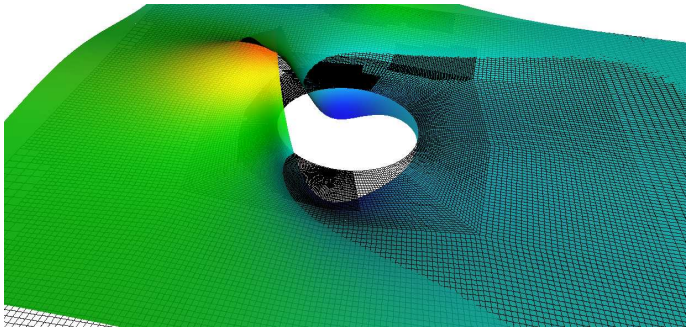
Finite Elements I & parts of II



Malte Braack

Mathematisches Seminar

Christian-Albrechts-Universität zu Kiel

Scriptum, 02.07.2018

Contents

Acknowledgement

Special thanks to Utku Kaya and Bastian Schroeter for their support in careful proof reading.

Chapter 1

One dimensional Boundary Value Problems

We consider the following linear Boundary Value Problem (BVP) of 2. order:

$$\begin{aligned} -(a(x)u'(x))' &= f(x) \qquad \forall x \in (0,1)\,, \\ u(0) &= g_0\,, \\ u(1) &= g_1\,, \end{aligned} \tag{1.1}$$

with given data on the right hand side $f : (0,1) \to \mathbb{R}$, a parameter function $a : (0,1) \to \mathbb{R}^+$, and boundary data $g_0, g_1 \in \mathbb{R}$. These are so-called *Dirichlet conditions*. In the case $g_0 = g_1 = 0$ the Dirichlet condition *homogeneous*. We consider the more general case of *inhomogeneous* Dirichlet conditions.

The system (1.1) is, for instance, a model for the temperature distribution u on a bar of length 1. The temperature g_0, g_1 is prescribed at the two ends of the bar. $a(x)$ describes the heat conduction coefficient which may vary in space. The right hand side f models the influence of an external heat source.

Now, we are going to precise the term 'solution' u. At first, we have to figure out which properties we assume for the data. A reasonable assumption is

$$a \in C^1[0,1]\,, \qquad f \in C[0,1]\,.$$

This means that the coefficient a is continuously differentiable in $(0,1)$ an the derivatives are continuously extenable to the closed intervall $[0,1]$. Although this assumption does not seem very restrictive, e.g. the case $a(x) = \sqrt{x}$ is excluded. The solution u should be a function

$$u \in C^2[0,1].$$

1.1 Finite Difference Discretization

A very straight forward approach to solve the boundary value problem above is the finite difference method. For ease of presentation we assume at this stage that the coefficient is constant, $a \equiv 1$. Hence, we obtain the equation:

$$\begin{aligned} -u''(x) &= f(x) \qquad \forall x \in (0,1)\,, \\ u(0) &= g_0\,, \\ u(1) &= g_1\,, \end{aligned}$$

We partition the intervall $(0,1)$ in n equidistant subintervalls of length $h = 1/n$. The second derivative $u''(x)$ can be approximated on interior nodes $x_i := hi$ by the central difference quotient:

$$u''(x_i) \approx \frac{1}{h^2}(u(x_{i+1}) - 2u(x_i) + u(x_{i-1}))\,, \quad i = 1, \ldots, n-1\,.$$

On the boundary nodes x_0 und x_n we do not need any difference quotient, because the solution is prescribed due to the Dirichlet conditions.

We obtain a linear equation system of the following form:

$$\frac{1}{h^2}\begin{pmatrix} h^2 & & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & & h^2 \end{pmatrix}\begin{pmatrix} U_0 \\ U_1 \\ \vdots \\ \vdots \\ U_N \end{pmatrix} = \begin{pmatrix} g_0 \\ f(x_1) \\ \vdots \\ \vdots \\ f(x_{n-1}) \\ g_1 \end{pmatrix}.$$

Here, U_i denotes the approximation of the solution u on the node x_i,

$$U_i \approx u(x_i)\,.$$

The empty fields in the matrix correspond to zero entries. We obtain a sparse matrix. In the current case of a one dimensional BVP the matrix is even triagonal. The boundary values can be determined directly:

$$U_0 = g_0 \qquad U_N = g_1\,.$$

It remains to solve the linear system

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{pmatrix}\begin{pmatrix} U_1 \\ \vdots \\ \vdots \\ U_{N-1} \end{pmatrix} = \begin{pmatrix} f(x_1) + h^{-2}g_0 \\ f(x_2) \\ \vdots \\ f(x_{n-1}) + h^{-2}g_1 \end{pmatrix}$$

In a subsequent section we will readdress this matrix. Then we will also address the question of solvability.

1.2 Variational Formulation

We will now develop an alternative approach that is much more complicated, but even more elegant. This effort is worthwhile, since the overall concept can be transferred then to more complicated equations and to more spatial dimensions.

Instead of BVP (1.1) we require that

$$-\int_0^1 (a(x)u'(x))'\phi(x)\,dx \;=\; \int_0^1 f(x)\phi(x)\,dx\,,$$

holds for a sufficient amount of so-called test functions ϕ. Moreover, we have the Dirichlet values $u(0) = g_0$ and $u(1) = g_1$. Now we also assume $\phi \in C^1(0,1)$. Then, we obtain through integration by parts

$$\begin{aligned} -\int_0^1 (a(x)u'(x))'\phi(x)\,dx &= -a(x)u'(x)\phi(x)\Big|_{x=0}^{x=1} + \int_0^1 a(x)u'(x)\phi(x)'\,dx \\ &= -a(1)u'(1)\phi(1) + a(0)u'(0)\phi(0) + \int_0^1 a(x)u'(x)\phi(x)'\,dx \end{aligned}$$

Let us now assume that the test function vanishes at the boundary points, so

$$\phi(0) = \phi(1) \;=\; 0\,,$$

then we obtain the equations

$$\begin{aligned} \int_0^1 au'\phi'\,dx &= \int_0^1 f\phi\,dx\,, \\ u(0) &= g_0 \\ u(1) &= g_1\,. \end{aligned} \tag{1.2}$$

for all "valid" test functions ϕ, which vanish at the boundary points.

Note that, there is no need to require $u \in C^2[0,1]$ in order to satisfy (1.2), since the second derivatives of u do not occure any more. Obviously, it is enough to have $u \in C^1[0,1]$ in addition to the assumption for the occurring integral to be well-defined. We will present this in more detail in the following chapter.

Chapter 2

Basic knowledge in Functional Analysis

2.1 Continuous linear operators between normed spaces

Definition 2.1 *A mapping $T : E \to F$ between two normed vector spaces $(E, \|\cdot\|_E), (F, \|\cdot\|_F)$ is called* bounded, *if*

$$\|T\|_{E;F} \quad := \quad \sup_{0 \neq x \in E} \frac{\|Tx\|_F}{\|x\|_E} \quad < \quad \infty \,.$$

Lemma 2.2 *If $(E, \|\cdot\|_E), (F, \|\cdot\|_F)$ are normed spaces and $T : E \to F$ is a linear mapping. Then the following properties are equivalent:*

(a) T is continuous.

(b) T is continuous at point zero.

(c) T is limited.

Proof. (a)⇒(b) is obvious. (b)⇒(c): Because of the continuity of T there exists $\epsilon = 1$ a $\delta > 0$, so that $T(B_{\delta,E}(0)) \subset B_{\epsilon,F}(0)$. With the homogeneity of the norm now follows:

$$\|T\|_{E;F} \quad = \quad \sup_{0 \neq x \in E} \frac{\|Tx\|_F}{\|x\|_E} = \sup_{x \in E, \|x\| = \delta} \frac{\|Tx\|_F}{\delta} \leq \frac{1}{\delta} \,.$$

(c)⇒(a): Let $x_n \to x$ be a convergent sequence in E. Then for the images holds $Tx_n \to Tx$, because:

$$\|Tx_n - Tx\|_F \quad = \quad \|T(x_n - x)\|_F \quad \leq \quad \|T\|_{E;F} \|x_n - x\| \quad \to \quad 0 \,.$$

□

Definition 2.3 *To two normed spaces $(E, \|\cdot\|_E), (F, \|\cdot\|_F)$ denotes $\mathcal{L}(E,F)$ the space of linear and continuous mappings. In case of $E = F$ we may also write briefly $\mathcal{L}(E)$.*

Lemma 2.4 *In the space $\mathcal{L}(E,F)$ $\|\cdot\|_{E;F}$ is a norm, so that $(\mathcal{L}(E,F), \|\cdot\|_{E;F})$ is also a normed vector space.*

Proof. From the linear algebra we know, that $\mathcal{L}(E,F)$ is also a $\mathbb{K}$-vector space. The standard properties of $\|\cdot\|_{E;F}$ can be easily verified. ∎

2.2 The dual space

Definition 2.5 *If $(E, \|\cdot\|)$ is a normed space, then the normed space $\mathcal{L}(E,\mathbb{K})$ is shortly denoted by E' and is called* dual space *to E. The norm $\|\cdot\|_{E';\mathbb{K}}$ on E' is briefly denoted by $\|\cdot\|_{E'}$. The elements of E' are called continuous linear functionals. For the elements $f \in E'$ and $x \in E$ we usually write*

$$\langle f, x\rangle \quad := \quad f(x)\,.$$

The dual norm referring to $\|\cdot\|$ is thus

$$\|f\|_{E'} \quad := \quad \sup_{x\in E, \|x\|\leq 1} |\langle f, x\rangle|\,.$$

In the case of $\mathbb{K} = \mathbb{R}$ obviously holds

$$\|f\|_{E'} \quad := \quad \sup_{x\in E, \|x\|\leq 1} \langle f, x\rangle\,.$$

Corollary 2.6 *Let G be a subspace of a normed space E and $g \in G'$. Then an extension $f \in E'$ of g exists with $\|f\|_{E'} = \|g\|_{G'}$.*

Proof. The proof yields from the extension theorem of Hahn-Banach. We refer to the relevant literature of functional analysis. ∎

Corollary 2.7 *For each $x_0 \in E$ there exists a $f_0 \in E'$ with $\|f_0\|_{E'} = \|x_0\|$ and $\langle f_0, x_0\rangle = \|x_0\|^2$.*

Proof. We apply Corollary 2.6 for the subspace $G := \mathbb{K}x_0$ and $\langle g, tx_0\rangle := t\|x_0\|^2$. This yields the existence of $f_0 \in E'$ with $\langle f_0, x_0\rangle = \langle g, x_0\rangle = \|x_0\|^2$ and

$$\|f_0\|_{E'} \;=\; \|g\|_{G'} \;=\; \sup_{t\in\mathbb{K}, \|tx_0\|\leq 1} |\langle g, tx_0\rangle| \;=\; \sup_{t\in\mathbb{K}, \|tx_0\|\leq 1} |t|\|x_0\|^2 \;=\; \|x_0\|\,.$$

∎

In general this f_0 is not unique.

Corollary 2.8 *In normed spaces E and each $x \in E$ its norm can be obtained by*

$$\|x\| \;=\; \sup_{f\in E',\|f\|_{E'}\le 1} |\langle f,x\rangle| \;=\; \max_{f\in E',\|f\|_{E'}\le 1} |\langle f,x\rangle| \,.$$

Proof. The case $x = 0$ is trivial. For $x \neq 0$ it holds on the one hand

$$\sup_{f\in E',\|f\|_{E'}\le 1} |\langle f,x\rangle| \;\le\; \|x\| \,.$$

On the other hand, due to Corollary 2.7 it exists $f_0 \in E'$ with $\langle f_0, x\rangle = \|x\|^2$ and $\|f_0\|_{E'} = \|x\|$. It follows that $f_1 := \|x\|^{-1} f_0 \in E'$ with $\|f_1\|_{E'} = 1$ and $\langle f_1, x\rangle = \|x\|$. We obtain as a result:

$$\|x\| = \langle f_1, x\rangle \le \sup_{f\in E',\|f\|_{E'}\le 1} \langle f,x\rangle \;\le\; \|x\| \,.$$

This implies the assertion. ∎

Corollary 2.9 *Let $G \subset E$ be a subspace of a normed space E. Furthermore we assume the following implication:*

$$f \in E' \,,\; f|_G = 0 \;\Longrightarrow\; f|_E = 0 \,.$$

Then G is dense in E.

Proof. We refer to standard literature in Functional Analysis (Separation theorem of Hahn-Banach). ∎

2.3 Lebesgue-integrable functions $L^p(\Omega)$

When we speak in the following about *measurable sets* $A \subset \mathbb{K}^n$, we always refer to *Lebesgue-measurable*, i.e. for all $E \subset \mathbb{K}^n$ it holds:

$$\lambda(E) \;=\; \lambda(A \cap E) + \lambda((\mathbb{K}^n \setminus A) \cap E) \,.$$

Here $\lambda : \mathcal{P}(\mathbb{K}^n) \to [0,\infty]$ denotes the exterior Lebesgue-measure. The measurable sets form a $\sigma-$algebra with measure $\lambda(A)$. A set A is a null-set, if its measure vanishes, $\lambda(A) = 0$. Accordingly, a function $u : \Omega \to \overline{\mathbb{R}}$ is called *measurable* (or *Lebesgue-measurable*, if the niveau sets $N_{>\alpha} := \{x \in \Omega : u(x) > \alpha\}$ are measurable for all $\alpha \in \mathbb{R}$. Here, $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$.

Definition 2.10 *Let $\Omega \subset \mathbb{R}^n$ be a measurable set. A function $u : \Omega \to \overline{\mathbb{R}}$ is called* Lebesgue-integrable *over Ω, if*

$$\int_\Omega u(x)\,dx \;<\; \infty \,.$$

For Lebesgue-integrable functions we allow the values out of the set $\overline{\mathbb{R}}$. With the expression "almost everywhere" (a.e.) is meant that a certain property holds pointwise everywhere with the exception of points out of a (Lebesgue-) null set $N \subset \Omega$:

$$u = v \text{ a.e. on } \Omega \quad :\Longleftrightarrow \quad u(x) = v(x) \quad \forall x \in \Omega \setminus N\,.$$

Furthermore, in $L^p(\Omega)$-spaces we consider equivalence classes of functions. Here, two functions $u, v : \Omega \to \overline{\mathbb{R}}$ are equivalent, iff

$$u \;=\; v \text{ a.e. on } \Omega\,.$$

Definition 2.11 *The sets $L^1(\Omega)$ is the set of all equivalence classes (as described above) of Lebesgue-integrable functions. For $1 \le p < \infty$ the space $L^p(\Omega)$ is defined by*

$$L^p(\Omega) \;:=\; \{u : \Omega \to \overline{\mathbb{R}} \text{ measurable } :\; |u|^p \in L^1(\Omega)\}\,,$$

with the norm

$$\|u\|_{L^p(\Omega)} \;:=\; \left(\int_\Omega |u(x)|^p dx\right)^{1/p}\,.$$

For $p = \infty$, $L^\infty(\Omega)$ contains all measurable functions $u : \Omega \to \overline{\mathbb{R}}$ with the property that a the null set $N \subset \Omega$ exists, such that $\sup_{x\in\Omega\setminus N} |u(x)| < \infty$. Its norm is then given by

$$\|u\|_{L^\infty(\Omega)} \;:=\; \inf_{\mu(N)=0} \sup_{x\in\Omega\setminus N} |u(x)| < \infty.$$

The following theorems show that these are indeed normed spaces. In particular, we have to verify the triangle inequality.

Lemma 2.12 (Young's inequality) *For $a, b, p, q \in \mathbb{R}$, $a, b \ge 0$, $p, q > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ it holds:*

$$ab \;\le\; \frac{1}{p}a^p + \frac{1}{q}b^q\,.$$

Proof. Exercise. ∎

Young's inequality is often used in the following form:

Corollary 2.13 *Let $a, b \in \mathbb{R}$ and $\epsilon > 0$. Then it holds*

$$|ab| \;\le\; \frac{\epsilon}{2}a^2 + \frac{1}{2\epsilon}b^2.$$

Proof. We just use the setting $p = q = 2$ and the fact that $|ab| = |\epsilon^{1/2}a||\epsilon^{-1/2}b|$. ∎

Theorem 2.14 (Hölder inequality) *Let $\Omega \subset \mathbb{R}^n$ be a measurable set, $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then it holds for all $u \in L^p(\Omega)$, $v \in L^q(\Omega)$:*

$$\int_\Omega |u(x)v(x)|\, dx \quad \leq \quad \|u\|_{L^p(\Omega)} \|v\|_{L^q(\Omega)}\,.$$

In particular $uv \in L^1(\Omega)$.

Proof. For $p = 1$ and $p = \infty$ the assertion is trivial. We consider for $1 < p < \infty$ the functions:

$$a(x) \quad := \quad \frac{|u(x)|}{\|u\|_{L^p(\Omega)}}\,, \qquad b(x) \;:=\; \frac{|v(x)|}{\|v\|_{L^q(\Omega)}}\,.$$

Youngs inequality yields for a.e. $x \in \Omega$:

$$a(x)b(x) \quad \leq \quad \frac{1}{p}a(x)^p + \frac{1}{q}b(x)^q\,.$$

Integration over Ω yields:

$$\int_\Omega a(x)b(x)\, dx \quad \leq \quad \frac{1}{p}\|a\|^p_{L^p(\Omega)} + \frac{1}{q}\|b\|^q_{L^q(\Omega)} \;=\; \frac{1}{p} + \frac{1}{q} \;=\; 1\,.$$

This implies the assertion. ∎

For $p > 1$, $u \in L^p(\Omega)$ and $1/p + 1/q = 1$ we obtain

$$\|u\|_{L^1(\Omega)} \quad = \quad \|1u\|_{L^1(\Omega)} \;\leq\; \|1\|_{L^q(\Omega)}\, \|u\|_{L^p(\Omega)}$$

Taking into account that $\|1\|_{L^q(\Omega)} = \mu(\Omega)^{1/q} = \mu(\Omega)^{1-1/p}$ we obtain

$$\|u\|_{L^1(\Omega)} \quad \leq \quad \mu(\Omega)^{1-1/p}\|u\|_{L^p(\Omega)}. \tag{2.1}$$

Theorem 2.15 (Inequality of Minkowski) *Let $\Omega \subset \mathbb{R}^n$ be a measurable set in $\mathbb{R}^n$, $1 \leq p \leq \infty$ and $u, v \in L^p(\Omega)$. Then the triangle inequality holds*

$$\|u + v\|_{L^p(\Omega)} \quad \leq \quad \|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}.$$

In particular, $u + v \in L^p(\Omega)$.

Proof. For $p = 1$ and $p = \infty$ the assertion is a direct consequence of the point-wise triangle inequality. For $1 < p < \infty$ one can easily derive the following estimate by help of the Jensen inequality for convex functions:

$$|u(x) + v(x)|^p \;\leq\; (|u(x)| + |v(x)|)^p \quad \leq \quad 2^{p-1}(|u(x)|^p + |v(x)|^p).$$

Hence, we have $w := u + v \in L^p(\Omega)$. To bound its L^p-norm from above we deduce as follows:

$$|w(x)|^p \quad \leq \quad |u(x)| \cdot |w(x)|^{p-1} + |v(x)| \cdot |w(x)|^{p-1}.$$

Now, we use this to integrate both parts over Ω:

$$\|w\|^p_{L^p(\Omega)} \quad \leq \quad \int_\Omega |u(x)| \cdot |w(x)|^{p-1} dx + \int_\Omega |v(x)| \cdot |w(x)|^{p-1} dx \,.$$

However, we have still to ensure that the two terms on the right hand side are finite. This is true due to the previously mentioned inequality of Hölder, because:

$$\begin{aligned} u, v \in L^p(\Omega) \quad &\Longrightarrow \quad |u|, |v| \in L^p(\Omega) \\ |w| \in L^p(\Omega) \quad &\Longrightarrow \quad |w|^{p-1} \in L^{p/(p-1)}(\Omega) = L^q(\Omega), \quad 1 = \frac{1}{p} + \frac{1}{q} \,. \end{aligned}$$

Application of Hölder's inequality two times yields

$$\|w\|^p_{L^p(\Omega)} \quad \leq \quad \|u\|_{L^p(\Omega)} \||w|^{p-1}\|_{L^q(\Omega)} + \|v\|_{L^p(\Omega)} \||w|^{p-1}\|_{L^q(\Omega)}.$$

Since further $\||w|^{p-1}\|_{L^q(\Omega)} = \|w\|^{p-1}_{L^{(p-1)q}(\Omega)}$ and $(p-1)q = p$ we obtain

$$\|u+v\|^p_{L^p(\Omega)} \quad \leq \quad (\|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}) \|u+v\|^{p-1}_{L^p(\Omega)}.$$

In the case of $\|u+v\|_{L^p(\Omega)} \neq 0$, we obtain the assertion by simply dividing by $\|u+v\|^{p-1}_{L^p(\Omega)}$. In the other case, the assertion is trivial. ∎

Theorem 2.16 *For $1 \leq p \leq \infty$ and measurable set $\Omega \subset \mathbb{R}^n$ the space $(L^p(\Omega), \|\cdot\|_{L^p(\Omega)})$ is complete, i.e. a Banach spaces.*

Proof. The inequality of Minkowski ensures that $L^p(\Omega)$ is closed with respect to summation. Hence, it is a linear space. Moreover, the triangle inequality implies that $\|\cdot\|_{L^p(\Omega)}$ is a semi-norm. For its definiteness we have to verify the implication:

$$\|u\|_{L^p(\Omega)} = 0 \quad \Longrightarrow \quad u = 0 \quad \text{a.e.}$$

For $1 \leq p < \infty$, $u \in L^p(\Omega)$ is equivalent to $|u|^p \in L^1(\Omega)$. Therefore, $\|u\|_{L^p(\Omega)} = 0$ implies

$$0 = \|u\|^{1/p}_{L^p(\Omega)} = \||u|^p\|_{L^1(\Omega)} \,.$$

Because $\|\cdot\|_{L^1(\Omega)}$ is a norm in $L^1(\Omega)$, it follows $|u|^p = 0$ a.e. This yields the assertion. For $p = \infty$ we obtain as well $u = 0$ a.e.. Moreover, we have to verify the completeness. Let $(u_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $L^p(\Omega)$. We investigate the cases $p = \infty$ and $1 \leq p < \infty$ separately.
$p = \infty$: We know that a null set $N \subset \Omega$ exists, such that $|u_n(x)| \leq \|u_n\|_{L^\infty(\Omega)} \leq C$ for all $x \in \Omega \setminus N$ and all $n \in \mathbb{N}$. We define

$$u(x) \quad := \quad \begin{cases} 0 & \text{if } x \in N \,, \\ \lim_{n\to\infty} u_n(x) & \text{if } x \in \Omega \setminus N \,. \end{cases}$$

This u is well defined, because for $x \notin N$ the sequence $(u_n(x))_{n\in\mathbb{N}}$ is a Cauchy sequence and hence convergent. As a consequence, u is a.e. a limes of measurable functions and therefore it is itself a measurable and bounded function in whole Ω. In other words, $u \in L^\infty(\Omega)$. Moreover, we have for all $x \in \Omega \setminus N$:

$$|u(x) - u_n(x)| \;=\; \lim_{m\to\infty} |u_m(x) - u_n(x)| \;\le\; \lim_{m\to\infty} \|u_m - u_n\|_{L^\infty(\Omega)} \to\; 0\,.$$

This implies $\lim_{n\to\infty} \|u - u_n\|_{L^\infty(\Omega)} = 0$.
$1 \le p < \infty$: In this case we seek a convergent subsequence. Since a Cauchy sequence has at most one cluster point, the convergence then follows immediately. The subsequence we chose in such a way that

$$\sum_{k\in\mathbb{N}} \|u_{n_{k+1}} - u_{n_k}\|_{L^p(\Omega)} \;<\; \infty. \tag{2.2}$$

Then, we define

$$u \;:=\; \sum_{k\in\mathbb{N}} (u_{n_{k+1}} - u_{n_k})\,.$$

We have to show:

(a) It exists a subsequence which fulfills (2.2),

(b) u converges a.e. and $u \in L^p(\Omega)$,

(c) $\lim_{n\to\infty} \|u - u_{n_k}\| = 0$.

To (a): This property is obtained e.g. by choosing for $k \in \mathbb{N}$ the number $n_k \in \mathbb{N}$ such that for all $n, m \ge n_k$ holds: $\|u_n - u_m\|_{L^p(\Omega)} \le 2^{-k}$. We can assume the sequence $(n_k)_{k\in\mathbb{N}}$ to be strictly monotonically increasing.
To (b): We set

$$h_n \;:=\; \sum_{k=1}^{n} |u_{n_{k+1}} - u_{n_k}|.$$

Obviously holds $h_n \in L^p(\Omega)$ and therefore $h_n^p \in L^1(\Omega)$ and $h_n^p \ge 0$. The sequence $(h_n)_{n\in\mathbb{N}}$ is monotonically increasing. The Lemma of Fatou implies that

$$\int_\Omega \liminf_{n\to\infty} h_n^p dx \;\le\; \liminf_{n\to\infty} \int_\Omega h_n(x)^p dx \;=\; \liminf_{n\to\infty} \|h_n\|_{L^p(\Omega)}^p < \infty.$$

The Theorem of Monotone Convergence yields that $h^p := \liminf_{n\to\infty} h_n^p \in L^1(\Omega)$. It follows that $|h^p(x)| < \infty$ and $|h(x)| < \infty$ a.e. in Ω. This implies (b).
To (c): Finally, we verify that $u(x) = \lim_{k\to\infty} u_{n_k}(x)$ a.e. and $\|u - u_{n_k}\|_{L^p(\Omega)} \to 0$. ∎

Lemma 2.17 *Let $\Omega \subset \mathbb{R}^n$ be measurable with finite measure $\mu(\Omega) < \infty$, and $1 \le p < \tilde{p} \le \infty$. Then it holds the inclusion $L^{\tilde{p}}(\Omega) \subset L^p(\Omega)$. In particular, for $u \in L^{\tilde{p}}(\Omega)$:*

$$\|u\|_{L^p(\Omega)} \;\le\; \mu(\Omega)^{\frac{1}{p}-\frac{1}{\tilde{p}}} \|u\|_{L^{\tilde{p}}(\Omega)} .$$

Here, we read the reciprocal of $\tilde{p}$ in the case $\tilde{p} = \infty$ as $\frac{1}{\tilde{p}} = 0$.

Proof. For $\tilde{p} = \infty$ we obtain the pretension immediately:

$$\|u\|^p_{L^p(\Omega)} \;=\; \int_\Omega |u(x)|^p dx \;\le\; \mu(\Omega)\, \|u\|^p_{L^\infty(\Omega)} .$$

For $1 < \tilde{p} < \infty$ we use the Hölder inequality with $q := \tilde{p}/(\tilde{p}-p)$ and use $\frac{1}{q}+\frac{p}{\tilde{p}} = \frac{\tilde{p}-p}{\tilde{p}}+\frac{p}{\tilde{p}} = 1$:

$$\|u\|^p_{L^p(\Omega)} \;=\; \int_\Omega 1 \cdot |u(x)|^p dx \;=\; \|1 \cdot |u|^p\|_{L^1(\Omega)} \le\; \|1\|_{L^q(\Omega)} \|u^p\|_{L^{\tilde{p}/p}(\Omega)}, .$$

This implies due to $\|1\|_{L^q(\Omega)} = \mu(\Omega)^{\frac{1}{q}}$ and $\|u^p\|_{L^{\tilde{p}/p}(\Omega)} = \|u\|^p_{L^{\tilde{p}}(\Omega)}$:

$$\|u\|^p_{L^p(\Omega)} \;\le\; \mu(\Omega)^{\frac{1}{q}} \|u\|^p_{L^{\tilde{p}}(\Omega)} .$$

This yields the pretension by taking on both sides the p-th root and the fact that $1/(qp) = 1/p - 1/\tilde{p}$. ∎

This Lemma implies for bounded measurable sets Ω:

$$L^\infty(\Omega) \;\subset\; \ldots \subset\; L^3(\Omega) \;\subset\; L^2(\Omega) \;\subset\; L^1(\Omega) .$$

Theorem 2.18 (Representation theorem of Riesz) *Let $f \in (L^p(\Omega))'$, $1 < p < \infty$. Then there exists exactly one $u \in L^q(\Omega)$, $1 = \frac{1}{p} + \frac{1}{q}$, such that*

$$\langle f, v\rangle \;=\; \int_\Omega u(x)v(x)\, dx \qquad \forall v \in L^p(\Omega) .$$

Furthermore, $\|u\|_{L^q(\Omega)} = \|f\|_{(L^p(\Omega))'}$.

Proof. We demonstrate that the mapping

$$T : L^q(\Omega) \to L^p(\Omega)',$$
$$\langle Tu, v\rangle = \int_\Omega u(x)v(x)\, dx$$

is an isometric isomorphism. The boundedness of T is obtained by Hölder's inequality:

$$|\langle Tu, v\rangle| \;\le\; \|u\|_{L^q(\Omega)} \|v\|_{L^p(\Omega)} .$$

Hence, $Tu \in L^p(\Omega)'$ and $\|Tu\|_{(L^p(\Omega))'} \le \|u\|_{L^q(\Omega)}$. For the proof of

$$\|Tu\|_{(L^p(\Omega))'} \;\ge\; \|u\|_{L^q(\Omega)}$$

we choose $v^*(x) := |u(x)|^{q-2}u(x)$ if $u(x) \neq 0$, and $v^*(x) = 0$ if $u(x) = 0$, respectively. This implies

$$\langle Tu, v^* \rangle \;=\; \int_\Omega |u(x)|^q \, dx \;=\; \|u\|^q_{L^q(\Omega)}.$$

Hence, due to $q - q/p = 1$ it follows

$$\|Tu\|_{(L^p(\Omega))'} \geq \langle Tu, v^* \rangle \|v^*\|^{-1}_{L^p(\Omega)} \;=\; \|u\|^q_{L^q(\Omega)} \|u\|^{-q/p}_{L^q(\Omega)} \;=\; \|u\|_{L^q(\Omega)}.$$

We still have to show the surjectivity: The set $E := T(L^q(\Omega))$ is a closed subspace. Therefore, it is sufficient to show that E is dense in $L^p(\Omega)'$. This can be shown by Corollary 2.9. Let $h \in (L^p(\Omega)')'$ a linear form which vanishes on E, i.e.

$$\langle h, Tu \rangle \;=\; 0 \qquad \forall u \in L^q(\Omega).$$

Due to the reflexivity of $L^p(\Omega)$ (see Theorem 2.19)) we deduce

$$\int_\Omega h(x)u(x) \, dx \;=\; \langle Tu, h \rangle \;=\; \langle h, Tu \rangle \;=\; 0.$$

We choose in particular $u^*(x) := |h(x)|^{p-2}h(x)$, which leads to

$$0 \;=\; \int_\Omega |h(x)|^{p-2}h(x)^2 \, dx \;=\; \int_\Omega |h(x)|^p \, dx \;=\; \|h\|^p_{L^p(\Omega)}.$$

Hence $h = 0$ in $L^p(\Omega)$, so that T is surjective. However, we still have to verify that $u^* \in L^q(\Omega)$:

$$\|u^*\|^q_{L^q(\Omega)} \;=\; \int_\Omega ||h(x)|^{p-2}h(x)|^q dx \;=\; \int_\Omega |h(x)|^{pq-q} dx$$

Because of $pq - q = p$, this integral is bounded, which yields $u^* \in L^q(\Omega)$. □

Theorem 2.19 *The spaces $L^p(\Omega)$ are reflexive if $1 < p < \infty$, i.e. $L^p(\Omega) = L^p(\Omega)''$.*

Proof. We refer to standard literature on Functional Analysis. □

Definition 2.20 *Let $\Omega \subset \mathbb{R}^n$ be a measurable set and $1 \leq p \leq \infty$. Then we define*

$$L^p_{loc}(\Omega) \;:=\; \{f : \Omega \to \mathbb{K} \,:\, f\chi_K \in L^p(\Omega) \quad \forall K \subset \Omega \text{ compact.}\},$$

where χ_K denotes the characteristic function of the set K.

Lemma 2.21 *Let $\Omega \subset \mathbb{R}^n$ be open and $f \in L^1_{loc}(\Omega)$ is assumed to have the property*

$$\int_\Omega fu \, dx \;=\; 0 \qquad \forall u \in C^\infty_0(\Omega). \tag{2.3}$$

Then $f = 0$ a.e. in Ω.

Proof. The proof can be divided into two steps:
(a) Let us assume firstly that $\lambda(\Omega)$ is bounded and $f \in L^1(\Omega)$. For arbitrary $f_\epsilon \in \mathcal{C}_0(\Omega)$ it holds by the triangle inequality

$$\|f\|_{L^1(\Omega)} \;\leq\; \|f - f_\epsilon\|_{L^1(\Omega)} + \|f_\epsilon\|_{L^1(\Omega)}\,.$$

Due to the density of $\mathcal{C}_0(\Omega)$ in $L^1(\Omega)$ there exists for arbitrary $\epsilon > 0$ a function $f_\epsilon \in \mathcal{C}_0(\Omega)$, such that

$$\|f - f_\epsilon\|_{L^1(\Omega)} \;\leq\; \epsilon.$$

Now, we bound $\|f_\epsilon\|_{L^1(\Omega)}$ properly. For arbitrary compactum $K \subset \Omega$ it holds

$$\|f_\epsilon\|_{L^1(\Omega)} \;=\; \|f_\epsilon\|_{L^1(K)} + \|f_\epsilon\|_{L^1(\Omega\setminus K)}.$$

We choose $K := K_1 \cup K_2$ with compact and disjoint sets $K_1 := \{x \in \Omega :\ f_\epsilon(x) \geq \epsilon\}$ and $K_2 := \{x \in \Omega :\ f_\epsilon(x) \leq -\epsilon\}$. The theorem of Tietze-Uryson ensures the existence of a function $u_0 \in \mathcal{C}_0(\Omega)$, such that $u_0|_{K_1} = 1$, $u_0|_{K_2} = -1$ and $|u_0(x)| \leq 1$ for all $x \in \Omega$. The following identities are valid:

$$\|f_\epsilon\|_{L^1(K)} \;=\; \int_K f_\epsilon u_0\,dx \;=\; \int_\Omega f_\epsilon u_0\,dx - \int_{\Omega\setminus K} f_\epsilon u_0\,dx.$$

Because of $|u_0| \leq 1$, it holds

$$\int_{\Omega\setminus K} f_\epsilon u_0\,dx \;\leq\; \|f_\epsilon\|_{L^1(\Omega\setminus K)} \;\leq\; \epsilon\lambda(\Omega\setminus K) \leq \epsilon\lambda(\Omega).$$

Hence, we have

$$\|f\|_{L^1(\Omega)} \;\leq\; \epsilon + \left|\int_\Omega f_\epsilon u_0\,dx\right| + 2\epsilon\lambda(\Omega).$$

It remains to bound the integral $\int_\Omega f_\epsilon u_0\,dx$. To this end we use the property (2.3) and Hölders inequality:

$$\begin{aligned}\left|\int_\Omega f_\epsilon u_0\,dx\right| &= \left|\int_\Omega f u_0\,dx + \int_\Omega (f_\epsilon - f)u_0\,dx\right| \\ &= \left|\int_\Omega (f_\epsilon - f)u_0\,dx\right| \\ &\leq \|f_\epsilon - f\|_{L^1(\Omega)}\|u_0\|_{L^\infty(\Omega)} \;\leq\; \epsilon.\end{aligned}$$

This yields

$$\|f\|_{L^1(\Omega)} \;\leq\; 2\epsilon(1 + \lambda(\Omega)).$$

Since $\epsilon > 0$ is arbitrary, we follow $\|f\|_{L^1(\Omega)} = 0$, i.e. $f \equiv 0$ a.e. in Ω.
(b) Now, let $\lambda(\Omega)$ be unbounded. We write Ω in the form $\Omega = \cup_{n\in\mathbb{N}}\Omega_n$ with open and bounded sets Ω_n, e.g.

$$\Omega_n \;:=\; \{x \in \Omega \,:\, \mathrm{dist}(x, \mathbb{K}^n \setminus \Omega) > 1/n \text{ and } |x| < n\}\,.$$

Application of case (a) yields $f|_{\Omega_n} = 0$ for all $n \in \mathbb{N}$, and hence $f|_\Omega = 0$. □

Theorem 2.22 *For any open set $\Omega \subset \mathbb{K}^n$ and any $1 \le p < \infty$ is the space of continuous functions with compact support $\mathcal{C}_0(\Omega)$ dense in $L^p(\Omega)$.*

Proof. $p = 1$: Firstly, the space of step functions is dense in $L^1(\Omega)$. Secondly, step functions can be approximated with arbitrary accuracy by continuous functions. Finally, we approximate $\mathcal{C}(\Omega)$ by $\mathcal{C}_0(\Omega)$ with arbitrary accuracy in the L^1 norm.
$1 < p < \infty$: Due to the representation theorem of Riesz all functionals on $L^p(\Omega)$ are given by the space $L^q(\Omega)$ with $1 = \frac{1}{p} + \frac{1}{q}$. Let $h \in L^q(\Omega)$ be given in such a way that

$$\int_\Omega h(x)u(x) \;=\; 0 \qquad \forall u \in \mathcal{C}_0(\Omega).$$

By Hölder's inequality is $h \in L^1_{loc}(\Omega)$, because

$$\int_\Omega h(x)\chi_K(x)\,dx \;\le\; \|h\|_{L^q(\Omega)}\|\chi_K\|_{L^p(\Omega)} \;=\; \|h\|_{L^q(\Omega)}\lambda(K)^{1/p} \;<\; \infty.$$

With Lemma 2.21 it follows that $h = 0$ in $L^q(\Omega)$. The density criterion of Corollary 2.9 yields that $\mathcal{C}_0(\Omega)$ is dense. □

2.4 Weak derivatives

In this section we introduce an extended form of derivatives. For the boundary value problem introduced before we only need to integrate over real intervals. However, we consider the more general case in $\mathbb{R}^n$, because that is needed in the following chapters.

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain, i.e. an open, non-empty and connected subset. Hence it is (Lebesgue-) measurable. The support $supp\,\varphi$ of a function $\varphi : \Omega \to \mathbb{R}$ is the closed set

$$supp\,\varphi \;:=\; \overline{\{x \in \Omega \,:\, \varphi(x) \neq 0\}}\,.$$

The set of C^∞-functions with compact support is denoted by

$$\mathcal{D}(\Omega) \;:=\; \{u \in C^\infty(\Omega) \,:\, supp\,u \text{ compact}\}$$

The derivatives of such functions to a multiindex $\alpha \in \mathbb{N}_0^n$ is denoted by $\partial^\alpha u$:

$$\partial^\alpha u \;:=\; \frac{\partial^{\alpha_n} \ldots \partial^{\alpha_1} u}{\partial x_n^{\alpha_n} \ldots \partial x_1^{\alpha_1}}$$

Definition 2.23 *Let $\alpha \in \mathbb{N}_0^n$ be a multi-index and $u \in L^1_{loc}(\Omega)$. A function $w_\alpha \in L^1_{loc}(\Omega)$ is called* weak derivative *of degree α of u, if*

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

Such functions w_α will be denoted by $D^\alpha u$.

Lemma 2.24 *For $u \in \mathcal{C}^m(\Omega)$, $m \in \mathbb{N}$, it holds $D^\alpha u = \partial^\alpha u$ for $|\alpha| \le m$.*

Proof. The proof follows immediately by integration by parts. ∎

However, the expressions '(classical) derivative' and 'weak derivative' are not equivalent: There exist functions which have weak derivative but not derivatives in the classical sense. We consider e.g. a bounded domain $\Omega \subseteq \mathbb{R}^2$ divided into two disjoint subdomains $\overline{\Omega} = \overline{\Omega_1 \cup \Omega_2}$. We denote the mutual boundary by $\Gamma := \partial\Omega_1 \cap \partial\Omega_2$. Now, we consider functions which are continuously differentiable on both subdomains, globally continuous, but on points on Γ not differentiable: $u|_{\Omega_i} \in C^1(\Omega_i)$ for $i \in \{1,2\}$, $u \in C(\Omega)$ but $u \notin C^1(\Omega)$. We may think e.g. in piecewise polynomials with the property to be continuous on Γ. We define functions

$$w|_{\Omega_i} := \frac{\partial u}{\partial x}\Big|_{\Omega_i} \quad \text{and} \quad w|_\Gamma \text{ arbitrary.}$$

It holds that $w \in L^1(\Omega) \subseteq L^1_{loc}(\Omega)$ and for arbitrary $\varphi \in \mathcal{D}(\Omega)$, because Γ (as a subset of Ω) is a null set:

$$\int_\Omega w\varphi \, dx \;=\; \int_{\Omega_1} w\varphi \, dx + \int_{\Omega_2} w\varphi \, dx \;=\; \int_{\Omega_1} \frac{\partial u}{\partial x}\varphi \, dx + \int_{\Omega_2} \frac{\partial u}{\partial x}\varphi \, dx \,.$$

Application of integration by parts and the notation $n^i = (n^i_x, n^i_y)$ for the normal vectors on Γ in direction from the interior of Ω_i to the exterior:

$$\int_{\Omega_i} \frac{\partial u}{\partial x}\varphi \, dx \;=\; -\int_{\Omega_i} u \frac{\partial \varphi}{\partial x} \, dx + \int_{\partial\Omega_i} u n^i_x \varphi \, ds$$

Since φ vanishes on $\partial\Omega$, $n^1 = -n^2$ and u is continuous on Γ, it follows:

$$\int_\Omega w\varphi \, dx \;=\; -\sum_{i=1}^{2} \int_{\Omega_i} u \frac{\partial \varphi}{\partial x} \, dx \;=\; -\int_\Omega u \frac{\partial \varphi}{\partial x} \, dx$$

Hence, w is a weak derivative of u in direction x, i.e. $w = D_x u$. For the derivative in direction of y we conclude similarly.

2.5 Sobolev spaces

Definition 2.25 *The Sobolev space of order $m \in \mathbb{N}_0$ in a domain $\Omega \subset \mathbb{R}^n$ and the power $1 \le p \le \infty$ is given by*

$$W^{m,p}(\Omega) \;:=\; \{u \in L^p(\Omega) \,:\, \exists \text{ weak derivative } D^\alpha u \in L^p(\Omega) \;\forall |\alpha| \le m\}.$$

On these spaces we define the semi-norms

$$|u|_{W^{m,p}(\Omega)} \;:=\; \left(\sum_{|\alpha|=m} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{1/p},$$

and the norms

$$\|u\|_{W^{m,p}(\Omega)} \;:=\; \left(\sum_{i=0}^{m} |u|^p_{W^{m,p}(\Omega)} \right)^{1/p} \;=\; \left(\sum_{|\alpha|\le m} \int_\Omega |D^\alpha u|^p \, dx \right)^{1/p}.$$

Sometimes the following equivalent norms are used:

$$\|u\|'_{W^{m,p}(\Omega)} \;:=\; \sum_{|\alpha|\le m} \|D^\alpha u\|_{L^p(\Omega)}.$$

The equivalence of these two norms is an easy exercise.

Theorem 2.26 *The spaces $W^{m,p}(\Omega)$ with $1 \le p \le \infty$ are Banach spaces.*

Proof. (a) At first, one has to verify that the properties of a norm are valid. The definiteness and the homogenity are obviously valid, but the triangle inequality is more tricky. We number all possible multi-indices with $|\alpha| \le m$ by $\alpha(0), \ldots, \alpha(l)$. For $u, v \in W^{m,p}(\Omega)$ we denote their derivatives by

$$\xi_i \;:=\; \|D^{\alpha(i)} u\|_{L^p(\Omega)} \quad \text{and} \quad \eta_i \;:= \|D^{\alpha(i)} v\|_{L^p(\Omega)}, \qquad i = 0, \ldots, l.$$

For $p = \infty$ is the triangle inequality trivial. For $1 \le p < \infty$ we use the inequality of Minkowski in l^p and the riangle inequality in $L^p(\Omega)$:

$$\begin{aligned}
\|u+v\|_{W^{m,p}(\Omega)} &= \left(\sum_{i=0}^{l} \|D^{\alpha(i)}(u+v)\|^p_{L^p(\Omega)} \right)^{1/p} \\
&\le \left(\sum_{i=0}^{l} |\xi_i + \eta_i|^p \right)^{1/p} \\
&\le \left(\sum_{i=0}^{l} |\xi_i|^p \right)^{1/p} + \left(\sum_{i=0}^{l} |\eta_i|^p \right)^{1/p} \\
&= \|u\|_{W^{m,p}(\Omega)} + \|v\|_{W^{m,p}(\Omega)}
\end{aligned}$$

(b) Completeness: Let $(u_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $W^{m,p}(\Omega)$, so that for $|\alpha| \le m$ the sequence $(D^\alpha u_n)_{n\in\mathbb{N}}$ is also Cauchy in $L^p(\Omega)$. Due to the completeness of $L^p(\Omega)$ there exists limits $w_\alpha \in L^p(\Omega)$:

$$\lim_{n\to\infty} D^\alpha u_n \;=\; w_\alpha \qquad \text{convergence in } L^p(\Omega)\,.$$

Let $\varphi \in \mathcal{D}(\Omega)$ be arbitrary given and let $supp\,\varphi \subset \omega \subset \Omega$ with a bounded domain ω, and hence with finite measure $\mu(\omega) < \infty$. We have the equality

$$\int_\omega D^\alpha u_n \varphi\, dx \;=\; \int_\Omega D^\alpha u_n \varphi\, dx \;=\; (-1)^{|\alpha|} \int_\Omega u_n \partial^\alpha \varphi \;=\; (-1)^{|\alpha|} \int_\omega u_n \partial^\alpha \varphi.$$

Due to Lemma 2.17 holds for $p \ge 1$ $D^\alpha u_n, w_\alpha \in L^1(\omega)$ and

$$D^\alpha u_n \;\to\; w_\alpha \quad \text{convergence in } L^1(\omega)\,.$$

Therefore, going to the limit $n \to \infty$ on both sides of the equation above yields

$$\int_\omega w_\alpha \varphi\, dx \;=\; (-1)^{|\alpha|} \int_\omega u \partial^\alpha \varphi.$$

Because of $supp\,\varphi \subset \omega$, the integration over ω can be replaced by the integration over Ω. Hence, w_α is weak derivative of u of degree α. This yields $u \in W^{m,p}(\Omega)$. ∎

Theorem 2.27 *For all $m \in \mathbb{N}$, $m \ge 1$, it holds $W^{m,p}(\Omega) \subset W^{m-1,p}(\Omega)$ with continuous embedding (identity):*

$$id : W^{m,p}(\Omega) \to W^{m-1,p}(\Omega)\,.$$

Proof. The inclusion $W^{m,p}(\Omega) \subset W^{m-1,p}(\Omega)$ is obvious. Continuity of the embedding is a consequence of the estimate

$$\|u\|_{W^{m-1,p}(\Omega)} \;\le\; \|u\|_{W^{m,p}(\Omega)}\,.$$

∎

Theorem 2.28 *The space $H^m(\Omega) := W^{m,2}(\Omega)$ has a scalar product*

$$(u,v)_{H^m(\Omega)} \;:=\; \sum_{0\le|\alpha|\le m} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}$$

which generates the same topology as its norm. Hence it is a Hilbert space.

Proof. The completeness was already shown. Verifying that this scalar product generates the norms is an obvious task. ∎

Theorem 2.29 (Meyers & Serrin) *Let $1 \le p < \infty$. Then $C^\infty(\Omega) \cap W^{1,p}(\Omega)$ is dense in $W^{1,p}(\Omega)$.*

Proof. The proof can be carried out by considering mollifiers and the consideration of bounds to compact supports. We refer to standard literature in Functional Analysis. ∎
In one dimension $\Omega \subset \mathbb{R}$ we state that $C[a,b]$ is dense in $H^1(a,b)$.

Chapter 3

Variational formulation

The variational formulation of the boundary value problem (BVP) (1.1) makes use of the following (affine) Hilbert spaces

$$\begin{aligned} V &:= H^1(0,1) \\ V_0 &:= H_0^1(0,1) \;=\; \{u \in H^1(0,1) : u(0) = u(1) = 0\}, \\ V_g &:= \{u \in H^1(0,1) : u(0) = g_0, u(1) = g_1\}. \end{aligned}$$

We like to mention that it is not clear yet whether the point values $u(0)$ and $u(1)$ exist. However, we will address this point in the following sections and give a legitimation to use such point values for functions in V.

The L^2-scalar product in the intervall $(0,1)$ will be denoted by

$$(u,v) \;:=\; \int_0^1 uv\,dx\,.$$

Since L^2 is identical to its dual space, $L^2(0,1) = L^2(0,1)'$, we can consider the right hand side $f \in L^2(0,1)$ as a functional in $L^2(0,1)' \subset V'$. The evaluation of the functional f for a function ϕ will be written in the form $\langle f, \phi\rangle$. Hence, in this particular case it holds

$$\langle f, \phi\rangle \;=\; (f, \phi).$$

Further, for a more condensed form of presentation, we introduce the bilinear form $A : V \times V \to \mathbb{R}$, defined as

$$A(u,\phi) \;:=\; \int_0^1 au'\phi'\,dx,$$

where the coefficient a in the integral may depend on x. Now the variational formulation reads

$$u \in V_g : \quad A(u,\phi) \;=\; \langle f, \phi\rangle \qquad \forall \phi \in V_0\,. \tag{3.1}$$

In order to have a well-defined integral, we request for the data

$$f \in L^2(0,1)\,, \quad a \in L^\infty(0,1;\mathbb{R}^+)$$

Definition 3.1 *A solution $u \in V_g$ of (3.1) is called* weak solution *of the BVP (1.1).*

Theorem 3.2 *We assume for the data that $a \in C^1[0,1]$, $f \in C(0,1)$. Then each classical solution $u \in C^2[0,1]$ of (1.1) is also a weak solution of (3.1).*

Proof. It is sufficient to show that $u \in C^2[0,1] \subset H^1(0,1)$. This is obvious, because I is a closed interval where u' takes its maximum. Equation (3.1) follows by integration by parts. ∎

Remark 3.3 *A classical solution $u \in C^2(0,1)$ is not necessarily a weak solution. For instance, $u(x) = \sqrt{2x - x^2}$ is a classical solution of*

$$\begin{aligned} -(a(x)u'(x))' &= 1 \qquad \forall x \in (0,1) \\ u(0) &= 0 \\ u(1) &= 1, \end{aligned}$$

with $a(x) = \sqrt{2x - x^2}$. But $\|u\|_{H^1(0,1)} = \infty$ so that $u \notin H^1(0,1)$.

Remark 3.4 *A weak solution $u \in V_g$ of (3.1) is not necessarily a classical solution, because (1.1) does not hold without existence of the second order derivatives of u in $(0,1)$.*

A slightly different formulation is as follows: We choose an arbitrary $\bar{g} \in V_g$ and define $f_{\bar{g}} \in V'$ by

$$\langle f_{\bar{g}}, \phi \rangle \;:=\; \langle f, \phi \rangle - A(\bar{g}, \phi).$$

Then $u \in V_g$ is a solution of (1.2), iff $u_0 := u - \bar{g} \in V_0$ is a solution of

$$A(u_0, \phi) \;=\; \langle f_{\bar{g}}, \phi \rangle \qquad \forall \phi \in V_0.$$

This problem is of the same type as previously with homogenenous Dirichlet data.. A partial differential equation (PDE) with inhomogenenous Dirichlet data can be reformulated by modification of the right hand side into an equation with homogenenous Dirichlet data. Therefore, we may concentrate in the following to the case of homogenenous boundary data. By a suitable reformulation ($u_0 \to u$, $f_{\bar{g}} \to f$) we arrive at a problem of the form:

$$u \in V_0 : \qquad A(u, \phi) \;=\; \langle f, \phi \rangle \qquad \forall \phi \in V_0\,. \tag{3.2}$$

It is an easy exercise that u does not depend on the particular choice $\bar{g} \in V_g$.

3.1 The trace operator in 1-D

As already mentioned above, we still don't know whether the requirement of Dirichlet data of H^1-functions is meaningful, since Lebesgue-integrable functions can be arbitrarily modified on Lebesgue null sets without changing its equivalence class. However, we will see that in the case of H^1-functions such Dirichlet data is justified. We will consider a continuous map

$$\gamma : H^1(0,1) \to \mathbb{R}^2, \quad u \mapsto (u(0), u(1))$$

which is uniquely defined in a certain sense.

Lemma 3.5 *We consider the intervall $[a,b]$ of lenght $l = b - a > 0$. Then it holds*

$$\max(|u(a)|, |u(b)|) \leq C_{tr} \|u\|_{H^1(a,b)} \qquad \forall u \in C[a,b],$$

with the constant $C_{tr} = \sqrt{2}\max(l^{1/2}, l^{-1/2})$.

Proof. It holds for arbitrary $x \in [a,b]$:

$$u(x) = u(a) + \int_a^x u'(y)\, dy.$$

Hence, we obtain the bound

$$|u(a)| \leq |u(x)| + \int_a^x |u'(y)|\, dy \leq |u(x)| + \|u'\|_{L^1(a,b)}.$$

Integration of both sides and the use of (2.1) yields

$$\begin{aligned} l|u(a)| = \int_a^b |u(a)|\, dx &\leq \int_a^b |u(x)|\, dx + \|u'\|_{L^1(a,b)} \int_a^b dx \\ &= \|u\|_{L^1(a,b)} + l\|u'\|_{L^1(a,b)} \\ &\leq l^{1/2}\|u\|_{L^2(a,b)} + l^{3/2}\|u'\|_{L^2(a,b)}. \end{aligned}$$

Now we divide both sides by l:

$$\begin{aligned} |u(a)| &\leq l^{-1/2}\|u\|_{L^2(a,b)} + l^{1/2}\|u'\|_{L^2(a,b)} \\ &\leq \max(l^{-1/2}, l^{1/2})\sqrt{2}(\|u\|^2_{L^2(a,b)} + \|u'\|^2_{L^2(a,b)})^{1/2} \\ &= C_{tr}\|u\|_{H^1(a,b)}. \end{aligned}$$

□

Theorem 3.6 (Trace theorem in 1-D) *We consider the intervall $[a,b]$ of lenght $l = b - a > 0$. It exists a continuous operator (called* trace operator*)*

$$\gamma : H^1(a,b) \to \mathbb{R}^2$$

with the property $\gamma(u) = (u(a), u(b))^T$ for every $u \in C[a,b]$.

Proof. As a result of the previous lemma we deduce that the functional

$$\tilde{\gamma}_0 : C^1[a,b] \to \mathbb{R}, \qquad u \mapsto \tilde{\gamma}_0(u) = u(a)$$

is linear and bounded with respect to the $H^1(a,b)$-norm and hence also continuous. Now we make use of the fact that $C^1[a,b]$ is dense in $H^1(a,b)$. Therefore, it exists a unique continuous extension

$$\gamma_0 : H^1(a,b) \to \mathbb{R}, \qquad u \mapsto \gamma_0(u),$$

with

$$\|\gamma_0\|_{H^1(a,b)} \quad = \quad \|\tilde{\gamma}_0\|_{C^1[a,b]}.$$

For this extension holds $\gamma_0(u) = u(a)$ if $u \in C^1[a,b]$. This is the reason why we can assign to $H^1(a,b)$-functions the point value $u(a) := \gamma_0(u)$ in a meaningful way. An analogous argumentation yields $u(b) := \gamma_1(u)$. In summary, we state that the operator $\gamma := \gamma_0 \times \gamma_1 : H^1(a,b) \to \mathbb{R}^2$ is continuous which yield the assertion. ∎

3.2 Trace theorem in several dimensions

Now we consider the Poisson problem in multi dimensions

$$-\Delta u = f \qquad \text{in } \Omega \tag{3.3}$$

$$u = u_0 \qquad \text{on } \partial\Omega. \tag{3.4}$$

Here, $\Omega \subset \mathbb{R}^d$ is a domain (open and connected), and Δ is the Laplace operator

$$\Delta u(x) \quad := \quad \sum_{i=1}^{d} \frac{\partial^2 u(x)}{\partial x_i^2}.$$

The corresponding variational space is the Sobolev space $H^1(\Omega)$. However, we should firstly analyze the traces of H^1-functions on the boundary $\partial\Omega$ of the domain.

For spatial dimensions $d \geq 2$ the embedding $H^1(\Omega)$ in $C(\Omega)$ does in general not hold anymore. But, however, the traces of H^1 functions are still L^2 functions on the boundary, if we assume a certain regularity of the boundary $\partial\Omega$.

Example. For $d \geq 3$ the functions

$$u(x) \quad = \quad \|x\|_2^{-\alpha}$$

are in $H^1(\Omega)$ if $0 < \alpha < d/2 - 1$, but they are not any more continuous in the origin. They are unbounded in L^∞ but in H^1.

In what follows we will make use of the following technical lemma:

Lemma 3.7 *Let $\epsilon > 0$ and $f : [0, \epsilon] \to \mathbb{R}$ be a continuous functions. Then it holds:*

$$\int_0^\epsilon \int_0^t f(s)\, ds\, dt \;=\; \int_0^\epsilon f(t)(\epsilon - t)\, dt.$$

Proof. Let F be the primitive of f. By integration by parts we obtain:

$$\begin{aligned}
\int_0^\epsilon \int_0^t f(s)\, ds\, dt &= \int_0^\epsilon (F(t) - F(0))\, dt \\
&= \int_0^\epsilon F(t) \cdot 1\, dt - \epsilon F(0) \\
&= F(\epsilon)\epsilon - \int_0^\epsilon f(t) t\, dt - \epsilon F(0) \\
&= \epsilon(F(\epsilon) - F(0)) - \int_0^\epsilon f(t) t\, dt \\
&= \int_0^\epsilon f(t)(\epsilon - t)\, dt.
\end{aligned}$$

□

In this section we will consider domains $\Omega \subset \mathbb{R}^d$ with certain properties of the boundary:

Definition 3.8 *Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a domain:*

- *We say that Ω has a* piecewise smooth boundary *$\partial\Omega$, if a finite partition $\gamma_1 \cup \ldots \cup \gamma_m = \partial\Omega$ exists, s.t. each part γ_i is a C^1-curve.*
- *It satisfies an* inner cone condition*, if for every point $x \in \partial\Omega$ it exists a normalized vector $\xi(x)$ and a non-trivial cone of the form*

  $$K_x \;=\; \left\{x + y : \|y\| \leq r,\ y^T \xi(x) \geq \|y\| \cos\alpha\right\} \;\subseteq\; \Omega.$$

  Here, $0 < \alpha < \pi/2$ and $r > 0$, are not allowed to depend on x.
- *It satisfies an* outer cone condition*, if non-trivial cones K_x exists, s.t. $\Omega \cap K_x = \emptyset$ for all $x \in \partial\Omega$.*

The inner cone condition implies for $d = 2$ that all inner angles at non-differentiable boundary points are positive. For instance, domains with polygonal boundaries and inner angles at the edges $0 < \omega < 2\pi$ satisfy the inner cone condition. However, the inner cone condition is a weaker restriction than a local Lipschitz condition, i.e. the boundary can be locally parameterized by a Lipschitz continuous function..

Lemma 3.9 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a piecewise smooth boundary $\partial\Omega$ and an inner cone condition. Then there exists $\epsilon > 0$ such that for every smooth part*

$\Gamma_i \subset \partial\Omega$ and a corresponding C^1-parameterization (probably after rotation of the Cartesian coordinates) in form of a graph of a function $\phi_i \in C^1(I_i, \mathbb{R})$,

$$\Gamma_i \;:=\; G_{\phi_i} \;=\; \{(x, \phi_i(x)) \,|\, x \in I_i\},$$

it holds

$$\omega_i \;:=\; \{(x,y) \in I_i \times \mathbb{R} \,:\, \phi_i(x) < y < \phi_i(x) + \epsilon\} \;\subseteq\; \Omega.$$

Theorem 3.10 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a piecewise smooth boundary $\partial\Omega$ and an inner cone condition. Then $C^1(\overline{\Omega})$ is densly embedded in $H^1(\Omega)$.*

Theorem 3.11 (Trace theorem) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with piecewise smooth boundary $\partial\Omega$ and an inner cone condition. Then there exists a linear and continuous map*

$$\gamma : H^1(\Omega) \to L^2(\partial\Omega)$$

with $\gamma|_{C(\overline{\Omega})} \equiv id$.

Proof. Because of the dense embedding $C^1(\overline{\Omega}) \subset H^1(\Omega)$ it is sufficient to show that the identity $\gamma := id$ satisfies the following bound for $C^1(\overline{\Omega})$-functions:

$$\|v\|_{L^2(\partial\Omega)} \;\leq\; c\|v\|_{H^1(\Omega)} \qquad \forall v \in C^1(\overline{\Omega}),$$

with a constant c depending on Ω only. For simplicity, we restrict to the two-dimensional case, $d = 2$.

We split $\partial\Omega$ in m smooth parts $\Gamma_1, \ldots, \Gamma_m$. Let ϕ_i be the parameterizations and ω_i be the sets according to Lemma 3.9. The boundary integral on Γ_i is given by

$$\|v\|^2_{L^2(\Gamma_i)} \;=\; \int_{\Gamma_i} v^2 \, ds \;=\; \int_{I_i} v(x, \phi_i(x))^2 \|(1, \phi'(x))\|_2 \, dx. \tag{3.5}$$

Now, we will bound the point values $v(x,y)^2$ with $y = \phi_i(x)$ from above in a proper way. On Γ_i we have the representation

$$v(x,y) \;=\; v(x, \phi_i(x) + t) - \int_0^t \partial_y v(x, \phi_i(x) + s) \, ds \qquad (0 \leq t \leq \epsilon).$$

Integration of both sides over $0 \leq t \leq \epsilon$ yields

$$\epsilon v(x,y) \;=\; \int_0^\epsilon v(x, \phi_i(x) + t) \, dt - \int_0^\epsilon \int_0^t \partial_y v(x, \phi_i(x) + s) \, ds \, dt.$$

By the previous Lemma and the notation $g_1(t) := v(x, \phi_i(x) + t)$, $g_2(t) := \partial_y v(x, \phi_i(x) + t)$ we obtain:

$$\epsilon v(x,y) \;=\; \int_0^\epsilon g_1(t) \, dt - \int_0^\epsilon g_2(t)(\epsilon - t) \, dt.$$

We take the square on both sides and apply the inequality of Hölder to transform the L^1-norm to the L^2-norm:

$$\begin{aligned} \epsilon^2 v(x,y)^2 &\leq 2\|g_1\|^2_{L^1(0,\epsilon)} + 2\|g_2(t)(\epsilon - t)\|^2_{L^1(0,\epsilon)} \\ &\leq 2(\|1\|^2_{L^2(0,\epsilon)}\|g_1\|^2_{L^2(0,\epsilon)} + \|t\|^2_{L^2(0,\epsilon)}\|g_2\|^2_{L^2(0,\epsilon)}) \\ &= 2(\epsilon\|g_1\|^2_{L^2(0,\epsilon)} + \frac{1}{3}\epsilon^3\|g_2\|^2_{L^2(0,\epsilon)}). \end{aligned}$$

We use this point-wise upper bound for the integral (3.5). With $c_i := \max\{(1+|\phi'(x)|)^{1/2} : x \in I_i\}$ we obtain:

$$\|v\|^2_{L^2(\Gamma_i)} \leq c_i \left(2\epsilon^{-1}\|v\|^2_{L^2(\omega_i)} + \frac{2}{3}\epsilon\|\partial_y v\|^2_{L^2(\omega_i)}\right).$$

We sum over all boundary parts and set $c := 2\sum_{i=1}^m c_i$:

$$\|v\|^2_{L^2(\gamma)} \leq c\left(\epsilon^{-1}\|v\|^2_{L^2(\Omega)} + \frac{1}{3}\epsilon|v|^2_{H^1(\Omega)}\right).$$

Here, we replaced the norm $\|\partial_y v\|_{L^2(\Omega)}$ by the norm of the full gradient in order to account for the previously mentioned coordinate transformation. □

Examples:

- The domain

  $$\Omega_1 = \{(x,y) \in \mathbb{R}^2 : 0 < y < x^5 < 1\}$$

  does not satisfies such inner cone condition. The function $u(x,y) = x^{-1}$ is in $H^1(\Omega_1)$, but his trace $\gamma(u)$ is not square integrable over $\partial\Omega$.

- A further counter example is the 'dotted disk' which does not satisfys an outer cone condition:

  $$\Omega_2 = \{(x_1,x_2) \in \mathbb{R}^2 : 0 < \|x\|_2 < 1\}.$$

  Let us consider the function

  $$w_0(r) := \ln\ln(e/r).$$

  with a singularity at the origin, and their regularizations $w_n \in H^1(\Omega_2)$ (Exercise!) for $n \in \mathbb{N}$:

  $$w_n(r) := \begin{cases} w_0(r) & \text{for } r \geq 1/n \\ \ln\ln(en) & \text{for } r < 1/n. \end{cases}$$

Then the functions

$$u_n(x) \;:=\; \frac{w_n(\|x\|_2)}{w_n(0)}$$

are in $H^1(\Omega)$ and satisfy the boundary condition

$$u_n(x) \;=\; 1 \quad \text{for } \|x\|_2 = 1,$$

For $n \to \infty$ the sequence $u_n \to 0$ converges a.e. in $\overline{\Omega}$. In particular holds $\lim_{n\to\infty} \|\nabla u_n\| = 0$. Hence, $u_n \to u = 0$ in $H^1(\Omega)$, but $\lim_{n\to\infty} u_n(0) = 1 \neq u(0) = 0$. We have constructed a sequence which converges in $H^1(\Omega)$ and with a limit which does not satisfy the boundary condition at the origin.

3.3 The inequality of Poincaré

3.3.1 Inequality of Poincaré in 1D

As a consequence of the trace theorem we can define the space

$$H_0^1(0,1) \;:=\; \{u \in H^1(0,1) : u(0) = u(1) = 0\} \;=\; Kern(\gamma) \tag{3.6}$$

as a subspace of $H^1(0,1)$. Furthermore, we will show that on $H_0^1(0,1)$ there exists another scalar product.

Theorem 3.12 (Inequality of Poincaré in 1-D) *For any intervall $[a,b]$ of lenght $l = b-a > 0$ there exists a constant $C_p = l^{3/4}$ such that the following bound is valid:*

$$\|u\|_{L^2(a,b)} \;\leq\; C_p \|u'\|_{L^2(a,b)} \qquad \forall u \in H_0^1(a,b)\,.$$

Proof. Similar to the proof of the trace theorem, we will show the assertion firstly for functions $u \in C^1[a,b] \cap H_0^1(a,b)$. Due to the trace theorem we deduce that $u(a) = 0$. Then, we have

$$u(x) \;=\; \int_a^x u'(y)\,dy.$$

This yields with $l := b - a$:

$$\|u\|_{L^\infty(a,b)} \;=\; \max_{x\in[a,b]} |u(x)| \;\leq\; \int_a^b |u'(y)|\,dy \;=\; \|u'\|_{L^1(a,b)} \leq\; l^{1/2}\|u'\|_{L^2(a,b)}.$$

Here, the last inequality holds due to the fact that $C^1[a,b] \subset L^2(a,b)$ and the Cauchy inequality. This implies for the L^2-norm:

$$\|u\|^2_{L^2(a,b)} \;\leq\; \int_a^b \|u\|^2_{L^\infty(a,b)}\,dx = l\|u\|^2_{L^\infty(a,b)} \;\leq\; l^{3/2}\|u'\|^2_{L^2(a,b)}.$$

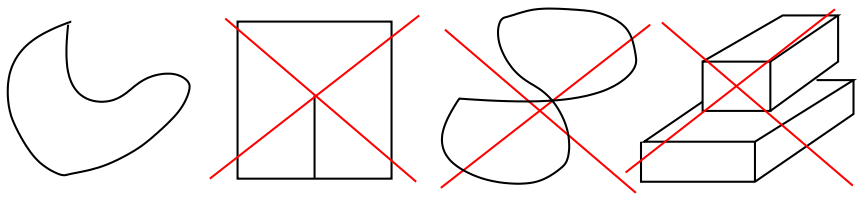


Figure 3.1: Examples of a Lipschitz domain (left) and three non-Lipschitz domains (right).

