\dots,M, \\ u_0^k = g_1(t^k),\ u_N^k = g_2(t^k), \qquad k=0,\dots,M, \\ u_i^0 = u^0(x_i), \qquad i=0,\dots,N. \end{cases}$$

Set $\gamma := \tau/h^2$ for short. From (6.2) we have an equivalent *σ-dependent scheme*:

$$(6.3)\quad (1+2\sigma\gamma)u_i^{k+1} - \sigma\gamma(u_{i-1}^{k+1} + u_{i+1}^{k+1}) \\ = (1-2(1-\sigma)\gamma)u_i^k + (1-\sigma)\gamma(u_{i-1}^k + u_{i+1}^k) + \tau f_i^k =: F_i^k.$$

In particular, it establishes different stencils (see Fig. 6.1):

- ($\sigma = 0$) ***explicit (forward) Euler scheme:***
$$u_i^{k+1} = (1-2\gamma)u_i^k + \gamma(u_{i-1}^k + u_{i+1}^k) + \tau f(t^k, x_i).$$
- ($\sigma = 1$) ***implicit (backward) Euler scheme:***
$$(1+2\gamma)u_i^{k+1} - \gamma(u_{i-1}^{k+1} + u_{i+1}^{k+1}) = u_i^k + \tau f(t^{k+1}, x_i).$$
- ($\sigma = 1/2$) ***Crank–Nicolson scheme:***

$$2(1+\gamma)u_i^{k+1} - \gamma(u_{i-1}^{k+1} + u_{i+1}^{k+1}) \\ = 2(1-\gamma)u_i^k + \gamma(u_{i-1}^k + u_{i+1}^k) + 2\tau f(t^k + \tau/2, x_i).$$

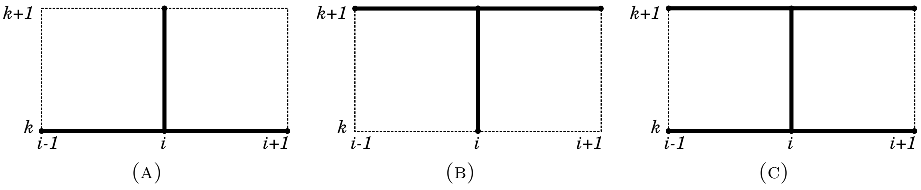


FIGURE 6.1. (A) explicit Euler, (B) implicit Euler, (C) Crank–Nicolson.

6.2. Error estimates.

LEMMA 6.1 (L^∞-consistency error). *(Grossmann et al. 2007, Lemma 2.73, p.105) Let the solution of* (6.1) *be smooth such that $u \in C^{3,4}(\overline{Q_T})$, then the consistency error estimate holds:*

$$\|[D_\tau^+ + L_h^\sigma](r_{(\tau,h)}u) - f_{(\tau,h)}\|_{\infty,(\tau,h)} = \mathrm{O}(|\sigma - 1/2|\tau + \tau^2 + h^2).$$

LEMMA 6.2. *For the linear system* (6.2)*, stability and a-priori estimates are equivalent.*

LEMMA 6.3 (L^∞-stability error). *(Grossmann et al. 2007, p.106-107) If the* ***Courant–Friedrichs–Lewy (CFL) condition*** *holds, namely:*

$$(1-\sigma)\gamma = (1-\sigma)\tau/h^2 \leqslant 1/2,$$

then the stability estimate holds:

$$\max_{i\in\{0,\dots,N\},\,k\in\{0,\dots,M-1\}} |u_i^{k+1}| \leqslant \max_{i\in\{0,\dots,N\}} |u_i^0| + \tau \sum_{j=0}^{M-1} \max_{i\in\{0,\dots,N\}} |f_i^j|.$$

Theorem 6.1 (L^∞-convergence error). *(Grossmann et al. 2007, Th.2.74, p.107) Let the solution of* (6.1) *be smooth* $u \in C^{3,4}(\overline{\Omega})$*, and* (τ, h)*-discretization* (6.2) *is such that the CFL-condition holds. The convergence error admits the estimate:*

$$\max_{i\in\{0,\dots,N\},\,k\in\{0,\dots,M-1\}} |u(t^k, x_i) - u_i^k| = \mathrm{O}(|\sigma - 1/2|\tau + \tau^2 + h^2).$$

6.3. Semi-discretization. The *vertical method of lines (MOL)*: Find the vector $u(t) = (u_1(t), \dots, u_{N-1}(t))$ such that

$$\begin{cases} \dfrac{du_i}{dt} - \dfrac{1}{h^2}(u_{i+1} + u_{i-1} - 2u_i) = f_i, & i = 1, \dots, N-1, \\ u_0 \equiv g(0), \quad u_N \equiv g(1), & \\ u_i(0) = u^0(x_i) & i = 1, \dots, N-1 \end{cases}$$

implies the system of $(N-1)$ first-order linear ODE supported by $(N-1)$ initial conditions, see illustration in Fig. 6.2.

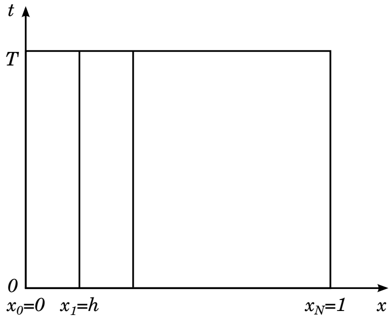


Figure 6.2. Vertical method of lines.

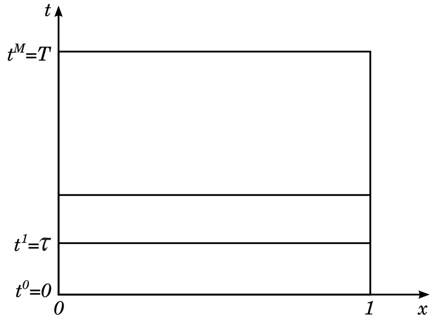


Figure 6.3. Horizontal methods of lines.

While the *horizontal (Rothe's) MOL*: Find the vector $u(x) = (u^1(x), \dots, u^M(x))$ such that

$$\begin{cases} \dfrac{u^{k+1} - u^k}{\tau} - \dfrac{d^2 u^{k+1}}{dx^2} = f^k, & k = 0, \dots, M-1, \\ u^{k+1}(0) = g(0), \quad u^{k+1}(1) = g(1), & k = 0, \dots, M-1, \\ u^0 \equiv u^0 \end{cases}$$

results in the system of M second-order elliptic PDE supported by the Dirichlet BC, see illustration in Fig. 6.3.

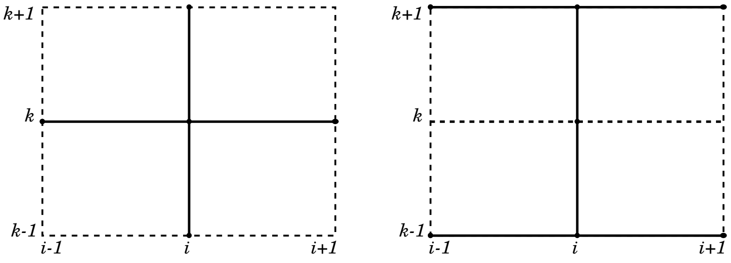
7. FDM for hyperbolic IBVP

7.1. The 1d wave equation on uniform grid. Consider the *wave equation*:

$$[\frac{\partial^2}{\partial t^2} - \Delta]u = f$$

and the respective hyperbolic IBVP: Find $u(t,x) \in C^2(Q_T) \cap C^{1,0}(\overline{Q_T})$ in the 1d-spatial domain $\Omega = (0,1)$ such that

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2} = f(t,x) & \text{for } t \in (0,T),\ x \in \Omega, \\ u(t,x) = g(t,x) & \text{for } t \in (0,T),\ x \in \Gamma := \{0,1\}, \\ u(0,x) = u^0(x),\ \frac{\partial u}{\partial t}(0,x) = v^0(x) & \text{for } x \in \Omega. \end{cases} \tag{7.1}$$



(A) (B)

FIGURE 7.1. (A) explicit and (B) implicit schemes for hyperbolic IBVP.

We discretize $[D^-_\tau D^+_\tau + L_h]u_{(\tau,h)} = f_{(\tau,h)}$ by means of the schemes, see Fig. 7.1:

(E) 5-point **explicit scheme**:

$$\frac{u_i^{k+1} + u_i^{k-1} - 2u_i^k}{\tau^2} = \frac{u_{i+1}^k + u_{i-1}^k - 2u_i^k}{h^2} + f_i^k, \tag{7.2}$$

or, in the equivalent form with the *Courant number* $\gamma = \tau/h$:

$$u_i^{k+1} = -u_i^{k-1} + 2(1-\gamma^2)u_i^k + \gamma^2(u_{i+1}^k + u_{i-1}^k) + \tau^2 f_i^k;$$

(I) 7-point **implicit scheme**:

$$\frac{u_i^{k+1} + u_i^{k-1} - 2u_i^k}{\tau^2} - \frac{u_{i+1}^{k+1} + u_{i-1}^{k+1} - 2u_i^{k+1}}{2h^2} - \frac{u_{i+1}^{k-1} + u_{i-1}^{k-1} - 2u_i^{k-1}}{2h^2} = f_i^k, \tag{7.3}$$

or, equivalently,

$$(1 + \tfrac{\gamma^2}{2})u_i^{k+1} - \tfrac{\gamma^2}{2}(u_{i+1}^{k+1} + u_i^{k+1}) = 2u_i^k - (1+\gamma^2)u_i^{k-1} + \tfrac{\gamma^2}{2}(u_{i+1}^{k-1} + u_{i-1}^{k-1}) + \tau^2 f_i^k.$$

The discrete equations are supported by the discrete initial data:

$$u_i^0 = u^0(x_i), \quad u_i^1 = u_i^0 + \tau v^0(x_i).$$

7.2. Three-term recurrence and stability. Consider the second order difference equation (three-term recurrence relation): given coefficients a_1, a_2, initial data U^0, U^1, and right-hand side $\tau^2 F^k \in \mathbb{R}$, find $U^k \in \mathbb{R}$ such that

$$U^k + a_1 U^{k-1} + a_2 U^{k-2} = \tau^2 F^k \quad \text{for } k \geq 2.$$

The characteristic equation $\rho^2 + a_1\rho + a_2 = 0$ has two complex roots $\rho_1, \rho_2 \in \mathbb{C}$. We express U^k as the sum $U^k = \overline{U}^k + \tilde{U}^k$, where $\overline{U}^0 = U^0$, $\overline{U}^1 = U^1$, and $\overline{U}^k$ solves the homogeneous equation

$$\overline{U}^k + a_1 \overline{U}^{k-1} + a_2 \overline{U}^{k-2} = 0 \quad \text{for } k \geq 2.$$

Whereas $\tilde{U}^k$ satisfies zero initial data $\tilde{U}^0 = \tilde{U}^1 = 0$ and the recurrence equation

$$\tilde{U}^k + a_1 \tilde{U}^{k-1} + a_2 \tilde{U}^{k-2} = \tau^2 F^k \quad \text{for } k \geq 2.$$

Lemma 7.1 (Stability of three-term recurrence). *(Rannacher 2008, Satz 6.1, S.225) Let* $|\rho_1| \leq 1$ *and* $|\rho_2| \leq 1$*. Then* ***stability error estimate*** *holds:*

$$|\overline{U}^k| \leqslant c \max\{|U^0|, |U^1|\}, \quad c > 0, \quad |\tilde{U}^k| \leqslant \tau^2 k^2 \max_{j=2,\dots,k} |F^j|. \tag{7.4}$$

With the help of Lemma, for the discretizations of the wave equation we establish consistency, stability, and the following convergence results.

7.3. von Neumann convergence analysis.

Theorem 7.1 (Convergence error). *(Rannacher 2008, Satz 6.2, S.223) Let the solution of* (7.1) *be smooth* $u \in C^4(\overline{Q_T})$ *and the Courant–Friedrichs–Lewy (CFL) condition* $\gamma := \tau/h \leqslant 1$ *hold for the explicit scheme* (7.2)*, then both discretizations* (7.2) *and* (7.3) *obey the convergence error*

$$\max_{t^k \in [0,T]} \|u(t^k) - u^k\|_{0,h} = \mathrm{O}(\tau^2 + h^2).$$

7.4. First order hyperbolic system. In 1d-spatial cylinder $Q_T = (0,T) \times (0,1)$, for given coefficient $a \in \mathbb{R}$, continuous functions f, initial data v^0, and continuously differentiable u^0, consider IP for the wave equation from (7.1):

$$\frac{\partial^2 u}{\partial t^2} - a^2 \frac{\partial^2 u}{\partial x^2} = f \text{ in } Q_T, \quad u = u^0, \; \frac{\partial u}{\partial t} = v^0 \text{ as } t = 0.$$

Introducing new variables $v_1 = \partial u/\partial t$ and $v_2 = a\,\partial u/\partial x$ such that

$$\frac{\partial v_1}{\partial t} = \frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} = a \frac{\partial v_2}{\partial x}, \quad \frac{\partial v_2}{\partial t} = a \frac{\partial^2 u}{\partial t \partial x} = a \frac{\partial v_1}{\partial x},$$

we can express it as a 1st-order linear hyperbolic system for the pair $v(t,x) = (v_1, v_2)^\top \in C^1(Q_T)^2 \cap C(\overline{Q_T})^2$ with a symmetric 2×2 matrix A:

$$\frac{\partial v}{\partial t} + A \frac{\partial v}{\partial x} = 0, \quad A = \begin{pmatrix} 0 & -a \\ -a & 0 \end{pmatrix}, \quad v = \begin{pmatrix} v^0 \\ a\frac{\partial u^0}{\partial x} \end{pmatrix} \text{ at } t = 0.$$

For periodic data such that $f(t, x+1) = f(t,x)$, $\partial u^0/\partial x(x+1) = \partial u^0/\partial x(x)$, $v^0(x+1) = v^0(x)$ in $\overline{Q_T}$, we complete the system with periodic boundary conditions:

$$v(t,0) = v(t,1) \quad \text{for } t \in (0,T).$$

In this way we arrive at the form of conservation laws following next.

8. FDM for linear conservation law

Let a vector-function $f = (f_1, \dots, f_d)^\top : \mathbb{R}^{d+2} \mapsto \mathbb{R}^d$ and the initial data $u^0 \in W^{1,\infty}(\mathbb{R}^d)$ be given. Consider the *first order, nonlinear PDE* problem: Find $u(t,x) \in W^{1,\infty}(\mathbb{R}_+ \times \mathbb{R}^d)$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t} + \mathrm{div}_x f(t,x,u) = 0 \quad \text{a.e. } t > 0, x \in \mathbb{R}^d, \\ u(0,x) = u^0(x) \quad \text{a.e. } x \in \mathbb{R}^d, \end{cases} \tag{8.1}$$

where

$$\mathrm{div}_x f(t,x,u) = \sum_{i=1}^{d} \Big(\frac{\partial f}{\partial x_i} + \frac{\partial f}{\partial u} \frac{\partial u}{\partial x_i} \Big).$$

If f is differentiable, then it can be rewritten as a nonlinear *transport problem (TP)*:

$$\frac{\partial u}{\partial t} + v \cdot \nabla_x u = F := -\sum_{i=1}^{d} \frac{\partial f}{\partial x_i}$$

with the *velocity* $v := \frac{\partial f}{\partial u}(t,x,u)$.

In the linear case, $v(t,x)$ and $F(t,x)$ do not depend on u.

8.1. Linear transport problem in 1d. For a constant velocity $v > 0$, consider the *linear transport problem in 1d*: Find $u(t,x) \in W^{1,\infty}(\mathbb{R}_+ \times \mathbb{R})$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t} + v \dfrac{\partial u}{\partial x} = 0 \quad \text{a.e. } t > 0, x \in \mathbb{R}, \\ u(0,x) = u^0(x) \quad \text{a.e. } x \in \mathbb{R}. \end{cases} \tag{8.2}$$

Lemma 8.1 (Solution formula and maximum principle). *A unique solution to* (8.2) *is given by the explicit formula*

$$u(t,x) = u^0(x - vt) \tag{8.3}$$

and satisfies the maximum principle:

$$\inf_{x \in \mathbb{R}} u^0(x) \leqslant u(t,x) \leqslant \sup_{x \in \mathbb{R}} u^0(x).$$

Exercise 8.1. Justify (8.3) by differentiation, take infimum and supremum.

On the uniform grid $(t^k, x_i) = (k\tau, ih)$ for $k \in \mathbb{N}$, $i \in \mathbb{Z}$, explicit discretization of (8.2) with respect to time implies *upwind schemes*:

$$D_\tau^+ u = -v[\sigma D_h^+ + (1-\sigma) D_h^-]u, \quad \sigma \in [0,1]$$

with the particular stencils illustrated in Fig. 8.1.

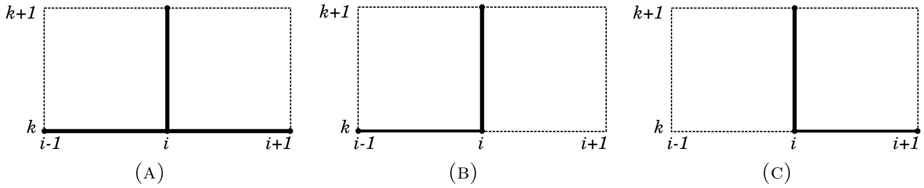


Figure 8.1. Stencils for (A) $\sigma = 1/2$, (B) $\sigma = 0$, (C) $\sigma = 1$.

8.2. L^∞-error analysis.

Theorem 8.1 (L^∞-error analysis). *(Grossmann et al. 2007, Th.2.7, p.39) Let $u^0 \in C^{1,1}(\mathbb{R})$ have a Lipschitz continuous derivative. In the maximum norm, the **explicit upwind scheme** ($\sigma = 0$): Find $u_{(\tau,h)} = (u_i^k)_{i\in\mathbb{Z},k\in\mathbb{N}}$ such that*

$$\begin{cases} [D_\tau^+ + vD_h^-]u_{(\tau,h)} = 0, \\ u_{(\tau,h)}(0,x_h) = u^0(x_h) \end{cases} \tag{8.4}$$

is consistent of order 1. Under the CFL-condition $\gamma := v\tau/h \leqslant 1$, the discrete maximum principle holds:

$$\inf_{i\in\mathbb{Z}} u_i^k \leqslant u_i^k \leqslant \sup_{i\in\mathbb{Z}} u_i^k,$$

then (8.4) *is stable and convergent of order 1 in any finite time interval $t \in [0,T]$.*

8.3. Fourier stability analysis. Take a countable complex-valued *basis* in the space $L^2(-\pi,\pi)$:

$$v_j(x) = \sqrt{\frac{h}{2\pi}}e^{\imath jx}, \quad j \in \mathbb{Z}, \quad \imath^2 = -1, \tag{8.5}$$

which is orthogonal such that $\int_{-\pi}^{\pi} v_j\overline{v_l}\,dx = h\delta_{jl}$ with the complex conjugate function $\overline{v_l(x)} = \sqrt{\frac{h}{2\pi}}e^{-\imath lx}$ and the Kronecker delta $\delta_{jl} = \begin{cases} 1, \ j = l \\ 0, \ j \neq l \end{cases}$. For a grid function $u_h = (u_j)_{j\in\mathbb{Z}}$ on the equidistant space grid $x_h = (jh)_{j\in\mathbb{Z}}$, we define the *discrete Fourier transform*:

$$\mathcal{F} : l^2 := \{u_h : \sum_{j\in\mathbb{Z}} |u_j|^2 < \infty\} \mapsto L^2(-\pi,\pi), \quad \mathcal{F}u_h(x) = \sum_{j\in\mathbb{Z}} u_j v_j(x),$$

and for $u \in L^2(-\pi,\pi)$ the *inverse Fourier transform* $\mathcal{F}^{-1} : L^2(-\pi,\pi) \mapsto l^2$ by

$$(\mathcal{F}^{-1}u)_h = ((\mathcal{F}^{-1}u)_j)_{j\in\mathbb{Z}}, \quad (\mathcal{F}^{-1}u)_j = \frac{1}{h}\int_{-\pi}^{\pi} u(x)\overline{v_j(x)}\,dx.$$

Lemma 8.2 (Parseval's identity). *The norms are equal:*

$$\|\mathcal{F}u_h\|^2_{L^2(-\pi,\pi)} = \|u_h\|^2_{0,h}.$$

Lemma 8.3 (Shift property).

$$\mathcal{F}[(u_{j+1})_{j\in\mathbb{Z}}](x) = e^{-\imath x}(\mathcal{F}u_h)(x), \quad \mathcal{F}[(u_{j-1})_{j\in\mathbb{Z}}](x) = e^{\imath x}(\mathcal{F}u_h)(x).$$

Theorem 8.2 (L^2-error). *(Grossmann et al. 2007, Th.2.10, p.43) For the initial data $u^0 \in C^{1,1}(\mathbb{R})$, let its derivative $(u^0)'$ have a compact support in interval $(-P,P)$ with a finite number $P > 0$.*

(i) *With respect to the norm $\max\limits_{k\in\{0,\dots,M\}} \|\cdot\|_{0,h}$, the **explicit upwind scheme** is consistent of order one.*

(ii) *Under the CFL-condition $\gamma \leqslant 1$, it is stable and convergent of order one.*

Remark 8.1. Consider a difference equation satisfying the generic estimate:

$$|u_i^{k+1}| \leqslant a|u_i^k| + \tau|f|, \quad k = 1,\dots,M-1, \tau = T/M,$$

where a is called an *amplification factor*. A sufficient condition for stability is

$$a \leqslant 1 + 1/k.$$

Indeed, we estimate iteratively

$$|u_i^{k+1}| \leqslant a(a|u_i^{k-1}| + \tau|f|) + \tau|f| \leqslant \ldots \leqslant a^k|u_i^1| + \tau \sum_{j=0}^{k-1} a^j |f|.$$

If $a \leqslant 1 + 1/k$, then $a^k \leqslant (1+1/k)^k < e$, and the geometric series yields:

$$\sum_{j=0}^{k-1} a^j = \begin{cases} \frac{1-a^{k-1}}{1-a} \leqslant 1, & \text{if } a < 1, \quad k \geq 2 \\ k, & \text{if } a = 1, \\ \frac{1-a^{k-1}}{1-a} < \frac{e-1}{a-1}, & \text{if } 1 < a < 1 + 1/k. \end{cases}$$

8.4. Instability analysis.

Remark 8.2. For $\gamma > 1$, the upwind scheme (8.4) is *unstable.*

If $\gamma > 1$, then for $\varepsilon \in (0, 2(\gamma-1))$ there exists $\delta \in (0, \pi/2)$ such that $a := |(1-\gamma) + \gamma e^{\imath x}| = 2\gamma - 1 > 1 + \varepsilon$. Therefore, the amplification factor $a > 1 + \varepsilon$ and using $e_h^0 = 0$, from (**??**) we estimate the error asymptotically for $\tau = M/k$:

$$\|e_h^{k+1}\|_{0,h} \geqslant -a\|e_h^k\|_{0,h} + \tau\|I_h^k\|_{0,h} \geqslant \ldots \geqslant (-a)^{k+1}\|e_h^0\|_{0,h} + \tau \sum_{j=0}^{k} a^j$$

$$= \tau \frac{a^{k+1}-1}{a-1} > \tau \frac{(1+\varepsilon)^{k+1}-1}{a-1} = \mathrm{O}\Big(\frac{(1+\varepsilon)^{k+1}}{k}\Big) \underset{k\to\infty}{\to} \infty.$$

Remark 8.3. The *upwind scheme* ($\sigma = 1$): $u_j^{k+1} = (1+\gamma)u_j^k - \gamma u_{j+1}^k$ is *unstable.*

Applying the Fourier transform we have $\mathcal{F}e_j^{k+1} = (1+\gamma)\mathcal{F}e_j^k - \gamma\mathcal{F}e_{j+1}^k + \tau\mathcal{F}I_j^k$, where $\mathcal{F}e_{j+1}^k = e^{-\imath x}\mathcal{F}e_j^k$ according to the shift property. Henceforth,

$$\|e_h^{k+1}\|_{0,h} \geqslant -|(1+\gamma) + \gamma e^{-\imath x}|\, \|e_h^k\|_{0,h} + \tau\|I_h^k\|_{0,h},$$

and the amplification factor $a = 1 + 2\gamma > 1$.

