Numerical Methods for Partial Differential Equations

Joachim Schöberl

May 12, 2022

Contents

Chapter 1

Introduction

Differential equations are equations for an unknown function involving differential operators. An *ordinary* differential equation (ODE) requires differentiation with respect to one variable. A *partial* differential equation (PDE) involves partial differentiation with respect to two or more variables.

1.1 Classification of PDEs

The general form of a linear PDE of second order is: find $u : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ such that

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

The coefficients $a_{i,j}(x), b_i(x), c(x)$ and the right hand side $f(x)$ are given functions. In addition, certain type of boundary conditions are required. The behavior of the PDE depends on the type of the differential operator

$$L := \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} a_{i,j} \frac{\partial}{\partial x_j} + \sum_{i=1}^{d} b_i \frac{\partial}{\partial x_i} + c.$$

Replace $\frac{\partial}{\partial x_i}$ by s_i. Then

$$\sum_{i,j=1}^{d} s_i a_{i,j} s_j + \sum_{i=1}^{d} b_i s_i + c = 0$$

describes a quartic shape in $\mathbb{R}^d$. We treat the following cases:

1. In the case of a (positive or negative) definite matrix $a = (a_{i,j})$ this is an ellipse, and the corresponding PDE is called elliptic. A simple example is $a = I$, $b = 0$, and $c = 0$, i..e.

$$-\sum_i \frac{\partial^2 u}{\partial x_i^2} = f.$$

Elliptic PDEs require boundary conditions.

2. If the matrix a is semi-definite, has the one-dimensional kernel span$\{v\}$, and $b \cdot v \neq 0$, then the shape is a parabola. Thus, the PDE is called parabolic. A simple example is
$$-\sum_{i=1}^{d-1} \frac{\partial^2 u}{\partial x_i^2} + \frac{\partial u}{\partial x_d} = f.$$
Often, the distinguished direction corresponds to time. This type of equation requires boundary conditiosn on the $d-1$-dimensional boundary, and initial conditions in the different direction.

3. If the matrix a has $d-1$ positive, and one negative (or vise versa) eigenvalues, then the shape is a hyperbola. The PDE is called hyperbolic. The simplest one is
$$-\sum_{i=1}^{d-1} \frac{\partial^2 u}{\partial x_i^2} + \frac{\partial^2 u}{\partial x_d^2} = f.$$
Again, the distinguished direction often corresponds to time. Now, two initial conditions are needed.

4. If the matrix a is zero, then the PDE degenerates to the first order PDE
$$b_i \frac{\partial u}{\partial x_i} + cu = f.$$
Boundary conditions are needed at a part of the boundary.

These cases behave very differently. We will establish theories for the individual cases. A more general classicfication, for more positive or negative eigenvalues, and systems of PDEs is possible. The type of the PDE may also change for different points x.

1.2 Weak formulation of the Poisson Equation

The most elementary and thus most popular PDE is the Poisson equation
$$-\Delta u = f \qquad \text{in } \Omega, \tag{1.2}$$
with the boundary conditions
$$\begin{array}{rcll} u &=& u_D & \text{on } \Gamma_D, \\ \frac{\partial u}{\partial n} &=& g & \text{on } \Gamma_N, \\ \frac{\partial u}{\partial n} + \alpha u &=& g & \text{on } \Gamma_R. \end{array} \tag{1.3}$$
The domain Ω is an open and bounded subset of $\mathbb{R}^d$, where the problem dimension d is usually 1, 2 or 3. For $d = 1$, the equation is not a PDE, but an ODE. The boundary

$\Gamma := \partial\Omega$ consists of the three non-overlapping parts Γ_D, Γ_N, and Γ_R. The outer unit normal vector is called n. The Laplace differential operator is $\Delta := \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$, the normal derivative at the boundary is $\frac{\partial}{\partial n} := \sum_{i=1}^d n_i \frac{\partial}{\partial x_i}$. Given are the functions f, u_D and g in proper function spaces (e.g., $f \in L_2(\Omega)$). We search for the unknown function u, again, in a proper function space defined later.

The boundary conditions are called

- Dirichlet boundary condition on Γ_D. The function value is prescribed,
- Neumann boundary condition on Γ_N. The normal derivative is prescribed,
- Robin boundary condition on Γ_R. An affine linear relation between the function value and the normal derivative is prescribed.

Exactly one boundary condition must be specified on each part of the boundary.

We transform equation (1.2) together with the boundary conditions (1.3) into its weak form. For this, we multiply (1.2) by smooth functions (called test functions) and integrate over the domain:

$$-\int_\Omega \Delta u v \, dx = \int_\Omega f v \, dx \tag{1.4}$$

We do so for sufficiently many test functions v in a proper function space. Next, we apply Gauss' theorem $\int_\Omega \operatorname{div} p \, dx = \int_\Gamma p \cdot n \, ds$ to the function $p := \nabla u \, v$ to obtain

$$\int_\Omega \operatorname{div}(\nabla u \, v) \, dx = \int_\Gamma \nabla u \cdot n \, v \, ds$$

From the product rule there follows $\operatorname{div}(\nabla u v) = \Delta u v + \nabla u \cdot \nabla v$. Together we obtain

$$\int_\Omega \nabla u \cdot \nabla v \, dx - \int_\Gamma \frac{\partial u}{\partial n} v \, ds = \int_\Omega f v \, dx.$$

Up to now, we only used the differential equation in the domain. Next, we incorporate the boundary conditions. The Neumann and Robin b.c. are very natural (and thus are called natural boundary conditions). We simply replace $\frac{\partial u}{\partial n}$ by g and $-\alpha u + g$ on Γ_N and Γ_R, respectively. Putting unknown terms to the left, and known terms to the right hand side, we obtain

$$\int_\Omega \nabla u \cdot \nabla v \, dx + \int_{\Gamma_R} \alpha u v \, ds - \int_{\Gamma_D} \frac{\partial u}{\partial n} v \, ds = \int f v \, dx + \int_{\Gamma_N + \Gamma_R} g v \, ds.$$

Finally, we use the information of the Dirichlet boundary condition. We work brute force and simple keep the Dirichlet condition in strong sense. At the same time, we only allow test functions v fulfilling $v = 0$ on Γ_D. We obtain the

Weak form of the Poisson equation:
Find u such that $u = u_D$ on Γ_D and

$$\int_\Omega \nabla u \cdot \nabla v \, dx + \int_{\Gamma_R} \alpha u v \, ds = \int_\Omega f v \, dx + \int_{\Gamma_N + \Gamma_R} g v \, ds \tag{1.5}$$

$$\forall\, v \text{ such that } v = 0 \text{ on } \Gamma_D.$$

We still did not define the function space in which we search for the solution u. A proper choice is

$$V := \{v \in L_2(\Omega) : \nabla u \in [L_2(\Omega)]^d \text{ and } u|_\Gamma \in L_2(\partial\Omega)\}.$$

It is a complete space, and, together with the inner product

$$(u, v)_V := (u, v)_{L_2(\Omega)} + (\nabla u, \nabla v)_{L_2(\Omega)} + (u, v)_{L_2(\Gamma)}$$

it is a Hilbert space. Now, we see that $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$ is useful. The Dirichlet b.c. u_D must be chosen such that there exists an $u \in V$ with $u = u_D$ on Γ_D. By definition of the space, all terms are well defined. We will see later, that the problem indeed has a unique solution in V.

1.3 The Finite Element Method

Now, we are developing a numerical method for approximating the weak form (1.5). For this, we decompose the domain Ω into triangles T. We call the set $\mathcal{T} = \{T\}$ triangulation. The set $\mathcal{N} = \{x_j\}$ is the set of nodes. By means of this triangulation, we define the finite element space, V_h:

$$V_h := \{v \in C(\Omega) : v|_T \text{ is affine linear } \ \forall\, T \in \mathcal{T}\}$$

This is a sub-space of V. The derivatives (in weak sense, see below) are piecewise constant, and thus, belong to $[L_2(\Omega)]^2$. The function $v_h \in V_h$ is uniquely defined by its values $v(x_j)$ in the nodes $x_j \in \mathcal{N}$. We decompose the set of nodes as

$$\mathcal{N} = \mathcal{N}_D \cup \mathcal{N}_f,$$

where $\mathcal{N}_\mathcal{D}$ are the nodes on the Dirichlet boundary, and $\mathcal{N}_f$ are all others (f as free). The finite element approximation is defined as

Find u_h such that $u_h(x) = u_D(x) \ \ \forall\, x \in \mathcal{N}_D$ and

$$\int_\Omega \nabla u_h \cdot \nabla v_h \, dx + \int_{\Gamma_R} \alpha u_h v_h \, ds = \int f v_h \, dx + \int_{\Gamma_N + \Gamma_R} g v_h \, ds \tag{1.6}$$

$$\forall\, v_h \in V_h \text{ such that } v_h(x) = 0 \ \ \forall\, x \in \mathcal{N}_D$$

Now it is time to choose a basis for V_h. The most convenient one is the nodal basis $\{\varphi_i\}$ characterized as

$$\varphi_i(x_j) = \delta_{i,j}. \tag{1.7}$$

The Kronecker-δ is defined to be 1 for $i = j$, and 0 else. These are the popular hat functions. We represent the finite element solution with respect to this basis:

$$u_h(x) = \sum u_i \varphi_i(x) \tag{1.8}$$

By the nodal-basis property (1.7) there holds $u_h(x_j) = \sum_i u_i \varphi_i(x_j) = u_j$. We have to determine the coefficients $u_i \in \mathbb{R}^N$, with $N = |\mathcal{N}|$. The $N_D := |\mathcal{N}_D|$ values according to nodes on Γ_D are given explicitely:

$$u_j = u_h(x_j) = u_D(x_j) \qquad \forall\, x_j \in \Gamma_D$$

The others have to be determined from the variational equation (1.6). It is equivalent to fulfill (1.6) for the whole space $\{v_h \in V_h : v_h(x_j) = 0 \;\forall\; x_j \in \mathcal{N}_D\}$, or just for its basis $\{\varphi_i : x_i \in \mathcal{N}_f\}$ associated to the free nodes:

$$\sum_i \Big\{ \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, dx + \int_{\Gamma_R} \alpha\varphi_i\varphi_j \, ds \Big\} u_i = \int f\varphi_j \, dx + \int_{\Gamma_N+\Gamma_R} g\varphi_j \, ds \tag{1.9}$$

$$\forall\, \varphi_j \;\text{ such that } x_j \in \mathcal{N}_f$$

We have inserted the basis expansion (1.8). We define the matrix $A = (A_{ji}) \in \mathbb{R}^{N\times N}$ and the vector $f = (f_j) \in \mathbb{R}^N$ as

$$\begin{aligned} A_{ji} &:= \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, dx + \int_{\Gamma_R} \alpha\varphi_i\varphi_j \, ds, \\ f_j &:= \int f\varphi_j \, dx + \int_{\Gamma_N+\Gamma_R} g\varphi_j \, ds. \end{aligned}$$

According to Dirichlet- and free nodes they are splitted as

$$A = \begin{pmatrix} A_{DD} & A_{Df} \\ A_{fD} & A_{ff} \end{pmatrix} \qquad \text{and} \qquad f = \begin{pmatrix} f_D \\ f_f \end{pmatrix}.$$

Now, we obtain the system of linear equations for $u = (u_i) \in \mathbb{R}^N$, $u = (u_D, u_f)$:

$$\begin{pmatrix} I & 0 \\ A_{fD} & A_{ff} \end{pmatrix} \begin{pmatrix} u_D \\ u_f \end{pmatrix} = \begin{pmatrix} u_D \\ f_f \end{pmatrix}. \tag{1.10}$$

At all, we have N coefficients u_i. N_D are given explicitely from the Dirichlet values. These are N_f equations to determine the remaining ones. Using the known u_D, we can reformulate it as symmetric system of equations for $u_f \in \mathbb{R}^{N_f}$:

$$A_{ff} u_f = f_f - A_{fD} u_D$$

Chapter 2

The abstract theory

In this chapter we develop the abstract framework for variational problems.

2.1 Basic properties

Definition 1. *A* **vector space** *V is a set with the operations $+ : V \times V \to V$ and $\cdot : \mathbb{R} \times V \to V$ such that for all $u, v \in V$ and $\lambda, \mu \in \mathbb{R}$ there holds*

- $u + v = v + u$
- $(u + v) + w = u + (v + w)$
- $\lambda \cdot (u + v) = \lambda \cdot u + \lambda \cdot v, \quad (\lambda + \mu) \cdot u = \lambda \cdot u + \mu \cdot u$

Examples are $\mathbb{R}^n$, the continuous functions C^0, or the Lebesgue space L_2.

Definition 2. *A* **normed** *vector space $(V, \|\cdot\|)$ is a vector space with the operation $\|.\| : V \to \mathbb{R}$ being a norm, i.e., for $u, v \in V$ and $\lambda \in \mathbb{R}$ there holds*

- $\|u + v\| \leq \|u\| + \|v\|$
- $\|\lambda\, u\| = |\lambda|\, \|u\|$
- $\|u\| = 0 \Leftrightarrow u = 0$

Examples are $(C^0, \|\cdot\|_{\sup})$, or $(C^0, \|\cdot\|_{L_2})$.

Definition 3. *In a* **complete** *normed vector space, Cauchy sequences $(u_n) \in V^{\mathbb{N}}$ converge to an $u \in V$. A complete normed vector space is called* **Banach space**.

Examples of Banach spaces are $(L_2, \|\cdot\|_{L_2})$, $(C^0, \|\cdot\|_{\sup})$, but not $(C^0, \|\cdot\|_{L_2})$.

Definition 4. *The* **closure** *of a normed vector-space $(W, \|\cdot\|_V)$, denoted as $\overline{W}^{\|\cdot\|_V}$ is the smallest complete space containing W.*

Example: $\overline{C}^{\|\cdot\|_{L_2}} = L_2$.

Definition 5. *A* **functional** *or a* **linear form** $l(\cdot)$ *on* V *is a linear mapping* $l(\cdot) : V \to \mathbb{R}$. *The canonical norm for linear forms is the* **dual norm**

$$\|l\|_{V^*} := \sup_{0 \neq v \in V} \frac{l(v)}{\|v\|}.$$

A linear form l *is called bounded if the norm is finite. The vector space of all bounded linear forms on* V *is called the dual space* V^*.

An example for a bounded linear form is $l(\cdot) : L_2 \to \mathbb{R} : v \to \int v \, dx$.

Definition 6. *A* **bilinear form** $A(\cdot, \cdot)$ *on* V *is a mapping* $A : V \times V \to \mathbb{R}$ *which is linear in* u *and in* v. *It is called* **symmetric** *if* $A(u, v) = A(v, u)$ *for all* $u, v \in V$.

Examples are the bilinear form $A(u, v) = \int uv \, dx$ on L_2, or $A(u, v) := u^T A v$ on $\mathbb{R}^n$, where A is a (symmetric) matrix.

Definition 7. *A symmetric bilinear form* $A(\cdot, \cdot)$ *is called an* **inner product** *if it satisfies*

- $A(v, v) \geq 0 \; \forall \, v \in V$
- $A(v, v) = 0 \Leftrightarrow v = 0$

Often, is is denoted as $(\cdot, \cdot)_A$, $(\cdot, \cdot)_V$, *or simply* $(\cdot, \cdot)$.

An example on $\mathbb{R}^n$ is $u^T A v$, where A is a symmetric and positive definite matrix.

Definition 8. *An* **inner product space** *is a vector space* V *together with an inner product* $(\cdot, \cdot)_V$.

Lemma 9. Cauchy Schwarz inequality. *If* $A(\cdot, \cdot)$ *is a symmetric bilinear form such that* $A(v, v) \geq 0$ *for all* $v \in V$, *then there holds*

$$A(u, v) \leq A(u, u)^{1/2} A(v, v)^{1/2}$$

Proof: For $t \in \mathbb{R}$ there holds

$$0 \leq A(u - tv, u - tv) = A(u, u) - 2tA(u, v) + t^2 A(v, v)$$

If $A(v, v) = 0$, then $A(u, u) - 2tA(u, v) \geq 0$ for all $t \in \mathbb{R}$, which forces $A(u, v) = 0$, and the inequality holds trivially. Else, if $A(v, v) \neq 0$, set $t = A(u, v)/A(v, v)$, and obtain

$$0 \leq A(u, u) - A(u, v)^2 / A(v, v),$$

which is equivalent to the statement. □

Lemma 10. $\|v\|_V := (v, v)_V^{1/2}$ *defines a norm on the inner product space* $(V, (\cdot, \cdot)_V)$.

Definition 11. *An inner product space $(V, (\cdot,\cdot)_V)$ which is complete with respect to $\|\cdot\|_V$ is called a* **Hilbert space**.

Definition 12. *A* **closed subspace** *S of an Hilbert space V is a subset which is a vector space, and which is complete with respect to $\|\cdot\|_V$.*

A finite dimensional subspace is always a closed subspace.

Lemma 13. *Let T be a continuous linear operator from the Hilbert space V to the Hilbert space W. The kernel of T, $\ker T := \{v \in V : Tv = 0\}$ is a closed subspace of V.*

Proof: First we observe that $\ker T$ is a vector space. Now, let $(u_n) \in \ker T^{\mathbb{N}}$ converge to $u \in V$. Since T is continuous, $Tu_n \to Tu$, and thus $Tu = 0$ and $u \in \ker T$. $\square$

Lemma 14. *Let S be a subspace (not necessarily closed) of V. Then*

$$S^{\perp} := \{v \in V : (v, w) = 0 \; \forall \, w \in S\}$$

is a closed subspace.

The proof is similar to Lemma 13.

Definition 15. *Let V and W be vector spaces. A* **linear operator** *$T : V \to W$ is a linear mapping from V to W. The operator is called* **bounded** *if its operator-norm*

$$\|T\|_{V\to W} := \sup_{0\neq v\in V} \frac{\|Tv\|_W}{\|v\|_V}$$

is finite.

An example is the differential operator on the according space $\frac{d}{dx} : (C^1(0,1), \|\cdot\|_{sup} + \|\frac{d}{dx}\cdot\|_{sup}) \to (C(0,1), \|\cdot\|_{sup})$.

Lemma 16. *A bounded linear operator is continuous.*

Proof. Let $v_n \to v$, i.e. $\|v_n - v\|_V \to 0$. Then $\|Tv_n - Tv\| \leq \|T\|_{V\to W}\|v_n - v\|_V$ converges to 0, i.e. $Tv_n \to Tv$. Thus T is continuous. $\square$

Definition 17. *A* **dense subspace** *S of V is such that every element of V can be approximated by elements of S, i.e.*

$$\forall \varepsilon > 0 \, \forall u \in V \, \exists v \in S \text{ such that } \|u - v\|_V \leq \varepsilon.$$

Lemma 18 (extension principle)**.** *Let S be a dense subspace of the normed space V, and let W be a complete space. Let $T : S \to W$ be a bounded linear operator with respect to the norm $\|T\|_{V\to W}$. Then, the operator can be uniquely extended onto V.*

Proof. Let $u \in V$, and let v_n be a sequence such that $v_n \to u$. Thus, v_n is Cauchy. Tv_n is a well defined sequence in W. Since T is continuous, Tv_n is also Cauchy. Since W is complete, there exists a limit w such that $Tv_n \to w$. The limit is independent of the sequence, and thus Tu can be defined as the limit w. □

Definition 19. *A bounded linear operator $T : V \to W$ is called* **compact** *if for every bounded sequence $(u_n) \in V^{\mathbb{N}}$, the sequence (Tu_n) contains a convergent sub-sequence.*

Lemma 20. *Let V, W be Hilbert spaces. An operator is compact if and only if there exists a complete orthogonal system (u_n) for $(\ker T)^\perp$ and values $\lambda_n \to 0$ such that*

$$(u_n, u_m)_V = \delta_{n,m} \qquad (Tu_n, Tu_m)_W = \lambda_n \delta_{n,m}$$

This is the eigensystem of the operator $K : V \to V^* : u \mapsto (Tu, T\cdot)_W$.

Proof. (sketch) There exists an maximizing element of $\frac{(Tv,Tv)_W}{(v,v)_V}$. Scale it to $\|v\|_V = 1$ and call it u_1, and $\lambda_1 = \frac{(Tu_1,Tu_1)_W}{(u_1,u_1)_V}$. Repeat the procedure on the V-complement of u_1 to generate u_2, and so on. □

2.2 Projection onto subspaces

In the Euklidean space $\mathbb{R}^2$ one can project orthogonally onto a line through the origin, i.e., onto a sub-space. The same geometric operation can be defined for closed subspaces of Hilbert spaces.

Theorem 21. *Let S be a closed subspace of the Hilbert space V. Let $u \in V$. Then there exists a unique closest point $u_0 \in S$:*

$$\|u - u_0\| \leq \|u - v\| \qquad \forall\, v \in S$$

There holds

$$u - u_0 \perp S$$

Proof: Let $d := \inf_{v \in S} \|u - v\|$, and let (v_n) be a minimizing sequence such that $\|u - v_n\| \to d$. We first check that there holds

$$\|v_n - v_m\|^2 = 2\,\|v_n - u\|^2 + 2\,\|v_m - u\|^2 - 4\,\|1/2(v_n + v_m) - u\|^2.$$

Since $1/2(v_n + v_m) \in S$, there holds $\|1/2(v_n + v_m) - u\| \geq d$. We proof that (v_n) is a Cauchy sequence: Fix $\varepsilon > 0$, choose $N \in \mathbb{N}$ such that for $n > N$ there holds $\|u - v_n\|^2 \leq d^2 + \varepsilon^2$. Thus for all $n, m > N$ there holds

$$\|v_n - v_m\|^2 \leq 2(d^2 + \varepsilon^2) + 2(d^2 + \varepsilon^2) - 4d^2 = 4\varepsilon^2.$$

Thus, v_n converge to some $u_0 \in V$. Since S is closed, $u_0 \in S$. By continuity of the norm, $\|u - u_0\| = d$.

Fix some $0 \neq w \in S$, and define $\varphi(t) := \|u - \underbrace{u_0 - tw}_{\in S}\|^2$. $\varphi(\cdot)$ is a convex function, it takes its unique minimum d at $t = 0$. Thus

$$0 = \frac{d\varphi(t)}{dt}|_{t=0} = \{2(u - u_0, w) - 2t(w, w)\}|_{t=0} = 2(u - u_0, w)$$

We obtained $u - u_0 \perp S$. If there were two minimizers $u_0 \neq u_1$, then $u_0 - u_1 = (u_0 - u) - (u_1 - u) \perp S$ and $u_0 - u_1 \in S$, which implies $u_0 - u_1 = 0$, a contradiction. $\square$

Theorem 21 says that given an $u \in V$, we can uniquely decompose it as

$$u = u_0 + u_1, \qquad u_0 \in S \quad u_1 \in S^\perp$$

This allows to define the operators $P_S : V \to S$ and $P_S^\perp : V \to S^\perp$ as

$$P_S u := u_0 \qquad P_S^\perp u := (I - P_S)u = u_1$$

Theorem 22. *P_S and $P_S^\perp$ are linear operators.*

Definition 23. *A linear operator P is called a* **projection** *if $P^2 = P$. A projector is called* **orthogonal**, *if $(Pu, v) = (u, Pv)$.*

Lemma 24. *The operators P_S and $P_S^\perp$ are both orthogonal projectors.*

Proof: For $u \in S$ there holds $P_S u = u$. Since $P_S u \in S$, there holds $P_S^2 u = P_S u$. It is orthogonal since

$$(P_S u, v) = (P_S u, v - P_S v + P_S v) = (\underbrace{P_S u}_{\in S}, \underbrace{v - P_S v}_{\in S^\perp}) + (P_S u, P_S v) = (P_S u, P_S v).$$

With the same argument there holds $(u, P_S v) = (P_S u, P_S v)$. The co-projector $P_S^\perp = I - P_S$ is a projector since

$$(I - P_S)^2 = I - 2P_S + P_S^2 = I - P_S.$$

It is orthogonal since $((I - P_S)u, v) = (u, v) - (P_S u, v) = (u, v) - (u, P_S v) = (u, (I - P_S)v)$
$\square$

2.3 Riesz Representation Theorem

Let $u \in V$. Then, we can define the related continuous linear functional $l_u(\cdot) \in V^*$ by

$$l_u(v) := (u, v)_V \qquad \forall\, v \in V.$$

The opposite is also true:

Theorem 25. Riesz Representation Theorem. *Any continuous linear functional l on a Hilbert space V can be represented uniquely as*

$$l(v) = (u_l, v) \tag{2.1}$$

for some $u_l \in V$. Furthermore, we have

$$\|l\|_{V^*} = \|u_l\|_V.$$

Proof: First, we show uniqueness. Assume that $u_1 \neq u_2$ both fulfill (2.1). This leads to the contradiction

$$\begin{aligned} 0 &= l(u_1 - u_2) - l(u_1 - u_2) \\ &= (u_1, u_1 - u2) - (u_2, u_1 - u_2) = \|u_1 - u_2\|^2. \end{aligned}$$

Next, we construct the u_l. For this, define $S := \ker l$. This is a closed subspace.
Case 1: $S^\perp = \{0\}$. Then, $S = V$, i.e., $l = 0$. So take $u_l = 0$.
Case 2: $S^\perp \neq \{0\}$. Pick some $0 \neq z \in S^\perp$. There holds $l(z) \neq 0$ (otherwise, $z \in S \cap S^\perp = \{0\}$). Now define

$$u_l := \frac{l(z)}{\|z\|^2} z \qquad \in S^\perp$$

Then

$$\begin{aligned} (u_l, v) &= (\underbrace{u_l}_{S^\perp}, \underbrace{v - l(v)/l(z)z}_{S}) + (u_l, l(v)/l(z)z) \\ &= l(z)/\|z\|^2 (z, l(v)/l(z)z) \\ &= l(v) \end{aligned}$$

Finally, we prove $\|l\|_{V^*} = \|u_l\|_V$:

$$\|l\|_{V^*} = \sup_{0 \neq v \in V} \frac{l(v)}{\|v\|} = \sup_v \frac{(u_l, v)_V}{\|v\|_V} \leq \|u_l\|_V$$

and

$$\|u_l\| = \frac{l(z)}{\|z\|^2} \|z\| = \frac{l(z)}{\|z\|} \leq \|l\|_{V^*}.$$

2.4 Symmetric variational problems

Take the function space $C^1(\Omega)$, and define the bilinear form

$$A(u, v) := \int_\Omega \nabla u \nabla v + \int_\Gamma uv \, ds$$

and the linear form

$$f(v) := \int_\Omega f v \, dx$$

The bilinear form is non-negative, and $A(u,u) = 0$ implies $u = 0$. Thus $A(\cdot,\cdot)$ is an inner product, and provides the norm $\|v\|_A := A(v,v)^{1/2}$. The normed vector space $(C^1, \|.\|_A)$ is not complete. Define

$$V := \overline{C^1(\Omega)}^{\|.\|_A},$$

which is a Hilbert space per definition. If we can show that there exists a constant c such that

$$f(v) = \int_\Omega f v \, dx \leq c\|v\|_A \qquad \forall \, v \in V$$

then $f(.)$ is a continuous linear functional on V. We will prove this later. In this case, the Riesz representation theorem tells that there exists an unique $u \in V$ such that

$$A(u,v) = f(v).$$

This shows that the weak form has a unique solution in V.

Next, take the finite dimensional ($\Rightarrow$ closed) finite element subspace $V_h \subset V$. The finite element solution $u_h \in V_h$ was defined by

$$A(u_h, v_h) = f(v_h) \qquad \forall \, v_h \in V_h,$$

This means

$$A(u - u_h, v_h) = A(u, v_h) - A(u_h, v_h) = f(v_h) - f(v_h) = 0$$

u_h is the projection of u onto V_h, i.e.,

$$\|u - u_h\|_A \leq \|u - v_h\|_A \qquad \forall \, v_h \in V_h$$

The error $u - u_h$ is orthogonal to V_h.

2.5 Coercive variational problems

In this chapter we discuss variational problems posed in Hilbert spaces. Let V be a Hilbert space, and let $A(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bilinear form which is

- coercive (also known as elliptic)

$$A(u,u) \geq \alpha_1 \|u\|_V^2 \qquad \forall \, u \in V, \tag{2.2}$$

- and continuous

$$A(u,v) \leq \alpha_2 \|u\|_V \, \|v\|_V \qquad \forall \, u, v \in V, \tag{2.3}$$

with bounds α_1 and α_2 in $\mathbb{R}^+$. It is not necessarily symmetric. Let $f(.) : V \to \mathbb{R}$ be a continuous linear form on V, i.e.,

$$f(v) \leq \|f\|_{V^*}\|v\|_V.$$

We are posing the variational problem: find $u \in V$ such that

$$A(u,v) = f(v) \qquad \forall\, v \in V.$$

Example 26. *Diffusion-reaction equation:*

Consider the PDE

$$-\operatorname{div}(a(x)\nabla u) + c(x)u = f \qquad \text{in } \Omega,$$

with Neumann boundary conditions. Let V be the Hilbert space generated by the inner product $(u,v)_V := (u,v)_{L_2} + (\nabla u, \nabla v)_{L_2}$. The variational formulation of the PDE involves the bilinear form

$$A(u,v) = \int_\Omega (a(x)\nabla u)\cdot \nabla v\, dx + \int_\Omega c(x)uv\, dx.$$

Assume that the coefficients $a(x)$ and $c(x)$ fulfill $a(x) \in \mathbb{R}^{d\times d}$, $a(x)$ symmetric and $\lambda_1 \leq \lambda_{\min}(a(x)) \leq \lambda_{\max}(a(x)) \leq \lambda_2$, and $c(x)$ such that $\gamma_1 \leq c(x) \leq \gamma_2$ almost everywhere. Then $A(\cdot,\cdot)$ is coercive with constant $\alpha_1 = \min\{\lambda_1, \gamma_1\}$ and $\alpha_2 = \max\{\lambda_2, \gamma_2\}$.

Example 27. *Diffusion-convection-reaction equation:*

The partial differential equation

$$-\Delta u + b\cdot\nabla u + u = f \qquad \text{in } \Omega$$

with Dirichlet boundary conditions $u = 0$ on $\partial\Omega$ leads to the bilinear form

$$A(u,v) = \int \nabla u\, \nabla v\, dx + \int b\cdot\nabla u\, v\, dx + \int uv\, dx.$$

If $\operatorname{div} b \leq 0$, what is an important case arising from incompressible flow fields ($\operatorname{div}\, b = 0$), then $A(\cdot,\cdot)$ is coercive and continuous w.r.t. the same norm as above.

Instead of the linear form $f(\cdot)$, we will often write $f \in V^*$. The evaluation is written as the duality product

$$\langle f, v\rangle_{V^*\times V} = f(v).$$

Lemma 28. *A continuous bilinear form $A(\cdot,\cdot) : V\times V \to \mathbb{R}$ induces a continuous linear operator $A : V \to V^*$ via*

$$\langle Au, v\rangle = A(u,v) \qquad \forall\, u,v \in V.$$

The operator norm $\|A\|_{V\to V^*}$ is bounded by the continuity bound α_2 of $A(\cdot,\cdot)$.

Proof: For every $u \in V$, $A(u, \cdot)$ is a bounded linear form on V with norm

$$\|A(u,\cdot)\|_{V^*} = \sup_{v \in V} \frac{A(u,v)}{\|v\|_V} \leq \sup_{v \in V} \frac{\alpha_2 \|u\|_V \, \|v\|_V}{\|v\|_V} = \alpha_2 \|u\|_V$$

Thus, we can define the operator $A : u \in V \to A(u, \cdot) \in V^*$. It is linear, and its operator norm is bounded by

$$\begin{aligned} \|A\|_{V \to V^*} &= \sup_{u \in V} \frac{\|Au\|_{V^*}}{\|u\|_V} = \sup_{u \in V} \sup_{v \in V} \frac{\langle Au, v \rangle_{V^* \times V}}{\|u\|_V \, \|v\|_V} \\ &= \sup_{u \in V} \sup_{v \in V} \frac{A(u,v)}{\|u\|_V \, \|v\|_V} \leq \sup_{u \in V} \sup_{v \in V} \frac{\alpha_2 \|u\|_V \|v\|_V}{\|u\|_V \, \|v\|_V} = \alpha_2. \end{aligned}$$

□

Using this notation, we can write the variational problem as operator equation: find $u \in V$ such that

$$Au = f \qquad (\text{in } V^*).$$

Theorem 29 (Banach's contraction mapping theorem)**.** *Given a Banach space V and a mapping $T : V \to V$, satisfying the Lipschitz condition*

$$\|T(v_1) - T(v_2)\| \leq L \, \|v_1 - v_2\| \qquad \forall \, v_1, v_2 \in V$$

for a fixed $L \in [0, 1)$. Then there exists a unique $u \in V$ such that

$$u = T(u),$$

i.e. the mapping T has a unique fixed point u. The iteration $u^1 \in V$ given, compute

$$u^{k+1} := T(u^k)$$

converges to u with convergence rate L:

$$\|u - u^{k+1}\| \leq L \|u - u^k\|$$

Theorem 30 (Lax Milgram)**.** *Given a Hilbert space V, a coercive and continuous bilinear form $A(\cdot, \cdot)$, and a continuous linear form $f(.)$. Then there exists a unique $u \in V$ solving*

$$A(u, v) = f(v) \qquad \forall \, v \in V.$$

There holds

$$\|u\|_V \leq \alpha_1^{-1} \|f\|_{V^*} \tag{2.4}$$

Proof: Start from the operator equation $Au = f$. Let $J_V : V^* \to V$ be the Riesz isomorphism defined by

$$(J_V g, v)_V = g(v) \qquad \forall \, v \in V, \; \forall \, g \in V^*.$$

Then the operator equation is equivalent to

$$J_V A u = J_V f \qquad (\text{in } V),$$

and to the fixed point equation (with some $0 \neq \tau \in \mathbb{R}$ chosen below)

$$u = u - \tau J_V(Au - f). \tag{2.5}$$

We will verify that

$$T(v) := v - \tau J_V(Av - f)$$

is a contraction mapping, i.e., $\|T(v_1) - T(v_2)\|_V \leq L\|v_1 - v_2\|_V$ with some Lipschitz constant $L \in [0, 1)$. Let $v_1, v_2 \in V$, and set $v = v_1 - v_2$. Then

$$\begin{aligned}
\|T(v_1) - T(v_2)\|_V^2 &= \|\{v_1 - \tau J_V(Av_1 - f)\} - \{v_2 - \tau J_V(Av_2 - f)\}\|_V^2 \\
&= \|v - \tau J_V Av\|_V^2 \\
&= \|v\|_V^2 - 2\tau(J_V Av, v)_V + \tau^2 \|J_V Av\|_V^2 \\
&= \|v\|_V^2 - 2\tau \langle Av, v\rangle + \tau^2 \|Av\|_{V^*}^2 \\
&= \|v\|_V^2 - 2\tau A(v, v) + \tau^2 \|Av\|_{V^*}^2 \\
&\leq \|v\|_V^2 - 2\tau\alpha_1 \|v\|_V^2 + \tau^2 \alpha_2^2 \|v\|_V^2 \\
&= (1 - 2\tau\alpha_1 + \tau^2\alpha_2^2)\|v_1 - v_2\|_V^2
\end{aligned}$$

Now, we choose $\tau = \alpha_1/\alpha_2^2$, and obtain a Lipschitz constant

$$L^2 = 1 - \alpha_1^2/\alpha_2^2 \in [0, 1).$$

Banach's contraction mapping theorem state that (2.5) has a unique fixed point. Finally, we obtain the bound (2.4) from

$$\|u\|_V^2 \leq \alpha_1^{-1} A(u, u) = \alpha_1^{-1} f(u) \leq \alpha_1^{-1} \|f\|_{V^*} \|u\|_V,$$

and dividing by one factor $\|u\|$. □

2.5.1 Approximation of coercive variational problems

Now, let V_h be a closed subspace of V. We compute the approximation $u_h \in V_h$ by the Galerkin method

$$A(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in V_h. \tag{2.6}$$

This variational problem is uniquely solvable by Lax-Milgram, since, $(V_h, \|.\|_V)$ is a Hilbert space, and continuity and coercivity on V_h are inherited from the original problem on V.

The next theorem says, that the solution defined by the Galerkin method is, up to a constant factor, as good as the best possible approximation in the finite dimensional space.

Theorem 31 (Cea)**.** *The approximation error of the Galerkin method is quasi optimal*

$$\|u - u_h\|_V \leq \alpha_2/\alpha_1 \inf_{v \in V_h} \|u - v_h\|_V$$

Proof: A fundamental property is the Galerkin orthogonality

$$A(u-u_h,w_h) = A(u,w_h) - A(u_h,w_h) = f(w_h) - f(w_h) = 0 \qquad \forall\, w_h \in V_h.$$

Now, pick an arbitrary $v_h \in V_h$, and bound

$$\begin{aligned}\|u-u_h\|_V^2 &\le \alpha_1^{-1} A(u-u_h,u-u_h)\\ &= \alpha_1^{-1}A(u-u_h,u-v_h) + \alpha_1^{-1}A(u-u_h,\underbrace{v_h-u_h}_{\in V_h})\\ &\le \alpha_2/\alpha_1\, \|u-u_h\|_V\|u-v_h\|_V.\end{aligned}$$

Divide one factor $\|u-u_h\|$. Since $v_h \in V_h$ was arbitrary, the estimation holds true also for the infimum in V_h. □

If $A(\cdot,\cdot)$ is additionally symmetric, then it is an inner product. In this case, the coercivity and continuity properties are equivalent to

$$\alpha_1\|u\|_V^2 \le A(u,u) \le \alpha_2\,\|u\|_V^2 \qquad \forall\, u \in V.$$

The generated norm $\|.\|_A$ is an equivalent norm to $\|.\|_V$. In the symmetric case, we can use the orthogonal projection with respect to $(.,.)_A$ to improve the bounds to

$$\|u-u_h\|_V^2 \le \alpha_1^{-1}\|u-u_h\|_A^2 \le \alpha_1^{-1} \inf_{v_h\in V_h} \|u-v_h\|_A^2 \le \alpha_2/\alpha_1\|u-v_h\|_V^2.$$

The factor in the quasi-optimality estimate is now the square root of the general, non-symmetric case.

2.6 Inf-sup stable variational problems

The coercivity condition is by no means a necessary condition for a stable solvable system. A simple, stable problem with non-coercive bilinear form is to choose $V = \mathbb{R}^2$, and the bilinear form $B(u,v) = u_1v_1 - u_2v_2$. The solution of $B(u,v) = f^Tv$ is $u_1 = f_1$ and $u_2 = -f_2$. We will follow the convention to call coercive bilinear forms $A(\cdot,\cdot)$, and the more general ones $B(\cdot,\cdot)$.

Let V and W be Hilbert spaces, and $B(\cdot,\cdot) : V \times W \to \mathbb{R}$ be a continuous bilinear form with bound

$$B(u,v) \le \beta_2\|u\|_V\|v\|_W \qquad \forall\, u\in V,\ \forall\, v\in W. \tag{2.7}$$

The general condition is the **inf-sup condition**

$$\inf_{\substack{u\in V\\ u\neq 0}} \sup_{\substack{v\in W\\ v\neq 0}} \frac{B(u,v)}{\|u\|_V\,\|v\|_W} \ge \beta_1. \tag{2.8}$$

Define the linear operator $B : V \to W^*$ by $\langle Bu, v\rangle_{W^*\times W} = B(u,v)$. The inf-sup condition can be reformulated as

$$\sup_{v\in W} \frac{\langle Bu, v\rangle}{\|v\|_W} \geq \beta_1 \|u\|_V, \qquad \forall\, u \in V$$

and, using the definition of the dual norm,

$$\|Bu\|_{W^*} \geq \beta_1 \|u\|_V. \tag{2.9}$$

We immediately obtain that B is one to one, since

$$Bu = 0 \Rightarrow u = 0$$

Lemma 32. *Assume that the continuous bilinear form $B(\cdot,\cdot)$ fulfills the inf-sup condition (2.8). Then the according operator B has closed range.*

Proof: Let Bu^n be a Cauchy sequence in W^*. From (2.9) we conclude that also u^n is Cauchy in V. Since V is complete, u_n converges to some $u \in V$. By continuity of B, the sequence Bu^n converges to $Bu \in W^*$. □

The inf-sup condition (2.8) does not imply that B is onto W^*. To insure that, we can pose an inf-sup condition the other way around:

$$\inf_{\substack{v\in W\\ v\neq 0}} \sup_{\substack{u\in V\\ u\neq 0}} \frac{B(u,v)}{\|u\|_V\,\|v\|_W} \geq \beta_1. \tag{2.10}$$

It will be sufficient to state the weaker condition

$$\sup_{\substack{u\in V\\ u\neq 0}} \frac{B(u,v)}{\|u\|_V\,\|v\|_W} > 0 \qquad \forall\, v \in W. \tag{2.11}$$

Theorem 33. *Assume that the continuous bilinear form $B(\cdot,\cdot)$ fulfills the inf-sup condition (2.8) and condition (2.11). Then, the variational problem: find $u \in V$ such that*

$$B(u,v) = f(v) \qquad \forall\, v \in W \tag{2.12}$$

has a unique solution. The solution depends continuously on the right hand side:

$$\|u\|_V \leq \beta_1^{-1} \|f\|_{W^*}$$

Proof: We have to show that the range $R(B) = W^*$. The Hilbert space W^* can be split into the orthogonal, closed subspaces

$$W^* = R(B) \oplus R(B)^\perp.$$

Assume that there exists some $0 \neq g \in R(B)^\perp$. This means that

$$(Bu, g)_{W^*} = 0 \qquad \forall\, u \in V.$$

Let $v_g \in W$ be the Riesz representation of g, i.e., $(v_g, w)_W = g(w)$ for all $w \in W$. This v_g is in contradiction to the assumption (2.11)

$$\sup_{u \in V} \frac{B(u, v_g)}{\|u\|_V} = \sup_{u \in V} \frac{(Bu, g)_{W^*}}{\|u\|_V} = 0.$$

Thus, $R(B)^\perp = \{0\}$ and $R(B) = W^*$. □

Example 34. *A coercive bilinear form is inf-sup stable.*

Example 35. *A complex symmetric variational problem:*

Consider the complex valued PDE

$$-\Delta u + iu = f,$$

with Dirichlet boundary conditions, $f \in L_2$, and $i = \sqrt{-1}$. The weak form for the real system $u = (u_r, u_i) \in V^2$ is

$$\begin{array}{rcll} (\nabla u_r, \nabla v_r)_{L_2} + (u_i, v_r)_{L_2} & = & (f, v_r) & \forall\, v_r \in V \\ (u_r, v_i)_{L_2} - (\nabla u_i, \nabla v_i)_{L_2} & = & -(f, v_i) & \forall\, v_i \in V \end{array} \tag{2.13}$$

We can add up both lines, and define the large bilinear form $B(\cdot, \cdot) : V^2 \times V^2 \to \mathbb{R}$ by

$$B((u_r, u_i), (v_r, v_i)) = (\nabla u_r, \nabla v_r) + (u_i, v_r) + (u_r, v_i) - (\nabla u_i, \nabla v_i)$$

With respect to the norm $\|v\|_V = (\|v\|_{L_2}^2 + \|\nabla v\|_{L_2}^2)^{1/2}$, the bilinear form is continuous, and fulfills the inf-sup conditions (exercises !) Thus, the variational formulation: find $u \in V^2$ such that

$$B(u, v) = (f, v_r) - (f, v_i) \qquad \forall\, v \in V^2$$

is stable solvable.

2.6.1 Approximation of inf-sup stable variational problems

Again, to approximate (2.12), we pick finite dimensional subspaces $V_h \subset V$ and $W_h \subset W$, and pose the finite dimensional variational problem: find $u_h \in V_h$ such that

$$B(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in W_h.$$

But now, in contrast to the coercive case, the solvability of the finite dimensional equation does not follow from the solvability conditions of the original problem on $V \times W$. E.g., take the example in $\mathbb{R}^2$ above, and choose the subspaces $V_h = W_h = \text{span}\{(1, 1)\}$.

We have to pose an extra inf-sup condition for the discrete problem:

$$\inf_{\substack{u_h \in V_h \\ u_h \neq 0}} \sup_{\substack{v_h \in W_h \\ v_h \neq 0}} \frac{B(u_h, v_h)}{\|u_h\|_V \, \|v_h\|_W} \geq \beta_{1h}. \tag{2.14}$$

On a finite dimensional space, one to one is equivalent to onto, and we can skip the second condition.

Theorem 36. *Assume that $B(\cdot,\cdot)$ is continuous with bound β_2, and $B(\cdot,\cdot)$ fulfills the discrete inf-sup condition with bound β_{1h}. Then there holds the quasi-optimal error estimate*

$$\|u - u_h\| \leq (1 + \beta_2/\beta_{1h}) \inf_{v_h \in V_h} \|u - v_h\| \tag{2.15}$$

Proof: Again, there holds the Galerkin orthogonality $B(u, w_h) = B(u_h, w_h)$ for all $w_h \in V_h$. Again, choose an arbitrary $v_h \in V_h$:

$$\begin{aligned}
\|u - u_h\|_V &\leq \|u - v_h\|_V + \|v_h - u_h\|_V \\
&\leq \|u - v_h\|_V + \beta_{1h}^{-1} \sup_{w_h \in W_h} \frac{B(v_h - u_h, w_h)}{\|w_h\|_V} \\
&= \|u - v_h\|_V + \beta_{1h}^{-1} \sup_{w_h \in W_h} \frac{B(v_h - u, w_h)}{\|w_h\|_V} \\
&\leq \|u - v_h\|_V + \beta_{1h}^{-1} \sup_{w_h \in W_h} \frac{\beta_2 \|v_h - u\|_V \, \|w_h\|_W}{\|w_h\|_W} \\
&= (1 + \beta_2/\beta_{1h}) \|u - v_h\|_V.
\end{aligned}$$

Chapter 3

Sobolev Spaces

In this section, we introduce the concept of generalized derivatives, we define families of normed function spaces, and prove inequalities between them. Let Ω be an open subset of $\mathbb{R}^d$, either bounded or unbounded.

3.1 Generalized derivatives

Let $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ be a multi-index, $|\alpha| = \sum \alpha_i$, and define the classical differential operator for functions in $C^\infty(\Omega)$

$$D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_d}.$$

For a function $u \in C(\Omega)$, the support is defined as

$$\operatorname{supp}\{u\} := \overline{\{x \in \Omega : u(x) \neq 0\}}.$$

This is a compact set if and only if it is bounded. We say u has compact support in Ω, if $\operatorname{supp} u \subset \Omega$. If Ω is a bounded domain, then u has compact support in Ω if and only if u vanishes in a neighbourhood of $\partial\Omega$.

The space of smooth functions with compact support is denoted as

$$\mathcal{D}(\Omega) := C_0^\infty(\Omega) := \{u \in C^\infty(\Omega) : \ u \text{ has compact support in } \Omega\}. \tag{3.1}$$

For a smooth function $u \in C^{|\alpha|}(\Omega)$, there holds the formula of integration by parts

$$\int_\Omega D^\alpha u \varphi \, dx = (-1)^{|\alpha|} \int_\Omega u D^\alpha \varphi \, dx \qquad \forall\, \varphi \in \mathcal{D}(\Omega). \tag{3.2}$$

The L_2 inner product with a function u in $C(\Omega)$ defines the linear functional on $\mathcal{D}$

$$u(\varphi) := \langle u, \varphi \rangle_{\mathcal{D}' \times \mathcal{D}} := \int_\Omega u \varphi \, dx.$$

We call these functionals in $\mathcal{D}'$ distributions. When u is a function, we identify it with the generated distribution. The formula (3.2) is valid for functions $u \in C^\alpha$. The strong regularity is needed only on the left hand side. Thus, we use the less demanding right hand side to extend the definition of differentiation for distributions:

Definition 37. *For $u \in \mathcal{D}'$, we define $g \in \mathcal{D}'$ to be the generalized derivative $D_g^\alpha u$ of u by*

$$\langle g, \varphi \rangle_{\mathcal{D}' \times \mathcal{D}} = (-1)^{|\alpha|} \langle u, D^\alpha \varphi \rangle_{\mathcal{D}' \times \mathcal{D}} \qquad \forall\, \varphi \in \mathcal{D}$$

If $u \in C^\alpha$, then D_g^α coincides with D^α.
The function space of **locally integrable** functions on Ω is called

$$L_1^{loc}(\Omega) = \{u : u_K \in L_1(K)\ \forall \text{ compact } K \subset \Omega\}.$$

It contains functions which can behave very badly near $\partial\Omega$. E.g., $e^{e^{1/x}}$ is in $L^1_{loc}(0,1)$. If Ω is unbounded, then the constant function 1 is in L_1^{loc}, but not in L_1.

Definition 38. *For $u \in L_1^{loc}$, we call g the weak derivative $D_w^\alpha u$, if $g \in L_1^{loc}$ satisfies*

$$\int_\Omega g(x)\varphi(x)\,dx = (-1)^{|\alpha|} \int_\Omega u(x) D^\alpha \varphi(x)\,dx \qquad \forall\, \varphi \in \mathcal{D}.$$

The weak derivative is more general than the classical derivative, but more restrictive than the generalized derivative.

Example 39. *Let $\Omega = (-1, 1)$ and*

$$u(x) = \left\{ \begin{array}{ll} 1 + x & x \le 0 \\ 1 - x & x > 0 \end{array} \right\}$$

Then,

$$g(x) = \left\{ \begin{array}{cl} 1 & x \le 0 \\ -1 & x > 0 \end{array} \right\}$$

is the first generalized derivative D_g^1 of u, which is also a weak derivative. The second generalized derivative h is

$$\langle h, \varphi \rangle = -2\varphi(0) \qquad \forall\, \varphi \in \mathcal{D}$$

It is not a weak derivative.

In the following, we will focus on weak derivatives. Unless it is essential we will skip the sub-scripts w and g.

3.2 Sobolev spaces

For $k \in \mathbb{N}_0$ and $1 \le p < \infty$, we define the Sobolev norms

$$\|u\|_{W_p^k(\Omega)} := \left(\sum_{|\alpha| \le k} \|D^\alpha u\|_{L_p}^p \right)^{1/p},$$

for $k \in \mathbb{N}_0$ we set

$$\|u\|_{W_\infty^k(\Omega)} := \max_{|\alpha| \le k} \|D^\alpha u\|_{L_\infty}.$$

In both cases, we define the **Sobolev spaces** via

$$W_p^k(\Omega) = \{u \in L_1^{loc} : \|u\|_{W_p^k} < \infty\}$$

In the previous chapter we have seen the importance of complete spaces. This is the case for Sobolev spaces:

Theorem 40. *The Sobolev space $W_p^k(\Omega)$ is a Banach space.*

Proof: Let v_j be a Cauchy sequence with respect to $\| \cdot \|_{W_p^k}$. This implies that $D^\alpha v_j$ is a Cauchy sequence in L_p, and thus converges to some v^α in $\|.\|_{L_p}$.

We verify that $D^\alpha v_j \to v^\alpha$ implies $\int_\Omega D^\alpha v_j \varphi \, dx \to \int_\Omega v^\alpha \varphi \, dx$ for all $\varphi \in \mathcal{D}$. Let K be the compact support of φ. There holds

$$\begin{aligned} \int_\Omega (D^\alpha v_j - v^\alpha)\varphi \, dx &= \int_K (D^\alpha v_j - v^\alpha)\varphi \, dx \\ &\le \|D^\alpha v_j - v^\alpha\|_{L_1(K)} \|\varphi\|_{L_\infty} \\ &\le \|D^\alpha v_j - v^\alpha\|_{L_p(K)} \|\varphi\|_{L_\infty} \to 0 \end{aligned}$$

Finally, we have to check that v^α is the weak derivative of v:

$$\begin{aligned} \int v^\alpha \varphi \, dx &= \lim_{j\to\infty} \int_\Omega D^\alpha v_j \varphi \, dx \\ &= \lim_{j\to\infty} (-1)^{|\alpha|} \int_\Omega v_j D^\alpha \varphi \, dx = \\ &= (-1)^\alpha \int_\Omega v D^\alpha \varphi \, dx. \end{aligned}$$

□

An alternative definition of Sobolev spaces were to take the closure of smooth functions in the domain, i.e.,

$$\widetilde{W}_p^k := \overline{\{C^\infty(\Omega) : \|.\|_{W_p^k} \le \infty\}}^{\|.\|_{W_p^k}}.$$

A third one is to take continuously differentiable functions up to the boundary

$$\widehat{W}_p^k := \overline{C^\infty(\overline{\Omega})}^{\|.\|_{W_p^k}}.$$

Under moderate restrictions, these definitions lead to the same spaces:

Theorem 41. *Let $1 \le p < \infty$. Then $\widetilde{W}_p^k = W_p^k$.*

Definition 42. *The domain Ω has a* **Lipschitz boundary**, *$\partial\Omega$, if there exists a collection of open sets O_i, a positive parameter ε, an integer N and a finite number L, such that for all $x \in \partial\Omega$ the ball of radius ε centered at x is contained in some O_i, no more than N of the sets O_i intersect non-trivially, and each part of the boundary $O_i \cap \Omega$ is a graph of a Lipschitz function $\varphi_i : \mathbb{R}^{d-1} \to \mathbb{R}$ with Lipschitz norm bounded by L.*

Theorem 43. *Assume that Ω has a Lipschitz boundary, and let $1 \le p < \infty$. Then $\widehat{W}_p^k = W_p^k$.*

The case W_2^k is special, it is a Hilbert space. We denote it by

$$H^k(\Omega) := W_2^k(\Omega).$$

The inner product is

$$(u, v)_{H^k} := \sum_{|\alpha| \le k} (D^\alpha u, D^\alpha v)_{L_2}$$

In the following, we will prove most theorems for the Hilbert spaces H^k, and state the general results for W_p^k.

3.3 Trace theorems and their applications

We consider boundary values of functions in Sobolev spaces. Clearly, this is not well defined for $H^0 = L_2$. But, as we will see, in H^1 and higher order Sobolev spaces, it makes sense to talk about $u|_{\partial\Omega}$. The definition of traces is essential to formulate boundary conditions of PDEs in weak form.