This shows the assertion for C^1-functions. For the more general case $u \in H_0^1(a,b)$ we can pass to the limit due to the density of $C^1[a,b] \cap H_0^1(a,b)$ in $H_0^1(a,b)$ and the continuity of the terms appearing:

$$\|u\|^2_{L^2(a,b)} \;=\; \lim_{n\to\infty} \|u_n\|^2_{L^2(a,b)} \;\le\; l^{3/2} \lim_{n\to\infty} \|u_n'\|^2_{L^2(a,b)} \;=\; l^{3/2}\|u'\|^2_{L^2(a,b)}.$$

□

In the case of the unity interval $(0,1)$ the Poincaré constant is simply $C_p = 1$.

3.3.2 Inequality of Poincaré in multi dimensions

An important class of domains are so-called Lipschitz domains:

Definition 3.13 *A domain $\Omega \subset \mathbb{R}^d$ is called Lipschitz domain, if for each point $x \in \partial\Omega$ a neighborhood U with $x \in U$ exists, s.t. $U \cap \partial\Omega$ can be represented as a graph of a Lipschitz-continuous function.*

Examples of Lipschitz and non-Lipschitz domains are depicted in Fig. 3.1. Every Lipschitz domain satisfies an outer cone condition, see Alt [1]. A domain with piecewise smooth boundary and an outer cone condition is automatically a Lipschitz domain. We state the following central theorem without proof:

Theorem 3.14 (Rellich embedding theorem) *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then the embeddding $H^{m+1}(\Omega) \subset H^m(\Omega)$ is compact for each $m \in \mathbb{N}_0$.*

This theorem leads to the following implication: The unit sphere

$$\{v \in H^{m+1}(\Omega) \,:\, \|v\|_{H^{m+1}(\Omega)} \le 1\}$$

is relatively compact in the $H^m(\Omega)$-norm.

Corollary 3.15 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then each weakly convergent sequence in $H^{m+1}(\Omega)$ is strongly convergent in $H^m(\Omega)$.*

Proof. Let $(v_n)_{n\in\mathbb{N}}$ be a weakly convergent sequence in $H^{m+1}(\Omega)$ with $v_n \rightharpoonup v \in H^{m+1}(\Omega)$. Because weakly convergent sequences are always bounded, this sequence forms a relatively compact set in $H^m(\Omega)$. Hence, its closure

$$\overline{\{v_k \,:\, k\in\mathbb{N}\}} \qquad \text{(closure w.r.t } H^m(\Omega)\text{-norm)}$$

is a compact set in $H^m(\Omega)$. This implies the existence of a subsequence (v_{n_k}) strongly converging in $H^m(\Omega)$ to $w \in H^m(\Omega)$, and also $v_{n_k} \rightharpoonup w$ in $H^m(\Omega)$. Because $H^m(\Omega)' \subseteq H^{m+1}(\Omega)'$ it also holds $v_{n_k} \rightharpoonup v$ in $H^m(\Omega)$. Since weak limits are unique, it follows $v = w$. The same argumentation is valid for arbitrary subsequences of (v_n). Therefore, v is the unique agglomeration point of the sequence, and $v_n \to v$ in $H^m(\Omega)$. ∎

Theorem 3.16 (General inequality of Poincaré) *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain, and $\Gamma \subset \partial\Omega$ be a $\mathbb{R}^{d-1}$ measurable sets with non-trivial measure. In the space $V_0 := \{u \in H^1(\Omega) \,:\, u = 0$ a.e. on $\Gamma\}$ it holds the inequality*

$$\|u\|_{L^2(\Omega)} \;\leq\; c_\Omega |u|_{H^1(\Omega)} \qquad \forall u \in V_0,$$

with a constant c_Ω only depending on Ω.

Proof. In this proof we use the short notation $\|\cdot\|$ for the $L^2(\Omega)$-norm. Assuming that the assertion does not hold, then there would exist a sequence $(u_k)_{k\in\mathbb{N}}$ in V_0 with

$$\|u_k\| \;\geq\; k\|\nabla u_k\|.$$

We consider the functions $w_k := u_k/\|u_k\| \in V_0$. We have $\|w_k\| = 1$ and

$$\|\nabla w_k\| \;=\; \|\nabla u_k\|/\|u_k\| \;\leq\; \frac{1}{k},$$

so that $\nabla w_k \to 0$ in $L^2(\Omega)$. The sequence $(w_k)_{k\in\mathbb{N}}$ is a bounded set in $V_0 \subset H^1(\Omega)$. Consequently, it exists a weakly convergent subsequence with weak limit $w \in V_0$. We denote this subsequence again by $(w_k)_{k\in\mathbb{N}}$ so that $w_k \rightharpoonup w$ in $H^1(\Omega)$. Due to Corollary 3.15 we know that $w_k \to w$ in $L^2(\Omega)$. Together with the L^2-convergence $\nabla w_k \to 0$ in $L^2(\Omega)$ we follow that $w_k \to w$ in H^1 with $\nabla w = 0$ a.e. in Ω. Since domains are connected, we know that w is a.e. equal to a constant w_0. Due to $1 = \lim_{k\to\infty} 1 = \lim_{k\to\infty} \|w_k\|_{L^2(\Omega)} = \|w\|_{L^2(\Omega)}$ it holds that $w_0 \neq 0$. But due to $w|_\Gamma = 0$ we obtain two different representatives of the same H^1-function with different traces on Γ, namely $w|_\Gamma = 0$ and $w|_\Gamma = w_0 \neq 0$. This is a contradiction to the trace theorem. ∎

3.4 Existence and uniqueness of weak solutions

Corollary 3.17 *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded domain with Lipschitz boundary $\partial\Omega$. The semi-norm $|\cdot|_{H^1(\Omega)}$ is indeed a norm on $H^1_0(\Omega)$ and it is equivalent to $\|\cdot\|_{H^1(\Omega)}$.*

Moreover,

$$(u,v)_{H_0^1(\Omega)} \quad := \quad \int_\Omega \nabla u \nabla v \, dx$$

is a scalar product on $H_0^1(\Omega)$ which induces the norm $|\cdot|_{H^1(\Omega)}$. Hence, $H_0^1(\Omega)$ in connection with the norm $|\cdot|_{H^1(\Omega)}$ is a Hilbert space.

Proof. Due to the inequality of Poincaré (Thm. 3.12 and 3.16) it holds

$$\begin{aligned} \|u\|^2_{H^1(\Omega)} &= \|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)} \\ &\leq (1+C_p^2)\|\nabla u\|^2_{L^2(\Omega)} \\ &\leq (1+C_p^2)|u|^2_{H^1(\Omega)}. \end{aligned}$$

This yields that $\|\nabla u\|_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)}$ and hence the equivalence of the norms. Checking that $(\cdot,\cdot)_{H_0^1(\Omega)}$ is a scalar product is an easy task. Furthermore, it is obvious that

$$(u,u)_{H_0^1(\Omega)} = \|\nabla u\|^2_{L^2(\Omega)}.$$

□

Now we are going to address the solvability of the problem (3.2). We firstly consider the case of a constant coefficient $a \equiv 1$. In this case, the bilinear form $A : V \times V \to \mathbb{R}$ is just the H_0^1 scalar product:

$$A(u,v) \quad = \quad (\nabla u, \nabla v) \quad = \quad \int_\Omega \nabla u(x) \nabla v(x) \, dx. \tag{3.7}$$

The solvability is now a consequence of the representation theorem of Riesz (Thm. 2.18).

Lemma 3.18 *Let H be a real Hilbert space and $f \in H'$. Then the following problems are equivalent:*

$$\begin{aligned} &(a) \qquad u \in H: \quad (u,\phi)_H = \langle f, \phi \rangle \qquad \forall \phi \in H \\ &(b) \qquad u \in H: \quad J(u) = \min_{v \in H} J(v) \text{ for the functional } J(v) := \frac{1}{2}\|v\|^2_H - \langle f, v \rangle. \end{aligned}$$

Proof. A solution u of problem (b) is characterised by

$$J(u) \quad \leq \quad J(u+tw)$$

for all $w \in H$ and all $t \in [0,1]$. It holds

$$J(u+tw) \quad = \quad J(u) + \frac{t^2}{2}\|w\|^2_H + t(u,w)_H - t\langle f, w \rangle.$$

Hence problem (b) is equivalent to

$$u \in H: \qquad \frac{t}{2}\|w\|_V^2 + (u,w)_H - \langle f, w\rangle \;\leq\; 0 \qquad \forall w \in H\,,\ \forall t \in [0,1],$$

or formulated in the form

$$u \in H: \qquad (u,w)_H \;\leq\; \langle f, w\rangle \qquad \forall w \in H.$$

Since with $w \in H$ it holds also $-w \in H$, this condition is exactly fulfilled, iff u solves problem (a). ∎

This lemma can be generalized in the form that, instead of (3.7), we request for the bilinear form $A(\cdot,\cdot)$ only:

- A is symmetric, i.e. $A(u,v) = A(v,u)$ for all $u, v \in H$, and
- A is positive, i.e. $A(u,u) > 0$ for all $u \in H \setminus \{0\}$.

It is easy to show that these conditions are sufficient to show that the equation $A(u,v) = \langle f, v\rangle$ for all $v \in H$ is equivalent to the minimization of

$$J(v) \;=\; \frac{1}{2}A(v,v) - \langle f, v\rangle \;\to\; \min.$$

Theorem 3.19 (Representation theorem of Riesz) *Let H be a real Hilbert space and $f \in H'$. Then the Problem*

$$u \in H: \quad (u,\phi)_H \;=\; \langle f, \phi\rangle \qquad \forall \phi \in H \tag{3.8}$$

has a unique solution $u \in H$ and it holds $\|u\|_H = \|f\|_{H'}$.

Proof. (a) Existence: The functional J in Lemma 3.18 is bounded from below:

$$J(v) \;\geq\; \frac{1}{2}\|v\|_H^2 - \|f\|_{H'}\|v\|_H \;= \frac{1}{2}(\|v\|_H - \|f\|_{H'})^2 - \frac{1}{2}\|f\|_{H'}^2 \geq\; -\frac{1}{2}\|f\|_{H'}^2.$$

Therefore there exists a minimal sequence $(u_k)_{k\in\mathbb{N}}$ in H with

$$\lim_{k\to\infty} J(u_k) \;=\; \inf_{v\in H} J(v) \;>\; -\infty.$$

Furthermore, this sequence forms a Cauchy sequence, because the parallelogram law yields

$$\begin{aligned}
\|u_m - u_n\|_H^2 &= 2\|u_m\|^2 + 2\|u_n\|^2 - \|u_m + u_n\|^2 \\
&= 4J(u_m) + 4J(u_n) - 8J((u_m + u_n)/2) \\
&\leq 4J(u_m) + 4J(u_n) - 8\inf_{v\in H} J(v).
\end{aligned}$$

Taking the limit $m, n \to \infty$ leads to

$$\lim_{m,n\to\infty} \|u_m - u_n\|_H^2 \;\leq\; 0$$

which is the Cauchy property. Because H (as Hilbert space) is complete, this sequence converges to a $u = \lim_{n\to\infty} u_n \in H$. By continuity of J it follows that

$$J(u) \;=\; \lim_{n\to\infty} J(u_n) \;=\; \inf_{v\in H} J(v)\,.$$

Due to Lemma 3.18, u is solution of the considered variational problem (3.8).
(b) Uniqueness: Let now $\tilde{u}$ be a secondary solution of (3.8). Then the difference $u - \tilde{u}$ satisfies the equation

$$(u-\tilde{u},\phi)_H \;=\; 0 \qquad \forall \phi \in H.$$

Setting in particular $\phi := u - \tilde{u}$ yields $\|u-\tilde{u}\|_H^2 = 0$ which implies $u = \tilde{u}$.
(c) $\|u\|_H = \|f\|_{H'}$:

$$\|f\|_{H'} \;=\; \sup_{0\neq v\in H} \frac{\langle f, v\rangle}{\|v\|_H} \;=\; \sup_{0\neq v\in H} \frac{(u,v)_H}{\|v\|_H} \;=\; \|u\|_H.$$

Here, the last equality is an implication of the inequality of Cauchy

$$(u,v)_H \;\leq\; \|v\|_H\,\|u\|_H$$

and the choice $v := u$:

$$\sup_{0\neq v\in H} (u,v)_H \;\geq\; \|u\|_H^2.$$

□

This result directly implies that H and H' are isomorph to each other: Each $f \in H'$ can be identified uniquely with $Rf \in H$ by the property

$$(Rf, v)_H \;=\; \langle f, v\rangle \qquad \forall v \in H.$$

We formulate this result in the following Corollary.

Corollary 3.20 *In every real Hilbert space H is the mapping $R : H' \to H$, $f \mapsto Rf = u$ of the previous Theorem an isomorphism, $H' \simeq H$. This isomorphism is called* isomorphism of Riesz.

3.5 Theorem of Lax-Milgram

In the representation theorem of Riesz it is important that the bilinear form $A(\cdot,\cdot)$ is just a scalar product in H. This will be generalized in this section. In particular, we consider bilinear forms with the following properties:

Definition 3.21 *A bilinear form $A : H \times H \to \mathbb{R}$ of a real Hilbert space H is called:*
*(a) H-*continuous *(or* bounded*), if a constant $\alpha_1 \geq 0$ exists, s.t.*

$$|A(u,v)| \quad \leq \quad \alpha_1 \|u\|_H \|v\|_H \qquad \forall u, v \in H,$$

*(b) H-*elliptic *(or H-*coercive*), if a constante $\alpha_2 > 0$ exists, s.t.*

$$A(u,u) \quad \geq \quad \alpha_2 \|u\|_H^2 \qquad \forall u \in H.$$

The continuity implies that for each fixed $v \in H$ the mapping

$$\mathcal{A}v : H \to \mathbb{R}\,, \qquad w \mapsto \langle \mathcal{A}v, w \rangle := A(v,w)$$

is continuous, $\mathcal{A}v \in H'$. This implies:

$$\|\mathcal{A}v\|_{H'} \quad = \quad \sup_{0 \neq w \in H} \frac{\langle \mathcal{A}v, w \rangle}{\|w\|_H} \quad \leq \quad \sup_{0 \neq w \in H} \frac{\alpha_1 \|v\|_H \|w\|_H}{\|w\|_H} \quad = \quad \alpha_1 \|v\|_H. \tag{3.9}$$

Therefore, we can consider $\mathcal{A}$ as a linear continuous mapping

$$\mathcal{A} : H \to H'.$$

Furthermore, it holds obviously that $0 < \alpha_2 \leq \alpha_1$. The following theorem of the two mathematicians Lax[1] and Milgram[2] now states that $\mathcal{A}$ is a bijective operator.

Theorem 3.22 (Lax-Milgram) *Let H be a real Hilbert space and $A : H \times H \to \mathbb{R}$ a H-continuous and H-elliptic bilinear form with corresponding constants $\alpha_1, \alpha_2 > 0$. Then to each functional $f \in H'$ there exists a unique solution $u \in H$ of the variational problem*

$$A(u,v) \quad = \quad \langle f, v \rangle \qquad \forall v \in H. \tag{3.10}$$

Further, the solution has the following bounds:

$$\frac{1}{\alpha_1} \|f\|_{H'} \quad \leq \quad \|u\|_H \quad \leq \quad \frac{1}{\alpha_2} \|f\|_{H'}.$$

Proof. (a) We write the equation (3.10) in form of a fixed point problem:

$$B(u) \quad = \quad u$$

with the operator

$$B(u) \quad = \quad u + \tau R(f - \mathcal{A}u).$$

[1] Peter David Lax, $*$ 1926, hungarian mathematician, Wolf award for mathematics in 1987, and Abel award 2005

[2] Arthur Norton Milgram, 1912-1961, US american mathematician.

Here, $R : H' \to H$ denotes the previously introduced isomorphism of Riesz. The parameter $\tau > 0$ is still arbitrary and will be fixed later. We want to apply the fixed point theorem of Banach, because this would yield a unique fixed point u. Since R is bijective, the fixed point would also be a solution of (3.10). We will show now that B is a contraction. With the notation $M := I - \tau R\mathcal{A}$ it holds

$$\|B(u) - B(v)\|_H \;=\; \|M(u-v)\|_H \;\le\; \|M\|_{H;H}\|u-v\|_H.$$

Therefore, it is sufficient to show that $\|M\|_{H;H} < 1$. For $v \in H$ it holds

$$\|Mv\|_H^2 \;=\; \|v\|_H^2 - 2\tau(R\mathcal{A}v, v) + \tau^2\|R\mathcal{A}v\|_H^2.$$

Due to the ellipticity, the representation theorem of Riesz implies

$$(R\mathcal{A}v, v) \;=\; \langle \mathcal{A}v, v\rangle \;=\; A(v,v) \;\ge\; \alpha_2\|v\|_H^2.$$

The H-continuity (3.9) yields

$$\|R\mathcal{A}v\|_H^2 \;=\; \|\mathcal{A}v\|_H^2 \;\le\; \alpha_1^2\|v\|_H^2.$$

In summary we obtain that

$$\|Mv\|_H^2 \;\le\; (1 - 2\tau\alpha_2 + \tau^2\alpha_1^2)\|v\|_H^2 \qquad \forall v \in H,$$

or in other words

$$\|M\|_{H;H} \;\le\; \sqrt{1 - 2\tau\alpha_2 + \tau^2\alpha_1^2} \;<\; 1,$$

if $0 < \tau < 2\alpha_2/\alpha_1^2$.
(b) The mentioned upper bound of the estimate of $\|u\|_H$ is obtained by

$$\alpha_2\|u\|_H^2 \;\le\; A(u,u) \;=\; \langle f, u\rangle \;\le\; \|f\|_{H'}\|u\|_H,$$

which implies $\alpha_2\|u\|_H \le \|f\|_{H'}$. The lower bound holds due to

$$\|u\|_H \;\ge\; \alpha_1^{-1}\|\mathcal{A}u\|_{H'} \;=\; \alpha_1^{-1}\|f\|_{H'}.$$

□

Definition 3.23 *Let $T \in \mathcal{L}(E,F)$ be a bijective operator (i.e. $T : E \to F$ linear, continuous and bijective) between two Banach spaces E, F. The* condition number *is the quantity*

$$\kappa(T) \;:=\; \|T\|_{E;F}\|T^{-1}\|_{F;E}.$$

Remark: By the principle of open mappings we know that the bijectivity and continuity of T automatically implies the continuity of T^{-1}. In this case, we have $\|T^{-1}\|_{F;E} < \infty$.

Lemma 3.24 *Let E, F be two normed spaces and $T \in \mathcal{L}(E,F)$ bijective. Then for all $u, v \in E$, $u \neq 0$, and $f := Tu, g := Tv \in F$ we have the bound*

$$\frac{\|u-v\|_E}{\|u\|_E} \quad \leq \quad \kappa(T)\frac{\|f-g\|_F}{\|f\|_F}.$$

Proof. Easy exercise. ∎

A 'small' condition number implies that small changes in the right hand side only imply small changes in the solution. This is a certain form of stability with respect to the data. In the contrary, a large condition number $\kappa(T) \gg 1$ may lead to the situation that certain small changes in the right hand side (data) may lead to substantial differences in the solution of the corresponding problem.

Now we apply the definition of the condition number to our bilinear form $A : H \times H \to \mathbb{R}$:

Corollary 3.25 *Let H be a real Hilbert space and $A : H \times H \to \mathbb{R}$ a H-continuous and H-elliptic bilinear form with corresponding constants $\alpha_1 \geq \alpha_2 > 0$. Then the condition number of the induced operator $\mathcal{A} : H \to H'$ is bounded by*

$$\alpha_2/\alpha_1 \quad \leq \quad \kappa(\mathcal{A}) \quad \leq \quad \alpha_1/\alpha_2.$$

Proof. The following bounds are easy to verify:

$$\alpha_2 \quad \leq \quad \|\mathcal{A}\|_{H;H'} \quad \leq \quad \alpha_1.$$

The upper bound of $\|\mathcal{A}\|_{H;H'}$ was already show previously. This implies for the inverse operator:

$$\alpha_1^{-1} \quad \leq \quad \|\mathcal{A}^{-1}\|_{H;H'} \quad \leq \quad \alpha_2^{-1}.$$

Taking the product of these two bound yields the assertion. ∎

Now we will check the conditions of the theorem of Lax-Milgram for the case of our BVP (1.1). We will furthermore assume that

$$a_- := \inf_{x\in\Omega} \operatorname{ess}\, a(x) \quad > \quad 0.$$

This means that the coefficient $a(x)$ may only vanish or be negative on a null set of Ω.

(a) Firstly, we note that $V_0 = H_0^1(\Omega)$ is as a subspace of $H^1(\Omega)$ a pre-Hilbert space. Furthermore, V_0 is the kernel of the continuous trace operator γ, cf. (3.6), and therefore it is closed, hence V_0 it complete.

(b) The bilinear form A is V_0-continuous, because

$$\begin{aligned} |A(u,v)| &= \left| \int_\Omega a(x)\nabla u(x)\nabla v(x)\,dx \right| \\ &\leq \|a\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}. \end{aligned}$$

(c) The V_0-ellipticity is also easy to verify, because the inequality of Poincaré yields

$$\begin{aligned} A(u,u) &= \int_\Omega a(x)\nabla u(x)^2\,dx \\ &\geq a_- \|\nabla u\|^2_{L^2(\Omega)} \\ &\geq (1+C_p^2)^{-1} a_- \|u\|^2_{H^1(\Omega)}. \end{aligned}$$

Hence we obtain the ellipticity constant $\alpha_2 = (1+C_p^2)^{-1} a_- > 0$.

3.6 Neumann boundary conditions

Now, let us shortly consider the case that on one part of the boundary $\Gamma_N \subset \partial\Omega$ Neumann conditions and in the remaining part $\Gamma_D = \partial\Omega \setminus \Gamma_N$ Dirichlet values are imposed:

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial n} &= g_N \quad \text{on } \Gamma_N, \\ u &= g_D \quad \text{on } \Gamma_D, \end{aligned}$$

with given $g_N \in L^2(\Gamma_N)$ and $g_D \in L^2(\Gamma_D)$. After integration by parts it holds for $v \in V_{\Gamma_D,0} := \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}$

$$\begin{aligned} (-\Delta u, v)_{L^2(\Omega)} &= (\nabla u, \nabla v)_{L^2(\Omega)} - (\nabla u \cdot n, v)_{L^2(\partial\Omega)} \\ &= (\nabla u, \nabla v)_{L^2(\Omega)} - (\nabla u \cdot n, v)_{L^2(\Gamma_N)}. \end{aligned}$$

Hence, the variational formulation becomes

$$\widehat{u} \in V_{\Gamma_D,0}: \qquad (\nabla \widehat{u}, \nabla v)_{L^2(\Omega)} = \langle \widehat{f}, v \rangle \qquad \forall v \in V_{\Gamma_D,0},$$

with

$$\langle \widehat{f}, v \rangle := \langle f, v \rangle - (\nabla \widehat{g}_D, \nabla v)_{L^2(\Omega)} + (g_N, v)_{L^2(\Gamma_N)}.$$

Now $u := \widehat{u} + \widehat{g}_D$ is the corresponding solution. We note that the bilinear form $A(\cdot,\cdot)$ only includes the (homogenenous) Dirichlet conditions, while the inhomogeneous Dirichlet and Neumann data enter into the right hand side $\widehat{f}$.

Chapter 4

Finite Elements for the Poisson problem

4.1 Galerkin method

Let V be an infinite-dimensional Hilbert space, $f \in V'$ a linear functional and $A : V \times V \to \mathbb{R}$ a bilinear form. The principal idea of the *Galerkin*[1] *method* for solving a variational problem of the form

$$u \in V : \quad A(u, v) \;=\; \langle f, v \rangle \qquad \forall v \in V \tag{4.1}$$

consists in substituting the Hilbert space V by a finite-dimensional space $V_h \subset V$. Usually h stands for a parameter for describing a mesh size. The approximative discrete solution will be denoted by u_h:

$$u_h \in V_h : \quad A(u_h, v_h) \;=\; \langle f, v_h \rangle \qquad \forall v_h \in V_h. \tag{4.2}$$

A fundamental feature of such Galerkin methods is the so-called Galerkin orthogonality

$$A(u - u_h, v_h) \;=\; 0 \qquad \forall v_h \in V_h. \tag{4.3}$$

Essential is that only discrete test functions v_h in V_h can be used. By *finite elements* a certain type of such finite-dimensional spaces V_h are meant: Firstly, there should exist a basis of V_h with small support. Secondly, the functions should have a simple structure. This leads to the choice of piecewise polynomials. We firstly concentrate on finite elements consisting of piecewise linear functions which are the most simple class of finite elements.

Definition 4.1 *A* finite element method *consists in solving the variational problem (4.2) with a finite-dimensional space V_h, consisting of piece-wise polynomials. The method is called* conforming *if $V_h \subset V$. Otherwise, the method is called* non-conforming.

[1] Boris Grigoryevich Galyorkin, 1871-1945, soviet mathematician and engineer

Existence and uniqueness are given by the following corollary:

Corollary 4.2 *The variational problem (4.2) with a V-elliptic and V-continuous bilinear form A with conforming finite elements V_h has always a unique solution $u_h \in V_h$.*

Proof. Since finite-dimensional subspaces $V_h \subset V$ are always complete, V_h is a Hilbert-space, too. Due to the conformity, the V-ellipticity and V-continuity of the bilinear form A is directly transmitted to V_h. Furthermore, $f \in V' \subset V_h'$. Existence and uniqueness now follows by applications of the Theorem of Lax-Milgram (Thm. 3.22). ∎

4.2 Linear finite elements in 1D

We decompose the intervall $[0,1]$ in n subintervalls $I_k = [x_{k-1}, x_k)$ with $h_k := x_k - x_{k-1}$,

$$0 = x_0 < x_1 < x_2 < \ldots < x_n = 1.$$

We denote the space of (affine) linear functions on the interval I by $P_1(I)$. Our finite element space now consists of continuous and piece-wise linear functions:

$$P_{h,1} \quad := \quad \{v \in C[0,1] \,:\, v|_{I_k} \in P_1(I_k)\ \forall k = 1, \ldots, n\}.$$

The discrete space can be chosen as $V_h = P_{h,1}$. These finite elements are also called *Courant elements* . We already demonstrated the conformity $V_h \subset V = H^1(0,1)$. $v \in V_h$ is uniquely defined by its *nodal values* $v(x_k)$, $k = 0, \ldots, n$. Vice-versa, to arbitrary $n+1$ nodal values we can find a unique finite element function $v \in V_h$. It holds $\dim V_h = n+1$.

The finite element space V_{hg} with piece-wise linear polynomials corresponding to the affine space $V_g = g + H_0^1(0,1)$ is:

$$\begin{aligned} V_{hg} \quad &:= \quad P_{h,1} \cap V_g \\ &= \quad \{v \in C[0,1] \,:\, v(x_0) = g_0,\ v(x_n) = g_1\,,\ v|_{I_k} \in P_1(I_k)\ \forall k = 1, \ldots, n\}. \end{aligned}$$

For the space V_h different choices of a *basis* is possible. By the *Lagrange*[2] *basis* we understand the "hat functions", which take the values 1 and 0 on the nodes x_j. The "inner" basis functions are

$$\phi_k(x) \quad = \quad X_{I_k}(x)\frac{x - x_{k-1}}{x_k - x_{k-1}} + X_{I_{k+1}}(x)\frac{x_{k+1} - x}{x_{k+1} - x_k} \quad k = 1, \ldots, n-1,$$

and those on the boundary

$$\phi_0(x) \quad = \quad X_{I_1}(x)\frac{x_1 - x}{x_1 - x_0} \quad \text{and} \quad \phi_n(x) \ = \ X_{I_n}(x)\frac{x - x_{n-1}}{x_n - x_{n-1}}.$$

[2] Joseph-Louis de Lagrange, 1736-1813, italian mathematician and astronom.

Here, X_J denotes the characteristic function of the interval J. The support of these functions are given by

$$\operatorname{supp}\phi_k \;=\; I_k \cup I_{k+1}, \qquad k = 1, \ldots, n-1,$$

and $\operatorname{supp}\phi_0 = I_1$, $\operatorname{supp}\phi_n = I_n$. We can express u_h in terms of this basis by

$$u_h(x) \;=\; \sum_{k=0}^{n} u_k \phi_k(x).$$

The coefficients u_k are just the nodal values $u_k = u_h(x_k)$. The discrete variational problem (3.2) with homogeneous Dirichlet values can be formulated in the form

$$\sum_{j=1}^{n-1} A(\phi_j, \phi_i) u_j \;=\; \langle f, \phi_i \rangle \qquad \forall i = 1, \ldots, n-1,$$

because the nodal values on the beginning and the end of the interval are set to zero, $u_0 = u_n = 0$. This corresponds to the linear equation system

$$AU \;=\; b,$$

with the nodal vector $U = (u_1, \ldots, u_{n-1})^T$, right hand side $b = (b_1, \ldots, b_{n-1})^T$, $b_k = \langle f, \phi_k \rangle$ and the *stiffness matrix* $A = (a_{ij})_{i,j=1,\ldots,n-1}$.

4.2.1 Stiffness matrix

The coefficients of the stiffness matrix are

$$a_{ij} \;=\; A(\phi_j, \phi_i).$$

Since the supports of the basis functions ϕ_k are local, the stiffness matrix is sparse. In particular it holds for $|i-j| > 1$ in the case of the bilinear form (3.7):

$$a_{ij} \;=\; \int_0^1 \phi_j'(x)\phi_i'(x)\,dx \;=\; 0.$$

For piece-wise linear finite elements the derivatives are piece-wise constant:

$$\phi_i'|_{I_i} \;=\; h_i^{-1}, \quad \phi_i'|_{I_{i+1}} \;=\; -h_{i+1}^{-1}.$$

For the diagonal entries $1 \le i \le n-1$ we obtain

$$\begin{aligned} a_{ii} \;&=\; \int_0^1 \phi_i'(x)\phi_i'(x)\,dx \;=\; \int_{x_{i-1}}^{x_i} \phi_i'(x)\phi_i'(x)\,dx + \int_{x_i}^{x_{i+1}} \phi_i'(x)\phi_i'(x)\,dx \\ &=\; h_i h_i^{-2} + h_{i+1}(-h_{i+1})^{-2} \;=\; h_i^{-1} + h_{i+1}^{-1}. \end{aligned}$$

The off-diagonal entries become

$$\begin{aligned} a_{i,i+1} &= \int_0^1 \phi_i'(x)\phi_{i+1}'(x)\,dx = \int_{x_i}^{x_{i+1}} \phi_i'(x)\phi_{i+1}'(x)\,dx \\ &= h_{i+1}(-h_{i+1}^{-1}h_{i+1}^{-1}) = -h_{i+1}^{-1}. \end{aligned}$$

Due to symmetry it holds in this particular case $a_{i+1,i} = a_{i,i+1} = -h_{i+1}$. In summary we obtain the following $(n+1)\times(n+1)$ matrix

$$A = \begin{pmatrix} h_1^{-1} & -h_1^{-1} & 0 & \cdots & \cdots \\ -h_1^{-1} & h_1^{-1}+h_2^{-1} & -h_2^{-1} & 0 & \cdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \vdots \\ 0 & \cdots & & -h_n^{-1} & h_n^{-1} \end{pmatrix}.$$

In the case of equidistant subintervalls $h = h_1 = \ldots = h_n$ we obtain

$$A = \frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & & 2 & -1 \\ 0 & \cdots & \cdots & -1 & 1 \end{pmatrix}.$$

Note that this matrix does not yet contain the (Dirichlet-) boundary conditions. In comparison, the finite-difference matrix differs only by the scaling with the mesh size h,

$$A_{FEM} = hA_{FDM}.$$

4.2.2 Right-hand side

The right-hand side $b = (b_1, \ldots, b_{n-1})$ consists of the components

$$\begin{aligned} b_k &= \int_0^1 f(x)\phi_k(x)\,dx \\ &= h_k^{-1}\int_{x_{k-1}}^{x_k} f(x)(x - x_{k-1})\,dx + h_{k+1}^{-1}\int_{x_k}^{x_{k+1}} f(x)(x_{k+1} - x)\,dx. \end{aligned}$$

For general f these integrals can not be integrated exactly. We have to use a numerical quadrature formula. Here, one has to guarantee a sufficient accuracy. We use here as an example the trapezoidal rule which is exact for linear functions. Since ϕ_k vanishes on the integration points x_{k-1} and x_{k+1}, and it takes the value 1 at the point x_k, the

approximation reduces for $k = 1, \ldots, n-1$ to

$$\begin{aligned} b_k &\approx h_k^{-1}\frac{1}{2}h_k(0 + f(x_k)(x_k - x_{k-1}) + h_{k+1}^{-1}\frac{1}{2}h_{k+1}(f(x_k)(x_{k+1} - x_k) + 0) \\ &= \frac{1}{2}f(x_k)\,(h_k + h_{k+1})\,. \end{aligned}$$

In the case of equidistant mesh width h we obtain

$$b_k \approx hf(x_k) = hb_k^{FDM}.$$

Hence, the right-hand side also corresponds to the one for a finite-difference (FD) scheme scaled by h. In summary we obtain on equidistant meshes with linear elements the same solution as for a FD scheme, as long as we integrate the right-hand side by the trapezoidal rule. However, we have the possibility to use other (more accurate) quadrature formulas. This can be useful for highly fluctuating f.

4.2.3 Mass matrix

Beside of the stiffness matrix A, another matrix is important. The so-called *mass matrix* assembles the zero order terms (e.g. appearing in the equation $-u'' + u = f$):

$$M = (m_{ij})_{i,j=1,\ldots,n}\,, \quad m_{ij} = (\phi_j, \phi_i).$$

The mass matrix is always symmetric. In the case of linear (P_1) elements, this matrix becomes:

$$M_h = \frac{1}{6}\begin{pmatrix} 2(h_1+h_2) & h_2 & & & & \\ h_2 & 2(h_2+h_3) & h_3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_n & \\ & & & h_{n-1} & 2h_n & \end{pmatrix}.$$

In the case of equidistant meshes, it reduces to the following tridiagonal matrix

$$M_h = \frac{h}{6}\,\mathrm{tridiag}(1, 4, 1).$$

4.3 P_r-elements in 1D

In the previous section we considered linear finite elements. If we use piecewise polynomials of higher order, we expect to enhance the accuracy of the method. This leads to P_r-elements, where $r \geq 1$ is the lokal degree of the polynoms:

$$P_{h,r} := \{v \in C[0,1] : v|_{I_k} \in P_r(I_k)\ \forall k = 1, \ldots, n\}\,.$$

We will have a look onto the quadratic elements, $r = 2$, an their Lagrange basis. Normalized on the interval $[0, 1]$ the three basis functions are (cf. Fig. 4.1):

$$\begin{aligned} \phi_1(x) &= 2\left(x - \frac{1}{2}\right)(x-1), \\ \phi_2(x) &= -4x(x-1), \\ \phi_3(x) &= 2x\left(x - \frac{1}{2}\right). \end{aligned}$$

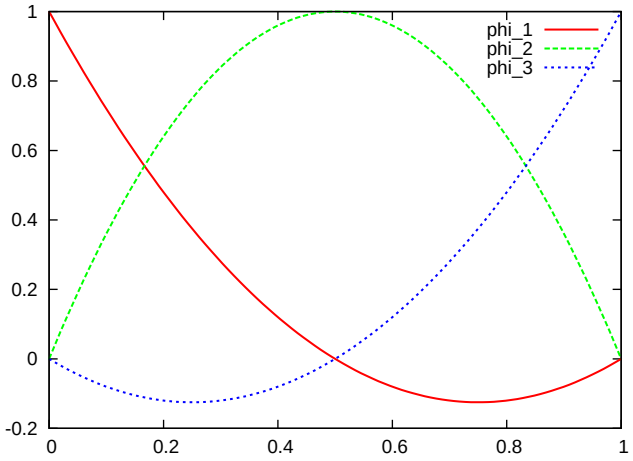


Figure 4.1: Lagrange basis of quadratic finite elements P_2 normalized to the unit interval.

4.4 C^0-finite Elements in 2D

For the Poisson problem in more than one dimension, $d \geq 2$, we need a triangulation of the domain $\Omega \subset \mathbb{R}^d$. A grid (or triangulation) $\mathcal{T} = \{T_1, \ldots, T_m\}$ of Ω consists of elements T_k, such that

$$\overline{\Omega} = \bigcup_{k=1}^{m} T_k.$$

The elements are closed sets so that the edges of the elements are parts of the element. For two-dimensionalen domains, $d = 2$, we distinguish between triangular and quadrilateral elements.

Definition 4.3 *A grid (or triangulation) $\mathcal{T} = \{T_1, \ldots, T_m\}$ of the domain $\Omega \subset \mathbb{R}^2$ consisting of triangular and quadrilateral elements is called* admissible, *if the intersection of two elements $T_j \cap T_j$, with $i \neq j$, is the empty set, consists only on one grid point of T_j and T_i, or consists of a common edge.*

For quadrilateral meshes the shape regularity is often relaxed in order to allow for mesh refinement. In this case, so called *hanging nodes* appear. If the shape regularity should be maintained for quadrilateral meshes, local mesh refinement should allow for mixing triangles and quadrilaterals, see Fig. 4.2.

Lemma 4.4 *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded domain and $r \in \mathbb{N}$. Then the P_r- and the Q_r-elements on admissible triangulations $\mathcal{T}$ are H^1-conforming.*

Proof. In section 2.4 we have shown that continuous, piece-wise polynomials are in $H^1(\Omega)$. ∎

4.4.1 Polynomials on triangular meshes

An essential feature of finite elements is that the ansatz and test functions are polynomial on the elements. On triangular meshes, P_r-elements are piecewise polynomials of degree $\leq r$:

$$P_r \quad := \quad \left\{ u : \mathbb{R}^2 \to \mathbb{R} \,|\, u(x,y) = \sum_{\substack{0 \leq i+j \leq r \\ 0 \leq i,\, j}} a_{ij} x^i y^j \right\}.$$

In the case Fall $r = 1$ we call them linear elements or Courant elements. In this case we obtain on each triangle three degrees of freedom:

$$u(x,y) \quad = \quad a_0 + a_1 x + a_2 y.$$

These can be identified with the nodes of the triangle. In the case of quadratic elements, $r = 2$, we have on each element six degrees of freedom:

$$u(x,y) \quad = \quad a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2.$$

The corresponding geometrical identities are the three nodal points and the three edges of the triangular element.

On triangular meshes $\mathcal{T}$ the corresponding finite elements space is

$$P_r(\mathcal{T}) \quad := \quad \left\{ u \in C(\overline{\Omega}) \,:\, u|_T \in P_r \quad \forall T \in \mathcal{T} \right\}. \tag{4.4}$$

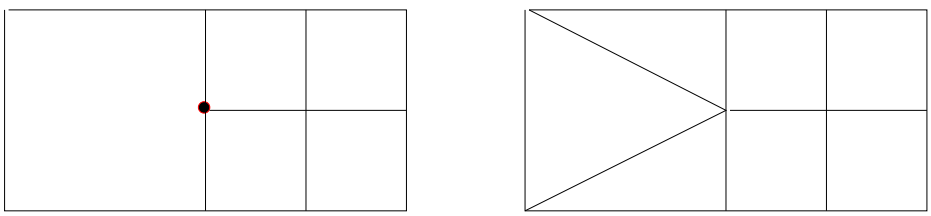


Figure 4.2: A hanging node for a quadrilateral mesh (left) and an admissible hybrid mesh (right) with triangles and quadrilaterals to compensate such hanging nodes.

4.4.2 Polynomials on tetrahedral meshes in 3D

The extension of triangles to the case $d = 3$ are *tetrahedrons*. The corresponding polynomials are of the form

$$u(x,y,z) = \sum_{\substack{0 \le i+j+k \le r \\ i,j,k \ge 0}} a_{ijk} x^i y^j z^k.$$

4.4.3 Polynomials on quadrilateral meshes

On quadrilateral we often use polynomials of partial degree $\le r$. These are the so-called Q_r-elements

$$Q_r \;:=\; \left\{ u(x,y) = \sum_{0 \le i,j \le r} a_{ij} x^i y^j \right\}.$$

In the case $r = 1$ we obtain *bilinear elements* with four degrees of freedom,

$$u(x,y) \;=\; a_0 + a_1 x + a_2 y + a_3 xy.$$

according to the four nodes of the element. For $r = 2$ we have *bi-quadratic* elements with nine degrees of freedom:

$$\begin{aligned} u(x,y) \;=\; & a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2 \\ & + a_6 x^2 y + a_7 x y^2 + a_8 x^2 y^2. \end{aligned}$$

We obtain on quadrilateral meshes $\mathcal{T}$ the following finite element space:

$$Q_r(\mathcal{T}) \;:=\; \left\{ u \in C(\overline{\Omega}) \,:\, u|_T \in Q_r \quad \forall T \in \mathcal{T} \right\}. \tag{4.5}$$

The extension to three dimensions (3D) are *hexahedrons*.

4.5 Finite element basis

4.5.1 Unisolvence

For a representation of the solution $u_h \in V_h$ and for assembling the stiffness matrix a basis of V_h is required. The most natural choice is the so-called Lagrange basis (or nodal basis) . In order to introduce them we define so-called nodal functionals $\chi \in P_r'$. These are continuous linear functionals of the form

$$\langle \chi, p \rangle \;=\; p(N_0) \quad \forall p \in P_r$$

with fixed $N_0 \in \mathbb{R}^d$. The idea is to define a suitable choice of such nodal functionals. Afterwards we determine for each nodal functional a corresponding basis function.

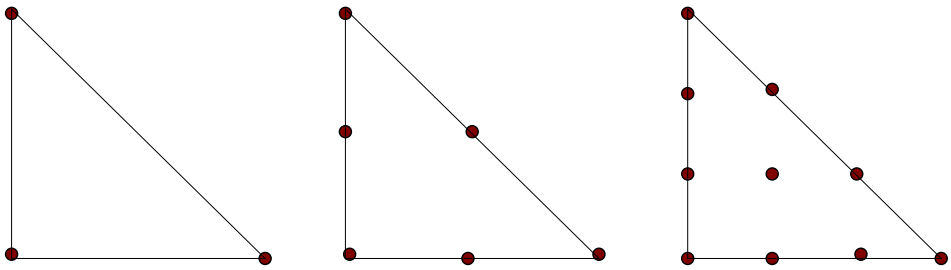


Figure 4.3: Suitable collocation points for nodal functionals leading to unisolvent polynomials for P_1, P_2 and P_3 (left to right).

Definition 4.5 *The space P_r of polynomials of degree r in combination of a set of nodal functionals $\{\chi_1, \ldots, \chi_s\} \subset P_r'$ is called* unisolvent*, if each $p \in P_r$ is uniquely defined by the values of these nodal functionals $\langle \chi_i, p \rangle$, $i = 1, \ldots, s$.*

4.5.2 Basis on triangular meshes

Lemma 4.6 *On the triangle T we consider the $r+1$ lines and $s := 1 + \ldots + (r+1)$ collocation points $N_1, \ldots, N_s$ as depicted in Fig. 4.3. Then P_r in combination with the corresponding s nodal functionals $\langle \chi_i, v \rangle = v(N_i)$ is unisolvent.*

Proof. Let $w_1, \ldots, w_s \in \mathbb{R}$ be given. We have to show that there exists only one polynomial $p \in P_r$ s.t.

$$p(N_i) \quad = \quad w_i \qquad \text{for } i = 1, \ldots, s.$$

We show this by induction over r. For $r = 0$ there is only one node ($s = 1$). Polynomials of degree $r = 0$ are constant, so that one function value determines the polynomial uniquely. Hence, the assertion is true for $r = 0$. We assume that the assertion is true for $r - 1$. The induction step is as follows:
(a) Due to the invariance with respect to affine linear transformations we can assume that the points $N_1, \ldots, N_{r+1}$ are collocated on the x-axis; $N_j = (x_j, 0)$. It exists a unique one-dimensional polynomial q of degree r with

$$q_1(x_i) \quad = \quad w_i \qquad \forall i \in \{1, \ldots, r+1\}.$$

According to the induction hypothesis it exists a unique polynomial $q_2 \in P_{r-1}$ with

$$q_2(N_j) \quad = \quad (w_j - q_1(x_j))/y_j \qquad \forall j \in \{r+2, \ldots, s\}.$$

Now we choose

$$p(x, y) \quad = \quad q_1(x) + y q_2(x, y).$$

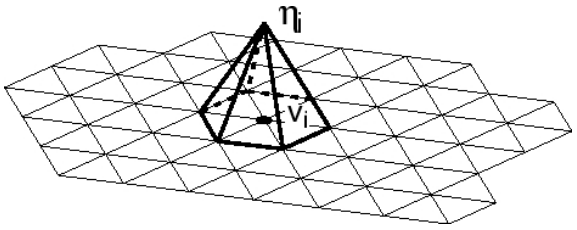


Figure 4.4: Lagrange basis function of P_1 finite elements at the node N_i.

Obviously holds for $i = 1, \ldots, r+1$: $p(N_i) = q_1(x_i) = w_i$ and $p(N_j) = q_1(x_j) + y_j q_2(x_j, y_j) = w_j$ for $j = r+2, \ldots, s$.
(b) For showing the uniqueness we set $w_1 = \ldots = w_s = 0$. Let $p \in P_r$ a corresponding interpolation polynomial. It can be written in the form

$$\begin{aligned} p(x,y) &= p_1(x,y) + \sum_{i+j=r} a_{ij} x^i y^j \\ &= p_1(x,y) + a_{r0} x^r + y \sum_{i+j=r, j\geq 1} a_{ij} x^i y^{j-1}, \end{aligned}$$

with a polynomial p_1 of degree $r-1$ and coefficients $a_{ij} \in \mathbb{R}$. The interpolation property on the points $N_1, \ldots, N_{r+1}$ (collocated on the x-axis) require

$$0 = w_j = p(x_j, 0) = p_1(x_j, 0) + b x_j^r.$$

The right hand side defines a polynomial in $\mathbb{R}$ of maximal degree r:

$$q(x) := p_1(x,0) + a_{r0} x^r.$$

The uniqueness of polynomial interpolation on the $r+1$ points, $q(x_1) = \ldots = q(x_{r+1}) = 0$ implies $q \equiv 0$. This yields to $p_1(\cdot, 0) \equiv 0$, $a_{r0} = 0$. This means that $p_1(x,y)$ has for $y = 0$ a root, so that the factor y can be factorized out: $p_1(x,y) = y q_1(x,y)$. This leads to

$$\begin{aligned} p(x,y) &= y q_1(x,y) + y \sum_{i+j=r, j\geq 1} a_{ij} x^i y^{j-1} \\ &= y q_2(x,y), \end{aligned}$$

with $q_2 \in P_{r-1}$. The interpolation property on the remaining points N_j for $j = r+2, \ldots, s$ reads

$$0 = p(N_j) = y_j q_2(x_j, y_j).$$

Since $y_j \neq 0$ for these j, q_2 should have at least $s - (r+1) = 1 + \ldots + r$ roots. Due to the induction hypothesis, we follow that $q_2 \equiv 0$, which finally leads to $p \equiv 0$. ∎

The Lagrange finite element basis consists of functions ϕ_i to the nodal functionals χ_i. These are uniquely defined by the Kronecker symbols

$$\phi_i(N_j) \quad = \quad \delta_{ij}.$$

For P_1-elements we depict such a basis function in Fig. 4.4. The uniqueness is a direct implication of the unisolvence. Without the consideration of the boundary conditions, the size of this P_1 basis is on triangular meshes equal to the number of mesh nodes. For quadratic elements (P_2) we have additional basis functions for each element edge.

4.5.3 Basis on quadrilateral meshes

On quadrilaterals the choice of the nodal functional is even simpler, because we can take advantage of a tensor-product structure. The corresponding nodes on a square are just collocated in equidistant fashion.

For admissible quadrilateral meshes (without hanging nodes) we count the degrees of freedom as follows: the dimension of V_h for Q_1 elements is equal to the number of mesh nodes. For bi-quadratic elements (Q_2) the number of degrees of freedom n corresponds to the sum of number of mesh nodes, number of element edges and number of elements. For $r = 3$ the elements on quadrilaterals are called *bicubic* elements. Choosing even higher order finite elements are usually only meaningful in some particular cases, because the advantage of better approximation properties require that the exact solution is very smooth, i.e. is element of a Sobolev space $H^m(\Omega)$ with large exponent m.

4.6 Transformation and geometrical properties

4.6.1 Transformation of triangular meshes

To describe a triangle in arbitrary orientation, rotation and scaling we use affine linear transformations. We denote such an affine linear transformation from a reference triangle $\widehat{T}$ with nodes $\xi_1 = (0,0)$, $\xi_2 = (1,0)$ and $\xi_3 = (0,1)$ onto an element $T \in \mathcal{T}$ (see Fig. 4.5) by

$$\Phi_T : \widehat{T} \to T\,,\ \widehat{x} \mapsto \Phi_T(\widehat{x}) = x.$$

For triangles such an affine linear transformation is sufficient to obtain an arbitrary triangle. Such transformation is also an element in P_1^2. It is important to verify that such transformation applied to a polynomial $\widehat{u}$ of degree $\leq r$ results again to a polynomial $u = \widehat{u} \circ \Phi_T^{-1}$ of degree $\leq r$ ($r \geq 1$). Therefore, is the requirement $u|_T \in P_r$ equivalent to $(u \circ \Phi_T)|_{\hat{T}} \in P_r$: P_r is invariant under affin-linear transformation.

Lemma 4.7 *A finite element function $u_h \in P_r(\mathcal{T})$ restricted to an edge of a triangular element $T \in \mathcal{T}$ can be expressed as a one-dimensional polynomial of degree r.*

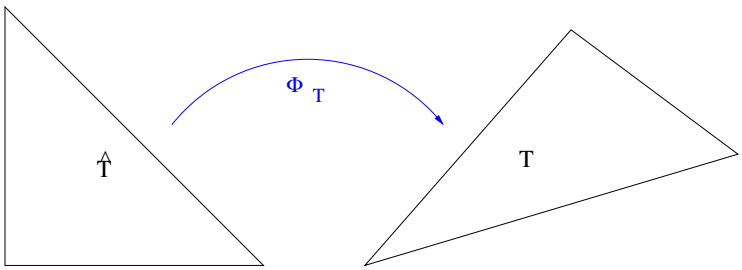


Figure 4.5: Transformation Φ_T from reference element $\widehat{T}$ to the element T.

Proof. At first we state that the P_r-Lagrange basis polynomials on the reference element $\hat{T}$ restricted to each of the three edges of $\hat{T}$ can be written as a polynomial of degree r in one variable. Let now $u_h \in P_r(\mathcal{T})$ be given. The representation on the reference element is given by $\hat{u}_h := u_h \circ \Phi_T$ and is a polynomial of degree r. Let $\hat{e} = \Phi_T^{-1} e$ be an edge of the reference triangle e of T. Then $\hat{u}_h|_{\hat{e}}$ is a one-dimensional polynomial of degree r, hence for $(\hat{x}, \hat{y}) \in \hat{e}$:

$$\hat{u}(\hat{x}, \hat{y}) \quad = \quad q(\xi)$$

with $\xi = \xi(\hat{x}, \hat{y}) = \alpha\hat{x} + \beta\hat{y}$ and $q \in P_r$. This means $\xi \in P_1$ and for $(x, y) \in e$ it holds:

$$\begin{aligned} u_h(x,y) &= (\hat{u}_h \circ \Phi_T^{-1})(x,y) \\ &= (q \circ \xi \circ \Phi_T^{-1})(x,y) \\ &= q((\xi \circ \Phi_T^{-1})(x,y)). \end{aligned}$$

Therefore, the function u_h on the edge e is a polynomial of degree r in the one-dimensional variable $(\xi \circ \Phi_T^{-1})(x, y)$. ∎

Corollary 4.8 *P_r-elements are globally continuous.*

Proof. We still have to check whether the functions of the form

$$u_h(x,y) \quad = \quad \sum_{i=1}^{n} \phi_i(x,y) u_i$$

are globally continuous. Each element edge is assigned to $r+1$ nodal functionals. The restriction of u_h onto one edge corresponds after coordinate transformation a one-dimensional polynomial of degree $\leq r$ (see later) and is therefore unique. The $r+1$ nodal functionals are interpolated by the finite element functions from both adjacent elements. Hence, the two polynomials are identical on the edge. This yields to $u_h \in C(\overline{\Omega})$. ∎

¿

4.6.2 Transformation of quadrilateral meshes

For quadrilateral meshes the transformation from the reference element (here denoted by $\widehat{K}$) can be realized in different ways. If we only allow for linear transformations

$$\Phi_K : \widehat{K} \to K$$

i.e. $\Phi_K \in P_1^2$, then only parallelograms are obtained for K. In the case of bilinear transformation, i.e. $\Phi_K \in Q_1^2$ for $r = 1$, we may obtain arbitrary distorted quadrilateral elements with straight edges. For curved edges we have to allow for transformations $\Phi_K \in Q_2^d$. We speak about *iso-parametric* transformations, if the transformation and the finite element function on the reference element are of the same type.

For a triangular element T we denote the outer radius by h_T, and by ϱ_T the innen radius, cf. Fig. 4.6. The quotient of the outer and inner radii is a measure of anisotropy of an element:

$$\kappa_T \quad = \quad \frac{h_T}{\rho_T}.$$

For triangular meshes, large values of κ_T lead to acute angles. This has an influence onto the approximation properties. On quadrilateral meshes, big anisotropies may not necessarily lead to acute angles, but possibly to stretched rectangles.

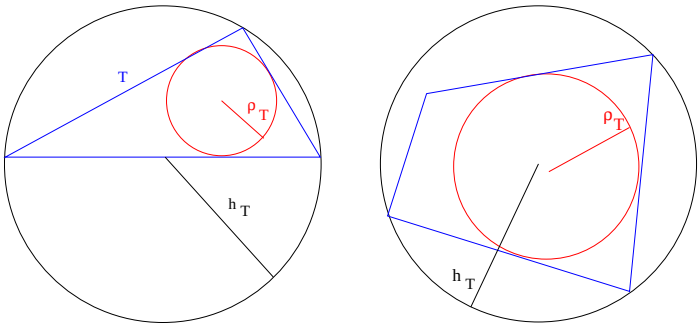


Figure 4.6: Inner radius ρ_T and outer radius h_T of a triangular elements (left) and a quadrilateral element (right).

Chapter 5

A priori error estimates

In this section we will compare the discretization error between the exact solution $u \in V$ and the discrete solution $u_h \in V_h$ in comparison to the so-called *approximation error*

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

The approximation error is the smallest error we can hope for. We always assume $h \leq 1$, because one is usually interested in small mesh sizes.

5.1 Cea's lemma

The following lemma ensures that we obtain with the Galerkin method this approximation error up to a constant, as long as the bilinear form is V-continuous and V-elliptic.

Lemma 5.1 (Cea's Lemma) *Let the bilinear form $A : V \times V \to \mathbb{R}$ satisfy the conditions of the Theorem of Lax-Milgram (V-continuous and V-elliptic with constants $\alpha_1 > 0$ and $\alpha_2 > 0$, respectively). Further let $V_h \subset V$ be a subspace. Then it holds for the discretization error:*

$$\|u - u_h\|_V \quad \leq \quad \frac{\alpha_1}{\alpha_2} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Usually the space V_h is finite dimensional. Therefore, we speak here about the discrete solution u_h and the discretization error.

Proof. Due to ellipticity it holds

$$\alpha_2 \|u - u_h\|_V^2 \quad \leq \quad A(u - u_h, u - u_h).$$

Now, we use the Galerkin orthogonality (4.3)

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

For arbitrary $v_h \in V_h$ we can choose $w_h := u_h - v_h \in V_h$ and obtain due to the linearity of $A(\cdot,\cdot)$ in the second argument

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

Dividing both sides by $\alpha_2\|u-u_h\|_V$ yields the assertion. ∎

The approximation error is usually difficult to determine. However, an upper bound is obtained by considering the *interpolation error* $\|u-P_hu\|_V$, with the interpolation operator

$$P_h : V \to V_h.$$

It holds obviously

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

As we will discuss later, several interpolations are possible. The most simple one is the so-called nodal interpoland.

Theorem 5.2 *Under the same assumptions on the bilinear form $A : V \times V \to \mathbb{R}$ as in Cea's Lemma (5.1) with $V = H^1(\Omega)$ and the additional smoothness assumption on the exact solution $u \in H^2(\Omega)$, it holds for the discretization error with linear finite elements (P_1)*

$$\|u-u_h\|_{H^1(\Omega)} \leq Ch|u|_{H^2(\Omega)}, \tag{5.1}$$

where $h := \max\{h_T : T \in \mathcal{T}_h\}$. The constant depends on the continuity and ellipticity constants α_1 and α_2, respectively, in the form $C \sim \alpha_1/\alpha_2$.

Proof. According to Cea's Lemma it is sufficent to show the existence of $v_h \in V_h$ with

$$|u-v_h|_{H^1(\Omega)} \leq C_I h|u|_{H^2(\Omega)}.$$

The existence of such v_h is given by the so-called nodal interpolation $v_h = I_hu$ which satisfies for $u \in H^2(\Omega)$ the following:

$$\|u-I_hu\|_{L^2(\Omega)} + h|u-I_hu|_{H^1(\Omega)} \leq C_I h^2 |u|_{H^2(\Omega)}.$$

This property will be proved in the section 5.3. ∎

5.2 A priori error estimate for symmetric bilinear forms

If A is symmetric, A generates a (perhaps different) scalar product by

$$(u,v)_A \quad := \quad A(u,v).$$

Theorem 5.3 *Let the bilinear form $A : V \times V \to \mathbb{R}$ and the solution u fulfill the same assumptions as in Theorem 5.1, and additional let A be symmetric. Then it holds :*

$$\|u-u_h\|_V \quad \leq \quad \sqrt{\frac{\alpha_1}{\alpha_2}} \inf_{v_h \in V_h} \|u-v_h\|_V.$$

Proof. Now we show that the discretization error and the approximation error are identical, i.e.

$$\|u-u_h\|_A \quad = \quad \inf_{v_h \in V_h} \|u-v_h\|_A.$$

This equation follows due to the Galerkin orthogonality and the following estimate:

$$\begin{aligned} \|u-u_h\|_A^2 \quad &= \quad A(u-u_h, u-u_h) \;=\; A(u-u_h, u-v_h) \\ &\leq \quad \|u-u_h\|_A \|u-v_h\|_A, \end{aligned}$$

and therefore for arbitrary $v_h \in V_h$:

$$\|u-u_h\|_A \quad \leq \quad \|u-v_h\|_A.$$

Taking the infimum over all $v_h \in V_h$ on the right hand side and the trivial bound in the other direction yields the equality of discretization error and approximation error. Denoting the continuity constant by $\alpha_1 > 0$ and the ellipticity constant by $\alpha_2 > 0$ leads to:

$$\begin{aligned} \alpha_2 \|u-u_h\|_V^2 \quad &\leq \quad A(u-u_h, u-u_h) \;=\; \|u-u_h\|_A^2 \;=\; \inf_{v_h \in V_h} \|u-v_h\|_A^2 \\ &= \quad \inf_{v_h \in V_h} A(u-v_h, u-v_h) \;\leq\; \alpha_1 \inf_{v_h \in V_h} \|u-v_h\|_V^2. \end{aligned}$$

This implies the assertion. ∎

5.3 Bramble-Hilbert lemma

Theorem 5.4 (Embedding theorem of Sobolev) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with an inner cone condition and let $k, m \in \mathbb{N}_0$ with $m < k - d/2$. Then the embedding $H^k(\Omega) \subseteq C^m(\Omega)$ is continuous. In particular holds for all multi-indices $|\alpha| \leq m$:*

$$\|D^\alpha u\|_{L^\infty(\Omega)} \quad \leq \quad C \|u\|_{H^k(\Omega)} \qquad \forall u \in H^k(\Omega).$$

The embedding should be understood in terms of a corresponding C^m-representative.

Proof. ∎

Lemma 5.5 *Let $T \subset \mathbb{R}^d$ be a Lipschitz domain and $k \in \mathbb{N}$ with $k > d/2$. Each function $u \in H^k(T)$ with $|u|_{H^k(T)} = 0$, is a.e. identical to a polynomial in $P_{k-1}(T)$.*

Proof. The embedding theorem of Sobolev (Thm. 5.4) ensures that $u \in C^m(T)$ for $m \in \mathbb{N}_0$ with $0 \leq m < k - d/2$. The assumption $|u|_{H^k(T)} = 0$ furthermore implies that $\|D^\alpha u\|_{L^2(T)} = 0$ for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \geq k$. Hence, $D^\alpha u \in L^2(T)$ and

$$u \quad \in \quad \bigcap_{j=0}^{\infty} H^j(T).$$

The embedding theorem of Sobolev now ensures that actually $u \in C^\infty(T)$. The proof that u is a polynomial can be obtained by 'classical' arguments: Obviously $D^\alpha u \equiv 0$ if $|\alpha| \geq k$. This implies $D^\beta u \in P_0(T)$ if $|\beta| \geq k - 1$. By induction we arrive at $u \in P_{k-1}(T)$. ∎

Lemma 5.6 *Let $T \subset \mathbb{R}^d$ be a Lipschitz domain and $m \in \mathbb{N}$, $m > d/2$. Furthermore, we consider s pairwise different points $z_1, \ldots, z_s \in \overline{T}$ in such a way that the interpolation to polynomials of degree $m-1$ are uniquely defined. Then the $H^m(T)$-norm is equivalent to the norm*

$$|\!|\!| u |\!|\!| \quad := \quad |u|_{H^m(T)} + \sum_{j=1}^{s} |u(z_j)|. \tag{5.2}$$

Proof. (a) The norm $|\!|\!| \cdot |\!|\!|$ can be bounded from above by the $H^m(T)$-norm: According to the embedding theorem of Sobolev (Thm. 5.4), the embedding $H^m(T) \subset C(T)$ is continuous. Hence

$$|u(z_j)| \quad \leq \quad \|u\|_{L^\infty(T)} \leq c\|u\|_{H^m(T)} \qquad \forall j \in \{1, \ldots, s\}.$$

(b) Now we demonstrate that the $H^m(T)$-norm can be bounded by the $|\!|\!| \cdot |\!|\!|$-norm. Assuming that such a bound does not hold. Then there exists a sequence $(u_n)_{n\in\mathbb{N}}$ in $H^m(T)$ with $\|u_n\|_{H^m(T)} = 1$ and

$$\lim_{n\to\infty} |\!|\!| u_n |\!|\!| \quad = \quad 0\,.$$

This sequence is bounded in $H^m(T)$ and due to the embedding theorem of Rellich (Thm. 3.14), the embedding $H^m(T)$ in $H^{m-1}(T)$ is compact. Therefore, the sequence should have a convergent subsequence $(u_{n_k})_{k\in\mathbb{N}}$ which converges strongly in $H^{m-1}(T)$. We denote this limit by $u \in H^{m-1}(T)$. This subsequence is also a Cauchy sequence in $H^m(T)$, because

$$\begin{aligned} \|u_{n_k} - u_{n_l}\|^2_{H^m(T)} &\leq \|u_{n_k} - u_{n_l}\|^2_{H^{m-1}(T)} + |u_{n_k} - u_{n_l}|^2_{H^m(T)} \\ &\leq \|u_{n_k} - u_{n_l}\|^2_{H^{m-1}(T)} + |\!|\!| u_{n_k} - u_{n_l} |\!|\!|^2 \\ &\to 0 \qquad (n_k, n_l \to 0). \end{aligned}$$

This implies $u_{n_k} \to u$ in $H^m(T)$ and

$$\|u\|_{H^m(T)} \;=\; \lim_{k\to\infty} \|u_{n_k}\|_{H^m(T)} \;=\; 1\,.$$

On the other hand we have

$$|u|_{H^m(T)} \;\le\; |\!|\!|u|\!|\!| \;=\; \lim_{k\to\infty} |\!|\!|u_{n_k}|\!|\!| \;=\; 0\,.$$

Here we used the fact that $|\!|\!|\cdot|\!|\!|$ is continuous with respect to $H^m(T)$-norm which is an implication of part (a) of this proof. Using the previous Lemma yields that u is in T a polynomial of degree $m-1$, i.e. $u \in P_{m-1}(T)$. Due to $|\!|\!|u|\!|\!| = 0$ we have in particular $u(z_j) = 0$ for all $1 \le j \le s$. The uniqueness of the nodal interpolation with respect to these s points implies that $u \equiv 0$. This is a contradiction to $\|u\|_{H^m(T)} = 1$. □

Lemma 5.7 *Let $T \subset \mathbb{R}^2$ be a Lipschitz domain, $m \ge 2$ and $I^{m-1} : H^m(T) \to P_{m-1}$ the nodal interpolation which is well-defined by the nodal values on s pairwise different points $z_1, \ldots, z_s \in \overline{T}$. Then there exists a constant $c = c(T, z_1, \ldots, z_s)$ s.t.*

$$\|u - I^{m-1}u\|_{H^m(T)} \;\le\; c|u|_{H^m(T)} \qquad \forall u \in H^m(T).$$

Proof. According to the previous lemma, the $H^m(T)$-norm is equivalent to the $|\!|\!|\cdot|\!|\!|$-norm defined in (5.2). Due to the property $u(z_j) = Iu(z_j)$ for $j = 1, \ldots, s$ we deduce for arbitrary $u \in H^m(T)$:

$$\begin{aligned} \|u - I^{m-1}u\|_{H^m(T)} &\le c|\!|\!|u - I^{m-1}u|\!|\!| \\ &= c|u - I^{m-1}u|_{H^m(T)} \\ &\le c|u|_{H^m(T)} + c|I^{m-1}u|_{H^m(T)}. \end{aligned}$$

Since $I^{m-1}u \in P_{m-1}$, is holds $|I^{m-1}u|_{H^m(T)} = 0$, so that we arrive at the assertion. □

Remark: In two dimensions, $d = 2$, we require $s := m(m+1)/2$ points to ensure the uniqueness of the interpolation to a polynomial of total degree $m-1$.

For two normed spaces E, F and $\Phi \in \mathcal{L}(E, F)$, the image of Φ in the F-norm can be bounded by the norm of the inverse image in E. For the Sobolev-Raum $E = H^m(T)$ this means in particular:

$$\|\Phi u\|_F \;\le\; c\|u\|_{H^m(T)} \qquad \forall u \in H^m(T).$$

The following lemma makes an even stronger statement in the sense that (under certain assumptions on Φ) only the H^m-seminorm is required on the right hand side.

Lemma 5.8 (Bramble-Hilbert-Lemma) [1] *Let $T \subset \mathbb{R}^d$ be a Lipschitz domain, F a normed space, $m \in \mathbb{N}$ with $m > d/2$ and $\Phi \in \mathcal{L}(H^m(T), F)$ (i.e. $\Phi : H^m(T) \to F$ linear and continuous) in such a way that*

$$P_{m-1}(T) \;\subset\; Ker(\Phi).$$

[1] James H. Bramble and Stephen R. Hilbert.