Exercise 8.2. Prove that the upwind scheme ($\sigma = 1/2$) is unstable.

9. FDM for nonlinear conservation law and systems

For $f \in C^2(\mathbb{R})$ with $f'' > 0$ and $u^0 \in W^{1,\infty}(\mathbb{R})$, consider the *1d nonlinear conservation problem*: Find $u(t,x) \in W^{1,\infty}(\mathbb{R}_+ \times \mathbb{R})$ such that

$$\text{(9.1)} \qquad \begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x} f(u) = 0 & \text{a.e. } t > 0, x \in \mathbb{R}, \\ u(0,x) = u^0(x) & \text{a.e. } x \in \mathbb{R}. \end{cases}$$

Lemma 9.1 (Solution formula). *Let the derivative* $(u^0)' \geqslant 0$*, then the unique solution to* (9.1) *is given by the implicit formula:*

$$u(t,x) = u^0\big(x - tf'(u(t,x))\big).$$

The explicit method suggests the discretization:

$$D_\tau^+ u + [\sigma D_h^+ + (1-\sigma)D_h^-]f(u) = 0, \quad \sigma \in [0,1].$$

For example, for $\sigma = 1/2$ we get:

$$u_i^{k+1} = u_i^k - \tfrac{\tau}{2h}[f(u_{i+1}^k) - f(u_{i-1}^k)].$$

It can be generalized to a ***conservation scheme***:

$$u_i^{k+1} = u_i^k - \frac{\tau}{h}[F(u_{i+1}^k, u_i^k) - F(u_i^k, u_{i-1}^k)] =: H(u_{i-1}^k, u_i^k, u_{i+1}^k) \tag{9.2}$$

with a *numerical flux* $F(v,w) : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $F(v,v) = f(v)$. Indeed, from

$$\begin{aligned}\tfrac{1}{h}[F(u(x+h),u(x)) - F(u(x),u(x-h))] &= \tfrac{\partial}{\partial u}F(u(x),u(x))\tfrac{\partial u}{\partial x} + \mathrm{O}(h)\\ &= f'(u(x))\frac{\partial u}{\partial x} + \mathrm{O}(h) = \frac{\partial}{\partial x}f(u(x)) + \mathrm{O}(h),\end{aligned}$$

it follows the consistency when $F(v,v) = f(v)$.

9.1. Lax–Friedrichs flux. The well-known *Lax–Friedrichs flux* has the form:

$$F(v,w) = \frac{1}{2}[f(v) + f(w) + \frac{h}{\tau}(w - v)].$$

With this flux the conservation scheme reads:

$$\begin{aligned}u_i^{k+1} &= u_i^k - \tfrac{\tau}{2h}[f(u_{i+1}^k) + f(u_i^k) + \tfrac{h}{\tau}(u_i^k - u_{i+1}^k)\\ &- f(u_i^k) - f(u_{i-1}^k) - \frac{h}{\tau}(u_{i-1}^k - u_i^k)] = \tfrac{u_{i+1}^k + u_{i-1}^k}{2} - \tfrac{\tau}{2h}[f(u_{i+1}^k) - f(u_{i-1}^k)].\end{aligned} \tag{9.3}$$

Definition 9.1. *A conservation scheme* (9.2) *is called monotone, if the function* $(u,v,w) \mapsto H(u,v,w)$ *is non-decreasing.*

Lemma 9.2. *The Lax–Friedrichs scheme* (9.3) *is monotone, if the CFL-condition* $|f'|\dfrac{\tau}{h} \leqslant 1$ *holds.*

Note that in the linear case of $f(u) = vu$, the CFL-condition implies $f' = v$.

Lemma 9.3. *(Grossmann et al. 2007, Lemma 2.19, p.55) From the monotone property, it follows the discrete maximum principle:*

$$m_i^k := \min(u_{i-1}^k, u_i^k, u_{i+1}^k) \leqslant u_i^{k+1} \leqslant \max(u_{i-1}^k, u_i^k, u_{i+1}^k) =: M_i^k.$$

Remark 9.1. (Grossmann et al. 2007, Th.2.21, p.56) A monotone conservation scheme is convergent of order at most 1.

9.2. Lax–Wendroff method. Begin with the *linear transport problem* (8.2):

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0 \text{ as } t > 0, \quad u = u^0 \text{ at } t = 0, \quad \text{for } x \in \mathbb{R}.$$

Look the explicit scheme in the form (Larsson and Thomee 2008, Sec.12.1):

$$u_i^{k+1} = au_{i-1}^k + bu_i^k + cu_{i+1}^k$$

with unknown coefficients $a, b, c \in \mathbb{R}$. To get the second-order consistency, we need $b = 1 - a - c$ such that solution of (8.2) after Taylor's expansion in (t^k, x_i) yields

$$\begin{aligned}&u(t^{k+1},x_i) - au(t^k,x_{i-1}) - bu(t^k,x_i) - cu(t^k,x_{i+1})\\ &= u(t^{k+1},x_i) - u(t^k,x_i) + a\big(u(t^k,x_i) - u(t^k,x_{i-1})\big) - c\big(u(t^k,x_{i+1}) - u(t^k,x_i)\big)\\ &\qquad = \Big[\tau\frac{\partial u}{\partial t} + (a-c)h\frac{\partial u}{\partial x} + \tau^2\frac{\partial^2 u}{\partial t^2} - (a+c)\frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}\Big](t^k,x_i) + \mathrm{O}(\tau^3 + h^3).\end{aligned}$$

Differentiating (8.2) we substitute here $\frac{\partial^2 u}{\partial t^2} = -v\frac{\partial^2 u}{\partial t \partial x}$ and $\frac{\partial^2 u}{\partial x^2} = -\frac{1}{v}\frac{\partial^2 u}{\partial t \partial x}$ and use the Courant number $\gamma = v\frac{\tau}{h}$, then the expression in square brackets is zero when

$a - c = \gamma$ and $a + c = \gamma^2$. Solving three relations we determine uniquely the coefficients

$$a = \frac{1}{2}\gamma(\gamma+1), \quad c = \frac{1}{2}\gamma(\gamma-1), \quad b = 1-\gamma^2,$$

which constitute the *Lax–Wendroff scheme*:

$$u_i^{k+1} = u_i^k - \tfrac{\gamma}{2}(u_{i+1}^k - u_{i-1}^k) + \tfrac{\gamma^2}{2}(u_{i+1}^k - 2u_i^k + u_{i-1}^k). \tag{9.4}$$

For stability, we apply the Fourier transform $\mathcal{F}$ to (9.4) and use (8.2) to obtain the following equation for the error $e_{(\tau,h)} := r_{(\tau,h)}u - u_{(\tau,h)}$:

$$\begin{aligned}\mathcal{F}e_h^{k+1}(x) &= \tfrac{\gamma(\gamma+1)}{2}\mathcal{F}[(e_{j-1}^k)_{j\in\mathcal{Z}}](x) + (1-\gamma^2)\mathcal{F}e_h^k(x) + \tfrac{\gamma(\gamma-1)}{2}\mathcal{F}[(e_{j+1}^k)_{j\in\mathcal{Z}}](x)\\ &= \big(\tfrac{\gamma(\gamma+1)}{2}e^{\imath x} + 1 - \gamma^2 + \tfrac{\gamma(\gamma-1)}{2}e^{-\imath x}\big)\mathcal{F}e_h^k(x)\end{aligned}$$

in virtue of the shift property. The amplification factor here

$$\begin{aligned}|1-\gamma^2+\gamma^2\cos x + \imath\gamma\sin x|^2 &= \big(1-\gamma^2(1-\cos x)\big)^2 + \gamma^2(1-\cos^2 x)\\ &= 1-\gamma^2(1-\gamma^2)(1-\cos x)^2 \le 1\end{aligned}$$

implies the stability, if the CFL condition $\gamma \le 1$ holds.

For the *nonlinear scalar conservation law* in (9.1):

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}f(u) = 0,$$

the Lax–Wendroff scheme (9.4) can be generalized as follows

$$\begin{aligned}u_i^{k+1} &= u_i^k - \tfrac{\tau}{2h}\big(f(u_{i+1}^k) - f(u_{i-1}^k)\big)\\ &\quad + \frac{\tau^2}{2h^2}\Big[f'\big(\tfrac{u_{i+1}^k+u_i^k}{2}\big)\big(f(u_{i+1}^k)-f(u_i^k)\big) - f'\big(\tfrac{u_i^k+u_{i-1}^k}{2}\big)\big(f(u_i^k)-f(u_{i-1}^k)\big)\Big].\end{aligned}$$

9.3. Symmetric hyperbolic system. We look for an unknown vector $U = (u_1,\dots,u_N)^\top(t,x)$ satisfying the *hyperbolic system of linear equations in 1d*:

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} = 0, \quad t>0,\ x\in\mathbb{R},$$

with given symmetric matrix $A \in \mathbb{R}^{N\times N}$. For the symmetric system, the Lax–Friedrichs scheme (9.3) reads (Larsson and Thomee 2008, Sec.12.1):

$$\begin{aligned}U_i^{k+1} &= \tfrac{1}{2}\big(I - \tfrac{\tau}{h}A\big)U_{i+1}^k - \tfrac{1}{2}\big(I + \tfrac{\tau}{h}A\big)U_{i-1}^k\\ &= \tfrac{1}{2}\big(U_{i+1}^k + U_{i-1}^k\big) - \tfrac{\tau}{2h}A\big(U_{i+1}^k - U_{i-1}^k\big),\end{aligned} \tag{9.5}$$

whereas the Lax–Wendroff scheme (9.4) turns into

$$\begin{aligned}U_i^{k+1} &= \tfrac{\tau}{h}A\big(\tfrac{\tau}{h}A + I\big)U_{i-1}^k + \big(I - \tfrac{\tau^2}{h^2}A^2\big)U_i^k + \tfrac{\tau}{h}A\big(\tfrac{\tau}{h}A - I\big)U_{i+1}^k\\ &= U_i^k - \tfrac{\tau}{2h}A\big(U_{i+1}^k - U_{i-1}^k\big) + \tfrac{\tau^2}{h^2}A^2\big(U_{i+1}^k - 2U_i^k + U_{i-1}^k\big).\end{aligned} \tag{9.6}$$

CHAPTER 2

Numerics of Variational Problems (VP)

10. Variational theory

Let V be a *Hilbert space* (a complete vector space, that is, every Cauchy sequence converges) with the *norm* $\| \cdot \|$ induced by a *scalar product* $(\cdot , \cdot)$. Its *dual space* V^* consists of linear continuous functionals $f : V \mapsto \mathbb{R}$ equipped with the *dual norm*

$$\|f\|_* := \sup_{0 \neq v \in V} \frac{|f(v)|}{\|v\|},$$

where $f(v) = \langle f, v \rangle$ implies the duality pairing $\langle \cdot , \cdot \rangle$ between V^* and V. For a given $f \in V^*$ and a *bilinear form* $a : V \times V \mapsto \mathbb{R}$, consider an abstract *variational problem* (VP): Find $u \in V$ such that

$$a(u, v) = f(v) \quad \text{for all } v \in V. \tag{10.1}$$

10.1. Lax–Milgram theorem.

Theorem 10.1 (Lax–Milgram). *(Grossmann et al. 2007, Lemma 3.25, p.145) If a is continuous:*

$$\text{there exists } \overline{a} > 0 : \quad |a(u, v)| \leqslant \overline{a}\|u\|\|v\| \tag{10.2}$$

for all $u, v \in V$, and coercive:

$$\text{there exists } 0 < \underline{a} \leqslant \overline{a} : \quad a(u, u) \geqslant \underline{a}\|u\|^2, \tag{10.3}$$

then VP (10.1) *has the unique solution satisfying the a-priori estimate:*

$$\|u\| \leqslant \frac{1}{\underline{a}}\|f\|_*.$$

10.2. Minimization problem.

Corollary 10.1. *(Grossmann et al. 2007, Lemma 3.26, p.146) Within the Lax–Milgram theorem, if a is symmetric such that $a(u, v) = a(v, u)$ for all $u, v \in V$, then* (10.1) *is equivalent to the minimization problem:*

$$J(u) = \min_{v \in V} J(v), \quad J : V \mapsto \mathbb{R}, \quad J(v) := \tfrac{1}{2}a(v, v) - f(v). \tag{10.4}$$

10.3. Second-order elliptic BVP. For example, consider the *elliptic BVP* in a domain $\Omega \subset \mathbb{R}^d$ with the Lipschitz boundary $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ such that $\Gamma_D \cap \Gamma_N = \emptyset$, under the *mixed Dirichlet–Neumann BC*:

$$\begin{cases} Lu := -\displaystyle\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\Big(\sum_{j=1}^{d} a_{ij}(x)\frac{\partial u}{\partial x_j}\Big) + \sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f \quad \text{in } \Omega, \\ u = 0 \quad \text{at } \Gamma_D, \quad \displaystyle\sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial u}{\partial x_j} n_i = 0 \quad \text{at } \Gamma_N, \end{cases} \tag{10.5}$$

which is *uniformly elliptic* with respect to $x \in \Omega$, when

$$\underline{a} \sum_{i=1}^{d} \xi_i^2 \leqslant \sum_{i,j=1}^{d} a_{ij}(x) \xi_i \xi_j \leqslant \overline{a} \sum_{i=1}^{d} \xi_i^2 \quad \text{for } \xi = (\xi_1, \dots, \xi_d)^\top \in \mathbb{R}^d.$$

Multiply (10.5) with a smooth test function v and integrate by parts in Ω such that

$$\int_\Omega (Lu) v \, dx = \int_\Omega \Big\{ \sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + \sum_{i=1}^{d} b_i(x) \frac{\partial u}{\partial x_i} v + c(x) u v \Big\} \, dx$$

$$- \int_{\Gamma_N} \Big(\sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_j} n_i \Big) v \, dS_x - \int_{\Gamma_D} \Big(\sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_j} n_i \Big) v \, dS_x = \int_\Omega f v \, dx =: f(v).$$

The boundary terms here are equal to zero due to the Neumann BC in (10.5) and for $v = 0$ on Γ_D, thus we arrive at the *bilinear form*

$$a(u, v) := \int_\Omega \big\{ \sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + \sum_{i=1}^{d} b_i(x) \frac{\partial u}{\partial x_i} v + c(x) u v \big\} \, dx.$$

Introduce the *Sobolev function space*:

$$V = \{ v \in H^1(\Omega) : \quad v = 0 \quad \text{a.e. on } \Gamma_D \},$$

which is equipped with the norm induced by the inner product:

$$\|v\|^2_{H^1(\Omega)} = \int_\Omega (v^2 + |\nabla v|^2) \, dx, \quad (u, v)_{H^1(\Omega)} = \int_\Omega (uv + \nabla u \cdot \nabla v) \, dx,$$

where $|\nabla v|^2 = \nabla v \cdot \nabla v$ and $\nabla u \cdot \nabla v = \sum_{i=1}^{d} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i}$. The norm is equivalent to the *semi-norm*:

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

due to the ***Poincare inequality*** (Clason 2013, Th.2.5, p.17):

$$\int_\Omega v^2 \, dx \leqslant K_{\mathrm{P}}(\Omega) \int_\Omega |\nabla v|^2 \, dx \tag{10.6}$$

which holds for all $v \in H^1(\Omega)$ such that $v = 0$ on Γ_D.

The right-hand side $f \in L^2(\Omega)$ is endowed with the respective norm

$$\|f\|^2_{L^2(\Omega)} = \int_\Omega f^2 \, dx.$$

In order to apply the Max–Milgram Theorem 10.1, we check properties of the bilinear form $a : V \times V \mapsto \mathbb{R}$.

• The *continuity* of a needs the coefficients to be uniformly bounded, i.e. $a_{ij}(x), b_i(x), c(x) \in L^\infty(\Omega)$.

• Consider

$$a(u, u) = \int_\Omega \sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_i} \, dx + \int_\Omega \sum_{i=1}^{d} b_i(x) \frac{\partial u}{\partial x_i} u \, dx + \int_\Omega c(x) u^2 \, dx. \tag{10.7}$$

We show *coercivity* of a. Due to the ellipticity of $(a_{ij})_{i,j=1}^d$ and applying the Poincare inequality, the first integral in (10.7) can be estimated as

$$\int_\Omega \sum_{i,j=1}^d a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial u}{\partial x_i}\,dx \geqslant \underline{a}\int_\Omega \sum_{i=1}^d \big(\frac{\partial u}{\partial x_i}\big)^2\,dx = \underline{a}\|\nabla u\|^2_{L^2(\Omega)}$$
$$= \underline{a}\big(\tfrac{1}{1+K_{\rm P}(\Omega)} + \tfrac{K_{\rm P}(\Omega)}{1+K_{\rm P}(\Omega)}\big)\|\nabla u\|^2_{L^2(\Omega)} \geqslant \tfrac{\underline{a}}{1+K_{\rm P}(\Omega)}\|u\|^2_{H^1(\Omega)}.$$

Integrating by parts and using the identity $\frac{\partial u}{\partial x_i}u = \frac{1}{2}\frac{\partial}{\partial x_i}(u^2)$, the second integral in the right-hand side of (10.7) is estimated as follows:

$$\int_\Omega \sum_{i=1}^d b_i(x)\frac{\partial u}{\partial x_i}u\,dx = -\frac{1}{2}\int_\Omega \sum_{i=1}^d \frac{\partial b_i}{\partial x_i}u^2\,dx + \frac{1}{2}\int_{\Gamma_N}\sum_{i=1}^d (b_i n_i)u^2\,dS_x$$
$$+\frac{1}{2}\int_{\Gamma_D}\sum_{i=1}^d (b_i n_i)u^2\,dS_x = -\frac{1}{2}\int_\Omega ({\rm div}\, b)u^2\,dx + \frac{1}{2}\int_{\Gamma_N}(b\cdot n)u^2\,dS_x,$$

where the integral over Γ_D is equal to zero due to the homogeneous Dirichlet BC. As the result, we proceed with the estimate:

$$a(u,u) \geqslant \frac{\underline{a}}{1+K_{\rm P}(\Omega)}\|u\|^2_{H^1(\Omega)} + \int_\Omega (c - \frac{1}{2}{\rm div}\, b)u^2\,dx + \frac{1}{2}\int_{\Gamma_N}(b\cdot n)u^2\,dS_x,$$

which follows the *sufficient conditions* for the coercivity:

$$b_i(x)\in W^{1,\infty}(\Omega): \quad c - \frac{1}{2}{\rm div}\, b \geqslant 0 \quad \text{in } \Omega, \quad b\cdot n \geqslant 0 \quad \text{on } \Gamma_N.$$

• (*a-priori estimate*) Applying the Cauchy–Schwarz inequality and the continuous embedding $H^1(\Omega)\hookrightarrow L^2(\Omega)$ provides the chain of inequalities:

$$\frac{\underline{a}}{1+K_{\rm P}(\Omega)}\|u\|^2_{H^1(\Omega)} \leqslant \int_\Omega fu\,dx \leqslant \|f\|_{L^2(\Omega)}\|u\|_{L^2(\Omega)} \leqslant \|f\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)},$$

which implies the a-priori estimate in the form: $\|u\|_{H^1(\Omega)} \leqslant \frac{1+K_{\rm P}(\Omega)}{\underline{a}}\|f\|_{L^2(\Omega)}$.

