Michele Ginesi

(Unofficial) Lecture Notes
Advanced Numerical Analysis II

May 7, 2025

Contents

Chapter 1

The Poisson problem

We start with the strong formulation of the Dirichlet problem: find u such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^2$ is a *domain* (i.e. an open, bounded and connected set) with boundary $\Gamma = \partial\Omega$, $f = f(x)$ is a given function and $\Delta = \partial_{xx} + \partial_{yy}$ is the *Laplacian operator*. In general, we require Ω to be Lipschitzian.

1.1 Minimization principle

The finite elements method is not based on the strong form, but rather on a *minimization statement* or, more generally, on a *weak formulation*: find

$$u = \arg\min_{w\in X} J(w),$$

where

$$X = \{v \text{ sufficiently smooth function : } v|_\Gamma = 0\}$$

is a linear space (we will specify later what "enough smooth" means), and

$$J(w) := \frac{1}{2}\int_\Omega \nabla w \cdot \nabla w \, dA - \int_\Omega f w \, dA$$

Note that $\nabla w \cdot \nabla w = (\partial_x w)^2 + (\partial_y w)^2$.

Proposition 1.1. *Over all functions $w \in X$, u which satisfies the Poisson problem* (1.1) *makes $J(w)$ as small as possible.*

Proof. Let $w = u + v$. Then

$$\begin{aligned} J(u+v) &= \frac{1}{2}\int_\Omega \nabla(u+v)\cdot\nabla(u+v)\, dA - \int_\Omega f(u+v)\, dA \\ &= \underbrace{\frac{1}{2}\int_\Omega \nabla u\cdot\nabla u\, dA - \int_\Omega f u\, dA}_{J(u)} + \underbrace{\int_\Omega \nabla u\cdot\nabla v\, dA - \int_\Omega f v\, dA}_{\delta J_v(u)} + \underbrace{\frac{1}{2}\int_\Omega \nabla v\cdot\nabla v\, dA}_{>0 \text{ for } v\neq 0}. \end{aligned}$$

Now,

$$
\begin{aligned}
\delta J_v(u) &= \int_\Omega \nabla u\cdot\nabla v\, dA - \int_\Omega f v\, dA\\
&= \int_\Omega \underbrace{\nabla(v\nabla u)}_{\nabla v\cdot\nabla u + v\Delta u}\, dA - \int_\Omega v\Delta u\, dA - \int_\Omega f v\, dA\\
&= \int_\Gamma v\nabla u\cdot\hat{n}\, dS - \int_\Omega v(\Delta u + f)\, dA\\
&= 0, \qquad \forall v\in X,
\end{aligned}
$$

since $v\equiv 0$ on Γ and $\Delta u + f = 0$ in Ω. Thus

$$J(u+v) = J(u) + \frac{1}{2}\int_\Omega \nabla v\cdot\nabla v\, dA, \qquad \forall v\in X;$$

whence $J(w) > J(u), \forall w\in X\setminus\{u\}$. Thus, u is a minimizer. □

Our weak formulation is: find $u\in X$ such that $\delta J_v(u) = 0, \forall v\in X$, i.e.

$$\int_\Omega \nabla u\cdot\nabla v\, dA = \int_\Omega f v\, dA, \qquad \forall v\in X.$$

Definition 1.1. A *linear form L* is a functional $L: Y\to\mathbb{R}$ such that

$$L(\alpha v_1 + \beta v_2) = \alpha L(v_1) + \beta L(v_2), \alpha,\beta\in\mathbb{R}, \quad v_1, v_2\in Y.$$

Definition 1.2. A *bilinear form B* is a functional $B: Y\times Z\to\mathbb{R}$ which is linear with respect to both arguments.
A bilinear form $B: Y\times Y\to\mathbb{R}$ is said to be *symmetric and positive definite* (SPD) if it is

1. *symmetric*: $B(v,w) = B(w,v)$, for any $v,w\in Y$;
2. *positive definite*: $B(w,w) > 0$ for any $w\in Y\setminus\{0\}$.

If we now consider the SPD bilinear form

$$a(w,v) = \int_\Omega \nabla w\cdot\nabla v\, dA,$$

and the linear form

$$l(v) = \int_\Omega f v\, dA,$$

we can rewrite the minimization principle and the weak statement as follows:

$$u = \operatorname*{argmin}_{w\in X}\frac{1}{2}a(w,w) - l(w) \Leftrightarrow a(u,v) = l(v), \qquad \forall v\in X.$$

Since a involves only first derivatives, we have $X = \mathscr{H}^1_0(\Omega) = \overline{\mathscr{C}^1_C}^{\mathscr{H}^1}$ (i.e. the closure of $\mathscr{C}^1_C$ with respect to the $\mathscr{H}^1$ norm), so we have an inner product with the induced norm

$$
\begin{aligned}
\langle w,v\rangle_{\mathscr{H}^1(\Omega)} &= \int_\Omega (\nabla w\cdot\nabla v + v w)\, dA,\\
\|w\|_{\mathscr{H}^1(\Omega)} &= \left(\int_\Omega \left(|\nabla w|^2 + w^2\right) dA\right)^{\frac{1}{2}}.
\end{aligned}
$$

Recall. In an Hilbert space, we have the *Cauchy-Schwarz inequality*

$$\langle w,v\rangle_Y \le \|w\|_Y\,\|v\|_Y.$$

Theorem 1.1. *Let $u \in \mathscr{H}^1(\Omega)$, Ω open in $\mathbb{R}$. Then there exists a unique $\tilde{u} \in \mathscr{C}(\Omega)$ such that*

$$\tilde{u}(x) - \tilde{u}(y) = \int_y^x u'(t)\, dt$$

and $\tilde{u} = u$ almost everywhere.

We call $\tilde{u}$ the *continuous representative* of u and in 1-dim we can make the language abuse $u = \tilde{u}$.

The scalar product which makes $\mathscr{H}^1$ an Hilbert space is

$$\langle u, v \rangle_{\mathscr{H}^1(\Omega)} = \int_\Omega uv\, dA + \int_\Omega u'v'\, dA,$$

which induces the norm

$$\| u \|_{\mathscr{H}^1(\Omega)} = \sqrt{\langle u, u \rangle_{\mathscr{H}^1(\Omega)}}.$$

We can extend these results: let Ω be open in $\mathbb{R}$, $\mathscr{H}^m(\Omega)$ is the space of L^2 functions which admit m weak derivatives still in L^2:

$$\mathscr{H}^m(\Omega) = \left\{ v : \int_\Omega v^2 < \infty, \int_\Omega (v')^2, \infty, \dots, \int_\Omega \left(v^{(m)}\right)^2 < \infty \right\},$$

with associated inner product and norm

$$\langle w, v \rangle_{\mathscr{H}^m(\Omega)} = \sum_{j=0}^m \int_\Omega \frac{d^j w}{dx^j} \frac{d^j v}{dx^j}\, dx,$$

$$\| w \|_{\mathscr{H}^m(\Omega)} = \left(\sum_{j=0}^m \int_\Omega \left(\frac{d^j w}{dx^j} \right)^2 dx \right)^{\frac{1}{2}}.$$

Another important space is $\mathscr{H}_0^1(\Omega)$: the closure of $\mathscr{C}_C^1(\Omega)$ in $\mathscr{H}^1(\Omega)$. Since $\mathscr{C}_C^\infty(\Omega)$ is dense in $\mathscr{H}^1(\Omega)$, we can see $\mathscr{H}_0^1(\Omega)$ as the closure of $\mathscr{C}_C^\infty(\Omega)$ in $\mathscr{H}^1(\Omega)$.

Theorem 1.2. *Let $u \in \mathscr{H}^1(\Omega)$, $\Omega \subset \mathbb{R}$. Then $u \in \mathscr{H}_0^1(\Omega)$ if and only if $u = 0$ on $\partial\Omega$. Moreover we have $u \in \mathscr{H}_0^1(\Omega)$ if and only if $\bar{u} \in \mathscr{H}^1(\mathbb{R})$, $u = \bar{u}$ on Ω and $\bar{u}(x) = 0$ for any $x \in \mathbb{R} \setminus \Omega$.*

In higher dimension, $\Omega \subset \mathbb{R}^n$, the definition is similar: we just need to consider partial derivatives instead of standard ones. The main difference is that, in general, *we can't always find a continuous representative.*

Example. For $k \in (0, 1/2)$, the function

$$u(x, y) = \left(\log \left(\sqrt{x^2 + y^2} \right) \right)^k$$

is in $\mathscr{H}^1(B(0,1))$.

The formal definition of $\mathscr{H}_0^1(\Omega)$ is the same: the closure of $\mathscr{C}_C^1(\Omega)$ in $\mathscr{H}^1(\Omega)$; and, again, $\mathscr{C}_C^\infty(\Omega)$ is dense in $\mathscr{H}_0^1(\Omega)$. The main difference is that, in general, *there is not a continuous representative, so we couldn't think about functions which are zero at the boundary.*

Proposition 1.2. *Let Ω be a bounded and sufficiently regular domain, and let $u \in \mathscr{H}^1(\Omega) \cap \mathscr{C}\left(\bar{\Omega}\right)$. Then $u \in \mathscr{H}_0^1(\Omega)$ if and only if $u = 0$ on $\partial\Omega$.*

Theorem 1.3. *Let Ω be bounded, $\partial\Omega$ sufficiently regular. Then there exists a unique $T : \mathscr{H}^1(\Omega) \to L^2(\partial\Omega)$ linear and continuous such that*

1. *$Tu = u|_{\partial\Omega}$ if $u \in \mathscr{H}^1(\Omega) \cap \mathscr{C}\left(\bar{\Omega}\right)$;*

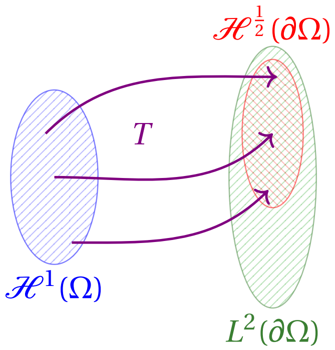
2. *$\| Tu \|_{L^2(\partial\Omega)} \le C \| u \|_{\mathscr{H}^1(\Omega)}$.*

Such an operator is called *trace*, Tu is called the trace of u. Such an operator is not surjective on $L^2(\partial\Omega)$. The set of functions in $L^2(\partial\Omega)$ which are traces of functions in $\mathcal{H}^1(\Omega)$ is a subspace of $L^2(\partial\Omega)$ denoted by $\mathcal{H}^{1/2}(\partial\Omega)$. We have the following chain of inclusions:

$$\mathcal{H}^1(\partial\Omega) \subset \mathcal{H}^{1/2}(\partial\Omega) \subset L^2(\partial\Omega) \equiv H^0(\partial\Omega).$$

Take now $u \in \mathcal{H}^k(\Omega)$. Then

$$Tu \in \mathcal{H}^{k-\frac{1}{2}}(\partial\Omega) \subset \mathcal{H}^{k-1}(\partial\Omega).$$



Example. For $n = 2$, $u \in \mathcal{H}^1(\Omega)$, $Tu \in \mathcal{H}^{1/2}(\partial\Omega)$ is not regular; but if we start with $v \in \mathcal{H}^2(\Omega)$, we get $Tv \in \mathcal{H}^{3/2}(\partial\Omega) \subset \mathcal{H}^1(\Omega)$, so tv has L^2 derivatives.

Remark. Take $n = 2$, $\mathcal{H}^1(\partial\Omega)$ contains only functions which admit continuous representative (since $\partial\Omega$ is isomorph to a subset of $\mathbb{R}$), $\mathcal{H}^{1/2}(\partial\Omega)$ unfortunately not.

Example. With $\Omega = \left\{(x,y) : x^2 + y^2 \le 1, x \ge 0\right\}$ as domain, $k \in (0, 1/2)$,the function $u(x,y) = \left(\log\left(\left(x^2+y^2\right)^{-1/2}\right)\right)^k$ is such that Tu is not continuous.

Remark. Functions in $\mathcal{H}^{1/2}(\Omega)$ does not admit jumps, but only other types of discontinuity.

We can now interpret

$$\mathcal{H}^1_0(\Omega) = \ker(T) = \left\{u \in \mathcal{H}^1(\Omega) : Tu \equiv 0\right\}.$$

Let's now consider $\Omega \subset \mathbb{R}^2$, Γ_l a line in Ω, then we can define

$$l(v) = \int_{\Gamma_l} Tv \, d\gamma$$

with $v \in \mathcal{H}^1$. The functional $l : \mathcal{H}^1 \to \mathbb{R}$ is clearly linear, but is it well defined? We can consider a new domain in which Γ_l is a part of the boundary, so $Tv \in L^2(\Gamma_l)$. Moreover we have

$$|l(v)| \le \int_{\Gamma_l} |Tv| \, d\gamma \le C \|Tv\|_{L^2(\Gamma_l)} \le C \|v\|_{\mathcal{H}^1},$$

so l is bounded. This is an interesting fact, since in the problem $-u'' = g$ we have already constructed

$$a(u,v) = \int u'v' = \int gv = l(v).$$

We now recap some results about embedding of Sobolev spaces. Let $\Omega \subset \mathbb{R}^n$ be open and bounded, $\mathscr{C}^j(\bar{\Omega}) \subset \mathscr{C}^j(\Omega)$ such that $D^\alpha u$ is bounded and uniformly continuous, with $0 \le |\alpha| \le n$, is a Banach space with norm

$$\|u\|_{\mathscr{C}^m(\Omega)} = \max_{0\le|\alpha|\le m} \sup_{x\in\Omega} \left|D^\alpha u(x)\right|.$$

If Ω is sufficiently regular, $m \ge n/2$, then

$$\mathcal{H}^{j+m} \hookrightarrow \mathscr{C}^j(\bar{\Omega}), \qquad\qquad j = 0, 1, 2, \ldots.$$

If $u \in \mathcal{H}^{j+m}(\Omega)$, there exists $\tilde{u} \in \mathscr{C}^j(\Omega)$ such that $\tilde{u} = u$ almost everywhere and

$$\|\tilde{u}\|_{\mathscr{C}^j(\bar{\Omega})} \le K \|u\|_{\mathcal{H}^{j+m}(\Omega)}.$$

Recall. Given two metric spaces (X, d_1) and (Y, d_2), a function $f : X \to Y$ is said to be *uniformly continuous* if for every real number $\varepsilon > 0$ there exists $\delta > 0$ such that for every $x, y \in X$ with $d_1(x,y) < \delta$, we have $d_2(f(x), f(y)) < \varepsilon$.

Example. $n = m = 1$, $\mathcal{H}^{j+1} \hookrightarrow \mathscr{C}^j$, functions in $\mathcal{H}^k$ admit representatives in $\mathscr{C}^{k-1}$.

Example. $n = m = 2$, $\mathscr{H}^2 \hookrightarrow \mathscr{C}^0$ ($\mathscr{H}^1$ functions are not continuous in dimension one), also for $n = 3$, $m = 2$ $\mathscr{H}^2 \hookrightarrow \mathscr{C}^0$.

We now introduce a *seminorm*

$$|w|_{\mathscr{H}^m} = \sqrt{\int_0^1 \left(\frac{d^m w}{dx^m}\right)^2 dx.}$$

In $\mathscr{H}^1_0(\Omega)$ this seminorm is equivalent to the norm:

$$C \| w \|_{\mathscr{H}^1} \le |w|_{\mathscr{H}^1} \le \| w \|_{\mathscr{H}^1}, \qquad \forall w \in \mathscr{H}^1_0$$

The second inequality is trivial ($\| w \|_{\mathscr{H}^1} = \|(\| w)_{L^2} + \left\| w' \right\|_{L^2} = |w|_{\mathscr{H}^1} + \| w \|_{L^2}$); the first one is called *Poincaré-Friedich inequality*. Moreover, in the case $n = 1$,

$$\begin{aligned} C^2 \left(\| w \|^2_{L^2} + \left\| w' \right\|^2_{L^2} \right) \le |w|^2_{\mathscr{H}^1} &\Leftrightarrow \| w \|^2_{L^2} + |w|^2_{\mathscr{H}^1} \le C^{-2} |w|^2_{\mathscr{H}^1} \\ &\Leftrightarrow \| w \|^2_{L^2} \le |w|^2_{\mathscr{H}^1} \left(C^{-2} - 1 \right) \end{aligned}$$

which is a sort of Poincaré's inequality.

Recalling the linear bounded operator $l : X \to \mathbb{R}$ ($l \in X'$) and our bilinear form a, $a(u, v) = l(v)$; l is called the *data*. If $X = \mathscr{H}^1_0$ ($X' = \mathscr{H}^{-1}$), then $|l(v)| \le C \| v \|_{\mathscr{H}^1}$ for any $v \in \mathscr{H}^1_0$ and we have the *dual norm*

$$\| l \|_{\mathscr{H}^{-1}} = \sum_{v \in \mathscr{H}^1_0 \setminus \{0\}} \frac{l(v)}{\| v \|_{\mathscr{H}^1}}.$$

Example. $\Omega = (0, 1)$, $Y = L^2(\Omega)$, $L(v) = \int_0^1 v$, then

$$L(v) = \int_0^1 1\, v \le \left(\int_0^1 1^2 \right)^{\frac{1}{2}} \left(\int_0^1 v^2 \right)^{\frac{1}{2}} = \| v \|_{L^2},$$

so $L \in \left(L^2\right)'$. Moreover $\left(L^2\right)' \cong L^2$ in the sense that any functional $L_\eta(v)$ of the form $L_\eta(v) = \int_0^1 \eta v$ for $\eta \in L^2$ is bounded: $\left\| L\eta \right\|_{(L^2)'} = \left\| \eta \right\|_{L^2}$.

Example. $Y = \mathscr{H}^1_0(\Omega)$, $L(v) = v(1/2)$ (we are talking about the continuous representative). We can write

$$L(v) = v\left(\frac{1}{2}\right) = \int_0^{\frac{1}{2}} v' \le \left(\int_0^{\frac{1}{2}} (v')^2 \right)^{\frac{1}{2}} \left(\int_0^{\frac{1}{2}} 1^2 \right)^{\frac{1}{2}} \le \sqrt{\frac{1}{2}} |v|_{\mathscr{H}^1}.$$

Therefore L is bounded also in the $\mathscr{H}^1$ norm. Moreover

$$L \in \mathscr{H}^{-1}(\Omega) \setminus \left(L^2(\Omega)\right)'.$$

Theorem 1.4 (Riesz representation)**.** *Let Y be an Hilber space. Then for any $L \in Y'$ there exists a unique $u \in Y$ such that $\langle u, v \rangle_Y = L(v)$, for any $v \in Y$.*

Example. Let $\eta \in L^2$, L_η the associated linear function $L_\eta(v) = \int_0^1 \eta v \, dx$, $v \in \mathscr{C}^\infty_C$. So by Cauchy-Schwarz, $L_\eta \in \left(L^2\right)' \cong L^2$ and also $L_\eta \in \left(\mathscr{C}^\infty_C\right)' = \mathscr{D}$ (the set of distributions).

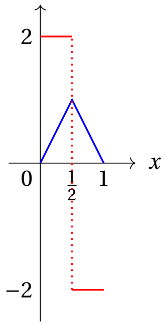
Example. We can consider the delta distribution $L_{\delta_{x_0}}(v) = v(x_0)$, $L_{\delta_{x_0}} \in \mathscr{D}$, but $L_{\delta_{x_0}} \notin L^2$ since we cannot write $v(x_0) = \int \eta v$.

Let us also recall the *distributional derivatives*: consider L, then

$$\begin{aligned} D^m L(v) &= \left\langle D^m L, v \right\rangle \\ &= (-1)^m \left\langle L, \frac{d^m v}{dx^m} \right\rangle \\ &= (-1)^m L \left(\frac{d^m v}{dx^m} \right). \end{aligned}$$

Take $\eta \in L^2$, we want to write $\left\langle D^m L_\eta, v \right\rangle = \int_0^1 \tilde{D}^m_\eta v$ for some $\tilde{D}^m_\eta \in L^2$. If it was possible, we could conclude that η has m (distributional) derivatives in L^2.

Example. Take $\eta = 1 - 2|x - 1/2|$. η is clearly in L^2. We can define $L_\eta(v) = \int_0^1 \eta v \, dx$. So, taking $m = 1$,



$$\begin{aligned}\langle D^1 L_\eta, v\rangle &= -\int_0^1 \eta \frac{dv}{dx}\, dx\\ &= -\lim_{\varepsilon\to 0}\left(\int_0^{\frac{1}{2}-\varepsilon} \eta\frac{dv}{dx}\, dx + \int_{\frac{1}{2}+\varepsilon}^{1} \eta\frac{dv}{dx}\, dx\right);\end{aligned}$$

from the properties of Lebesgue integration. But, by integrating by parts

$$\begin{aligned}\int_0^{\frac{1}{2}-\varepsilon} \eta\frac{dv}{dx}\, dx &= \eta\left(\frac{1}{2}-\varepsilon\right) v\left(\frac{1}{2}-\varepsilon\right) - \int_0^{\frac{1}{2}-\varepsilon} 2v\, dx\\ \int_{\frac{1}{2}+\varepsilon}^{1} \eta\frac{dv}{dx}\, dx &= -\eta\left(\frac{1}{2}+\varepsilon\right) v\left(\frac{1}{2}+\varepsilon\right) - \int_{\frac{1}{2}+\varepsilon}^{1} -2v\, dx\end{aligned}$$

since $v(0) = v(1) = 0$. Thus, as $\varepsilon \to 0$, we find

$$\langle D^1 L_\eta, v\rangle = \int_0^1 \tilde{D}^1_\eta v\, dx$$

where $\tilde{D}^1_\eta$ is a Heaviside-like function:

$$\tilde{D}^1_\eta = \begin{cases} 2 & [0, \frac{1}{2}] \\ -2 & (\frac{1}{2}, 1] \end{cases}, \qquad \tilde{D}^1_\eta \in L^2,$$

and we call it η', so $\eta \in \mathscr{H}^1$. Note that the value of $\tilde{D}^1_\eta$ at $x = 1/2$ is irrelevant since it will not affect the integral. Now we want to see if we can go furthere and find a second derivative. By definition it would be

$$\begin{aligned}\langle D^2 L_\eta, v\rangle &= \int_0^1 \eta\frac{d^2v}{dx^2}\, dx\\ &= -\int_0^{\frac{1}{2}} 2\frac{dv}{dx}\, dx + \int_{\frac{1}{2}}^{1} 2\frac{dv}{dx}\, dx\\ &= -2v\left(\frac{1}{2}\right) - 2v\left(\frac{1}{2}\right)\\ &= -4v\left(\frac{1}{2}\right)\\ &= -4\left\langle L_{\delta_{\frac{1}{2}}}, v\right\rangle,\end{aligned}$$

but $L_{\delta_{1/2}}$ is not square integrable, so $\eta \notin \mathscr{H}^2$.

Now go back to the weak formulation of the Poisson problem: $a(u, v) = l(v)$, for any $v \in \mathscr{H}^1_0$, $u \in \mathscr{H}^1_0$, $l \in \mathscr{H}^{-1}$. We would like to use Lax-Milgram theorem, so we need the coercivity and the continuity (i.e. boundedness) of the bilinear form. Both conditions are satisfied for the Poisson equation. It is then simple to prove stability: since $u \in \mathscr{H}^1_0$, taking $u = v$ gives

$$\alpha\|u\|^2_{\mathscr{H}^1} \le a(u, u) = l(u).$$

So

$$\begin{aligned}\alpha\|u\|_{\mathscr{H}^1} &\le \frac{l(u)}{\|u\|_{\mathscr{H}^1}}\\ &\le \sup_{v\in\mathscr{H}^1_0} \frac{l(v)}{\|v\|_{\mathscr{H}^1}}\\ &= \|l\|_{\mathscr{H}^{-1}},\end{aligned}$$

thus $C = \alpha^{-1}$.

Theorem 1.5 (Lax-Milgram)**.** *Let V be an Hilbert space and $a(\cdot,\cdot)$ a bilinear form on V which is bounded ($a(u,v) \leq C\|u\|\|v\|$) and coercive ($a(u,u) \geq \alpha\|u\|^2$). Then for any $f \in V'$ there exists a unique $u \in V$ solution to $a(u,v) = f(v)$ and it holds $\|u\| \leq {}^1\!/_C \|f\|_{V'}$.*

Remark. If u exists, it is unique. Indeed, take u_1, u_2 solutions, then

$$\begin{array}{rcll} a(u_1,v) & = & l(v) & \\ & & & - \\ a(u_2,v) & = & l(v) & \\ \hline & & & \\ a(u_1-u_2,v) & = & 0 & \end{array}$$

Take a particular $v = u_1 - u_2$, then $a(u_1-u_2, u_1-u_2) = 0$. Since a is coercive, $a(u_1-u_2,u_1-u_2) \geq \alpha\|u_1-u_2\|^2_{\mathscr{H}^1}$, so $\|u_1-u_2\|_{\mathscr{H}^1} = 0$, hence $u_1 = u_2$.

Cinsider now

$$\begin{cases} -u_{xx} = f & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

The weak formulation is: find $u \in \mathscr{H}^1_0(0,1)$ such that $\int_0^1 u_x v_x \, dx = \int_0^1 f v \, dx$ for any $v \in \mathscr{H}^1_0(0,1)$. The left hand side is

$$\begin{aligned} \left\langle -D^2 L_u, v \right\rangle &= \langle -Lu, v_{xx} \rangle \\ &= -\int_0^1 u v_{xx} \, dx \\ &= -[u v_x]_0^1 + \int_0^1 u_x v_x \, dx. \end{aligned}$$

The right hand side can be written as $\left\langle L_f, v \right\rangle$. So

$$\left\langle -D^2 L_u, v \right\rangle = \left\langle L_f, v \right\rangle \Rightarrow -D^2 L_u = L_f$$

hence $-u_{xx} = f$. This means that *the strong formulation is the distributional equivalent of the weak formulation.*

Example. Consider the problem

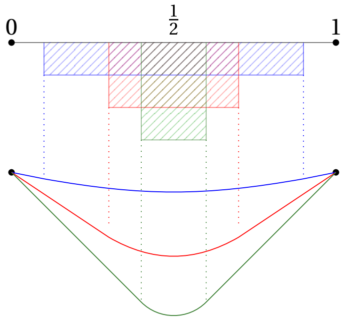
$$\begin{cases} -u_{xx} = f_n & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

where

$$f_n = -2n\chi_{B\left(\frac{1}{2},\frac{1}{n}\right)}(x) \in L^2(0,1).$$

The line-parabola-line solution has not continuous second derivative:

$$u_n(x) = \begin{cases} -2nx & x \leq \frac{1}{2} - \frac{1}{n} \\ n\left(x - \frac{1}{2}\right)^2 + \frac{1}{n} - 1 & x \in B\left(\frac{1}{2}, \frac{1}{n}\right) \\ -2n(1-x) & x \geq \frac{1}{2} + \frac{1}{n} \end{cases}$$



What about hte limit case? There is no convergence in L^2, but in $\mathscr{H}^{-1}$ (the delta function)

$$\left\langle L_{f_n}, v \right\rangle = \int_0^1 f_n v.$$

Now we can consider the limit

$$\begin{aligned}\lim_{n\to\infty}\langle L_{f_n}, v\rangle = \lim_{n\to\infty}\int_{\frac{1}{2}-\frac{1}{n}}^{\frac{1}{2}+\frac{1}{n}} -2nv\,dx \\ &= \lim_{n\to\infty} -2n\left(V\left(\frac{1}{2}+\frac{1}{n}\right)-V\left(\frac{1}{2}-\frac{1}{n}\right)\right) \\ &= -4\lim_{n\to\infty}\frac{V\left(\frac{1}{2}+\frac{1}{n}\right)-V\left(\frac{1}{2}-\frac{1}{n}\right)}{\frac{2}{n}} \\ &= 4V'\left(\frac{1}{2}\right) \\ &= 4v\left(\frac{1}{2}\right) \\ &= \left\langle L_{\delta_{\frac{1}{2}}}, v\right\rangle\end{aligned}$$

where V is the primitive of v. So there is no limit for f_n in $\mathscr{H}^2$, but there is in $\mathscr{H}_0^1$. Then

$$\lim_{n\to\infty} u_n(x) = \begin{cases} -2x & x<\frac{1}{2} \\ 2(x-1) & x>\frac{1}{2}\end{cases} \in \mathscr{H}_0^1\setminus\mathscr{H}^2,$$

but $u_n \in \mathscr{H}^2$, so we can take the limit, but only in a weak sense.

1.2 The Neumann problem

The strong formulation is: find u such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma^D \\ \frac{\partial u}{\partial n} = g & \text{on } \Gamma^N \end{cases} \tag{1.5}$$

where $\overline{\Gamma} = \overline{\Gamma^D} \cup \overline{\Gamma^N}$, with Γ^D is non empty. The condition $\gamma^D \neq \emptyset$ is necessary to have unicity of the solution. Indeed, if u was solution of

$$\begin{cases} u_{xx} = f & \text{in } \Omega \\ u_x = g & \text{on } \partial\Omega \end{cases},$$

then also $u + c$ would be a solution, for any $c \in \mathbb{R}$. Moreover, if $\Gamma^D = \emptyset$, the Neumann problem could have no solution. In particular, the equation tells us

$$-\int_\Omega \Delta u\, dA = \int_\Gamma -\nabla u\cdot \hat{n}\, dS = \int_\Omega f\, dA,$$

but the boundary condition tells us that

$$\int_\Gamma \partial_n u\, dS = \int_\Gamma \nabla u\cdot]hatn\, dS = \int_\Gamma g\, dS.$$

So we have the *solvability condition*

$$\int_\Omega f\, dA + \int_\Gamma g\, dS = 0.$$

If it is satisfied, our problem will have a solution, but not just one (the will be infinite, and all of them will differ by a constant).
As in the Dirichlet problem, we will solve a minimization statement

$$u = \arg\min_{w\in X} J(w)$$

with

$$X = \left\{ v \in \mathscr{H}^1(\Omega) : \, v|_{\Gamma^D} = 0 \right\}$$

and

$$J(w) = \frac{1}{2}\int_\Omega \nabla w \cdot \nabla w \, dA - \int_\Omega f w \, dA - \int_{\Gamma^N} g w \, dS$$

Remark. In X we don't put Neumann conditons because otherwise it wouldn't be a linear space.

Theorem 1.6. *If u is a strong solution to* (1.5)*, then u satisfies* (1.6).