Then there exists a constant c (depending on T and Φ) s.t.

$$\|\Phi u\|_F \leq c|u|_{H^m(T)} \qquad \forall u \in H^m(T).$$

Proof. Let $u \in H^m(T)$ and $\widetilde{u} = I^{m-1}u \in P_{m-1}(T)$ the nodal interpolant according to the previous lemma. Due to the assumption $P_{m-1}(T) \subset Ker(\Phi)$ we have $\Phi\widetilde{u} = 0$. This yields in combination with the previous lemma:

$$\begin{aligned} \|\Phi u\|_F &= \|\Phi(u-\widetilde{u})\|_F \\ &\leq \|\Phi\|_{H^m(T);F}\|u-\widetilde{u}\|_{H^m(T)} \\ &\leq c\|\Phi\|_{H^m(T);F}|u|_{H^m(T)}. \end{aligned}$$

□

5.4 Nodal interpolation

In the first step we consider the nodal interpolation on the reference element:

Corollary 5.9 *For the nodal interpolation $I^r : C(\hat{T}) \to P_r(\hat{T})$ on a reference element $\hat{T}$ with polynomials of total degree $r > d/2 - 1$ it holds with a constant $c = c(\hat{T}, r)$:*

$$\|\widehat{u} - I^r\widehat{u}\|_{L^\infty(\hat{T})} \leq c|\widehat{u}|_{H^{r+1}(\hat{T})} \qquad \forall \widehat{u} \in H^{r+1}(\hat{T}).$$

Proof. The assertion is a direct consequence of the Bramble-Hilbert-Lemma with $\Phi := id - I^r$ and $m = r + 1$. □

In the second step we consider affine linear transformations from the reference element $\widehat{T} \subset \mathbb{R}^d$ to $T \subset \mathbb{R}^d$, hence $\Phi_T \in P_1(\widehat{T})^d$, of the form

$$\Phi_T\widehat{x} = x_0 + B_T\widehat{x}.$$

If B is an invertible matrix, $B \in GL(d, \mathbb{R})$, this transformation is not degenerate

Lemma 5.10 *For the spectral norm $\|\cdot\|_2$ of the transformation it holds*

$$\|B_T\|_2 \leq \frac{h_T}{\rho_{\widehat{T}}} \quad \textit{und} \quad \|B_T^{-1}\|_2 \leq \frac{h_{\widehat{T}}}{\rho_T}.$$

Because, the inner and outer radius of the reference element is usually considered of order 1, it holds

$$\|B_T\|_2 \sim h_T \quad \text{and} \quad \|B_T^{-1}\|_2 \sim \rho_T^{-1}.$$

Lemma 5.11 (Transformation formula) *Let $\Phi_T : \widehat{T} \to T$ be an affine-linear transformation from the reference element $\widehat{T} \subset \mathbb{R}^d$ to $T \subset \mathbb{R}^d$, and $v \in H^m(T)$. Then $\widehat{v} = v \circ \Phi_T \in H^m(\widehat{T})$ and it holds*

$$|\widehat{v}|_{H^m(\widehat{T})} \leq \|B_T\|_2^m |\det(B_T)|^{-1/2}|v|_{H^m(T)}.$$

Proof. The partial derivatives on a point $x = \Phi_T \widehat{x}$ is given by:

$$\frac{\partial \widehat{v}(\widehat{x})}{\partial \widehat{x}_i} \;=\; \frac{\partial (v \circ \Phi_T)(\widehat{x})}{\partial \widehat{x}_i} \;=\; \sum_{j=1}^{d} \frac{\partial v(x)}{\partial x_j} \frac{\partial (\Phi_T)_j}{\partial \widehat{x}_i}(\widehat{x}) \;=\; (B_T \nabla v(x))_i.$$

This leads to the upper bound

$$\|\widehat{\nabla} \widehat{v}(\widehat{x})\|_2 \;\leq\; \|B_T\|_2 \|\nabla v(x)\|_2.$$

Per induction we obtain for the derivatives of order m:

$$\left(\sum_{|\alpha|=m} |\partial^\alpha \widehat{v}(\widehat{x})|^2 \right)^{1/2} \;\leq\; \|B_T\|_2^m \Big(\sum_{|\alpha|=m} |\partial^\alpha v(x)|^2 \Big)^{1/2}.$$

Thiy yields to the follwing bounds of the semi norm:

$$\begin{aligned} |\widehat{v}|^2_{H^m(\widehat{T})} &= \int_{\widehat{T}} \sum_{|\alpha|=m} |\partial^\alpha \widehat{v}(\widehat{x})|^2 \, d\widehat{x} \\ &\leq \|B_T\|_2^{2m} \int_{\widehat{T}} \sum_{|\alpha|=m} |\partial^\alpha v(\Phi_T \widehat{x})|^2 d\widehat{x}\,. \end{aligned}$$

The integral on the right hand side can be expressed by substitution ($dx/d\widehat{x} = |det\, D\Phi_T|$):

$$\begin{aligned} \int_{\widehat{T}} \sum_{|\alpha|=m} |\partial^\alpha v(\Phi_T \widehat{x})|^2 d\widehat{x} &= \int_T \sum_{|\alpha|=m} |\partial^\alpha v(x)|^2 |\det D\Phi_T|^{-1} \, dx \\ &= |\det B_T|^{-1} |v|^2_{H^m(T)}. \end{aligned}$$

The assertion is now obtained by using the previous inequality and taking the square root on both sides. ∎

In what follows we will use the broken H^m-seminorm:

$$|u|_{H^m(\mathcal{T}_h)} \;:=\; \left(\sum_{T \in \mathcal{T}_h} |u|^2_{H^m(T)} \right)^{1/2}.$$

For $u \in H^m(\Omega)$ it obviously hold $|u|_{H^m(\mathcal{T}_h)} = |u|_{H^m(\Omega)}$. However, we can apply this broken seminorm also on finite element functions which are only element-wise in H^m. The broken norm $\|u\|_{H^m(\mathcal{T}_h)}$ is defined accordingly.

Theorem 5.12 (Interpolation error) *Let $\mathcal{T}_h$ be an admissible triangulation of $\Omega \subset \mathbb{R}^d$ by simplices with maximal mesh width $h := \max\{h_T : T \in \mathcal{T}_h\}$. Then it holds for the finite element nodal interpolation I_h on P_r-elements, $r \geq d/2 - 1$, and $0 \leq m \leq r+1$:*

$$\|u - I_h u\|^2_{H^m(\mathcal{T}_h)} \;\leq\; c|h^{r+1-m} u|^2_{H^{r+1}(\mathcal{T}_h)} \qquad \forall u \in H^{r+1}(\Omega),$$

where $c = c(r, \kappa)$ is a constant, which depend on the maximal element anisotropy κ, but does not depend on h.

The right hand side in this estimate reads

$$\sum_{T\in\mathcal{T}_h} h_T^{2(r+1-m)}|u|^2_{H^{r+1}(T)}.$$

Proof. It is sufficient to show on each element $T \in \mathcal{T}$ the bound

$$\|u - I_h u\|_{H^m(T)} \leq c h_T^{r+1-m}|u|_{H^{r+1}(T)} \qquad \forall u \in H^{r+1}(T). \tag{5.3}$$

We use the transformation formula of Lemma 5.11 for $0 \leq l \leq m$ in opposite direction and apply Lemma 5.7 ($c = c(l,r)$):

$$\begin{aligned} |u - I_h u|_{H^l(T)} &\leq c\|B_T^{-1}\|_2^l |\det(B_T)|^{1/2} |\widehat{u} - I\widehat{u}|_{H^l(\widehat{T})} \\ &\leq c\|B_T^{-1}\|_2^l |\det(B_T)|^{1/2} \|\widehat{u} - I\widehat{u}\|_{H^{r+1}(\widehat{T})} \\ &\leq c\|B_T^{-1}\|_2^l |\det(B_T)|^{1/2} |\widehat{u}|_{H^{r+1}(\widehat{T})}. \end{aligned}$$

Application of the transformation formula of Lemma 5.11 a second time yields to

$$|\widehat{u}|_{H^{r+1}(\widehat{T})} \leq c\|B_T\|_2^{r+1} |\det(B_T)|^{-1/2} |u|_{H^{r+1}(T)}.$$

Using this in the estimate above gives

$$|u - I_h u|_{H^l(T)} \leq c\|B_T^{-1}\|_2^l \|B_T\|_2^{r+1} |u|_{H^{r+1}(T)}.$$

Furthermore, Lemma 5.10 ensures the existence of a constant $c = c(m)$ s.t.

$$\|B_T^{-1}\|_2^l \|B_T\|_2^{r+1} \leq \left(\frac{h_{\widehat{T}}}{\rho_T}\right)^l \left(\frac{h_T}{\rho_{\widehat{T}}}\right)^{r+1} \leq c\kappa_T^l h_T^{r+1-l}.$$

Hence, we arrive at

$$|u - I_h u|_{H^l(T)} \leq c\kappa_T^l h_T^{r+1-l} |u|_{H^{r+1}(T)},$$

which directly leads to (5.3). ∎

Remark 5.13 *The same estimate is valid for Q_r-elements under affine-linear transformation, i.e. on parallelogram meshes.*

5.5 Sequence of triangulations

In practice we often use a (finite) sequence of meshes in order to successively improve the accuracy of the approximation. Such a sequence may consist of globally refined meshes, where all elements are (basically) of the same size, or locally refined meshes, where a large variety of element sizes may occur. For theoretical considerations we use the following terms:

Definition 5.14 *A sequence of admissible triangulations $\{\mathcal{T}_h\}$ of a domain $\Omega \subset \mathbb{R}^2$ is called* shape regular*, if a constant $\kappa > 0$ exists, such that for each element $T \in \mathcal{T}_h$ of any triangulation $\mathcal{T}_h$ with outer radius h_T and inner radius ρ_T it holds:*

$$h_T \quad \leq \quad \kappa \rho_T.$$

A shape-regular sequence is called quasi-uniform*, if for each triangulation $\mathcal{T}_h$ satisfies:*

$$\max_{K \in \mathcal{T}_h} h_K \quad \leq \quad \kappa \rho_T \qquad \forall T \in \mathcal{T}_h.$$

In shape-regular triangulations, the relation between outer and inner radius is bounded. This prevents, in particular, that the cells degenerate on finer meshes. The anisotropy relations are bounded. However, the sizes (outer radii) of the cells may vary a lot. The parameter κ_T in the proof of Theorem 5.12 is then bounded from above independent of the particular triangulation $\mathcal{T}_h$. Hence, the constant in the upper bound in Theorem 5.12 is independent of the particular $\mathcal{T}_h \in \{\mathcal{T}_h\}$.

On quasi-uniform meshes, the relation between maximal outer radius and minimal inner radius remains bounded:

$$\frac{\max_{K \in \mathcal{T}_h} h_K}{\min_{K \in \mathcal{T}_h} \rho_K} \quad \leq \quad \kappa.$$

On such quasi-uniform meshes we directly obtain the following result:

Corollary 5.15 *Let $\{\mathcal{T}_h\}$ be a family of quasi-uniform triangulations of $\Omega \subset \mathbb{R}^d$. Then it holds for the finite element nodal interpolation $I_h : C(\bar{\Omega}) \to V_h$ with P_r-elements, $r \geq d/2 - 1$, and $0 \leq m \leq r + 1$:*

$$\|v - I_h v\|_{H^m(\mathcal{T}_h)} \quad \leq \quad c h^{r+1-m} |v|_{H^{r+1}(\Omega)} \qquad \forall v \in H^{r+1}(\Omega)\,,$$

where $h = \max_{T \in \mathcal{T}_h} h_T$ and $c = c(r, \kappa)$.

Proof. The proof follows directly from Theorem 5.12. ∎

5.6 A priori error estimate in the H^1-norm

The following theorem holds for domains Ω with C^2-boundary, i.e. it exists a (bijective) parameterization $\varphi \in C^2([0,1), \partial\Omega)$.

Theorem 5.16 (Stability of Poisson problem) *Let $\Omega \subset \mathbb{R}^d$, $d \in 2, 3$, be a convex domain or a domain with C^2-boundary, $V = H_0^1(\Omega)$, $A : V \times V \to \mathbb{R}$ a V-continuous and V-elliptic bilinear form with sufficient smooth coefficient functions, and $f \in L^2(\Omega)$. Then the unique solution $u \in V$ to the corresponding variational problem (4.1) satisfies $u \in H^2(\Omega)$ and*

$$\|u\|_{H^2(\Omega)} \quad \leq \quad c \|f\|_{L^2(\Omega)},$$

with a constant $c = c(\Omega, A)$.

Proof. We only give the principal idea of the proof. At first, one shows the H^2-regularity for rectangles Ω. In the second step, one splits the domain in a finite number of rectangles (in the interior if Ω) and special square-like elements with only one curved edge. The latter ones can be transformed to squares with bounded determinant of the mapping. The final result is then obtained by a partition of unity. ∎
Domains with non-convex corners are not covered by this theorem. Such domains usually allow for solutions $u \notin H^2(\Omega)$.

Corollary 5.17 (a priori error estimate) *We make the same assumptions as in Theorem 5.16 and consider a family $\{\mathcal{T}_h\}$ of shape regular triangulations of $\Omega \subset \mathbb{R}^d$. The corresponding finite element solutions u_h of (4.1) with P_1-elements or Q_1-elements on parallelogram meshes satisfy ($c = c(\Omega, A, \kappa)$)*

$$\|u - u_h\|_{H^1(\Omega)} \;\leq\; ch|u|_{H^2(\Omega)} \;\leq\; ch\|f\|_{L^2(\Omega)}.$$

Proof. According to Cea's Lemma 5.1, the use of the nodal interpolation $I_h u$ and the corresponding interpolation estimate in Theorem 5.12 with $m = r = 1$ we obtain with $c = c(\kappa)$

$$\begin{aligned}
\|u - u_h\|^2_{H^1(\Omega)} &\leq \frac{\alpha_1}{\alpha_2}\|u - Iu_h\|^2_{H^1(\Omega)} \\
&= \frac{\alpha_1}{\alpha_2}\left(\|u - Iu_h\|^2_{L^2(\Omega)} + |u - Iu_h|^2_{H^1(\Omega)}\right) \\
&\leq \frac{\alpha_1}{\alpha_2}(c^2h^4 + c^2h^2)|u|^2_{H^2(\Omega)}.
\end{aligned}$$

Taking the square root on both sides we obtain with another constant $c > 0$:

$$\|u - u_h\|_{H^1(\Omega)} \;\leq\; ch|u|_{H^2(\Omega)}.$$

The stability estimate in Theorem 5.16 leads to the assertion. ∎
The naturally arising question is whether the use of higher order finite elements, e.g. P_2, may lead to improved estimates. The answer in general is: no, because a convergence order $O(h^2)$ in the energy norm $|\cdot|_{H^1(\Omega)}$ requires H^3-regularity of the exact solution u. This is only valid for smooth boundaries which can usually not be triangulated with triangular meshes and affine-linear transformations. Special boundary approximations are needed. We will (hopefully) come back to this point at a later occasion.

5.7 A priori error estimate in the L^2-norm

So far we considered the energy norm (H^1-seminorm) for the error estimate. One may expect that the error in the L^2-norm is of one order better, because the interpolation error behaves like this. However, to derive such an estimate we need a special technique, the so-called duality argument. We express the L^2-norm by the norm of the dual space.

For a given right hand side $g \in V'$ let z_g the dual solution of the problem

$$z_g \in V: \qquad A(v, z_g) \;=\; \langle g, v\rangle \qquad \forall v \in V. \tag{5.4}$$

Compared to the first problem, here the test and ansatz functions in the bilinear form are obviously exchanged. In the particular case of the Poisson problem we obtain due to the self-adjointness the same bilinear form.

Theorem 5.18 (Aubin-Nitsche) *Let V, W be two Hilbert spaces with continuous embedding $V \subseteq W$. Furthermore, let the bilinear form $A : V \times V \to \mathbb{R}$ be V-continuous. Then it holds for the finite element solution $u_h \in V_h \subset V$ of (4.1):*

$$\|u - u_h\|_W \;\leq\; \alpha_1 \sup_{g\in W', \|g\|_{W'}=1} \left\{ \inf_{z_h \in V_h} \|z_g - z_h\|_V \right\} \|u - u_h\|_V,$$

where z_g is the solution of the dual problem (5.4) and α_1 the continuity constant of $A(\cdot,\cdot)$ with respect to the V-norm.

Proof. Let $S := \{g \in W', \|g\|_{W'} = 1\}$. Then we have

$$\|u - u_h\|_W \;=\; \sup_{g\in S} \langle g, u - u_h\rangle. \tag{5.5}$$

Due to the continuous embedding $V \subset W$ we conclude $W' \subseteq V'$, so that in particular $g \in S \subseteq V'$. Hence, we can use g as right hand side in the dual problem (5.4). By Galerkin orthogonality for the dual problem and the continuity of the bilinear form it holds for arbitrary $z_h \in V_h$:

$$\begin{aligned} \langle g, u - u_h\rangle &= A(u - u_h, z_g) \\ &= A(u - u_h, z_g - z_h) \\ &\leq \alpha_1 \|u - u_h\|_V \|z_g - z_h\|_V. \end{aligned}$$

Taking the infimum over all possible $z_h \in V_h$ gives

$$\langle g, u - u_h\rangle \;\leq\; \alpha_1 \inf_{z_h \in V_h} \|z_g - z_h\|_V \|u - u_h\|_V.$$

Using this in (5.5) implies the assertion. ∎

In the particular case of the Poisson problem we obtain:

Corollary 5.19 (A priori error estimate in L^2) *Let $\{\mathcal{T}_h\}$ be a family of shape-regular triangulations of a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, which is convex or has a C^2 boundary. Then it holds for the P_1 (and Q_1 on parallelogram meshes) finite element solutions $u_h \in V_h \subset H_0^1(\Omega)$ of (4.1):*

$$\|u - u_h\|_{L^2(\Omega)} \;\leq\; ch^2 \|f\|_{L^2(\Omega)}.$$

Proof. Here we have $V := H_0^1(\Omega) \subset W := L^2(\Omega)$ with continuous embedding. The continuity constant is in the case of the Laplace operator $\alpha_1 = 1$. According to Theorem 5.18 we obtain:

$$\|u - u_h\|_{L^2(\Omega)} \quad \leq \sup_{g \in L^2(\Omega), \|g\|=1} \left\{ \inf_{z_h \in V_h} \|\nabla(z_g - z_h)\|_{L^2(\Omega)} \right\} \|\nabla(u - u_h)\|_{L^2(\Omega)}.$$

Application of Corollary 5.17 to the dual problem with right hand side $g \in L^2(\Omega)$, $\|g\|_{L^2(\Omega)} = 1$, yields the bound for the solution z_g of the dual problem (5.4):

$$\|\nabla(z_g - z_h)\|_{L^2(\Omega)} \quad \leq \quad ch.$$

Application of this Corollary to the primal problem leads to

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \quad \leq \quad ch\|f\|_{L^2(\Omega)}.$$

The combination of these two estimates leads to the assertion. □

Chapter 6

A posteriori error estimates

In this chapter we investigate techniques to estimate the discretization error on the basis of the computed discrete solution. This is needed to perform local mesh refinement, because we have to know in which elements the error is relatively large. In the subsequent analysis some interpolations to the finite element spaces are needed. However, the nodal interpolation requires $H^2(\Omega)$-regularity of the solution and does not work for H^1-regularity only. The L^2-projection require less regularity but is a global process so that the maximal mesh size would appear in the estimates. This makes it not usable for local mesh refinement. Therefore, we start with a local projection without the requirement of H^2-regularity.

6.1 Clément interpolation

The operator of Clément maps H^1-functions onto continuous P_1-elements:

$$C_h : H^1(\Omega) \to P_1(\mathcal{T}_h).$$

The construction is as follows: To each mesh vertex N_i we define the union of all adjacent elements

$$\omega_i \quad := \quad \bigcup_{T \in \mathcal{T}_h, N_i \in T} T,$$

and let $\Pi_i^0 : L^2(\omega_i) \to \mathbb{R}$ be the L^2-projection onto the constants, defined by

$$\int_{\omega_i} (v - \Pi_i^0 v)\, dx \quad = \quad 0 \qquad \forall v \in L^2(\omega_i).$$

Now, we define

$$C_h v(x) \quad := \quad \sum_{i=1}^{N} \Pi_i^0 v \cdot \phi_i(x),$$

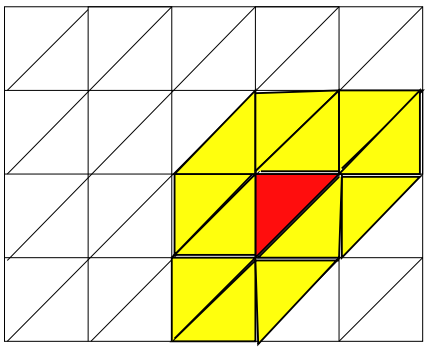


Figure 6.1: The patch ω_T (yellow and red) for the element T (red).

where ϕ_i denotes the P_1 Lagrange basis function corresponding to the vertex N_i.

In the following interpolation estimate further element patches ω_T for each element $T \in \mathcal{T}_h$ are used. These are unions of all elements which have a least one vertex in common with T, see Fig. 6.1:

$$\omega_T := \bigcup_{T' \in \mathcal{T}_h, T \cap T' \neq \emptyset} T' = \bigcup_{N_j \in T} \omega_j \,.$$

Theorem 6.1 (Approximation property of the Clément operator) *Let $\{\mathcal{T}_h\}$ be a shape-regular family of triangulations of a bounded domain $\Omega \subset \mathbb{R}^2$. Then the operator of Clément C_h is well-defined, linear and has the following property: It exists a constant C, s.t. for all $m \in \{0,1\}$, all $v \in H^1(\Omega)$ and all elements $T \in \mathcal{T}_h$ it holds:*

$$\begin{aligned} \|v - C_h v\|_{H^m(T)} &\leq C h_T^{1-m} \|v\|_{H^1(\omega_T)}, \\ \|v - C_h v\|_{L^2(\partial T)} &\leq C h_T^{1/2} \|v\|_{H^1(\omega_T)}. \end{aligned}$$

Proof. For a proof we refer to [6]. It uses in particular the following estimate:

$$\|v - \Pi_i^0 v\|_{L^2(\omega_i)} \leq C \mathrm{diam}\,(\omega_i) \|v\|_{H^1(\omega_i)},$$

with a suitable constant C, and the property of the patches ω_T on shape-regular meshes it holds with $C = C(\kappa)$:

$$|\omega_T| \leq C|T| \leq C h_T^d.$$

□

6.2 A posteriori error estimate in the energy norm

Now we address the so-called a posteriori error estimate. These are error estimates which can be used for numerical evaluation on basis of the computed discrete solution $u_h \in$

V_h. The aims are twofold: The estimate can be used (a) to get an upper bound of the computed solution, and (b) to obtain local information where the largest error is 'produced'. Such information is essential to perform local mesh refinement. In particular, for solutions with low regularity (e.g. singularities in certain derivatives of u) and to resolve interior or boundary layers, the concept of local mesh refinement essential to obtain efficient numerical methods. We present such concepts in the framework of the Poisson problem, although the extension and generalizations to more difficult types of equations is not always straightforward.

We firstly introduce some notations. As cell residuals of the Poisson problem we understand the quantities

$$\rho_T(u_h) \quad := \quad \|\Delta u_h + f\|_{L^2(T)}.$$

For the exact solution u with enough regularity these residuals vanish, $\rho_T(u) = 0$. For P_1- and for Q_1-elements on parallelograms, these residuals are actually independent of u_h, because $\Delta u_h|_T = 0$ and hence $\rho_T(u_h) = \|f\|_{L^2(T)}$. The second important quantity consists of the jump terms across inner edges (faces) e of two elements T_1, T_2 of the triangulation:

$$\begin{aligned} \rho_e(u_h) \quad &:= \quad \|[\partial u_h/\partial n]_e\|_{L^2(e)}, \\ \text{where} \qquad \left[\frac{\partial u_h}{\partial n}\right]_e \quad &:= \quad \left.\frac{\partial u_h}{\partial n}\right|_{T_1} - \left.\frac{\partial u_h}{\partial n}\right|_{T_2}. \end{aligned}$$

These vanish for C^1-functions. However, for the discrete solution u_h these are usually not zero. The set of all inner edges (i.e. edges which are not part of the boundary $\partial\Omega$) will be denoted by $\mathcal{E}_h$ and the set of those edges of an element T by $\mathcal{E}_T$.

Theorem 6.2 *Let $\{\mathcal{T}_h\}$ be a family of shape-regular triangulations of the domain $\Omega \subset \mathbb{R}^2$, and let u_h be the P_1-solution of the Poisson problem (4.1). Then it holds*

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \quad \leq \quad c \left(\sum_{T \in \mathcal{T}_h} h_T^2 \rho_T(u_h)^2 + \sum_{e \in \mathcal{E}_h} h_e \rho_e(u_h)^2 \right)^{1/2},$$

with a constant $c = c(\Omega, \kappa)$.

Proof. The duality argument with $B_1 := \{g \in H_0^1(\Omega)' : \|g\|_{H_0^1(\Omega)'} = 1\}$ again yields

$$|u - u_h|_{H^1(\Omega)} \quad = \quad \sup_{g \in B_1} \langle g, u - u_h \rangle.$$

Let $z_g \in H_0^1(\Omega)$ be the solution of the dual problem (5.4). With the Galerkin orthogonality of the primal problem we obtain for arbitrary function $\psi_h \in V_h$ and the notation $w :=$

$z_g - \psi_h$:

$$\begin{aligned}\langle g, u-u_h\rangle &= (\nabla(u-u_h), \nabla z_g)_{L^2(\Omega)} = (\nabla(u-u_h), \nabla w)_{L^2(\Omega)} \\ &= (f,w)_{L^2(\Omega)} - (\nabla u_h, \nabla w)_{L^2(\Omega)} \\ &= \sum_{T\in\mathcal{T}_h} \left\{ (f,w)_{L^2(T)} - (\nabla u_h, \nabla w)_{L^2(T)} \right\}.\end{aligned}$$

On each element T it holds due to integration by parts:

$$(f,w)_{L^2(T)} - (\nabla u_h, \nabla w)_{L^2(T)} = (f + \Delta u_h, w)_{L^2(T)} - \int_{\partial T} \partial_n u_h w \, ds.$$

The boundary integrals on the right hand side yield after summation over all elements and all corresponding edges:

$$\sum_{T\in\mathcal{T}_h} \int_{\partial T} \partial_n u_h w \, ds = \sum_{e\in\mathcal{E}_h} \int_e [\partial_n u_h] w \, ds \leq \sum_{e\in\mathcal{E}_h} \|[\partial_n u_h]\|_e \|w\|_e.$$

Note that (a) the intergrals over boundary edges vanish, because w has vanishing trace on $\partial\Omega$, and (b) the directions of the normals change depending on the cell where we perform the integration by parts. Using this in the bound above and applying Cauchy-Schwarz inequality we obtain the bound:

$$\begin{aligned}\langle g, u-u_h\rangle &\leq \sum_{T\in\mathcal{T}_h} \|f + \Delta u_h\|_{L^2(T)} \|w\|_{L^2(T)} + \sum_{e\in\mathcal{E}_h} \|[\partial_n u_h]\|_{L^2(e)} \|w\|_{L^2(e)} \\ &= \sum_{T\in\mathcal{T}_h} \left\{ \rho_T(u_h) \|w\|_{L^2(T)} + \frac{1}{2} \sum_{e\in\mathcal{E}_T} \|[\partial_n u_h]\|_{L^2(e)} \|w\|_{L^2(e)} \right\} \\ &= \sum_{T\in\mathcal{T}_h} \left\{ \rho_T(u_h) \|w\|_{L^2(T)} + \frac{1}{2} \sum_{e\in\mathcal{E}_T} \rho_e(u_h) \|w\|_{L^2(e)} \right\}.\end{aligned}$$

For the interpolation errors $w = z_g - \psi_h$ we dont use the nodal interpolation, because this would lead to the second derivatives of z_g. But these probably not even exist, depending on the type of domain Ω. Usually further restriction would be needed, e.g. convexity. If we are only able to use first derivatives, i.e. the $|\cdot|_{H^1(\Omega)}$-norm, we can use the Clément operator of the previous section, $\psi_h := C_h z_g$:

$$\|w\|_{L^2(T)} = \|z_g - C_h z_g\|_{L^2(T)} \leq Ch_T |z_g|_{H^1(\omega(T))}.$$

For the interpolation error on the edges we use the approximations property of the Clément operator on $e \in \mathcal{E}_T$:

$$\|w\|_{L^2(e)} = Ch_e^{1/2} |z_g|_{H^1(\omega(T))}.$$

This leads to

$$\begin{aligned}\langle g, u-u_h\rangle &\leq C\sum_{T\in\mathcal{T}_h}\left\{\rho_T(u_h)h_T|z_g|_{H^1(\omega(T))}+\frac{1}{2}|z_g|_{H^1(\omega(T))}\sum_{e\in\mathcal{E}_T}\rho_e(u_h)h_e^{1/2}\right\}\\ &= C\sum_{T\in\mathcal{T}_h}\left\{\rho_T(u_h)h_T+\frac{1}{2}\sum_{e\in\mathcal{E}_T}\rho_e(u_h)h_e^{1/2}\right\}|z_g|_{H^1(\omega(T))}\\ &\leq C\left(\sum_{T\in\mathcal{T}_h}\left\{\rho_T(u_h)h_T+\frac{1}{2}\sum_{e\in\mathcal{E}_T}\rho_e(u_h)h_e^{1/2}\right\}^2\right)^{1/2}\left(\sum_{T\in\mathcal{T}_h}|z_g|^2_{H^1(\omega(T))}\right)^{1/2}.\end{aligned}$$

Because we assumed the triangulations to be shape-regular, the sets $\omega(T)$ contain a maximal number of elements of $\mathcal{T}_h$. This yields to

$$\sum_{T\in\mathcal{T}_h}|z_g|^2_{H^1(\omega(T))} \leq C|z_g|_{H^1(\Omega)}.$$

In combination with the duality argument we arrive at

$$|u-u_h|_{H^1(\Omega)} \leq C\sum_{T\in\mathcal{T}_h}\left\{h_T^2\rho_T(u_h)^2+\frac{1}{2}\sum_{e\subset\mathcal{E}_T}h_e\rho_e(u_h)^2\right\}\sup_{g\in B_1}|z_g|_{H^1(\Omega)}.$$

A boundedness of the dual solution is still needed:

$$|z_g|^2_{H^1(\Omega)} = (\nabla z_g,\nabla z_g) = \langle g, z_g\rangle \leq \|g\|_{H_0^1(\Omega)'}|z_g|_{H^1(\Omega)} = |z_g|_{H^1(\Omega)}.$$

This implies $\sup_{g\in B_1}|z_g|_{H^1(\Omega)}\leq 1$. The assertion follows. ∎

6.3 Error estimation by dual weighted residuals

6.3.1 Dual weighted residuals for linear problems

Now we derive an error estimator which gives information about the differences of the continuous and discrete solution with respect to functional output. We consider a linear functional

$$J: V\to\mathbb{R}.$$

Examples of such functionals are

$$\begin{aligned}\text{point errors} \quad & J(u)=u(x_0),\\ \text{or} \quad & J(u)=\|\nabla u(x_0)\|_2,\\ \text{boundary fluxes} \quad & J(u)=\int_\Gamma\frac{\partial u}{\partial n}\,ds,\end{aligned}$$

for subdomains $\Omega' \subset \Omega$ or boundary parts $\Gamma \subset \partial\Omega$. As generalization one may (see next subsection) also nonlinear functionals , as for instance

$$\begin{array}{ll} \text{energy norm} & J(u) = \|\nabla u\|_{L^2(\Omega')} \\ L^2\text{-norm} & J(u) = \|u\|_{L^2(\Omega')}. \end{array}$$

For systems of partial differential equations often certain components of the solution are of interest. We assume that J is linear and continuous, hence $J \in V'$.

We write the underlying continuous and the discrete equations in abstract form:

$$u \in V: \qquad A(u,v) = \langle f, v\rangle \quad \forall v \in V, \tag{6.1}$$

$$u_h \in V_h: \qquad A(u_h,v) = \langle f, v\rangle \quad \forall v \in V_h, \tag{6.2}$$

with a continuous bilinear form $A : V \times V \to \mathbb{R}$. The residual of the equation will be denoted by

$$\varrho(u_h, v) := \langle f, v\rangle - A(u_h, v)\,.$$

Theorem 6.3 *Let V be a Hilbert space, $A : V \times V \to \mathbb{R}$ a continuous bilinear form, $J \in V'$ and $V_h \subset V$. Then the discretization error of the finite element solution of the variational problems (6.1) and (6.2) is given by*

$$J(u) - J(u_h) = \varrho(u_h, z - i_h z),$$

where $i_h : V \to V_h$ is an arbitrary map and $z \in V$ is the solution of the dual (or adjoint) problem

$$z \in V: \qquad A(v,z) = J(v) \qquad \forall v \in V.$$

Proof. The error with J can be expressed by the dual solution z:

$$J(u) - J(u_h) = A(u,z) - A(u_h,z) = \langle f, z\rangle - A(u_h,z) = \varrho(u_h,z).$$

Since u_h is the solution of the discrete primal problem, we can subtract an arbitrary interpolant $i_h z \in V_h$ from the test function (Galerkin orthogonality). This gives the assertion. ∎

Corollary 6.4 *Let $V = H_0^1(\Omega)$, $J \in V'$, $f \in L^2(\Omega)$ and $A(u,v) = (\nabla v, \nabla v)_{L^2(\Omega)}$ the bilinear form of the Laplace operator. Then it holds for the discretization error of a conforming finite element solution $u_h \in V_h$:*

$$J(u) - J(u_h) = \sum_{T\in\mathcal{T}_h} \left\{ (f + \Delta u_h, z - z_h)_{L^2(T)} - \frac{1}{2}([\partial_n u_h], z - z_h)_{L^2(\partial T)} \right\}.$$

Proof. We use the abbreviation $w := z - i_h z$, express the residuals as sum over local residuals and integrate by parts

$$\begin{aligned}\varrho(u_h, z - i_h z) &= \sum_{T\in\mathcal{T}_h} \left\{(f, w)_{L^2(T)} - (\nabla u_h, \nabla w)_{L^2(T)}\right\} \\ &= \sum_{T\in\mathcal{T}_h} \left\{(f + \Delta u_h, w)_{L^2(T)} - (\partial_n u_h, w)_{L^2(\partial T)}\right\}.\end{aligned}$$

Expressing the edge integrals as jumps over the derivatives in normal directions over the edges leads to the desired form. ∎

This now leads directly to the error estimator

$$|J(u) - J(u_h)| \approx \eta_J := \sum_{T\in\mathcal{T}_h} \eta_T(u_h),$$

with local error indicators

$$\eta_T(u_h) := \|f + \Delta u_h\|_{L^2(T)}\omega_T + \frac{1}{2}\|[\partial_n u_h]\|_{L^2(\partial T)}\omega_{\partial T},$$

and suitable approximations with can be numerically evaluated $\omega_T \approx \|z - z_h\|_{L^2(T)}$ and $\omega_{\partial T} \approx \|z - z_h\|_{L^2(\partial T)}$. In the next subsection we present several possibilities for this.

6.3.2 Approximation of the weights

We present three possibilities to compute the weighting functions $\omega_T, \omega_{\partial T}$ in the dual weighted residual error estimator.

1.) *Computation of the dual solution with higher accuracy:* One may compute the dual problem by higher-order finite elments or with a finer mesh size. If $V_h = P_r(\mathcal{T}_h)$ one may choose $\widetilde{V}_h = P_{r'}(\mathcal{T}_h)$ with $r' > r$ or $\widetilde{V}_h = P_r(\mathcal{T}_{h'})$ with $h' < h$, and the compute $\widetilde{z}_h \in \widetilde{V}_h$:

$$\widetilde{z}_h \in \widetilde{V}_h : \qquad A(v, \widetilde{z}_h) = J(v) \qquad \forall v \in \widetilde{V}_h.$$

Now a possible approximation of the interpolation error is:

$$z - i_h z \approx \widetilde{z}_h - i_h \widetilde{z}_h,$$

where an interpolation $i_h : \widetilde{V}_h \to V_h$ has to be computed. Possible choices are in particular $r' = 2r$ or $h' = h/2$. The most expensive part is to determine $\widetilde{z}_h$. It is even more expensive than the computation of the primal solution u_h.

2.) *Interpolation by higher order:* Instead of computing one may solve the dual solution by the same method as the primal problem and perform afterwards an interpolation $P_{2r}(\mathcal{T}_{2h})$:

$$i_{2h}^{2r} : P_r(\mathcal{T}_h) \to P_{2r}(\mathcal{T}_{2h}).$$

Now the approximation reads

$$z - z_h \approx i_{2h}^{2r} z_h - z_h.$$

The evaluation of $i_{2h}^{2r} z_h$ must be carried out on patches $T' \in \mathcal{T}_{2h}$ of several elements $T_1, \ldots, T_{2^d} \in \mathcal{T}_h$. The idea is that by the splitting

$$\|z - z_h\|_{L^2(T)} \leq \|z - i_{2h}^{2r} z_h\|_{L^2(T)} + \|i_{2h}^{2r} z_h - z_h\|_{L^2(T)}$$

one can show that the term $\|z - i_{2h}^{2r} z_h\|_{L^2(T)}$ is of higher order accuracy than $\|i_{2h}^{2r} z_h - z_h\|_{L^2(T)}$, at least under certain uniformity conditions of the mesh $\mathcal{T}_h$. However, in practice this method also works robustly on non-uniform meshes.

3.) *Discrete difference quotients:* We estimate and approximate for $r \geq 1$:

$$\begin{aligned} \|z - i_h z\|_{L^2(T)} &\leq C h_T^{r+1} |z|_{H^{r+1}(T)} \\ &\leq C h_T^{r+1+d/2} \|\nabla^{r+1} z\|_{L^\infty(T)} \\ &\approx C h_T^{r+1+d/2} |D_h^{r+1} z_h(x_T)|, \end{aligned}$$

where D_h^{r+1} builds a discrete difference quotient of order $r+1$ on the mesh $\mathcal{T}_h$, and assuming $z \in W^{r+1,\infty}(\Omega)$. In the case of linear or bilinear elements, $r = 1$, in two dimensions we obtain

$$\|z - i_h z\|_{L^2(T)} \approx C h_T^3 |D_h^2 z_h(x_T)|,$$

with (center) point $x_T \in T$. The evaluation of the second order difference quotients has to be carried out on patches of elements as well, because on a single element u_h is (bi-)linear ($r = 1$).

6.3.3 Examples of output functionals

In this section we give some examples of possible output functionals:

(a) Mean over a subdomain $\Omega' \subset \Omega$:

$$J(u) := \int_{\Omega'} u \, dx.$$

The dual problem of the Laplace operator reads in strong form:

$$\begin{aligned} -\Delta z &= \mathbb{1}_{\Omega'} \qquad \text{in } \Omega, \\ z &= 0 \qquad \text{auf } \partial\Omega. \end{aligned}$$

with the characteristic function $\mathbb{1}_{\Omega'} : \Omega \to \{0, 1\}$ with respect to Ω'.

(b) Point value at $x_0 \in \Omega$:

$$J(u) := u(x_0).$$

The dual problem of the Laplace operator reads in strong form:

$$\begin{aligned} -\Delta z &= \delta_{x_0} \qquad \text{in } \Omega\,, \\ z &= 0 \qquad \text{auf } \partial\Omega, \end{aligned}$$

with the Dirac distribution $\delta_{x_0} : \Omega \to \{0, \infty\}$, hence $\delta_{x_0}(y) = 0$ for $y \neq x_0$ and $\int_\Omega \delta_{x_0}(x)dx = 1$. In this case, the discrete problem must be regularized:

$$z_h \in V_h : \qquad (\nabla z_h, \nabla v) = \int_{R_h} \phi_h v \, dx \qquad \forall v \in V_h,$$

with a function ϕ_h, with the property $\lim_{h\to 0} \phi_h = \delta_{x_0}$ and support $R_h := |\mathrm{supp}\, \phi_h| \sim h^d$.
(c) Error in the energy norm: In this case the corresponding functional reads

$$J(v) := \|\nabla e\|^{-1}_{L^2(\Omega)} (\nabla v, \nabla e)_{L^2(\Omega)}.$$

This functional is also dependent on e, if it should be formulated as a *linear* functional. The adjoint problem is

$$z \in V : \qquad (\nabla v, \nabla z)_{L^2(\Omega)} = |e|^{-1}_{H^1(\Omega)} (\nabla v, \nabla e)_{L^2(\Omega)}.$$

If we take as test function $v = z$, we obtain

$$\begin{aligned} |z|^2_{H^1(\Omega)} &= |e|^{-1}_{H^1(\Omega)} (\nabla z, \nabla e)_{L^2(\Omega)} \\ &\leq |e|^{-1}_{H^1(\Omega)} |z|_{H^1(\Omega)} |e|_{H^1(\Omega)} \\ &= |z|_{H^1(\Omega)}. \end{aligned}$$

This implies $|z|_{H^1(\Omega)} \leq 1$. Applying the Clément interpolation we obtain as upper bound for the local approximation error of the adjoint solution in the case of P_1-elements:

$$\|z - C_h z\|_{L^2(T)} + h_T^{1/2} \|z - C_h z\|_{L^2(\partial T)} \leq C h_T \|z\|_{H^1(\omega_T)}.$$

Plugging this into the error estimate above yields to

$$\begin{aligned} & |u - u_h|_{H^1(\Omega)} \\ \leq\ & C \sum_{T\in\mathcal{T}_h} \left\{ \|f + \Delta u_h\|_{L^2(T)} \|z - C_h z\|_{L^2(T)} + \frac{1}{2} \|[\partial_n u_h]\|_{L^2(\partial T)} \|z - C_h z\|_{L^2(\partial T)} \right\} \\ \leq\ & C \sum_{T\in\mathcal{T}_h} \left\{ h_T \rho_T(u_h) + \frac{1}{2} h_T^{1/2} \rho_{\partial T}(u_h) \right\} \|z\|_{H^1(\omega_T)} \\ \leq\ & C \left(\sum_{T\in\mathcal{T}_h} \left\{ h_T^2 \rho_T(u_h)^2 + \frac{1}{4} h_T \rho_{\partial T}(u_h)^2 \right\} \right)^{1/2} \left(\sum_{T\in\mathcal{T}_h} \|z\|^2_{H^1(\omega_T)} \right)^{1/2} \\ \leq\ & C \left(\sum_{T\in\mathcal{T}_h} \left\{ h_T^2 \rho_T(u_h)^2 + \frac{1}{4} h_T \rho_{\partial T}(u_h)^2 \right\} \right)^{1/2} \|z\|_{H^1(\Omega)}. \end{aligned}$$

Usage of the inequality of Poincaré leads to $\|z\|_{H^1(\Omega)} \leq C|z|_{H^1(\Omega)} \leq C$. This corresponds to the estimator developed in Section 6.2.

(d): Error in the L^2-norm: In this case the corresponding functional reads

$$J(v) \;:=\; \|e\|_{L^2(\Omega)}^{-1}(v,e)_{L^2(\Omega)}.$$

We choose the nodal interpolation $I_h z$ and obtain with an analogous estimate as before, if $z \in H^2(\Omega)$:

$$\begin{aligned}
&\|u-u_h\|_{L^2(\Omega)} \\
\leq\;& C\sum_{T\in\mathcal{T}_h}\Big\{\|f+\Delta u_h\|_{L^2(T)}\|z-I_h z\|_{L^2(T)} \\
&\qquad +\frac{1}{2}\|[\partial_n u_h]\|_{L^2(\partial T)}\|z-I_h z\|_{L^2(\partial T)}\Big\} \\
\leq\;& C\left(\sum_{T\in\mathcal{T}_h}\Big\{h_T^4\rho_T(u_h)^2+\frac{1}{4}h_T^3\rho_{\partial T}(u_h)^2\Big\}\right)^{1/2}|z|_{H^2(\Omega)}.
\end{aligned}$$

On convex domains Ω or domains with C^2-boundary it holds according to Theorem 5.16 $z \in H^2(\Omega)$ with

$$|z|_{H^2(\Omega)} \leq C_\Omega \|J\|_{L^2(\Omega)}.$$

The functional J has to be interpreted as a L^2-function. It holds

$$\begin{aligned}
\|J\|_{L^2(\Omega)} &= \sup_{v\in L^2(\Omega)}\frac{|J(v)|}{\|v\|_{L^2(\Omega)}} \\
&= \|e\|_{L^2(\Omega)}^{-1}\sup_{v\in L^2(\Omega)}\|v\|_{L^2(\Omega)}^{-1}|(v,e)| \\
&\leq \|e\|_{L^2(\Omega)}^{-1}\|e\|_{L^2(\Omega)} \;=\; 1.
\end{aligned}$$

In summary we obtain the following result:

Theorem 6.5 (a posteriori in L^2) *Let $\Omega \subset \mathbb{R}^2$ be a convex domain or a domain with C^2-boundary. Then we have for the P_1 finite element solution the a posteriori error estimate*

$$\|u-u_h\|_{L^2(\Omega)} \;\leq\; C_\Omega\left(\sum_{T\in\mathcal{T}_h}\Big\{h_T^4\rho_T(u_h)^2+\frac{1}{4}h_T^3\rho_{\partial T}(u_h)^2\Big\}\right)^{1/2}.$$

6.3.4 Alternative approach via a saddle-point problem

In this section we present a more abstract approach. This will be useful for the analysis of sesqui-linear forms which come into play for nonlinear partial differential equations.

However, let A be here still bilinear. We consider the auxiliary functional $L : V \times V \to \mathbb{R}$ given by

$$L(u,z) \;=\; J(u) - A(u,z) + \langle f, z\rangle.$$

This affin-linear L is chosen in such a way that the solutions u, z and u_h, z_h of the underlying problems appear as argument in L in the form J:

$$\begin{aligned} J(u) &= L(u,z) \qquad \forall z \in V, \\ J(u_h) &= L(u_h, z_h) \qquad \forall z_h \in V_h. \end{aligned}$$

The Gateaux[1]-derivative of this functional is now given by

$$\begin{aligned} \partial_z L(u,z)(v) &= \lim_{\epsilon\to 0} \frac{L(u, z+\epsilon v) - L(u,z)}{\epsilon} \;=\; -A(u,v) + \langle f, v\rangle\,, \\ \partial_u L(u,z)(v) &= \lim_{\epsilon\to 0} \frac{L(u+\epsilon v, z) - L(u,z)}{\epsilon} \;=\; J(v) - A(v,z). \end{aligned}$$

In the following presentation it is very useful to use the variables $x = (u,z), x_h = (u_h, z_h) \in V \times V$. The difference will be denoted by $e = x - x_h$. Since L is bilinear, the Gateaux derivative is identical to the Fréchet derivative[2]. Obviously it holds for the solution u, the discrete solution u_h and arbitrary $z \in V$

$$\begin{aligned} \partial_z L(x)(v) &= 0 \qquad \forall v \in V, \\ \partial_z L(x_h)(v) &= 0 \qquad \forall v \in V_h. \end{aligned}$$

Now we can express the error $u - u_h$ with respect to J by using L:

$$J(u) - J(u_h) \;=\; L(x) - L(x_h) \;=\; \int_0^1 L'(x_h + \lambda e)(e)\, d\lambda.$$

Here L' denotes the derivative

$$L'(x+\lambda e)(e) \;=\; \partial_z L(x+\lambda e)(e_z) + \partial_u L(x+\lambda e)(e_u).$$

Since L is affine linear, L' is linear. Therefore, the integral above is identical to numerical quadrature with the trapezoidal rule:

$$J(u) - J(u_h) \;=\; \frac{1}{2}(L'(x)(e) + L'(x_h)(e))$$

Until now the function $z \in V$ was still arbitrary. Now we define $z \in V$ and $z_h \in V_h$ as the solutions of the dual problems:

$$\begin{aligned} z \in V: \qquad A(v,z) &= J(v) \qquad \forall v \in V\,, \\ z_h \in V_h: \qquad A(v,z_h) &= J(v) \qquad \forall v \in V_h. \end{aligned}$$

[1] René Eugéne Gateaux, 1889-1914, french mathematician, scholar of von Hadamard

[2] Maurice René Fréchet, 1878-1973, french mathematician, scholar of von Hadamard as well

For this particular z and $x = (u, z)$ we obtain

$$\partial_u L(x)(v) \;=\; 0 \qquad \forall v \in V.$$

Hence, the derivative $L'(x)(e)$ vanishes completely, $L'(x)(e) = 0$, and the error representation reduces to

$$\begin{aligned} J(u) - J(u_h) &= \frac{1}{2} L'(x_h)(e) \\ &= \frac{1}{2}(\partial_z L(x_h)(z - z_h) + \partial_u L(x_h)(u - u_h)) \\ &= \frac{1}{2}\left\{\varrho(u_h, z - z_h) + \varrho^*(z_h, u - u_h)\right\}, \end{aligned}$$

with the residuals

$$\begin{aligned} \varrho(u_h, v) &:= \partial_z L(x_h)(v) \;=\; \langle f, v\rangle - A(u_h, v)\,, \\ \varrho^*(z_h, v) &:= \partial_u L(x_h)(v) \;=\; J(v) - A(v, z_h). \end{aligned}$$

On the first glance it seems that this is not really useful, because $u - u_h$ appears on both sides of the equation. But, due to Galerkin orthogonality for conforming finite elements $V_h \subset V$, we can add arbitrary discrete test functions to the residuals ϱ and ϱ^*. We obtain the following error representation:

Theorem 6.6 *Let V be a Hilbert space, $A : V \times V \to \mathbb{R}$ a continuous bilinear form and $J \in V'$. The discretization error of the finite element solution of the variational problem (6.1) and (6.2) is given by:*

$$J(u) - J(u_h) \;=\; \frac{1}{2}\left\{\varrho(u_h, z - i_h z) + \varrho^*(z_h, u - i_h u)\right\},$$

for arbitrary projections $i_h : V \to V_h$.

Corollary 6.7 *Under the same assumptions as the previous theorem it holds*

$$J(u) - J(u_h) \;=\; \varrho(u_h, z - i_h z) \;=\; \varrho^*(z_h, u - i_h u).$$

Proof. The assertion follows directly by Theorems 6.3 and 6.6. ∎

6.3.5 Weighted residual error estimation for nonlinear problems

In this section we consider a semi-linear form $A : V \times V \to \mathbb{R}$ and a nonlinear functional $J : V \to \mathbb{R}$. Of course, A remains linear in the second argument. The corresponding dual problem is obtained by linearization of A and J at u_h:

$$z \in V : \qquad A'(u_h)(v, z) \;=\; J'(u_h)(v) \qquad \forall v \in V\,.$$

Such a semi-linear form results e.g. from a semilinear equation of the type:

$$\begin{aligned} -\epsilon\Delta u + u^2 &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{auf } \partial\Omega. \end{aligned}$$

The corresponding A becomes

$$A(u,v) := (\epsilon\nabla u, \nabla v)_{L^2(\Omega)} + (u^2, v)_{L^2(\Omega)}.$$

The Fréchet derivative needed for the dual problem reads in this case

$$A'(u)(v,z) = (\epsilon\nabla v, \nabla z)_{L^2(\Omega)} + (2uv, z)_{L^2(\Omega)}.$$

We now show the following error representation:

Theorem 6.8 *Let V be a Hilbert space, $A : V \times V \to \mathbb{R}$ a continuous and differentiable semilinearform and $J : V \to \mathbb{R}$ a continuously differentiable functional. Then it holds for the discretization error for a conforming finite elemente solution of the variational problems (6.1) snd (6.2):*

$$J(u) - J(u_h) = \frac{1}{2}\left\{\varrho(u_h, z - i_h z) + \varrho^*(z_h, u - i_h u)\right\} + R,$$

where $i_h : V \to V_h$ is an arbitrary projection and R a remainder term

$$\begin{aligned} R = \int_0^1 \Big\{ & J'''(u_h + \lambda e_u)(e_u, e_u, e_u) \\ & -\partial_{uuu} A(u_h + \lambda e_u)(e_u, e_u, e_u; z_h + \lambda e_z) p(\lambda) \Big\} d\lambda, \end{aligned}$$

with $e_u := u - u_h$.

Proof. The proof is very similar to the one of Theorem 6.6 with the following modification, because the application of the trapezoidal rule is not any more exact:

$$J(u) - J(u_h) = \int_0^1 L'(x_h + \lambda e)(e)\, d\lambda.$$

We use the polynomial $p(\lambda) = \frac{1}{2}\lambda(\lambda - 1)$. Integration by parts yields due to $p'' \equiv 1$, $p'(0) = -\frac{1}{2}$, $p'(1) = \frac{1}{2}$ and $p(0) = p(1) = 0$ to:

$$\begin{aligned} J(u) - J(u_h) &= \int_0^1 L'(x_h + \lambda e)(e) p''(\lambda)\, d\lambda \\ &= L'(x_h + \lambda e)(e) p'(\lambda)\big|_{\lambda=0}^{\lambda=1} - \int_0^1 L''(x_h + \lambda e)(e, e) p'(\lambda)\, d\lambda \\ &= \frac{1}{2}\{L'(x)(e) + L'(x_h)(e)\} - L''(x_h + \lambda e)(e, e) p(\lambda)\big|_{\lambda=0}^{\lambda=1} + \\ &\quad \int_0^1 L'''(x_h + \lambda e)(e, e, e) p(\lambda)\, d\lambda \\ &= \frac{1}{2}\{L'(x)(e) + L'(x_h)(e)\} + R, \end{aligned}$$

with

$$R := \int_0^1 L'''(x_h + \lambda e)(e,e,e)p(\lambda)\, d\lambda.$$

□

6.4 Strategies for local mesh refinement

In this section we address the question how to adapt a given mesh on basis of local error indicators η_T in order to obtain a sequence of efficient triangulations. Let us suppose that we have information of the form

$$|J(u) - J(u_h)| \leq \eta(u_h) = \sum_{T\in\mathcal{T}_h} \eta_T.$$

The local contributions η_T depend on u_h. The question is how to obtain an adapted mesh $\mathcal{T}_{h'}$, such that the resulting error on that mesh is smaller than the one before, but it should also contain only a few (or moderate) number of new cells. This process will be repeated until a given tolerance TOL is archived, i.e. $\eta(u_h) \leq TOL$. This is the so-called *stopping criterion*. Therefore we should select certain cells of the mesh $\mathcal{T}_h$ for refinement. It is reasonable that an efficient strategy equilibrates the error among all cells. Hence, we should aim to equilibrate all local error indicators which are considered as representatives of the local error contributions:

$$\min_{T\in\mathcal{T}_h} \eta_T \approx \max_{T\in\mathcal{T}_h} \eta_T.$$

In the ideal case all error indicators are exactly the same.

In the following refinement strategies we will always select those cells for refinement with the biggest values of the error indicators. Therefore, we consider the cells ordered according to the sizes of η_T:

$$\eta_{T_1} \leq \eta_{T_2} \leq \ldots \leq \eta_{T_N}.$$

It is reasonable to choose those cells which have the largest contributions to the error. Hence, we ask for the number $r \in \mathbb{N}$, $0 \leq r \leq N$, and refine the cells T_1 to T_r.

6.4.1 Error equilibration

To equilibrate the local error contributions we set as target value

$$\eta^* := \frac{Tol}{N},$$

where N denotes the number of elements and Tol the given tolerance to be archived. If all cells have the same contributions η^*, the total error would just be Tol. A cell T will be refined, if holds

$$\eta_T \quad > \quad \alpha\eta^*,$$

with a 'tuning' parameter $\alpha > 0$. The choice of this parameter is of course not trivial. A small value of α leads to a relatively large number of cells to be refined in each adaptation step. An disadvantage is that the number of selected cells can vary a lot. However, for $\alpha \approx 1$ we have the principal possibility to equilibrate the errors: an equilibrated mesh which does not yet satisfy the stopping criterion will be globally refined. This is in this case a reasonable procedure.

6.4.2 Fixed-fraction

The principal idea here is that

(a) the number of cells is increased by a certain rate, or

(b) the number of cells is determined by the requirement that the sum of the corresponding error indicators is a certain fraction of the total error.

In case (a) the refinement criterion reads

$$\text{(a) Refine } T_1, \ldots, T_r \text{ mit } r = \texttt{int}(\alpha N).$$

Her, `int` denotes the integer value and $0 < \alpha < 1$ is a free parameter. The case (b) reads

$$\text{(b) Refine } T_1, \ldots, T_r \text{ with } r \in \mathbb{N} \text{ in such away that } \sum_{i=1}^{r} \eta_i \leq \alpha\eta < \sum_{i=1}^{r+1} \eta_i.$$

$0 < \alpha < 1$ is also a free parameter. Both method usually do not lead to equilibrated meshes.

6.5 Adaptivity with quadrilateral elements

For the use of quadrilateral (or hexahedral elements in 3D) in combination with local mesh refinement we have two principal possibilities:

The first possibility is the combination of quadrilateral and tetrahedral elements in order to avoid the appearance of hanging nodes, see Fig. 6.2. The resulting finite element space remains H^1-conforming, because the finite element functions are still globally continuous. The numerical realization (implementation) of this procedure is of cause very involved, because at least two types of finite elements (Q_1 and P_1) are needed to be implemented.

The second possibility for quadrilateral meshes is to allow such kind of nodes as geometrical identities, but modify the corresponding finite elements in such a way that the finite element functions remain globally continuous. This means for Q_1-elements that we do not have degrees of freedom at these hanging nodes. 6.2. The red marked nodes are hanging nodes on which the solution (and all ansatz and test functions) are interpolated by its values on the adjacent green nodes. This procedure is in 3D with hexahedral elements very similar.

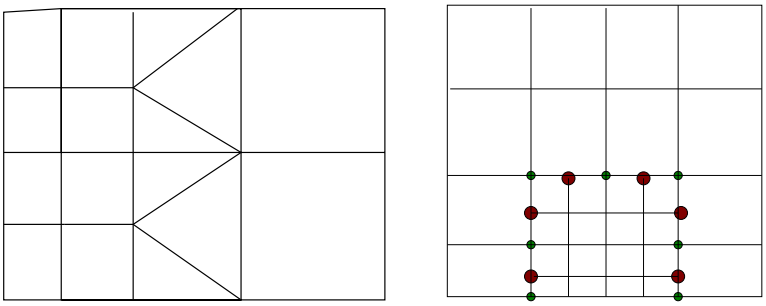


Figure 6.2: Left: Mixing of quadrilateral and triangular elements. Right: Permitting the occurence of hanging nodes on a quadrilateral mesh.

Chapter 7

Algebraic properties of the stiffness matrix

7.1 Lemma of Stampacchia

Lemma 7.1 *Let $\Omega \subset \mathbb{R}$ be open and bounded, $1 \leq p < \infty$, and $\varphi \in C^1(\mathbb{R})$ with $\|\varphi'\|_{L^\infty(\mathbb{R})} < \infty$. Then the chain rule reads*

$$\nabla(\varphi \circ u) \;=\; \varphi'(u)\nabla u \quad \forall u \in W^{1,p}(\Omega).$$

Proof. We choose a $C^1(\Omega)$-sequence with the convergence $u_m \to u$ in $W^{1,p}(\Omega)$. This implies the point-wise convergence $(\varphi \circ u_m)(x) \to (\varphi \circ u)(x)$ for a.e. $x \in \Omega$ and by the inequality of Minkowski a uniform bound of the following L^p-norm ($c := \|\varphi'\|_{L^\infty(\mathbb{R})}$, $d := |\varphi(0)|$):

$$\begin{aligned} \|\varphi \circ u_m\|_{L^p(\Omega)} &= \left(\int_\Omega |\varphi(u_m(x))|^p\right)^{1/p} \leq \left(\int_\Omega (d + c|u_m(x)|)^p\right)^{1/p} \\ &= \|d + cu_m\|_{L^p(\Omega)} \leq d|\Omega|^{1/p} + c\|u_m\|_{L^p(\Omega)} \leq C. \end{aligned}$$

These properties are sufficient to imply the convergence

$$\varphi \circ u_m \;\to\; \varphi \circ u \qquad \text{in } L^p(\Omega).$$

The bound $\|\varphi' \circ u_m - \varphi' \circ u\|_{L^p(\Omega)} \leq \|\varphi'\|_{L^\infty(\mathbb{R})}\|u_m - u\|_{L^p(\Omega)} \to 0$ yields

$$\varphi' \circ u_m \;\to\; \varphi' \circ u \qquad \text{in } L^p(\Omega).$$

This implies the weak convergence

$$\nabla(\varphi \circ u_m) \;\rightharpoonup\; \nabla(\varphi \circ u) \qquad \text{in } \mathcal{D}'(\Omega). \tag{7.1}$$

Moreover, we obtain the strong convergence

$$\varphi'(u_m)\nabla u_m \;\to\; \varphi'(u)\nabla u \qquad \text{in } L^p(\Omega), \tag{7.2}$$

as a consequence of the boundedness

$$\|\varphi'(u_m)\nabla u_m\|_{L^p(\Omega)} \;\le\; \|\varphi'\|_{L^\infty(\Omega)}\|\nabla u_m\|_{L^p(\Omega)} \;<\; \infty,$$

and the point-wise convergence (a.e.)

$$\begin{aligned} & |(\varphi'(u_m)\nabla u_m - \varphi'(u)\nabla u)(x)| \\ \le\; & |\underbrace{(\varphi'(u_m(x)) - \varphi'(u(x)))}_{\to 0}|\underbrace{\nabla u_m(x)}_{\to u(x)}| + |(\varphi'(u(x))||\underbrace{\nabla(u_m - u)(x)}_{\to 0}| \\ \to\; & 0, \end{aligned}$$

for a.e. $x \in \Omega$ and a suitable subsequence. The left hand sides of (7.1)-(7.2) are identical due to the chain rule for $C^1(\Omega)$-functions:

$$\nabla(\varphi \circ u_m) \;=\; \varphi'(u_m)\nabla u_m.$$

Hence, the limits in (7.1)-(7.2) are identical in the sense of L^p-functions:

$$\nabla(\varphi \circ u) \;=\; \varphi'(u)\nabla u.$$

□

We now define the positive and negative part of a function $u \in W^{1,p}(\Omega)$:

$$u_+(x) \;:=\; \begin{cases} u(x), & \text{if } u(x) > 0, \\ 0 & \text{else.} \end{cases}$$

$$u_-(x) \;:=\; \begin{cases} 0, & \text{if } u(x) > 0, \\ u(x) & \text{else.} \end{cases}$$

We furthermore use

$$\mathbb{1}_{u>0}(x) \;:=\; \begin{cases} 1, & \text{if } u(x) > 0, \\ 0 & \text{else.} \end{cases}$$

The functions $\mathbb{1}_{u\ge 0}$ and $\mathbb{1}_{u<0}$ are defined analogously.

Lemma 7.2 (Lemma of Stampacchia) *Let $\Omega \subset \mathbb{R}$ open, $1 \le p < \infty$, and $u \in W^{1,p}(\Omega)$. Then it holds $u_+, u_- \in W^{1,p}(\Omega)$ and*

$$\nabla u_+ \;=\; \nabla u \cdot \mathbb{1}_{u\ge 0} \;=\; \nabla u \cdot \mathbb{1}_{u>0}.$$

Proof. (a) We consider the function

$$\theta(t) \;:=\; \begin{cases} t & \text{for } t > 0 \\ 0 & \text{else.} \end{cases}$$

and its C^1-regularization for $\epsilon > 0$:

$$\theta_\epsilon(t) \;:=\; \begin{cases} \sqrt{t^2+\epsilon^2} - \epsilon & \text{for } t > 0 \\ 0 & \text{else.} \end{cases}$$

We obviously obtain the limits for $\epsilon \to 0$:

$$\begin{aligned} \lim_{\epsilon\to 0} \|\theta - \theta_\epsilon\|_{L^\infty(\mathbb{R})} &= 0 \\ \lim_{\epsilon\to 0} \|\mathbb{1}_{x>0} - \theta'_\epsilon\|_{L^\infty(\mathbb{R})} &= 0. \end{aligned}$$

This yields for $\epsilon \to 0$ the point-wise convergence:

$$\begin{aligned} (\theta_\epsilon \circ u)(x) &\to (\theta \circ u)(x) \\ (\theta'_\epsilon \circ u)(x)\nabla u(x) &\to (\theta' \circ u)\nabla u(x) \end{aligned}$$

In combination with the boundedness of $\|(\theta'_\epsilon \circ u)\nabla u\|_{L^p(\Omega)}$ we obtain the strong convergence

$$(\theta'_\epsilon \circ u)\nabla u \;\to\; (\theta' \circ u)\nabla u \quad \text{in } L^p(\Omega)$$

and the weak convergence

$$\nabla(\theta_\epsilon \circ u) \;\rightharpoonup\; \nabla(\theta \circ u) \qquad \text{in } \mathcal{D}'(\Omega).$$

It holds $u_+ = \theta \circ u$ and by Lemma 7.1:

$$\nabla(\theta_\epsilon \circ u) \;=\; \theta'_\epsilon(u)\nabla u.$$

The two limits must be identical (in the sense of L^p-functions):

$$\nabla u_+ \;=\; (\theta' \circ u)\nabla u \;\in\; L^p(\Omega).$$

This means $u_+ \in W^{1,p}(\Omega)$. The final form for ∇u_+ is obtained due to $\theta' \circ u = \mathbb{1}_{u>0}$.
(b) The negative part u_- can be expressed by $u_- = -(-u)_+ \in W^{1,p}(\Omega)$ and

$$\begin{aligned} \nabla u_- &= -\nabla((-u)_+) \;=\; -(\mathbb{1}_{-u>0})\nabla(-u) \;=\; (\mathbb{1}_{u<0})\nabla u \\ &= \nabla u \cdot \mathbb{1}_{u<0}. \end{aligned}$$

(c)

$$\begin{aligned} (\mathbb{1}_{u<0} + \mathbb{1}_{u=0} + \mathbb{1}_{u>0})\nabla u &= \nabla u \;=\; \nabla(u_+ + u_-) \\ &= \nabla u \cdot (\mathbb{1}_{u>0} + \mathbb{1}_{u<0}) \end{aligned}$$

This implies $\mathbb{1}_{u=0}\nabla u = 0$ and $\nabla u_+ = \mathbb{1}_{u\geq 0}\nabla u$. ∎

7.2 Maximum principles

7.2.1 Maximum principle for the infinite dimensional problem

Let $\Omega \subset \mathbb{R}^d$ be a domain and $L : H^1(\Omega) \to H^{-1}(\Omega)$ be a linear operator. We consider the boundary value problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \qquad (7.3)\\ u &= g \quad \text{on } \partial\Omega \qquad (7.4) \end{aligned}$$

Definition 7.3 *A linear differential operator $L : H^1(\Omega) \to H^{-1}(\Omega)$ satisfies the* maximum principle*, if the following implication holds: Let $f \in H^{-1}(\Omega)$ with $f \leq 0$, and $g \in H^1(\Omega)$ bounded by a constant $m \in \mathbb{R}$, $m \geq 0$, i.e.*

$$g \leq m \text{ a.e. in } \Omega.$$

Then every solution $u \in H^1(\Omega)$ of (7.3)-(7.4) satisfies

$$u \leq m \quad \text{a.e. in } \Omega.$$

In this definition, the inequality $f \leq 0$ for $f \in H^{-1}(\Omega)$ means:

$$\langle f, \varphi \rangle \leq 0 \text{ for all } \varphi \in H_0^1(\Omega) \text{ with } \varphi \geq 0.$$

Corollary 7.4 *Solutions of (7.3)-(7.4) with L satisfying the maximum principle are always unique.*

Proof. Let $u_1, u_2 \in H^1(\Omega)$ be two solutions of (7.3)-(7.4). Then $w := u_1 - u_2 \in H_0^1(\Omega)$ is solution of the homogeneous equation

$$Lw = 0 \quad \text{in } \Omega.$$

Due to the maximum principle holds $w \leq 0$ a.e. in Ω. The same arguments for $-w$ yields $-w \leq 0$ a.e., and hence $w \equiv 0$ in $L^2(\Omega)$. ∎

Theorem 7.5 *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz-domain, $A = (a_{ij})$ be a positive definite matrix with coefficients $a_{ij}, c \in L^\infty(\Omega)$, $c \geq 0$. Then the differential operator*

$$Lu := div(A\nabla u) + cu \qquad (7.5)$$

satisfies the maximum principle.