• For symmetric $a_{ij} = a_{ji}$ and $b_i = 0$, we get the equivalent minimization problem (10.4) with the *objective function* $J : H^1(\Omega)\mapsto \mathbb{R}$,

$$J(v) = \frac{1}{2}\int_\Omega \Big\{\sum_{i,j=1}^d a_{ij}(x)\frac{\partial v}{\partial x_j}\frac{\partial v}{\partial x_i} + c(x)v^2\Big\}\,dx - \int_\Omega fv\,dx,$$

which coercivity is provided by the lower bound:

$$J(v) \geqslant \begin{cases} \dfrac{1}{2}\dfrac{\underline{a}}{1+K_{\rm P}(\Omega)}\|v\|^2_{H^1(\Omega)} - \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}, & \text{if } c(x)\geqslant 0,\\ \dfrac{1}{2}\Big[\dfrac{\underline{a}}{1+K_{\rm P}(\Omega)} - K_{\rm P}(\Omega)\|c\|_{L^\infty(\Omega)}\Big]\|v\|^2_{H^1(\Omega)} - \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}, & \text{otherwise.}\end{cases}$$

11. Galerkin method

Let $V_h \subset V$ (conforming approximation) be a *finite dimensional subspace* of V, then V_h is also a Hilbert space endowed with the topology of V. The bilinear form $a : V_h \times V_h \mapsto \mathbb{R}$ has the same properties. From the Lax–Milgram Theorem 10.1, there exists a unique solution $u_h \in V_h$ of the discrete variational problem:

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in V_h. \tag{11.1}$$

A discrete approximation is called *consistent* if the exact solution $u \in V$ satisfies the discrete variational formulation. Since $V_h \subset V$, we can test (10.1) with the test function $v = w_h \in V_h$:

$$a(u, w_h) = f(w_h) \quad \text{for all } w_h \in V_h, \tag{11.2}$$

thus justifying that (11.1) is consistent.

11.1. Cea's lemma on discretization error.

Lemma 11.1 (Cea). *(Grossmann et al. 2007, Th.3.31, p.152) The discretization error for solutions of the problems* (10.1) *and* (11.1) *is estimated by the best approximation error:*

$$\|u - u_h\| \leqslant \frac{\overline{a}}{\underline{a}} \inf_{v_h \in V_h} \|u - v_h\|.$$

Corollary 11.1. *If a is symmetric, then the improved estimate holds*

$$\|u - u_h\| \leqslant \sqrt{\frac{\overline{a}}{\underline{a}}} \inf_{v_h \in V_h} \|u - v_h\|.$$

Exercise 11.1. Prove the corollary by checking the estimate $a(u-u_h, u-u_h) \leqslant a(u - v_h, u - v_h)$ and using $J(u_h) \leqslant J(v_h)$.

Let the dimention of V_h be finite, say $N \in \mathbb{N}$, then there exists a basis $(\varphi_i)_{i=1}^N$ such that $V_h = \operatorname{span}\{\varphi_1, \ldots, \varphi_N \in V\}$. Taking the ansatz for a *trial function*

$$u_h = \sum_{j=1}^N u_j \varphi_j(x), \quad u_j \in \mathbb{R},$$

and *test functions* $v_h = \varphi_i$ in (11.1), we get the linear system:

$$\sum_{j=1}^N a(\varphi_j, \varphi_i) u_j = f(\varphi_i), \quad i = 1, \ldots, N, \tag{11.3}$$

with the *stiffness matrix* $L = (L_{ij})_{i,j=1}^N \in \mathbb{R}^{N \times N}$, where $L_{ij} := a(\varphi_j, \varphi_i)$.

Lemma 11.2. *L is positive definite: $\xi^\top L \xi > 0$ for $0 \neq \xi \in \mathbb{R}^N$.*

Corollary 11.2. *If a is symmetric, then L is a* Spd*-matrix.*

There are two general approaches to the basis choice:

- by *spectral elements*, every basis function $\varphi_i(x)$ has a global support;
- by *finite elements*, every basis function $\varphi_i(x)$ has a local support.

11.2. Discretizations of Poisson equation in 1d. For example, consider the *mixed BVP for the Poisson equation in 1d*:

$$\begin{cases} -u'' = f \quad \text{in } \Omega = (0,1), \\ u(0) = 0 \text{ (Dirichlet BC)}, \quad u'(1) = 0 \text{ (Neumann BC)}. \end{cases}$$

In the function space $V = \{v \in H^1(0,1) : \ v(0) = 0\}$ the corresponding *variational problem* reads: Find $u \in V$ such that

$$\int_0^1 u'v' \, dx = \int_0^1 fv \, dx \quad \text{for all } v \in V.$$

1. Discretization by monomial basis. Let $\varphi_i = \frac{1}{i}x^i$, $\varphi_i' = x^{i-1}$, then the stiffness matrix with entries

$$L_{ij} = \int_0^1 x^{i-1}x^{j-1}\,dx = \int_0^1 x^{i+j-2}\,dx = \frac{1}{i+j-1}x^{i+j-1}\Big|_0^1 = \frac{1}{i+j-1}$$

is called the *Hilbert matrix*: $L = \begin{pmatrix} 1 & 1/2 & 1/3 & \iddots \\ 1/2 & 1/3 & \iddots & \\ 1/3 & \iddots & & \\ \iddots & & & \end{pmatrix}$. The condition number $\varkappa(L) = \frac{\lambda_{\max}}{\lambda_{\min}} \gg 1$, hence the method is unstable.

2. Discretization by eigenbasis. Solving the eigenvalue problem:

$$\begin{cases} -\varphi'' = \lambda\varphi \quad \text{in } (0,1), \\ \varphi(0) = 0, \quad \varphi'(1) = 0, \end{cases}$$

we get the eigenvectors $\varphi_i(x) = \sin[\pi(\frac{1}{2}+i)x]$ for $i \in \mathbb{N}$, such that

$$\varphi_i'(x) = \pi(\frac{1}{2}+i)\cos[\pi(\frac{1}{2}+i)x], \quad \varphi_i''(x) = -(\pi(\frac{1}{2}+i))^2\sin[\pi(\frac{1}{2}+i)x],$$

and the eigenvalues $\lambda_i = -(\pi(\frac{1}{2}+i))^2$.

The corresponding *stiffness matrix* can be calculated as follows, for $i \neq j$:

$$\begin{aligned} L_{ij} &= \pi^2(\tfrac{1}{2}+i)(\tfrac{1}{2}+j)\int_0^1 \cos[\pi(\tfrac{1}{2}+i)x]\cos[\pi(\tfrac{1}{2}+j)x]\,dx \\ &= \pi^2(\tfrac{1}{2}+i)(\tfrac{1}{2}+j)\int_0^1 \big(\cos[\pi(i-j)x] + \cos[\pi(1+i+j)x]\big)\,dx \\ &= \tfrac{1}{2}\pi^2(\tfrac{1}{2}+i)(\tfrac{1}{2}+j)\Big\{\frac{1}{\pi(i-j)}\sin[\pi(i-j)x]\Big|_0^1 + \frac{1}{\pi(1+i+j)}\sin[\pi(1+i+j)x]\Big|_0^1\Big\} = 0, \end{aligned}$$

and for $i = j$:

$$\begin{aligned} L_{ii} &= \pi^2(\tfrac{1}{2}+i)^2\int_0^1 \cos^2[\pi(\tfrac{1}{2}+i)x]\,dx \\ &= \tfrac{1}{2}\pi^2(\tfrac{1}{2}+i)^2\int_0^1 (1+\cos[\pi(1+2i)x])\,dx = \tfrac{1}{2}\pi^2(\tfrac{1}{2}+i)^2. \end{aligned}$$

The equation (11.3) implies $\frac{1}{2}\pi^2(\frac{1}{2}+i)^2 u_i = f_i := \int_0^1 f\varphi_i\,dx$ following $u_i = \frac{2f_i}{\pi^2(\frac{1}{2}+i)^2}$ and the analytic solution given explicitly by the series:

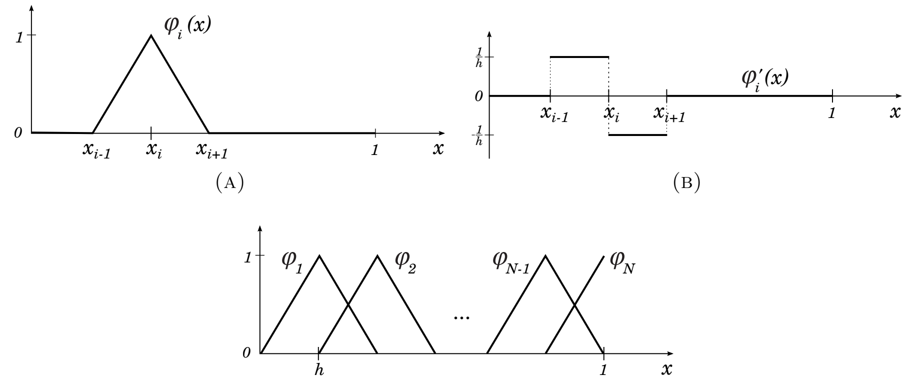
$$u_h(x) = \sum_{i=1}^N \tfrac{2}{\pi^2(\frac{1}{2}+i)^2} f_i\varphi_i(x).$$

Here $u_h \to 0$ as $i \to \infty$ provides stability. However, the disadvantage is that the method needs a-priori to know an eigenbasis.

3. Discretization by finite element method (FEM). On the uniform grid $\overline{\Omega_h} = \{x_i = ih,\ i = 0, \dots, N\}$ in $\overline{\Omega} = [0, 1]$ of the mesh size $h = 1/N$, construct continuous piecewise-linear basis functions (so-called "hat" functions):

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h}, & x \in (x_{i-1}, x_i), \\ \frac{x_{i+1} - x}{h}, & x \in (x_i, x_{i+1}), \\ 0, & \text{otherwise}, \end{cases} \quad \varphi_i'(x) = \begin{cases} \frac{1}{h}, & x \in (x_{i-1}, x_i), \\ -\frac{1}{h}, & x \in (x_i, x_{i+1}), \\ 0, & \text{otherwise}, \end{cases}$$

as illustrated in Figure 11.1.



(A)

(B)

(C)

FIGURE 11.1. (A) function φ_i, (B) derivative φ_i', (C) basis $(\varphi_i)_{i=1}^N$.

The corresponding stiffness matrix computes

$$L_{(i-1,i)} = \int_{x_{i-1}}^{x_i} \left(-\tfrac{1}{h}\right)\left(\tfrac{1}{h}\right) dx = -\tfrac{1}{h}, \quad L_{(i,i)} = \int_{x_{i-1}}^{x_i} \left(\tfrac{1}{h}\right)^2 dx + \int_{x_i}^{x_{i+1}} \left(-\tfrac{1}{h}\right)^2 dx = \tfrac{2}{h}$$

resulting in the tridiagonal system:

$$\begin{cases} -u_{i-1} + 2u_i - u_{i+1} = hf_i \quad \text{for } i = 1, \dots, N-1, \\ u_0 = 0, \quad -u_{N-1} + u_N = 0, \end{cases} \tag{11.4}$$

which coincides with the discretization by finite differences.

LEMMA 11.3. *(Clason 2013, Section 1.3, p.7–8) Let the solution of* (11.4) *be smooth $u \in H^2(0,1)$. The discretization error admits the estimate:*

$$\|u - u_h\|_{H^1(0,1)} \leqslant \tfrac{h}{\sqrt{2}} \|u\|_{H^2(0,1)}.$$

12. Finite element (FE) method

DEFINITION 12.1 (FE). *A finite element (FE) is called the triple (T, P, Ψ):*

- (i) *element $T \subset \mathbb{R}^d$ as a simply-connected domain with a piecewise-smooth boundary,*
- (ii) *space of shape functions P of dimension $k \in \mathbb{N}$,*
- (iii) *basis $\Psi = \{\Psi_1, \dots, \Psi_k\} : P \mapsto \mathbb{R}$ in the dual space P^* of linear continuous functionals.*

For example, the *elements* are: line segment, arc in 1d; triangle, quadrilateral in 2d; tetrahedron, polyhedron in 3d.

The space $P = \mathrm{span}\{\varphi_1, \dots, \varphi_k\}$ is composed by *shape functions* $(\varphi_i)_{i=1}^k$ which are, typically, polynomials.

12.1. Polynomial bases of shape functions. The standard basis is defined with the help of 1d *Lagrangian polynomials* on a grid $(\xi_i)_{i=1}^k$ in $[0,1]$:

$$S_i^k(\xi) = \prod_{j=1,\dots,k,\, j\neq i} \frac{\xi - \xi_i}{\xi_j - \xi_i}.$$

- On 2 points $\{0,1\}$, the linear interpolation $u(\xi) = u(0)S_1^2(\xi) + u(1)S_2^2(\xi)$ is provided by the shape functions:

$S_1^2(0) = 1,\ S_1^2(1) = 0$, then $S_1^2(\xi) = 1 - \xi$; $\quad S_2^2(0) = 0,\ S_2^2(1) = 1$, then $S_2^2(\xi) = \xi$.

- On 3 points $\{0, 1/2, 1\}$, the quadratic interpolation holds:

$$u(\xi) = u(0)S_1^3(\xi) + u(1/2)S_2^2(\xi) + u(1)S_3^3(\xi),$$

where $S_1^3(0) = 1$, $S_1^3(1/2) = S_1^3(1) = 0$, then $S_1^3(\xi) = (2\xi - 1)(\xi - 1)$;
$S_2^3(1/2) = 1$, $S_2^3(0) = S_2^3(1) = 0$, then $S_2^3(\xi) = 4\xi(1-\xi)$;
$S_3^3(1) = 1$, $S_3^3(0) = S_3^3(1/2) = 0$, then $S_3^3(\xi) = (2\xi - 1)\xi$.

The *hierarchical* is called a basis such that $(\varphi_i)_{i=1}^k \subset (\varphi_i)_{i=1}^{k+1}$. The example is based on the Gegenbauer polynomials in $[0,1]$ which are defined recursively as

$$(n+1)G_{n+1}^{(-1/2)}(\xi) = (2n-1)(2\xi - 1)G_n^{(-1/2)}(\xi) - (n-2)G_{n-1}^{(-1/2)}(\xi)$$

starting from $G_0^{(-1/2)}(\xi) := 1$, $G_1^{(-1/2)}(\xi) := 1 - 2\xi$, then the next are calculated as $G_2^{(-1/2)}(\xi) = 2\xi(1-\xi)$, $G_3^{(-1/2)}(\xi) = 2\xi(1-\xi)(2\xi - 1)$, etc.

In this case, the shape functions are: $S_1 = \frac{1}{2}(G_0^{(-1/2)} + G_1^{(-1/2)}) = 1 - \xi$, $S_2 = \frac{1}{2}(G_0^{(-1/2)} - G_1^{(-1/2)}) = \xi$, $S_3 = G_2^{(-1/2)}$ (the bubble mode), etc.

The *Hermit (cubic) basis* in $[0,1]$ is defined such that

ij	00	01	10	11
$h_{ij}(0)$	1	0	0	0
$h_{ij}(1)$	0	1	0	0
$h'_{ij}(0)$	0	0	1	0
$h'_{ij}(1)$	0	0	0	1

$$h_{00} = (1+2\xi)(1-\xi)^2,\ h'_{00} = -3\xi(1-\xi),$$
$$h_{01} = \xi^2(3 - 2\xi),\ h'_{01} = 3\xi(1-\xi),$$
$$h_{10} = \xi(1-\xi)^2,\ h'_{10} = (\xi - 1)(3\xi - 1),$$
$$h_{11} = \xi^2(\xi - 1),\ h'_{11} = \xi(3\xi - 2).$$

It follows the *spline interpolation*: $u(\xi) = h_{00}u(0) + h_{01}u(1) + h_{10}u'(0) + h_{11}u'(1)$.

12.2. Dual basis. For a *nodal basis* $(\varphi_i)_{i=1}^k$ of shape functions, the dual basis is characterized by the property:

$$\Psi_i(\varphi_j) = \delta_{ij}, \quad i,j = 1,\dots,k. \tag{12.1}$$

For example,

- the *Lagrange element*: $\Psi_1(\varphi) = \varphi(0)$, $\Psi_2(\varphi) = \varphi(1)$;
- the *Hermite element*: $\dots$, $\Psi_3(\varphi) = \varphi'(0)$, $\Psi_4(\varphi) = \varphi'(1)$.

Lemma 12.1 (Criterion of dual basis). *(Clason 2013, Lemma 4.3, p.30) The functions $(\Psi_i)_{i=1}^k$ form a basis in P^* if and only if the equalities $\Psi_i(u) = 0$ for all $i = 1,\dots,k$ and $u \in P$ follow that $u = 0$.*

12.3. FE interpolation.

Definition 12.2 (Triangulation). *The decomposition $(T_i)_{i\in I}$ is called an* ***admissible triangulation*** *of a domain $\Omega\subset\mathbb{R}^d$, when*

(i) *T_i are elements of the same type (e.g. triangles);*
(ii) *they are disjoint: $T_i\cap T_j=\emptyset$ for $i\neq j$, and $\bigcup_{i\in I}\overline{T}_i=\overline{\Omega}$;*
(iii) *hanging nodes are excluded, that is, for $i\neq j$, the intersection $\overline{T}_i\cap\overline{T}_j$ is either empty or a nodal point (in 1d, 2d, 3d), or a complete edge (in 2d, 3d), or a complete face (in 3d).*

Definition 12.3 (FE interpolation). *For every FE (T_i,P^i,Ψ^i) with a nodal basis $(\varphi_j)_{j=1}^k$, the operator $I:P^i\mapsto P^i$ defined as*

$$Iu:=\sum_{j=1}^k\Psi_j(u)\varphi_j \tag{12.2}$$

determines a local interpolant of u. The map $I:C^m(\overline{\Omega})\mapsto C^m(\overline{\Omega})$, $m\in\mathbb{N}_0$, such that its restriction on every element T_i coincides with (12.2), *determines a global interpolant. The space $V_h=\{Iu:\ u\in C^m(\overline{\Omega})\}$ is called C^m-finite element space.*

Lemma 12.2 (Properties of FE interpolation). *(Clason 2013, Lemma 4.6, p.35)*

(i) *The operator I is linear.*
(ii) *Interpolation property: $\Psi_j(Iu)=\Psi_j(u)$, $j=1,\dots,k$, on T_i.*
(iii) *I is a projection (that is $I^2=I$) such that $Iu=u$ for $u\in V_h$.*

The following lemma is useful for conforming FEM later on.

Lemma 12.3. *(Grossmann et al. 2007, Lemma 4.1, p.179) If $u\in C^m(\overline{\Omega})$ and $u\in C^{m+1}(T_i)$ for all $i\in I$, then $u\in H^{m+1}(\Omega)$.*

13. Simplex FE

We consider a triangulation $\overline{\Omega}=\bigcup_{i\in I}\overline{T}_i\subset\mathbb{R}^d$, which vertexes x^j compose the grid $\overline{\Omega_h}=(x^j)_{j=1}^{\bar N}$. Every element T_i given by a *convex polytope* (bounded by flat sides) with vertexes $(x^j)_{j\in J_i}$, where the indexes $J_i=\{j_1,\dots,j_{|J_i|}\}\subset\{1,\dots,\bar N\}$, can be described by a convex hull/envelope in the form

$$T_i=\mathrm{conv}\{(x^j)_{j\in J_i}\}:=\{x=\sum_{l=1}^{|J_i|}\lambda_l x^l:\quad \lambda_l\geqslant 0 \text{ for all } l,\quad \sum_{l=1}^{|J_i|}\lambda_l=1\}, \tag{13.1}$$

where $(\lambda_l)_{l=1}^{|J_i|}$ are called *barycentric coordinates*.

13.1. The d-simplex.

Definition 13.1. *Based on the formula* (13.1),

(i) *the element $T_i\subset\mathbb{R}^d$ is called d-simplex, if the cardinality $|J_i|=d+1$.*
(ii) *The unit d-simplex is the set*

$$T=\{\xi=\sum_{l=1}^{d+1}\lambda_l e^l:\quad \lambda_l\geqslant 0 \textit{ for all } l,\quad \sum_{l=1}^{d+1}\lambda_l=1\},$$

with the following $d+1$ vertexes $(e^l)_{l=1}^{d+1}$ in $\mathbb{R}^d$:

$$e^1 = (0,0,\dots,0), \quad e^2 = (1,0,\dots,0), \quad \dots, \quad e^{d+1} = (0,0,\dots,1).$$

In Figure 13.1 the unit simplexes are illustrated in local coordinates $(\xi_1,\dots,\xi_d)$: *1-simplex* which is segment; *2-simplex* which is triangle; *3-simplex* which is tetrahedron.