We start in one dimension. Let $u \in C^1([0, h])$ with some $h > 0$. Then, we can bound

$$\begin{aligned} u(0) &= \left(1 - \frac{x}{h}\right) u(x)|_{x=0} = -\int_0^h \left\{\left(1 - \frac{x}{h}\right) u(x)\right\}' dx \\ &= -\int_0^h \frac{-1}{h} u(x) + \left(1 - \frac{x}{h}\right) u'(x)\, dx \\ &\le \left\|\frac{1}{h}\right\|_{L_2} \|u\|_{L_2} + \left\|1 - \frac{x}{h}\right\|_{L_2} \|u'\|_{L_2} \\ &\simeq h^{-1/2} \|u\|_{L_2(0,h)} + h^{1/2} \|u'\|_{L_2(0,h)}. \end{aligned}$$

This estimate includes the scaling with the interval length h. If we are not interested in the scaling, we apply Cauchy-Schwarz in $\mathbb{R}^2$, and combine the L_2 norm and the H^1 semi-norm $\|u'\|_{L_2}$ to the full H^1 norm and obtain

$$|u(0)| \le \sqrt{h^{-1/2} + h^{1/2}} \sqrt{\|u\|_{L_2}^2 + \|u'\|_{L_2}^2} = c\, \|u\|_{H^1}.$$

Next, we extend the trace operator to the whole Sobolev space H^1:

Theorem 44. *There is a well defined and continuous trace operator*

$$\operatorname{tr} : H^1((0,h)) \to \mathbb{R}$$

whose restriction to $C^1([0,h])$ *coincides with*

$$u \to u(0).$$

Proof: Use that $C^1([0,h])$ is dense in $H^1(0,h)$. Take a sequence u_j in $C^1([0,h])$ converging to u in H^1-norm. The values $u_j(0)$ are Cauchy, and thus converge to an u_0. The limit is independent of the choice of the sequence u_j. This allows to define $\operatorname{tr} u := u_0$. $\square$

Now, we extend this 1D result to domains in more dimensions. Let Ω be bounded, $\partial\Omega$ be Lipschitz, and consists of M pieces Γ_i of smoothness C^1.

We can construct the following covering of a neighbourhood of $\partial\Omega$ in Ω: Let $Q = (0,1)^2$. For $1 \leq i \leq M$, let $s_i \in C^1(Q, \Omega)$ be invertible and such that $\|s_i'\|_{L_\infty} \leq c$, $\|(s_i')^{-1}\|_{L_\infty} \leq c$, and $\det s_i' > 0$. The domains $S_i := s_i(Q)$ are such that $s_i((0,1) \times \{0\}) = \Gamma_i$, and the parameterizations match on $s_i(\{0,1\} \times (0,1))$.

Theorem 45. *There exists a well defined and continuous operator*

$$\operatorname{tr} : H^1(\Omega) \to L_2(\partial\Omega)$$

which coincides with $u|_{\partial\Omega}$ *for* $u \in C^1(\overline{\Omega})$.

Proof: Again, we prove that

$$\operatorname{tr} : C^1(\overline{\Omega}) \to L_2(\partial\Omega) : u \to u|_{\partial\Omega}$$

is a bounded operator w.r.t. the norms $\|.\|_{H^1}$ and L_2, and conclude by density. We use the partitioning of $\partial\Omega$ into the pieces Γ_i, and transform to the simple square domain $Q = (0,1)^2$. Define the functions u_i on $Q = (0,1)^2$ as

$$\tilde{u}_i(\tilde{x}) = u(s_i(\tilde{x}))$$

We transfer the L_2 norm to the simple domain:

$$\begin{aligned}
\|\operatorname{tr} u\|^2_{L_2(\partial\Omega)} &= \sum_{i=1}^{M} \int_{\Gamma_i} u(x)^2 \, dx \\
&= \sum_{i=1}^{M} \int_0^1 u(s_i(\xi,0))^2 \left| \frac{\partial s_i}{\partial \xi}(\xi,0) \right| d\xi \\
&\leq c \sum_{i=1}^{M} \int_0^1 \tilde{u}_i(\xi,0)^2 \, d\xi
\end{aligned}$$

To transform the H^1-norm, we differentiate with respect to $\tilde{x}$ by applying the chain rule

$$\frac{d\tilde{u}_i}{d\tilde{x}}(\tilde{x}) = \frac{du}{dx}(s_i(\tilde{x}))\frac{ds_i}{d\tilde{x}}(\tilde{x}).$$

Solving for $\frac{du}{dx}$ is

$$\frac{du}{dx}(s_i(\tilde{x})) = \frac{d\tilde{u}_i}{d\tilde{x}}(\tilde{x}) \left(\frac{ds}{d\tilde{x}}\right)^{-1} (\tilde{x})$$

The bounds onto s' and $(s')^{-1}$ imply that

$$c^{-1} |\nabla_x u| \leq |\nabla_{\tilde{x}} \tilde{u}| \leq c\, |\nabla_x u|$$

We start from the right hand side of the stated estimate:

$$\begin{aligned}
\|u\|^2_{H^1(\Omega)} &\geq \sum_{i=1}^M \int_{S_i} |\nabla_x u|^2 \, dx \\
&= \sum_{i=1}^M \int_Q |\nabla_x u(s_i(\tilde{x}))|^2 \det(s')\, d\tilde{x} \\
&\geq c \sum_{i=1}^M \int_Q |\nabla_{\tilde{x}} \tilde{u}(\tilde{x})|^2 \; d\tilde{x}
\end{aligned}$$

We got a lower bound for $\det(s') = (\det(s')^{-1})^{-1}$ from the upper bound for $(s')^{-1}$.

It remains to prove the trace estimate on Q. Here, we apply the previous one dimensional result

$$|u(\xi,0)|^2 \leq c \int_0^1 \left\{ u(\xi,\eta)^2 + \left(\frac{\partial u(\xi,\eta)}{\partial \eta}\right)^2 \right\} d\eta \qquad \forall\, \xi \in (0,1)$$

The result follows from integrating over ξ

$$\begin{aligned}
\int_0^1 |u(\xi,0)|^2 \, d\xi &\leq c \int_0^1 \int_0^1 \left\{ u(\xi,\eta)^2 + \left(\frac{\partial u(\xi,\eta)}{\partial \eta}\right)^2 \right\} d\eta \, d\xi \\
&\leq c \, \|u\|^2_{H^1(Q)}.
\end{aligned}$$

□

Considering the trace operator from $H^1(\Omega)$ to $L_2(\partial\Omega)$ is not sharp with respect to the norms. We will improve the embedding later.

By means of the trace operator we can define the sub-space

$$H^1_0(\Omega) = \{u \in H^1(\Omega) : \text{tr } u = 0\}$$

It is a true sub-space, since $u = 1$ does belong to H^1, but not to H_0^1. It is a closed sub-space, since it is the kernel of a continuous operator.

By means of the trace inequality, one verifies that the linear functional

$$g(v) := \int_{\Gamma_N} g \text{ tr } v \, dx$$

is bounded on H^1.

Integration by parts

The definition of the trace allows us to perform integration by parts in H^1:

$$\int_\Omega \nabla u \varphi \, dx = -\int_\Omega u \operatorname{div} \varphi \, dx + \int_{\partial\Omega} \operatorname{tr} u \, \varphi \cdot n \, dx \qquad \forall \varphi \in [C^1(\overline{\Omega})]^2$$

The definition of the weak derivative (e.g. the weak gradient) looks similar. It allows only test functions φ with compact support in Ω, i.e., having zero boundary values. Only by choosing a normed space, for which the trace operator is well defined, we can state and prove integration by parts. Again, the short proof is based on the density of $C^1(\overline{\Omega})$ in H^1.

Sobolev spaces over sub-domains

Let Ω consist of M Lipschitz-continuous sub-domains Ω_i such that

- $\overline{\Omega} = \cup_{i=1}^M \overline{\Omega}_i$
- $\Omega_i \cap \Omega_j = \emptyset \quad$ if $i \neq j$

The interfaces are $\gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$. The outer normal vector of Ω_i is n_i.

Theorem 46. *Let $u \in L_2(\Omega)$ such that*

- *$u_i := u|_{\Omega_i}$ is in $H^1(\Omega_i)$, and $g_i = \nabla u_i$ is its weak gradient*
- *the traces on common interfaces coincide:*

$$\operatorname{tr}_{\gamma_{ij}} u_i = \operatorname{tr}_{\gamma_{ij}} u_j$$

Then u belongs to $H^1(\Omega)$. Its weak gradient $g = \nabla u$ fulfills $g|_{\Omega_i} = g_i$.

Proof: We have to verify that $g \in L_2(\Omega)^d$, defined by $g|_{\Omega_i} = g_i$, is the weak gradient of u, i.e.,

$$\int_\Omega g \cdot \varphi \, dx = -\int_\Omega u \operatorname{div} \varphi \, dx \qquad \forall \varphi \in [C_0^\infty(\Omega)]^d$$

We are using Green's formula on the sub-domains

$$\begin{aligned}
\int_\Omega g \cdot \varphi \, dx &= \sum_{i=1}^M \int_{\Omega_i} g_i \cdot \varphi \, dx = \sum_{i=1}^M \int_{\Omega_i} \nabla u_i \cdot \varphi \, dx \\
&= \sum_{i=1}^M - \int_{\Omega_i} u_i \operatorname{div} \varphi \, dx + \int_{\partial\Omega_i} \operatorname{tr} u_i \, \varphi \cdot n_i \, ds \\
&= - \int_\Omega u \operatorname{div} \varphi \, dx + \sum_{\gamma_{ij}} \int_{\gamma_{ij}} \left\{ \operatorname{tr}_{\gamma_{ij}} u_i \, \varphi \cdot n_i + \operatorname{tr}_{\gamma_{ij}} u_j \, \varphi \cdot n_j \right\} \, ds \\
&= - \int_\Omega u \, \operatorname{div} \varphi \, dx
\end{aligned}$$

We have used that $\varphi = 0$ on $\partial\Omega$, and $n_i = -n_j$ on γ_{ij}. □

Applications of this theorem are (conforming nodal) finite element spaces. The partitioning Ω_i is the mesh. On each sub-domain, i.e., on each element T, the functions are polynomials and thus in $H^1(T)$. The finite element functions are constructed to be continuous, i.e., the traces match on the interfaces. Thus, the finite element space is a sub-space of H^1.

Extension operators

Some estimates are elementary to verify on simple domains such as squares Q. One technique to transfer these results to general domains is to extend a function $u \in H^1(\Omega)$ onto a larger square Q, apply the result for the square, and restrict the result onto the general domain Ω. This is now the motivation to study extension operators.

We construct a non-overlapping covering $\{S_i\}$ of a neighbourhood of $\partial\Omega$ on both sides. Let $\partial\Omega = \cup \Gamma_i$ consist of smooth parts. Let $s : (0,1) \times (-1,1) \to S_i : (\xi, \eta) \to x$ be an invertible function such that

$$\begin{aligned}
s_i((0,1) \times (0,1)) &= S_i \cap \Omega \\
s_i((0,1) \times \{0\}) &= \Gamma_i \\
s_i((0,1) \times (-1,0)) &= S_i \setminus \overline{\Omega}
\end{aligned}$$

Assume that $\| \frac{ds_i}{dx} \|_{L_\infty}$ and $\| \left(\frac{ds_i}{dx} \right)^{-1} \|_{L_\infty}$ are bounded.

This defines an invertible mapping $x \to \hat{x}(x)$ from the inside to the outside by

$$\hat{x}(x) = s_i(\xi(x), -\eta(x)).$$

The mapping preserve the boundary Γ_i. The transformations s_i should be such that $x \to \hat{x}$ is consistent at the interfaces between S_i and S_j.

With the flipping operator $f : (\xi, \eta) \to (\xi, -\eta)$, the mapping is the composite $\hat{x}(x) = s_i(f(s_i^{-1}))$. From that, we obtain the bound

$$\left\| \frac{d\hat{x}}{dx} \right\| \leq \left\| \frac{ds}{dx} \right\| \left\| \left(\frac{ds}{dx} \right)^{-1} \right\|.$$

Define the domain $\widetilde{\Omega} = \Omega \cup S_1 \cup \ldots \cup S_M$.

We define the extension operator by

$$\begin{aligned} (Eu)(\hat{x}) &= u(x) \qquad \forall\, x \in \cup S_i \\ (Eu)(x) &= u(x) \qquad \forall\, x \in \Omega \end{aligned} \tag{3.3}$$

Theorem 47. *The extension operator $E : H^1(\Omega) \rightarrow H^1(\widetilde{\Omega})$ is well defined and bounded with respect to the norms*

$$\|Eu\|_{L_2(\widetilde{\Omega})} \leq c\,\|u\|_{L_2(\Omega)}$$

and

$$\|\nabla Eu\|_{L_2(\widetilde{\Omega})} \leq c\,\|\nabla u\|_{L_2(\Omega)}$$

Proof: Let $u \in C^1(\overline{\Omega})$. First, we prove the estimates for the individual pieces S_i:

$$\int_{S_i \setminus \Omega} Eu(\hat{x})^2 \, d\hat{x} = \int_{S_i \cap \Omega} u(x)^2 \, \det\left(\frac{d\hat{x}}{dx}\right) dx \leq c\|u\|^2_{L_2(S_i \cap \Omega)}$$

For the derivatives we use

$$\frac{dEu(\hat{x})}{d\hat{x}} = \frac{du(x(\hat{x}))}{d\hat{x}} = \frac{du}{dx}\frac{dx}{d\hat{x}}.$$

Since $\frac{dx}{d\hat{x}}$ and $(\frac{dx}{d\hat{x}})^{-1} = \frac{d\hat{x}}{dx}$ are bounded, one obtains

$$|\nabla_{\hat{x}} Eu(\hat{x})| \simeq |\nabla_x u(x)|,$$

and

$$\int_{S_i \setminus \Omega} |\nabla_{\hat{x}} Eu|^2 \, d\hat{x} \leq c \int_{S_i \cap \Omega} |\nabla u|^2 \, dx$$

These estimates prove that E is a bounded operator into H^1 on the sub-domains $S_i \setminus \Omega$. The construction was such that for $u \in C^1(\overline{\Omega})$, the extension Eu is continuous across $\partial\Omega$, and also across the individual S_i. By Theorem 46, Eu belongs to $H^1(\widetilde{\Omega})$, and

$$\|\nabla Eu\|^2_{L_2(\widetilde{\Omega})} = \|\nabla u\|^2_{\Omega} + \sum_{i=1}^{M} \|\nabla u\|^2_{S_i \setminus \Omega} \leq c\|\nabla u\|^2_{L_2(\Omega)},$$

By density, we get the result for $H^1(\Omega)$. Let $u_j \in C^1(\overline{\Omega}) \rightarrow u$, than u_j is Cauchy, Eu_j is Cauchy in $H^1(\widetilde{\Omega})$, and thus converges to $u \in H^1(\widetilde{\Omega})$.

The extension of functions from $H^1_0(\Omega)$ onto larger domains is trivial: Extension by 0 is a bounded operator. One can extend functions from $H^1(\Omega)$ into $H^1_0(\widetilde{\Omega})$, and further, to an arbitrary domain by extension by 0.

For $\hat{x} = s_i(\xi, -\eta)$, $\xi, \eta \in (0,1)^2$, define the extension

$$E_0 u(\hat{x}) = (1 - \eta)\, u(x)$$

This extension vanishes at $\partial\widetilde{\Omega}$

Theorem 48. *The extension E_0 is an extension from $H^1(\Omega)$ to $H_0^1(\widetilde{\Omega})$. It is bounded w.r.t.*

$$\|E_0 u\|_{H^1(\widetilde{\Omega})} \leq c\|u\|_{H^1(\Omega)}$$

Proof: Exercises

In this case, it is not possible to bound the gradient term only by gradients. To see this, take the constant function on Ω. The gradient vanishes, but the extension is not constant.

3.3.1 The trace space $H^{1/2}$

The trace operator is continuous from $H^1(\Omega)$ into $L_2(\partial\Omega)$. But, not every $g \in L_2(\partial\Omega)$ is a trace of some $u \in H^1(\Omega)$. We will motivate why the trace space is the fractional order Sobolev space $H^{1/2}(\partial\Omega)$.

We introduce a stronger space, such that the trace operator is still continuous, and onto. Let $V = H^1(\Omega)$, and define the trace space as the range of the trace operator

$$W = \{\text{tr } u : u \in H^1(\Omega)\}$$

with the norm

$$\|\operatorname{tr} u\|_W = \inf_{\substack{v\in V \\ \text{tr } u = \text{tr } v}} \|v\|_V. \tag{3.4}$$

This is indeed a norm on W. The trace operator is continuous from $V \to W$ with norm 1.

Lemma 49. *The space $(W, \|.\|_W)$ is a Banach space. For all $g \in W$ there exists an $u \in V$ such that* tr $u = g$ *and* $\|u\|_V = \|g\|_W$

Proof: The kernel space $V_0 := \{v : \text{tr } v = 0\}$ is a closed sub-space of V. If tr $u =$ tr v, then $z := u - v \in V_0$. We can rewrite

$$\|\text{ tr } u\|_W = \inf_{z\in V_0} \|u - z\|_V = \|u - P_{V_0} u\|_V \qquad \forall\, u \in V$$

Now, let $g_n = \text{tr } u_n \in W$ be a Cauchy sequence. This does not imply that u_n is Cauchy, but $P_{V_0^\perp} u_n$ is Cauchy in V:

$$\|P_{V_0^\perp}(u_n - u_m)\|_V = \|\operatorname{tr}(u_n - u_m)\|_W.$$

The $P_{V_0^\perp} u_n$ converge to some $u \in V_0^\perp$, and g_n converge to $g := \text{tr } u$. $\square$

The minimizer in (3.4) fulfills

$$\text{tr } u = g \qquad \text{and} \qquad (u, v)_V = 0 \qquad \forall\, v \in V_0.$$

This means that u is the solution of the weak form of the Dirichlet problem

$$\begin{aligned} -\Delta u + u &= 0 && \text{in } \Omega \\ u &= g && \text{on } \partial\Omega. \end{aligned}$$

To give an explicit characterization of the norm $\|.\|_W$, we introduce **Hilbert space interpolation**:

Let $V_1 \subset V_0$ be two Hilbert spaces, such that V_1 is dense in V_0, and the embedding operator $id : V_1 \to V_0$ is compact. We can pose the eigen-value problem: Find $z \in V_1$, $\lambda \in \mathbb{R}$ such that

$$(z, v)_{V_1} = \lambda (z, v)_{V_0} \qquad \forall\, v \in V_1.$$

There exists a sequence of eigen-pairs (z_k, λ_k) such that $\lambda_k \to \infty$. The z_k form an orthonormal basis in V_0, and an orthogonal basis in V_1.

The converse is also true. If z_k is a basis for V_0, and the eigenvalues $\lambda_k \to \infty$, then the embedding $V_1 \subset V_0$ is compact.

Given $u \in V_0$, it can be expanded in the orthonormal eigen-vector basis:

$$u = \sum_{k=0}^{\infty} u_k z_k \qquad \text{with} \qquad u_k = (u, z_k)_{V_0}$$

The $\|.\|_{V_0}$ - norm of u is

$$\|u\|_{V_0}^2 = (\sum_k u_k z_k, \sum_l u_l z_l)_{V_0} = \sum_{k,l} u_k u_l (z_k, z_l)_{V_0} = \sum_k u_k^2.$$

If $u \in V_1$, then

$$\|u\|_{V_1}^2 = (\sum_k u_k z_k, \sum_l u_l z_l)_{V_1} = \sum_{k,l} u_k u_l (z_k, z_l)_{V_1} = \sum_{k,l} u_k u_l \lambda_k (z_k, z_l)_{V_0} = \sum_k u_k^2 \lambda_k$$

The sub-space space V_1 consists of all $u = \sum u_k z_k$ such that $\sum_k \lambda_k u_k^2$ is finite. This suggests the definition of the interpolation norm

$$\|u\|_{V_s}^2 = \sum_k (u, z_k)_{V_0}^2 \lambda_k^s,$$

and the interpolation space $V_s = [V_0, V_1]_s$ as

$$V_s = \{u \in V_0 : \|u\|_{V_s} < \infty\}.$$

We have been fast with using infinite sums. To make everything precise, one first works with finite dimensional sub-spaces $\{u : \exists n \in \mathbb{N} \text{ and } u = \sum_{k=1}^n u_k z_k\}$, and takes the closure.

In our case, we apply Hilbert space interpolation to $H^1(0,1) \subset L_2(0,1)$. The eigen-value problem is to find $z_k \in H^1$ and $\lambda_k \in \mathbb{R}$ such that

$$(z_k, v)_{L_2} + (z_k', v')_{L_2} = \lambda_k\, (z_k, v)_{L_2} \qquad \forall\, v \in H^1$$

By definition of the weak derivative, there holds $(z_k')' = (1-\lambda_k)z_k$, i.e., $z^k \in H^2$. Since $H^2 \subset C^0$, there holds also $z \in C^2$, and a weak solution is also a solution of the strong form

$$\begin{aligned} z_k - z_k'' &= \lambda_k z_k \qquad \text{on } (0,1) \\ z_k'(0) = z_k'(1) &= 0 \end{aligned} \tag{3.5}$$

All solutions, normalized to $\|z_k\|_{L_2} = 1$, are

$$z_0 = 1 \qquad \lambda_0 = 1$$

and, for $k \in \mathbb{N}$,

$$z_k(x) = \sqrt{2}\cos(k\pi x) \qquad \lambda_k = 1 + k^2\pi^2.$$

Indeed, expanding $u \in L_2$ in the cos-basis $u = u_0 + \sum_{k=1}^{\infty} u_k \sqrt{2}\cos(k\pi x)$, one has

$$\|u\|_{L_2}^2 = \sum_{k=0}^{\infty} (u, z_k)_{L_2}^2$$

and

$$\|u\|_{H^1}^2 = \sum_{k=0}^{\infty} (1 + k^2\pi^2)(u, z_k)_{L_2}^2$$

Differentiation adds a factor $k\pi$. Hilbert space interpolation allows to define the fractional order Sobolev norm ($s \in (0,1)$)

$$\|u\|_{H^s(0,1)}^2 = \sum_{k=0}^{\infty} (1 + k^2\pi^2)^s (u, z_k)_{L_2}^2$$

We consider the trace $\operatorname{tr}|_E$ of $H^1((0,1)^2)$ onto one edge $E = (0,1) \times \{0\}$. For $g \in W_E := \operatorname{tr} H^1((0,1)^2)$, the norm $\|g\|_W$ is defined by

$$\|g\|_W = \|u_g\|_{H^1}.$$

Here, u_g solves the Dirichlet problem $u_g|_E = g$, and $(u_g, v)_{H^1} = 0\ \forall v \in H^1$ such that $\operatorname{tr}_E v = 0$.

Since $W \subset L_2(E)$, we can expand g in the L_2-orthonormal cosine basis z_k

$$g(x) = \sum g_n z_k(x)$$

The Dirichlet problems for the z_k,

$$\begin{aligned} -\Delta u_k + u_k &= 0 \qquad \text{in } \Omega \\ u_k &= z_k \qquad \text{on } E \\ \tfrac{\partial u_k}{\partial n} &= 0 \qquad \text{on } \partial\Omega \setminus E, \end{aligned}$$

have the explicit solution

$$u_0(x,y) = 1$$

and

$$u_k(x,y) = \sqrt{2}\cos(k\pi x)\frac{e^{k\pi(1-y)} + e^{-k\pi(1-y)}}{e^{k\pi} + e^{-k\pi}}.$$

The asymptotic is

$$\|u_k\|_{L_2}^2 \simeq (k+1)^{-1}$$

and

$$\|\nabla u_k\|_{L_2}^2 \simeq k$$

Furthermore, the u_k are orthogonal in $(.,.)_{H^1}$. Thus $u_g = \sum_n g_n u_k$ has the norm

$$\|u_g\|_{H^1}^2 = \sum g_n^2 \|u_k\|_{H^1}^2 \simeq \sum g_n^2(1+k).$$

This norm is equivalent to $H^{1/2}(E)$.

We have proven that the trace space onto one edge is the interpolation space $H^{1/2}(E)$. This is also true for general domains (Lipschitz, with piecewise smooth boundary).

3.4 Equivalent norms on H^1 and on sub-spaces

The intention is to formulate 2^{nd} order variational problems in the Hilbert space H^1. We want to apply the Lax-Milgram theory for continuous and coercive bilinear forms $A(.,.)$. We present techniques to prove coercivity.

The idea is the following. In the norm

$$\|v\|_{H^1}^2 = \|v\|_{L_2}^2 + \|\nabla v\|_{L_2}^2,$$

the $\|\nabla \cdot\|_{L_2}$-semi-norm is the dominating part up to the constant functions. The L_2 norm is necessary to obtain a norm. We want to replace the L_2 norm by some different term (e.g., the L_2-norm on a part of Ω, or the L_2-norm on $\partial\Omega$), and want to obtain an equivalent norm.

We formulate an abstract theorem relating a norm $\|.\|_V$ to a semi-norm $\|.\|_A$. An equivalent theorem was proven by Tartar.

Theorem 50 (Tartar)**.** *Let $(V,(.,.)_V)$ and $(W,(.,.)_W)$ be Hilbert spaces, such that the embedding $id : V \to W$ is compact. Let $A(.,.)$ be a non-negative, symmetric and V-continuous bilinear form with kernel $V_0 = \{v : A(v,v) = 0\}$. Assume that*

$$\|v\|_V^2 \simeq \|v\|_W^2 + \|v\|_A^2 \qquad \forall\, v \in V \tag{3.6}$$

Then there holds

1. *The kernel V_0 is finite dimensional. On the factor space V/V_0, $A(.,.)$ is an equivalent norm to the quotient norm*

$$\|u\|_A \simeq \inf_{v\in V_0} \|u-v\|_V \qquad \forall\, u\in V \tag{3.7}$$

2. *Let $B(.,.)$ be a continuous, non-negative, symmetric bilinear form on V such that $A(.,.)+B(.,.)$ is an inner product. Then there holds*

$$\|v\|_V^2 \simeq \|v\|_A^2 + \|v\|_B^2 \qquad \forall\, v\in V$$

3. *Let $V_1 \subset V$ be a closed sub-space such that $V_0 \cap V_1 = \{0\}$. Then there holds*

$$\|v\|_V \simeq \|v\|_A \qquad \forall\, v \in V_1$$

Proof: 1. Assume that V_0 is not finite dimensional. Then there exists an $(.,.)_V$-orthonormal sequence $u_k \in V_0$. Since the embedding $id : V \to W$ is compact, it has a sub-sequence converging in $\|.\|_W$. But, since

$$2 = \|u_k - u_l\|_V^2 \simeq \|u_k-u_l\|_W^2 + \|u_k-u_l\|_A = \|u_k-u_l\|_W^2$$

for $k \neq l$, u_k is not Cauchy in W. This is a contradiction to an infinite dimensional kernel space V_0. We prove the equivalence (3.7). To bound the left hand side by the right hand side, we use that $V_0 = \ker A$, and norm equivalence (3.6):

$$\|u\|_A = \inf_{v\in V_0} \|u-v\|_A \le \inf_{v\in V_0}\|u-v\|_V$$

The quotient norm is equal to $\|P_{V_0^\perp} u\|$. We have to prove that $\|P_{V_0^\perp}u\|_V \le \|P_{V_0^\perp}u\|_A$ for all $u\in V$. This follows after proving $\|u\|_V \le \|u\|_A$ for all $u \in V_0^\perp$. Assume that this is not true. I.e., there exists a V-orthogonal sequence (u_k) such that $\|u_k\|_A \le k^{-1}\|u_k\|_V$. Extract a sub-sequence converging in $\|.\|_W$, and call it u_k again. From the norm equivalence (3.6) there follows

$$2 = \|u_k-u_l\|_V^2 \preceq \|u_k-u_l\|_W + \|u_k-u_l\|_A \to 0$$

2. On V_0, $\|.\|_B$ is a norm. Since V_0 is finite dimensional, it is equivalent to $\|.\|_V$, say with bounds

$$c_1\,\|v\|_V^2 \le \|v\|_B^2 \le c_2\,\|v\|_V^2 \qquad \forall\, v\in V_0$$

From 1. we know that

$$c_3\,\|v\|_V^2 \le \|v\|_A^2 \le c_4\,\|v\|_V^2 \qquad \forall\, v\in V_0^\perp.$$

Now, we bound

$$\begin{aligned}
\|u\|_V^2 &= \|P_{V_0}u\|_V^2 + \|P_{V_0^\perp}u\|_V^2 \\
&\leq \frac{1}{c_1}\| \underbrace{P_{V_0}u}_{u-P_{V_0^\perp}u} \|_B^2 + \|P_{V_0^\perp}\|_V^2 \\
&\leq \frac{2}{c_1}\left(\|u\|_B^2 + c_2\|P_{V_0^\perp}u\|_V^2\right) + \|P_{V_0^\perp}u\|_V^2 \\
&= \frac{2}{c_1}\|u\|_B^2 + \frac{1}{c_2}\left(1+\frac{2c_2}{c_1}\right)\|P_{V_0^\perp}u\|_A^2 \\
&\preceq \|u\|_B^2 + \|u\|_A^2
\end{aligned}$$

3. Define $B(u,v) = (P_{V_1}^\perp u, P_{V_1}^\perp u)_V$. Then $A(.,.) + B(.,.)$ is an inner product: $A(u,u) + B(u,u) = 0$ implies that $u \in V_0$ and $u \in V_1$, thus $u = \{0\}$. From 2. there follows that $A(.,.) + B(.,.)$ is equivalent to $(.,.)_V$. The result follows from reducing the equivalence to V_1.

□

We want to apply Tartar's theorem to the case $V = H^1$, $W = L_2$, and $\|v\|_A = \|\nabla v\|_{L_2}$. The theorem requires that the embedding $id : H^1 \to L_2$ is compact. This is indeed true for bounded domains Ω:

Theorem 51. *The embedding of $H^k \to H^l$ for $k > l$ is compact.*

We sketch a proof for the embedding $H^1 \subset L_2$. First, prove the compact embedding $H_0^1(Q) \to L_2(Q)$ for a square Q, w.l.o.g. set $Q = (0,1)^2$. The eigen-value problem: Find $z \in H_0^1(Q)$ and λ such that

$$(z,v)_{L_2} + (\nabla z, \nabla v)_{L_2} = \lambda (u,v)_{L_2} \qquad \forall\, v \in H_0^1(Q)$$

has eigen-vectors $z_{k,l} = sin(k\pi x) sin(l\pi y)$, and eigen-values $1 + k^2\pi^2 + l^2\pi^2 \to \infty$. The eigen-vectors are dense in L_2. Thus, the embedding is compact.

On a general domain $\Omega \subset Q$, we can extend $H^1(\Omega)$ into $H_0^1(Q)$, embed $H_0^1(Q)$ into $L_2(Q)$, and restrict $L_2(Q)$ onto $L_2(\Omega)$. This is the composite of two continuous and a compact mapping, and thus is compact. □

The kernel V_0 of the semi-norm $\|\nabla v\|$ is the constant function.

Theorem 52 (Friedrichs inequality)**.** *Let $\Gamma_D \subset \partial\Omega$ be of positive measure $|\Gamma_D|$. Let $V_D = \{v \in H^1(\Omega) : \operatorname{tr}_{\Gamma_D} v = 0\}$. Then*

$$\|v\|_{L_2} \preceq \|\nabla v\|_{L_2} \qquad \forall\, v \in V_D$$

Proof: The intersection $V_0 \cap V_D$ is trivial $\{0\}$. Thus, Theorem 50, 3. implies the equivalence

$$\|v\|_V^2 = \|v\|_{L_2}^2 + \|\nabla v\|_{L_2}^2 \simeq \|\nabla v\|_{L_2}.$$

□

Theorem 53 (Poincaré inequality)**.** *There holds*

$$\|v\|_{H^1(\Omega)}^2 \preceq \|\nabla v\|_{L_2}^2 + (\int_\Omega v\,dx)^2$$

Proof: $B(u,v) := (\int_\Omega u\,dx)(\int_\Omega v\,dx)$ is a continuous bilinear form on H^1, and $(\nabla u, \nabla v) + B(u,v)$ is an inner product. Thus, Theorem 50, 2. implies the stated equivalence. □

- Let $\omega \subset \Omega$ have positive measure $|\omega|$ in $\mathbb{R}^d$. Then

$$\|u\|_{H^1(\Omega)}^2 \simeq \|\nabla v\|_{L_2(\Omega)}^2 + \|v\|_{L_2(\omega)},$$

- Let $\gamma \subset \partial\Omega$ have positive measure $|\gamma|$ in $\mathbb{R}^{d-1}$. Then

$$\|u\|_{H^1(\Omega)}^2 \simeq \|\nabla v\|_{L_2(\Omega)}^2 + \|v\|_{L_2(\gamma)},$$

Theorem 54 (Bramble Hilbert lemma)**.** *Let U be some Hilbert space, and $L : H^k \to U$ be a continuous linear operator such that $Lq = 0$ for polynomials $q \in P^{k-1}$. Then there holds*

$$\|Lv\|_U \leq |v|_{H^k}.$$

Proof: The embedding $H^k \to H^{k-1}$ is compact. The V-continuous, symmetric and non-negative bilinear form $A(u,v) = \sum_{\alpha:|\alpha|=k}(\partial^\alpha u, \partial^\alpha v)$ has the kernel P^{k-1}. Decompose $\|u\|_{H^k}^2 = \|u\|_{H^{k-1}}^2 + A(u,u)$. By Theorem 50, 1, there holds

$$\|u\|_A \simeq \inf_{v \in V_0} \|u - v\|_{H^k}$$

The same holds for the bilinear-form

$$A_2(u,v) := (Lu, Lv)_U + A(u,v)$$

Thus

$$\|u\|_{A_2} \simeq \inf_{v \in V_0} \|u - v\|_{H^k} \qquad \forall\, u \in V$$

Equalizing both implies that

$$(Lu, Lu)_U \leq \|u\|_{A_2}^2 \simeq \|u\|_A^2 \qquad \forall\, u \in V,$$

i.e., the claim.

We will need point evaluation of functions in Sobolev spaces H^s. This is possible, we $u \in H^s$ implies that u is continuous.

Theorem 55 (Sobolev's embedding theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary. If $u \in H^s$ with $s > d/2$, then $u \in L_\infty$ with*

$$\|u\|_{L_\infty} \preceq \|u\|_{H^s}$$

There is a function in C^0 within the L_∞ equivalence class.

3.5 Interpolation Spaces

3.5.1 Hilbert space interpolation

Let $V_1 \subset V_0$ be two Hilbert spaces with dense embedding. For simplicity we assume that the embedding is compact. Then there exists a system of eigenvalues λ_k and eigenvectors z_k such that

$$(z_k, v)_1 = \lambda_k^2 \, (z_k, v)_0 \qquad \forall \, v \in V_1.$$

The eigenvectors are orthogonal and are normalized such that

$$(z_k, z_l)_0 = \delta_{k,l} \qquad \text{and} \qquad (z_k, z_l)_1 = \lambda_k^2 \delta_{k,l}.$$

Eigenvalues are ascening, by compactness there holds $\lambda_k \to \infty$.

The set of eigenvectors is a complete system. Thus $u \in V_0$ can be expanded as

$$u = \sum_{k=1}^{\infty} u_k z_k \qquad \text{with } u_k = (u, z_k)_0.$$

There holds

$$\begin{aligned} \|u\|_0^2 &= \sum u_k^2 \\ \|u\|_1^2 &= \sum \lambda_k^2 u_k^2 \quad < \infty \;\text{ for } u \in V_1. \end{aligned}$$

For $s \in (0,1)$ we define the interpolation norm

$$\|u\|_{\tilde{s}} := \Big(\sum_{k=1}^{\infty} \lambda_k^{2s} u_k^2 \Big)^{1/2} \tag{3.8}$$

and the interpolation space

$$V_s := [V_0, V_1]_s := \{ u \in V_0 : \|u\|_{\tilde{s}} < \infty \}.$$

There holds

$$V_1 \subset V_s \subset V_0.$$

Example: Let $V_0 = L_2(0,1)$ and $V_1 = H_0^1(0,1)$. Then

$$z_k = \sqrt{2} \sin(k \pi x) \qquad \text{and} \qquad \lambda_k = k$$

3.5.2 Banach space interpolation

We give an alternative definition of interpolation spaces, which is also applicable for Banach spaces. It is known as Banach space interpolation, K-functional method, real method of interpolation, or Peetre's method.

Let $V_1 \subset V_0$ be Banach spaces with dense and continuous embedding. We define the K-functional $K : \mathbb{R}^+ \times V_0 \to \mathbb{R}$ as

$$K(t,u) := \inf_{v_1 \in V_1} \sqrt{\|u - v_1\|_0^2 + t^2 \|v_1\|_1^2}.$$

Note that

$$\begin{aligned} K(t,u) &\leq \|u\|_0, \\ K(t,u) &\leq t\,\|u\|_1 \qquad \text{for } u \in V_1. \end{aligned}$$

The decay in t measures the *smoothness* of u. For $s \in (0,1)$ we define the interpolation norm as

$$\|u\|_s := \Big(\int_0^\infty t^{-2s} K(t,u)^2 dt/t \Big)^{1/2} \tag{3.9}$$

and the interpolation spaces $V_s := \{u \in V_0 : \|u\|_s < \infty\}$.

The K-functional method is more general. If the spaces are Hilbert, then both interpolation methods coincide:

Theorem 56. *Let $V_1 \subset V_0$ be Hilbert spaces with compact embedding. Then*

$$\|u\|_s = C_s \, \|u\|_{\tilde{s}},$$

where $C_s^2 = \int_0^\infty \frac{\tau^{1-2s}}{1+\tau^2}\, d\tau$.

Proof. For $u = \sum u_k z_k$ we calculate the K-functional as

$$\begin{aligned} K(t,u)^2 &= \inf_{v \in V_1} \|u - v\|_0^2 + t^2 \|v\|_1^2 \\ &= \inf_{\substack{(v_k) \in \ell_2 \\ (\lambda_k v_k) \in \ell_2}} \sum_k (u_k - v_k)^2 + t^2 \lambda_k^2 v_k^2 \\ &= \sum_k \inf_{v_k \in \mathbb{R}} (u_k - v_k)^2 + t^2 \lambda_k^2 v_k^2. \end{aligned}$$

The minimum of each summand is taken for

$$v_k = \frac{1}{1 + t^2 \lambda_k^2} u_k$$

and its value is

$$\frac{t^2 \lambda_k^2}{1 + t^2 \lambda_k^2} u_k^2.$$

Thus

$$K(t,u)^2 = \sum_{k=1}^\infty \frac{t^2 \lambda_k^2}{1 + t^2 \lambda_k^2} u_k^2$$

and

$$\begin{aligned}\|u\|_s^2 &= \int_0^\infty t^{-2s} K(t,u)^2\,dt/t = \int_0^\infty \sum_k \frac{t^2\lambda_k^2}{1+t^2\lambda_k^2} u_k^2\,dt/t \\ &= \sum_k \int_0^\infty t^{-2s} \frac{t^2\lambda_k^2}{1+t^2\lambda_k^2} u_k^2\,dt/t\end{aligned}$$

Substitution $\tau = \lambda_k t$ gives

$$\begin{aligned}\|u\|_s^2 &= \sum_k \int_0^\infty \left(\frac{\tau}{\lambda_k}\right)^{-2s} \frac{\tau^2}{1+\tau^2} u_k^2\,d\tau/\tau \\ &= \sum_k \lambda_k^{2s} u_k^2 \quad \int_0^\infty \frac{\tau^{1-2s}}{1+\tau^2}\,d\tau \\ &= C_s^2 \,\|u\|_{\tilde{s}}^2\end{aligned}$$

□

Theorem 57. *For $u \in V_1$ there holds*

$$\|u\|_s \preceq \|u\|_0^{1-s}\,\|u\|_1^s$$

Proof: Excercise

3.5.3 Operator interpolation

Let $V_1 \subset V_0$ and $W_1 \subset W_0$ with dense embedding.

Theorem 58. *Let $T : V_0 \to W_0$ be a linear operator such that $TV_1 \subset W_1$ with norms*

$$\|T\|_{V_0\to W_0} \leq c_0 \qquad \textit{and} \quad \|T\|_{V_1\to W_1} \leq c_1.$$

Then

$$T : [V_0, V_1]_s \to [W_0, W_1]_s$$

with norm

$$\|T\|_{[V_0,V_1]_s\to[W_0,W_1]_s} \leq c_0^{1-s} c_1^s$$

Proof. We use the definition of the interpolation norm, $TV_1 \subset W_1$, operator norms and substitution $\tau = c_1 t / c_0$

$$\begin{aligned}
\|Tu\|_{[W_0,W_1]_s} &= \int_0^\infty t^{-2s} K_W(t,Tu)^2 \, dt/t \\
&= \int_0^\infty t^{-2s} \inf_{w_1 \in W_1} \{\|Tu - w_1\|_{W_0} + t^2 \|w_1\|_{W_1}^2\} \, dt/t \\
&\leq \int_0^\infty t^{-2s} \inf_{v_1 \in V_1} \{\|Tu - Tv_1\|_{W_0} + t^2 \|Tv_1\|_{W_1}^2\} \, dt/t \\
&\leq \int_0^\infty t^{-2s} \inf_{v_1 \in V_1} \{c_0^2 \|u - v_1\|_{V_0} + t^2 c_1^2 \|v_1\|_{V_1}^2\} \, dt/t \\
&\leq \int_0^\infty \left(\frac{c_0 \tau}{c_1}\right)^{-2s} \inf_{v_1 \in V_1} \{c_0^2 \|u - v_1\|_{V_0}^2 + c_0^2 \tau^2 \|v_1\|_{V_1}^2\} \, d\tau/\tau \\
&= c_0^{2-2s} c_1^{2s} \int_0^\infty \tau^{-2s} K_V(t,u)^2 \, d\tau/\tau \\
&= c_0^{2-2s} c_1^{2s} \|u\|_{[V_0,V_1]_s}^2
\end{aligned}$$

□

3.5.4 Interpolation of Sobolev Spaces

As an example of interpolation spaces we show the following:

Theorem 59. *Let Ω be a Lipschitz domain. Then*

$$[L_2(\Omega), H^2(\Omega)]_{1/2} = H^1(\Omega).$$

Proof. Let Q be a square containing Ω, w.l.o.g. $Q = (0, 2\pi)^2$, and $z_{k,l} = e^{ikx} e^{ily}$ be the trigonometric basis for (complex-valued) periodic Sobolev Spaces $H^m_{per}(Q)$. Then

$$\|u\|_{H^m}^2 \simeq \sum_{k,l} (k^2 + l^2)^m |u_{k,l}|^2,$$

and thus $H^1_{per}(Q) = [H^0_{per}(Q), H^2_{per}(Q)]_{1/2}$ by Hilbert space interpolation.

Now let $E : L_2(\Omega) \to L_2(Q)$ be an extension operator such that

$$E : H^m(\Omega) \to H^m_{per}(Q)$$

is continuous for all $m \in \{0, 1, 2\}$. Furthermore, let

$$R : L_2(Q) \to L_2(\Omega) : u \mapsto u|_\Omega$$

be the restriction operator. Trivially, $R : H^m_{per}(Q) \to H^m(\Omega)$ is continuous for $m \in \mathbb{N}_0$.

We show that

$$\|u\|_{H^1(\Omega)} \simeq \|u\|_{[L_2(\Omega), H^2(\Omega)]_{1/2}}.$$

Using operator interpolation we get

$$
\begin{aligned}
\|u\|_{H^1(\Omega)} &= \|REu\|_{H^1(\Omega)} \le \|R\|\|Eu\|_{H^1(Q)} \\
&\simeq \|Eu\|_{[L_2(Q),H^2_{per}(Q)]_{1/2}} \\
&\le \|E\|^{1/2}_{L_2(\Omega)\to L_2(Q)} \|E\|^{1/2}_{H^2(\Omega)\to H^2_{per}(Q)} \|u\|_{[L_2(\Omega),H^2(\Omega)]_{1/2}} \\
&\simeq \|u\|_{[L_2(\Omega),H^2(\Omega)]_{1/2}},
\end{aligned}
$$

and similarly the other way around. □

Theorem 60. *Let Ω be a Lipschitz domain. Then*

$$[L_2(\Omega), H^2_0(\Omega)]_{1/2} = H^1_0(\Omega).$$

Proof. Exercise □

Literature:

1. J. Bergh and J. Lofstrom. *Interpolation spaces.* Springer, 1976

2. J. H. Bramble. *Multigrid Methods.* Chapman and Hall, 1993

3.6 The weak formulation of the Poisson equation

We are now able to give a precise definition of the weak formulation of the Poisson problem as introduced in Section 1.2, and analyze the existence and uniqueness of a weak solution.

Let Ω be a bounded domain. Its boundary $\partial\Omega$ is decomposed as $\partial\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_R$ according to Dirichlet, Neumann and Robin boundary conditions.

Let

- $u_D \in H^{1/2}(\Gamma_D)$,
- $f \in L_2(\Omega)$,
- $g \in L_2(\Gamma_N \cup \Gamma_R)$,
- $\alpha \in L_\infty(\Gamma_D), \alpha \ge 0$.

Assume that there holds

(a) The Dirichlet part has positive measure $|\Gamma_D| > 0$,

(b) or the Robin term has positive contribution $\int_{\Gamma_R} \alpha \, dx > 0$.

Define the Hilbert space

$$V := H^1(\Omega),$$

the closed sub-space

$$V_0 = \{v : \operatorname{tr}_{\Gamma_D} v = 0\},$$

and the linear manifold

$$V_D = \{u \in V : \operatorname{tr}_{\Gamma_D} u = u_D\}.$$

Define the bilinear form $A(.,.) : V \times V \to \mathbb{R}$

$$A(u,v) = \int_\Omega \nabla u \, \nabla v \, dx + \int_{\Gamma_R} \alpha uv \, ds$$

and the linear form

$$f(v) = \int_\Omega f v \, dx + \int_{\Gamma_N \cup \Gamma_R} gv \, dx.$$

Theorem 61. *The weak formulation of the Poisson problem*

Find $u \in V_D$ such that

$$A(u,v) = f(v) \qquad \forall\, v \in V_0 \tag{3.10}$$

has a unique solution u.

Proof: The bilinear-form $A(.,.)$ and the linear-form $f(.)$ are continuous on V. Tartar's theorem of equivalent norms proves that $A(.,.)$ is coercive on V_0.

Since u_D is in the closed range of $\operatorname{tr}_{\Gamma_D}$, there exists an $\tilde{u}_D \in V_D$ such that

$$\operatorname{tr} \tilde{u}_D = u_D \qquad \text{and} \qquad \|\tilde{u}_D\|_V \preceq \|u_D\|_{H^{1/2}(\Gamma_D)}$$

Now, pose the problem: Find $z \in V_0$ such that

$$A(z,v) = f(v) - A(\widetilde{u}_D, v) \qquad \forall\, v \in V_0.$$

The right hand side is the evaluation of the continuous linear form $f(.) - A(\widetilde{u}_D, .)$ on V_0. Due to Lax-Milgram, there exists a unique solution z. Then, $u := \widetilde{u}_D + z$ solves (3.10). The choice of $\widetilde{u}_D$ is not unique, but, the constructed u is unique. □

3.6.1 Shift theorems

Let us restrict to Dirichlet boundary conditions $u_D = 0$ on the whole boundary. The variational problem: Find $u \in V_0$ such that

$$A(u,v) = f(v) \qquad \forall\, v \in V_0$$

is well defined for all $f \in V_0^*$, and, due to Lax-Milgram there holds

$$\|u\|_{V_0} \leq c\|f\|_{V_0^*}.$$

Vice versa, the bilinear-form defines the linear functional $A(u,.)$ with norm

$$\|A(u,.)\|_{V_0^*} \leq c\|u\|_{V_0}$$

This dual space is called H^{-1}:

$$H^{-1} := [H_0^1(\Omega)]^*$$

Since $H_0^1 \subset L_2$, there is $L_2 \subset H^{-1}(\Omega)$. All negative spaces are defined as $H^{-s}(\Omega) := [H_0^s]^*(\Omega)$, for $s \in \mathbb{R}^+$. There holds

$$\ldots H_0^2 \subset H_0^1 \subset L_2 \subset H^{-1} \subset H^{-2} \ldots$$

The solution operator of the weak formulation is smoothing twice. The statements of shift theorem are that for $s > 0$, the solution operator maps also

$$f \in H^{-1+s} \rightarrow u \in H^{1+s}$$

with norm bounds

$$\|u\|_{H^{1+s}} \preceq \|f\|_{H^{-1+s}}.$$

In this case, we call the problem H^{1+s} - regular.

Theorem 62 (Shift theorem)**.**

(a) Assume that Ω is convex. Then, the Dirichlet problem is H^2 regular.

(b) Let $s \geq 2$. Assume that $\partial\Omega \in C^s$. Then, the Dirichlet problem is H^s-regular.

We give a proof of (a) for the square $(0,\pi)^2$ by Fourier series. Let

$$V_N = \text{span}\{\sin(kx)\sin(ly) : 1 \leq k,l \leq N\}$$

For an $u = \sum_{k,l=1}^N u_{kl}\sin(kx)\sin(ly) \in V_N$, there holds

$$\begin{aligned} \|u\|_{H^2}^2 &= \|u\|_{L_2}^2 + \|\partial_x u\|_{L_2}^2 + \|\partial_y u\|_{L_2}^2 + \|\partial_x^2 u\|_{L_2}^2 + \|\partial_x\partial_y u\|_{L_2}^2 + \|\partial_y^2 u\|^2 \\ &\simeq \sum_{k,l=1}^N (1 + k^2 + l^2 + k^4 + k^2 l^2 + l^4) u_{kl}^2 \\ &\simeq \sum_{k,l=1}^N (k^4 + l^4) u_{kl}^2, \end{aligned}$$

and, for $f = -\Delta u$,

$$\| - \Delta u\|_{L_2}^2 = \sum_{k,l=1}^{N} (k^2 + l^2)^2 u_{kl}^2 \simeq \sum_{k,l=1}^{N} (k^4 + l^4) u_{kl}^2.$$

Thus we have $\|u\|_{H^2} \simeq \|\Delta u\|_{L_2} = \|f\|_{L_2}$ for $u \in V_N$. The rest requires a closure argument: There is $\{-\Delta v : v \in V_N\} = V_N$, and V_N is dense in L_2. $\square$

Indeed, on non-smooth non-convex domains, the H^2-regularity is not true. Take the sector of the unit-disc

$$\Omega = \{(r \cos \phi, r \sin \phi) : 0 < r < 1,\ 0 < \phi < \omega\}$$

with $\omega \in (\pi, 2\pi)$. Set $\beta = \pi/\omega < 1$. The function

$$u = (1 - r^2) r^\beta sin(\phi\beta)$$

is in H_0^1, and fulfills $\Delta u = -(4\beta + 4) r^\beta sin(\phi\beta) \in L_2$. Thus u is the solution of a Dirichlet problem. But $u \notin H^2$.

On non-convex domains one can specify the regularity in terms of weighted Sobolev spaces. Let Ω be a polygonal domain containing M vertices V_i. Let ω_i be the interior angle at V_i. If the vertex belongs to a non-convex corner ($\omega_i > \pi$), then choose some

$$\beta_i \in (1 - \frac{\pi}{\omega}, 1)$$

Define

$$w(x) = \prod_{\substack{\text{non-convex} \\ \text{Vertices } V_i}} |x - V_i|^{\beta_i}$$

Theorem 63. *If f is such that $wf \in L_2$. Then $f \in H^{-1}$, and the solution u of the Dirichlet problem fulfills*

$$\|wD^2u\|_{L_2} \preceq \|wf\|_{L_2}.$$

Chapter 4

Finite Element Method

Ciarlet's definition of a finite element is:

Definition 64 (Finite element)**.** *A finite element is a triple* (T, V_T, Ψ_T)*, where*

1. T *is a bounded set*
2. V_T *is function space on* T *of finite dimension* N_T
3. $\Psi_T = \{\psi_T^1, \ldots, \psi_T^{N_T}\}$ *is a set of linearly independent functionals on* V_T*.*

The nodal basis $\{\varphi_T^1 \ldots \varphi_T^{N_T}\}$ for V_T is the basis dual to Ψ_T, i.e.,

$$\psi_T^i(\varphi_T^j) = \delta_{ij}$$

Barycentric coordinates are useful to express the nodal basis functions.

Finite elements with point evaluation functionals are called Lagrange finite elements, elements using also derivatives are called Hermite finite elements.

Usual function spaces on $T \subset \mathbb{R}^2$ are

$$\begin{aligned} P^p &:= \operatorname{span}\{x^i y^j : 0 \leq i,\, 0 \leq j,\, i + j \leq p\} \\ Q^p &:= \operatorname{span}\{x^i y^j : 0 \leq i \leq p,\, 0 \leq j \leq p\} \end{aligned}$$

Examples for finite elements are

- A linear line segment
- A quadratic line segment
- A Hermite line segment
- A constant triangle
- A linear triangle
- A non-conforming triangle

- A Morley triangle
- A Raviart-Thomas triangle

The local nodal interpolation operator defined for functions $v \in C^m(\overline{T})$ is

$$I_T v := \sum_{\alpha=1}^{N_T} \psi_T^\alpha(v)\varphi_T^\alpha$$

It is a projection.

Two finite elements (T, V_T, Ψ_T) and $(\widehat{T}, V_{\widehat{T}}, \Psi_{\widehat{T}})$ are called *equivalent* if there exists an invertible function F such that

- $T = F(\widehat{T})$
- $V_T = \{\hat{v} \circ F^{-1} : \hat{v} \in V_{\widehat{T}}\}$
- $\Psi_T = \{\psi_i^T : V_T \to \mathbb{R} : v \to \psi_i^{\hat{T}}(v \circ F)\}$

Two elements are called affine equivalent, if F is an affine-linear function.

Lagrangian finite elements defined above are equivalent. The Hermite elements are not equivalent.

Two finite elements are called interpolation equivalent if there holds

$$I_T(v) \circ F = I_{\widehat{T}}(v \circ F)$$

Lemma 65. *Equivalent elements are interpolation equivalent*

The Hermite elements define above are also interpolation equivalent.

A regular triangulation $\mathcal{T} = \{T_1, \ldots, T_M\}$ of a domain Ω is the subdivision of a domain Ω into closed triangles T_i such that $\overline{\Omega} = \cup T_i$ and $T_i \cap T_j$ is

- either empty
- or an common edge of T_i and T_j
- or $T_i = T_j$ in the case $i = j$.

In a wider sense, a triangulation may consist of different element shapes such as segments, triangles, quadrilaterals, tetrahedra, hexhedra, prisms, pyramids.

A finite element complex $\{(T, V_T, \Psi_T)\}$ is a set of finite elements defined on the geometric elements of the triangulation $\mathcal{T}$.

It is convenient to construct finite element complexes such that all its finite elements are affine equivalent to one *reference finite element* $(\widehat{T}, \hat{V}_T, \hat{\Psi}_T)$. The transformation F_T is such that $T = F_T(\widehat{T})$.

Examples: linear reference line segment on $(0, 1)$.

The finite element complex allows the definition of the global interpolation operator for C^m-smooth functions by

$$I_{\mathcal{T}} v_{|T} = I_T v_T \qquad \forall\, T \in \mathcal{T}$$

The finite element space is

$$V_{\mathcal{T}} := \{v = I_{\mathcal{T}} w : w \in C^m(\overline{\Omega})\}$$

We say that $V_{\mathcal{T}}$ has regularity r if $V_{\mathcal{T}} \subset C^r$. If $V_{\mathcal{T}} \neq C^0$, the regularity is defined as -1.
Examples:

- The P^1 - triangle with vertex nodes leads to regularity 0.
- The P^1 - triangle with edge midpoint nodes leads to regularity -1.
- The P^0 - triangle leads to regularity -1.