Proof. We consider $\varphi := (u - m)_+$. Due to the Lemma von Stampacchia it holds $\varphi \in H^1(\Omega)$. Moreover $\varphi \geq 0$ and $\varphi|_{\partial\Omega} = (g - m)_+ \leq 0$. Hence $\varphi \in H_0^1(\Omega)$ is a suitable test function:

$$0 \geq \langle f, \varphi \rangle = \langle Lu, \varphi \rangle = \int_\Omega (A\nabla u \nabla \varphi + cu\varphi) dx$$

Me may write φ also in the form $\varphi = (u-m)\chi$ with $\chi := \mathbb{1}_{u>m}$. Hence the equation above can be expressed as

$$\begin{aligned} 0 &\geq \int_\Omega (A\nabla u \nabla[(u-m)\chi] + cu(u-m)\chi)dx \\ &= \int_\Omega (A\nabla[(u-m)\chi]\nabla[(u-m)\chi] + c(u-m)\chi(u-m)\chi)dx + \int_\Omega cm\chi(u-m)\chi dx \\ &= \int_\Omega (A\nabla\varphi\nabla\varphi) + c\varphi^2 dx + \int_\Omega \underbrace{c}_{\geq 0}\underbrace{m}_{\geq 0}\underbrace{\chi}_{\geq 0}\underbrace{\varphi}_{\geq 0} dx \\ &\geq 0 + 0 \;=\; 0. \end{aligned}$$

Now we deduce that $A\nabla\varphi\nabla\varphi = 0$. Due to the assumed positivity of A it follows $\nabla\varphi = 0$ and by Poincare $\varphi = 0$. This implies $u \leq m$. ∎

7.2.2 Discrete maximum principle

Definition 7.6 *Let $V_h \subset H_0^1(\Omega)$ a finite dimensional subspace, and $L_h : V_h \to V_h'$ a linear operator. L_h satisfies the* discrete maximum principle*, if the following implication holds: Let $f_h \in V_h'$ with $f_h \leq 0$, then every solution $u_h \in V_h$ of $L_h u_h = f_h$ satisfies $u_h \leq 0$.*

This property is closely related to the M-matrix property:

Definition 7.7 *A quadratic matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is called* M-matrix,[1] *if A is regular, all off-diagonal components are non-positive ($a_{ij} \leq 0$ for $i \neq j$), and $A^{-1} \geq 0$ (componentwise).*

An immediate consequence of the M-matrix property is the discrete inverse monotone property of the solution x of $Ax = b$:

$$b \leq 0 \quad \Rightarrow \quad x \leq 0.$$

Lemma 7.8 *Let V_h be a P_1 finite element space, $L_h : V_h \to V_h'$ a linear operator and A be the stiffness matrix corresponding to the Lagrange basis of V_h. If A is a M-matrix, then L_h satisfies the discrete maximum principle.*

Proof. Let $L_h u_h = f_h \leq 0$ and $\{\phi_1, \ldots, \phi_n\}$ be the Lagrange basis. The corresponding linear system reads $Ax = b \in \mathbb{R}^n$, with $b_i = (f_h, \phi_i)$ and $u_h = \sum_i x_i\phi_i$. It is sufficient to show that u_h is non-positive at the nodal points. In other words: we shall verify that $x \leq 0$. Since A is a M-matrix, it holds $x = A^{-1}b$. Furthermore, $f_h \leq 0$ implies $b \leq 0$. For arbitrary $i \in \{1, \ldots, n\}$:

$$x_i \;=\; (A^{-1}b)_i \leq 0.$$

∎

[1] The name M-matrix is related to the german mathematician Hermann Minkowski, 1864-1909.

Lemma 7.9 *The following three conditions are sufficient for a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ to be a M-matrix:*

(*a*) *A is* weakly diagonal-dominant*, i.e.*

$$\begin{aligned} |a_{ii}| &\geq \sum_{j\neq i} |a_{ij}| \qquad \forall i \in \{1,\dots,n\} \\ |a_{ii}| &> \sum_{j\neq i} |a_{ij}| \qquad \textit{for at least one } i \in \{1,\dots,n\}. \end{aligned}$$

(*b*) *A is* irreducibel*, i.e. it does not exist a permutation matrix $P \in \{0,1\}^{n\times n}$, s.t. $P^T AP$ is a block-triangular matrix of the following form*

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

with quadratic matrices A_{11}, A_{22}, and

(*c*) *A is of* non-negative type*, , i.e. all off-diagonal elements are non-positive and all diagonal elements are positive.*

The linear equation system based on irreducible matrices cannot be transformed into two smaller systems which can be solved sequentially (of type $A_{22}x_2 = b_2$ and $A_{11}x_1 = b_1 - A_{12}x_2$).

Proof. It is possible to show that quadratic matrices $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ are already regular, if they are irreducible and weakly diagonal-dominant (see [11]). Hence, it exists A^{-1}. We write the matrix in the form $A = D + L + R$, with the diagonal matrix D, a lower triangular matrix L and an upper triangular matrix R. Since A is of non-negative type, it holds $D > 0$, $L, R \leq 0$. Therefore the Jacobi iteration matrix is non-negative, $J := -D^{-1}(L + R) \geq 0$, and also every power $J^k \geq 0$ for $k \in \mathbb{N}$. It is well-known (due to weakly diagonal-dominance) that this iteration matrix has only eigenvectors λ with $|\lambda| < 1$. Hence, for its spectral norm holds $\|J\|_2 < 1$ and the Neuman series of J converges. This implies

$$\begin{aligned} A^{-1}D &= (D^{-1}A)^{-1} = (I + D^{-1}(L+R))^{-1} = (I-J)^{-1} \\ &= \sum_{k=0}^{\infty} J^k \geq 0. \end{aligned}$$

This implies $A^{-1}D \geq 0$. Due to $D > 0$, we obtain $A^{-1} \geq 0$. ∎

We already know that for indices $i \neq j$, which do not correspond to a common triangle T, the associated entry α_{ij} in the stiffness matrix vanishes, because the intersection of the corresponding Lagrange basis functions ψ_i and ψ_j is a Lebesgue null set. Therefore, we now consider indices $i \neq j$, which correspond to a common triangle T of the triangulation. The angle $\alpha_{T,i,j}$ is that one built by those edges which do not contain the nodes x_i and x_j as vertices, see Fig. 7.1.

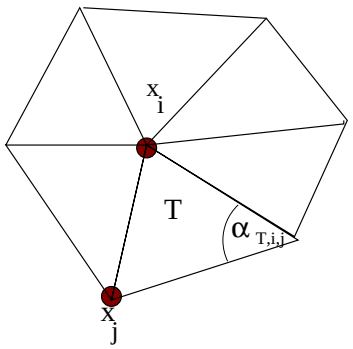


Figure 7.1: Notations for the proof of the M-matrix property.

Lemma 7.10 *We consider the Lagrange basis $\{\psi_1, \ldots, \psi_N\}$ of P_1-elements. Let $i \neq j$ two indices corresponding to a common triangle T of the triangulation and $\alpha_{T,i,j}$ the associated angle to the edges. Then it holds*

$$\int_T \nabla\psi_j \nabla\psi_i \, dx \quad = \quad -\|\nabla\psi_i\|_{L^2(T)} \|\nabla\psi_j\|_{L^2(T)} \cos\alpha_{T,i,j}.$$

In particular, the left-hand side is non-positive, if $-\pi/2 \leq \alpha_{T,i,j} \leq \pi/2$.

Proof. For P_1-elements the gradients $\nabla\psi_j$ and $\nabla\psi_i$ are constant on T. Let n_i be the normal vector of the edge of T, which does not contain the vertex x_i. Then it holds

$$\nabla\psi_i \quad = \quad -\|\nabla\psi_i\|_{L^2(T)} \cdot n_i^T |T|^{-1/2},$$

and accordingly $\nabla\psi_j = -\|\nabla\psi_j\|_{L^2(T)} \cdot n_j^T |T|^{-1/2}$. This implies

$$\int_T \nabla\psi_j \nabla\psi_i \, dx \quad = \quad \|\nabla\psi_i\|_{L^2(T)} \|\nabla\psi_j\|_{L^2(T)} \langle n_j, n_i \rangle.$$

The assertion now follows due to $\langle n_j, n_i \rangle = \cos(\pi - \alpha_{T,i,j}) = \cos(\alpha_{T,i,j} - \pi) = -\cos(\alpha_{T,i,j})$. □

Due to this result we will consider triangulations, where all arising angles are acute. This will give us the inverse monotonicity on the discrete level.

Definition 7.11 *A triangulation $\mathcal{T}_h$ of a domain $\Omega \subset \mathbb{R}^2$ is called* weakly acute, *if all arising angles of the elements are $\leq \pi/2$.*

Theorem 7.12 *Let $\mathcal{T}_h$ be a weakly acute triangulation of a polygonal bounded domain Ω with the additional property that each element has at least one vertex which does not belong to the boundary $\partial\Omega$. Then the associated stiffness matrix of the Laplace operator $-\Delta$ discretized by P_1 finite elements is a M-matrix and the corresponding discrete operator satisfies the discrete maximum principle.*

Proof. We show that the stiffness matrix A is irreducible, weakly diagonal-dominant and of non-negative type. With Lemma 7.9 follows the M-matrix property.
(a) The irreducibility of A follows from the condition that each element of $\mathcal{T}_h$ has at least one vertex which is an inner node and the property that Ω is connected (Exercise!).
(b) The previous lemma ensures that A is of non-negative type.
(c) It remains to verify the weak diagonal-dominance. To this end we denote the Lagrange basis by $\{\psi_j : j = 1, \dots, N\}$. Let $i \in \{1, \dots, N\}$ be arbitrary but fixed, and the corresponding support $\omega := \operatorname{supp} \psi_i$. We differentiate between two cases of indices:
Case1: The index i corresponds to an inner vertex, such that all elements which contain this vertex do not intersect the boundary. Then it holds for $\varphi := \sum_{j=1}^N \psi_j$, $\varphi|_\omega = 1$, and for the vector $\mathbb{1} = (1, \dots, 1)^T$:

$$\sum_{j=1}^N a_{ij} \;=\; (A\mathbb{1})_i \;=\; \left(\nabla \sum_{j=1}^N \psi_j, \nabla\psi_i\right)_{L^2(\Omega)} \;=\; (\nabla\varphi, \nabla\psi_i)_{L^2(\omega)} \;=\; 0.$$

This yields

$$|a_{ii}| \;=\; \left|\sum_{j\neq i} \underbrace{a_{ij}}_{\leq 0}\right| \;=\; \sum_{j\neq i} |a_{ij}|.$$

Case 2: Let i be an index to a boundary triangle. Let $N+1, \dots, N+r$ be the nodal indices of the boundary vertices which lay in ω. By augmentation of the test functions of the boundary vertices we obtain for $x \in \omega$:

$$\sum_{j=1}^{N+r} \psi_j(x) \;=\; 1.$$

Therefore, the previous Lemma yields to:

$$\begin{aligned}\sum_{j=1}^N a_{ij} \;&=\; \left(\nabla \sum_{j=1}^N \psi_j, \nabla\psi_i\right)_{L^2(\Omega)} \;=\; -\left(\nabla \sum_{j=N+1}^{N+r} \psi_j, \nabla\psi_i\right)_{L^2(\Omega)} \\ &=\; -\sum_{j=N+1}^{N+r} (\nabla\psi_j, \nabla\psi_i)_{L^2(\Omega)} \;=\; -\sum_{j=N+1}^{N+r} a_{ij} \;\geq\; 0.\end{aligned}$$

Furthermore, at most one angle in T is of size $\pi/2$, so that we find a boundary index j such that $\alpha_{T,i,j} < \pi/2$ and $a_{ij} < 0$. Therefore, we obtain

$$\sum_{j=1}^N a_{ij} \;>\; 0,$$

and hence

$$|a_{ii}| \;=\; a_{ii} \;=\; \sum_{j=1}^{N} a_{ij} - \sum_{j\neq i} a_{ij} \;>\; -\sum_{j\neq i} a_{ij} \;=\; \sum_{j\neq i} |a_{ij}|.$$

□

7.3 Condition number

We know that the spectral condition number $\kappa(A)$ of a symmetric, positive definite matrix $A \in \mathbb{R}^{n\times n}$ is given by the quotient of the maximal and minimal eigenvalue λ_{max} and λ_{min}, respectively:

$$\kappa(A) \;=\; \frac{\lambda_{max}}{\lambda_{min}}.$$

This condition number is important to determine the convergence rate of linear solvers. E.g., the linear convergence rate of the gradient method is bounded by

$$\rho_{grad} \;=\; \frac{\kappa(A)-1}{\kappa(A)+1},$$

while the bound for the conjugate gradient (cg) method is

$$\rho_{cg} \;=\; \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}.$$

The following theorem considers the condition number for a P_1 discretization of the one-dimensional Laplace operator:

Theorem 7.13 *The stiffness matrix A_h of P_1-finite elements for the one-dimensional Laplace operator is symmetric positive definite and the condition number is bounded as*

$$\kappa(A_h) \;\leq\; 12C_p^2 h_{min}^{-2},$$

where $h_{min} = \min\{h_k : k = 1,\dots,n\}$ and C_p is the Poincaré constant.

Proof. Without loss of generality we consider $\Omega = (0,1)$.
(a) The symmetry of $A_h = (a_{ij})$ follows by

$$a_{ij} \;=\; (\phi_j', \phi_i')_{L^2(0,1)} = a_{ji}.$$

(b) The largest and smallest eigenvalues $\lambda_{max}, \lambda_{min}$ can be bounded by the Rayleigh quotients:

$$\min_{0\neq x\in\mathbb{R}^{n_h}} \frac{(A_h x, x)_{l_2}}{\|x\|_{l_2}^2} \;\leq\; \lambda_{min} \;\leq\; \lambda_{max} \;\leq\; \max_{0\neq x\in\mathbb{R}^{n_h}} \frac{(A_h x, x)_{l_2}}{\|x\|_{l_2}^2}.$$

The assertion follows by showing:

$$\begin{aligned} \min_{0\neq x\in\mathbb{R}^{n_h}} \frac{(A_h x, x)_{l_2}}{\|x\|_{l_2}^2} &\geq \frac{1}{3C_p^2} h_{min}, \\ \max_{0\neq x\in\mathbb{R}^{n_h}} \frac{(A_h x, x)_{l_2}}{\|x\|_{l_2}^2} &\leq 4h_{min}^{-1}. \end{aligned}$$

The nominators in the Rayleigh quotients become (with the convention $x_0 = x_n := 0$):

$$\begin{aligned} (A_h x)_i &= -h_i^{-1}x_{i-1} + (h_i^{-1} + h_{i+1}^{-1})x_i - h_{i+1}^{-1}x_{i+1}, \\ (A_h x, x)_{l_2} &= \sum_{i=1}^{n-1} (A_h x)_i x_i \\ &= \sum_{i=1}^{n-1} \left(-h_i^{-1}x_{i-1}x_i + (h_i^{-1} + h_{i+1}^{-1})x_i^2 - h_{i+1}^{-1}x_i x_{i+1}\right) \\ &= -2\sum_{i=1}^{n} h_i^{-1}x_{i-1}x_i + \sum_{i=1}^{n-1}(h_i^{-1} + h_{i+1}^{-1})x_i^2. \end{aligned}$$

The upper bound of the maximal Rayleigh quotient is obtained as follows:

$$\begin{aligned} |(A_h x, x)_{l_2}| &\leq \sum_{i=1}^{n} h_i^{-1}(x_{i-1}^2 + x_i^2) + \sum_{i=1}^{n-1}(h_i^{-1} + h_{i+1}^{-1})x_i^2 \\ &\leq 2\sum_{i=1}^{n-1}(h_i^{-1} + h_{i+1}^{-1})x_i^2 \\ &\leq 4h_{min}^{-1}\|x\|_{l_2}^2 . \end{aligned}$$

For the lower bound of the minimal Rayleigh quotient we use the presentation $u_h = \sum_{i=1}^{n-1} x_i\phi_i$ and use the Poincaré inequality (Thm. 3.12) with the mass matrix M_h:

$$(A_h x, x)_{l_2} = A(u_h, u_h) = |u_h|_{H^1(0,1)}^2 \geq C_p^{-2}\|u_h\|_{L^2(0,1)}^2 = C_p^{-2}(M_h x, x)_{l_2}.$$

Therefore it is sufficient to show

$$(M_h x, x)_{l_2} \geq \frac{1}{3} h_{min}\|x\|_{l_2}^2.$$

We use the form of the mass matrix in Section 4.2.3:

$$\begin{aligned}
(M_h x)_i &= \frac{1}{6}\left(h_i x_{i-1} + 2(h_i + h_{i+1})x_i + h_{i+1}x_{i+1}\right), \\
(M_h x, x)_{l_2} &= \frac{1}{6}\sum_{i=1}^{n-1}\left(h_i x_{i-1}x_i + 2(h_i + h_{i+1})x_i^2 + h_{i+1}x_{i+1}x_i\right) \\
&\geq \frac{1}{6}\sum_{i=1}^{n-1}\left(-\frac{1}{2}h_i x_{i-1}^2 - \frac{1}{2}h_i x_i^2 + 2(h_i + h_{i+1})x_i^2 - \frac{1}{2}h_{i+1}x_{i+1}^2 - \frac{1}{2}h_{i+1}x_i^2\right) \\
&= \frac{1}{6}\sum_{i=1}^{n-1}(h_i + h_{i+1})x_i^2 \\
&\geq \frac{1}{3}h_{min}\|x\|_{l_2}^2.
\end{aligned}$$

(c) The positive definite property is a result of the coercivity of the bilinear form $A(\cdot,\cdot)$. □

Chapter 8

Crouzeix-Raviart element

Until now, the discrete finite element space was always conforming, i.e. $V_h \subset V$. In this section we consider an example of a non-conforming method for the Poisson problem.

8.1 Definition of the Crouzeix-Raviart element

Let $\mathcal{T}_h$ be an admissible triangulation of triangles. Let M_h be the set of all edge middle points of inner edges and N_h be the set of all edge middle points of edges laying on the boundary $\partial\Omega$. We consider the linear space of all linear polynomials without the requirement of global continuity. We only assume continuity on the edge middle points:

$$\mathcal{CR}_h \;:=\; \{v \in L^2(\Omega) \,:\, v|_T \in P_1\,,\; v|_{M_h} \text{ continuous and } v|_{N_h} = 0\}\,.$$

These elements are called *Crouzeix-Raviart elements*. The discrete variational problem now reads

$$u_h \in \mathcal{CR}_h : \qquad A_h(u_h, v) \;=\; \langle f, v\rangle \qquad \forall v \in \mathcal{C}R_h, \tag{8.1}$$

where the bilinear form A_h is given by

$$A_h(u, v) \;:=\; \sum_{T\in\mathcal{T}_h} (\nabla u, \nabla v)_{L^2(T)}. \tag{8.2}$$

This bilinear form can be applied to elements in $\mathcal{CR}_h$ but also on $V := H_0^1(\Omega)$. But we keep in mind that $\mathcal{CR}_h \not\subset V$.

Theorem 8.1 *The discretization of the Poisson problem with the (lowest order) Crouzeix-Raviart element (8.1) has always for every $f \in H^{-1}(\Omega)$ a unique solution.*

Proof. We will again apply the Lax-Milgram theorem. We use the scalar product $(u, v)_h := A_h(u, v)$ in $\mathcal{CR}_h$. Then A_h is $\mathcal{C}R_h$-continuous and coercive with constants $\alpha_1 = \alpha_2 = 1$. ∎

Before we analyze the discretization error of this method a more general Lemma is required.

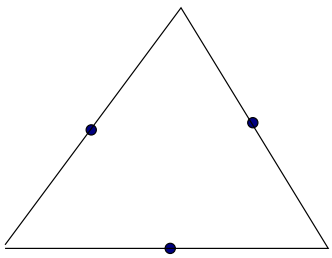


Figure 8.1: Crouzeix-Raviart element; a the marked edge middle points the FE function is continuous.

8.2 Lemma of Strang

Let V and W_h be Hilbert spaces with a h-dependent norm $\|\cdot\|_{W_h}$. We consider a discrete problem of the form

$$u_h \in W_h : \qquad A_h(u_h, v_h) \;=\; \langle f_h, v_h\rangle \qquad \forall v_h \in W_h. \tag{8.3}$$

The bilinear form A_h and the right hand side $f_h \in W_h'$ may depend on h. Furthermore we assume that $A_h(\cdot,\cdot)$ is defined for $u \in V$ as first argument. To be more precise we asume to have a linear mapping

$$\begin{aligned} A_h : (W_h \oplus V) \times W_h &\to \mathbb{R} \\ (w_h + u, v_h) &\mapsto A_h(w_h + u, v_h). \end{aligned}$$

Here, W_h and V are considered to be linear subspaces of a larger linear space, where the direct sum $\oplus$ is well-defined. We ask for an upper bound of the discretization error for u_h and the solution $u \in V$:

$$u \in V : \qquad A(u, v) \;=\; \langle f, v\rangle \qquad \forall v \in V. \tag{8.4}$$

We further assume that A_h is $(V \oplus W_h)$-continuous and W_h-coercive in the form:

$$\begin{aligned} |A_h(u + u_h, v_h)| &\le \alpha_1 \|u + u_h\|_{W_h} \|v_h\|_{W_h} \qquad \forall u_h, v_h \in W_h \, \forall u \in V, \\ A_h(u_h, u_h) &\ge \alpha_2 \|u_h\|_{W_h}^2 \qquad\qquad\qquad \forall u_h \in W_h. \end{aligned}$$

Here, $\alpha_1, \alpha_2 > 0$ are positive constants independent of h. The following lemma of Berger, Scott and Strang[1] makes a connection between the discretization error in W_h, the approximation error and the consistency error.

Lemma 8.2 (Berger, Scott and Strang) *Let the bilinear form A_h be W_h-coercive and $(V \otimes W_h)$-continuous. Then the following bound for the discretization error of the discrete solution $u_h \in W_h$ and (8.3) the continuous solution $u \in V$ of (8.4) holds*

$$\|u - u_h\|_{W_h} \;\le\; \left(1 + \frac{\alpha_1}{\alpha_2}\right) \inf_{v_h \in W_h} \|u - v_h\|_{W_h} + \frac{1}{\alpha_2} \sup_{v_h \in W_h} \frac{\langle f_h, v_h\rangle - A_h(u, v_h)}{\|v_h\|_{W_h}}.$$

[1] William Gilbert Strang, born 1944, US-american mathematician at the MIT

Proof. The triangle inequality leads to

$$\|u-u_h\|_{W_h} \leq \inf_{v_h\in W_h} \left(\|u-v_h\|_{W_h} + \|v_h-u_h\|_{W_h}\right).$$

We have to estimate the last arising term. Due to the W_h-coercivity and W_h-continuity it holds

$$\begin{aligned}
\alpha_2\|u_h-v_h\|_{W_h}^2 &\leq A_h(u_h-v_h,u_h-v_h)\\
&= A_h(u-v_h,u_h-v_h)+A_h(u_h-u,u_h-v_h)\\
&= A_h(u-v_h,u_h-v_h)+\langle f_h,u_h-v_h\rangle - A_h(u,u_h-v_h)\\
&\leq \alpha_1\|u-v_h\|_{W_h}\|u_h-v_h\|_{W_h}+\langle f_h,u_h-v_h\rangle - A_h(u,u_h-v_h).
\end{aligned}$$

We divide by $\alpha_2\|u_h-v_h\|_{W_h}$ and obtain

$$\|u_h-v_h\|_{W_h} \leq \frac{\alpha_1}{\alpha_2}\|u-v_h\|_{W_h} + \frac{\langle f_h,u_h-v_h\rangle - A_h(u,u_h-v_h)}{\alpha_2\|u_h-v_h\|_{W_h}}.$$

Using this in the bound above we arrive at

$$\begin{aligned}
\|u-u_h\|_{W_h} &\leq \inf_{v_h\in W_h}\left(\left(1+\tfrac{\alpha_1}{\alpha_2}\right)\|u-v_h\|_{W_h} + \frac{\langle f_h,u_h-v_h\rangle - A_h(u,u_h-v_h)}{\alpha_2\|u_h-v_h\|_{W_h}}\right)\\
&\leq \inf_{v_h\in W_h}\left(1+\tfrac{\alpha_1}{\alpha_2}\right)\|u-v_h\|_{W_h}\\
&\quad + \frac{1}{\alpha_2}\sup_{v_h\in W_h}\frac{\langle f_h,u_h-v_h\rangle - A_h(u,u_h-v_h)}{\|u_h-v_h\|_{W_h}}.
\end{aligned}$$

This gives us the assertion. ∎

The first term on the right hand side is the already know *approximation error*. The second term is called *consistency error*.

8.3 A priori error analysis of the Crouzeix-Raviart element

Following the previous section we need a norm for the space $W_h = \mathcal{CR}_h$. We define it as in the proof for existence and uniqueness of the Crouzeix-Raviart formulation, namely by

$$\|v_h\|_h := \left(\sum_{T\in\mathcal{T}_h}|v_h|_{H^1(T)}^2\right)^{1/2}.$$

Theorem 8.3 *Let $\Omega\subset\mathbb{R}^2$ be a convex or C^2-bounded, $\{\mathcal{T}_h\}$ a family of shape regular triangulations of Ω and $f\in L^2(\Omega)$. Then the discretization error for the Poisson problem discretized with the (lowest order) Crouzeix-Raviart element is bounded as*

$$\begin{aligned}
\|u-u_h\|_{L^2(\Omega)} &\leq ch^2|u|_{H^2(\Omega)},\\
\|u-u_h\|_h &\leq ch|u|_{H^2(\Omega)},
\end{aligned}$$

with a constant $c=c(\Omega,\kappa)$.

Proof. We will firstly show the second bound. The L^2-bound is obtained afterwards by a duality argument.
(a) The bilinear form A_h is continuous and coercive as requested in Strang's Lemma 8.2. Hence, we have to bound the approximation error and the consistency error (separately). For the first one we state that the continuous P_1 elements are a subset of the Crouzeix-Raviart elements, $P_{1,h} \subset \mathcal{C}R_h$. Therefore,

$$\inf_{v_h \in \mathcal{C}R_h} \|u - v_h\|_h \;\leq\; \inf_{v_h \in P_{1,h}} \|u - v_h\|_h \;\leq\; ch|u|_{H^2(\Omega)}.$$

Note that $u \in H^2(\Omega)$ is ensured due to the assumptions on Ω. The consistency error does not vanish in general, because the continuous solution $u \in V \cap H^2(\Omega)$ does not necessarily satisfy the discrete equation (8.2). For the residual it holds with $v_h \in V_h$,

$$\begin{aligned}
\rho_h(u, v_h) &:= \langle f, v_h\rangle - A_h(u, v_h) \\
&= (f, v_h)_{L^2(\Omega)} - \sum_{T\in\mathcal{T}_h} \{(-\Delta u, v_h)_{L^2(T)} + (\partial_n u, v_h)_{L^2(\partial T)}\} \\
&= \sum_{T\in\mathcal{T}_h} \{(f + \Delta u, v_h)_{L^2(T)} - (\partial_n u, v_h)_{L^2(\partial T)}\} \\
&= -\sum_{T\in\mathcal{T}_h} (\partial_n u, v_h)_{L^2(\partial T)}.
\end{aligned}$$

Because $\partial_n u$ is just changing the sign dependent from which element we consider the inner edge, it holds for functions $\phi \in L^2(\mathcal{E}_h)$, where $\mathcal{E}_h$ denotes the set of all inner edges, and ϕ vanishes on boundary edges:

$$\sum_{T\in\mathcal{T}_h} (\partial_n u, \phi)_{L^2(\partial T)} \;=\; 0.$$

Hence, we obtain for such ϕ:

$$\rho_h(u, v_h) \;=\; -\sum_{T\in\mathcal{T}_h} (\partial_n u, v_h - \phi)_{L^2(\partial T)}.$$

We choose such $\phi = \phi(v_h)$ which is constant on each inner edge $e \in \mathcal{E}_h$ with the value (M_e denotes the center point of edge e):

$$\phi|_e \;=\; \frac{1}{2|e|} \sum_{i=1}^{2} \int_e v_h|_{T_i}\, ds \;=\; v_h(M_e).$$

This choice implies for $i = 1, 2$:

$$\int_e (v_h|_{T_i} - \phi)\, ds \;=\; 0.$$

Furthermore, we denote the nodal interpolation operator to P_1-elements by $I_h : V \to P_{1,h}$. Then $(\partial_n I_h u)|_{T_i}$ is constant on the edges (but different for each T_i), so that

$$\sum_{i=1}^{2}((\partial_n I_h u)|_{T_i}, v_h|_{T_i} - \phi)_{L^2(e)} \quad = \quad 0.$$

Since this holds for all inner edges, we deduce

$$\rho_h(u, v_h) \quad = \quad -\sum_{T\in\mathcal{T}_h} (\partial_n(u - I_h u)|_T, v_h|_T - \phi)_{L^2(\partial T)}.$$

Application of the Cauchy-Schwarz inequality leads to

$$|\rho_h(u, v_h)| \quad \leq \quad \sum_{T\in\mathcal{T}_h} \|\partial_n(u - I_h u)|_T\|_{L^2(\partial T)} \|v_h|_T - \phi\|_{L^2(\partial T)}.$$

The appearing terms will be bound separately. We use the trace theorem and the transformation theorem in the form

$$\|v\|_{L^2(\partial T)} \quad \leq \quad c(h_T^{1/2}\|\nabla v\|_{L^2(T)} + h_T^{-1/2}\|v\|_{L^2(T)}) \quad \forall v \in H^1(T),$$

with a constant $c = c(\kappa)$ (depending on the anisotropy κ of the element). In combination with the Bramble-Hilbert Lemma we arrive at

$$\begin{aligned}
\|\partial_n(u - I_h u)\|_{L^2(\partial T)} &\leq \|\nabla(u - I_h u)\|_{L^2(\partial T)} \\
&\leq c(h_T^{1/2}|\nabla(u - I_h u)|_{H^1(T)} + h_T^{-1/2}\|\nabla(u - I_h u)\|_{L^2(T)}) \\
&\leq c h_T^{1/2}|u|_{H^2(T)}\,, \\
\|v_h|_T - \phi\|_{L^2(\partial T)} &\leq c h_T |v_h|_{H^1(\partial T)} \\
&\leq c h_T (h_T^{-1/2}\|\nabla v_h\|_{L^2(T)} + h_T^{1/2} \underbrace{|\nabla v_h|_{H^1(T)}}_{=0}) \\
&= c h_T^{1/2}|v_h|_{H^1(T)}.
\end{aligned}$$

We use these interpolation estimates in the bound above for the residuals:

$$\begin{aligned}
|\rho_h(u, v_h)| &\leq c\sum_{T\in\mathcal{T}_h} h_T |u|_{H^2(T)}|v_h|_{H^1(T)} \\
&\leq ch \left(\sum_{T\in\mathcal{T}_h} |u|^2_{H^2(T)}\right)^{1/2} \left(\sum_{T\in\mathcal{T}_h} |v_h|_{H^1(T)}\right)^{1/2} \\
&= ch|u|_{H^2(\Omega)}\|v_h\|_h.
\end{aligned}$$

As before, we use here the notation $h := \max\{h_T | T \in \mathcal{T}_h\}$. We arrive at

$$\sup_{v_h\in W_h} \frac{\rho_h(u, v_h)}{\|v_h\|_h} \quad \leq \quad ch|u|_{H^2(\Omega)}.$$

In summary, we see that both parts of the error, the approximation- and the consistency error are qualitatively similar bounded, so that we get the second estimate.
(b) As already mentioned, we use a duality argument.. With $B := \{g \in L^2(\Omega) : \|g\|_{L^2(\Omega)} = 1\}$ and the dual solutions $z_g \in V$ and $z_{g,h} \in W_h$ for the right hand side g it holds

$$\|u - u_h\| = \sup_{g \in B} \langle g, u - u_h \rangle.$$

$$\begin{aligned} \langle g, u - u_h \rangle &= \langle g, u - u_h \rangle - A_h(u - u_h, z_g) + A_h(u - u_h, z_g) \\ &= \langle g, u - u_h \rangle - A_h(u - u_h, z_g) + A_h(u - u_h, z_g - z_{g_h}) + A_h(u - u_h, z_{g_h}). \end{aligned}$$

We estimate these terms separately:

- The first two terms are a residual and can be treated by the techniques from part (a) of this proof: (ρ_h^* denotes the dual residual):

$$\begin{aligned} |A_h(u - u_h, z_g) - \langle g, u - u_h \rangle| &= |\rho_h^*(z_g, u - u_h)| \\ &\leq ch|z_g|_{H^2(\Omega)} \|u - u_h\|_h \\ &\leq ch^2 |z_g|_{H^2(\Omega)} |u|_{H^2(\Omega)} \\ &\leq ch^2 |u|_{H^2(\Omega)} \|g\|_{L^2(\Omega)}. \end{aligned}$$

- The third term on the right hand side can be bounded by

$$\begin{aligned} A_h(u - u_h, z_g - z_{g,h}) &\leq \|u - u_h\|_h \|z_g - z_{g,h}\|_h \\ &\leq ch^2 \|u\|_{H^2(\Omega)} \|g\|_{L^2(\Omega)}. \end{aligned}$$

- The last term can also be written as a residual term:

$$\begin{aligned} A_h(u - u_h, z_{g_h}) &= A_h(u, z_{g_h}) - A_h(u_h, z_{g_h}) \\ &= A_h(u, z_{g_h} - z_g) + A_h(u, z_g) - \langle f, z_{g_h} \rangle \\ &= A_h(u, z_{g_h} - z_g) + \langle f, z_g - z_{g_h} \rangle \\ &= |\rho_h(u, z_g - z_{g,h})| \end{aligned}$$

  This residual can be treated by the techniques from part (a) of this proof:

$$\begin{aligned} |\rho_h(u, z_g - z_{g,h})| &\leq ch|u|_{H^2(\Omega)} \|z_g - z_{g,h}\|_h \\ &\leq ch^2 |u|_{H^2(\Omega)} |z_g|_{H^2(\Omega)} \\ &\leq ch^2 |u|_{H^2(\Omega)} \|g\|_{L^2(\Omega)} \end{aligned}$$

Assembling all terms together yields the assertion:

$$\|u - u_h\| \leq ch^2 |u|_{H^2(\Omega)}.$$

□

Chapter 9

Mixed formulations and inf-sup conditions

9.1 Closed range theorem

Let E, F be two Banach spaces and $T \in \mathcal{L}(E, F)$, hence a continuous (and bounded) linear operator $T : E \to F$. The adjoint operator $T^* : F' \to E'$ is given for $f \in F'$ and $x \in E$ by

$$\langle T^* f, x\rangle \quad = \quad \langle f, Tx\rangle.$$

The corresponding kernel will be denoted by $N(T^*)$,

$$N(T^*) \quad := \quad \{f \in F' \ : \ T^* f = 0\}.$$

The orthogonal space is defined by

$$N(T^*)^\perp \quad = \quad \{y \in F \ : \ \langle f, y\rangle = 0 \ \forall f \in N(T^*)\}.$$

In the following be use a central result of functional analysis. For a proof we refer to [3].

Theorem 9.1 (Closed range theorem) *Let E, F be two Banach spaces and $T \in \mathcal{L}(E, F)$. Then $T(E)$ is closed in F, iff $T(E) = N(T^*)^\perp$.*

Furthermore, we will use the following easy result:

Lemma 9.2 *Let E, F be two Banach spaces and $T \in \mathcal{L}(E, F)$ injective. Furthermore we assume that $T^{-1} : T(E) \to E$ is continuous. Then $T(E)$ is closed.*

Proof. Let $(y_n)_{n\in\mathbb{N}} \subseteq T(E)$ a convergent sequence with $y_n \to y \in F$. We have to demonstrate that $y \in T(E)$. Due to convergence, this sequence is a Cauchy sequence. By continuity of T^{-1} the sequence of inverse images $x_n := T^{-1}(y_n)$ is also Cauchy and, hence, convergent: $x_n \to x \in E$. By continuity of T we deduce $T(x) = T(\lim x_n) = \lim T(x_n) = \lim y_n = y$. This shows $y \in T(E)$. ∎

As a simple consequence of the Closed range theorem we obtain the following Corollary:

Corollary 9.3 *Let E, F be two Banach spaces and $T \in \mathcal{L}(E,F)$ injective. Furthermore we assume that $T^{-1} : T(E) \to E$ is continuous and T^* injective. Then T is an isomorphism from E to F.*

Proof. Due to the continuity of T and T^{-1} (on the image $T(E)$) we can use Lemma 9.2 which gives us that $T(E)$ is closed. Theorem 9.1 and the injectivity of T^* yields to

$$T(E) \;=\; N(T^*)^\perp \;=\; \{0\}^\perp \;=\; F.$$

Hence, T is bijective. In combination with the continuity properties of T and T^{-1} we deduce that T is an isomorphism. ∎

9.2 Inf-sup condition for Petrov-Galerkin methods

In this chapter we consider problems where test and ansatz function are from different spaces. Let U, V two Hilbert spaces and

$$A : U \times V \to \mathbb{R}$$

a bilinear form. We ask for solvaility and uniqueness of linear problems of the form

$$u \in U : \qquad A(u,v) \;=\; \langle f, v\rangle \qquad \forall v \in V\,. \tag{9.1}$$

In the case of finite-dimensional spaces U, V, such problems are called of *Petrov-Galerkin* type. In order to obtain existence and uniqueness, $dim\, U = dim\, V$ is a necessary condition. However, this is not sufficient, because the Theorem of Lax-Milgram (Satz 3.22) is not applicable if $U \neq V$. Hence, we need a generalization.

For U, V as introduced above we have the following definition.

Definition 9.4 *$A : U \times V \to \mathbb{R}$ is called* continuous*, if a constant $\alpha_1 > 0$ exists, s.t.*

$$|A(u,v)| \;\leq\; \alpha_1 \|u\|_U \|v\|_V \qquad \forall u \in U\ \forall v \in V.$$

A satisfies an inf-sup condition*, if $\gamma > 0$ exists, s.t.*

$$\inf_{u\in U\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{A(u,v)}{\|u\|_U \|v\|_V} \;\geq\; \gamma.$$

Lemma 9.5 *Let U be a Hilbert space and let $A : U \times U \to \mathbb{R}$ be coercive. Than A satisfies an inf-sup condition with $U = V$.*

Proof. Let $\alpha_2 > 0$ be the coercivity constant. Then we have

$$\inf_{u\in U\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{A(u,v)}{\|u\|_U \|v\|_V} \;\geq\; \inf_{u\in U\setminus\{0\}} \frac{A(u,u)}{\|u\|_V^2} \;\geq\; \inf_{u\in U\setminus\{0\}} \frac{\alpha_2 \|u\|_V^2}{\|u\|_V^2} \;=\; \alpha_2.$$

∎

Theorem 9.6 *Let U, V be two Hilbert spaces and let $A : U \times V \to \mathbb{R}$ be (bi-)linear and satisfies an inf-sup condition. Then for $f \in V'$, any solution of problem (9.1) is unique and satisfies the following stability property:*

$$\|u\|_U \quad \leq \quad \frac{1}{\gamma}\|f\|_{V'}\,.$$

Proof. (a) Let $u, \tilde{u} \in U$ be two solutions of (9.1). Then it holds for the difference $e = u - \tilde{u}$ by linearity of A:

$$A(e, v) \quad = \quad 0 \qquad \forall v \in V.$$

This implies

$$\sup_{v \in V \setminus \{0\}} \frac{A(e, v)}{\|v\|_V} \quad = \quad 0.$$

The inf-sup property of A yields to $e = 0$, i.e. $u = \tilde{u}$.
(b) The stability property also follows from the inf-sup property:

$$\gamma\|u\|_U \leq \sup_{v \in V \setminus \{0\}} \frac{A(u, v)}{\|v\|_V} \quad = \quad \sup_{v \in V \setminus \{0\}} \frac{\langle f, v\rangle}{\|v\|_V} \quad \leq \quad \sup_{v \in V \setminus \{0\}} \frac{\|f\|_{V'}\|v\|_V}{\|v\|_V} \quad = \quad \|f\|_{V'}.$$

□

This theorem does not yet include any information about the existence of solutions. We need the Theorem of the Closed Image given in the previous subsection.

Theorem 9.7 *We assume that the bilinear form $A : U \times V \to \mathbb{R}$ has the following properties:*

(a) A is continuous,

(b) A satisfies an inf-sup condition, and

(c) for each $v \in V \setminus \{0\}$ it should exist a $u \in U$ s.t. $A(u, v) \neq 0$.

Then (9.1) has for each $f \in V'$ a unique solution $u \in U$.

Proof. We introduce the linear operator $L : U \to V'$, which is for $u \in U$ and $v \in V$ defined by

$$\langle Lu, v\rangle \quad := \quad A(u, v)$$

and show that L is surjective. The continuity of A implies the continuity of L, hence $L \in \mathcal{L}(U, V')$. Theorem 9.6 ensures that L in injective and the stability $\gamma\|u\|_U \leq \|Lu\|_{V'}$. In other words:

$$\|L^{-1}f\|_U \quad \leq \quad \gamma^{-1}\|f\|_{V'} \qquad \forall f \in L(U),$$

which implies the continuity of L^{-1} on the image $L(U)$. Lemma 9.2 yields that $L(U)$ is closed in V'.
We now consider the adjoint operator $L^* : V \to U'$, $\langle L^* v, u\rangle = \langle Lu, v\rangle$. Here, we used $V'' = V$, because Hilbert spaces are reflexive. Due to $\langle L^* w, u\rangle = \langle Lu, w\rangle = A(u, w)$ we deduce

$$\begin{aligned} N(L^*) &= \{w \in V \ : \ \langle L^* w, u\rangle = 0 \quad \forall u \in U\} \\ &= \{w \in V \ : \ A(u, w) = 0 \quad \forall u \in U\} \\ &= \{0\}, \end{aligned}$$

where the last equality is a consequence of the assumption (c). The closed range theorem now yields

$$L(U) \ = \ N(L^*)^\perp \ = \ \{0\}^\perp \ = \ V'.$$

This is the surjectivity of L. The solution is obtained by $u := L^{-1} f$. ∎
This theorem is a generalization of the Theorem of Lax-Milgramm (Thm. 3.22) in the following context: (i) In the case $U \neq V$ and (ii) in the case $U = V$, but with non-elliptic bilinear form A which at least satisfies an inf-sup condition.

Theorem 9.8 *The conditions of Thm. 9.7 should be satisfied for the Hilbert spaces U, V and also for conforming finite-element spaces $U_h \subset U$, $V_h \subset V$ with constants $\alpha_1, \gamma > 0$. Then it holds for the unique solution $u_h \in U_h$:*

$$\|u - u_h\|_U \ \leq \ (1 + \alpha_1/\gamma) \inf_{w_h \in U_h} \|u - w_h\|_U.$$

Proof. We firstly use the triangle inequality:

$$\|u - u_h\|_U \ \leq \ \|u - w_h\|_U + \|u_h - w_h\|_U \qquad \forall w_h \in U_h.$$

We should obviously bound the last term. We define the following functional $l \in V'$ by

$$\langle l, v\rangle \ := \ A(u - w_h, v) \qquad \forall v \in V,$$

and the linear operator $L \in \mathcal{L}(U_h, V_h')$ by

$$\langle Lu_h, v_h\rangle \ := \ A(u_h, v_h).$$

Due to Galerkin orthogonality it holds for arbitrary $v_h \in V_h$

$$\langle l, v_h\rangle \ = \ A(u - w_h, v_h) \ = \ A(u_h - w_h, v_h) \ = \ \langle L(u_h - w_h), v_h\rangle.$$

In other words:

$$l|_{V_h} \ = \ L(u_h - w_h) \quad \text{or} \quad u_h - w_h \ = \ L^{-1} l.$$

Due to the boundedness of A we have $\|l\|_{V'} \leq \alpha_1 \|u - w_h\|_U$. As shown in the proof of Theorem 9.7, it holds $\|L^{-1}\|_{V_h'; U_h} \leq \gamma^{-1}$. Now follows

$$\|u_h - w_h\|_U \ = \ \|L^{-1} l\|_U \ \leq \ \gamma^{-1} \|l\|_{V_h'} \ \leq \ \gamma^{-1} \|l\|_{V'} \ \leq \ \gamma^{-1} \alpha_1 \|u - w_h\|_U.$$

This gives the assertion. ∎

9.3 Inf-sup condition for saddle point problems

Let V and Q be two Hilbert spaces and $X := V \times Q$ the product space. We consider a general saddle point problem of the following form:

$$\begin{aligned} a(u,\phi) - b(\phi,p) &= \langle f,\phi\rangle_V \qquad \forall \phi \in V, &(9.2)\\ b(u,\xi) &= \langle g,\xi\rangle_Q \qquad \forall \xi \in Q, &(9.3)\end{aligned}$$

with continuous bilinear forms

$$a : V \times V \to \mathbb{R} \quad \text{and} \quad b : V \times Q \to \mathbb{R}.$$

This system can be written in a more compressed for by using the bilinear form $A : X \times X \to \mathbb{R}$ with the product space $X := V \times Q$,

$$A(u,p;\phi,\xi) \;:=\; a(u,\phi) - b(\phi,p) + b(u,\xi),$$

and the right hand side by the functional

$$\langle l;(\phi,\xi)\rangle \;:=\; \langle f,\phi\rangle_V + \langle g,\xi\rangle_Q.$$

We seek a pair $(u,p) \in X$, s.t.

$$A(u,p;\phi,\xi) \;=\; \langle l;(\phi,\xi)\rangle \qquad \forall (\phi,\xi) \in X. \tag{9.4}$$

A natural norm on the product space X is simply

$$\|(u,p)\|_X \;:=\; \left(\|u\|_V^2 + \|p\|_Q^2\right)^{1/2}.$$

The bilinear form A is obviously not X-coercive, because

$$A(u,p;u,p) \;=\; a(u,u) - b(u,p) + b(u,p) \;=\; a(u,u).$$

That means, we loose all control over the variable p. Even if a is V-coercive, we only arrive at

$$A(u,p;u,p) \;\geq\; \alpha_2 \|u\|_V^2 \;\not\geq\; c\|(u,p)\|_X^2$$

for $p \in Q$. Therefore, the Lax-Milgram theorem is not applicable.

The following Theorem is called LBB-condition due to its inventors Ladyzhenskaja[1], Babuška[2] and Brezzi[3]. It is nothing else than an inf-sup condition for the non-coercive part $b(\cdot,\cdot)$ of the saddle point problem.

[1] Olga Alexandrovna Ladyzhenskaja, 1922-2004, russian mathematician and physicist.
[2] Ivo Babuška, born 1926 in Prag, czech mathematician.
[3] Franko Brezzi, born 1945 in Vimercate, italian mathematician.

Theorem 9.9 *For the continuous bilinear forms $a : V \times V \to \mathbb{R}$ and $b : V \times Q \to \mathbb{R}$ should hold:*
(a) a is symmetric and V_0-elliptic, where V_0 is the kernel of b:

$$V_0 \;=\; \{v \in V \;:\; b(v,\xi) = 0 \qquad \forall \xi \in Q\}.$$

(b) b satisfies an inf-sup condition, i.e. it exists a constant $\gamma > 0$, s.t.

$$\inf_{p\in Q\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{b(v,p)}{\|v\|_V \|p\|_Q} \;\geq\; \gamma. \tag{9.5}$$

Then the saddle point problem (9.2)-(9.3) has for arbitrary $f \in V'$ and $g \in Q'$ always a unique solution $(u,p) \in X$.

Proof. The proof consists of three steps:
Step 1: We consider the case $g = 0$. The V_0-ellipticity of the bilinear form a ensures the existence and uniqueness of a solution $u \in V_0$ of

$$a(u,\phi) \;=\; \langle f,\phi\rangle \qquad \forall \phi \in V_0. \tag{9.6}$$

It remains to find a (unique) $p \in Q$ with

$$b(\phi,p) \;=\; \langle l,\phi\rangle \;:=\; a(u,\phi) - \langle f,\phi\rangle \qquad \forall \phi \in V. \tag{9.7}$$

Here, the right hand side is a functional $l \in V'$, Due to (9.6) it holds for $\phi \in V_0$: $\langle l,\phi\rangle = 0$. Hence, $l \in V_0^{\perp}$. It exists a decomposition of the form $V = V_0 \oplus V_1$ with a subspace $V_1 \subseteq V$. Now, we consider the restricted bilinear form $\tilde{b} = b|_{V_1\times Q}$:

$$\tilde{b} : V_1 \times Q \to \mathbb{R}.$$

The inf-sup condition (9.5) directly implies the inf-sup condition of $\tilde{b}$:

$$\inf_{p\in Q\setminus\{0\}} \sup_{v\in V_1\setminus\{0\}} \frac{\tilde{b}(v,p)}{\|v\|_V \|p\|_Q} \;\geq\; \gamma. \tag{9.8}$$

For the application of Theorem 9.7 for $\tilde{b}$ we have to verify the condition (c): We should prove that for arbitrary $\phi \in V_1 \setminus \{0\}$ there exists $q \in Q$ s.t. $\tilde{b}(\phi,q) \neq 0$. This property holds by construction of V_0 and the fact that $V_0 \cap V_1 = \{0\}$. Theorem 9.7 now yields a unique solution $p \in Q$ of

$$\tilde{b}(\phi,p) \;=\; \langle l,\phi\rangle \qquad \forall \phi \in V_1.$$

This implies (9.7) by definition of V_0.
Step 2: For general g the proof is obtained by reformulation of the problem. We firstly ask for existence and uniqueness of $u_g \in V_1$, s.t.

$$b(u_g,\xi) \;=\; \langle g,\xi\rangle \qquad \forall \xi \in Q. \tag{9.9}$$

In this situation, Theorem 9.7 cannot be directly applied, because the roles of u_g and ξ are exchanged. However, in the proof of Theorem 9.7 the operator $B : Q \to V_1'$, defined by

$$\langle Bq, v\rangle \quad := \quad b(v,q),$$

is an isomorphism. Then, the adjoint operator $B^* : V_1 \to Q'$ is also an isomorphism. Due to $\langle B^*v, q\rangle = b(v,q)$, the existence of a unique solution $u_g \in V_1$ of (9.9) follows. Now, the saddle point problem is equivalent to seeking $w := u - u_g \in V$ and $p \in Q$, s.t.

$$\begin{aligned} a(w,\phi) - b(\phi,p) &= \langle f,\phi\rangle - a(u_g,\phi) \qquad \forall \phi \in V, \\ b(w,\xi) &= 0 \qquad\qquad\qquad\qquad\quad \forall \xi \in Q. \end{aligned}$$

This problem is of the same type as in Step 1 of this proof, and hence, it has a unique solution.

Step 3: For proving the uniqueness of the solution (u,p) we show its stability. Theorem 9.6 (formulated for B^*) yields the stability of u_g:

$$\|u_g\|_V \quad \leq \quad \gamma^{-1}\|g\|_{Q'}.$$

With the V_0-ellipticity of the bilinear form a it holds:

$$\begin{aligned} \|w\|_V \;=\; \|w\|_{V_0} \quad &\leq \quad \alpha_2^{-1}(\|f\|_{V_0'} + \|a(u_g,\cdot)\|_{V_0'}) \\ &\leq \quad \alpha_2^{-1}(\|f\|_{V'} + \alpha_1\|u_g\|_V) \\ &\leq \quad \alpha_2^{-1}\left(\|f\|_{V'} + \frac{\alpha_1}{\gamma}\|g\|_{Q'}\right). \end{aligned}$$

Here, $\alpha_1 > 0$ is the continuity constant, and $\alpha_2 > 0$ the ellipticity (coercivity) constant of a. Now, the uniqueness follows from the uniqueness of the solution of the homogeneous equation with vanishing right hand side. ∎

9.4 Mixed formulation of the Poisson problem

We want to write the Poisson problem as a system of 1. order differential equations and discretize this system afterwards by finite elements. We will see that this leads to a saddle point problem. We recapitulate that

$$\Delta p \quad = \quad \operatorname{div} \nabla p,$$

with the divergence operator $\operatorname{div} : H^1(\Omega)^d \to L^2(\Omega)$ definied by

$$\operatorname{div} u(x) \quad := \quad \sum_{i=1}^{d} \frac{\partial u_i(x)}{\partial x_i}.$$

This motivates us to introduce the new vector-valued variable $u := \nabla p$. We obtain the following system of 1. order:

$$\begin{aligned} u - \nabla p &= 0, \\ -\text{div } u &= f. \end{aligned}$$

9.4.1 Primal-mixed formulation of the Poisson problem

For a variational formulation of the system above, we integrate the second equation by parts. We obtain the so-called *primal-mixed formulation* of the Poisson problem. We seek $u \in L^2(\Omega)^d$ and $p \in H_0^1(\Omega)$, s.t.

$$(u, \phi)_{L^2(\Omega)^d} - (\nabla p, \phi)_{L^2(\Omega)^d} = 0 \qquad \forall \phi \in L^2(\Omega)^d, \tag{9.10}$$

$$(u, \nabla \xi)_{L^2(\Omega)^d} = \langle f, \xi \rangle \qquad \forall \xi \in H_0^1(\Omega). \tag{9.11}$$

We introduce the following notations for the Hilbert spaces

$$V := L^2(\Omega)^d \quad \text{and} \quad Q := H_0^1(\Omega),$$

and the bilinear forms $a : V \times V \to \mathbb{R}$, $b : V \times Q \to \mathbb{R}$ given by:

$$\begin{aligned} a(u, \phi) &:= (u, \phi)_V, \\ b(u, \xi) &:= (u, \nabla \xi)_V. \end{aligned}$$

The system (9.10)-(9.11) is then obviously of saddle point form (9.2)-(9.3) (the role of the right hand sides is changed).

Theorem 9.10 *For the primal-mixed formulation of the Poisson problem (9.10)-(9.11) exists for each $f \in Q'$ a unique solution $(u, p) \in V \times Q$.*

Proof. The bilinear form a is V-elliptic, because for $u \in V$ it holds $a(u, u) = \|u\|_V^2$. Hence, it is also elliptic on the kernel $V_0 \subseteq V$ of b. For varifying the inf-sup property of b we consider $p \in Q \setminus \{0\}$. Due to the Dirichlet conditions, p cannot be constant in Ω, i.e. $\nabla p \not\equiv 0$. We choose $v := \nabla p$:

$$\sup_{v \in V \setminus \{0\}} \frac{b(v, p)}{\|v\|_V \|p\|_Q} \geq \frac{b(\nabla p, p)}{\|\nabla p\|_V \|p\|_Q} = \frac{(\nabla p, \nabla p)_V}{\|\nabla p\|_V^2} = 1.$$

Therefore, the inf-sup property holds with constant $\gamma = 1$. Theorem 9.9 now ensures existence and uniqueness of the solution. ∎

9.4.2 Dual-mixed formulation of the Poisson problem

Another variational formulation can be obtained by performing integration by parts of the pressure gradient instead of the divergence. We use the Hilbert spaces:

$$\begin{aligned} V &:= H_{div}(\Omega) = \{v \in L^2(\Omega)^d \,:\, \text{div } v \in L^2(\Omega)\}, \\ Q &:= L^2(\Omega). \end{aligned}$$

The corresponding scalar product and norm of V are given by

$$\begin{aligned} (u,v)_V &:= (u,v)_{Q^d} + (\text{div } u, \text{div } v)_Q, \\ \|v\|_V &:= \left(\|v\|^2_{Q^d} + \|\text{div } v\|^2_Q\right)^{1/2}. \end{aligned}$$

The alternative mixed formulation is called *dual-mixed formulation* and reads: Seek $(u,p) \in V \times Q$, s.t.

$$(u,\phi)_{Q^d} + (p, \text{div } \phi)_Q = 0 \qquad \forall \phi \in V, \tag{9.12}$$

$$-(\text{div } u, \xi)_Q = \langle f, \xi \rangle \qquad \forall \xi \in Q. \tag{9.13}$$

At the first glance, the homogeneous Dirichlet conditions $p|_{\partial\Omega} = 0$ do not appear in this variational formulation. They are not implemented in the space Q, because L^2-functions do not necessarily have traces on the boundary. However, we will see later that the Dirichlet conditions are nevertheless included in this formulation.

Theorem 9.11 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. The dual-mixed formulation of the Poisson problem (9.12)-(9.13) has for each $f \in L^2(\Omega)$ a unique solution $(u,p) \in V \times Q$. For p holds the regularity $p \in H^1_0(\Omega)$.*

Proof. (a) The formulation (9.12)-(9.13) is also of saddle point form (9.2)-(9.3) with the bilinear forms

$$a(u,\phi) := (u,\phi)_{Q^d} \quad \text{and} \quad b(u,\xi) := -(\text{div } u, \xi)_Q.$$

a is V-continuous. The kernel of b is given by

$$V_0 := \{v \in V \,:\, (\text{div } v, \xi)_Q = 0 \ \forall \xi \in Q\}.$$

For $u \in V_0$ we have $\xi := \text{div } u \in Q$ and therefore $\|\text{div } u\|_Q = 0$. It follows

$$a(u,u) = \|u\|^2_{Q^d} = \|u\|^2_{Q^d} + \|\text{div } u\|^2_Q = \|u\|^2_V.$$

Hence, the bilinear form a is also V_0-elliptic.

Now we show the inf-sup property of b. Let $p \in Q \setminus \{0\}$ be arbitrary. We construct a suitable $v \in V \setminus \{0\}$. Since $C_0^\infty(\Omega)$ is dense in $Q = L^2(\Omega)$, it exists a $\tilde{p} \in C_0^\infty(\Omega)$ with

$$\|p - \tilde{p}\|^2_{L^2(\Omega)} \quad \leq \quad \frac{1}{2}\|p\|^2_Q.$$

We choose the first component of v as follows

$$v_1(x_1, \ldots, x_d) \quad := \quad -\int_{-\infty}^{x_1} \tilde{p}(t, x_2, \ldots, x_d)\, dt.$$

This integral is well defined, because Ω is assumed to be bounded. All other components of v are set to zero, $v_2 = \ldots = v_d := 0$. Due to this construction it holds point-wise

$$\text{div } v \quad = \quad \frac{\partial v_1}{\partial x_1} \quad = \quad -\tilde{p}.$$

With the Poincare inequality we have with a constant $c = c(\Omega)$ for the $L^2(\Omega)$-norm of v:

$$\|v\|_{Q^d} \quad \leq \quad c|v|_{H^1(\Omega)} \quad = \quad c\|\text{div } v\|,$$

and for the norm in V:

$$\|v\|^2_V \quad = \quad \|v\|^2_{Q^d} + \|\text{div } v\|^2 \quad \leq \quad (1 + c^2)\|\text{div } v\|^2 \quad = \quad (1 + c^2)\|\tilde{p}\|^2.$$

By use of the parallelogram equation and the inequality of Young to derive:

$$\begin{aligned} b(v, p) \quad &= \quad -(\text{div } v, p) = \ (\tilde{p}, p) \\ &= \quad \frac{1}{2}\left(\|\tilde{p}\|^2 + \|p\|^2 - \|\tilde{p} - p\|^2\right) \\ &\geq \quad \frac{1}{2}\left(\|\tilde{p}\|^2 + \frac{1}{2}\|p\|^2\right) \\ &\geq \quad \frac{1}{4}\left(\|\tilde{p}\|^2 + \|p\|^2\right) \\ &\geq \quad \frac{1}{2}\|\tilde{p}\|\|p\|. \end{aligned}$$

This yields with another constant $\gamma = \frac{1}{2}(1 + c^2)^{-1/2} > 0$:

$$\frac{b(v, p)}{\|v\|_V \|p\|_Q} \quad \geq \quad \frac{1}{2\sqrt{1 + c^2}} \frac{\|\tilde{p}\|\|p\|}{\|\tilde{p}\|\|p\|} \quad = \gamma.$$

This is the LBB-condition. Theorem 9.9 yields the unique solution $u \in V$ and $p \in Q$.
(b) Now we show that $p \in H^1_0(\Omega)$. Since every $C_0^\infty(\Omega)$-function is also in Q, it holds for the solution p in particular

$$(u, \phi) \quad = \quad -(p, \text{div } \phi) \qquad \forall \phi \in \mathcal{C}_0^\infty(\Omega).$$

This is just the criterion that p has the weak gradient, $\nabla p = u$. This shows that $p \in H^1(\Omega)$. The trace theorem now ensures that p has a L^2-trace on $\partial\Omega$. In order to see that this trace vanishes, let $\phi \in C^\infty(\Omega)$ be arbitrary. We take this function as test function in (9.12). Integration by parts leads us to

$$\begin{aligned} 0 &= (u, \phi) + (p, \operatorname{div} \phi) \\ &= (\nabla p, \phi) + (p, \operatorname{div} \phi) \\ &= \int_{\partial\Omega} p(\phi \cdot n)\, ds. \end{aligned}$$

Since ϕ was arbitrary, the trace of p must vanish on $\partial\Omega$, i.e. $p \in H_0^1(\Omega)$. ∎

9.5 Inf-sup condition for discrete saddle point problems

The discrete saddle point problem reads: Seek $u_h \in V_h$, $p_h \in Q_h$ s.t.

$$a(u_h, \phi) - b(\phi, p_h) = \langle f, \phi \rangle_V \qquad \forall \phi \in V_h, \tag{9.14}$$

$$b(u_h, \xi) = \langle g, \xi \rangle_Q \qquad \forall \xi \in Q_h. \tag{9.15}$$

Corollary 9.12 *We consider a saddle point problem (9.2)-(9.3) which satisfies the assumptions in Theorem 9.9. Let furthermore $\{\mathcal{T}_h\}$ be a family of quasi-uniform triangulations of Ω, $V_h \subset V$ and $Q_h \subset Q$ conforming finite element spaces with a discrete inf-sup condition*

$$\inf_{p_h \in Q_h \setminus \{0\}} \sup_{v_h \in V_h \setminus \{0\}} \frac{b(v_h, p_h)}{\|v_h\|_V \|p_h\|_Q} \geq \gamma_d > 0, \tag{9.16}$$

and the property that the bilinear form $a(\cdot, \cdot)$ is $V_{0,h}$-ellptic, where $V_{0,h} := \{v_h \in V_h : b(v_h, \xi) = 0 \ \forall \xi \in Q_h\}$. Then there exists a unique solution $(u_h, p_h) \in V_h \times Q_h$. Furthermore, if $\gamma_d > 0$ is independent of h, we have the a priori error estimate

$$\|u - u_h\|_V + \|p - p_h\|_Q \leq C \left(\inf_{w_h \in V_h} \|u - w_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right).$$

Proof. Is an immediate consequence of Theorem 9.9. ∎

The following Theorem gives a criterion for having a discrete inf-sup condition.

Theorem 9.13 (Fortin criterion) *Let $b : V \times Q \to \mathbb{R}$ be a bilinear form, which satisfies an inf-sup condition (9.5), $V_h \times Q_h \subset V \times Q$ a family of conforming finite element spaces such that a linear projection $\Pi_h : V \to V_h$ exists with the orthogonality property*

$$b(v - \Pi_h v, p_h) = 0 \qquad \forall p_h \in Q_h \ \forall v \in V,$$

and the stability property

$$\|\Pi_h\|_{V;V_h} \leq C$$

with a h independent constant C. Then b satisfies also the discrete inf-sup condition (9.16) with $\gamma_d > 0$ independent of h.

Proof. For $p_h \in Q_h \setminus \{0\}$ let $v \in V$ there is due to the (continuous) inf-sup condition a function $v \in V$ with

$$b(v, p_h) \quad \geq \quad \gamma \|v\|_V \|p_h\|_Q.$$

We choose $v_h := \Pi_h v$. It follows by the stability of Π_h:

$$\begin{aligned} b(v_h, p_h) &= b(v, p_h) + b(\Pi_h v - v, p_h) \\ &= b(v, p_h) \\ &\geq \gamma \|v\|_V \|p_h\|_Q \\ &\geq C^{-1} \gamma \|v_h\|_V \|p_h\|_Q. \end{aligned}$$

This is the discrete inf-sup condition with $\gamma_0 := C^{-1}\gamma$. ∎

9.6 Raviart-Thomas-Element

We now introduce the Raviart-Thomas element for the discretization of the Poisson problem in the dual-mixed formulation (9.12)-(9.13) in two spatial dimensions. As finite-dimensional subspaces of $Q = L^2(\Omega)$ we choose element-wise constant functions:

$$Q_h \quad = \quad \{p \in L^2(\Omega) \,:\, p|_T \in P_0 \quad \forall T \in \mathcal{T}_h\}.$$

The space V_h consists of element-wise linear polynomials of P_1^2 of the particular form:

$$v_h(x,y)|_T \quad = \quad \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T \begin{pmatrix} x \\ y \end{pmatrix} \tag{9.17}$$

with coefficients $a_T, b_T, c_T \in \mathbb{R}$. Instead of enforcing global continuity, only continuity in normal direction across cell edges are demanded, i.e.

$$\begin{aligned} V_h &= \mathcal{RT}_h \\ &:= \{v_h \in L^2(\Omega)^2 \,:\, v_h(x,y)|_T \text{ according to (9.17) with } a_T, b_T, c_T \in \mathbb{R} \ \forall T \in \mathcal{T}_h \\ &\quad \text{and } [v_h \cdot n_e] = 0 \text{ for all inner edges } e\}. \end{aligned}$$

This is the Raviart-Thomas element of lowest order. For Raviart-Thomas elements of order $k \in \mathbb{N}_0$, we choose element-wise coefficients of polynomial order k, i.e. $a_T, b_T, c_T \in P_k$. However, here we consider only the case $k = 0$.

On each element $T \in \mathcal{T}_h$ the finite element function $v_h \in \mathcal{RT}_h$ has three degrees of freedom. For a given edge $e \in \mathcal{E}_h$ all points $(x, y) \in e$ differ by a multiple of the tangential

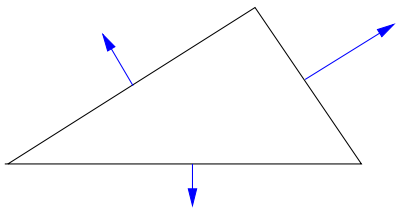


Figure 9.1: In the Raviart-Thomas element; the normal components $v_h^T n_e$ on the three edges are prescribed.

t_e of the edge. Therefore, $(x,y)^T n_e = c_e$ with a constant c_e. Hence, we obtain for the normal component of v_h:

$$\begin{aligned} v_h(x,y)^T n_e &= (a_T + c_T x, b_T + c_T y)^T n_e \\ &= (a_T, b_T)^T n_e + c_T (x,y)^T n_e \\ &= (a_T, b_T)^T n_e + c_T c_e. \end{aligned}$$

This means that the normal components are constant on each edge (independent of x and y) of the triangle. Therefore, the three degrees of freedom can be represented by the three normal components $v_h^T n_e$ for the three edges $e \in \mathcal{E}_T$.

Now we want to describe how to identify the polynomial ansatz of $v_h|_T$ by knowing only these three normal components: For each vertex N_i, $i = 1,2,3$ of the triangle the two components of $v_h(N_i)$ are determined by the two normal components of the adjacent edges. Knowing the six values $v_h(N_i)_1$, $v_h(N_i)_2$, $i = 1,2,3$, we can determine the unique linear polynomial $v_h|_T \in P_1^2$.