Proof. Let $w = u + v$, then

$$\begin{aligned} J(u+v) &= \frac{1}{2}\int_\Omega \nabla(u+v)\cdot\nabla(u+v) \, dA - \int_\Omega f(u+v) \, dA - \int_{\gamma^N} g(u+v) \, dS \\ &= \frac{1}{2}\int_\Omega \nabla u \cdot \nabla v \, dA - \int_\Omega f u \, dA - \int_{\Gamma^N} g u \, dS + \underbrace{\int_\Omega \nabla u \cdot \nabla v \, dA - \int_\Omega f v \, dA - \int_{\Gamma^N} g v \, dS}_{\delta J_v(u)} + \frac{1}{2}\int_\Omega \nabla v \cdot \nabla v \, dA. \end{aligned}$$

Now

$$\begin{aligned} \delta J_v(u) &= \int_\Omega \nabla\cdot(v \nabla u) \, dA - \int_\Omega v \Delta u \, dA - \int_\Omega f v \, dA - \int_{\Gamma^N} g v \, dS \\ &= \int_{\Gamma^D} v \nabla u \cdot \hat{n} \, dS + \int_\Omega v(-\Delta u - f) \, dA - \int_{\Gamma^N} g v \, dS, \\ &= 0, \qquad \forall v \in X, \end{aligned}$$

whence

$$J(u+v) = J(u) + \frac{1}{2}\int_\Omega \nabla v \cdot]nablav \, dA \geq J(u) \qquad\qquad \forall v \in X,$$

and equality holds if and only if $v \equiv 0$. □

Talking about weak formulation, we want to find $u \in X$ such that $\delta J_v(u) = 0$ for any $v \in X$, i.e.

$$\int_\Omega \nabla u \cdot \nabla v \, dA = \int_\Omega f v \, dA + \int_{\Gamma^N} g v \, dS, \qquad\qquad \forall v \in X.$$

See how to recover quickly the weak formulation from the strong one:

$$\begin{aligned} -\Delta u = f &\to \int_\Omega -\Delta u v = \int_\Omega f v \\ &\to \int_\Gamma (-\nabla u \cdot \hat{n}) v + \int_\Omega \nabla u \cdot \nabla y = \int_\Omega f v \\ &= \int_{\Gamma^N} -g v + \int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v. \end{aligned}$$

Remark. Since $u \in X \subset \mathscr{H}^1$, $\nabla u \in L^2$, we cannot really impose $\nabla u \cdot \hat{n} = g$ in a strong sense (the boundary limit of an L^2 function makes little sense since jumps are permitted and individual points are ignored).

1.3 Inhomogeneous Dirichlet conditions

In strong formulation, the problem is: find u such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = u^D & \text{on } \Gamma = \Gamma^D \end{cases}$$

The boundary data u^D must satisfies certain regularity conditions on Γ^D. In fact u^D must be a little more than $L^2\left(\Gamma^D\right)$, but need not be quite as much as $\mathscr{H}^1\left(\Gamma^D\right)$ (discontinuites should be avoided). So we require $u^D \in \mathscr{H}^{1/2}$.

The minimization statement is: find

$$u = \arg\min_{w \in X^D} J(w)$$

where

$$X^D = \left\{v \in \mathscr{H}^1(\Omega) : v|_{\Gamma^D} = u^D\right\}$$

is not a linear space, but the difference of any two members in X^D is a member of

$$X = \left\{v \in \mathscr{H}^1(\Omega) : v|_{\Gamma^D} = 0\right\},$$

which is a linear space. The functional J is

$$J(w) = \frac{1}{2}\underbrace{\int_\Omega \nabla w \cdot \nabla w \, dA}_{a(w,w)} - \underbrace{\int_\Omega f w \, dA}_{l(w)}.$$

the weak formulation is: find $u \in X^D$ such that $\delta J_v(u) = 0$ for any $v \in X = \mathscr{H}_0^1(\Omega)$, i.e.

$$\underbrace{\int_\Omega \nabla u \cdot \nabla v \, dA}_{a(u,v)} = \underbrace{\int_\Omega f v \, dA}_{l(v)}, \qquad \forall v \in X.$$

Chapter 2

Discretization in $\mathbb{R}^1$

Let us consider the Poisson problem with homogeneous Dirichlet boundary conditions in $\Omega = (0,1)$:

$$\begin{cases} -u_{xx} = f & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

for given f.
The space is $X = \mathscr{H}_0^1(\Omega)$, the "energy" function is

$$J(w) = \frac{1}{2}\int_0^1 w_x^2\,dx - \int_0^1 f w\,dx,$$

while the bilinear and linear forms are

$$a(u,v) = \int_0^1 u_x v_x\,dx, \qquad l(u) = \int_0^1 f u\,dx,$$

so that

$$J(w) = \frac{1}{2}a(w,w) - l(w).$$

We now talk about regularity.
If $l \in \mathscr{H}^{-1}(\Omega)$,

$$\|u\|_{\mathscr{H}^1(\Omega)} \le C\,\|l\|_{\mathscr{H}^{-1}(\Omega)}\,.$$

If $l \in L^2(\Omega)$, $l(v) = \int_0^1 f v\,dx$, then

$$\|u\|_{\mathscr{H}^2(\Omega)} \le C_0\,\|f\|_{L^2(\Omega)}$$

where, we recall

$$\|v\|^2_{\mathscr{H}^2} = |v|^2_{\mathscr{H}^2} + \|v\|^2_{\mathscr{H}^1} = \int_0^1 \left(v_{xx}^2 + v_x^2 + v^2\right)dx.$$

Morally, we need more regularity in order to approximate solutions (as for finite differences of second order) we needed $\mathscr{C}^4$).
So we have

$$\|u\|^2_{\mathscr{H}^1} \le C\,\|l\|^2_{\mathscr{H}^{-1}} \le C\,\|f\|^2_{L^2},$$

where

$$\|l\|_{\mathscr{H}^{-1}} = \sup_{v\in\mathscr{H}_0^1} \frac{l(v)}{\|v\|_{\mathscr{H}^1}} \le \|l\|_{(L^2)'} = \sup_{v\in L^2} \frac{|l(v)|}{\|v\|_{L^2}}.$$

By Riesz, the last quantity is equal to $\|f\|_{L^2}$. We are solving $-u_{xx} = f$, which means

$$|u|_{\mathscr{H}^2} \le \|f\|_{L^2} \Leftrightarrow \sqrt{\int |u_{xx}|^2} \le \sqrt{\int f^2}.$$

Rewriting

$$\|u\|_{\mathcal{H}^2}^2 = \|u\|_{\mathcal{H}^1}^2 + |u|_{\mathcal{H}^2}^2 \le \left(1 - C^2\right) \|f\|_{L^2}^2 .$$

For the Neumann problem, we have to solve

$$\begin{cases} -u_{xx} = f & \text{in } \Omega = (0,1) \\ u(0) = 0 \\ u_x(1) = g \end{cases}$$

for given f and g.
for the minimization statement, we want to find

$$u = \arg\min_{w \in X} J(w)$$

where J is

$$J(w) = \frac{1}{2}\int_0^1 w_x^2 \, dx - \int_0^1 f w \, dx - g\, w(1)$$

and the space X is

$$X = \left\{ w \in \mathcal{H}^1(\Omega) : v(0) = 0 \right\}.$$

The weak formulation is: find $u \in X$ such that $\delta J_v(u) = 0$ for any $v \in X$, i.e.

$$\int_0^1 u_x v_x \, dx = \int_0^1 f v \, dx + g\, v(1), \qquad \forall v \in X$$

Defining the bilinear form and linear form as

$$a(w, v) = \int_0^1 w_x v_x \, dx, \qquad l(v) = \int_0^1 f v \, dx + g\, v(1),$$

the minimization statement and weak formulation are

$$u = \arg\min_{w \in X} \frac{1}{2} a(w, w) - l(w)$$

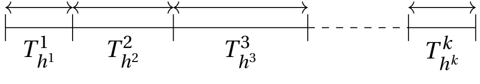
and

$$a(u, v) = l(v), \qquad \forall v \in X$$

respectively.

2.1 Rayleigh-Ritz approach

A small warning is needed at this stage: the notation may seem a bit cumbersome at a first sight, but this is needed for easily generalising it in higher dimension spaces.



An interval is denoted by $T^k_{h^k}$ and it is called *element* (the T stays for "triangle") and h^k is the diameter (in 1D, the length of the interval, in 2D the longest edge of the triangle). The union of all the triangles is called *mesh* and it is denoted by $\mathcal{T}_h$.
We will not accept all kinds of discretization. If we limit the ratio $h_{\min}/h_{\text{MAX}}$ we say it is *bounded from below when* $h \to 0$. There can also be a limit on the ratio radius/diameter (this permit to avoid to have too small minimum angle of the discretization). With abbreviate notation, we define

$$\overline{\Omega} = \bigcup_{T_h \in \mathcal{T}_h} \overline{T_h}.$$

We introduce the space

$$X_h = \left\{ v \in X = \mathscr{H}_0^1 : v|_{T_h^k} \in \mathbb{P}_1\left(T_h^k\right), k = 1, 2, \dots, K \right\}$$

where $\mathbb{P}_1(A)$ is the space of linear[1] polynomials over A. It can be, equivalently, defined as

$$X_h = \left\{ v \in X = \mathscr{H}_0^1 : v|_{T_h^k} \in \mathbb{P}_1\left(T_h^k\right), \forall T_h^k \in \mathscr{T}_h \right\}.$$

Definition 2.1. Given an interval $[0,1]$ and a discretization $\mathscr{T}_n$ of $n+2$ points $\{x_i\}_{i=0}^{n+1}$, we say it is a *quasi-uniform discretization* if

$$(n+1)h_{\min} > \varepsilon > 0,$$

where $h_{\min} = \min_{0 \le i \le n+1} \{x_{i+1} - x_i\}$ is the length of the shortest interval, and ε does not depend on n.

Example. The uniform discretization $x_i = i/(n+1)$ is, trivially, quasi-uniform.

Example. A parabolic discretization

$$x_i = \left(\frac{i}{n+1}\right)^2$$

is not quasi-uniform since

$$(n+1)h_{\min} = \frac{n+1}{(n+1)^2} = (n+1)^{-1}.$$

Example. A Čebyšëv-like discretization

$$x_i = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{i\pi}{n+1}\right)$$

is not quasi-uniform since from

$$(n+1)\frac{-\cos\left(\frac{i\pi}{n+1}\right) + \cos\left(\frac{(i-1)\pi}{n+1}\right)}{2} = (n+1)\sin\left(\frac{(2i-1)\pi}{2(n+1)}\right)\sin\left(\frac{\pi}{2(n+1)}\right)$$

follows

$$(n+1)h_{\min} = (n+1)\sin^2\left(\frac{\pi}{2(n+1)}\right).$$

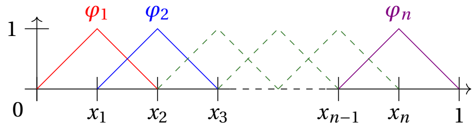
Example. Consider a discretization given by $n+1$ intervals of length $h_0, h_0 r, h_0 r^2, \dots, h_0 r^n$ with $r > 1$. We find that

$$h_0 = \frac{r-1}{r^{n+1}-1}$$

is the minimum interval and $(n+1)h_0$ is not bounded from below.

Consider now X_h such that $\dim(X_h) = n$. The *nodal basis* is the set $\{\varphi_j\}$ of hat functions, where φ_j is non-zero only on $\overline{T_h^i} \cup \overline{T_h^{i+1}}$.
Let $v \in X_h$, then

$$v(x) = \sum_{i=1}^{n} v_i \varphi_i(x).$$



[1] It should be more precise to say "affine", but we will continue with this, more intuitive, terminology.

Now consider $w_h \in X_h$, we look for w_h which minimizes J. Of course $J(w_h) \geq J(u)$ since u is the unique minimizer. Write

$$\begin{aligned}
J(w_h) &= J\left(\sum_{j=1}^{n} w_j\varphi_j(x)\right)\\
&= \frac{1}{2}a\left(\sum_{j=1}^{n} w_j\varphi_j(x), \sum_{j=1}^{n} w_j\varphi_j(x)\right) - l\left(\sum_{j=1}^{n} w_j\varphi_j(x)\right)\\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j a\big(\varphi_i(x),\varphi_j(x)\big) - \sum_{i=1}^{n} w_i l\big(\varphi_i(x)\big)\\
&= \frac{1}{2}\mathbf{w}^T A\mathbf{w} - \mathbf{w}^T F\\
&= J^{\mathbb{R}}(\mathbf{w}),
\end{aligned}$$

where $J^{\mathbb{R}} : \mathbb{R}^n \to \mathbb{R}$ (while the domain of J is X or X_h), and

$$A = \left[a_{ij}\right] = \left[a\big(\varphi_i(x),\varphi_j(x)\big)\right], \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}, \qquad F = \begin{bmatrix} l\big(\varphi_1(x)\big) \\ \vdots \\ l\big(\varphi_n(x)\big) \end{bmatrix}.$$

Remark. A is SPD.

Now we want to minimize $J^{\mathbb{R}}$ in $\mathbb{R}^n$ (it is equivalent to minimize J in X_h). So we want to find

$$\mathbf{u}_h = \arg\min_{\mathbf{w}\in\mathbb{R}^n} J^{\mathbb{R}}(\mathbf{w}),$$

where, by notation, $\mathbf{u}_h = \left[\mathbf{u}_{h_1},\ldots,\mathbf{u}_{h_n}\right]^T$ is a vector, while $u_h(x) = \sum_j u_{h_j}\varphi_j(x)$ is a function. Now, by using the fact that A is SPD,

$$\begin{aligned}
J^{\mathbb{R}}(\mathbf{w}) &= J^{\mathbb{R}}(\mathbf{u}_h + \mathbf{v})\\
&= \frac{1}{2}\mathbf{u}_h^T A\mathbf{u}_h - \mathbf{u}_h^T F + \frac{1}{2}\mathbf{v}^T A\mathbf{u}_h + \frac{1}{2}\mathbf{u}_h A\mathbf{v} - \mathbf{v}^T F + \frac{1}{2}\mathbf{v}^T A\mathbf{v}\\
&= \frac{1}{2}\mathbf{u}_h^T A\mathbf{u}_h - \mathbf{u}_h^T F + \mathbf{v}^T A\mathbf{u}_h - \mathbf{v}^T F + \frac{1}{2}\mathbf{v}^T A\mathbf{v}.
\end{aligned}$$

So $\mathbf{u}_h$ is the minimizer if and only if $\mathbf{v}^T A\mathbf{u}_h - \mathbf{v}^T F = 0$, i.e. $A\mathbf{u}_h - F = 0$ if and only if $J^{\mathbb{R}}(\mathbf{u}_h + \mathbf{v}) \geq J^{\mathbb{R}}(\mathbf{u}_h)$, so we just need to solve a linear system.

Remark. If A is SPD, then there is aunique solution.

This approach (minimization in a finite dimensional subspace) is called *Rayleight-Ritz approach.*

2.2 Galerkin approach

The Galerkin approach is based on the weak formulation, which will exists even when the minimization statement does not.

Find $u_h \in X_h$ such that $a(u_h, v) = l(v)$ for any $v \in X_h$, i.e.

$$a\left(u_h, \sum_{i=1}^{n} v_i\varphi_i(x)\right) = l\left(\sum_{i=1}^{n} v_i\varphi_i(x)\right), \qquad \forall\mathbf{v}\in\mathbb{R}^n,$$

but $\mathbf{u}_h = \sum_j \mathbf{u}_{h_j}\varphi_j(x)$, so

$$a\left(\sum_j u_{h_j}\varphi_j, \sum_i v_i\varphi_i\right) = l\left(\sum_i v_i\varphi_i\right), \qquad \forall\mathbf{v}\in\mathbb{R}^n,$$

which can be written as

$$\mathbf{v}^T A\mathbf{u}_h = \mathbf{v}^T F, \quad \forall\mathbf{v}\in\mathbb{R}^n \qquad \Leftrightarrow \qquad A\mathbf{u}_h = F,$$

where A is the *stiffness matrix.*

2.2.1 Weighted residual techniques

Given some general operator $\mathscr{L}$ and the associated PDE $\mathscr{L}u = f$, a *weighted residual technique* looks for a $\hat{u} \in X_1 \subset X$ such that

$$\int_\Omega v\,(\mathscr{L}\hat{u} - f)\; dA = 0, \qquad\qquad \forall v \in X_2.$$

In particular, we no longer require $\mathscr{L}u = f$ in a pointwise sense, but rather in an integral sense relative to test functions v. We expect as X_1 and X_2 become richer, $\hat{u}$ should approach u. Many different procedures can be derived based on different choices of X_1, X_2 and their associated basis. In the Galerkin procedure, $X_1 = X_2$.

Remark. The weak formulation is more general than the minimization one: in this we need the symmetry of a, while Lax-Milgram doesn't require it.

2.3 Quadrature and Assembly in 1D

Let us consider the problem to approximate

$$\int_0^1 f(x)\varphi_i(x)\; dx = \int_{x_{i-1}}^{x_{i+1}} f(x)\varphi_i(x)\; dx$$

for an "inner" hat test function $\varphi_i(x)$. Let us suppose for simplicity that $h_i = h$, $i = 1,2,\dots,n-1$.

2.3.1 Interpolation

If we consider the approximation

$$f(x) \approx \sum_{j=1}^{n} f\left(x_j\right)\varphi_j(x)$$

then

$$\begin{aligned}
\int_{x_{i-1}}^{x_{i+1}} f(x)\varphi_i(x)\; dx &\approx \int_{x_{i-1}}^{x_{i+1}} \left(\sum_{j=i-1}^{i+1} f(x_j)\varphi_j(x)\right)\varphi_i(x)\; dx \\
&= \int_{x_{i-1}}^{x_i} \big(f(x_{i-1})\varphi_{i-1}(x) + f(x_i)\varphi_i(x)\big)\,\varphi_i(x)\; dx + \int_{x_i}^{x_{i+1}} \big(f(x_i)\varphi_i(x) + f(x_{i+1})\varphi_{i+1}(x)\big)\,\varphi_i(x)\; dx \\
&= \frac{h}{6} f\left(x_{i-1}\right) + \left(\frac{h}{3} + \frac{h}{3}\right) f\left(x_i\right) + \frac{h}{6} f\left(x_{i+1}\right),
\end{aligned}$$

where the previous coefficients are given by the so called *weights matrix*.

The error in the approximation of a function $f \in \mathscr{C}^2$ by piecewise linear polynomials is proportional to h^2. Therefore

$$\begin{aligned}
\int_{x_{i-1}}^{x_{i+1}} f(x)\varphi_i(x)\; dx &= \int_{x_{i-1}}^{x_{i+1}} \left(\sum_{j=1}^{m} f(x_j)\varphi_j(x) + \mathscr{O}(h^2)\right)\varphi_i(x)\; dx \\
&= \frac{h}{6} f(x_{i-1}) + \left(\frac{h}{3} + \frac{h}{3}\right) f(x_i) + \frac{h}{6} f(x_{i+1}) + \mathscr{O}(h^2) \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x)\; dx \\
&= \frac{h}{6} f(x_{i-1}) + \frac{2h}{3} f(x_i) + \frac{h}{6} f(x_{i+1}) + \mathscr{O}(h^2)h.
\end{aligned}$$

Thus, the global error is $\mathscr{O}(h^3)$.

2.3.2 Trapezoidal rule

We can approximate

$$\begin{aligned}\int_{x_{i-1}}^{x_{i+1}} f(x)\varphi_i(x)\,dx &= \int_{x_{i-1}}^{x_i} f(x)\varphi_i(x)\,dx + \int_{x_i}^{x_{i+1}} f(x)\varphi_i(x)\,dx \\ &= (0 + f(x_i))\frac{h}{2} + (f(x_i) + 0)\frac{h}{2}.\end{aligned}$$

The quadrature error for the first integral is

$$\frac{h^3}{12}(f(x)\varphi_i(x))''\Big|_{\xi_{i-1}} = \frac{h^3}{12}\left(f''(\xi_{i-1})\frac{\xi_{i-1} - x_{i-1}}{h} + 2\frac{f'(\xi_{i-1}}{h}\right)$$

where ξ_{i-1} is a point in (x_{i-1}, x_1). For the second integral, analogously,

$$\frac{h^3}{12}(f(x)\varphi_i(x))''\Big|_{\xi_i} = \frac{h^3}{12}\left(f''(\xi_i)\frac{x_{i+1} - \xi_i}{h} - 2\frac{f'(\xi_i}{h}\right)$$

where ξ_{i-1} is a point in (x_i, x_{i+1}). Their sum is

$$\begin{aligned}&\mathcal{O}(h^3) + \frac{h^3}{12}\left(\frac{2f'(\xi_{i-1})}{h} - \frac{2f'(\xi_i)}{h}\right) \\ &\qquad = \mathcal{O}(h^3) + \frac{h^3}{12}\frac{1}{h}\left(2f'(x_i) + 2f''(\eta_{i-1})(\xi_{i-1} - x_i) - 2f'(x_i) - 2f''(\eta_i)(\xi_i - x_i)\right) \\ &\qquad = \mathcal{O}(h^3).\end{aligned}$$

Therefore the global error is $\mathcal{O}(h^3)$ if $f \in \mathscr{C}^2$. The same considerations hold if we consider quasi-uniform meshes.
Let us compute the *mass matrix*. We have

$$\begin{aligned}M_{ij} &= \int_{x_{i-1}}^{x_{i+1}} \varphi_j(x)\varphi_i(x)\,dx \\ &= \int_{x_{i-1}}^{x_i} \varphi_j(x)\varphi_i(x)\,dx + \int_{x_i}^{x_{i+1}} \varphi_j(x)\varphi_i(x)\,dx \\ &= \begin{cases} \dfrac{h_{i-1}}{6} & \text{if } j = i-1 \\ \dfrac{h_{i-1} + h_i}{3} & \text{if } j = i \\ \dfrac{h_i}{6} & \text{if } j = i+1 \end{cases}\end{aligned}$$

If we try to approximate by the trapezoidal rule the computation of the mass matrix, we get

$$\begin{aligned}M_{ij} &= \int_{x_{i-1}}^{x_{i+1}} \varphi_j(x)\varphi_i(x)\,dx \\ &= \int_{x_{i-1}}^{x_i} \varphi_j(x)\varphi_i(x)\,dx + \int_{x_i}^{x_{i+1}} \varphi_j(x)\varphi_i(x)\,dx \\ &\approx \varphi_j(x_i)\frac{h_{i-1}}{2} + \varphi_j(x_i)\frac{h_i}{2} \\ &= \delta_i^j\frac{h_{i-1} + h_i}{2}.\end{aligned}$$

It is equivalent to the operation of *lumping*, that is to sum up all the elements of each row of the exact mass matrix.

2.3.3 Barycentric formula

Let us start with the midpoint rule to approximate

$$\int_{x_{i-1}}^{x_i} f(x)\varphi_i(x)\,dx + \int_{x_i}^{x_{i+1}} f(x)\varphi_i(x)\,dx \approx f\left(\frac{x_{i-1}+x_i}{2}\right)\frac{h}{2} + f\left(\frac{x_i+x_{i+1}}{2}\right)\frac{h}{2}.$$

With the same argument above, the error is $\mathscr{O}(h^3)$ if $f \in \mathscr{C}^2$. Then we substitute

$$\begin{aligned} f\left(\frac{x_{i-1}+x_i}{2}\right) &= \frac{f(x_{i-1})+f(x_i)}{2} + \mathscr{O}\left(h^2\right) = \bar{f}_{i-1} + \mathscr{O}\left(h^2\right), \\ f\left(\frac{x_i+x_{i+1}}{2}\right) &= \frac{f(x_i)+f(x_{i+1})}{2} + \mathscr{O}\left(h^2\right) = \bar{f}_i + \mathscr{O}\left(h^2\right), \end{aligned}$$

keeping the same $\mathscr{O}\left(h^3\right)$ global error. The same consideration holds if we consider quasi-uniform meshes. The final form can be written as

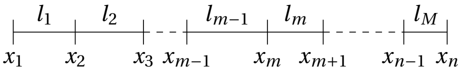
$$\begin{aligned} \int_{x_{i-1}}^{x_i} f(x)\varphi_i(x)\,dx + \int_{x_i}^{x_{i+1}} f(x)\varphi_i(x)\,dx &\approx \bar{f}_{i-1}\int_{x_{i-1}}^{x_i} \varphi_i(x)\,dx + \bar{f}_i \int_{x_i}^{x_{i+1}} \varphi_i(x)\,dx \\ &= \bar{f}_{i-1}\frac{h}{2} + \bar{f}_i\frac{h}{2}. \end{aligned}$$

2.3.4 Gauss-Legendre quadrature

Gauss-Legendre quadrature with one node coincides with the midpoint rule. With two nodes, the nodes in the interval $(-1,1)$ are $\pm\sqrt{1/3}$ with associated weights both equal to 1. The global error is $\mathscr{O}(h^5)$ if $f \in \mathscr{C}^4$. With the nodes 0 and $\pm\sqrt{3/5}$ the weights are $8/9$ and $5/9$.

2.4 Assembly

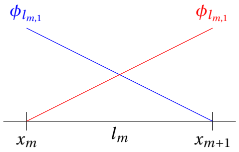
Let us see a general implementation strategy for FEM. Suppose we have M elements $(l_m)_{m=1}^M$ (in the 1D case, the intervals) with the associate points. With respect to the scheme, where $n = M+1$, we have,



$$l_{m,1} = m, \qquad l_{m,2}, \qquad\qquad 1 \le m \le M,$$

which means that *points x_m and x_{m+1} are associated to element l_m*.

The two basis functions which have value 1 on node $l_{m,k}$ and 0 on node $l_{m,3-k}$, for $k = 1,2$, have the form (on l_m)



$$\phi_{l_{m,1}}(x) = \frac{\alpha_{m,1}+\beta_{m,1}x}{\Delta_m} = \frac{\begin{vmatrix} 1 & 1 \\ x & x_{l_{m,2}} \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ x_{l_{m,1}} & x_{l_{m,2}} \end{vmatrix}} = \frac{x_{l_{m,2}} - x}{x_{l_{m,2}} - x_{l_{m,1}}},$$

$$\phi_{l_{m,2}}(x) = \frac{\alpha_{m,2}+\beta_{m,2}x}{\Delta_m} = \frac{\begin{vmatrix} 1 & 1 \\ x_{l_{m,1}} & x \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ x_{l_{m,1}} & x_{l_{m,2}} \end{vmatrix}} = \frac{x - x_{l_{m,1}}}{x_{l_{m,2}} - x_{l_{m,1}}},$$

obtaining

$$\phi_{l_{m,1}}(x) = \frac{x_{m+1}-x}{h_m}, \quad \phi_{l_{m,2}}(x) = \frac{x-x_m}{h_m},$$

(we mean $\varphi_{l_{m,k}}(x) \equiv \phi_{l_{m,k}}(x)$ on l_m). These will contribute to the elements $a_{l_{m,k},l_{m,k}}$ and $a_{l_{m,k},l_{m,3-k}}$ (and its symmetric) of the stiffness matrix

$$\begin{aligned} a_{l_{m,k},l_{m,k}} &= \int_0^1 \varphi'_{l_{m,k}}(x)\varphi'_{l_{m,k}}(x)\,dx, \\ a_{l_{m,k},l_{m,3-k}} &= \int_0^1 \varphi'_{l_{m,k}}(x)\varphi'_{l_{m,3-k}}(x)\,dx, \end{aligned} \qquad k = 1,2,$$

and to the elemnent $\tilde{f}_{l_{m,k}}$ of the right hand side: $\tilde{f}_{l_{m,k}}$ = approximation of $\int_0^1 f(x)\varphi_{l_{m,k}}\,dx$. In this way

$$\begin{aligned} a_{i,j} &= \sum_{\substack{l_{m,k}=i \\ l_{m,h}=j}} A_{l_{m,k},l_{m,h}}, \\ \tilde{f}_i &= \sum_{l_{m,k}=i} \tilde{F}_{l_{m,k}}, \end{aligned}$$

where

$$\begin{aligned} A_{l_{m,k},l_{m,h}} &= \int_{l_m} \phi'_{l_{m,k}}(x)\phi'_{l_{m,h}}(x)\,dx \\ \tilde{F}_{l_{m,k}} &\approx \int_{l_m} f(x)\phi_{l_{m,k}}(x)\,dx. \end{aligned}$$

Hence, for inner nodes,

$$\begin{aligned} a_{l_{m,1},l_{m,1}} &= A_{l_{m-1,2},l_{m-1,2}} + A_{l_{m,1},l_{m,1}} \\ &= \int_{l_{m-1}} \left(\frac{\beta_{m-1,2}}{\Delta_{m-1}}\right)^2 dx + \int_{l_m} \left(\frac{\beta_{m,1}}{\Delta_m}\right)^2 dx \\ &= \int_{l_{m-1}} \left(\frac{1}{\Delta_{m-1}}\right)^2 dx + \int_{l_m} \left(\frac{-1}{\Delta_m}\right)^2 dx \\ &= \frac{1}{\Delta_{m-1}} + \frac{1}{\Delta_m}, \end{aligned}$$

$$\begin{aligned} a_{l_{m,1},l_{m,2}} = a_{l_{m,2},l_{m,1}} &= A_{l_{m,1},l_{m,2}} \\ &= \int_{l_m} \frac{\beta_{m,1}}{\Delta_m}\frac{\beta_{m,2}}{\Delta_m}\,dx \\ &= \int_{l_m} -\frac{1}{\Delta_m}\frac{1}{\Delta_m}\,dx \\ &= -\frac{1}{\Delta_m}, \end{aligned}$$

$$\tilde{f}_{l_{m,1}} = \tilde{F}_{l_{m-1,2}} + \tilde{F}_{l_{m,1}}.$$

Hence, the assembly is done by

```
$a_{i,j} = 0$ ,  $\tilde{f}_i = 0$ ,  $1 \le i,j \le n$
FOR  $m = 1,\dots,M$
   FOR  $k = 1,\dots,2$
      $a_{l_{m,k},l_{m,k}} += 1/\Delta_m$
      $\tilde{f}_{l_{m,k}} += \tilde{F}_{l_{m,k}}$
      FOR  $i = k+1,\dots,2$
         $a_{l_{m,k},l_{m,i}} += -1/\Delta_m$
         $a_{l_{m,i},l_{m,k}} = a_{l_{m,k},l_{m,i}}$
      END
   END
END
```

2.4.1 Barycentric coordinates

Given the element l_m, it is possible to define its *barycentric coordinates* in this way: a point x in l_m is defined by the couple $\left(\lambda_{l_{m,1}}(x), \lambda_{l_{m,2}}(x)\right)$ such that

$$x = x_{l_{m,1}} \lambda_{l_{m,1}}(x) + x_{l_{m,2}} \lambda_{l_{m,2}}(x).$$

The coordinates $\lambda_{l_{m,k}}(x)$ satisfy $\lambda_{l_{m,k}}(x_{l_{m,h}}) = \delta_h^k$, $\lambda_{l_{m,1}}(x) + \lambda_{l_{m,2}}(x) \equiv 1$. Therefore

$$\lambda_{l_{m,k}}(x) = \varphi_{l_{m,k}}(x).$$

2.5 Projection

Given Hilbert spaces Y and $Z \subset Y$, the relation

$$\langle \Pi y, v \rangle_Y = \langle y, v \rangle_Y, \qquad \forall v \in Z,$$

defines the projection Πy of $y \in Y$ onto Z.