For smooth functions, functionals $\psi_{T,\alpha}$ and $\psi_{\widetilde{T},\tilde{\alpha}}$ sitting in the same location are equivalent. The set of global functionals $\Psi = \{\psi_1, \ldots, \psi_N\}$ is the linearly independent set of functionals containing all (equivalence classes of) local functionals.

The connectivity matrix $C_T \in \mathbb{R}^{N \times N_T}$ is defined such that the local functionals are derived from the global ones by

$$\Psi_T(u) = C_T^t \Psi(u)$$

Examples in 1D and 2D

The nodal basis for the global finite element space is the basis in $V_{\mathcal{T}}$ dual to the global functionals ψ_j, i.e.,

$$\psi_j(\varphi_i) = \delta_{ij}$$

There holds

$$\begin{aligned}
\varphi_i|_T &= I_T \varphi_i = \sum_{\alpha=1}^{N_T} \psi_T^\alpha(\varphi_i) \varphi_T^\alpha \\
&= \sum_{\alpha=1}^{N_T} (C_T^t \psi(\varphi_i))_\alpha \varphi_T^\alpha \\
&= \sum_{\alpha=1}^{N_T} (C_T^t e_i)_\alpha \varphi_T^\alpha = \sum_{\alpha=1}^{N_T} C_{T,i\alpha} \varphi_T^\alpha
\end{aligned}$$

4.1 Finite element system assembling

As a first step, we assume there are no Dirichlet boundary conditions. The finite element problem is

$$\text{Find } u_h \in V_{\mathcal{T}} \text{ such that } : A(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in V_{\mathcal{T}} \tag{4.1}$$

The nodal basis and the dual functionals provides the one to one relation between $\mathbb{R}^N$ and $V_{\mathcal{T}}$:

$$\mathbb{R}^N \ni \underline{u} \leftrightarrow u_h \in V_{\mathcal{T}} \qquad \text{with} \qquad u_h = \sum_{i=1}^{N} \varphi_i \underline{u}_i \qquad \text{and} \qquad \underline{u}_i = \psi_i(u_h).$$

Using the nodal basis expansion of u_h in (4.1), and testing only with the set of basis functions, one has

$$A(\sum_{i=1}^{N} u_i \varphi_i, \varphi_j) = f(\varphi_j) \qquad \forall\, j = 1 \ldots N$$

With

$$A_{ji} = A(\varphi_i, \varphi_j) \qquad \text{and} \qquad \underline{f}_j = f(\varphi_j),$$

one obtains the linear system of equations

$$A\underline{u} = \underline{f}$$

The preferred way to compute the matrix A and vector f is a sum over element contributions. The restrictions of the bilinear and linear form to the elements are

$$A_T(u, v) = \int_T \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega \cap T} \alpha u v \, ds$$

and

$$f_T(v) = \int_T f v \, dx + \int_{\partial\Omega \cap T} g v \, ds$$

Then

$$A(u, v) = \sum_{T \in \mathcal{T}} A_T(u, v) \qquad f(v) = \sum_{T \in \mathcal{T}} f_T(v)$$

On each element, one defines the $N_T \times N_T$ **element matrix** and **element vector** in terms of the local basis on T:

$$A_{T,\alpha\beta} = A_T(\varphi_\beta^T, \varphi_\alpha^T) \qquad\qquad \underline{f}_{T,\alpha} = \underline{f}_T(\varphi_\alpha^T)$$

Then, the global matrix and the global vector are

$$A = \sum_{T \in \mathcal{T}} C_T A_T C_T^t$$

and

$$\underline{f} = \sum_{T \in \mathcal{T}} C_T \underline{f}_T$$

Namely,

$$\begin{aligned} \underline{f}_i &= f(\varphi_i) = \sum_{T \in \mathcal{T}} f_T(\varphi_i|_T) = \sum_{T \in \mathcal{T}} f_T(\sum_\alpha C_{T,i\alpha} \varphi_T^\alpha) \\ &= \sum_{T \in \mathcal{T}} \sum_\alpha C_{T,i\alpha} f_T(\varphi_T^\alpha) = \sum_{T \in \mathcal{T}} \sum_\alpha C_{T,i\alpha} \underline{f}_\alpha \end{aligned}$$

and

$$\begin{aligned}A_{ji} &= \sum_{T\in\mathcal{T}} A(\varphi_i|_T, \varphi_j|_T) = \sum_{T\in\mathcal{T}} A(\sum_\alpha C_{T,i\alpha}\varphi_T^\alpha, \sum_\beta C_{T,j\beta}\varphi_T^\beta) \\ &= \sum_{T\in\mathcal{T}}\sum_\alpha\sum_\beta C_{T,i\alpha}A_{T,\alpha\beta}C_{T,j\beta}\end{aligned}$$

On the elements T, the integrands are smooth functions. Thus, numerical integration rules can be applied.

In the case of Dirichlet boundary conditions, let $\gamma_D \subset \{1,\ldots,N\}$ correspond to the vertices x_i at the Dirichlet boundary, and $\gamma_f = \{1,\ldots N\} \setminus \gamma_D$.

We have the equations

$$\sum_{i\in\gamma_D} A_{ji}u_i + \sum_{i\in\gamma_f} A_{ji}u_i = f_j \qquad \forall\, j \in \gamma_f$$

Inserting $u_i = u_D(x_D)$ for $i \in \gamma_i$ results in the reduced system

$$\sum_{i\in\gamma_f} A_{ji}u_i = f_j - \sum_{i\in\gamma_D} A_{ji}u_D(x_i)$$

An alternative approach is to approximate Dirichlet boundary conditions by Robin b.c., $\frac{\partial u}{\partial n} + \alpha u = \alpha u_D$, with large parameter α.

4.2 Finite element error analysis

Let u be the solution of the variational problem, and u_h its Galerkin approximation in the finite element sub-space V_h. Cea's Lemma bounds the finite element error $u - u_h$ by the best approximation error

$$\|u - u_h\|_V \leq C \inf_{v_h\in V_h} \|u - v_h\|_V.$$

The constant factor C is the ratio of the continuity bound and the coercivity bound of the bilinear form $A(.,.)$.

Provided that the solution u is sufficiently smooth, we can take the finite element interpolant to bound the best approximation error:

$$\inf_{v\in V_h} \|u - v_h\|_V \leq \|u - I_{\mathcal{T}}u\|_V$$

In the following, we will bound the interpolation error.

Lemma 66. *Let $\widehat{T}$ and T be d-dimensional domains related by the invertible affine linear transformation $F_T : \widehat{T} \to T$*

$$F_T(x) = a + Bx,$$

where $a \in \mathbb{R}^d$ and B is a regular matrix in $\mathbb{R}^{d \times d}$. Then there holds:

$$\|u \circ F_T\|_{L_2(\widehat{T})} = (\det B)^{-1/2} \|u\|_{L_2(T)} \tag{4.2}$$

$$\frac{\partial}{\partial x_{i_m}} \cdots \frac{\partial}{\partial x_{i_1}} (u \circ F_T) = \sum_{j_m=1}^{d} \cdots \sum_{j_1=1}^{d} \left(\frac{\partial}{\partial x_{j_m}} \cdots \frac{\partial}{\partial x_{j_1}} u \right) \circ F_T \;\; B_{j_m,i_m} \ldots B_{j_1,i_1} \tag{4.3}$$

$$|u \circ F|_{H^m(\widehat{T})} \preceq (\det B)^{-1/2} \|B\|^m \, |u|_{H^m(T)} \tag{4.4}$$

Proof: Transformation of integrals, chain rule. □

We define the diameter of the element T

$$h_T = \operatorname{diam} T$$

A triangulation is called **shape regular**, if all its elements fulfill

$$|T| \succeq h_T^2$$

with a "good" constant ~ 1. If one studies convergence, one considers families of triangulations with decreasing element sizes h_T. In that case, the family of triangulations is called shape regular, if there is a common constant C such that all elements of all triangulations fulfill $|T| \geq C h_T^2$.

Lemma 67. *Let $F_T = a + Bx$ be the mapping from the reference triangle to the triangle T. Let $|T| \succeq h_T^2$. Then there holds*

$$\begin{aligned} \|B_T\| &\simeq h_T \\ \|B_T^{-1}\| &\simeq h_T^{-1} \end{aligned}$$

The following lemma is the basis for the error estimate. This lemma is the main application for the Bramble Hilbert lemma. Sometimes, it is called the Bramble Hilbert lemma itself:

Lemma 68. *Let (T, V_T, Ψ_T) be a finite element such that the element space V_T contains polynomials up to order P^k. Then there holds*

$$\|v - I_T v\|_{H^1} \leq C |v|_{H^m} \qquad \forall\, v \in H^m(T)$$

for all $m > d/2$, $m \geq 1$, and $m \leq k + 1$.

Proof: First, we prove that $id - I_T$ is a bounded operator from H^m to H^1:

$$\begin{aligned} \|v - I_T v\|_{H^1} &\leq \|v\|_{H^1} + \|I_T v\|_{H^1} = \|v\|_{H^1} + \|\sum_\alpha \psi_\alpha(v)\varphi_\alpha\|_{H^1} \\ &\leq \|v\|_{H^1} + \sum_\alpha \|\varphi_\alpha\|_{H^1}|\psi_\alpha(v)| \\ &\preceq \|v\|_{H^m} \end{aligned}$$

The last step used that for H^m, with $m > d/2$, point evaluation is continuous. Now, let $v \in P^k(T)$. Since $P^k \subset V_T$, and I_T is a projection on V_T, there holds $v - I_T v = 0$. The Bramble Hilbert Lemma applied for $U = H^1$ and $L = id - I_T$ proves the result. □

To bound the finite element interpolation error, we will transform functions from the elements T to the reference element $\widehat{T}$.

Theorem 69. *Let $\mathcal{T}$ be a shape regular triangulation of Ω. Let $V_\mathcal{T}$ be a C^0-regular finite element space such that all local spaces contain P^1. Then there holds*

$$\|v - I_\mathcal{T} v\|_{L_2(\Omega)} \preceq \left\{ \sum_{T\in\mathcal{T}} h_T^4 |v|^2_{H^2(T)} \right\}^{1/2} \qquad \forall\, v \in H^2(\Omega)$$

$$|v - I_\mathcal{T} v|_{H^1(\Omega)} \preceq \left\{ \sum_{T\in\mathcal{T}} h_T^2\, |v|^2_{H^2(T)} \right\}^{1/2} \qquad \forall\, v \in H^2(\Omega)$$

Proof: We prove the H^1 estimate, the L_2 one follows the same lines. The interpolation error on each element is transformed to the interpolation error on one reference element:

$$\begin{aligned} |v - I_\mathcal{T} v|^2_{H^1(\Omega)} &= \sum_{T\in\mathcal{T}} |(id - I_T)v_T|^2_{H^1(T)} \\ &\preceq \sum_{T\in\mathcal{T}} (\det B_T)\, \|B_T^{-1}\|^2\, |(id - I_T)v_T \circ F_T|^2_{H^1(\widehat{T})} \\ &= \sum_{T\in\mathcal{T}} (\det B_T) \|B_T^{-1}\|^2 \|(id - I_{\widehat{T}})(v_T \circ F_T)\|^2_{H^1(\widehat{T})} \end{aligned}$$

On the reference element $\widehat{T}$ we apply the Bramble-Hilbert lemma. Then, we transform back to the individual elements:

$$\begin{aligned} |v - I_\mathcal{T} v|^2_{H^1(\Omega)} &\preceq \sum_{T\in\mathcal{T}} (\det B_T) \|B_T^{-1}\|^2 |v_T \circ F_T|^2_{H^2(\widehat{T})} \\ &\preceq \sum_{T\in\mathcal{T}} (\det B_T)\, \|B_T^{-1}\|^2\, (\det B_T^{-1})\, \|B_T\|^4\, |v_T|^2_{H^2(T)} \\ &\simeq \sum_{T\in\mathcal{T}} h_T^2 \|v\|^2_{H^2(T)}. \end{aligned}$$

□

A triangulation is called **quasi − uniform**, if all elements are essentially of the same size, i.e., there exists one global h such that

$$h \simeq h_T \qquad \forall T \in \mathcal{T}.$$

On a quasi-uniform mesh, there hold the interpolation error estimates

$$\begin{aligned} \|u - I_{\mathcal{T}} u\|_{L_2(\Omega)} &\preceq h^2 \, |u|_{H^2} \\ |u - I_{\mathcal{T}} u|_{H^1(\Omega)} &\preceq h \, |u|_{H^2} \end{aligned}$$

We are interested in the rate of the error in terms of the mesh-size h.

Theorem 70 (Finite element error estimate)**.** *Assume that*

- *the solution u of the weak bvp is in H^2,*
- *the triangulation $\mathcal{T}$ is quasi-uniform of mesh-size h,*
- *the element spaces contain P^1.*

Then, the finite element error is bounded by

$$\|u - u_h\|_{H^1} \preceq h \, |u|_{H^2}$$

Error estimates in L_2-norm

The above theorem bounds the error in the L_2-norm of the function, and the L_2-norm of the derivatives with the same rate in terms of h. This is obtained by the natural norm of the variational formulation.

The interpolation error suggests a faster convergence in the weaker norm L_2. Under certain circumstances, the finite element error measured in L_2 also decays faster. The considered variational problem is

$$\text{Find } u \in V : A(u, v) = f(v) \qquad \forall v \in V.$$

We define the *dual problem* as

$$\text{Find } w \in V : A(v, w) = f(v) \qquad \forall v \in V.$$

In the case of a symmetric bilinear form, the primal and the dual problem coincide.

Theorem 71 (Aubin-Nitsche)**.** *Assume that*

- *the dual weak bvp is H^2 regular*
- *the triangulation $\mathcal{T}$ is quasi-uniform of mesh-size h,*

- *the element spaces contain P^1.*

Then, there holds the L_2-error estimate

$$\|u - u_h\|_{L_2} \preceq h^2 \, |u|_{H^2}$$

Proof: Solve the dual problem with the error $u - u_h$ as right hand side:

$$\text{Find } w \in V : A(v, w) = (u - u_h, v)_{L_2} \qquad \forall \, v \in V.$$

Since the dual problem is H^2 regular, there holds $w \in H^2$, and $\|w\|_{H^2} \preceq \|u - u_h\|_{L_2}$. Choose the test function $v := u - u_h$ to obtain the squared norm

$$A(u - u_h, w) = (u - u_h, u - u_h)_{L_2}.$$

Using the Galerkin orthogonality $A(u - u_h, v_h) = 0$ for all $v_h \in V_h$, we can insert $I_{\mathcal{T}} w$:

$$\|u - u_h\|_{L_2}^2 = A(u - u_h, w - I_{\mathcal{T}} w).$$

Next we use continuity of $A(.,.)$ and the interpolation error estimates:

$$\|u - u_h\|_{L_2}^2 \preceq \|u - u_h\|_{H^1} \, \|w - I_{\mathcal{T}} w\|_{H^1} \preceq \|u - u_h\|_{H^1} \, h \, |w|_{H^2}.$$

From H^2 regularity:

$$\|u - u_h\|_{L_2}^2 \preceq h \|u - u_h\|_{H^1} \, \|u - u_h\|_{L_2},$$

and, after dividing one factor

$$\|u - u_h\|_{L_2} \preceq h \|u - u_h\|_{H^1} \preceq h^2 |u|_{H^2}.$$

$\square$

Approximation of Dirichlet boundary conditions

Till now, we have neglected Dirichlet boundary conditions. In this case, the continuous problem is

$$\text{Find } u \in V_D : \qquad A(u, v) = f(v) \qquad \forall \, v \in V_0,$$

where

$$V_D = \{ v \in H^1 : \operatorname{tr}_{\Gamma_D} v = u_D \} \qquad \text{and} \qquad V_0 = \{ v \in H^1 : \operatorname{tr}_{\Gamma_D} v = 0 \}.$$

The finite element problem is

$$\text{Find } u_h \in V_{hD} : \qquad A(u_h, v_h) = f(v_h) \qquad \forall \, v_h \in V_{h0},$$

where

$$V_{hD} = \{ I_{\mathcal{T}} v : v \in V_D \} \qquad \text{and} \qquad V_{h0} = \{ I_{\mathcal{T}} v : v \in V_0 \}.$$

The definition of V_{hD} coincides with $\{ v_h \in V_h : v_h(x_i) = u_D(x_i) \;\; \forall \text{ vertices } x_i \text{ on } \Gamma_D \}$.

There holds $V_{h0} \subset V_0$, but, in general, there does not hold $V_{hD} \subset V_D$.

Theorem 72 (Error estimate for Dirichlet boundary conditions)**.** *Assume that*

- $A(.,.)$ *is coercive on* V_{h0}*:*

$$A(v_h, v_h) \geq \alpha_1 \|v_h\|_V^2 \qquad \forall\, v_h \in V_{h0}$$

- $A(.,.)$ *is continuous on* V*:*

$$A(u, v) \leq \alpha_2 \|u\|_V \|v\|_V \qquad \forall\, u, v \in V$$

Then there holds the finite element error estimate

$$\|u - u_h\|_{H^1} \preceq h|u|_{H^2}$$

Proof: To make use of the coercivity of $A(.,.)$, we need an element in V_{h0}. There holds Galerkin orthogonality $A(u - u_h, v_h) = 0\ \forall\, v_h \in V_{h0}$:

$$\begin{aligned}
\|u - u_h\|_V^2 &= \|u - I_h u + I_h u - u_h\|_V^2 \leq 2\,\|u - I_h u\|_V^2 + 2\,\|I_h u - u_h\|_V^2 \\
&\leq 2\,\|u - I_h u\|_V^2 + \frac{2}{\alpha_1} A(I_h u - u_h, I_h u - u_h) \\
&\leq 2\|u - I_h u\|_V^2 + \frac{2}{\alpha_1} A(I_h u - u, I_h u - u_h) + \frac{2}{\alpha_1} A(u - u_h, I_h u - u_h) \\
&\leq 2\,\|u - I_h u\|_V^2 + \frac{2\alpha_2}{\alpha_1}\,\|I_h u - u\|\|I_h u - u_h\| + 0 \\
&\leq 2\,\|u - I_h u\|_V^2 + \frac{2\alpha_2}{\alpha_1}\,\|I_h u - u\|(\|I_h u - u\| + \|u - u_h\|) \\
&= (2 + \frac{2\alpha_2}{\alpha_1})\,\|u - I_h u\|_V^2 + \frac{2\alpha_2}{\alpha_1}\,\|u - I_h u\|_V \|u - u_h\|_V
\end{aligned}$$

Next, we apply $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ for $a = \frac{2\alpha_2}{\alpha_1}\|u - I_h u\|_V$ and $b = \|u - u_h\|_V$:

$$\|u - u_h\|_V^2 \leq (2 + \frac{2\alpha_2}{\alpha_1})\|u - I_h u\|_V^2 + 2\frac{\alpha_2^2}{\alpha_1^2}\|u - I_h u\|_V^2 + \frac{1}{2}\,\|u - u_h\|_V^2$$

Moving the term $\frac{1}{2}\|u - u_h\|$ to the left, we obtain

$$\|u - u_h\|_V^2 \preceq \|u - I_h u\|_V^2 \preceq h\,|u|_{H^2}$$

□

High order elements

One can obtain faster convergence, if the solution is smooth, and elements of higher order are used:

Theorem 73. *Assume that*

- *the solution is smooth:* $u \in H^m$ *for* $m \geq 2$
- *all element spaces* V_T *contain polynomials* P^p *for* $p \geq 1$
- *the mesh is quasi-uniform*

Then there holds

$$h^{-1}\|u - I_h u\|_{L_2} + \|u - I_h u\|_{H^1} \preceq h^{\min\{m-1,p\}}\|u\|_{H^m}$$

The proof is analogous to the case $m = 2$ and $p = 1$. The constants in the estimates depend on the Sobolev index m and on the polynomial order p. Nodal interpolation is instable (i.e., the constant grow with p) for increasing order p. There exist better choices to bound the best approximation error.

Graded meshes around vertex singularities

On non-convex meshes domains, the solution is in general not in H^2, but in some weighted Sobolev space. The information of the weight can be used to construct proper locally refined meshes.

On a sector domain with a non-convex corner of angle $\omega > \pi$, the solution is bounded in the weighted Sobolev norm

$$\|r^\beta D^2 u\|_{L_2} \leq C,$$

with $\beta = \frac{\pi}{\omega}$. One may choose a mesh such that

$$h_T \simeq \underline{h} r_T^\beta, \qquad \forall T \in \mathcal{T}$$

where r_T is the distance of the center of the element to the singular corner, and $\underline{h} \in \mathbb{R}^+$ is a global mesh size parameter.

We bound the interpolation error:

$$\begin{aligned}\|u - I_{\mathcal{T}} u\|_{H^1}^2 &\preceq \sum_{T\in\mathcal{T}} h_T^2 |u|_{H^2(T)} \simeq \sum_{T\in\mathcal{T}} \underline{h}^2\, |r^\beta D^2 u|_{L_2(T)} \\ &\simeq \underline{h}^2\, \|r^\beta D^2 u\|_{L_2(\Omega)}^2 \preceq C\, \underline{h}^2\end{aligned}$$

The number of elements in the domain can be roughly estimated by the integral over the density of elements. The density is number of elements per unit volume, i.e., the inverse of the area of the element:

$$N_{el} \simeq \int_\Omega |T|^{-1}\, dx = \int_\Omega \underline{h}^{-2} r^{-2\beta}\, dx = \underline{h}^{-2} \int r^{-2\beta}\, dx \simeq C\underline{h}^{-2}$$

In two dimensions, and $\beta \in (0, 1)$, the integral is finite.

Combining the two estimates, one obtains a relation between the error and the number of elements:

$$\|u - I_{\mathcal{T}} u\|_V^2 \preceq N_{el}^{-1}$$

This is the same order of convergence as in the H^2 regular case !

4.3 A posteriori error estimates

We will derive methods to estimate the error of the computed finite element approximation. Such *a posteriori* error estimates may use the finite element solution u_h, and input data such as the source term f.

$$\eta(u_h, f)$$

An error estimator is called *reliable*, if it is an upper bound for the error, i.e., there exists a constant C_1 such that

$$\|u - u_h\|_V \leq C_1\, \eta(u_h, f) \tag{4.5}$$

An error estimator is *efficient*, if it is a lower bound for the error, i.e., there exists a constant C_2 such that

$$\|u - u_h\|_V \geq C_2\, \eta(u_h, f). \tag{4.6}$$

The constants may depend on the domain, and the shape of the triangles, but may not depend on the source term f, or the (unknown) solution u.

One use of the a posteriori error estimator is to know the accuracy of the finite element approximation. A second one is to guide the construction of a new mesh to improve the accuracy of a new finite element approximation.

The usual error estimators are defined as sum over element contributions:

$$\eta^2(u_h, f) = \sum_{T \in \mathcal{T}} \eta_T^2(u_h, f)$$

The local contributions should correspond to the local error. For the common error estimators there hold the local efficiency estimates

$$\|u - u_h\|_{H^1(\omega_T)} \geq C_2\, \eta_T(u_h, f).$$

The patch ω_T contains T and all its neighbor elements.

In the following, we consider the Poisson equation $-\Delta u = f$ with homogenous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. We choose piecewise linear finite elements on triangles.

The Zienkiewicz Zhu error estimator

The simplest a posteriori error estimator is the one by Zienkiewicz and Zhu, the so called ZZ error estimator.

The error is measured in the H^1-semi norm:

$$\|\nabla u - \nabla u_h\|_{L_2}$$

Define the gradient $p = \nabla u$ and the discrete gradient $p_h = \nabla u_h$. The discrete gradient p_h is a constant on each element. Let $\tilde{p}_h$ be the p.w. linear and continuous finite element function obtained by averaging the element values of p_h in the vertices:

$$\tilde{p}_h(x_i) = \frac{1}{|\{T : x_i \in T\}|} \sum_{T : x_i \in T} p_{h|T} \qquad \text{for all vertices } x_i$$

The hope is that the averaged gradient is a much better approximation to the true gradient, i.e.,

$$\|p - \tilde{p}_h\|_{L_2} \leq \alpha \, \|p - p_h\|_{L_2} \tag{4.7}$$

holds with a small constant $\alpha \ll 1$. This property is known as *super-convergence*.It is indeed true on (locally) uniform meshes, and smoothness assumptions onto the source term f.

The ZZ error estimator replaces the true gradient in the error $p - p_h$ by the good approximation $\tilde{p}_h$:

$$\eta(u_h) = \|\tilde{p}_h - p_h\|_{L_2(\Omega)}$$

If the super-convergence property (4.7) is fulfilled, than the ZZ error estimator is reliable:

$$\begin{aligned} \|\nabla u - \nabla u_h\|_{L_2} &= \|p - p_h\|_{L_2} \leq \|p_h - \widetilde{p}_h\|_{L_2} + \|p - \widetilde{p}_h\|_{L_2} \\ &\leq \|p_h - \widetilde{p}_h\|_{L_2} + \alpha \|p - p_h\|_{L_2}, \end{aligned}$$

and

$$\|\nabla u - \nabla u_h\|_{L_2} \leq \frac{1}{1 - \alpha} \|p_h - \widetilde{p}_h\|_{L_2}.$$

It is also efficient, a similar short application of the triangle inequality.

There is a rigorous analysis of the ZZ error estimator, e.g., by showing equivalence to the following residual error estimator.

The residual error estimator

The idea is to compute the residual of the Poisson equation

$$f + \Delta \, u_h,$$

in the natural norm H^{-1}. The classical Δ-operator cannot be applied to u_h, since the first derivatives, ∇u_h, are non-continuous across element boundaries. One can compute the residuals on the elements

$$f_{|T} + \Delta \, u_{h|T} \qquad \forall T \in \mathcal{T},$$

and one can also compute the violation of the continuity of the gradients on the edge $E = T_1 \cap T_2$. We define the normal-jump term

$$\left[\frac{\partial u_h}{\partial n} \right] := \frac{\partial u_h}{\partial n_1} |_{T_1} + \frac{\partial u_h}{\partial n_2} |_{T_2}.$$

The residual error estimator is

$$\eta^{res}(u_h, f)^2 := \sum_T \eta_T^{res}(u_h, f)^2$$

with the element contributions

$$\eta_T^{res}(u_h, f)^2 := h_T^2 \|f + \Delta u_h\|_{L_2(T)}^2 + \sum_{\substack{E:E\subset T\\ E\subset\Omega}} h_E \left\| \left[\frac{\partial u_h}{\partial n} \right] \right\|_{L_2(E)}^2 .$$

The scaling with h_T corresponds to the natural H^{-1} norm of the residual.

To show the reliability of the residual error estimator, we need a new *quasi*-interpolation operator, the Clément- operator Π_h. In contrast to the interpolation operator, this operator is well defined for functions in L_2.

We define the vertex patch of all elements connected with the vertex x

$$\omega_x = \bigcup_{T:x\in T} T,$$

the edge patch consisting of all elements connected with the edge E

$$\omega_E = \bigcup_{T:E\cap T\neq\emptyset} T,$$

and the element patch consisting of the element T and all its neighbors

$$\omega_T = \bigcup_{T':T\cap T'\neq\emptyset} T'.$$

The nodal interpolation operator I_h was defined as

$$I_h v = \sum_{x_i\in\mathcal{V}} v(x_i)\varphi_i,$$

where φ_i are the nodal basis functions. Now, we replace the nodal value $v(x_i)$ by a local mean value.

Definition 74 (Clément quasi-interpolation operator)**.** *For each vertex x,let $\overline{v}^{\omega_x}$ be the mean value of v on the patch ω_x, i.e.,*

$$\overline{v}^{\omega_x} = \frac{1}{|\omega_x|} \int_{\omega_x} v\, dx.$$

The Clément operator is

$$\Pi_h v := \sum_{x_i\in\mathcal{V}} \overline{v}^{\omega_{x_i}} \varphi_i.$$

In the case of homogeneous Dirichlet boundary values, the sum contains only inner vertices.

Theorem 75. *The Clément operator satisfies the following continuity and approximation estimates:*

$$
\begin{aligned}
\|\nabla \Pi_h v\|_{L_2(T)} &\preceq \|\nabla v\|_{L_2(\omega_T)} \\
\|v - \Pi_h v\|_{L_2(T)} &\preceq h_T \|\nabla v\|_{L_2(\omega_T)} \\
\|v - \Pi_h v\|_{L_2(E)} &\preceq h_E^{1/2} \|\nabla v\|_{L_2(\omega_E)}
\end{aligned}
$$

Proof: First, choose a reference patch $\widehat{\omega}_T$ of dimension $\simeq 1$. The quasi-interpolation operator is bounded on $H^1(\omega_T)$:

$$
\|v - \Pi_h v\|_{L_2(\widehat{T})} + \|\nabla(v - \Pi_h v)\|_{L_2(\widehat{T})} \preceq \|v\|_{H^1(\widehat{\omega}_T)} \tag{4.8}
$$

If v is constant on ω_T, then the mean values in the vertices take the same values, and also $(\Pi_h v)_{|T}$ is the same constant. The constant function (on ω_T) is in the kernel of $\|v - \Pi_h v\|_{H^1(T)}$. Due to the Bramble-Hilbert lemma, we can replace the norm on the right hand side of (4.8) by the semi-norm:

$$
\|v - \Pi_h v\|_{L_2(\widehat{T})} + \|\nabla(v - \Pi_h v)\|_{L_2(\widehat{T})} \preceq \|\nabla v\|_{L_2(\widehat{\omega}_T)} \tag{4.9}
$$

The rest follows from scaling. Let $F : x \to hx$ scale the reference patch $\widehat{\omega}_T$ to the actual patch ω_T. Then

$$
\|v - \Pi_h v\|_{L_2(T)} + h \, \|\nabla(v - \Pi_h v)\|_{L_2(T)} \preceq h \, \|\nabla v\|_{L_2(\omega_T)}
$$

The estimate for the edge term is similar. One needs the scaling of integrals from the reference edge $\widehat{E}$ to E:

$$
\|v\|_{L_2(E)} = h_E^{1/2} \|v \circ F\|_{L_2(\hat{E})}
$$

Theorem 76. *The residual error estimator is reliable:*

$$
\|u - u_h\| \preceq \eta^{res}(u_h, f)
$$

Proof: From the coercivity of $A(.,.)$ we get

$$
\|u - u_h\|_{H^1} \preceq \frac{A(u - u_h, u - u_h)}{\|u - u_h\|_{H^1}} \leq \sup_{0 \neq v \in H^1} \frac{A(u - u_h, v)}{\|v\|_{H^1}}.
$$

The Galerkin orthogonality $A(u - u_h, v_h) = 0$ for all $v_h \in V_h$ allows to insert the Clément interpolant in the numerator. It is well defined for $v \in H^1$:

$$
\|u - u_h\|_{H^1} \leq \sup_{0 \neq v \in H^1} \frac{A(u - u_h, v - \Pi_h v)}{\|v\|_{H^1}}.
$$

We use that the true solution u fulfills $A(u,v) = f(v)$, and insert the definitions of $A(.,.)$ and $f(.)$:

$$\begin{aligned} A(u-u_h, v-\Pi_h v) &= f(v-\Pi_h v) - A(u_h, v-\Pi_h v) \\ &= \int_\Omega f(v-\Pi_h v)\,dx - \int_\Omega \nabla u_h \nabla(v-\Pi_h v)\,dx \\ &= \sum_{T\in\mathcal{T}} \int_T f(v-\Pi_h v)\,dx - \sum_{T\in\mathcal{T}} \int_T \nabla u_h \nabla(v-\Pi_h v)\,dx \end{aligned}$$

On each T, the finite element function u_h is a polynomial. This allows integration by parts on each element:

$$A(u-u_h, v-\Pi_h v) = \sum_{T\in\mathcal{T}} \int_T f(v-\Pi_h v)\,dx - \sum_{T\in\mathcal{T}} \left\{ -\int_T \Delta u_h (v-\Pi_h v)\,dx + \int_{\partial T} \frac{\partial u_h}{\partial n}(v-\Pi_h v)\,ds \right\}$$

All inner edges E have contributions from normal derivatives from their two adjacent triangles $T_{E,1}$ and $T_{E,2}$. On boundary edges, $v - \Pi_h v$ vanishes.

$$\begin{aligned} &A(u-u_h, v-\Pi_h v) \\ &\quad= \sum_T \int_T (f+\Delta u_h)(v-\Pi_h v)\,dx + \sum_E \int_E \left\{ \frac{\partial u_h}{\partial n}|_{T_{E,1}} + \frac{\partial u_h}{\partial n}|_{T_{E,2}} \right\} (v-\Pi_h v)\,ds \\ &\quad= \sum_T \int_T (f+\Delta u_h)(v-\Pi_h v)\,dx + \sum_E \int_E \left[\frac{\partial u_h}{\partial n} \right] (v-\Pi_h v)\,ds \end{aligned}$$

Applying Cauchy-Schwarz first on $L_2(T)$ and $L_2(E)$, and then in $\mathbb{R}^n$:

$$\begin{aligned} &A(u-u_h, v-\Pi_h v) \\ &\leq \sum_T \|f+\Delta u_h\|_{L_2(T)} \|v-\Pi_h v\|_{L_2(T)} + \sum_E \left\| \left[\frac{\partial u_h}{\partial n} \right] \right\|_{L_2(E)} \|v-\Pi_h v\|_{L_2(E)} \\ &= \sum_T h_T \|f+\Delta u_h\|_{L_2(T)} h_T^{-1} \|v-\Pi_h v\|_{L_2(T)} + \sum_E h_E^{1/2} \left\| \left[\frac{\partial u_h}{\partial n} \right] \right\|_{L_2(E)} h_E^{-1/2} \|v-\Pi_h v\|_{L_2(E)} \\ &\leq \left\{ \sum_T h_T^2 \|f+\Delta u_h\|^2_{L_2(T)} \right\}^{1/2} \left\{ \sum_T h_T^{-2} \|v-\Pi_h v\|^2_{L_2(T)} \right\}^{1/2} + \\ &\quad + \left\{ \sum_E h_E \left\| \left[\frac{\partial u_h}{\partial n} \right] \right\|^2_{L_2(E)} \right\}^{1/2} \left\{ \sum_E h_E^{-1} \|v-\Pi_h v\|^2_{L_2(E)} \right\}^{1/2} \end{aligned}$$

We apply the approximation estimates of the Clément operator, and use that only a bounded number of patches are overlapping:

$$\sum_T h_T^{-2} \|v-\Pi_h v\|^2_{L_2(T)} \preceq \sum_T \|\nabla v\|^2_{L_2(\omega_T)} \preceq \|\nabla v\|^2_{L_2(\Omega)},$$

and similar for the edges

$$\sum_E h_E^{-1} \|v - \Pi_h v\|^2_{L_2(E)} \leq \|\nabla v\|^2_{L_2(\Omega)}.$$

Combining the steps above we observe

$$\begin{aligned}
\|u-u_h\|_V &\preceq \sup_{v \in H^1} \frac{A(u-u_h, v - \Pi_h v)}{\|v\|_H^1} \\
&\preceq \sup_{V \in H^1} \frac{\left\{ \sum_T h_T^2 \|f + \Delta u_h\|^2_{L_2(T)} + \sum_E h_E \left\| \left[\frac{\partial u_h}{\partial n} \right] \right\|^2_{L_2(E)} \right\}^{1/2} \|\nabla v\|_{L_2(\Omega)}}{\|v\|_{H^1}} \\
&\leq \left\{ \sum_T h_T^2 \|f + \Delta u_h\|^2_{L_2(T)} + \sum_E h_E \left\| \left[\frac{\partial u_h}{\partial n} \right] \right\|^2_{L_2(E)} \right\}^{1/2},
\end{aligned}$$

what is the reliability of the error estimator $\eta^{res}(u_h, f)$

Theorem 77. *If the source term f is piecewise polynomial on the mesh, then the error estimator η^{res} is efficient:*

$$\|u - u_h\|_V \succeq \eta^{res}(u_h, f)$$

Goal driven error estimates

The above error estimators estimate the error in the energy norm V. Some applications require to compute certain values (such as point values, average values, line integrals, fluxes through surfaces, ...). These values are descibed by linear functionals $b : V \to \mathbb{R}$. We want to design a method such that the error in this goal, i.e.,

$$b(u) - b(u_h)$$

is small. The technique is to solve additionally the dual problem, where the right hand side is the goal functional:

$$\text{Find } w \in V: \qquad A(v, w) = b(v) \qquad \forall\, v \in V.$$

Usually, one cannot solve the dual problem either, and one applies a Galerkin method also for the dual problem:

$$\text{Find } w_h \in V_h: \qquad A(v_h, w_h) = b(v_h) \qquad \forall\, v_h \in V_h.$$

In the case of point values, the solution of the dual problem is the Green function (which is not in H^1). The error in the goal is

$$b(u - u_h) = A(u - u_h, w) = A(u - u_h, w - w_h).$$

A rigorous upper bound for the error in the goal is obtained by using continuity of the bilinear-form, and energy error estimates η^1 and η^2 for the primal and dual problem, respectively:

$$|b(u-u_h)| \preceq \|u-u_h\|_V \|w-w_h\|_V \preceq \eta^1(u_h,f)\,\eta^2(w_h,b).$$

A good heuristic is the following (unfortunately, not correct) estimate

$$b(u-u_h) = A(u-u_h, w-w_h) \preceq \sum_{T\in\mathcal{T}} \|u-u_h\|_{H^1(T)}\,\|w-w_h\|_{H^1(T)} \preceq \sum_T \eta_T^1(u_h,f)\,\eta_T^2(w_h,b) \tag{4.10}$$

The last step would require a local reliability estimate. But, this is not true.

We can interpret (4.10) that way: The local estimators $\eta_T^2(w_h)$ provide a way for weighting the primal local estimators according to the desired goal.

Mesh refinement algorithms

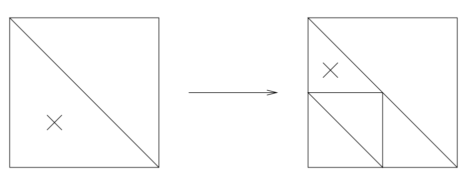
A posteriori error estimates are used to control recursive mesh refinement:

Start with initial mesh $\mathcal{T}$
Loop
  compute fe solution u_h on $\mathcal{T}$
  compute error estimator $\eta_T(u_h,f)$
  if $\eta \leq$ tolerance then stop
  refine elements with large η_T to obtain a new mesh

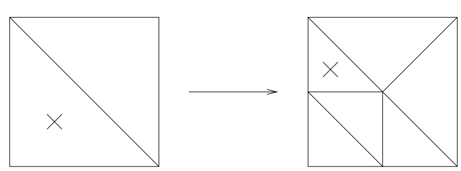
The mesh refinement algorithm has to take care of

- generating a sequence of regular meshes
- generating a sequence of shape regular meshes

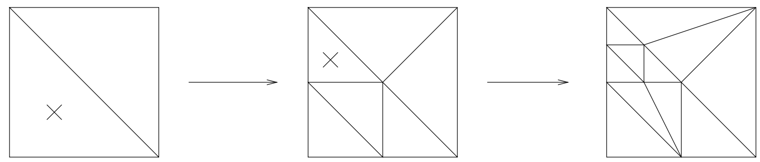
Red-Green Refinement:
A marked element is split into four equivalent elements (called red refinement):



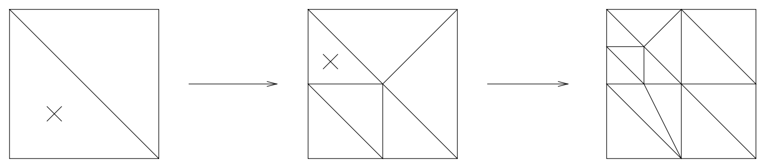
But, the obtained mesh is not regular. To avoid such irregular nodes, also neighboring elements must be split (called green closure):



If one continues to refine that way, the shape of the elements may get worse and worse:

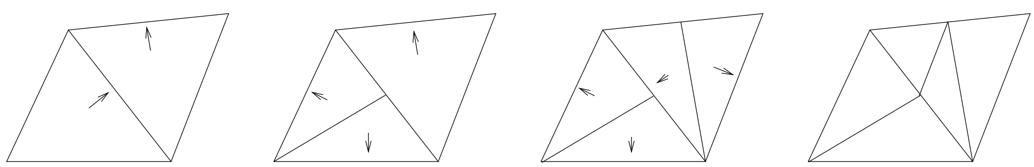


A solution is that elements of the green closure will not be further refined. Instead, remove the green closure, and replace it by red refinement.



Marked edge bisection:
Each triangle has one marked edge. The triangle is only refined by cutting from the middle of the marked edge to the opposite vertex. The marked edges of the new triangles are the edges of the old triangle.

If there occurs an irregular node, then also the neighbor triangle must be refined.



To ensure finite termination, one has to avoid cycles in the initial mesh. This can be obtained by first sorting the edges (e.g., by length), end then, always choose the largest edges as marked edge.

Both of these refinement algorithms are also possible in 3D.

4.4 Equilibrated Residual Error Estimates

4.4.1 General framework

Equilibrated residual error estimators provide upper bounds for the discretization error in energy norm without any generic constant. We consider the standard problem: find $u \in V := H_0^1(\Omega)$ such that

$$\int_\Omega \lambda \nabla u \cdot \nabla v = \int_\Omega f v \qquad \forall\, v \in V$$

The left hand side defines the bilinear-form $A(\cdot,\cdot)$, the right hand side the linear-form $f(\cdot)$. We define a finite element sub-space $V_h \subset V$ of order k, and the finite element solution

$$\text{find } u_h \in V_h : \quad A(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in V_h.$$

We assume that f is element-wise polynomial of order $k-1$, and λ is element-wise constant and positive.

The residual $r(\cdot) \in V^*$ is

$$r(v) = f(v) - A(u_h, v) \qquad v \in V$$

Since

$$\|u - u_h\|_A = \sup_{v \in V} \frac{A(u - u_h, v)}{\|v\|_A} = \sup_{v \in V} \frac{r(v)}{\|v\|_A},$$

we aim in estimating $\|r\|$ in the norm dual to $\|\cdot\|_A$, which is essentially the H^{-1}-norm. In general, the direct evaluation of this norm is not feasible. Using the structure of the problem, we can represent the residual as

$$r(v) = \sum_{T \in \mathcal{T}} \int_T r_T v + \sum_{E \in \mathcal{E}} \int_E r_E v,$$

where r_T and r_E are given given as

$$r_T = f_T + \operatorname{div} \lambda_T \nabla u_{h|T} \qquad \text{and} \qquad r_E = \left[\lambda \frac{\partial u_h}{\partial n}\right]_E$$

The element-residual r_T is a polynomial of order $k-1$ on the element T, and the edge residual (the normal jump) is a polynomial of order $k-1$ on the edge E.

The *residual error estimator* estimates the residual in terms of weighted L_2-norms:

$$\|r\|^2 \simeq \eta^{res}(u_h, f)^2 := \sum_T \frac{h_T^2}{\lambda_T} \|r_T\|^2_{L_2(T)} + \sum_E \frac{h_E}{\lambda_E} \|r_E\|^2_{L_2(E)}$$

Here, λ_E is some averaging of the coefficients on the two elements containing the edge E. The equivalence holds with constants depending on the shape of elements, the relative jump of the coefficient, and the polynomial order k.

The *equilibrated residual error estimator* η^{er} is defined in terms of the same data r_T and r_E. It satisfies

$$\begin{aligned} \|u - u_h\|_A &\leq \eta^{er} && \text{reliable with constant 1} \\ \|u - u_h\|_A &\geq c\,\eta^{er} && \text{efficient with a generic constant } c \end{aligned}$$

The lower bound depends on the shape of elements and the coefficient λ, but is robust with respect to the polynomial order k.

The main idea is the following: Instead of calculating the H^{-1}-norm of r, we compute a lifting σ^Δ such that $\operatorname{div} \sigma^\Delta = r$, and calculate the L_2-norm of σ^Δ. Since r is not a regular function, the equation must be posed in distributional form:

$$\int_\Omega \sigma^\Delta \cdot \nabla \varphi = -r(\varphi) \qquad \forall \varphi \in V$$

Then, the residual can be estimated without envolving any generic constant:

$$\begin{aligned}
\|r\|_{A^*} &= \sup_{v \in V} \frac{r(v)}{\|v\|_A} = \sup_v \frac{\int \sigma^\Delta \cdot \nabla v}{\|v\|_A} \\
&= \sup_v \frac{\int \lambda^{-1/2} \sigma^\Delta \cdot \lambda^{1/2} \nabla v}{\|v\|_A} \leq \sup_v \frac{\sqrt{\int \lambda^{-1} |\sigma^\Delta|^2} \sqrt{\int \lambda |\nabla v|^2}}{\|v\|_A} = \|\sigma^\Delta\|_{L_2, 1/\lambda}
\end{aligned}$$

The norm $\|\sigma^\Delta\| := \int \lambda^{-1} |\sigma^\Delta|^2$ can be evaluated easily.

Remark: The flux-postprocessing $\sigma := \lambda \nabla u_h + \sigma^\Delta$ provides a flux $\sigma \in H(\operatorname{div})$ such that $\operatorname{div} \sigma = f$, i.e. the flux is in exact equilibrium with the source f. Thus the name.

4.4.2 Computation of the lifting $\|\sigma^\Delta\|$

The residual is a functional of the form

$$r(v) = \sum_T (r_T, v)_{L_2(T)} + \sum_E (r_E, v)_{L_2(E)},$$

where r_T and r_E are polynomials of order $k-1$. We search for σ^Δ which is element-wise a vector-valued polynomial of order k, and not continuous across edges. Element-wise integration by parts gives

$$\int_\Omega \sigma \cdot \nabla \varphi = -\sum_T \int_T \operatorname{div} \sigma_{|T} \varphi + \sum_E \int_E [\sigma \cdot n]_E \varphi.$$

Thus $\operatorname{div} \sigma = r$ in distributional sense reads as

$$\operatorname{div} \sigma | T = r_T \qquad \text{and} \qquad [\sigma \cdot n]_E = -r_E$$

for all elements T and edges E. We could now pose the problem

$$\min_{\substack{\sigma \in P^k(\mathcal{T})^2 \\ \operatorname{div} \sigma = r}} \|\sigma\|_{L_2, 1/\lambda}$$

We minimize the weighted-L_2 norm since we want to find the smallest possible upper bound for the error. This is already a computable approach. But, the problem is global, and its solution is of comparable cost as the solution of the original finite element system. The existence of a σ such that div $\sigma = r$ also needs a proof.

We want to localize the construction of the flux. Local problems are associated with vertex-patches $\omega_V = \cup_{T : V \in T} T$. We proceed in two steps:

1. localization of the residual: $r = \sum_V r^V$
2. local liftings: find σ^V such that $\operatorname{div} \sigma^V = r^V$ on the vertex patch

Then, for $\sigma := \sum \sigma^V$ there holds $\operatorname{div} \sigma = r$

The localization is given by multiplication of the P^1 vertex basis functions (hat-functions) ϕ_V:

$$r^V(v) := r(\phi_V v)$$

Since $\sum_V \phi_V = 1$, there holds $\sum r^V(\cdot) = r(\cdot)$. The localized residual has the same structure of element and edge terms:

$$r^V(v) = \sum_{T \subset \omega_V} (r_T^V, v)_{L_2(T)} + \sum_{E \subset \omega_V} (r_E^V, v)_{L_2(E)},$$

with

$$r_T^V = \phi_V r_T \qquad \text{and} \qquad r_E^V = \phi_V r_E$$

The local residual vanishes on constants on the patch:

$$r^V(1) = r(\phi_V \, 1) = A(u - u_h, \phi_V) = 0$$

The last equality follows from the Galerkin-orthogonality.

We give an explicit construction of the lifting σ^V in terms of the Brezzi-Douglas-Marini (BDM) element. The k^{th} order BDM element on a triangle is given by $V_T = [P^k]^2$ and the degrees of freedom:

(i) $\int_E \sigma \cdot n \, q_i$ with q_i a basis for $P^k(E)$

(ii) $\int_T \operatorname{div} \sigma \, q_i$ with q_i a basis for $P^{k-1}(T) \cap L_2^0(T)$

(iii) $\int_T \sigma \cdot \operatorname{curl} q_i$ with q_i a basis for $P_0^{k+1}(T)$

Exercise: Show that these dofs are unisolvent. Count dimensions, and prove that $[\forall i : \psi_i(\sigma) = 0] \Rightarrow \sigma = 0$.

Now, we give an explicit construction of equilibrated fluxes on a vertex patch. Label elements $T_1, T_2, \ldots T_n$ in a counter-clock-wise order. Edge E_i is the common edge between triangle T_{i-1} and T_i (with identifying $T_0 = T_n$). We define σ by specifying the dofs of the BDM element:

1. Start on T_1. We set $\sigma_n = -r_{E_1}^V$ on edge E_1. On the edge on the patch-boundary we set $\sigma_n = 0$, and on E_2 we set $\sigma_n = \mathit{const}$ such that $\int_{\partial T_1} \sigma_n = \int_{T_1} r_T^V$. We use the dofs of type (ii) to specify $\int_T \operatorname{div} \sigma \, q = \int_T r_T^V q \;\forall q \in P^{k-1} \cap L_2^0(T)$. Together with get $\operatorname{div} \sigma = r_T$. Dofs of type (iii) are not needed, and set 0. There holds

$$\int_{E_2} \sigma_n = \int_{T_1} r_T^V - \int_{E_1} \sigma_n = \int_{T_1} r_{T_1}^V + \int_{E_1} r_{E_1}^V$$

2. Continue with element T_2. On edge E_2 common with T_1 set σ_n such that $[\sigma \cdot n]_{E_2} = r_{E_2}$. Otherwise, proceed as on T_1. Thus
$$\int_{E_3} \sigma_n = \int_{T_1} r^V_{T_1} + \int_{E_1} r^V_{E_1} + \int_{T_2} r^V_{T_2} + \int_{E_2} r^V_{E_2}$$

3. Continue to element T_n. Observe that on T_n:
$$\int_{E_1} \sigma_n = \sum_{i=1}^n \int_{T_i} r^V_{T_i} + \sum_{i=1}^n \int_{E_i} r^V_{E_i} = 0,$$
which follows from $r^V(1) = 0$. Thus, also $[\sigma \cdot n]_{E_1} = r^V_{E_1}$ is satisfied.

This explicit construction proves the existence of an equilibrated flux. Instead of this explicit construction, one may solve a local constrained optimization problem
$$\min_{\sigma^V : \operatorname{div} \sigma^V = r^V} \|\sigma\|_{L_2, \lambda^{-1}}$$
This applies also for 3D. Furthoer notes

- mixed boundary conditions are possible
- the efficiency for the h-FEM is shown by scaling arguments, and equivalence to the residual error estimator
- efficiency is also proven to be robust with respect to polynomial order k, examples show overestimation less than 1.5

Literature:

1. D. Braess and J. Schöberl. Equilibrated Residual Error Estimator for Maxwell's Equations. *Mathematics of Computation*, Vol 77(262), 651-672, 2008

2. D. Braess, V. Pillwein and J. Schöberl: Equilibrated Residual Error Estimates are p-Robust. Computer *Methods in Applied Mechanics and Engineering.* Vol 198, 1189-1197, 2009

4.5 Non-conforming Finite Element Methods

In a conforming finite element method, one chooses a sub-space $V_h \subset V$, and defines the finite element approximation as
$$\text{Find } u_h \in V_h : \qquad A(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in V_h$$
For reasons of simpler implementation, or even of higher accuracy, the conforming framework is often violated. Examples are:

- The finite element space V_h is not a sub-space of $V = H^m$. Examples are the non-conforming P^1 triangle, and the Morley element for approximation of H^2.
- The Dirichlet boundary conditions are interpolated in the boundary vertices.
- The curved domain is approximated by straight sided elements
- The bilinear-form and the linear-form are approximated by inexact numerical integration

The lemmas by Strang are the extension of Cea's lemma to the non-conforming setting.

The First Lemma of Strang

In the first step, let $V_h \subset V$, but the bilinear-form and the linear-form are replaced by mesh-dependent forms

$$A_h(.,.) : V_h \times V_h \to \mathbb{R}$$

and

$$f_h(.) : V_h \to \mathbb{R}.$$

We do not assume that A_h and f_h are defined on V. We assume that the bilinear-forms A_h are uniformly coercive, i.e., there exists an α_1 independent of the mesh-size such that

$$A_h(v_h, v_h) \geq \alpha_1 \|v_h\|_V^2 \qquad \forall\, v_h \in V_h$$

The finite element problem is defined as

$$\text{Find } u_h \in V_h : \qquad A_h(u_h, v_h) = f_h(v_h) \qquad \forall\, v_h \in V_h$$

Lemma 78 (First Lemma of Strang)**.** *Assume that*

- $A(.,.)$ *is continuous on* V
- $A_h(.,.)$ *is uniformly coercive*

Then there holds

$$\begin{aligned} \|u - u_h\| \quad &\preceq \quad \inf_{v_h \in V_h} \left\{ \|u - v_h\| + \sup_{w_h \in V_h} \frac{|A(v_h, w_h) - A_h(v_h, w_h)|}{\|w_h\|} \right\} \\ &\quad + \sup_{w_h \in V_h} \frac{f(w_h) - f_h(w_h)}{\|w_h\|} \end{aligned}$$

Proof: Choose an arbitrary $v_h \in V_h$, and set $w_h := u_h - v_h$. We use the uniform coercivity, and the definitions of u and u_h:

$$\begin{aligned} \alpha_1 \|u_h - v_h\|_V^2 \quad &\leq \quad A_h(u_h - v_h, u_h - v_h) = A_h(u_h - v_h, w_h) \\ &= \quad A(u - v_h, w_h) + [A(v_h, w_h) - A_h(v_h, w_h)] + [A_h(u_h, w_h) - A(u, w_h)] \\ &= \quad A(u - v_h, w_h) + [A(v_h, w_h) - A_h(v_h, w_h)] + [f_h(w_h) - f(w_h)] \end{aligned}$$

Divide by $\|u_h - v_h\| = \|w_h\|$, and use the continuity of $A(.,.)$:

$$\|u_h - v_h\| \preceq \|u - v_h\| + \frac{|A(v_h, w_h) - A_h(v_h, w_h)|}{\|w_h\|} + \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|} \tag{4.11}$$

Using the triangle inequality, the error $\|u - u_h\|$ is bounded by

$$\|u - u_h\| \leq \inf_{v_h \in V_h} \|u - v_h\| + \|v_h - u_h\|$$

The combination with (4.11) proves the result. □

Example: Lumping of the L_2 bilinear-form:
Define the H^1 - bilinear-form

$$A(u, v) = \int_\Omega \nabla u \cdot \nabla v + \int_\Omega uv\, dx,$$

and perform Galerkin discretization with P^1 triangles. The second term leads to a non-diagonal matrix. The vertex integration rule

$$\int_T v\, dx \approx \frac{|T|}{3} \sum_{\alpha=1}^{3} v(x_{T,\alpha})$$

is exact for $v \in P^1$. We apply this integration rule for the term $\int uv\, dx$:

$$A_h(u, v) = \int \nabla u \cdot \nabla v + \sum_{T \in \mathcal{T}} \frac{|T|}{3} \sum_{\alpha=1}^{3} u(x_{T,\alpha}) v(x_{T,\alpha})$$

The bilinear form is now defined only for $u, v \in V_h$. The integration is not exact, since $uv \in P^2$ on each triangle.