The Raviart-Thomas finite element method for the Poisson problem now seeks for $(u_h, p_h) \in \mathcal{RT}_h \times Q_h$, s.t.

$$(u_h, \phi)_{Q^d} + (p_h, \text{div } \phi)_Q = 0 \qquad \forall \phi \in \mathcal{RT}_h, \tag{9.18}$$

$$-(\text{div } u_h, \xi)_Q = \langle f, \xi \rangle \qquad \forall \xi \in Q_h. \tag{9.19}$$

For analyzing the solvability and uniqueness of discrete solutions we need the following preparatory Lemma:

Lemma 9.14 *The mapping div : $\mathcal{RT}_h \to Q_h$ is surjective. Furthermore, we have*

$$\|v_h\|_{H_{div}(\Omega)} \leq C \|div\, v_h\|_{L^2(\Omega)} \qquad \forall v_h \in \mathcal{RT}_h. \tag{9.20}$$

Proof. (a) Let $p_h \in Q_h$ be given. We choose a convex domain $\widetilde{\Omega}$ with $\Omega \subset \widetilde{\Omega}$ and extend p_h to $\widetilde{\Omega}$ trivially, i.e. $p_h|_{\tilde{\Omega}\setminus\Omega} \equiv 0$. Now we consider the Poisson problem

$$u \in H_0^1(\tilde{\Omega}): \quad \Delta u = p_h \quad \text{on } \tilde{\Omega}$$

Due to $p_h \in L^2(\tilde{\Omega})$ and the convexity of $\widetilde{\Omega}$ we have a unique solution with $u \in H^2(\widetilde{\Omega}) \cap H_0^1(\widetilde{\Omega})$. We define $v := \nabla u \in H^1(\Omega)$. This implies

$$\text{div } v = p_h \quad \text{a.e. in } \Omega.$$

Now we have to perform a suitable projection of v onto the finite element space $\mathcal{RT}_h$: Let $v_h \in \mathcal{RT}_h$ be the Raviart-Thomas function which has as normal component on each edge $e \in \mathcal{E}_h$ just the mean of the normal component of v:

$$v_h^T n_e \;=\; \frac{1}{|e|}\int_e v^T n_e dS.$$

This construction has the following property on each element $T \in \mathcal{T}_h$

$$\int_T p_h\,dx \;=\; \int_T \text{div } v\,dx \;=\; \int_{\partial T} v\cdot n\,ds \;=\; \int_{\partial T} v_h\cdot n\,ds \;=\; \int_T \text{div } v_h\,dx.$$

Taking into account that $p_h \in Q_h$ and div v_h are constant on each cell T, it follows that div $v_h(x) = p_h(x)$ for all inner points $x \in \bigcup_{T\in\mathcal{T}_h} T$.
(b) By transformation to the reference element one shows for each $T \in \mathcal{T}_h$:

$$\|v_h\|_{L^2(T)^d} \;\le\; |T|^{1/2}\|\hat{v}_h\|_{L^2(\hat{T})^d} \;\le\; c|T|^{1/2}\|\hat{v}\|_{H^1(\hat{T})^d} \;\le\; c\|v\|_{H^1(T)^d},$$

with a constant $c = c(\kappa)$. This implies

$$\begin{aligned}\|v_h\|^2_{H_{div}(\Omega)} &= \|v_h\|^2_{L^2(\Omega)^d} + \|\text{div } v_h\|^2_{L^2(\Omega)}\\ &= \sum_{T\in\mathcal{T}_h}\|v_h\|^2_{L^2(T)^d} + \|p_h\|^2_{L^2(\Omega)}\\ &\le c\|v\|^2_{H^1(\Omega)^d} + \|p_h\|^2_{L^2(\Omega)}.\end{aligned}$$

Due to the convexity of $\widetilde{\Omega}$ and Theorem 5.16 we deduce

$$\|v\|_{H^1(\Omega)^d} \;\le\; \|u\|_{H^2(\Omega)} \;\le\; \|u\|_{H^2(\widetilde{\Omega})} \;\le\; C\|p_h\|_{L^2(\widetilde{\Omega})} \;=\; C\|p_h\|_{L^2(\Omega)}.$$

This yields the desired bound. ∎
We already verified that the saddle point problem for the Poisson problem in the spaces $H_{div}(\Omega) \times L^2(\Omega)$ ensures unique solvability. The next Theorem addresses the discrete formulation in the spaces $\mathcal{RT}_h \times Q_h$.

Lemma 9.15 *There exists a linear projection $\Pi_h : H_{div}(\Omega) \to \mathcal{RT}_h$ with the following properties:*

1. *Orthogonality property:*

$$(div\,(v - \Pi_h v), q_h)_{L^2(\Omega)} \;=\; 0 \qquad \forall q_h \in Q_h\ \forall v \in H_{div}(\Omega). \tag{9.21}$$

2. Stability:

$$\|\Pi_h\|_{V;V} \quad \leq \quad C$$

3. Approximability:

$$\|u - \Pi_h u\|_{L^2(\Omega)^d} \quad \leq \quad C h_{max} |u|_{H^1(\Omega)^d} \qquad \forall u \in H^1(\Omega)^d.$$

Proof. 1. Since Q_h consists of element-wise constant functions, the orthogonality property can also be formulated locally:

$$\int_T \text{div } \Pi_h v \, dx \quad = \quad \int_T \text{div } v \, dx \qquad \forall v \in H_{div}(\Omega) \; \forall T \in \mathcal{T}_h.$$

This can be achieved by the previous Lemma as follows: For given $v \in H_{div}(\Omega)$ we define $p_h \in Q_h$ element-wise as

$$p_h|_T \quad := \quad \frac{1}{|T|} \int_T \text{div } v \, dx,$$

and $\Pi_h v$ as its inverse image of the mapping div :

$$\Pi_h v \quad := \quad \text{div}^{-1} p_h \; \in \; \mathcal{RT}_h.$$

This implies div $\Pi_h v = p_h$ and

$$\int_T \text{div } \Pi_h v \, dx \quad = \quad \int_T p_h \, dx \; = \; \int_T \text{div } v \, dx.$$

2. The mesh-size independency of $\|\Pi_h\|_{V;V}$ results from the estimate (9.20):

$$\|\Pi_h\|_{V;V} \quad = \quad \sup_{v \in V} \frac{\|\Pi_h v\|_{H_{div}(\Omega)}}{\|v\|_{H_{div}(\Omega)}} \; \leq \; C \sup_{v \in V} \frac{\|p_h\|_{L^2(\Omega)}}{\|v\|_{H_{div}(\Omega)}},$$

and

$$\begin{aligned} \|p_h\|^2_{L^2(\Omega)} \quad &= \quad \sum_{T \in \mathcal{T}_h} \|p_h\|^2_{L^2(T)} \; \leq \; \sum_{T \in \mathcal{T}_h} \frac{1}{|T|} \|\text{div } v\|^2_{L^1(T)} \\ &\leq \quad \sum_{T \in \mathcal{T}_h} \frac{\|1\|^2_{L^2(T)}}{|T|} \|\text{div } v\|^2_{L^2(T)} \; = \; \|\text{div } v\|^2_{L^2(\Omega)} \\ &\leq \quad \|v\|_{H_{div}(\Omega)}. \end{aligned}$$

This yields $\|\Pi_h\|_{V;V} \leq C$.

3. We show on the reference triangle and each normal $\hat{n}_j$

$$\|(\hat{u} - \hat{\Pi}\hat{u}) \cdot \hat{n}_j\|_{L^2(\hat{T})} \quad \leq \quad C |\hat{u} - \hat{\Pi}\hat{u}|_{H^1(\hat{T})^d} \qquad \forall \hat{u} \in H^1(\hat{T})^d. \tag{9.22}$$

By variable transformation this estimate yields the following approximation result on each triangle T:

$$\|u - \Pi_h u\|_{L^2(T)^d} \leq Ch_T |u - \Pi_h u|_{H^1(T)^d} \qquad \forall u \in H^1(T)^d.$$

In combination with (Exercise!)

$$|\Pi_h u|_{H^1(T)^d} \leq C|u|_{H^1(T)^d}$$

we obtain the desired estimate. In order to verify (9.22) let $w := \hat{u} \cdot n_j - \hat{\Pi}\hat{u} \cdot \hat{n}_j = (I - \hat{\Pi})(\hat{u} \cdot \hat{n}_j)$ and $w_0 := P_0 w$ its L^2-projection onto constants. We firstly use the triangle inequality

$$\|w\|_{L^2(\hat{T})} \leq \|w_0\|_{L^2(\hat{T})} + \|w - w_0\|_{L^2(\hat{T})}.$$

The first term on the right hand side is bounded by the equivalence of norms (note that $\|\cdot\|_{L^2(\hat{e})}$ is a norm on $\mathcal{P}_0$)

$$\|w_0\|_{L^2(\hat{T})} \leq C\|w_0\|_{L^2(\hat{e})}$$

and in combination with the trace theorem

$$\begin{aligned} \|w_0\|^2_{L^2(\hat{e})} &= (w_0, w_0)_{\hat{e}} \\ &= (w - w_0, w_0)_{\hat{e}} \\ &\leq \|w - w_0\|_{L^2(\hat{e})} \|w_0\|_{L^2(\hat{e})} \\ &\leq C\|w_0 - w\|_{H^1(\hat{T})} \|w_0\|_{L^2(\hat{e})} \end{aligned}$$

we arrive at

$$\|w\|_{L^2(\hat{T})} \leq C\|w_0 - w\|_{H^1(\hat{T})}.$$

Finally, we use the known result for the L^2-projection:

$$\|w_0 - w\|_{H^1(\hat{T})} \leq 2(\|w_0 - w\|_{L^2(\hat{T})} + |w|_{H^1(\hat{T})}) \leq C|w|_{H^1(\hat{T})}$$

□

Theorem 9.16 *The Raviart-Thomas finite element method yield for the Poisson problem (9.18)-(9.19) for each right hand side $f \in L^2(\Omega)$ a unique discrete solution $(u_h, p_h) \in \mathcal{RT}_h \times Q_h$.*

Proof. We are going to check the assumptions in Theorem 9.9:
(a) $\mathcal{RT}_h \times Q_h \subset H_{div}(\Omega) \times L^2(\Omega)$: The relation $Q_h \subset L^2(\Omega)$ is obvious. We have to check if for each $v_h \in \mathcal{RT}_h$ holds div $v_h \in L^2(\Omega)$. Since v_h is element-wise polynomial, it holds

div $v_h|_T \in L^2(T)$ for all $T \in \mathcal{T}_h$. This implies div $v_h \in L^2(\Omega)$.
(b) Ellipticity of the bilinear form $a(\cdot,\cdot)$: The bilinear form $a(v,\phi) = (v,\phi)_{L^2(\Omega)}$ is continuous on V and on V_0 elliptic. Therefore, it is also continuous on $\mathcal{RT}_h$ and elliptic on

$$V_{0,h} \quad := \quad \mathcal{RT}_h \cap V_0.$$

Furthermore, it is easy to check that

$$V_{0,h} \quad = \quad \{v_h \in \mathcal{RT}_h \,|\, (\text{div } v_h, q_h) = 0 \quad \forall q_h \in Q_h\}.$$

(c) The inf-sup condition for $b(v_h, p_h)$ is obtained by the criterion of Fortin in Theorem 9.13 and the previous Lemma which ensures the existence of $\Pi_h : H_{div}(\Omega) \to \mathcal{RT}_h$ with the desired properties. ∎

Theorem 9.17 *We assume that $(u,p) \in H^1(\Omega)^d \times H^1(\Omega)$ is the solution of (9.12)-(9.13) and that $\{\mathcal{T}_h\}$ is a family of shape-regular triangulations. Then it holds for the discretization error of the Raviart-Thomas elements of lowest order*

$$\begin{aligned}\|u-u_h\|_{H_{div}(\Omega)} + \|p-p_h\|_{L^2(\Omega)} \quad &\leq \quad ch_{max}\left(|u|_{H^1(\Omega)^d} + |p|_{H^1(\Omega)}\right) \\ &\quad + \inf_{f_h \in Q_h} \|f - f_h\|_{L^2(\Omega)},\end{aligned}$$

where $h_{max} := \max\{h_T : T \in \mathcal{T}_h\}$.

Proof. Due to Corollary 9.12 we know that

$$\|u-u_h\|_{H_{div}(\Omega)} + \|p-p_h\|_{L^2(\Omega)} \leq C\left(\inf_{w_h \in V_h} \|u-w_h\|_{H_{div}(\Omega)} + \inf_{q_h \in Q_h} \|p-q_h\|_{L^2(\Omega)}\right).$$

It is easy to verify that (by transformation onto the reference triangle)

$$\inf_{q_h \in \mathbb{P}_0} \|p-q_h\|_{L^2(T)} \quad \leq \quad ch_T |p|_{H^1(T)},$$

so that we obtain

$$\inf_{q_h \in Q_h} \|p-q_h\|_{L^2(\Omega)} \quad \leq \quad ch_{max}|p|_{H^1(\Omega)}.$$

Hence it remains to derive an appropriate bound for the interpolation of $u \in H^1(\Omega)$. As shown in Lemma 9.15 it exists a projection $\Pi_h u \in \mathcal{RT}_h$ with the property

$$\|u - \Pi_h u\|_{L^2(\Omega)} \quad \leq \quad Ch_{max}|u|_{H^1(\Omega)^d}.$$

In order to bound the divergence in the norm $\|\cdot\|_V$, we remind that the orthogonality property (9.21) is the characterization that div $\Pi_h u$ is the L^2-projection of div u. Hence, it features the minimizing property:

$$\|\text{div } (u - \Pi_h u)\|_{L^2(\Omega)} \quad = \quad \inf_{f_h \in Q_h} \|\text{div } u - f_h\|_{L^2(\Omega)}.$$

Taking into account that div $u = f$ gives the assertion. ∎

Chapter 10

Darcy equation

The Darcy equations[1] describes flow through porous media. An example is the behavior of ground water flow (interesting, e.g., for resolved pollutants). These equations are also extremely interesting for petrochemical engineering problems, because the subterranean flow of oil and the saturation of the ground are valuable parameters.

A central property of the Darcy equations is that the flow velocity v is proportional to the pressure gradient ∇p plus external forces. We consider the Darcy equation with the usual Dirichlet boundary conditions for the normal components of the velocity:

$$\begin{aligned} K^{-1}v + \nabla p &= \rho g \qquad \text{in } \Omega, \\ \operatorname{div} v &= l \qquad \text{in } \Omega, \\ v \cdot n &= v_0 \qquad \text{on } \partial\Omega. \end{aligned}$$

The matrix K contains the permeability values of the porous medium. K is assumed to be symmetric and positive definite. Hence we have the positive constants

$$0 \quad < \quad k_- := \inf_{x\in\Omega} \frac{\langle Kx, x\rangle}{\|x\|^2} \quad \leq \quad k_+ := \sup_{x\in\Omega} \frac{\langle Kx, x\rangle}{\|x\|^2} \quad < \quad \infty.$$

In heterogeneous media this matrix may not only vary in space, $K = K(x)$, but can be discontinuous, hence $K \in L^\infty(\Omega)^{d\times d}$. In some cases (e.g. in geophysics), this matrix is unknown and subject to its own investigation (inverse problems). Because the medium may accumulate and release the fluid, the velocity is not divergence free, but may contain sources and sinks. The density of the fluid will be denoted by ρ and is assumed to be constant. Moreover, the external fource g (gravitation) is assumed to be constant.

By use of the theorem of Gauß we immediately obtain the following compatibility condition:

$$\int_\Omega l\,dx \;=\; \int_\Omega \operatorname{div} v\,dx \;=\; \int_{\partial\Omega} v\cdot n\,ds \;=\; \int_{\partial\Omega} v_0\,ds.$$

[1] Henry Darcy, 10.06.1803–03.01.1858, french engineer, famous for his work on porous media flow.

Note that v is not necessarily a H^1-function, but v and div v must be in $L^2(\Omega)$. Th trace $tr\,(v \cdot n)$ vof $v \cdot n$ onn the boundary $\partial\Omega$ should also be L^2-integrable on the boundary. The natural Hilbert spaces for the fluid velocity and the pressure are given by

$$\begin{aligned} V &:= H_{div}(\Omega) := \{v \in L^2(\Omega)^d : \text{div } v \in L^2(\Omega)\}, \\ Q &:= L^2_0(\Omega). \end{aligned}$$

10.1 Primal formulation for the pressure

Applying (formally) the divergence operator onto the first equation, the velocity becomes eliminated:

$$\text{div } (K\nabla p) = \text{div } (\rho K g) - l \qquad \text{in } \Omega.$$

Hence, we obtain the Poisson problem with varying coefficients

$$\begin{aligned} -\text{div}\,(K\nabla p) &= f \qquad \text{in } \Omega\,, \\ \frac{\partial p}{\partial n} &= w_0 \qquad \text{on } \partial\Omega\,, \end{aligned}$$

with the right-hand side $f := K^{-1}l - \text{div } (\rho g)$, $w_0 := \rho g \cdot n - K^{-1}v_0$. Methods to solve this kind of problem have been discussed extensively in section **??**.

After determination of the pressure p (or its discrete representative p_h on a triangulation $\mathcal{T}_h$), the velocity field can be obtained afterwards by the equation $v = K(\rho g - \nabla p)$. In the discrete setting, an L^2-projection has to be applied:

$$(v_h, \phi) = (K(\rho g - \nabla p_h), \phi) \qquad \forall \phi \in V_h\,.$$

This can be done, e.g., on the same mesh. However, this step requires inversion of the mass matrix , which is in general not a hard because the condition number is moderate. One may choose e.g. $p_h \in P_1(\mathcal{T}_h)$ and $V_h = P_1(\mathcal{T}_h)^d$.

This approach is possible in certain occasions. However, in many cases, the most important quantity of interest is the flow field and not the pressure. An approximation of p_h by finite elements usually results in discontinuous gradients ∇p_h across cell edges. Discontinuous permeabilities K may introduce further discontinuities. In these cases, the L^2-projection can be used to generate continuous velocities. Another possibility is to consider the coupled system of p and v. The resulting saddle-point problem will be discussed in the next section.

10.2 Mixed variational formulation for Darcy

Considering the coupled system of Darcy leads to the variational formulation:

$$\begin{aligned}(K^{-1}v,\phi) - (p,\operatorname{div}\,\phi) + \int_{\partial\Omega} p\phi\cdot n\,ds &= (\rho g,\phi) \qquad \forall\phi\in V,\\ (\operatorname{div}\,v,\xi) - \int_{\partial\Omega} v\cdot n\xi\,ds &= (l,\xi) - \int_{\partial\Omega} v_0\xi\,ds \qquad \forall\xi\in Q.\end{aligned}$$

In the L^2-scalar product involving the pressure, the pressure gradient was integrated by parts, so that the derivative is shifted to the test function. The boundary condition for the velocities is incorporated in a weak sense by adding the corresponding boundary integral containing $v\cdot n$.

This is a saddle-point problem which requires an inf-sup condition for its well-posedness. Comparing this system with (9.12)-(9.13) we see that the only differences are the right-hand sides and the boundary integrals. Therefore, we can use the already derived inf-sup condition of the dual-mixed formulation of the Poisson problem. We obtain unique solvability.

We write the variables and test functions in the form of pairs, $u=\{v,p\}\in X:=V\times Q$, $\varphi=\{\phi,\xi\}\in X$. The corresponding bilinear form reads now

$$\begin{aligned}A(u,\varphi) &:= (K^{-1}v,\phi)_{L^2(\Omega)} - (p,\operatorname{div}\,\phi)_{L^2(\Omega)} + (\operatorname{div}\,v,\xi)_{L^2(\Omega)}\\ &\quad + \int_{\partial\Omega}(p\phi\cdot n - v\cdot n\xi)\,ds.\end{aligned} \tag{10.1}$$

The variational problem is now of the form

$$u\in X: \qquad A(u,\varphi) = F(\varphi) \qquad \forall\varphi\in X,$$

where

$$F(\varphi) := (\rho g,\phi) + (l,\xi) - \int_{\partial\Omega} v_0\xi\,ds.$$

Wir haben bereits in Kapitel 9.6 ein LBB-stabiles Element kennengelernt, nämlich dass Raviart-Thomas-Element.

10.3 A general a priori error estimate for stabilized finite elements

We have seen that the use of equal-order polynomials are not possible for the Darca-problem due to the absence of an appropriate discrete inf-sup condition. This is a pity, because equal-order finite elements have great advantages from the point of view of implementation. Moreover, efficient linear solvers are easier to design for equal-order elements,

because the coupling between the variables can then be treated by efficient block solvers (e.g. block Gauss-Seidel etc.).

Therefore, we will present here the possibility of stabilization techniques. The principal idea is to add stabilizing terms to the Galerkin formulation leading to a coercive discrete bilinear form. In this section we formulate such an approach in a very general manner. Afterwards we will use the general result for a discrete treatment of the Darcy equations.

Let us consider a linear variational problem given by

$$u \in X: \qquad A(u,\varphi) \;=\; F(\varphi) \qquad \forall \varphi \in X, \tag{10.2}$$

and a stabilized discrete version

$$u_h \in X_h: \qquad A_h(u_h,\varphi) \;=\; F_h(\varphi) \qquad \forall \varphi \in X_h,$$

with a stabilized bilinear form

$$A_h(u,\varphi) \;:=\; A(u,\varphi) + S_h(u,\varphi).$$

Furthermore, the right hand side may be different in order to obtain strong consistency (explained below). We use a suffix h for the modified right hand side. Strong consistency means that the exact solution also satisfies the discrete equation:

$$A_h(u,\varphi) \;=\; F_h(\varphi) \qquad \forall \varphi \in X_h. \tag{10.3}$$

This implies the Galerkin orthogonality for the discretization error $e := u - u_h$:

$$A_h(e,\varphi) \;=\; 0 \qquad \forall \varphi \in X_h\,. \tag{10.4}$$

We will measure the discretization error in the semi-norm $|\cdot| : X \to [0,\infty)$ and assume that A_h is coercive with respect to this semi-norm:

$$A_h(u,u) \;\geq\; \alpha_2 |u|^2 \qquad \forall u \in X_h, \tag{10.5}$$

with an h-independent constant $\alpha_2 > 0$.

Theorem 10.1 *Let X be a Hilbertr space, $X_h \subset X$ be a closed subspace, and $A, A_h : X \times X \to \mathbb{R}$ bilinear forms. On X we have a semi-norm $|\cdot| : X \to [0,\infty)$. Furthermore we assume that:*

1. *A_h is strongly consistent (10.3),*
2. *A_h ist $|\cdot|$-coercive, i.e (10.5) holds,*
3. *for the solution $u \in X$ of (10.2) it holds an interpolation property of the form*

$$|u - I_h u| \;\leq\; H_h(u), \tag{10.6}$$

with a (mesh dependent) functional $H_h : X \to \mathbb{R}$.

4. For a suitable $\alpha_1 \geq 0$ it holds:

$$A_h(w - I_h w, \varphi_h) \quad \leq \quad \alpha_2 H_h(w)|\varphi_h| \qquad \forall \varphi_h \in X_h,\, \forall w \in X.$$

Then we have the a priori estiamte

$$|u - u_h| \quad \leq \quad \left(1 + \frac{\alpha_2}{\alpha_1}\right) H_h(u).$$

Proof. We split the error $e = u - u_h$ in interpolation error $\eta := u - I_h u$ and projections error $\xi := I_h u - u_h \in X_h$. Obviously, it is sufficient to show the following upper bound for the projection error :

$$|\xi| \quad \leq \quad \frac{\alpha_2}{\alpha_1} H_h(u). \tag{10.7}$$

With the triangle inequality we then obtain the assertion. For showing (10.7) we use the coercivity (10.5), the linearity of A_h, and the Galerkin orthogonality (10.4):

$$\begin{aligned} \alpha_2|\xi|^2 &\leq A_h(I_h u - u_h, \xi) \\ &= A_h(u - u_h, \xi) + A_h(I_h u - u, \xi) \\ &= A_h(u - I_h u, -\xi) \\ &\leq \alpha_1 H_h(u)|\xi|. \end{aligned}$$

Dividing by $\alpha_2|\xi|$ yields the bound (10.7). ∎

Remark: In the case that the consistency (10.3) is only valid for sufficiently regular exact solutions, then the a priori estimate is still valid for such regular solutions u.

10.4 A special stabilized finite element scheme for Darcy.

In this section we use P_s (or Q_s)-elements for the pressure, and P_r (or Q_r)-elements for the velocities, $r, s \geq 1$. The stabilization term we are going to use is of the form (c.f. Masud and Hughes [12]):

$$\begin{aligned} S_h(u_h, \varphi) &= \frac{1}{2}(K^{-1}v + \nabla p, \phi + K\nabla\xi)_{L^2(\Omega)^d}, \\ F_h(\varphi) &= F(\varphi) + \frac{1}{2}(\rho g, \phi + K\nabla\xi)_{L^2(\Omega)^d}. \end{aligned} \tag{10.8}$$

We will use the norm

$$|\!|\!| u |\!|\!| \quad := \quad \left(\|K^{-1/2}v\|^2_{L^2(\Omega)} + \frac{1}{2}\|K^{1/2}\nabla p\|^2_{L^2(\Omega)}\right)^{1/2}.$$

We then immediately obtain the following stability result:

Lemma 10.2 *The bilinear form A_h given by the sum of (10.1) and (10.8) is $\|\cdot\|$-coercive and strongly consistent (10.3). Hence, the discrete solution $u_h \in X_h$ exists and is unique.*

Proof. (a) The Galerkin part is only partially coercive:

$$A(u,u) \;=\; (K^{-1}v,v) \;=\; \|K^{-1/2}v\|^2_{L^2(\Omega)^d}.$$

For the stabilization term we use Hölder inequality and Young's inequality:

$$\begin{aligned}
2S_h(u,u) &= (K^{-1}v+\nabla p, v+K\nabla p)\\
&= \|K^{-1/2}v\|^2 + 2(v,\nabla p) + \|K^{1/2}\nabla p\|^2\\
&\geq \|K^{-1/2}v\|^2 - 2\|K^{-1/2}v\|\|K^{1/2}\nabla p\| + \|K^{1/2}\nabla p\|^2\\
&\geq \|K^{-1/2}v\|^2 - 2\|K^{-1/2}v\|^2 - \frac{1}{2}\|K^{1/2}\nabla p\|^2 + \|K^{1/2}\nabla p\|^2\\
&= -\|K^{-1/2}v\|^2 + \frac{1}{2}\|K^{1/2}\nabla p\|^2 .
\end{aligned}$$

The sum of these two terms yield the coercivity

$$A_h(u,u) \;\geq\; \frac{1}{2}|\!|\!|u|\!|\!|^2.$$

(b) In the case of sufficient regularity, v and p are classical solutions, i.e. $K^{-1}v+\nabla p = \rho g$. This implies

$$S_h(u,\varphi) \;=\; \frac{1}{2}(\rho g, \phi + K\nabla\xi) \;=\; F_h(\varphi) - F(\varphi),$$

which yields the strong consistency (10.3). ∎

For the a priori analysis we firstly look onto the interpolation error.

Lemma 10.3 *For the interpolation error $u - I_h u$ with the nodal interpoland $I_h u$ on P_r (or Q_r) elements of v and P_s (or Q_s) elements of p it holds on quasi uniform meshes*

$$|\!|\!|u - I_h u|\!|\!| \;\leq\; c(k_-^{-1/2}h^{r+1}|v|_{H^{r+1}} + k_+^{1/2}h^s|p|_{H^{s+1}}),$$

with the notation $h = \max\{h_T : T \in \mathcal{T}_h\}$, and a constant $c = c(\Omega,\kappa)$ depending on Ω and the anisotropy κ of the mesh.

Proof. The estimate is a direct consequence of Thm. 5.12. ∎

Theorem 10.4 *Under the same assumptions of Lemma 10.3 it holds for the discretization error for the Darcy problem with the stabilization (10.8)*

$$|\!|\!|u - u_h|\!|\!| \;\leq\; c(k_-^{-1/2}h^{r+1}|v|_{H^{r+1}} + k_+^{1/2}h^s|p|_{H^{s+1}}).$$

Proof. Ww want to use Thm. 10.1. Due to Lemma 10.3 the property (10.6) is valid with

$$H_h(u) \;:=\; c(k_-^{-1/2}h^{r+1}|v|_{H^{r+1}} + k_+^{1/2}h^s|p|_{H^{s+1}}).$$

Hence, it is sufficient to validate the following bound for $u \in X$ and $\varphi \in X_h$:

$$A_h(u - I_h u, \varphi) \;\leq\; \alpha_1(k_-^{-1/2}h^{r+1}|v|_{H^{r+1}} + k_+^{1/2}h^s|p|_{H^{s+1}})|\!|\!|\varphi|\!|\!|.$$

We will bound each arising term separately. The scalar product and norms have to be understood in the $L^2(\Omega)$ sense. The Galerkin terms are integrated by parts:

$$\begin{aligned} A(\eta,\varphi) &= (K^{-1}\eta_v,\phi) + (\nabla\eta_p,\phi) - (\eta_v,\nabla\xi) \\ &\leq \|K^{-1/2}\eta_v\|\,\|K^{-1/2}\phi\| + \|K^{1/2}\nabla\eta_p\|\,\|K^{-1/2}\phi\| + \|K^{-1/2}\eta_v\|\,\|K^{1/2}\nabla\xi\| \\ &\leq 2\left(\|K^{-1/2}\eta_v\| + \|K^{1/2}\nabla\eta_p\|\right)|\!|\!|\varphi|\!|\!| \\ &\leq c\left(k_-^{-1/2}h^{r+1}|v|_{H^{r+1}(\Omega)} + k_+^{1/2}h^s|p|_{H^{s+1}(\Omega)}\right)|\!|\!|\varphi|\!|\!|\,. \end{aligned}$$

The stabilization terms can be bounded accordingly:

$$\begin{aligned} S_h(\eta,\varphi) &= \frac{1}{2}(K^{-1}\eta_v + \nabla\eta_p, \phi + K\nabla\xi) \\ &\leq \frac{1}{2}\Big(\|K^{-1/2}\eta_v\|\,\|K^{-1/2}\phi\| + \|K^{-1/2}\eta_v\|\,\|K^{1/2}\nabla\xi\| \\ &\quad + \|K^{1/2}\nabla\eta_p\|\,\|K^{-1/2}\phi\| + \|K^{1/2}\nabla\eta_p\|\,\|K^{1/2}\nabla\xi\|\Big) \\ &\leq \left(\|K^{-1/2}\eta_v\| + \|K^{1/2}\nabla\eta_p\|\right)|\!|\!|\varphi|\!|\!| \\ &\leq c\left(k_-^{-1/2}h^{r+1}|v|_{H^{r+1}(\Omega)} + k_+^{1/2}h^s|p|_{H^{s+1}(\Omega)}\right)|\!|\!|\varphi|\!|\!|\,. \end{aligned}$$

This gives the assertion. ∎

Note that the right hand side depends on the properties of the permeability matrix $K(x)$. However, this is not surprising because the norm $|\!|\!|\cdot|\!|\!|$ includes these parameters as well.

This estimate (and the discretization method itself) has still some drawbacks. Obviously, we do not have any control about any derivatives of the flow field. One may expect high oscillatory perturbations in the velocity. This can be improved by adding further control to the divergence:

$$\begin{aligned} S_h(u,\varphi) &:= \frac{1}{2}(K^{-1}v + \nabla p, \phi + K\nabla\xi)_{L^2(\Omega)^d} + \gamma(\text{div } v, \text{div } \xi), \\ F_h(\varphi) &:= F(\varphi) + \frac{1}{2}(\rho g, \phi + K\nabla\xi)_{L^2(\Omega)^d} + \gamma(l, \text{div } \xi). \end{aligned}$$

We obtain coercivity $A_h(u,u) \geq \frac{1}{2}|\!|\!|u|\!|\!|_*^2$ in the norm

$$|\!|\!|u|\!|\!|_* \;:=\; \left(\|K^{-1/2}v\|_{L^2(\Omega)}^2 + \frac{1}{2}\|K^{1/2}\nabla p\|_{L^2(\Omega)}^2 + \gamma\|\text{div } v\|^2\right)^{1/2}.$$

The corresponding interpolation estimate becomes

$$\|\!|u - I_h u|\!\|_* \quad \leq \quad c((1 + hk_-^{-1/2})h^r|v|_{H^{r+1}} + k_+^{1/2}h^s|p|_{H^{s+1}}).$$

Now, we see that for equal-order elements, $r = s$, and for h sufficiently small, we arrive at

$$\|\!|u - u_h|\!\|_* \quad \leq \quad ch^r|u|_{H^{r+1}},$$

which is optimal for the error in the divergence of v and the gradient of p. The error in the L^2-norm of v should be obtained by a duality argument.

Chapter 11

Stokes equation

Die mitunter einfachsten Modellgleichungen zur Beschreibung einer Strömung sind die sogenannten Stokes-Gleichungen. Durch diese Gleichungen wird ein extrem zähflüssiges Fluid, z.B. Honig, beschrieben. Im Gegensatz zu einen weniger viskosen Fluid treten hier keine kleinskaligen Wirbel auf. Das Gleichungssystem hierzu ist linear und daher relativ einfach zu analysieren. Dennoch zeigen sich an diesem System grundsätzliche numerische Schwierigkeiten.

We consider a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, a velocity field $v : \overline{\Omega} \to \mathbb{R}^d$ and a pressure $p : \overline{\Omega} \to \mathbb{R}$. The velocities are vector-valued $v = (v_1, \ldots, v_d)$. The external force is denoted by $f : \overline{\Omega} \to \mathbb{R}^d$. Now the Stokes system reads

$$\begin{aligned} \operatorname{div} v &= 0 \quad \text{in } \Omega, && (11.1)\\ -\Delta v + \nabla p &= f \quad \text{in } \Omega, && (11.2)\\ v &= v_0 \quad \text{on } \partial\Omega. && (11.3) \end{aligned}$$

For simplicity we used here Dirichlet conditions for v. In order to speak about classical solutions one requires $v \in C^2(\Omega)^d \cap C(\overline{\Omega})^d$ and $p \in C^1(\Omega)$.

The first equation of this system (i.e. (11.1)) ensures that the flow field is the divergence free. This is necessary to guarantee conservation of mass: Applying the Gauss Theorem for an arbitrary (Lipschitz) domain $\omega \subset \Omega$, we obtain

$$0 \;=\; \int_\omega \operatorname{div} v \, dx \;=\; \int_{\partial\omega} v \cdot n \, ds,$$

where n denotes the normal vector on $\partial\omega$. Splitting the boundary $\partial\Omega$ into an inflow part and a outflow part,

$$\begin{aligned} \Gamma_{in} &= \{x \in \partial\Omega \,:\, v \cdot n < 0\},\\ \Gamma_{out} &= \{x \in \partial\Omega \,:\, v \cdot n > 0\}, \end{aligned}$$

leads to

$$\int_{\Gamma_{in}} v \cdot n \, ds \;=\; -\int_{\Gamma_{out}} v \cdot n \, ds.$$

This means that the amount of inflow is identical to the amount of outflow. Hence, in Ω there are no sinks and no sources of material.

Note that obviously a compatibility condition is required:

$$\int_{\partial\omega} v_0 \cdot n \, ds \;=\; 0.$$

Moreover, the gradient is obviously only determined up to a constant. Therefore, we use a normalizing condition of the pressure of the form

$$\int_{\Omega} p \, dx \;=\; 0. \tag{11.4}$$

Now we will consider an appropriate variational formulation.

11.1 Variational formulation for Stokes

For the variational formulation the diffusive term and the pressure gradient will be integrated by parts:

$$\begin{aligned}(-\Delta v, \phi)_{L^2(\Omega)^d} &= (\nabla v, \nabla\phi)_{L^2(\Omega)^d} - (\partial_n v, \phi)_{L^2(\partial\Omega)^d}\,,\\ (\nabla p, \phi)_{L^2(\Omega)^d} &= -(p, \operatorname{div}\,\phi)_{L^2(\Omega)} + (pn, \phi)_{L^2(\partial\Omega)^d}\,.\end{aligned}$$

The arising boundary integrals vanish, if the test functions are zero on the boundary. We obtain the system

$$(\operatorname{div}\, v, \xi)_{L^2(\Omega)} \;=\; 0 \qquad \forall \xi \in Q, \tag{11.5}$$

$$(\nabla v, \nabla\phi)_{L^2(\Omega)^d} - (p, \operatorname{div}\,\phi)_{L^2(\Omega)} \;=\; \langle f, \phi\rangle \quad \forall \phi \in V. \tag{11.6}$$

The corresponding variational spaces are

$$\begin{aligned}\phi \in V &:= H_0^1(\Omega)^d,\\ \xi \in Q &:= L_0^2(\Omega) = \{q \in L^2(\Omega) : (q,1)_{L^2(\Omega)} = 0\}.\end{aligned}$$

The pressure normalization is build-in into the space Q. This system consists of $d+1$ partial differential equations. For $d = 2$, these can be written component-wise as:

$$\begin{aligned}\int_\Omega (\partial_x v^1 + \partial_y v^2)\xi \, dx &= 0 && \forall \xi \in Q\,,\\ \int_\Omega (\partial_x v^1 \partial_x \phi^1 + \partial_y v^1 \partial_y \phi^1) - \int_\Omega p\partial_x\phi^1 \, dx &= \int_\Omega f^1\phi^1 \, dx && \forall \phi^1 \in H_0^1(\Omega),\\ \int_\Omega (\partial_x v^2 \partial_x \phi^2 + \partial_y v^2 \partial_y \phi^2) - \int_\Omega p\partial_y\phi^2 \, dx &= \int_\Omega f^2\phi^2 \, dx && \forall \phi^2 \in H_0^1(\Omega).\end{aligned}$$

The components of each function are here denoted by an upper index, v^1, v^2, ϕ^1 und ϕ^2. The product space will be denoted by

$$X \;=\; V \times Q.$$

We denote the elements of X in the form $\{v, p\} \in X$. This type of parenthesis is used in order to avoid confusion with the brackets for the L^2-scalar product.

The following theorem makes an assessment of weak and classical solutions.

Theorem 11.1 *For right hand sides $f \in L^2(\Omega)^d$ weak solutions $\{v, p\} \in X$ of (11.5)-(11.6) with $v \in C^2(\Omega)^d \cap C(\overline{\Omega})$ and $p \in C^1(\Omega)$ are also classical solutions of the Stokes problem (11.1)-(11.3).*

Proof. We choose as test function $\xi := \operatorname{div} v \in L^2(\Omega)$. For a suitable constant c we have $\xi - c \in Q$. Due to (11.5) we derive

$$\|\operatorname{div} v\|^2_{L^2(\Omega)} \;=\; (\operatorname{div}\, v, \xi)_{L^2(\Omega)} \;=\; (\operatorname{div}\, v, c)_{L^2(\Omega)} \;=\; c \int_\Omega \operatorname{div}\, v \, dx.$$

Use of the Gauß Theorem and due to $v|_{\partial\Omega} = 0$ a.e. we obtain

$$\int_\Omega \operatorname{div}\, v \, dx \;=\; \int_{\partial\Omega} v \cdot n \, ds \;=\; 0.$$

This implies $\|\operatorname{div} v\|_{L^2(\Omega)} = 0$ and therefore div $v(x) = 0$ a.e. in Ω. Due to $v \in C^1(\Omega)$ the divergence vanishes in every point $x \in \Omega$.
The point-wise validity of equation (11.2) is obtained by tracking back to the elliptic case (Poisson problem): For the weak solution $v \in V$ it holds

$$(\nabla v, \nabla \phi)_{L^2(\Omega)^d} \;=\; \langle g, \phi \rangle \qquad \forall \phi \in V,$$

with $\langle g, \phi \rangle := \langle f, \phi \rangle + (p, \operatorname{div}\, \phi)_{L^2(\Omega)}$. Hence, v is classical solution of the Poisson problem

$$-\Delta v \;=\; g \;:=\; f - \nabla p \qquad \text{in } \Omega,$$

with associated homogeneous Dirichlet condition . But this is exactly equation (11.2). □

The solvability of the Stokes problem can not be shown by the Lax-Milgram Theorem, because the bilinear form $A : X \times X \to \mathbb{R}$, defined by

$$A(\{v, p\}, \{\phi, \xi\}) \;:=\; (\operatorname{div}\, v, \xi)_{L^2(\Omega)} + (\nabla v, \nabla \phi)_{L^2(\Omega)^d} - (p, \operatorname{div}\, \phi)_{L^2(\Omega)}$$

is not X-coercive:

$$A(\{v, p\}, \{v, p\}) \;=\; |v|_{H^1(\Omega)^d} \;\not\geq\; \alpha \|\{v, p\}\|_X \;=\; \alpha(|v|_{H^1(\Omega)^d} + \|p\|_{L^2(\Omega)}).$$

We are in a very similar situation as for the mixed formulations of the Poisson problem in Chapter 6: The Stokes problem (11.5)-(11.6) is of saddle-point type (9.2)-(9.3) with the bilinear forms $a : V \times V \to \mathbb{R}$ und $b : V \times Q \to \mathbb{R}$, given by

$$\begin{aligned} a(v,\phi) &= (\nabla v, \nabla \phi)_{L^2(\Omega)^d}, \\ b(v,\xi) &= (\operatorname{div} v, \xi)_{L^2(\Omega)}. \end{aligned}$$

We have to verify the corresponding inf-sup property in order to obtain existence and uniqueness of solutions.

11.2 The gradient operator for L^2-functions

We know that the gradient operator maps H^1-functions onto L^2-functions. In this section we investigate the gradient acting on L^2-functions, because this is important to show this inf-sup property.

$$\nabla \;:\; L^2(\Omega) \to H^{-1}(\Omega)^d$$

with the definition for $q \in L^2(\Omega)$:

$$\langle \nabla q, v \rangle \;:=\; -(q, \operatorname{div} v) \qquad \forall v \in H_0^1(\Omega)^d.$$

Lemma 11.2 *[Lemma of J.L. Lions] Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then it holds*

$$L^2(\Omega) \;=\; \left\{ q \in H^{-1}(\Omega) \,|\, \nabla q \in H^{-1}(\Omega)^d \right\}.$$

Proof. The direction $\subseteq$ is easy to verify. The other direction, i.e. that $q, \partial_i q \in H^{-1}(\Omega)$ for $i \in \{1, \ldots, d\}$ implies that $q \in L^2(\Omega)$ is very difficult. We refer to [9]. ∎

Lemma 11.3 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then it holds*

$$\|p\|_{L^2(\Omega)} \;\cong\; \|p\|_{H^{-1}(\Omega)} + \|\nabla p\|_{H^{-1}(\Omega)^d}.$$

Proof. The bound

$$\|p\|_{H^{-1}(\Omega)} \;\leq\; c\|p\|_{L^2(\Omega)} \tag{11.7}$$

is obvious due to the continuous embedding $id : L^2(\Omega) \to H^{-1}(\Omega)$. The boundedness of the gradient is obtained as follows:

$$\begin{aligned} \|\nabla p\|_{H^{-1}(\Omega)^d} &= \sup_{v \in H_0^1(\Omega)^d} \frac{(p, \operatorname{div} v)_{L^2(\Omega)}}{\|\nabla v\|_{L^2(\Omega)^{d\times d}}} \\ &\leq \sup_{v \in H_0^1(\Omega)^d} \frac{\|p\|_{L^2(\Omega)} \|\operatorname{div} v\|_{L^2(\Omega)}}{\|\nabla v\|_{L^2(\Omega)^{d\times d}}} \;\leq\; c\|p\|_{L^2(\Omega)}\,. \end{aligned} \tag{11.8}$$

Hence, the gradient operator is continuous and $\nabla \in \mathcal{L}(L^2(\Omega); H^{-1}(\Omega))$. The bound into the other direction can be obtained by Lemma 11.2. The equality of the spaces imply that the identity

$$id : L^2(\Omega) \to \mathcal{H} := \{q \in H^{-1}(\Omega) \,|\, \nabla q \in H^{-1}(\Omega)^d\}$$

is bijective. $\mathcal{H}$ is in combination with a suitable norm (exactly that one defined by the right hand side of (11.9) a Banach space. The bijection *id* acts as a map between Banach spaces. The estimates (11.7)-(11.8) ensures the continuity of *id*. The Theorem of the open image now gives us the continuity of the inverse map. This implies (11.9). □

The estimate

$$\|p\|_{L^2(\Omega)} \quad \leq \quad c\left(\|p\|_{H^{-1}(\Omega)} + \|\nabla p\|_{H^{-1}(\Omega)^d}\right). \tag{11.9}$$

is also called *Inequality of Nečas*, see [13].

In the next lemma we consider the quotient space $\dot{E} := E/Ker(A)$ for E, F Banach spaces, and $A \in \mathcal{L}(E, F)$. The quotient space becomes with the norm

$$\|\dot{u}\|_{\dot{E}} \quad := \quad \inf_{u \in \dot{u}} \|u\|_E \qquad \forall \dot{u} \in \dot{E}$$

also a Banach space.

Lemma 11.4 *Let E, F Banach spaces, $A \in \mathcal{L}(E, F)$, and $\dot{E} := E/Ker(A)$ be the quotient space. Then*

$$\begin{aligned} \dot{A} \quad : \quad & \dot{E} \to R(A) \subseteq F, \\ & \dot{u} \mapsto \dot{A}\dot{u} = Au, \end{aligned}$$

is linear, continuous and bijective, $\dot{A} \in \mathcal{L}(\dot{E}, R(A))$.

Proof. We leave the proof as an exercise. □

Theorem 11.5 *Let E, F, G three Banach spaces, $A \in \mathcal{L}(E, F)$ and $B \in \mathcal{L}(E, G)$. Furthermore, let B a compact operator and it should hold the following equivalence of norms:*

$$\|u\|_E \quad \cong \quad \|Au\|_F + \|Bu\|_G.$$

Then, the kernel $Ker(A)$ is finite-dimensional, and $R(A)$ is closed.

Proof. (a) We know that a normed space is finite-dimensional, iff its unit ball is compact. Therefore, we will show that $S := \{u \in Ker(A) \,|\, \|u\|_E \leq 1\}$ is compact. Let $(u_n)_{n\in\mathbb{N}}$ an arbitrary sequence in S. For the elements of S the norms $\|u\|_E$ and $\|Bu\|_G$ are equivalent. Hence $(u_n)_{n\in\mathbb{N}}$ is also bounded with respect to the norm $\|Bu_n\|_G$. Since B was assumed to be compact, the sequence $(Bu)_{n\in\mathbb{N}}$ is relatively compact. Hence it exists

a subsequence which is convergent in G, also denoted by $(Bu_n)_{n\in\mathbb{N}}$. This sequence is then also a Cauchy sequence in E. Due to the completeness of E there is a $u \in E$ with

$$S \ni u_n \to u \in E.$$

By continuity of A we have $u \in Ker(A)$ and $u \in S$. This implies the compactness of S.
(b) According to the previous lemma, we know that $\dot{A}$ is linear, continuous and bijective. Now we show that it is an isomorphism. Hence, we have to show that its inverse is continuous, i.e.

$$\|\dot{u}\|_{\dot{E}} \;\leq\; c\|\dot{A}\dot{u}\|_F \qquad \forall \dot{u} \in \dot{E}.$$

For given $\dot{u} \in \dot{E}$ let $u \in E$ be an arbitrary representer. It holds

$$\|\dot{u}\|_{\dot{E}} \;:=\; \inf_{e\in Ker(A)} \|u+e\|_E.$$

According to part (a) we know that the kernel is finite dimensional. Therefore, the infimum is attained, i.e. there exists a representer $\tilde{u} \in E$ with $\|\dot{u}\|_{\dot{E}} = \|\tilde{u}\|_E$. We deduce that it is sufficient to demonstrate

$$\|\tilde{u}\|_E \;\leq\; c\|A\tilde{u}\|_F \qquad \forall \dot{u} \in \dot{E}. \tag{11.10}$$

We show this by contradiction. Assuming that this bound does not hold. Then it exists a sequence $(\tilde{u}_n)_{n\in\mathbb{N}}$ with

$$\|\tilde{u}_n\|_E \;=\; 1 \quad \text{and} \quad \lim_{n\to\infty} \|A\tilde{u}_n\|_F \;=\; 0.$$

This sequence is bounded in E beschränkt. By compactness of B, their images $(B\tilde{u}_n)_{n\in\mathbb{N}}$ include a convergent subsequence, also denoted by $(B\tilde{u}_n)_{n\in\mathbb{N}}$. Hence, both sequences $(A\tilde{u}_n)_{n\in\mathbb{N}}$ and $(B\tilde{u}_n)_{n\in\mathbb{N}}$ are convergent. By the equivalence of the norms we deduce that $(\tilde{u}_n)_{n\in\mathbb{N}}$ forms a Cauchy sequence in E and hence $\tilde{u}_n \to \tilde{u}$ in E. Now we have on the one hand

$$\|\tilde{u}\|_E \;=\; \lim_{n\to\infty} \|\tilde{u}_n\|_E \;=\; 1,$$

and on the other hand

$$\|A\tilde{u}\|_F \;=\; \lim_{n\to\infty} \|A\tilde{u}_n\|_F \;=\; 0.$$

This implies that $\tilde{u} \in Ker(A)$. Because $\tilde{u}$ is a representer of $\dot{u}$, we deduce $\dot{u} = 0$ and $\|\dot{u}\|_{\dot{E}} = 0$. This yields the contradiction $\|\tilde{u}\|_E = 0$. ∎

Corollary 11.6 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then the image $R(\nabla)$ of the mapping*

$$\nabla : L^2(\Omega) \to H^{-1}(\Omega)^d$$

is a Banach space.

Proof. We apply the previous Theorem with the setting $E := L^2(\Omega)$, $F := H^{-1}(\Omega)$, $G := H^{-1}(\Omega)^d$, $A := id$ and $B := \nabla$. Lemma 11.3 ensures

$$\|u\|_E \quad \cong \quad \|u\|_F + \|\nabla u\|_G.$$

Furthermore, we know that $\nabla : L^2(\Omega) \to H^{-1}(\Omega)^d$ is compact, so that we may apply the previous Theorem. This shows that $R(\nabla)$ is complete. □

Corollary 11.7 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then the mapping*

$$\nabla : L^2(\Omega) \to R(\nabla) \subseteq H^{-1}(\Omega)^d$$

is open.

Proof. $\nabla : Q \to R(\nabla)$ is surjective by construction. $R(\nabla)$ ist a normed space and due to Corollary 11.6 complete. It means that ∇ is a linear, continuous and surjective mapping between Banach spaces. The Open Mapping Theorem implies that this mapping is open. □

11.3 The gradient operator for L^2_0-functions

Now we restrict the gradient operator to functions in $Q = L^2_0(\Omega)$:

$$\nabla : L^2_0(\Omega) \to R(\nabla) \subseteq H^{-1}(\Omega)^d$$

Corollary 11.8 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then it holds*

$$\|p\|_{L^2(\Omega)} \quad \leq \quad c_\Omega \|\nabla p\|_{H^{-1}(\Omega)^d} \quad \forall p \in Q\,. \tag{11.11}$$

Proof. (a) We show that $\nabla : Q \to R(\nabla)$ is injective. We consider firstly the gradient operator for smooth functions

$$\nabla : C^\infty(\Omega) \cap Q \to C^\infty(\Omega).$$

The kernel of this ∇ is (due to the connectivity of Ω) trivial $\{0\} = Ker(\nabla) \subset C^\infty(\Omega)$ and therefore a closed set in $C^\infty(\Omega) \cap Q$. Due to the continuity of the extended operator $\nabla : Q \to R(\nabla)$ we know that the kernel $A := Ker(\nabla) \subset Q$ is also closed. Now we know that $C^\infty(\Omega) \cap Q$ is dense in Q. We conclude that the two kernels are identical: $A = \{0\}$ (Exercise).

The mapping

$$\begin{array}{rcl} \dot\nabla : L^2(\Omega)/Ker(\nabla) & \to & H^{-1}(\Omega)^d \\ \dot p & \mapsto & \nabla p \end{array}$$

is injective by definition. Now we demonstrate that

$$L^2(\Omega)/Ker(\nabla) \quad \cong \quad L^2(\Omega)/\mathbb{R} \;\cong\; Q = L^2_0(\Omega).$$

The corresponding isomorphism $\Phi : L^2(\Omega)/Ker(\nabla) \to Q$ is given by $\Phi(\dot p) := p - \bar p$ with $\bar p := |\Omega|^{-1} \int_\Omega p\, dx$. We have to show:

- Φ is well defined, i.e., $\Phi(\dot{p})$ is independent of the choice of $p \in \dot{p}$. To this end, we have to prove that $q \in Ker(\nabla)$ implies $q = const.$
- Φ is injective: $\Phi(\dot{p}) = 0 \Rightarrow p = \bar{p} \Rightarrow p \equiv const. \Rightarrow \nabla p = 0$,
- Φ is surjective:
- $\|\dot{p}\|_{L^2(\Omega)/Ker(\nabla)} = \|\Phi(p)\|$.

Let $p \in Q$ with $\|\nabla p\|_{H^{-1}(\Omega)^d} = 0$. In order to show that $\|p\|_{L^2(\Omega)} = 0$ it is sufficient to show $\|p\|_{H^{-1}(\Omega)} = 0$, because of the Lemma of Lions. Hence, we have to show

$$(p, w) = 0 \qquad \forall w \in H_0^1(\Omega).$$

STILL TO SHOW

(b) By the same arguments as in Corollary 11.7, we know that $R(\nabla) = \nabla(Q)$ is complete and, hence, the gradient operator as mapping of functions in Q is an open mapping. Moreover, $\nabla : Q \to R(\nabla)$ is surjective by construction and therefore bijective. This implies that Q and $R(\nabla) = \nabla(Q)$ are isomorph, and hence (11.11) holds. ∎

Lemma 11.9 *Es gilt $Q = Q'$.*

Proof. Since $Q \subseteq L^2(\Omega)$ we know that $Q \subseteq L^2(\Omega) = L^2(\Omega)' \subseteq Q'$. Hence, it remains to show that $Q' \subseteq Q$. Let $q \in Q'$ be given. We will firstly show that $q \in L^2(\Omega)$: For $p \in L^2(\Omega)$ we can apply q onto $p - \bar{p} \in Q$. We define $\langle q, p\rangle := \langle q, p - \bar{p}\rangle$. This q is a linear functional on $L^2(\Omega)$ and is continuous because

$$\begin{aligned}\|q\|_{L^2(\Omega)'} &= \sup_{p\in L^2(\Omega)} \frac{\langle q,p\rangle}{\|p\|} = \sup_{p\in L^2(\Omega)} \frac{\langle q,p-\bar{p}\rangle}{\|p\|} \le \|q\|_{Q'} \sup_{p\in L^2(\Omega)} \frac{\|p-\bar{p}\|}{\|p\|} \\ &\le \|q\|_{Q'} \sup_{p\in L^2(\Omega)} \frac{\|p\|}{\|p\|} = \|q\|_{Q'}.\end{aligned}$$

This yields that $q \in L^2(\Omega)' = L^2(\Omega)$. It remains to show $q \in Q$. Let $c \in \mathbb{R}$ an arbitrary constant. We have for arbitrary $p \in Q$:

$$\langle q + c, p\rangle = \langle q, p\rangle + \langle c, p\rangle = \langle q, p\rangle + c\int_\Omega p\,dx = \langle q, p\rangle$$

This implies that q and $q + c$ are identical functionals on Q. Therefore, we can identify q with $q - \bar{q} \in Q$. ∎

11.4 Inf-sup property of the Stokes system

Theorem 11.10 (inf-sup for Stokes) *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then there exists $\gamma > 0$ s.t.*

$$\inf_{p\in Q\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{(div\ v, p)}{\|\nabla v\|\|p\|} \ge \gamma.$$

Proof. [of Theorem 11.10] Let $p \in Q \setminus \{0\}$ be given. The Lemma of Lions implies $\nabla p \in H^{-1}(\Omega)$. Its norm is defined as

$$\|\nabla p\|_{H^{-1}(\Omega)^d} \quad = \quad \sup_{v \in V \setminus \{0\}} \frac{\langle \nabla p, v \rangle}{\|\nabla v\|}$$

The action of ∇p as a functional on V is defined as

$$\langle \nabla p, v \rangle \quad = \quad -(\text{div } v, p) \quad \forall v \in V.$$

This implies

$$\sup_{v \in V \setminus \{0\}} \frac{(\text{div } v, p)}{\|\nabla v\| \|\nabla p\|_{H^{-1}(\Omega)^d}} \quad = \quad 1.$$

The inf-sup property now follows due to

$$\|p\| \quad \leq \quad \frac{1}{\gamma} \|\nabla p\|_{H^{-1}(\Omega)^d} \quad \forall p \in Q,$$

with a positive constant γ. The proof of this upper bound will be demonstrated in the following. ∎

Theorem 11.11 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain, and $f \in H^{-1}(\Omega)^d$. Then the Stokes problem (11.5)-(11.6) has a unique solution $\{v, p\} \in H_0^1(\Omega)^d \times L_0^2(\Omega)$.*

Proof. Follows directly from the inf-sup property of $b(p, \phi) = (p, \text{div}\, \phi)$ (Thm. 11.10) and the coercivity of $a(v, \phi) = (\nabla v, \nabla \phi)$. ∎
We are going to use for the kernel of b the notation

$$V_0 \quad := \quad \{v \in H_0^1(\Omega) \, : \, (\text{div } v, \xi)_{L^2(\Omega)} = 0 \quad \forall \xi \in Q\}$$

and for its polare

$$V_0^0 \quad := \quad \{\phi \in V' \, : \, \langle \phi, v \rangle = 0 \quad \forall v \in V_0\} \; \subseteq \; H^{-1}(\Omega)^d.$$

Lemma 11.12 *Let $V_0^\perp$ be the orthogonal complement to V_0 in V, i.e. $V = V_0 \oplus V_0^\perp$. Then it holds $V_0^0 \cong (V_0^\perp)'$.*

Proof. Each functional $g \in (V_0^\perp)'$ can be extended to a functional $\tilde{g} \in V'$ with $\tilde{g}|_{V_0} \equiv 0$ by defining $\langle \tilde{g}, v \rangle := \langle g, v^\perp \rangle$ where $v^\perp$ is the projection of v onto $V_0^\perp$. ∎

Theorem 11.13 *[Theorem of de Rham] Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then it holds:*

(a) For every $\Phi \in V_0^0$ it exists a $p \in L^2(\Omega)$ with $-\nabla p = \Phi$ in the $H^{-1}(\Omega)$ sense. In other word:

$$\nabla : L^2(\Omega) \to V_0^0 \subseteq H^{-1}(\Omega)^d$$

is surjective.

(b) The p in (a) is uniquely defined up to a constant.

Proof. (a): The assertion makes a statement about the gradient operator

$$\nabla : L^2(\Omega) \to R(\nabla) \subseteq H^{-1}(\Omega)^d.$$

In particular, we have to show that $V_0^0 \subseteq R(\nabla)$. Due to Theorem 11.5 we know that $R(\nabla)$ is a closed subspace of $H^{-1}(\Omega)^d$. Hence, we can apply the Closed Range Theorem which ensures that

$$R(\nabla) = Ker(\nabla^*)^0.$$

Here, ∇^* is the adjoint operator of the gradient, namely the divergence:

$$\nabla^* = \text{div} : V \to L^2(\Omega).$$

Note that $V'' = V$ and $L^2(\Omega)' = L^2(\Omega)$. By definition we have $Ker(\nabla^*) = V_0$ and therefore $R(\nabla) = V_0^0$.
(b): Let $p_1, p_2 \in L^2(\Omega)$ have the property that $-\nabla p_1 = -\nabla p_2 = \Phi$. Then $q := p_1 - p_2 \in L^2(\Omega)$ has the property $\nabla q = 0$ (as functional in $H^{-1}(\Omega)^d$). Let $\bar{q} = |\Omega|^{-1} \int_\Omega q \, dx$. Then $q - \bar{q} \in Q$ and $\nabla(q - \bar{q}) = 0$. As shown in the proof of Corollary 11.8, we have $q - \bar{q} = 0$. This means that $q = \bar{q}$ is constant. ∎

Corollary 11.14 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then there exists a constant $c_\Omega > 0$, s.t. for every pressure $p \in Q$ it exists a velocity $v \in V$ with div $v = p$ and $|v|_{H^1(\Omega)^d} \leq c_\Omega \|p\|_{L^2(\Omega)}$.*

Proof. We have previously seen that $\nabla : Q \to V_0^0$ is an isomorphism. This is equivalent to the fact that the adjoint operator

$$\nabla^* : (V_0^0)' \to Q', \quad \phi \mapsto \text{div } \phi$$

is an isomorphism. Moreover, we have seen in Lemma 11.12 that $V_0^0 \cong (V_0^\perp)'$. This implies $(V_0^0)' \cong (V_0^\perp)'' \cong V_0^\perp$. Since $Q' = Q$ according to Lemma 11.9 and $\nabla^* = \text{div}$ we obtain the isomorphism

$$\text{div} : V_0^\perp \to Q.$$

This gives the assertion. ∎

11.5 The discrete LBB-condition for Stokes

The inf-sup property in Theorem 11.10 for the space X usually do not transfer to the discrete counterpart X_h. Such a discrete inf-sup condition reads:

Definition 11.15 (LBB condition) *Discrete subspaces $V_h \subset V$ and $Q_h \subset Q$ fulfill the discrete inf-sup condition (LBB condition) for the Stokes system, iff there exists $\gamma > 0$ s.t.*

$$\inf_{p_h \in Q_h \setminus \{0\}} \sup_{v_h \in V_h \setminus \{0\}} \frac{(p_h, div\ v_h)}{|v_h|_{H^1(\Omega)^d} \|p_h\|} \quad \geq \quad \gamma. \tag{11.12}$$

This condition is called LBB condition in honor of the mathematicians Ladyshenskaya, Babuska and Brezzi.

The LBB condition is in particular not fulfilled, if there exists a discrete pressure $p_h \in Q_h$ which lies in the kernel of the divergence operator, i.e.

$$(p_h, \operatorname{div} v_h) \quad = \quad 0 \qquad \forall v_h \in V_h.$$

11.6 Examples of unstable Stokes elements

Unstable elements are for instance all equal order elements as P_r/P_r-elements or Q_r/Q_r-elements for v_h and p_h.

Another unstable combination is the combination of Q_1-elements for the velocities and P_0-elements (i.e. discontinuous, but element-wise constant) for the pressure:

$$\begin{aligned} V_h \quad &:= \quad Q_1(\mathcal{T}_h)^d \quad = \quad \{v_h \in C(\overline{\Omega})^d \,:\, v_h|_T \in Q_1^d(T)\ \forall T \in \mathcal{T}_h\}\,, \\ Q_h \quad &:= \quad P_0(\mathcal{T}_h) \quad = \quad \{p_h \in L^2(\Omega) \,:\, p_h|_T \in P_0(T)\ \forall T \in \mathcal{T}_h\}\,. \end{aligned}$$

This combination is known as the unstable Q_1/P_0 element. The degrees of freedom can be distributed according to Fig. 11.1. For simplicity we consider an equdistant tensor grid in 2D. We will construct a discrete pressure p_h^* which is in the kernel of the divergence operator. In the quadratic cell T_{ij} with the vertices (i,j), $(i{+}1,j)$, $(i,j{+}1)$ and $(i{+}1,j{+}1)$ the corresponding pressure will be denoted by $p^*_{i+1/2,j+1/2}$ and is defined as:

$$p^*_{i+1/2,j+1/2} \quad = \quad \begin{cases} 1 & \text{if } i+j \text{ pair,} \\ -a & \text{if } i+j \text{ unpair.} \end{cases}$$

The parameter $a > 0$ is chosen such that

$$\int_\Omega p_h^*\, d\mathbf{x} \quad = \quad 0,$$

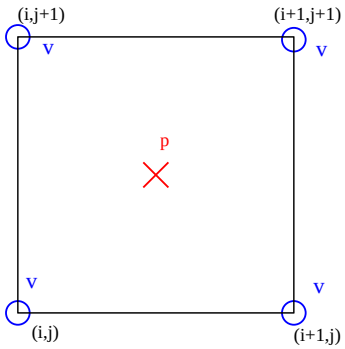


Figure 11.1: Schematische Darstellung des instabilen Q_1/P_0-Elements für Stokes.

hence $p_h^* \in Q_h$. This pressure has the character of checkerboard modes. Let x_T denote the midpoint of the element T_{ij}. The L^2-norm of p_h^* on that element becomes

$$\begin{aligned}\int_{T_{ij}} p_h^* \mathrm{div}\ v_h\, d\mathbf{x} &= p^*_{i+1/2,j+1/2} \int_{T_{ij}} (\partial_x v_h^1 + \partial_y v_h^2)\, d\mathbf{x} \\ &= \frac{1}{2h} h^2 p^*_{i+1/2,j+1/2} \left(v^1_{i+1,j+1} + v^1_{i+1,j} - v^1_{i,j+1} - v^1_{i,j} + \right. \\ &\qquad \left. v^2_{i+1,j+1} + v^2_{i,j+1} - v^2_{i+1,j} - v^2_{i,j}\right).\end{aligned}$$

Here we used the fact that $\mathrm{div}\, v_h$ is linear on the element. Therefore, the trapezoidal rule for approximation of the integral is indeed exact. Taking the sum over all elements and sorting the $v^1_{i,j}$ and $v^2_{i,j}$ terms results in a sum over all inner vertices:

$$\int_\Omega p_h^* \mathrm{div}\ v_h\, d\mathbf{x} = h^2 \sum_{i,j} (v^1_{i,j} \nabla^1_{ij} p_h^* + v^2_{i,j} \nabla^2_{ij} p_h^*),$$

with difference quotients of the pressure:

$$\begin{aligned}\nabla^1_{ij} p_h^* &= \frac{1}{2h}(p_{i+1/2,j+1/2} + p_{i+1/2,j-1/2} - p_{i-1/2,j+1/2} - p_{i-1/2,j-1/2}), \\ \nabla^2_{ij} p_h^* &= \frac{1}{2h}(p_{i+1/2,j+1/2} + p_{i-1/2,j+1/2} - p_{i+1/2,j-1/2} - p_{i-1/2,j-1/2}).\end{aligned}$$

These difference quotients vanish for the upper chosen pressure p_h^* of checkerboard type. Hence, we obtain $(p_h^*, \mathrm{div}\ v_h)_{L^2(\Omega)} = 0$ independent of the discrete velocity $v_h \in Q_1(\mathcal{T}_h)^2$.

11.6.1 Weakly stable Stokes elements

A finite element pair is called *weakly stable* for the Stokes equation, if a weaker version of inf-sup property of the following type is valid: For each mesh size parameter h there exists a constant $\gamma_h > 0$ s.t.

$$\forall p_h \in Q_h\ \exists v_h \in V_h: \quad (\mathrm{div}\ v_h, p_h)_{L^2(\Omega)} \geq \gamma_h |v_h|_{H^1(\Omega)^d} \|p_h\|_{L^2(\Omega)}.$$

Here the constant γ_h is not bounded from below away from zero, but

$$\lim_{h\to 0} \gamma_h \quad = \quad 0.$$

Such a property is obtained, for instance, if one changes the previous unstable pair Q_1/P_0 by eliminating the upper mentioned particular pressure p_h^* from the pressure space. To be more specific, the reduced pressure space becomes

$$Q_h \quad := \quad \{p_h^*\}^\perp \quad = \quad \{p_h \in P_0(\mathcal{T}_h) : (p_h, p_h^*)_{L^2(\omega)} = 0\}.$$

This is motivated by the fact that the discrete pressure space P_0 was obviously too large for the bilinear velocities Q_1^d. However, the resulting finite element pair is not yet inf-sup stable, but only weakly stable in the sense described above. $Q_1/(P_0 \cap \{p_h^*\}^\perp)$ nahezu instabil ist.