Example. Consider $Y = \mathbb{R}^n$, $Z = \langle w \rangle$, $w \in \mathbb{R}^n$, then

$$\Pi y = \frac{\langle y, w \rangle}{\langle w, w \rangle} w.$$

Note that

$$\langle \Pi y, \alpha w \rangle = \frac{\langle y, w \rangle}{\langle w, w \rangle} \langle w, \alpha w \rangle = \alpha \langle y, w \rangle = \langle y, \alpha w \rangle.$$

We now recall some basic facts on the projection.
The projection Πy minimizes $\|y - z\|_Y^2$ for any $z \in Z$. Indeed

$$\begin{aligned} \left\|y - (\Pi y + v)\right\|_Y^2 &= \langle (y - \Pi y) - v, (y - \Pi y) - v \rangle_Y \\ &= \|y - \Pi y\|_Y^2 - 2\langle y - \Pi y, v \rangle_Y + \|v\|_Y^2 \qquad \forall v \in Z. \end{aligned}$$

Note that $z = \Pi y + v \in Z$ and $\Pi y \in Z$ implies that $v \in Z$ and hence, since $\langle \Pi y, v \rangle_Y = \langle y, v \rangle_Y$ for any $v \in Z$ we have $\langle y - \Pi y, v \rangle_Y = 0$. This result states that

$$\|y - \Pi y\|_Y^2 < \|y - z\|_Y^2, \qquad \forall z \in Z \setminus \{\Pi y\},$$

so Πy *is the best approximation in Z of y in the $\|\cdot\|_Y$ norm.* We have also

$$\|\Pi y\| \leq \|y\|,$$

indeed $\langle \Pi y, v \rangle = \langle y, v \rangle$ for any $v \in Z$, in particular for $v = \Pi y$. Then

$$\|\Pi y\|^2 = \langle \Pi y, \Pi y \rangle = \langle y, \Pi y \rangle \leq \|y\| \, \|\Pi y\|$$

$$\Rightarrow \|\Pi y\| \leq \|y\|.$$

Moreover, we also have

$$\|y - \Pi y\| \leq \|y\|,$$

indeed $\|y - \Pi y\| \leq \|y - z\|$ for ay $z \in Z$, in particular for $z = 0$. Finally, we trivially have

$$\Pi(\Pi y) = \Pi y.$$

2.6 The interpolant

Recall

$$X_h = \left\{ v \in X : \, v|_{T_h} \in \mathbb{P}_1 \left(T_h\right), \forall T_h \in \mathscr{T}_h \right\}.$$

Given $w \in X$, we define the interpolant $\mathscr{I}_h w$ as

$$\mathscr{I}_h w(x) = \sum_{i=1}^{n} w\left(x_i\right) \phi_i(x).$$

It satisfies $\mathscr{I}_h w \in X_h$ and $\mathscr{I}_h w(x_i) = w(x_i)$, $i = 0, 1, \ldots, n+1$.

If $w \in \mathscr{H}_0^1(0,1)$ and $w|_{T_h^k} \in \mathscr{C}^2(T_h^k)$, for each $T_h^k \in \mathscr{T}_h$, then by Rolle's theorem

$$\begin{aligned}
\left| (w - \mathscr{I}_h w)'|_{T_h^k}(x) \right| &= \left| (w - \mathscr{I}_h w)'|_{T_h^k}(z_k) + \int_{z_k}^{x} (w - \mathscr{I}_h w)''|_{T_h^k}(t)\, dt \right| \\
&= \left| \int_{z_k}^{x} w''|_{T_h^k}(t)\, dt \right| \\
&\le h_k \max_{x \in T_h^k} \left| w''(x) \right|
\end{aligned}$$

where z_k is the point in which $(w - \mathscr{I}_h w)' = 0$ and h_k is the biggest of the length of the intervals of integration. We have

$$\left| (w - \mathscr{I}_h w)'(x) \big|_{T_h^k} \right| \le \int_{x_{k-1}}^{x_k} \left| w''(x) \right| dx,$$

which is true even if $w \in \mathscr{H}^1(T_h^k)$ since w is absolutely continuous. Hence

$$\begin{aligned}
|w - \mathscr{I}_h w|_{\mathscr{H}^1}^2 &= \sum_{k=1}^{K} \int_{T_h^k} \left| (w - \mathscr{I}_h w)'(x) \right|^2 dx \\
&\le \sum_{k=1}^{K} h_k \left(h_k \max_{x \in T_h^k} \left| w''(x) \right| \right)^2
\end{aligned}$$

If we take $h = \max_{1 \le k \le K} h_k$, we get

$$\begin{aligned}
\sum_{k=1}^{K} h_k \left(h_k \max_{x \in T_h^k} \left| w''(x) \right| \right)^2 &\le \left(h \max_{1 \le k \le K} \max_{x \in T_h^k} \left| w''(x) \right| \right)^2 \sum_{k=1}^{K} h_k \\
&= \left(h \max_{1 \le k \le K} \max_{x \in T_h^k} \left| w''(x) \right| \right)^2
\end{aligned}$$

and hence

$$|w - \mathscr{I}_h w|_{\mathscr{H}^1} \le h \max_{1 \le k \le K} \max_{x \in T_h^k} \left| w''(x) \right|. \tag{2.3}$$

We can also write

$$\begin{aligned}
(w - \mathscr{I}_h w)|_{T_h^k}(x) &= \int_{x_{k-1}}^{x_k} (w - \mathscr{I}_h w)'|_{T_h^k}(t)\, dt \\
&\le h_k h_k \max_{x \in T_h^k} \left| w''(x) \right|
\end{aligned}$$

and therefore

$$\begin{aligned}
\| w - \mathscr{I}_h w \|_{L^2}^2 &= \sum_{k=1}^{K} \int_{T_h^k} |(w - \mathscr{I}_h w)(x)|^2 dx \\
&\le \sum_{k=1}^{K} h_k \left(h_k^2 \max_{x \in T_h^k} \left| w''(x) \right| \right)^2,
\end{aligned}$$

whence

$$\|w - \mathscr{I}_h w\|_{L^2} \le h^2 \max 1 \le k \le K \max_{x \in T_h^k} \left| w''(x) \right|. \tag{2.4}$$

We now introduce the *broken space*

$$\mathscr{H}^2(\Omega, \mathscr{T}_h) = \left\{ w \in \mathscr{H}_0^1(\Omega) : \; w|_{T_h^k} \in \mathscr{H}^2\left(T_h^k\right), \forall k = 1, 2, \ldots, K \right\}$$

with the relative *broken seminorm* and *broken norm*

$$|w|^2_{\mathscr{H}^2(\Omega,\mathscr{T}_h)} = \sum_{k=1}^K |w|^2_{\mathscr{H}^2\left(T_h^k\right)},$$
$$\|w\|^2_{\mathscr{H}^2(\Omega,\mathscr{T}_h)} = \sum_{k=1}^K \|w\|^2_{\mathscr{H}^2\left(T_h^k\right)},$$

Remark. Belonging to $\mathscr{H}^2(\Omega, \mathscr{T}_h)$ depends on the triangulation.

We already have estimations (2.3) and (2.4). We now want some others estimates on the broken space $\mathscr{H}^2(\Omega, \mathscr{T}_h)$.

$$\begin{aligned} \left| (w - \mathscr{I}_h w)'|_{T_h^k} \right| &\le \int_{x_{k-1}}^{x_k} \left| w''(x) \right| dx \\ &\le \sqrt{\int_{x_{k-1}}^{x_k} 1^2} \sqrt{\int_{x_{k-1}}^{x_k} |w''|^2} \\ &\le h_k^{\frac{1}{2}} |w|_{\mathscr{H}^2\left(T_h^k\right)}. \end{aligned}$$

Therefore

$$\begin{aligned} \int_{x_{k-1}}^{x_k} \left| (w - \mathscr{I}_h w)' \right|^2 &\le \int_{x_{k-1}}^{x_k} \left(h_k \int_{x_{k-1}}^{x_k} \left| w'' \right|^2 \right) \\ &= h_k^2 \int_{x_{k-1}}^{x_k} \left| w'' \right|^2 \end{aligned}$$

hence

$$\begin{aligned} |w - \mathscr{I}_h w|_{\mathscr{H}^1} &\le \sqrt{\sum_{k=1}^K h_k^2 |w|^2_{\mathscr{H}^2}} \\ &\le h |w|_{\mathscr{H}^2(\Omega,\mathscr{T}_h)} \\ &\le h \|w\|_{\mathscr{H}^2(\Omega,\mathscr{T}_h)}. \end{aligned}$$

Moreover

$$(w - \mathscr{I}_h w)|_{T_h^k} = \int_{x_{k-1}}^{x_k} (w - \mathscr{I}_h w)'(t) \, dt.$$

Therefore

$$\begin{aligned} \left| (w - \mathscr{I}_h w)|_{T_h^k} \right| &\le \int_{x_{k-1}}^{x_k} \left| (w - \mathscr{I}_h w)' \right| \\ &\le h_k^{\frac{1}{2}} \sqrt{\int_{x_{k-1}}^{x_k} \left| (w - \mathscr{I}_h w)' \right|^2} \\ &\le h_k^{\frac{1}{2}} \sqrt{h_k^2 \int_{x_{k-1}}^{x_k} |w''|^2}. \end{aligned}$$

Hence

$$\int_{x_{k-1}}^{x_k} |w - \mathscr{I}_h w|^2 \, dx \le h_k h_k^3 \int_{x_{k-1}}^{x_k} \left|w''\right|^2 \, dx$$

and

$$\begin{aligned}
\| w - \mathscr{I}_h w \|_{L^2} &\le \sqrt{\sum_{k=1}^{K} h_k^4 |w|^2_{\mathscr{H}^2\left(T_h^k\right)}} \\
&\le h^2 |w|_{\mathscr{H}^2(\Omega,\mathscr{T}_h)} \\
&\le h^2 \, \| w \|_{\mathscr{H}^2(\Omega,\mathscr{T}_h)} .
\end{aligned}$$

Finally

$$\begin{aligned}
\| w - \mathscr{I}_h w \|_{\mathscr{H}^1(\Omega)} &\le \sqrt{\sum_{k=1}^{K} \left(h_k^2 + h_k^4\right) |w|^2_{\mathscr{H}^2(\Omega,\mathscr{T}_h)}} \\
&\le h\sqrt{1+h^2} |w|_{\mathscr{H}^2(\Omega,\mathscr{T}_h)} \\
&\le h\sqrt{1+h^2} \, \| w \|_{\mathscr{H}^2(\Omega,\mathscr{T}_h)} .
\end{aligned}$$

2.7 Energy norm

Definition 2.2. Let a be an SPD bilinear form. We define the *energy norm* (called also *a-norm*) $|||v|||$ as

$$|||v|||^2 = a(v,v).$$

Example. For the Poisson problem

$$|||v|||^2 = \int_0^1 v_x^2 \, dx = |v|^2_{\mathscr{H}^1(\Omega)} .$$

We want now to compute $|||u - u_h|||$. First of all, from the weak formulation,

$$\begin{array}{rclcl}
a(u,v) & = & l(v) & & \forall v \in X = \mathscr{H}_0^1 \\
 & & & - & \\
a(u_h,v) & = & l(v) & & \forall v \in X_h \\
\hline
a(u-u_h,v) & = & 0 & & \forall v \in X_h
\end{array}$$

Take an element $w_h \in X_h$. Since we are in a linear space, we can write $w_h = u_h + v_h$, hence

$$\begin{aligned}
a(u - w_h, u - w_h) &= a(u - u_h - v_h, u - u_h - v_h) \\
&= a(u - u_h, u - u_h) - 2a(u - u_h, v_h) + a(v_h, v_h) .
\end{aligned}$$

By orthogonality, $a(u - u_h, v_h) = 0$. Moreover $a(v_h, v_h) > 0$ for $v_h \neq 0$. Hence

$$\inf_{w_h \in X_h} |||u - w_h||| \ge |||u - u_h|||.$$

From this property, we have that $u_h = \Pi^a u$ is the orthogonal projection given by the bilinear form a. So we have the following fact:

$$l(v) = a\left(\Pi^a u, v\right) = a(u,v), \qquad \forall v \in X_h.$$

Remark. The fact that $u_h = \Pi^a u$ is true for any SPD bilinear form a, any boundary condition, any finite element space X_h and any dimension. For any particular SPD problem (i.e. any linear problem for which a in the weak formulation is SPD) the only thing that changes is the definition of the norm.

We know

$$|u - \mathcal{I}_h u|_{\mathcal{H}^1} \leq h \, \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)},$$

thus, in general,

$$|||u - u_h||| \leq |||u - \mathcal{I}_h u|||.$$

For the Poisson problem

$$|||u - \mathcal{I}_h u||| = |u - \mathcal{I}_h u|_{\mathcal{H}^1} \leq Ch \, \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h}.$$

Proposition 2.1. *For the Poisson problem, u_h is the interpolant, i.e.*

$$u_h = \mathcal{I}_h u.$$

Proof.

$$\begin{array}{rcll} a(u,\varphi_i) & = & l(\varphi_i) & \\ & & & - \\ a(u_h,\varphi_i) & = & l(\varphi_i) & \\ \hline \int_0^1 (u_h - u)'\varphi_i' & = & 0 & \end{array}$$

But

$$\begin{aligned} \int_0^1 (u_h - u)'\varphi_i' &= \sum_{k=1}^{K} \int_{T_h^k} (u_h - u)'\varphi_i' \\ &= \sum_{k=1}^{K} \left(\left[(u_h - u)\varphi_i' \right]_{x_{k-1}}^{x_k} - \int_{T_h^k} (u_h - u)\varphi_i'' \right) \\ &= \left[(u_h - u)\varphi_i' \right]_{x_{i-1}}^{x_i} + \left[(u_h - u)\varphi_i' \right]_{x_i}^{x_{i+1}}, \end{aligned}$$

where in the last step we used the fact that $\varphi_i'' = 0$ (since φ_i is piecewise linear) and that $\varphi_i(x) = 0$ if $x \notin (x_{i-1}, x_{i+1})$. The last right hand side can be written as

$$\frac{(u_h(x_i) - u(x_i)) - (u_h(x_{i-1}) - u(x_{i-1}))}{h_i} - \frac{(u_h(x_{i+1}) - u(x_{i+1})) - (u_h(x_i) - u(x_i))}{h_{i+1}}.$$

So we can now write

$$\underbrace{\begin{bmatrix} \ddots & \ddots & & & \\ \ddots & \ddots & \ddots & & \\ & \frac{1}{h_{i-1}} & \frac{1}{h_{i-1}} + \frac{1}{h_i} & \frac{1}{h_i} & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{bmatrix}}_{A} \begin{bmatrix} u_h(x_1) - u(x_1) \\ \vdots \\ \vdots \\ \vdots \\ u_h(x_n) - u(x_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

where A is the stiffness matrix, which is positive definite. Thus $u_h(x_i) = u(x_i)$ for any i. □

For any $w_h \in X_h$,

$$a(u - u_h, u - u_h) = a(u - u_h, u - w_h) + a(u - u_h, w_h - u_h) = a(u - u_h, u - w_h).$$

By coercivity and continuity of $a(\cdot,\cdot)$,

$$\begin{aligned} \alpha \, \|u - u_h\|_{\mathcal{H}^1}^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - w_h) \\ &\leq \beta \, \|u - u_h\|_{\mathcal{H}^1} \, \|u - w_h\|_{\mathcal{H}^1}, \end{aligned}$$

hence

$$\|u-u_h\|_{\mathcal{H}^1} \le \frac{\beta}{\alpha} \inf_{w_h\in X_h} \|u-w_h\|_{\mathcal{H}^1}, \qquad \beta > \alpha.$$

If a is SPD

$$a(u-u_h, u-u_h) \le \inf_{w_h\in X_h} a(u-w_h, u-w_h)$$

then

$$\alpha \|u-u_h\|^2_{\mathcal{H}^1} \le \inf_{w_h\in X_h} \beta \|u-w_h\|^2_{\mathcal{H}^1},$$

hence

$$\|u-u_h\|_{\mathcal{H}^1} \le \sqrt{\frac{\beta}{\alpha}} \inf_{w_h\in X_h} \|u-w_h\|_{\mathcal{H}^1} \le \frac{\beta}{\alpha} \inf_{w_h\in X_h} \|u-w_h\|_{\mathcal{H}^1}.$$

In the general case (a non symmetric)

$$\begin{aligned}\|u-u_h\|_{\mathcal{H}^1} &\le \frac{\beta}{\alpha} \inf_{w_h\in X_h} \|u-w_h\|_{\mathcal{H}^1} \\ &\le \frac{\beta}{\alpha} \|u-\mathcal{I}_h u\|_{\mathcal{H}^1} \\ &\le C\frac{\beta}{\alpha} h \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)}.\end{aligned}$$

We introduce an auxilary problem: find $\phi \in X = \mathcal{H}^1_0(\Omega)$ such that

$$a(v,\phi) = \int_0^1 ev\, dx,$$

where e is the error $u-u_h$. Set $v=e$:

$$\begin{aligned}\|e\|^2_{L^2} &= \int_0^1 e\,e\, dx \\ &= a(e,\phi) \\ &= a(e,\phi-\mathcal{I}_h\phi) \\ &\le \beta \|e\|_{\mathcal{H}^1} \left\|\phi-\mathcal{I}_h\phi\right\|_{\mathcal{H}^1} \\ &\le \beta \|e\|_{\mathcal{H}^1} Ch \left\|\phi\right\|_{\mathcal{H}^2} \\ &\le Ch \|e\|_{\mathcal{H}^1} \|e\|_{L^2}.\end{aligned}$$

Hence

$$\|e\|_{L^2} \le Ch \|e\|_{\mathcal{H}^1} \le Ch^2 \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)}.$$

2.8 Linear functionals

Definition 2.3. A *linear functional "output"* is defined by

$$s(u) = l^0(u) + c^0$$

where $l^0 : \mathcal{H}^1_0(\Omega) \to \mathbb{R}$ is a bounded linear functional.

Example.

$$l^0(v) = \int_D v\, dx, \qquad D \subset \Omega$$

Example.

$$l^0(v) = -\int_0^1 (1-x)_x v_x \, dx, \quad c^0 = \int_0^1 (1-x) f \, dx.$$

This last example is very useful since $s(u) = u_x(0)$ for the Poisson problem. Indeed

$$l^0(u) + c^0 = -\int_0^1 ((1-x)_x u_x - (1-x) f) \, dx = -[(1-x)u_x]_0^1 - \int_0^1 (1-x)(-u_{xx} - f) \, dx = u_x(0).$$

Remark. We can't just set $l^0(v) = v_x(0)$ because it wouldn't be bounded, while our choice is. Indeed

$$l^0(v) = -\int_0^1 (1-x)_x v_x \, dx = \int_0^1 v_x \, dx \leq \|v\|_{\mathcal{H}^1(\Omega)}.$$

We now see some general results

$$\begin{aligned} l^0 \in \mathcal{H}^{-1} &\Rightarrow l^0(e) \leq C \|e\|_{\mathcal{H}^1} \leq Ch \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)}, \\ l^0 \in (L^2)' &\Rightarrow l^0(e) \leq C \|e\|_{L^2} \leq Ch^2 \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)}. \end{aligned}$$

In fact, for any $l^0 \in \mathcal{H}^{-1}$,

$$\begin{aligned} \left| l^0(e) \right| &\leq C \|e\|_{\mathcal{H}^1} \left\| \psi - \psi_h \right\|_{\mathcal{H}^1} \\ &\leq Ch \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)} \, h \left\| \psi \right\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)} \\ &= Ch^2 \|u\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)} \left\| \psi \right\|_{\mathcal{H}^2(\Omega,\mathcal{T}_h)}, \end{aligned}$$

where

$$\begin{aligned} a(v,\psi) &= -l^0(v), & \forall v \in X, \\ a(v,\psi_h) &= -l^0(v), & \forall v \in X_h, \end{aligned}$$

and ψ is an adjoint, or dual, variable.

If we want to consider infinity norm, it is like to use the linear output functional

$$l^0(e) = \max_i |e(x_i)|.$$

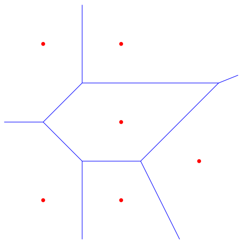
It is bounded: in fact

$$\begin{aligned} l^0(e) &= e(x_{\bar{\imath}}) \\ &= \int_0^{\bar{\imath}} e'(x) \\ &\leq \sqrt{\int_0^{\bar{\imath}} 1} \sqrt{\int_0^{\bar{\imath}} (e'(x))^2} \\ &= \sqrt{x_{\bar{\imath}}} \, |e|_{\mathcal{H}^1} \\ &\leq \|e\|_{\mathcal{H}^1}. \end{aligned}$$

Chapter 3

Triangulation

We start this Chapter by introducing some definition in order to fix the notation.



We consider the general situation in which we have a set of *sites* $S = \{s_1, s_2, \ldots, s_m\}$ in $\mathbb{R}^2$, assuming that no four sites are cocircular.

Definition 3.1 (circumcenter)**.** The *circumcenter* of a triangle is the point in which the three perpendicular bisectors meet. It is the center of the triangle's circumcircle.

Definition 3.2 (graph)**.** A *graph* is a set of sites and arcs (called edges) connecting some of the sites. A graph is said to be planar if it can be drawn without intersecting edges; and it is said to be *connected* if any pair of sites is connected by a finite sequence of edges.

Definition 3.3 (convex hull)**.** The *convex hull* of a set of sites is the smallest (in the sense of inclusion) convex set containing all the sites.

Definition 3.4 (planar straight line graph)**.** A *planar straight line graph* is a planar graph whose edges are segments.

Definition 3.5 (Voronoi polygon)**.** The bisectors between pairs of sites are straight lines that partition the plane into m convex regions, one corresponding to each site s_i. Each region is called *Voronoi polygon* of s_i: it is the set of points which are closer to s_i than to the remaining sites in S. The partition of the plane is called *Voronoi diagram* of S. The vertices and the edges of the convex regions are called *Voronoi vertices* and *Voronoi edges* respectively.

Proposition 3.1. *The number of Voronoi vertices is* $2(m-1)-h$ *and the number of Vorooi edges is* $3(m-1)-h$*, where h is the number of vertices of the convex hull of S.*

Proof. First of all, we consider a circle intersecting the unbounded edges of the Voronoi diagram and consider the resulting planar connected graph made of Voronoi vertices, Voronoi edges, and the additional edges for each unbounded Voronoi polygon. Of course, they correspond to the number of vertices in the convex hull of S, and therefore they are h. For this graph $2E = 3V$, where E is the number of edges, and V the number of vertices (there are three edges departing from any point and in this way each edge is counted twice). Then, using Euler's formula $V - E + F = 2$ (F is the number of faces, including the external one), we get $V = 2(F-2)$ and $E = 3(F-2)$. The number of Voronoi vertices is $V - h$ and the number of Voronoi edges is $E - h$. Since $F = m+1$, we have $2(m+1-2)-h$ Voronoi vertices and $3(m+1-2)-h$ Voronoi edges. □

Proposition 3.2. *The circle with center a Voronoi vertex and passing through the (three) sites that define its center has no other sites in its interior.*

Proof. By definition the center is the Voronoi vertex closer to the three sites. If another site would be inside the circle, then it would be the closest to the vertex, in contraddiction with its property. □

Definition 3.6. The *dual* of a Voronoi diagram is obtained by joining pairs of sites whose Voronoi polygon are adjacent.

Definition 3.7. A *triangulation* of sites is a set of straight line segment graph which intersect only at the sites ans so that every region internal to the convex hull is a trianlge.

Proposition 3.3. *The dual of a Voronoi diagram is a triangulation.*

It is called *Delaunay triangulation. It is unique if no four sites are cocircular.* Otherwise, the dual of a Voronoi diagram contains regions which are not triangle. In this case, any triangulation obtained by adding edges to the dual of a Voronoi diagram is a Delaunay triangulation. A Delaunay triangulation *maximizes the minimum angle of the triangles among all the triangulations* (more precisely, it has the least lexicography order of the angles).

Definition 3.8. A family of tringlations $\mathscr{T}_h$ is said to be *regular* if there exists a constant $\delta > 0$, independent of h, such that

$$\frac{h_k}{\rho_k} \le \delta, \qquad \forall T_h^k \in \mathscr{T}_h,$$

where h_k is the diameter of the triangle, and ρ_k is the radius.

Regularity excludes very deformed triangles, that is excessive small angles. In this sense, Delaunay triangulations are optimal.

Given ρ the radius of the inscribed circle in a triangle of edges a, b and c, area A and semiperimeter p, we have

$$A = \frac{(a+b+c)\rho}{2} = p\rho.$$

On the other hand, by Erone's formula

$$\rho = \frac{\sqrt{p(p-a)(p-b)(p-c)}}{p} = (p-a)\sqrt{\frac{(p-b)(p-c)}{p(p-a)}}.$$

Now, from

$$\sin\left(\frac{\alpha}{2}\right) = \sqrt{\frac{1-\cos(\alpha)}{2}}, \qquad \cos\left(\frac{\alpha}{2}\right) = \sqrt{\frac{1+\cos(\alpha)}{2}}, \qquad \cos(\alpha) = \frac{b^2+c^2-a^2}{2bc},$$

we get

$$\tan\left(\frac{\alpha}{2}\right) = \sqrt{\frac{(p-b)(p-c)}{p(p-a)}},$$

and therefore

$$\rho = (p-a)\tan\left(\frac{\alpha}{2}\right).$$

Then

$$\frac{h}{(p-a)\tan\left(\frac{\alpha}{2}\right)} \le \delta.$$

If the diameter h is a, the minimum value for $a/(p-a)$ is 2, attained for $a = b = c$. If not, $h/(p-a) \ge 2$. Therefore

$$\tan\left(\frac{\alpha}{2}\right) \ge \frac{2}{\delta}.$$

Since this reasoning is independent of the choice of the angle, we conclude that any angle has the same property.
A related question is: how many adjacent triangles are there? There are the three which share an edge with the given triangle. Then, any adjacent triangle insiste with an angle (bounded as seen) on one of the three vertices. Therefore, the number of adjacent tirangles is

$$n < \frac{5\pi}{2\arctan\left(\frac{2}{\delta}\right)} - 3.$$

The number 5π is the sum of external angles of a triangle, and -3 somes from the fact that the three triangles sharing an edge insist with two angles.

Definition 3.9. A *constrained Delaunay triangulation* of a planar straight line graph is a triangulation in which each edge of the graph is present as a simple edge of the triangulation. It is not truly a Delaunay triangulation.
A *conforming Delaunay triangulation* of a planar straight line graph is a true Delaunay triangulation in which each segment may have been subdivided into several edges by the insertion of additional vertices, called *Steiner points*. Steiner points are inserted also to meet constraints on the minimum angle and maximum triangle area.
A *constrained conforming Delaunay triangulation* of a planar straight line graph is a constrained Delaunay triangulation that includes Steiner points. It is not truly a Delaunay triangulation, but usually takes fewer vertices.

There are several algorithms (e.g. the *Fortune algorithm*) to construct a Delaunay triangulation. It holds the following.

Proposition 3.4. *The Delaunay triangulation of a set of m sites can be computed in* $\mathcal{O}(m\log(m))$ *operations, using* $\mathcal{O}(m)$ *storage.*

3.1 Quadrature in 2D

In two dimensions, an element l_m is a triangle, whose vertices are $l_{m,k}$, $k = 1,2,3$.

3.1.1 Trapezoidal rule

We approximate

$$\int_{l_m} g(x,y)\,dx\,dy \approx |\Delta_m| \frac{g\left(x_{l_{m,1}}, y_{l_{m,1}}\right) + g\left(x_{l_{m,2}}, y_{l_{m,2}}\right) + g\left(x_{l_{m,3}}, y_{l_{m,3}}\right)}{3}.$$

For the mass matrix, let us compute

$$\begin{aligned} M_{i,j} &= \sum_{\substack{l_{m,k}=i\\ l_{m,h}=j}} \int_{l_m} \phi_{l_{m,h}}(x,y)\phi_{l_{m,k}}(x,y)\,dx\,dy \\ &= \begin{cases} \sum_{l_{m,k}=i} \dfrac{|\Delta_m|}{6} & \text{if } i = j \\ \sum_{\substack{l_{m,k}=i\\ l_{m,h}=j}} \dfrac{|\Delta_m|}{12} & \text{if } i \neq j \end{cases} \end{aligned}$$

If we take the sum over j, we get

$$\sum_j M_{i,j} = \sum_{l_{m,k}=i} \frac{|\Delta_m|}{6} + 2\sum_{l_{m,k}=i} \frac{|\Delta_m|}{12} = \sum_{l_{m,k}=i} \frac{|\Delta_m|}{3}.$$

The factor 2 comes from the fact that if a triangle has vertices i and j, then there is another triangle with the same vertices. If we try to approximate by trapezoidal rule the computation of the mass matrix, we get

$$M_{i,i} \approx \sum_{l_{m,k}=i} \frac{|\Delta_m|}{3}, \qquad M_{i,j} \approx 0, \quad i \neq j.$$

It is equivalent to the operation of lumping (i.e. sum up all the elements of each row of the exact mass matrix).