Inserting the nodal basis φ_i, we obtain a diagonal matrix for the second term:

$$\varphi_i(x_{T,\alpha}) \varphi_j(x_{T,\alpha}) = \begin{cases} 1 & \text{for } x_i = x_j = x_{T,\alpha} \\ 0 & \text{else} \end{cases}$$

To apply the first lemma of Strang, we have to verify the uniform coercivity

$$\sum_T \frac{|T|}{3} \sum_{\alpha=1}^{3} |v_h(x_{T,\alpha})|^2 \geq \alpha_1 \sum_T \int_T |v_h|^2\, dx \qquad \forall\, v_h \in V_h, \tag{4.12}$$

which is done by transformation to the reference element. The consistency error can be estimated by

$$\left| \int_T u_h v_h\, dx - \frac{|T|}{3} \sum_{\alpha=1}^{3} u_h(x_\alpha) v_h(x_\alpha) \right| \preceq h_T^2\, \|\nabla u_h\|_{L_2(T)}\, \|\nabla v_h\|_{L_2(T)} \tag{4.13}$$

Summation over the elements give

$$A(u_h, v_h) - A_h(u_h, v_h) \preceq h^2 \|u_h\|_{H^1(\Omega)} \, \|v_h\|_{H^1(\Omega)}$$

The first lemma of Strang proves that this modification of the bilinear-form preserves the order of the discretization error:

$$\begin{aligned}
\|u - u_h\|_{H^1} &\preceq \inf_{v_h \in V_h} \left\{ \|u - v_h\|_{H^1} + \sup_{w_h \in V_h} \frac{|A(v_h, w_h) - A_h(v_h, w_h)|}{\|w_h\|_{H^1}} \right\} \\
&\preceq \|u - I_h u\|_{H^1} + \sup_{w_h \in V_h} \frac{|A(I_h u, w_h) - A_h(I_h u, w_h)|}{\|w_h\|_{H^1}} \\
&\preceq h \, \|u\|_{H^2} + \sup_{w_h \in V_h} \frac{h^2 \, \|I_h u\|_{H^1} \|w_h\|_{H^1}}{\|w_h\|_{H^1}} \\
&\preceq h \, \|u\|_{H^2}
\end{aligned}$$

A diagonal L_2 matrix has some advantages:

- It avoids oscillations in boundary layers (exercises!)
- In explicit time integration methods for parabolic or hyperbolic problems, one has to solve linear equations with the L_2-matrix. This becomes cheap for diagonal matrices.

The Second Lemma of Strang

In the following, we will also skip the requirement $V_h \subset V$. Thus, the norm $\|.\|_V$ cannot be used on V_h, and it will be replaced by mesh-dependent norms $\|.\|_h$. These norms must be defined for $V + V_h$. As well, the mesh-dependent forms $A_h(.,.)$ and $f_h(.)$ are defined on $V + V_h$. We assume

- uniform coercivity:
$$A_h(v_h, v_h) \geq \alpha_1 \|v_h\|_h^2 \qquad \forall \, v_h \in V_h$$
- continuity:
$$A_h(u, v_h) \leq \alpha_2 \|u\|_h \|v_h\|_h \qquad \forall \, u \in V + V_h, \; \forall \, v_h \in V_h$$

The error can now be measured only in the discrete norm $\|u - u_h\|_{V_h}$.

Lemma 79. *Under the above assumptions there holds*

$$\|u - u_h\|_h \preceq \inf_{v_h \in V_h} \|u - v_h\|_h + \sup_{w_h \in V_h} \frac{|A_h(u, w_h) - f_h(w_h)|}{\|w_h\|_h} \tag{4.14}$$

Remark: The first term in (4.14) is the approximation error, the second one is called consistency error.
Proof: Let $v_h \in V_h$. Again, set $w_h = u_h - v_h$, and use the V_h-coercivity:

$$\begin{aligned}
\alpha_1 \, \|u_h - v_h\|_h^2 &\leq A_h(u_h - v_h, u_h - v_h) = A_h(u_h - v_h, w_h) \\
&= A_h(u - v_h, w_h) + [f_h(w_h) - A_h(u, w_h)]
\end{aligned}$$

Again, divide by $\|u_h - v_h\|$, and use continuity of $A_h(.,.)$:

$$\|u_h - v_h\|_h \preceq \|u - v_h\|_h + \frac{A_h(u, w_h) - f_h(w_h)}{\|w_h\|_h}$$

The rest follows from the triangle inequality. □

The non-conforming P^1 triangle

The non-conforming P^1 triangle is also called the Crouzeix-Raviart element.

The finite element space generated by the non-conforming P^1 element is

$$V_h^{nc} := \{v \in L_2 : v_{|T} \in P^1(T), \text{and } v \text{ is continuous in edge mid-points}\}$$

The functions in V_h^{nc} are not continuous across edges, and thus, V_h^{nc} is not a sub-space of H^1. We have to extend the bilinear-form and the norm in the following way:

$$A_h(u, v) = \sum_{T \in \mathcal{T}} \int_T \nabla u \nabla v \, dx \qquad \forall \, u, v \in V + V_h^{nc}$$

and

$$\|v\|_h^2 := \sum_{T \in \mathcal{T}} \|\nabla v\|_{L_2(T)}^2 \qquad \forall \, v \in V + V_h^{nc}$$

We consider the Dirichlet-problem with $u = 0$ on Γ_D.

We will apply the second lemma of Strang.

The continuous P^1 finite element space V_h^c is a sub-space of V_h^{nc}. Let $I_h : H^2 \to V_h^c$ be the nodal interpolation operator.

To bound the approximation term in (4.14), we use the inclusion $V_h^c \subset V_h^{nc}$:

$$\inf_{v_h \in V_h^{nc}} \|u - v_h\|_h \leq \|u - I_h u\|_{H^1} \preceq h \, \|u\|_{H^2}$$

We have to bound the consistency term

$$\begin{aligned}
r(w_h) &= A_h(u, w_h) - f(w_h) \\
&= \sum_T \int_T \nabla u \nabla w_h - \sum_T \int_T f w_h \, dx \\
&= \sum_T \int_{\partial T} \frac{\partial u}{\partial n} w_h \, ds - \sum_T \int_T (\Delta u + f) \, w_h \, ds \\
&= \sum_T \int_{\partial T} \frac{\partial u}{\partial n} w_h \, ds
\end{aligned}$$

Let E be an edge of the triangle T. Define the mean value $\overline{w_h}^E$. If E is an inner edge, then the mean value on the corresponding edge of the neighbor element is the same. The normal derivative $\frac{\partial u}{\partial n}$ on the neighbor element is (up to the sign) the same. If E is an edge on the Dirichlet boundary, then the mean value is 0. This allows to subtract edge mean values:

$$r(w_h) = \sum_T \sum_{E \subset T} \int_E \frac{\partial u}{\partial n}(w_h - \overline{w_h}^E)\, ds$$

Since $\int_E w_h - \overline{w_h}^E\, ds = 0$, we may insert the constant function $\frac{\partial I_h u}{\partial n}$ on each edge:

$$r(w_h) = \sum_T \sum_{E \subset T} \int_E \left(\frac{\partial u}{\partial n} - \frac{\partial I_h u}{\partial n} \right)(w_h - \overline{w_h}^E)\, ds$$

Apply Cauchy-Schwarz on $L_2(E)$:

$$r(w_h) = \sum_T \sum_{E \subset T} \|\nabla(u - I_h u)\|_{L_2(E)} \|w_h - \overline{w_h}^E\|_{L_2(E)}$$

To estimate these terms, we transform to the reference element $\widehat{T}$, where we apply the Bramble Hilbert lemma. Let $T = F_T(\widehat{T})$, and set

$$\widehat{u} = u \circ F_T \qquad \widehat{w}_h = w_h \circ F_T$$

There hold the scaling estimates

$$\begin{aligned}
|w_h|_{H^1(T)} &\simeq |\widehat{w}_h|_{H^1(\widehat{T})} \\
\|w_h - \overline{w_h}^E\|_{L_2(E)} &\simeq h_E^{1/2} \|\widehat{w}_h - \overline{\widehat{w}_h}^{\widehat{E}}\|_{L_2(\widehat{E})} \\
|u|_{H^2(T)} &\simeq h_T^{-1} |\widehat{u}|_{H^2(\widehat{T})} \\
\|\nabla(u - I_h u)\|_{L_2(E)} &\simeq h_E^{-1/2} \|\nabla(\widehat{u} - \widehat{I}_h \widehat{u})\|_{L_2(E)}
\end{aligned}$$

On the reference element, we apply the Bramble Hilbert lemma, once for w_h, and once for u. The linear operator

$$L : H^1(\widehat{T}) \to L_2(\widehat{E}) : \widehat{w}_h \to \widehat{w}_h - \overline{\widehat{w}_h}^{\widehat{E}}$$

is bounded on $H^1(\widehat{T})$ (trace theorem), and $Lw = 0$ for $w \in P_0$, thus

$$\|\widehat{w}_h - \overline{\widehat{w}_h}^{\widehat{E}}\|_{L_2(\widehat{E})} \preceq |\widehat{w}_h|_{H^1(\widehat{T})}$$

Similar for the term in u: There is $\|\nabla(u - I_h u)\|_{L_2(E)} \preceq \|u\|_{H^2(T)}$, and $u - I_h u$ vanishes for $u \in P^1$.

Rescaling to the element T leads to

$$\begin{aligned}
\|w_h - \overline{w_h}^E\|_{L_2(E)} &\preceq h^{1/2}\, |w_h|_{H^1(T)} \\
\|\nabla(u - I_h u)\|_{L_2(E)} &\preceq h^{1/2}\, |u|_{H^2(T)}
\end{aligned}$$

This bounds the consistency term

$$r(w_h) \preceq \sum_T h\,|u|_{H^2(T)}|w_h|_{H^1(T)} \preceq h\,\|u\|_{H^2(\Omega)}\,\|w_h\|_h.$$

The second lemma of Strang gives the error estimate

$$\|u - u_h\| \preceq h\,\|u\|_{H^2}$$

There are several applications where the non-conforming P^1 triangle is of advantage:

- The L_2 matrix is diagonal (exercises)
- It can be used for the approximation of problems in fluid dynamics described by the Navier Stokes equations (see later).
- The finite element matrix has exactly 5 non-zero entries in each row associated with inner edges. That allows simplifications in the matrix generation code.

4.6 hp - Finite Elements

Let V_{hp} be a p-th order finite element sub-space of H^1. By scaling and Bramble-Hilbert technique one obtains the best-approxiamtion error estimate

$$\inf_{v_{hp}\in V_{hp}} \|u - v_{hp}\|_{H^1} \leq c h^{m-1}\|u\|_{H^m}$$

for $m \leq p+1$. The constant c depends on the order p. If m is fixed, we do obtain reduction of the approximation error as we increase p. Next we develop methods to obtain so called p-version error estimates

$$\inf_{v_{hp}\in V_{hp}} \|u - v_{hp}\|_{H^1} \leq c\left(\frac{h}{p}\right)^{m-1}\|u\|_{H^m},$$

where c is independent of h and p. This estimate proves also convergence of the p-version finite element method: One may fix the mesh, and increase the order p.

A detailed analyis of local H^m norms allows an optimal balance of mesh-size h and polynomial order p. This hp-version leads to exponential convergence

$$\inf_{v_{hp}\in V_{hp}} \|u - v_{hp}\|_{H^1} \leq c e^{-N^\alpha},$$

where N is the number of unknowns.

We will prove the p-version estimate, but not the hp-result.

4.6.1 Legendre Polynomials

Orthogonal polynomials are important to construct stable basis functions for the p-FEM, as well as for error estimates.

Let Π_n denote the space of polynomials up to order n. We write π_n for a generic polynomial in Π_n, with a different value any time it appears.

Definition of Legendre polynomials via Rodrigues' formula:

$$P_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2-1)^n.$$

It is a polynomial of degree n. The first few Legendre polynomials are

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \tfrac{3}{2}x^2 - \tfrac{1}{2} \end{aligned}$$

P_n is even if n is even, and P_n is odd if n is odd. Since $(x^2-1)^n = x^{2n} - n x^{2n-2} + \pi_{2n-4}$ (with proper modification for small n) we have

$$P_n(x) = \frac{1}{2^n n!} \frac{(2n)!}{n!} x^n - \frac{n}{2^n n!} \frac{(2n-2)!}{(n-2)!} x^{n-2} + \pi_{n-4} \tag{4.15}$$

Lemma 80. *There holds*

$$\int_{-1}^{1} P_n(x) P_m(x)\, dx = \frac{2}{2n+1} \delta_{n,m}. \tag{4.16}$$

Proof. W.l.o.g. let $n \leq m$. Multiple integration by parts gives

$$\begin{aligned} 2^{n+m} n! m! \int_{-1}^{1} P_n(x) P_m(x)\, dx &= \int_{-1}^{1} \frac{d^n}{dx^n}(x^2-1)^n \frac{d^m}{dx^m}(x^2-1)^n\, dx \\ &= \int_{-1}^{1} \frac{d^{n+1}}{dx^{n+1}}(x^2-1)^n \frac{d^{m-1}}{dx^{m-1}}(x^2-1)^m + \left[\frac{d^n}{dx^n}(x^2-1)^n \underbrace{\frac{d^{m-1}}{dx^{m-1}}(x^2-1)^m}_{=0 \text{ for } x \in \{-1,1\}} \right]_{-1}^{1} \\ &= \cdots \\ &= \int_{-1}^{1} \frac{d^{n+m}}{dx^{n+m}}(x^2-1)^n (x^2-1)^m\, dx \end{aligned}$$

For $n < m$, the first factor of the integrand vanishes, and we have orthogonality. For

$n = m$ this equals

$$\begin{aligned}
2^{2n}(n!)^2 \|P_n\|_{L_2}^2 &= \int_{-1}^{1} (2n)!(x^2-1)^n \, dx = (2n)! \int_{-1}^{1} (x-1)^n (x+1)^n \\
&= -(2n)! \int_{-1}^{1} \frac{n}{n+1} (x-1)^{n+1} (x+1)^{n-1} \\
&= (2n)! \int_{-1}^{1} \frac{n(n-1)}{(n+1)(n+2)} (x-1)^{n+2} (x+1)^{n-2} = \ldots \\
&= (2n)! \frac{n!}{2n(2n-1)\cdots(n+1)} \int_{-1}^{1} (x-1)^{2n} \, dx = (n!)^2 \frac{1}{2n+1} \, 2^{2n+1},
\end{aligned}$$

which proves the scaling. $\square$

Next we prove the 3-term recurrency, which can be used for efficient evaluation.

Lemma 81. *There holds*

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x). \tag{4.17}$$

Proof. Set $r(x) = (n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x)$. Using (4.15), we see that the leading coefficients cancel, and thus $r \in \Pi_{n-2}$. From Lemma 80 we get for any $q \in \Pi_{n-2}$

$$\int_{-1}^{1} r(x)q(x) \, dx = (n+1) \int_{-1}^{1} P_{n+1} \, q - (2n+1) \int_{-1}^{1} P_n \underbrace{xq}_{\in \Pi_{n-1}} + n \int_{-1}^{1} P_{n-1} \, q = 0,$$

and thus $r = 0$. $\square$

Lemma 82. *Legendre polynomials satisfy the Sturm-Liouville differential equation*

$$\frac{d}{dx}\left[(x^2-1)\frac{d}{dx}P_n(x)\right] = n(n+1)P_n(x)$$

Proof. Both sides are in Π_n. We compare leading coefficients, for this set $P_n = a_n x^n + \pi_{n-2}$ (with $a_n = \frac{1}{2^n n!} \frac{(2n)!}{n!}$).

$$\begin{aligned}
lhs &= \frac{d}{dx}\left[(x^2-1)\frac{d}{dx}\left(a_n x^n + \pi_{n-2}\right)\right] \\
&= \frac{d}{dx}\left[(x^2-1)(a_n n x^{n-1} + \pi_{n-3})\right] \\
&= \frac{d}{dx}\left[a_n n x^{n+1} + \pi_{n-1}\right] \\
&= n(n+1) a_n x^n + \pi_{n-2},
\end{aligned}$$

and we get the same leading coefficient for rhs. Furthermore, for $q \in \Pi_{n-1}$ there holds

$$\begin{aligned}
\int_{-1}^{1} lhs\, q &= -\int_{-1}^{1} (x^2-1)P_n' q'\, dx + \underbrace{\left[(x^2-1)P_n' q\right]_{-1}^{1}}_{=0} \\
&= \int_{-1}^{1} P_n \underbrace{\left((x^2-1)q'\right)'}_{\in \Pi_{n-1}}\, dx - \left[P_n (x^2-1) q'\right]_{-1}^{1} = 0,
\end{aligned}$$

and the same for the rhs. Thus the identity is proven. $\square$

Lemma 82 implies that the Legendre polynomials are also orthogonal w.r.t. $(u', v')_{L_2, 1-x^2}$, i.e.

$$\int_{-1}^{1} (1-x^2) P_n' P_m' = n(n+1) \|P_n\|_{L_2}^2 \delta_{n,m}$$

4.6.2 Error estimate of the L_2 projection

Since polynomials are dense in $L_2(-1,1)$, we get

$$u = \sum_{n=0}^{\infty} a_n P_n$$

with the generalized Fourier coefficients

$$a_n = \frac{(u, P_n)_{L_2}}{\|P_n\|_{L_2}^2},$$

and

$$\|u\|_{L_2}^2 = \sum_{n=0}^{\infty} a_n^2 \|P_n\|^2.$$

Let $P_{L_2}^{\Pi_p}$ denote the L_2-projection onto Π_p. There holds

$$P_{L_2}^{\Pi_p} u = \sum_{n=0}^{p} a_n P_n$$

The projection error is

$$\|u - P_{L_2}^{\Pi_p} u\|_{L_2}^2 = \sum_{n=p+1}^{\infty} a_n^2 \|P_n\|^2$$

Lemma 83. *The L_2-projection error satisfies*

$$\|u - P_{L_2}^{\Pi_p} u\|_{L_2(-1,1)} \leq \frac{1}{\sqrt{(p+1)(p+2)}} |u|_{H^1(-1,1)} \tag{4.18}$$

Proof. Since P_n are orthogonal also w.r.t. $(u', v')_{L_2,1-x^2}$, there holds

$$\|u'\|^2_{L_2,1-x^2} = \sum_{n\in\mathbb{N}} a_n^2 \, \|P_n'\|^2_{1-x^2},$$

provided that u is in H^1. The projection error satisfies

$$\begin{aligned}
\|u - P_{L_2}^{\Pi_p} u\|^2_{L_2} &= \sum_{n>p} a_n^2 \|P_n\|^2_{L_2} = \sum_{n>p} a_n^2 \frac{1}{n(n+1)} \|P_n'\|^2_{1-x^2} \\
&\leq \frac{1}{(p+1)(p+2)} \sum_{n>p} a_n^2 \|P_n'\|^2_{1-x^2} \leq \frac{1}{(p+1)(p+2)} \sum_{n\in\mathbb{N}} a_n^2 \|P_n'\|^2_{1-x^2} \\
&= \frac{1}{(p+1)(p+2)} \|u'\|^2_{1-x^2}
\end{aligned}$$

Finally, the result follows from

$$\int_{-1}^{1} (1-x^2)(u')^2 \, dx \leq \int_{-1}^{1} (u')^2 \, dx.$$

$\square$

Similar as in Lemma 82 on shows also

$$\frac{d^m}{dx^m}\left[(x^2-1)^m \frac{d^m}{dx^m} P_n(x)\right] = (n+m)(n+m-1)\dots(n-m+1)P_n(x)$$

for $m \leq n$, and, as in Lemma 83

$$\|u - P_{L_2}^{\Pi_p} u\|_{L_2} \leq \sqrt{\frac{(p-m+1)!}{(p+m+1)!}} |u|_{H^m}.$$

4.6.3 Orthogonal polynomials on triangles

Orthogonal polynomials on tensor product elements are simply constructed by tensorization. Orthogonal polynomials on simplicial elements are more advanced. They are based on Jacobi polynomials:

For $\alpha, \beta > -1$, Jacobi polynomials are defined by

$$P_n^{(\alpha,\beta)}(x) := \frac{(-1)^n}{2^n n!} \frac{1}{w(x)} \frac{d^n}{dx^n}\left(w(x)(1-x^2)^n\right)$$

with the weight function

$$w(x) = (1-x)^\alpha (1+x)^\beta.$$

Jacobi polynomials are orthogonal w.r.t. the weighted inner product

$$\int_{-1}^{1} w(x) P_n^{(\alpha,\beta)}(x) P_m^{(\alpha,\beta)}(x) dx = \delta_{n,m} \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\,\Gamma(n+\beta+1)}{n!\Gamma(n+\alpha+\beta+1)}.$$

Note that $P_n^{(0,0)} = P_n$.

Define the unit-triangle T with vertices $(-1,0)$, $(1,0)$ and $(0,1)$.

Lemma 84 (Dubiner basis)**.** *The functions*

$$\varphi_{i,j}(x,y) := P_i\left(\tfrac{x}{1-y}\right)(1-y)^i P_j^{(2i+1,0)}(2y-1) \qquad i+j \leq p$$

form an $L_2(T)$-orthogonal basis for $\Pi_p(T)$.

Proof. Note that $\varphi_{i,j} \in \Pi_{i+j}$. Substitution $\xi = \frac{x}{1-y}$ leads to

$$\begin{aligned}
&\int_T \varphi_{ij}(x,y)\varphi_{kl}(x,y)\,d(x,y) = \\
&= \int_0^1 \int_{-1+y}^{1-y} P_i\left(\tfrac{x}{1-y}\right)(1-y)^i P_j^{(2i+1,0)}(2y-1) P_k\left(\tfrac{x}{1-y}\right)(1-y)^k P_l^{(2k+1,0)}(2y-1)\,dxdy \\
&= \int_0^1 \int_{-1}^{1} P_i(\xi)P_k(\xi)(1-y)^{i+k+1} P_j^{(2i+1,0)}(2y-1) P_l^{(2k+1,0)}(2y-1)\,d\xi dy \\
&= \delta_{i,k}\|P_i\|_{L_2}^2 \int_0^1 (1-y)^{2i+1} P_j^{(2i+1,0)}(2y-1) P_l^{(2i+1,0)}(2y-1)\,dy \\
&= C_{ij}\delta_{i,k}\delta_{j,l}
\end{aligned}$$

□

4.6.4 Projection based interpolation

By means of the orthogonal polynomials one shows approximation error estimates of the form

$$\inf_{q \in \Pi_p(T)} \|u - q\|_{H^k(T)} \leq c p^{k-m} |u|_{H^m(T)} \qquad m \geq k,$$

with $c \neq c(p)$, easily in 1D and tensor product elements, and also on n-dimensional simplices [Braess+Schwab: Approximation on simplices with respect to weighted Sobolev norms, J. Approximation Theory 103, 329-337 (2000)].

But, an interpolation operator to an H^1-conforming finite element space has to satisfy continuity constraints across element boundaries. We show that we get the same rate of convergence under these constraints.

The 1D case

Let $I = (-1, 1)$. We define the operator $I_p : H^1(I) \to \Pi_p$ such that

$$\begin{aligned} I_p u(x) &= u(x) \qquad x \in \{-1, 1\} \qquad &(4.19) \\ \int_I (I_p u)' q' &= \int_I u' q' \qquad \forall\, q \in \Pi_{p,0}(I), \qquad &(4.20) \end{aligned}$$

where $\Pi_{p,0}(D) := \{q \in \Pi_p(D) : q = 0 \text{ on } \partial D\}$. This procedure is exactly a p-version Galerkin-method for the Dirichlet problem. Since boundary values are preserved, the interpolation operator produces a globally continuous function. The operator I_p is a kind of mixture of interpolation and projection, thus the term *projection based interpolation* introduced by Demkowicz has been established.

Lemma 85 (Commuting diagram). *There holds*

$$\Pi_{L_2}^{\Pi_{p-1}} u' = (I_p u)'$$

Proof. The range of both sides is Π_{p-1}. We have to show that $(I_p u)'$ is indeed the L_2-projection of u', i.e.

$$\int_I (I_p u)' q = \int_I u' q \qquad \forall\, q \in \Pi_{p-1}.$$

This holds since $\{q' : q \in \Pi_{p,0}\} = \{q \in \Pi_{p-1} : \int q = 0\}$ and (4.20), and

$$\int_{-1}^{1} (I_p u)'\, 1 = (I_p u)(1) - (I_p u)(-1) = u(1) - u(-1) = \int_{-1}^{1} u'\, 1.$$

$\square$

The H^1-error estimate follows directly from the commuting diagram property:

$$|u - I_p u|_{H^1(I)} = \|u' - (I_p u)'\|_{L_2} = \|u' - P_{L_2}^{\Pi_{p-1}} u'\|_{L_2} \leq \frac{c}{p^{m-1}} |u'|_{H^{m-1}}.$$

By the Aubin-Nitsche technique one obtains an extra p for the L_2-error:

$$\|u - I_p u\|_{L_2(I)} \preceq \frac{1}{p} |u - I_p u|_{H^1(I)} \leq \frac{1}{p^m} |u|_{H^m(I)}$$

One also gets for $q \in \Pi_p$

$$|u - I_p u|_{H^1} = |u - q - I_p(u - q)|_1 \leq ||Id - I_p||_{H^1 \to H^1} \, |u - q|_{H^1}$$

Projection based interpolation on triangles

We define the operator $I_p : H^2(T) \to \Pi_p(T)$ as follows:

$$\begin{aligned} I_p u(x) &= u(x) \qquad \forall \text{ vertices } x && (4.21) \\ \int_E \partial_\tau (I_p u) \partial_\tau q &= \int_E \partial_\tau u \partial_\tau q \qquad \forall \text{ edges } E, \ \forall\, q \in \Pi_{p,0}(E) && (4.22) \\ \int_T \nabla (I_p u) \nabla q &= \int_T \nabla u \nabla q \qquad \forall\, q \in \Pi_{p,0}(T) && (4.23) \end{aligned}$$

Note that $I_p u$ on the edge E depends only on $u|_E$, and thus the interpolant is continuous across neighbouring elements.

Lemma 86. *Let $v \in C(\partial T)$ such that $v|_E \in \Pi_p(E)$. Then there exists an extension $\tilde{v} \in \Pi_p(T)$ such that $\tilde{v}|_{\partial T} = v$ and*

$$|\tilde{v}|_{H^1} \leq c \, |v|_{H^{1/2}(\partial T)},$$

where c is independent of p.

Major steps have been shown in exercises 5.2 and 6.6. Note that the minimal-norm extension $\tilde{v}$ is the solution of the Dirichlet problem, i.e.

$$\int_T \nabla \tilde{v} \nabla w = 0 \qquad \forall\, w \in \Pi_{p,0}$$

Theorem 87 (error estimate)**.** *There holds*

$$|u - I_p u|_{H^1} \preceq \inf_{q \in \Pi_p} |u - q|_{H^1(T)} + \sum_{E \subset \partial T} \frac{1}{\sqrt{p}} \inf_{q \in \Pi_p(E)} |u - q|_{H^1(E)} \preceq \frac{1}{p^{m-1}} |u|_{H^m}$$

for $u \in H^m, m \geq 2$.

Proof. Let u_p be the $|\cdot|_{H^1}$ best approximation to u, i.e.

$$\int_T \nabla u_p \nabla v = \int_T \nabla u \nabla v \qquad \forall\, v \in \Pi_p,$$

and, for uniqueness, mean values are preserved: $\int_T u_p = \int_T u$. There holds

$$|u - u_p|_{H^1} \leq \frac{c}{p^{m-1}} |u|_{H^m}.$$

We apply the triangle inequality:

$$|u - I_p u|_{H^1} \leq |u - u_p|_{H^1} + |u_p - I_p u|_{H^1}$$

Since

$$\int_T \nabla u_p \nabla v = \int_T \nabla u \nabla v = \int_T \nabla I_p u \nabla v \qquad \forall\, v \in \Pi_{p,0}(T),$$

we have that

$$u_p - I_p u \perp_{H^1} \Pi_{p,0},$$

i.e. $u_p - I_p u$ is solution of the Dirichlet problem with boundary values $(u_p - I_p u)|_{\partial T}$. Lemma 86 implies that

$$|u_p - I_p u|_{H^1(T)} \preceq |u_p - I_p u|_{H^{1/2}(\partial T)}$$

We insert an u on the boundary to obtain

$$\begin{aligned}
|u - I_p u|_{H^1} &\preceq |u - u_p|_{H^1(T)} + |u_p - I_p u|_{H^{1/2}(\partial T)} \\
&\leq |u - u_p|_{H^1(T)} + |u_p - u|_{H^{1/2}(\partial T)} + |u - I_p u|_{H^{1/2}(\partial T)} \\
&\leq |u - u_p|_{H^1(T)} + \|u - I_p u\|_{L_2(\partial T)}^{1/2} |u - I_p u|_{H^1(\partial T)}^{1/2}.
\end{aligned}$$

In the last step we used that $H^{1/2}(\partial T) = [L_2, H^1]_{1/2}$ (i.e. the interpolation space). Next, we observe that I_p restricted to one edge E is exactly the 1D operator. Using Aubin-Nitsche we get

$$\begin{aligned}
|u - I_p u|_{H^1} &\preceq |u - u_p|_{H^1(T)} + p^{-1/2} \|u - I_p u\|_{H^1(\partial T)} \\
&\preceq |u - u_p|_{H^1(T)} + \sum_E p^{1-m} \, |u|_{H^{m-1/2}(E)} \\
&\preceq p^{1-m} |u|_{H^m(T)}
\end{aligned}$$

In the last step we used the trace theorem. □

Chapter 5

Linear Equation Solvers

The finite element method, or other discretization schemes, lead to linear systems of equations

$$Au = f.$$

The matrices are typically

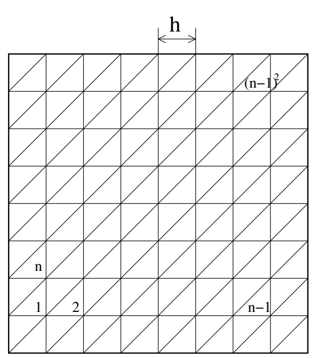
- of large dimension N ($10^4 - 10^8$ unknowns)
- and sparse, i.e., there are only a few non-zero elements per row.

A matrix entry A_{ij} is non-zero, if there exists a finite element connected with both degrees of freedom i and j.

A 1D model problem: Dirichlet problem on the interval. A uniform grid with n elements. The matrix is

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}_{(n-1)\times(n-1)},$$

A 2D model problem: Dirichlet problem on a unit-square. A uniform grid with $2n^2$ triangles. The unknowns are enumerated lexicographically:



The FEM - matrix of dimension $N = (n-1)^2$ is

$$A = \begin{pmatrix} D & -I & & & \\ -I & D & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & D & -I \\ & & & -I & D \end{pmatrix} \qquad \text{with} \qquad D = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}_{(n-1)\times(n-1)},$$

and the $(n-1) \times (n-1)$ identity matrix I.

5.1 Direct linear equation solvers

Direct solvers are factorization methods such as LU-decomposition, or Cholesky factorization. They require in general $O(N^3) = O(n^6)$ operations, and $O(N^2) = O(n^4)$ memory. A fast machine can perform about 10^9 operations per second[1]. This corresponds to

n	∼N	time	memory
10	10^2	1 ms	80 kB
100	10^4	16 min	800 MB
1000	10^6	30 years	8 TB

A band-matrix of (one-sided) band-width b is a matrix with

$$A_{ij} = 0 \qquad \text{for } |i-j| > b$$

The LU-factorization maintains the band-width. L and U are triangular factors of band-width b. A banded factorization method costs $O(Nb^2)$ operations, and $O(Nb)$ memory. For the 1D example, the band-with is 1. Time and memory are $O(n)$. For the 2D example, the band width is $O(n)$. The time complexity is $O(n^4)$, the memory complexity is $O(n^3)$.

This corresponds to

n	time	memory
10	10 μs	8 kB
100	0.1 s	8 MB
1000	16 min	8 GB

[1] time of writing was 2003

Block-elimination methods

By splitting the unknowns into two groups, we rewrite the equation $Au = f$ as a block system

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

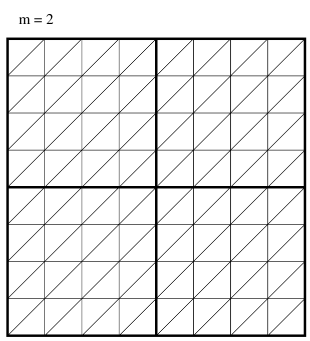
First, expressing u_1 from the first row gives

$$u_1 = A_{11}^{-1}(f_1 - A_{12}u_2),$$

and the Schur-complement equation to determine u_2

$$(\underbrace{A_{22} - A_{21}A_{11}^{-1}A_{12}}_{=:S})u_2 = f_2 - A_{21}A_{11}^{-1}f_1.$$

This block-factorization is used in sub-structuring algorithms: Decompose the domain into $m \times m$ sub-domains, each one containing $2\frac{n}{m} \times \frac{n}{m}$ triangles. Split the unknowns into interior (I), and coupling (C) unknowns.



The interior ones corresponding to different sub-domains have no connection in the matrix. The block matrix is

$$\begin{pmatrix} A_I & A_{IC} \\ A_{CI} & A_C \end{pmatrix} = \begin{pmatrix} A_{I,1} & & & A_{IC,1} \\ & \ddots & & \vdots \\ & & A_{I,m^2} & A_{IC,m^2} \\ A_{CI,1} & \cdots & A_{CI,m^2} & A_C \end{pmatrix}$$

Factorizing the block-diagonal interior block A_I splits into m^2 independent factorization problems. If one uses a banded factorization, the costs are

$$m^2 \left(\frac{n}{m}\right)^4 = \frac{n^4}{m^2}$$

Computing the Schur complement

$$S = A_C - A_{CI}A_I^{-1}A_{IC} = A_C - \sum_{i=1}^{m^2} A_{CI,i}A_{I,i}^{-1}A_{IC,i}$$

is of the same cost. The Schur complement is of size mn, and has band-width n. Thus, the factorization costs $O(mn^3)$. The total costs are of order

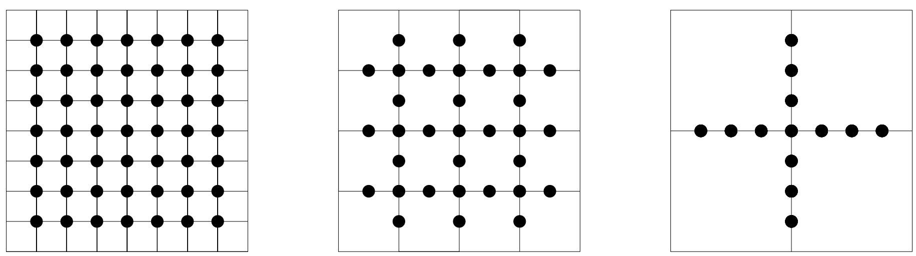
$$\frac{n^4}{m^2} + mn^3$$

Equilibrating both terms lead to the optimal number of sub-domains $m = n^{1/3}$, and to the asymptotic costs

$$n^{3.33}$$

If a parallel computer is used, the factorization of A_I and the computation of Schur complements can be performed in parallel.

The hierarchical sub-structuring algorithm, known as nested dissection, eliminates interior unknowns hierarchically:



Let $n = 2^L$. On level l, with $1 \leq l \leq L$, one has 4^l sub-domains. Each sub-domain has $O(2^{L-l})$ unknowns. The factorization of the inner blocks on level l costs

$$4^l \, (2^{L-l})^3 = 2^{3L-l}$$

Forming the Schur-complement is of the same cost. The sum over all levels is

$$\sum_{l=1}^{L} 2^{3L-l} = 2^{3L} \left(\frac{1}{2} + \frac{1}{4} + \ldots \right) \approx 2^{3L}$$

The factorization costs are $O(n^3)$. Storing the matrices on each level costs

$$4^l \, (2^{L-l})^2 = 2^{2L}.$$

The total memory is $O(L \times 2^{2L}) = O(n^2 \, \log n)$.

This corresponds to

n	time	memory
10	1 μs	3 kB
100	1 ms	500 kB
1000	1 s	150 MB

A corresponding sparse factorization algorithm for matrices arising from unstructured meshes is based on minimum degree ordering. Successively, the unknowns with the least connections in the matrix graph are eliminated.

In 2D, a direct method with optimal ordering is very efficient. In 3D, the situation is worse for the direct solver. There holds $N = n^3$, time complexity $= O(N^2)$, and memory $= O(N^{1.33})$.

5.2 Iterative equation solvers

Iterative equation solvers improve the accuracy of approximative solution by an successive process. This requires in general much less memory, and, depending on the problem and on the method, may be (much) faster.

The Richardson iteration

A simple iterative method is the preconditioned Richardson iteration (also known as simple iteration, or Picard iteration):

start with arbitrary u^0
for $k = 0, 1, \ldots$ convergence
  $d^k = f - Au^k$
  $w^k = C^{-1} d^k$
  $u^{k+1} = u^k + \tau w^k$

Here, τ is a damping parameter which may be necessary to ensure convergence. The matrix C is called a preconditioner. It should fulfill

1. C is a good approximation to A
2. the matrix-vector multiplication $w = C^{-1} d$ should be cheap

A simple choice is $C = \operatorname{diag} A$, the Jacobi preconditioner. The application of C^{-1} is cheap. The quality of the approximation $C \approx A$ will be estimated below. The optimal choice for the first criterion would be $C = A$. But, of course, $w = C^{-1} d$ is in general not cheap.

Combining the steps, the iteration can be written as

$$u^{k+1} = u^k + \tau C^{-1}(f - Au^k)$$

Let u be the solution of the equation $Au = f$. We are interested in the behavior of the error $u^k - u$:

$$\begin{aligned} u^{k+1} - u &= u^k - u + \tau C^{-1}(f - Au^k) \\ &= u^k - u + \tau C^{-1}(Au - Au^k) \\ &= (I - \tau C^{-1} A)(u^k - u) \end{aligned}$$

We call the matrix

$$M = I - \tau C^{-1} A$$

the *iteration matrix*. The error transition can be estimated by

$$\|u^{k+1} - u\| \leq \|M\| \, \|u^k - u\|.$$

The matrix norm is the associated matrix norm to some vector norm. If $\rho := \|M\| < 1$, then the error is reduced. The error after k steps is

$$\|u^k - u\| \leq \rho^k \|u^0 - u\|$$

To reduce the error by a factor ε (e.g., $\varepsilon = 10^{-8}$), one needs

$$N_{its} = \frac{\log \varepsilon}{\log \rho}$$

iterations.

We will focus on the symmetric ($A = A^T$) and positive definite ($u^T A u > 0$ for $u \neq 0$) case (short: SPD). Then it makes sense to choose symmetric and positive definite preconditioners $C = C^T$. Eigenvalue decomposition allows a sharp analysis. Pose the generalized eigenvalue problem

$$Az = \lambda C z.$$

Let (λ_i, z_i) be the set of eigen-pairs. The spectrum is $\sigma\{C^{-1}A\} = \{\lambda_i\}$. The eigen-vectors z_i are normalized to

$$\|z_i\|_C = 1$$

The eigenvalues can are bounded from below and from above by the Rayleigh quotient:

$$\min_v \frac{v^T A v}{v^T C v} \leq \lambda_i \leq \max_v \frac{v^T A v}{v^T C v}$$

The ratio of largest to smallest eigen-value is the relative spectral condition number

$$\kappa = \frac{\lambda_N}{\lambda_1}$$

We will establish the spectral bounds

$$\gamma_1 \, v^T C v \leq v^T A v \leq \gamma_2 \, v^T C v \qquad \forall \, v \in \mathbb{R}^N,$$

which allow to bound the eigenvalues

$$\lambda_i \in [\gamma_1, \gamma_2],$$

and the condition number $\kappa \leq \frac{\gamma_2}{\gamma_1}$.

A vector v can be expressed in terms of the eigen-vector basis z_i as $v = \sum v_i e_i$. There holds

$$\begin{aligned} \|v\|_C^2 &= \sum v_i^2 \\ \|v\|_A^2 &= \sum \lambda_i v_i^2 \end{aligned}$$

Lemma 88. *The iteration matrix M can be bounded in A-norm and in C-norm:*

$$\|M\|_A \leq \sup_{\lambda\in[\gamma 1,\gamma_2]} |1-\tau\lambda|$$

$$\|M\|_C \leq \sup_{\lambda\in[\gamma 1,\gamma_2]} |1-\tau\lambda|$$

Proof: Express $v = \sum v_i z_i$. Then

$$Mv = (I-\tau C^{-1}A)v = \sum v_i (I-\tau C^{-1}A) z_i = \sum v_i (1-\tau\lambda_i) z_i$$

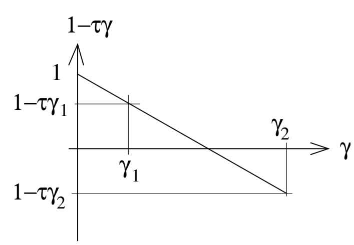
The norm is

$$\begin{aligned} \|Mv\|_A^2 &= \sum \lambda_i v_i^2 (1-\tau\lambda_i)^2 \\ &\leq \sup_i (1-\tau\lambda_i)^2 \sum \lambda_i v_i^2 \\ &\leq \sup_{\lambda\in[\gamma_1,\gamma_2]} (1-\tau\lambda)^2 \, \|v\|_A^2 \end{aligned}$$

and thus

$$\|M\|_A = \sup_{v\in\mathbb{R}^n} \frac{\|Mv\|_A}{\|v\|_A} \leq \sup_{\lambda\in[\gamma_1,\gamma_2]} |1-\tau\lambda|.$$

The proof is equivalent for $\|M\|_C$. $\square$



The optimal choice of the relaxation parameter τ is such that

$$1-\tau\gamma_1 = -(1-\tau\gamma_2),$$

i.e.,

$$\tau = \frac{2}{\gamma_1+\gamma_2}$$

The convergence factor is

$$1-\tau\gamma_1 = \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}.$$

Assume we knew sharp spectral bounds $\gamma_1 = \lambda_1$ and $\gamma_2 = \lambda_N$. Then the convergence factor is

$$\|M\| = \frac{\kappa-1}{\kappa+1} \approx 1 - \frac{2}{\kappa}$$

The number of iterations to reduce the error by a factor ε is

$$N_{its} = \frac{\log \varepsilon}{\log \rho} \approx \frac{\log \varepsilon}{-2/\kappa} = \log \varepsilon^{-1} \frac{\kappa}{2}$$

Take the 1D stiffness matrix of dimension $N \times N$:

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

and the trivial preconditioner $C = I$. The eigen-vectors z_i and eigen-values λ_i are

$$\begin{aligned} z_i &= \left(\sin \frac{ij\pi}{N+1} \right)_{j=1,\ldots N} \\ \lambda_i &= 2 - 2\cos(\frac{i\pi}{N+1}) \end{aligned}$$

The extremal eigenvalues are

$$\begin{aligned} \lambda_1 &= 2 - 2\cos(\frac{\pi}{N+1}) \approx \frac{\pi^2}{N^2} \\ \lambda_N &= 2 - 2\cos(\frac{N\pi}{N+1} \approx 4 - \frac{\pi^2}{N^2}. \end{aligned}$$

The optimal damping is

$$\tau = \frac{2}{\lambda_1 + \lambda_N} = \frac{1}{2},$$

and the convergence factor is

$$\|M\| \approx 1 - \frac{2\lambda_1}{\lambda_N} \approx 1 - \frac{2\pi^2}{N^2}$$

The number of iterations is

$$N_{its} \simeq \log \varepsilon^{-1} N^2$$

For the 2D model problem with $N = (n-1)^2$, the condition number behaves like

$$\kappa \simeq n^2.$$

The costs to achieve a relative accuracy ε are

$$N_{its} \times \text{Costs-per-iteration} \simeq \log \varepsilon^{-1} n^2 N \simeq \log \varepsilon^{-1} n^4$$

The costs per digit are comparable to the band-factorization. The memory requirement is optimal $O(N)$.

The gradient method

It is not always feasible to find the optimal relaxation parameter τ a priori. The gradient method is a modification to the Richardson method to find automatically the optimal relaxation parameter τ:

The first steps are identic:

$$d^k = f - Au^k \qquad w^k = C^{-1} d^k$$

Now, perform the update

$$u^{k+1} = u^k + \tau w^k$$

such that the error is minimal in energy norm:

$$\text{Find } \tau \text{ such that } \|u - u^{k+1}\|_A = \min!$$

Although the error cannot be computed, this minimization is possible:

$$\begin{aligned} \|u - u^{k+1}\|_A^2 &= \|u - u^k - \tau w^k\|_A^2 \\ &= (u - u^k)^T A (u - u_k) - 2\tau (u - u^k)^T A w^k + \tau^2 (w^k)^T A w^k \end{aligned}$$

This is a convex function in τ. It takes its minimum at

$$0 = 2(u - u^k)^T A w^k + 2\tau_{opt} (w^k)^T A w^k,$$

i.e.,

$$\tau_{opt} = \frac{w^k A (u - u^k)}{(w^k)^T A w^k} = \frac{w^k d^k}{(w^k)^T A w^k}$$

Since the gradient method gives optimal error reduction in energy norm, its convergence rate can be estimated by the Richardson iteration with optimal choice of the relaxation parameter:

$$\|u - u^{k+1}\|_A \leq \frac{\kappa - 1}{\kappa + 1} \|u - u^k\|_A$$

The Chebyshev method

We have found the optimal choice of the relaxation parameter for one step of the iteration. If we perform m iterations, the overall rate of convergence can be improved by choosing variable relaxation parameters $\tau_1, \dots \tau_m$.

The m-step iteration matrix is

$$M = M_m \dots M_2 M_1 = (I - \tau_m C^{-1} A) \dots (I - \tau_1 C^{-1} A).$$

By diagonalization, the A-norm and C-norm are bounded by

$$\|M\| \leq \max_{\lambda \in [\gamma_1, \gamma_N]} |(1 - \tau_1 \lambda) \dots (1 - \tau_m \lambda)|$$

The goal is to optimize $\tau_1, \dots \tau_m$:

$$\min_{\tau_1,\dots\tau_m} \max_{\lambda\in[\gamma_1,\gamma_N]} |(1-\tau_1\lambda)\dots(1-\tau_m\lambda)|$$

This is a polynomial in λ, of order m, and $p(0) = 1$:

$$\min_{\substack{p\in P^m\\ p(0)=1}} \max_{\lambda\in[\gamma_1,\gamma_N]} |p(\lambda)|. \tag{5.1}$$

This optimization problem can be solved explicitely by means of Chebyshev polynomials. These are the polynomials defined by

$$T_m(x) = \begin{cases} \cos(m \arccos(x)) & |x| \le 1 \\ \cosh(m \operatorname{arccosh}(x)) & |x| > 1 \end{cases}$$

The T_m fulfill the recurrence relation

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{m+1}(x) &= 2xT_m(x) - T_{m-1}(x) \end{aligned}$$

The T_m fulfill also

$$T_m(x) = \frac{1}{2}\left[(x+\sqrt{x^2-1})^m + (x+\sqrt{x^2-1})^{-m}\right]$$

The optimum of (5.1) is

$$p(x) = \frac{T_m\left(\frac{2x-\gamma_1-\gamma_2}{\gamma_2-\gamma_1}\right)}{T_m\left(\frac{-\gamma_1-\gamma_2}{\gamma_2-\gamma_1}\right)} = C_m\, T_m\left(\frac{2x-\gamma_1-\gamma_2}{\gamma_2-\gamma_1}\right)$$

The numerator is bounded by 1 for the range $\gamma_1 \le x \le \gamma_2$. The factor C_m can be computed as

$$C_m = \frac{2c^m}{1+c^{2m}} \qquad \text{with} \qquad c = \frac{\sqrt{\gamma_2}-\sqrt{\gamma_1}}{\sqrt{\gamma_2}+\sqrt{\gamma_1}}$$

Using the condition number we have

$$c \approx 1 - \frac{2}{\sqrt{\kappa}},$$

and

$$C_m \approx (1 - \frac{2}{\sqrt{\kappa}})^m$$

Now, an error reduction by a factor of ε can be achieved in

$$N_{its} \approx \log \varepsilon^{-1} \sqrt{\kappa}$$

steps. The original method by choosing m different relaxation parameters τ_k is not a good choice, since

- it is not numerically stable
- one has to know a priori the number of iterations

The recurrence relation for the Chebyshev polynomials leads to a practicable iterative method called Chebyshev iteration.

The conjugate gradient method

The conjugate gradient algorithm automatically finds the optimal relaxation parameters for the best k-step approximation.

Let $p_0, p_1, \ldots$ be a finite sequence of A-orthogonal vectors, and set

$$V_k = \operatorname{span}\{p_0, \ldots p_{k-1}\}$$

We want to approximate the solution u in the linear manifold $u_0 + V_k$:

$$\min_{v \in u_0 + V_k} \|u - v\|_A$$

We represent u_k as

$$u_k = u_0 + \sum_{l=0}^{k-1} \alpha_l p_l$$

The optimality criteria are

$$0 = (u - u_k, p_j)_A = (u - u_0 - \sum_{l=0}^{k-1} \alpha_l p_l, p_j)_A \qquad 0 \leq j < k.$$

The coefficients α_l follow from the A-orthogonality:

$$\alpha_l = \frac{(u - u_0)^T A p_l}{p_l^T A p_l} = \frac{(f - Au_0)^T p_l}{p_l^T A p_l}$$

The α_l are computable, since the A-inner product was chosen. The best approximations can be computed recursively:

$$u_{k+1} = u_k + \alpha_k p_k$$

Since $u_k - u_0 \in V_k$, and $p_k \perp_A V_k$, there holds

$$\alpha_k = \frac{(f - Au_k)^T p_k}{p_k^T A p_k}.$$

Any k-step simple iteration approximates the solution u_k in the manifold

$$u_0 + \mathcal{K}_k(d_0)$$

with the *Krylov space*

$$\mathcal{K}_k(d_0) = \{C^{-1} d_0, C^{-1} A C^{-1} d_0, \ldots, C^{-1} (A C^{-1})^{k-1} d_0\}.$$

Here, $d_0 = f - Au_0$ is the initial residual. The conjugate gradient method computes an A-orthogonal basis of the Krylov-space. The term *conjugate* is equivalent to A-orthogonal.

Conjugate Gradient Algorithm:
Choose u_0, compute $d_0 = f - Au_0$, set $p_0 = C^{-1}d_0$.
for $k = 0, 1, 2, \dots$ compute

$$\begin{aligned}
\alpha_k &= \frac{d_k^T p_k}{p_k^T A p_k} \\
u_{k+1} &= u_k + \alpha_k\, p_k \\
d_{k+1} &= d_k - \alpha_k\, A p_k \\
\beta_k &= -\frac{d_{k+1}^T C^{-1} A p_k}{p_k^T A p_k} \\
p_{k+1} &= C^{-1} d_{k+1} + \beta_k p_k
\end{aligned}$$

Remark 89. *In exact arithmetic, the conjugate gradient algorithm terminates at a finite number of steps $\overline{k} \leq N$.*

Theorem 90. *The conjugate gradient algorithm fulfills for $k \leq \overline{k}$*

1. *The sequence p_k is A-orthogonal. It spans the Krylov-space $\mathcal{K}_k(d_0)$*
2. *The u_k minimizes*
$$\min_{v \in u_0 + \mathcal{K}_k(d_0)} \|u - v\|_A$$
3. *There holds the orthogonality*
$$d_k^T p_l = 0 \qquad \forall\, l < k$$

Proof: Per induction in k. We assume

$$\begin{aligned}
p_k^T A p_l &= 0 \qquad \forall\, l < k \\
d_k^T p_l &= 0 \qquad \forall\, l < k
\end{aligned}$$

This is obvious for $k = 0$. We prove the property for $k+1$: For $l < k$ there holds

$$d_{k+1}^T p_l = (d_k - \alpha_k A p_k)^T p_l = d_k^T p_l - \alpha_k p_k^T A p_l = 0$$

per induction. For $l = k$ there is

$$d_{k+1}^T p_k = (d_k - \alpha_k A p_k)^T p_k = d_k^T p_k - \frac{d_k^T p_k}{p_k^T A p_k}\, p_k^T A p_k = 0.$$

Next, prove the A-orthogonality of the p_k. For $l < k$ we have

$$\begin{aligned}
(p_{k+1}, p_l)_A &= (C^{-1} d_{k+1} + \beta_k p_k, p_l)_A \\
&= d_{k+1}^T C^{-1} A p_l
\end{aligned}$$

There is

$$C^{-1}Ap_l \in \operatorname{span}\{p_0, \dots p_k\},$$

and $d_{k+1}^T p_j = 0$ for $j \le k$. For $l = k$ there is

$$\begin{aligned}(p_{k+1}, p_k)_A &= (C^{-1}d_{k+1} + \beta_k p_k, p_k)_A \\ &= (C^{-1}d_{k+1}, p_k)_A - \frac{d_{k+1}^T C^{-1}Ap_k}{p_k^T Ap_k}\, p_k^T Ap_k = 0\end{aligned}$$

□

The coefficients α_k and β_k should be computed by the equivalent, and numerically more stable expressions

$$\alpha_k = \frac{d_k^T C^{-1} d_k}{p_k^T A p_k} \qquad \beta_k = \frac{d_{k+1}^T C^{-1} d_{k+1}}{d_k^T C^{-1} d_k}.$$

Theorem 91. *The conjugate gradient iteration converges with the rate*

$$\|u - u_k\|_A \le \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k$$

Proof: The conjugate gradient gives the best approximation in the Krylov space. Thus, it can be bounded by the Chebyshev method leading to that rate.

The conjugate gradient iteration is stopped as soon as a convergence criterion is fulfilled. Ideally, one wants to reduce the error in the energy norm by a factor ε:

$$\|u - u_k\|_A \le \varepsilon\, \|u - u_0\|_A$$

But, the energy error cannot be computed. We rewrite

$$\|u - u_k\|_A^2 = \|A^{-1}(f - Au_k)\|_A^2 = \|A^{-1}d_k\|_A^2 = d_k^T A^{-1} d_k$$

If C is a good approximation to A, then also C^{-1} is one to A^{-1}. The error can be approximated by

$$d_k^T C^{-1} d_k.$$

This scalar is needed in the conjugate gradient iteration, nevertheless.