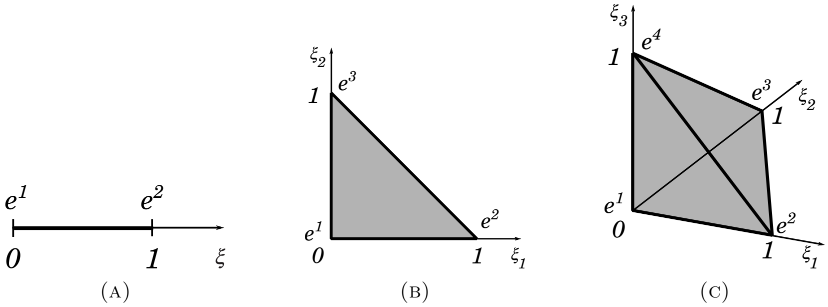


FIGURE 13.1. (A) *1-simplex*, (B) *2-simplex*, (C) *3-simplex.*

LEMMA 13.1 (Parametrization). *(Hackbusch 1992, Exercise 8.3.14, p.177) Any d-simplex T_i with vertexes $(x^{j_l})_{l=1}^{d+1}$ can be uniquely mapped to the unit simplex T by means of the affine transformation:*

$$X^i : T \mapsto T_i, \quad x = X^i(\xi) := x^{j_1} + \sum_{l=1}^{d} \xi_l (x^{j_{l+1}} - x^{j_1})$$

with the Jacobian determinant

$$|X^i| := det \begin{pmatrix} \frac{\partial X_1^i}{\partial \xi_1} & \dots & \frac{\partial X_1^i}{\partial \xi_d} \\ \dots & \ddots & \dots \\ \frac{\partial X_d^i}{\partial \xi_1} & \dots & \frac{\partial X_d^i}{\partial \xi_d} \end{pmatrix} = \begin{vmatrix} x_1^{j_2} - x_1^{j_1} & \dots & x_1^{j_{d+1}} - x_1^{j_1} \\ \dots & \ddots & \dots \\ x_d^{j_2} - x_d^{j_1} & \dots & x_d^{j_{d+1}} - x_d^{j_1} \end{vmatrix} > 0.$$

EXERCISE 13.1. Check that $X^i(e^l) = x^{j_l}$ for all $l = 1,\dots,d+1$.

13.2. Triangular FE. In 2d, consider a *triangular FE* consisting of element T_i which is the triangle with vertexes (x^1, x^2, x^3), P being the space of linear functions, $\Psi_j(\varphi) = \varphi(x^j)$, $j = 1,2,3$ implying a nodal basis. We assume the *uniform triangulation* in the unit square $\Omega = (0,1)^2$ endowed with the grid

$$\overline{\Omega_h} = \{x^j \in \overline{\Omega},\ j = 1,\dots,\bar{N} : \quad x^j = (kh, lh),\ k,l \in \mathbb{N}_0\}$$

of the mesh size $h = 1/\sqrt{\bar{N}} - 1$, as illustrated in Figure 13.2 for $\bar{N} = 16$.

For the *lower triangles* we calculate the affine transformation:

$$X^i = x^j + \xi_1(h,0) + \xi_2(0,h) = x^j + h\xi,\ |X^i| = \begin{vmatrix} h & 0 \\ 0 & h \end{vmatrix} = h^2, \quad (X^i)^{-1} = \frac{x - x^j}{h},$$

similarly, for the *upper triangles*:

$$X^i = x^j - \xi_1(h,0) - \xi_2(0,h) = x^j - h\xi,\ |X^i| = \begin{vmatrix} -h & 0 \\ 0 & -h \end{vmatrix} = h^2, \quad (X^i)^{-1} = \frac{x^j - x}{h}.$$

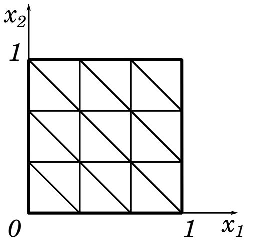


FIGURE 13.2. Example triangulation with $h = 1/3$ for $\bar{N} = 16$.

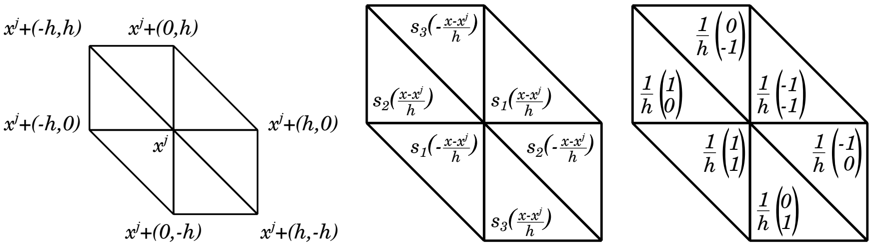
The *shape functions* in the unit 2-simplex T should satisfy $S_i(e^j) = \delta_{ij}$ for the nodal basis, thus

$$S_1(\xi) = 1 - \xi_1 - \xi_2, \quad S_2(\xi) = \xi_1, \quad S_3(\xi) = \xi_2.$$

The *transformed shape functions* in T_i are

$$\varphi_j(x) = \begin{cases} S_j(\frac{x-x^j}{h}) & \text{for lower triangles;} \\ S_j(\frac{x^j-x}{h}) & \text{for upper triangles.} \end{cases}$$

For every node $x^j \in \overline{\Omega_h}$, consider a *patch* $\overline{\Pi}_j = \bigcup_{i \in I_j} \overline{T}_i$ over the index set $I_j = \{i \in I : j \in J_i\}$ consisted of maximum 6 triangles $T_{i_1}, \ldots, T_{i_6}$ adjacent to x^j. From local hat-functions we compose a *global basis function* (which are continuous piecewise linear) $\hat{\varphi}_j(x)$ on every patch $\Pi_j \subset \overline{\Omega}$ as illustrated in Figure 13.3.


x^j+(-h,h)
x^j+(0,h)
x^j+(-h,0)
x^j+(h,0)
x^j
x^j+(0,-h)
x^j+(h,-h)
$s_3(-\frac{x-x^j}{h})$
$s_2(\frac{x-x^j}{h})$
$s_1(\frac{x-x^j}{h})$
$s_1(-\frac{x-x^j}{h})$
$s_2(-\frac{x-x^j}{h})$
$s_3(\frac{x-x^j}{h})$
$\frac{1}{h}\begin{pmatrix}0\\-1\end{pmatrix}$
$\frac{1}{h}\begin{pmatrix}1\\0\end{pmatrix}$
$\frac{1}{h}\begin{pmatrix}-1\\-1\end{pmatrix}$
$\frac{1}{h}\begin{pmatrix}1\\1\end{pmatrix}$
$\frac{1}{h}\begin{pmatrix}-1\\0\end{pmatrix}$
$\frac{1}{h}\begin{pmatrix}0\\1\end{pmatrix}$

(A) (B) (C)

FIGURE 13.3. (A) patch $\overline{\Pi}_j$, (B) function $\hat{\varphi}_j$, (C) gradient $\nabla\hat{\varphi}_j(x)$.

13.3. Stiffness and mass matrices for Poisson equation.

EXAMPLE 13.1. Consider the *Dirichlet problem for the Poisson equation*: Find $u \in H^1(\Omega)$ such that $u|_{\Gamma_D} = 0$ and

$$\int_\Omega (\nabla u \cdot \nabla v - fv)\, dx = 0$$

for all $v \in H^1(\Omega)$ such that $v|_{\Gamma_D} = 0$. The Galerkin ansatz

$$u_h = \sum_{l=1}^{\bar{N}} u_l \hat{\varphi}_l, \quad f_h = \sum_{l=1}^{\bar{N}} f_l \hat{\varphi}_l,$$

and the test function $v_h = \hat{\varphi}^j$ result in the discrete problem:

$$(13.2)\qquad \sum_{l\in I_j}\int_{\Pi_l\cap\Pi_j} (u_l \nabla\hat{\varphi}_l\cdot\nabla\hat{\varphi}_j - f_l\hat{\varphi}_l\hat{\varphi}_j\,dx) = 0 \quad \text{for } j=1,\dots,\bar{N}, \quad u_j|_{\Gamma_D}=0.$$

We compute the gradient of the global shape functions

$$\nabla_x S_1\Big(\frac{x-x^j}{h}\Big) = \begin{pmatrix} \frac{\partial\xi_1}{\partial x_1} & \frac{\partial\xi_1}{\partial x_2} \\ \frac{\partial\xi_2}{\partial x_1} & \frac{\partial\xi_2}{\partial x_2}\end{pmatrix} \nabla_\xi S_1 = \begin{pmatrix} 1/h & 0 \\ 0 & 1/h\end{pmatrix}\nabla_\xi S_1 = \frac{1}{h}\begin{pmatrix}-1\\-1\end{pmatrix},$$

$$\nabla_x S_2\Big(\frac{x-x^j}{h}\Big) = \frac{1}{h}\begin{pmatrix}1\\0\end{pmatrix}, \quad \nabla_x S_3\Big(\frac{x-x^j}{h}\Big) = \frac{1}{h}\begin{pmatrix}0\\1\end{pmatrix},$$

$$\nabla_x S_1\Big(-\frac{x-x^j}{h}\Big) = \begin{pmatrix} \frac{\partial\xi_1}{\partial x_1} & \frac{\partial\xi_1}{\partial x_2} \\ \frac{\partial\xi_2}{\partial x_1} & \frac{\partial\xi_2}{\partial x_2}\end{pmatrix} \nabla_\xi S_1 = \begin{pmatrix} -1/h & 0 \\ 0 & -1/h\end{pmatrix}\nabla_\xi S_1 = \frac{1}{h}\begin{pmatrix}1\\1\end{pmatrix},$$

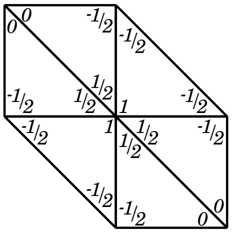
$$\nabla_x S_2\Big(-\frac{x-x^j}{h}\Big) = \frac{1}{h}\begin{pmatrix}-1\\0\end{pmatrix}, \quad \nabla_x S_3\Big(-\frac{x-x^j}{h}\Big) = \frac{1}{h}\begin{pmatrix}0\\-1\end{pmatrix}$$

entering the integral in (13.2) by the mean of $\sum_{l\in I_j}\int_{\Pi_l\cap\Pi_j}\nabla\hat{\varphi}_l\cdot\nabla\hat{\varphi}_j\,dx$ which can be directly calculated for every element T_i as follows

$$\int_{T_i}|\nabla\varphi_1|^2\,dx = \frac{h^2}{2}\Big(\frac{1}{h^2}+\frac{1}{h^2}\Big) = 1, \qquad \int_{T_i}|\nabla\varphi_l|^2\,dx = \frac{h^2}{2}\frac{1}{h^2} = \frac{1}{2} \quad \text{for } l=2,3,$$

$$\int_{T_i}\nabla\varphi_1\cdot\nabla\varphi_l\,dx = \frac{h^2}{2}\Big(-\frac{1}{h^2}\Big) = -\frac{1}{2} \quad \text{for } l=2,3, \qquad \int_{T_i}\nabla\varphi_2\cdot\nabla\varphi_3\,dx = 0.$$

Consequently, we assemble the *stiffness matrix* and get the corresponding *interior stencil* for the Laplace operator as illustrated below in Figure 13.4.



0	-1	0
-1	4	-1
0	-1	0

FIGURE 13.4. Interior stencil for the stiffness matrix for $-\Delta_h$.

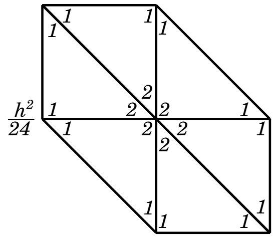
Next we calculate the *mass matrix* $\sum_{l\in I_j}\int_{\Pi_l\cap\Pi_j}\hat{\varphi}_l\hat{\varphi}_j\,dx$ entering (13.2) for every element T_i as

$$\int_{T_i}(\varphi_j)^2\,dx = |X^i|\int_T S_j^2\,d\xi = \frac{h^2}{12}, \qquad \int_{T_i}\varphi_j\varphi_l\,dx = |X^i|\int_T S_jS_l\,d\xi = \frac{h^2}{24}$$

as shown in Figure 13.5, using the *exact integration formula over the unit d-simplex* (Ciarlet and Lions 1991, (25.14), p.187):

$$(13.3)\qquad \int_T \xi_1^{n_1}\xi_2^{n_2}\dots\xi_d^{n_d}\,d\xi = \frac{n_1!n_2!\dots n_d!}{(d+n_1+n_2+\dots+n_d)!}.$$

EXERCISE 13.2. Calculate integrals in the mass matrix using formula (13.3).



$$\frac{h^2}{24}\begin{array}{|c|c|c|}\hline 1 & 1 & 0 \\ \hline 1 & 6 & 1 \\ \hline 0 & 1 & 1 \\ \hline \end{array}$$

FIGURE 13.5. Interior stencil for the mass matrix.

14. Rectangular FE in 2d

For the unit square $T = (0,1)^2$ with vertexes

$$e^1 = (0,0), \quad e^2 = (1,0), \quad e^3 = (0,1), \quad e^4 = (1,1),$$

a special parametrization such that $\lambda_l(e^j) = \delta_{lj}$, given by

$$\lambda_1 = (1-\xi_1)(1-\xi_2), \quad \lambda_2 = \xi_1(1-\xi_2), \quad \lambda_3 = (1-\xi_1)\xi_2, \quad \lambda_4 = \xi_1\xi_2,$$

follows that $(\lambda_l)_{l=1}^4$ are *barycentric coordinates* with the properties:

- $\lambda_l \geqslant 0$ for all l,
- $\sum_{l=1}^4 \lambda_l = (1 - \xi_1 - \xi_1 + \xi_1\xi_2) + (\xi_1 - \xi_1\xi_2) + (\xi_2 - \xi_1\xi_2) + \xi_1\xi_2 = 1$,
- $\sum_{l=1}^4 \lambda_l e^l = \begin{pmatrix} \xi_1(1-\xi_2) + \xi_1\xi_2 \\ (1-\xi_1)\xi_2 + \xi_1\xi_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$.

LEMMA 14.1 (Bilinear mapping). *(Brenner and Scott 2008, Ex.4.x.21, p.127) Any quadrilateral T_i with vertexes $(x^{j_l})_{l=1}^4$ can be transformed to the unit square by the bilinear coordinate transformation (isomorphismus) $X^i : T \mapsto T_i$,*

$$x = X^i(\xi) := (1-\xi_1)(1-\xi_2)x^{j_1} + \xi_1(1-\xi_2)x^{j_2} + (1-\xi_1)\xi_2 x^{j_3} + \xi_1\xi_2 x^{j_4}$$

with the Jacobian determinant $\left|\frac{\partial X^i}{\partial \xi}\right|$

$$= \det\begin{pmatrix} (1-\xi_2)(x_1^{j_2} - x_1^{j_1}) + \xi_2(x_1^{j_4} - x_1^{j_3}), & (1-\xi_1)(x_1^{j_3} - x_1^{j_1}) + \xi_1(x_1^{j_4} - x_1^{j_2}) \\ (1-\xi_2)(x_2^{j_2} - x_2^{j_1}) + \xi_2(x_2^{j_4} - x_2^{j_3}), & (1-\xi_1)(x_2^{j_3} - x_2^{j_1}) + \xi_1(x_2^{j_4} - x_2^{j_2}) \end{pmatrix}.$$

For convex T_i, the mapping is a diffeomorphism (i.e. the both X^i and $(X^i)^{-1}$ are differentiable).

EXERCISE 14.1. Give an example of a nonconvex quadrilateral T_i such that $\left|\frac{\partial X^i}{\partial \xi}(\xi)\right| = 0$ in some $\xi \in \overline{T}$.

14.1. Bilinear elements. Let $\mathbb{P}_k$ denote the set of *polynomials of degree at most* $k \in \mathbb{N}_0$:

$$\mathbb{P}_k = \{p(\xi) = \sum_{i=0}^k c_i \xi^i, \quad c_i \in \mathbb{R}, \; i = 0, \dots, k\}.$$

In any element T in $\mathbb{R}^2$ define the set of polynomials of degree k in each variable:

$$\mathbb{Q}_k(T) = \{p(\xi) = \sum_{i_1, i_2 = 0}^k c_{(i_1,i_2)} \xi_1^{i_1} \xi_2^{i_2}, \quad c_{(i_1,i_2)} \in \mathbb{R}, \; i_1, i_2 = 0, \dots, k\},$$

and the smaller set $\mathbb{P}_k(T) \subset \mathbb{Q}_k(T)$ of polynomials of the total degree k:

$$\mathbb{P}_k(T) = \{p(\xi) = \sum_{i_1+i_2=0}^{k} c_{(i_1,i_2)} \xi_1^{i_1} \xi_2^{i_2}, \quad c_{(i_1,i_2)} \in \mathbb{R},\ i_1 + i_2 = 0, \dots, k\}.$$

On the unit square T, we consider the bilinear $\mathbb{Q}_1$-basis $(S_l)_{l=1}^4$ formed by linear $\mathbb{P}_1$-polynomials:

l	1	2	3	4
S_l	$(1-\xi_1)(1-\xi_2)$	$\xi_1(1-\xi_2)$	$(1-\xi_1)\xi_2$	$\xi_1\xi_2$
∇S_l	$\begin{pmatrix} -(1-\xi_2) \\ -(1-\xi_1) \end{pmatrix}$	$\begin{pmatrix} 1-\xi_2 \\ -\xi_1 \end{pmatrix}$	$\begin{pmatrix} -\xi_2 \\ 1-\xi_1 \end{pmatrix}$	$\begin{pmatrix} \xi_2 \\ \xi_1 \end{pmatrix}$

For a *rectangle* T_i of the size $h_1 \times h_2$, the bilinear mapping $X^i : T \mapsto T_i$ from Lemma 14.1 is *affine*:

$$X^i(\xi) = x^j + (\xi_1 h_1, \xi_2 h_2), \quad \Big|\frac{\partial X^i}{\partial \xi}\Big| = \begin{vmatrix} h_1 & 0 \\ 0 & h_2 \end{vmatrix} = h_1 h_2, \quad \xi_l = \frac{x_l - x_l^j}{h_l},\ l = 1, 2.$$

Definition 14.1. *The rectangular FE consists of*

- *$h_1 \times h_2$-rectangle T_i with vertexes $(x^{j_l})_{l=1}^4$,*
- *the space $P = \mathbb{Q}_1(T_i)$ of bilinear functions,*
- *a dual basis $(\Psi^l)_{l=1}^4$ in P^* such that $\Psi_l(\varphi) = \varphi(x^{j_l})$ for $l = 1, \dots, 4$.*

On the uniform triangulation by $h_1 \times h_2$-rectangles in $\Omega = (0, a_1) \times (0, a_2)$ endowed with the grid:

$$\overline{\Omega_h} = \{x^j \in \overline{\Omega},\ j = 1, \dots, \bar{N} : \quad x^j = (kh_1, lh_2),\ k, l \in \mathbb{N}_0\},$$

for every vertex $x^j \in \overline{\Omega_h}$ consider the *patch*

$$\overline{\Pi}_j = \bigcup_{i \in I_j} \overline{T}_i \quad \text{for } I_j = \{i \in I : j \in J_i\}$$

consisting of maximum 4 rectangles adjacent to x^j. The nodal basis $(\hat{\varphi}_j)_{j=1}^{\bar{N}}$ is composed of continuous bilinear functions in $\overline{\Omega}$ and implies that $\hat{\varphi}_j(x^l) = \delta_{jl}$ for all vertexes x^l of Π_j (maximum 9), thus forming the Galerkin ansatz

$$u_h = \sum_{j=1}^{\bar{N}} u_j \hat{\varphi}_j \in V_h \subset H^1(\Omega). \tag{14.1}$$

14.2. Stiffness and mass matrices. To compute the stiffness and mass matrices for u_h given by (14.1), the *Gauss–Legendre quadrature* is useful which is exact for $\mathbb{P}_3(0,1)$:

$$\int_0^1 f(\xi)\, d\xi \approx \frac{1}{2}\Big[f\Big(\frac{1 - 1/\sqrt{3}}{2}\Big) + f\Big(\frac{1 + 1/\sqrt{3}}{2}\Big)\Big].$$

Using it, we can calculate the *stiffness matrix* for the Laplace operator (see (13.2)):

$$\sum_{l \in I_j} \int_{\Pi_l \cap \Pi_j} \nabla\hat{\varphi}_l \cdot \nabla\hat{\varphi}_j\, dx = \Big|\frac{\partial x}{\partial \xi}\Big| \int_T \Bigg[\begin{pmatrix} \frac{\partial \xi_1}{\partial x_1} & \frac{\partial \xi_1}{\partial x_2} \\ \frac{\partial \xi_2}{\partial x_1} & \frac{\partial \xi_2}{\partial x_2} \end{pmatrix} \nabla_\xi S_l\Bigg]^\top \Big(\frac{\partial \xi}{\partial x}\Big) \nabla_\xi S_j\, d\xi$$

$$= h_1 h_2 \int_T \Bigg[\begin{pmatrix} 1/h_1 & 0 \\ 0 & 1/h_2 \end{pmatrix} \nabla_\xi S_l\Bigg]^\top \begin{pmatrix} 1/h_1 & 0 \\ 0 & 1/h_2 \end{pmatrix} \nabla_\xi S_j\, d\xi$$

on the interior patch $\begin{array}{|cc|cc|}\hline A_4 & A_3 & A_3 & A_4\\ A_2 & A_1 & A_1 & A_2\\ \hline A_2 & A_1 & A_1 & A_2\\ A_4 & A_3 & A_3 & A_4\\ \hline\end{array}$ with the following coefficients due to (Thomson and Pinsky 1995):

$$A_1 = \frac{h_1h_2}{3}\Big(\frac{1}{h_1^2}+\frac{1}{h_2^2}\Big), \quad A_2 = -\frac{h_1h_2}{6}\Big(\frac{2}{h_1^2}-\frac{1}{h_2^2}\Big),$$
$$A_3 = -\frac{h_1h_2}{6}\Big(-\frac{1}{h_1^2}+\frac{2}{h_2^2}\Big), \quad A_4 = -\frac{h_1h_2}{6}\Big(\frac{1}{h_1^2}+\frac{1}{h_2^2}\Big).$$

The resulting *9-point stencil* $\begin{array}{|c|c|c|}\hline A_4 & 2A_3 & A_4\\ \hline 2A_2 & 4A_1 & 2A_2\\ \hline A_4 & 2A_3 & A_4\\ \hline\end{array}$ is equal to $\frac{1}{3}$ $\begin{array}{|c|c|c|}\hline -1 & -1 & -1\\ \hline -1 & 8 & -1\\ \hline -1 & -1 & -1\\ \hline\end{array}$ in the case of square when $h_1 = h_2$.