In the following sections we will discuss some inf-sup stable elements for the Stokes system

11.7 Mini element

The Mini element is based on the P_1/P_1-discretization, where the velocity space is augmented in a proper way [4]. The augmentation consists of element-wise *bubble functions*. In two dimensions we have on each triangular element $T \in \mathcal{T}_h$ the cubic function

$$b_T(x) \quad := \quad \lambda_{T,1}\lambda_{T,2}\lambda_{T,3} \qquad \forall x \in T,$$

where $\lambda_{T,i}$ denote the barycentric coordinate associated to the edge number i of the element. On the reference triangle, these are give by $\lambda_{\widehat{T},1} = x_1$, $\lambda_{\widehat{T},2} = x_2$ and $\lambda_{\widehat{T},3} = 1 - x_1 - x_2$. Outside of T we set $b_T(x) = 0$. Since b_T already vanishes on all edges of T, these bubble functions are globally continuous with support $\operatorname{supp} b_T \subset T$. In Fig. 11.2 such a bubble function is illustrated. The velocity space is chosen as

$$V_h \quad := \quad \left(P_1(\mathcal{T}_h) \oplus \operatorname{span} \bigcup_{T\in\mathcal{T}_h} \{b_T\} \right)^2.$$

We denote this space by P_1^{bubble}. The pressure is sought in $Q_h = P_1(\mathcal{T}_h) \cap Q$.

Theorem 11.16 *The Mini element P_1^{bubble}/P_1 is inf-sup stable.*

Proof. We show the inf-sup property by the criterion of Fortin in Thm. 9.13. To this end we have to show the existence of a projection $\Pi_h : V \to V_h$ with the orthogonality property

$$(\operatorname{div}(v - \Pi_h v), p_h) \quad = \quad 0 \qquad \forall p_h \in Q_h \; \forall v \in V,$$

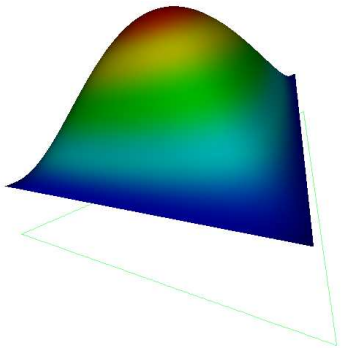


Figure 11.2: Lokal bubble function b_T which is used to augment the P_1-space of the velocities (Mini element).

and the boundedness with an h-independent constant

$$\|\Pi_h v\|_{H^1(\Omega)^d} \leq C\|v\|_{H^1(\Omega)^d} \qquad \forall v \in V.$$

(a) Definition of the projection: The projection will be of the form

$$\Pi_h = C_h + \pi_h(I - C_h),$$

with two further projections $C_h, \pi_h : V \to V_h$. The projection C_h will be a Clément-type interpolation . The second projection will ensure the orthogonality property: We define on each element

$$\beta_{i,T} := \int_T v_i \, dx \left(\int_T b_T \, dx \right)^{-1}, \qquad i = 1, 2.$$

Further, we define component wise

$$(\pi_h v)_i := \sum_{T \in \mathcal{T}_h} \beta_{i,T} b_T.$$

(b) Orthogonality: Due to this construction, the π_h conserves the mean value on each element:

$$\int_T (\pi_h v - v) \, dx = 0 \qquad \forall T \in \mathcal{T}_h.$$

This implies by integration by parts and due to $(v - \pi_h v)|_{\partial\Omega} = 0$ the desired orthogonality:

$$\begin{aligned}(\operatorname{div}(v - \pi_h v), p_h)_{L^2(\Omega)} &= \sum_{T \in \mathcal{T}_h} (\operatorname{div}(v - \pi_h v), p_h)_{L^2(T)} \\ &= \sum_{T \in \mathcal{T}_h} \left((\pi_h v - v, \nabla p_h)_{L^2(T)^d} + \int_{\partial T} (v - \pi_h v) \cdot n p_h \, ds \right).\end{aligned}$$

Now we use the fact that the sum of the integrals over the inner edges vanish due to changing signs of the normals. The integrals over boundary edges vanish, because $v|_{\partial\Omega} = 0$ a.e. and $\pi_h v|_{\partial\Omega} = 0$. The pressure gradients ∇p_h are element-wise constant, so that we obtain:

$$(\operatorname{div}(v - \pi_h v), p_h)_{L^2(\Omega)} \;=\; \sum_{T\in\mathcal{T}_h} (\nabla p_h)|_T \int_T (\pi_h v - v)\, dx \;=\; 0. \tag{11.13}$$

In order to validate a similar property for Π_h we state that

$$v - \Pi_h v \;=\; v - (C_h v + \pi_h(v - C_h v)) \;=\; (I - \pi_h)(v - C_h v).$$

Hence, we obtain

$$(\operatorname{div}(v - \Pi_h v), p_h)_{L^2(\Omega)} \;=\; (\operatorname{div}(I - \pi_h)(v - C_h v), p_h)_{L^2(\Omega)} \;=\; 0.$$

Here we need that $(v - \Pi_h v)|_{\partial\Omega} = 0$ a.e., because it was used for getting (11.13). This property does not hold for the original Clément interpolation, because in general we have $C_h v|_{\partial\Omega} \neq 0$. However, one may use a modified Clément interpolation according to Scott and Zhang [15]. This is based on building mean values along element edges and remains constant traces along boundary edges unchanged.

(c) Stability: The modified Clément interpolation mentioned above has similar stability and approximation properties as the original one: For $v \in H_0^1(\Omega)$ we have

$$\|C_h v\|_{H^1(\Omega)^d} \;\leq\; c\|v\|_{H^1(\Omega)^d}, \tag{11.14}$$

$$\|v - C_h v\|_{L^2(T)^d} \;\leq\; c h_T \|v\|_{H^1(\omega_T)^d}. \tag{11.15}$$

Here, ω_T is the "Clément" patch of the element T. The stability of Π_h is now obtained as follows:

$$\begin{aligned}\|\Pi_h v\|_{H^1(\Omega)^d} &\leq \|C_h v\|_{H^1(\Omega)^d} + \|\pi_h(I - C_h)v\|_{H^1(\Omega)^d} \\ &\leq \|v\|_{H^1(\Omega)^d} + \left(\sum_{T\in\mathcal{T}_h} \|\pi_h(I - C_h)v\|^2_{H^1(T)^d}\right)^{1/2}.\end{aligned}$$

On each element we apply an inverse estimate (see Theorem afterwards):

$$\|\pi_h(I - C_h)v\|_{H^1(T)^d} \;\leq\; c h_T^{-1} \|\pi_h(I - C_h)v\|_{L^2(T)^d}.$$

Furthermore it is easy to show that π_h is L^2-stable, so that we arrive at

$$\begin{aligned}\|\pi_h(I - C_h)v\|_{H^1(T)^d} &\leq c h_T^{-1} \|v - C_h v\|_{L^2(T)^d} \\ &\leq c\|v\|_{H^1(\omega_T)^d}.\end{aligned}$$

This yields

$$\|\Pi_h v\|_{H^1(\Omega)^d} \;\leq\; c\|v\|_{H^1(\Omega)^d}.$$

□

Theorem 11.17 *On shape-regular meshes, the discretization of the Stokes problem with the Mini element P_1^{bubble}/P_1 yields for arbitrary right hand sides $f \in L^2(\Omega)^d$ always a unique solution $\{v_h, p_h\} \in V_h \times Q_h$. This satisfies the a priori error estimate*

$$|v - v_h|_{H^1(\Omega)^d} + \|p - p_h\|_{L^2(\Omega)} \quad \leq \quad Ch\{|v|_{H^2(\Omega)^d} + |p|_{H^1(\Omega)}\},$$

with $h := \max\{h_T : T \in \mathcal{T}_h\}$ and an h-independent constant C.

Proof. Existence and uniqueness follow directly by the inf-sup property according to the previous Theorem. For the a priori estimate we apply Corollary 9.12. Since $P_1(\mathcal{T}_h)^d \subset V_h$ and $P_1(\mathcal{T}_h) = Q_h$ we have

$$\begin{aligned} \inf_{w_h \in V_h} |v - w_h|_{H^1(\Omega)^d} &\leq Ch|v|_{H^2(\Omega)^d}, \\ \inf_{q_h \in Q_h} \|p - q_h\|_{L^2(\Omega)} &\leq Ch|p|_{H^1(\Omega)}. \end{aligned}$$

□

11.8 Inverse estimate

The following theorem can be used to bound stronger norms by weaker norms, as long as the considered function is finite dimensional. For each difference with respect to the order of derivative we "loose" one power of h. We note that the inverse estimate is not valid for general Sobolev functions.

Theorem 11.18 (Inverse estimate) *Let $\{\mathcal{T}_h\}$ be a sequence of shape-regular triangulations of the a domain $\Omega \subset \mathbb{R}^d$, $\{V_h\}$ associated finite element spaces, and $m \geq k \geq 1$. Then it holds the inverse estimates*

$$\begin{aligned} \|u_h\|_{H^{m+k}(T)} &\leq \mu_{inv} h_T^{-k} \|u_h\|_{H^m(T)} \qquad \forall u_h \in V_h, \\ \|u_h\|_{H^{m+k}(\Omega)} &\leq \mu_{inv} h_{min}^{-k} \|u_h\|_{H^m(\Omega)} \qquad \forall u_h \in V_h. \end{aligned}$$

The constant μ_{inv} depend on the (local) polynomial order of V_h, on the maximal aspect ratio κ and on m, but is independent of k and h.

Proof. (a) At first we prove such an estimate on the reference element $\widehat{T}$ with respect to the semi-norm:

$$|u|_{H^{m+k}(\widehat{T})} \quad \leq \quad c|u|_{H^m(\widehat{T})},$$

for polynomials of degree r, i.e. $u \in P_r$. This bound if obviously valid for $r \leq m$. Hence, we assume $r > m$. Let $I : C(\widehat{T}) \to P_{m-1}$ be the nodal interpolation operator at

$s = m(m+1)/2$ (in 2D) pairwise different points $z_1, \ldots, z_s \in \overline{T}$. We have $Iu(z_j) = u(z_j)$ for $j = 1, \ldots, s$, and $|Iu|_{H^{m+k}(\widehat{T})} = |Iu|_{H^m(\widehat{T})} = 0$. This yields

$$|u|_{H^{m+k}(\widehat{T})} \;=\; |u - Iu|_{H^{m+k}(\widehat{T})} \;\leq\; \|u - Iu\|_{H^{m+k}(\widehat{T})}.$$

As already shown in Lemma 5.6,

$$\|\!|u|\!\| \;:=\; |u|_{H^m(\widehat{T})} + \sum_{j=1}^{s} |u(z_j)|.$$

is a norm on $H^m(\widehat{T})$, and therefore also on . Because $\|\cdot\|_{H^{m+k}(\widehat{T})}$ is also a norm on the space $P_r \oplus P_{m-1}$. Due to the finite dimensionality of $P_r \oplus P_{m-1}$, the norms $\|\cdot\|_{H^{m+k}(\widehat{T})}$ and $\|\!|\cdot|\!\|_{H^{m+k}(\widehat{T})}$ are equivalent for $u - Iu \in P_r \oplus P_{m-1}$. This implies

$$\begin{aligned}
|u|_{H^{m+k}(\widehat{T})} &\leq c\|\!|u - Iu|\!\| \\
&= c|u - Iu|_{H^m(\widehat{T})} + c\sum_{j=1}^{s} |(u - Iu)(z_j)| \\
&= c|u - Iu|_{H^m(\widehat{T})} \\
&\leq c|u|_{H^m(\widehat{T})} + c|Iu|_{H^m(\widehat{T})} \\
&= c|u|_{H^m(\widehat{T})}\,.
\end{aligned}$$

(b) By the transformation rule we obtain from (a) an inverse estimate on arbitrary cells $T \in \mathcal{T}_h$:

$$|u_h|_{H^{n+k}(T)} \;\leq\; ch_T^{-k}|u_h|_{H^n(T)} \qquad \forall u_h \in P_r(T)\,.$$

Summation of the squares of such terms of type $|\cdot|_{H^{n+k}(T)}$ with $n \leq m$ and over all elements $T \in \mathcal{T}_h$ yields the assertion. ∎

11.9 The Crouzeix-Raviart element

We already know the Crouzeix-Raviart element from Chapter **??** for the Poisson problem. One may take this element for each velocity component and cell-wise constant pressures:

$$\begin{aligned}
\mathcal{C}R_h \;&:=\; \Big\{ v_h \in L^2(\Omega)^d \,:\, v_h|_T \in P_1 \;\forall T \in \mathcal{T}_h, \\
&\qquad v_h \text{ continuous at midpoints of each (inner) edge} \Big\}\,, \\
Q_h \;&:=\; P_0(\mathcal{T}_h).
\end{aligned}$$

The number of degrees of freedom for the pressure corresponds to the number of elements, and the number of degrees of freedom for the velocities corresponds to the number of inner edges (faces in 3D) multiplied by the dimension (to obtain vector-valued velocities). The discrete velocities will have the following properties:

- $v_h \in L^2(\Omega)^d$,
- $v_h|_T$ linear $\forall T \in T_h$,
- v_h is continuous at the mid points of every (inner) edge of the triangles,
- $v_h(x) = 0$ for mid points x of outer edges.

Since the velocities are discontinuous, this is a non-conforming element (with respect to H^1). However, one can show that this element is inf-sup stable.

11.10 The divergence-free Crouzeix-Raviart element

Now, we consider a Crouzeix-Raviart element for the Stokes system which includes the divergence-free condition by construction. An important aspect is that a restriction to divergence-free velocities, i.e. $V_h \subset V_0$, implies that the Stokes problem is of type (**??**).

Additionally to the properties listed in the previous subsection, the discrete velocities will have the following additional property:

- div $v_h(x) = 0$ for each x in the inner of an element $T \in T_h$,

Hence, the finite element space consists of discontinuos polynomials and remains to be a non-conforming space (with respect to H^1). We denote this space by $\mathcal{C}R_h^{div}$. The formal definition is easier by defining a basis:

(a) To each inner vertex x_i we assign a corresponding basis function u_i^n as follows: Its support is just the patch of triangles having this vertex as node x_i.

$$\begin{aligned} u_i^n(x_j) \cdot n_j &= |E_j|^{-1}, && \text{for } x_i \in E_j \in \mathcal{E}_h(x_j), \\ u_i^n(x_j) \cdot n_j &= 0, && \text{for } x_i \in E_i \in \mathcal{E}_h^{all} \setminus \mathcal{E}_h(x_j), \\ u_i^n(x_j) \cdot t_j &= 0, && \text{for } E_j \in \mathcal{E}_h^{all}. \end{aligned}$$

Here x_j denotes the midpoint of edge E_j, and $\mathcal{E}_h(x_j)$ denotes the set of all inner edges of $\mathcal{T}_h$, which contains x_j as end point. Let $\mathcal{E}_h^{all}$ be the set of all inner and outer edges of the triangulation. The normalized normal vector (in math. positive direction) of vertex x_j is denoted by n_j and the normalized tangent vector (in direction x_j) by t_j.

(b) Furthermore, we assign to each inner edge E_j a basis function u_j^t. The normalized vector u_j^t has the same direction like the tangent t_j at this edge:

$$\begin{aligned} u_j^t(x_j) &= t_j, \\ u_j|_{E_i} &= 0 \qquad \text{if } i \neq j. \end{aligned}$$

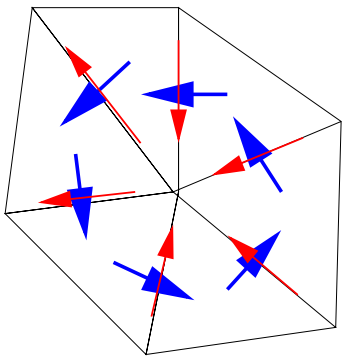


Figure 11.3: Basis functions of the divergence-free Crouzeix-Raviart element. The red arrow demonstrate the tangential velocities u_j^t, the blue arrows demonstrate the velocity u_i^n in normal direction of the edges.

The dimension of $\mathcal{C}R_h^{div}$ corresponds (in the case of Dirichlet conditions for v_h on the entire boundary $\partial\Omega$) to the number of inner vertices N_{nodes} plus the number of inner edges N_{edges}. These basis functions are depict in Figure 11.3. The corresponding finite element space reads

$$\mathcal{C}R_h^{div} \quad := \quad \text{span}\left\{ \bigcup_{1\le i\le N_{nodes}} \{u_i^n\} \cup \bigcup_{1\le j\le N_{edges}} \{u_j^t\} \right\}.$$

On a structured triangular mesh the number of edges is similar to three times the number of vertices (modulo boundary vertices). Therefore this discretization consists on such grids of $4N_{nodes}$ degrees of freedom:

$$\dim \mathcal{C}R_h^{div} \quad \approx \quad 4N_{nodes}.$$

The corresponding discrete variational form reads:

$$v_h \in \mathcal{C}R_h^{div}: \qquad \sum_{T\in\mathcal{T}_h} (\nabla v_h, \nabla\phi)_{L^2(T)^d} \quad = \quad \langle f, \phi\rangle \qquad \forall \phi \in \mathcal{C}R_h^{div}.$$

This problem ist elliptic but non-conforming, because $\mathcal{C}R_h^{div} \not\subset H_0^1(\Omega)$:

$$\sum_{T\in\mathcal{T}_h} (\nabla v_h, \nabla v_h)_{L^2(T)^d} \quad = \quad \sum_{T\in\mathcal{T}_h} |v_h|^2_{H^1(T)} \quad =: \quad \|v_h\|^2_{\mathcal{C}R_h^{div}} \qquad \forall v_h \in \mathcal{C}R_h^{div}.$$

Hence, existence and uniqueness follows directly by the Representation Theorem of Riesz (Thm. 3.19) and also from the Theorem of Lax-Milgram (Thm. 3.22).

In contrast to the Crouzeix-Raviart element for the Poisson problem in Section **??** we have for these elements $P_1(\mathcal{T}_h)^d \not\subset \mathcal{C}R_h^{div}$, because the P_1-vector fields are not necessarily divergence-free. However, one can show the following a priori error estimate:

Theorem 11.19 *On shape regular meshes, we have the following a priori estimate for the solution $v_h \in \mathcal{CR}_h^{div}$:*

$$\|v - v_h\|_h \quad \leq \quad Ch(|v|_{H^2(\Omega)^d} + |p|_{H^1(\Omega)}),$$

supposing the regularity $v \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$.

We are not going to state the proof here, but we refer to the original literature [7]. An important step is to show the following property (formulated for these elements of lowest order):

$$\int_E [v_h]_E \, ds \quad = \quad 0 \qquad \forall v_h \in \mathcal{CR}_h^{div} \quad \forall E \in \mathcal{E}_h \, .$$

Hiere, $[v_h]_E$ is the jump of v_h across the edge E. This property is called *patch test*.

11.11 Taylor-Hood element

The Taylor-Hood element consists of different polynomial degrees for velocity and pressure. For $k \in \mathbb{N}$, the velocities v_h are element-wise of order $k+1$ and the pressure p_h element-wise of order k:

$$\begin{aligned} V_h &:= (H_0^1(\Omega) \cap P_{k+1}(\mathcal{T}_h))^d, \\ Q_h &:= L_0^2(\Omega) \cap P_k(\mathcal{T}_h). \end{aligned}$$

We now demonstrate the inf-sup stability of the lowest-order Taylor-Hood element, $k = 1$.

Theorem 11.20 *On shape-regular tetrahedral meshes the Taylor-Hood element $P_2 - P_1$ satiesfies the inf-sup property (11.12) for the Stokes system.*

Proof. Let $p_h \in Q_h \setminus \{0\}$ be given. Let $v \in V$ the corresponding velocity to the inf-sup property in the infinite dimensional space V, i.e.

$$(\operatorname{div} v, p_h) \quad \geq \quad \gamma |v|_{H^1(\Omega)^d} \|p_h\|_{L^2(\Omega)}.$$

As the discrete counterpart, we chose the Clement interpolant $v_h := \mathcal{C}_h v$. Now, we will show the following bounds:

$$(a) \qquad \frac{(\operatorname{div} v_h, p_h)}{|v_h|_{H^1(\Omega)^d}} \geq \frac{1}{c_1} \left(\gamma \|p_h\|_{L^2(\Omega)} - |p_h|_h \right),$$

$$(b) \qquad \exists w_h \in V_h \text{ s.t. } |p_h|_h \leq c_2 \frac{(\operatorname{div} w_h, p_h)}{|w_h|_{H^1(\Omega)^d}},$$

with a seminorm $|\cdot|_h$. (a) and (b) imply

$$\begin{aligned} S &:= \sup_{v_h \in V_h} \frac{(\operatorname{div} v_h, p_h)}{|v_h|_{H^1(\Omega)^d}} \\ &\geq \frac{1}{c_1} \left(\gamma \|p_h\|_{L^2(\Omega)} - |p_h|_h \right) \\ &\geq \frac{1}{c_1} \left(\gamma \|p_h\|_{L^2(\Omega)} - c_2 S \right), \end{aligned}$$

which yields the discrete inf-sup property

$$S \geq \gamma_h \|p_h\|_{L^2(\Omega)},$$

with the constant $\gamma_h := \gamma/(c_1 + c_2)$. Now, we show (a) and (b).
(a) Due to the H^1-stability of the Clement interpolation, $\|v_h\|_{H^1(\Omega)^d} \leq c_1 \|v\|_{H^1(\Omega)^d}$, and the approximation property $\|v - v_h\|_{H^1(\Omega)^d} \leq c_3 h |v|_{H^1(\Omega)^d}$, we deduce

$$\begin{aligned} \frac{(\operatorname{div} v_h, p_h)}{|v_h|_{H^1(\Omega)^d}} &\geq c_1^{-1} \frac{(\operatorname{div} v_h, p_h)}{|v|_{H^1(\Omega)^d}} \\ &= c_1^{-1} \left(\frac{(\operatorname{div} v, p_h)}{|v|_{H^1(\Omega)^d}} + \frac{(\operatorname{div} (v_h - v), p_h)}{|v|_{H^1(\Omega)^d}} \right) \\ &= c_1^{-1} \left(\gamma \|p_h\|_{L^2(\Omega)} - \frac{(v_h - v, \nabla p_h)}{|v|_{H^1(\Omega)^d}} \right) \\ &\geq c_1^{-1} \left(\gamma \|p_h\|_{L^2(\Omega)} - c_3 \sum_{T \in \mathcal{T}_h} h_T \|\nabla p_h\|_T \right). \end{aligned}$$

Defining the semi-norm as

$$|q|_h := c_3 \left(\sum_{T \in P_k} h_T^2 |q|_{H^1(T)}^2 \right)^{1/2}$$

yields the bound (a).
(b) For the proof of (b) we assume that the triangulation $\mathcal{T}_h$ has a certain patch-structure, although this is not mandatory for the proof, but it makes the proof much simpler. Let $\mathcal{T}_h$ consists of triangles according to the following construction: There exists *special points* $\{a_1, \ldots, a_r\} \subset \{x_1, \ldots, x_{N_{nodes}}\}$, s.t. the patches

$$P_k := \bigcup \{T \in \mathcal{T}_h \,:\, a_k \in T\}$$

form a partition of Ω. Furthermore, each element $T \in \mathcal{T}_h$ has exactly one edge on the boundary of a patch and has one *special point* as vertex. Such a mesh is depicted in Figure 11.4.

We set on the *special points* $w_h(a_i) = 0$ for all $i = 1, \ldots, r$. We define for each $k \in \{1, \ldots, r\}$ the pressure $q_k := p_h|_{P_k}$, and extend it by zero outside of P_k so that $\operatorname{supp} q_k \subset P_k$. Furthermore, we will define $w_k \in V_h$ with $\operatorname{supp} w_k \subset P_k$ and the P_1-part of w_k is set to zero. Hence, the only non-vanishing degrees of freedom of w_k are the quadratic ones. Analogously to the proof of Thm. 11.17 we have

$$(\operatorname{div} w_k, q_k) \quad = \quad -\sum_{T \in P_k} \nabla q_k|_T \cdot \int_T w_k \, dx.$$

Since $w_k|_T$ is a pure quadratic polynomial, it holds (Exercise)

$$\int_T w_k \, dx \quad = \quad \frac{1}{3}|T|(w_k(\alpha_{k,T}) + w_k(\beta_{k,T})),$$

where $\alpha_{k,T}, \beta_{k,T}$ are the edge midpoints of the element T with the joint endpoint x_k. The associated normalized tangential vectors to the edges e_α and e_β will be denoted by t_α and t_β. Now we set

$$w_k(\alpha_{k,T}) \quad := \quad -\frac{\partial q_k}{\partial t_\alpha} t_\alpha |e_\alpha|^2 \quad \text{and} \quad w_k(\beta_{k,T}) \quad := \quad -\frac{\partial q_k}{\partial t_\beta} t_\beta |e_\beta|^2.$$

Note, that q_k is continuous on the edges, which imply that $\partial q_k / \partial t_\alpha$ and $\partial q_k / \partial t_\beta$ do not depend on the special choice of element T, but only on its edges. This feature makes $w_k(\alpha_{k,T})$ and $w_k(\beta_{k,T})$ well-defined. Now, it holds by the inverse estimate (Thm. 11.18):

$$|w_k|_{H^1(T)^d} \quad \leq \quad c h_T^{-1} \|w_k\|_{L^2(T)^d} \quad \leq \quad c h_T |q_k|_{H^1(T)}. \tag{11.16}$$

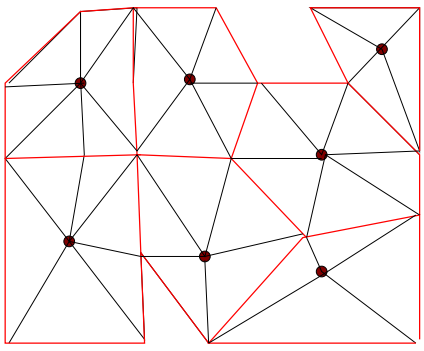


Figure 11.4: Patches of the mesh for proving the inf-sup property of the Taylor-Hood element. The *special points* are colored in dark red.

Now, we have

$$\begin{aligned}(\text{div } w_k, q_k) &= \frac{1}{3}\sum_{T\in P_K} |T|\nabla q_k|_T \cdot \left(\frac{\partial q_k}{\partial t_\alpha} t_\alpha |e_\alpha|^2 + \frac{\partial q_k}{\partial t_\beta} t_\beta |e_\beta|^2\right) \\ &= \frac{1}{3}\sum_{T\in P_K} |T| \left(\frac{\partial q_k}{\partial t_\alpha}\nabla q_k|_T \cdot t_\alpha |e_\alpha|^2 + \frac{\partial q_k}{\partial t_\beta}\nabla q_k|_T \cdot t_\beta |e_\beta|^2\right) \\ &\geq \frac{c}{3}\sum_{T\in P_K} |T| \left(\left(\frac{\partial q_k}{\partial t_\alpha}\right)^2 |e_\alpha|^2 + \left(\frac{\partial q_k}{\partial t_\beta}\right)^2 |e_\beta|^2\right).\end{aligned}$$

Due to the uniform boundedness of the inner angles of all $T \in \mathcal{T}_h$ away from zero and from π, it holds

$$c_1|q_k|^2_{H^1(T)} \leq |T| \left(\left(\frac{\partial q_k}{\partial t_\alpha}\right)^2 + \left(\frac{\partial q_k}{\partial t_\beta}\right)^2\right) \leq c_2|q_k|^2_{H^1(T)}.$$

In combination with (11.16) we obtain:

$$(\text{div } w_k, q_k) \geq c|q_k|^2_h.$$

Furthermore, note that $|w_k|_{H^1(T)^d} \leq c|q_k|_h$. For the function $w_h := \sum_{k=1}^r w_k$ we get

$$(\text{div } w_h, p_h) = \sum_{k=1}^r (\text{div } w_k, q_k) \geq c|p_h|^2_h \geq c|w_h|_{H^1(\Omega)^d}|p_h|_h.$$

This completes the proof. ∎

Theorem 11.21 *For the Taylor-Hood solution $\{v_h, p_h\} \in (H^1_0(\Omega) \cap P_2(\mathcal{T}_h)^d) \times (L^2_0(\Omega) \cap P_1(\mathcal{T}_h))$ holds on shape-regular meshes of a polygonal bounded domain Ω the a priori error estiamte*

$$|v - v_h|_{H^1(\Omega)^d} + \|p - p_h\|_{L^2(\Omega)} \leq Ch^k(|v|_{H^{k+1}(\Omega)^d} + |p|_{H^k(\Omega)}),$$

for $k \in \{1, 2\}$, if $v \in H^{k+1}(\Omega)^d$ and $p \in H^k(\Omega)$.

Proof. The assertion follows from Corollary 9.12 and the approximation property

$$\begin{aligned}\inf_{w_h\in P_2(\mathcal{T}_h)^d} |v - w_h|_{H^1(\Omega)^d} &\leq Ch^k|u|_{H^{k+1}(\Omega)^d}, \\ \inf_{q_h\in P_1(\mathcal{T}_h)} \|p - q_h\|_{L^2(\Omega)} &\leq Ch^k|p|_{H^k(\Omega)}.\end{aligned}$$

∎

11.12 Residual-free bubbles

Motivated by the inf-sup stable Mini-element, we consider in this section a much wider class of "bubble"-functions. We replace the space of the cubic bublles span $\bigcup\{b_T : T \in \mathcal{T}_h\}$ by

$$B_{RF} \;:=\; \bigoplus_{T\in\mathcal{T}_h} \{H_0^1(T)^d\}.$$

The velocity space reads now formally $V_h := (P_1(\mathcal{T}_h)^d) \oplus B_{RF}$ and the pressure space remains to be $Q_h := P_1(\mathcal{T}_h) \cap Q$. Since the velocity space includes the Mini-element, we already now that this combination is inf-sup stable. However, note that the velocity space is still infinite dimensional. We make the following assumptions:

- The right hand side f is cell-wise constant.

The velocities can be uniquely decomposed into

$$v_h \;=\; v_L + v_B$$

with $v_L \in P_1(\mathcal{T}_h)^d$ and $v_B \in B_{RF}$. Now, we will condensate the part v_B out of the equations. The equations can also be split into:

$$\begin{aligned}
(\nabla(v_L+v_B),\nabla\phi_L)_{L^2(\Omega)^d} - (p_h,\mathrm{div}\;\phi_L)_{L^2(\Omega)} &= \langle f,\phi_L\rangle \qquad \forall\phi_L\in P_1(\mathcal{T}_h)^d, &(11.17)\\
(\nabla(v_L+v_B),\nabla\phi_B)_{L^2(\Omega)^d} - (p_h,\mathrm{div}\;\phi_B)_{L^2(\Omega)} &= \langle f,\phi_B\rangle \qquad \forall\phi_B\in B_{RF}, &(11.18)\\
(\mathrm{div}\;(v_L+v_B),\xi)_{L^2(\Omega)} &= 0 \qquad \forall\xi\in Q_h. &(11.19)
\end{aligned}$$

We have a closer look onto the second equation with test function $\phi_{B,T} \in B_{RF}$ with $\mathrm{supp}\,\phi_{B,T} \subset T$, i.e. $\phi_{B,T} \in H_0^1(T)^d$:

$$(\nabla v_L,\nabla\phi_{B,T})_{L^2(T)^d} \;=\; (-\Delta v_L,\phi_{B,T})_{L^2(T)^d} + \int_{\partial T} \nabla v_L\,\phi_{B,T}\cdot n\,ds \;=\; 0,$$

because $\Delta v_L|_T = 0$ and $\phi_{B,T}|_{\partial T} = 0$. Due to the same reason we have

$$-(p_h,\mathrm{div}\;\phi_B)_{L^2(T)} \;=\; (\nabla p_h,\phi_B)_{L^2(T)^d} \;=\; \nabla p_h|_T \int_T \phi_B\,dx.$$

Due to the assumption on $f \in L^2(\Omega)$ we can write

$$\langle f,\phi_B\rangle \;=\; f|_T \int_T \phi_B\,dx.$$

Therefore, equation (11.18) can be written in the form

$$(\nabla v_B,\nabla\phi_B)_{L^2(T)^d} \;=\; (f-\nabla p_h)|_T \int_T \phi_B\,dx \qquad \forall\phi_{B,T}\in H_0^1(T)^d.$$

This equation should hold for all elements $T \in \mathcal{T}_h$ and corresponds to the elliptic partial differential equation

$$\begin{aligned} -\Delta v_B &= f - \nabla p_h \qquad \text{in } T, \\ v_B &= 0 \qquad\qquad \text{auf } \partial T, \end{aligned}$$

where $f - \nabla p_h$ is assumed to be element-wise constant. Let us denote the solution operator of this equation by L_T, i.e. $v_B|_T = L_T(f - \nabla p_h)$.

In equation (11.17) we have

$$(\nabla v_B, \nabla \phi_L)_{L^2(\Omega)^d} = \sum_{T \in \mathcal{T}_h} \left((v_B, -\Delta \phi_L)_{L^2(T)^d} + \int_{\partial T} v_B \cdot n \, \phi_L \, ds \right) = 0,$$

because $\Delta \phi_L|_T = 0$ and $v_B|_{\partial T} = 0$. Therefore, the system (11.17)-(11.19) becomes now

$$\begin{aligned} (\nabla v_L, \nabla \phi)_{L^2(\Omega)^d} - (p_h, \operatorname{div} \phi)_{L^2(\Omega)} &= \langle f, \phi \rangle \qquad \forall \phi \in P_1(\mathcal{T}_h)^d, \\ (\operatorname{div} v_L, \xi)_{L^2(\Omega)} - \sum_{T \in \mathcal{T}_h} (\operatorname{div} L_T(\nabla p_h - f), \xi)_{L^2(T)} &= 0 \qquad \forall \xi \in Q_h. \end{aligned}$$

We write this system in a more compact form by seeking $\{v_L, p_h\} \in X_h$, s.t.

$$A(\{v_L, p_h\}, \{\phi, \xi\}) + S_h(p_h, \xi) = \langle F_h, \{\phi, \xi\} \rangle \qquad \forall \{\phi, \xi\} \in X_h, \tag{11.20}$$

with $X_h := (V \cap P_1(\mathcal{T}_h)^d) \times Q_h$ and the bilinear form A, including all Galerkin parts, and S_h, including the additional terms due to the condensation of the bubble parts:

$$\begin{aligned} A(\{v_L, p_h\}, \{\phi, \xi\}) &:= (\nabla v_L, \nabla \phi)_{L^2(\Omega)^d} - (p_h, \operatorname{div} \phi)_{L^2(\Omega)} + (\operatorname{div} v_L, \xi)_{L^2(\Omega)}, \\ S_h(p_h, \xi) &:= -\sum_{T \in \mathcal{T}_h} (\operatorname{div} L_T \nabla p_h, \xi)_{L^2(T)}, \\ \langle F_h, \{\phi, \xi\} \rangle &:= \langle f_h, \phi \rangle - \sum_{T \in \mathcal{T}_h} (\operatorname{div} L_T f, \xi)_{L^2(T)}. \end{aligned}$$

Due to the property of element-wise constant $\nabla p_h - f$, we know that the solution of the cell-problems can be written as

$$L_T(\nabla p_h - f)|_T := b_T(\nabla p_h - f),$$

where $b_T \in C_0^2(T)$ is the solution of the equation

$$\begin{aligned} -\Delta b_T &= 1 \qquad \text{in } T, \\ b_T &= 0 \qquad \text{on } \partial T. \end{aligned}$$

If we integrate the terms of $S_h(\cdot, \cdot)$ by parts, the boundary integrals vanish (due to the homogeneous Dirichlet values of b_T), so that

$$S_h(p_h, \xi) = \sum_{T \in \mathcal{T}_h} (b_T \nabla p_h, \nabla \xi)_{L^2(T)} = \sum_{T \in \mathcal{T}_h} \alpha_T (\nabla p_h, \nabla \xi)_{L^2(T)},$$

with the parameter

$$\alpha_T \;:=\; |T|^{-1}\int_T b_T\,dx.$$

Here, we used the fact that $\nabla p_h - f$ and $\nabla\xi$ are element-wise constant, so that we can take the mean value of b_T out of the integral. As an exercise we leave the fact that on shape-regular meshes holds

$$c_1 h_T^2 \;\le\; \alpha_T \;\le\; c_2 h_T^2,$$

with constants $c_1, c_2 > 0$ only dependent of the maximal aspect ratio κ.

The resulting system (11.20) coincides with a stabilized Finite-Element method. We will analyze such a system in the next section.

11.13 The PSPG method

11.13.1 PSPG of lowest order

We consider in this sction the following stabilized finite element method for Stokes, which is called *Pressure-Stabilized-Petrov-Galerkin* (PSPG) method:

$$u_h \in X_h: \qquad A_h(u_h,\varphi) \;=\; \langle F_h,\varphi\rangle \qquad \forall\varphi=\{\phi,\xi\}\in X_h, \tag{11.21}$$

with

$$\begin{aligned} A_h(u_h,\varphi) &:= A(u_h,\varphi) + S_h(p_h,\xi),\\ S_h(p_h,\xi) &:= c\sum_{T\in\mathcal{T}_h} h_T^2(\nabla p_h,\nabla\xi)_{L^2(T)}\,,\\ \langle F_h,\{\phi,\xi\}\rangle &:= \langle f_h,\phi\rangle + c\sum_{T\in\mathcal{T}_h} h_T^2(f,\nabla\xi)_{L^2(T)}. \end{aligned}$$

Lemma 11.22 *The stabilized P_1/P_1-Stokes system of the form (11.21), $r\ge 1$, satisfies the following modified inf-sup property: It exists $\gamma>0$, s.t.*

$$\inf_{p_h\in Q_h\setminus\{0\}}\left(\sup_{v_h\in V_h}\frac{(div\ v_h,p_h)}{|v_h|_{H^1(\Omega)}\|p_h\|_{L^2(\Omega)}}+\frac{S_h(p_h,p_h)^{1/2}}{\|p_h\|_{L^2(\Omega)}}\right)\;\ge\;\gamma.$$

Proof. Let $p_h\in Q_h$ be given. Due to the continuous inf-sup property, there exists $\gamma>0$ and $v\in V$, s.t. for arbitrary $v_h\in V_h$ it holds

$$\begin{aligned}\gamma &\le \frac{(\mathrm{div}\ v,p_h)}{|v|_{H^1(\Omega)}\|p_h\|_{L^2(\Omega)}}\\ &= \frac{(\mathrm{div}\ v_h,p_h)}{|v_h|_{H^1(\Omega)}\|p_h\|_{L^2(\Omega)}}\frac{|v_h|_{H^1(\Omega)}}{|v|_{H^1(\Omega)}}+\frac{(\mathrm{div}\ (v-v_h),p_h)}{|v|_{H^1(\Omega)}\|p_h\|_{L^2(\Omega)}}\,.\end{aligned}$$

Now, we are looking for a special interpolation operator $C_h : V \to V_h$ s.t. we have for $v_h := C_h v$ the following properties:

$$\begin{aligned} |C_h v|_{H^1(\Omega)} &\leq c|v|_{H^1(\Omega)}, \\ |(\text{div } (v - C_h v), p_h)| &\leq cS_h(p_h, p_h)^{1/2}|v|_{H^1(\Omega)}. \end{aligned}$$

By showing the validity of these estimates, we obtain the claim. We take once more the Scott-Zhang variant of the Clément interpoland (see proof of Thm. 11.17 or [15]). Hence, we have the upper mentioned H^1-stability. It remains to show the second bound. Due to the properties (11.14) and (11.15) we deduce:

$$\begin{aligned} (\text{div } (v - C_h v), p_h) &= (v - C_h v, \nabla p_h) \\ &\leq \sum_{T \in \mathcal{T}_h} \|v - C_h v\|_T \|\nabla p_h\|_T \\ &\leq c \sum_{T \in \mathcal{T}_h} h_T |v|_{H^1(T)} \|\nabla p_h\|_T \\ &\leq c \left(\sum_{T \in \mathcal{T}_h} |v|^2_{H^1(T)} \right)^{1/2} \left(\sum_{T \in \mathcal{T}_h} h_T^2 \|\nabla p_h\|_T^2 \right)^{1/2} \\ &= c|v|_{H^1(\Omega)} S_h(p_h, p_h)^{1/2}. \end{aligned}$$

□

Theorem 11.23 *The PSPG system with P_1/P_1 elements (11.21) has for each $f \in L^2(\Omega)$ and $c > 0$ a unique solution $u_h = \{v_h, p_h\} \in X_h$. Furthermore, the following stability property holds:*

$$|v_h|_{H^1(\Omega)^d} + \|p_h\|_{L^2(\Omega)} + c \sum_{T \in \mathcal{T}_h} h_T |p_h|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)^d},$$

with a constant $C = C(\Omega, \nu)$.

Proof. (a) We have a finite-dimensional linear system with the same number of unknowns (ansatz functions) as equations (test functions). Hence, solvability and uniqueness is equivalent to the uniqueness of the solution of the homogeneous system, i.e. with $f \equiv 0$. By diagonal testing we obtain

$$\begin{aligned} 0 &= A(\{v_h, p_h\}, \{v_h, p_h\}) + S_h(p_h, p_h) \\ &= (\nabla v_h, \nabla v_h)_\Omega - (p_h, \text{div } v_h)_\Omega + (\text{div } v_h, p_h)_\Omega + c \sum_{T \in \mathcal{T}_h} h_T^2 (\nabla p_h, \nabla p_h)_T \\ &= \|\nabla v_h\|^2_\Omega + c \sum_{T \in \mathcal{T}_h} h_T^2 \|\nabla p_h\|^2_T. \end{aligned}$$

This implies $\|\nabla v_h\|_{L^2(\Omega)^d} = 0$ and hence $v_h \equiv 0$. For the discrete pressure we obtain $\|\nabla p_h\|^2_{L^2(T)} = 0$ for each element $T \in \mathcal{T}_h$. This means that p_h is element-wise constant. Due to continuity of p_h, the pressure is globally constant. Due to vanishing mean of $p_h \in Q$ we deduce $p_h \equiv 0$.
(b) Stability: Diagonal testing as in part (a) but with not necessarily vanishing f yields:

$$\|\nabla v_h\|^2_\Omega + c \sum_{T\in\mathcal{T}_h} h_T^2(\|\nabla p_h\|^2_T - (f, \nabla p_h)_T) \quad = \quad (f, v_h)_\Omega.$$

By the inequality of Hölder we obtain the upper bound

$$\begin{aligned}\|\nabla v_h\|^2_\Omega + c \sum_{T\in\mathcal{T}_h} h_T^2\|\nabla p_h\|^2_T \quad &= \quad (f, v_h)_\Omega + c \sum_{T\in\mathcal{T}_h} h_T^2(f, \nabla p_h)_T \\ &\leq \quad \|f\|_\Omega\|v_h\|_\Omega + c \sum_{T\in\mathcal{T}_h} h_T^2\|f\|_T\|\nabla p_h\|_T.\end{aligned}$$

Using the inequality of Poincaré $\|v_h\|_\Omega \leq c_\Omega\|\nabla v_h\|_\Omega$ and Young's inequality yields

$$\|\nabla v_h\|^2_\Omega + c \sum_{T\in\mathcal{T}_h} h_T^2\|\nabla p_h\|^2_T \quad \leq \quad \frac{1}{2}c_\Omega^2\|f\|^2_\Omega + \frac{1}{2}\|\nabla v_h\|^2_\Omega + c \sum_{T\in\mathcal{T}_h} h_T^2 \left(\frac{1}{2}\|f\|^2_T + \frac{1}{2}\|\nabla p_h\|^2_T\right).$$

Subtracting identical terms on both sides leads to

$$\frac{1}{2}\|\nabla v_h\|^2_\Omega + \frac{c}{2} \sum_{T\in\mathcal{T}_h} h_T^2\|\nabla p_h\|^2_T \quad \leq \quad \frac{1}{2}c_\Omega^2\|f\|^2_\Omega + \frac{c}{2} \sum_{T\in\mathcal{T}_h} h_T^2\|f\|^2_T.$$

The sum on the right hand side can also be bounded by $\|f\|^2_\Omega$. Taking the square root on both side implies

$$|v_h|_{H^1(\Omega)^d} + c \sum_{T\in\mathcal{T}_h} h_T|p_h|_{H^1(\Omega)} \quad \leq \quad C\|f\|_\Omega.$$

(c) It remains to show an upper bound for the L^2-norm of the pressure. This is obtained by the modified inf-sup property of Lemma 11.22: Basic calculus yields the form

$$\gamma\|p_h\|_{L^2(\Omega)} \quad \leq \quad S_h(p_h, p_h)^{1/2} + \sup_{\phi_h\in V_h} \frac{(\text{div } \phi_h, p_h)}{|\phi_h|_{H^1(\Omega)}}.$$

We use the fact that p_h solves the discrete variational formulation of the Stokes equation:

$$\frac{(\text{div } \phi_h, p_h)}{|\phi_h|_{H^1(\Omega)}} \quad = \quad \frac{-(f, \phi_h) + (\nabla v_h, \nabla\phi_h)}{|\phi_h|_{H^1(\Omega)}} \quad \leq \quad \|f\|_{H^{-1}(\Omega)} + |v_h|_{H^1(\Omega)}.$$

Because of $\|f\|_{H^{-1}(\Omega)} \leq c_\Omega\|f\|_{L^2(\Omega)}$ we obtain

$$\|p_h\|_{L^2(\Omega)} \quad \leq \quad \frac{1}{\gamma}\left(S_h(p_h, p_h)^{1/2} + c_\Omega\|f\|_\Omega + |v_h|_{H^1(\Omega)}\right).$$

Using the bound derived in part (b) of this proof yields the assertion. ∎

Theorem 11.24 *The a priori estimate of the PSPG-method for linear elements (P_1/P_1) reads*

$$\|\nabla(v-v_h)\| + \|p-p_h\| + \left(\sum_T h_T^2 \|\nabla(p-p_h)\|_T^2\right)^{1/2} \quad \leq \quad ch\left(|v|_{H^2(\Omega)^d} + |p|_{H^1(\Omega)}\right),$$

if $(v,p) \in H^2(\Omega)^d \times H^1(\Omega)$.

Proof. As usual, we split the error $e_v = v - v_h$ in interpolation error $\eta_v = v - i_h v$ and projection error $\xi_v = i_h v - v_h$; analogously for the pressure variable, $p - p_h = e_p = \eta_p + \xi_p$, but with the Clément interpolation. The bound of the interpolation error is an immediate consequence of the Bramble-Hilbert lemma:

$$\|\nabla\eta_v\| + \|\eta_p\| + \sum_T h_T \|\nabla(p-p_h)\|_T \quad \leq \quad ch\left(|v|_{H^2(\Omega)^d} + |p|_{H^1(\Omega)}\right).$$

Now, we are going to bound $\|\nabla\xi_v\|$ properly. It holds for $\xi = (\xi_v, \xi_p)$ by Galerkin orthogonality:

$$\begin{aligned}
\|\nabla\xi_v\|^2 + S_h(\xi_p,\xi_p) &= A_h(\xi,\xi) \\
&= A_h(\xi,\xi) - A_h(e,\xi) \\
&= A_h(\eta,\xi) \\
&= (\nabla\eta_v, \nabla\xi_v) - (\eta_p, \operatorname{div}\xi_v) + (\operatorname{div}\eta_v, \xi_p) + S_h(\eta_p,\xi_p).
\end{aligned}$$

The individual terms can be bounded as:

$$\begin{aligned}
(\nabla\eta_v, \nabla\xi_v) &\leq \|\nabla\eta_v\|^2 + \frac{1}{4}\|\nabla\xi_v\|^2, \\
-(\eta_p, \operatorname{div}\xi_v) &\leq c\|\eta_p\|^2 + \frac{1}{4}\|\nabla\xi_v\|^2, \\
(\operatorname{div}\eta_v, \xi_p) &= (\eta_v, \nabla\xi_p) \leq \sum_T h_T^{-2}\|\eta_v\|_T h_T \|\nabla\xi_p\|, \\
&\leq \sum_T h_T^{-2}\|\eta_v\|_T^2 + \frac{1}{4}S_h(\xi_p,\xi_p), \\
S_h(\eta_p,\xi_p) &\leq S_h(\eta_p,\eta_p) + \frac{1}{4}S_h(\xi_p,\xi_p).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\|\nabla\xi_v\|^2 + S_h(\xi_p,\xi_p) &\leq c\left(\|\nabla\eta_v\|^2 + \|\eta_p\|^2 + \sum_T h_T^{-2}\|\eta_v\|_T^2 + S_h(\eta_p,\eta_p)\right) \\
&\leq ch^2\left(|v|^2_{H^2(\Omega)^d} + |p|^2_{H^1(\Omega)}\right).
\end{aligned}$$

In combination with the interpolation estimate we arrive at the desired upper bound for $\|\nabla(v - v_h)\|$ and $\sum_T h_T^2 \|\nabla(p - p_h)\|_T^2$.
The bound for the L^2-error of the pressure is obtained by the discrete inf-sup property:

$$\gamma \|e_p\| \quad \leq \quad \sup_{\phi_h \in V_h} \frac{(\text{div } \phi_h, p - p_h)}{\|\nabla \phi_h\|} + S_h(e_p, e_p)^{1/2}.$$

Galerkin orthogonality yields

$$(\text{div } \phi_h, p - p_h) \quad = \quad (\nabla e_v, \nabla \phi_h) \quad \leq \quad \|\nabla e_v\| \|\nabla \phi_h\|.$$

Hence, we obtain the last remaining bound

$$\gamma \|e_p\| \quad \leq \quad \|\nabla e_v\| + S_h(e_p, e_p)^{1/2}.$$

□

11.13.2 PSPG of arbitrary order

This PSPG method can also be used for higher polynomial degrees $r \geq 1$, but the concrete form of the stabilization term must be slightly modified. For P_r/P_r or Q_r/Q_r elements the form reads

$$S_h(u_h, \xi) \quad := \quad c \sum_{T \in \mathcal{T}_h} h_T^2 (-\Delta v_h + \nabla p_h, \nabla \xi)_{L^2(T)}.$$

The stabilization is not only dependent on the pressure but also on the velocity. In the case of P_1-elements on triangles or tetrahedrons, or Q_1-elements on parallelogram meshes, the velocity term vanishes, because then $\partial_{xx}^2 v_h|_T = \partial_{yy}^2 v_h|_T = 0$.

If we formulate the resulting system in a strong form, we arrive at

$$\begin{aligned} -\Delta v + \nabla p &= f \qquad \text{in } \Omega, \\ \text{div } v - h^2 \text{div } (-\Delta v + \nabla p - f) &= 0 \qquad \text{in } \Omega, \\ v &= 0 \qquad \text{auf } \partial\Omega. \end{aligned}$$

This indicates that the exakt solution of the Stokes system also fulfills this equation, because the additional stabilization terms are based on the strong residuum of one equation.

Lemma 11.25 *The PSPG-method is* strongly consistent*, i.e. the exact solution $u \in X$ satiesfies in the case of sufficient regularity also the discrete equations.*

Proof. In the case that the exact solution $u = \{v, p\} \in X$ is sufficiently regular, it is also a classical solution. Then it satisfies in particular the equation

$$-\Delta v + \nabla p \quad = \quad f \qquad \text{in } \Omega.$$

Hence, we have for $\varphi = \{\phi, \xi\}$:

$$S_h(u_h, \xi) - \langle F_h, \varphi \rangle \quad = \quad -(f, \phi),$$

and therefore

$$A_h(u, \varphi) + S_h(u, \xi) \quad = \quad \langle F_h, \varphi \rangle.$$

□

11.14 Algebraic system for the discrete Stokes problem

In this section we formulate the discrete Stokes problem as a linear algebraical system of equations. To this end, we express the solution in terms of a basis of ansatz functions. Here, we assume that the system is inf-sup stable so that no additional stabilization terms appear. Possible examples of finite elements are the Mini-element or Taylor-Hood elements. For the Lagrange-basis we use the following notation

$$\begin{aligned} V_h &= \text{span}\,\{\phi_i \,:\, 1 \le i \le N_v\}, \\ Q_h &= \text{span}\,\{\xi_i \,:\, 1 \le i \le N_p\}. \end{aligned}$$

N_v and N_p are the number of degrees of freedom for v_h and p_h, respectively. The vector for describing the entire system consists of velocity and pressure components $u_h = \{v_h, p_h\}$, i.e.

$$v_h(x) \quad = \quad \sum_{i=1}^{N_v} \phi_i(x) v_i, \quad p_h(x) \; = \; \sum_{i=1}^{N_p} \xi_i(x) p_i.$$

From now on we neglect the subsript h, because we do not expect any confusion with the continuous solution. Let $f \in \mathbb{R}^{Nv+N_p}$ be the vector of the right hand side. The linear system reads

$$\mathcal{A}u \quad = \quad f.$$

We order the degrees of freedom for v and p separately, so that the stiffness matrix of the Stokes problem becomes block structured:

$$\mathcal{A} \quad = \quad \begin{pmatrix} A & B^T \\ -B & 0 \end{pmatrix}.$$

The coefficients of the blocks $A = (a_{ij})$ and $B = (b_{ij})$ are given by

$$a_{ij} \quad := \quad \int_\Omega \nabla\phi_j \, \nabla\phi_i \, dx, \qquad b_{ij} \; := \; -\int_\Omega \text{div}\ \phi_j \, \xi_i \, dx.$$

Hence, the linear system is of saddle-point form

$$\begin{pmatrix} A & B^T \\ -B & 0 \end{pmatrix} \begin{pmatrix} v \\ p \end{pmatrix} \;=\; \begin{pmatrix} f \\ 0 \end{pmatrix},$$

because the diagonal blocks for the pressure are zero. The stiffness matrix $\mathcal{A}$ is obviously not a M-matrix. It is not positive definite neither, because

$$\langle \mathcal{A}u, u\rangle \;=\; \langle Av, v\rangle + \langle B^T p, v\rangle - \langle Bv, p\rangle \;=\; \langle Av, v\rangle \;=\; \|v\|_A^2.$$

In particular, for vanishing velocities we have $\langle \mathcal{A}u, u\rangle = 0$. Hence, iterative standard methods as Jacobi, Gauss-Seidel or the conjugate-gradient method (CG) are not appropriate for this stiffness matrix $\mathcal{A}$.

11.14.1 Schur complement of saddle-point problems

In this section we show how the positive-definiteness of the matrix block $A \in \mathbb{R}^{dN_v \times dN_v}$ (Laplace matrix) can be used, to solve the discrete Stokes system. The idea consists in inverting A within an outer iteration. Due the regularity of A, the linear system

$$\begin{aligned} Av + B^T p &= f, \\ -Bv &= g, \end{aligned}$$

is equivalent to

$$\begin{aligned} BA^{-1}B^T p &= BA^{-1}f - Bv, \\ -Bv &= g. \end{aligned}$$

Note that for inf-sup stable elements for Stokes, the right hand side g vanishes. However, we keep this term for later use (not inf-sup stable, stabilized finite elements). We introduce the so-called *Schur complement*

$$S \;:=\; BA^{-1}B^T.$$

We obtain the following reduced system for the pressure:

$$Sp \;=\; BA^{-1}f + g. \tag{11.22}$$

Knowing p, the velocities are obtained by the regular system (Poisson problem in the case of the Stokes system)

$$Av \;=\; f - B^T p.$$

However, the difficulty is that A^{-1} is usually not given and cannot be obtained by Gaussian elimination (or other method), because it not sparse (although A is sparse). Usually, only the action of A^{-1} to vectors, as e.g. $A^{-1}f$, can be computed.

Lemma 11.26 *For inf-sup stable finite elements for the Stokes system, the Schur complement $S = BA^{-1}B^T$ is symmetric positive definite.*

Proof. Since A is symmetric positive definite (spd), A^{-1} is spd. This implies the symmetry of S. Each pressure vector $p \in \mathbb{R}^{N_p} \setminus \{0\}$ does not lie in the kernel of B^T, because we assumed discrete inf-sup stability. Therefore, we have $q := B^T p \neq 0$ and

$$\langle p, Sp\rangle \;=\; \langle p, BA^{-1}B^T p\rangle \;=\; \langle q, A^{-1}q\rangle \;> 0.$$

□

11.14.2 Uzawa algorithm with constant step lenght

The idea of the Uzawa algorithm is to solve the equation (11.22) iteratively by a gradient method, because S is symmetric positive definite. In order to describe this method we define the defect of the pressure $p^{(k-1)}$:

$$d^{(k)} \;:=\; BA^{-1}f + g - Sp^{(k-1)}.$$

This defect corresponds to the negative gadient of the corresponding quadratic funktional

$$F(p) \;:=\; \frac{1}{2}\langle Sp, p\rangle - \langle BA^{-1}f + g, p\rangle.$$

The new iterate $p^{(k)}$ is obtained by adding the defekt, multiplied by a step length α, to the previous iterate, $p^{(k)} := p^{(k-1)} + \alpha d^{(k)}$. However, it is recommended to use the mass matrix $M = (m_{ij})$, $m_{ij} := (\xi_j, \xi_i)_{L^2(\Omega)}$, as preconditioner for the pressure. Hence, the new pressure becomes .

$$p^{(k)} \;=\; p^{(k-1)} + \alpha M^{-1} d^{(k)}.$$

The defect can also be written as

$$\begin{aligned} d^{(k)} &= BA^{-1}f + g - BA^{-1}B^T p^{(k-1)} \\ &= BA^{-1}(f - B^T p^{(k-1)}) + g \\ &= Bv^{(k)} + g, \end{aligned}$$

if $v^{(k)}$ solves the equation

$$Av^{(k)} \;=\; f - B^T p^{(k-1)}.$$

In combination with a stopping criterion with a desired tolerance $TOL > 0$ we obtain the entire algorithm, see Table 11.1. Here, $\|\cdot\|$ denotes a vector norm in $\mathbb{R}^{N_p}$. For the convergence of the Uzawa algorithm, the choice of the step lenght $\alpha > 0$ is crucial. It must be small enough to obtain convergence. We now state a convergence criterion:

`Uzawa algorithm (with constant step lenght)`

(i) Choose an initial value for the pressure, e.g. $p^{(0)} \equiv 0$. Set $k = 0$.

(ii) Increase $k \to k+1$ and solve the Poisson problem

$$Av^{(k)} = f - B^T p^{(k-1)}.$$

(iii) Determine the defect $d^{(k)} := Bv^{(k)} + g$.

(iv) Invert the mass matrix: $Me^{(k)} = d^{(k)}$.

(v) Make a pressure correction

$$p^{(k)} := p^{(k-1)} + \alpha e^{(k)}.$$

(vi) If $\alpha \|e^{(k)}\| > TOL$, goto (ii).

Table 11.1: Uzawa algorithm with constant step length.

Theorem 11.27 *Let λ_{max} be the maximal eigenvalue of the matrix $M^{-1}S$. The Uzawa algorithm with constant step lenght $0 < \alpha < 2/\lambda_{max}$ is convergent for inf-sup stable finite elements and the limit is the exact solution $\{v, p\}$.*

Proof. The gradient method with constant step lenght is a Richardson iteration[1]. This is convergent for symmetric positive definite matrices if

$$\alpha < \frac{2}{\lambda_{max}}.$$

□

The optimal choice of a constant step lenght α for a Richardson iteration is

$$\alpha = \frac{2}{\lambda_{min} + \lambda_{max}},$$

where λ_{min} is the minimal eigenvalue of $M^{-1}S$.

Lemma 11.28 *Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive definite. Then we have the following identity:*

$$\|x\|_A = \max_{y \in \mathbb{R}^n} \frac{\langle Ax, y \rangle}{\|y\|_A}$$

[1]Lewis Fry Richardson, 11.10.1881-30.09.1953, British mathematician and physician. He particularly contributed to the question of metereological forecast by help of mathematical models.

for the vector norm

$$\|y\|_A \;:=\; \sqrt{\langle Ay, y\rangle}.$$

Proof. Due to the symmetry of $A = A^{1/2}A^{1/2}$ we have

$$\begin{aligned}\langle Ax, y\rangle &= \langle A^{1/2}x, A^{1/2}y\rangle \;\le\; \|A^{1/2}x\|\,\|A^{1/2}y\| \\ &= \Big(\langle A^{1/2}x, A^{1/2}x\rangle\langle A^{1/2}y, A^{1/2}y\rangle\Big)^{1/2} \\ &= (\langle Ax, x\rangle\langle Ay, y\rangle)^{1/2} \;=\; \|x\|_A\|y\|_A.\end{aligned}$$

This implies

$$\|x\|_A \;\ge\; \max_{y\in\mathbb{R}^n}\frac{\langle Ax, y\rangle}{\|y\|_A}.$$

With the particular choice $y := x$, we obtain the opposite direction:

$$\max_{y\in\mathbb{R}^n}\frac{\langle Ax, y\rangle}{\|y\|_A} \;\ge\; \frac{\langle Ax, x\rangle}{\|x\|_A} \;=\; \|x\|_A.$$

□

Lemma 11.29 *For the maximal and minimal eigenvalue of $M^{-1}S$ it holds:*

$$0 \;<\; \gamma^2 \;\le\; \lambda_{min} \;\le\; \lambda_{max} \;\le\; c.$$

Here, γ is the discrete inf-sup constant and c another constant, which can be chosen as $c = 1$ in the case of conforming finite elements, and $c = 4$ elsewhere.

Proof. (a) Firstly, we express the eigenvalues λ in an appropriate form. For each eigenvalue λ of $M^{-1}S$ exists an eigenvector y_λ, s.t. $M^{-1}Sy_\lambda = \lambda y_\lambda$. This implies due to the symmetry of M:

$$\lambda \;=\; \frac{\|M^{-1}Sy_\lambda\|_M}{\|y_\lambda\|_M} \;=\; \frac{\langle Sy_\lambda, M^{-1}Sy_\lambda\rangle^{1/2}}{\|y_\lambda\|_M} \;=\; \lambda^{1/2}\frac{\langle Sy_\lambda, y_\lambda\rangle^{1/2}}{\|y_\lambda\|_M}.$$

And therefore

$$\lambda \;=\; \frac{\langle Sy_\lambda, y_\lambda\rangle}{\|y_\lambda\|_M^2}.$$

Using the notation $x_\lambda := A^{-1}B^Ty_\lambda \in \mathbb{R}^{N_v}$, the nominator can be written in the form

$$\langle Sy_\lambda, y_\lambda\rangle \;=\; \langle BA^{-1}B^Ty_\lambda, y_\lambda\rangle \;=\; \langle A^{-1}B^Ty_\lambda, B^Ty_\lambda\rangle \;=\; \langle x_\lambda, Ax_\lambda\rangle \;=\; \|x_\lambda\|_A^2.$$

By help of Lemma 11.28 we have $\|x_\lambda\|_A = \max_z(\langle Ax_\lambda, z\rangle/\|z\|_A)$. Therefore,

$$\langle Sy_\lambda, y_\lambda\rangle \;=\; \max_{z\in\mathbb{R}^{N_v}}\frac{\langle Ax_\lambda, z\rangle^2}{\|z\|_A^2} \;=\; \max_{z\in\mathbb{R}^{N_v}}\frac{\langle B^Ty_\lambda, z\rangle^2}{\|z\|_A^2}.$$

Hence, each eigenvalue λ can be expressed in the form

$$\lambda \;=\; \max_{z\in\mathbb{R}^{N_v}} \frac{\langle B^T y_\lambda, z\rangle^2}{\|z\|_A^2 \|y_\lambda\|_M^2}.$$

(b) For the smallest eigenvalue λ_{min} this leads to:

$$|\lambda_{min}| \;\geq\; \min_{y\in\mathbb{R}^{N_p}} \max_{z\in\mathbb{R}^{N_v}} \frac{\langle B^T y, z\rangle^2}{\|z\|_A^2 \|y\|_M^2} \;=\; \left(\min_{y\in\mathbb{R}^{N_p}} \max_{z\in\mathbb{R}^{N_v}} \frac{\langle y, Bz\rangle}{\|z\|_A \|y\|_M} \right)^2.$$

Now, we express this correspondence in terms of the finite element functions $y \leftrightarrow p_h$ and $z \leftrightarrow v_h$. We have

$$\|z\|_A \;=\; \|\nabla v_h\|_{L^2(\Omega)}\,, \quad \|y\|_M \;=\; \|p_h\|_{L^2(\Omega)}\,, \quad \langle y, Bz\rangle \;=\; (p_h, \mathrm{div}\; v_h)_{L^2(\Omega)}.$$

We obtain with the discrete inf-sup constant γ:

$$|\lambda_{min}| \;\geq\; \left(\min_{p_h\in Q_h} \max_{v_h\in V_h} \frac{(p_h, \mathrm{div}\; v_h)_{L^2(\Omega)}}{\|\nabla v_h\|_{L^2(\Omega)} \|p_h\|_{L^2(\Omega)}} \right)^2 \;\geq\; \gamma^2.$$

(c) Analogously, we deduce

$$|\lambda_{max}| \;\leq\; \left(\max_{p_h\in Q_h} \max_{v_h\in V_h} \frac{(p_h, \mathrm{div}\; v_h)_{L^2(\Omega)}}{\|\nabla v_h\|_{L^2(\Omega)} \|p_h\|_{L^2(\Omega)}} \right)^2 \;\leq\; \left(\max_{v_h\in V_h} \frac{\|\mathrm{div}\; v_h\|_{L^2(\Omega)}}{\|\nabla v_h\|_{L^2(\Omega)}} \right)^2.$$

For $d \leq 3$ it holds $\|\mathrm{div}\; v_h\| \leq c\|\nabla v_h\|$. In the case of conforming finite elements, one can show such a bound with $c = 1$, and in the non-conforming case with $c = 2$. ∎

The numerical costs of each Uzawa iteration consists in solving the Poisson problem in step (ii), a matrix-vector product in step (iii) to compute the defect, and inversion of the mass matrix M in step (iv). The Poisson problem is usually solved by an iterative method as well, e.g. by an efficient multigrid method. In the case that step (ii) is solved only approximatively, the tolerance TOL must be chosen small enough. Otherwise, the residual may stagnate. This can be improved by formulating step (ii) in the algorithm of Table 11.1 in form of a defect correction :

$$\begin{aligned} Aw^{(k)} &= f - B^T p^{(k-1)} - Av^{(k-1)}, \\ v^{(k)} &:= v^{(k-1)} + w^{(k)}. \end{aligned}$$

Here, the right hand side for the equation of $w^{(k)}$ can also be expressed in different form requirering less algebraic operations:

$$\begin{aligned} f - B^T p^{(k-1)} - Av^{(k-1)} &= -B^T p^{(k-1)} + B^T p^{(k-2)} \\ &= -B^T (p^{(k-1)} - p^{(k-2)}) \\ &= -\alpha B^T e^{(k-1)} \\ &= -\alpha B^T M^{-1} d^{(k-1)}. \end{aligned}$$

Such inner defect correction leads to further possibilities to reduce the numerical costs. For instance, instead of inverting the matrix A a matrix $\tilde{A}$, which is simpler to invert, can be used:

$$v^{(k)} := v^{(k-1)} - \alpha \tilde{A}^{-1} B^T M^{-1} d^{(k-1)}.$$

11.14.3 Uzawa algorithm with optimal step lenght

The optimal step length for the gradient method is obtained by a variable step length with line search. For solving an equation of the form $\mathcal{A}x = b$, the approximative solution $x^{(k-1)}$ and correction direction $e^{(k)}$ in iteration k reads with optimal step lenght

$$\alpha_k = \frac{\langle b - \mathcal{A}x^{(k-1)}, e^{(k)} \rangle}{\langle \mathcal{A}e^{(k)}, e^{(k)} \rangle}.$$

In our case, $b = BA^{-1}f - g$ and $\mathcal{A} = S$ leads in iteration k to the step lenght

$$\alpha_k = \frac{\langle d^{(k)}, e^{(k)} \rangle}{\langle A^{-1}B^T e^{(k)}, B^T e^{(k)} \rangle}. \tag{11.23}$$

However, in order to use this step length given in (11.23) requires additional numerical costs due to the two scalar products and due to the need to compute $A^{-1}B^T e^{(k)}$ which corresponds to a solution of an additional Poisson problem. The latter can be circumvented by reordering the iteration stated above. The corresponding pseudo-code is given in Table 11.2.