For solving the 2D model problem with $C = I$, the time complexity is

$$\log \varepsilon^{-1}\, N \sqrt{\kappa} = \log \varepsilon^{-1}\, n^3$$

The costs for one digit are comparable to the recursive sub-structuring algorithm. In 3D, the conjugate gradient method has better time complexity.

5.3 Preconditioning

In the following, let the symmetric and positive definite matrix A arise from the finite element discretization of the H^1-elliptic and continuous bilinear-form $A(.,.)$. We construct preconditioners C such that the preconditioning action

$$w = C^{-1} \times d$$

is efficiently computable, and estimate the spectral bounds

$$\gamma_1 \, u^T C u \leq u^T A u \leq \gamma_2 \, u^T C u \qquad \forall \, u \in \mathbb{R}^N$$

The analysis of the preconditioner is performed in the finite element framework. For this, define the Galerkin isomorphism

$$G : \mathbb{R}^N \to V_h : \underline{u} \to u = \sum u_i \varphi_i,$$

where φ_i are the fe basis functions. Its dual is

$$G^* : V_h^* \to \mathbb{R}^N : d(\cdot) \to (d(\varphi_i))_{i=1,\ldots,N}.$$

To distinguish vectors and the corresponding finite element functions, we write vectors $\underline{u} \in \mathbb{R}^N$ with underlines (when necessary).

The evaluation of the quadratic form is

$$\underline{u}^T A \underline{u} = A(G\underline{u}, G\underline{u}) \simeq \|G\underline{u}\|_{H^1}^2$$

The Jacobi Preconditioner

The Jacobi preconditioner C is

$$C = \operatorname{diag} A.$$

The preconditioning action is written as

$$C^{-1} \times \underline{d} = \sum_{i=1}^{N} e_i (e_i^T A e_i)^{-1} e_i^T \ \underline{d}$$

Here, e_i is the i^{th} unit-vector. Thus, $e_i^T A e_i$ gives the i^{th} diagonal element A_{ii} of the matrix, which is

$$A_{ii} = A(\varphi_i, \varphi_i) \simeq \|\varphi_i\|_{H^1}^2.$$

The quadratic form generated by the preconditioner is

$$\underline{u}^T C \underline{u} = \sum_{i=1}^{N} u_i^2 \|\varphi_i\|_A^2 \simeq \sum_{i=1}^{N} u_i^2 \|\varphi_i\|_{H^1}^2$$

Theorem 92. *Let h be the minimal mesh-size of a shape-regular triangulation. Then there holds*

$$h^2\ \underline{u}^T C \underline{u} \preceq u^T A u \preceq\ u^T C u \tag{5.2}$$

Proof: We start to prove the right inequality

$$\underline{u}^T A \underline{u} = \| \sum_i u_i \varphi_i \|_A^2 \preceq \underline{u}^T C \underline{u} = \sum_i u_i^2 \|\varphi_i\|_A^2.$$

We define the interaction matrix O with entries

$$O_{ij} = \begin{cases} 1 & A(\varphi_i, \varphi_j) \neq 0 \\ 0 & \text{else} \end{cases}$$

On a shape regular mesh, only a (small) finite number of basis functions have overlapping support. Thus, O has a small number of entries 1 per row. There holds

$$\begin{aligned}
\| \sum_i u_i \varphi_i \|_A^2 &= \sum_i \sum_j u_i u_j A(\varphi_i, \varphi_j) \\
&= \sum_i \sum_j u_i u_j O_{ij} A(\varphi_i, \varphi_j) \\
&\leq \sum_i \sum_j (u_i \|\varphi_i\|_A) O_{ij} (u_j \|\varphi_j\|_A) \\
&\leq \rho(O) \sum_i (u_i \|\varphi_i\|_A)^2 \\
&= \rho(O) \underline{u}^T C \underline{u}.
\end{aligned}$$

The spectral radius $\rho(O) = \max_{x \in \mathbb{R}^N} \frac{x^T O x}{\|x\|^2}$ is bounded by the (small) finite row-sum norm of O.

The other estimate is proven element by element. Note that

$$\underline{u}^T A \underline{u} \simeq \|u\|_{H^1(\Omega)}^2 = \sum_T \| \sum_i u_i \varphi_i \|_{H^1(T)}^2$$

and

$$\underline{u}^T C \underline{u} \simeq \sum_i \|u_i \varphi_i\|_{H^1(\Omega)}^2 = \sum_T \sum_i \|u_i \varphi_i\|_{H^1(T)}^2.$$

We prove the inequality for each individual element. The triangle T has diameter h_T. On T, we expand u in terms of the element shape functions φ_α, namely $u|_T = \sum_{\alpha=1}^3 u_\alpha \varphi_\alpha$. We transform to the reference element $\widehat{T}$:

$$\begin{aligned}
\| \sum_\alpha u_\alpha \varphi_\alpha \|_{H^1(T)}^2 &= \| \sum_\alpha u_\alpha \varphi_\alpha \|_{L_2(T)}^2 + \|\nabla \sum_\alpha u_\alpha \varphi_\alpha \|_{L_2(T)}^2 \\
&\simeq h_T^2 \| \sum_\alpha u_\alpha \widehat{\varphi}_\alpha \|_{L_2(\widehat{T})}^2 + \|\nabla \sum_\alpha u_\alpha \widehat{\varphi}_\alpha \|_{L_2(\widehat{T})}^2 \\
&\geq h_T^2 \| \sum_\alpha u_\alpha \widehat{\varphi}_\alpha \|_{L_2(\widehat{T})}^2
\end{aligned}$$

and

$$\begin{aligned}
\sum_\alpha \|u_\alpha \varphi_\alpha\|^2_{H^1(T)} &= \sum_\alpha \|u_\alpha \varphi_\alpha\|^2_{L_2(T)} + \sum_\alpha \|\nabla u_\alpha \varphi_\alpha\|^2_{L_2(T)} \\
&\simeq h_T^2 \sum_\alpha \|u_\alpha \widehat{\varphi}_\alpha\|^2_{L_2(\widehat{T})} + \sum_\alpha \|\nabla u_\alpha \widehat{\varphi}_\alpha\|^2_{L_2(\widehat{T})} \\
&\preceq \sum_\alpha \|u_\alpha \widehat{\varphi}_\alpha\|^2_{L_2(\widehat{T})}
\end{aligned}$$

Both, $(u)_\alpha \rightarrow \| \sum_\alpha u_\alpha \widehat{\varphi}_\alpha\|_{L_2(\widehat{T})}$ and $u \rightarrow \left\{ \sum_\alpha \|u_\alpha \widehat{\varphi}_\alpha\|^2_{L_2(\widehat{T})} \right\}^{1/2}$ are norms on $\mathbb{R}^3$. Since all norms in $\mathbb{R}^3$ are equivalent, we have

$$\sum_\alpha \|u_\alpha \varphi_\alpha\|^2_{H^1(T)} \preceq h_T^{-2} \| \sum_\alpha u_\alpha \varphi_\alpha \|^2_{H^1(T)}. \tag{5.3}$$

By summing over all elements and choosing $h = \min h_T$, we have proven the left inequality of (5.2). $\square$

Remark: Inequality (5.3) is sharp. To prove this, choose $u_\alpha = 1$.

Block-Jacobi preconditioners

Instead of choosing the diagonal, one can choose a block-diagonal of A, e.g.,

- In the case of systems of PDEs, choose blocks consisting of all degrees of freedom sitting in one vertex. E.g., mechanical deformations (u_x, u_y, u_z).
- For high order elements, choose blocks consisting of all degrees of freedom associated to the edges (faces, inner) of the elements.
- On anisotropic tensor product meshes, choose blocks consisting of unknowns in the short direction
- Domain decomposition methods: Choose blocks consisting of the unknowns in a sub-domain

Decompose the unknowns into M blocks, the block i has dimension N_i. Define the rectangular embedding matrices

$$E_i \in \mathbb{R}^{N \times N_i} \qquad i = 1, \ldots, M.$$

E_i consists of N_i unit vectors corresponding to the unknowns in the block i. Each $\underline{u} \in \mathbb{R}^N$ can be uniquely written as

$$\underline{u} = \sum_{i=1}^M E_i \underline{u}_i \qquad \text{with} \quad \underline{u}_i \in \mathbb{R}^{N_i}$$

The diagonal blocks are

$$A_i = E_i^T A E_i \qquad i = 1, \ldots, M.$$

The block Jacobi preconditioner is

$$C^{-1} \times \underline{d} = \sum_{i=1}^{M} E_i A_i^{-1} E_i^T \, \underline{d}$$

The quadratic form induced by C can be written as

$$\underline{u}^T C \underline{u} = \sum_i \underline{u}_i^T A_i \underline{u}_i = \sum_i \|G E_i \underline{u}_i\|_A^2$$

where $u = \sum E_i u_i$.

Example: Discretize the unit interval $I = (0,1)$ into n elements of approximate size $h \simeq 1/n$. Split the unknowns into two blocks, the left half and the right half, and define the corresponding block-Jacobi preconditioner.

Set

$$I = I_1 \cup T_{n/2} \cup I_2,$$

with $I_1 = (0, x_{n/2})$, $T_{n/2} = [x_{n/2}, x_{n/2+1}]$, and $I_2 = (x_{n/2+1}, 1)$. Decompose

$$\underline{u} = E_1 \underline{u}_1 + E_2 \underline{u}_2.$$

The corresponding finite element functions are $u_i = G E_i \underline{u}_i$. There holds

$$u_1(x) = \left\{ \begin{array}{cl} G\underline{u}(x) & x \in I_1 \\ \text{linear} & x \in T \\ 0 & x \in I_2 \end{array} \right. ,$$

and u_2 vice versa. The quadratic form is

$$\underline{u}^T C \underline{u} = \sum_i u_i A_i u_i = \sum_i \|G E_i u_i\|_A^2$$

Evaluation gives

$$\begin{array}{rcl} \|u_1\|_A^2 & = & \|u_1\|_{H^1(I_1)}^2 + \|u_1\|_{H^1(T)}^2 \\ & \simeq & \|u_1\|_{H^1(I_1)}^2 + h^{-1} |u(x_{n/2})|^2 \\ & \preceq & \|u\|_{H^1(I)}^2 + h^{-1} \|u\|_{H^1(I)}^2 \qquad \text{(trace theorem)} \\ & \simeq & h^{-1} \, \|u\|_A^2, \end{array}$$

and thus

$$\underline{u}^T C \underline{u} = \sum_i \|u_i\|_A^2 \preceq h^{-1} \|u\|_A^2 \simeq h^{-1} \underline{u}^T A \underline{u}.$$

The situation is the same in $\mathbb{R}^d$.

Exercise: Sub-divide the interval I into M sub-domains of approximative size $H \approx 1/M$. What are the sprectral bounds of the block-Jacobi preconditioner ?

Additive Schwarz preconditioners

The next generalization is an *overlapping* block Jacobi preconditioner. For $i = 1, \ldots, M$ let $E_i \in \mathbb{R}^{N \times N_i}$ be rectangular matrices such that each $u \in \mathbb{R}^N$ can be (not necessarily uniquely) written as

$$\underline{u} = \sum_{i=1}^{M} E_i \underline{u}_i \qquad \text{with} \quad \underline{u}_i \in \mathbb{R}^{N_i}$$

Again, the overlapping block-Jacobi preconditioning action is

$$C^{-1} \times \underline{d} = \sum_{i=1}^{M} E_i A_i^{-1} E_i^T \, \underline{d}$$

Example: Choose the unit-interval problem from above. The block 1 contains all nodes in $(0, 3/4)$, and the block 2 contains nodes in $(1/4, 1)$. The blocks overlap, the decomposition is not unique.

The columns of the matrices E_i are not necessarily unit-vectors, but are linearly independent. In this general setting, the preconditioner is called *Additive Schwarz preconditioner*. The following lemma gives a useful representation of the quadratic form. It was proven in similar forms by many authors (Nepomnyaschikh, Lions, Dryja+Widlund, Zhang, Xu, Oswald, Griebel, ...) and is called also Lemma of many fathers, or Lions' Lemma:

Lemma 93 (Additive Schwarz lemma)**.** *There holds*

$$\underline{u}^T C \underline{u} = \inf_{\substack{\underline{u}_i \in \mathbb{R}^{N_i} \\ \underline{u} = \sum E_i \underline{u}_i}} \sum_{i=1}^{M} \underline{u}_i^T A_i \underline{u}_i$$

Proof: The right hand side is a constrained minimization problem of a convex function. The feasible set is non-empty, the CMP has a unique solution. It is solved by means of Lagrange multipliers. Define the Lagrange-function (with Lagrange multipliers $\lambda \in \mathbb{R}^N$):

$$L((u_i), \lambda) = \sum u_i^T A u_i + \lambda^T (u - \sum E_i u_i).$$

Its stationary point (a saddle point) is the solution of the CMP:

$$\begin{aligned} 0 = \nabla_{u_i} L((u_i), \lambda) &= 2 A_i u_i + E_i^T \lambda \\ 0 = \nabla_{\lambda} L((u_i), \lambda) &= u - \sum E_i u_i \end{aligned}$$

The first line gives

$$u_i = \frac{1}{2} A_i^{-1} E_i^T \lambda.$$

Use it in the second line to obtain

$$0 = u - \frac{1}{2} \sum E_i A_i^{-1} E_i \lambda = u - \frac{1}{2} C^{-1} \lambda,$$

i.e., $\lambda = 2Cu$, and

$$u_i = A_i^{-1} E_i^T C u.$$

The minimal value is

$$\begin{aligned} \sum u_i^T A_i u_i &= \sum u^T C E_i A_i^{-1} A_i A_i^{-1} E_i^T C u \\ &= \sum u^T C E_i A_i^{-1} E_i^T C u \\ &= u^T C C^{-1} C u = u^T C u \end{aligned}$$

□

Next, we rewrite the additive Schwarz iteration matrix

$$I - \tau C^{-1} A = I - \tau \sum_{i=1}^{M} E_i A_i^{-1} E_i^T A$$

in the fe framework. Let

$$V_i = G E_i \mathbb{R}^{N_i} \subset V_h$$

be the sub-space corresponding to the range of E_i, and define the A-orthogonal projection

$$P_i : V_h \to V_i : \qquad A(P_i u, v_i) = A(u, v_i) \qquad \forall\, v_i \in V_i$$

Lemma 94. *Set $u = G\underline{u}$, the application of the iteration matrix is $\underline{\hat{u}} = (I - \tau C^{-1} A)\underline{u}$, and set $\hat{u} = G\underline{\hat{u}}$. Then there holds*

$$\hat{u} = \left(I - \tau \sum_{i=1}^{M} P_i \right) u.$$

Proof: Let $\underline{w}_i = A_i^{-1} E_i^T A \underline{u}$. Then

$$\hat{u} = u - \tau G E_i \underline{w}_i.$$

There holds $w_i := G E_i \underline{w}_i \in V_i$, and

$$\begin{aligned} A(G E_i \underline{w}_i, G E_i \underline{v}_i) &= \underline{v}_i^T E_i^T A E_i \underline{w}_i \\ &= \underline{v}_i^T A_i \underline{w}_i = \underline{v}_i^T E_i^T A \underline{u} \\ &= A(G\underline{u}, G E_i \underline{v}_i) \qquad\qquad \forall\, v_i \in \mathbb{R}^{N_i}, \end{aligned}$$

i.e., $w_i = P_i u$. □

The additive Schwarz preconditioner is defined by the space splitting

$$V = \sum_{i=1}^{M} V_i$$

If the spaces V_i are A-orthogonal, then $\sum_i P_i = I$, and (with $\tau = 1$), and the iteration matrix is $M = 0$.

The reformulation of the additive Schwarz lemma 93 in the finite element framework is

Lemma 95 (Additive Schwarz lemma)**.** *Let $u = G\underline{u}$. There holds*

$$\underline{u}^T C \underline{u} = \inf_{\substack{u_i \in V_i \\ u = \sum u_i}} \sum_{i=1}^{M} \|u_i\|_A^2$$

Example: Let

$$A(u,v) = \int_0^1 u'v' + \varepsilon \int uv\,dx$$

with $0 \leq \varepsilon \ll 1$. The bilinear-form is H^1-elliptic and continuous, but the bounds depend on the parameter ε. Let C_J be the Jacobi preconditioner. The proof of Theorem 92 shows that

$$\varepsilon h^2 \underline{u}^T C_J \underline{u} \preceq \underline{u}^T A \underline{u} \preceq \underline{u}^T C_J \underline{u}.$$

The non-robust lower bound is sharp: Take $\underline{u} = (1, \ldots, 1)^T$.

The solution is to add the additional sub-space

$$V_0 = \text{span}\{1\} = G E_0 \mathbb{R}^1$$

to the AS preconditioner (with $E_0 \in \mathbb{R}^{N \times 1}$ consisting of 1-entries). The preconditioning action is

$$C^{-1} \times d = \text{diag}\{A\}^{-1} d + E_0 (E_0^T A E_0)^{-1} E_0^T d.$$

The spectral bounds are robust in ε:

$$h^2 \underline{u}^T C \underline{u} \preceq \underline{u}^T A \underline{u} \preceq \underline{u}^T C \underline{u},$$

namely

$$\begin{aligned}
\underline{u}^T C \underline{u} &= \inf_{\substack{u_i \in V_i \\ u = \sum_0^M u_i}} \sum_{i=0}^{M} \|u_i\|_A^2 \\
&= \inf_{u_0 \in V_0} \left\{ \|u_0\|_A^2 + \inf_{\substack{u_i \in V_i \\ u - u_0 = \sum_1^M u_i}} \sum_{i=1}^{M} \|u_i\|_A^2 \right\} \\
&\preceq \inf_{u_0 \in V_0} \|u_0\|_A^2 + h^{-2} \|u - u_0\|_{H^1}^2
\end{aligned}$$

The last step was the result of the Jacobi preconditioner applied to $(u,v)_{H^1}$. Finally, we choose $u_0 = \int_0^1 u\,dx$ to obtain

$$\begin{aligned}
\underline{u}^T C \underline{u} &\preceq \|u_0\|_A^2 + h^{-2} \|u - u_0\|_{H^1}^2 \\
&\preceq \varepsilon \|u_0\|_{L_2}^2 + h^{-2} \|\nabla (u - u_0)\|_{L_2}^2 \\
&\preceq \varepsilon \|u\|_{L_2}^2 + h^{-2} \|\nabla u\|_{L_2}^2 \\
&= h^{-2} \|u\|_A^2
\end{aligned}$$

Overlapping domain decomposition preconditioning

Let $\Omega = \cup_{i=1}^M \Omega_i$ be a decomposition of Ω into M sub-domains of diameter H. Let $\widetilde{\Omega}_i$ be such that

$$\Omega_i \subset \widetilde{\Omega}_i \qquad \operatorname{dist}\{\partial\widetilde{\Omega}_i \setminus \partial\Omega, \partial\Omega_i\} \succeq H,$$

and only a small number of $\widetilde{\Omega}_i$ are overlapping. Choose a finite element mesh of mesh size $h \leq H$, and the finite element space is V_h. The overlapping domain decomposition preconditioner is the additive Schwarz preconditioner defined by the sub-space splitting

$$V_h = \sum V_i \qquad \text{with} \qquad V_i = V_h \cap H_0^1(\widetilde{\Omega}_i).$$

The bilinear-form $A(.,.)$ is H^1-elliptic and continuous. The implementation takes the sub-matrices of A with nodes inside the enlarged sub-domains $\widetilde{\Omega}_i$.

Lemma 96. *The overlapping domain decomposition preconditioner fulfills the spectral estimates*

$$H^2 \underline{u}^T C \underline{u} \preceq \underline{u} A \underline{u} \preceq \underline{u}^T C \underline{u}.$$

Proof: The upper bound is generic. For the lower bound, we construct an explicit decomposition $u = \sum u_i$.

There exists a *partition of unity* $\{\psi_i\}$ such that

$$0 \leq \psi_i \leq 1, \qquad \operatorname{supp}\{\psi_i\} \subset \widetilde{\Omega}_i, \qquad \sum_{i=1}^M \psi_i = 1$$

and

$$\|\nabla \psi_i\|_{L_\infty} \preceq H^{-1}.$$

Let $\Pi_h : L_2 \to V_h$ be a Clément-type quasi-interpolation operator such that Π_h is a projection on V_h, and

$$\|\Pi_h v\|_{L_2} \preceq \|v\|_{L_2}, \qquad \text{and} \qquad \|\nabla \Pi_h v\|_{L_2} \preceq \|\nabla v\|_{L_2}.$$

For given $u \in V_h$, we choose the decomposition

$$u_i = \Pi_h(\psi_i u).$$

Indeed $u_i \in V_i$ is a decomposition of $u \in V_h$:

$$\sum u_i = \sum \Pi_h(\psi_i u) = \Pi_h \left((\sum \psi_i) u \right) = \Pi_h u = u$$

The lower bound follows from

$$\begin{aligned}
\underline{u}^T C \underline{u} &= \inf_{u=\sum v_i} \sum_i \|v_i\|_A^2 \\
&\leq \sum_i \|u_i\|_A^2 \preceq \sum_i \|u_i\|_{H^1}^2 \\
&= \sum_i \|\Pi_h(\psi_i u)\|_{H^1}^2 \\
&\preceq \sum_i \|\psi_i u\|_{H^1}^2 \\
&= \sum_i \left\{ \|\psi_i u\|_{L_2(\tilde{\Omega}_i)}^2 + \|\nabla(\psi_i u)\|_{L_2(\tilde{\Omega}_i)}^2 \right\} \\
&\preceq \sum_i \left\{ \|\psi_i u\|_{L_2(\tilde{\Omega}_i)}^2 + \|(\nabla\psi_i) u\|_{L_2(\tilde{\Omega}_i)}^2 + \|\psi_i \nabla u\|_{L_2(\tilde{\Omega}_i)}^2 \right\} \\
&\preceq \sum_i \left\{ \|u\|_{L_2(\tilde{\Omega}_i)}^2 + H^{-2}\|u\|_{L_2(\tilde{\Omega}_i)}^2 + \|\nabla u\|_{L_2(\tilde{\Omega}_i)}^2 \right\} \\
&\preceq \|u\|_{L_2(\Omega)}^2 + H^{-2}\,\|u\|_{L_2(\Omega)}^2 + \|\nabla u\|_{L_2(\Omega)}^2 \\
&\preceq H^{-2}\,\|u\|_A^2.
\end{aligned}$$

□

Overlapping DD preconditioning with coarse grid correction

The local DD preconditioner above gets worse, if the number of sub-domains increases. In the limit, if $H \simeq h$, the DD preconditioner is comparable to the Jacobi preconditioner.

To overcome this degeneration, we add one more subspace. Let $\mathcal{T}_H$ be a coarse mesh of mesh-size H, and $\mathcal{T}_h$ is the fine mesh generated by sub-division of $\mathcal{T}_H$. Let V_H be the finite element space on $\mathcal{T}_H$. The sub-domains of the domain decomposition are of the same size as the coarse grid.

The sub-space decomposition is

$$V_h = V_H + \sum_{i=1}^{M} V_i.$$

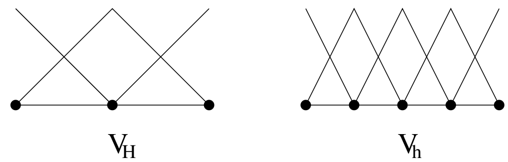
Let $G_H : \mathbb{R}^{N_H} \to V_H$ be the Galerkin isomorphism on the coarse grid, i.e.,

$$G_H \underline{u}_H = \sum_{i=1}^{N_H} u_{H,i} \varphi_i^H$$

The coarse space fulfills $V_H \subset V_h$. Thus, every coarse grid basis φ_i^H can be written as linear combination of fine grid basis functions φ_j^h:

$$\varphi_i^H = \sum_{j=1}^{N} E_{H,ji} \varphi_j^h.$$

Example:



The first basis function φ_1^H is

$$\varphi_1^H = \varphi_1^h + \frac{1}{2}\varphi_2^h$$

The whole matrix is

$$E_H = \begin{pmatrix} 1 & & \\ 1/2 & 1/2 & \\ & 1 & \\ & 1/2 & 1/2 \\ & & 1 \end{pmatrix}.$$

There holds

$$G_H \underline{u}_H = G_h E_H \underline{u}_H.$$

Proof:

$$\begin{aligned} G_H \underline{u}_H &= \sum_{i=1}^{N_H} u_{H,i}\varphi_i^H = \sum_{i=1}^{N_H}\sum_{j=1}^{N_h} u_{H,i}E_{H,ji}\varphi_j^h \\ &= \sum_{j=1}^{N_h} \varphi_j^h (E_H \underline{u}_H)_j = GE\underline{u}_H \end{aligned}$$

The matrix E_H transforms the coefficients $\underline{u}_H$ w.r.t. the coarse grid basis to the coefficients $\underline{u}_h = E_H \underline{u}_H$ w.r.t. the fine grid basis. It is called *prolongation matrix*.

The DD preconditioner with coarse grid correction is

$$C^{-1} \times d = \sum_i E_i A_i^{-1} E_i^T d + E_H (E_H^T A E_H)^{-1} E_H^T d$$

The first part is the local DD preconditioner from above. The second part is the coarse grid correction step. The matrix E_H^T (called *restriction matrix*) transfers the defect d from the fine grid to a defect vector on the coarse grid. Then, the coarse grid problem with matrix $E_H^T A E_H$ is solved. Finally, the result is prolongated to the fine grid.

The matrix $A_H := E_H^T A E_H$ is the Galerkin matrix w.r.t. the coarse grid basis:

$$\begin{aligned} A_{H,ij} &= \underline{e}_j^T E_H^T A E_H \underline{e}_i = A(G_h E_H \underline{e}_i, G_h E_H \underline{e}_j) \\ &= A(G_H \underline{e}_i, G_H \underline{e}_j) = A(\varphi_i^H, \varphi_j^H). \end{aligned}$$

Theorem 97. *The overlapping domain decomposition preconditioner with coarse grid system fulfills the optimal spectral estimates*

$$\underline{u}^T C \underline{u} \preceq \underline{u}^T A \underline{u} \preceq \underline{u}^T C \underline{u}.$$

Proof: The quadratic form generated by the preconditioner is

$$\underline{u}^T C \underline{u} = \inf_{\substack{u_H \in V_H, u_i \in V_i \\ u = u_H + \sum u_i}} \|u_H\|_A^2 + \sum_{i=1}^M \|u_i\|_A^2.$$

Again, the upper bound $\underline{u}^T A \underline{u} \preceq \underline{u}^T C \underline{u}$ follows from the finite overlap of the spaces $V_H, V_1, \ldots V_M$. To prove the lower bound, we come up with an explicit decomposition. We split the minimization into two parts:

$$\underline{u}^T C \underline{u} = \inf_{u_H \in V_H} \inf_{\substack{u_i \in V_i \\ u - u_H = \sum u_i}} \|u_H\|_A^2 + \sum_{i=1}^M \|u_i\|_A^2 \tag{5.4}$$

In the analysis of the DD precondition without coarse grid system we have observed that

$$\inf_{\substack{u_i \in V_i \\ u - u_H = \sum u_i}} \sum_{i=1}^M \|u_i\|_A^2 \preceq H^{-2} \|u - u_H\|_{L_2}^2 + \|\nabla(u - u_H)\|_{L_2}^2$$

Using this in (5.4) gives

$$\begin{aligned} \underline{u}^T C \underline{u} &\preceq \inf_{u_H \in V_H} \left\{ \|u_H\|_A^2 + H^{-2} \|u - u_H\|_{L_2}^2 + \|\nabla(u - u_H)\|_{L_2}^2 \right\} \\ &\preceq \inf_{u_H \in V_H} \left\{ \|\nabla u_H\|_{L_2}^2 + H^{-2} \|u - u_H\|_{L_2}^2 + \|\nabla u\|_{L_2}^2 \right\} \end{aligned}$$

To continue, we introduce a Clément operator $\Pi_H : H^1 \to V_H$ being continuous in the H^1-semi-norm, and approximating in L_2-norm:

$$\|\nabla \Pi_H u\|_{L_2}^2 + H^{-2} \|u - \Pi_H u\|_{L_2}^2 \preceq \|\nabla u\|_{L_2}^2$$

Choosing now $u_H := \Pi_H u$ in the minimization problem we obtain the result:

$$\begin{aligned} \underline{u}^T C \underline{u} &\preceq \|\nabla \Pi_H u\|_A^2 + H^{-2} \|u - \Pi_H u\|_{L_2}^2 + \|\nabla u\|_{L_2}^2 \\ &\preceq \|\nabla u\|^2 \simeq \|u\|_A^2 \end{aligned}$$

□

The inverse factor H^{-2} we have to pay for the local decomposition could be compensated by the approximation on the coarse grid.

The costs for the setup depend on the underlying direct solver for the coarse grid problem and the local problems. Let the factorization step have time complexity N^α. Let

N be the number of unknowns at the fine grid, and M the number of sub-domains. Then the costs to factor the coarse grid problem and the M local problems are of order

$$M^\alpha + M\left(\frac{N}{M}\right)^\alpha$$

Equilibrating both terms gives the optimal choice of number of sub-domains

$$M = N^{\frac{\alpha}{2\alpha-1}},$$

and the asymptotic costs

$$N^{\frac{\alpha^2}{2\alpha-1}}.$$

Example: A Cholesky factorization using bandwidth optimization for 2D problems has time complexity N^2. The optimal choice is $M = N^{2/3}$, leading to the costs of

$$N^{4/3}.$$

Multi-level preconditioners

The preconditioner above uses two grids, the fine one where the equations are solved, and an artificial coarse grid. Instead of two grids, one can use a whole hierarchy of grids $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_\mathcal{L} = \mathcal{T}$. The according finite element spaces are

$$V_0 \subset V_1 \subset \ldots \subset V_L = V_h.$$

Let E_l be the prolongation matrix from level l to the finest level L. Define

$$A_l = E_l^T A E_l \qquad \text{and} \qquad D_l = \operatorname{diag}\{A_l\}.$$

Then, the multi-level preconditioner is

$$C^{-1} = E_0 A_0^{-1} E_0^T + \sum_{l=1}^{L} E_l D_l^{-1} E_l^T$$

The setup, and the application of the preconditioner takes $O(N)$ operations. One can show that the multi-level preconditioner fulfills optimal spectral bounds

$$\underline{u}^T C \underline{u} \preceq \underline{u}^T A \underline{u} \preceq \underline{u}^T C \underline{u}.$$

An iterative method with multi-level preconditioning solves the matrix equation $Au = f$ of size N with $O(N)$ operations !

5.4 Analysis of the multi-level preconditioner

We want to solve a finite element system on $V_L := V_h \subset H^1$. To define the multi-level preconditioner $C = C_L$, we use also finite element spaces on coarser meshes $\mathcal{T}_0, \mathcal{T}_1, \ldots \mathcal{T}_L$:

$$V_0 \subset V_1 \subset \ldots \subset V_L$$

Assume $h_l \simeq 2^{-l}$. Let $\{\varphi_{l,i} : 1 \leq i \leq N_l\}$ be the hat-basis for V_l, with $N_l = \dim V_l$. Let A_l be the finite element matrix on V_l.

$E_l \in \mathbb{R}^{N_l \times N_{l-1}}$ is the prolongation matrix from level $l-1$ to level l.

The multi-level preconditioner is defined recursively:

$$\begin{aligned} C_0^{-1} &:= A_0^{-1} \\ C_l^{-1} &:= (\operatorname{diag} A_l)^{-1} + E_l C_{l-1}^{-1} E_l^T \quad 1 \leq l \leq L. \end{aligned}$$

The computational complexity of one application of C_L^{-1} is $O(N)$ operations.

(An extended version of) the Additive Schwarz Lemma allows to rewrite

$$\begin{aligned} \|u_l\|_{C_l}^2 &= \inf_{\substack{u_l = u_{l-1} + \sum_{i=1}^{N_l} u_{l,i} \\ u_{l-1} \in V_{l-1}, u_{l,i} \in \operatorname{span}\{\varphi_{l,i}\}}} \|u_{l-1}\|_{C_{l-1}}^2 + \sum_{i=1}^{N_l} \|u_{l,i}\|_A^2 \\ &= \inf_{u = u_0 + \sum_{l=1}^{L} \sum_{i=1}^{N_l} u_{l,i}} \sum_l \sum_i \|u_{l,i}\|_A^2 + \|u_0\|_A^2 \end{aligned}$$

Reordering the minimization we obtain

$$\begin{aligned} \|u\|_{C_L}^2 &= \inf_{\substack{u = \sum_{l=0}^{L} u_l \\ u_l \in V_l}} \|u_0\|_A^2 + \sum_{l=1}^{L} \inf_{u_l = \sum u_{l,i}} \sum_{i=1}^{N_l} \|u_{l,i}\|_A^2 \\ &\simeq \inf_{\substack{u = \sum_{l=0}^{L} u_l \\ u_l \in V_l}} \|u_0\|_A^2 + \sum_{l=1}^{L} h_l^{-2} \|u_l\|_{L_2}^2 \end{aligned}$$

Lemma 98 (simple analysis)**.**

$$\frac{1}{L} C \preceq A \preceq LC$$

Proof. $A \preceq LC$ follows from maximal overlap of spaces and the inverse estimate $\|\nabla u_l\|_{L_2} \preceq h_l^{-1} \|u_l\|_{L_2}$. Let $u = \sum_{l=0}^{L} u_l$ be an arbitrary decomposition:

$$\|\sum_{l=0}^{L} u_l\|_A^2 \leq (L+1) \sum_{l=0}^{L} \|u_l\|_A^2 \preceq L\big(\|u_0\|_A^2 + \sum_{l=1}^{L} h_l^{-2} \|u_l\|_{L_2}^2\big).$$

Since the estimate holds for any decompositon, it also holds for the infimum.

To show $C \preceq LA$ we come up with an explicit decomposition of $u \in V_L$. Let $\Pi_l : L_2 \to V_l$ be a Clément-type operator which is a projection and satisfies

$$\|\Pi_l u\|_{H^1} + h_l^{-1} \|u - \Pi_l u\|_{L_2} \preceq \|u\|_{H^1} \qquad \forall\, u \in H^1.$$

Define

$$\begin{aligned} u_0 &:= \Pi_0 u \\ u_l &:= \Pi_l u - \Pi_{l-1} u \qquad 1 \leq l \leq L. \end{aligned}$$

Then $u = \sum_{l=0}^L u_l$ and

$$\|u\|_C^2 \preceq \|\Pi_0 u\|_A^2 + \sum_{l=1}^L h_l^{-2} \|\Pi_l u - \Pi_{l-1} u\|_{L_2}^2 \preceq L\, \|u\|_{H^1}^2 \approx L\, \|u\|_A^2$$

We have bound each of the $L+1$ terms by the H^1-norm of u, thus the factor L. □

Next we show an improved estimate leading to the optimal condition number $\kappa(C^{-1}A) \preceq 1$, independent of the number of refinement levels:

Lemma 99. *There holds*

$$C \preceq A \preceq C$$

Proof. We show $A \preceq C$. Let $u = \sum_{l=0}^L u_l$ an arbitrary decomposition. First, we split up the coarsest level:

$$\|u\|_A^2 \leq \|u_0\|_A^2 + \|\sum_{l=1}^L u_l\|_A^2$$

Next we show the estimate

$$A(u_l, v_k) \leq 2^{-\frac{|l-k|}{2}} h_l^{-1} \|u_l\| h_k^{-1} \|v_k\| \qquad \forall\, u_l \in V_l,\ v_k \in V_k$$

We assume $l \leq k$. We perform integration by parts on the level-l triangles, and apply Cauchy-Schwarz and scaling techniques:

$$\begin{aligned} A(u_l, v_k) &= \sum_{T \in \mathcal{T}_l} \int_T \nabla u_l \nabla v_k \\ &\leq \sum_T \int_T \underbrace{-\Delta u_l}_{=0} v_k + \int_{\partial T} \frac{\partial u_l}{\partial n} v_k \\ &\leq \sum_T \left\| \frac{\partial u_l}{\partial n} \right\|_{\partial T_l} \|v_k\|_{\partial T_l} \\ &\leq h_l^{-3/2} \|u_l\|_{L_2} h_k^{-1/2} \|v_k\|_{L_2} \\ &= \underbrace{\sqrt{h_k/h_l}}_{\simeq 2^{-|k-l|/2}} h_l^{-1} \|u_l\|_{L_2} h_k^{-1} \|v_k\|_{L_2} \end{aligned}$$

We define the overlap - matrix $\mathcal{O} \in \mathbb{R}^{L \times L}$ as

$$\mathcal{O}_{kl} = 2^{-|k-l|/2}.$$

Then

$$\begin{aligned}
\| \sum_{l=1}^{L} u_l \|_A^2 &= \sum_{l,k=1}^{N} A(u_l, u_k) \preceq \sum_{l,k} \mathcal{O}_{kl} h_k^{-1} \|u_k\|_{L_2} \, h_l^{-1} \|u_l\|_{L_2} \\
&\leq \rho(\mathcal{O}) \sum_{l=1}^{L} h_l^{-2} \|u_l\|_{L_2}^2
\end{aligned}$$

The spectral radius $\rho(\mathcal{O})$ can be estimated by the row-sum-norm, which is bounded by a convergent geometric sequence

$$\sum_{k=1}^{L} 2^{-|k-l|/2} \leq 2 \sum_{k=0}^{\infty} \sqrt{2}^{-k} \leq \frac{2}{1 - \sqrt{2}}.$$

Since the decomposition was arbitrary, the estimate holds for the minimal decomposition.

Now we show $C \preceq A$. We procede similar as above. Let $\Pi_l : L_2 \rightarrow V_l$ be an Clément-type operator such that

$$\begin{aligned}
\|\Pi_l u\|_{L_2} &\leq \|u\|_{L_2} \qquad \forall\, u \in L_2 \\
\|u - \Pi_l u\|_{L_2} &\preceq h_l^2 \|u\|_{H^2} \quad \forall\, u \in H^2.
\end{aligned}$$

We define $u_0 = \Pi_0 u$ and $u_l = \Pi_l u - \Pi_{l-1} u$. We obtain the 2 estimates

$$\begin{aligned}
h_l^{-2} \|u_l\|_{L_2}^2 &\preceq h_l^{-2} \|u\|_{L_2}^2, \\
h_l^{-2} \|u_l\|_{L_2}^2 &\preceq h_l^2 \|u\|_{H^2}.
\end{aligned}$$

The idea of the proof is that H^1 is the interpolation space $[L_2, H^2]_{1/2}$. We define the K-functional

$$K(t, u)^2 = \inf_{\substack{u = u_0 + u_2 \\ u_0 \in L_2, u_2 \in H^2}} \left\{ \|u_0\|_{L_2}^2 + t^2 \|u_2\|_{H^2}^2 \right\}.$$

Combining the 2 estimates above we get

$$h_l^{-2} \|u_l\|_{L_2}^2 \preceq h_l^{-2} K^2(h_l^2, u)$$

Thus, the sum over L levels is

$$\sum_{l=1}^{L} h_l^{-2} \|u_l\|_{L_2}^2 \leq \sum_{l=1}^{L} h_l^{-2} K^2(h_l^2, u) \simeq \sum_{l=1}^{L} 2^l K^2(2^{-l}, u)$$

Next we use that $K(s,.) \simeq K(t,.)$ for $t \leq s \leq 2t$ and replace the sum by an integral, and substitute $t := 2^{-l}, dt \simeq -2^{-l} dl = -t dl$:

$$\begin{aligned}
\sum_{l=1}^{L} h_l^{-2} \|u_l\|_{L_2}^2 &\leq \int_{l=1}^{L+1} 2^l K^2(2^{-l}, u) dl \\
&\simeq \int_{2^{-L-1}}^{1} t^{-1} K^2(t, u) \frac{dt}{t} \\
&\preceq \int_{0}^{\infty} t^{-1} K^2(t, u) \frac{dt}{t} \\
&= \|u\|_{[L_2, H^2]_{1/2}}^2 \simeq \|u\|_{H^1}^2
\end{aligned}$$

$\square$

An intuitive explanation of the proof is that different terms of the sum $\sum_{l=1}^{L} h_l^{-2} \|\Pi_l u - \Pi_{l-1} u\|_{L_2}^2$ are dominated by different frequency components of u. The squared H^1-norm is the sum over H^1-norms of the individual frequency components.

Chapter 6

Mixed Methods

A mixed method is a variational formulation involving two function spaces, and a bilinear-form of a special saddle point structure. Usually, it is obtained from variational problems *with constraints.*

6.1 Weak formulation of Dirichlet boundary conditions

We start with the Poisson problem

$$-\Delta u = f \qquad \text{in } \Omega, \tag{6.1}$$

and boundary conditions

$$\begin{array}{rcll} u & = & u_D & \text{on } \Gamma_D, \\ \frac{\partial u}{\partial n} & = & 0 & \text{on } \Gamma_N. \end{array}$$

In contrast to the earlier method, we multiply equation (6.1) with test functions $v \in H^1$ (without imposing Dirichlet constraints), and integrate by parts. Using the Neumann boundary conditions, we obtain

$$\int_\Omega \nabla u \nabla v \, dx - \int_{\Gamma_D} \frac{\partial u}{\partial n} v \, ds = \int_\Omega f v \, dx$$

The normal derivative $\frac{\partial u}{\partial n}$ is not known on Γ_D. We simply call it $-\lambda$:

$$\lambda := -\frac{\partial u}{\partial n}$$

To pose the Dirichlet boundary condition, we multiply $u = u_D$ by sufficiently many test functions, and integrate over Γ_D:

$$\int_{\Gamma_D} u \mu \, ds = \int_{\Gamma_D} u_D \mu \, ds \qquad \forall \, \mu \in ?$$

Combining both equations, we get the system of equations: Find $u \in V = H^1(\Omega)$ and $\lambda \in Q =?$ such that

$$\begin{array}{lcll} \int_\Omega \nabla u \cdot \nabla v \, dx \;+\; \int_{\Gamma_D} v\lambda \, ds & = & \int f v \, dx & \forall\, v \in V, \\ \int_{\Gamma_D} u\mu \, ds & = & \int_{\Gamma_D} u_D \mu \, ds & \forall\, \mu \in Q. \end{array} \tag{6.2}$$

A similar formulation can be obtained for interface conditions.

6.2 A Mixed method for the flux

We start from the second order pde

$$\operatorname{div}(a\nabla u) = f \qquad \text{in } \Omega,$$

and boundary conditions

$$\begin{aligned} u &= u_D \qquad \text{on } \Gamma_D \\ a\frac{\partial u}{\partial n} &= g \qquad \text{on } \Gamma_N \end{aligned}$$

Next, we introduce the flux variable $\sigma := a\nabla u$ to rewrite the equations as: Find u and σ such that

$$a^{-1}\sigma - \nabla u = 0, \tag{6.3}$$

$$\operatorname{div}\,\sigma = -f, \tag{6.4}$$

and boundary conditions

$$\begin{aligned} u &= u_D \qquad \text{on } \Gamma_D \\ \sigma \cdot n &= g \qquad \text{on } \Gamma_N. \end{aligned}$$

We want to derive a variational formulation for the system of equations. For that, we multiply the first equations by vector-valued test functions τ, the second equation by test functions v, and integrate:

$$\begin{array}{lcll} \int_\Omega (a^{-1}\sigma)\cdot\tau \, dx \;-\; \int_\Omega \tau \cdot \nabla u \, dx & = & 0 & \forall\, \tau \\ \int_\Omega \operatorname{div}\,\sigma\, v \, dx & = & -\int f v \, dx & \forall\, v \end{array}$$

We would like to have the second term of the first equation of the same structure as the first term in the second equation. This can be obtained by integration by parts applied to either one of them. The interesting case is to integrate by parts in the first line to obtain:

$$\int_\Omega (a^{-1}\sigma)\cdot\tau \, dx + \int_\Omega \operatorname{div}\,\tau\, u \, dx - \int_{\Gamma_D} \tau_n\, u \, ds - \int_{\Gamma_N} \tau_n\, u \, ds = 0.$$

Here, we make use of the boundary conditions. On the Dirichlet boundary, we know $u = u_D$, and use that in the equation. The Neumann boundary condition $\sigma \cdot n = g$ must be put into the approximation space, it becomes an essential boundary condition. Thus, it is enough to choose test functions of the sub-space fulfilling $\tau \cdot n = 0$. The problem is now the following. The space V will be fixed later. Find $\sigma \in V, \sigma_n = g$ on Γ_N, and $u \in Q$ such that

$$\begin{array}{lclcll} \int_\Omega (a^{-1}\sigma) \cdot \tau \, dx & + & \int_\Omega \operatorname{div} \tau \, u \, dx & = & \int_{\Gamma_D} u_D \tau_n \, ds & \forall \, \tau, \ \tau_n = 0 \text{ on } \Gamma_N \\ \int_\Omega \operatorname{div} \sigma \, v \, dx & & & = & -\int f v \, dx & \forall \, v \end{array}$$

The derivatives are put onto the flux unknown σ (and its test function τ). We don't have to derive the primal unknown u. This will give us better approximation for the fluxes than for the scalar. That is one of the reasons to use this mixed method.

6.3 Abstract theory

A mixed variational formulation involves two Hilbert spaces V and Q, bilinear-forms

$$\begin{array}{rcl} a(u,v) & : & V \times V \to \mathbb{R}, \\ b(u,q) & : & V \times Q \to \mathbb{R}, \end{array}$$

and continuous linear-forms

$$\begin{array}{rcl} f(v) & : & V \to \mathbb{R}, \\ g(q) & : & Q \to \mathbb{R}. \end{array}$$

The problem is to find $u \in V$ and $p \in Q$ such that

$$\begin{array}{lclcll} a(u,v) & + & b(v,p) & = & f(v) & \forall \, v \in V, \\ b(u,q) & & & = & g(q) & \forall \, q \in Q. \end{array} \tag{6.5}$$

The two examples from above are of this form.

Instead of considering this as a system of equations, one can look at the mixed method as one variational problem on the product spaces $V \times Q$. For this, simply add both lines, and search for $(u,p) \in V \times Q$ such that

$$a(u,v) + b(u,q) + b(v,p) = f(v) + g(q) \qquad \forall \, (v,q) \in V \times Q.$$

Define the big bilinear-form $B(.,.) : (V \times Q) \times (V \times Q) \to \mathbb{R}$ as

$$B((u,p),(v,q)) = a(u,v) + b(u,q) + b(v,p),$$

to write the whole system as single variational problem

$$\text{Find } (u,p) \in V \times Q : \qquad B((u,p),(v,q)) = f(v) + g(q) \qquad \forall \, (v,q) \in V \times Q$$

By the Riesz-representation theorem, we can define operators:

$$\begin{array}{llllll} A: & V \to V: & u \to Au: & (Au,v)_V & = & a(u,v) \quad \forall\, v \in V \\ B: & V \to Q: & u \to Bu: & (Bu,q)_Q & = & b(u,q) \quad \forall\, q \in Q \\ B^*: & Q \to V: & p \to B^*p: & (B^*p,v)_V & = & b(v,p) \quad \forall\, v \in V. \end{array}$$

By means of these operators, we can write the mixed variational problem as operator equation

$$\begin{array}{lll} Au & + \quad B^*p & = J_V f, \\ Bu & & = J_Q g. \end{array} \tag{6.6}$$

Here, we used the Riesz-isomorphisms $J_V : V^* \to V$ and $J_Q : Q^* \to Q$.

In the interesting examples, the operator B has a large kernel:

$$V_0 := \{v : Bv = 0\}$$

Lemma 100. *Assume that B^*Q is closed in V. Then there holds the V-orthogonal decomposition*

$$V = V_0 + B^*Q$$

Proof: There holds

$$\begin{array}{lll} V_0 & = & \{v : Bv = 0\} \\ & = & \{v : (Bv,q)_Q = 0 \quad \forall\, q \in Q\} \\ & = & \{v : (v, B^*q)_V = 0 \quad \forall\, q \in Q\}. \end{array}$$

This means, V_0 is the V-orthogonal complement to B^*Q. □

Now, we will give conditions to ensure a unique solution of a mixed problem:

Theorem 101 (Brezzi's theorem)**.** *Assume that $a(.,.)$ and $b(.,.)$ are continuous bilinear-forms*

$$a(u,v) \leq \alpha_2 \|u\|_V \|v\|_V \qquad \forall\, u,v \in V, \tag{6.7}$$

$$b(u,q) \leq \beta_2 \|u\|_V \|q\|_Q \qquad \forall\, u \in V,\ \forall\, q \in Q. \tag{6.8}$$

Assume there holds coercivity of $a(.,.)$ on the kernel,i.e.,

$$a(u,u) \geq \alpha_1 \|u\|_V^2 \qquad \forall\, u \in V_0, \tag{6.9}$$

and there holds the LBB (Ladyshenskaja-Babuška-Brezzi) condition

$$\sup_{u \in V} \frac{b(u,q)}{\|u\|_V} \geq \beta_1 \|q\|_Q \quad \forall\, q \in Q. \tag{6.10}$$

Then, the mixed problem is uniquely solvable. The solution fulfills the stability estimate

$$\|u\|_V + \|p\|_Q \leq c\{\|f\|_{V^*} + \|g\|_{Q^*}\},$$

with the constant c depending on $\alpha_1, \alpha_2, \beta_1, \beta_2$.

Proof: The big bilinear-form $B(.,.)$ is continuous

$$B((u,p),(v,q)) \preceq (\|u\| + \|p\|)\,(\|v\| + \|q\|).$$

We prove that it fulfills the $\inf - \sup$ condition

$$\inf_{v,q} \sup_{u,p} \frac{B((u,p),(v,q))}{(\|v\|_V + \|q\|_Q)(\|u\|_V + \|p\|_Q)} \geq \beta.$$

Then, we use Theorem 33 (by Babuška-Aziz) to conclude continuous solvability.

To prove the $\inf - \sup$-condition, we choose arbitrary $v \in V$ and $q \in Q$. We will construct $u \in V$ and $p \in Q$ such that

$$\|u\|_V + \|p\|_Q \preceq \|v\|_V + \|q\|_Q$$

and

$$B((u,p),(v,q)) = \|v\|_V^2 + \|q\|_Q^2.$$

First, we use (6.10) to choose $u_1 \in V$ such that

$$b(u_1, q) = \|q\|_Q^2 \qquad \text{and} \qquad \|u_1\|_V \leq 2\beta_1^{-1}\,\|q\|_Q.$$

Next, we solve a problem on the kernel:

$$\text{Find } u_0 \in V_0: \quad a(u_0, w_0) = (v, w_0)_V - a(u_1, w_0) \qquad \forall\, w_0 \in V_0$$

Due to assumption (6.9), the left hand side is a coercive bilinear-form on V_0. The right hand side is a continuous linear-form. By Lax-Milgram, the problem has a unique solution fulfilling

$$\|u_0\|_V \preceq \|v\|_V + \|u_1\|_V$$

We set

$$u = u_0 + u_1.$$

By the Riesz-isomorphism, we define a $z \in V$ such that

$$(z, w)_V = (v, w)_V - a(u, w) \qquad \forall\, w \in V$$

By construction, it fulfills $z \perp_V V_0$. The LBB condition implies

$$\|p\|_Q \leq \beta_1^{-1} \sup_v \frac{b(v,p)}{\|v\|_V} = \beta_1^{-1} \sup_v \frac{(v, B^*p)_V}{\|v\|_V} = \beta_1^{-1} \|B^*p\|_V,$$

and thus B^*Q is closed, and $z \in B^*Q$. Take the $p \in Q$ such that

$$z = B^*p.$$

It fulfills

$$\|p\|_Q \leq \beta_1^{-1} \|z\|_V \preceq \|v\|_V + \|q\|_Q$$

Concluding, we have constructed u and p such that

$$\|u\|_V + \|p\|_Q \preceq \|v\|_V + \|q\|_Q,$$

and

$$\begin{aligned} B((u,p),(v,q)) &= a(u,v) + b(v,p) + b(u,q) \\ &= a(u,v) + (z,v)_V + b(u,q) \\ &= a(u,v) + (v,v)_V - a(u,v) + b(u,q) \\ &= \|v\|_V^2 + b(u_1,q) \\ &= \|v\|_V^2 + \|q\|_Q^2. \end{aligned}$$

□

6.4 Analysis of the model problems

Now, we apply the abstract framework to the two model problems.

Weak formulation of Dirichlet boundary conditions

The problem is well posed for the spaces

$$V = H^1(\Omega) \qquad \text{and} \qquad Q = H^{-1/2}(\Gamma_D)$$

Remember, $H^{-1/2}(\Gamma_D)$ is the dual to $H^{1/2}(\Gamma_D)$. The later one is the trace space of $H^1(\Omega)$, the norm fulfills

$$\|u_D\|_{H^{1/2}(\Gamma_D)} \simeq \inf_{\substack{w\in H^1 \\ \text{tr } w = u_D}} \|w\|_{H^1(\Omega)}.$$

The bilinear-forms are

$$\begin{aligned} a(u,v) &= \int_\Omega \nabla u \, \nabla v \, dx \\ b(u,\lambda) &= \langle \lambda, \text{tr } u \rangle_{H^{-1/2}\times H^{1/2}} \end{aligned}$$

To be precise, the integral $\int_{\Gamma_D} \lambda u \, dx$ is extended to the duality product $\langle \lambda, u \rangle$. For regular functions ($\lambda \in L_2(\Gamma_D)$), we can write the L_2-inner product.

Theorem 102. *The mixed problem (6.2) has a unique solution $u \in H^1(\Omega)$ and $\lambda \in H^{-1/2}(\Gamma_D)$.*

Proof: The spaces V and Q, and the bilinear-forms $a(.,.)$ and $b(.,.)$ fulfill the assumptions of Theorem 101. The kernel space V_0 is

$$V_0 = \{u : \int_{\Gamma_D} u\mu \, dx = 0 \quad \forall \, \mu \in L_2(\Gamma_D)\} = \{u : \text{tr}_{\Gamma_D} u = 0\}$$

The continuity of $a(.,.)$ on V is clear. It is not coercive on V, but, due to Friedrichs inequality, it is coercive on V_0.