For example, if $j = l$, then

$$\int_{\Pi^j} |\nabla\hat{\varphi}_j| = h_1h_2\sum_{l=1}^{4}\int_T \left|\begin{pmatrix}1/h_1 & 0\\ 0 & 1/h_2\end{pmatrix}\nabla S_l\right|^2 d\xi,$$

and for $l = 1$:

$$\begin{aligned} A_1 &= h_1h_2\int_0^1\int_0^1\Big|\Big(-\frac{1-\xi_2}{h_1}, -\frac{1-\xi_1}{h_2}\Big)\Big|^2 d\xi_1\,d\xi_2\\ &= h_1h_2\int_0^1\Big\{\frac{1}{h_1^2}(1-\xi_2)^2+\int_0^1\frac{1}{h_2^2}(1-\xi_1^2)\,d\xi_1\Big\}\,d\xi_2 = h_1h_2\Big(\frac{1}{3h_1^2}+\frac{1}{3h_2^2}\Big)\end{aligned}$$

due to formula $\int_0^1 \frac{1}{h_2^2}(1-\xi_1^2)\,d\xi_1 = \frac{1}{2h_2^2}\big[\big(1-\frac{1-1/\sqrt{3}}{2}\big)^2+\big(1-\frac{1+1/\sqrt{3}}{2}\big)^2\big] = \frac{1}{3h_2^2}$.

EXERCISE 14.2. Calculate the coefficients A_2, A_3, A_4 by the Gauss–Legendre quadrature.

Similarly, the *mass matrix* according to (13.2):

$$\sum_{l\in I_j}\int_{\Pi_l\cap\Pi_j}\hat{\varphi}_l\hat{\varphi}_j\,dx = \Big|\frac{\partial x}{\partial\xi}\Big|\sum_{l=1}^{4}\int_T S_lS_j\,d\xi$$

can be computed on the interior patch as

$$h_1h_2\begin{array}{|cc|cc|}\hline 1/36 & 1/18 & 1/18 & 1/36\\ 1/18 & 1/9 & 1/9 & 1/18\\ \hline 1/18 & 1/9 & 1/9 & 1/18\\ 1/36 & 1/18 & 1/18 & 1/36\\ \hline\end{array} \quad \text{following the stencil} \quad \frac{h_1h_2}{36}\begin{array}{|c|c|c|}\hline 1 & 4 & 1\\ \hline 4 & 16 & 4\\ \hline 1 & 4 & 1\\ \hline\end{array}.$$

EXERCISE 14.3. Calculate the mass matrix by the Gauss–Legendre quadrature.

14.3. Helmholtz equation. Consider Dirichlet problem for the *Helmholtz equation*: Fixed a wave number $k\in\mathbb{R}$, find $u\in H^1(\Omega)$ such that $u = u^0$ on $\partial\Omega$ and

$$\int_\Omega(\nabla u\cdot\nabla v - k^2uv)\,dx = 0$$

for all $v\in H_0^1(\Omega)$. The discretization on the rectangular FE with the trial functions

$$u_h = \sum_{l=1}^{\bar{N}} u_l\hat{\varphi}_l, \quad u_h^0 = \sum_{l=1}^{\bar{N}} u_l^0\hat{\varphi}_l$$

and the test functions $v = \hat{\varphi}_j$ leads to the relations:

$$\sum_{l \in I_j} \int_{\Pi_l \cap \Pi_j} u_l \big[\nabla\hat{\varphi}_l \cdot \nabla\hat{\varphi}_j - k^2 \hat{\varphi}_l \hat{\varphi}_j\big]\, dx \quad \text{for } j = 1, \ldots, \bar{N}, \quad u_j = u_j^0 \quad \text{on } \partial\Omega$$

yielding the system matrix stencil:

$A_4 - \frac{k^2 h_1 h_2}{36}$	$2A_3 - \frac{k^2 h_1 h_2}{9}$	$A_4 - \frac{k^2 h_1 h_2}{36}$
$2A_2 - \frac{k^2 h_1 h_2}{9}$	$4A_1 - \frac{4k^2 h_1 h_2}{9}$	$2A_2 - \frac{k^2 h_1 h_2}{9}$
$A_4 - \frac{k^2 h_1 h_2}{36}$	$2A_3 - \frac{k^2 h_1 h_2}{9}$	$A_4 - \frac{k^2 h_1 h_2}{36}$

.

We note numerical difficulty (pollution effect, instability) for large wave number k.

15. Interpolation error estimate

We specify FE-interpolation (see Definition 12.3) for the *Sobolev spaces* $W^{k,p}$, $k, p \in \mathbb{N}$ in a domain Ω with the Lipschitz boundary $\partial\Omega$, equipped with the *norm*

$$\|u\|^p_{W^{k,p}(\Omega)} := \sum_{0 \leqslant |\alpha| \leqslant k} \|D^\alpha u\|^p_{L^p(\Omega)}, \quad \|u\|^p_{L^p(\Omega)} := \int_\Omega |u|^p \, dx,$$

where the derivative $D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ is determined by the *multi-index* $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ of the length $|\alpha| = \alpha_1 + \ldots + \alpha_d$.

Generalization of the Poincare inequality (10.6) is called ***Poincare–Friedrichs inequality*** (Brenner et al. 2008, Th.10.6.12, p.299): For $v \in W^{k,p}(\Omega)$ it holds

$$\text{(15.1)} \quad \|v\|^p_{W^{k,p}(\Omega)} \leqslant K_{\mathrm{P}}(\Omega) \Big\{ \sum_{0 \leqslant |\alpha| \leqslant k-1} \Big| \int_\Omega D^\alpha v \, dx \Big|^p + |v|^p_{W^{k,p}(\Omega)} \Big\}, \quad K_{\mathrm{P}}(\Omega) > 0$$

with the *seminorm* $|v|^p_{W^{k,p}(\Omega)} := \sum_{|\alpha| = k} \|D^\alpha u\|^p_{L^p(\Omega)}$.

15.1. Bramble–Hilbert lemma.

Lemma 15.1 (Bramble–Hilbert). *(Grossmann et al. 2007, Th.4.25, p.224) Let a functional $f : W^{k,p}(\Omega) \mapsto \mathbb{R}$ be*

(i) *bounded:* $|f(u)| \leqslant c_1 \|u\|_{W^{k,p}(\Omega)}$, $c_1 > 0$,

(ii) *sublinear:* $|f(u + v)| \leqslant c_2(|f(u)| + |f(v)|)$, $c_2 > 0$.

If $f(u) = 0$ for all polynomials $u \in \mathbb{P}_{k-1}$, then there exists $c > 0$ such that

$$\text{(15.2)} \qquad |f(u)| \leqslant c|u|_{W^{k,p}(\Omega)}.$$

15.2. Interpolation by simplex FE. Consider a triangulation $\overline{\Omega} = \bigcup_{i \in I} \overline{T}_i$ by a family of FE (T_i, P, Ψ). Let $h_i := \max_{x,y \in T_i} \|x - y\|$ be the *diameter*, and $\rho_i > 0$ be the largest *radius of a ball* inscribed in the element T_i.

Definition 15.1. (i) *The triangulation is called affine-equivalent, if there exists an affine bijection $X^i : T \mapsto T_i$ for all $i \in I$.*

(ii) *Triangulation is called quasi-uniform, if there exists $\sigma > 0$ such that*

$$\text{(15.3)} \qquad \frac{h_i}{\rho_i} \leqslant \sigma \quad \textit{for all } i \in I.$$

The mesh-size is defined as $h := \max_{i \in I} h_i$.

For a nodal basis $(\varphi_j)_{j=1}^k$ in P and a dual basis $(\Psi_j)_{j=1}^k$ in P^* we recall a *global FE-interpolation* $I : W^{k,p}(\Omega) \mapsto V_h \subset V$ such that

$$Iu = \sum_{j=1}^{k} \Psi_j(u)\varphi_j \quad \text{for every element } T_i.$$

THEOREM 15.1 (Interpolation error). *(Grossmann et al. 2007, Th.4.28, p.225) For an affine-equivalent, quasi-uniform triangulation, let the interpolation I be a projection (i.e. if $u \in V_h$, then $Iu = u$) on the space V_h of polynomials of degree at most $k-1$ defined piecewisely over the triangulation. Then the interpolation error:*

$$\sum_{i\in I} \|u - Iu\|_{W^{l,p}(T_i)} \leqslant ch^{k-l}|u|_{W^{k,p}(\Omega)} \quad \textit{for } 0 \leqslant l \leqslant k. \tag{15.4}$$

EXAMPLE 15.1. Let $u \in H^2(\Omega) := W^{2,2}(\Omega)$ (i.e. $k = p = 2$), and Iu be the linear interpolant (since $k-1 = 1$) on simplexes. In this case, the estimate (15.4) has the form as $l = 1$ and $l = 0$, respectively:

$$\|u - Iu\|_{H^1(\Omega)} \leqslant ch|u|_{H^2(\Omega)}, \quad \|u - Iu\|_{L^2(\Omega)} \leqslant ch^2|u|_{H^2(\Omega)}.$$

15.3. Interpolation by rectangular FE. We consider the triangulation $\overline{\Omega} = \bigcup_{i\in I} \overline{T}_i$ by rectangles. Each rectangle T_i allows affine transformation to the unit square T as shown in Figure 15.1.

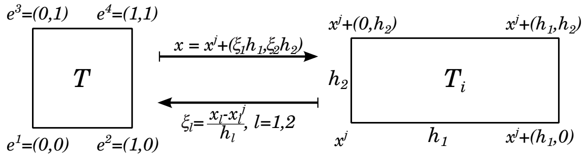


FIGURE 15.1. Rectangular element bijection.

Introduce the *FE-interpolation operator* $I : H^2(\Omega) \subset \overline{C(\Omega)} \mapsto V_h$, where V_h is the space of continuous in $\overline{\Omega}$ and piecewise polynomials $\mathbb{Q}_1(T_i) = \mathbb{P}_1(x_1^j, x_1^j + h_1) \times \mathbb{P}_1(x_2^j, x_2^j + h_2)$ on every T_i. The interpolation I is uniquely defined by the vertexes $(x^j)_{j=1}^N$, therefore,

LEMMA 15.2. *The interpolation I is a projection such that $Iu_h = u_h$ for all $u_h \in V_h$. Moreover, since $\mathbb{P}_1(T_i) \subset \mathbb{Q}_1(T_i)$, then $Iu_h = u_h$ for $u_h \in \mathbb{P}_1(T_i)$.*

EXERCISE 15.1. Take $u(\xi) = c_0 + c_1\xi_1 + c_2\xi_2$ and calculate $Iu(\xi)$.

THEOREM 15.2 (Interpolation error). *(Le Dret 2012, Th.5.32, p.135) For a quasi-uniform triangulation by rectangular FE such that*

$$\frac{\max\limits_{i\in I}(h_1,h_2)}{\min\limits_{i\in I}(h_1,h_2)} \leqslant \sigma \quad (\textit{with fixed } \sigma > 0),$$

where the mesh-size $h := \max\limits_{i\in I}(h_1, h_2)$, the interpolation error admits the estimate:

$$\sum_{i\in I} \|u - Iu\|_{L^2(T_i)} \leqslant ch^2|u|_{H^2(\Omega)}, \quad \sum_{i\in I} |u - Iu|_{H^1(T_i)} \leqslant ch|u|_{H^2(\Omega)}, \quad c > 0.$$

16. Approximation error estimate

For a suitable subspace V of $H^m(\Omega)$, $m \in \mathbb{N}$, which is a Hilbert space again, consider the *continuous and bilinear form* $a : V \times V \mapsto \mathbb{R}$ satisfying assumptions (10.2) and (10.3), *linear continuous functional* $f : V \mapsto \mathbb{R}$, and recall the *elliptic VP* (10.1): Find $u \in V$ such that

$$a(u, v) = f(v) \quad \text{for all } v \in V. \tag{16.1}$$

For a *finite dimensional subspace* $V_h \subset V$ (implying the conforming method), recall the *discrete problem* (11.1): Find $u_h \in V_h$ such that

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in V_h. \tag{16.2}$$

Theorem 16.1 (Approximation error). *Let the interpolation operator $I : V \mapsto V_h$ satisfy the assumptions of Theorem 15.1. If the solution of* (16.1) *is smooth $u \in H^{k+1}(\Omega)$ with $k \geqslant m$, then the discrete solution of* (16.2) *has the error:*

$$\|u - u_h\|_{H^m(\Omega)} \leqslant ch^{k+1-m}|u|_{H^{k+1}(\Omega)}, \quad c > 0. \tag{16.3}$$

16.1. Aubin–Nitsche lemma. For the linear continuous functional $f(v) := (f, v)_{L^2(\Omega)} = \int_\Omega fv\,dx$ recalling the notation of the scalar product in $L^2(\Omega)$, we consider the ***adjoint problem***: Find $u_f \in V_h$ such that

$$a(v, u_f) = (f, v)_{L^2(\Omega)} \quad \text{for all } v \in V. \tag{16.4}$$

Lemma 16.1 (Aubin–Nitsche). *(Ciarlet and Lions 1991, Th.19.1, p.141) For the solutions u, u_h, u_f of problems* (16.1), (16.2), (16.4), *the error estimate holds:*

$$\|u-u_h\|_{L^2(\Omega)} \leqslant \overline{a}\|u-u_h\|_{H^m(\Omega)} \sup_{f\in L^2(\Omega)} \left\{ \frac{1}{\|f\|_{L^2(\Omega)}} \inf_{v_h\in V_h} \|u_f - v_h\|_{H^m(\Omega)} \right\}. \tag{16.5}$$

16.2. Duality-based error estimates.

Corollary 16.1 (Duality-based L^2-estimate). *(Ciarlet and Lions 1991, Th.19.2, p.142) Within the Aubin–Nitsche Lemma 16.1, if the solutions $u \in H^{k+1}(\Omega)$, $k \geqslant 1$, and $u_f \in H^2(\Omega)$ with $|u_f|_{H^2(\Omega)} \leqslant c\|f\|_{L^2(\Omega)}$, then the L^2-error estimate holds:*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant ch^{k+1}|u|_{H^{k+1}(\Omega)}, \quad c > 0. \tag{16.6}$$

Lemma 16.2 (L^∞-error estimate). *(Ciarlet and Lions 1991, Th.17.2, p.135) In the statement of Corollary 16.1, let the triangulation satisfy the **inverse assumption**: there exists $\nu > 0$ such that the diameter h_i of each element T_i is uniformly bounded from below as follows:*

$$h_i \geqslant \frac{h}{\nu} \quad \textit{for all } i \in I. \tag{16.7}$$

If the solution of (16.1) *is smooth $u \in H^2(\Omega) \cap W^{1,\infty}(\Omega)$, then the L^∞-error is estimated as:*

$$\|u - u_h\|_{L^\infty(\Omega)} \leqslant c_1 h|u|_{W^{1,\infty}(\Omega)} + c_2 h^{2-d/2}|u|_{H^2(\Omega)} \quad c_1, c_2 > 0.$$

17. A-posteriori (residual-based) error estimate

17.1. Clement quasi-interpolant. For a non-continuous function $u \in V$, e.g. $u \in H^1(\Omega) \not\hookrightarrow C(\Omega)$ for $d \geqslant 2$, when a point-wise interpolation is not determined, we will define the *Clement quasi-interpolant*.

Let $(\varphi_j)_{j=1}^{\bar N}$ be a nodal basis in V_h associated to nodes $(x^j)_{j=1}^{\bar N}$. Locally, for every FE (T_i, P, Ψ) there exist *nodes* $x^j \in \bar T_i$ such that for $u_h \in V_h \cap C(\overline{\Omega})$ it holds:

$$\Psi_l(u_h) = u_h(x^j), \quad l = 1, \dots, k.$$

For Lagrange elements, nodes can be vertexes, middles of edges, or centers of the mass. For the triangulation $\overline{\Omega} = \bigcup\limits_{i\in I} \overline{T}_i$, for every such node x^j we define the *patch*

$$\overline{\Pi}^j = \bigcup_{i\in I} \{\overline{T}_i : x^j \in \overline{T}_i\}.$$

Definition 17.1 (local L^2-projection). *The projection $\pi : L^2(\Pi^j) \mapsto \mathbb{P}_k$ is defined by*

$$\begin{cases} \text{for } x^j \notin \partial\Omega : & \displaystyle\int_{\Pi^j} (u - \pi u) q \, dx = 0 \quad \text{for all } q \in \mathbb{P}_k(\Pi^j), \\ \text{for } x^j \in \partial\Omega : & \pi u = 0. \end{cases}$$

Let V_h be the space of piece-wise polynomials of order at most k. For FE $(T_i, \mathbb{P}_k(T_i), \Psi)$ with the nodal basis $(\varphi_l)_{l=1}^k$ and the dual basis $(\Psi_l)_{l=1}^k$, the local *Clement quasi-interpolant* is defined by formula

$$Iu = \sum_{l=1}^{k} \Psi_l(\pi u) \varphi_l.$$

Example 17.1. In the case of piecewise constant u ($k = 0$), we have constant q and $\int_{\Pi^j} \pi u \, dx = \int_{\Pi^j} u \, dx$, then the Clement quasi-interpolant implies the average:

$$\pi u = \frac{1}{|\Pi^j|} \int_{\Pi^j} u \, dx.$$

17.2. A-posteriori error estimate for Poisson equation. For *adaptive meshing*, it needs to evaluate a-posteriori error depending on u_h.

Consider the *Poisson equation* in the form (10.1): For $f \in L^2(\Omega)$, find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega (\nabla u \cdot \nabla v - f v) \, dx = 0 \quad \text{for all } v \in H_0^1(\Omega),$$

and its discrete counterpart (11.1): Find a piecewise linear function $v_h \in V_h \subset H_0^1(\Omega)$ such that

$$\int_\Omega (\nabla u_h \cdot \nabla v_h - f v_h) \, dx = 0 \quad \text{for all } v_h \in V_h.$$

Theorem 17.1 (A-posteriori error estimate). *(Clason 2013, Section 6.2) Let I be (quasi-)interpolant. If $\Delta u_h \in L^2(T_i)$ and $\frac{\partial u_h}{\partial \nu} \in L^2(\Gamma_{ij})$ at the joint edges:*

$$\Gamma_{ij} := \{\overline{T}_i \cap \overline{T}_j : \quad (d-1)\text{-}measure \ |\overline{T}_i \cap \overline{T}_j| \neq 0\}, \quad i, j \in I,$$

then the a-posteriori error estimate holds with some $c_1, c_2 > 0$*:*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant c_1 \sum_{i\in I} h_i \|f + \Delta u_h\|_{L^2(T_i)} + c_2 \sum_{i,j\in I} \sqrt{h_i} \big\| [\![\frac{\partial u_h}{\partial \nu}]\!] \big\|_{L^2(\Gamma_{ij})}.$$

More details concerning discontinuous functions with jumps $[\![\cdot]\!]$ and the respecive Green's formula see in (Khludnev and Kovtunenko, 2000, Section 1.4).

18. Generalization of FEM

Recall the VP (10.1) stated in the abstract form: Find $u \in V$ such that

$$a(u,v) = f(v) \quad \text{for all } v \in V. \tag{18.1}$$

18.1. Non-consistent FEM. After discretization $a_h \neq a$, $f_h \neq f$ when using *numerical integration* typical for inhomogeneous coefficients, curved boundaries.

Example 18.1. (Ciarlet and Lions 1991, (25.19), p.188) The following *quadrature* for nodes $(x^j)_{j=1}^7$ such that x^1, x^2, x^3 are vertexes, x^4, x^5, x^6 are middles of edges, and x^7 is the center of mass, is exact for $\mathbb{P}_3(T_i)$:

$$\int_{T_i} u(x)\,dx \approx \frac{|T_i|}{60} \big\{ 3\sum_{j=1}^{3} u(x^j) + 8\sum_{j=4}^{6} u(x^j) + 27 u(x^7) \big\}.$$

For a subspace $V_h \subset V$, let a *bilinear form* $a_h : V_h \times V_h \mapsto \mathbb{R}$ be

- *uniform continuous*: $\exists \tilde{a} > 0$ such that $|a_h(u_h, v_h)| \leqslant \tilde{a} \|u_h\|_V \|v_h\|_V$,
- *uniform coercive*: $\exists 0 < \underset{\sim}{a} \leq \tilde{a}$ such that $a_h(u_h, u_h) \geqslant \underset{\sim}{a} \|u_h\|_V^2$,

and $f_h \in V_h^*$ be a linear continuous functional. The Lax–Milgram Theorem 10.1 provides uniqueness of a solution to the *discrete problem*: Find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = f_h(v_h) \quad \text{for all } v_h \in V_h. \tag{18.2}$$

The first Strang lemma generalizes Cea's Lemma 11.1.