3.1.2 Baricentric formulas

We approximate

$$\int_\Omega f(x,y)\varphi_i(x,y)\,dx\,dy \approx \sum_{l_{m,k}=i} \bar{f}_m \frac{|\Delta_m|}{3}$$

where

$$\bar{f}_m = \frac{f\left(x_{l_{m,1}}, y_{l_{m,1}}\right) + f\left(x_{l_{m,2}}, y_{l_{m,2}}\right) + f\left(x_{l_{m,3}}, y_{l_{m,3}}\right)}{3}.$$

3.1.3 Gauss-Legendre quadrature

the first Gauss-Legendre quadrature formula is

$$\int_{l_m} g(x,y)\,dx\,dy \approx |\Delta_m|\, g\left(\frac{x_{l_{m,1}} + x_{l_{m,2}} + x_{l_{m,3}}}{3}, \frac{y_{l_{m,1}} + y_{l_{m,2}} + y_{l_{m,3}}}{3}\right) = |\Delta_m|\, g\left(x_{l_m}, y_{l_m}\right),$$

which is exact for $g \in \mathbb{P}_1$. In fact, due to the property of the centroid

$$\left(\bar{x}_m, \bar{y}_m\right) = \left(x_{l_m}, y_{l_m}\right) = \frac{\int_{l_m} x\,dx\,dy + \int_{l_m} y\,dx\,dy}{|\Delta_m|}$$

(you can see it even if you apply the trapezoidal rule to the linear functions x and y). Now we have

$$g(x,y) = g\left(x_{l_m}, y_{l_m}\right) + \nabla g\left(x_{l_m}, y_{l_m}\right) \cdot \left((x,y) - \left(x_{l_m}, y_{l_m}\right)\right),$$

and therefore

$$\begin{aligned}
\int_{l_m} g(x,y)\,dx\,dy &= \int_{l_m} g\left(x_{l_m}, y_{l_m}\right)dx\,dy + \int_{l_m} \nabla g\left(x_{l_m}, y_{l_m}\right) \cdot \left((x,y) - \left(x_{l_m}, y_{l_m}\right)\right) \\
&= |\Delta_m|\, g\left(x_{l_m}, y_{l_m}\right) + \nabla g\left(x_{l_m}, y_{l_m}\right) \cdot \int_{l_m} \left((x,y) - \left(x_{l_m}, y_{l_m}\right)\right) dx\,dy \\
&= |\Delta_m|\, g\left(x_{l_m}, y_{l_m}\right) + \nabla g\left(x_{l_m}, y_{l_m}\right) \cdot \left(|\Delta_m| \left(x_{l_m}, y_{l_m}\right) - \int_{l_m} \left(x_{l_m}, y_{l_m}\right) dx\,dy\right) \\
&= |\Delta_m|\, g\left(x_{l_m}, y_{l_m}\right).
\end{aligned}$$

There exist higher order Gauss-Legendre quadrature formulas for triangles, involving three, seven or more points.

3.2 Assembly

The assembly in the two dimensional case is not much different from the one dimensional case. First of all, the number of points is m and the number of triangles is n. Then we consider the basis functions $\varphi_{l_{m,k}}$ which has value 1 on node $l_{m,k}$ and 0 on nodes $l_{m,h}$, $h \in \{1,2,3\} \setminus \{k\}$ of the triangle l_m. It has the form (on l_m)

$$\phi_{l_{m,k}}(x,y) = \frac{\alpha_{m,k} + \beta_{m,k}x + \gamma_{m,k}y}{2\Delta_m} = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x_{l_{m,1}} & x & x_{l_{m,3}} \\ y_{l_{m,1}} & y & y_{l_{m,3}} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_{l_{m,1}} & x_{l_{m,2}} & x_{l_{m,3}} \\ y_{l_{m,1}} & y_{l_{m,2}} & y_{l_{m,3}} \end{vmatrix}}$$

where Δ_m is the area (with sign) of the triangle l_m. We need to compute

$$\int_{l_m} \left(\frac{\partial \phi_{l_{m,k}}(x,y)}{\partial x} \frac{\partial \phi_{l_{m,h}}(x,y)}{\partial x} + \frac{\partial \phi_{l_{m,k}}(x,y)}{\partial y} \frac{\partial \phi_{l_{m,h}}(x,y)}{\partial y} \right) dx\, dy$$

for $h,k = 1,2,3$ for the stiffness matrix; and

$$\int_{l_m} f(x,y)\phi_{l_{m,k}}(x,y)\, dx\, dy$$

for the right hand side. We have

$$\int_{l_m} \frac{\partial \phi_{l_{m,k}}(x,y)}{\partial x} \frac{\partial \phi_{l_{m,h}}(x,y)}{\partial x}\, dx\, dy = \int_{l_m} \frac{\beta_{m,k}}{2\Delta_m} \frac{\beta_{m,h}}{2\Delta_m}\, dx\, dy = \frac{\beta_{m,k}\beta_{m,h}}{4|\Delta_m|},$$

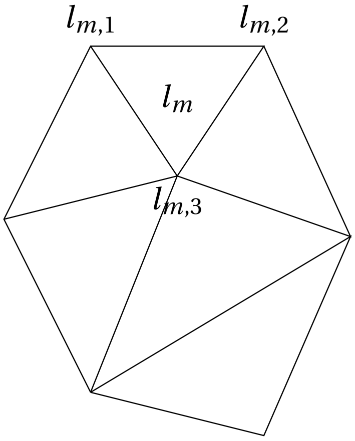
$$\int_{l_m} \frac{\partial \phi_{l_{m,k}}(x,y)}{\partial y} \frac{\partial \phi_{l_{m,h}}(x,y)}{\partial y}\, dx\, dy = \int_{l_m} \frac{\gamma_{m,k}}{2\Delta_m} \frac{\gamma_{m,h}}{2\Delta_m}\, dx\, dy = \frac{\gamma_{m,k}\gamma_{m,h}}{4|\Delta_m|},$$

and their sum corresponds to $A_{l_{m,k},l_{m,h}}$; and

$$\int_{l_m} f(x,y)\phi_{l_{m,k}}(x,y)\, dx\, dy \approx \tilde{F}_{l_{m,k}}.$$

The algorithm for the assembly is

```
$a_{ij} = 0$ , $\tilde{f}_i = 0$ , $1 \le i,j \le n$
FOR $m = 1,\ldots,M$
   FOR $k = 1,\ldots,3$
      $a_{l_{m,k},l_{m,k}} += \frac{\beta_{m,k}\beta_{m,k}}{4|\Delta_m|} + \gamma_{m,k}\gamma_{m,k}4|\Delta_m|$
      $\tilde{f}_{l_{m,k}} += \tilde{F}_{l_{m,k}}$
      FOR $h = k+1,\ldots,3$
         $a_{l_{m,k},l_{m,h}} += \frac{\beta_{m,k}\beta_{m,h}}{4|\Delta_m|} + \gamma_{m,k}\gamma_{m,h}4|\Delta_m|$
         $a_{l_{m,k},l_{m,h}} = a_{l_{m,h},l_{m,k}}$
      END
   END
END
```



3.2.1 Barycentric coordinates

The barycentric coordinates on element l_m are $\lambda_{l_{m,k}}(x,y) = \varphi_{l_{m,k}}(x,y)$, $k = 1,2,3$.

Chapter 4

Iterative methods for sparse linear systems

Given an Hilbert space H and subspaces M and L, the *projection* Px of $x \in H$ onto M orthogonally to L is defined by

$$Px \in M, \qquad \langle x - Px, y \rangle_H = 0, \qquad \forall y \in L.$$

If $L = M$, then P is called *orthogonal projection* and in this case the following is true

$$\arg\min_{y \in M} \|x - y\|_H = Px.$$

If the projection is not orthogonal, then it is called *oblique*.

Let us consider the linear system $Ax = b$ whose exact solution is denoted by $\bar{x} = x_0 + \bar{\delta}$.

Proposition 4.1. *If $A \in \mathbb{R}^{n\times n}$ is SPD, then a vector $\tilde{x}$ is the result of an orthogonal projection from $\mathbb{R}^n$ onto $\mathcal{K} \subset \mathbb{R}^n$ with the starting vector x_0, that is*

$$\begin{aligned} \tilde{x} &= x_0 + \tilde{\delta}, & \tilde{\delta} &\in \mathcal{K}, \\ \langle b - A\tilde{x}, \delta \rangle &= 0, & \forall \delta &\in \mathcal{K} \end{aligned}$$

if and only if

$$\tilde{x} = \arg\min_{x \in x_0 + \mathcal{K}} E(x),$$

where, given $x = x_0 + \delta$,

$$E(x) = \sqrt{\langle A(\bar{x} - x), \bar{x} - x \rangle} = \sqrt{\langle A(\bar{\delta} - \delta), \bar{\delta} - \delta \rangle}.$$

Proof. First of all, A can be written, using the Cholesky decomposition, as $A = R^T R$. If $\tilde{x}$ is the minimizer of E, we have

$$\begin{aligned} E(\tilde{x}) &= \min_{x \in x_0 + \mathcal{K}} E(x) \\ &= \min_{\delta \in \mathcal{K}} \sqrt{\langle A(\bar{\delta} - \delta), \bar{\delta} - \delta \rangle} \\ &= \min_{\delta \in \mathcal{K}} \sqrt{\langle R(\bar{\delta} - \delta), R(\bar{\delta} - \delta) \rangle} \\ &= \min_{\delta \in \mathcal{K}} \|R(\bar{\delta} - \delta)\|_2 \\ &= \min_{\delta \in \mathcal{K}} \|R\bar{\delta} - R\delta\|_2 \\ &= \min_{w \in R\mathcal{K}} \|R\bar{\delta} - w\|_2, \end{aligned}$$

which, by hypothesis, is taken by $\tilde{w} = R\tilde{\delta}$, where $\tilde{x} = x_0 + \tilde{\delta}$. But the minimum in $R\mathcal{K}$ is taken by the orthogonal projection of $R\bar{\delta}$ onto $R\mathcal{K}$ too. Therefore $\tilde{w}$ is such a projection and satisfies, for any $w = R\delta$, $\delta \in \mathcal{K}$,

$$\begin{aligned}
0 &= \left\langle R\bar{\delta} - \tilde{w}, w \right\rangle \\
&= \left\langle R(\bar{\delta} - \tilde{\delta}), w \right\rangle \\
&= \left\langle R(\bar{\delta} - \tilde{\delta}), R\delta \right\rangle \\
&= \left\langle A(\bar{\delta} - \tilde{\delta}), \delta \right\rangle \\
&= \langle A(\bar{x} - \tilde{x}), \delta \rangle \\
&= \langle b - A\tilde{x}, \delta \rangle .
\end{aligned}$$

If, on the contrary, $\tilde{x}$ is the result of an orthogonal projection, then the previous argument can be used starting from the end. □

In this case, $\tilde{\delta}$ is the orthogonal projection of $\bar{\delta}$ onto $\mathcal{K}$ through the scalar product $\langle \cdot, \cdot \rangle_A$. In fact, for any $\delta\ in\mathcal{K}$

$$\left\langle \bar{\delta} - \tilde{\delta}, \delta \right\rangle_A = \delta^T A(\bar{x} - \tilde{x}) = \langle b - A\tilde{x}, \delta \rangle = 0.$$

This is not true for $\tilde{x}$ and $\bar{x}$ since $\tilde{x} \notin \mathcal{K}$.

Proposition 4.2. *If A is non singular and $\mathcal{L} = A\mathcal{K}$, then a vector $\tilde{x}$ is the result of an oblique projection method onto $\mathcal{K}$ orthogonally to $\mathcal{L}$ with the starting vector x_0, that is*

$$\begin{aligned}
\tilde{x} &= x_0 + \tilde{\delta}, & \tilde{\delta} &\in \mathcal{K}, \\
\langle b - A\tilde{x}, w \rangle &= 0, & \forall\, w &\in \mathcal{L} = A\mathcal{K},
\end{aligned}$$

if and only if

$$\tilde{x} = \arg \min_{x \in x_0 + \mathcal{K}} R(x),$$

where, given $x = x_0 + \delta$

$$R(x) = \|b - Ax\|_2 = \sqrt{\langle b - Ax, b - Ax \rangle} = \sqrt{\langle A(\bar{x} - x), A(\bar{x} - x) \rangle} = \sqrt{\left\langle A(\bar{\delta} - \delta), A(\bar{\delta} - \delta) \right\rangle}.$$

Proof. We have

$$\begin{aligned}
R(\tilde{x}) &= \min_{x \in x_0 + \mathcal{K}} R(x) \\
&= \min_{\delta \in \mathcal{K}} \sqrt{\left\langle A(\bar{\delta} - \delta), A(\bar{\delta} - \delta) \right\rangle} \\
&= \min_{\delta \in \mathcal{K}} \left\| A(\bar{\delta} - \delta) \right\|_2 \\
&= \min_{\delta \in \mathcal{K}} \left\| A\bar{\delta} - A\delta \right\|_2 \\
&= \min_{w \in \mathcal{L}} \left\| A\bar{\delta} - w \right\|_2 ,
\end{aligned}$$

which is taken by $\tilde{w} = A\tilde{\delta}$, where $\tilde{x} = x_0 + \tilde{\delta}$. But the minimum in $A\mathcal{K} = \mathcal{L}$ is taken by the orthogonal projection of $A\bar{\delta}$ onto $\mathcal{L}$, too. Therefore $\tilde{w}$ is such a projection and satisfies, for any $w \in \mathcal{L}$,

$$\begin{aligned}
0 &= \left\langle A\bar{\delta} - \tilde{w}, w \right\rangle \\
&= \left\langle A(\bar{\delta} - \tilde{\delta}, w \right\rangle \\
&= \langle A(\bar{x} - \tilde{x}), w \rangle \\
&= \langle b - A\tilde{x}, w \rangle .
\end{aligned}$$

□

4.1 Conjugate gradient (C.G.) method

Given an SPD matrix A of dimension n, the idea is to solve $A\bar{x} = b$ by minimizing the quadratic functional

$$J(x) = x^T Ax - 2b^T x$$

whose gradient id $\nabla J(x) = 2Ax - 2b = -2r(x)$. If we introduce the error $e(x) = x - \bar{x}$, we have $r(x) = -Ae(x)$. Moreover, if we consider the functional

$$E(x) = e(x)^T Ae(x) = r(x)^T A^{-1} r(x),$$

we have $\nabla E(x) = \nabla J(x)$ and $E(x) \geq 0$, $E(\bar{x}) = 0$. So *the minimization of $J(x)$ is equivalent to the minimization of $E(x)$.*

Starting from an initial vector x_0, we can use a descent method to find a sequence

$$x_m = x_{m-1} + \alpha_{m-1} p_{m-1}$$

in such a way that $E(x_m) < E(x_{m-1})$.
Given p_{m-1}, we can compute an optimal α_{m-1} in such a way that

$$\alpha_{m-1} = \arg\min_{\alpha} E(x_{m-1} + \alpha p_{m-1}).$$

It is

$$E(x_{m-1} + \alpha p_{m-1}) = E(x_{m-1}) - 2\alpha p_{m-1}^T r_{m-1} + \alpha^2 p_{m-1}^T A p_{m-1}$$

so we can easily see that is nothing but a parabola in α. The minimum is therefore achieved at $-b/2a$. The minimum of $E(x_{m-1} + \alpha p_{m-1})$ is then

$$\alpha_{m-1} = \frac{p_{m-1}^T r_{m-1}}{p_{m-1}^T A p_{m-1}}.$$

Proposition 4.3. *If α_{m-1} is optimal, then*

$$r_m^T p_{m-1} = p_{m-1}^T r_m = 0. \tag{4.1}$$

Proof. First of all, we have

$$\begin{aligned} r_m &= b - Ax_m \\ &= b - A(x_{m-1} + \alpha_{m-1} p_{m-1}) \\ &= r_{m-1} - \alpha_{m-1} A p_{m-1} \end{aligned} \tag{4.2}$$

and then

$$\begin{aligned} r_m^T p_{m-1} &= r_{m-1}^T p_{m-1} - \alpha_{m-1} p_{m-1}^T A p_{m-1} \\ &= r_{m-1}^T p_{m-1} - p_{m-1}^T r_{m-1} \\ &= 0. \end{aligned}$$

□

The equation $E(x) = E(x_{m-1})$ is that of an ellipsoid passing through x_{m-1}, with r_{m-1} a vector orthogonal to the surface pointing inside.
Given p_{m-1} and α_{m-1}, we can compute x_m and r_m. Now, we are ready for the next direction p_m. It has to be “simple” to compute, so we may require

$$p_m = r_m + \beta_m p_{m-1} \tag{4.3}$$

with β_m to be found in such a way to have the maximum reduction of $E(x)$ starting from $E(x_m)$. Therefore

$$E(x_{m+1}) = E(x_m + \alpha_m p_m) = E(x_m) - 2\alpha_m p_m^T r_m + \alpha_m^2 p_m^T A p_m$$

and using the definition of α_m

$$E(x_{m+1}) = E(x_m)\left(1 - \frac{(p_m^T r_m)^2}{E(x_m)(p_m^T A p_m)}\right) = E(x_m)\left(1 - \frac{(p_m^T r_m)^2}{(r_m^T A^{-1} r_m)(p_m^T A p_m)}\right).$$

We observe that, by using (4.1)

$$p_m^T r_m = (r_m + \beta_m p_{m-1})^T r_m = r_m^T r_m,$$

and this relation holds always true if $p_0 = r_0$. Therefore, the only possibility to minimize $E(x_{m+1})$ is to take $p_m^T A p_m$ as small as possible, and hence, from

$$p_m^T A p_m = r_m^T A r_m + 2\beta_m r_m^T A p_{m-1} + \beta_m^2 p_{m-1}^T A p_{m-1}$$

we get

$$\beta_m = -\frac{r_m^T A p_{m-1}}{p_{m-1}^T A p_{m-1}}$$

With this choice we obtain

$$p_m^T A p_{m-1} = 0.$$

Using again (4.1) we get

$$p_{m-1}^T r_{m-1} = r_{m-1}^T r_{m-1} + \beta_{m-1} p_{m-2}^T r_{m-1} = r_{m-1}^T r_{m-1}$$

and therefore

$$\alpha_{m-1} = \frac{p_{m-1}^T r_{m-1}}{p_{m-1}^T A p_{m-1}} = \frac{r_{m-1}^T r_{m-1}}{p_{m-1}^T A p_{m-1}}.$$

Finally, from definition (4.3) for p_{m-1} we have

$$A p_{m-1} = A r_{m-1} + \beta_{m-1} A p_{m-2},$$

and therefore

$$p_{m-1}^T A p_{m-1} = p_{m-1}^T A r_{m-1} = r_{m-1}^T A p_{m-1}.$$

Taking expression (4.2) for r_m, if we multiply by r_{m-1}^T we get

$$r_m^T r_{m-1} = r_{m-1}^T r_m = r_{m-1}^T r_{m-1} - \underbrace{\frac{r_{m-1}^T r_{m-1}}{p_{m-1}^T A p_{m-1}}}_{\alpha_{m-1}} r_{m-1}^T A p_{m-1} = 0$$

and if we multiply by r_m^T we get

$$r_m^T r_m = r_m^T r_{m-1} - \frac{r_{m-1}^T r_{m-1}}{p_{m-1}^T A p_{m-1}} r_m^T A p_{m-1} = -r_{m-1}^T r_{m-1} \frac{r_m^T A p_{m-1}}{p_{m-1}^T A p_{m-1}} = r_{m-1}^T r_{m-1} \beta_m,$$

from which

$$\beta_m = \frac{r_m^T r_m}{r_{m-1}^T r_{m-1}}.$$

We have therefore the following implementation of the method, known as Hestenes-Stiefel

```
$x_0$ given ,  $p_0 = r_0 = b - Ax_0$
FOR  $m = 1,2,\ldots$  UNTIL  $\|r_{m-1}\|_2 \leq$  tol  $\|b\|_2$
    $w_{m-1} = Ap_{m-1}$
    $\alpha_{m-1} = \frac{r_{m-1}^T r_{m-1}}{p_{m-1}^T w_{m-1}}$
    $x_m = x_{m-1} + \alpha_{m-1} p_{m-1}$
    $r_m = r_{m-1} - \alpha_{m-1} w_{m-1}$
    $\beta_{m-1} = \frac{r_m^T r_m}{r_{m-1}^T r_{m-1}}$
    $p_m = r_m + \beta_m p_{m-1}$
END
```

We now see some properties.

Theorem 4.1. *For $m > 1$, if $r_i \neq 0$ for $0 \leq i \leq m-1$, then*

$$\begin{aligned} p_i^T r_{m-1} &= 0, && i < m-1, \\ p_i^T A p_{m-1} &= 0, && i < m-1, \\ r_i^T A r_{m-1} &= 0, && i < m-1 \\ \text{span}\,\{r_0, r_1, \ldots, r_{m-1}\} &= \text{span}\,\left\{r_0, Ar_0, \ldots, A^{m-1} r_0\right\} \\ \text{span}\,\left\{p_0, p_1, \ldots, p_{m-1}\right\} &= \text{span}\,\left\{r_0, Ar_0, \ldots, A^{m-1} r_0\right\} \end{aligned}$$

proof (sketch). The proof goes by induction and on the first step we need all the five equalities to be true. □

Definition 4.1. The space $\mathcal{K}_m = \text{span}\left(r_0, Ar_0, \ldots, A^{m-1} r_0\right)$ is called *Krylov space.*

The set $(r_0, r_1, \ldots, r_{m-1})$ is an orthogonal basis for the Krylov space, which has therefore dimension m. It follows that the set $\left(p_0, p_1, p_{m-1}\right)$ is a set of linearly independent vectors. Since A is SPD, the property $p_i^T A p_{m-1} = 0$, $i < m-1$ means $p_i^T A p_j = 0$ for $i, j < m-1$, $i \neq j$.

Definition 4.2. A set of vectors different from 0 and satisfying

$$v_i^T A v_j = 0, \qquad\qquad i, j < m, \quad i \neq j,$$

is called a set of *conjugate* (with respect to A) *vectors.*

By construction, the approximate solution x_m produced by the algorithm is in the space $x_0 + \mathcal{K}_m$. By the way, it is possible to prove independently the following.

Proposition 4.4. *A set of conjugate vectors for an SPD matrix is a set of linear independent vectors.*

Proof. Let us suppose that $\sum_{i=1}^k c_i v_i = 0$ with $c_j \neq 0$. Then

$$\left(\sum_{i=1}^k c_i v_i\right)^T A v_j = 0 = \sum_{i=1}^k c_i \left(v_i^T A v_j\right) = c_j v_j^T A v_j.$$

Since A is SPD, the result cannot be zero, unless $v_j \neq 0$ (absurd). □

Theorem 4.2. *The approximate solution x_m produced by the algorithm satisfies*

$$E(x_m) = \inf_{x \in x_0 + \mathcal{K}_m} E(x).$$

Proof. Let us take a vector $x \in x_0 + \mathcal{K}_m$. It is of the form $x_0 + \sum_{i=0}^{m-1} \lambda_i p_i$, and therefore, taking into account that p_i, $i = 0, 1, \dots, m-1$ are conjugate vectors

$$E(x) = E\left(x_0 + \sum_{i=0}^{m-1} \lambda_i p_i\right) = E(x_0) - 2\sum_{i=0}^{m-1} \lambda_i p_i^T r_0 + \lambda_i^2 p_i^T A p_i.$$

Now, we observe that

$$\begin{aligned} p_i^T r_0 &= p_i^T (r_1 + \alpha_0 A p_0) \\ &= p_i^T r_1 \\ &= p_i^T (r_2 + \alpha_1 A p_1) \\ &= p_i^T r_2 \\ &\vdots \\ &= p_i^T r_i. \end{aligned}$$

Therefore

$$E(x) = E(x_0) - 2\sum_{i=0}^{m-1} \lambda_i p_i^T r_i + \lambda_i^2 p_i^T A p_i,$$

and the minimum is taken for $\lambda_i = \alpha_i$, $i \le m-1$. □

This is a remarkable property: we started with looking for $u \in X$ such that $a(u, v) = l(v)$, for any $v \in X$; then we selected a proper $X_h \subset X$ and discovered that u_h satisfying $a(u_h, v) = l(v)$ for any $v \in X_h$ satisfies

$$|||u_h - u||| = \inf_{v \in X_h} |||v - u|||$$

too. Now, if $u_h = \sum_{j=1}^n \bar{x}_j \varphi_j$ and $v = \sum_{j=1}^n x_j \varphi_j$ with $x \in x_0 + \mathcal{K}_m$

$$\begin{aligned} E(x) &= (x - \bar{x})^T A (x - \bar{x}) \\ &= a(v - u_h, v - u_h) \\ &= |||v - u_h|||. \end{aligned}$$

The C.G. method can find the infimum of $E(x)$ on the space $x_0 + \mathcal{K}_m$. Therefore, the solution x_m of the C.G. method is the result of an orthogonal projection method onto $\mathcal{K}_m$. This is clear also from the properties of the method, since

$$0 = r_m^T r_i = \langle b - A x_m, r_i \rangle, \qquad 0 \le i \le m-1,$$

and $(r_0, r_1, \dots, r_{m-1})$ is a basis for $\mathcal{K}_m$.

Proposition 4.5. *The C.G. algorithm converges in n iterations at maximum*

Proof. The Krylov spaces $\mathcal{K}_k = \text{span}\left(p_0, p_1, \dots, p_{k-1}\right)$ has dimension n at maximum. □

In practice, since it is not possible to compute truly conjugate directions in machine arithmetic, usually the C.G. algorithm is used as an iterative method (and is sometimes called *semi-iterative* method).

It is possible to prove the following convergence estimate

$$|||e(x_k)||| = \sqrt{E(x_k)} \le 2\left(\frac{\sqrt{\text{cond}_2(A)} - 1}{\sqrt{\text{cond}_2(A)} + 1}\right)^k |||e(x_0)|||,$$

where $\text{cond}_2(A)$[1] is the condition number, in the 2−norm, of A.

[1] We recall that $\text{cond}_2(A) = \|A\|_2 \left\|A^{-1}\right\|_2 \ge 1$ since $\|I\| = \left\|A \cdot A^{-1}\right\|_2 \le \|A\| \left\|A^{-1}\right\|$

There exists a slightly better estimate

$$\|e(x_k)\| \leq 2\left(\frac{c^k}{1+c^{2k}}\right) |||e(x_0)|||$$

where

$$c = \frac{\sqrt{\text{cond}_2(A)}-1}{\sqrt{\text{cond}_2(A)}+1}.$$

For the computational cost, if we want to reduce the initial error E_0 by a quantity ε, we have to take

$$2\left(\frac{\sqrt{\text{cond}_2(A)}-1}{\sqrt{\text{cond}_2(A)}+1}\right)^k = \varepsilon$$

from which

$$\begin{aligned} k &= \frac{\log\left(\frac{\varepsilon}{2}\right)}{\log\left(\frac{\sqrt{\text{cond}_2(A)}-1}{\sqrt{\text{cond}_2(A)}+1}\right)} \\ &= \frac{\log\left(\frac{\varepsilon}{2}\right)}{\log\left(1-\frac{2}{\sqrt{\text{cond}_2(A)}+1}\right)} \\ &\approx \frac{\log\left(\frac{\varepsilon}{2}\right)}{-\frac{2}{\sqrt{\text{cond}_2(A)}+1}} \\ &= \frac{1}{2}\log\left(\frac{2}{\varepsilon}\right)\sqrt{\text{cond}_2(A)}. \end{aligned}$$

For a matrix with $\text{cond}_2(A) \approx h^{-2}$ the number of expected iterations is therefore $\mathcal{O}(h^{-1})$. The cost of a single iteration is $\mathcal{O}(n)$ if A is sparse. The algorithm does not explicitely require the entries of A, but only the "action" of A to a vector v. For instance, if A is the stiffness matrix of the 1D Poisson problem and

$$v_h(x) = \sum_{j=1}^{n} v_j \varphi_j(x)$$

then the $i-$th row of Av can be obtained by

$$\int_\Omega v_h'(x)\varphi_i'(x)\,dx.$$

4.2 Arnoldi's method

We have seen that the C.G. method produces in practice an orthogonal basis $\left(r_j\right)_{j=0}^{m-1}$ of the Krylov space $\mathcal{K}_m$ and therefore the solution can be written as $x_m = x_0 + \sum_{j=1}^{m-1}$. With non-symmetric matrices, we would like to do the same (that is, to construct an orthogonal basis for the Krylov space). It is possible with the *Arnoldi's algorithm.*

Arnoldi's method is an orthogonal projection method onto $\mathcal{K}_m$ for general non hermitian matrices.

Arnoldi's procedure is an algorithm for building an orthogonal basis of the Krylov subspace $\mathcal{K}_m$. In exact arithmetic, one variant of the algorithm is as follows:

```
choose a vecctor $v_1$ such that $\|v_1\|_2 = 1$
FOR $j = 1,2,\dots,m$
    $h_{i,j} = \langle Av_j, w_i \rangle, \quad i = 1,2,\dots,j$
    $w_j = Av_j - \sum_{i=1}^{j} h_{i,h} v_i$
    $h_{j+1,j} = \| w_j \|_2$
    IF $h_{j+1,j} = 0$ STOP
    $v_{j+1} = \frac{w_j}{h_{j+1,j}}$
END
```

At each step, the algorithm multiplies the previous Arnoldi vector v_j by A and then orthonormalizes the resulting vector w_j against all previous v_i's by a standard Gram-Schmidt procedure. It will stop if the vector w_j computed on line 4 vanishes.
We now see few simple properties

Proposition 4.6. *Assume the algorithm does not stop before the $m-th$ step. Then the vectors $v_1, v_2, \dots, v_m$ form an orthonormal basis of the Krylov subspace $\mathcal{K}_m = \operatorname{span}\left(v_1, Av_1, \dots A^{m-1} v_1\right)$.*

Proof. The vectors $v_j,\ j = 1,2,\dots,m$ are orthonormal by construction. That they span $\mathcal{K}_m$ follows from the fact that each vector v_j os of the form $q_{j-1}(A)v_1$ where q_{j-1} is a polynomial of degree $j-1$. This can be shown by induction on j. Trivially, for $j = 1$, $v_1 = q_0(A)v_1$ with $q_1(t) \equiv 1$. Assume now the result is true for all integer less or equal to j and consider v_{j+1}. We have

$$\begin{aligned} h_{j+1,j} v_{j+1} &= Av_j - \sum_{i=1}^{j} h_{i,j} v_i \\ &= Aq_{j-1}(A)v_1 - \sum_{i=1}^{j} h_{i,j} q_{i-1}(A) v_1, \end{aligned}$$

which shows that v_{j+1} can be expressed as $q_j(A)v_1$ where q_j is of degree j. □

Proposition 4.7. *Denote by V_m the $n \times m$ matrix with column vectors $v_1, v_2, \dots, v_m$; by $\bar{H}_m$ the $(m+1) \times m$ Hessemberg matrix whose nonzero entries $h_{i,j}$ are defined by Arnoldi's algorithm; and by H_m the matrix obtained from $\bar{H}_m$ by deleting its last row. Then the following relations hold:*

$$AV_m = V_m H_m + w_m e_m^T = V_{m+1} \bar{H}_m \tag{4.4a}$$

$$V_m^T A V_m = H_m \tag{4.4b}$$

Remark. If A is symmetric, from (4.4b) we get that also H_m is, and since H_m is Hessemberg, we obtain that it is also tridiagonal. Therefore, the Gram-Schmidt procedure has cost $\mathcal{O}(m)$.

As we noted earlier, the algorithm may break down in case the norm of w_j vanishes at a certain step j. In this case, the vector v_{j+1} cannot be computed, and the algorithm stops. Still to be determined the conditions under which this situation occurs.