The bilinear-form $b(.,.)$ is continuous on $V \times Q$:

$$b(u, \mu) = \langle \mu, \text{tr } u \rangle_{H^{-1/2} \times H^{1/2}} \leq \|\mu\|_{H^{-1/2}} \| \text{tr } u\|_{H^{1/2}(\Gamma_D)} \preceq \|\mu\|_Q \|u\|_{H^1} = \|\mu\|_Q \|u\|_V$$

The LBB - condition of $b(.,.)$ follows more or less from the definition of norms:

$$\begin{aligned}
\|q\|_Q &= \sup_{u \in H^{1/2}} \frac{\langle q, u \rangle}{\|u\|_{H^{1/2}}} \\
&\simeq \sup_{u \in H^{1/2}} \frac{\langle q, u \rangle}{\inf_{\substack{w \in H^1(\Omega) \\ \text{tr } w = u}} \|w\|_{H^1(\Omega)}} \\
&= \sup_{u \in H^{1/2}} \sup_{\substack{w \in H^1(\Omega) \\ \text{tr } w = u}} \frac{\langle q, u \rangle}{\|w\|_{H^1}} \\
&= \sup_{w \in H^1} \frac{\langle q, \text{tr } w \rangle}{\|w\|_{H^1}} = \sup_{w \in V} \frac{b(w, q)}{\|w\|_V}
\end{aligned}$$

Mixed method for the fluxes

This mixed method requires the function space $H(\text{div}, \Omega)$:

Definition 103. *A measurable function g is called the weak divergence of σ on $\Omega \subset \mathbb{R}^d$ if there holds*

$$\int_\Omega g \, \varphi \, dx = - \int_\Omega \sigma \cdot \nabla \varphi \, dx \qquad \forall \, \varphi \in C_0^\infty(\Omega)$$

The function space $H(\text{div})$ is defined as

$$H(\text{div}, \Omega) := \{ \sigma \in [L_2(\Omega)]^d : \text{div } \sigma \in L_2 \},$$

its norms is

$$\|\sigma\|_{H(\text{div})} = \left\{ \|\sigma\|_{L_2}^2 + \| \text{div } \sigma\|_{L_2}^2 \right\}^{1/2}$$

The mixed method is formulated on the spaces

$$V = H(\text{div}) \qquad Q = L_2$$

The bilinear-forms are

$$\begin{aligned}
a(\sigma, \tau) &= \int a^{-1} \sigma \tau \, dx \qquad \forall \, \sigma, \tau \in V \\
b(\sigma, v) &= \int \text{div } \sigma \, v \, dx \qquad \forall \, \sigma \in V, \; \forall \, v \in Q
\end{aligned}$$

We assume that the symmetric matrix $a \in \mathbb{R}^{d \times d}$ and its inverse a^{-1} are bounded.

Theorem 104. *The mixed problem for the fluxes is well posed.*

Proof: We check the conditions of the theorem of Brezzi: The bilinear-forms are bounded, namely

$$a(\sigma, \tau) = \int a^{-1} \sigma \tau \, dx \leq \|a^{-1}\|_{L_\infty} \|\sigma\|_{L_2} \|\tau\|_{L_2} \preceq \|\sigma\|_V \|\tau\|_V$$

and

$$b(\sigma, v) = \int \operatorname{div} \, \sigma v \, dx \leq \|\operatorname{div} \sigma\|_{L_2} \|v\|_{L_2} \leq \|\sigma\|_V \|v\|_Q.$$

The kernel space $V_0 = \{\tau : b(\tau, v) = 0 \; \forall \, v \in Q\}$ is

$$V_0 = \{\tau \in H(\operatorname{div}) : \operatorname{div} \, \tau = 0\}$$

There holds the kernel-ellipticity of $a(.,.)$. Let $\tau \in V_0$. Then

$$a(\tau, \tau) = \int \tau^T a^{-1} \tau \, dx \geq \inf_{x \in \Omega} \lambda_{\min}(a^{-1}) \int |\tau|^2 \, dx \succeq \|\tau\|_{L_2}^2 = \|\tau\|_{H(\operatorname{div})}^2$$

We are left to verify the LBB condition

$$\sup_{\sigma \in H(\operatorname{div})} \frac{\int \operatorname{div} \, \sigma \, v \, dx}{\|\sigma\|_{H(\operatorname{div})}} \succeq \|v\|_{L_2} \qquad \forall \, v \in L_2. \tag{6.11}$$

For given $v \in L_2$, we will construct a flux σ satisfying the inequality. For this, we solve the artificial Poisson problem $-\Delta \varphi = v$ with Dirichlet boundary conditions $\varphi = 0$ on $\partial\Omega$. The solution satisfies $\|\nabla \varphi\|_{L_2} \preceq \|v\|_{L_2}$. Set $\sigma = -\nabla \varphi$. There holds div $\sigma = v$. Its norm is

$$\|\sigma\|_{H(\operatorname{div})}^2 = \|\sigma\|_{L_2}^2 + \|\operatorname{div} \, \sigma\|_{L_2}^2 = \|\nabla \varphi\|_{L_2}^2 + \|v\|_{L_2}^2 \preceq \|v\|_{L_2}^2.$$

Using it in (6.11), we get the result

$$\frac{\int \operatorname{div} \, \sigma \, v \, dx}{\|\sigma\|_{H(\operatorname{div})}} = \frac{\int v^2 \, dx}{\|\sigma\|_{H(\operatorname{div})}} \succeq \|v\|_{L_2}.$$

The function space $H(\operatorname{div})$

The mixed formulation has motivated the definition of the function space $H(\operatorname{div})$. Now, we will study some properties of this space. We will also construct finite elements for the approximation of functions in $H(\operatorname{div})$. In Section 3.3.1, we have investigated traces of functions in H^1. Now, we apply similar techniques to the space $H(\operatorname{div})$. Again, the proofs are based on the density of smooth functions.

For a function in H^1, the boundary values are well defined by the trace operator. For a vector-valued function in $H(\operatorname{div})$, only the normal-component is well defined on the boundary:

Theorem 105. *There exists a normal-trace operator*

$$\operatorname{tr}_n : H(\operatorname{div}) \to H^{-1/2}(\partial\Omega)$$

such that for $\sigma \in H(\operatorname{div}) \cap [C(\overline{\Omega})]^d$ *it coincides with its normal component*

$$\operatorname{tr}_n \sigma = \sigma \cdot n \qquad \text{on } \partial\Omega.$$

Proof: For smooth functions, the trace operator gives the normal component on the boundary. We have to verify that this operator is bounded as operator from $H(\operatorname{div})$ to $H^{-1/2}(\partial\Omega)$. Then, by density, we can extend the trace operator to $H(\operatorname{div})$. Let $\sigma \in H(\operatorname{div}) \cap [C^1(\overline{\Omega})]^d$:

$$\begin{aligned}
\| \operatorname{tr}_n \sigma \|_{H^{-1/2}} &= \sup_{\varphi \in H^{1/2}(\partial\Omega)} \frac{\int_{\partial\Omega} \sigma \cdot n \, \varphi \, ds}{\|\varphi\|_{H^{1/2}}} \simeq \sup_{\varphi \in H^1(\Omega)} \frac{\int_{\partial\Omega} \sigma \cdot n \operatorname{tr} \varphi \, ds}{\|\varphi\|_{H^1}} \\
&= \sup_{\varphi \in H^1(\Omega)} \frac{\int_{\partial\Omega} (\sigma \operatorname{tr} \varphi) \cdot n \, ds}{\|\varphi\|_{H^1}} = \sup_{\varphi \in H^1(\Omega)} \frac{\int_\Omega \operatorname{div}(\sigma\varphi) \, dx}{\|\varphi\|_{H^1}} \\
&= \sup_{\varphi \in H^1(\Omega)} \frac{\int_\Omega (\operatorname{div}\sigma)\varphi \, dx + \int_\Omega \sigma \cdot \nabla\varphi \, dx}{\|\varphi\|_{H^1}} \leq \sup_{\varphi \in H^1(\Omega)} \frac{\|\operatorname{div}\sigma\|_{L_2} \|\varphi\|_{L_2} + \|\sigma\|_{L_2} \|\nabla\varphi\|_{L_2}}{\|\varphi\|_{H^1}} \\
&\leq \left\{ \|\sigma\|_{L_2}^2 + \|\operatorname{div} \sigma\|_{L_2}^2 \right\}^{1/2} = \|\sigma\|_{H(\operatorname{div})}
\end{aligned}$$

□

Lemma 106. *There holds integration by parts*

$$\int_\Omega \sigma \cdot \nabla\varphi \, dx + \int_\Omega (\operatorname{div} \sigma)\varphi \, dx = \langle \operatorname{tr}_n \sigma, \operatorname{tr} \varphi \rangle_{H^{-1/2} \times H^{1/2}}$$

for all $\sigma \in H(\operatorname{div})$ *and* $\varphi \in H^1(\Omega)$.

Proof: By density of smooth functions, and continuity of the trace operators.

Now, let $\Omega_1, \ldots \Omega_M$ be a non-overlapping partitioning of Ω. In Section 3.3.1, we have proven that functions which are in $H^1(\Omega_i)$, and which are continuous across the boundaries $\gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$, are in $H^1(\Omega)$. A similar property holds for functions in $H(\operatorname{div})$.

Theorem 107. *Let* $\sigma \in [L_2(\Omega)]^d$ *such that*

- $$\sigma|_{\Omega_i} \in H(\operatorname{div}, \Omega_i)$$

- $$\operatorname{tr}_{n,i} \sigma|_{\Omega_i} = -\operatorname{tr}_{n,j} \sigma|_{\Omega_j} \qquad \text{on } \gamma_{ij}.$$

Then $\sigma \in H(\operatorname{div}, \Omega)$, *and*

$$(\operatorname{div} \sigma)|_{\Omega_i} = \operatorname{div} (\sigma|_{\Omega_i}).$$

The proof follows the lines of Theorem 46.

We want to compute with functions in $H(\mathrm{div})$. For this, we need finite elements for this space. The characterization by sub-domains allows the definition of finite element sub-spaces of $H(\mathrm{div})$. Let $\mathcal{T} = \{T\}$ be a triangulation of Ω. One family of elements are the BDM (Brezzi-Douglas-Marini) elements. The space is

$$V_h = \{\sigma \in [L_2]^2 : \sigma|_T \in [P^k]^d,\ \sigma \cdot n \text{ continuous across edges}\}.$$

This finite element space is larger than the piece-wise polynomial H^1-finite element space of the same order. The finite element functions can have non-continuous tangential components across edges.

The cheapest element for $H(\mathrm{div})$ is the lowest order Raviart-Thomas element RT0. The finite element $(T, V_T, \{\psi_i\})$ is defined by the space of shape functions V_T, and linear functionals ψ_i. The element space is

$$V_T = \left\{ \begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x \\ y \end{pmatrix} : a, b, c \in \mathbb{R} \right\},$$

the linear functionals are the integrals of the normal components on the three edges of the triangle

$$\psi_i(\sigma) = \int_{e_i} \sigma \cdot n \, ds \qquad i = 1, 2, 3$$

The three functionals are linearly independent on V_T. This means, for each choice of $\sigma_1, \sigma_2, \sigma_3$, there exists three unique numbers $a, b, c \in \mathbb{R}$ such that

$$\sigma = \begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x \\ y \end{pmatrix}.$$

satisfies $\psi_i(\sigma) = \sigma_i$.

Exercise: Compute the shape functions for the RT0 - reference triangle.

The global finite element functions are defined as follows. Given one value σ_i for each edge e_i of the triangulation. The corresponding RT0 finite element function σ is defined by

$$\sigma|_T \in V_T \qquad \text{and} \qquad \int_{e_i} \sigma|_T \cdot n_{e_i} \, ds = \sigma_i$$

for all edges $e_i \subset T$ and all triangles $T \in \mathcal{T}$.

We have to verify that this construction gives a function in $H(\mathrm{div})$. For each element, $\sigma|_T$ is a linear polynomial, and thus in $H(\mathrm{div}, T)$. The normal components must be continuous. By construction, there holds

$$\int_e \sigma|_{T,i} \cdot n \, ds = \int_e \sigma|_{T,j} \cdot n \, ds$$

for the edge $e = T_i \cap T_j$. The normal component is continuous since $\sigma \cdot n_e$ is constant on an edge: Points (x, y) on the edge e fulfill $xn_x + yn_y$ is constant. There holds

$$\sigma \cdot n_e = \left[\begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x \\ y \end{pmatrix} \right] \cdot \begin{pmatrix} n_x \\ n_y \end{pmatrix} = an_x + bn_y + c(xn_x + yn_y) = \text{constant}$$

The global RT0-basis functions φ_i^{RT} are associated to the edges, and satisfy

$$\int_{e_i} \varphi_j^{RT} \cdot n_e \, ds = \delta_{ij} \qquad \forall\, i, j = 1, \ldots N_{edges}$$

By this basis, we can define the RT - interpolation operator

$$I_h^{RT} \sigma = \sum_{edges\ e_i} \left(\int_{e_i} \sigma \cdot n_e \, ds \right) \varphi_i^{RT}$$

It is a projection on V_h. The interpolation operator preserves the divergence *in mean*:

Lemma 108. *The RT0 - interpolation operator satisfies*

$$\int_T \operatorname{div} I_h \sigma \, dx = \int_T \operatorname{div} \, \sigma dx$$

for all triangles $T \in \mathcal{T}$.

Let P_h be the L_2 projection onto piece-wise constant finite element functions. This is: Let $Q_h = \{q \in L_2 : q|_T = \text{const} \ \forall T \in \mathcal{T}\}$. Then $P_h p$ is defined by $P_h p \in Q_h$ and $\int_\Omega P_h p \, q_h \, dx = \int_\Omega P p \, q_h \, dx \ \forall\, q_h \in Q_h$. This is equivalent to $P_h p$ satisfies $P_h p \in Q_h$ and

$$\int_T P_h p \, dx = \int_T p \, dx \qquad \forall\, T \in \mathcal{T}.$$

The Raviart-Thomas finite elements are piecewise linear. Thus, the divergence is piecewise constant. From div $I_h \sigma \in Q_h$ and Lemma 108 there follows

$$\operatorname{div} \, I_h \sigma = P_h \operatorname{div} \, \sigma.$$

This relation is known as commuting diagram property:

$$\begin{array}{ccc} H(\operatorname{div}) & \xrightarrow{\operatorname{div}} & L^2 \\ \Big\downarrow I_h & & \Big\downarrow P_h \\ V_h^{RT} & \xrightarrow{\operatorname{div}} & Q_h \end{array} \tag{6.12}$$

The analysis of the approximation error is based on the transformation to the reference element. For H^1 finite elements, interpolation on the element T is equivalent to interpolation on the reference element $\widehat{T}$, i.e., $(I_h v) \circ F_T = \hat{I}_h(v \circ F_T)$. This is not true for the $H(\operatorname{div})$ elements: The transformation F changes the direction of the normal vector. Thus $\int_e \sigma \cdot n \, ds \neq \int_{\hat{e}} \hat{\sigma} \cdot \hat{n} \, ds$.

The Piola transformation is the remedy:

Definition 109 (Piola Transformation)**.** *Let $F : \widehat{T} \to T$ be the mapping from the reference element $\widehat{T}$ to the element T. Let $\hat\sigma \in L_2(\widehat{T})$. Then, the Piola transformation*

$$\sigma = \mathcal{P}\{\hat\sigma\}$$

is defined by

$$\sigma(F(\hat x)) = (\det F')^{-1} F' \hat\sigma(\hat x).$$

The Piola transformation satisfies:

Lemma 110. *Let $\hat\sigma \in H(\operatorname{div}, \widehat{T})$, and $\sigma = \mathcal{P}\{\hat\sigma\}$. Then there holds*

$$(\operatorname{div}\,\sigma)(F(\hat x)) = (\det F')^{-1} \operatorname{div}\,\hat\sigma$$

Let $\hat e$ be an edge of the reference element, and $e = F(\hat e)$. Then

$$\int_e \sigma \cdot n\, ds = \int_{\hat e} \hat\sigma \cdot \hat n\, ds$$

Proof: Let $\widehat\varphi \in C_0^\infty(\widehat{T})$, and $\varphi(F(\hat x)) = \widehat\varphi(x)$. Then there holds

$$\begin{aligned}
\int_T \operatorname{div}\,\sigma\,\varphi\, dx &= \int_T \sigma \cdot \nabla\varphi\, dx \\
&= \int_{\widehat{T}} \left[(\det F')^{-1} F'\sigma\right] \cdot \left[(F')^{-T} \nabla\hat\varphi\right] (\det F')\, d\hat x \\
&= \int_{\widehat{T}} \hat\sigma \nabla\hat\varphi\, d\hat x = \int_{\widehat{T}} \operatorname{div}\hat\sigma\,\hat\varphi\, d\hat x \\
&= \int_T (\det F')^{-1} (\operatorname{div}\hat\sigma)\varphi\, dx.
\end{aligned}$$

Since C_0^∞ is dense in $L_2(T)$, there follows the first claim. To prove the second one, we show that

$$\int_e (\sigma \cdot n)\varphi\, ds = \int_{\hat e} (\hat\sigma \cdot n)\hat\varphi\, dx$$

holds for all $\varphi \in C^\infty(T)$, $\varphi = 0$ on $\partial T \setminus e$. Then, let $\varphi \to 1$ on the edge e:

$$\int_e (\sigma\cdot n)\varphi\, ds = \int_T \operatorname{div}(\sigma\varphi)\, dx = \int_{\widehat{T}} \operatorname{div}(\hat\sigma\hat\varphi)\, d\hat x = \int_{\hat e} (\hat\sigma \cdot \hat n)\hat\varphi\, d\hat s.$$

□

Lemma 111. *The Raviart-Thomas triangle T and the Raviart-Thomas reference triangle are interpolation equivalent:*

$$I_h^{RT} \mathcal{P}\{\hat\sigma\} = \mathcal{P}\{\widehat{I}_h^{RT} \hat\sigma\}$$

Proof: The element spaces are equivalent, i.e., $V_T = \mathcal{P}\{V_{\widehat{T}}\}$, and the functionals $\psi_i(\sigma) = \int_e \sigma \cdot n\, ds$ are preserved by the Piola transformation.

Theorem 112. *The Raviart-Thomas interpolation operator satisfies the approximation properties*

$$\begin{aligned} \|\sigma - I_h^{RT}\sigma\|_{L_2(\Omega)} &\preceq h\,\|\nabla\sigma\|_{L_2(\Omega)} \\ \|\operatorname{div}\sigma - \operatorname{div}\, I_h^{RT}\sigma\|_{L_2(\Omega)} &\preceq h\,\|\nabla\operatorname{div}\sigma\|_{L_2(\Omega)} \end{aligned}$$

Proof: Transformation to the reference element, using that the interpolation preserves constant polynomials, and the Bramble Hilbert lemma. The estimate for the divergence uses the commuting diagram property

$$\|\operatorname{div}(I - I_h^{RT})\sigma\|_{L_2} = \|(I - P_h)\operatorname{div}\,\sigma\|_{L_2} \preceq h\,\|\nabla\,\operatorname{div}\sigma\|_{L_2}$$

□

6.5 Approximation of mixed systems

We apply a Galerkin-approximation for the mixed system. For this, we choose (finite element) sub-spaces $V_h \subset V$ and $Q_h \subset Q$, and define the Galerkin approximation $(u_h, p_h) \in V_h \times Q_h$ by

$$B((u_h, p_h), (v_h, q_h)) = f(v_h) + g(q_h) \qquad \forall\, v_h \in V_h\; \forall\, q_h \in Q_h.$$

Theorem 113. *Assume that the finite element spaces fulfill the discrete stability condition*

$$\inf_{v \in V_h, q \in Q_h} \sup_{u \in V_h, p \in Q_h} \frac{B((u,p),(v,q))}{(\|v\|_V + \|q\|_Q)(\|u\|_V + \|p\|_Q)} \geq \beta. \tag{6.13}$$

Then the discretization error is bounded by the best-approximation error

$$\|u - u_h\|_V + \|p - p_h\|_Q \preceq \inf_{v_h \in V_h, q_h \in Q_h} \{\|u - v_h\|_V + \|p - q_h\|_Q\}$$

Proof: Theorem 36 applied to the big system $B((u,p),(v,q))$. □

The stability on the continuous level $V \times Q$ does not imply the discrete stability ! Usually, one checks the conditions of Brezzi on the discrete level to prove stability of $B(.,.)$ on the discrete levels. The continuity of $a(.,.)$ and $b(.,.)$ are inherited from the continuous levels. The stability conditions have to be checked separately. The discrete kernel ellipticity

$$a(v_h, v_h) \succeq \|v_h\|_V^2 \qquad \forall\, v_h \in V_{0h} = \{v_h \in V_h : b(v_h, q_h) = 0\; \forall\, q_h \in Q_h\}, \tag{6.14}$$

and the discrete LBB condition

$$\sup_{u_h \in V_h} \frac{b(u_h, q_h)}{\|u_h\|_V} \succeq \|q_h\|_Q \qquad \forall\, q_h \in Q_h. \tag{6.15}$$

The discrete LBB condition is posed for less dual variables q_h in $Q_h \subset Q$, but the space in the supremum is also smaller. It does not follow from the LBB condition on the continuous levels.

There is a canonical technique to derive the discrete LBB condition from the continuous one:

Lemma 114. *Assume there exists a quasi-interpolation operator*

$$\Pi_h : V \to V_h$$

which is continuous

$$\|\Pi_h v\|_V \preceq \|v\|_V \qquad \forall\, v \in V,$$

and which satisfies

$$b(\Pi_h v, q_h) = b(v, q_h) \qquad \forall\, q_h \in Q_h.$$

Then, the continuous LBB condition implies the discrete one.

Proof: For all $p_h \in Q_h$ there holds

$$\sup_{v_h \in V_h} \frac{b(v_h, p_h)}{\|v_h\|_V} \geq \sup_{v \in V} \frac{b(\Pi_h v, p_h)}{\|\Pi_h v\|_V} \succeq \sup_{v \in V} \frac{b(v, p_h)}{\|v\|_V} \succeq \|p_h\|_Q$$

$\square$

Approximation of the mixed method for the flux

Choose the pair of finite element spaces, the Raviart Thomas spaces

$$V_h = \{v \in H(\text{div}) : v|_T \in V_T^{\text{RT}}\} \quad \subset \quad V = H(\text{div})$$

and the space of piece-wise constants

$$Q_h = \{q \in L_2 : q|_T \in P^0\} \subset Q = L_2.$$

Pose the discrete mixed problem: Find $(\sigma_h, u_h) \in V_h \times Q_h$ such that

$$\begin{array}{lclcll}
\int_\Omega (a^{-1}\sigma_h) \cdot \tau_h \, dx & + & \int_\Omega \text{div } \tau_h \, u_h \, dx & = & \int_{\Gamma_D} u_D \tau_n \, ds & \forall\, \tau_h \in V_h \\
\int_\Omega \text{div } \sigma_h \, v_h \, dx & & & = & -\int_\Omega f v_h \, dx & \forall\, v_h \in Q_h.
\end{array} \tag{6.16}$$

Lemma 115 (Discrete Stability)**.** *The discrete mixed variational problem (6.16) is well posed.*

Proof: By Brezzi's theorem. Continuity of the bilinear-form and the linear-form follow from the continuous level. We prove the kernel ellipticity: Since

$$\text{div } V_h \subset Q_h,$$

there holds

$$\int \operatorname{div} \sigma_h \, q_h \, dx = 0 \quad \forall \, q_h \in Q_h \quad \Rightarrow \quad \operatorname{div} \sigma_h = 0,$$

and thus $V_{0h} \subset V_0$. In this special case, the discrete kernel ellipticity is simple the restriction of the continuous one to V_{0h}. We are left with the discrete LBB condition. We would like to apply Lemma 114. The quasi-interpolation operator is the Raviar-Thomas interpolation operator I_h^{RT}. The abstract condition

$$b(I_h^{RT} \sigma, v_h) = b(\sigma, v_h) \qquad v_h \in Q_h$$

reads as

$$\int_T \operatorname{div} I_h^{RT} \sigma \, dx = \int_T \operatorname{div} \sigma dx,$$

which was proven in Lemma 108. But, the interpolation operator is not continuous on $H(\operatorname{div})$. The edge-integrals are not well defined on $H(\operatorname{div})$. We have to include the sub-space $[H^1]^d \subset H(\operatorname{div})$. There holds

$$\|I_h^{RT} \sigma\|_{H(\operatorname{div})} \preceq \|\sigma\|_{H^1} \qquad \forall \, \sigma \in [H^1]^d,$$

and the stronger LBB condition (see Section on Stokes below)

$$\sup_{\sigma \in [H^1]^d} \frac{(\operatorname{div} \, \sigma, v)_{L_2}}{\|\sigma\|_{H^1}} \geq \beta \|v\|_{L_2} \qquad \forall \, v \in L_2.$$

We follow the proof of Lemma 114: For all $v_h \in Q_h$ there holds

$$\sup_{\sigma_h \in V_h} \frac{b(\sigma_h, v_h)}{\|\sigma_h\|_V} \geq \sup_{\sigma \in [H^1]^d} \frac{(\operatorname{div} I_h^{RT} \, \sigma, v_h)}{\|I_h^{RT} \sigma\|_V} \succeq \sup_{\sigma \in [H^1]^d} \frac{(\operatorname{div} \, \sigma, v_h)}{\|\sigma\|_{H^1}} \succeq \|v_h\|_{L_2}.$$

Brezzi's theorem now proves that the discrete problem is well posed, i.e., it fulfills the discrete inf-sup condition.

□

Theorem 116 (A priori estimate)**.** *The mixed finite element method for the flux satisfies the error estimates*

$$\|\sigma - \sigma_h\|_{L_2} + \| \operatorname{div}(\sigma - \sigma_h)\|_{L_2} + \|u - u_h\|_{L_2} \preceq h \, \left(\|\sigma\|_{H^1} + \|u\|_{H^1} + \|f\|_{H^1}\right) \tag{6.17}$$

Proof: By discrete stability, one can bound the discretization error by the best approximation error

$$\|\sigma - \sigma_h\|_{H(\operatorname{div})} + \|u - u_h\|_{L_2} \preceq \inf_{\substack{\tau_h \in V_h \\ v_h \in Q_h}} \left\{ \|\sigma - \tau_h\|_{H(\operatorname{div})} + \|u - v_h\|_{L_2} \right\}.$$

The best approximation error is bounded by the interpolation error. The first term is (using the commuting diagram property and $\operatorname{div} \sigma = f$)

$$\inf_{\tau_h \in V_h} \left\{ \|\sigma - \tau_h\|_{L_2} + \| \operatorname{div}(\sigma - \tau_h)\|_{L_2} \right\} \leq \|\sigma - I_h^{RT} \sigma\|_{L_2} + \|(I - P^0) \operatorname{div} \sigma\|_{L_2} \preceq h \, \left(\|\sigma\|_{H^1} + \|f\|_{H^1} \right).$$

The second term is

$$\inf_{v_h\in Q_h}\|u-v_h\|_{L_2}\leq\|u-P^0u\|_{L_2}\preceq h\,\|u\|_{H^1}.$$

$\square$

The smoothness requirements onto the solution of (6.17) are fulfilled for problems on convex domains, and smooth (constant) coefficients a. There holds $\|u\|_{H^2}\preceq\|f\|_{L_2}$. Since $\sigma=a\nabla u$, there follows $\|\sigma\|_{H^1}\preceq\|f\|_{L_2}$. The mixed method requires more smoothness onto the right hand side data, $f\in H^1$. It can be reduced to H^1 on sub-domains, what is a realistic assumption. On non-convex domains, u is in general not in H^2 (and σ not in H^1). Again, weighted Sobolev spaces can be used to prove similar estimates on properly refined meshes.

Approximation of the mixed method for Dirichlet boundary conditions

A possibility is to choose continuous and piece-wise linear finite element spaces on the domain and on the boundary

$$V_h=\{v\in C(\Omega):v|_T\in P^1\qquad\forall\,T\},$$

$$Q_h=\{\mu\in C(\partial\Omega):\mu|_E\in P^1\qquad\forall\,E\subset\partial\Omega\}.$$

Theorem 117. *The discrete mixed method is well posed.*

Proof: Exercises.

6.6 Supplement on mixed methods for the flux : discrete norms, super-convergence and implementation techniques

6.6.1 Primal and dual mixed formulations

A mixed method for the flux can be posed either in the so called primal form: find $\sigma\in V=[L_2]^2$, $u\in H^1$ with $u=u_D$ on Γ_D such that

$$\begin{array}{lclcll}\int_\Omega(a^{-1}\sigma)\cdot\tau\,dx & - & \int_\Omega\tau\cdot\nabla u\,dx & = & 0 & \forall\,\tau,\\ -\int_\Omega\sigma\cdot\nabla v\,dx & & & = & -\int fv\,dx-\int_{\Gamma_N}gv\,ds & \forall\,v,\ v=0\text{ on }\Gamma_D,\end{array}$$

or in the so called dual mixed form: find $\sigma\in V=H(\mathrm{div})$, $u\in L_2$ with $\sigma\cdot n=g$ on Γ_N

$$\begin{array}{lclcll}\int_\Omega(a^{-1}\sigma)\cdot\tau\,dx & + & \int_\Omega\operatorname{div}\,\tau\,u\,dx & = & \int_{\Gamma_D}u_D\tau_n\,ds & \forall\,\tau,\ \tau_n=0\text{ on }\Gamma_N\\ \int_\Omega\operatorname{div}\,\sigma\,v\,dx & & & = & -\int fv\,dx & \forall\,v.\end{array}$$

Both are formally equivalent: If the solutions are smooth enough for integration by parts, both solutions are the same. In both cases, the big-B bilinear-form is $inf-sup$ stable with respect to the corresponding norms.

The natural discretization for the primal-mixed formulation uses standard H^1-finite elements of order k for u, and discontinuous L_2 elements of order $k-1$ for σ. Here, the discrete Brezzi conditions are trivial. The dual one requires Raviart-Thomas (RT) or Brezzi-Douglas-Marini (BDM) elements for σ, and L_2 elements for u. This pairing delivers locally exact conservation ($\int_{\partial T} \sigma_n = -\int_T f$). In particular this property makes the method interesting by itself, but often this scheme is a part of a more complex problem (e.g. Navier-Stokes equations).

Our plan is as follows: We want to use the dual finite element method, but analyze it in a primal - like setting. Since Q_h is no sub-space of H^1, we have to use a discrete counterpart of the H^1-norm:

$$\begin{aligned}
\|\tau\|_{V_h}^2 &:= \|\tau\|_{L_2}^2 \\
\|v\|_{Q_h}^2 &:= \|v\|_{H^1,h}^2 := \sum_T \|\nabla v\|_{L_2(T)}^2 + \sum_{E\subset\Omega} \tfrac{1}{h}\|[v]\|_{L_2(E)}^2 + \sum_{E\subset\Gamma_D} \tfrac{1}{h}\|v\|_{L_2(E)}^2
\end{aligned}$$

The factor $\frac{1}{h}$ provides correct scaling: If we transform an element patch to the reference patch, the jump term scales like the H^1-semi-norm. This norm is called discrete H^1-norm, or DG-norm (as it is essential for Discontinuous Galerkin methods discussed later).

There holds a discrete Friedrichs inequality

$$\|v\|_{L_2} \preceq \|v\|_{H^1,h}.$$

Theorem 118. *The dual-mixed discrete problem satisfies Brezzi's conditions with respect to L_2 and discrete H^1-norms.*

Proof. The $a(.,.)$ bilinear-form is continuous and coercive on $(V_h, \|\cdot\|_{L_2})$. Now we show continuity of $b(.,.)$ on the finite element spaces. We integrate by parts on the elements, and rearrange boundary terms:

$$\begin{aligned}
b(\sigma_h, v_h) &= \int_\Omega \operatorname{div}\,\sigma_h\, v_h = \sum_T \int_T \operatorname{div}\,\sigma_h\, v_h \\
&= \sum_T -\int_T \sigma_h \cdot \nabla v_h + \int_{\partial T} \sigma_h \cdot n\, v_h \\
&= \sum_T -\int_T \sigma_h \cdot \nabla v_h + \sum_{E\subset\Omega} \int_E \sigma_h n_E [v] + \sum_{E\subset\Gamma_D} \int_E \sigma_h n_E v + \sum_{E\subset\Gamma_N} \int_E \underbrace{\sigma_h n_E}_{=0} v
\end{aligned}$$

The jump term is defined as $[v](x) = \lim_{t\to 0^+} v(x+tn_E) - v(x-tn_E)$. Thus, $\sigma_h n_E [v]$ is independent of the direction of the normal vector. Next we apply Cauchy-Schwarz, and use that $h\|\sigma_h \cdot n\|_{L_2(E)}^2 \preceq \|\sigma_h\|_{L_2(T)}$ for some $E \subset T$ (scaling and equivalence of norms on finite dimensional spaces):

$$\begin{aligned} b(\sigma_h, v_h) &\leq \sum_T \|\sigma_h\|_{L_2(T)} \|\nabla v_h\|_{L_2(T)} + \sum_{E \subset \Omega} h^{1/2}\|\sigma_h\|_{L_2(E)}\, h^{-1/2}\|[v_h]\|_{L_2(E)} + \sum_{E \subset \Gamma_D} \ldots \\ &\preceq \|\sigma_h\|_{L_2(\Omega)}\|v_h\|_{H^1,h} \end{aligned}$$

The linear-forms are continuous with norms $h^{-1/2}\|u_D\|_{L_2(\Gamma_D)}$ and $\|f\|_{L_2(\Omega)}$, respectively.

Finally, we show the LBB - condition: Given an $v_h \in Q_h$, we define σ_h as follows:

$$\begin{aligned} \sigma_h \cdot n_E &= \tfrac{1}{h}[v_h] \qquad \text{on } E \subset \Omega \\ \sigma_h \cdot n_E &= \tfrac{1}{h} v_h \qquad \text{on } E \subset \Gamma_D \\ \sigma_h \cdot n_E &= 0 \qquad \text{on } E \subset \Gamma_N \\ \int_T \sigma_h \cdot q &= -\int_T \nabla v_h \cdot q \quad \forall\, q \in [P^{k-1}]^2. \end{aligned}$$

This definition mimics $\sigma = -\nabla v$. This construction is allowed by the definition of Raviart-Thomas finite elements. Thus we get

$$\begin{aligned} b(\sigma_h, v_h) &= \sum_T -\int \sigma_h \underbrace{\nabla v_h}_{\in [P^{k-1}]^2} + \sum_{E \subset \Omega} \tfrac{1}{h}\|[v_h]\|_{L_2(E)}^2 + \sum_{E \subset \Gamma_D} \tfrac{1}{h}\|v_h\|_{L_2(E)}^2 \\ &= \sum_T \int \nabla v_h \cdot \nabla v_h + \sum_{E \subset \Omega} \tfrac{1}{h}\|[v_h]\|_{L_2(E)}^2 + \sum_{E \subset \Gamma_D} \tfrac{1}{h}\|v_h\|_{L_2(E)} \\ &= \|v_h\|_{H^1,h}^2 \end{aligned}$$

By scaling arguments we see that $\|\sigma_h\|_{L_2} \preceq \|v_h\|_{H^1,h}$. Thus we got σ_h such that

$$\frac{b(\sigma_h, v_h)}{\|\sigma_h\|_{L_2}} \succeq \|v_h\|_{H^1,h},$$

and we have constructed the candidate for the LBB condition. □

6.6.2 Super-convergence of the scalar

Typically, the discretization error of mixed methods depend on best-approximation errors in all variables:

$$\|\sigma - \sigma_h\|_{L_2} + \|u - u_h\|_{H^1,h} \preceq \inf_{\tau_h, v_h} \|\sigma - \tau_h\|_{L_2} + \|u - v_h\|_{H^1,h}$$

By the usual Bramble-Hilbert and scaling arguments we see that (using the element-wise L_2-projection P_h):

$$\|u - P_h u\|_{H^1,h} \preceq h^k \|u\|_{H^{1+k}} \qquad k \geq 0.$$

In the lowest order case ($k = 0$) we don't get any convergence !!

But, we can show error estimates for σ in terms of approximability for σ only. Furthermore, we can perform a local postprocessing to improve also the scalar variable.

Theorem 119. *There holds*

$$\|\sigma - \sigma_h\|_{L_2} + \|P_h u - u_h\| \preceq \|\sigma - I_h \sigma\|_{L_2},$$

where I_h is the canonical RT interpolation operator satisfying the commuting diagram property $\operatorname{div} I_h = P_h \operatorname{div}$.

Proof. As usual for a priori estimates, we apply stability of the discrete problem, and use the Galerkin orthogonality:

$$\|I_h\sigma - \sigma_h\|_{L_2} + \|P_h u - u_h\|_{H^1,h} \preceq \sup_{\tau_h, v_h} \frac{B((I_h\sigma - \sigma_h, P_h u - u_h), (\tau_h, v_h))}{\|\tau_h\|_{L_2} + \|v_h\|_{H^1,h}} \tag{6.18}$$

$$= \sup_{\tau_h, v_h} \frac{B((I_h\sigma - \sigma, P_h u - u), (\tau_h, v_h))}{\|\tau_h\|_{L_2} + \|v_h\|_{H^1,h}} \tag{6.19}$$

Now we elaborate on the terms of $B((I_h\sigma - \sigma, P_h u - u), (\tau_h, v_h)) = \int a^{-1}(I_h\sigma - \sigma) \cdot \tau_h + \int \operatorname{div}(I_h\sigma - \sigma) v_h + \int \operatorname{div} \tau_h (P_h u - u)$: For the first one we use Cauchy-Schwarz, and bounds for the coefficient a:

$$\int a^{-1}(I_h\sigma - \sigma) \cdot \tau_h \preceq \|\sigma - I_h\sigma\|_{L_2} \|\tau_h\|_{L_2}$$

For the second one we use the commuting diagram, and orthogonality:

$$\int \operatorname{div}(I_h\sigma - \sigma) v_h = \int \underbrace{(P_h - Id) \operatorname{div} \sigma}_{\in Q_h^\perp} \underbrace{v_h}_{\in Q_h} = 0$$

For the third one we use that $\operatorname{div} V_h \subset Q_h$:

$$\int \underbrace{\operatorname{div} \tau_h}_{\in Q_h} \underbrace{(P_h u - u)}_{Q_h^\perp}$$

Thus, the right hand side of equation (6.19) can be estimated by $\|\sigma - I_h\sigma\|_{L_2}$. Finally, an application of the triangle inequality proves the result. □

Remark 120. *This technique applies for BDM elements as well as for RT. Both satisfy* $\operatorname{div} V_h = Q_h$*, and the commuting diagram. For RT_k elements, i.e. $[P^k]^2 \subset RT_k \subset [P^{k+1}]^2$ as well as BDM_k elements, i.e. $BDM_k = [P^k]^d$ we get the error estimate*

$$\|\sigma - I_h\sigma\|_{L_2} \preceq h^{k+1} \|\sigma\|_{H^k}$$

with $k \geq 0$ for RT and $k \geq 1$ for BDM.

Remark 121. *The scalar variable shows super-convergence: A filtered error, i.e. $P_h u - u_h$ is of higer order than the error $u - u_h$ itself: One order for RT and two orders for BDM*

We can apply a local post-processing to compute the scalar part with higher accuracy: We use the equation $\sigma = a\nabla u$, and the good error estimates for σ and $P_h u$. We set $\widetilde{Q}_h := P^{k+1}$, and solve a local problem on every element:

$$\min_{\substack{\tilde{v}_h \in \widetilde{Q}_h \\ \int_T v_h = \int_T \tilde{v}_h}} \|a\nabla \tilde{v}_h - \sigma_h\|^2_{L_2(T), a^{-1}}$$

6.6.3 Solution methods for the linear system

The finite element discretization leads to the linear system for the coefficient vectors (called σ and u again):

$$\left(\begin{array}{cc} A & B^t \\ B & 0 \end{array}\right)\left(\begin{array}{c} \sigma \\ u \end{array}\right) = \left(\begin{array}{c} 0 \\ -f \end{array}\right)$$

This matrix is indefinite, it has $\dim V_h$ positive and $\dim Q_h$ negative eigenvalues. This causes difficulties for the linear equation solver.

The first possibility is a direct solver, which must (in contrast to positive definite systems) apply Pivot strategies.

A second possibility is block-elimination: eliminate σ from the first equation. The regularity of A follows from L_2-coercivity of $a(.,.)$:

$$\sigma = -A^{-1}B^t u$$

and insert it into the second equation:

$$-BA^{-1}B^t u = -f$$

Thanks to the LBB-condition, B has full rank, and thus the Schur complement matrix is regular. Since B is the discretization of the div-operator, B^t of the negative gradient, and A of $a(x)^{-1}I$, the equation can be interpreted as a discretization of

$$\operatorname{div} a \nabla u = -f$$

This approach is not feasible, since A^{-1} is not a sparse matrix anymore.

One can use extensions of the conjugate gradient (CG) method for symmetric but indefinit matrices (e.g. MINRES). Here, preconditioners are important. Typically, for block-systems one uses block-diagonal preconditioners to rewrite the system as

$$\left(\begin{array}{cc} \widetilde{G}_V^{-1} & 0 \\ 0 & \widetilde{G}_Q^{-1} \end{array}\right)\left(\begin{array}{cc} A & B^T \\ B & 0 \end{array}\right)\left(\begin{array}{c} \sigma \\ u \end{array}\right) = \left(\begin{array}{cc} \widetilde{G}_V^{-1} & 0 \\ 0 & \widetilde{G}_Q^{-1} \end{array}\right)\left(\begin{array}{c} 0 \\ -f \end{array}\right)$$

where $\widetilde{G}_V$ and $\widetilde{G}_Q$ are approximations to the Gramien matrices in V_h and Q_h:

$$G_{V,ij} = (\varphi_i^\sigma, \varphi_j^\sigma)_V \quad \text{and} \quad G_{Q,ij} = (\varphi_i^u, \varphi_j^u)_Q$$

One can either choose the $H(\text{div})$-L_2, or the $[L_2]^2$-H^1 setting, which leads to different kind of preconditioners. Here, the later one is much simpler. This is a good motivation for considering the alternative framework.

6.6.4 Hybridization

Hybridization is a technique to derive a new variational formulation which obtains the same solution, but its system matrix is positive definit. For this, we break the normal-continuity of the flux functions, and re-inforce it via extra equations. We obtain new variables living on the element-edges (or faces in 3D).

We start from the first equation $a^{-1}\sigma - \nabla u = 0$, multiply with element-wise discontinuous test-functions τ, and integrate by parts on the individual elements:

$$\int_\Omega a^{-1}\sigma\tau + \sum_T \{ \int_T u \operatorname{div}\tau - \int_{\partial T} u\, \tau_n \} = 0$$

We now introduce the new unknown variable $\hat{u}$ which is indeed the restriction of u onto the mesh skeleton $u_{|\cup E}$.

We set $V := \prod_T H(\operatorname{div}, T)$ and $Q := L_2(\Omega) \times \prod_E H^{1/2}(E)$, and pose the so called hybrid problem: find $\sigma \in V$ and $(u, \hat{u}) \in Q$ such that

$$\begin{array}{lllllll}
\int_\Omega (a^{-1}\sigma)\cdot\tau\, dx & + & \sum_T \int_T \operatorname{div}\, \tau\, u\, dx & + & \sum_T \int_{\partial T} \hat{u}\, \tau_n & = & 0 & \forall\, \tau \in V \\
\sum_T \int_T \operatorname{div}\, \sigma\, v\, dx & & & & & = & -\int f v\, dx & \forall\, v \in L_2(\Omega) \\
\sum_T \int_{\partial T} \hat{v}\, \sigma_n & & & & & = & 0 & \forall\, \hat{v} \in \prod_E H^{1/2}(E).
\end{array}$$

The last equation can be rearranged edge by edge:

$$\sum_{E\subset\Omega} \int_E [\sigma_n]\, \hat{v} + \sum_{E\subset\partial\Omega} \int_E \sigma_n\, \hat{v} = 0 \qquad \forall\, \hat{v} \in \prod_E H^{1/2}(E),$$

which implies normal-continuity of σ. Dirichlet/Neumann boundary conditions are posed now for the skeleton variable $\hat{u}$.

This system is discretized by discontinuous RT/BDM elements for σ, piecewise polynomials on elements T for u, and piecewise polynomials on edges for $\hat{u}$ such that the order matches with $\sigma \cdot n$.

1. This discrete system is well-posed with respect to the norms $\|\sigma\|_V := \|\sigma\|_{L_2}$ and $\|u, \hat{u}\|^2 = \sum_T \|\nabla u\|^2_{L_2(T)} + \frac{1}{h}\|u - \hat{u}\|^2_{\partial T}$. Similar proof as for Theorem 118.

2. The components σ_h and u_h of the solution of the hybrid problem correspond to the solution of the mixed method.

3. Since the σ_h is discontinuous across elements, the arising matrix A is block-diagonal. Now it is cheap to form the Schur complement

$$-BA^{-1}B^T \binom{u}{\hat{u}} = \binom{f}{0}$$

This is now a system with a positive definite matrix for unknowns in the elements and on the edges. Since the matrix block for the element-variables is still block-diagonal, they can be locally eliminated, and only the skeleton variables are remaining. In the lowest order case, the matrix is the same as for the non-conforming P^1 element.

Chapter 7

Discontinuous Galerkin Methods

Discontinuous Galerkin (DG) methods approximate the solution with piecewise functions (polynomials), which are discontinuous across element interfaces. Advantages are

- block-diagonal mass matrices which allow cheap explicit time-stepping
- upwind techniques for dominant convection
- coupling of non-matching meshes
- more flexibility for stable mixed methods

DG methods require more unknowns, and also have a denser stiffness matrix. The last disadvantage can be overcome by hybrid DG methods (HDG).

7.1 Transport equation

We consider the first order equation

$$\operatorname{div}(bu) = f \qquad \text{on } \Omega,$$

where b is the given wind, and f is the given source. Boundary conditions are specified

$$u = u_D \qquad \text{on } \Gamma_{in},$$

where the inflow boundary is

$$\Gamma_{in} = \{x \in \partial\Omega : b \cdot n < 0\},$$

and the outflow boundary $\Gamma_{out} = \partial\Omega \setminus \Gamma_{in}$.

The instationary transport equation

$$\frac{\partial u}{\partial t} + \operatorname{div}(bu) = f \qquad \text{on } \Omega \times (0, T)$$

with initial conditions $u = u_0$ for $t = 0$ can be considered as stationary transport equation in space-time:

$$\operatorname{div}_{x,t}(\tilde{b}u) = f,$$

where $\tilde{b} = (b, 1)$. The inflow boundary consists now of the lateral boundary $\Gamma_{in} \times (0, T)$ and the bottom boundary $\Omega \times \{0\}$, which is an inflow boundary according to $(b, 1) \cdot (0, -1) < 0$.

The equation in conservative form leads to a conservation principle. Let V be an arbitrary control volume. From the Gauß theorem we get

$$\int_{\partial V} b \cdot nu = \int_V f$$

The total outflow is in balance with the production inside V.

For stability, we assume $\operatorname{div} b = 0$. This is a realistic assumption, since the wind is often the solution of the incompressible Navier Stokes equation.

A variational formulation is

$$\int \operatorname{div}(bu)v = \int fv \qquad \forall v$$

If we set $v = u$, and use

$$\operatorname{div}(bu)u = \tfrac{1}{2}\operatorname{div}(bu^2)$$

We obtain

$$\int \operatorname{div}(bu)u = \tfrac{1}{2}\int_{\partial\Omega} b_n u^2 = \int fu$$

For $f = 0$ we obtain

$$\int_{\Gamma_{out}} |b_n|u^2 = \int_{\Gamma_{in}} |b_n|u^2.$$

This inflow-outflow isometry is a stability argument. For time dependent problems (with $b_n = 0$ on $\partial\Omega$), it ensures the conservation of L_2-norm in time.

7.1.1 Solvability

We assume $b \in L_\infty$ with $\operatorname{div} b = 0$. We consider the problem: find $u \in V$, $u = u_D$ on Γ_{in} and

$$B(u, v) = f(v) \qquad \forall\, v \in W$$

with

$$B(u, v) = \int \operatorname{div}(bu)v \qquad \text{and} \qquad f(v) = \int fv$$

The space V shall be defined by the (semi)-norm

$$\|u\|_V = \|b\nabla u\|_{L_2}$$

Depending on b, it is a norm for $\{u = 0 \text{ on } \Gamma_{in}\}$. Roughly speaking, if every point in Ω can be reached by a finite trajectory along b, then $\|u\|_{L_2} \preceq \|u\|_V$, by Friedrichs' inequality.

It does not hold if b has vortices. Then $\|\cdot\|_V$ is only a semi-norm. Note, for space-time problems $\tilde{b}$ cannot have vortices. For theory, we will assume that

$$\|u\|_{L_2} \preceq \|u\|_V \qquad \forall\, u = 0 \text{ on } \Gamma_{in}$$

The test space is

$$W = L_2$$

The forms $B(.,.)$ and $f(.)$ are continuous. The $\inf-\sup$ conditions is trivial:

$$\sup_v \frac{B(u,v)}{\|v\|_{L_2}} \underbrace{\geq}_{v:=b\nabla u} \frac{\int (b\nabla u)^2}{\|b\nabla u\|} = \|b\nabla u\|_{L_2}$$

7.2 Discontinuous Galerkin Discretization

A DG method is a combination of finite volume methods and finite element methods. We start with a triangulation $\{T\}$. On every element we multiply the equation by a test-function:

$$\int_T b\nabla uv = \int_T fv \qquad \forall\, v$$

We integrate by parts:

$$-\int_T bu\nabla v + \int_{\partial T} b_n uv = \int_T fv$$

On the element-boundary we replace $b_n u$ by its up-wind limit $b_n u^{up}$. On the element inflow boundary $\partial T_{in} = \{x \in \partial T : bn_T < 0\}$, the upwind value is the value from the neighbour element, while on the element outflow boundary it is the value from the current element T. For elements on the domain inflow boundary, the upwind value is taken as the boundary value u_D. For the continuous solution there holds

$$-\int_T bu\nabla v + \int_{\partial T} b_n u^{up} v = \int_T fv$$

Now we integrate back:

$$\int_T b\nabla uv + \int_{\partial T} b_n(u^{up} - u)v = \int_T fv$$

On the outflow boundary, the boundary integral cancels out, on the inflow boundary we can write it as a jump term $[u] = u^{up} - u$:

$$\int_T b\nabla uv + \int_{\partial T_{in}} b_n[u]v = \int_T fv$$

We define the DG bilinear-form

$$B^{DG}(u,v) = \sum_T \left\{ \int_T b\nabla uv + \int_{\partial T_{in}} b_n[u]v \right\}.$$

The true solution is consistent with

$$B^{DG}(u,v) = f(v) \qquad \forall\, v \text{ piece-wise continuous.}$$

We define DG finite element spaces:

$$V_h := W_h := \{v \in L_2 : v|_T \in P^k\}$$

The DG formulation is: find $u_h \in V_h$ such that

$$B^{DG}(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in W_h$$

For the discontinuous space, the jump-term is important. If we use continous spaces, the jump-term disappears. The discrete norms are defined as

$$\begin{aligned} \|u_h\|_{V_h}^2 &:= \sum_T \|b \nabla u\|_{L_2(T)}^2 + \sum_E \tfrac{1}{h} \|b_n [u]\|_{L_2(E)}^2 \\ \|v_h\|_{W_h} &= \|v_h\|_{L_2} \end{aligned}$$

The part with the jump-term mimics the derivative as kind of finite difference term across edges.

We prove solvability of the discrete problem by showing a discrete $\inf-\sup$ condition. But, in general, one order in h is lost due to a mesh-dependent $\inf-\sup$ constant. This factor shows up in the general error estimate by consistency and stability. It can be avoided in 1D, and on special meshes.

Theorem 122. *There holds the discrete* $\inf-\sup$ *condition*

$$\sup_{v_h} \frac{B(u_h, v_h)}{\|v_h\|_{W_h}} \succeq h\, \|u_h\|_{V_h}$$

Proof. We take two different test-functions: $v_1 = u_h$ and $v_2 := b \cdot \nabla_T u$, and combine them properly. The second test-function would not be possible in the standard C^0 finite element space.

There holds (dropping sub-scripts h):

$$\begin{aligned} B(u_h, v_1) &= B(u_h, u_h) = \sum_T \int_T b \nabla u\, u + \int_{\partial T_{in}} b_n [u] u \\ &= \sum_T \frac{1}{2} \int_{\partial T} b_n u^2 + \int_{\partial T_{in}} b_n [u] u \end{aligned}$$

We reorder the terms edge-by-edge. On the edge E we get contributions from two elements: On the inflow-boundary of the down-wind element we get

$$\tfrac{1}{2} \int_E b_n (u^d)^2 + \int_E b_n (u^u - u^d) u^d = \int_E |b_n| \left(\tfrac{1}{2} (u^d)^2 - u^u u^d \right)$$

We used that $b \cdot n$ is negative on the inflow boundary. From the up-wind element we get on its outflow-boundary:

$$\int_E \tfrac{1}{2}|b_n|\,(u^u)^2$$

Summing up, we have the square

$$\int_E \tfrac{1}{2}|b_n|(u^u - u^d)^2,$$

and summing over elements we get the non-negative term

$$B(u_h, u_h) = \tfrac{1}{2}\sum_E \int_E |b_n|[u]^2.$$

An extra treatment of edges on the whole domain boundary gives that the jump must be replaced by the function values on $\partial\Omega$.

We plug in the second test function v_2:

$$B(u_h, v_2) \quad = \quad \sum_T \int_T (b\nabla u)^2 + \int_{\partial T_{in}} b_n[u]\, b\nabla u$$

We use Young's inequality to bound the second term from below:

$$B(u_h, v_2) \geq \sum_T \int_T (b\nabla u)^2 - \frac{1}{2\gamma}\|b_n[u]\|^2_{\partial L_2(T_{in})} - \frac{\gamma}{2}\|b\nabla u\|^2_{L_2(\partial T_{in})}$$

By the choice $\gamma \simeq h$ and a inverse trace inequality (which needs smoothness assumptions onto b) we can bound the last term by the first one on the right hand. Thus

$$B(u_h, v_2) \succeq \sum_T \int_T (b\nabla u)^2 - \sum_E \frac{1}{h}\|[u]\|^2_{L_2(E),b_n}$$

Finally, we set

$$v_h = \frac{1}{h}v_1 + v_2$$

to obtain

$$B(u_h, v_h) \succeq \sum_T \int_T (b\nabla u)^2 + \sum_E \frac{1}{h}\|[u]\|^2_{L_2(E),b_n} \simeq \|u_h\|^2_{V_h}.$$

But, for this choice we get

$$\|v_h\|_{W_h} \preceq h^{-1}\|u_h\|_{V_h},$$

and thus the h-dependent $\inf - \sup$-*constant.* □

7.3 Nitsche's method for Dirichlet boundary conditions

We build in Dirichlet b.c. in a weak sense. In constrast to a mixed method, we obtain a positive definit matrix.