Lemma 18.1 (Strang). *(Ciarlet and Lions 1991, Th.26.1, p.192) For the solutions u,* u_h *of problems* (18.1), (18.2)*, discretization error admits the estimate:*

$$\begin{aligned}\|u - u_h\|_V \leqslant \inf_{v_h \in V_h} \Big\{ &\big(1 + \frac{\bar{a}}{\underset{\sim}{a}}\big) \|u - v_h\|_V \\ &+ \frac{1}{\underset{\sim}{a}} \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_V} + \frac{1}{\underset{\sim}{a}} \sup_{w_h \in V_h} \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|_V} \Big\}.\end{aligned}$$

18.2. Poisson equation with inhomogeneous coefficient. For example, consider the variational problem: For the given right-hand side $f \in W^{m,\infty}(\Omega)$, $m \geqslant 1$, and uniformly bounded coefficients $0 < \alpha_0 \leqslant \alpha(x) \leqslant \alpha_1$ for all $x \in \overline{\Omega}$, find $u \in H_0^1(\Omega)$ such that

$$\int_{\Omega} \big(\alpha(x) \nabla u(x) \cdot \nabla v(x) - f(x) v(x)\big)\,dx = 0 \quad \text{for all } v \in H_0^1(\Omega),$$

and its discretization by some *Gaussian quadrature* with weights $\omega_j \geq 0$, $j \in K$:

$$\sum_{i\in I} |T_i| \sum_{j\in K} \omega_j [\alpha(x^j) \nabla u_h(x^j) \cdot \nabla v_h(x^j) - f(x^j) v_h(x^j)] = 0 \quad \text{for all } v_h \in V_h.$$

For short we denote the (local) *error*:

$$E_i(u) = \int_{T_i} u(x)\,dx - |T_i| \sum_{j\in K} \omega_j u(x^j)$$

and assume that $E_i(u) = 0$ for $u \in \mathbb{P}_{m-1}$, that means that the quadrature is exact for polynomials of degree less than or equal to $m-1$.

Since it holds the *uniform continuity*:

$$\sum_{j\in K} \omega_j [\alpha \nabla u_h \cdot \nabla v_h](x^j) \leqslant \alpha_1 \sqrt{\sum_{j\in K} \omega_j |\nabla u_h(x^j)|^2} \sqrt{\sum_{j\in K} \omega_j |\nabla v_h(x^j)|^2},$$

and the *uniform coercivity*:

$$\sum_{j\in K} \omega_j \alpha |\nabla u_h|^2(x^j) \geqslant \alpha_0 \sum_{j\in K} \omega_j |\nabla u_h|^2,$$

we can apply the Lax–Milgram Theorem 10.1 providing existence of the unique discrete solution $u_h \in V_h$.

COROLLARY 18.1. *(Grossmann et al. 2007, Section 4.5.3) If $u \in H^2(\Omega)$, then from the first Strang Lemma 18.1 it follows the estimate:*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant c\{h|u|_{H^2(\Omega)} + h^m \|f\|_{W^{m,\infty}(\Omega)}\}, \quad c > 0.$$

18.3. Non-conforming FEM. We consider the variational problem (18.1) and the discretized problem (18.2) in the general case when $V_h \not\subset V$.

Set the sum of the vector spaces V and V_h as

$$Z_h := V + V_h = \{z = v + v_h : \quad v \in V,\, v_h \in V_h\},$$

and extend the bilinear form a_h to $a_h : Z_h \times Z_h \mapsto \mathbb{R}$ *continuously*:

$$\text{there exists } \hat{a} > 0 \text{ such that} \quad |a_h(w_h, z_h)| \leqslant \hat{a} |w_h|_{Z_h} |z_h|_{Z_h}, \quad w_h, z_h \in Z_h$$

with respect to the *seminorm*: $|w_h|^2_{Z_h} := a_h(w_h, w_h)$.

LEMMA 18.2 (Strang). *(Ciarlet and Lions 1991, Th.31.1, p.212) For the solutions u, u_h of problems* (18.1), (18.2) *and $V_h \not\subset V$, the error estimate holds:*

$$|u - u_h|_{Z_h} \leqslant (1 + \hat{a}) \inf_{w_h \in V_h} |u - w_h|_{Z_h} + \frac{1}{\sqrt{\underset{\sim}{a}}} \sup_{v_h \in V_h} \frac{|f_h(v_h) - a_h(u, v_h)|}{\|v_h\|_{V_h}}.$$

19. Petrov–Galerkin method

In two different Banach spaces U, V (where V is reflexive), given the bilinear form $a : U \times V \mapsto \mathbb{R}$ and the linear continuous functional $f \in V^*$, set the *generalized VP* extending (10.1): Find $u \in U$ such that

$$a(u, v) = f(v) \quad \text{for all } v \in V. \tag{19.1}$$

Generalization of the Lax–Milgram Theorem 10.1 is the following

19.1. Ladyzhenskaya–Babuška–Brezzi–Nečas (LBBN) theorem.

Theorem 19.1 (Ladyzhenskaya–Babuška–Brezzi–Nečas (LBBN)). *(Clason 2013, Th.8.1, p.58) Under the following assumptions of*

(i) *continuity:* $|a(u,v)| \leqslant \bar{a}\|u\|_U\|v\|_V$ *for* $u \in U$, $v \in V$,

(ii) *inf-sup condition:* $\displaystyle\inf_{u\in U}\sup_{v\in V}\frac{a(u,v)}{\|u\|_U\|v\|_V} \geqslant \underline{a} > 0,$

(iii) *injectivity:* $a(u,v) = 0$ *for all* $u \in U$ *follows* $v = 0$,

there exists a unique solution to problem (19.1) *satisfying the a-priori estimate*

$$\|u\|_U \leqslant \tfrac{1}{\underline{a}}\|f\|_{V^*}.$$

Remark 19.1. The ***inf-sup condition*** is equivalent to the following one

$$\sup_{v\in V}\frac{a(u,v)}{\|v\|_V} \geqslant \underline{a}\|u\|_U \quad \text{for all } u \in U.$$

Remark 19.2. In the case of $U = V$, the *coercivity* follows

- *inf-sup condition*: $\underline{a}\|u\| \leqslant \frac{a(u,u)}{\|u\|} \leqslant \sup\limits_{v\in V}\frac{a(u,v)}{\|v\|}$;
- *injectivity*: if $a(u,v) = 0$ for all $u \in V$, substitute here $u = v$, then $\underline{a}\|v\|^2 \leqslant a(v,v) = 0$ concludes $\|v\| = 0$.

Remark 19.3. In the case of $U = V$ and the symmetric form $a(u,v) = a(v,u)$, the inf-sup condition also follows *injectivity*: $\sup\limits_{u\in U}\frac{a(v,u)}{\|u\|} \geqslant \underline{a}\|v\|$ implies $v = 0$.

Take the *test space* $V_h \subset V$ and the *trial space* $U_h \subset U$. Even for $U = V$, it can be different $U_h \neq V_h$ (e.g. for different shape functions). The *Petrov–Galerkin approximation* reads: Find $u_h \in U_h$ such that

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in V_h. \tag{19.2}$$

Theorem 19.2 (Discrete version of LBBN). *(Clason 2013, Th.8.2, p.61) Let* $\dim U_n = \dim V_n$ *and the discrete inf-sup condition hold:*

$$\inf_{u_h\in U_h}\sup_{v_h\in V_h}\frac{a(u_h,v_h)}{\|u_h\|_U\|v_h\|_V} \geqslant \underset{\sim}{a} > 0.$$

Then the unique solution of (19.2) *exists and satisfies the stability estimate:*

$$\|u_h\|_U \leqslant \tfrac{1}{\underset{\sim}{a}}\|f\|_{V^*}.$$

19.2. Generalized Cea's lemma.

Lemma 19.1 (Generalization of Cea's lemma). *For solutions* u, u_h *of problems* (19.1) *and* (19.2), *discretization error is estimated by the best approximation error:*

$$\|u - u_h\|_U \leqslant \Big(1 + \frac{\bar{a}}{\underset{\sim}{a}}\Big)\inf_{w_h\in U_h}\|u - w_h\|_U.$$

20. Mixed VP

Let V, Λ be two Hilbert spaces, $f \in V^*$ be a linear continuous functional, and $a : V \times V \mapsto \mathbb{R}$, $b : V \times \Lambda \mapsto \mathbb{R}$ be two bilinear forms. Consider the *mixed primed-dual VP*: Find a solution pair $(u,v) \in V \times \Lambda$ such that

$$a(u,v) + b(v,\lambda) = f(v) \quad \text{for all } u \in V, \tag{20.1a}$$

$$b(u,\mu) = 0 \quad \text{for all } \mu \in \Lambda. \quad (= g(\mu) \text{ can be extended for } g \in \Lambda^*) \tag{20.1b}$$

20.1. Well-posedness theorem.

Theorem 20.1 (Well-posedness). *(Brezzi and Fortin 1991, Th.1.1, p.42)(Clason 2013, Th.10.1, p.75) Let the symmetric bilinear form a*

(i) *be continuous:* $|a(u,v)| \leqslant \bar{a}\|u\|_V\|v\|_V$ *for* $u \in U$, $v \in V$ *and*
(ii) *satisfy the inf-sup condition:* $\inf\limits_{u\in V} \sup\limits_{v\in V} \frac{a(u,v)}{\|u\|_V\|v\|_V} \geqslant \underline{a} > 0$,

the bilinear form b:

(iii) *be continuous:* $|b(v,\mu)| \leqslant \bar{b}\|v\|_V\|\mu\|_\Lambda$ *for* $v \in V$, $\mu \in \Lambda$ *and*
(iv) *satisfy the Ladyzhenskaya–Babuška–Brezzi (LBB) condition:*

$$\inf_{\mu\in\Lambda} \sup_{v\in V} \frac{b(v,\mu)}{\|v\|_V\|\mu\|_\Lambda} \geqslant \underline{b} > 0.$$

Then there exists the unique solution to (20.1) *satisfying the a-priori estimate*

$$\|u\|_V + \|\lambda\|_\Lambda \leqslant \frac{1}{\min\{\underline{a},\underline{b}\}}\|f\|_{V^*}.$$

20.2. Primal-dual variational formulation. Introduce a *convex cone* (i.e. $0 \in K$, and $u, v \in K$ follows $\alpha u + \beta v \in K$ for all $\alpha, \beta \in \mathbb{R}$) treating (20.1b) as a constraint:

$$K := \{v \in V : \quad b(v,\mu) = 0 \quad \text{for all } \mu \in \Lambda\},$$

its *orthogonal complement* with respect to a scalar product in V:

$$K^\perp := \{u \in V : \quad (u,v)_V = 0 \quad \text{for all } v \in K\},$$

and the *dual cone*:

$$K^* := \{\mu \in \Lambda : \quad b(v,\mu) = 0 \quad \text{for all } v \in V\}.$$

Corollary 20.1. *(Brenner et al. 2008, Lemma 12.2.12, p.335) By using the closed range theorem (Steinbach 2008, Th.3.6, p.48), the inf-sup condition for* $u, v \in V$ *in Theorem 20.1 can be relaxed to* $u, v \in K$.

Introduce the *objective function* $J : V \mapsto \mathbb{R}$ given by

$$J(v) := \frac{1}{2}a(v,v) - f(v),$$

and the *Lagrangian* $\mathcal{L} : V \times \Lambda \mapsto \mathbb{R}$ defined as the sum:

$$\mathcal{L}(u,v) := J(v) + b(v,\mu).$$

Theorem 20.2 (Optimality). *(Grossmann et al. 2007, Section 4.6.1) The problem* (20.1) *is equivalent to*

(i) ***minimax (saddle-point) problem**: Find* $(u,v) \in V \times \Lambda$ *such that* $\mathcal{L}(u,\mu) \leqslant \mathcal{L}(u,\lambda) \leqslant \mathcal{L}(v,\mu)$ *for all* $(v,\mu) \in V \times \Lambda$, *that is*

$$\mathcal{L}(u,\lambda) = \min_{u\in V}\max_{\mu\in\Lambda} \mathcal{L}(v,\mu); \tag{20.2}$$

(ii) ***constrained minimization (primal) problem**: Find* $u \in K$ *such that* $J(u) = \min\limits_{v\in K} J(v)$, *which is equivalent to*

$$a(u,v) = f(v) \quad \textit{for all } v \in K, \tag{20.3a}$$

*and its **dual/adjoint problem**: Find* $\lambda \in \Lambda$ *such that*

$$b(v,\lambda) = f(v) - a(u,v) \quad \textit{for all } v \in K^\perp. \tag{20.3b}$$

21. Mixed FEM

Within conforming approximation by finite dimensional subspaces $V_h \subset V$ and $\Lambda_h \subset \Lambda$, the *discrete mixed VP* reads: Find $(u_h, \lambda_h) \in V_h \times \Lambda_h$ such that

$$\text{(21.1a)} \qquad a(u_h, v_h) + b(v_h, \lambda_h) \;=\; f(v_h) \quad \text{for all } v_h \in V_h,$$

$$\text{(21.1b)} \qquad b(u_h, \mu_h) \;=\; 0 \quad \text{for all } \mu_h \in \Lambda_h.$$

Introduce the *discrete primal cone* (which is also convex):

$$K_h := \{v_h \in V_h : \quad b(v_h, \mu_h) = 0 \quad \text{for all } \mu_h \in \Lambda_h\}$$

such that (21.1) leads to the *discrete constrained VP*: Find $u_h \in K_h$ such that

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in K_h.$$

It is important to note that:

- since $K_h \not\subset K$ the constrained formulation is not conforming with (20.3a),
- the continuous inf-sup condition over $u, v \in K$ does not follow any discrete inf-sup condition over $u_h, v_h \in K_h$.

21.1. Well-posedness and error estimate.

Theorem 21.1 (Well-posedness and discretization error). *(Clason 2013, Th.10.4, p.78) If the discrete inf-sup and LBB conditions hold, respectively:*

$$\begin{cases} \sup\limits_{v_h \in K_h} \dfrac{a(u_h, v_h)}{\|v_h\|_V} \geqslant \underset{\sim}{a}\|u_h\|_V & \textit{for all } u_h \in K_h, \\ \sup\limits_{v_h \in K_h} \frac{b(v_h, \mu_h)}{\|v_h\|_V} \geqslant \underset{\sim}{b}\|\mu_h\|_\Lambda & \textit{for all } \mu_h \in \Lambda_h, \end{cases}$$

then there exists the unique solution to (21.1) *satisfying the stability estimate:*

$$\|u_h\|_V + \|\lambda_h\|_\Lambda \leqslant \frac{1}{\min\{\underset{\sim}{a}, \underset{\sim}{b}\}} \|f\|_{V^*},$$

and the discretization error can be estimated as

$$\text{(21.2)} \quad \|u - u_h\|_V + \|\lambda - \lambda_h\|_\Lambda \\ \leqslant \sqrt{2}\Big(1 + \frac{\max\{\overline{a}, \overline{b}\}}{\min\{\underset{\sim}{a}, \underset{\sim}{b}\}}\Big)\Big\{\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{\mu_h \in \Lambda_h} \|\lambda - \mu_h\|_\Lambda\Big\}.$$

Remark 21.1. Introducing the dual cone

$$K_h^* := \{\mu_h \in \Lambda_h : \quad b(v_h, \mu_h) = 0 \quad \text{for all } v_h \in V_h\}$$

and its *orthogonal complement* with respect to a scalar product in Λ:

$$(K_h^*)^\perp := \{\lambda_h \in \Lambda_h : \quad (\lambda_h, \mu_h)_\Lambda = 0 \quad \text{for all } \mu_h \in K_h^*\},$$

the LBB condition over $\mu_h \in \Lambda_h$ in Theorem 21.1 can be relaxed to $\mu_h \in (K_h^*)^\perp$. However, in this case we lose uniqueness and estimate for the dual variable λ_h.

21.2. Mixed FEM for Poisson equation. For $F \in L^2(\Omega)$ consider the Dirichlet problem written in the *primal form*: Find $\lambda \in H_0^1(\Omega)$ such that

$$\int_\Omega (\nabla\lambda \cdot \nabla\mu - F\mu)\,dx = 0 \quad \text{for all } \mu \in H_0^1(\Omega), \tag{21.3}$$

and in the *primal-dual form*: Find $u \in H(\operatorname{div},\Omega) =: V$, where the space

$$H(\operatorname{div},\Omega) = \{u = (u_1,\dots,u_d)^\top \in L^2(\Omega;\mathbb{R}^d): \quad \operatorname{div} u \in L^2(\Omega)\},$$

and $\lambda \in L^2(\Omega) =: \Lambda$ such that

$$\int_\Omega (u\cdot v + \lambda \operatorname{div} v)\,dx = 0 \quad \text{for all } v = (v_1,\dots,v_d)^\top \in H(\operatorname{div},\Omega), \tag{21.4a}$$

$$\int_\Omega \mu \operatorname{div} u\,dx = -\int_\Omega F\mu\,dx \quad \text{for all } \mu \in L^2(\Omega). \tag{21.4b}$$

The latter equation is equivalent to $\operatorname{div} u = -F$ by the fundamental lemma of calculus of variations. The notation of bilinear and linear forms as

$$a(u,v) := \int_\Omega u\cdot v\,dx, \quad b(v,\lambda) := \int_\Omega \lambda \operatorname{div} v\,dx, \quad g(\mu) := -\int_\Omega F\mu\,dx,$$

agrees the general form (21.1).

Lemma 21.1 (Equivalence). *(Grossmann et al. 2007, Lemma 4.81, p.273) Let $\lambda \in H^2(\Omega) \cap H_0^1(\Omega)$ exist and satisfy the Poisson equation $-\Delta\lambda = F$ in Ω. Then problems* (21.3) *and* (21.4) *are equivalent with $u = \nabla\lambda$.*

21.3. Well-posedness of mixed FEM for Poisson equation.

Lemma 21.2 (Well-posedness). *(Clason 2013, Lemma 10.6, p.82) The primal-dual VP* (21.4) *is uniquely solvable with the a-priori estimate:*

$$\|u\|_{H(\operatorname{div},\Omega)} + \|\lambda\|_{L^2(\Omega)} \leqslant \big(1 + \frac{1}{\underline{b}} + \frac{1}{\underline{b}^2}\big)\|F\|_{L^2(\Omega)}, \quad \frac{1}{\underline{b}} = \sqrt{1 + K_{\mathrm{P}}(\Omega)^2}.$$

21.4. Raviart–Thomas FE. Let $\overline{\Omega} = \bigcup_{i\in I} \overline{T_i}$ be an affine-equivalent triangulation. We define a *conforming FE* by the mean of finite-dimensional spaces:

$$\Lambda_h = \{\mu_h \in L^2(\Omega): \quad \mu_h \in \mathbb{P}_0(T_i),\, i \in I\} \quad \text{(i.e. piece-wise constant)},$$

$$V_h = \{v_h \in H(\operatorname{div},\Omega): \quad v_h(x) = b + cx,\ b \in \mathbb{R}^d,\ c \in \mathbb{R} \ \text{ for } x \in T_i,\ i \in I\}.$$

To determine b and c we aim at the two key properties:

(i) $\operatorname{div} v_h \in L^2(T_i)$ for $i \in I$,

(ii) the jump $[\![v_h]\!]\cdot n_{ij} = 0$ across the joint edges Γ_{ij} for $i,j \in I$.

The first property holds since $\operatorname{div} v_h = \operatorname{div}(b + cx) = cd$ is constant in each T_i. The second property is possible by the choice of coefficients b, c in V_h for plane edges

$$\Gamma_{ij} = \{x \in \mathbb{R}^d: \quad n_{ij}\cdot x = \gamma, \quad \gamma \in \mathbb{R}\},$$

because $v_h \cdot n_{ij} = (b + cx)\cdot n_{ij} = b\cdot n_{ij} + c\gamma$ is constant for $x \in \Gamma_{ij}$. Therefore, for every element T_i, we take shape functions $(\varphi_l)_{l=1}^k$ associated to the edge Γ_l midpoints y^l in the form $\varphi_l(x) = b^l + c_l x$ such that after multiplication with the normal vector n_m at Γ_m:

$$\varphi_l(y^m)\cdot n_m = \frac{1}{|\Gamma_m|}\delta_{lm}. \tag{21.5}$$

It forms a nodal basis with the *dual basis*: $\Psi_m(\varphi_l) := \int_{\Gamma_m} \varphi_l \cdot n^m \, dS_x = \delta_{lm}$.

Lemma 21.3. *(Grossmann et al. 2007, p.274) The Raviart–Thomas FE given by $(T_i, (\varphi_l)_{l=1}^k, (\Psi_l)_{l=1}^k)$ is a valid FE.*

Example 21.1. Consider the *unit triangle* in 2d with

- the vertexes $(0,0), (1,0), (0,1)$;
- the midpoints: $y^1 = (\frac{1}{2}, 0)$, $y^2 = (0, \frac{1}{2})$, $y^3 = (\frac{1}{2}, \frac{1}{2})$;
- the unit normal vector: $n_1 = (0,-1)$, $n_2 = (-1,0)$, $n_3 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$;
- the shape functions: $\varphi_1(x) = x + n_1$, $\varphi_2(x) = x + n_2$, $\varphi_3(x) = x$.