Proposition 4.8. *Arnoldi's algorithm breaks down at step j if and only if the minimal polynomial of v_1 is of degree j. Moreover, in this case, the space $\mathcal{K}_j$ is invariant under A.*

4.2.1 Implicit restarted Arnoldi's algorithm

Let us analyze the method under the popular *ARPACK* package for eigenvalue problems. It allows to compute "some" eigenvalues of large sparse matrices (such as the largest in magnitude, the smallest,...). We start with an Arnoldi factorization

$$V_m^T A V_m = H_m, \qquad \text{with} \quad m \leqq n.$$

If (θ, s) is an eigenpair for H_m, that is $H_m s = \theta s$, then

$$\langle v, Ax - \theta x \rangle = 0, \qquad\qquad \forall v \in \mathcal{K},$$

where $x = V_m s$ and $\mathscr{K}$ is the Krylov space spanned by the columns of V_m. In fact, v can be rewritten as $V_m y$ and therefore

$$\begin{aligned}\langle V_m y, Ax - \theta x\rangle &= y^T V_m^T A V_m s - y^T \underbrace{V_m^T V_m}_{I_m} s\theta \\ &= y^t (H_m s - \theta s) \\ &= 0.\end{aligned}$$

The couple (θ, s) is called *Ritz pair* and it is close to an eigenpair of A. In fact

$$\|Ax - \theta x\|_2 = \|(AV_m - V_m H_m)s\|_2 = \left|\beta_m e_m^T s\right|.$$

We can compute the eigenvalues of H_m, for instance by the QR method, and select the “unwanted” eigenvalue μ_m (which is an approximation of an eigenvalue μ of A). Then, we apply one iteration of shifted QR algorithm, that is

$$\begin{aligned}H_m - \mu_m I_m &= Q_1 R_1, \\ H_m^+ &= R_1 Q_1 + \mu_m I_m.\end{aligned}$$

Of course, $Q_1 H_m^+ = H_m Q_1$. Now we right multiply the Arnoldi factorization, in order to get

$$AV_m Q_1 = V_m H_m Q_1 + w_m e_m^T Q_1. \tag{4.5}$$

With few manipulations

$$\begin{aligned}AV_m Q_1 &= V_m Q_1 H_m^+ + w_m e_m^T Q_1 \\ &= V_m Q_1 (R_1 Q_1 + \mu_m I_m) + w_m e_m^T Q_1, \\ (A - \mu_m I_n) V_m Q_1 &= (V_m Q_1)(R_1 Q_1) + w_m e_m^T Q_1, \\ (A - \mu_m I_n) V_m &= V_m Q_1 R_1 + w_m e_m^T,\end{aligned}$$

and, by setting $V_m^+ = V_m Q_1$, we have that the first column of the last expression is

$$(A - \mu_m I_n) v_1 = V_m^+ R_1 e_1 = v_1^+ (e_1^T R_1 e_1),$$

that is, *the first column of* V_m^+ *is a multiple of* $((A - \mu_m I_n) v_1$. If v_1 was a linear combination of the eigenvectors x_j of A, then

$$v_1^+ /\!/ (A - \mu_m I_n) v_1 = \sum_j (\alpha_j \lambda_j x_j - \alpha_j \mu_m x_j).$$

Since μ_m is close to a $\lambda_{\bar{\jmath}}$, v_1^+ lacks the component parallel to $x_{\bar{\jmath}}$. Relation (4.5) can be rewritten as

$$AV_m^+ = V_m^+ H_m^+ + w_m e_m^+ Q_1,$$

and if we consider the first column, it is an Arnoldi factorization with a starting vector v_1^+ (which is of unitary norm) lacking the unwanted component. In particular, given the m eigenvalues of H_m, they are split into the k wanted and the $p = m - k$ unwanted; and p shifted QR decompositions (with each of the unwanted eigenvalues) are performed. Then, the Arnoldi factorization is right multiplied by $Q = Q_1 Q_2 \cdots Q_p$ and the first k colums kept. This turns out to be an Arnoldi factorization. In fact

$$AV_m^+ I_{m,k} = AV_k^+$$

and

$$V_m^+ H_m^+ I_{m,k} = V_k^+ H_k^+ + (V_m^+ e_{k+1} h_{k+1,k}) e_k^T$$

and since the Q_j's are Hessemberg matrices, the last row of Q, that is $e_m^T Q$, has the first $k-1$ entries equal to zero and then a value σ (and then something else). Therefore

$$w_m e_m^T Q I_{m,k} = w_m \sigma e_k^T.$$

All together, the first k columns are

$$\begin{aligned} AV_k^+ &= V_k^+ H_k^+ w_k^+ e_k^T \\ w_k^+ &= V_m^+ e_{k+1} h_{k+1,k} + w_m \sigma \end{aligned}$$

that is an Arnoldi factorization applied to a initial vector lacking the unwanted components. Then, the factorization is continued up to m columns.

The easiest to compute eigenvalues with a Krylov method are the largest in magnitude (as for the power method). Therefore, if some other eigenvalues are desired, it is necessary to apply propr transformations. Let us consider the generalized problem

$$Ax = \lambda Mx.$$

If we are interested into eigenvalues around σ, first we notice that

$$(A - \sigma M)x = (\lambda - \sigma)Mx \qquad \Rightarrow \qquad x = (\lambda - \sigma)(A - \sigma M)^{-1} Mx,$$

from which

$$(A - \sigma M)^{-1} Mx = \nu x, \qquad \nu = \frac{1}{\lambda - \sigma}.$$

Therefore, if we apply the Krylov method (or the power method) to the operator $OP^{-1}B = (A - \sigma M)^{-1} M$ we end up with the eigenvalues closer to σ. In order to do that, we need to be able to solve linear systems with $(A - \sigma M)$ and multiply vectors with M.

4.2.2 Solution to overdetermined systems

Suppose we want to "solve" the linear system

$$\bar{H} y_m = b, \qquad \bar{H} \in \mathbb{R}^{(m+1)\times m}, \quad y_m \in \mathbb{R}^m, \quad b \in \mathbb{R}^{m+1},$$

with $\bar{H}$ of rank m. Since it is *overdetermined*, we can look for the following least squares solution

$$y_m = \arg\min \|b - \bar{H} y\|_2^2.$$

Since the gradient is

$$\nabla_y \|b - \bar{H} y\|_2^2 = -2\bar{H}^T b + 2\bar{H}^T \bar{H} y,$$

the minimum is taken at the solution of

$$\bar{H}^T \bar{H} y_m = \bar{H}^T b.$$

This is called *normal equation* and it is not usually used in order to compute y_m.
A second possibility is to compute the QR factorization of $\bar{H}$. If $\bar{H} = QR$, with $Q \in \mathbb{R}^{(m+1)^2}$ orthogonal and $R \in \mathbb{R}^{(m+1)\times m}$ upper triangular of rank m, then

$$\begin{aligned} \bar{H}^T \bar{H} y_m = \bar{H}^T b &\Leftrightarrow R^T Q^T Q R y_m = R^T Q^T b \\ &\Leftrightarrow R^T (R y_m - Q^T b) = 0. \end{aligned}$$

Since the last column of R^T is zero, we can consider only the first m rows of the linear system $Ry_m = Q^T b$, thus getting a square linear system.
Yet another possibility is to consider the SVD decomposition. We have

$$\bar{H} = USV^T,$$

with $U \in \mathbb{R}^{(m+1)^2}$, $V \in \mathbb{R}^{m^2}$ orthogonal matrices, and

$$S = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & s_m \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(m+1)\times m}.$$

Therefore

$$\begin{aligned} \|b - \bar{H}y_m\|_2 &= \|U^T(b - \bar{H}VV^T)y_m\|_2 \\ &= \|U^T b - U^T\bar{H}V(V^T y_m)\|_2 \\ &= \|f - Sz\|_2, & z = V^T y, \quad f = U^T b. \end{aligned}$$

Now, clearly $\arg\min \|f - Sz\|_2$ has components $z_i = {f_i}/{s_i}$, $i = 1, 2, \dots, m$ and $y_m = Vz$.

4.2.3 Arnoldi's method for linear systems (FOM)

Given an initial guess x_0 to the original linear system $Ax = b$, we now consider an *orthogonal projection method* which takes $\mathscr{L} = \mathscr{K} = \mathscr{K}_m(A, r_0)$, with

$$\mathscr{K}_m(A, r_0) = \text{span}\left\{r_0, Ar_0, A^2 r_0, \dots A^{m-1} r_0\right\},$$

in which $r_0 = b - Ax_0$. This method seek an approximate solution x_m from the affine subspace $x_0 + \mathscr{K}_m$ of dimension m by imposing the *Galerkin condition*

$$b - Ax_m \perp \mathscr{K}_m.$$

If $v_1 = {r_0}/{\|r_0\|_2}$, in Arnoldi's method, and we set $\beta = \|r_0\|_2$, then

$$V_m^T A V_m = H_m \qquad \text{and} \qquad V_m^T r_0 = V_m^T(\beta v_1) = \beta e_1.$$

As a result, the approximate solution using the above $m-$dim subspace is given by

$$x_m = x_0 + V_m y_m, \qquad y_m = H_m^{-1}(\beta e_1).$$

A method based on this approach is the *full orthogonalization method* (*FOM*):

```
compute $r_0 = b - Ax_0$ ,  $\beta = \|r_0\|_2$ ,  $v_1 = r_0/\beta$
define $H_m = \{h_{i,j}\}_{i,j=1,\dots,m}$ ,  $H_m = 0$
FOR $j = 1, 2, \dots, m$
   $w_j = Av_j$
   FOR $i = 1, 2, \dots j$
      $h_{i,j} = \langle w_j, v_i \rangle$
      $w_j -= h_{i,j} v_i$
   END
   compute $h_{j+1,j} = \|w_j\|_2$
   IF $h_{j+1,j} = 0$
```

```
            $m = j$
            GO TO [1]
        END
        $v_{j+1} = \frac{w_j}{h_{j+1,j}}$
    END
[1] $y_m H_m^{-1}(\beta e_1)$,
    $x_m = x_0 + V_m y_m$
```

This algorithm depends on a parameter m which is the dimension of the Krylov subspace. In practice, it is desiderable to select m in a dynamic fashion. This would be possible if the residual norm of the solution x_m is available inexpensively (without to compute x_m itself). Then the algorithm can be stopped at the appropriate step using this information. The following proposition gives a result in this direction.

Proposition 4.9. *The residual vector of the approximate solution x_m computed by the FOM algorithm is such that*

$$b - Ax_m = -h_{m+1,m} e_m^T y_m v_{m+1} \left| e_m^T y_m \right|.$$

Proof. We have the relations

$$\begin{aligned} b - Ax_m &= b - A(x_0 + V_m y_m) \\ &= r_0 - AV_m y_m \\ &= \beta v_1 - V_m H_m y_m - h_{m+1,m} e_m^T y_m v_{m+1}. \end{aligned}$$

By the definition of y_m, $H_m y_m = \beta e_1$, and so $\beta v_1 - V_m H_m y_m = 0$ from which the result follows. □

A rough estimate of the cost of each step of the algorithm is determined as follows. If $N_z(A)$ is the number of nonzero elements of A, then m steps of the Arnoldi's procedure will require m matrix-vector products at the cost of $2m \times N_z(A)$. Each of the Gram-Schmidt steps cost approximatively $4 \times j \times n$ operations, which brings the total over the m steps to approximatively $2m^2 n$. Thus, on the average, a step of FOM costs approximatively

$$2N_z(A) + 2mn.$$

Regarding the storage, m vectors of length n are required to save the basis V_m. Additional vectors must be used to keep the current solution and right hand side, and a scratch vector for the matrix-vector product. In addition, the Hessemberg matrix H_m must be saved. The total is therefore roughly

$$(m+3)n + \frac{m^2}{2}.$$

In most situations, m is small relative to n, so this cost is dominated by the first term.

4.3 Preconditioning

The idea is to change $A\bar{x} = b$ to

$$P^{-1}A\bar{x} = P^{-1}b$$

in such a way that $P^{-1}A$ is better conditioned than A. The main problem for the C.G. algorithm is that even if P is SPD, $P^{-1}A$ is not, in general, SPD. We can therefore factorize P into $R^T R$ (Choleski) and consider the linear system

$$P^{-1}AR^{-1}\bar{y} = P^{-1}b \qquad \Leftrightarrow \qquad R^{-T}AR^{-1}\bar{y} = R^{-T}b, \qquad R^{-1}\bar{y} = \bar{x}.$$

Now, $\tilde{A} = R^{-T}AR^{-1}$ is SPD and we can solve the linear system $\tilde{A}\bar{y} = \tilde{b}$, $\tilde{b} = R^{-T}b$ with the C.G. method. Setting $\tilde{x}_m = Rx_m$, we have $\tilde{r}_m = \tilde{b} - \tilde{A}\tilde{x}_m = R^{-T}b - R^{-T}Ax_m = R^{-T}r_m$. It is possible then to arrange the C.G. algorithm for $\tilde{A}$, $\tilde{x}_0$ and $\tilde{b}$ as

```
$x_0$ given , $r_0 = b - Ax_0$ , $Pz_0 = r_0$ , $p_0 = z_0$
FOR $m = 1, 2, \ldots,$ UNTIL $\|r_m\|_2 \le$ tol $\|b\|_2$
    $w_{m-1} = Ap_{m-1}$
    $\alpha_{m-1} = \frac{z_{m-1}^T r_{m-1}}{p_{m-1}^T w_{m-1}}$
    $x_m = x_{m-1} + \alpha_m p_{m-1}$
    $r_m = r_{m-1} - \alpha_{m-1} w_{m-1}$
    $Pz_m = r_m$
    $\beta_m = \frac{z_m^T r_m}{z_{m-1}^T r_{m-1}}$
    $p_m = z_m + \beta_m p_{m-1}$
END
```

The directions p_m are still A conjugate directions (with $Pp_0 = r_0$). It is easy to see that if $P = A$, then $x_1 = A^{-1}b = \bar{x}$. This algorithm requires the solution of the linear system $Pz_m = r_m$ at each iteration. From one side, P should be as close as possible to A, from the other it should be "easy" to invert.
The simplest choice is $P = \mathrm{diag}(A)$. It is called *Jacobi preconditioner.* If P is not diagonal, usually it is factorized once and for all into $P = R^T R$, R the triangula Cholesky factor, in such a way that z_m can be recovered by two simple triangular linear systems. A possible choice is the *incomplete Cholesky factorization* of A. That is $P = \tilde{R}^T \tilde{R} \approx A$ where

$$\begin{cases} \left(A - \tilde{R}^T \tilde{R}\right)_{i,j} = 0 & \text{if } a_{i,j} \neq 0 \\ \tilde{r}_{i,j} = 0 & \text{if } a_{i,j} = 0 \end{cases}.$$

The preconditioned C.G. (PCG) method does not explicitely require the entries of P, but only the action of P^{-1} (which can be $R^{-1}R^{-T}$) to a vector z_m (that is, the solution of a linear system with matrix P).

4.3.1 Differential preconditioners

If $u(x) \approx \bar{u}(x) \approx \tilde{u}(x)$ with

$$\bar{u}(x) = \sum_{i=1}^{m} \bar{u}_i \phi_i(x)$$

with $\bar{u}_i \approx u(x_i)$ and

$$\tilde{u}(x) = \sum_{i=1}^{n} \tilde{u}_j \psi_j(x), \qquad n \le m,$$

with $\tilde{u}_j \approx u(y_j)$, then it is possible to evaluate $\tilde{u}(x_i)$ by

$$[\tilde{u}(x_1), \ldots, \tilde{u}(x_m)]^T = R\tilde{u}, \qquad R \in \mathbb{R}^{m \times n}, \quad R_{i,j} = \psi_j(x_i)$$

and $\bar{u}(y_j)$ by

$$\left[\bar{u}(y_1), \ldots, \bar{u}(y_n)\right]^T = Q\bar{u}, \qquad Q \in \mathbb{R}^{n \times m}, \quad Q_{i,j} = \phi_i(y_j).$$

We also have

$$\begin{aligned}
[u(x_1), \ldots, u(x_m)]^T &\approx \bar{u} \approx R\tilde{u}, \\
\left[u(y_1), \ldots, u(y_n)\right]^T &\approx \tilde{u} \approx Q\bar{u}, \\
[u(x_1), \ldots, u(x_m)]^T &\approx RQ\bar{u}, \\
\left[u(y_1), \ldots, u(y_n)\right]^T &\approx QR\tilde{u}.
\end{aligned}$$

Therefore

$$RQ \approx I_m, \qquad\qquad QR \approx I_n.$$

Thus, in order to solve the “difficult” problem $\bar{A}\bar{u} = \bar{b}$, we may want to compute $\bar{A}$ of the “easy” problem $\tilde{A}\tilde{u} = \tilde{b}$ and then use the approximation

$$\bar{A}\bar{u} \approx R\tilde{A}Q\bar{u} \Leftrightarrow \bar{A} \approx R\tilde{A}Q$$

to compute a preconditioner:

$$\bar{A}^{-1} \approx (R\tilde{A}Q)^{-1} \approx R\tilde{A}^{-1}Q.$$

4.4 Optimization methods

We consider a couple of methods for the minimization of a function.

4.4.1 Nonlinear conjugate gradient method

We can extend the C.G. method for the minimization of $f(x)$, $x \in \mathbb{R}^n$ in the following way:

```
$x_0$ given, $d_0 = g_0 = -\nabla f(x_0)$
FOR $m = 1,2,\ldots$ UNTIL $\|d_{m-1}\| \leq$ tol $\|d_0\|$
    $\alpha_{m-1} = \arg\min_\alpha f(x_{m-1} + \alpha d_{m-1})$
    $x_m = x_{m-1} + \alpha_{m-1} d_{m-1}$
    $g_m = -\nabla f(x_m)$
    $\beta_m = \frac{g_m^T g_m}{g_{m-1}^T g_{m-1}}$
    $d_m = g_m + \beta_m d_{m-1}$
END
```

It is in general not necessary to compute exactly α_{m-1}. In this case we speack about *inexact line-search.* It can be performed, for instance, by few steps of golden search of f or by few steps on a nonlinear solver for ∇f. The choice of β_m corresponds to *Hetcher - Reeves.* It is possible to use a preconditioner. In fact, $-\nabla f(x_m)$ is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}$, that is

$$-\nabla f(x_m)^T y = (b - Ax_m)^T y \in \mathbb{R}$$

for some $A : \mathbb{R}^n \to \mathbb{R}^n$ and b. It is then possible to use $\tilde{A}$ as a preconditioner and g_m is computed as

$$g_m = -\tilde{A}^{-1}\nabla f(x_m).$$

4.4.2 (Quasi)-Newton methods

It is possible to approximate f (if regular enough and with an SPD Hessian) by a quadratic model

$$f(x) \approx f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T H_f(x_0)(x - x_0).$$

The minimum of the model is given by

$$x - x_0 = -H_f(x_0)^{-1}\nabla f(x_0)$$

and therefore we can define

$$x_1 = x_0 - H_f(x_0)^{-1}\nabla f(x_0).$$

In this form, it is equivalent to the first step of newton method for the solution of the nonlinear systems of equations

$$\nabla f(x) = 0.$$

Instead of the exact Hessian, it is possible to approximate it.

4.5 Generalize minimal residual method

The *Generalize minimal residual method* (*GMRES*) is a projection mwthod based on taking $\mathscr{K} = \mathscr{K}_m$ and $\mathscr{L} = A\mathscr{K}_m$, in which $\mathscr{K}_m$ is the $m-$th Krylov subspace with $v_1 = {r_0}/{\|r_0\|_2}$. Such a technique minimizes the residual norm over all vectors in $x_0 + \mathscr{K}_m$. There are two ways to derive the algorithm

First way. Any vector x in $x_0 + \mathscr{K}_m$ can be written as $x = x_0 + V_m y$, where y is an $m-$vector. Defining

$$J(y) = \|b - A_x\|_2 = \left\|b - A(x_0 + V_m y)\right\|_2,$$

the relation (4.4a) results in

$$\begin{aligned} b - Ax &= b - A(x_0 + V_m y) \\ &= r_0 - AV_m y \\ &= \beta v_1 - V_{m+1}\bar{H}_m y \\ &= V_{m+1}(\beta e_1 - \bar{H}_m y). \end{aligned}$$

Since the column vectors of V_{m+1} are orthonormal, then

$$J(y) = \left\|b - A(x_0 + V_m y)\right\|_2 = \left\|\beta e_1 - \bar{H}_m y\right\|_2.$$

The GMRES approximation is the unique vector of $x_0 + \mathscr{K}_m$ which minimizes $J(y)$. This approximation can be obtained quite simply as $x_m = x_0 + V_m y_m$, where

$$y_m = \arg\min_y \left\|\beta e_1 - \bar{H}_m y\right\|_2.$$

The minimizer y_m is inexpensive to compute since it requires the solution of a $(m+1) \times m$ least squares probelm, where m is tipically small. This gives the following algorithm

```
    $r_0 = b - Ax_0$ ,  $\beta = \|r_0\|_2$ ,  $v_1 = r_0/\beta$
    FOR $j = 1,2,\dots,m$
       $w_j = Av_j$
       FOR $i = 1,\dots,j$
          $h_{i,j} = \langle w_j, v_i\rangle$
          $w_j -= h_{i,j} v_i$
       END
       $h_{j-1,j} = \|w_j\|_2$
       IF $h_{j+1,j} = 0$
          $m = 1$
          GO TO [1]
       END
       $v_{j-1} = w_j/h_{j+1,j}$
    END
[1] $\bar{H}_m = \{h_{i,j}\}_{1\le i\le m+1}^{1\le j\le m}$
    $y_m = \arg\min_y \|\beta e_1 - \bar{H}_m y\|_2$
    $x_m = x_0 + V_m y_m$
```

Second way. Use equation

$$\tilde{x} = x_0 + V(W^T AV)^{-1} W^T r_0,$$

with $W_m = AV_m$.

Chapter 5

Higher order basis functions

We consider, for simplicity, the homogeneous Dirichlet problem.

5.1 1D case

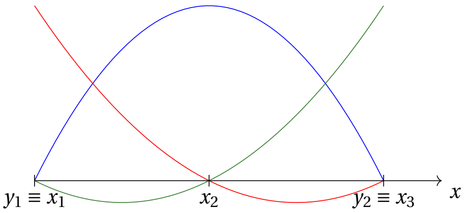
In the 1D case, Ω is an open interval and $X = \mathscr{H}_0^1(\Omega)$. We consider the space

$$X_h^2 = \left\{ v_h \in X : \, v_h|_{T_h^k} \; in \mathbb{P}_2(T_h^k) \right\}.$$

A polynomial of degree two on an interval is defined by three points, usually the two extreme points and the middle one.

Therefore, given an original set of nodes $\left(y_j\right)_{j=1}^m \subset \Omega$, we have to consider the new set of nodes $\{x_i\}_{i=1}^{2m-1} \subset \Omega$ given by

$$\begin{cases} x_i = y_{\frac{i+1}{2}} & \text{if } i \text{ is odd} \\ x_i = \frac{y_{\frac{i}{2}} + y_{\frac{i}{2}+1}}{2} & \text{if } i \text{ is even} \end{cases}$$



and the set of basis functions $\varphi_i(x) \in X_h^2$, $\varphi_i(x_j) = \delta_i^j$ for $1 \le i, j \le 2m-1$. The explicit expression is

$$\varphi_i(x) = \begin{cases} \dfrac{(x - x_{i-1})(x - x_{i-2})}{(x_i - x_{i-1})(x_i - x_{i-2})} & x \in [x_{i-2}, x_i] \\[2ex] \dfrac{(x_{i+1} - x)(x_{i+2} - x)}{(x_{i+1} - x_i)(x_{i+2} - x_i)} & x \in [x_i, x_{i+2}] \\[2ex] 0 & \text{otherwise} \end{cases}$$

if i is odd. For the midpoint intervals (i even) we have

$$\varphi_i(x) = \begin{cases} \dfrac{(x_{i+1} - x)(x - x_{i-1})}{(x_{i+1} - x_1)(x_i - x_{i-1})} & x \in [x_{i-1}, x_{i+1}] \\[2ex] 0 & \text{otherwise} \end{cases}$$

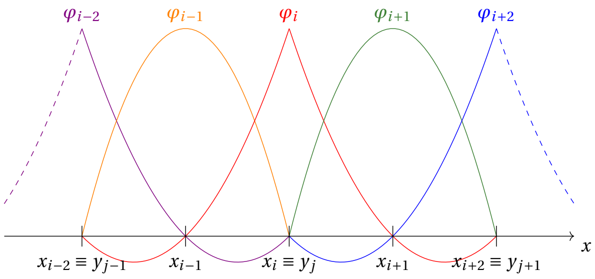
Alternatively, on the element l_j, with endpoints $l_{j,1}$ and $l_{j,3}$, and middle point $l_{j,2}$, the form of $\varphi_{l_{j,k}}$ is

$$\phi_{l_{j,1}}(x) = \frac{\begin{vmatrix} 1 & 1 \\ x & x_{l_{j,2}} \end{vmatrix} \cdot \begin{vmatrix} 1 & 1 \\ x & x_{l_{j,3}} \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ x_{l_{j,1}} & x_{l_{j,2}} \end{vmatrix} \cdot \begin{vmatrix} 1 & 1 \\ x_{l_{j,1}} & x_{l_{j,3}} \end{vmatrix}},$$

$$\phi_{l_{j,2}}(x) = \frac{\begin{vmatrix} 1 & 1 \\ x_{l_{j,1}} & x \end{vmatrix} \cdot \begin{vmatrix} 1 & 1 \\ x & x_{l_{j,3}} \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ x_{l_{j,1}} & x_{l_{j,2}} \end{vmatrix} \cdot \begin{vmatrix} 1 & 1 \\ x_{l_{j,2}} & x_{l_{j,3}} \end{vmatrix}},$$

$$\phi_{l_{j,3}}(x) = \frac{\begin{vmatrix} 1 & 1 \\ x_{l_{j,1}} & x \end{vmatrix} \cdot \begin{vmatrix} 1 & 1 \\ x_{l_{j,2}} & x \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ x_{l_{j,1}} & x_{l_{j,3}} \end{vmatrix} \cdot \begin{vmatrix} 1 & 1 \\ x_{l_{j,2}} & x_{l_{j,3}} \end{vmatrix}}.$$

Clearly, now some basis functions φ_i shares its support with φ_{i-2}, φ_{i-1}, φ_{i+1}, φ_{i+2} and therefore the stiffness matrix is, for instance, a pentadiagonal matrix.



5.1.1 Error estimates

The weak formulation is: find $u \in \mathscr{H}^1(\Omega)$ such that $a(u,v) = l(v)$ for any $v \in \mathscr{H}^1(\Omega)$ with a bilinear, coercive and continuous, and l linear and bounded. Therefore we assume that $u \in \mathscr{H}^1(\Omega)$. Let us denote the generic triangle (edge) by T_h^k and its length by h_k. The maximum lenght of the triangle is h.

$\mathscr{H}^1$ norm, X_h^r space. Let $u_h \in X_h^r$. Then

- if $u \in \mathscr{H}^{p+1}(\Omega, \mathscr{T}_h)$ (u "piecewise regular") and $s = \min\{p,r\}$

$$\|u - u_h\|_{\mathscr{H}^1(\Omega)} \le C \sum_{T_h^k \in \mathscr{T}_h} \sqrt{h_k^{2s} |u|^2_{\mathscr{H}^{s+1}(T_h^k)}} \le C h^s |u|_{\mathscr{H}^{s+1}(\Omega, T_h^k)};$$

- if $u \in \mathscr{H}^{p+1}(\Omega)$ (u "regular" therefore also "piecewise regular") and $s = \min\{p,r\}$

$$\|u - u_h\|_{\mathscr{H}^1(\Omega)} \le C \sum_{T_h^k \in \mathscr{T}_h} \sqrt{h_k^{2s} |u|^2_{\mathscr{H}^{s+1}(T_h^k)}} \le C h^s |u|_{\mathscr{H}^{s+1}(\Omega)}.$$

Of course, the seminorms on the right hand sides can be overstimated by the corresponding norms.

L^2 norm, X_h^r space. Let $u_h \in X_h^r$. If from $l(v) = l_f(v) = \int_\Omega f v$ (therefore $f \in L^2(\Omega)$) it follows that $u \in \mathscr{H}^2(\Omega)$ (it is called *elliptic regularity*, for instance the Poisson problem), then

- if $u \in \mathscr{H}^{p+1}(\Omega, \mathscr{T}_h)$, $s = \min\{p,r\}$,

$$\|u - u_h\|_{L^2(\Omega)} \le C h^{s+1} |u|_{\mathscr{H}^{s+1}(\Omega, \mathscr{T}_h)};$$

- if $u \in \mathscr{H}^{p+1}(\Omega)$, $s = \min\{p,r\}$,

$$\|u - u_h\|_{L^2(\Omega)} \le C h^{s+1} |u|_{\mathscr{H}^{s+1}(\Omega)}.$$

Of course, again, the seminorms on the right hand sides can be oversimated by the corresponding norms.

5.2 2D case

In the $2D$ case, Ω is a polygon and $X = \mathscr{H}^1(\Omega)$, we consider the space

$$X_h^2 = \left\{ v_h \in X \cap \mathscr{C}^0(\bar{\Omega}) : v_h|_{T_h} \in \mathbb{P}_2(\mathscr{T}_h) \right\}.$$

A polynomial of degree two on a triangle is defined by six points in general position. Usually the three vertices and the three middle points are taken. We introduce the *barycentric coordinates*: any point x in a triangle l_j with vertices $\{x_1, x_2, x_3\} \subset \Omega$ can be written in a unique way as convex combination of the vertices:

$$x = \lambda_1(x) + \lambda_2(x) + \lambda_3(x), \qquad \lambda_1(x) + \lambda_2(x) + \lambda_3(x) \equiv 1.$$

We have that $\lambda_k(x)$ coincides, on the triangle, with the piecewise linear function $\phi_{l_{j,k}}(x)$.

Six distinct point in the plane and six corresponding values *are not enough* for the uniqueness of the interpolation polynomial of degree two. Even in the simpler case of degree one, there is no polynomial of such degree which takes the values $(0,0,1)$ in the distinct points $(0,0)$, $(0,1)$ and $(0,2)$. On the other hand, there are infinite polynomials of degree one taking the values $(0,0,0)$ on the same points.