We consider the equation

$$-\Delta u = f \qquad \text{and} \qquad u = u_D \text{ on } \partial\Omega.$$

A diffusion coefficient, or mixed boundary conditions are possible as well. We multiply with testfunctions, integrate and integrate by parts:

$$\int_\Omega \nabla u \nabla v - \int \partial_n u v = \int f v \qquad \forall\, v$$

We do not restrict test functions to $v = 0$. To obtain a symmetric bilinear-form, we add a consistent term

$$\int_\Omega \nabla u \nabla v - \int_{\partial\Omega} \partial_n u v - \int_{\partial\Omega} \partial_n v u = \int f v - \int_{\partial\Omega} \partial_n v u_D \qquad \forall\, v$$

Finally, to obtain stability (as proven below), we add the so called stabilization term

$$\int_\Omega \nabla u \nabla v - \int_{\partial\Omega} \partial_n u v - \int_{\partial\Omega} \partial_n v u + \frac{\alpha}{h} \int u v = \int f v - \int_{\partial\Omega} \partial_n v u_D + \frac{\alpha}{h} \int u_D v \qquad \forall\, v$$

These are the forms of Nitsche's method:

$$\begin{aligned} A(u,v) &= \int_\Omega \nabla u \nabla v - \int_{\partial\Omega} \partial_n u v - \int_{\partial\Omega} \partial_n v u + \frac{\alpha}{h} \int_{\partial\Omega} u v \\ f(v) &= \int_\Omega f v - \int_{\partial\Omega} \partial_n v u_D + \frac{\alpha}{h} \int_{\partial\Omega} u_D v \end{aligned}$$

$A(.,.)$ is not defined for $u, v \in H^1$, but it requires also well defined normal derivatives. This is satisfied for the flux $\nabla u \in H(\text{div})$ of the solution, and finite element test functions v.

We define the Nitsche norm:

$$\|u\|_{1,h}^2 := \|\nabla u\|_{L_2}^2 + \tfrac{1}{h} \|u\|_{L_2(\partial\Omega)}^2$$

Lemma 123. *If $\alpha = O(p^2)$ is chosen sufficiently large, then $A(.,.)$ is elliptic on the finite element space:*

$$A(u_h, u_h) \succeq \|u_h\|_{1,h}^2 \quad \forall\, u_h \in V_h$$

Proof. On one element there holds the inverse trace inequality

$$\|u_h\|_{L_2(\partial T)}^2 \leq c\,\frac{p^2}{h} \|u_h\|_{L_2}^2 \qquad \forall\, u_h \in P^p(T)$$

The h-factor is shown by transformation to the reference element, the p-factor (polynomial order) is proven by expansion in terms of orthogonal polynomials. Using the element-wise estimate for all edges on the domain boundary, we obtain

$$\|u_h\|^2_{L_2(\partial\Omega)} \leq c\,\frac{p^2}{h}\|u\|^2_{L_2(\Omega)} \tag{7.1}$$

Evaluating the bilinear-form, and applying Young's inequality for the mixed term we get

$$\begin{aligned} A(u_h, u_h) &= \|\nabla u_h\|^2_{L_2} - 2\int_{\partial\Omega} \partial_n u u + \frac{\alpha}{h}\|u\|^2_{L_2(\partial\Omega)} \\ &\geq \|\nabla u_h\|^2_{L_2} - \tfrac{1}{\gamma}\|n\cdot\nabla u_h\|^2_{L_2(\partial\Omega)} - \gamma\|u\|^2_{L_2(\partial\Omega)} + \frac{\alpha}{h}\|u\|^2_{L_2(\partial\Omega)} \end{aligned}$$

The inverse trace inequality applied to ∇u_h gives

$$\|n\cdot\nabla u_h\|_{\partial\Omega} \leq \|\nabla u_h\|_{\partial\Omega} \leq c\frac{p^2}{h}\|\nabla u_h\|_\Omega$$

By choosing

$$\gamma > c\frac{p^2}{h} \qquad \text{and} \qquad \gamma \leq \frac{\alpha}{h}$$

we can absorb the negative terms into the positive ones. Therefore it is necessary to choose

$$\alpha > cp^2$$

$\square$

For the error analysis we apply the discrete stability and consistency:

$$\begin{aligned} \|u_h - I_h u\|_{1,h} &\preceq \sup_{v_h} \frac{A(u_h - I_h u, v_h)}{\|v_h\|_{1,h}} \\ &= \sup_{v_h} \frac{A(u - I_h u, v_h)}{\|v_h\|_{1,h}} \end{aligned}$$

We cannot argue with continuity of $A(.,.)$ on H^1 (which is not true), but we can estimate the interpolation error $u - I_h u$ for all four terms of $A(u_h - I_h u, v_h)$.

7.3.1 Nitsche's method for interface conditions

We give now a variational formulation for interface conditions $u_1 = u_2$, $\partial_{n_1} u_1 + \partial_{n_2} u_2 = 0$ on the interface γ separating Ω_1 and Ω_2. Boundary conditions on the outer boundary are treated as usual. Integration by parts on the sub-domains leads to

$$\int_{\Omega_1} \nabla u \nabla v - \int_\gamma \partial_{n_1} u_1 v_1 + \int_{\Omega_2} \nabla u \nabla v - \int_\gamma \partial_{n_2} u_2 v_2 = \int f v$$

We define the mean value

$$\{\partial_{n_1} u\} = \tfrac{1}{2}(\partial_{n_1} u_1 + \partial_{n_2} u_2)$$

and jump

$$[v] = v_1 - v_2.$$

Using continuity of the normal flux (taking the orientation into account) we get

$$\sum_i \int_{\Omega_i} \nabla u \nabla v - \int_\gamma \{\partial_{n_1} u\}[v] = \int fv.$$

Note that both terms, mean of normal derivative and jump, change sign if we exchange the enumeration of sub-domains.

We procede as before and add consistent symmetry and stabilization terms:

$$\sum_i \int_{\Omega_i} \nabla u \nabla v - \int_\gamma \{\partial_{n_1} u\}[v] - \int_\gamma \{\partial_{n_1} v\}[u] + \frac{\alpha}{h}\int_\gamma [u][v] = \int fv.$$

The variational formulation is consistent on the solution, and elliptic on V_h, which is proven as before. This approach is an alternative to the mixed method (mortar method), since it leads to positive definite matrices (called also gluing method).

7.4 DG for second order equations

Nitsche's method for interface conditions can be applied element-by-element. This is the (independently developed) Discontinuous Galerkin (DG) method. Precisely, it's called SIP-DG (symmetric interior penalty) DG:

$$A(u,v) = \sum_T \left\{ \int_T \nabla u \nabla v - \tfrac{1}{2}\int_{\partial T} \partial_n u [v] - \tfrac{1}{2}\int_{\partial T} \partial_n v [u] + \tfrac{\alpha}{h}\int_{\partial T} [u][v] \right\}$$

(and proper treatment of integrals on the domain boundary). The factor $\frac{1}{2}$ is coming from splitting the consistent terms to the two elements on the edge.

Convergence analysis similar to Nitsche's method.

Beside the SIP-DG, also different version are in use: The NIP-DG (non-symmetric interior penalty) DG:

$$A(u,v) = \sum_T \left\{ \int_T \nabla u \nabla v - \tfrac{1}{2}\int_{\partial T} \partial_n u [v] + \tfrac{1}{2}\int_{\partial T} \partial_n v [u] + \tfrac{\alpha}{h}\int_{\partial T} [u][v] \right\}$$

The term formerly responsible for symmetry is added with a different sign. The variational problem is still consistent on the true solution. The advantage of the NIP-DG is that

$$A(u,u) = \sum_T \|\nabla u\|^2_{L_2(T)} + \tfrac{\alpha}{h}\|[u]\|^2_{L_2(\partial T)},$$

i.e. $A(.,.)$ is elliptic in any case $\alpha > 0$. The disadvantage is that $A(.,.)$ is not consistent for the dual problem, i.e. the Aubin-Nitsche trick cannot be applied. It is popular for convection-diffusion problems, where the bi-form is non-symmetric anyway. The IIP-DG (incomplete) skips the third term completely. Advantages are not known to the author.

7.4.1 Hybrid DG

One disadvantage of DG - methods is that the number of degrees of freedom is much higher than a continuous Galerkin method on the same mesh. Even worse, the number of non-zero entries per row in the system matrix is higher. The second disadvantage can be overcome by hybrid DG methods: One adds additional variables $\hat{u}$, $\hat{v}$ on the inter-element facets (edges in 2D, faces in 3D). The derivation is very similar:

$$\sum_T \int_T \nabla u \nabla v - \int_{\partial T} \partial_n u v = \sum_T f v \qquad \forall\, v \in P^k(T), \forall\, T$$

Using continuity of the normal flux, we may add $\sum_T \int_{\partial T} \partial_n u \hat{v}$ with a single-valued test-function on the facets:

$$\sum_T \int_T \nabla u \nabla v - \int_{\partial T} \partial_n u (v - \hat{v}) = \sum_T f v \qquad \forall\, v \in P^k(T), \forall\, T$$

Again, we smuggle in consistent terms for symmetry and coercivity:

$$\sum_T \int_T \nabla u \nabla v - \int_{\partial T} \partial_n u (v - \hat{v}) - \int_{\partial T} \partial_n v (u - \hat{u}) + \frac{\alpha}{h} \int_{\partial T} (u - \hat{u})(v - \hat{v}) = \sum_T f v \qquad \forall\, v \in P^k(T), \forall\, T$$

The jump between neighbouring elements is now replaced by the difference of element-values and facet values. The natural norm is

$$\|u, \hat{u}\|^2 = \sum_T \|\nabla u\|^2 + \tfrac{1}{h} \|u - \hat{u}\|_{\partial T}$$

The HDG methods allows for static condensation of internal variables which results in a global system for the edge-unknowns, only.

The lowest order method uses $P^1(T)$ and $P^1(E)$, and we get $O(h)$ convergence. When comparing with the non-conforming P^1-method, HDG has more unknowns on the edges, but the same order of convergence. The Lehrenfeld-trick is to smuggle in a projector:

$$\sum_T \int_T \nabla u \nabla v - \int_{\partial T} \partial_n u (v - \hat{v}) - \int_{\partial T} \partial_n v (u - \hat{u}) + \frac{\alpha}{h} \int_{\partial T} P^{k-1} (u - \hat{u})(v - \hat{v}) = \sum_T f v \qquad \forall\, v \in P^k(T), \forall\, T$$

This allows to reduce the order on edges by one, while maintaining the order of convergence.

7.4.2 Bassi-Rebay DG

One disadvantage of IP-DG is the necessary penalty term with α sufficiently large. For well-shaped meshes $\alpha = 5p^2$ is usually enough. But, for real problems the element deformation may become large, and then a fixed α is not feasible. Setting α too large has a negative effect for iterative solvers.

An alternative is to replace the penalty term by

$$\|[u]\|_{BR}^2 := \sup_{\sigma_h \in [P^{k-1}]^d} \frac{([u], n \cdot \sigma_h)_{L_2(\partial T)}^2}{\|\sigma_h\|_{L_2(T)}^2}$$

It can be implemented by solving a local problem with L_2-bilinear-form.

In the coercivity proof, the *bad* term is now estimated as

$$\int_{\partial T} n \cdot \nabla u_h \, [u_h] \leq \sup_{\sigma_h} \frac{\int_{\partial T} n \cdot \sigma_h \, [u_h]}{\|\sigma_h\|} \|\nabla u_h\| \leq \|\nabla u_h\| \, \|[u]\|_{BR}$$

The BR-norm scales like the IP - norm (in h and p), but the (typically unknown) constant in the inverse trace inequality can be avoided.

7.4.3 Matching integration rules

Another method to avoid guessing the sufficiently large α is to use integration rules, such that the integration points for the boundary integral are a sub-set of the integration points of the volume term. Now, Young's inequality can be applied for the numerical integrals. The $\frac{\alpha}{h}$ factor is now replaced by the largest relative scaling of weights for the boundary integrals and volume integrals. The pro is the simplicity, the con is the need of numerical integration rules which need more points.

7.4.4 (Hybrid) DG for Stokes and Navier-Stokes

DG or HDG methods allow the construction of numercal methods for incompressible flows, which obtain exactly divergence free discrete velocities. We discretize Stokes's equation as follows:

$$V_h = BDM^k \qquad Q_h = P^{k-1,dc},$$

and the bilinear-forms

$$a(u_h, v_h) = a^{DG}(u_h, v_h)$$

and

$$b(u_h, q_h) = \int \operatorname{div} u_h \, q_h.$$

Since the space V_h is not conforming for H^1, the DG - technique is applied. The $b(.,.)$ bilinear-form is well defined for the $H(\operatorname{div})$-conforming finite element space V_h. There holds

$$\operatorname{div} V_h = Q_h,$$

and thus the discrete divergence free condition

$$\int \operatorname{div} u_h q_h = 0 \qquad \forall q_h$$

implies

$$\| \operatorname{div} u_h \|_{L_2}^2 = \int \operatorname{div} u_h \underbrace{\operatorname{div} u_h}_{\in Q_h} = 0.$$

Hybridizing the method leads to facet variables for the tangential components, only. This method can be applied to the Navier Stokes equations. Here, the exact divergence-free discrete solution leads to a stable method for the nonlinear transport term (References: Master thesis Christoph Lehrenfeld: HDG for Navier Stokes, h-version LBB, Master thesis Philip Lederer: p-robust LBB for triangular elements).

Chapter 8

Applications

We investigate numerical methods for equations describing real life problems.

8.1 The Navier Stokes equation

The Navier Stokes equation describe the flow of a fluid (such as water or air). The incompressible Navier Stokes equation models incompressible fluids (such as water). The stationary N.-St. equation models a solution in steady state (no change in time).

The field variables are the fluid velocity $u = (u_x, u_y, u_z)$, and the pressure p. Conservation of momentum is

$$-\nu \Delta u + \rho (u \cdot \nabla) u - \nabla p = f$$

The first term describes friction of the fluid (ν is called viscosity). The second one arises from conservation of momentum of moving particles. It is called the convective term (ρ is the density). The source term f models forces, mainly gravity. The incompressibility of the fluid is described by

$$\text{div } u = 0.$$

Different types of boundary conditions onto u and p are possible.

The Navier Stokes equation is nonlinear. In general, no unique solution is guaranteed. The common approach to find a solution is the so called Oseen iteration: Given u^k, find the next iterate (u^{k+1}, p^{k+1}) by solving

$$\begin{aligned} -\nu \Delta u^{k+1} + \rho (u^k \cdot \nabla) u^{k+1} - \nabla p^{k+1} &= f \\ \text{div } u^{k+1} &= 0. \end{aligned}$$

Under reasonable conditions, this Oseen equation is uniquely solvable. Since u^k is the solution of the old step, it satisfies div $u^k = 0$. Furthermore, we assume that the velocity u^k is bounded in L_∞-norm.

From now on, we continue to investigate the Oseen equation. Given a vector-field $w = (w_x, w_y, w_z) \in [L_\infty]^3$ such that div $w = 0$. Find u and p such that

$$\begin{aligned} -\Delta u + (w \cdot \nabla) u - \nabla p &= f \\ \text{div } u &= 0. \end{aligned}$$

We have removed the viscosity by rescaling the equation. The factor ρ/ν is incorporated into the vector-field w.

As usual, we go over to the weak formulation: Find $u \in V = [H^1]^3$ and $p \in Q = L_2$ such that

$$\begin{array}{rcll} \int \{\nabla u \, \nabla v + (w \cdot \nabla) u \, v\} \, dx \; + \; \int \text{div } v \, p \, dx & = & \int f v \, dx & \forall \, v \in V \\ \int \text{div} \, u \, q \, dx & = & 0 & \forall \, q \in Q. \end{array} \tag{8.1}$$

This variational problem is a mixed formulation. It satisfies the conditions of Brezzi: The bilinear forms are

$$\begin{aligned} a(u, v) &= \int \{\nabla u \nabla v + (w \cdot \nabla) u v\} \, dx, \\ b(u, q) &= \int \text{div } u \, q \, dx. \end{aligned}$$

Both forms are continuous. The form $a(.,.)$ is non-symmetric. In $a(.,.)$, the x, y, and z components of u and v are independent. To investigate $a(.,.)$, it is enough to consider scalar bilinear-forms. We define the inflow and outflow boundaries

$$\begin{aligned} \Gamma_i &= \{x \in \partial\Omega : w \cdot n < 0\}, \\ \Gamma_o &= \{x \in \partial\Omega : w \cdot n \geq 0\}. \end{aligned}$$

If we pose Dirichlet boundary conditions on Γ_i, then $a(.,.)$ is coercive (see example 27, and exercises). The ratio of the continuity bound and the coercivity bound depends on the norm of the convection w. With increasing w, the problem is getting worse.

The form $b(.,.)$ satisfies the LBB condition:

$$\sup_{u \in [H^1_{0,D}]^3} \frac{\int \text{div} \, u \, q \, dx}{\|u\|_{H^1}} \succeq \|q\|_{L_2} \qquad \forall \, q \in L_2.$$

In the case of (partial) Dirichlet boundary conditions ($H^1_{0,D} = \{u : u = 0 \text{ on } \Gamma_D\}$), this condition is very nontrivial to prove. If there are only Dirichlet b.c., one has to use $Q = L_2^0 = \{q : \int_\Omega q \, dx = 0\}$.

Under these conditions, Brezzi's theorem proves a unique solution of the Oseen equation.

Finite elements for Navier-Stokes equation

We want to approximate the Oseen equation by a Galerkin method: Find $u_h \in V_h$ and $p_h \in Q_h$ such that

$$\begin{array}{lclcll} \int \{\nabla u_h \nabla v_h + (w\cdot\nabla) u_h\, v_h\}\, dx & + & \int \mathrm{div}\; v_h\, p_h\, dx & = & \int f v_h\, dx & \forall\, v_h \in V_h \\ \int \mathrm{div}\, u_h\, q_h\, dx & & & = & 0 & \forall\, q_h \in Q. \end{array} \tag{8.2}$$

To obtain convergence $u_h \to u$ and $p_h \to p$, it is important to choose proper approximation spaces V_h and Q_h. Using the simplest elements, namely continuous and piece-wise linear elements for $V_h \subset [H^1]^3$, and piece-wise constants for $Q_h \subset L_2$ does not work. The discrete LBB condition is not fulfilled: In 2D, there are asymptotically twice as many triangles than vertices, i.e., dim $V_h \approx$ dim Q_h, and $\int \mathrm{div}\, u_h\, q_h\, dx = 0\ \forall\, q_h \in Q_h$ implies $u_h \approx 0$.

The simplest spaces which lead to convergence are the non-conforming P_1 element for the velocities, and piece-wise constant elements for the pressure. The arguments are

- There are unknowns on the edges to construct a Fortin operator satisfying
$$\int_e u \cdot n\, ds = \int_e (I_h u)\cdot n\, ds,$$
and thus proving the discrete LBB condition.
- The error due to the non-conforming space $V_h \not\subset V$ is of the same order as the approximation error (see Section 4.5).

8.1.1 Proving LBB for the Stokes Equation

Stability of the continuous equation

We consider Stokes equation: find $u \in [H^1_0]^d$ and $p \in L^0_2$ such that

$$\begin{array}{lclcll} \int \nabla u \cdot \nabla v & + & \int \mathrm{div}\, v\, p & = & \int f v & \forall\, v \in [H^1_0]^d \\ \int \mathrm{div}\, u\, q & & & = & 0 & \forall\, q \in L^0_2. \end{array} \tag{8.3}$$

Solvability follows from Brezzi's theorem. The only non-trivial part is the LBB condition:

$$\sup_{v \in [H^1_0]^d} \frac{\int \mathrm{div}\, v\, p}{\|v\|_{H^1}} \geq \beta \|p\|_{L_2} \qquad \forall\, p \in L^0_2$$

We sketch two different proofs:

Proof 1: The LBB condition becomes simple if we skip the Dirichlet conditions:

$$\sup_{v \in [H^1]^d} \frac{\int \mathrm{div}\, v\, p}{\|v\|_{H^1}} \geq \beta \|p\|_{L_2} \forall\, p \in L_2$$

Take $p \in L_2(\Omega)$, extend it by 0 to $L_2(\mathbb{R}^d)$. Now compute a right-inverse of div via Fourier transform:

$$\begin{aligned} \hat{p}(\xi) &= \mathcal{F}(p) \\ \hat{u}(\xi) &= \frac{-i\xi}{|\xi|^2}\hat{p}(\xi) \\ u(x) &= \mathcal{F}^{-1}(\hat{u}) \end{aligned}$$

Since $\operatorname{div} u = p$ translates to $i\xi \cdot \hat{u} = \hat{p}$, we found a right-inverse to the divergence. Furthermore, $|u|_{H^1(\Omega)} = \|i\xi\hat{u}\|_{L_2} \preceq \|\hat{p}\|_{L_2} = \|p\|_{L_2}$. We restrict this u to Ω. The L_2-part of $\|u\|_{H^1}$ follows from the Poincare inequality after subtracting the mean value.

The technical part is to ensure Dirichlet - boundary conditions. One can build an extension operator $\mathcal{E}$ from $L_2(\mathbb{R}^d \setminus \Omega)$ onto $\mathbb{R}^d$, which commutes with the div-operator: $\operatorname{div} \mathcal{E}u = \mathcal{E}^p \operatorname{div} u$, and sets

$$u_{final} := u - \mathcal{E}u$$

This u satisfies $u = 0$ on $\partial\Omega$. Since $\operatorname{div} u = p = 0$ outside of Ω, the correction did not change the divergence inside Ω. A self-contained proof is given in J. Bramble: *A proof of the inf-sup condition for the Stokes equation on Lipschitz domains*, Mathematical Models and Methods in Applied Sciences, Vol 13, pp 361-371 (2003).

Proof 2: Directly construct a right-inverse for the div-operator via integration. We assume that Ω is star-shaped w.r.t. ω, and $a \in \omega$. Extend p by 0 to $L_2(\mathbb{R}^d)$:

$$u_a(x) := -(x-a)\int_1^\infty t^{d-1} p(a + t(x-a))\,dt \qquad x \neq a$$

and $u_a(a) = 0$. If $\int_\Omega p = 0$, then $\operatorname{div} u_a = p$. Furthermore, $u = 0$ outside Ω. Next, we average over star-points in ω:

$$u := \frac{1}{|\omega|}\int_\omega u_a\,da$$

There is still $\operatorname{div} u = p$. Now, one can show that $\|u\|_{H^1} \preceq \|p\|_{L_2}$. See M. Costabel and A.McIntosh: *On Bogovskiĭ and regularized Poincaré integral operators for de Rham complexes on Lipschitz domains*, Mathematische Zeitschrift 265, 297-320 (2010).

8.1.2 Discrete LBB

Now, we turn to the discrete system posed on $V_h \subset V$ and $Q_h \subset Q$. The discrete LBB condition follows from the continuous one by construction of a Fortin operator (Lemma 99).

Elements with discontinuous pressure

The simplest pair is the non-conforming P^1 element, and constant pressure:

$$V_h = P^{1,nc}, \qquad Q_h = P^{0,dc}$$

We have to extend the V-norm and forms by the sum over element-wise norms and forms. The Fortin-operator $I_F : V \to V_h$ is defined via

$$\int_E I_F u = \int_E u \quad \forall \,\text{edges } E$$

It is continuous from H^1 to broken H^1 (via mapping), and satisfies

$$\int_T \operatorname{div}(I_F u) = \int_{\partial T} (I_F u) \cdot n = \int_{\partial T} u \cdot n = \int_T \operatorname{div} u,$$

and thus

$$b_h(I_F u, q_h) = b(u, q_h) \qquad \forall \, q_h \in Q_h$$

The error estimate follows similar as in the second Lemma by Strang:

$$\begin{aligned}
&\|u - u_h\|_{H^1,nc} + \|p - p_h\|_{L_2} \\
&\quad \preceq \inf_{v_h, q_h} \|u - v_h\|_{H^1,nc} + \|p - q_h\|_{L_2} + \sup_{w_h} \frac{\sum_T \int \nabla u \nabla w_h + p \operatorname{div} w_h - f w_h}{\|w_h\|_{H^1,nc}} \\
&\quad \preceq ch \left(\|u\|_{H^2} + \|p\|_{H^1} \right)
\end{aligned}$$

This convergence rate $O(h)$ is considered to be optimal for these elements.

Next we consider

$$V_h = P^2 \qquad Q_h = P^{0,dc}$$

We would like to define the Fortin operator similar as before:

$$\begin{aligned}
I_F u(V) &= u(V) && \forall \,\text{vertices } V \\
\textstyle\int_E I_F u &= \textstyle\int_E u && \forall \,\text{edges } E
\end{aligned}$$

But, the vertex evaluation is not allowed in H^1. We proceed now in two steps: First approximate u in the finite element space via a Clément operator Π_h:

$$u_h^1 := \Pi_h u,$$

and modify this u_h^1 via a correction term:

$$u_h := I_F u := u_h^1 + I_F^2 (u - u_h^1)$$

The correction operator I_F^2 is defined as

$$\begin{aligned}
I_F^2 u(V) &= 0 && \forall \,\text{vertices } V, \\
\textstyle\int_E I_F^2 u &= \textstyle\int_E u && \forall \,\text{edges } E.
\end{aligned}$$

It preserves edge-integrals and thus satisfies $b(u - I_F^2 u, q_h) = 0 \; \forall \, u \; \forall \, q_h$. Furthermore, it is continuous with respect to

$$\|I_F^2 u\|_{H^1} \preceq \|u\|_{H^1} + h^{-1} \|u\|_{L_2}$$

Thus, the combined operator I_F is continuous:

$$\begin{aligned} \|I_F u\|_{H^1} &\preceq \|\Pi_h u\|_{H^1} + \|I_F^2(u - \Pi_h u)\|_{H1} \\ &\preceq \|u\|_{H^1} + \|u - \Pi_h u\|_{H^1} + h^{-1}\|u - \Pi_h u\|_{L_2} \\ &\preceq \|u\|_{H^1} \end{aligned}$$

It also satisfies the constraints:

$$b(u - I_F u, q_h) = b(u - \Pi_h u - I_F^2(u - \Pi_h u), q_h) = b((Id - I_F^2)(u - \Pi_h u), q_h) = 0$$

Error estimates are

$$\|u - u_h\|_{H^1} + \|p - p_h\|_{L_2} \preceq \inf_{v_h, q_h} \|u - v_h\|_{H^1} + \|p - q_h\|_{L_2} = O(h)$$

Although we approximate u_h with P^2-elements, the bad approximation of p leads to first order convergence, only. This element is considered to be sub-optimal.

A solution is the the pairing

$$V_h = P^{2+} \qquad Q_h = P^{1,nc},$$

where P^{2+} is the second order space enriched with cubic bubbles:

$$P^{2+}(\mathcal{T}) = \{v_h \in H^1 : v_{h|T} \in P^3(T), v_{h|E} \in P^2(E)\}$$

It leads to second order convergence. Since the costs of a method depend mainly on the coupling dofs, the price for the additional bubble is low.

Elements with continuous pressure

Although the pressure p is only in L_2, we may approximate it with continuous elements. The so called mini-element is

$$V_h = P^{1+} \qquad Q_h = P^{1,cont},$$

where P^{1+} is P^1 enriched by the cubic bubble. The continuous pressure allows integration by parts:

$$\int \operatorname{div} u \, q_h = -\int u \nabla q_h$$

The gradient of q_h is element-wise constant. We thus construct a Fortin-operator preserving element-wise mean values. Again, we use the Clément operator and a correction operator:

$$u_h := I_F u = \Pi_h u + I_F^2(u - \Pi_h u)$$

The correction is now defined as

$$\begin{aligned} I_F^2 u &= 0 && \text{on} \cup E \\ \textstyle\int_T I_F^2 u &= \textstyle\int_T u && \forall \, \text{elements } T. \end{aligned}$$

It satisfies $b(u - I_F^2 u, q_h) = 0 \qquad \forall\, u\, \forall\, q_h$, and, as above:

$$\|I_F^2 u\|_{H^1} \preceq \|u\|_{H^1} + h^{-1}\|u\|_{L_2}$$

Thus, the combined operator is a Fortin operator. This method is $O(h)$ convergent.

Another (essentially) stable pair is $P^2 \times P^{1,cont}$ (the popular Taylor Hood element). Its analysis is more involved. It requires the additional assumption that no two edge of one element are on the domain boundary. Its convergence rate is $O(h^2)$.

8.2 Elasticity

We start with a one-dimensional model. Take a beam which is loaded by a force density f in longitudinal (x) direction. We are interested in the displacement $u(x)$ in x direction.

The variables are

- The *strain* ε: It describes the elongation. Take two points x and y on the beam. After deformation, their distance is $y + u(y) - (x + u(x))$. The relative elongation of the beam is
$$\frac{\{y + u(y) - (x + u(x))\} - (y - x)}{y - x} = \frac{u(y) - u(x)}{y - x}.$$
In the limit $y \to x$, this is u'. We define the strain ε as
$$\varepsilon = u'.$$
- The *stress* σ: It describes internal forces. If we cut the piece (x, y) out of the beam, we have to apply forces at x and y to keep that piece in equilibrium. This force is called stress σ. Equilibrium is
$$\sigma(y) - \sigma(x) + \int_x^y f(s)\, ds = 0,$$
or
$$\sigma' = -f$$

Hook's law postulates a linear relation between the strain and the stress:

$$\sigma = E\varepsilon.$$

Combining the three equations

$$\varepsilon = u' \qquad \sigma = E\varepsilon \qquad \sigma' = -f$$

leads to the second order equation for the displacement u:

$$-(Eu')' = f.$$

Boundary conditions are

- Dirichlet b.c.: Prescribe the displacement at the boundary
- Neumann b.c: Prescibe the stress at the boundary

Elasticity in more dimensions

We want to compute the deformation of the body $\Omega \subset \mathbb{R}^d$.

- The body is loaded with a volume force density $f : \Omega \to \mathbb{R}^d$.
- The displacement is described by a the vector-valued function

$$u : \Omega \to \mathbb{R}^d.$$

- The strain ε becomes a symmetric tensor in $\mathbb{R}^{d\times d}$. The elongation in the direction of the unit-vector n is

$$n^T \varepsilon n.$$

The (linearized!) relation between the displacement u and the strain is now

$$\varepsilon_{ij} = \frac{1}{2}\left\{\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right\},$$

or, in compact form

$$\varepsilon = \varepsilon(u) = \frac{1}{2}\left\{\nabla u + (\nabla u)^T\right\}.$$

If the displacement is a pure translation (u = const), then the strain vanishes. Also, if the displacement is a linearized (!) rotation, (in two dimensions $u = (u_x, u_y) = (y, -x)$, the strain vanishes. We call these deformations the rigid body motions:

$$\begin{aligned} R^{2D} &= \left\{\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + b\begin{pmatrix} y \\ -x \end{pmatrix} : a_1, a_2, b \in \mathbb{R}\right\} \\ R^{3D} &= \left\{a + b \times x : a, b \in \mathbb{R}^3\right\} \end{aligned}$$

- The stress becomes a tensor $\sigma \in \mathbb{R}^{d\times d}$. Consider the part $V \subset \Omega$. To keep V in equilibrium, on has to apply the surface force density σn at ∂V:

$$\int_{\partial V} \sigma n \, ds + \int_V f \, dx = 0.$$

Apply Gauss theorem to obtain the differential form

$$\operatorname{div} \sigma = -f.$$

The div-operator is applied for each row of σ. A further hypothesis, equilibrium of angular momentum, implies that σ is symmetric.

- Hook's law is now a relation between two second order tensors:

$$\sigma_{ij} = \sum_{kl} D_{ijkl} \varepsilon_{kl},$$

in short

$$\sigma = D\varepsilon,$$

where D is a fourth order tensor. For an isotropic material (same properties in all directions), the matrial law has the special structure

$$\sigma = 2\mu\,\varepsilon + \lambda \operatorname{tr}\{\varepsilon\}\, I.$$

The two parameters μ and λ are called Lamé's parameters. The trace tr is defined as $\operatorname{tr}\{\varepsilon\} = \sum_{i=1}^{d} \varepsilon_{ii}$.

Collecting the equations

$$\varepsilon = \varepsilon(u) \qquad \sigma = D\varepsilon \qquad \operatorname{div} \sigma = -f$$

leads to

$$-\operatorname{div} D\,\varepsilon(u) = f.$$

Multiplication with test-functions $v : \Omega \to \mathbb{R}^d$, and integrating by parts leads to

$$\int_\Omega D\,\varepsilon(u) : \nabla v\, dx = \int f \cdot v\, dx \qquad \forall\, v$$

The operator ':' is the inner product for matrices, $A : B = \sum_{ij} A_{ij} B_{ij}$. Next, we use that $\sigma = D\varepsilon(u)$ is symmetric. Thus, $\sigma : \nabla v = \sigma : (\nabla v)^T = \sigma : \frac{1}{2}\{\nabla v + (\nabla v)^T\}$.

The equations of elasticity in weak form read as: Find $u \in V = [H^1_{0,D}(\Omega)]^d$ such that

$$\int_\Omega D\varepsilon(u) : \varepsilon(v)\, dx = \int_\Omega f \cdot v\, dx \qquad \forall\, v \in V.$$

Displacement (Dirichlet) boundary conditions ($u = u_D$ at Γ_D) are essential b.c., and are put into the space V. Neumann boundary conditions (natural b.c.) model surface forces $sigman = g$, and lead to the additional term $\int_{\Gamma_N} g \cdot v\, ds$ on the right hand side.

The bilinear-form in the case of an isotropic material reads as

$$\int 2\mu\,\varepsilon(u) : \varepsilon(v) + \lambda \operatorname{div} u\, \operatorname{div} v\, dx.$$

We assume a positive definite material law

$$D\varepsilon : \varepsilon \succeq \varepsilon : \varepsilon \qquad \forall \text{symmetric } \varepsilon \in \mathbb{R}^{d\times d}$$

Theorem 124. *Assume that the Dirichlet boundary Γ_D has positive measure. Then the equations of elasticity are well posed in $[H^1]^d$.*

Proof: Continuity of the bilinear-form and the linear-form are clear. Ellipticity of the bilinear-form follows from the positive definite matrial law, and the (non-trivial) Korn inequality

$$\int_\Omega \varepsilon(u) : \varepsilon(v)\, dx \succeq \|u\|^2_{H^1(\Omega)} \qquad \forall\, u \in [H^1_{0,D}]^d$$

The Lax-Milgram theorem proves a unique solution u. □

The discretization of the elasticity problem is straight forward. Take a finite dimensional sub-space $V_h \subset V$, and perform Galerkin projection. One may use the 'standard' nodal finite elements for each component.

Structural mechanics

Many engineering applications involve thin structures (walls of a building, body of a car, ...). On thin structures, the standard approach has a problem: One observed that the simulation results get worse as the thickness decreases. The explanation is that the constant in Korn's inequality gets small for thin structures. To understand and overcome this problem, we go over to beam, plate and shell models.

We consider a thin ($t \ll 1$) two-dimensional body

$$\Omega = I \times (-t/2, t/2) \qquad \text{with} \qquad I = (0,1)$$

The goal is to derive a system of one-dimensional equations to describe the two-dimensional deformation. This we obtain by a semi-discretization. Define

$$\widetilde{V}_M = \left\{ \begin{pmatrix} u_x(x,y) \\ u_y(x,y) \end{pmatrix} \in V : u_x(x,y) = \sum_{i=0}^{M_x} u_x^i(x) y^i,\ u_y(x,y) = \sum_{i=0}^{M_y} u_y^i(x) y^i \right\}.$$

This function space on $\Omega \subset \mathbb{R}^2$ is isomorph to a one-dimensional function space with values in $\mathbb{R}^{M_x+M_y+2}$. We perform semi-discretization by searching for $\tilde{u} \in \widetilde{V}_M$ such that

$$A(\tilde{u}, \tilde{v}) = f(\tilde{v}) \qquad \forall\, \tilde{v} \in \widetilde{V}_M.$$

As $M_x, M_y \to \infty$, $\widetilde{V}_M \to V$, and we obtain convergence $\tilde{u} \to u$.

The lowest order (qualitative) good approximating semi-discrete space is to set $M_x = 1$ and $M_y = 0$. This is

$$\widetilde{V} = \left\{ \begin{pmatrix} U(x) - \beta(x) y \\ w(x) \end{pmatrix} \right\}$$

Evaluating the bilinear-form (of an isotropic material) leads to

$$\begin{aligned} A\left(\begin{pmatrix} U - y\beta \\ w \end{pmatrix}, \begin{pmatrix} \tilde{U} - y\tilde{\beta} \\ \tilde{w} \end{pmatrix} \right) &= (2\mu + \lambda) t \int_0^1 U' \tilde{U}'\, dx + \\ & (2\mu + \lambda) \frac{t^3}{12} \int_0^1 \beta' \tilde{\beta}' + 2\mu \frac{t}{2} \int_0^1 (w' - \beta)(\tilde{w}' - \tilde{\beta})\, dx \end{aligned}$$

The meaning of the three functions is as follows. The function $U(x)$ is the average (over the cross section) longitudinal displacement, $w(x)$ is the vertical displacement. The function β is the linearized rotation of the normal vector.

We assume that the load $f(x, y)$ does not depend on y. Then, the linear form is

$$f\begin{pmatrix} \tilde{U} - y\tilde{\beta} \\ \tilde{w} \end{pmatrix} = t\int_0^1 f_x \tilde{U}\,dx + t\int_0^1 f_y \tilde{w}\,dx$$

The semi-discretization in this space leads to two decoupled problems. The first one describes the longitudinal displacement: Find $U \in H^1(I)$ such that

$$(2\mu + \lambda)t\int_0^1 U'\tilde{U}'\,dx = t\int_0^1 f_x \tilde{U}\,dx \qquad \forall\,\tilde{U} \in H^1(I).$$

The small thickness parameter t cancels out. It is a simple second order problem for the longitudinal displacement.

The second problems involves the 1D functions w and β: Find $(w, \beta) \in V = ?$ such that

$$(2\mu + \lambda)\frac{t^3}{12}\int_0^1 \beta'\tilde{\beta}'\,dx + \mu t\int_0^1 (w' - \beta)(\tilde{w}' - \tilde{\beta})\,dx = t\int_0^1 f_y \tilde{w}\,dx \qquad \forall\,(\tilde{w}, \tilde{\beta}) \in V$$

The first term models bending. The derivative of the rotation β is (approximative) the curvature of the deformed beam. The second one is called the shear term: For thin beams, the angle $\beta \approx \tan\beta$ is approximatively w'. This term measures the difference $w' - \beta$. This second problem is called the Timoshenko beam model.

For simplification, we skip the parameters μ and λ, and the constants. We rescale the equation by dividing by t^3: Find (w, β) such that

$$\int \beta'\tilde{\beta}'\,dx + \frac{1}{t^2}\int (w' - \beta)(\tilde{w}' - \tilde{\beta})\,dx = \int t^{-2} f\tilde{w}\,dx. \tag{8.4}$$

This scaling in t is natural. With $t \to 0$, and a force density $f \sim t^2$, the deformation converges to a limit. We define the scaled force density

$$\tilde{f} = t^{-2} f$$

In principle, this is a well posed problem in $[H^1]^2$:

Lemma 125. *Assume boundary conditions $w(0) = \beta(0) = 0$. The bilinear-form $A((w, \beta), (\tilde{w}, \tilde{\beta}))$ of (8.4) is continuous*

$$A((w, \beta), (\tilde{w}, \tilde{\beta})) \preceq t^{-2}(\|w\|_{H^1} + \|\beta\|_{H^1})(\|\tilde{w}\|_{H^1} + \|\tilde{\beta}\|_{H^1})$$

and coercive

$$A((w, \beta), (w, \beta)) \geq \|w\|_{H^1}^2 + \|\beta\|_{H^1}^2$$

Proof: ...

As the thickness t becomes small, the ratio of the continuity and coercivity bounds becomes large ! This ratio occurs in the error estimates, and indicates problems. Really, numerical computations show bad convergence for small thickness t.

The large coefficient in front of the term $\int (w' - \beta)(\tilde{w}' - \tilde{\beta})$ forces the difference $w' - \beta$ to be small. If we use piece-wise linear finite elements for w and β, then w_h' is a piece-wise constant function, and β_h is continuous. If $w_h' - \beta_h \approx 0$, then β_h must be a constant function !

The idea is to weaken the term with the large coefficient. We plug in the projection P^0 into piece-wise constant functions: Find (w_h, β_h) such that

$$\int \beta_h' \tilde{\beta}_h' \, dx + \frac{1}{t^2} \int P^0(w_h' - \beta_h) \, P^0(\tilde{w}_h' - \tilde{\beta}_h) \, dx = \int \tilde{f} \tilde{w}_h \, dx. \tag{8.5}$$

Now, there are finite element functions w_h and β_h fulfilling $P^0(w_h' - \beta_h) \approx 0$.

In the engineering community there are many such tricks to modify the bilinear-form. Our goal is to understand and analyze the obtained method.

Again, the key is a mixed method. Start from equation (8.4) and introduce a new variable

$$p = t^{-2}(w' - \beta). \tag{8.6}$$

Using the new variable in (8.4), and formulating the definition (8.6) of p in weak form leads to the bigger system: Find $(w, \beta) \in V$ and $p \in Q$ such that

$$\begin{array}{rcrcll} \int \beta' \tilde{\beta}' \, dx & + & \int (\tilde{w}' - \tilde{\beta}) p \, dx & = & \int \tilde{f} \tilde{w} \, dx & \forall \, (\tilde{w}, \tilde{\beta}) \in V \\ \int (w' - \beta) \tilde{p} \, dx & - & t^2 \int p \, \tilde{p} \, dx & = & 0 & \forall \, \tilde{p} \in Q. \end{array} \tag{8.7}$$

This is a mixed formulation of the abstract structure: Find $u \in V$ and $p \in Q$ such that

$$\begin{array}{rcrcll} a(u, v) & + & b(v, p) & = & f(v) & \forall \, v \in V, \\ b(u, q) & - & t^2 c(p, q) & = & 0 & \forall \, q \in Q. \end{array} \tag{8.8}$$

The big advantage now is that the parameter t does not occur in the denominator, and the limit $t \to 0$ can be performed.

This is a family of well posed problems.

Theorem 126 (extended Brezzi). *Assume that the assumptions of Theorem 101 are true. Furthermore, assume that*

$$a(u, u) \geq 0,$$

and $c(p, q)$ is a symmetric, continuous and non-negative bilinear-form. Then, the big form

$$B((u, p), (v, q)) = a(u, v) + b(u, q) + b(v, p) - t^2 \, c(p, q)$$

is continuous and stable uniformly in $t \in [0, 1]$.

We check Brezzi's condition for the beam model. The spaces are $V = [H^1]^2$ and $Q = L_2$. Continuity of the bilinear-forms $a(.,.)$, $b(.,.)$, and $c(.,.)$ is clear. The LBB condition is

$$\sup_{w,\beta} \frac{\int (w' - \beta) q \, dx}{\|w\|_{H^1} + \|\beta\|_{H^1}} \succeq \|q\|_{L_2}$$

We construct a candidate for the supremum:

$$w(x) = \int_0^x q(s) \, ds \qquad \text{and} \qquad \beta = 0$$

Then

$$\frac{\int (w' - \beta) q \, dx}{\|w\|_{H^1} + \|\beta\|_{H^1}} \succeq \frac{\int q^2 \, dx}{\|w'\|} = \|q\|_{L_2}$$

Finally, we have to check kernel ellipticity. The kernel is

$$V_0 = \{(w, \beta) : \beta = w'\}.$$

On V_0 there holds

$$\begin{aligned} \|w\|_{H^1}^1 + \|\beta\|_{H^1}^2 &\preceq \|w'\|^2 + \|\beta\|_{H^1}^2 = \|\beta\|_{L_2}^2 + \|\beta\|_{H^1}^2 \\ &\preceq \|\beta'\|_{L_2} = a((w, \beta), (w, \beta)) \end{aligned}$$

The lowest order finite element discretization of the mixed system is to choose continuous and piece-wise linear elements for w_h and β_h, and piecewise constants for p_h. The discrete problem reads as: Find $(w_h, \beta_h) \in V_h$ and $p_h \in Q_h$ such that

$$\begin{array}{cccclc} \int \beta_h' \tilde{\beta}_h' \, dx & + & \int (\tilde{w}_h' - \beta_h) p_h \, dx & = & \int \tilde{f} \tilde{w}_h \, dx & \forall \, (w_h, \beta_h) \in V_h \\ \int (w_h' - \beta_h) \tilde{p}_h \, dx & - & t^2 \int p_h \, \tilde{p}_h \, dx & = & 0 & \forall \, \tilde{p}_h \in Q_h. \end{array} \tag{8.9}$$

This is a inf-sup stable system on the discrete spaces V_h and Q_h. This means, we obtain the **uniform** a priori error estimate

$$\begin{aligned} \|(w - w_h, \beta - \beta_h)\|_{H_1} + \|p - p_h\|_{L_2} &\preceq \inf_{\tilde{w}_h, \tilde{\beta}_h, \tilde{p}_h} \|(w - \tilde{w}_h, \beta - \tilde{\beta}_h)\|_{H_1} + \|p - \tilde{p}_h\|_{L_2} \\ &\preceq h \, \{\|w\|_{H^2} + \|\beta\|_{H^2} + \|p\|_{H^1}\} \end{aligned}$$

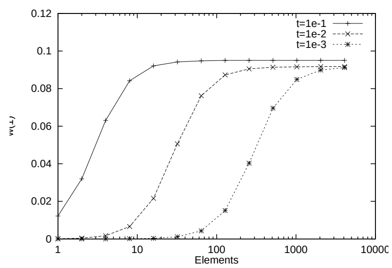
The required regularity is realistic.

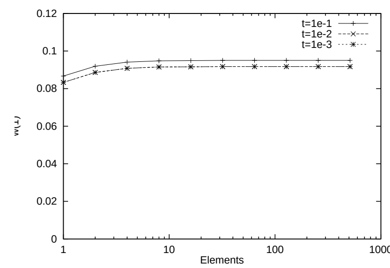
The second equation of the discrete mixed system (8.9) states that

$$p_h = t^{-2} P^0 (w_h' - \beta_h)$$

If we insert this observation into the first row, we obtain exactly the discretization method (8.5) ! Here, the mixed formulation is a tool for analyzing a non-standard (primal) discretization method. Both formulations are equivalent. They produce exactly the same finite element functions. The mixed formulation is the key for the error estimates.

The two pictures below show simulations of a Timoshenko beam. It is fixed at the left end, the load density is constant one. We compute the vertical deformation $w(1)$ at the right boundary. We vary the thickness t between 10^{-1} and 10^{-3}. The left pictures shows the result of a standard conforming method, the right picture shows the results of the method using the projection. As the thickness decreases, the standard method becomes worse. Unless h is less than t, the results are completely wrong ! The improved method converges uniformly well with respect to t:





8.3 Maxwell equations

Maxwell equations describe electro-magnetic fields. We consider the special case of stationary magnetic fields. Maxwell equations are three-dimensional.

A magnetic field is caused by an electric current. We suppose that a current density

$$j \in [L_2(\Omega)]^3$$

is given. (Stationary) currents do not have sources, i.e., div $j = 0$.

The involved (unknown) fields are

- The magnetic flux B (in German: Induktion). The flux is free of sources, i.e.,

$$\operatorname{div} B = 0.$$

- The magnetic field intensity H (in German: magnetische Feldstärke). The field is related to the current density by Henry's law:

$$\int_S j \cdot n \, ds = \int_{\partial S} H \cdot \tau \, ds \qquad \forall \text{ Surfaces } S$$

By Stokes´ Theorem, one can derive Henry's law in differential form:

$$\operatorname{curl} H = j$$

The differential operator is curl = rot = $\nabla\times$. Both fields are related by a material law. The coefficient μ is called permeability:

$$B = \mu H$$

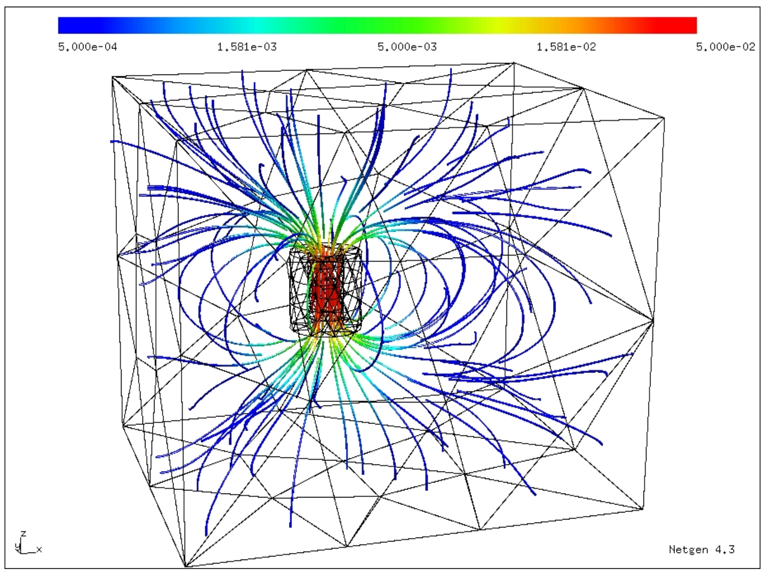
The coefficient μ is 10^3 to 10^4 times larger in iron (and other ferro-magnetic metals) as in most other media (air). In a larger range, the function $B(H)$ is also highly non-linear.

Collecting the equations we have

$$\text{div } B = 0 \qquad B = \mu H \qquad \text{curl } H = j \tag{8.10}$$

In principle, Maxwell equations are valid in the whole $\mathbb{R}^3$. For simulation, we have to truncate the domain and have to introduce artificial boundary conditions.

The picture below shows the magnetic field caused by a tangential current density in a coil:



Compare these equations to the diffusion equation $-\text{div } a\nabla u = f$. Here, we could introduce new unknowns $g = \nabla u$ and $\sigma = ag$. On simply connected domains, g is a gradient field if and only if $\text{curl}\, g = 0$. We could reformulate the equations as: Find vector fields g and σ such that

$$\text{curl } g = 0 \qquad \sigma = ag \qquad \text{div } \sigma = -f.$$

The system of magnetostatic equations looks similar. Only, the right hand side data is applied to the curl-equation, instead of the div-equation. In a similar way as curl $g = 0$ allows to introduce a scalar field u such that $g = \nabla u$, div $B = 0$ allows to introduce a vector potential A such that

$$B = \text{curl } A.$$

Inserting the vector-potential into the equations (8.10), one obtains the second order equation

$$\text{curl } \mu^{-1} \text{curl } A = j. \tag{8.11}$$

The two original fields B and H can be obtained from the vector potential A.

The vector-potential A is not uniquely defined by (8.11). One may add a gradient field to A, and the equation is still true. To obtain a unique solution, the so called Coloumb-Gauging can be applied:

$$\operatorname{div} A = 0. \tag{8.12}$$

As usual, we go over to the weak form. Equations (8.11) and (8.12) together become: Find A such that

$$\int_\Omega \mu^{-1} \operatorname{curl} A \operatorname{curl} v \, dx = \int_\Omega j \cdot v \, dx \qquad \forall\, v \in ?$$

and

$$\int_\Omega A \cdot \nabla \psi \, dx = 0.$$

We want to choose the same space for A and the according test functions v. But, then we have more equations than unknowns. The system is still solvable, since we have made the assumption $\operatorname{div} j = 0$, and thus j is in the range of the curl- operator. To obtain a symmetric system, we add a new scalar variable φ. The problem is now: Find $A \in V = ?$ and $\varphi \in Q = H^1/\mathbb{R}$ such that

$$\begin{array}{llllll} \int \mu^{-1} \operatorname{curl} A \cdot \operatorname{curl} v \, dx & + & \int \nabla\varphi \cdot v \, dx & = & \int j \cdot v \, dx & \forall\, v \in V \\ \int A \cdot \nabla\psi \, dx & & & = & 0 & \forall\, \psi \in Q \end{array} \tag{8.13}$$

The proper space V is the $H(\operatorname{curl})$:

$$H(\operatorname{curl}) = \{ v \in [L_2(\Omega)]^3 : \operatorname{curl} v \in [L_2(\Omega)]^3 \}$$

Again, the differential operator curl is understood in the weak sense. The canonical norm is

$$\|v\|_{H(\operatorname{curl})} = \left\{ \|v\|_{L_2}^2 + \|\operatorname{curl} v\|_{L_2}^2 \right\}^{1/2}.$$

Similar to H^1 and $H(\operatorname{div})$, there exists a trace operator for $H(\operatorname{curl})$. Now, only the tangential components of the boundary values are well defined:

Theorem 127 (Trace theorem)**.** *There exists a tangential trace operator* $\operatorname{tr}_\tau v : H(\operatorname{curl}) \to W(\partial\Omega)$ *such that*

$$\operatorname{tr}_\tau v = (v|_{\partial\Omega})_\tau$$

for smooth functions $v \in [C(\overline{\Omega})]^3$.

Theorem 128. *Let* $\Omega = \cup \Omega_i$. *Assume that* $u|_{\Omega_i} \in H(\operatorname{curl}, \Omega_i)$, *and the tangential traces are continuous across the interfaces* γ_{ij}. *Then* $u \in H(\operatorname{curl}, \Omega)$.

The theorems are according to the ones we have proven for $H(\operatorname{div})$. But, the proofs (in $\mathbb{R}^3$) are more involved.

The gradient operator ∇ relates the space H^1 and $H(\operatorname{curl})$:

$$\nabla : H^1 \to H(\operatorname{curl})$$

Furthermore, the kernel space

$$H^0(\text{curl}) = \{v \in H(\text{curl}) : \text{curl } v = 0\}$$

is exactly the range of the gradient:

$$H^0(\text{curl}) = \nabla H^1$$

Theorem 129. *The mixed system (8.13) is a well posed problem on* $H(\text{curl}) \times H^1/\mathbb{R}$.

Proof: The bilinear-forms

$$a(A, v) = \int \mu^{-1} \operatorname{curl} A \cdot \operatorname{curl} v \, dx$$

and

$$b(v, \varphi) = \int v \cdot \nabla\varphi \, dx$$

are continuous w.r.t. the norms of $V = H(\text{curl})$ and $Q = H^1/\mathbb{R}$.

The LBB-condition in this case is trivial. Choose $v = \nabla\varphi$:

$$\sup_{v \in H(\text{curl})} \frac{\int v \nabla\varphi \, dx}{\|v\|_{H(\text{curl})}} \geq \frac{\int \nabla\varphi \cdot \nabla\varphi \, dx}{\|\nabla\varphi\|_{H(\text{curl})}} = \frac{\|\nabla\varphi\|_{L_2}^2}{\|\nabla\varphi\|_{L_2}} = \|\nabla\varphi\|_{L_2} \simeq \|\varphi\|_Q$$

The difficult part is the kernel coercivity of $a(.,.)$. The norm involves also the L_2-norm, while the bilinear-form only involves the semi-norm $\| \text{curl } v\|_{L_2}$. Coercivity cannot hold on the whole V: Take a gradient function $\nabla\psi$. On the kernel, the L_2-norm is bounded by the semi-norm:

$$\|v\|_{L_2} \preceq \| \text{curl } v\| \qquad \forall\, v \in V_0,$$

where

$$V_0 = \{v \in H(\text{curl}) : \int v \nabla\varphi \, dx = 0 \ \ \forall\, \varphi \in H^1\}$$

This is a Friedrichs-like inequality.