Now we calculate the scalar products

(l=1): $\varphi_1(y^1) \cdot n_1 = y^1 \cdot n_1 + |n_1|^2 = 1, \quad \varphi_1(y^2) \cdot n_2 = (y^2 + n_1) \cdot n_2 = 0,$
$\varphi_1(y^3) \cdot n_3 = (y^3 + n_1) \cdot n_3 = (\frac{1}{2}, -\frac{1}{2}) \cdot (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = 0;$
(l=2): $\varphi_2(y^1) \cdot n_1 = y^1 \cdot n_1 = 0, \quad \varphi_2(y^2) \cdot n_2 = y^2 \cdot n_2 + |n_2|^2 = 1,$
$\varphi_2(y^3) \cdot n_3 = (y^3 + n_2) \cdot n_3 = (-\frac{1}{2}, \frac{1}{2}) \cdot (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = 0;$
(l=3): $\varphi_3(y^1) \cdot n_1 = y^1 \cdot n_1 = 0, \quad \varphi_3(y^2) = y^1 \cdot n_2 = 0, \quad \varphi_3(y^3) \cdot n_3 = y^3 \cdot n_3 = \frac{1}{\sqrt{2}}$

which agree (21.5) with $|\Gamma_1| = |\Gamma_2| = 1$ and $|\Gamma_3| = \sqrt{2}$.

22. Variational theory of parabolic problems

Let index $p \in [1, \infty]$ and V be a reflective Banach space.

Definition 22.1. *Functions $u : (0,T) \mapsto V$ with the norm $\|u(t)\|_V \in L^p(0,T)$ form the Lebesgue–Bochner space $L^p(0,T;V)$ equipped with the norm*

$$\|u\|^p_{L^p(0,T:V)} := \int_0^T \|u(t)\|^p_V \, dt,$$

which is also a Banach space.

22.1. Gelfand triple.

Lemma 22.1 (Gelfand triple). *(Roubíček 2005, Lemma 7.3, p.191) Let H be a Hilbert space, the embedding $V \hookrightarrow H = H^* \hookrightarrow V^*$ be continuous such that $\|u\|_{V^*} \leqslant c_1 \|u\|_H \leqslant c_2 \|u\|_V$ with some $0 < c_1 \leq c_2$. For the dual index $q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$, the Sobolev–Bochner space*

$$W^{1,p,q}(V, V^*) := \{u \in L^p(0,T;V), \ \frac{du}{dt} \in L^q(0,T;V^*)\} \hookrightarrow C(0,t;H)$$

is embedded continuously, and dense when the embedding $V \hookrightarrow H$ is dense. The following formula for integration by parts holds for $u, v \in W^{1,p,q}(V, V^*)$:

$$\int_0^T \langle \frac{du}{dt}, v \rangle_{(V^*,V)} \, dt = -\int_0^T \langle u, \frac{dv}{dt} \rangle_{(V,V^*)} + (u,v)_H \big|_{t=0}^T.$$

In a Hilbert space V, for the given bilinear form $a(t, \cdot, \cdot) : V \times V \mapsto \mathbb{R}$, right-hand side $f(t, \cdot) \in L^2(0,T;V^*)$ and initial data $u^0 \in H$, consider the *parabolic VP*: Find $u \in W^{1,2,2}(V, V^*)$ such that

$$\int_0^T \big\{ \langle \frac{du}{dt}, v \rangle_{(V^*,V)} + a(u,v) - f(v) \big\} \, dt + (u(0) - u^0, v^0)_H = 0 \tag{22.1}$$

for all test functions $(v, v^0) \in L^2(0,T;V) \times H$.

22.2. Well-posedness theorem.

THEOREM 22.1 (Well-posedness). *(Schöberl 2016, Ern and Guermond 2004, Th.6.6., p.282) Let the mapping $t \mapsto a(t,\cdot,\cdot)$ for all $t \in (0,T)$ be*

(i) *continuous;*
(ii) *uniformly bounded: $\exists \overline{a} > 0$ such that $|a(t,u,v)| \leqslant \overline{a}\|u\|_V\|v\|_V$;*
(iii) *uniformly coercive: $\exists 0 < \underline{a} \leq \overline{a}$ such that $\underline{a}\|u\|_V^2 \leqslant a(t,u,v)$.*

Then, a unique solution to (22.1) exists and satisfies the a-priori estimate:

$$\|u\|_{W^{1,2,2}(V,V^*)} := \|u\|_{L^2(0,T;V)} + \Big\|\frac{du}{dt}\Big\|_{L^2(0,T;V^*)} \leqslant \frac{1}{\underline{a}}\big(\|f\|_{L^2(0,T;V^*)} + \|u^0\|_H\big).$$

EXERCISE 22.1. Prove the a-priori estimate, which follows in a usual way from the inf-sup condition.

22.3. Space-time Galerkin method. Within the conforming approximation, choose *finite dimensional* subspaces $V_k \subset V$ for every t_k from the *time-grid*: $0 = t_0 < t_1 < \ldots < t_M = T$. Denoting $\tau_k := t_k - t_{k-1}$ we consider:

- *Piecewise-linear in time trial functions*:

$$u_h(t) := \frac{t_k - t}{\tau_k} u_h(t_{k-1}) + \frac{t - t_{k-1}}{\tau_k} u_h(t_k), \quad t \in [t_{k-1}, t_k], \quad k = 1, \ldots, M.$$

Note that, if all $u_h(t) \in V_k \subset V$, then $u_h(t) \in W^{1,\infty}(0,T;V) \hookrightarrow W^{1,2,2}(V,V^*)$ because of the embedding $V_k \subset V \subset V^*$.

- *Piecewise-constant in time test functions*:

$$v_h(t) := v_h(t_{k-1}), \quad t \in (t_{k-1}, t_k], \quad k = 1, \ldots, M.$$

Note that, if $v_k(t_k) \in V_k \subset V$, then $v_h(t) \in L^\infty(0,T;V) \hookrightarrow L^2(0,T;V)$.

By inserting the ansatz $u = u_h$, $v = v_h$ we discretize problem (22.1) as follows

$$\sum_{k=1}^M \Big\{ \int_{t_{k-1}}^{t_k} \langle \frac{du_h(t)}{dt}, v_h \rangle_{(V^*,V)}\, dt + \int_{t_{k-1}}^{t_k} a(u_h(t), v_h)\, dt \Big\} = \sum_{k=1}^M \int_{t_{k-1}}^{t_k} f(v_h)\, dt.$$

Expressing the discrete time derivative $\frac{du_h(t)}{dt} = \frac{u_h(t_k) - u_h(t_{k-1})}{t_k}$ and using the trapezoidal rule $\int_{t_{k-1}}^{t_k} a(u_h(t), v_h)\, dt = a(\int_{t_{k-1}}^{t_k} u_h(t)\, dt, v_h) = a(\tau_k \frac{u_h(t_k) + u_h(t_{k-1})}{2}, v_h)$, we can rewrite the scheme as iterations:

$$\begin{cases} \langle u_h(t_k) - u_h(t_{k-1}), v_h \rangle_{(V^*,V)} + \dfrac{\tau_k}{2} a(u_h(t_k) + u_h(t_{k-1}), v_h) = \displaystyle\int_{t_{k-1}}^{t_k} f(v_h)\, dt \\ \text{for } k = 1, \ldots, M; \qquad u_h(t_0) = u^0. \end{cases}$$

This implies a semi-discrete Rothe method with the midpoint rule. But no any discrete inf-sup condition holds.

23. Discontinuous Galerkin method for parabolic VP

For motivation we consider a non-differentiable function $u \in L^2(0,T;V)$ which time-derivative $\frac{du}{dt}$ can be still defined in the weak sense:

$$\int_0^T \langle \frac{du}{dt}, v \rangle_{(V^*,V)}\, dt = -\int_0^T \langle u, \frac{dv}{dt} \rangle_{(V,V^*)}\, dt + (u,v)\big|_{t=0}^T$$

for $v \in W^{1,2,2}(V,V^*)$, thus allowing *discontinuity in time.*

23.1. Space of discontinuous in time functions. Discretize $(0,T)$ by a time grid $0=t_0<t_1<\ldots<t_M=T$ with $\tau_k=t_k-t_{k-1}$ for $k=1,\ldots,M$, and consider *continuous from the left* functions $u(t_k):=u(t_k^-)$, where $t_k^\pm=\lim_{s\to 0}(t_k\pm s)$, allowing *jumps* $[\![u(t_k)]\!]:=u(t_k^+)-u(t_k)$.

Lemma 23.1 (Positivity property). *For a function $u\in W^{1,2}(t_{k-1},t_k;V)$ such that $\frac{du}{dt}\in W^{1,2}(t_{k-1},t_k;V^*)$, the following formula holds:*

$$(23.1)\quad b_0(u,u):=\sum_{k=1}^M\int_{t_{k-1}}^{t_k}\langle\frac{du}{dt},u\rangle_{(V^*,V)}\,dt+\sum_{k=2}^M([\![u(t_{k-1})]\!],u(t_{k-1}^+))_H$$
$$+\|u(t_0^+)\|_H^2=\frac{1}{2}\sum_{k=1}^M\|[\![u(t_{k-1})]\!]\|_H^2+\frac{1}{2}\|u(t_M)\|_H^2+\frac{1}{2}\|u(t_0^+)\|_H^2\geq 0.$$

Now define the ***discrete test space*** $Y_h\subset L^2(0,T;V)$ supported with the norm

$$\|v_h\|_{Y_h}^2:=\sum_{k=1}^M\int_{t_{k-1}}^{t_k}\|v_h\|_V^2\,dt=\int_{t_0}^{t_M}\|v_h\|_V^2\,dt=:\|v_h\|_{L^2(0,T;V)}^2$$

and the ***discrete trial space*** $X_h\subset L^2(0,T;V)$ with the mesh-dependent norm

$$\|u_h\|_{X_h}^2$$
$$:=\sum_{k=1}^M\int_{t_{k-1}}^{t_k}[\|u_h\|_V^2+\|\frac{du_h}{dt}\|_{V^*}^2]\,dt+\sum_{k=2}^M\frac{1}{\tau_k}\|[\![u_h(t_{k-1})]\!]\|_H^2+\frac{1}{\tau_1}\|u_h(t_0^+)\|_H^2.$$

A typical approximation is realized by piecewise p-polynomials:

$$X_h=Y_h=\{v_h:\,(t_{k-1}^+,t_k)\mapsto\mathbb{R},\quad v_h\in\mathbb{P}_p\quad\text{for }k=1,\ldots,M\}.$$

Since for polynomials $v_h\sim\big(\frac{t_k-t}{\tau_k}\big)^pu_h(t_{k-1}^+)+\big(\frac{t-t_{k-1}}{\tau_k}\big)^pu_h(t_k)$, then the integral

$$\int_{t_{k-1}}^{t_k}\|v_h\|_V^2\,dt\sim\frac{1}{\tau_k^{2p}}\int_{t_{k-1}}^{t_k}t^{2p}\,dt\,\|v_h(t_{k-1}^+)\|_V^2\sim\tau_k\|v_h(t_{k-1}^+)\|_V^2.$$

Therefore, there exists $c_p>0$ such that the estimate holds for $k=1,\ldots,M$:

$$(23.2)\qquad \tau_k\|v_h(t_{k-1}^+)\|_V^2\leqslant c_p\int_{t_{k-1}}^{t_k}\|v_h\|_V^2\,dt.$$

The discretized continuous problem (22.1) implies iterations: Find $u_h\in X_h$ such that

$$\begin{cases}\displaystyle\int_{t_{k-1}}^{t_k}[a(v_h,v_h)+\big\langle\tfrac{du_h}{dt},v_h\big\rangle_{(V^*,V)}-f(v_h)]\,dt+\big([\![u_h(t_{k-1})]\!],v_h(t_{k-1}^+)\big)_H=0\\ \text{for all }v_h\in Y_h,\quad k=1,\ldots,M;\qquad u_h(t_0)=u^0.\end{cases}$$

After summation over k we get an equivalent, *discrete Galerkin (DG)* scheme:

$$(23.3)\quad b_1(u_h,v_h):=\sum_{k=1}^M\int_{t_{k-1}}^{t_k}[a(u_h,v_h)+\big\langle\frac{du_h}{dt},v_h\big\rangle_{(V^*,V)}-f(v_h)]\,dt$$
$$+\sum_{k=2}^M\big([\![u_h(t_{k-1})]\!],v_h(t_{k-1}^+)\big)_H+(u_h(t_0^+),v_h(t_0^+))_H=\int_0^Tf(v_h)\,dt+(u^0,v_h(t_0^+))_H.$$

Example 23.1. The *piecewise-constant approximation* ($p = 0$) implying FVM:

$$X_h = Y_h = \{v_h(t) = v_h(t_k) \text{ for } t \in (t_{k-1}^+, t_k],\ k = 1, \dots, M\}$$

follows $\frac{du_h}{dt} = 0$, and (23.3) results in the *explicit Euler scheme* for $k = 1, \dots, M$:

$$(u_h(t_k) - u_h(t_{k-1}), v_h(t_k))_H + \tau_k a(u_h(t_k), v_h(t_k)) = \int_{t_{k-1}}^{t_k} f(v_h(t_k))\, dt.$$

23.2. Existence based on discrete LBBN theorem.

Theorem 23.1 (Existence). *(Clason 2013, Th.12.1, p.96) There exists a unique solution to DG problem* (23.3)*.*

23.3. Stability estimate.

Remark 23.1. From the inf-sup condition it follows the stability estimate in the mesh-dependent norm:

$$\underset{\sim}{b}\|u_h\|_{X_h} \leqslant \sup_{v_h \in Y_h} \frac{b_1(u_h, v_h)}{\|v_h\|_{Y_h}} = \sup_{v_h \in Y_h} \frac{\int_0^T f(v_h)\, dt + (u^0, v_h(t_0^+))_H}{\|v_h\|_{Y_h}}$$
$$\leqslant \|f\|_{L^2(0,T:V^*)} + c_2\sqrt{c_p}\frac{1}{\sqrt{\tau_1}}\|u^0\|_H.$$

Alternatively, using (23.1) and inserting $v_h = u_h$ in (23.3) provides the lower bound:

$$b_1(u_h, u_h) \ \geqslant\ \underline{a}\|u_h\|_{Y_h}^2 \ +\ \frac{1}{2}\sum_{k=2}^{M} \|[\![u_h(t_{k-1})]\!]\|_H^2 \ +\ \frac{1}{2}\|u_h(t_k)\|_H^2 \ +\ \frac{1}{2}\|u_h(t_0^+)\|_H^2,$$

and the upper bound:

$$b_1(u_h, u_h) \leqslant \frac{1}{2\underline{a}}\|f\|_{L^2(0,T;V^*)} + \frac{\underline{a}}{2}\|u_h\|_{Y_h}^2 + \|u^0\|_H^2 + \frac{1}{4}\|u_h(t_0^+)\|_H^2.$$

23.4. Discretization error estimate. In the following we will utilize the *local projection* operator $\pi : C((t_{k-1}^+, t_k]) \mapsto \mathbb{P}_p$ given by

$$\begin{cases} (i)\ \pi u(t_{k-1}^+) = u(t_{k-1}^+), \\ (ii) \text{ if } p \geqslant 1, \text{ then } \quad \int_{t_{k-1}}^{t_k} (u - \pi u)\varphi\, dt = 0 \quad \text{for all } \varphi \in \mathbb{P}_{p-1}((t_{k-1}^+, t_k]). \end{cases}$$

For $t \mapsto u : C^{p+1}((t_{k-1}^+, t_k])$, Theorem 15.1 provides the local *error estimate*:

$$\|u - \pi u\|_V \leqslant c\tau_k^{p+1} \int_{t_{k-1}}^{t_k} \|\frac{d^{p+1}u}{dt^{p+1}}\|_{V^*}\, dt = \mathrm{O}(\tau_k^{p+1}). \tag{23.4}$$

Theorem 23.2 (Discretization error). *(Clason 2013, Th.12.4, p.102) For a sufficiently smooth solution u of* (22.1) *such that the interpolation error estimate* (23.4) *holds, the error of descretization by* (23.3) *is estimated as*

$$\|u - u_h\|_{Y_h} = \mathrm{O}(\tau^{p+1}), \quad \tau := \max_{k=1,\dots,M} \tau_k.$$

24. FEM for second order hyperbolic VP

In the time-space cylinder $Q_T := (0,T)\times\Omega$ which base $\Omega \subset \mathbb{R}^d$ has the Lipschitz boundary $\partial\Omega =: \Gamma$, consider the elliptic operator in the divergence form:

$$L := -\sum_{j=1}^{d} \frac{\partial}{\partial x_j}\big(a_{ij}(t,x)\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\big)$$

with uniformly bounded and elliptic coefficients possessing a_{ij}, $\frac{\partial a_{ij}}{\partial t} \in L^\infty(Q_T)$, and the *hyperbolic IBVP* (e.g. the *wave equation* when $L = -\Delta$):

$$\frac{\partial^2 u}{\partial t^2} + Lu = f \text{ in } Q_T; \quad u = 0 \text{ on } (0,T)\times\Gamma; \quad u(0) = u^0,\ \frac{\partial u}{\partial t}(0) = v^0 \text{ in } \Omega.$$

We set the ***Gelfand triple*** of spaces $V := H_0^1(\Omega) \subset H := L^2(\Omega) \subset V^* = H^{-1}(\Omega)$.

For the given right-hand side $f \in L^2(0,T;H)$, initial data $u^0 \in V$ and $v^0 \in H$, the *weak variational formulation* reads: Find $u \in L^\infty(0,T;V)$, $v := \frac{du}{dt} \in L^\infty(0,T;H)$, $\frac{dv}{dt} = \frac{d^2u}{dt^2} \in L^2(0,T;V^*)$ such that

$$\langle\frac{d^2u}{dt^2},\bar u\rangle_{(V^*,V)} + \int_\Omega \sum_{i,j=1}^{d} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial \bar u}{\partial x_j}\,dx = (f,\bar u)_H \quad \text{for all } \bar u \in V, \tag{24.1a}$$

$$u(0) = u^0; \quad \frac{\partial u}{\partial t}(0) = v^0 \quad \text{in } \Omega. \tag{24.1b}$$

We denote by $a(u,\bar u) := \int_\Omega \sum_{i,j=1}^{d} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial \bar u}{\partial x_j}\,dx$, and $(f,\bar u)_H := \int_\Omega f\bar u\,dx$.

24.1. Well-posedness theorem. Later on, the notion of *energy* will be used.

Definition 24.1 (Energy). *The* ***energy*** *for the hyperbolic equation* (24.1a) *is determined as follows:*

$$E(u(t)) := \frac{1}{2}\big\|\frac{du}{dt}(t)\big\|_H^2 + \frac{1}{2}a(u(t),u(t)).$$

Theorem 24.1 (Well-posedness). *(Hunter 2014, Th.7.3, p.213) A unique solution to the hyperbolic VP* (24.1) *exists and satisfies the a-priori estimate:*

$$\sup_{t\in(0,T)} E(u(t)) + \big\|\frac{d^2u}{dt^2}\big\|_{L^2(0,T;V^*)} \leqslant c\,r(T), \quad c > 0,$$

where we have marked $r(t) := \|u^0\|_V^2 + \|v^0\|_H^2 + \int_0^t \|f\|_H^2\,ds$ *for short.*

24.2. Semi-discretization by FE. Within conforming approximation, set *finite-dimensional subspaces* $V_h \subset V$ by $V = \operatorname{span}\{\varphi_j\}_{j=1}^N$ by truncating the basis $(\varphi_j)_{j=1}^\infty$ in V. Now insert in (24.1) the truncated series

$$u_h(t,x) = \sum_{j=1}^{N} u_j(t)\varphi_j(x)$$

to get the *discrete problem*: Find $u_h \in V_h$ such that

$$\frac{d^2u_h}{dt^2} + L_h u_h = f_h; \quad u_h(0) = u_h^0, \quad \frac{du_h}{dt}(0) = v_h^0$$

by means of the following *ODE system*:

$$\sum_{j=1}^{N}\big[(\varphi_j,\varphi_i)_H\frac{d^2u_j}{dt^2}+a(\varphi_j,\varphi_i)u_j\big]=(f,\varphi_i)_H,\quad i=1,\dots,N. \tag{24.2}$$

The property of linear independence of the basis provides the system matrix to be invertible, hence there exists a unique solution.