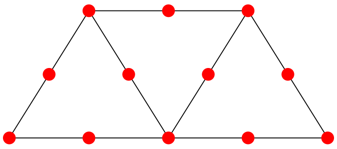
Proposition 5.1. *Given three non collinear points $x_1, x_2, x_3 \in \Omega$ and the corresponding middle points $x_{1,2}, x_{1,3}$ and $x_{2,3}$, a polynomial $p(x)$ of total degree two is well defined by the values of $p(x)$ at these six points.*

Proof. It is enough to prove that if $p(x_i) = p(x_{i,j}) = 0$ for any $1 \le i, j \le 3$, $i \ne j$, then $p(x) \equiv 0$. Along the edge x_2x_3, p is a quadratic polynomial in one variable which is zero at three points. Therefore it is zero on the whole edge and we can write $p(x) = \lambda_1(x)w_1(x)$ with $w_1(x) \in \mathbb{P}_1$ (take $p(x)$, divide by $\lambda_1(x)$ and observe that the reminder is zero). In the same way, p is zero along the edge x_1x_3 and therefore $p(x) = \lambda_1(x)\lambda_2(x)w_0(x)$ with $w_0(x) = \gamma \in \mathbb{P}_0$. If we now take the point $x_{1,2}$, we have

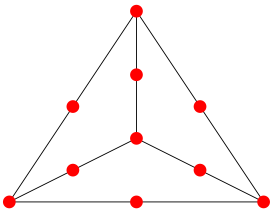
$$0 = p(x_{1,2}) = \lambda(1)(x_{1,2})\lambda_2(x_{1,2})\gamma = \frac{1}{2}\frac{1}{2}\gamma,$$

and therefore $\gamma = 0$. □

The degree of freedom is quite complicate in general. Given the number m of original nodes and the number n of triangles, by Euler's formula we have that the number of edges is $m + (n+1) - 2 = m + n - 1$ (in Euler's formula it has to be counted also the unbounded region outside the triangulation as a single face). Therefore, the dimension of X_h^2 is



12 degrees of freedom



10 degrees of freedom

$$\underbrace{m}_{\text{original nodes}} + \underbrace{m+n-1}_{\text{middle points}} = 2m+n-1.$$

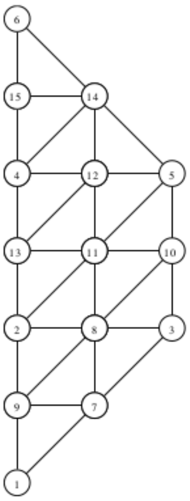
It is not possible to know a priori the structure of the stiffness matrix.

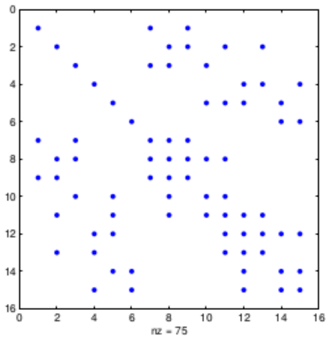
5.2.1 Bandwidth reduction

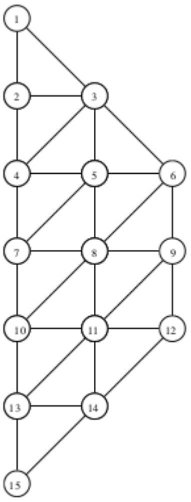
Even in the simplest case of piecewise linear functions, we can get an “ugly” ordening of the nodes that would yield to an “ugly” sparsity pattern.
The *degree* of a node is the number of nodes adjacent to it. We can consider the following heuristic algorithm, called *Cuthill - McKee reordering*.

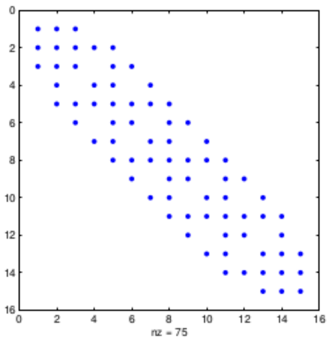
```
Select a node $i$ and set
the first element of the
array R to $i$.

Put the adjacent nodes of $i$
in the increasing order of
their degree in the array Q.

DO UNTIL Q is empty
  take the first node in Q: if
  it is already in R, delete it,
  otherwise add it to R, delete
  it from Q and add to Q the
  adjacent nodes of it which are
  not already in R or Q, in the
  increasing order of their
  degree.
END
```

The new label of node $R(j)$ is j. A variant is the so-called *reverse Cuthill-McKee ordering,* in which the final ordering produced by the previous algorithm is reversed.









5.2.2 Error estimate

The weak formulation is: find $u \in \mathcal{H}^1(\Omega)$ such that $a(u,v) = l(v)$ for any $v \in \mathcal{H}^1(\Omega)$ with a bilinear, coercive and continuous, and l linear and bounded. Therefore we assume that $u \in \mathcal{H}^1(\Omega)$. Let us denote the generic triangle by K and its diameter by h_k, the maximum diameter of the triangle is h.

$\mathcal{H}^1$ norm, X_h^r space. Let $\{\mathcal{T}_h\}_h$ be a family of regular triangulations of Ω, convex polygon and $u_h \in X_h^r$. Then

- if $u \in \mathcal{H}^{p+1}(\Omega, \mathcal{T}_h)$, $s = \min\{p, r\}$, then

$$\|u-u_h\|_{\mathcal{H}^1(\Omega)} \le C \sum_{T_h^k \in \mathcal{T}_h} \sqrt{h_k^{2s} |u|_{\mathcal{H}^{s+1}(T_h^k)}} \le C h^s |u|_{\mathcal{H}^{s+1}(\Omega,\mathcal{T}_h)};$$

- if $u \in \mathscr{H}^{p+1}(\Omega)$, $s = \min\left\{p, r\right\}$, then

$$\|u - u_h\|_{\mathscr{H}^1(\Omega)} \le C \sum_{T_h^k \in \mathscr{T}_h} \sqrt{h_k^{2s} \,|u|_{\mathscr{H}^{s+1}(T_h^k}} \le C h^s \,|u|_{\mathscr{H}^{s+1}(\Omega)} \,.$$

Of course, the seminorms on the right hand sides can be overestimated by the corresponding norms.

L^2 **norm,** X_h^r **space.** Let $\{\mathscr{T}_h\}$ be a family of regular triangulations of Ω convex polygon, and $u_h \in X_h^r$. If from $l(v) = l_f(v) = \int_\Omega f v$ (therefore $f \in L^2(\Omega)$) and Ω convex it follows that $u \in \mathscr{H}^2(\Omega)$ (it is called elliptic regularity, fro instance, the Poisson problem), then

- if $u \in \mathscr{H}^{p+1}(\Omega, \mathscr{T}_h)$, $s = \min\{p, r\}$, then

$$\|u - u_h\|_{L^2(\Omega)} \le C h^{s+1} \,|u|_{\mathscr{H}^{s+1}(\Omega, \mathscr{T}_h)} \,;$$

- if $u \in \mathscr{H}^{p+1}(\Omega)$, $s = \min\{p, r\}$, then

$$\|u - u_h\|_{L^2(\Omega)} \le C h^{s+1} \,|u|_{\mathscr{H}^{s+1}(\Omega)} \,.$$

Of course, the seminorms can be overestimate by the corresponding norms.

Chapter 6

ADR equations

In this part we consider problem of the following form

$$\begin{cases} Lu = -\nabla\cdot(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{6.1}$$

where μ, σ, f and $\mathbf{b}$ are given functions (pr constants). In the most general case we will suppose

$$\begin{aligned} &\mu \in L^\infty(\Omega), && \mu(x) \geq \mu_0 > 0, \\ &\sigma \in L^2(\Omega), && \sigma(x) \geq 0\text{a.e. in } \Omega, \\ &\mathbf{b} \in \left[L^\infty(\Omega)\right]^2, && \nabla\cdot\mathbf{b} \in L^2(\Omega), \\ &f \in L^2(\Omega). \end{aligned}$$

In many practical applications, the *diffusion term* $-\nabla\cdot(\mu\nabla u)$ is dominated by the *convection term* $\mathbf{b}\cdot\nabla u$ (also called *transport term*) or by the *reaction term* σu (also called the *absorption term* when σ is non-negative). In such cases, the solution can give rise to *boundary layers*, that is regions, generally close to the boundary of Ω, where the solution is characterized by strong gradients.

6.1 Weak problem formulation

Let $V = \mathscr{H}_0^1(\Omega)$. By introducing the bilinear form $a: V\times V \to \mathbb{R}$,

$$a(u,v) = \int_\Omega \mu\nabla u\cdot\nabla v \, d\Omega + \int_\Omega v\mathbf{b}\cdot\nabla u \, d\Omega + \int_\Omega \sigma u v \, d\Omega,$$

for any $u, v \in V$, the weak formulation of the problem (6.1) becomes: find $u \in V$ such that $a(u,v) = \langle f, v\rangle$, for any $v \in V$.

In order to prove existence and uniqueness of the weak solution, we will put ourselves in the conditions to apply Lax-Milgram lemma.

To verify the coercivity of the bilinear form $a(\cdot,\cdot)$, we proceed separately on the single terms.

For the first term, we have

$$\int_\Omega \mu\nabla v\cdot\nabla v \, d\Omega \geq \mu_0 \left\|\nabla v\right\|^2_{L^2(\Omega)}.$$

As $v \in \mathscr{H}_0^1(\Omega)$, Poincaré inequality holds:

$$\left\|v\right\|_{L^2(\Omega)} \leq C_\Omega \left\|\nabla v\right\|^2_{L^2(\Omega)}$$

for a suitable $C_\Omega > 0$, independent of v. Thus

$$\left\|v\right\|^2_{\mathscr{H}^1(\Omega)} = \left\|v\right\|^2_{L^2(\Omega)} + \left\|\nabla v\right\|^2_{L^2(\Omega)} \leq \left(1 + C_\Omega^2 \left\|\nabla v\right\|^2_{L^2(\Omega)}\right),$$

therefore

$$\int_\Omega \mu\nabla v\cdot\nabla v\; d\Omega \geq \frac{\mu_0}{1+C_\Omega^2}\,\|v\|^2_{L^2(\Omega)}\,.$$

Using Green's formula on the convective term,

$$\begin{aligned}\int_\Omega v\mathbf{b}\cdot\nabla v\; d\Omega &= \frac{1}{2}\int_\Omega \mathbf{b}\cdot\nabla(v^2)\; d\Omega \\ &= -\frac{1}{2}\int_\Omega v^2(\nabla\cdot\mathbf{b})\; d\Omega + \frac{1}{2}\int_{\partial\Omega}\mathbf{b}\cdot\hat{n}v^2\; d\gamma \\ &= \frac{1}{2}\int_\Omega v^2(\nabla\cdot\mathbf{b})\; d\Omega,\end{aligned}$$

as $v \equiv 0$ on $\partial\Omega$, whence

$$\int_\Omega v\mathbf{b}\cdot\nabla v\; d\Omega + \int_\Omega \sigma v^2\; d\Omega = \int_\Omega v^2\left(-\frac{1}{2}\nabla\cdot]bb+\sigma\right)\; d\Omega.$$

The last integral is certainly positive if we suppose that

$$-\frac{1}{2}\nabla\cdot\mathbf{b}+\sigma \geq 0, \qquad\qquad \text{a.e. } \in\Omega.$$

Consequently, the bilinear form $a(\cdot,\cdot)$ is coercive, as

$$a(v,v)\geq\alpha\,\|v\|_{\mathcal{H}^1(\Omega)}\,, \qquad\qquad \forall v\in V, \qquad\qquad \alpha=\frac{\mu_0}{1+C_\Omega^2}.$$

Let us prove that the bilinear form is continuous, that is there exists $M>0$ such that

$$|a(u,v)|\leq M\,\|u\|_{\mathcal{H}^1(\Omega)}\,\|v\|_{\mathcal{H}^1(\Omega)}\,, \qquad\qquad \forall u,v\in V.$$

The first term of the bilinear form can be bounded as follows:

$$\left|\int_\Omega \mu\nabla u\cdot\nabla v\; d\Omega\right| \leq \left\|\mu\right\|_{L^\infty(\Omega)}\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \leq \left\|\mu\right\|_{L^\infty(\Omega)}\|u\|_{\mathcal{H}^1(\Omega)}\|v\|_{\mathcal{H}^1(\Omega)}\,,$$

having used the Hölder and the Cauchy-Schwarz inequalities, as well as $\|\nabla w\|_{L^2}\leq\|w\|_{\mathcal{H}^1}$ for any $w\in\mathcal{H}^1$.
For the second term, analogously, we find

$$\left|\int_\Omega v\mathbf{b}\cdot\nabla u\; d\Omega\right| \leq \|\mathbf{b}\|_{L^\infty}\|v\|_{L^2}\|\nabla u\|_{L^2} \leq \|\mathbf{b}\|_{L^\infty}\|v\|_{\mathcal{H}^1}\|u\|_{\mathcal{H}^1}\,.$$

For the third term, thanks to the Cauchy-Schwarz inequality, we obtain

$$\left|\int_\Omega \sigma uv\; d\Omega\right| \leq \|\sigma\|_{L^2}\|uv\|_{L^2} \leq \|\sigma\|_{L^2}\|u\|_{\mathcal{H}^1}\|v\|_{\mathcal{H}^1}\,.$$

Summing up, the boundedness follows by taking

$$M=\left\|\mu\right\|_{L^\infty}+\|\mathbf{b}\|_{L^\infty}+\|\sigma\|_{L^2}\,.$$

On the other hand, $\langle f,v\rangle$ define a bounded and linear functional thanks to the Cauchy-Shwarz inequality.

As the Lax-Milgram lemma hypotheses are verified, it follows that the solution to the weak problem exists and it is unique. Moreover, the following a-priori esimates hold

$$\|u\|_{\mathcal{H}^1(\Omega)}\leq\frac{1}{\alpha}\left\|f\right\|_{L^2(\Omega)}\,, \qquad\qquad \|\nabla u\|_{L^2(\Omega)}\leq\frac{C_\Omega}{\mu_0}\left\|f\right\|_{L^2(\Omega)}\,.$$

The Galerkin approximation is: find $u_h \in V_h$ such that $a(u_h, v_h) = \langle f, v_h \rangle$ for any $v_h \in V_h$, where $\{V_h\}_{h>0}$ is a suitable family of subspaces of $\mathscr{H}_0^1(\Omega)$. By replicating the proof carried out above for the exact problem, the following estimates can be proved

$$\|u_h\|_{\mathscr{H}^1} \leq \frac{1}{\alpha}\,\|f\|_{L^2}, \qquad\qquad \|\nabla u\|_{L^2} \leq \frac{C_\Omega}{\mu_0}\,\|f\|_{L^2}.$$

Moreover, the Galerkin error inequality gives

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \int_{v_h \in V_h} \|u - v_h\|_V .$$

The constant M/α becomes as large (and, consequently, the estimate meaningless) as the ratio $\|\mathbf{b}\|_\infty/\|\mu\|_\infty$ (respectively $\|\sigma\|_2/\|\mu\|_\infty$) grows, which happens when the convective (respectively reactive) term dominates over the diffusive one. In such cases *the Galerkin method can give inaccurate solutions, unless a extremely small discretization step h is used.*

Remark. Problem (6.1) is known as the non-conservative form of the ADR problem. The *conservative form* is

$$\begin{cases} Lu = \nabla\cdot(-\mu\nabla u + \mathbf{b}u) + \sigma u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{6.2}$$

If $\mathbf{b}$ is constant, then the two formulations (6.1) and (6.2) are equivalent.
The bilinear form associated to (6.2) is

$$a(u, v) = \int_\Omega (\mu\nabla u - \mathbf{b}u)\cdot\nabla v \, d\Omega + \int_\Omega \sigma u v \, d\Omega,$$

for any $u, v \in V$. The coercivity is satisfied if

$$\frac{1}{2}\nabla\cdot\mathbf{b} + \sigma \geq 0.$$

Under these assumptions, the conclusions drawn for problem (6.1) (and for its approximations) also hold for problem (6.2).

6.2 Analysis of a 1D diffusion transport problem

Let us consider the following 1D transport-diffusion problem

$$\begin{cases} -\mu u'' + bu' = 0 & 0 < x < 1 \\ u(0) = 0, u(1) = 1 \end{cases}$$

μ and b being two positive constants.
Its weak formulation is: find $u \in \mathscr{H}^1(0,1)$ such that $a(u, v) = 0$ for any $v \in \mathscr{H}_0^1(0,1)$, with $u(0) = 0$, $u(1) = 1$ and

$$a(u, v) = \int_0^1 (\mu u' v' + b u' v)\, dx.$$

We can reformulate the problem by introducing a suitable lifting (or extension) of the boundary data. In this particular case, we can choose $Rg = x$. Having set $\mathring{u} = u - Rg = u - x$, we can reformulate the weak problem as: find $\mathring{u} \in \mathscr{H}_0^1(0,1)$ such that $a(\mathring{u}, v) = F(v)$ for any $v \in \mathscr{H}_0^1(0,1)$ where

$$F(v) = -a(x, v) = -\int_0^1 bv \, dx$$

represent the contribution due to the data lifting.
We define the *global Péclet number* as

$$\mathbb{P}\mathrm{e}_g = \frac{bL}{2\mu}, \tag{6.3}$$

L being the linear dimension of the domain (1 in our case). This ratio *provides a “measure” of how the conservative term dominates the diffusive one.* For a negative b, its absolute value should be used in (6.3).

We start by computing the exact solution of such a problem. Its associated characteristic equation is

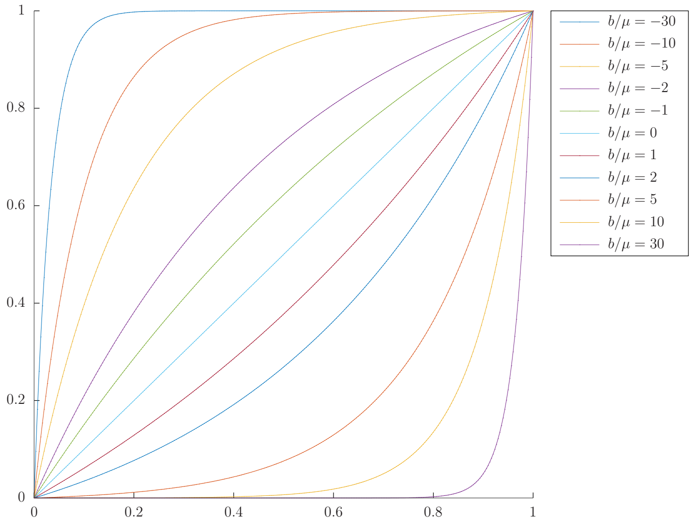
$$-\mu\lambda^2 + b\lambda = 0$$

which has roots $\lambda_1 = 0$, $\lambda_2 = b/\mu$. The general solution is therefore

$$u(x) = C_1\,\mathrm{e}^{\lambda_1 x} + C_2\,\mathrm{e}^{\lambda_2 x} = C_1 + C_2\,\mathrm{e}^{\frac{b}{\mu}x}.$$

By imposing the boundary conditions we find the constant C_1 and C_2, and therefore the solution

$$u(x) = \frac{\exp\frac{b}{\mu}x - 1}{\exp\frac{b}{\mu} - 1}.$$



Using Taylor expansion for the exponentials, if $b/\mu \ll 1$, we obtain

$$\begin{aligned} u(x) &= \frac{1 + \frac{b}{\mu}x + \ldots - 1}{1 + \frac{b}{\mu} + \ldots - 1} \\ &\approx \frac{\frac{b}{\mu}x}{\frac{b}{\mu}} \\ &= x. \end{aligned}$$

Thus the solution lies near the straight line interpolating the boundary data (which is the solution corresponding to the case $b = 0$).
Conversely, if $b/\mu \gg 1$, the exponentials are very large, hence

$$u(x) \approx \frac{\exp\left(\frac{b}{\mu}x\right)}{\exp\left(\frac{b}{\mu}\right)} = \exp\left(-\frac{b}{\mu}(1-x)\right),$$

and the solution is close to zero on almost all of the interval, except for a neighbourhood of the point $x = 1$, where it tends to 1 exponentially. Such a neighbourhood has a width of the order of μ/b and is therefore very small: the solution exhibits a *boundary layer* of width $\mathcal{O}(\mu/b)$ in proximity of $x = 1$, where the derivative behaves like b/μ and is therefore unbounded as $\mu \to 0$.

Let us now suppose to use the Galerkin finite elements method with piecewise linear polynomials: find $u_h \in X_h^1$ such that

$$a(u_h, v_h) = 0, \qquad \forall\, v_h \in \mathring{X}_h^1, \qquad u_h(0) = 0, \qquad u_h(1) = 1,$$

where, denoting by x_i, $i = 0, 1, \ldots, M$ the vertices of the partition of $(0,1)$, we have set

$$X_h^r = \left\{v_h \in \mathscr{C}^0([0,1]) : v_h|_{[x_{i-1}, xi]} \in \mathbb{P}_r, i = 1, \ldots, M\right\}, \qquad \mathring{X}_h^r = \left\{v_h \in X_h^r : v_h(0) = v_h(1) = 0\right\}$$

for $r \geq 1$. Having chosen, for each $i = 1, \dots, M-1$, $v_h = \varphi_i$ (the $i-$th basis function for X_h^1), we have

$$\int_0^1 \mu u_h' \varphi_i' \, dx + \int_0^1 b u_h' \varphi_i' \, dx = 0.$$

Put differently, if we suppose the support of φ_i to be $[x_{i-1}, x_{i+1}]$ and writing

$$u_h = \sum_{j=1}^{M-1} u_j \varphi_j(x),$$

we have

$$\mu \left(u_{i-1} \int_{x_{i-1}}^{x_i} \varphi_{i-1}' \varphi_i' \, dx + u_i \int_{x_{i-1}}^{x_{i+1}} (\varphi_i')^2 \, dx + u_{i+1} \int_{x_i}^{x_{i+1}} \varphi_{i+1}' \varphi_i' \, dx \right) +$$
$$b \left(u_{i-1} \int_{x_{i-1}}^{x_i} \varphi_{i-1}' \varphi_i \, dx + u_i \int_{x_{i-1}}^{x_{i+1}} \varphi_i' \varphi_i \, dx + u_{i+1} \int_{x_i}^{x_{i+1}} \varphi_{i+1}' \varphi_i' \, dx \right) = 0$$

for any $i = 1, \dots, M-1$. If the partition is uniform, that is $x_0 = 0$, $x_i = x_{i-1} + h$, we obtain

$$\mu \left(-\frac{u_{i-1}}{h} + \frac{2u_i}{h} - \frac{u_{i+1}}{h} \right) + b \left(-\frac{u_{i-1}}{h} \frac{h}{2} + \frac{u_{i+1}}{h} \frac{h}{2} \right) = 0,$$

that is

$$\left(\frac{b}{2} - \frac{\mu}{h} \right) u_{i+1} + \frac{2\mu}{h} u_i - \left(\frac{b}{2} + \frac{\mu}{h} \right) u_{i-1} = 0.$$

Dividing by μ/h and defining the *local* (or *grid*) *Péclet number*

$$\mathbb{P}\mathrm{e} = \frac{|b|\, h}{2\mu},$$

we finally have

$$(\mathbb{P}\mathrm{e} - 1) u_{i+1} + 2u_i - (\mathbb{P}\mathrm{e} + 1) u_{i+1} = 0.$$

This is a linear difference equation that admits exponential solutions of the form $u_i = \rho^i$. Replacing such expression, we get

$$(\mathbb{P}\mathrm{e} - 1)\rho^2 + 2\rho - (\mathbb{P}\mathrm{e} + 1) = 0,$$

which has roots

$$\rho_{1,2} = \frac{-1 \pm \sqrt{1 + \mathbb{P}\mathrm{e}^2 - 1}}{\mathbb{P}\mathrm{e} - 1} \qquad \Rightarrow \qquad \rho_1 = \frac{1 + \mathbb{P}\mathrm{e}}{1 - \mathbb{P}\mathrm{e}}, \qquad \rho_2 = 1.$$

The seneral solution is

$$u_i = A_1 \rho_1^i + A_2 \rho_2^i,$$

with A_1, A_2 arbitrary constants. By imposing the boundary conditions $u_0 = 0$, $u_M = 1$ we find

$$A_1 = -A_2, \qquad A_2 = \left(1 - \left(\frac{1 + \mathbb{P}\mathrm{e}}{1 - \mathbb{P}\mathrm{e}} \right)^M \right)^{-1}.$$

To conclude, the solution has the following nodal values:

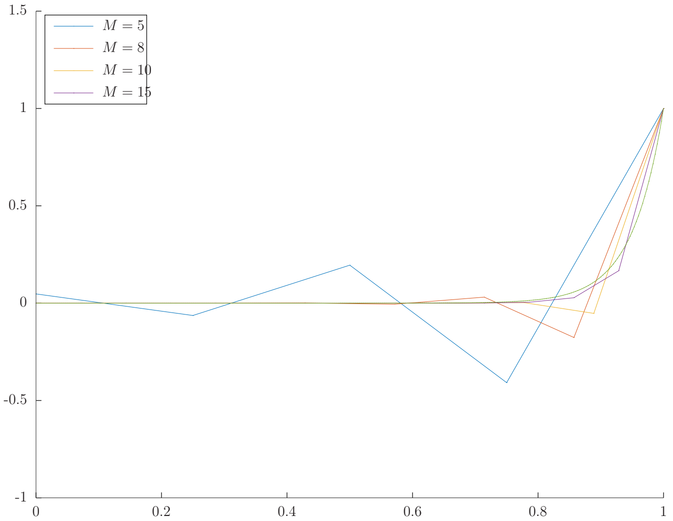
$$u_i = \frac{1 - \left(\frac{1+\mathbb{P}\mathrm{e}}{1-\mathbb{P}\mathrm{e}} \right)^i}{1 - \left(\frac{1+\mathbb{P}\mathrm{e}}{1-\mathbb{P}\mathrm{e}} \right) M}.$$

If $\mathbb{P}\mathrm{e} > 1$, the term within brackets is negative and the solution becomes oscillatory, as opposed to the exact solution that is monotone. The most abvious remedy would be to choose h sufficiently small, in order to ensure $\mathbb{P}\mathrm{e} < 1$. However this strategy is not always convenient: it would require an unreasonably high number of nodal points, in particular in higher dimension. A more suitable remedy consist in using an a-priori adaptive procedure that refines the grid only in proximity of the boundary layer.

6.3 Decentered FD schemes and artificial diffusion

We discretize the first derivative at the point x_i with a decentred incremental ratio where the value at x_{i-1} intervenes if the field is positive, and at x_{i+1} in the opposite case. This technique is called *upwinding* and the resulting scheme, called *upwind scheme* (*FDUP* in short) in the case $b > 0$ writes

$$-\mu\frac{u_{i+1}-2u_i+u_{i+1}}{h^2}+b\frac{u_i-u_{i-1}}{h}=0. \tag{6.4}$$



The price to pay is a reduction of the order of convergence, because the decentred incremental ratio introduces a local discretization error $\mathcal{O}(h)$ as opposed to $\mathcal{O}(h^2)$ as in the CDF case. We remark that

$$\frac{u_i-u_{i-1}}{h}=\frac{u_{i+1}-u_{i-1}}{2h}-\frac{h}{2}\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}.$$

Thus, the upwind scheme can be reinterpreted as a CFD scheme where an artificial diffusion term proportional to h has been introduced. As a matter of fact, (6.4) is equivalent to

$$-\mu_h\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}+b\frac{u_{i+1}-u_{i-1}}{2h}=0, \tag{6.5}$$

where $\mu_h=\mu(1+\mathbb{P}\mathrm{e})$, $\mathbb{P}\mathrm{e}$ being the local Péclet number. This scheme corresponds to the discretization using a CFD scheme of the *perturbed problem*

$$-\mu_h u''+bu'=0.$$

The viscosity correction $\mu_h-\mu=\mu\mathbb{P}\mathrm{e}={}^{bh}/_{2}$ is called *numerical* (or *artificial*) *viscosity*. The new local Péclet number associated to the scheme (6.5) is

$$\mathbb{P}\mathrm{e}^\star=\frac{bh}{2\mu_h}=\frac{\mathbb{P}\mathrm{e}}{1+\mathbb{P}\mathrm{e}},$$

so $\mathbb{P}\mathrm{e}^\star<1$ for all possible values of $h>0$. This interpretation allows to extend the upwind technique to finite elements, and also to the 2D case, where the notion of decentered differentiation is not trivial.

More generally, in a CFD scheme of the form (6.5) we can use the numerical viscosity coefficient

$$\mu_h=\mu(1+\phi(\mathbb{P}\mathrm{e}))$$

where ϕ is a suitable function that must satisfy $\lim_{t\to 0^+}\phi(t)=0$.
A possible choice of ϕ could be:

$$\phi(t)=t-1+B(2t)$$

where B is the *Bernoulli function*

$$B(t)=\begin{cases}\dfrac{t}{\mathrm{e}^t-1} & \text{if } t>0\\[2ex] 1 & \text{if } t=0\end{cases}$$

with which we obtain the *exponential fitting* (or *Scharfetter and Gummel*) *scheme*. Having denoted by ϕ^U and ϕ^{SG} the two functions determined by $\phi(t) = t$ and (6.6), we observe that $\phi^{SG} \approx \phi^U$ if $\mathbb{P}\mathrm{e} \to \infty$, while $\phi^{SG} = \mathcal{O}(\mathbb{P}\mathrm{e}^2)$, $\phi^U = \mathcal{O}(\mathbb{P}\mathrm{e})$ if $\mathbb{P}\mathrm{e} \to 0^+$.
For each given μ and b, the Scharfetter-Gummel scheme is a second order (w.r.t. h) scheme, and, because of this, is also called *upwind scheme with optimal viscosity.* It can be also verified that, if f is constant, then the numerical solution is exact.

We observe that the local Péclet number associated with the coeffcient (6.6) is

$$\mathbb{P}\mathrm{e}^\star = \frac{bh}{2\mu_h} = \frac{\mathbb{P}\mathrm{e}}{1+\phi(\mathbb{P}\mathrm{e})}$$

and is therefore always less than 1, for each value of h.

Remark. The matrix associated with the upwind and S.G. scheme is an M-matrix regardless of the value of h, hence the numerical solution is monotone.

Recall. A *Z-matrix* is a matrix whose off-diagonal entries are less or equal to zero. An *M-matrix* is a Z-matrix whose eigenvalues have positive real part.