The following theorem ensures linear convergent in the energy norm

$$\|p\|_S := \sqrt{\langle Sp, p \rangle}.$$

Theorem 11.30 *The Uzawa algorithm with optimal step length is for inf-sup stable finite elements linearly convergent to the exact solution $\{v, p\}$. The convergent rate with respect to the pressure error is given by $r = (\kappa - 1)/(\kappa + 1)$, i.e.*

$$\|p - p^{(k)}\|_S \leq r^k \|p - p^{(0)}\|_S.$$

Proof. By the theory of the gradient method we know that this method is linearly convergent with a rate dependent of the spectral condition number $\kappa = \kappa(M^{-1}S)$ of the preconditioned Schur complement $M^{-1}S$. It is defined as the ratio of the maximal and minimal eigenvalue of $M^{-1}S$: $\kappa = \lambda_{max}/\lambda_{min}$. The rate r can also be expressed in the form:

$$r = \frac{(\lambda_{max}/\lambda_{min}) - 1}{(\lambda_{max}/\lambda_{min}) + 1} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}.$$

□

Now we ask if this rate depends on the mesh size. It results that κ is independent of h.

`Uzawa algorithmus (with optimal step lenght)`

(i) Set $k = 0$, $\alpha_0 = 1$, $v^{(0)} = 0$. Choose start value for the pressure, e.g.. $p^{(0)} \equiv 0$, solve Poisson problem: $Aw^{(0)} = f - B^T p^{(0)}$.

(ii) Increase $k \to k+1$ and compute velocity:

$$v^{(k)} := v^{(k-1)} + \alpha_{k-1} w^{(k-1)}.$$

(iii) Compute defect: $d^{(k)} := Bv^{(k)} + g$.

(iv) Invert the mass matrix: $Me^{(k)} = d^{(k)}$.

(v) Solve Poisson problem: $Aw^{(k)} = -B^T e^{(k)}$.

(vi) Determine step lenght

$$\alpha_k := \frac{\langle d^{(k)}, e^{(k)} \rangle}{\langle -w^{(k)}, B^T e^{(k)} \rangle}.$$

(vii) Pressure update

$$p^{(k)} := p^{(k-1)} + \alpha_k e^{(k)}.$$

(viii) If $\alpha_k \|e^{(k)}\| > TOL$, goto (ii).

Table 11.2: Uzawa algorithm with optimal step lenght.

Corollary 11.31 *The spectral condition number $\kappa = \kappa(M^{-1}S)$ of the Schur complement of an inf-sup stable finite element discretization on shape-regular meshes bounded from above independently of h (with notation as in Lemma 11.29):*

$$\kappa \leq \frac{c}{\gamma^2}.$$

Proof. The assertion follows directly from Lemma 11.29. □

Although the condition number does not depend on the mesh size, it is still possible that κ is large. This is particularly the case, if the domain or the cells are anisotropic which leads to small inf-sup constants γ. In such cases, an even faster iteration is the cg-iteration.

11.14.4 CG-method for the Stokes-Schur complement

The conjugate-gradient (cg) method requires two additional auxiliary vectors. In the pseudo-code in Table 11.3, these are named $q^{(k)}$ and $q^{(k+1)}$. The theory of the cg method

gives us the following linear convergent rate:

$$r \;=\; \frac{1-\kappa^{-1/2}}{1+\kappa^{-1/2}} \;=\; \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \;=\; \frac{\lambda_{max}^{1/2}-\lambda_{min}^{1/2}}{\lambda_{max}^{1/2}+\lambda_{min}^{1/2}}.$$

The iteration error is bounded as

$$\|p-p^{(k)}\|_S \;\leq\; 2r^k\|p-p^{(0)}\|_S.$$

11.14.5 Uzawa method for stabilized Stokes systems

Now we consider the Stokes system with pressure stabilization:

$$\begin{aligned} Av+B^Tp &= f\,,\\ -Bv+Cp &= g\,, \end{aligned}$$

Here, the symmetric matrix $C=(c_{ij})_{1\leq i,j\leq N_p}$ for the pressure-pressure coupling is added as an additional matrix block. In the case of the PSPG-method with P_1-elements, the matrix entries read

$$c_{ij} \;:=\; c\sum_{T\in\mathcal{T}_h} h_T^2(\nabla\xi_j,\nabla\xi_i)_{L^2(T)}.$$

The corresponding factorization is now given by

$$\begin{pmatrix} A & B^T\\ -B & C\end{pmatrix} \;=\; \begin{pmatrix} A & 0\\ -B & S\end{pmatrix}\begin{pmatrix} I & A^{-1}B^T\\ 0 & I\end{pmatrix},$$

with a modified Schur complement

$$S \;:=\; BA^{-1}B^T+C.$$

The reduced equation now reads

$$(BA^{-1}B^T+C)p \;=\; BA^{-1}f+g.$$

Now, we try to figure out what kind of changes are needed in the Uzawa algorithm: Step (iii) in Algorithm 11.1 and 11.2 to determine the defect should be replaced by

$$\text{(iii')}\quad d^{(k)} \;:=\; g+Bv^{(k)}-Cp^{(k-1)}.$$

Instead of the pure mass matrix as preconditioner, now $(M+C)^{-1}$ should be used. Hence, step (iv) should be replaced by

$$\text{(iv')}\quad (M+C)e^{(k)} \;:=\; d^{(k)}.$$

Furthermore, the step length in step (vi) of Algorithm 11.2 should be modified as

$$\text{(vi')}\quad \alpha_k \;:=\; \frac{\langle d^{(k)},e^{(k)}\rangle}{\langle -Bw^{(k)}+Cp^{(k-1)},e^{(k)}\rangle}.$$

`Uzawa-CG algorithm`

(i) Set $k = 0$, $\alpha_0 = 1$, $v^{(0)} = 0$, $d^{(0)} = 0$, $p^{(0)} \equiv 0$. Compute defect:

$$\begin{aligned} \text{solve } Aw^{(0)} &= f - B^T p^{(0)}, \\ q^{(1)} &:= g + Bw^{(0)}. \end{aligned}$$

(ii) Increase $k \to k+1$ and compute velocity:

$$v^{(k)} := v^{(k-1)} + \alpha_{k-1} w^{(k-1)}.$$

(iii) Compute search direction:

$$d^{(k)} := q^{(k)} + \frac{\langle q^{(k)}, M^{-1} q^{(k)} \rangle}{\langle q^{(k-1)}, M^{-1} q^{(k-1)} \rangle} d^{(k-1)}.$$

(iv) Invert the mass matrix: $Me^{(k)} = d^{(k)}$.

(v) Solve Poisson problem: $Aw^{(k)} = -B^T e^{(k)}$.

(vi) Determine step length

$$\alpha_k := \frac{\langle q^{(k)}, e^{(k)} \rangle}{\langle -w^{(k)}, B^T e^{(k)} \rangle}.$$

(vii) Pressure update

$$p^{(k)} := p^{(k-1)} + \alpha_k e^{(k)}.$$

(viii) Determine new defect

$$q^{(k+1)} := q^{(k)} - \alpha_k B w^{(k)}.$$

(ix) If $\|q^{(k+1)}\| > TOL$, goto step (ii).

Table 11.3: Uzawa-CG algorithm.

11.15 Boundary conditions for Stokes

In the sections before we considered only Dirichlet conditions for the velocities, $v|_{\partial\Omega} = v_0$. In the analysis we even restricted to homogeneous Dirichlet conditions, $v|_{\partial\Omega} = 0$. However, this is not a restriction, because inhomogeneous Dirichlet conditions can be reformulated

into homogeneous ones, by modofication of the forcing term f. This was already shown in Section 3 for the Poisson problem. In this section, we will present another meaningful boundary condition for Stokes.

11.15.1 Outflow boundary conditions

A very important boundary condition is the so-called *do-nothing* condition to model outflow boundaries. The derivative of the velocity in normal direction is coupled to the pressure by

$$\frac{\partial v}{\partial n} - p \cdot n \;=\; 0 \qquad \text{auf } \Gamma_N \subset \partial\Omega. \tag{11.24}$$

Dirichlet conditions can be used on the remaining part of $\Gamma_D \subseteq \partial\Omega$, so that $\partial\Omega = \Gamma_D \cup \Gamma_N$. This outflow condition is often used in combination with an inflow condition, e.g. on $\Gamma_D = \Gamma_{D,1} \cup \Gamma_{D,0}$,

$$v|_{\Gamma_{D,1}} \;=\; v_{in} \quad \text{and} \quad v|_{\Gamma_{D,0}} \;=\; 0.$$

The corresponding variational formulation is still of the form (11.5)-(11.6). The terms in (11.24) are automatically included in the variational formulation due to the integration by parts of the pressure gradient and the viscose term. However, the function spaces are different: Since the pressure p is part of the boundary condition (11.24), the mean value of the pressure cannot be normalized as before. The variational spaces are now given by

$$V \;:=\; \{v \in H^1(\Omega) \,:\, v|_{\Gamma_D} = 0 \text{ a.e.}\} \qquad \text{and} \quad Q \;:=\; L^2(\Omega)\,. \tag{11.25}$$

Note, that the Poiseuille flow

$$v(x,y) \;=\; (y(H-y),0)^T$$

in a rectangular channel $\Omega = (0,L) \times (0,H)$ still satisfies the boundary condition (11.24) at the right boundary $x = L$.

Lemma 11.32 *Assuming that the weak solution $\{v,p\}$ of the variational problem (11.5)-(11.6), formulated in the spaces (11.25) is sufficiently regular, i.e. $v \in H^2(\Omega)$ and $p \in H^1(\Omega)$. Then the mean value of the pressure vanishes on straight boundaries Γ_N:*

$$\int_{\Gamma_N} p\, ds \;=\; 0.$$

Proof. For arbitrary test function $\phi \in V$ we have with the solution $v \in H^2(\Omega)$ and $p \in H^1(\Omega)$:

$$\begin{aligned}(f,\phi) &= (\nabla v, \nabla\phi) - (p, \operatorname{div}\,\phi) \\ &= -(\Delta v + \nabla p, \phi) + \int_{\Gamma_N} \left(\frac{\partial v}{\partial n} - pn\right)\phi\, ds.\end{aligned}$$

We choose a sequence of test functions $(\phi_k)_{k\in\mathbb{N}}$ with $\|\phi_k\|_{L^2(\Omega)} \to 0$ and $\phi_k|_{\Gamma_N} \equiv 1$. Then, the domain integrals above become arbitrary small, so that we deduce

$$\int_{\Gamma_N} pn\,ds \;=\; \int_{\Gamma_N} \frac{\partial v}{\partial n}\,ds\,.$$

We now express the vector of the normal derivatives of v in terms of the normal and tangential component, v_n and v_t, respectively:

$$\frac{\partial v}{\partial n} \;=\; \frac{\partial}{\partial n}(v_n n + v_t t) \;=\; \frac{\partial v_n}{\partial n} n + \frac{\partial v_t}{\partial n} t.$$

Note that $\frac{\partial n}{\partial n} = \frac{\partial t}{\partial t} = 0$ on Γ_N, because Γ_N is assumed to have zero curvature. We further deduce

$$\int_{\Gamma_N} pn\,ds \;=\; \int_{\Gamma_N} \left(\frac{\partial v_n}{\partial n} n + \frac{\partial v_t}{\partial n} t \right) ds.$$

This implies in particular

$$\int_{\Gamma_N} p\,ds \;=\; \int_{\Gamma_N} \frac{\partial v_n}{\partial n}\,ds.$$

Now we write the divergence in terms of the tangential and normal component:

$$0 \;=\; \mathrm{div}\; v \;=\; \frac{\partial v_n}{\partial n} + \frac{\partial v_t}{\partial t}.$$

We obtain by applying the main theorem of integration and differential calculus:

$$\int_{\Gamma_N} p\,ds \;=\; -\int_{\Gamma_N} \frac{\partial v_t}{\partial t}\,ds \;=\; v_t(x_1) - v_t(x_2),$$

where x_1 and x_2 are the initial- and end-point of Γ_N, respectively. Because these are points on Γ_D, and hence $v_t(x_1) = v_t(x_2) = 0$, the proof is complete. ∎

Gravitational forces. Due to the presrciption of the pressure mean on outflow boundaries, we have to take particular caution in the case of gravitational forces in the right hand side. Let us consider the 2D case with right hand side of the form $f = (0, g)^T$ with a constant g. For homogenous Dirichlet values for the velocities $v = 0$ on Γ_D, the right hand side of the equation $-\Delta v + \nabla p = f$ can be compensated by a pressure gradient. Hence, the solution becomes $v \equiv 0$ and $\partial_y p = g$, which implies

$$p(x, y) \;=\; p_0 + g \cdot (y - y_{max}),$$

with y_{max}, the maximal y-coordinate of Ω, and p_0 an arbitrary constant. This pressure at rest is layered. For negative g (e.g. on the earth $g = -9.81\,m/s^2$), the maximal pressure

is located at the lowest point with $y = y_{min}$. In combination with the outflow condition on a straight boundary $\Gamma_{out} = \{(x_0, y) : y_{min} \le y \le y_{max}\}$ we obtain:

$$\int_{y_{min}}^{y_{max}} (p_0 + g \cdot (y - y_{max}))\, dy \quad = \quad 0.$$

In the case that we only have one outflow boundary part, this condition leads to a normalization of p_0. In the case of multiple outflow boundary parts, such a vertical layer of the pressure may only be possible by employing suitable mean pressures on the right hand side of the outflow boundary condition.

Stability. Testing the bilinear form diagonally, yields for the solution $|v|^2_{H^1(\Omega)} \le \|f\|_{H^{-1}(\Omega)}$. However, since we do not have $v \in H^1_0(\Omega)$, we cannot use the standard Poincaré inequality to bound the L^2-norm of v by its gradient ∇v. We have to apply the more general Theorem 3.16. This requires that the Dirichlet boundary Γ_D is not a Lebesgue null set in $\mathbb{R}^{d-1}$. Now we can deduce

$$\|v\|_{H^1(\Omega)} \quad \le \quad c_\Omega \|f\|_{H^{-1}(\Omega)},$$

which ensures existence and uniqueness in the case of the do-nothing condition (11.24):

Theorem 11.33 *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz-domain, $f \in L^2(\Omega)^d$, and the Dirichlet boundary $\Gamma_D \subset \partial\Omega$ should have a positive Lebesgue-measure (in $\mathbb{R}^{d-1}$). Then there exists a unique solution of the Stokes problem (11.5)-(11.6), formulated in the variational spaces (11.25).*

Proof. (a) The existence of a solution is obtained as in the proof of Theorem 11.1.
(b) Due to the linearity of the problem, for uniqueness it is sufficient to consider the homogeneous problem, i.e. $f = 0$. The norm of the solution v of the homogeneous Stokes Problem is bounded by the right hand side f, hence $\|v\|_{H^1(\Omega)} = 0$. ∎

11.15.2 Boundary condition with prescribed pressure difference

In certain configurations, we want to prescribe a pressure difference between inlet Γ_{in} and outlet Γ_{out} of the domain Ω, instead of prescribing the velocity. If we consider no-slip boundary conditions on the remaining boundary Γ_D, i.e. $v|_{\Gamma_D} = 0$, the variational formulation reads

$$\begin{aligned} (\text{div } v, \xi) \quad &= \quad 0 && \forall \xi \in Q, \\ (\nabla v, \nabla \phi) - (p, \text{div } \phi) \quad &= \quad (f, \phi) - p_{in} \int_{\Gamma_{in}} \phi \cdot n \, ds - p_{out} \int_{\Gamma_{out}} \phi \cdot n \, ds \qquad && \forall \phi \in V. \end{aligned}$$

Here, $p_{in}, p_{out} \in \mathbb{R}$ are given pressure mean values for the boundary Γ_{in} and Γ_{out}. The space V becomes in this case

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

Chapter 12

Parabolic problems

12.1 Heat equation

Let $\Omega \subset \mathbb{R}^d$ be a domain and $I = (0, T]$ with $T > 0$ a given time interval. The variabla $x \in \Omega$ denotes a point in space, and $t \in I$ a time instant. We seek a function $u : \Omega \times I \to \mathbb{R}, (x, t) \mapsto u(x, t)$ s.t.

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta u &= f && \text{in } \Omega \times I, \\ u &= 0 && \text{on} \partial\Omega \times I, \\ u(\cdot, 0) &= u_0 && \text{in } \Omega. \end{aligned}$$

With respect to time we have an initial value problem (IVP) , and with respect to space we have a boundary value problem (BVP). This equation is called heat equation *heat equation*, because it describes the temporal and spatial distribution of heat (temperature). However, this is only a prototypical example. Many other physical phenomena (e.g., diffusion processes) can be described by such type of partial differential equations . In the particular case that the solution u is constant in time, i.e. $\partial_t u = 0$, we again obtain the previously discussed Poisson problem. In this case, we speak about a stationary process.

In order to discretize in time we introduce a finite number of time points $0 = t_0 < t_1 < \ldots < t_M = T$. The corresponding time steps are denoted by $k_m := t_m - t_{m-1}$ and the sub-intervals by $I_m := (t_{m-1}, t_m]$.

12.2 Backward Euler method

Assuming that we already know $u_{m-1} := u(\cdot, t_{m-1})$ for $m \geq 1$, the backward Euler method uses the following backward difference quotient for the time derivative:

$$\frac{\partial u}{\partial t}(t_m) \approx \frac{1}{k_m}(u(t_m) - u(t_{m-1})).$$

Moreover, we use as (approximative) right hand side $\overline{f}_m := k_m^{-1} \int_{t_{m-1}}^{t_m} f(\cdot,t)\,dt$. At time t_m, we now consider for $u_m \approx u(t_m)$ the equation

$$\begin{aligned} \frac{u_m - u_{m-1}}{k_m} - \Delta u_m &= \overline{f}_m \qquad \text{in } \Omega, \\ u_m &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

The corresponding weak formulation in $V = H_0^1(\Omega)$ now reads

$$u_m \in V: \qquad k_m^{-1}(u_m, v)_\Omega + (\nabla u_m, \nabla v)_\Omega = k_m^{-1}(u_{m-1}, v)_\Omega + (\overline{f}_m, v)_\Omega \qquad \forall v \in V.$$

Here, we used the short notation $(\cdot,\cdot)_\Omega$ for the L^2-scalar product in Ω. Using finite elements with a triangulation $\mathcal{T}_m$ and corresponding spaces $V_m \subset V$, we seek $u_{h,m} \in V_m$ s.t.

$$k_m^{-1}(u_{h,m}, v)_\Omega + (\nabla u_{h,m}, \nabla v)_\Omega = k_m^{-1}(u_{h,m-1}, v)_\Omega + (\overline{f}_m, v)_\Omega \qquad \forall v \in V_m.$$

Let $\{\phi_i\}$ be a basis of V_m and U the corresponding vector representing $u_{h,m}$ in this basis, the linear system reads

$$(k_m^{-1} M + A)U = b.$$

Here, $M = (m_{ij})$ denotes the mass matrix, and $A = (a_{ij})$ the stiffness matrix with the entries

$$\begin{aligned} m_{ij} &= (\phi_j, \phi_i)_\Omega, \\ a_{ij} &= (\nabla\phi_j, \nabla\phi_i)_\Omega. \end{aligned}$$

The resulting linear system is still symmetric. Note that in the case of small time steps, $k_m^{-1} \gg 1$, the mass matrix becomes the dominant term. For large time steps, the Laplace part dominates.

12.2.1 A priori error estimate

In the following analysis, we use the semi-norm

$$\|u\|_{k,\infty} := \max_{1\le m\le M} \|u(t_m)\|_{L^2(\Omega)},$$

which is actually a norm for discrete functions (in time). Furthermore, we use the notation

$$\|u\|_{L^2(I_m;H_0^1(\Omega))} := \left(\int_I |u(t)|_{H^1(\Omega)}^2 \, dt \right)^{1/2} = \left(\int_I \int_\Omega |\nabla u(x,t)|^2 \, dx\, dt \right)^{1/2}.$$

In the following we will use the so-called *Ritz projection* $R_h^m u$ of $u_m := u(t_m)$, i.e.

$$R_h^m u \in V_h: \qquad (\nabla R_h^m u, \nabla v)_\Omega = (\nabla u_m, \nabla v)_\Omega \qquad \forall v \in V_h.$$

This projection will be used in the algorithm to project the initial value u_0 to $u_{h,0} := R_h^0 u_0$.

Theorem 12.1 *Let Ω be a convex domain or a domain with C^2-boundary, $\{\mathcal{T}_h\}$ is a family of shape-regular meshes (constant in time) with maximal element size $h := \max_{T\in\mathcal{T}_h} h_T$. For the exact solution of the heat equation we assume $u \in H^1(I, H_0^1(\Omega))$ and $\|u\|_{k,\infty} < \infty$. Then we have for the discrete backward Euler solution with P_1- or Q_1-finite elements the following a-priori estimate:*

$$\|u - u_h\|_{k,\infty} \;\leq\; c\max\left(1, \left(\frac{T}{k}\right)^{1/2}\right) h^2 \|\nabla^2 u\|_{k,\infty} + \left(\sum_{i=1}^{M} k_i^2 \|\partial_t u\|^2_{L^2(I_i;H_0^1(\Omega))}\right)^{1/2}.$$

This estimate states in the case of a uniform time step $k = k_m$:

$$\|u - u_h\|_{h,\infty} \;=\; O(h^2 k^{-1/2} + k).$$

This is not satisfactory due to the factor $k^{-1/2}$. However, one can avoid this factor for triangulations independent of time. We give some details after the proof.

Proof. Let $R_h^m u$ be the *Ritz projection* of $u_m := u(t_m)$. We split the error into $\xi_m := u_m - R_h^m u$ and $\eta_m := u_{h,m} - R_h^m u \in V_h$. According to Corollary 5.19 we can express the error $\|\xi_m\|_{L^2(\Omega)}$ in terms of the H^2-semi-norm of u_m:

$$\|\xi_m\|_{L^2(\Omega)} \;\leq\; ch^2 \|\Delta u_m\|_{L^2(\Omega)} \;\leq\; ch^2 |u_m|_{H^2(\Omega)} \;\leq\; ch^2 \|\nabla^2 u\|_{k,\infty}.$$

We will show the following bound:

$$\|\eta_m\| \;\leq\; c_1(\|\xi_0\| + \|\xi_m\|) + A_m \tag{12.1}$$

with

$$A_m^2 \;\leq\; \sum_{i=1}^{m} k_i^2 \|\partial_t u\|^2_{L^2(I_i;H_0^1(\Omega))} + cTk^{-1}h^4 \|\nabla^2 u\|^2_{k,\infty}.$$

This bound of $\|\eta_m\|$ will give us the assertion via

$$\begin{aligned}
\|u - u_h\|_{k,\infty} &\leq \max_{1\leq m\leq M} (\|\xi_m\|_{L^2(\Omega)} + \|\eta_m\|_{L^2(\Omega)}) \\
&\leq c\|\xi_0\| + \max_{1\leq m\leq M} ((c_1+1)\|\xi_m\|_{L^2(\Omega)} + A_m) \\
&\leq c \max_{0\leq m\leq M} \|\xi_m\|_{L^2(\Omega)} + \max_{1\leq m\leq M} A_m \\
&\leq c\left(1 + (T/k)^{1/2}\right) h^2 \|\nabla^2 u\|_{k,\infty} + \left(\sum_{i=1}^{M} k_i^2 \|\partial_t u\|^2_{L^2(I_i;H_0^1(\Omega))}\right)^2
\end{aligned}$$

Hence, it remains to verify (12.1). To this end we use the telescope series

$$\|\eta_m\|^2 \;=\; \sum_{i=1}^{m} (\|\eta_i\|^2 - \|\eta_{i-1}\|^2),$$

where we used that the initial value is given by the Ritz projection as well, so that $\eta_0 = 0$. For the arising terms in the sum the following identity is used which holds for arbitrary scalar products $(\cdot,\cdot)$ with corresponding norm $\|\cdot\|$:

$$\|a\|^2 - \|b\|^2 = 2(a-b,a) - \|a-b\|^2.$$

Here, we chose $a := \eta_i$, $b := \eta_{i-1}$, and ($\|\cdot\| := \|\cdot\|_{L^2(\Omega)}$, $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$):

$$\|\eta_i\|^2 - \|\eta_{i-1}\|^2 = 2(\eta_i - \eta_{i-1}, \eta_i) - \|\eta_i - \eta_{i-1}\|^2.$$

The last term will appear later with positive sign, so that this term will be compensated. For the first term on the right hand side we use for arbitrary $\phi \in V_h$ the property of the Ritz projection:

$$\begin{aligned}(\eta_i - \eta_{i-1}, \phi) &= k_m(\overline{f}_i, \phi) - k_i(\nabla u_{h,i}, \nabla\phi) - (R_h^i u - R_h^{i-1}u, \phi)\\ &= k_i(\overline{f}_m, \phi) - k_i(\nabla(\eta_i + u_i), \nabla\phi) - (u_i - u_{i-1}, \phi)\\ &\quad -(R_h^i u - u_m - R_h^{i-1}u + u_{i-1}, \phi).\end{aligned}$$

Taking the particular choice $\phi := \eta_m$ yields

$$\|\eta_i\|^2 - \|\eta_{i-1}\|^2 = 2(T_1 + T_2) - \|\eta_i - \eta_{i-1}\|^2.$$

with

$$\begin{aligned}T_1 &:= k_i(\overline{f}_i, \eta_i) - k_i(\nabla(\eta_i + u_i), \nabla\eta_i) - (u_i - u_{i-1}, \eta_i),\\ T_2 &:= (R_h^i u - u_m - R_h^{i-1}u + u_{i-1}, \eta_i).\end{aligned}$$

The term T_1 can be reformulated by using the heat equation and integration by parts as:

$$\begin{aligned}T_1 &= \int_{t_{i-1}}^{t_i} (f - \partial_t u, \eta_i)\, dt - k_i(\nabla u_i, \nabla\eta_i) - k_m\|\nabla\eta_i\|^2\\ &= \int_{t_{i-1}}^{t_i} (\nabla(u - u_i), \nabla\eta_i)\, dt - k_i\|\nabla\eta_i\|^2.\end{aligned}$$

We now use the identity

$$\int_{t_{i-1}}^{t_i} (u - u_i)\, dt = \int_{t_{i-1}}^{t_i} (t_{i-1} - t)\partial_t u\, dt,$$

leading to

$$\begin{aligned}T_1 &= \int_{t_{i-1}}^{t_i i} (t_{i-1} - t)(\nabla\partial_t u, \nabla\eta_i)\, dt - k_i\|\nabla\eta_i\|^2\\ &\le \int_{t_{i-1}}^{t_i} (\frac{1}{4}k_i^2\|\nabla\partial_t u\|^2 + \|\nabla\eta_i\|^2)\, dt - k_i\|\nabla\eta_i\|^2\\ &\le \frac{1}{4}k_i^2 \int_{t_{i-1}}^{t_i} \|\partial_t\nabla u\|^2\, dt + k_i\|\nabla\eta_i\|^2 - k_i\|\nabla\eta_i\|^2\\ &= \frac{1}{4}k_i^2 \int_{t_{i-1}}^{t_i} \|\partial_t\nabla u\|^2\, dt.\end{aligned}$$

The term T_2 involving the Ritz projection can be reformulated as:

$$T_2 = (R_h^i u - u_i, \eta_i) - (R_h^{i-1} u - u_{i-1}, \eta_{i-1}) + (R_h^{i-1} u - u_{i-1}, \eta_{i-1} - \eta_i).$$

Here, the last term can be bounded by

$$|(R_h^{i-1} u - u_{i-1}, \eta_{i-1} - \eta_i)| \leq \frac{1}{2}\|\eta_{i-1} - \eta_i\|^2 + ch^4 |u_{i-1}|^2_{H^2(\Omega)}.$$

The first term can be compensated by the previous appearance of the same term but with the opposite sign. In sum we obtain:

$$\begin{aligned}\|\eta_i\|^2 - \|\eta_{i-1}\|^2 &\leq -2(R_h^i u - u_i, \eta_i) + 2(R_h^{i-1} u - u_{i-1}, \eta_{i-1}) \\ &\quad + \frac{1}{2} k_i^2 \int_{t_{i-1}}^{t_i} \|\partial_t \nabla u\|^2 \, dt + ch^4 |u_{i-1}|^2_{H^2(\Omega)}.\end{aligned}$$

Summation over $i = 1, \ldots, m$ and applying the Cauchy-Schwarz inequality and Young's inequality leads to

$$\begin{aligned}\|\eta_m\|^2 &\leq -2(R_h^m u - u_m, \eta_m) + 2(R_h^0 u - u_0, \eta_0) \\ &\quad + \frac{1}{2} \sum_{i=1}^m k_i^2 \|\partial_t u\|^2_{L^2(I_i; H_0^1(\Omega))} + ch^4 \sum_{i=1}^m k_{i-1} |k_{i-1}^{-1/2} u_{i-1}|^2_{H^2(\Omega)} \\ &\leq 2\|\xi_m\|^2 + \frac{1}{2}\|\eta_m\|^2 + 2\|\xi_0\|^2 + \sum_{i=1}^m k_i^2 \|\partial_t u\|^2_{L^2(I_i; H_0^1(\Omega))} + cTh^4 \|k^{-1/2} \nabla^2 u\|^{1/2}_{h,\infty}.\end{aligned}$$

We arrive at

$$\|\eta_m\|^2 \leq 4(\|\xi_m\|^2 + \|\xi_0\|^2) + 2\sum_{i=0}^m k_i^2 \|\partial_t u\|^2_{L^2(I_i; H_0^1(\Omega))} + cTh^4 \|k^{-1/2} \nabla^2 u\|^2_{h,\infty}.$$

This coincides with (12.1) and finalizes the proof. ∎

Theorem 12.2 *Under the same assumptions as the previous Theorem and the additional assumption of constant spatial meshes, i.e. $V_h^m = V_h^{m-1}$, the a priori estimate can be improved as*

$$\|u - u_h\|_{k,\infty} \leq c \max\left(1, T^{1/2}\right) h^2 \|\nabla^2 u\|_{k,\infty} + \left(\sum_{m=1}^M k_m^2 \|\partial_t u\|^2_{L^2(I_m; H_0^1(\Omega))} \right)^{1/2}.$$

Proof. One can avoid the negative power $k^{-1/2}$ in this estimate by taking different weights in Young's inequality to derive a bound on T_2:

$$|(R_h^{i-1} u - u_{i-1}, \eta_{i-1} - \eta_i)| \leq \frac{1}{2} k_i^{-1} \|\eta_{i-1} - \eta_i\|^2 + ck_i h^4 |u_{i-1}|^2_{H^2(\Omega)}.$$

The negative power k_i^{-1} now appears in another term. Therefore, one has to bound the following sum properly:

$$\sum_{i=1}^{m} k_i^{-1}\|\eta_{i-1}-\eta_i\|^2 \;\leq\; c\left(\|\nabla^2 u_0\|^2 + \sum_{i=0}^{m}(k_i+h^2)\|\partial_t u\|^2_{L^2(I_i;H_0^1(\Omega))}\right). \qquad (12.2)$$

To this end, one firstly derive the following bound

$$\|\eta_{i-1}-\eta_i\|^2 + k_i\|\nabla\eta_i\|^2 - k_i\|\nabla\eta_{i-1}\|^2 \;\leq\; (k_i^2+ch^2k_i)\|\partial_t u\|^2_{L^2(I_m;H_0^1(\Omega))},$$

so that multiplication by k_i and summation over i yields (12.2). ∎

12.3 Trapezoidal rule / Crank-Nicholson scheme

An alternative to the backward Euler method, which is only of first order with respect to the time step k, is the trapezoidal rule (in the context of PDEs sometimes also called Crank[1]-Nicholson[2]-scheme). Here the Laplacian at $t=\frac{1}{2}(t_{m-1}+t_m)$ is approximated by the trapezoidal rule:

$$\Delta u \;\approx\; \frac{1}{2}\Delta(u_m+u_{m-1}).$$

The discretization error is here $O(k^2+h^2)$, as long as the spatial discretization is of 2. order. Semi-discretization in time leads to

$$\begin{aligned} k_m^{-1}u_m - \tfrac{1}{2}\Delta u_m &= k_m^{-1}u_{m-1} + \overline{f}_m - \tfrac{1}{2}\Delta u_{m-1} && \text{in } \Omega,\\ u_m &= 0 && \text{on } \partial\Omega. \end{aligned}$$

The corresponding finite element discretization on an mesh $\mathcal{T}_m$ with spaces $V_m\subset V$ we seek for $u_{h,m}\in V_m$ s.t.

$$\begin{aligned} k_m^{-1}(u_{h,m},\phi)_\Omega + \tfrac{1}{2}(\nabla u_{h,m},\nabla\phi)_\Omega &= k_m^{-1}(u_{h,m-1},\phi)_\Omega + (\overline{f}_m,\phi)_\Omega \\ &\quad -\tfrac{1}{2}(\nabla u_{h,m-1},\nabla\phi)_\Omega \qquad \forall\phi\in V_m. \end{aligned}$$

The corresponding linear system with stiffness matrix A and mass matrix M reads

$$\left(k_m^{-1}M+\tfrac{1}{2}A\right)U \;=\; b,$$

where the right hand side b depends on u_{m-1}. Therefore, the numerical costs for the computation of the right hand side is slightly higher than for the backward Euler method. However, the main costs is the assembling of the matrices A and M, and solving the linear

[1] John Crank, 1916-2006, englisch mathematician, working at the Brunel University, Uxbridge, close to London.

[2] Phyllis Nicolson, 1917-1968, englisch mathematician.

system. For this, the numerical costs are the same as for the backward Euler method. But the accuracy of the trapezoidal rule is usually much better, so that this method should be prefered. An disadvantage of the trapezoidal rule is sometimes less stable than backward Euler. In particular, the trapezoidal rule is (only) *A-stable*, but not *strongly A-stable*. This means that high-frequent parts of the solutions do not become damped and therefore remain in the solution for all time. Also local perturbations of the data (i.e. of f or u_0) are much less damped. Therefore, the trapezoidal rule is very sensible to non-smooth initial conditions u_0. In contrast to this, the backward Euler method is much more diffusive in time, leading to a lot of robustness. das implizite Euler-Verfahren aber sehr robust macht.

For the exact definition of the expressions *A-stable* and *strongly A-stable*, we refer to standard literature on time integrators (for ODEs).

12.4 θ-one step methods

The two previously presented time integration schemes can be viewed as particular choices of the more general θ-scheme. Here, $\theta \in [0,1]$ is a fixed parameter. We define

$$\begin{aligned} u_{m+\theta} &:= \theta u_{m+1} + (1-\theta) u_m, \\ f_{m+\theta} &:= \theta f_{m+1} + (1-\theta) f_m. \end{aligned}$$

We denote by A the bilinear form associated to the Laplace operator. The θ-scheme to define $u_{m+1} \in V$ semi-discrete in time now reads

$$k_m^{-1}(u_{m+1} - u_m, \phi)_\Omega + A(u_{m+\theta}, \phi) = (f_{m+\theta}, \phi)_\Omega \qquad \forall \phi \in V. \tag{12.3}$$

The corresponding bilinear form is still V-elliptic:

$$k_m^{-1}(u,u)_\Omega + \theta A(u,u) = k_m^{-1}\|u\|^2_{L^2(\Omega)} + \theta |u|^2_{H^1(\Omega)}.$$

The existence of a solution now follows by the Theorem of Lax-Milgram. The step to a spatial-discrete formulation is as before. The linear algebraical system become due to the linearity of A as:

$$(k_m^{-1} M + \theta A) U_m = (k_m^{-1} M + (\theta - 1) A) U_{m-1}.$$

We obtain the particular methods:

- $\theta = 1$: backward Euler method,
- $\theta = 1/2$: trapezoidal rule,
- $\theta = 0$: forward Euler method.

The following theorem now gives a relatively strong stability property of the semi-discrete problem. The stability is independent of the size of the time step k_m. We call this property *unconditionally stable*.

Theorem 12.3 *For the heat equation and* $\frac{1}{2} \leq \theta \leq 1$ *the one-step-*θ *method is* unconditionally stable*:*

$$\begin{aligned} & \|u_m\|^2_{L^2(\Omega)} + \sum_{l=0}^{M-1} \left(k_l |u^2_{l+\theta}|_{H^1(\Omega)} + (2\theta - 1)\|u_l - u_{l-1}\|^2_{L^2(\Omega)} \right) \\ \leq \quad & \|u_0\|^2_{L^2(\Omega)} + c \sum_{l=0}^{M-1} k_l \|f_{l+\theta}\|^2_{L^2(\Omega)}. \end{aligned}$$

Proof. We chose in (12.3) $\phi := u_{m+\theta}$

$$k_m^{-1}(u_{m+1} - u_m, u_{m+\theta})_\Omega + A(u_{m+\theta}, u_{m+\theta}) \quad = \quad (f_{m+\theta}, u_{m+\theta})_\Omega.$$

Moreover, we use that

$$\begin{aligned} (u_{m+1} - u_m, u_{m+\theta})_\Omega &= (u_{m+1} - u_m, \theta u_{m+1} + (1-\theta)u_m)_\Omega \\ &= (u_{m+1} - u_m, \tfrac{1}{2}u_{m+1} + \tfrac{1}{2}u_m + (\theta - \tfrac{1}{2})(u_{m+1} - u_m))_\Omega \\ &= \tfrac{1}{2}\|u_{m+1}\|^2 - \tfrac{1}{2}\|u_m\|^2 + (\theta - \tfrac{1}{2})\|u_{m+1} - u_m\|^2. \end{aligned}$$

Application of the Cauchy-Schwarz inequality, the Poincaré inequality and Young's inequality yields to

$$\begin{aligned} & \|u_{m+1}\|^2 - \|u_m\|^2 + (2\theta - 1)\|u_{m+1} - u_m\|^2 + 2k_m |u_{m+\theta}|^2_{H^1(\Omega)} \\ = \quad & 2k_m (f_{m+\theta}, u_{m+\theta})_\Omega \\ \leq \quad & 2k_m c_\Omega \|f_{m+\theta}\| |u_{m+\theta}|_{H^1(\Omega)} \\ \leq \quad & k_m c_\Omega^2 \|f_{m+\theta}\|^2 + k_m |u_{m+\theta}|^2_{H^1(\Omega)}. \end{aligned}$$

This implies

$$\|u_{m+1}\|^2 - \|u_m\|^2 + (2\theta - 1)\|u_{m+1} - u_m\|^2 + k_m |u_{m+\theta}|^2_{H^1(\Omega)} \quad \leq \quad k_m c_\Omega^2 \|f_{m+\theta}\|^2.$$

Summation over m finalizes the proof. ∎

12.5 Discontinuous Galerkin in time (dG)

Now we formulate the differential equation also in time in a variational form. Afterwards, we discretize it and obtain a discrete variational formulation. To this end, we chose semi-open sub-intervals $I_m = [t_{m-1}, t_m)$ and

$$V_k(I) \quad := \quad \{v \in L^2(I, V) \,:\, v|_{I_m} \in C^1(I_m, V)\,, 1 \leq m \leq M\}.$$

These functions are not neccessarily continuous at the end points of the sub-intervals. Therefore, we introduce the following notations for $v \in V(I)$:

$$v_m^- \;=\; \lim_{t \nearrow t_m} v(x)\,, \qquad v_m^+ \;=\; \lim_{t \searrow t_m} v(x)\,, \qquad [v]_m \;=\; v_m^+ - v_m^-.$$

In the variational formulation we seek $u \in V_k(I)$, s.t. the initial condition $u_0^- = u_0$ holds and for all $v \in V_k(I)$ it must hold

$$\sum_{m=1}^{M} \left\{ \int_{I_m} [(\partial_t u, v)_{L^2(V)} + A(u,v)]\, dt + ([u]_{m-1}, v_{m-1}^+)_{L^2(V)} \right\} \;=\; (f,v)_{L^2(I,V)}.$$

For a time-discrete formulation we now chose approximative finite-dimensional subspaces

$$V_{r,k}(I) \;:=\; \{v \in V_k(I) \,:\, v|_{I_m} \in P_r(I_m, V)\,, 1 \le m \le M\}.$$

The discontinuous Galerkin method $dG(r)$ of order r now consists in seeking $u_k \in V_{r,k}(I)$ s.t.

$$\sum_{m=1}^{M} \left\{ \int_{I_m} [(\partial_t u_k, v)_{L^2(V)} + A(u_k,v)]\, dt + ([u_k]_{m-1}, v_{m-1}^+)_{L^2(V)} \right\} \;=\; (f,v)_{L^2(I,V)}, \tag{12.4}$$

for all $v \in V_{r,k}(I)$. For the exact solution, the jump terms vanish, $[u]_{m-1} = 0$. The derivative $\partial_t u$ is in $L^2(I,V)$, so that we can express the sum as

$$\int_I [(\partial_t u, v)_{L^2(V)} + A(u,v)]\, dt \;=\; (f,v)_{L^2(I,V)}.$$

This is exactly the weak formulation of the heat equation. Therefore, the $dG(r)$-method is *consistent* in the sense that the exact solution $u \in H^1(I,V)$ still fulfills the discrete equation.

12.6 dG(0)

In the case of polynomial order $k = 0$ with respect to time, the test- and ansatz functions are constant in time, $\partial_t u_m|_{I_m} = 0$ and $u_m := u_{m-1}^+ = u_m^-$. The system (12.4) then reduces to

$$(u_m, v)_{L^2(\Omega)} + k_m A(u_m, v)?\text{`} \;=\; (u_{m-1}, v)_{L^2(\Omega)} + \int_{I_m} (f,v)_{L^2(\Omega)} \qquad \forall v \in P_r.$$

For numerical quadrature of the integral on the right hand side by the box rule, $\int_{I_m} f\, dt \approx k_m f(t_m)$ we recover exactly the backward Euler method.

Chapter 13

Convection-diffusion-reaction equations

In this section we investigate convection-diffusion-reactions equations for a scalar quantity $u : \Omega \to \mathbb{R}$:

$$\begin{aligned} -\epsilon \Delta u + (b \cdot \nabla) u + cu &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{auf } \partial\Omega. \end{aligned}$$

Here, $\epsilon > 0$ is a diffusion parameter (often a very small parameter). $c = c(x)$ is a real-valued function, describing a reaction, and $b = b(x)$ a vector-valued function (a vector field) $b : \Omega \to \mathbb{R}^d$. The *convection term* $(b \cdot \nabla)u$ has to be understood as

$$(b \cdot \nabla)u := \sum_{i=1}^{d} b_i(x) \frac{\partial u(x)}{\partial x_i}.$$

Such kind of equation appear very often in such form or a similar form. We will see in the following that for small diffusion parameter $\epsilon > 0$ non-trivial numerical problems arise. We will start with the one-dimensional case, $d = 1$.

13.1 Convection-diffusion equation in 1D

We firstly consider the upper mentioned convection-diffusion-reaction equation for the particular case $d = 1$, $b = 1$, $c = 0$ and $f = 1$. The computational domain Ω is the unit interval $I := (0, 1)$:

$$\begin{aligned} -\epsilon u'' + u' &= 1 \qquad \text{für } x \in I, \\ u(0) &= u(1) = 0. \end{aligned} \tag{13.1}$$

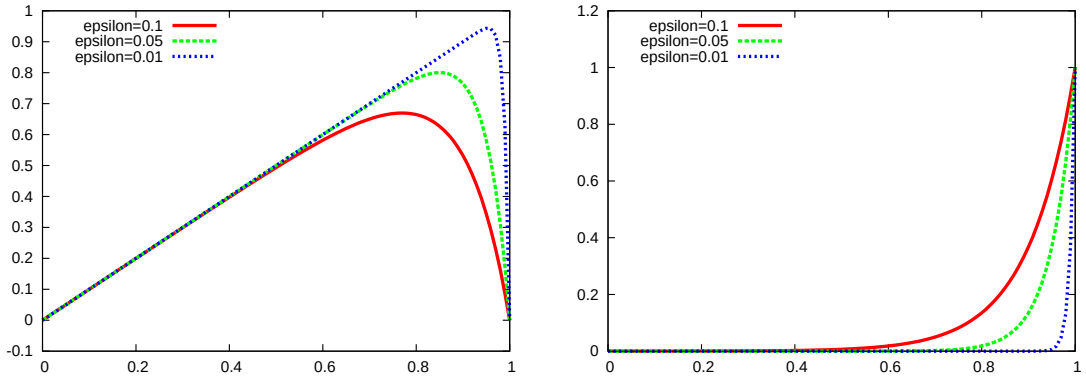


Figure 13.1: Solutions of the one-dimensional convections-diffusions equations $-\epsilon u'' + u' = 1$ (left) and $-\epsilon u'' + u' = 0$ (right) with corresponding Dirichlet boundary conditions for $\epsilon \in \{0.1, 0.05, 0.01\}$.

It is easy to check that the exact solution is given by

$$u(x) \quad = \quad x - \frac{\delta}{1-\delta}(\exp(x/\epsilon) - 1),$$

with $\delta := \exp(-1/\epsilon)$. In Fig. 13.1 we show this solution for three different values of ϵ. In the case of small ϵ, u is obviously in a wide area very close to the identity $x \mapsto x$. However, for x close to 1 strong gradients appear. This phenomena is called *boundary layer*. In the present case, the boundary layer is of exponential type. For $x \in [0,1)$ it holds $\lim_{\epsilon \to 0} u(x) = x$, and therefore for $a \in [0,1)$

$$\lim_{x \to a} \lim_{\epsilon \to 0} u(x) \quad = \quad a \quad = \quad \lim_{\epsilon \to 0} \lim_{x \to a} u(x).$$

But for $a = 1$:

$$1 \quad = \quad \lim_{x \to 1} \lim_{\epsilon \to 0} u(x) \quad \neq \quad \lim_{\epsilon \to 0} \lim_{x \to 1} u(x) \quad = \quad 0.$$

Hence, the function has a qualitative different behaviour in the point $x = 1$. This phenomena is the reason why the problem (13.1) is called *singularly perturbed*. Later we will need the behavior of the derivatives of u in dependence of ϵ. Therefore, we calculate here:

$$\begin{aligned}
\|u\|_{L^\infty(I)} &\leq 1, \\
\|u'\|_{L^\infty(I)} &= \sup_{x \in (0,1)} \left| 1 - \frac{\delta}{\epsilon(1-\delta)} \exp(x/\epsilon) \right| = \frac{\delta \exp(1/\epsilon)}{\epsilon(1-\delta)} - 1 = \frac{1}{\epsilon(1-\delta)} - 1 \leq \frac{2}{\epsilon}, \\
\|u''\|_{L^\infty(I)} &\leq \frac{2}{\epsilon^2}.
\end{aligned}$$

For the derivatives of higher order $k \geq 2$ we obtain the following asymptotic upper bound:

$$\|u^{(k)}\|_{L^\infty(I)} \quad = \quad \frac{\delta}{\epsilon^k(1-\delta)} \sup_{x \in (0,1)} e^{x/\epsilon} \quad = \quad \frac{1}{\epsilon^k(1-\delta)} \quad \leq \quad 2\epsilon^{-k}.$$

13.1.1 Central difference quotients

We will firstly discretize the model problem (13.1) with finite differences. For simplicity, we use an equidistant mesh of mesh width $h = 1/N$ and nodes $x_i = ih$. The nodal values $u(x_i)$ will be denoted by u_i.

For the approximation of the first derivatives $u'(x_i)$ we apply central difference quotients:

$$u'(x_i) \approx \frac{1}{2h}(u_{i+1} - u_{i-1}).$$

For the second derivatives we use central difference quotients of 2. order:

$$u''(x_i) \approx \frac{1}{h^2}(u_{i+1} - 2u_i + u_{i-1}).$$

The approximative system now reads, $u_0 = u_N = 0$ and

$$\begin{aligned}
\frac{2\epsilon}{h^2}u_1 + \left(-\frac{\epsilon}{h^2} + \frac{1}{2h}\right)u_2 &= 1, \\
\left(-\frac{\epsilon}{h^2} - \frac{1}{2h}\right)u_{i-1} + \frac{2\epsilon}{h^2}u_i + \left(-\frac{\epsilon}{h^2} + \frac{1}{2h}\right)u_{i+1} &= 1\,, \qquad i \in \{2, \ldots, N-2\}, \\
\left(-\frac{\epsilon}{h^2} - \frac{1}{2h}\right)u_{N-2} + \frac{2\epsilon}{h^2}u_{N-1} &= 1.
\end{aligned}$$

This approximation is (formaly) of second order in h. The following Lemma shows the conditional stability of this scheme:

Lemma 13.1 *The finite difference scheme of (13.1) with central difference quotients leads to a M-matrix, if $h \leq 2\epsilon$.*

Proof. (a) Independent of the mesh size, we will see that A is generalized diagonal dominant: For the row sum in row number i, $2 \leq i \leq N-1$, we get

$$\sum_{j=1}^{n} a_{ij} = -\frac{\epsilon}{h^2} - \frac{1}{2h} + \frac{2\epsilon}{h^2} - \frac{\epsilon}{h^2} + \frac{1}{2h} = 0.$$

For the first and last row ($i \in \{1, N-1\}$) it holds

$$\sum_{j=1}^{n} a_{ij} = \frac{2\epsilon}{h^2} - \frac{\epsilon}{h^2} + \frac{1}{2h} \geq \frac{\epsilon}{h^2} + \frac{1}{2h} > 0.$$

Hence, A is generalized diagonal dominant.
(b) A is irreducible, because the rows $i-1$, i and $i+1$ are mutually coupled due to

non-vanishing entries in the corresponding columns. (c) The mesh size restriction $h \leq 2\epsilon$ is needed to ensure that A is of non-negative type: For $1 \leq i, j \leq N$ and $j \neq i$:

$$\begin{aligned} a_{ii} &= \frac{2\epsilon}{h^2} > 0, \\ a_{ij} &= -\frac{\epsilon}{h^2} \pm \frac{1}{2h} \leq -\frac{\epsilon}{h^2} + \frac{1}{2h} \leq 0. \end{aligned}$$

The M-matrix property now follows by Lemma 7.9. □

The condition $h \leq 2\epsilon$ in the previous lemma is essential, as we will see in the following example. Unfortunately, this restriction is often violated in real applications due to computational resources.

Example. We consider the following example with known exact solution. The Dirichlet conditions are inhomogeneous, but the right hand side vanishes:

$$\begin{aligned} -\epsilon u'' + u' &= 0 \qquad x \in I, \\ u(0) = 0\,, \quad u(1) &= 1. \end{aligned}$$

The exact solution is given by

$$u(x) = (\exp(x/\epsilon) - 1)\frac{\delta}{1-\delta},$$

with $\delta = \exp(-1/\epsilon)$. The corresponding linear system obtained by central differences reads

$$r_i u_{i-1} + s_i u_i + t_i u_{i+1} = 0\,, \qquad i \in \{1, \ldots, N-1\}, \tag{13.2}$$

with

$$r_i = -\frac{\epsilon}{h^2} - \frac{1}{2h}, \quad s_i = \frac{2\epsilon}{h^2}, \quad t_i = -\frac{\epsilon}{h^2} + \frac{1}{2h}.$$

For small $\epsilon \ll 2h$, this linear system is of the form

$$\frac{1}{2h}(u_{i+1} - u_{i-1}) = \mathcal{O}(\epsilon)\,, \qquad i \in \{1, \ldots, N-1\}.$$

Due to the boundary conditions $u_0 = 0$ and $u_N = 1$, we get $u_4 \approx u_2 \approx u_0 = 0$ and $u_{N-4} \approx u_{N-2} \approx u_N = 1$. Hence, we expect for odd N strong oscillations. That this is indeed the case we have a look onto the exact discrete solution

$$u_i = \frac{q_h^i - 1}{q_h^N - 1}, \tag{13.3}$$

with $q_h := (2\epsilon + h)/(2\epsilon - h)$. We show this solution in Figure 13.2 for different mesh sizes h. In the case of $h \gg 2\epsilon$ we have $q_h \approx -1$, leading to strong oscillations of the values u_i.

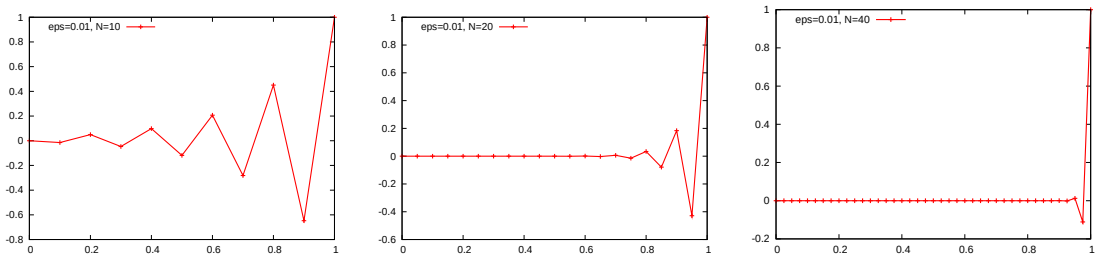


Figure 13.2: Discrete solutions of the one-dimensional convection-diffusion equation $-\epsilon u'' + u' = 0$ (right) discretized by central difference quotients, $\epsilon = 0.01$ and $N = 10$ (left), $N = 20$ (middle) and $N = 40$ (right).

13.1.2 Upwinding

In the so-called upwinding method the first order derivatives are not approximated by central difference quotients, but by one-sided difference quotients. In the case of left-sided difference quotients we use

$$u'(x_i) \approx \frac{1}{h}(u_i - u_{i-1}).$$

The resulting linear system is once more of the form (13.2), but with entries

$$r_i = -\frac{\epsilon}{h^2} - \frac{1}{h}, \quad s_i = \frac{2\epsilon}{h^2} + \frac{1}{h}, \quad t_i = -\frac{\epsilon}{h^2}.$$

The diagonal is always positive, and all non-vanishing off-diagonal entries are negative, independently of the ratio of ϵ and h. Moreover, the matrix is diagonally dominant:

Lemma 13.2 *The finite difference discretization of (13.1) with upwinding always leads to a M-matrix.*

The solution u_i is obtained by (13.3), but now with $q_h := 1 + h/\epsilon$, see Figure 13.3. This solution is obviously much better than the one obtained by central differences, although the one-sided difference quotient is only of first order.

In the general case of a linear equation of the form

$$-\epsilon u'' + bu' + cu = f \qquad x \in I,$$

the question whether we choose the right or left difference quotient, is dependent of the sign of the coefficient $b(x_i)$:

$$u'(x_i) \approx \begin{cases} h^{-1}(u_i - u_{i-1}), & \text{if } b(x_i) > 0, \\ h^{-1}(u_{i+1} - u_i), & \text{if } b(x_i) < 0. \end{cases}$$

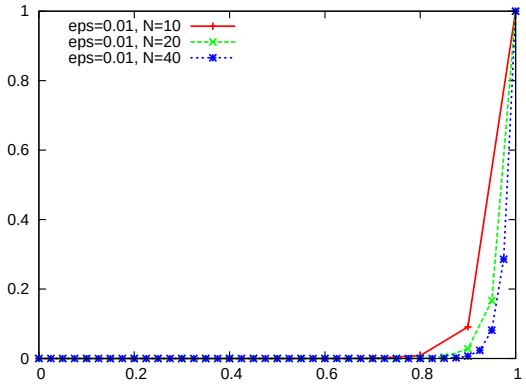


Figure 13.3: Discrete solution of the one-dimensional convection-diffusion equation $-\epsilon u'' + u' = 0$ (right) with upwinding, $\epsilon = 0.01$ and $N = 10$ (red), $N = 20$ (green) and $N = 40$ (blue).

This procedure is called *upwinding*, because the difference quotient is always oriented towards the opposite direction of the flow field b. The method is only of first order. In more than one dimension, the direction of the convection field must be appropriately determined on the mesh. Here, we explicitly mention the *finite volume* techniques without going into detail.

Theorem 13.3 *The finite difference scheme with upwinding is inverse monotone.*

Proof. The difference scheme is of the form

$$\begin{pmatrix} I & 0 \\ B & A \end{pmatrix} \begin{pmatrix} u_{\partial\Omega} \\ u \end{pmatrix} = \begin{pmatrix} u_{h,0} \\ f \end{pmatrix}.$$

Here, $u_{\partial\Omega}$ denote the degrees of freedom at the boundary (those should be exactly $u_{h,0}$) and u are the degrees of freedom in the interior of Ω. For non-negative boundary data $u_{h,0}$ and non-negative forcing f we deduce by the M-matrix property:

$$\begin{pmatrix} u_{\partial\Omega} \\ u \end{pmatrix} = \begin{pmatrix} I & 0 \\ B & A \end{pmatrix}^{-1} \begin{pmatrix} u_{h,0} \\ f \end{pmatrix} = \begin{pmatrix} u_{h,0} \\ A^{-1}(f - Bu_{h,0}) \end{pmatrix} \geq 0.$$

□

13.1.3 Artificial diffusion

An alternative method to obtain the M-matrix structure is the *artificial diffusion*. The idea is to make the diffusion coefficient artificially larger, if needed, but maintain the central difference quotient. Instead of ϵ the used diffusion coefficient is given by

$$\epsilon_h := \max(\epsilon, h/2).$$

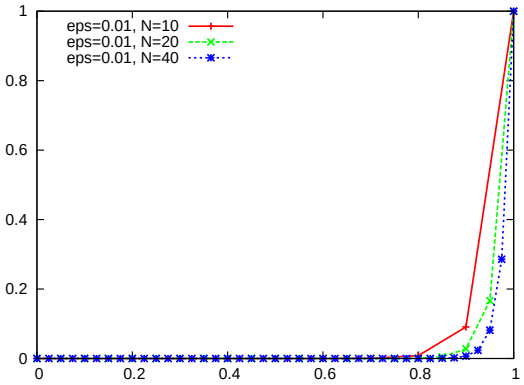


Figure 13.4: Discrete solution of the one-dimensional convection-diffusion equation $-\epsilon u'' + u' = 0$ (right) with artificial diffusion, $\epsilon = 0.01$ and $N = 10$ (red), $N = 20$ (green) and $N = 40$ (blue).

The result is depicted in Fig. 13.4.

13.2 Convection-diffusion-reaction equation in several dimensions

Now we come back to the case of several spatial dimensions:

$$-\epsilon\Delta u + (b\cdot\nabla)u + cu = f \qquad \text{in } \Omega, \tag{13.4}$$

$$u = 0 \qquad \text{auf } \partial\Omega. \tag{13.5}$$

A classical solution is a function $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfying this equation point-wise.

The variational formulation reads for homogeneous Dirichlet data in $V := H_0^1(\Omega)$:

$$u \in V: \quad A(u,\phi) = (f,\phi) \qquad \forall \phi \in V, \tag{13.6}$$

with the bilinear form $A : V \times V \to \mathbb{R}$

$$A(u,\phi) := \epsilon(\nabla u, \nabla\phi) + ((b\cdot\nabla)u,\phi) + (cu,\phi). \tag{13.7}$$

The regularity requirements for the coefficients b and c needed here are only to be bounded functions, i.e. $b \in L^\infty(\Omega)^d$ and $c \in L^\infty(\Omega)$. At first we focus on the convective term by diagonal testing. For divergence-free b vamishes $((b\cdot\nabla)u,u)$:

Lemma 13.4 *For $u \in H_0^1(\Omega)$ and $b \in H^1(\Omega)^d$ it holds*

$$((b\cdot\nabla)u,u) = -\frac{1}{2}((div\ b)u,u).$$

In particular, we obtain for divergence-free b: $((b\cdot\nabla)u,u) = 0$.

Proof. By integration by parts we obtain due to vanishing boundary integrals:

$$((b\cdot\nabla)u,u) \;=\; (\nabla u, bu) \;=\; -(u,\mathrm{div}\,(bu)) \;=\; -(u,(\mathrm{div}\ b)u) - (u,(b\cdot\nabla)u).$$

This implies

$$((b\cdot\nabla)u,u) \;=\; -\frac{1}{2}(u,(\mathrm{div}\ b)u)).$$

□

13.2.1 Existence theory by Lax-Milgram

In the following we use the Banach space

$$W^{1,\infty}(\Omega) \;:=\; \{v\in L^\infty(\Omega)\,:\,\nabla v\in L^\infty(\Omega)\}.$$

Theorem 13.5 *Assuming $c\in L^\infty(\Omega)$, $b\in W^{1,\infty}(\Omega)^d$, $\epsilon>0$ and*

$$\inf_\Omega \mathrm{ess}(c-\tfrac{1}{2}div\ b) \;\geq\; 0. \tag{13.8}$$

Then there exists a unique weak solution $u\in H_0^1(\Omega)$ of (13.6).

The essential infimum for $f:\Omega\to\mathbb{R}$ with respect to a Lebesgue measure μ is defined as

$$\inf_\Omega \mathrm{ess} f \;:=\; \sup\{m\in\mathbb{R}\,:\,\mu(x\in\Omega : f(x)<m)=0\}.$$

Proof. We will apply the Theorem of Lax-Milgram 3.22. The continuity of the associated bilinear form (13.7) results with a constant $C=C(\epsilon,b,c)$ and the Poincaré constant c_Ω by

$$\begin{aligned} |A(u,\phi)| &\leq \epsilon\|\nabla u\|\|\nabla\phi\| + \|b\|_\infty\|\nabla u\|\|\phi\| + \|c\|_\infty\|u\|\|\phi\| \\ &\leq C\|u\|_{H^1(\Omega)}\|\phi\|_{H^1(\Omega)} \\ &\leq Cc_\Omega|u|_{H^1(\Omega)}|\phi|_{H^1(\Omega)}. \end{aligned}$$

The V-ellipticity results from the previous lemma:

$$\begin{aligned} A(u,u) &= \epsilon(\nabla u,\nabla u) + ((b\cdot\nabla)u,u) + (cu,u) \\ &= \epsilon\|\nabla u\|^2 - \frac{1}{2}((\mathrm{div}\ b)u,u) + (cu,u) \\ &= \epsilon\|\nabla u\|^2 + ((c-\tfrac{1}{2}(\mathrm{div}\ b))u,u) \\ &\geq \epsilon\|\nabla u\|^2 + d\|u\|^2, \end{aligned}$$

where $d:=\inf_{x\in\Omega}\mathrm{ess}\left(c(x)-\frac{1}{2}\mathrm{div}\ b(x)\right)^{1/2}$. Due to the assumption this root is real, hence $d\geq 0$. The ellipticity/coercivity and continuity of the bilinear form now ensures existence and uniqueness of the solution by Lax-Milgram (Theorem 3.22). □

13.2.2 Weak maximum principle

Theorem 13.6 (Weak Maximum Principle) *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ a domain, $f \in C(\overline{\Omega})$, $b \in C(\overline{\Omega}; \mathbb{R}^d)$, $c \in C(\overline{\Omega})$ and $u \in C^2(\Omega) \cap C(\overline{\Omega})$ a solution of the equation*

$$Lu \;:=\; -\Delta u + (b \cdot \nabla)u + cu \;=\; f \quad \text{in } \Omega.$$

Then holds:

(i) If $c \geq 0$, $f \leq 0$ and $\max_{x\in\Omega} u(x) \geq 0$, then:

$$\max_{x\in\overline{\Omega}} u(x) \;=\; \max_{x\in\overline{\partial}\Omega} u(x)$$

(ii) If $c = 0$, $f \leq 0$, then

$$\max_{x\in\overline{\Omega}} u(x) \;=\; \max_{x\in\overline{\partial}\Omega} u(x)$$

(iii) If $c = 0$, $u|_{\partial\Omega} = 0$, then

$$\|u\|_{L^\infty(\Omega)} \;\leq\; C\|f\|_{L^\infty(\Omega)}$$

Proof. (i) Let $x_0 \in \overline{\Omega}$ s.t. $u(x_0) = \max_{x\in\overline{\Omega}} u(x)$. We firstly assume $f < 0$. We proof the assertion by contradiction assuming that

$$u(x_0) \;>\; m := \max_{x\in\overline{\partial}\Omega} u(x).$$

Due to the maximality of u in the inner point x_0 we know that $\nabla u(x_0) = 0$, $\partial^2_{x_i} u(x_0) \leq 0$ for all directions i, and $c(x_0)u(x_0) \geq c(x_0)m \geq 0$. This implies the contradiction

$$\begin{aligned} 0 \;&>\; f(x_0) = -\Delta u(x_0) + (b \cdot \nabla)u(x_0) + c(x_0)u(x_0) \\ &\geq\; 0 + 0 + 0 = 0 \end{aligned}$$

Now let $f \leq 0$. We consider for $\epsilon, \mu > 0$ the function

$$v_\epsilon(x) \;:=\; u(x) + \epsilon e^{\mu x_1}.$$

For this v_ϵ it holds for μ sufficiently large:

$$\begin{aligned} Lv_\epsilon(x) \;&=\; f(x) + \epsilon L(e^{\mu x_1}) \\ &\leq\; \epsilon L(e^{\mu x_1}) \\ &=\; \underbrace{\epsilon e^{\mu x_1}}_{>0}(-\mu^2 + \mu \underbrace{b_1(x)}_{\leq C} + \underbrace{c(x)}_{\leq C}) \;<\; 0. \end{aligned}$$

According to the maximum principle shown above for $f < 0$ we know that v_ϵ has his maximum at the boundary (for μ sufficiently large). Now we assume $x_0 \in \Omega$. Then there exists $\delta > 0$ s.t

$$\max_{x\in\partial\Omega} v_\epsilon(x) \;\geq\; v_\epsilon(x_0) \;>\; u(x_0) \;\geq\; \max_{x\in\partial\Omega} u(x) + \delta.$$

On the other hand we have

$$\max_{x\in\partial\Omega} v_\epsilon(x) \;\leq\; \max_{x\in\partial\Omega} u(x) + \epsilon C_\mu$$

However, this implies $\delta \leq C_\mu \epsilon$, which is not possible for ϵ sufficiently small.
(ii) The proof follows the lines of case (i) with the only difference that no sign-condition is needed, because the term $c(x)u(x)$ does not arise.
(iii) We can assume without loss of generality that $\Omega \subseteq [0,\infty[\times\mathbb{R}^{d-1}$. Let us consider the function

$$v(x) \;:=\; -\lambda e^{-\mu x_1} \;<\; 0,$$

with parameters $\lambda, \mu \geq 0$ to be determined. It holds $\|v\|_{L^\infty} \leq \lambda$. For μ sufficiently large we have

$$Lv(x) \;=\; -\mu^2 v(x) + b_1 \mu v(x) \;\geq\; \frac{1}{2}\mu^2 |v(x)| \;\geq\; 0.$$

Furthermore, we have for $\lambda > 0$ sufficiently large (e.g. $\lambda = 2\mu^{-2}\|f\|_{L^\infty}$):

$$L(u-v)(x) \;=\; f(x) - Lv(x) \;\leq\; f(x) - \frac{1}{2}\mu^2 |v(x)| \;\leq\; 0.$$

By part (ii) we deduce

$$\lambda \;\geq\; \lambda \max_{x\in\partial\Omega} e^{-\mu x_1} \;=\; \max_{x\in\partial\Omega}(-v)(x) \;=\; \max_{x\in\partial\Omega}(u-v)(x) \;=\; \max_{x\in\overline{\Omega}}(u-v)(x)$$

This implies

$$\max_{x\in\overline{\Omega}} u(x) \;\leq\; \max_{x\in\overline{\Omega}} v(x) + \max_{x\in\overline{\Omega}}(u-v)(x) \;\leq\; \lambda + \lambda \;=\; 2\lambda.$$

One shows analogously $\max_{x\in\overline{\Omega}}(\tilde{v}-u)(x) \leq \lambda$ for $\tilde{v}(x) := \lambda e^{-\mu x_1}$, and hence

$$\max_{x\in\overline{\Omega}}(-u(x)) \;\leq\; \max_{x\in\overline{\Omega}}(-\tilde{v}(x)) + \max_{x\in\overline{\Omega}}(\tilde{v}-u)(x) \;\leq\; 0 + \lambda \;=\; \lambda.$$

This implies $\|u\|_{L^\infty} \leq 2\lambda \leq C\|f\|_{L^\infty}$. ∎

13.2.3 Existence theory for regular domains

Theorem 13.7 *Let $k \in \mathbb{N}_0$, $f \in H^k(\Omega)$ and $\Omega \subset \mathbb{R}^d$ a domain with C^{k+2} boundary or a cuboid. Then for every weak solution $u \in H_0^1(\Omega)$ of $-\Delta u = f$ holds $u \in H^{k+2}(\Omega)$ and $\|u\|_{H^{k+2}(\Omega)} \leq C\|f\|_{H^k(\Omega)}$.*

Lemma 13.8 *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ a domain with C^4-boundary or a cuboid, $b \in C^2(\overline{\Omega};\mathbb{R}^d)$ and $f \in H^2(\Omega) \cap L^\infty(\Omega)$. Then any solution $u \in H^4(\Omega) \cap H_0^1(\Omega)$ of the equation*

$$Lu \;:=\; -\Delta u + (b\cdot\nabla)u \;=\; f$$

satisfies $\|u\|_{H^4(\Omega)} \leq C\|f\|_{H^2(\Omega)} + C_b\|f\|_{L^\infty(\Omega)})$ with a constant C independent of f and b, and a constant $C_b \leq C\|\mathrm{div}\, b\|_{L^\infty}$. In particular, the solution (if existent) is unique.