Finite elements in $H(\text{curl})$

We construct finite elements in three dimensions. The trace theorem implies that functions in $H(\text{curl})$ have continuous tangential components across element boundaries (=faces).

We design tetrahedral finite elements. The pragmatic approach is to choose the element space as $V_T = P^1$, and choose the degrees of freedom as the tangential component along the edges in the end-points of the edges. The dimension of the space is $3 \times \dim\{P^1\} = 3 \times 4 = 12$, the degrees of freedom are 2 per edge, i.e., $2 \times 6 = 12$. They are also linearly independent. In each face, the tangential component has 2 components, and is linear. Thus, the tangential component has dimension 6. These 6 values are defined by the 6 degrees of freedom of the 3 edges in the face. Neighboring elements share this 6 degrees of freedom in the face, and thus have the same tangential component.

There is a cheaper element, called Nédélec, or edge-element. It has the same accuracy for the curl-part (the B-field) as the P^1-element. It is similar to the Raviart-Thomas element. It contains all constants, and some linear polynomials. All 3 components are defined in common. The element space is

$$V_T = \{a + b \times x : a, b \in \mathbb{R}^3\}.$$

These are 6 coefficients. For each of the 6 edges of a tetrahedron, one chooses the integral of the tangential component along the edge

$$\psi_{E_i}(u) = \int_{E_i} u \cdot \tau_{E_i} \, ds.$$

Lemma 130. *The basis function φ_{E_i} associated with the edge E_i is*

$$\varphi_{E_i} = \lambda_{E_i^1} \nabla \lambda_{E_i^2} - \nabla \lambda_{E_i^2} \lambda_{E_i^1},$$

where E_i^1 and E_i^2 are the two vertex numbers of the edge, and $\lambda_1, \ldots \lambda_4$ are the vertex shape functions.

Proof:

- These functions are in V_T
- If $i \neq j$, then $\psi_{E_j}(\varphi_{E_i}) = 0$.
- $\psi_{E_i}(\varphi_{E_i}) = 1$

Thus, edge elements belong to $H(\text{curl})$. Next, we will see that they have also very interesting properties.

The de'Rham complex

The spaces H^1, $H(\text{curl})$, $H(\text{div})$, and L_2 form a sequence:

$$H^1 \xrightarrow{\nabla} H(\text{curl}) \xrightarrow{\text{curl}} H(\text{div}) \xrightarrow{\text{div}} L^2$$

Since $\nabla H^1 \subset [L_2]^3$, and $\operatorname{curl} \nabla = 0$, the gradients of H^1 functions belong to $H(\text{curl})$. Similar, since $\operatorname{curl} H(\text{curl}) \subset [L_2]^3$, and $\operatorname{div} \operatorname{curl} = 0$, the curls of $H(\text{curl})$ functions belong to $H(\text{div})$.

The sequence is a *complete* sequence. This means that the kernel of the right differential operator is exactly the range of the left one (on simply connected domains). We have used this property already in the analysis of the mixed system.

The same property holds on the discrete level: Let

W_h	be the nodal finite element sub-space of H^1
V_h	be the Nédélec (edge) finite element sub-space of $H(\text{curl})$
Q_h	be the Raviart-Thomas (face) finite element sub-space of $H(\text{div})$
S_h	be the piece-wise constant finite element sub-space of L_2

Theorem 131. *The finite element spaces form a complete sequence*

$$W_h \xrightarrow{\nabla} V_h \xrightarrow{\text{curl}} Q_h \xrightarrow{\text{div}} S_h$$

Now, we discretize the mixed formulation (8.13) by choosing edge-finite elements for $H(\text{curl})$, and nodal finite elements for H^1: Find $A_h \in V_h$ and $\varphi_h \in W_h$ such that

$$\begin{array}{lclll} \int \mu^{-1} \operatorname{curl} A_h \cdot \operatorname{curl} v_h \, dx & + & \int \nabla \varphi_h \cdot v_h \, dx & = \int j \cdot v_h \, dx & \forall\, v_h \in V_h \\ \int A_h \cdot \nabla \psi_h \, dx & & & = 0 & \forall\, \psi_h \in W_h \end{array} \tag{8.14}$$

The stability follows (roughly) from the discrete sequence property. The verification of the LBB condition is the same as on the continuous level. The kernel of the $a(.,.)$- form are the discrete gradients, the kernel of the $b(.,.)$-form is orthogonal to the gradients. This implies solvability. The discrete kernel-coercivity (with h-independent constants) is true (nontrivial).

The complete sequences on the continuous level and on the discrete level are connected in the de'Rham complex: Choose the canonical interpolation operators (vertex-interpolation I^W, edge-interpolation I^V, face-interpolation I^Q, L_2-projection I^S). This relates the continuous level to the discrete level:

$$\begin{array}{ccccccc} H^1 & \xrightarrow{\nabla} & H(\text{curl}) & \xrightarrow{\text{curl}} & H(\text{div}) & \xrightarrow{\text{div}} & L^2 \\ \downarrow I^W & & \downarrow I^V & & \downarrow I^Q & & \downarrow I^S \\ W_h & \xrightarrow{\nabla} & V_h & \xrightarrow{\text{curl}} & Q_h & \xrightarrow{\text{div}} & S_h\,. \end{array} \tag{8.15}$$

Theorem 132. *The diagram (8.15)* commutes*:*

$$I^V \nabla = \nabla I^W \qquad I^Q \operatorname{curl} = \operatorname{curl} I^V \qquad I^S \operatorname{div} = \operatorname{div} I^Q$$

Proof: We prove the first part. Note that the ranges of both, ∇I^W and $I^V \nabla$, are in V_h. Two functions in V_h coincide if and only if all functionals coincide. It remains to prove that

$$\int_E (\nabla I^W w) \cdot \tau \, ds = \int_E (I^V \nabla w) \cdot \tau \, ds$$

Per definition of the interpolation operator I^V there holds

$$\int_E (I^V \nabla w) \cdot \tau \, ds = \int_E \nabla w \cdot \tau \, ds$$

Integrating the tangential derivative gives the difference

$$\int_E \nabla w \cdot \tau \, ds = \int_E \frac{\partial w}{\partial \tau} \, ds = w(E^2) - w(E^1)$$

Starting with the left term, and using the property of the nodal interpolation operator, we obtain

$$\int_E (\nabla I^W w) \cdot \tau \, ds = (I^W w)(E^2) - (I^W w)(E^1) = w(E^2) - w(E^1).$$

We have already proven the commutativity of the $H(\text{div}) - L_2$ part of the diagram. The middle one involves Stokes´ theorem. $\square$

This is the key for interpolation error estimates. E.g., in $H(\text{curl})$ there holds

$$\begin{aligned} \|u - I^V u\|^2_{H(\text{curl})} &= \|u - I^V u\|^2_{L_2} + \|\operatorname{curl}(I - I^V)u\|^2_{L_2} \\ &= \|u - I^V u\|^2_{L_2} + \|(I - I^Q)\operatorname{curl} u\|^2_{L_2} \\ &\preceq h^2 \, \|u\|^2_{H^1} + h^2 \|\operatorname{curl} u\|^2_{H^1} \end{aligned}$$

Since the estimates for the L_2-term and the curl-term are separate, one can also scale each of them by an arbitrary coefficient.

The sequence is also compatible with transformations. Let $F : \widehat{T} \to T$ be an (element) transformation. Choose

$$\begin{aligned} w(F(x)) &= \hat{w}(x) && \\ v(F(x)) &= (F')^{-T} \hat{v}(x) && \text{(covariant transformation)} \\ q(F(x)) &= (\det F')^{-1} (F') q(x) && \text{(Piola-transformation)} \\ s(F(x)) &= (\det F')^{-1} \hat{s}(x) && \end{aligned}$$

Then

$$\begin{aligned} \hat{v} = \nabla \hat{w} \quad &\Rightarrow \quad v = \nabla w \\ \hat{q} = \operatorname{curl} \hat{v} \quad &\Rightarrow \quad q = \operatorname{curl} v \\ \hat{s} = \operatorname{div} \hat{q} \quad &\Rightarrow \quad s = \operatorname{div} q \end{aligned}$$

Using these transformation rules, the implementation of matrix assembling for $H(\text{curl})$-equations is very similar to the assembling for H^1 problems (mapping to reference element).

Chapter 9

Parabolic partial differential equations

PDEs involving first order derivatives in time, and an elliptic differential operator in space, are called parabolic PDEs. For example, time dependent heat flow is described by a parabolic PDE.

Let $\Omega \subset \mathbb{R}^d$, and $Q = \Omega \times (0, T)$. Consider the initial-boundary value problem

$$\frac{\partial u(x,t)}{\partial t} - \operatorname{div}(a(x) \nabla_x u(x,t)) = f(x,t) \qquad (x,t) \in Q,$$

with boundary conditions

$$\begin{aligned} u(x,t) &= u_D(x,t) \qquad (x,t) \in \Gamma_D \times (0,T), \\ a(x) \frac{\partial u}{\partial n} &= g(x,t) \qquad (x,t) \in \Gamma_N \times (0,T), \end{aligned}$$

and initial conditions

$$u(x,0) = u_0(x) \qquad x \in \Omega.$$

Weak formulation in space: Find $u : [0,T] \to H^1_{0,D}(\Omega)$ such that

$$\int_\Omega \partial_t u(x,t) v(x)\, dx + \int_\Omega a \nabla u(x,t) \cdot \nabla v(x,t)\, dx = \int_\Omega f(x,t) v(x,t)\, dx + \int_{\Gamma_N} g(x,t) v(x,t)\, dx$$

$$\forall\, v \in H^1_{0,D},\ t \in (0,T]$$

In abstract form: Find $u : [0,T] \to V$ s.t.

$$(u'(t), v)_{L_2} + a(u(t), v) = \langle f(t), v \rangle \qquad \forall\, v \in V,\ t \in (0,T]$$

In operator form (with $\langle Au, v \rangle = a(u,v)$):

$$u'(t) + Au(t) = f(t) \qquad \in V^*$$

Function spaces:

$$X = L_2((0,T),V) \qquad X^* = L_2((0,T),V^*)$$

with norms

$$\|v\|_X = \left(\int_0^T \|v(t)\|_V^2 \, dt\right)^{1/2} \qquad \|v\|_{X^*} = \left(\int_0^T \|v(t)\|_{V^*}^2 \, dt\right)^{1/2}$$

Definition 133. *Let $u \in L_2((0,T),V)$. It has a weak derivative $w \in L_2((0,T),V^*)$ if*

$$\int_0^T \varphi(t) \left\langle w, v \right\rangle_{V^* \times V} \, dt = -\int_0^T \varphi'(t)(u,v)_{L_2} \, dt \qquad \forall\, v \in V,\ \forall\, \varphi \in C_0^\infty(0,T)$$

Definition 134.

$$H^1((0,T),V;L_2) = \{v \in L_2((0,T),V) : v' \in L_2((0,T),V^*)\}$$

with norm

$$\|v\|_{H^1}^2 = \|v\|_X^2 + \|v'\|_{X^*}^2.$$

This space is a one-dimensional Sobolev space with range in a Hilbert space.

Theorem 135 (Trace theorem)**.** *Point evaluation is continuous:*

$$\max_{t \in [0,T]} \|v(t)\|_{L_2} \preceq \|v\|_{H^1}$$

This allows the formulation of the initial value $u(0) = u_0$.

Theorem 136. *Assume that $a(.,.)$ is coercive*

$$a(u,u) \geq \mu_1 \|u\|_V^2 \quad \forall\, u \in V$$

and continuous

$$a(u,v) \leq \mu_2 \|u\|_V \, \|v\|_V \qquad \forall\, u, v \in V.$$

Then, the parabolic problem has a unique solution depending continuously on the right hand side and the initial conditions:

$$\|u\|_{H^1((0,T),V;L_2)} \quad \preceq \quad \|u_0\|_{L_2} + \|f\|_{L_2((0,T),V^*}.$$

We only prove stability: Choose test functions $v = u(t)$:

$$(u'(t), u(t))_{L_2} + a(u(t), u(t)) = \langle f(t), u(t) \rangle$$

Use that

$$\frac{d}{dt}\|u(t)\|_{L_2}^2 = 2(u'(t), u(t))_{L_2},$$

and integrate the equation over $(0,T)$:

$$\begin{aligned}\frac{1}{2}\left\{\|u(T)\|_{L_2}^2 - \|u_0\|_{L_2}^2\right\} &= \int_0^T \langle f(s), u(s)\rangle - a(u(s),u(s))\,ds \\ &\leq \int_0^T \|f(s)\|_{V^*}\|u(s)\|_V - \mu_1\|u(s)\|_V^2\,ds \\ &\leq \|f\|_{X^*}\|u\|_X - \mu_1\|u\|_X^2\end{aligned}$$

Since $\|u(T)\| \geq 0$, one has

$$\mu_1\|u\|_X^2 - \|f\|_{X^*}\|u\|_X \leq \frac{1}{2}\|u_0\|_{L_2}$$

Solving the quadatic inequality, one obtains the bound

$$\|u\|_X \leq \frac{1}{2\mu_1}\left\{\|f\|_{X^*} + \sqrt{\|f\|_{X^*}^2 + 2\mu_1\|u_0\|_{L_2}^2}\right\}$$

The bound $\|u'\|_{L_2((0,T),V^*)}$ follows from $u'(t) = f(t) - Au(t)$.

9.1 Semi-discretization

We start with a discretization in space. Choose a (finite element) sub-space $V_h \subset V$. The Galerkin discretiztaion is: Find $u : [0,T] \to V_h$ such that

$$(u_h'(t), v_h)_{L_2} + a(u_h(t), v_h) = \langle f(t), v_h\rangle \qquad \forall\, v_h \in V_h,\ \forall\, t \in (0,T],$$

and initial conditions

$$(u_h(0), v_h)_{L_2} = (u_0, v_h)_{L_2} \qquad \forall\, v_h \in V_h.$$

Choose a basis $\{\varphi_1, \ldots \varphi_N\}$ of V_h. Expand the solution w.r.t. this basis:

$$u_h(x,t) = \sum_{i=1}^N u_i(t)\varphi_i(x),$$

and choose test functions $v = \varphi_j$. With the matrices

$$M = \left((\varphi_j, \varphi_i)_{L_2}\right)_{i,j=1,\ldots,N} \qquad A = \left(a(\varphi_j, \varphi_i)\right)_{i,j=1,\ldots,N},$$

and the t-dependent vector

$$f(t) = \left(\langle f(t), \varphi_j\rangle\right)_{i=1,\ldots,N},$$

one obtains the system of ordinary differential equations (ODEs)

$$Mu'(t) + Au(t) = f(t), \qquad u(0) = u_0$$

In general, the (mass) matrix M is non-diagonal. In the case of the (inexact) vertex integration rules, or non-conforming P_1-elements, M is a diagonal matrix. Then, this ODE can be efficiently reduced to explicit form

$$u'(t) + M^{-1}Au(t) = f(t)$$

Theorem 137. *There holds the error estimate*

$$\|u - u_h\|_{H^1((0,T),V;L_2)} \preceq \|(I - R_h)u\|_{H^1((0,T),V;L_2)},$$

where R_h is the Ritz projector

$$R_h : V \to V_h : \qquad a(R_h u, v_h) = a(u, v_h) \qquad \forall\, u \in V, \forall\, v_h \in V_h.$$

Proof: The error is split into two parts:

$$u(t) - u_h(t) = \underbrace{u(t) - R_h u(t)}_{\rho(t)} + \underbrace{R_h u(t) - u_h(t)}_{\Theta_h}$$

The first part, $u(t) - R_h u(t)$ is the elliptic discretization error, which can be bounded by Cea's lemma. To bound the second term, we use the properties for the continuous and the discrete formulation:

$$\begin{aligned} \langle f, v_h \rangle &= (u', v_h) + a(u, v_h) = (u', v_h) + a(R_h u, v_h) \\ &= (u_h', v_h) + a(u_h, v_h), \end{aligned}$$

i.e.,

$$(u' - u_h', v_h) + a(R_h u - u_h, v_h) = 0,$$

or

$$(R_h u' - u_h', v_h) + a(R_h u - u_h, v_h) = (R_h u' - u', v_h).$$

With the abbreviations from above we obtain the discrete parabolic equation for Θ_h:

$$\begin{aligned} (\Theta_h', v_h) + a(\Theta_h, v_h) &= (\rho', v_h) \\ \Theta_h(0) &= (I - R_h)u(0). \end{aligned}$$

The stability estimate, and the trace theorem bounds

$$\begin{aligned} \|\Theta_h\|_{H^1((0,T),V;L_2)} &\preceq \|(I - R_h)u(0)\|_{L_2(\Omega)} + \|\rho'\|_{L_2((0,T),V^*)} \\ &\preceq \|(I - R_h)u\|_{H^1((0,T),V;L_2)} \end{aligned}$$

$\square$

9.2 Time integration methods

Next, we discuss methods for solving the system of ODEs:

$$\begin{aligned} Mu'(t) + Au(t) &= f(t) \\ u(0) &= u_0 \end{aligned} \tag{9.1}$$

We focus on simple time integration rules and the specific properties arising from the space-discretization of parabolic PDEs. Let

$$0 = t_0 < t_1 < t_m = T,$$

a partitioning of the interval $[0, T]$. Define $\tau_j = t_{j+1} - t_j$. Integrating (9.1) over the intervalls leads to

$$M \{u(t_{j+1}) - u(t_j)\} + \int_{t_j}^{t_{j+1}} Au(s)\, ds = \int_{t_j}^{t_{j+1}} f(s)\, ds.$$

Next, we replace the integrals by numerical integration rules. The left-sided rectangle rule leads to

$$M\{u(t_{j+1}) - u(t_j)\} + \tau_j Au(t_j) = \tau_j f(t_j)$$

With the notation $u_j = u(t_j)$, this leads to the sequence of linear equations

$$Mu_{j+1} = Mu_j + \tau_j(f_j - Au_j)$$

In the case of a diagoal M-matrix, this is an explicit formulae for the new time step !

Using the right-sided rectangle rule leads to

$$M\{u_{j+1} - u_j\} + \tau_j Au_{j+1} = \tau_j f_{j+1},$$

or

$$(M + \tau_j A)u_{j+1} = Mu_j + \tau_j f_{j+1}.$$

In case of the right-side rule, a linear system must be solve in any case. Thus, this method is called an implicit time integration method. These two special cases are called the explicit Euler method, and the implicit Euler method. A third simple choice is the trapezoidal rule leading to

$$(M + \frac{\tau_j}{2} A)u_{j+1} = Mu_j + \frac{\tau_j}{2}(f_j + f_{j+1} - Au_j)$$

It is also an implcit method. Since the trapezoidal integration rule is more accurate, we expect a more accurate method for approximating the ODE.

All single-step time integration methods can be written in the form

$$u_{j+1} = G_j(u_j, f_j),$$

where G_j is linear in both arguments and shall be continuous with bounds

$$\|G_j(u_j, f_j)\|_M \leq L\, \|u_j\|_M + \tau_j\, l\, \|f_j\|_{M^{-1}},$$

with $L \geq 1$.

Lemma 138. *The time integration method fulfills the stability estimate*

$$\|u_j\|_M \leq L^j \, \|u_0\|_M + lL^j \sum_{i=0}^{j-1} \tau_i \|f_i\|_{M^{-1}} \tag{9.2}$$

The explicit Euler method is written as

$$u_{j+1} = (I - \tau M^{-1} A) u + \tau M^{-1} f_j,$$

and has bounds

$$\begin{aligned} L &= \max\{1, \tau \lambda_{max}(M^{-1}A) - 1\} \simeq \max\{1, \frac{\tau}{h^2}\} \\ l &= 1 \end{aligned}$$

If $\tau > h^2$, the powers L^j become very large. This means that the explicit Euler method becomes instable. Thus, for the explicit Euler method, the time-step τ must not be greater than ch^2.

The implicit Euler method is written as

$$u_{j+1} = (M + \tau A)^{-1} M u_j + \tau (M + \tau A)^{-1} f_j,$$

and has the bounds

$$\begin{aligned} L &= 1 \\ l &= 1 \end{aligned}$$

The method is stable for any time-step τ. Such a method is called A-stable.

Lemma 139. *The time discretization error $e_j := u(t_j) - u_j$ of the implicit Euler method satisfies the difference equation*

$$M\{e_{j+1} - e_j\} + \tau A e_{j+1} = d_j,$$

where the d_j satisfy

$$d_j = \int_{t_j}^{t_{j+1}} \{f(s) - Au(s)\} \, ds - \tau_j \{f(t_{j+1}) - Au(t_{j+1})\}.$$

Lemma 140. *The error of the integration rule can be estimated by*

$$\|d_j\| \preceq \tau \, \|(f - Au)'\|_{L_\infty} = \tau \, \|u''\|_{L_\infty}$$

Convergence of the time-discretization method follows from stability plus approximation:

Theorem 141. *The error of the implicit Euler method satisfies*

$$\|u(t_j) - u_j\|_M \leq \sum_{i=0}^{j} \tau \|d_j\| \preceq \tau \, \|u''\|_{L_\infty(0,T)}$$

The trapezoidal rule is A-stable, too. It is based on a more accurate integration rule, and leads to second order convergence $O(\tau^2)$. Convergence of higher order can be obtained by Runge-Kutta methods.

9.3 Space-time formulation of Parabolic Equations

In the previous section we have discretized in space to obtain an ordinary differential equation, which is solved by some time-stepping method. This approach is known as method of lines. Now we formulate a space-time variational problem. This is discretized in time and space by a (discontinuous) Galerkin method. We obtain time-slabs which are solved one after another. This approach is more flexible, since it allows to use different meshes in space on different time-slabs.

9.3.1 Solvability of the continuous problem

Let $V \subset H$ be Hilbert spaces, typically $H = L_2(\Omega)$ and $V = H^1(\Omega)$. Duality is defined with respect to H. For $t \in (0,T)$ we define the familiy $A(t) : V \to V^*$ of uniformely continuous and elliptic operators:

(a) $\langle A(t)u, u\rangle \geq \alpha_1 \|u\|_V^2$

(b) $\langle A(t)u, v\rangle \leq \alpha_2 \|u\|_V \, \|v\|_V$

We assume that $\langle A(t)u, v\rangle$ is integrable with respect to time. We do not assume that $A(t)$ is symmetric. We consider the parabolic equation: Find $u : [0,T] \to V$ such that

$$\begin{aligned} u' + Au &= f \quad \forall\, t \in (0,T) \\ u(0) &= u_0 \end{aligned}$$

We define $X = \{v \in L_2(V) : v' \in L_2(V^*)\}$ and $Y = L_2(V)$, with its dual $Y^* = L_2(V^*)$. A variational formulation is: Find $u \in X$ such that

$$\int_0^T \langle u' + Au, v\rangle = \int_0^T \langle f, v\rangle \qquad \forall\, v \in Y \tag{9.3}$$

$$(u(0), v_0)_H = (u_0, v_0) \qquad \forall v_0 \in H \tag{9.4}$$

Adding up both equations leads to the variational problem $B(u,v) = f(v)$ with the bilinear-form $B(.,.) : X \times (Y \times H) \to \mathbb{R}$:

$$B(u,(v,v_0)) = \int_0^T \langle u' + Au, v\rangle + (u(0), v_0)_H$$

and the linear-form $f : Y \times H \to \mathbb{R}$:

$$f(v,v_0) = \int_0^T \langle f, v\rangle + (u_0, v_0)_H$$

We assume that $f \in Y^*$ and $u_0 \in H$

Theorem 142 (Lions). *Problem (9.3)-(9.4) is uniquely solvable.*

Proof. We apply the theorem by Babuška-Aziz. We observe that all forms are continuous (trace-theorem). We have to verify both inf-sup conditions.

First, we show

$$\inf_{u\in X} \sup_{(v,v_0)\in Y\times H} \frac{B(u,v)}{\|u\|_X \|(v,v_0)\|_{Y\times H}} \geq \beta > 0 \tag{9.5}$$

We fix some $u \in X$ and set (with A^{-T} the inverse of the adjoint operator)

$$\begin{aligned} v &:= A^{-T}u' + u \\ v_0 &:= u(0) \end{aligned}$$

and obtain

$$\begin{aligned} B(u,v) &= \int \left\langle u' + Au, A^{-T}u' + u \right\rangle\, dt \;+ (u(0), u(0))_H \\ &= \int \left\langle A^{-1}u', u' \right\rangle + \langle Au, u\rangle + \langle u', u\rangle + \langle u, u'\rangle\, dt \;+ \|u_0\|_H^2 \\ &= \int \left\langle A^{-1}u', u' \right\rangle + \langle Au, u\rangle + \frac{d}{dt}\|u\|_H^2 + \|u_0\|_H^2 \\ &\geq \alpha_2^{-1}\|u'\|_{L_2(V^*)}^2 + \alpha_1\|u\|_{L_2(V)}^2 + \|u(T)\|_H^2 \\ &\succeq \|u\|_X^2 \end{aligned}$$

Since $\|(v, v_0)\|_{Y\times H} \preceq \|u\|_X$ the first $\inf-$sup-condition is proven. For the other one, we show

$$\forall\, 0 \neq (v, v_0) \in Y \times H \;\; \exists\, u \in X \; : \; B(u, v) > 0 \tag{9.6}$$

We fix some v, v_0. We define u by solving the parabolic equation

$$u' + \gamma L u = A^T v, \qquad u(0) = v_0,$$

where L is a symmetric, constant-in-time, continuous and elliptic operator on V. The parameter $\gamma > 0, \gamma = O(1)$ will be fixed later. The equation has a unique solution, which can be constructed by spectral theory. If $(v, v_0) \neq 0$, then also $u \neq 0$.

$$\begin{aligned} B(u,v) &= \int \left\langle u' + Au, A^{-T}(u' + \gamma L u)\right\rangle + \|v_0\|_H^2 \\ &= \int \left\langle u', A^{-T}u'\right\rangle + \langle u, u'\rangle + \left\langle u', A^{-T}\gamma L u\right\rangle + \gamma\, \langle u, Lu\rangle\, dt + \|v_0\|_H^2 \\ &\geq \int \tfrac{1}{\alpha_2}\|u'\|_{V^*}^2 + \tfrac{1}{2}\tfrac{d}{dt}\|u\|_H^2 - \|u'\|_{V^*}\|A^{-T}\gamma L u\|_V + \gamma\, \langle u, Lu\rangle + \|v_0\|_H^2 \end{aligned}$$

The second term is integrated in time, and we apply Young's inequality for the negative term:

$$\begin{aligned} B(u,v) &\geq \int \tfrac{1}{\alpha_2}\|u'\|_{V^*}^2 - \tfrac{1}{2\alpha_2}\|u'\|_{V^*}^2 - \tfrac{\alpha_2}{2}\|A^{-T}\gamma L u\|_V^2 + \gamma\, \langle u, Lu\rangle + \tfrac{1}{2}\|v_0\|_H^2 + \tfrac{1}{2}\|v(T)\|_H^2 \\ &\geq \int \tfrac{1}{2\alpha_2}\|u'\|_{V^*}^2 - \tfrac{\alpha_2\gamma^2}{2}\|A^{-T}L\|_{V\to V}^2\|u\|_V^2 + \gamma\, \langle u, Lu\rangle + \tfrac{1}{2}\|u(0)\|_H^2 \end{aligned}$$

We fix now γ sufficiently small such that $\frac{\alpha_2\gamma^2}{2}\|A^{-T}\gamma L\|^2_{V\to V}\|u\|^2_V \le \gamma\,\langle u, Lu\rangle$ to obtain

$$B(u,v) \succeq \|u'\|^2_{L_2(V^*)} + \|u_0\|^2_H > 0.$$

$\square$

A similar proof of Lions's theorem is found in Ern + Guermond.

9.3.2 A first time-discretization method

We discretize in time, but keep the spatial function space infinit dimensional. A first reasonable attempt is to use $X_h = P^1(V)$, and $Y_h = P^{0,disc}(V)$. Evaluation of $B(.,.)$ leads to

$$\begin{aligned} B(u_h, v_h) &= \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} \langle u_h' + Au_h, v_h\rangle + (u_h(0), v_h(0)) \\ &= \sum_{j=1}^{n} \langle u_j - u_{j-1}, v_j\rangle + \tfrac{\tau_j}{2}\,\langle A(u_{j-1}+u_j), v_j\rangle + (u_0, v_0) \end{aligned}$$

Here, the time derivative evaluates to finite differences of point values in t_j. Since $u_j \in V$, the duality pairs coincide with inner products in H. Thus, for every time-step we get the equation

$$u_j - u_{j-1} + \frac{\tau}{2}A(u_j + u_{j-1}) = \tau f_j$$

This is the trapezoidal method (Crank-Nicolson). From numerics for odes we remember it is A-stable, but not L-stable. We cannot prove a discrete $\inf-\sup$ condition.

9.3.3 Discontinuous Galerkin method

We give an alternative, formally equivalent variational formulation for the parabolic equation by integration by parts in time

$$\int -\langle u, v'\rangle + \langle Au, v\rangle + (u(T), v(T))_H - (u(0), v(0))_H = \int \langle f, v\rangle$$

Now, we plug in the given initial condition $u(0) = u_0$:

$$\int -\langle u, v'\rangle + \langle Au, v\rangle + (u(T), v(T))_H = \int \langle f, v\rangle + (u_0, v(0))_H$$

The higher H^1-regularity is now put onto the test-space, which validates point-evaluation at $t = 0$ and $t = T$. The trial-space is now only L_2, which gives no meaning for $u(T)$. There are two possible remedies, either to introduce a new variable for $u(T)$, or, to restrict the test space:

1. Find $u \in L_2(V)$, $u_T \in H$ such that

$$\int -\langle u, v'\rangle + \langle Au, v\rangle + (u_T, v(T))_H = \int \langle f, v\rangle + (u_0, v(0))_H \quad \forall\, v \in L_2(V), v' \in L_2(V^*) \tag{9.7}$$

2. Find $u \in L_2(V)$ such that

$$\int -\langle u, v'\rangle + \langle Au, v\rangle = \int \langle f, v\rangle + (u_0, v(0))_H \quad \forall\, v \in L_2(V), v' \in L_2(V^*), v(T) = 0 \tag{9.8}$$

Both problems are well posed (continuity and $\inf - \sup$ conditions, exercise). Now, the initial condition was converted from an essential to a natural boundary condition.

Next, we integrate back, but we do not substitute the initial condition back:

$$\int \langle u' + Au, v\rangle + (u(0), v(0))_H = \int \langle f, v\rangle + (u_0, v(0))$$

The initial condition is again a part of the variational formulation. Note that this formulation is fulfilled for $u \in H^1$, and smooth enough test functions providing the trace $v(0)$.

This technique to formulate initial conditions is used in the Discontinuous Galerkin (DG) method. For every time-slab (t_{j-1}, t_j) we define a parabolic equation, where the initial value is the end value of the previous time-slab.

Here, we first define a mesh $\mathcal{T} = \{t_0, t_1, \dots t_n\}$, and then the mesh-dependent formulation:

$$\int_{t_{j-1}}^{t_j} \langle u' + Au, v\rangle + (u(t_{j-1}^+), v(t_{j-1}^+))_H = \int_{t_{j-1}}^{t_j} \langle f, v\rangle + (u(t_{j-1}^-), v(t_{j-1}^+))_H \qquad \forall\, j \in \{1, \dots n\}$$

(with the notation $u(t_0^-) := u_0$). By using left and right sided limits, we get the u from the current time-slab, and the end-value from the previous time-slab, respectively. The variational formulation is valid for the solution $u \in H^1$, and piece-wise regular test-functions on the time-intervals.

The bilinear-form is defined as

$$B(u, v) = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \langle u' + Au, v\rangle + ([u]_{t_{j-1}}, v(t_{j-1}^+))_H$$

where the jump is defined as $[u]_{t_j} = u(t_j^+) - u(t_j^-)$, and the special case $[u]_{t_0} = u(t_0^+)$. The solution satisfies

$$B(u, v) = \int \langle f, v\rangle + (u_0, v(0)) \qquad \forall\, \text{p.w. smooth}\, v$$

The bilinear-form is defined for discontinuous trial and discontinuous test functions. It allows to define

$$X_h = Y_h = P^{k,dc}(V)$$

Let us elaborate the case of piece-wise constants in time:

$$\tau \langle Au_j, v_j\rangle + (u_j - u_{j-1}, v_j) = \tau \langle f_j, v_j\rangle$$

which leads to the implicit Euler method

$$\frac{u_j - u_{j-1}}{\tau} + Au_j = f_j$$

The imlicit Euler method is A and L-stable.

We define the mesh-dependent norms

$$\begin{aligned}
\|u\|_{X_h}^2 &= \sum_j \|u\|^2_{L_2(t_{j-1},t_j,V)} + \|u'\|^2_{L_2(t_{j-1},t_j,V^*)} + \tfrac{1}{t_j-t_{j-1}} \|[u]_{t_{j-1}}\|^2_{V^*} \\
\|v\|_{Y_h}^2 &= \sum_j \|v\|^2_{L_2(t_{j-1},t_j,V)}
\end{aligned}$$

Since $v|_{[t_{j-1},t_j]}$ is a polynomial, we can bound

$$\|v(t_{j-1}^+)\|_V^2 \leq \frac{c}{t_j - t_{j-1}} \|v\|^2_{L_2(t_{j-1},t_j,V)},$$

where the constant c deteriors with the polynomial degree. Thus, the bilinear-form is well defined and continuous on $X_h \times Y_h$.

Theorem 143. *The discrete problem is* $\inf - \sup$ *stable on* $X_h \times Y_h$

Proof. We mimic the first $\inf - \sup$ condition in Theorem 142, where we have set $v = u + A^{-T}u'$. We give the proof for the lowest order case (k=0)

$$v_h := u_h + \tfrac{\gamma}{t_j - t_{j-1}} A(t_{j-1})^{-1}[u_h]_{j-1},$$

with $\gamma = O(1)$ to be fixed later. Thanks to the discontinuous test-space, this is a valid test-function.

In the following we skip the subsripts h, and set $\tau = t_j - t_{j-1}$. There holds

$$\|v\|_{Y_h} \preceq \|u\|_{X_h}.$$

$$\begin{aligned}
B(u_h, v_h) &= \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \left\langle Au, u + \tfrac{\gamma}{\tau} A_{j-1}^{-T}[u]_{j-1} \right\rangle + \sum_j ([u]_{j-1}, u + \tfrac{\gamma}{\tau} A_{j-1}^{-1}[u]_{j-1})_H \\
&= \int \langle Au, u\rangle + \sum_j \int \left\langle Au, \tfrac{\gamma}{\tau} A^{-T}[u_j] \right\rangle + \sum_j (u_j - u_{j-1}, u_j)_H + \sum_j \tfrac{\gamma}{\tau} \|[u]_{t_{j-1}}\|_{A^{-1}}
\end{aligned}$$

The second term is split by Young's inequality:

$$\int_{t_{j-1}} t_j \left\langle Au, \tfrac{\gamma}{\tau} A^1_{j-1}[u]_{j-1} \right\rangle \leq \int \tfrac{1}{2} \langle Au, u\rangle + \tfrac{\gamma^2}{2\tau^2} \int \|A^{-1}[u]\|_A$$

Thus, for sufficiently small γ it can be absorbed into the first and last term.

We reorder the summation of the third term:

$$\begin{aligned} & (u_1, u_1)^2 - (u_1, u_2) + (u_2, u_2)^2 - (u_2, u_3) + \dots \\ = \; & \tfrac{1}{2}\|u_1\|_H^2 + \tfrac{1}{2}\|u_1 - u_2\|_H^2 + \dots \end{aligned}$$

Thus, we got (for piecewise constants in time):

$$B(u_h, v_h) \geq \|u\|^2_{L_2,V} + \sum \|[u]_{t_j}\|_H^2 + \frac{1}{\tau}\|[u]\|^2_{V^*} \succeq \|u_h\|^2_{X_h}$$

$\square$

By stability, we get for the discrete error

$$\begin{aligned} \|I_h u - u_h\|_{X_h} & \preceq \sup_{v_h} \frac{B_h(I_h u - u_h, v_h)}{\|v_h\|_{Y_h})} \\ & = \sup_{v_h} \frac{B_h(I_h u - u, v_h)}{\|v_h\|_{Y_h})} \\ & = \sup_{v_h} \frac{\sum_j \int \langle u' + Au - (I_h u)' + A I_h u, v_h\rangle + \sum_j ([u] - [I_h u], v_h)_H}{\|v_h\|_{L_2(V)}} \\ & \preceq \dots \end{aligned}$$

where the convergence rate depends as usual on the regulariy of the exact solution

Chapter 10

Second order hyperbolic equations: wave equations

We consider equations second order in time

$$\ddot{u} + Au = f$$

with initial conditions

$$u(0) = u_0 \qquad \text{and} \quad \dot{u}(0) = v_0,$$

with a symmetric, elliptic operator A.

10.1 Examples

- scalar wave equation (acoustic waves)

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

- electromagnetic wave equation:

$$\begin{aligned} \mu \frac{\partial H}{\partial t} &= -\operatorname{curl} E \\ \varepsilon \frac{\partial E}{\partial t} &= \operatorname{curl} H \end{aligned}$$

with the magnetic field H and the electric field E, and material parameters permeability μ and permittivity ε. By differentiating the first equation in space, and the second one in time, we obtain

$$\varepsilon \frac{\partial^2 E}{\partial t^2} + \operatorname{curl} \frac{1}{\mu} \operatorname{curl} E = 0$$

- elastic waves: We consider the hyperelastic elastic energy

$$J(u) = \int_\Omega W(C(u)) - fu$$

A body is in equilibrium, if $J' = 0$. If not, then $J' \in V^*$ acts as an accelerating force. Newton's law is

$$\rho \ddot{u} = -J'(u),$$

in variational form

$$\int \rho \ddot{u} v + \langle J'(u), v \rangle = 0 \qquad \forall\, v$$

In non-linear elasticity we have $J'(u) = \operatorname{div} P - f$, where P is the first Piola-Kirchhoff stress tensor. In linearized elasticity we obtain

$$\int \rho \ddot{u} v + \int D\varepsilon(u) : \varepsilon(v) = \int fv$$

We observe conservation of energy in the following sense for elasticity, and similar for the other cases. We define the kinetic energy as $\frac{1}{2}\|\dot{u}\|_\rho^2$ and the potential energy as $J(u)$. Then

$$\frac{d}{dt}\left\{\tfrac{1}{2}\|\dot{u}\|_\rho^2 + J(u)\right\} = (\dot{u}, \ddot{u})_\rho + \langle J'(u), \dot{u} \rangle = 0$$

For the linear equation set $J(u) = \frac{1}{2}\langle Au, u \rangle - \langle f, u \rangle$

10.2 Time-stepping methods for wave equations

We consider the method of lines, where we first discretize in space, and then apply some time-stepping method for the ODE. In principal, one can reduce the second order ODE to a first order system, and apply some Runge-Kutta method for it. This will in general require the solution of linear systems of twice the size. In addition, the structure (symmetric and positive definite) may be lost, which makes it difficult to solve.

We consider two approaches specially taylored for wave equations.

(a) for the second order equation

(b) for first order systems

10.2.1 The Newmark time-stepping method

We consider the ordinary differential equation

$$M\ddot{u} + Ku = f$$

We consider single-step methods: From given state $u_n \approx u(t_n)$ and velocity $\dot{u}_n \approx \dot{u}(t_n)$ we compute u_{n+1} and $\dot{u}_{n+1}$. The acceleration $\ddot{u}_n = M^{-1}(f_n - Ku_n)$ follows from the equation.

The Newmark method is based on a Taylor expansion for u and $\dot{u}$, where second order derivatives are approximated from old and new accelerations. The real parameters β and γ will be fixed later, τ is the time-step:

$$
\begin{aligned}
u_{n+1} &= u_n + \tau\dot{u}_n + \tau^2\left[(\tfrac{1}{2}-\beta)\ddot{u}_n + \beta\ddot{u}_{n+1}\right] && (10.1)\\
\dot{u}_{n+1} &= \dot{u}_n + \tau\left[(1-\gamma)\ddot{u}_n + \gamma\ddot{u}_{n+1}\right] && (10.2)
\end{aligned}
$$

Inserting the formula for u_{n+1} into $M\ddot{u} + Ku = f$ at time t_{n+1} we obtain

$$M\ddot{u}_{n+1} + K\left(u_n + \tau\dot{u}_n + \tau^2\left[(\tfrac{1}{2}-\beta)\ddot{u}_n + \beta\ddot{u}_{n+1}\right]\right) = f_{n+1}$$

Now we keep unknows left and put known variables to the right:

$$\left[M + \beta\tau^2 K\right]\ddot{u}_{n+1} = f_{n+1} - K\left(u_n + \tau\dot{u}_n + \tau^2(\tfrac{1}{2}-\beta)\ddot{u}_n\right)$$

The Newmark method requires to solve one linear system with the spd matrix $M + \tau^2\beta K$, for which efficient direct or iterative methods are available. After computing the new acceleration, the new state u_{n+1} and velocity $\dot{u}_{n+1}$ are computed from the explicit formulas (10.1) and (10.2).

The Newmark method satisfies a discrete energy conservation. See [Steen Krenk: "Energy conservation in Newmark based time integration algorithms" in *Compute methods in applied mechanics and engineering*, 2006, pp 6110-6124] for the calculations and various extensions:

$$\left[\tfrac{1}{2}\dot{u}M\dot{u} + \tfrac{1}{2}u^T K_{eq} u\right]_n^{n+1} = -(\gamma - \tfrac{1}{2})(u_{n+1}-u_n)K_{eq}(u_{n+1}-u_n)$$

where

$$K_{eq} = K + (\beta - \tfrac{1}{2}\gamma)\tau^2 K M^{-1} K,$$

and the notation $[E]_a^b := E(b) - E(a)$. Here, the right hand side f is skipped. From this, we get the conservation of a modified energy with the so called *equivalent* stiffness matrix K_{eq}. Depending on the parameter γ we get

- $\gamma = \frac{1}{2}$: conservation
- $\gamma > \frac{1}{2}$: damping
- $\gamma < \frac{1}{2}$: growth of energy (unstable)

If K_{eq} is positive definite, then this conservation proves stability. This is unconditionally true if $\beta \geq \frac{1}{2}\gamma$ (the method is called unconditionally stable). If $\beta < \frac{1}{2}\gamma$, the allowed time step is limited by

$$\tau^2 \leq \lambda_{max}(M^{-1}K)^{-1}\frac{1}{\frac{1}{2}\gamma - \beta}.$$

For second order problems we have $\lambda_{max}(M^{-1}K) \simeq h^{-2}$, and thus $\tau \preceq h$ which is a reasonable choice also for accuracy.

Choices for β and γ of particular interests are:

- $\gamma = \frac{1}{2}, \beta = \frac{1}{4}$: unconditionally stable, conservation of original energy ($K_{eq} = K$)
- $\gamma = \frac{1}{2}, \beta = 0$: conditionally stable. We have to solve

  $$M\ddot{u}_{n+1} = f_{n+1} - K(u_n + \tau_n \dot{u}_n + \tfrac{\tau^2}{2}\ddot{u}_n)$$

  which is explicit iff M is cheaply invertible (mass lumping, DG).

10.2.2 Methods for the first order system

We reduce the wave equation

$$\ddot{u} - \Delta u = f$$

to a first order system of pdes. We introduce $\sigma = \int_0^t \nabla u$. Then

$$\begin{aligned} \dot{\sigma} &= \nabla u \\ \dot{u} - \operatorname{div} \sigma &= \tilde{f} \end{aligned}$$

with the integrated source $\tilde{f} = \int_0^t f$. In the following we skip the source f.

A mixed variational formulation in $H(\text{div}) \times L_2$, for given initial conditions $u(0)$ and $\sigma(0)$, is:

$$\begin{aligned} (\dot{\sigma}, \tau) &= -(u, \operatorname{div} \tau) \qquad \forall\, \tau \\ (\dot{u}, v) &= (v, \operatorname{div} \sigma) \qquad \forall\, v \end{aligned}$$

After space discretization we obtain the ode system

$$\begin{pmatrix} M_\sigma & 0 \\ 0 & M_u \end{pmatrix} \begin{pmatrix} \dot{\sigma} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \sigma \\ u \end{pmatrix}$$

We get a similar struture from the Maxwell system:

$$\begin{aligned} (\mu \dot{H}, \tilde{H}) &= (\operatorname{curl} E, \tilde{H}) \qquad \forall\, \tilde{H} \\ (\varepsilon \dot{E}, \tilde{E}) &= -(\operatorname{curl} \tilde{E}, H) \qquad \forall\, \tilde{E} \end{aligned}$$

Conservation of energy is now seen from

$$\frac{d}{dt}\left[\tfrac{1}{2}\sigma^T M_\sigma \sigma + \tfrac{1}{2} u^T M_u u\right] \;=\; \sigma^T M_\sigma \dot{\sigma} + u^T M_u \dot{u} = -\sigma^T B^T u + u^T B \sigma = 0$$

A basis transformation with $M^{1/2}$ leads to the transformed system (the transformed B is called B again):

$$\begin{pmatrix} \dot{\sigma} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \sigma \\ u \end{pmatrix}$$

The matrix is skew-symmetric, and thus the eigenvalues are imaginary. They are contained in $i\,[-\rho(B), \rho(B)]$, where the spectral radious $\rho(B) \simeq h^{-1}$ for the first order operator. Using

Runge-Kutta methods, we need methods such that $i\,[-\tau\rho(B),\tau\rho(B)]$ is in the stability region. For large systems, explicit methods (M cheaply invertible !) are often preferred. While the stability region for the explicit Euler and improved Euler method do not include an interval on the imaginary axis, the RK4 method does.

Methods taylored for the skew-symmetric (Hamiltonian) structure are symplectic methods: The symplectic Euler method is

$$\begin{aligned} M_\sigma \frac{\sigma_{n+1}-\sigma_n}{\tau} &= -B^T u_n \\ M_u \frac{u_{n+1}-u_n}{\tau} &= B\sigma_{n+1} \end{aligned}$$

For updating the second variable, the new value of the first variable is used. For the analysis, we can reduce the large system to 2×2 systems, where β are singular values of $M_\sigma^{-1/2} B M_u^{-1/2}$:

$$\dot{\sigma} = -\beta u \qquad \dot{u} = \beta\sigma$$

The symplectic Euler method can be written as

$$\begin{pmatrix} \sigma_{n+1} \\ u_{n+1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ \tau\beta & 1 \end{pmatrix} \begin{pmatrix} 1 & -\tau\beta \\ 0 & 1 \end{pmatrix}}_{T = \begin{pmatrix} 1 & -\tau\beta \\ \tau\beta & 1-(\tau\beta)^2 \end{pmatrix}} \begin{pmatrix} \sigma_n \\ u_n \end{pmatrix}$$

The eigenvalues of T satisfy $\lambda_1\lambda_2 = \det(T) = 1$, and iff $\tau\beta \leq \sqrt{2}$ they are conjugate complex, and thus $|\lambda_1| = |\lambda_2| = 1$. Thus, the discrete solution is oscillating without damping or growth.

Again, diagonal mass matrices M_u and M_σ render explicit methods efficient.

Chapter 11

Hyperbolic Conservation Laws

We consider the equation

$$\frac{\partial u}{\partial t} + \operatorname{div} f(u) = 0$$

in space dimension n, with the state $u \in \mathbb{R}^m$, and the flux $f : \mathbb{R}^m \to \mathbb{R}^{m \times n}$. We need initial conditions $u(x, 0) = u_0(x)$, and proper boundary conditions.

Examples:

- Transport equation $m = 1$, $n \in \{1, 2, 3, \ldots\}$.

  $$f(u) = b^T u$$

  with $b \in \mathbb{R}^n$ the given wind.

- Burgers' equation $m = 1$, $n = 1$

  $$f(u) = \tfrac{1}{2} u^2$$

  Burgers' equation is a typical model problem to study effects of non-linear conservation laws.

- Wave equation in $\mathbb{R}^n$: $u = (p, v_1, \ldots v_n)$, here $m = n + 1$:

  $$f(p, u) = \begin{pmatrix} u_1 & \cdots & u_n \\ p & & \\ & \ddots & \\ & & p \end{pmatrix}$$

- Euler equations (model for compressible flows) in $\mathbb{R}^n$: $m = n + 2$, state $u = (\rho, m_1, \ldots, m_n, E)$, with density ρ, momentum $m = \rho v$, and energy. The flux function is

  $$f = \begin{pmatrix} \rho v \\ \rho v \otimes v + pI \\ (E + p) v \end{pmatrix}$$

  with the internal energy $e = E/\rho - \frac{1}{2}|v|^2$ (proportional to temperature), and state equation $p = p(\rho, e)$. Equations are conservation of mass, momentum and energy.

11.1 A little theory

Set $n = 1$, $m = 1$. We assume that f is convex, i.e. f' is strictly monotone increasing. For linear fluxes $f = bu$, the solution is the traveling wave

$$u(x,t) = u_0(x - bt).$$

It is constant along the characteristic lines $x(t) = x_0 + bt$.

For smooth fluxes f, the solution is constant along characteristic lines $x(t) = x_0 + f'(u_0(x_0))t$:

$$u(x(t),t) = u_0(x_0)$$

proof:

$$0 = \frac{d}{dt}u(x(t),t) = \frac{\partial u}{\partial t} + \frac{dx}{dt}\frac{\partial u}{\partial x} = \frac{\partial u}{\partial t} + f'(u)\frac{\partial u}{\partial x} = f_t + (f(u))_x$$

The smooth solution exists as long as characteristic lines don't intersect.

Example: Burgers equation. The velocity of the characteristic is $f'(u) = (\frac{1}{2}u^2)' = u$, i.e. the solution itself.

11.1.1 Weak solutions and the Rankine-Hugoniot relation

If characteristic lines intersect, the solution forms a shock. the Rankine-Hugoniot relation is a equation for the speed of the shock.

We assume the solution is piecewise smooth. To have meaningful discontinuous solutions, we have to consider weak solutions in space-time:

$$\int_{\Omega\times(0,T)} u\varphi_t + f(u)\nabla\varphi = 0 \qquad \forall\varphi\in C_0^\infty$$

(initial condition is skipped here, easily covered by non-vanishing test-functions for $t = 0$).

The weak form states that for $F = (f,u)$ there holds

$$\operatorname{div}_{x,t} F = 0$$

in weak senses. Thus, $F \in H(\operatorname{div})$. This requires that $F \cdot n$ is continuous across discontinuities. Let $s(t)$ the position of the shock. The normal vector satisfies

$$n \sim (1, -s')$$

Thus, $[F \cdot n] = 0$ reads as

$$f(u_l) - u_l s' = f(u_r) - u_r s'$$

and we get the Rankine-Hugoniot relation

$$s' = \frac{f(u_l) - f(u_r)}{u_l - u_r}$$

Example Burgers: The speed of the shocks is

$$s' = \frac{\frac{1}{2}u_l^2 - \frac{1}{2}u_r^2}{u_l - u_r} = \frac{u_l + u_r}{2}$$

11.1.2 Expansion fans

Assume $u(x+) > u(x-)$, then, due to convexity of the flux there is also $f'(u(x+)) > f'(u(x-))$. The speed on the right is higher than on the left. Here, all monotone increasing functions (between $u(x-)$ and $u(x+)$), constant along lines $x + f'(u)t$ are weak solutions.

Another conditions is necessary to pick the meaningful physical solution. Two choices are

Viscosity solutions. Consider the regularized equation

$$u_t^\varepsilon + f(u^\varepsilon)_x - \varepsilon u_{xx}^\varepsilon = 0$$

The limit (if existent) $\lim_{\varepsilon\to 0} u^\varepsilon$ is called viscosity solution.

Entropy solutions. We define some quantity $E(u)$ called entropy, where, for physical reasons, the total amount should not increase:

$$\frac{d}{dt}\int_\Omega E(u) \le 0$$

To localize it, we define the entropy flux F such that

$$F' = E'f'$$

If the solution is smooth, then

$$E(u)_t + F(u)_x = E'u_t + F'u_x = E'(u_t + f'u_x) = 0$$

Thus

$$\frac{d}{dt}\int_\Omega E(u) = \int_\Omega E(u)_t = -\int_\Omega F(u)_x = -\int_{\partial\Omega} F(u)\cdot n$$

No entropy changes for smooth solutions with isolated boundary. But, this is not true for discontinuous solutions.

We pose the entropy decrease $E(u)_t + F(u)_x \le 0$ in weak sense:

$$-\int_{\Omega\times(0,T)} E(u)\varphi_t + F(u)\nabla\varphi \le 0 \qquad \forall\,\varphi \in C_0^\infty, \varphi \ge 0$$

Similar to the Rankine-Hugoniot relation we integrate back on smooth regions in space-time

$$\sum_{(\Omega\times(0,T))_i} \int \big(\underbrace{E(u)_t + \operatorname{div} F(u)}_{=0}\big)\varphi - \int_\gamma \big([E(u)]s' - [F(u)]\big)\varphi \le 0 \qquad \forall\,\varphi \ge 0$$

Example Burgers: Choose the entropy $E(u) = u^2$. Then

$$F(u) = \int E'f' = \int 2uu = \frac{2}{3}u^3$$

Calculating

$$\begin{aligned}[E]s' - [F] &= (u_r^2 - u_l^2)\frac{u_r + u_l}{2} + \frac{2}{3}(u_r^3 - u_l^3) = \dots \\ &= -(u_r - u_l)\frac{(u_r - u_l)^2}{6}\end{aligned}$$

Now, posing the non-negative condition for $[E]s' - [F]$ we allow jumps only for $u_r < u_l$.

11.2 Numerical Methods

The natural methods for conservation laws are finite volume / discontinuous Galerkin methods:

$$\int_T \frac{\partial u}{\partial t} v - f(u)\nabla v + \int_{\partial T} g(u_l, u_r) v = 0 \qquad \forall\, T\ \forall\, v \in P^k(T)$$

Here, g is the numerical flux on the element boundary, calculated from left- and right sided states. For continuous $u = u_l = u_r$, it satisfies

$$g(u, u) = f(u)n$$

Otherwise, up-wind like fluxes (many different choices !) are used.

Finite volume methods can be designed such that entropy is non-increasing, often the calculations are technical. They are also used to prove the existence of solutions.

Higher order methods do not satisfy the maximum principle, which may lead to problems for non-linear equations (Euler: divide by ρ). Here, *limiters* are used: If the solution produces oscillations, it is smoothed (somehow). E.g., switch back to a finite volume method.

Recent development (by Guermond+Pasquetti+Popov) is the so-called *entropy viscosity method*: If the entropy relation is violated, artificial viscosity is switched on. Ideally, this happens only close to shocks.

Space-time methods include special mesh-generation related to the finite speed of propagation (front tracking methods, tent-pitching methods). [Gopalakrishnan, Schöberl, Wintersteiger 2016, master thesis Wintersteiger].