6.4 Stabilization methods

Here we deal with the method of *stabilized finite elements*. More precisely, instead of using the Galerkin finite element method, we consider the *generalized Galerkin method*: find $\mathring{u}_h \in V_h$ such that $a(\mathring{u}_h, v_h) = F_h(v_h)$, for any $v_h \in V_h$ where

$$\begin{aligned} a_h(\mathring{u}_h, v_h) &= a(\mathring{u}_h, v_h) + b_h(\mathring{u}_h, v_h), \\ F_h(v_h) &= F(v_h) + G_h(v_h). \end{aligned}$$

The additional terms b_h, G_h have the purpose of eliminating (or at least reducing) the numerical oscillations produced by the Galerkin method (when the grid is not fine enough) and are therefore named *stabilization terms.*

Remark. The term “stabilization” is inexact. The Galerkin method is stable, in the sense of the continuity of the solution with respect to the data of the problem. In this case, stabilization must be undestrood as the aim of reducing the oscillation of the numerical solution when $\mathbb{P}\mathrm{e} > 1$.

6.4.1 Artificial diffusion and decentered finite element schemes

Based on what we have seen for finite differences, we can apply the Galerkin method by replacing μ with $\mu_h = \mu(1+\phi(\mathbb{P}\mathrm{e}))$. This corresponds to choose

$$G_h(v_h) = 0, \qquad\qquad b_h(\mathring{u}_h, v_h) = \mu\phi(\mathbb{P}\mathrm{e}) \int_0^1 \mathring{u}_h' v_h' \, dx.$$

Since $a_h(\mathring{u}_h, \mathring{u}_h) \geq \mu_h |\mathring{u}_h|^2_{\mathcal{H}^1(\Omega)}$ and $\mu_h \geq \mu$, we can say that this formulation is “more coercive” than the standard Galerkin formulation, which corresponds to take $a_h = a$, $F_h = F$.

Theorem 6.1. *If $u \in \mathcal{H}^{r+1}(\Omega)$, then*

$$\|\mathring{u} - \mathring{u}_h\|_{\mathcal{H}^1(\Omega)} \leq C \frac{h^r}{\mu(1+\phi(\mathbb{P}e))} \|\mathring{u}\|_{\mathcal{H}^{r+1}(\Omega)} + \frac{\phi(\mathbb{P}e)}{1+\phi(\mathbb{P}e)} \|\mathring{u}\|_{\mathcal{H}^1(\Omega)},$$

where C is a suitable positive constant independent of h and μ.

Proof. By Strang's lemma we obtain

$$\|\mathring{u}-\mathring{u}_h\|_{\mathcal{H}^1} \le \inf_{w_h\in V_h}\left\{\left(1+\frac{M}{\mu_h}\right)\|\mathring{u}-w_h\|_{\mathcal{H}^1}+\frac{1}{\mu_h}\sup_{v_h\in V_h\setminus\{0\}}\frac{|a(w_h,v_h)-a_h(w_h,v_h)|}{\|v_h\|_{\mathcal{H}^1}}\right\}. \tag{6.7}$$

We choose $w_h = P_h^r\mathring{u}_h$: the orthogonal projection of $\mathring{u}$ on V_h with respect to the scalar product $\int_0^1 u'v'\,dx$ of $\mathcal{H}_0^1(\Omega)$, that is

$$P_h^r\mathring{u}\in V_h : \int_0^1 (P_h^r\mathring{u}-\mathring{u})'v_h'\,dx \qquad v_h\in V_h.$$

It can be proved that

$$\left\|(P_h^r\mathring{u})'\right\|_{L^2}\le\left\|\mathring{u}'\right\|_{L^2}, \qquad \left\|P_h^r\mathring{u}-\mathring{u}\right\|_{\mathcal{H}^1}\le Ch^r\,\|\mathring{u}\|_{\mathcal{H}^{r+1}},$$

C being a constant indepedent of h. Thus we can bound the first addendum of right hand side in (6.7) by $C/\mu_h h^r\,\|\mathring{u}\|_{\mathcal{H}^{r+1}}$.
By the form of b_h, we obtain

$$\frac{1}{\mu_h}\frac{|a(w_h,v_h)-a_h(w_h,v_h)|}{\|v_h\|_{\mathcal{H}^1}}\le\frac{\mu}{\mu_h}\phi(\mathbb{P}\mathrm{e})\frac{1}{\|v\|_{\mathcal{H}^1}}\left|\int_0^1 w_h'v_h'\,dx\right|.$$

Using the Cauchy-Schwarz inequality and observing

$$\left\|v_h'\right\|_{l^2}\le\|v_h\|_{\mathcal{H}^1}, \qquad \left\|(P_h^r\mathring{u})'\right\|_{l^2}\le\left\|P_h^r\mathring{u}\right\|_{\mathcal{H}^1}\le\|\mathring{u}\|_{\mathcal{H}^1},$$

we obtain

$$\frac{1}{\mu_h}\sup_{v_h\in V_h\setminus\{0\}}\frac{\left|a(P_h^r\mathring{u},v_h)-a_h(P_h^r\mathring{u},v_h)\right|}{\|v_h\|_{\mathcal{H}^1}}\le\frac{\phi(\mathbb{P}\mathrm{e})}{1+\phi(\mathbb{P}\mathrm{e})}\|\mathring{u}\|_{\mathcal{H}^1}.$$

□

Corollary 6.1. *For a given μ and for h tending to zero we have*

$$\|\mathring{u}-\mathring{u}_h\|_{\mathcal{H}^1}\le C_1\left(h^r\,\|\mathring{u}\|_{\mathcal{H}^{r+1}}+\phi(\mathbb{P}e)\,\|\mathring{u}\|_{\mathcal{H}^1}\right),$$

where C_1 is a positive constant independent of h; while for a given h and μ tending to zero we have

$$\|\mathring{u}-\mathring{u}_h\|_{\mathcal{H}^1}\le C_2\left(h^{r-1}\,\|\mathring{u}\|_{\mathcal{H}^{r+1}}+\|\mathring{u}\|_{\mathcal{H}^1}\right)$$

where C_2 is a positive constant independent of h and μ.

In particular, for a given μ, the error decays linearly in h when using the upwind viscosity, and quadraticaly with a S.G. scheme if $r\ge 2$.

6.4.2 The Petrov Galerkin method

An equivalent way to write the generalized Galerkin problem with numerical viscosity is to reformulate it as a *Petrov Galerkin* method, that is a method in which the space of test functions is different from the space in which the solution is sought. Precisely, the approximation takes the following form: find $\mathring{u}_h\in V_h$ such that $a(\mathring{u}_h,v_h)=F(v_h)$ for any $v_h\in V_h$, where $V_h\ne W_h$, while the bilinear form $a(\cdot,\cdot)$ is the same as in the initial problem. It can be verified that in the case of linear finite elements (i.e. $r=1$), the generalized Galerkin method can be rewritten in this form, where W_h is the space generated by the functions $\psi_i(x)=\varphi_i(x)+B_i^\alpha$. Here $B_i^\alpha=\alpha B_i(x)$ are the so-called *bubble functions*, with

$$B_i(x)=\begin{cases} g\left(1-\frac{x-x_{i-1}}{h}\right) & x\in[x_{i-1},x_i]\\ -g\left(\frac{x-x_i}{h}\right) & x\in[x_i,x_{i+1}]\\ 0 & \text{otherwise}\end{cases} \qquad g(\xi)=3\xi(\xi-1), \qquad \xi\in[0,1].$$

In the case of upwind scheme, we have $\alpha=1$, while in the case of S.G. scheme, $\alpha=\coth(\mathbb{P}\mathrm{e})-1/\mathbb{P}\mathrm{e}$.

6.4.3 The artificial diffusion and streamline-diffusion methods in the 2D case

The upwind artificial viscosity method can be generalized in the case where we consider a 2D or 3D problem of type (6.1). In such case, it will suffice to modify the Galerkin approximation by adding to the bilinear form a term like

$$Qh\int_\Omega \nabla u_h\cdot\nabla v_h\,d\Omega, \qquad\qquad Q>0,$$

which corresponds to adding the artificial diffusion term $-Qh\Delta u$ to (6.1). The corresponding method is called *upwind artificial diffusion.* In this way, an additional diffusion is introduced, not only in the direction of the field **b**, as one should rightly do in order to stabilize the oscillations generated by the Galerkin method, but also in the orthogonal direction, which is not all necessary. For instance, if we consider the 2D problem

$$\begin{cases} -\mu\Delta u+\frac{\partial u}{\partial x}=f & \text{in }\Omega\\ u=0 & \text{on }\partial\Omega\end{cases}$$

where the transport field is given by the vector $\mathbf{b}=[1,0]^T$, the artificial diffusion term we would add is

$$Qh\frac{\partial^2 u}{\partial x^2}, \qquad \text{and not} \qquad -Qh\Delta u=-Qh\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right).$$

More generally, we can add the following stabilization term

$$Qh\nabla\cdot((\mathbf{b}\cdot\nabla u)\mathbf{b})=Qh\nabla\cdot\left(\frac{\partial u}{\partial\mathbf{b}}\mathbf{b}\right),$$

with $Q=|\mathbf{b}|^{-1}$. In the Galerkin problem, the latter yields the following term

$$b_h(u_h,v_h)=Qh\langle\mathbf{b}\cdot\nabla u_h,\mathbf{b}\cdot\nabla v_h\rangle=Qh\left\langle\frac{\partial u_h}{\partial\mathbf{b}},\frac{\partial v_h}{\partial\mathbf{b}}\right\rangle.$$

The resulting discrete problem is therefore a modification of the Galerkin problem, called *streamline-diffusion* problem and reads: find $u_h\in V_h$ such that $a_h(u_h,v_h)=\langle f,v_h\rangle$ for any $v_h\in V_h$, where

$$a_h(u_h,v_h)=a(u_h,v_h)+b_h(u_h,v_h).$$

6.4.4 Consistency and truncation error

Let us consider a generalized Galerkin problem and replace $\mathring{u}_h$ by u_h to recover more familiar notations. Note that this formulation can refer to a problem in any spatial dimension. We define the functional

$$\tau_h(u,v_h)=a_h(u,v_h)-F_h(v_h),$$

whose norm

$$\tau_h(u)=\sup_{v_h\in V_h\setminus\{0\}}\frac{|\tau_h(u,v_h)|}{\|v_h\|_V}$$

is called *truncation error* associated to the generalized Galerkin method.

The Generalized Galerkin method is said to be *consistent* if $\lim_{h\to0}\tau_h(u)=0$; and is said to be *strongly* (or *fully*) *consistent* if the truncation error is zero for each value of h. The standard Galerkin method is strongly consistent since for any $v_h\in V_h$ we have

$$\tau_h(u,v_h)=a(u,v_h)-F(v_h)=0.$$

On the contrary, the Generalized Galerkin method is only consistent. Concerning the upwind and the streamline-diffusion methods, we have

$$\tau_h(u,v_h)=a_h(u,v_h)-F(v_h)=a_h(u,v_h)-a(u,v_h)=\begin{cases} Qh\langle\nabla u,\nabla v_h\rangle & \text{upwind}\\ Qh\left\langle\frac{\partial u}{\partial\mathbf{b}},\frac{\partial v_h}{\partial\mathbf{b}}\right\rangle & \text{streamline-diffusion}\end{cases}$$

hence both are consistent but not strongly consistent.

6.4.5 Symmetric and skew-symmetric part of an operator

Let V be an Hilbert space and V' its dual. We will say that an operator $L\colon V \to V'$ is *symmetric* if

$$\langle Lu, v\rangle_{V',V} = \langle u, Lv\rangle_{V,V'}, \qquad \forall u, v \in V,$$

and *skew-symmetric* if

$$\langle Lu, v\rangle_{V',V} = -\langle u, Lv\rangle_{V,V'}, \qquad \forall u, v \in V.$$

An operator can be split into the sum of a symmetric part L_S and a skew-symmetric part L_{SS}: $Lu = L_S u + L_{SS} u$.
Let us consider, for instance, the following diffusion-transport-reaction operator

$$Lu = -\mu\Delta u + \nabla\cdot(\mathbf{b}u) + \sigma u, \qquad x \in \Omega \subset \mathbb{R}^d, \quad d \geq 2,$$

operating on the space $V = \mathscr{H}_0^1(\Omega)$. Since

$$\begin{aligned}\nabla\cdot(\mathbf{b}u) &= \frac{1}{2}\nabla\cdot(\mathbf{b}u) + \frac{1}{2}\nabla\cdot(\mathbf{b}u)\\ &= \frac{1}{2}\nabla\cdot(\mathbf{b}u) + \frac{1}{2}u\nabla\cdot\mathbf{b} + \frac{1}{2}\mathbf{b}\cdot\nabla u,\end{aligned}$$

we can split L in the following way

$$Lu = \underbrace{-\mu\Delta u + \left(\sigma + \frac{1}{2}\nabla\cdot\mathbf{b}\right)u}_{L_S u} + \underbrace{\frac{1}{2}\left(\nabla\cdot(\mathbf{b}u) + \mathbf{b}\cdot\nabla u\right)}_{L_{SS}u}.$$

Note that the reaction coefficient has become $\sigma^\star = \sigma + \frac{1}{2}\nabla\cdot\mathbf{b}$. We now show that the two operators are respectively symmetric and skew-symmetric:

$$\begin{aligned}\langle L_S u, v\rangle_{V',V} &= \mu\langle\nabla u, \nabla v\rangle_{L^2} + \left\langle\sigma^\star u, v\right\rangle_{L^2}\\ &= -\mu\langle u, \Delta v\rangle_{V,V'} + \left\langle u, \sigma^\star v\right\rangle_{L^2}\\ &= \langle u, L_S v\rangle_{V,V'},\\ \langle L_{SS}u, v\rangle_{V',V} &= \frac{1}{2}\langle\nabla\cdot(\mathbf{b}u), v\rangle_{L^2} + \frac{1}{2}\langle\mathbf{b}\cdot\nabla u, v\rangle_{L^2}\\ &= -\frac{1}{2}\langle\mathbf{b}u, \nabla v\rangle_{L^2} + \frac{1}{2}\langle\nabla u, \mathbf{b}v\rangle_{L^2}\\ &= -\frac{1}{2}\langle u, \mathbf{b}\cdot\nabla v\rangle_{L^2} - \frac{1}{2}\langle u, \nabla\cdot(\mathbf{b}u)\rangle_{L^2}\\ &= -\langle u, L_{SS}v\rangle_{V,V'}.\end{aligned}$$

Remark. For matrices, A is symmetric if $A = A^T$, and skew symmetric if $A = -A^T$. We can always write $A = A_S + A_{SS}$, with

$$\begin{aligned}A_S &= \frac{1}{2}(A + A^T) && \text{symmetric part of } A\\ A_S &= \frac{1}{2}(A + A^T) && \text{skew-symmetric part of } A\end{aligned}$$

6.4.6 Strongly consistent methods

We consider an ADR problem that we write in the abstract form

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

Let us consider the corresponding weak formulation with $a(\cdot,\cdot)$ being the bilinear form associated to L. A stabilized and strongly consistent method can be obtained by adding a further term to the Galerkin approximation, that is by considering the problem: find $u_h \in V_h$ such that $a(u_h, v_h) = \mathscr{L}_h(u_h, f; v_h) = \langle f, v_h\rangle$, for any $v_h \in V_h$, for a suitable form $\mathscr{L}_h$ satisfying $\mathscr{L}_h(u, f; v_h) = 0$ for any $v_h \in V_h$. A possible choice is

$$\mathscr{L}_h(u_h, f; v_h) = \mathscr{L}_h^{(\rho)}(u_h, f; v_h) = \sum_{K\in\mathscr{T}_h} \langle Lu_h - f, \tau_K S^{(\rho)}(v_h)\rangle_{L^2(K)},$$

where ρ and τ_k are parameters to be determined, and S is defined as

$$S^{(\rho)}(v_h) = L_{SS} v_h + \rho L_S v_h.$$

A possible choice for τ_K is

$$\tau_K = \delta \frac{h_K}{|\mathbf{b}(x)|}, \qquad\qquad \forall x \in K, \quad \forall K \in \mathscr{T}_h$$

where h_K is the iameter of K, and δ is a dimensionless coefficient to be prescribed. To verify that this formulation is fully consistent, we note that

$$\tau_h(u, v_h) = a(u, v_h) + \mathscr{L}_h^{(\rho)}(u, f; v_h) - f(v_h)$$

is zero for all $v_h \in V_h$; thus the truncation error is null.
Let us see some particular cases associated to three different choices of the parameter ρ

$\rho = 1$ *Galerkin least square* (*GLS*): $S^{(1)}(v_h) = Lv_h$;

$\rho = -1$ *Douglas - Wang* (*DW*) method: $S^{(-1)}(v_h) = (L_{SS} - L_S)v_h$;

$\rho = 0$ *Stramline upwind Petrov-Galerkin* (*SUPG*): $S^{(0)}(v_h) = L_{SS} v_h$.

Remark. If $\sigma = 0$ and we use $\mathbb{P}_1$ finite elements, the three previous methods coincide, as $-\Delta u_h|_K = 0$ for any $K \in \mathscr{T}_h$.

Let us now limit ourselves to the two most classical procedures (GLS and SUPG) and to the problem written in the conservative form. We define the "ρ norm"

$$\|v\|_{(\rho)} = \sqrt{\mu\|\nabla v\|_{L^2}^2 + \left\|\sqrt{\gamma} v\right\|_{L^2}^2 + \sum_{K\in\mathscr{T}_h} \langle (L_{SS} + \rho L_S)v, \tau_K S^{(\rho)}(v)\rangle_{L^2}},$$

where γ is a function given by $1/2\nabla\cdot\mathbf{b}+\sigma$ where we use the conservative form of L, and $\gamma = -1/2\nabla\cdot\mathbf{b}+\sigma$ when L is in non-conservative form. In either case, we assume that γ is a non-negative function.

The following (stability) inequality holds: there exist $\alpha^\star$ depending on γ and on the coercivity constant α of $a(\cdot,\cdot)$ such that

$$\|u_h\|_{(\rho)} \le \frac{C}{\alpha^\star} \|f\|_{L^2(\Omega)},$$

where C is a suitable constant. Moreover, the following error estimate holds

$$\|u - u_h\|_{(\rho)} \le Ch^{r+\frac{1}{2}} |u|_{\mathscr{H}^{r+1}(\Omega)},$$

hence the order of accuracy of the method increases when the degree r of the polynomial we employ increases, as in the standard Galerkin method.

The choice of the stabilization parameter δ measuring the amount of artificial viscosity, is extremely important

$$\text{SUPG}: \delta \in \left(0, \frac{1}{C_0}\right), \qquad \text{GLS}: \delta > 0, \qquad \text{DW}: \delta \in \left(0, \frac{1}{2C_0}\right),$$

where C_0 is the constant of the following inverse inequality

$$\sum_{K \in \mathcal{T}_h} h_K^2 \int_K |\Delta v_h|^2 \, dK \leq C_0 \, \|\nabla v_h\|^2_{L^2(\Omega)}, \qquad \forall v_h \in X_h^r.$$

Remark. $C_0 = C_0(r)$, for linear finite elements, $C_0 = 0$ and δ has no upper bound.

6.5 Discontinuous Galerkin(DG) method

Up to now we have considered Galerkin methods with subspaces of continuous polynomial functions. We now deal with approximation techniques based on subspaces of polynomials that are discontinuous between elements.

6.5.1 DG method for the Poisson problem

Let us consider the Poisson problem together with homogeneous Dirichlet boundary conditions in a domain $\Omega \subset \mathbb{R}^2$ dvided in the union of M adjoint elements Ω_m, $m = 1, \dots, M$. We wish to obtain an alternative weak formulation to the usual one, that will serve as starting point for the DG method. To simpligy the discussion, we assume the exact solution to be sufficiently regular, e.g. $u \in \mathcal{H}_0^1(\Omega) \cap \mathcal{H}^2(\Omega)$, so that all operations below make sense.
Define the space

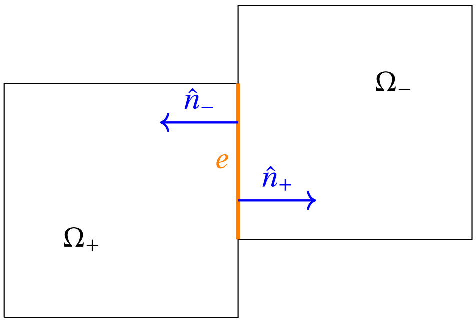
$$W^0 = \{v \in W : v|_{\partial\Omega} = 0\} \qquad \text{where} \qquad W = \left\{v \in L^2(\Omega) : \, v|_{\Omega_m} \in \mathcal{H}^1(\Omega_m), m = 1, \dots, M\right\}.$$

By Green's formula we have, for every $v \in W^0$

$$\sum_{m=1}^{M} \langle -\Delta u, v \rangle_{L^2(\Omega_m)} = \sum_{m=1}^{M} \left(\langle \nabla u, \nabla v \rangle_{L^2(\Omega_m)} - \int_{\partial\Omega_m} v \nabla u \cdot \hat{n}_m \right),$$

where $\hat{n}_m$ is the outward unit normal to $\partial\Omega_m$. Calling $\mathcal{E}_\delta$ the union of all the internal edges, i.e. the interfaces separating the subdomains, we can rearrange the terms to obtain

$$-\sum_{m=1}^{M} \int_{\partial\Omega_m} v \nabla u \cdot \hat{n}_m = -\sum_{e \in \mathcal{E}_\delta} \int_e \left(v^+ \nabla u^+ \cdot \hat{n}^+ + v^- \nabla u^- \cdot \hat{n}^- \right)\Big|_e,$$



in which the signa '+' and '-' label the informations according to the two possible normal orientations.
We will use the following notations to denote mean values and jumps on elements' edges:

$$\{v\} = \frac{v^+ + v^-}{2}, \qquad \{\{\nabla w\}\} = \frac{(\nabla w)^+ + (\nabla w)^-}{2},$$
$$[v] = v^+ \hat{n}^+ + v^- \hat{n}^-, \qquad [[\nabla w]] = (\nabla w)^+ \cdot \hat{n}^+ + (\nabla w)^- \cdot \hat{n}^-.$$

A little algebraic manipulation yields

$$\begin{aligned} v^+ \nabla u^+ \cdot \hat{n}^+ + v^- \nabla u^- \cdot \hat{n}^- &= 2[v] \cdot \{\{\nabla u\}\} - (v^+ \nabla u^- \cdot \hat{n}^+ + v^- \nabla u^+ \cdot \hat{n}^-) \\ &= 2[v] \cdot \{\{\nabla u\}\} + 2[[\nabla u]]\{v\} - (v^+ \nabla u^+ \cdot \hat{n}^+ + v^- \nabla u^- \cdot \hat{n}^-) \\ &= [v]\{\{\nabla u\}\} + [[\nabla u]]\{v\}. \end{aligned}$$

We have that the solution to the Poisson problem satisfies

$$\sum_{m=1}^{M} \langle \nabla u, \nabla v \rangle_{L^2(\Omega_m)} - \sum_{e\in\mathscr{E}_\delta} \int_e ([v]\{\{\nabla u\}\} + [[\nabla u]]\{v\}) = \sum_{m=1}^{M} \left\langle f, v \right\rangle_{L^2(\Omega_m)}, \qquad \forall v \in W^0.$$

Now we introduce the discrete space

$$W_\delta^0 = \{v_\delta \in W_\delta : v_\delta|_{\partial\Omega} = 0\} \qquad \text{with} \qquad W_\delta = \left\{v_\delta \in W : v_\delta|_{\Omega_m} \in \mathbb{P}_r(\Omega_m), m = 1, \dots, M\right\}.$$

Not that, since $u \in \mathscr{H}_0^1(\Omega) \cap \mathscr{H}^2(\Omega)$, then $[[\nabla u]]$ is null on every edge $e \in \mathscr{E}_\delta$. This fact motivates the following DG approximation for the 2D Poisson problem: find $u_\delta \in W_\delta^0$ satisfying

$$\sum_{m=1}^{M} \langle \nabla u_\delta, \nabla v_\delta \rangle_{L^2(\Omega_m)} - \sum_{e\in\mathscr{E}_\delta} \int_e [v_\delta]\{\{\nabla u_\delta\}\} - \tau \sum_{e\in\mathscr{E}_\delta} \int_e [u_\delta] \cdot \{\{\nabla v_\delta\}\} + \sum_{e\in\mathscr{E}_\delta} \gamma |e|^{-1} \int_e [u_\delta] \cdot [v_\delta]$$
$$= \sum_{m=1}^{M} \left\langle f, v_\delta \right\rangle_{L^2(\Omega_m)}, \qquad \forall v_\delta \in W_\delta^0, \quad (6.8)$$

where $\gamma = \gamma(r)$ is a suitable positive constant, $|e|$ is the length of $e \in \mathscr{E}_\delta$ and τ is a suitable fixed number.

Remark. The additional new terms $\tau[u_\delta]\{\{\nabla v_\delta\}\}$ and $\gamma|e|^{-1}[u_\delta] \cdot [v_\delta]$ do not undermine strong consistency (since $[u] = 0$ if u is the exact strong solution fo the Poisson problem), beside warranting greater generality and improved stability features.

Formulation (6.8) is called *Interior Penality* (*IP*). If $\tau = 1$ the method remains symmetric and it is known as *SIPG* (*Symmetric Interior Penality Galerkin*) *method*. For $\tau \neq 1$ the bilinear form is no longer symmetric and the values $\tau = -1$ and $\tau = 0$ respectively lead to the *NIPG* (*Non-symmetric Interior Penality Galerkin*) *method* and to the *IIPG* (*Incomplete Interior Penality Galerkin*) *method*.
Whereas the former is stable for any $\gamma > 0$, SIPG and IIPG require, in order to reach a stable formulation, a sufficiently large penalty parameter γ.

Several variants of formulation (6.8) have been proposed. A first version consist in replacing the last term of the left hand side with the stabilization term

$$\sum_{e\in\mathscr{E}_\delta} \gamma \int_e r_e([u_\delta]) \cdot r_e([v_\delta])$$

where r_e is a suitable extension operator which from the jump of a function $[v_\delta]$ across $e \in \mathscr{E}_\delta$ generates a function $r_e([v_\delta])$ with non-zero support on the elements having e as an edge.
A second variant replace $\{\{\nabla w\}\}$ with

$$\{\{\nabla w\}\}_\theta = \theta \nabla w^+ + (1-\theta)\nabla w^-, \qquad \theta \in [0,1].$$

Up to now we imposed the homogeneous Dirichlet conditions “strongly”. In order to add an inhomogeneous boundary constraint $u = g$ on $\partial\Omega$ in weak form, we write the discrete formulation in W_δ rather than in W_δ^0, and add to the left hand side the following contribution to the boundary edges $e \subset \partial\Omega$

$$-\sum_{e\in\mathscr{E}_\delta} \int_e v_\delta \nabla u_h \cdot \hat{n} - \tau \sum_{e\in\mathscr{E}_\delta} \int_e (u_\delta - g_delta) \nabla v_\delta \cdot \hat{n} + \sum_{e\subset\partial\Omega} \gamma |e|^{-1} \int_e (u_\delta - g_\delta) v_\delta, \qquad u_\delta, v_\delta \in W_\delta,$$

where the positive constant $\gamma = \gamma(r)$ is the same of (6.8), and g_δ is a convenient approximation of g. Te first term, arising from integration by parts formula, ensures that the method is strongly consistent, while the second term make the formulation symmetric if $\tau = 1$ and non-symmetric if $\tau \in \{0, -1\}$.

The last term penalizes the trace of the discrete solution u_δ and makes it "approach" the Dirichlet datum. Observe how incorporate these terms does not affect the method's strong consistency. The DG formulation with boundary conditions imposed weakly thus becomes: find $u_\delta \in W_\delta$ such that

$$\begin{aligned}\sum_{m=1}^{M} \langle \nabla u_\delta, \nabla v_\delta \rangle_{L^2(\Omega_m)} - \sum_{e\in\mathscr{E}_\delta} \int_e [v_\delta]\cdot\{\{\nabla u_\delta\}\} - \tau \sum_{e\in\mathscr{E}_\delta} \int_e [u_\delta]\cdot\{\{\nabla v_\delta\}\} - \sum_{e\subset\partial\Omega} \int_e v_\delta \nabla u_\delta \cdot \hat{n} \\ - \tau \sum_{e\subset\partial\Omega} \int_e u_\delta \nabla v_\delta \cdot \hat{n} + \sum_{e\subset\partial\Omega} \gamma |e|^{-1} \int_e u_\delta v_\delta = \sum_{m=1}^{M} \langle f, v_\delta \rangle_{L^2(\Omega_m)} - \tau \sum_{e\subset\partial\Omega} \int_e g_\delta \nabla v_\delta \cdot \hat{n} \\ + \sum_{e\subset\partial\Omega} \gamma |e|^{-1} \int_e g_\delta v_\delta, \qquad \forall v_\delta \in W_\delta. \quad (6.9)\end{aligned}$$

Concerning the accuracy of this method for discretizing the Poisson problem with homogenous Dirichlet boundary conditions, let us introduce the so-called *energy norm*

$$|||u_\delta||| = \sqrt{\sum_{m=1}^{M} \int_{\Omega_m} |\nabla u_\delta|^2 + \sum_{e\in\mathscr{E}_\delta} \gamma |e|^{-1} \int_e [u_\delta]^2 + \sum_{e\in\mathscr{E}_\delta} \gamma |e|^{-1} \int_e |u_\delta|^2}.$$

For formulation (6.8), where boundary conditions are imposed strongly, the last term is missing. If the exact solution is enough regular, SIPG ($\tau = 1$)converges with optimal convergence rate both in L^2 and energy norm as long as γ is large enough. Better said, for finite elements of degree r one has

$$h|||u - u_\delta||| + \|u - u_\delta\|_{L^2(\Omega)} \le C h^{r+1} |u|_{\mathscr{H}^{r+1}(\Omega)},$$

where C is an appropriate positive number that depends on r, the polynomial degree employed on each element Ω_m.

For the non-symmetric methods NIPG and IIPG, as these schemes are not strongly consistent on the adjoint problem, one cannot get optimal L^2 estimates. In many cases, nevertheless, both methods exhibit optimal rate of convergence when the degree of the approximation is odd and grids are sufficiently regular.

For all variants of the DG-N method (6.9), one can prove that if $u \in \mathscr{H}^{s+1}(\Omega)$, $s \ge 1$, and if the polynomial degree r satisfies $r \ge s$, the error can be estimated in energy norm as follows

$$|||u - u_\delta||| \le C \left(\frac{h}{r}\right)^s r^{\frac{1}{2}} |u|_{\mathscr{H}^{s+1}(\Omega)},$$

where C is a suitable positive constant that does not depend on r.