Proof. (a) By diagonal testing of the equation we obtain

$$\begin{aligned}\|\nabla u\|^2 &= -\frac{1}{2}((\mathrm{div}\, b)u, u) + (f,u)\\ &\leq C_b\|u\|^2 + \|f\|_{L^\infty}\|u\|\end{aligned}$$

By the discrete maximum principle we know $\|u\| \leq C\|f\|_{L^\infty}$, so that $\|\nabla u\| \leq C_b\|f\|_{L^\infty}$. By Poincare we obtain $\|u\|_{H^1} \leq C_b\|f\|_{L^\infty}$.
(b) This u solves $-\Delta u = \tilde{f} := f - (b\cdot\nabla)u$. By (a) we have $\tilde{f} \in L^2(\Omega)$ with $\|\tilde{f}\| \leq \|f\| + C_b\|f\|_{L^\infty} \leq C_b\|f\|_{L^\infty}$. By the regularity property of the Laplace operator (Thm. 13.7) we obtain $u \in H^2(\Omega)$ and $\|u\|_{H^2} \leq C\|\tilde{f}\| \leq C_b\|f\|_{L^\infty}$.
(c) We apply the regularity trick another time: Due to (b) we have $\tilde{f} \in H^1(\Omega)$ and hence $u \in H^3(\Omega)$ and $\|u\|_{H^3} \leq C\|\tilde{f}\|_{H^1} \leq C\|f\|_{H^1} + C_b\|u\|_{H^1} \leq C\|f\|_{H^1} + C_b\|f\|_{L^\infty}$.
(d) Making this regularity trick a last time yields $\tilde{f} \in H^2(\Omega)$ and hence $u \in H^3(\Omega)$ and $\|u\|_{H^4} \leq C\|\tilde{f}\|_{H^2} \leq C\|f\|_{H^2} + C_b\|u\|_{H^2} \leq C\|f\|_{H^2} + C_b\|f\|_{L^\infty}$. ∎

Theorem 13.9 *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ a domain with C^4-boundary or a cuboid, $b \in C^2(\overline{\Omega};\mathbb{R}^d)$ and $f \in H^2(\Omega) \cap L^\infty(\Omega)$. Then the equation*

$$Lu \;:=\; -\Delta u + (b\cdot\nabla)u \;=\; f$$

has a weak solution $u \in H^4(\Omega) \cap H_0^1(\Omega)$.

Remark: Due to the Sobolev embeddding

$$H^l(\Omega) \subset C^m(\Omega) \quad \text{for } l \geq m + \frac{d}{2} \text{ and } m > 0$$

we have $H^4(\Omega) \subset C^2(\Omega)$. Due to $u|_{\partial\Omega} = 0$ we have $u \in C^2(\Omega) \cap C(\overline{\Omega})$. this solution is also a classical solution.

Proof. We use the notations $F := H^2(\Omega) \cap L^\infty(\Omega)$ and $H := H^4(\Omega) \cap H_0^1(\Omega)$. The proof uses the so-called continuity method. For $\lambda \in [0,1]$ we consider the operator $L_\lambda := -\Delta + \lambda(b \cdot \nabla)$, and for $f \in F$ the equation

$$u \in H \text{ s.t} \qquad L_\lambda u = f. \tag{13.9}$$

and the set

$$\Lambda \;:=\; \left\{ \lambda \in [0,1] \;\middle|\; \forall f \in F \text{ equation (13.9) is in } H \text{ solvable} \right\}.$$

We know the existence of a solution u of the original equation iff $1 \in \Lambda$. Now, we are going to show:

(a) $\Lambda \neq \emptyset$,

(b) Λ is closed,

(c) Λ is open in $[0,1]$.

These three points imply that $\Lambda = [0,1]$, and in particular $1 \in \Lambda$.
(a): For $\lambda = 0$ we obtain the Laplace operator $L_\lambda = -\Delta$. The previous Theorem 13.7 ensures existence of a solution $u \in H$.
(b): Let $\lambda_k \in \Lambda$ for all $k \in \mathbb{N}$ and $\lambda := \lim \lambda_k \in [0,1]$. We have to verify $\lambda \in \Lambda$. Let $f \in F$ and $u_k \in H$ the solution for $L_{\lambda_k} u_k = f$. By Lemma 13.8 we have

$$\|u_k\|_{H^4(\Omega)} \;\leq\; C(\|f\|_{H^2(\Omega)} + \|f\|_{L^\infty(\Omega)}),$$

with a constant C independent of f and k (but possibly depending on $\|\operatorname{div} b\|_{L^\infty}$). Since H^4 is a reflexive space, there exists a subsequence $u_k \rightharpoonup u_\lambda \in H^4(\Omega)$. Due to the compact embedding $H^4(\Omega) \subset\subset C^2(\Omega)$ we conclude $u_k \to u_\lambda \in C^2(\Omega)$. This implies pointwise convergence (a.e.)

$$-\Delta u_k \;\to\; -\Delta u_\lambda \quad \text{and} \quad (b \cdot \nabla) u_k \to (b \cdot \nabla) u_\lambda$$

Taking into mind that $\lambda_k \to \lambda$, we conclude pointwise convergence (a.e.)

$$f \;=\; -\Delta u_k + \lambda_k (b \cdot \nabla) u_k \;\to\; -\Delta u_\lambda + \lambda (b \cdot \nabla) u_\lambda.$$

Hence, $u_\lambda \in C^2(\Omega)$ is a classical solution of $L_\lambda u_\lambda = f$. By the trace theorem we further obtain

$$\begin{aligned} \|u_\lambda\|_{L^2(\partial\Omega)} &\leq \underbrace{\|u_k\|_{L^2(\partial\Omega)}}_{=0} + \|u_\lambda - u_k\|_{L^2(\partial\Omega)} \\ &\leq c_\Omega \|u_\lambda - u_k\|_{H^1(\Omega)} \to 0. \end{aligned}$$

This shows $\lambda \in \Lambda$.
(c): Let $\lambda_0 \in \Lambda \cap]0,1[$. We seek for $\epsilon > 0$ s.t. $]\lambda_0 - \epsilon, \lambda + \epsilon[\subseteq \Lambda$. For $\lambda := \lambda_0 + \epsilon$ the equation $L_\lambda u = f$ is equivalent to

$$L_{\lambda_0} u \quad = \quad f - \epsilon(b \cdot \nabla)u. \tag{13.10}$$

Because $\lambda_0 \in \Lambda$ the equation

$$L_{\lambda_0} u_v \quad = \quad f - \epsilon(b \cdot \nabla)v \tag{13.11}$$

has for arbitrary $v \in H_0^1(\Omega)$ a solution $u_v \in H$. This solution is unique, because of Lemma 13.8. This implies the existence of an operator

$$T : H \to H, \; v \mapsto u_v = T(v),$$

where $T(v)$ is the solution of (13.11). We have the stability property for $v, w \in H$

$$\begin{aligned} \|T(v) - T(w)\|_{H^4(\Omega)} &\leq C\epsilon(\|(b \cdot \nabla)(v-w)\|_{H^2(\Omega)} + \|(b \cdot \nabla)(v-w)\|_{L^\infty(\Omega)}) \\ &\leq C\epsilon\|b\|_{C^2(\Omega)} \left(\|\nabla(v-w)\|_{H^2(\Omega)} + \|\nabla(v-w)\|_{L^\infty(\Omega)}\right) \\ &\leq C_b\epsilon\|\nabla(v-w)\|_{H^3(\Omega)} \\ &\leq C_b\epsilon\|v-w\|_{H^4(\Omega)}, \end{aligned}$$

where we used the embedding $H^3(\Omega) \subset C^1(\Omega)$. For $\epsilon < 1/(2C_b)$, the mapping $T : H \to H$ is a contraction. Application of the Fixed-Point-Theorem of Bannach ensures a fixed point $u_\lambda = T(u_\lambda) \in \overline{B}_R \subset H$. This u_λ solves (13.10). ∎

13.3 Convection-reaction equation

In the case of vanishing viscosity, $\epsilon = 0$, the equation previously discussed become of the form:

$$(b \cdot \nabla)u + cu \quad = \quad f \quad \text{in } \Omega. \tag{13.12}$$

We now look for appropriate boundary conditions. We split the boundary $\partial\Omega$ with outer normal vector $n(x) \in \mathbb{R}^d$ into the following parts:

$$\begin{aligned} \Gamma_+ &:= \{x \in \partial\Omega \,:\, b(x) \cdot n(x) > 0\}, \\ \Gamma_- &:= \{x \in \partial\Omega \,:\, b(x) \cdot n(x) < 0\}, \\ \Gamma_0 &:= \{x \in \partial\Omega \,:\, b(x) \cdot n(x) = 0\}. \end{aligned}$$

Γ_+ is usually named *outflow boundary*, Γ_- wis the *inflow boundary* and Γ_0 is called *characteristic boundary.*

We assume $b \in C(\overline{\Omega}, \mathbb{R}^d)$ and consider for $x_0 \in \overline{\Omega}$ the following system of ordinary differential equations:

$$\begin{aligned} \frac{\partial x}{\partial t} &= b(x)\,, \quad 0 < t \leq t_{max}(x_0), \\ x(0) &= x_0. \end{aligned}$$

Here, $t_{max}(x_0)$ has to be chosen such that $x(t)$ does not leave the closure of the domain Ω. Otherwise, b would not be any more defined. The solutions $x(t)$ are curves in $\overline{\Omega}$ intersecting the starting point x_0. These curves are called *characteristics*.

Theorem 13.10 *For the solution u of the convection-reaction equation (13.12) holds:*

$$u(x(t)) = u(x_0) + \int_0^t [f(x(s)) - c(x(s))u(x(s))]\, ds.$$

Proof. The chain rule yields

$$\frac{du}{dt}(x(t)) = \sum_{i=1}^{d} \frac{\partial u}{\partial x_i}(x(t))\frac{\partial x_i}{\partial t}(t) = \sum_{i=1}^{d} b_i(x(t))\frac{\partial u}{\partial x_i}(x(t)) = (b \cdot \nabla)u(x(t)).$$

This implies

$$\begin{aligned} u(x(t)) &= u(x_0) + \int_0^t \frac{du}{ds}(x(s))\, ds \\ &= u(x_0) + \int_0^t (b \cdot \nabla)u(x(s))\, ds. \end{aligned}$$

Now, we replace the term $(b \cdot \nabla)u(x(s))$ by the corresponding terms according to the differential equation (13.12). This completes the proof. ∎

This result shows that it is not possible to employ Dirichlet values at points $x_1 \in \Gamma_-$ and $x_2 \in \Gamma_+$ simultaneously, which are located on the same characteristic. Usually, one considers values on the inflow boundary Γ_- only.

13.4 Galerkin formulation

We now have a first look onto the Galerkin formulation of (13.6). Due to the previous results of the continuous formulation we go back to the following assumptions:

- L^∞-coefficients $b(x)$, $\nabla b(x)$ and $c(x)$,
- non-vanishing viscosity $\epsilon > 0$,
- with a constant c_0 it holds $c - \frac{1}{2}\mathrm{div}\, b \geq c_0 \geq 0$ pointwise a.e., hence (13.8).

We have the hope that then the discrete version is also bounded and elliptic. We chose the following ϵ-depending norm

$$\|u\|_\epsilon \;:=\; \left(\epsilon|u|^2_{H^1(\Omega)} + c_0\|u\|^2_{L^2(\Omega)}\right)^{1/2}.$$

In this norm we obtain the following a priori error estimate for a pure Galerkin formulation of the convection-diffusion equation:

Theorem 13.11 *Having the same assumtions for the coefficients of (13.6) as in Theorem 13.5, we have for P_r- or Q_r-elements on shape regular meshes for the Galerkin solution u_h the following error bound:*

$$\|u - u_h\|_\epsilon \;\leq\; Ch^r|u|_{H^{r+1}(\Omega)}. \tag{13.13}$$

The constant depends (as usual) on the parameters $C = C(\epsilon, b, c, r, \kappa)$.

Proof. We already mentioned that the bilinear form in (13.7) under assumption (13.8) is elliptic with respect to the $\|\cdot\|_\epsilon$-norm:

$$A(u,u) \;\geq\; \|u\|^2_\epsilon.$$

However, the boundedness of A is valid in a different norm:

$$\begin{aligned}
|A(u,\phi)| &= |\epsilon(\nabla u, \nabla\phi) + ((b\cdot\nabla)u, \phi) + (cu, \phi)| \\
&\leq \epsilon\|\nabla u\|\|\nabla\phi\| + |-((\operatorname{div} b)u, \phi) - (u, (b\cdot\nabla)\phi) + (cu,\phi)| \\
&\leq \epsilon\|\nabla u\|\|\nabla\phi\| + \|(c - \operatorname{div} b)u\|\|\phi\| + \|u\|\|(b\cdot\nabla)\phi\| \\
&\leq \epsilon\|\nabla u\|\|\nabla\phi\| + \|c - \operatorname{div} b\|_\infty\|u\|\|\phi\| + \|b\|_\infty\|u\|\|\nabla\phi\| \\
&\leq \alpha_1\|u\|_\epsilon\|\phi\|_{H^1(\Omega)},
\end{aligned}$$

with an ϵ, b and c-dependent constant $\alpha_1 := \sqrt{\epsilon} + \|b\|_\infty + \|c - \operatorname{div} b\|_\infty$. In the case of $c_0 = 0$ we apply the Poincaré inequality. This already ensures existence of a unique discrete solution u_h. Analogously to Cea's Lemma we deduce for arbitrary $v_h \in V_h$:

$$\begin{aligned}
\|u - u_h\|^2_\epsilon &\leq A(u - u_h, u - u_h) \;=\; A(u - u_h, u - v_h) \\
&\leq \alpha_1\|u - u_h\|_\epsilon\|u - v_h\|_{H^1(\Omega)}.
\end{aligned}$$

Hence

$$\|u - u_h\|_\epsilon \;\leq\; \alpha_1 \inf_{v_h\in V_h}\|u - v_h\|_{H^1(\Omega)}.$$

In combination with the standard interpolationestimate of Theorem 5.12 we obtain the assertion. ∎

Let us consider the simplest case, $r = 1$. The right-hand side in (13.13) now contains the H^2-norm of u. However, this quantity becomes unbounded for $\epsilon \to 0$:

$$\lim_{\epsilon \to 0} |u|_{H^2(\Omega)} \quad = \quad \infty.$$

As a consequence, it is doubtful whether (13.13) states an ϵ-uniform convergence. Later we will consider finite element methods of degree r but with a convergence order $h^{r+1/2}$ instead of h^r.

13.5 Upwinding with finite element

Let us consider the trapezoidal rule for the numerical quadrature of the convective term:

$$((b \cdot \nabla)\psi_j, \psi_i) \quad = \quad \sum_T \int_T (b \cdot \nabla)\psi_j \cdot \psi_i \, dx \;\approx\; \frac{1}{3} \sum_{T \ni x_i} |T| b(x_i) \cdot \nabla \psi_j|_T(x_i).$$

The gradient $\nabla \psi_j(x_i)$ is here not well defined, because $\nabla \psi_j$ is in general discontinuous at mesh points x_i. However, in order to make an upwinding as presented before for finite differences, the gradient $\nabla \psi_j|_T(x_i)$ is substituted by the gradient in a neighbour element T_k with $k = k(i)$ that is located in upwind direction of the node x_i. The exact definition of T_k is as follows, see the work of Tabata [16]:

$$x_i \in T_k \quad \text{and} \quad x_i - \sigma b(x_i) \in T_k \text{ for sufficiently small } \sigma > 0.$$

In case that $b(x_i) = 0$, we chose an arbitrary element T_k which contains x_i as node. Although this k is not uniquely defined, we always find at least on k satisfying this condition. This leads to the approximation

$$(b \cdot \nabla \psi_j, \psi_i) \quad \approx \quad \frac{1}{3} b(x_i) \cdot \nabla \psi_j|_{T_k} \sum_{T \ni x_i} |T| \,. \tag{13.14}$$

As positive result we obtain

$$b(x_i) \cdot \nabla \psi_i|_{T_k} \quad \geq \quad 0,$$

so that the dot product in (13.14) is always non-negative. This is important to arrive at a M-matrix:

Theorem 13.12 *The convection-diffusion-reaction problem (13.6)-(13.7) with $c \geq 0$ discretized with P_1 finite elements leads to a M-matrix, and hence to a discrete inverse monotone operator, if:*

- *the reaction term is approximated by mass lumping,*

- *the convective term is approximated by the upwinding (13.14), and*
- *the mesh $\mathcal{T}_h$ is* weakly acute.

Proof. We show that each individual part (convection, diffusion, reaction) of the equation leads to a M-matrix contribution:
(a) Diffusion: the M-matrix structure we already know due to the restriction to weakly acute meshes.
(b) Convection: For the diagonal elements we obtain

$$(b \cdot \nabla\psi_i, \psi_i) \;\approx\; a_{ii} \;=\; \frac{1}{3} b(x_i) \cdot \nabla\psi_i|_{T_k} \sum_{T \ni x_i} |T| \;\geq\; 0,$$

with $k = k(i, b(x_i))$. With the same notation as previously used for the Laplacian part it holds for inner nodes x_i due to $\varphi|_{T_k} = \sum_j \psi_j|_{T_k} \equiv 1$ for the row sum:

$$\sum_{j=1}^{N} a_{ij} \;=\; \frac{1}{3} b(x_i) \cdot \nabla \left(\sum_{j=1}^{N} \psi_j|_{T_k} \right) \sum_{T \ni x_i} |T| \;=\; \frac{1}{3} b(x_i) \cdot \nabla\varphi|_{T_k} \sum_{T \ni x_i} |T| \;=\; 0.$$

For nodes x_i of boundary triangles we have

$$\sum_{j=1}^{N} a_{ij} \;=\; -\sum_{j=N+1}^{N+r} a_{ij} \;=\; -\frac{1}{3} b(x_i) \cdot \nabla \left(\sum_{j=N+1}^{N+r} \psi_j|_{T_k} \right) \sum_{T \ni x_i} |T| \geq\; 0,$$

where the final inequality results from the fact that for $j = N+1, \ldots, N+r$ holds:

$$(b(x_i) \cdot \nabla)\psi_j|_{T_k} \;=\; -\|\psi_j\|_2 \langle b(x_i), n_j \rangle \;\leq\; 0.$$

(c) Reaction: For $c > 0$ holds $(c\psi_j, \psi_i) > 0$. This produces positive entries on the diagonals, but also on the off-diagonals. This may harm the M-matrix property. Numerical quadrature by mass lumping (trapezoidal rule) on nodal points ξ_k leads to

$$\int_T c\psi_j\psi_i \, dx \;\approx\; \frac{1}{3}|T| \sum_{k=1}^{3} c(\xi_k)\psi_j(\xi_k)\psi_i(\xi_k) = \frac{1}{3}|T| c(\xi_i)\delta_{i,j}.$$

Therefore, the off-diagonal entries of the resulting mass matrix vanish, $M_{ij} = 0$ for $i \neq j$, and $M_{ii} = \frac{1}{3}|T|c(x_i) > 0$. ∎

Remark: The mass lumping is a numerical integration of 2. order. Therefore, such approximation for reaction terms does not harm the overall convergence order if linear elements are used.

13.6 Streamline diffusion

In contrast to the LBB condition (inf-sup condition), the problem of dominant convection can not be circumvented by just a clever choice of appropriate finite elements. Rather, the Galerkin formulation itself should be modified. The prominent method is the streamline diffusion method, also called SUPG (Streamline Upwind Petrov Galerkin). This method makes use of the following stabilized bilinear form and functional:

$$\begin{aligned} S_h(u,\phi) &:= \sum_{T\in\mathcal{T}_h} \delta_T(-\epsilon\Delta u + (b\cdot\nabla)u + cu, (b\cdot\nabla)\phi)_{L^2(T)} \\ F_h(\phi) &:= (f,\phi) + \sum_{T\in\mathcal{T}_h} \delta_T(f,(b\cdot\nabla)\phi)_{L^2(T)}. \end{aligned}$$

The element-wise summation is needed, because the Laplacian Δ can not be applied on discrete functions $u_h \in V_h$ across cell boundaries. However, for linear P_1 elements, the element-wise Laplacian vanishes, so that in this case S_h looks much simpler. The parameter δ_T is on each element a constant real number, depending on the cell size H_T. Later, we will see that the optimal value is given by:

$$\delta_T := \min\left(\frac{h_T^2}{\epsilon}, \frac{h_T}{\|b\|_{L^\infty(T)}}\right). \tag{13.15}$$

This can also be expressed in terms of the so-called *lokal Péclet number*

$$Pe_T := \frac{\|b\|_{L^\infty(T)} h_T}{\epsilon},$$

by $\delta_T = h_T \min(1, Pe_T)/\|b\|_{L^\infty(T)}$. In the case of $Pe_T > 1$ we say that the situation is convection dominated. Otherwise, we speak about the diffusion-dominant regime. For numerical reason it is sometimes beneficial to use differentiable stabilization parameters δ_T, for instance

$$\delta_T := h_T\left(\frac{\epsilon}{h_T} + \|b\|_{L^\infty(T)}\right)^{-1}.$$

This choice differs from the previous definition (13.15) only by a small factor. Note that δ_T can always be scaled by a positive constant without loosing stability.

The stabilization terms contain the strong residual of the equation. Hence, this method is obviously strong consistent. Most of these terms are only needed for this consistency. The only stabilizing term is

$$\sum_{T\in\mathcal{T}_h} \delta_T((b\cdot\nabla)u, (b\cdot\nabla)\phi)_{L^2(T)}.$$

This term acts as diffusion into the direction of the convection b. This becomes more transparent by writing the gradient in the form

$$\nabla u = \frac{1}{\|b\|^2_{L^\infty(T)}}\Big((b\cdot\nabla u)b + (b^\perp\cdot\nabla u)b^\perp\Big),$$

assuming here that b is constant on each element T.

$$(\nabla u, \nabla \phi) \;=\; \frac{1}{\|b\|^2_{L^\infty(T)}} \Big((b\cdot\nabla u, b\cdot\nabla\phi) + (b^\perp\cdot\nabla u, b^\perp\cdot\nabla\phi) \Big)$$

The resulting ellipticity is obtained for P_1-elements V_h by

$$\begin{aligned}
S_h(u_h,u_h) &= \sum_{T\in\mathcal{T}_h} \delta_T((b\cdot\nabla)u_h + cu_h, (b\cdot\nabla)u_h)_{L^2(T)} \\
&\geq \sum_{T\in\mathcal{T}_h} \delta_T \Big(\|(b\cdot\nabla)u_h\|^2_{L^2(T)} - \|cu_h\|_{L^2(T)}\|(b\cdot\nabla)u_h\|_{L^2(T)} \Big) \\
&\geq \sum_{T\in\mathcal{T}_h} \delta_T \left(\|(b\cdot\nabla)u_h\|^2_{L^2(T)} - \frac{1}{2}\|cu_h\|^2_{L^2(T)} - \frac{1}{2}\|(b\cdot\nabla)u_h\|^2_{L^2(T)} \right) \\
&= \frac{1}{2}\sum_{T\in\mathcal{T}_h} \delta_T \Big(\|(b\cdot\nabla)u_h\|^2_{L^2(T)} - \|cu_h\|^2_{L^2(T)} \Big),
\end{aligned}$$

so that we obtain in combination with the ellipticity result in Section **??**:

$$\begin{aligned}
A(u_h,u_h) + S_h(u_h,u_h) &\geq \Big(\epsilon|u_h|^2_{H^1(\Omega)} + c_0\|u_h\|^2_{L^2(\Omega)} \Big) + \\
&\qquad \frac{1}{2}\sum_{T\in\mathcal{T}_h} \delta_T \Big(\|(b\cdot\nabla)u_h\|^2_{L^2(T)} - \|cu_h\|^2_{L^2(T)} \Big) \\
&\geq \frac{1}{2}\left(\epsilon|u_h|^2_{H^1(\Omega)} + c_0\|u_h\|^2_{L^2(\Omega)} + \sum_{T\in\mathcal{T}_h} \delta_T \|(b\cdot\nabla)u_h\|^2_{L^2(T)} \right),
\end{aligned}$$

for h sufficiently small so that $\delta_T\|c\|_{L^\infty(T)} \leq c_0$. This leads to the following Lemma:

Lemma 13.13 *The SUPG method is for the convection-diffusion equation under the same conditions as in Theorem 13.5 and $\delta_T\|c\|_{L^\infty(T)} \leq c_0$ coercive in the mesh-dependent norm*

$$|\!|\!| u |\!|\!|_h \;:=\; \left(\epsilon|u|^2_{H^1(\Omega)} + c_0\|u\|^2_{L^2(\Omega)} + \sum_{T\in\mathcal{T}_h} \delta_T \|(b\cdot\nabla)u_h\|^2_{L^2(T)} \right)^{1/2}.$$

Theorem 13.14 *The SUPG method for the convection-diffusion equation with the choice of parameters as in (13.15), with $u \in H^{r+1}(\Omega)$, the same conditions as in Theorem 13.5 and $\delta_T\|c\|_{L^\infty(T)} \leq c_0$, leads for P_r- or Q_r-elements on shape regular meshes to the following a priori error estimate:*

$$|\!|\!| u - u_h |\!|\!|_h \;\leq\; C \sum_{T\in\mathcal{T}_h} (h_T\|b\|_{L^\infty(T)} + \epsilon)^{1/2} h_T^r |u|_{H^{r+1}(T)}.$$

Proof. We make again use of Theorem 10.1. The coercivity of the stabilized bilinear form with respect to $\|\!| \cdot |\!\|_h$ was already shown in Lemma. We define

$$H_h(u) \;:=\; C \sum_{T\in\mathcal{T}_h} (h_T \|b\|_{L^\infty(T)} + \epsilon)^{1/2} h_T^r |u|_{H^{r+1}(T)}.$$

Therefore, we have just to show

$$\begin{aligned} |\!\| u - I_h u |\!\|_h &\leq c H_h(u) \qquad \forall u \in H^{r+1}, && (13.16)\\ A_h(u - I_h u, \varphi_h) &\leq \alpha_1 H_h(u) |\!\| \varphi_h |\!\|_h \qquad \forall \varphi_h \in X_h. && (13.17) \end{aligned}$$

For proving (13.16) we note that

$$\begin{aligned} \epsilon^{1/2} |u - I_h u|_{H^1(\Omega)} &\leq C \epsilon^{1/2} h^r |u|_{H^{r+1}(\Omega)}\,,\\ \|u - I_h u\|_{L^2(\Omega)} &\leq C h^{r+1} |u|_{H^{r+1}(\Omega)},\\ \sum_{T\in\mathcal{T}_h} \delta_T^{1/2} \|(b\cdot\nabla)(u_h - I_h u)\|_{L^2(T)} &\leq C \sum_{T\in\mathcal{T}_h} \|b\|_{L^\infty(T)} \delta_T^{1/2} h_T^r |u|_{H^{r+1}(T)}. \end{aligned}$$

The norm $|\!\| u - I_h u |\!\|_h$ can be bounded by the sum of all terms appearing on the left hand side and all terms on the right hand side are bounded by $H_h(u)$, as long as $\delta_T \leq h_T/\|b\|_{L^\infty(T)} + \epsilon/\|b\|^2_{L^\infty(T)}$. This condition is for the choice of δ_T according to (13.15) obviously fulfilled. For the constant holds $C = C(b, \kappa, r)$. Hence, (13.16) is shown.
For showing the estimate (13.17), we first bound the Galerkin terms:

$$\begin{aligned} (\epsilon\nabla(u - I_h u), \nabla\varphi_h) &\leq \epsilon^{1/2} C h^r |u|_{H^{r+1}(\Omega)} \epsilon^{1/2} \|\nabla\varphi_h\|,\\ ((b\cdot\nabla)(u - I_h u), \varphi_h) &\leq -(u - I_h u, (b\cdot\nabla)\varphi_h) + (u - I_h u, (\operatorname{div} b)\varphi_h),\\ (u - I_h u, (b\cdot\nabla)\varphi_h)_{L^2(T)} &\leq \delta_T^{-1/2} \|u - I_h u\|_{L^2(T)} \delta_T^{1/2} \|(b\cdot\nabla)\varphi_h\|_{L^2(T)}\,,\\ &\leq \delta_T^{-1/2} C h^{r+1} |u|_{H^{r+1}(T)} |\!\| \varphi_h |\!\|_h,\\ (u - I_h u, (\operatorname{div} b)\varphi_h) + (c(u - I_h u), \varphi_h) &\leq C h^{r+1} |u|_{H^{r+1}(\Omega)} |\!\| \varphi_h |\!\|_h. \end{aligned}$$

All these terms, except the one with the weight $\delta_T^{-1/2}$, are bounded by $H_h(u)|\!\|\varphi_h|\!\|_h$. This particular term will be considered later. Furthermore, for the stabilization terms we note that

$$(\epsilon\Delta(u - I_h u), \delta_T(b\cdot\nabla)\varphi_h)_{L^2(T)} \;\leq\; \epsilon\delta_T^{1/2} \|\Delta(u - I_h u)\| |\!\| \varphi_h |\!\|_h.$$

Moreover, we have

$$\begin{aligned} r = 1: \qquad \|\Delta(u - I_h u)\| &= \|\Delta u\| \leq C|u|_{H^{r+1}(\Omega)},\\ r > 1: \qquad \|\Delta(u - I_h u)\| &\leq h^{r-1} |u|_{H^{r+1}(\Omega)}. \end{aligned}$$

The other stabilization terms can be bounded analogously:

$$\begin{aligned} ((b\cdot\nabla)(u - I_h u), \delta_T(b\cdot\nabla)\varphi_h)_{L^2(T)} &\leq C \|b\|_{L^\infty(T)} \delta_T^{1/2} h^r |u|_{H^{r+1}(\Omega)} |\!\| \varphi_h |\!\|_h,\\ (c(u - I_h u), \delta_T(b\cdot\nabla)\varphi_h)_{L^2(T)} &\leq C \delta_T^{1/2} h^{r+1} |u|_{H^{r+1}(\Omega)} |\!\| \varphi_h |\!\|_h. \end{aligned}$$

Summation over all these terms yields

$$S_h(u - I_h u, \varphi_h) \quad \leq \quad C(\epsilon + h\|b\|_{L^\infty(T)} + h^2)\delta_T^{1/2} h^{r-1} |u|_{H^{r+1}(\Omega)} |\!|\!| \varphi_h |\!|\!|_h.$$

It remains to bound the upper mentioned term with weighting factor $\delta_T^{-1/2}$ as

$$\delta_T^{-1/2} h_T^{r+1} + (\epsilon + h_T\|b\|_{L^\infty(T)} + h_T^2)\delta_T^{1/2} h_T^{r-1} \quad \leq \quad C(h_T\|b\|_{L^\infty(T)} + \epsilon)^{1/2} h_T^r.$$

This is true, if $(h_T < 1)$

$$\delta_T^{-1/2} h_T^2 + (\epsilon + h_T\|b\|_{L^\infty(T)})\delta_T^{1/2} \quad \leq \quad C(\epsilon + h_T\|b\|_{L^\infty(T)})^{1/2} h_T.$$

These bounds are valid for

$$ch_T \;\leq\; (\epsilon + h_T\|b\|_{L^\infty(T)})^{1/2}\delta_T^{1/2} \quad \leq \quad Ch_T.$$

The choice

$$\delta_T \quad = \quad \delta_0 \frac{h_T^2}{\epsilon + h_T\|b\|_{L^\infty(T)}}$$

with arbitrary constant $\delta_0 > 0$ leads to the desired bounds. ∎

Note that the SUPG method does in general not ensure that the resulting stiffness matrix is a M-matrix.

13.7 Shock capturing

Since the streamline diffusion method (SUPG) does not lead to an inverse monotone operator, additional terms are needed. From the practical point of view, this is specially needed in the case that the solution exhibits sharp fronts (large gradients) but the monotonicity should be maintained on the discrete level. The upwind strategy and artificial diffusion iare monotone, but are only of first order. For better accuracy non-linear methods are needed. These are usually embraced under the name *shock capturing*. Here we present some of them.

13.7.1 Crosswind diffusion

One possibility is to introduce diffusion in *crosswind* direction $b^\perp$. In 2D it holds $b^\perp := (-b_2, b_1)/\|b\|$. A further testfunction of the form $(b^\perp \cdot \nabla)\varphi$ with appropriate weights are added. In combination with SUPG the stabilization becomes:

$$S_h(u_h, \varphi) \quad := \quad \sum_{T \in \mathcal{T}_h} (-\epsilon \Delta u_h + (b \cdot \nabla)u_h + cu_h, \delta_T[(b \cdot \nabla) + \gamma_T(b^\perp \cdot \nabla)]\varphi)_{L^2(T)}.$$

Here, δ_T is chosen as before in (13.15). The parameter γ_T depends on u_h in the following form:

$$\gamma_T \;:=\; \gamma_0 \tan\theta_T,$$

where θ_T denotes the angle between β_T and $\nabla u_h|_T$. The vector β_T is an appropriate element-wise projection of b on a constant direction. In order to obtain a fully consistent method, the entire residual has to enter into the stabilization term. Therefore, the corresponding right-hand side is given by

$$F_h(\varphi) \;:=\; \sum_{T\in\mathcal{T}_h} (f, \varphi + \delta_T[(b\cdot\nabla) + \gamma_T(b^\perp\cdot\nabla)]\varphi)_{L^2(T)}.$$

This method is not any more linear, because the parameter γ_T depends on the solution u_h. In [5] the following resultat was shown for $A_h = A + S_h$t:

Theorem 13.15 *In the case of vanishing reaction, i.e. $c\equiv 0$, and the choice*

$$\begin{aligned}
\delta_T &\geq ((d+1)\sin\alpha_T |b|_T \min(|\nabla\psi_{K,1}|,\ldots,|\nabla\psi_{K,d}|))^{-1}, \\
\gamma_T &:= \frac{1+\tan\alpha_K(1-|\cos\theta_T|)^{1/2}(1+|\cos\theta_T|)^{-1/2}}{1+\tan\alpha_K|\tan\theta_T|}\tan\theta_T,
\end{aligned}$$

the resulting semi-linear form A_h is inverse monoton for P_1-elements on triangular strictly acute *meshes $\mathcal{T}_h$.*

In this theorem the angles α_T of an element T is defined as follows: Let $\omega_{K,l}$ denote all appearing inner angles of element K. Then

$$0 \;<\; \alpha_T \;:=\; \frac{\pi}{2} - \max(\omega_{1,T},\ldots,\omega_{L,T}).$$

13.7.2 Non-linear isotropic diffusion

$$S_h(u_h,\varphi) \;:=\; \sum_{T\in\mathcal{T}_h} (-\epsilon\Delta u_h + (b\cdot\nabla)u_h + cu_h, [\delta_T(b\cdot\nabla) + \gamma_T(b^{||}\cdot\nabla)]\varphi)_{L^2(T)}$$

with the diffusion in direction of the gradient

$$b^{||} \;:=\; \frac{(b\cdot\nabla)u_h}{\|u_h\|^2}\nabla u_h,$$

as long as $\nabla u_h \neq 0$, and otherwise $b^{||} := 0$.

13.8 Galerkin least-squares method

In this section we present a general method to discretize partial differential equations. This concept will afterwards applied to convection-diffusion-reaction equations and to the Stokes problem. Let V be a real Hilbert space, $W \subset V$ a closed subspace, and $L : W \to V'$ a linear operator. Furthermore, we assume that V' is a Hilbert space.

Example: $V = H_0^1(\Omega)$, $W = H^2(\Omega) \cap H_0^1(\Omega)$, $H = L^2(\Omega)$ and

$$Lu \quad := \quad (b \cdot \nabla)u - \epsilon \Delta u + cu \tag{13.18}$$

For given $f \in V'$ we consider the equation

$$u \in W : \qquad Lu \quad = \quad f.$$

We have seen before that the discrete counterpart of equation (13.18) has bad stability properties.

The corresponding Least-Squares formulation with conforming finite elements $W_h \subset W$ now reads: Seek $u_h \in W_h$ s.t.

$$\|Lu_h - f\|_{V'}^2 \quad = \quad \min_{v_h \in W_h} \|Lv_h - f\|_{V'}^2.$$

Hence, this corresponds to the minimization of the quadratic functional $J(u) := \frac{1}{2}\|Lu_h - f\|_{V'}^2$. The directional derivative is given by

$$J'(u)(v) \quad = \quad (Lu, Lv)_{V'} - (f, Lv)_{V'}.$$

Therefore, the minimization in W_h is equivalent to the equation

$$u_h \in W_h : \qquad (Lu_h, Lv)_{V'} \quad = \quad (f, Lv)_{V'} \qquad \forall v \in W_h.$$

The advantage of this formulation is its symmetry and positive definiteness of the corresponding stiffness matrix. The disadvantage of this equation is that the corresponding finite element discretization require higher regularity elements. For instance, in the upper mentioned case $W_h \subset W = H^2(\Omega)$, we need C^1-elements. Their construction on general triangulations is very complicated. Moreover, the condition number of the least-squares problem is identical to the square of the original problem, $\operatorname{cond}(L^*L) = (\operatorname{cond} L)^2$. Therefore, we will discuss in the following a combination of Galerkin and least-squares.

13.8.1 Galerkin-Least-Squares Method (GLS)

We consider a bilinear form

$$A(u, v) \quad := \quad \langle Lu, v \rangle$$

which is assumed to be V-coercive with respect to a norm $\|\cdot\|_V$, i.e

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

Because we work with element-wise polynomials, we can apply L on u_h at least element-wise. We only need that $u_h \in V_h \subset V$ instead of $u_h \in W_h \subset W$. Therefore, we introduce the stabilized operator

$$\begin{aligned} A_h(u,v) &:= A(u,v) + S_h(u,v) \\ &= \langle Lu, v\rangle + \sum_{T\in\mathcal{T}_h} \delta_T (Lu, Lv)_{L^2(T)}, \end{aligned}$$

with coefficients δ_T for each element T. The discrete problem now reads:

$$u_h \in V_h : \qquad A_h(u_h, v) \;=\; (f,v) + \sum_{T\in\mathcal{T}_h} \delta_T (f, Lv)_{L^2(T)} \quad \forall v \in V_h.$$

The following lemma shows that this formulation is in a certain sense more stable than the original Galerkin formulation.

Lemma 13.16 *The GLS method is strongly consistent. In case of a coercive bilinear form (with ellipticity constant $\alpha_2 > 0$) we have the following stability property of the discrete GLS operator:*

$$A_h(u_h,u_h) \quad \geq \quad \alpha_2 \|u_h\|_V^2 + \sum_{T\in\mathcal{T}_h} \delta_T \|Lu_h\|_{L^2(T)}^2 \qquad \forall u_h \in V_h.$$

Proof. The strong consistency results from $Lu = f$ for smooth solutions. The stability result is obtained by simple calculus. ∎

13.8.2 Galerkin-Least-Squares method for convection-diffusion

In the case of (13.18) we have the norm

$$\|u\|_V \quad := \quad \Big(\epsilon \|\nabla u\|_{L^2(\Omega)}^2 + c_0^{1/2} \|u\|_{L^2(\Omega)^2} \Big)^2,$$

with $c_0 = \inf_{x\in\Omega} \operatorname{ess}(c(x) - \frac{1}{2}\operatorname{div} b(x)) \geq 0$ and the constant $\alpha_2 = 1$ for coercivity. The GLS stabilization term becomes in this case

$$S_h(u_h,\varphi) \quad = \quad \sum_{T\in\mathcal{T}_h} \delta_T (-\epsilon\Delta u_h + b\cdot\nabla u_h + c u_h, -\epsilon\Delta\varphi + (b\cdot\nabla)\varphi + c\varphi)_{L^2(T)}.$$

In the case of P_1-elements, the second derivatives vanish, so that

$$S_h(u_h,\varphi) \quad = \quad \sum_{T\in\mathcal{T}_h} \delta_T (b\cdot\nabla u_h + c u_h, b\cdot\nabla\varphi + c\varphi)_{L^2(T)}.$$

This stabilization term has a certain similarity with the streamline-diffusion method (SUPG): The difference is the additional test function $\delta_T c\varphi$.

Corollary 13.17 *For the convection-diffusion-reaction equation and parameters δ_T with the bound*

$$0 \leq \delta_T \leq \frac{1}{4} \min\left(\frac{c_0}{\|c\|^2_{L^\infty(T)}}, \frac{h_T^2}{\epsilon \mu_{inv}^2} \right),$$

where μ_{inv} is the constant from the inverse estimate for V_h in Theorem 11.18, we have for all $u_h \in W_h$:

$$A_h(u_h, u_h) \geq \frac{1}{2} \left(\|u_h\|^2_\epsilon + \sum_{T \in \mathcal{T}_h} \delta_T (\|b \cdot \nabla u_h\|^2_{L^2(T)} + \|cu_h - \epsilon \Delta u_h\|^2_{L^2(T)}) \right).$$

Proof. We use the previous lemma and derive a lower bound for $\|Lu_h\|$:

$$\begin{aligned} \|Lu_h\|^2_{L^2(T)} &= (-\epsilon \Delta u_h + b \cdot \nabla u_h + cu_h, -\epsilon \Delta u_h + b \cdot \nabla u_h + cu_h)_{L^2(T)} \\ &= \| - \epsilon \Delta u_h + cu_h\|^2_{L^2(T)} + \|b \cdot \nabla u_h\|^2_{L^2(T)} + 2(cu_h - \epsilon \Delta u_h, b \cdot \nabla u_h)_{L^2(T)}. \end{aligned}$$

The first two term are non-negative, but the third term on the right hand side does not have a sign. Therefore, we bound it modulus from above by

$$\begin{aligned} 2|(cu_h - \epsilon \Delta u_h, b \cdot \nabla u_h)_{L^2(T)}| &\leq 2\|cu_h - \epsilon \Delta u_h\|^2_{L^2(T)} + \frac{1}{2}\|b \cdot \nabla u_h\|^2_{L^2(T)} \\ &\leq 2\frac{\mu_{inv}^2}{h_T^2}\epsilon^2 \|\nabla u_h\|^2_{L^2(T)} + 2\|cu_h\|^2_{L^2(T)} + \frac{1}{2}\|b \cdot \nabla u_h\|^2_{L^2(T)}. \end{aligned}$$

For the assumed upper bound on δ_T we get

$$\begin{aligned} \sum_{T \in \mathcal{T}_h} 2\delta_T \left(\mu_{inv}^2 h_T^{-2} \epsilon^2 \|\nabla u_h\|^2_{L^2(T)} + \|cu_h\|^2_{L^2(T)} \right) &\leq \frac{\epsilon}{2}\|\nabla u_h\|^2 + \frac{c_0}{2}\|u_h\|^2_{L^2(T)} \\ &\leq \frac{1}{2}\|u_h\|^2_\epsilon. \end{aligned}$$

In summary, this yields

$$\sum_{T \in \mathcal{T}_h} \delta_T \|Lu_h\|^2_{L^2(T)} \geq \sum_{T \in \mathcal{T}_h} \delta_T \left(\|cu_h - \epsilon \Delta u_h\|^2_{L^2(T)} + \frac{1}{2}\|b \cdot \nabla u_h\|^2_{L^2(T)} \right) - \frac{1}{2}\|u_h\|^2_\epsilon.$$

The assertion now follows from the previous lemma and the coercivity with $\alpha_2 = 1$. ∎

The element-wise stabilization parameters δ_T can for instance be defined similar to SUPG as:

$$\delta_T := \frac{1}{4} \min\left(\frac{c_0}{\|c\|^2_{L^\infty(T)}}, \frac{h_T}{\|b\|_{L^\infty(T)}}, \frac{h_T^2}{\epsilon \mu_{inv}^2} \right).$$

13.8.3 Galerkin-Least-Squares for Stokes

In this subsection we shortly discuss the GLS method in the case of the Stokes problem. We consider the solution vector $u = (v, p)$ with velocity field v and pressure p, and the operator

$$Lu = \begin{pmatrix} -\Delta v + \nabla p \\ \text{div } v \end{pmatrix}.$$

The GLS-stabilization for vector-valued test functions $\varphi = (\phi, \xi)$ and a diagonal matrix as stabilization parameter $\delta_T = \text{diag}(\delta_T^p, \delta_T^v)$ reads:

$$\begin{aligned} S_h(u,\varphi) &= \sum_{T\in\mathcal{T}_h} (Lu, \delta_T L\varphi)_{L^2(T)} \\ &= \sum_{T\in\mathcal{T}_h} \left(\delta_T^p(-\Delta v + \nabla p, -\Delta\phi + \nabla\xi)_{L^2(T)} + \delta_T^v(\text{div } v, \text{div } \phi)_{L^2(T)}\right). \end{aligned}$$

For P_1-elements this reduces to

$$S_h(u,\varphi) = \sum_{T\in\mathcal{T}_h} \left(\delta_T^p(\nabla p, \nabla\xi)_{L^2(T)} + \delta_T^v(\text{div } v, \text{div } \phi)_{L^2(T)}\right).$$

This contains a element-wise Laplacian for the pressure, which is already known form Section **??**. There is one additional term, called *div-div*–stabilization or *grad-div*–stabilization (due to its analogon in strong form). This type of stabilization will be important for the nonlinear Navier-Stokes equations. However, for Stokes this stabilization term is not necessary and does not give any advantages. The opposite is the case: This term mutually couples all velocity components and therefore introduces additional couplings into the stiffness matrix.

13.9 Discontinuous Galerkin

Another frequently used method to solve equations with dominant convection is the discontinuous Galerkin method (dG). We firstly present this concept in the context of a convection-reaction equation (13.12) with Dirichlet conditions at the inflow boundary Γ_-:

$$\begin{aligned} (b\cdot\nabla)u + cu &= f \qquad \text{in } \Omega, \\ u &= u_0 \qquad \text{on } \Gamma_-. \end{aligned}$$

The variational formulation with weak implementation of the Dirichlet conditions reads

$$u \in V: \quad A(u,\varphi) = (f,\varphi)_{L^2(\Omega)} + \int_{\Gamma_-} |b\cdot n| u_0 \varphi \, ds \qquad \forall \varphi \in V,$$

with the bilinearform given by

$$A(u,\varphi) \;:=\; \sum_{T\in\mathcal{T}_h} ((b\cdot\nabla)u + cu, \varphi)_{L^2(T)} + \int_{\Gamma_-} |b\cdot n| u\varphi \, ds,$$

and the Hilbert space $V := H^1(\Omega)$.

In the discrete setting we will use test- and ansatz-functions of piecewise polynomials, which can be discontinuous at element boundaries:

$$W_h \;:=\; \{v \in L^2(\Omega) \,:\, v|_T \in P_r(T) \quad \forall T \in \mathcal{T}_h\}.$$

This is a non-conforming method, because $W_h \not\subset V$. The main disadvantage of dG methods is that much more degrees of freedom are needed compared to globally continuous test- and ansatz fuctions. These additional degrees do not enhance the approximation quality. They are rather responsible for better stability properties.

Due to the discontinuities at element boundaries we introduce the following notation for the limits from both sides:

$$v^{\pm}(x) \;:=\; \lim_{\epsilon\to 0} v(x \pm \epsilon b(x)).$$

Here, we assume that $b(x)$ is not tangential to the edges for points x on edges of the triangulation. Furthermore, we set the limes to zero, if $x \in \partial\Omega$ and $\pm\epsilon b(x)$ is directed to outside of Ω. As already done in Section 6.2, we use the notation $\mathcal{E}_h$ for the set of all inner edges of the triangulation $\mathcal{T}_h$. The jump over an edge $e \in \mathcal{E}_h$ will be denoted by

$$[\![v]\!]_e(x) \;:=\; v^+(x) - v^-(x)\,, \qquad x \in e.$$

With these notations, the discrete bilinearform reads

$$A_h(u,\varphi) \;:=\; A(u,\varphi) + \sum_{e\in\mathcal{E}_h} \int_e |b\cdot n| [\![u]\!]_e \varphi^+ \, ds. \tag{13.19}$$

The linear system to be solved reads

$$u_h \in W_h: \quad A_h(u_h,\varphi) \;=\; (f,\varphi)_{L^2(\Omega)} + \int_{\Gamma_-} |b\cdot n| u_0 \varphi \, ds \quad \forall \varphi \in W_h.$$

We now show coercivity with respect to the norm:

$$\|u\|_{dG,h} \;:=\; \left(c_0^{1/2} \|u\|^2_{L^2(\Omega)} + \frac{1}{2}\int_{\partial\Omega} |b\cdot n| u^2 \, ds + \frac{1}{2} \sum_{e\in\mathcal{E}_h} \int_e |b\cdot n| [\![u]\!]^2 \, ds \right)^{1/2} \tag{13.20}$$

with the constant $c_0 := \inf_{x\in\Omega} \mathrm{ess}\,(c(x) - \frac{1}{2}\mathrm{div}\; b(x))$.

Lemma 13.18 *The dG-method presented above is strongly consistent and the bilinearform (13.19) has the following coercivity property*

$$A_h(u,u) \;\geq\; \|u\|_{dG,h}^2 \qquad \forall u \in W_h \oplus V.$$

Proof. (a) Consistency: For the exact solution $u \in H^1(\Omega)$ of the infinite dimensional problem all jump terms over inner edges vanish, $[u]_e \equiv 0$. Therefore, it holds for this u:

$$A_h(u,\varphi) \;=\; A(u,\varphi) \;=\; (f,\varphi)_{L^2(\Omega)} + \int_{\Gamma_-} |b\cdot n| u_0 \varphi \, ds.$$

(b) Coercivity: Analogously to the proof of Theorem 13.5 we obtain for each element T and arbitrary $u \in W_h \oplus V$:

$$((b\cdot\nabla)u + cu, u)_{L^2(T)} \;\geq\; d_T \|u\|^2_{L^2(T)} + (\tfrac{1}{2}(b\cdot n)u, u)_{L^2(\partial T)},$$

with $d_T := \inf_{x\in T} \operatorname{ess}\,(c(x) - \frac{1}{2}\operatorname{div} b(x))^{1/2}$. Here, the boundary integral on the right hand side has to be evaluated with values of u as limits from the interior of T. Summation over all elements yields

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} ((b\cdot\nabla)u + cu, u)_{L^2(T)} \;&\geq\; d\|u\|^2_{L^2(\Omega)} + \frac{1}{2}\sum_{T\in\mathcal{T}_h}\int_{\partial T}(b\cdot n)u^2\,ds \\ &=\; d\|u\|^2_{L^2(\Omega)} - \frac{1}{2}\sum_{e\in\mathcal{E}_h}\int_e |b\cdot n|[\![u^2]\!]\,ds + \frac{1}{2}\int_{\partial\Omega}(b\cdot n)u^2\,ds,\end{aligned}$$

where we used $\int_e b\cdot n u^2|_{T_1} + \int_e b\cdot n u^2|_{T_2} = -\int_e |b\cdot n|[\![u^2]\!]$ for an edge $e = \partial T_1 \cap \partial T_2 \in \mathcal{E}_h$. The remaining boundary integrals of $A_h(\cdot,\cdot)$ become by diagonal testing:

$$\begin{aligned}\sum_{e\in\mathcal{E}_h}\int_e |b\cdot n|[\![u]\!]_e u^+\,ds \;&=\; \sum_{e\in\mathcal{E}_h}\int_e |b\cdot n|((u^+)^2 - u^-u^+)\,ds\,, \\ \int_{\Gamma_-}|b\cdot n|u^2\,ds \;&=\; -\int_{\Gamma_-}(b\cdot n)u^2\,ds.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}A_h(u,u) \;\geq\; & d\|u\|^2_{L^2(\Omega)} + \sum_{e\in\mathcal{E}_h}\int_e |b\cdot n|((u^+)^2 - u^+u^- - \tfrac{1}{2}[\![u^2]\!])\,ds \\ & + \frac{1}{2}\int_{\Gamma_+}(b\cdot n)u^2\,ds - \frac{1}{2}\int_{\Gamma_-}(b\cdot n)u^2\,ds.\end{aligned}$$

Taking into account that

$$\begin{aligned}(u^+)^2 - u^-u^+ - \tfrac{1}{2}[\![u^2]\!] \;&=\; \tfrac{1}{2}(u^+)^2 - u^+u^- + \tfrac{1}{2}(u^-)^2 \\ &=\; \tfrac{1}{2}(u^+ - u^-)^2 \;=\; \frac{1}{2}[\![u]\!]^2,\end{aligned}$$

that $b \cdot n > 0$ on Γ_+, $-b \cdot n > 0$ on Γ_- and $b \cdot n = 0$ on $\partial\Omega \setminus (\Gamma_+ \cup \Gamma_-)$, we complete the proof. ∎

In the following we derive an a priori error estimate by using Strangs Lemma. To this end we introduce the following two stronger norms on the discrete space W_h:

$$\begin{aligned} \|u\|_{dG,1}^2 &:= \|u\|_{dG,h}^2 + \sum_{T \in \mathcal{T}_h} h_T \|b \cdot \nabla u\|_{L^2(T)}^2, \\ \|u\|_{dG,2}^2 &:= \|u\|_{dG,h}^2 + \sum_{T \in \mathcal{T}_h} h_T^{-1} \|u\|_{L^2(T)}^2 + \sum_{e \in \mathcal{E}_h} \||b \cdot n| u^-\|_{L^2(e)}^2. \end{aligned}$$

Lemma 13.19 *For $c \in L^\infty(\Omega)$, a Hölder-continuous vector field $b \in \mathcal{C}^{0,1/2}(\overline{\Omega})^d$, with div $b \in L^\infty(\Omega)$ and (13.8) the bilinearform A_h in (13.19) is continuous with respect to the norm on W_h, i.e. there exists a constant $c_1 > 0$ s.t.*

$$A_h(v, w_h) \leq c_1 \|v\|_{dG,2} \|w_h\|_{dG,1} \qquad \forall v \in V \oplus W_h\,,\ \forall w_h \in W_h.$$

Proof. By partial integration on each element we obtain:

$$\begin{aligned} & ((b \cdot \nabla)v + cv, w_h)_{L^2(T)} \\ = \ & (v, (c - \tfrac{1}{2}\text{div } b) w_h))_{L^2(T)} - (v, b \cdot \nabla w_h)_{L^2(T)} - \int_{\partial T} v (b \cdot n) w_h \, ds \\ \leq \ & d_T \|v\|_{L^2(T)} \, d_T \|w_h\|_{L^2(T)} + h_T^{-1} \|v\|_{L^2(T)} \, h_T \|b \cdot \nabla w_h\|_{L^2(T)} - \int_{\partial T} v (b \cdot n) w_h \, ds, \end{aligned}$$

with $d_T := \inf_{x \in T} \operatorname{ess} (c(x) - \frac{1}{2}\text{div } b(x))^{1/2}$. Summation over all elements yields

$$\sum_{T \in \mathcal{T}_h} (b \cdot \nabla v + cv, w_h)_{L^2(T)} \leq c \|v\|_{dG,2} \|w_h\|_{dG,1}.$$

Furthermore, for the boundary integral in the Galerkin term, we have

$$\begin{aligned} \int_{\Gamma_-} |b \cdot n| v w_h \, ds &\leq \|b \cdot n v\|_{L^2(\Gamma_-)} \|b \cdot n w_h\|_{L^2(\Gamma_-)} \\ &\leq \|v\|_{dG,h} \|w_h\|_{dG,h}. \end{aligned}$$

For the jump terms we use the equality $[\![v]\!] w_h^+ = -v^- [\![w_h]\!] + [\![v w_h]\!]$:

$$\begin{aligned} \sum_{e \in \mathcal{E}_h} \int_e |b \cdot n| [\![v]\!]_e w_h^+ \, ds &= \sum_{e \in \mathcal{E}_h} \left(- \int_e |b \cdot n| v^- [\![w_h]\!] \, ds + \int_e |b \cdot n| [\![v w_h]\!] \, ds \right) \\ &\leq \|v\|_{dG,2} \|w_h\|_{dG,h}. \end{aligned}$$

In sum we arrive at the assertion. ∎

We will use the following stability property of A_h with respect to the norm $\|\cdot\|_{dG,1}$:

Lemma 13.20 *Under the same assumptions as in Lemma 13.19, the following discrete inf-sup condition holds for the dG method: It exists $\gamma > 0$ s.t.*

$$\inf_{v_h \in W_h} \sup_{w_h \in W_h} \frac{A_h(v_h, w_h)}{\|v_h\|_{dG,1} \|w_h\|_{dG,1}} \quad \geq \quad \gamma.$$

Proof. We choose $w_h := v_h + c_1 \sum_T h_T \bar{b}_T \cdot \nabla v_h|_T$ where $\bar{b}_T$ is a element-wise linear interpolation of $b|_T$. We obtain

$$\begin{aligned}
A_h(v_h, w_h) &= A_h(v_h, v_h) + c_1 A_h(v_h, \sum_T h_T \bar{b}_T \cdot \nabla v_h|_T) \\
&\geq \|v_h\|^2_{dG,h} + c_1 \sum_{T \in \mathcal{T}_h} h_T (b \cdot \nabla v_h + c v_h, \bar{b}_T \cdot \nabla v_h)_{L^2(T)} \\
&\quad + c_1 \sum_{T \in \mathcal{T}_h} h_T \int_{\Gamma_- \cap T} |b \cdot n| v_h \bar{b}_T \cdot \nabla v_h \, ds \\
&\quad + c_1 \sum_{T \in \mathcal{T}_h} h_T \sum_{e \in \mathcal{E}_h} \int_{e \cap T} |b \cdot n| [\![v_h]\!]_e (\bar{b}_T \cdot \nabla v_h)^+ \, ds \\
&= \|v_h\|^2_{dG,1} + c_1 \sum_{T \in \mathcal{T}_h} h_T \left((b \cdot \nabla v_h, (\bar{b}_T - b) \cdot \nabla v_h)_{L^2(T)} + (c v_h, \bar{b}_T \cdot \nabla v_h)_{L^2(T)} \right) \\
&\quad + c_1 \sum_{T \in \mathcal{T}_h} h_T \int_{\Gamma_- \cap T} |b \cdot n| v_h \bar{b}_T \cdot \nabla v_h \, ds + .. \\
&\geq \frac{1}{2} \|v_h\|^2_{dG,1} - c_1 \sum_{T \in \mathcal{T}_h} h_T \left(\|\bar{b}_T - b\|^2_{L^\infty(T)} \|\nabla v_h\|^2_{L^2(T)} + \|c v_h\|^2_{L^2(T)} \right) \\
&\quad + c_1 \sum_{T \in \mathcal{T}_h} h_T \int_{\Gamma_- \cap T} |b \cdot n| v_h \bar{b}_T \cdot \nabla v_h \, ds + ...
\end{aligned}$$

Siehe [10]. ∎

The following result states the optimality of the dG method for pure convection-reaction equations. Since the solution of such an equation (i.e. without diffusion) is usually not smooth, the case $r = 0$ and $r = 1$ are the most relevant ones.

Theorem 13.21 *Under the same assumptions as in Lemma 13.19, the dG method is quasi-optimal in the following sense:*

$$\|u - u_h\|_{dG,1} \quad \leq \quad \inf_{w_h \in W_h} \left(\|u - w_h\|_{dG,1} + c \|u - w_h\|_{dG,2} \right).$$

If $u \in H^{k+1}(\Omega)$, we have in particular for $dG(r)$ with $0 \leq k \leq r$:

$$\|u - u_h\|_{dG,1} \quad \leq \quad c h^{k+1/2} |u|_{H^{k+1}(\Omega)}.$$

Proof. (a) As usual, we split the total error $u - u_h$ into the interpolation error $\eta = u - w_h$, with arbitrary $w_h \in W_h$, and the projection error $\xi = w_h - u_h$. We use the

discrete inf-sup property, Galerkin orthogonality, and continuity of A_h:

$$\gamma\|\xi\|_{dG,1} \;\leq\; \sup_{\phi_h\in W_h} \frac{A_h(\xi,\phi_h)}{\|\phi_h\|_{dG,1}} \;=\; \sup_{\phi_h\in W_h} \frac{-A_h(\eta,\phi_h)}{\|\phi_h\|_{dG,1}} \;\leq\; C\|\eta\|_{dG,2}.$$

This yields

$$\|u-u_h\|_{dG,1} \;\leq\; \|\eta\|_{dG,1} + C\gamma^{-1}\|\eta\|_{dG,2}.$$

Since w_h was an arbitrary admissible finite element function, we obtain the first estimate.
(b) We choose the nodal interpolation $w_h = I_h u \in W_h$. We firstly consider the individual terms of $\|\eta\|_{dG,h}$: The L^2 terms are bounded easily by

$$\|\eta\|^2_{L^2(\Omega)} \;\leq\; h_T^{2(k+1)}|u|^2_{H^{k+1}(T)}.$$

Because of the continuity of $I_h u$, the jump terms vanish:

$$\int_e |b\cdot n|[\![\eta]\!]^2\,ds \;=\; 0.$$

For the boundary terms we use the multiplicative trace estimate stated below:

$$\begin{aligned}\int_e |b\cdot n|\eta^2\,ds &\leq C\|\eta\|_{L^2(e)}\\ &\leq C(\|\eta\|_{L^2(T)}|\eta|_{H^1(T)} + h_T^{-1}\|\eta\|^2_{L^2(T)})\\ &\leq Ch_T^{2k+1}|u|^2_{H^{k+1}(T)}.\end{aligned}$$

The constant C depends on $\|b\|_{L^\infty(T)}$. Hence, $\|\eta\|_{dG,h}$ is properly bounded. The additional terms of $\|\eta\|_{dG,1}$ are bounded as follows:

$$h_T\|b\cdot\nabla\eta\|^2_{L^2(T)} \;\leq\; ch_T^{2k+1}|u|^2_{H^{k+1}(T)}.$$

The additional terms of $\|\eta\|_{dG,2}$ are bounded as follows:

$$h_T^{-1}\|\eta\|^2_{L^2(T)} \;\leq\; h_T^{2(k+1)-1}|u|^2_{H^{k+1}(T)},$$

with a constant $c = c(\|b\|_{L^\infty(\Omega)})$. This shows

$$\|\eta\|_{dG,1} + \|\eta\|_{dG,2} \;\leq\; C\left(\sum_{T\in\mathcal{T}_h} h_T^{2(k+1)}|u|^2_{H^{k+1}(T)}\right)^{1/2}$$

which, in combination with (a), finalizes the proof. ∎

The following variant of the trace theorem was used in the previous proof to obtain the factor $h^{1/2}$ more than in the usual trace theorem 3.11:

Theorem 13.22 (Multiplicative trace estimate) *On shape regular meshes $\mathcal{T}_h$ there exists a constant $c > 0$ s.t. we have for each element $T \in \mathcal{T}_h$:*

$$\|v\|^2_{L^2(\partial T)} \leq c(\|v\|_{L^2(T)}|v|_{H^1(T)} + h_T^{-1}\|v\|^2_{L^2(T)}) \qquad \forall v \in H^1(T).$$

Proof. We present the proof of [8]. Without lost of generality we can assume that the midpoint of T is the origin. Let ρ_T denote the inner radius of T, h_T the outer radius, and n the normal of ∂T. For $x \in \partial T$ we have for the Euklidean norm $\rho_T \leq \|x\|_2 \leq h_T$. This yields the bound

$$\|v\|^2_{L^2(\partial T)} \leq \rho_T^{-1} \int_{\partial T} x \cdot n v^2 \, ds.$$

We now apply the Gauss integration Theorem:

$$\begin{aligned} \|v\|^2_{L^2(\partial T)} &\leq \rho_T^{-1} \int_T \text{div } (v^2 x) \, dx \\ &= \rho_T^{-1} \int_T (v^2 \text{div } x + x \cdot \nabla(v^2)) \, dx \\ &= \rho_T^{-1} \int_T (v^2 d + 2vx \cdot \nabla v) \, dx. \end{aligned}$$

By the inequality of Cauchy-Schwarz we now obtain

$$\begin{aligned} \|v\|^2_{L^2(\partial T)} &\leq \rho_T^{-1}(d\|v\|^2_{L^2(T)} + 2h_T\|v \cdot \nabla v\|_{L^1(T)}) \\ &\leq C h_T^{-1}(\|v\|^2_{L^2(T)} + h_T\|v\|_{L^2(T)}\|\nabla v\|_{L^2(T)}). \end{aligned}$$

This finalizes the proof. ∎

Bibliography

[1] H. W. Alt. *Lineare Funktionalanalysis*. Berlin: Springer. v, 431 p., 2006.

[2] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer, 4.ed., 2007.

[3] H. Brézis. *Análysis Functional*. Alianza Editorial. 233 p., 1984.

[4] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, Berlin, 1991.

[5] E. Burman and A. Ern. Continuous interior penalty *hp*-finite element methods for advection and advection-diffusion equations. *Math. Comput.*, 76(259):1119–1140, 2007.

[6] P. Clément. Approximation by finite element functions using local regularization. *R.A.I.R.O. Anal. Numer.*, 9:77–84, 1975.

[7] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Franc. Automat. Inform. Rech. Operat.*, 7(R-3):33–76, 1974.

[8] V. Dolejsi, M. Feistauer, and C. Schwab. A finite volume discontinuous Galerkin scheme for nonlinear convection-diffusion problems. *Calcolo*, 39(1), 2002.

[9] G. Duvaut and J. Lions. Inequalities in mechanics and physics. Translated from the French by C.W. John. Grundlehren der mathematischen Wissenschaften. Band 219. Berlin-Heidelberg-New York: Springer-Verlag. XVI, 397 p., 1976.

[10] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Applied Mathematical Sciences, 159, Springer, 2004.

[11] W. Hackbusch. *Multi-grid methods and applications*. Springer Series in Computational Mathematics, 4. Berlin etc.: Springer- Verlag. XIV, 377 p., 1985.

[12] T. J. Hughes, A. Masud, and J. Wan. A stabilized mixed discontinuous Galerkin method for Darcy flow. *Comput. Methods Appl. Mech. Eng.*, 195(25-28):3347–3381, 2006.

[13] J. Nečas. L'application de l'egalite de Rellich sur les systèmes elliptiques du deuxième ordre. *J. Math. Pures Appl. (9)*, 44:133–147, 1965.

[14] R. Rannacher. Numerische Mathematik 2 (Numerik Partieller Differentialgleichungen). Technical report, Universität Heidelberg, `http://numerik.iwr.uni-heidelberg.de/~lehre`, 2008.

[15] L. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comput.*, 54(190):483–493, 1990.

[16] M. Tabata. A finite element approximation corresponding to the upwind difference. *Memoirs of Numerical Mathematics*, 1:47–63, 1977.

[17] W. Zulehner. *Numerische Mathematik, Eine Einführung anhand von Differentialgleichungen, Bd. 1.* Birkhäuser, 2008.

Index