We define the *Ritz projector* in the discrete space $V_h\subset V$ of continuous in $\overline{\Omega}$, piecewise-linear functions endowed by the h-dependent norm:

$$\|u-u_h\|^2_{V_h}:=\|u-u_h\|^2_{L^2(\Omega)}+h^2\|\nabla(u-u_h)\|^2_{L^2(\Omega)}.$$

Definition 24.2. *The Ritz projector: $R_h:V\mapsto V_h$ is defined by the function $R_hu\in V_h$ solving the discrete VP:*

$$a(R_hu,\bar u_h)=a(u,\bar u_h)\quad \textit{for all } \bar u_h\in V_h.$$

If $u\in H^2(\Omega)$, then the interpolation error of the Ritz projector is estimated due to Theorem 15.1 as follows

$$\|u-R_hu\|^2_{V_h}=\|u-R_hu\|^2_{L^2(\Omega)}+h^2\|\nabla(u-R_hu)\|^2_{L^2(\Omega)}\leqslant c\,h^4\|u\|^2_{H^2(\Omega)}. \tag{24.3}$$

Theorem 24.2 (Space-discretization error). *(Grossmann et al. 2007, Th.5.35, p.361) Let the solution of the continuous problem* (24.1) *be smooth such that it holds $u(t),v(t)\in H^2(\Omega)$ and $\frac{d^2u}{dt^2}\in L^2(0,T;H^2(\Omega))$. The error by discretization* (24.2) *admits the estimate:*

$$\|u-u_h\|^2_{V_h}+\big\|\frac{d}{dt}(u-u_h)\big\|^2_{L^2(\Omega)}\leqslant c\big\{\|\nabla(R_hu^0-u^0_h)\|^2_{L^2(\Omega)}+\|R_hv^0-v^0_h\|^2_{L^2(\Omega)}$$
$$+h^4\big[\|u\|^2_{H^2(\Omega)}+\big\|\frac{du}{dt}\big\|^2_{H^2(\Omega)}+\int_0^T\big\|\frac{d^2u}{dt^2}\big\|_{H^2(\Omega)}\,ds\big]\big\},\quad c>0.$$

24.3. Semi-discretization in time. Now consider a *semi-discretization by Rothe's method*: For equidistant grid points $t_k=\tau k$, $k=1,\dots,M$ of the size $\tau=\frac{T}{M}$, find $u^k\in V$ solving

$$(D^+_\tau D^-_\tau u^k,\bar u)_H+a(\frac{1}{2}(u^{k+1/2}+u^{k-1/2}),\bar u)=(f(t_k),\bar u)_H \tag{24.4}$$

for all $\bar u\in V$ and $k=1,\dots,M-1$, with the notation for means:

$$u^{k+1/2}:=\frac{1}{2}\big(u^k+u^{k+1}\big),\quad u^{k-1/2}:=\frac{1}{2}\big(u^k+u^{k-1}\big).$$

In other words, solve iteratively

$$\frac{1}{\tau^2}(u^{k+1}+u^{k-1}-2u^k,\bar u)+a(\frac{1}{4}(u^k+u^{k+1})+\frac{1}{4}(u^k+u^{k-1}),\bar u)=(f(t_k),\bar u)_H$$

starting from $u^0,u^1\in V$, which implies the *elliptic BVP*:

$$\frac{1}{\tau^2}(u^{k+1},\bar u)_H+\frac{1}{4}a(u^{k+1},\bar u)=(f(t_k),\bar u)_H+\frac{1}{\tau^2}(2u^k-u^{k-1})-\frac{1}{4}a(2u^k+u^{k-1},\bar u).$$

Its right-hand side presents a linear continuous functional, which we denote by $b^k(\bar u)$ for short. The means in (24.4) are motivated by the following result.

Lemma 24.1 (Conservation of energy). *(Grossmann et al. 2007, Lemma 5.39, p.369) If $f\equiv 0$, then for all $k=1,\dots,M-1$ the energy is conserved:*

$$E_\tau(u^k):=\frac{1}{2}\|D^+_\tau u^k\|^2_H+\frac{1}{2}a(u^{k+1/2},u^{k+1/2})=E_\tau(u^0).$$

24.4. Time-discretization error.

THEOREM 24.3 (Time-discretization error). *(Grossmann et al. 2007, Th.5.40, p.370) If the solution of problem* (24.1) *is smooth such that* $t \mapsto u : C^3([0,T])$*, then the error is estimated in the energy norm as:*

$$E_\tau(u(t_k) - u^k) \leqslant c(\tau^4 + \|D_\tau^+(u(t_0) - u^0)\|_H^2), \quad k = 1, \dots, M-1, \quad c > 0.$$

REMARK 24.1. The full time-space discretization can be treated by combination of Theorem 24.2 and Theorem 24.3 together.

25. Spectral Methods

In a Hilbert space V with inner product $(\cdot, , \cdot)_V$, consider an abstract *variational problem*: For given $f \in V^*$ find $u \in V$ such that

$$(u, v)_V = f(v) \quad \text{for all } v \in V. \tag{25.1}$$

Two essential ingredients of any discretization are:

- choice of a *finite-dimensional subspace* $V_N \subset V$ (of dimension $N \in \mathbb{N}$);
- choice of a suitable *projection* $P_N : V \mapsto V_N$ satisfying $P_N^2 = P_N$ and the approximation property $\|u - P_N u\|_V \to 0$ as $N \to \infty$.

The *Galerkin approximation* implies: for a basis $(\varphi_j)_{j=1}^\infty$ given in V, set the finite subspace $V_N = \operatorname{span}\{\varphi_1, \dots, \varphi_N\}$ and approximate u by the truncated series

$$u^N := \sum_{j=1}^N u_j \varphi_j \in V_N, \quad u_j \in \mathbb{R},\ j = 1, \dots, N. \tag{25.2}$$

Inserting (25.2) in (25.1) and testing it with $v = \varphi_i$, this gives the *Galerkin equation*: Find $u^N \in V_N$ such that

$$\sum_{j=1}^N (\varphi_j, \varphi_i)_V u_j = f(\varphi_i), \quad i = 1, \dots, N. \tag{25.3}$$

The coefficients $p_{ij} := (\varphi_i, \varphi_j)_V$ compose the system matrix $P = (p_{ij})_{i,j=1}^N$. Since the basis is linearly independent, then P is nonsingular and there exists the inverse matrix $P^{-1} = (q_{ij})_{i,j=1}^N$ such that $\sum_{l=1}^N q_{kl} p_{li} = \delta_{ki}$. In this case, the projection can be defined as follows:

$$P_N u = \sum_{j=1}^N u_j \varphi_j, \quad \text{where } u_j := \sum_{i=1}^N (u, \varphi_i)_V q_{ij}.$$

EXERCISE 25.1. Check that $P_N^2 u = P_N u$ by inserting $u = P_N u$.

Further we consider typical choices of basis and projection by spectral methods.

25.1. Approximation by the Fourier method. For piecewise-continuous functions $u(x)$, $x \in (0, 2\pi]$, the complex-valued *Fourier series*:

$$u^\infty(x) = \sum_{j \in \mathbb{Z}} u_j e^{\imath jx}, \quad e^{\imath jx} = \cos(jx) + \imath \sin(jx), \quad u_j := \int_0^{2\pi} u(x) e^{-\imath jx}\, dx$$

is *periodic* such that $u^\infty(0) = u^\infty(2\pi) = \frac{1}{2}[u(0^+) - u(2\pi^-)]$ for $x^\pm = \lim_{s \to 0}(x \pm s)$, where $\imath$ stands for the imaginary unit such that $\imath^2 = -1$.

If we split it in the real and the imaginary parts, then we obtain

- *Fourier sine series*: $u^{\sin}(x) = -u^{\sin}(-x) = \sum_{j\in\mathbb{Z}} u_j \sin(jx)$ with $u_j := \frac{2}{\pi}\int_0^\pi u(x)\sin(jx)\,dx$;
- *Fourier cosine series*: $u^{\cos}(x) = u^{\cos}(-x) = \sum_{j\in\mathbb{Z}} u_j \cos(jx)$ with $u_j := \frac{2}{\pi(1+\delta_{j0})}\int_0^\pi u(x)\cos(jx)\,dx$.

Consider the *truncated Fourier series*:

$$u^N(x) := \sum_{j=-N}^{N} u_j e^{\imath jx}.$$

Lemma 25.1 (Fourier approximation). *(Gottlieb and Orszag 1977, (3.18)–(3.19), p.26–27) If the derivatives* $\frac{\partial^k u}{\partial x^k}$ *are continuous, periodic (i.e.* $\frac{\partial^k u}{\partial x^k}(0) = \frac{\partial^k u}{\partial x^k}(2\pi)$*) for* $k = 0, \dots, n-1$*, and* $\frac{\partial^n u}{\partial x^n}$ *is piecewise continuously differentiable in* $x \neq x_0$*, then*

$$u_j = \mathrm{O}(\tfrac{1}{j^n}), \quad u^N(x) - u(x) = \begin{cases} \mathrm{O}(\frac{1}{N^n}), & x \neq x_0, \\ \mathrm{O}(\frac{1}{N^{n-1}}), & x - x_0 = \mathrm{O}(\frac{1}{N}). \end{cases}$$

Consider the IBVP for *1d heat equation*:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, \quad t > 0,\, x \in \Omega; \\ u = 0, \quad x \in \partial\Omega; \quad u = u^0, \quad t = 0. \end{cases} \tag{25.4}$$

There exists a smooth solution $u(t) \in C^1(0,T;L^2(\Omega)) \cap C(0,T;V)$ in the space $V = H_0^1(\Omega) \cap H^2(\Omega)$ satisfying the variational equation in the strong form:

$$\int_\Omega \Big(\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}\Big) v\,dx = 0 \quad \text{for all } v \in L^2(\Omega).$$

Example 25.1. In the interval $\Omega = (0,\pi)$ with the boundary $\partial\Omega = \{0,\pi\}$, analytical solution of (25.4) is given by the Fourier sine series:

$$u(t,x) = \sum_{j=1}^{\infty} u_j(t)\sin(jx), \quad u_j(t) = e^{-j^2 t}u_j^0,$$

where u_j^0 are coefficients in the expansion of initial data:

$$u^0(x) = \sum_{j=1}^{\infty} u_j^0 \sin(jx), \quad u_j^0 := \frac{2}{\pi}\int_0^\pi u^0(x)\sin(jx)\,dx.$$

For $V_N = \mathrm{span}\{\sin(x), \dots, \sin(Nx)\}$, every basis function $\sin(jx) \in V$ for $j = 1, \dots, N$. The corresponding *spectral approximation*:

$$u^N(t,x) = \sum_{j=1}^{N} u_j(t)\sin(jx)$$

solves the *semi-discrete problem*: $\frac{\partial}{\partial t}(P_N u) - P_N \frac{\partial^2}{\partial x^2}(P_N u) = 0, \quad P_N u(0) = P_N u^0$. Here the projection $P_N u = u^N$ implies component-wisely the ODE for coefficients:

$$\frac{du_j}{dt} = -j^2 u_j, \quad u_j(0) = u_j^0, \quad j = 1, \dots, N.$$

Since all derivatives of the solution are continuous and periodic in $(0,2\pi]$, according to Lemma 25.1 it follows the *exponential convergence* (Gottlieb and Orszag 1977, p.2): $u^N(t,x)-u(t,x)=\mathrm{O}(e^{-N^2t})$.

25.2. Chebyshev polynomial approximation. The Chebyshev polynomials can be defined recursively for $x\in[-1,1]$:

$$T_0(x)=1,\quad T_1(x)=x,\quad T_{j+1}(x)=2xT_j(x)-T_{j-1}(x)\quad \text{for } j\geq 1.$$

They are orthogonal with respect to the weighted scalar product

$$(T_j,T_l)_\omega:=\int_{-1}^1\frac{1}{\sqrt{1-x^2}}T_j(x)T_l(x)\,dx=\frac{\pi(1+\delta_{j0})}{2}\delta_{jl}=\begin{cases}0, & j\neq l,\\ \pi, & j=l=0,\\ \frac{\pi}{2}, & j=l\neq 0.\end{cases}$$

Since $T_j(\cos\theta)=\cos(j\theta)$, it is equivalent to the Fourier cosine series for $\theta\in[0,\pi]$.

Lemma 25.2. *(Gottlieb and Orszag 1977, p.28) The Chebyshev polynomial approximation defined by*

$$u^{\mathrm{T}}(x)=\sum_{j=0}^{\infty}u_jT_j(x),\quad u_j=\frac{2}{\pi(1+\delta_{j0})}(u,T_j)_\omega$$

has the following properties:

(i) $u^{\mathrm{T}}(x)=\frac{1}{2}[u(x^+)+u(x^-)]$, $\quad u^{\mathrm{T}}(-1)=u(-1^+)$, $u^{\mathrm{T}}(1)=u(1^-)$;

(ii) *if* $\frac{\partial^k u}{\partial x^k}(x)$ *are continuous for* $k=0,\dots,n-1$ *and* $\frac{\partial^n u}{\partial x^n}(x)$ *is integrable, since* $|T_j(x)|\leqslant 1$, *then* $u_j=\mathrm{O}(\frac{1}{j^n})$ *and* $u^{\mathrm{T}}(x)-u^N(x)=\mathrm{O}(\frac{1}{N^{n-1}})$.

These properties are directly inherited from the Fourier cosine series.

Example 25.2. Consider the 1d heat equation (25.4) in $\Omega=(-1,1)$.

According to *τ-method*, we use the ansatz: $u^N(t,x)=\sum\limits_{j=0}^{N}u_j(t)T_j(x)=:P_Nu$.

Insert the ansatz in BC: $\sum_{j=0}^N u_jT_j(-1)=\sum_{j=0}^N u_jT_j(1)=0$, then

$$\sum_{j=0}^{N}u_j=\sum_{j=0}^{N}(-1)^ju_j=0, \tag{25.5}$$

that reduces two degrees of freedom, next insert it in the variational equation:

$$0=\big(\frac{\partial(P_Nu)}{\partial t}-P_N\frac{\partial^2(P_Nu)}{\partial x^2},P_Nv\big)_\omega=\big(\frac{\partial u^N}{\partial t}-\frac{\partial^2u^N}{\partial x^2},v^N\big)_\omega.$$

For $v^N=\sum_{l=0}^N v_lT_l(x)$, because of $v_N=v_{N-1}=0$, we get $N-1$ equations:

$$\sum_{j=0}^{N}\frac{du_j}{dt}(T_j,T_l)_\omega=\big(\frac{d^2}{dt^2}\big(\sum_{j=0}^{N}u_jT_j\big),T_l\big)_\omega,\quad l=0,\dots,N-2.$$

Using the differentiation formula (Gottlieb and Orszag 1977, (A.10), p.160):

$$\frac{d^2}{dt^2}\big(\sum_{j=0}^{N}u_jT_j\big)=\sum_{j=0}^{N}\big(\frac{1}{1+\delta_{j0}}\sum_{k=j+2,\;k+j\text{ even}}^{N-2}k(k^2-j^2)u_k\big)T_j=:u_j^{(2)}T_j,$$

$$j=0,\dots,N-2;\quad u_{N-1}^{(2)}=u_N^{(2)}:=0,$$

it follows the ODE system for $l = 0, \dots, N-2$:

$$\frac{du_l}{dt} = u_l^{(2)} + (N-1)[(N-1)^2 - l^2]u_{N-1} + N[N^2 - l^2]u_N. \tag{25.6}$$

Summing and subtracting (25.5), we can exclude u_{N-1} and u_N from (25.6) by

$$\sum_{j=0,\text{ even}}^{N-2} u_j = -\begin{cases} u_N, & N \text{ even},\\ u_{N-1}, & N \text{ odd}; \end{cases} \qquad \sum_{j=0,\text{ odd}}^{N-2} u_j = -\begin{cases} u_{N-1}, & N \text{ even},\\ u_N, & N \text{ odd}. \end{cases}$$

25.3. Collocation (pseudo-spectral approximation). Choose collocation points $(x^j)_{j=1}^N \in \Omega$ and the projection $P_N u := u^N(t,x) = \sum_{j=1}^N u_j(t)\varphi_j(x)$, where $u_j(t)$ solves the *collocation equation*:

$$\sum_{l=1}^N u_l(t)\varphi_l(x^j) = u(t, x^j),$$

which implies that $u^N(t,x^j) = u(t,x^j)$ for all $j = 1, \dots, N$.

• For the *Fourier sine series* $\varphi_j(x) = \sin(jx)$, set the collocation points $x^j = \frac{\pi j}{N+1}$, $j = 1, \dots, N+1$. The corresponding collocation equation:

$$u(t, \tfrac{\pi j}{N+1}) = \sum_{l=1}^N u_l(t) \sin(\tfrac{\pi j l}{N+1})$$

is solved analytically:

$$u_j(t) = \frac{2}{N+1} \sum_{l=1}^N u(t, x^l) \sin(\tfrac{\pi j l}{N+1}), \quad j = 1, \dots, N,$$

due to the identities (Gottlieb and Orszag 1977, p.14):

$$\sum_{l=1}^N \sin(\tfrac{\pi k l}{N+1}) \sin(\tfrac{\pi j l}{N+1}) = \tfrac{N+1}{2}\delta_{kj} \quad \text{for } k, j = 1, \dots, N.$$

• For the *Chebyshev polynomials* $\varphi_j(x) = T_j(x)$, take the extreme points $x^j = \cos(\frac{\pi j}{N})$, $j = 0, \dots, N-1$. Similarly, the collocation equation:

$$u(t, \cos(\tfrac{\pi j}{N})) = \textstyle\sum_{l=0}^{N-1} u_l(t) T_l(\cos(\tfrac{\pi j}{N}))$$

admits analytical solution as follows (Mason and Handscomb 2003, (11.41), p.276):

$$u_j(t) = \frac{2}{N(1+\delta_{j0})} \sum_{j=0}^{N-1} u(t, x^l) T_j(x^l), \quad j = 0, \dots, N-1.$$

Bibliography

Ch. Bernardi and E. Süli, Time and space adaptivity for the second-order wave equation, *Math. Models Methods Appl. Sci.* **15** (2005), 199–225. http://eprints.maths.ox.ac.uk/1178/1/NA-04-12.pdf

S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods.* Springer, New York, 2008. https://link.springer.com/book/10.1007%2F978-0-387-75934-0

F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods.* Springer, New York, 1991. https://ru.scribd.com/document/134779593/brezzi-fortin-mixed-and-hybrid-finite-elements-methods-pdf

Ch. Clason, *Numerical Partial Differential Equations.* Lecture Notes, KF-University of Graz, 2013. https://www.uni-due.de/~adf040p/skripte/NumPDENotes12.pdf

P.G. Ciarlet and J.L. Lions (eds.), *Handbook of Numerical Analysis. Vol. II: Finite Element Methods (Part 1).* North Holland, Amsterdam, 1991. https://www.ices.utexas.edu/sites/oden/wp-content/uploads/2013/06/1989-010.finite_element.pdf

A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements.* Springer, New York, 2004.

D. Gottlieb and S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications.* SIAM, Philadelphia, 1977. http://epubs.siam.org/doi/book/10.1137/1.9781611970425

Ch. Grossmann, H.-G. Roos and M. Stynes, *Numerical Treatment of Partial Differential Equations.* Springer, Berlin, 2007. https://link.springer.com/book/10.1007%2F978-3-540-71584-9

W. Hackbusch, *Elliptic Differential Equations. Theory and Numerical Treatment.* Springer, Berlin, 1992. https://link.springer.com/book/10.1007%2F978-3-658-15358-8

J.K. Hunter, *Partial Differential Equations.* Lecture Notes, University of California, 2014. https://www.math.ucdavis.edu/~hunter/pdes/pde_notes.pdf

F. Kappel, *Lineare Algebra I & II.* Lecture Notes, KF-University of Graz, 2006. https://imsc.uni-graz.at/kappel/files/linalg.pdf

S. Keeling, *Numerics for Partial Differential Equations.* Lecture Notes, KF-University of Graz, 2016. http://imsc.uni-graz.at/keeling/numpde_ss16/numpde.pdf

A.M. Khludnev and V.A. Kovtunenko, *Analysis of Cracks in Solids.* WIT-Press, Southampton, 2000. https://static.uni-graz.at/fileadmin/_Persoenliche_Webseite/kovtunenko_victor/kk1.pdf

P. Knabner and L. Angerman, *Numerical Methods for Elliptic and Parabolic Partial Differential Equations.* Springer, New York, 2003. http://www.lib.ysu.am/disciplines_bk/4cd0696bb9f1e1a65c5da8d743583b35.pdf

V.A. Kovtunenko, *Proseminar aus Partielle Differentialgleichungen.* Script, KF-University of Graz, 2010a. https://static.uni-graz.at/fileadmin/_Persoenliche_Webseite/kovtunenko_victor/proseminarPDG.pdf

V.A. Kovtunenko, *Proseminar aus Mathematische Modellierung.* Script, KF-University of Graz, 2010b. https://static.uni-graz.at/fileadmin/_Persoenliche_Webseite/kovtunenko_victor/proseminarMM.pdf

K. Kuttler, *Partial Differential Equations.* Lecture Notes, Brigham Young University, 2003. https://math.byu.edu/~klkuttle/547notesB.pdf

S. Larsson and V. Thomee, *Partial Differential Equations with Numerical Methods.* Springer, Berlin, Heildelberg, 2008. https://de.1lib.at/book/2420916/f80069

H. Le Dret, *Numerical Approximation of PDEs.* Lecture Notes, UPMC Sorbonne University, 2012. https://www.ljll.math.upmc.fr/~ledret/M1ApproxPDE.html

J.C. Mason and D.C. Handscomb, *Chebyshev Polynomials.* Chapman & Hall/CRC, Boca Raton, FL, 2003.

R. Rannacher, *Numerische Mathematik 2 (Numerik partieller Differentialgleichungen)*. Vorlesungsskriptum, RK-Universität Heidelberg, 2008. `http://ganymed.math.uni-heidelberg.de/~lehre/notes/num2/numerik2.pdf`

T. Roubíček, *Nonlinear Partial Differential Equations with Applications*. Birkhäuser, Basel, Boston, Berlin, 2005. `https://link.springer.com/book/10.1007%2F978-3-0348-0513-1#authorsandaffiliationsbook`

J. Schöberl, *Space-Time Formulation of Parabolic Equations*. Lecture Notes, Vienna University of Technology, 2016. `http://www.asc.tuwien.ac.at/~schoeberl/wiki/lva/numpde15/parabolic.pdf`

O. Steinbach, *Numerical Approximation Methods for Elliptic Boundary Value Problems*. Springer, New York, 2008. `http://dl.iranidata.com/book/daneshgahi/Numerical_approximation_methods_for%20Elliptic.Boundary.Value%20Problems(www.IraniData.com).pdf`

L.L. Thomson and P.M. Pinsky, A Galerkin Least-Squares finite element method for the two dimensional Helmholtz equation, *Int. J. Num. Meth. Engrg.* **38** (1995), 371–397. `https://www.researchgate.net/publication/227612503`

Z. Wu, J. Yin and C. Wang, *Elliptic & Parabolic Equations*. World Scientific, Singapore, 2006.