6.5.2 DG methods for diffusion-transport equations

We extend the DG method to the ADR problem as follows: find $u_\delta \in W_\delta^0$ s.t.

$$\begin{aligned}\sum_{m=1}^{M} \langle \mu \nabla u_\delta, \nabla v_\delta \rangle_{L^2(\Omega_m)} - \sum_{e\in\mathscr{E}_\delta} \int_e [v_\delta]\cdot\{\{\mu\nabla u_\delta\}\} - \tau \sum_{e\in\mathscr{E}_\delta} \int_e [u_\delta]\{\{\mu\nabla v_\delta\}\} + \sum_{e\in\mathscr{E}_\delta} \int_e \gamma |e|^{-1} [u_\delta]\cdot[v_\delta] \\ - \sum_{m=1}^{M} \langle \mathbf{b} u_\delta, \nabla v_\delta \rangle_{L^2(\Omega_m)} + \sum_{e\in\mathscr{E}_\delta} \int_e \{\{\mathbf{b} u_\delta\}\}_{\mathbf{b}} \cdot [v_\delta] + \sum_{m=1}^{M} \langle \sigma u_\delta, v_\delta \rangle_{L^2(\Omega_m)} \\ = \sum_{m=1}^{M} \langle f, v_\delta \rangle_{L^2(\Omega_m)}, \qquad \forall v_\delta \in W_\delta^0, \quad (6.10)\end{aligned}$$

where

$$\{\{\mathbf{b} u_\delta\}\}_{\mathbf{b}} = \begin{cases} \mathbf{b} u_\delta^+ & \text{if } \mathbf{b}\cdot\hat{n}^+ > 0 \\ \mathbf{b} u_\delta^- & \text{if } \mathbf{b}\cdot\hat{n}^+ < 0 \\ \mathbf{b}\{u_\delta\} & \text{if } \mathbf{b}\cdot\hat{n}^+ = 0 \end{cases}$$

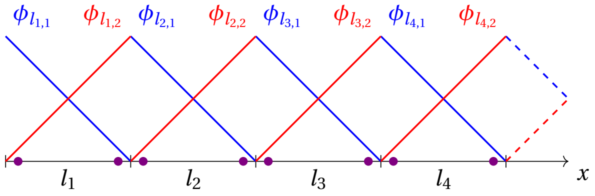
Observe that $\{\{\mathbf{b}u_\delta\}\}_\mathbf{b}[v_\delta] = 0$ if $\mathbf{b}\hat{n}^+ = 0$. If the ADR is written in non-conservative form, it is sufficient to substitute

$$\sum_{m=1}^{M} \langle \sigma u_\delta, v_\delta \rangle_{L^2(\Omega_m)} \qquad \text{with} \qquad \sum_{m=1}^{M} \langle \eta u_\delta, v_\delta \rangle_{L^2(\Omega_m)},$$

where $\eta(\mathbf{x}) = \sigma(\mathbf{x}) - \nabla \cdot \mathbf{b}(\mathbf{x})$. This time we suppose that there exists a positive constant $\eta_0 > 0$ such that $\eta(\mathbf{x}) \geq \eta_0$ for almost every $\mathbf{x} \in \Omega$.

The DG method can easily be localized to every subdomain Ω_m. Indeed, since the test functions do not have to be continuous, for every $m = 1, \dots, M$, we can choose test functions to vanish outside the element Ω_m. In this way, the sum in (6.10) is reduced to the single index m, while the one on the edges is reduced to the edges of the boundary of Ω_m.
We spot here another peculiarity of the DG method: it fits well with local refinements, element by element, either grid-wise or polynomial-wise.

6.5.3 Basis functions

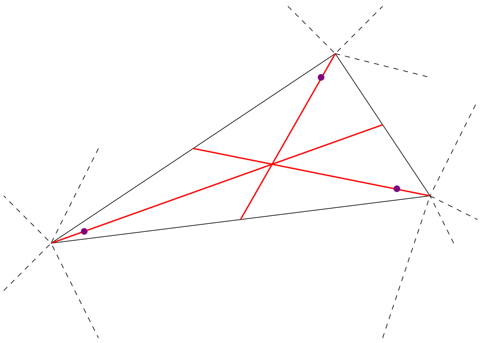


We first consider the space of discontinuous piecewise linear basis functions in 1D. For the choice of basis functions, we have at least two possibilities. If we consider the interval $l_m = [x_m, x_{m+1}]$ we can use $\phi_{l_{m,k}}$, $k = 1, 2$. Given a function $u(x)$, it can be approximated by

$$\hat{u}(x) = \sum_{m=1}^{M} \sum_{k=1}^{2} u_m^k \phi_{l_{m,k}}(x). \tag{6.11}$$

In order to recover the coefficients u_m^k we need six conditions. For instance, one could prescribe interpolation at the two extreme points for each interval. since in general the function to approximate is discontinuous, we can prescribe $u_m^1 = u(x_{l_{m,1}}^+)$, $u_m^2 = u(x_{l_{m,1}}^-)$, but then $\hat{u}(x_m) = 2u(x_m)$. The problem is that *at the common points between two adjacent intervals there are two basis functions which take value one.* In 1D, it would be possible to remedy by restricting $\phi_{l_{m,2}}$ to the interval $[x_{l_{m,1}}, x_{l_{m,1}})$ except the last $\phi_{l_{M,2}}$. But in a 2D triangulation it is not possible to specify in an easy way which single basis function should take value one at a vertex. In this sense, representation (6.11) should be understood as

$$\hat{u}(x)|_{l_m} = \sum_{k=1}^{2} u_m^k \phi_{l_{m,k}}(x).$$



In 2D, it is even not easy, given a discontinuous function to represent, to associate the correct value to the coefficients. Therefore, the common way is to prescribe six interpolation conditions at inner points close to the vertices. For instance, in 1D

$$y_{l_{m,k}} = \frac{x_{l_{m1,}} + x_{l_{m,2}}}{2} + 0.99\left(x_{l_{m,k}} - \frac{x_{l_{m1,}} + x_{l_{m,2}}}{2}\right).$$

In this way, the approximation $\hat{u}(x)$ would not be continuous in general and the coefficients u_m^k would preserve the meaning of "almost" the values of $u(x)$ at the discretization points.
Another completely different way is to abandon the idea of retrieving the values at the discretization

points (which, in the framework of discontinuous methods is not that important). First of all, we need the normalized Legendre polynomials of degree zero and one, namely

$$L_0(x) = 1, \qquad\qquad L_1(x) = \sqrt{3}(2x-1).$$

They satisfy

$$\int_0^1 L_h(x)L_k(x)\,dx = \delta_h^k.$$

Then, given the interval $I_i[x_i, x_{i+1}]$, it is

$$\frac{1}{x_{i+1}-x_i}\int_{I_i} L_h\left(\frac{x-x_i}{x_{i+1}-x_i}\right)L_k\left(\frac{x-x_i}{x_{i+1}-x_i}\right)\,dx = \delta_h^k.$$

We can therefore define

$$L_h^i(x)|_{I_j} = \delta_i^j \frac{1}{\sqrt{x_{i+1}-x_i}} L_h\left(\frac{x-x_i}{x_{i+1}-x_i}\right)$$

and we have

$$\int_{I_l} L_h^i(x)L_k^j(x)\,dx = \delta_h^k\delta_i^j\delta_i^l\delta_j^l.$$

The set $\{L_h^i\}_{0\le h\le 1, 1\le i\le M}$ is the set of basis functions. Given $v(x) \in L^2([0,1])$, its approximation in the finite element space is

$$\hat{v}(x) = \sum_{m=1}^{M}\sum_{k=0}^{1} \hat{v}_m^k L_k^m(x), \qquad \text{with} \qquad \hat{v}_m^k = \int_{I_m} v(x)L_k^m(x)\,dx.$$

In general, $\hat{v}(x)$ does not interpolate $v(x)$ at the nodes and the coefficients $\hat{v}_m^k$ have no physical meaning.
In 2D, an appropriate orthonormal basis for $1, x, y$ can be obtained by the Gram-Schmidt procedure.

6.5.4 Computation of local error estimator

We consider the following local error estimator:

$$\rho_K(u_h) = h_K \left\|f + \Delta u_h\right\|_{L^2(K)} + \frac{1}{2}\sum_{e\in K}\sqrt{h_e}\left\|\left[\frac{\partial u_h}{\partial n}\right]\right\|_{L^2(e)}.$$

the jump notation means $(\nabla u_h)^+ \cdot \hat{n}^+ + (\nabla u_h)^- \cdot \hat{n}^-$ on the internal edges and its L^2 norm corresponds to the L^2 norm of $\texttt{jump}\nabla u_h \cdot \vec{v}_k)$. Otherwise it is zero (but `jump` is not zero). Therefore

$$\begin{aligned}\rho_K(u_\delta) &= \sqrt{\int_\Omega h_K^2\left|f+\Delta u_h\right|^2\chi_K} + \frac{1}{2}\sum_{e\in\mathscr{E}_\delta}\sqrt{\int_e h_e\left[\frac{\partial u_h}{\partial n}\right]^2\chi_K}\\ &= \sqrt{\int_\Omega h_K^2\left|f+\Delta u_h\right|^2\chi_K} + \frac{1}{2}\sum_e\sqrt{\int_e (\texttt{nTonEdge}-1)h_e\texttt{jump}(\nabla u_h\cdot\vec{v}_k)^2\chi_K}\end{aligned}$$

where `nTonEdge` is 2 for internal edges and 1 otherwise.
Since in the local estimates some constants are anyway present, different formulation are available, using classical inequality

$$\sum_{i=1}^n a_i \le \sqrt{n}\sum_{i=1}^n a_i^2.$$

Chapter 7

Grid adaptivity

In order to have an efficient grid that minimizes the number of elements necessary to obtain the desired accuracy, we can equidistribute the error on each element $K \in \mathcal{T}_h$. In particular, we would like to obtain

$$h_K^r |u|_{\mathcal{H}^{r+1}(K)} \approx \eta, \qquad \forall K \in \mathcal{T}_h,$$

where η is a well-chosen constant that only depends on the desired accuracy and on the number of elements of the grid. A larger contribution from $|u|_{\mathcal{H}^{r+1}(K)}$ will need to be balanced by a smaller grid size h_K or by an higher polynomial degree r. In the first case, we will talk of *h-adaptivity*, in the second one of *p-adaptivity*. We will now focus on the first one.

What said up to now are of little use as the solution u is not known. We can therefore proceed according two different strategies:

- *a priori adaptivity*: replace the exact solution with a well-chosen approximation;
- *a posteriori adaptivity*: we link the approximation error to the behaviour of the approximate solution u_h known after solving the problem numerically. In such case, the optimal grid will be constructed through an iterative process where solution, error estimate and modification of the grid are recomputed until reaching the requested accuracy.

A priori and a posteriori adaptivities are not mutually exclusive, actually they can coexist.

7.1 A priory adaptivity based on derivatives reconstruction

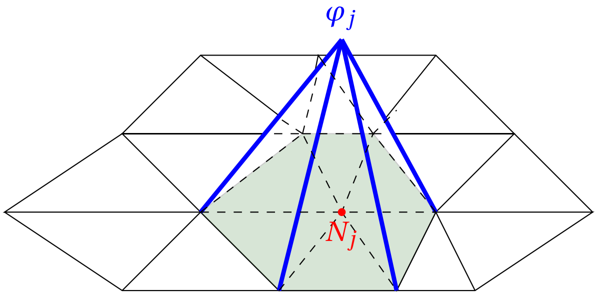
an a priory adaptivity technique is based on estimate

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq \frac{M}{\alpha} C \sqrt{\sum_{K \in \mathcal{T}_h} h_K^{2r} |u|_{\mathcal{H}^{r+1}(K)}} \tag{7.1}$$

where the derivatives of u are carefully approximated on each element, with the purpose of estimating the local seminorms of u. To do this, a approximate solution $u_{h^\star}$ is used, computed on a tentative grid with step size $h^\star$ large enough so that the computation is cheap, but not too large to generate an excessive error in the approximation of the derivatives, which could affect the effectiveness of the whole procedure.
We exemplify the algorithm for linear finite elements in which (7.1) writes

$$\|u - u_h\|_{\mathcal{H}^1(\Omega)} \leq C \sqrt{\sum_{K \in \mathcal{T}_h} h_K^2 |u|^2_{\mathcal{H}^2(K)}}. \tag{7.2}$$

Our aim is eventually to solve our problem on a grid $\mathcal{T}_h$ guaranteeing that the right hand side stands below a predefinite tolerance $\varepsilon > 0$.



Let us suppose we have computed a solution $u_{h^\star}$ on a preliminary grid $\mathcal{T}_{h^\star}$ with $N^\star$ triangles. We use $u_{h^\star}$ to approximate the second derivatives of u that intervene in the seminorm $|u|_{\mathcal{H}^2(K)}$. Since $u_{h^\star}$ does not have any continuous second derivatives in Ω, it is necessary to proceed with an adequate reconstruction technique. For each node N_j we consider the set K_{N_j} of the elements sharing N_j as a node (called patch). It is the set of the elements forming the support to the basis function φ_j. We then find the planes $\pi_i^j(x) = \mathbf{a}_i^j \cdot \mathbf{x} + b_i^j$ by minimizing

$$\int_{K_{N_i}} \left| \pi_i^j(\mathbf{x}) - \frac{\partial u_{h^\star}}{\partial x_j} \right|^2 d\mathbf{x}, \qquad j = 1,2;$$

solving a two-equations system for the coefficients $\mathbf{a}_i^j$ and b_i^j. This can be regarded as the *local projection phase*. We thus build a piecewise linear approximation $\mathbf{g}_{h^\star} \in (X^1_{h^\star})^2$ of the gradient $\nabla u_{h^\star}$

$$[\mathbf{g}_{h^\star}(\mathbf{x})]^j = \sum_i \pi_i^j(\mathbf{x}_i)\varphi_i(\mathbf{x}), \qquad j = 1,2,$$

where the sums span over all the nodes N_i of the grid. Once the gradient is reconstructed, we can proceed in two different ways, depending on the type of reconstruction we want to obtain for the second derivatives. We recall the concept of *Hessian matrix* $D^2(u) = \nabla(\nabla u)$

$$[D^2(u)]_{i,j} = \frac{\partial^2 u}{\partial x_i \partial x_j}, \qquad i,j = 1,2.$$

A piecewise constant approximation of the latter is obtained by setting, for each $K^\star \in \mathcal{T}_{h^\star}$

$$D_h^2|_{K^\star} = \frac{1}{2}\left(\nabla \mathbf{g}_h + (\nabla \mathbf{g}_h)^T\right)\Big|_{K^\star}.$$

Notice the use of the symmetric form of the gradient, which is necessary for Hessian symmetry.
We are now able to compute an approximation of $|u|_{\mathcal{H}^2(K^\star)}$ on a generic triangle $K^\star \in \mathcal{T}_{h^\star}$. From (7.2) we deduce that, to obtain the approximate solution u_h with an error smaller or equal to a tolerance ε, the new grid $\mathcal{T}_h^{NEW}$ must satisfy

$$\sum_{K \in \mathcal{T}_h^{NEW}} h_K^2 |u|^2_{\mathcal{H}^2(K)} \approx \sum_{K \in \mathcal{T}_h^{NEW}} h_k^2 \sum_{i,j=1}^{2} \left\| [D_h^2]_{i,j} \right\|^2_{L^2(K)} \le \left(\frac{\varepsilon}{C}\right)^2.$$

Idelly, one would wish the error to be equidistributed on each element K of the new grid.
A possible adaptation procedure then consist in generating the new grid by appropriately partitioning all of the $N^\star$ triangles $K^\star$ of $\mathcal{T}_{h^\star}$ for which we have

$$\eta^2_{K^\star} = h^2_{K^\star} \sum_{i,j=1}^{2} \left\| [D_h^2]_{i,j} \right\|^2_{L^2(K^\star)} > \frac{1}{N^\star}\left(\frac{\varepsilon}{C}\right)^2. \tag{7.3}$$

This method is said to be a *refinement* and it only aims at creating a finer grid than the initial one, but it clearly does not allow to fully satisfy the equidistribution condition.
More sophisticated algorithms also allow to derefine the grid in which the inequality (7.3) is verified with '≪' instead of '>'. However, derefinement is more difficult to implement than refinement; hence one often prefers to construct the new grid from scratch (procedure named *remeshing*). For this

purpose, on the basis of error estimate, the following *spacing function H* (constant on each element) is introduced

$$H|_{K^\star} = \frac{\varepsilon}{C\sqrt{N^\star}\sqrt{\sum_{i,j=1}^{2}\left\|\left[D_h^2\right]_{i,j}\right\|_{L^2(K^\star)^2}}|u_{h^\star}|_{\mathscr{H}^2(K^\star)}}, \qquad \forall K^\star \in \mathscr{T}_{h^\star},$$

and is used to construct the adapted grid.
The adaptation can be repeated for the solution computed on the new grid, until inequality (7.3) is inverted on all the elements.

Remark. The C constant can be estimated by applying (7.2) to known functions. An alternative that does not require explicitly knowing C consist in realizing the grid that equally distributes the error for a number $N^\star$ of a priori fixed elements.

7.2 A posteriori adaptivity

The procedure just seen can be unsatisfactory because the reconstruction of u's derivatives starting from $u_{h^\star}$ is often subject to errors that are not easy to quantify.
A radical alternative consist in adopting *a posteriori estimates* of the error, that are obtained as functions of computable quantities, usually based on the so-called *residue* $R \in V'$ of the approximate solution defined by

$$\langle R, v\rangle = F(v) - a(u_h, v) = a(u - u_h, v), \qquad \forall v \in V.$$

Then

$$\alpha\, \|u - u_h\|_V \le \|R\|_{V'} \le M\, \|u - u_h\|_V .$$

Indeed, by continuity of $a(\cdot,\cdot)$,

$$\|R\|_{V'} = \sup_{v\in V} \frac{\langle R, v\rangle}{\|v\|_V} \le M\, \|u - u_h\|_V .$$

On the other hand, taking $v = u - u_h$ and using the coercivity of $a(\cdot,\cdot)$

$$\alpha\, \|u - u_h\|_V^2 \le a(u - u_h, u - u_h) = \langle R, u - u_h\rangle \le \|R\|_{V'}\, \|u - u_h\|_V .$$

Now our goal is to express R in terms of computable quantities on every element K of the finite element triangulation.
For the sake of exposition, let us consider the Poisson problem, where $V_h = \mathring{X}_h^r$. In this specific case, $V = \mathscr{H}_0^1(\Omega)$, $V' = \mathscr{H}^{-1}(\Omega)$, $\alpha = M = 1$. For every $v \in \mathscr{H}_0^1(\Omega)$, $v_h \in V_h$, we have

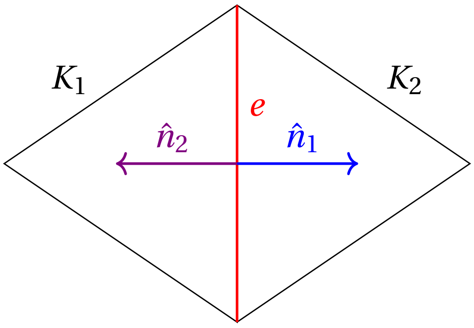
$$\begin{aligned}
\langle R, v\rangle &= \int_\Omega \nabla(u - u_h)\cdot\nabla v \, d\Omega \\
&= \int_\Omega \nabla(u - u_h)\cdot\nabla(v - v_h)\, d\Omega \\
&= \int_\Omega \underbrace{f(v - v_h)}_{\nabla u\cdot\nabla(v - v_h)}\, d\Omega - \int_\Omega \nabla u_h\cdot\nabla(v - v_h)\, d\Omega \\
&= \int_\Omega f(v - v_h)\, d\Omega + \sum_{K\in\mathscr{T}_h}\int_K \Delta u_h(v - v_h)\, d\Omega - \sum_{K\in\mathscr{T}_h}\int_{\partial K}\frac{\partial u_h}{\partial n}(v - v_h)\, d\gamma \\
&= \sum_{K\in\mathscr{T}_h}\int_K (f + \Delta u_h)(v - v_h)\, d\Omega - \sum_{K\in\mathscr{T}_h}\int_{\partial K}\frac{\partial u_h}{\partial n}(v - v_h)\, d\gamma.
\end{aligned}$$

We remark that all the local integrals make sense.

Having denoted by e a side of the generic triangle K, we define the *jump* of the normal derivative of u_h through the internal side e the quantity

$$\left[\frac{\partial u_h}{\partial n}\right]_e = \nabla u_h|_{K_1}\cdot \hat{n}_1 + \nabla u_h|_{K_2}\cdot\hat{n}_2$$
$$= \left(\nabla u_h|_{K_1} - \nabla u_h|_{K_2}\right)\cdot\hat{n}_1,$$

where K_1 and K_2 are the two triangles sharing e and $\hat{n}_1$, $\hat{n}_2 = -\hat{n}_1$ are the normal outgoing vectors. In order to extend such definition also to the boundary sides, we introduce the so-called *generalized jump*



$$\left[\frac{\partial u_h}{\partial n}\right] = \begin{cases}\left[\frac{\partial u_h}{\partial n}\right]_e & \text{if } e \in \mathscr{E}_\delta\\ 0 & \text{if } e\in\partial\Omega\end{cases}$$

Thank to this, we can therefore write

$$-\sum_{K\in\mathscr{T}_h}\int_{\partial K}\frac{\partial u_h}{\partial n}(v-v_h)\,d\gamma = -\sum_{K\in\mathscr{T}_h}\sum_{e\in\partial K}\int_e\frac{\partial u_h}{\partial n}(v-v_h)\,d\gamma$$
$$= -\sum_{K\in\mathscr{T}_h}\sum_{e\in\partial K}\frac{1}{2}\int_e\left[\frac{\partial u_h}{\partial n}\right](v-v_h)\,d\gamma$$
$$= -\frac{1}{2}\sum_{K\in\mathscr{T}_h}\int_{\partial K}\left[\frac{\partial u_h}{\partial n}(v-v_h)\right]d\gamma,$$

where the factor 1/2 takes into account that each internal side e of the grid is shared by tho elements. We now have (using the Cauchy-Schwarz inequality)

$$\langle R, v\rangle \le \sum_{K\in\mathscr{T}_h}\left(\|f+\Delta u_h\|_{L^2(K)}\|v-v_h\|_{L^2(K)} + \frac{1}{2}\left\|\left[\frac{\partial u_h}{\partial n}\right]\right\|_{L^2(\partial K)}\|v-v_h\|_{L^2(\partial K)}\right).$$

Now we look for $v_h \in V_h$ that allows to express the norm of $v-v_h$ as a function of a well chosen norm of v. Moreover, we want this norm to be “local”, i.e. computed over a region $\tilde{K}$ containing K, but as little as possible. If v where continuous, we could take as v_h the Lagrangian interpolant of v and use the previously cited interpolation error estimates on K. Unfortunately, in our case $v\in\mathscr{H}^1(\Omega)$ is not necessarily continuous. However, if $\mathscr{T}_h$ is a regular grid, we can introduce the so-called *Clément interpolation operator* $\mathscr{R}_h : \mathscr{H}^1(\Omega)\to V_h$ defined, in the case of linear finite elements, as

$$\mathscr{R}_h v(\mathbf{x}) = \sum_{N_j}(P_j v)(N_j)\varphi_j(\mathbf{x}), \qquad \forall v\in\mathscr{H}^1(\Omega),$$

where $P_j v$ denotes a local L^2 projection of v.
More precisely, it is a linear function defined on the patch K_{N_j} of the grid elements that share N_j, which is determined by the relation

$$\int_{K_{N_j}}(P_j v - v)\psi\,d\mathbf{x} = 0, \qquad \psi = 1, x, y.$$

As usual, the φ_j are the characteristic Lagrangian basis functions.
For each $v\in\mathscr{H}^1(\Omega)$ and each $K\in\mathscr{T}_h$, we have

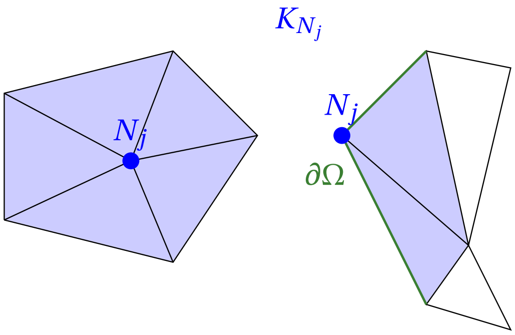
$$\|v-\mathscr{R}_h v\|_{L^2(K)} \le C_1 h_k |v|_{\mathscr{H}^1(\tilde{K})},$$
$$\|v-\mathscr{R}_h v\|_{L^2(\partial K)} \le C_2\sqrt{h_k}\,|v|_{\mathscr{H}^1(\tilde{K})},$$

where C_1, C_2 are positive constants that depend on the minimal angle of the elements of the triangulation, while

$$\tilde{K} = \left\{K_j\in\mathscr{T}_h : K_j\cap K\ne\emptyset\right\}.$$

By choosing $v_h = \mathscr{R}_h v$, $C = \max\{C_1, C_2\}$ and using the Cauchy-Schwarz inequality, we obtain

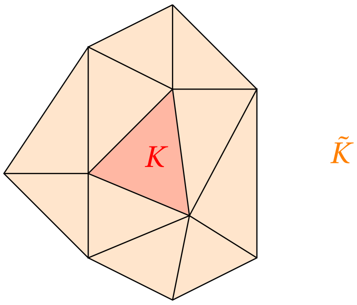
$$\begin{aligned}\langle R, v\rangle &\leq C \sum_{K\in\mathscr{T}_h} \rho_K(u_h)\,\|v\|_{\mathscr{H}^1(\tilde{K})}\\ &\leq C\sqrt{\sum_{K\in\mathscr{T}_h}\left[\rho_K(u_h)\right]^2}\sqrt{\sum_{K\in\mathscr{T}_h}\|v\|^2_{\mathscr{H}^1(\tilde{K}}}\end{aligned}$$



where

$$\rho_K(u_h) = h_k\,\|f + \Delta u_h\|_{L^2(K)} + \frac{1}{2}\sqrt{h_K}\left\|\left[\frac{\partial u_h}{\partial n}\right]\right\|_{L^2(\partial K)}$$

is the so-called *local residue*, composed by *internal residue* and *external residue*.



Since $\mathscr{T}_h$ is regular, the number of elements in $\tilde{K}$ is bounded by a certain $n \in \mathbb{N}$ independent of h. Thus

$$\|v\|_{\mathscr{H}^1(\Omega)} \leq \sqrt{\sum_{K\in\mathscr{T}_h}\|v\|^2_{\mathscr{H}^1(\tilde{K})}} \leq \sqrt{n}\,\|v\|_{\mathscr{H}^1(\Omega)}.$$

By Poincaré's inequality,

$$\|v\|_{L^2(\Omega)} \leq C\,\|v\|_{\mathscr{H}^1_0(\Omega)}, \qquad C = \sqrt{1 + C_\Omega^2},$$

whence

$$\|\mathscr{R}\|_{\mathscr{H}^{-1}(\Omega)} = \sup_{v\in\mathscr{H}^1_0(\Omega)} \frac{\langle\mathscr{R}, v\rangle}{\|v\|_{\mathscr{H}^1_0(\Omega)}} \leq C\sqrt{n}\sqrt{\sum_{K\in\mathscr{T}_h}\left[\rho_K(u_h)\right]^2}.$$

We conclude (since $\alpha = 1$ in this case) with the following *residual-based* a posteriori error estimate

$$\|u - u_h\|_{\mathscr{H}^1(\Omega)} \leq C\sqrt{n}\sqrt{\sum_{K\in\mathscr{T}_h}\left[\rho_K(u_h)\right]^2}.$$

Note that $\rho_K(u_h)$ is computable, since it is a function of the datum f and of h_K and u_h. The most delicate point is the estimate of C and n.
The a posteriori estimate can be used in order to guarantee

$$\frac{\varepsilon}{2} \leq \frac{\|u - u_h\|_{\mathscr{H}^1(\Omega)}}{\|u_h\|_{\mathscr{H}^1(\Omega)}} \leq \frac{3\varepsilon}{2}.$$

Chapter 8

Parabolic equations

We consider the parabolic equation of the form

$$\frac{\partial u}{\partial t} + Lu = f, \qquad x \in \Omega, \quad t > 0, \tag{8.1}$$

where Ω is a domain of $\mathbb{R}^d$, $d = 1,2,3$, $f = f(\mathbf{x}, t)$ is a given function and L is a generic elliptic operator. When solved only for a bounded temporal interval $0 < t < T$, the region $Q_T = \Omega \times (0, T)$ is called *cylinder* in the space $\mathbb{R}^d \times \mathbb{R}^+$. If $T = \infty$, $Q = \{(x, t) : x \in \Omega, t > 0\}$ is an infinite cylinder.
Equation (8.1) must be completed by assigning an initial condition $u(x,0) = u_0(x)$, $x \in \Omega$, together with boundary conditions which can take the following form:

$$\begin{aligned} u(x,t) &= \varphi(x,t), & x \in \Gamma^D, \quad t > 0, \\ \partial_n u(x,t) &= \psi(x,t), & x \in \Gamma^N, \quad t > 0, \end{aligned}$$

where u_0, φ and ψ are given functions and $\{\Gamma^D, \Gamma^N\}$ provides a boundary partition:

$$\Gamma^D \cup \Gamma^N = \partial\Omega, \qquad \mathring{\Gamma}^D \cap \mathring{\Gamma}^N = \emptyset.$$

8.0.1 Weak formulation and its approximation

We proceed formally by multiplying for each $t > 0$ the differential equation by a test function $v = v(\mathbf{x})$ and integrating on Ω. We set $V = \mathscr{H}^1_{\Gamma^D}(\Omega)$ and for each $t > 0$ we seek $u(t)\ inV$ such that

$$\int_\Omega \frac{\partial u(t)}{\partial t} v \, d\Omega + a(u(t), v) = \int_\Omega f(t) v \, d\Omega, \qquad \forall v \in V,$$

where $u(0) = u_0$, $a(\cdot,\cdot)$ is the bilinear form associated to L and, fro simplicity $\varphi = \psi = 0$. A sufficient condition for existence and uniqueness of the solution is that the following hypotheses hold:

- $a(\cdot,\cdot)$ continuous and weakly coercive: there exists $\lambda \geq 0$, $\alpha > 0$ such that $a(v,v) + \lambda \|v\|^2_{L^2(\Omega)} \geq \alpha \|v\|^2_V$ for any $v \in V$;
- $u_0 \in L^2(\Omega)$;
- $f \in L^2(Q)$.

Then the problem admits a unique solution $u \in L^2(\mathbb{R}^+; V) \cap \mathscr{C}^0(\mathbb{R}^0; L^2(\Omega))$ with $\mathscr{H}^1_{\Gamma^D}(\Omega)$.

We now consider the Galerkin approximation of the (weak) problem: for each $t > 0$, find $u_h(t) \in V_h$ such that

$$\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + a(u_h(t), v_h) = \int_\Omega f(t) v_h \, d\Omega, \qquad \forall v_h \in V_h, \tag{8.2}$$

and $u_h(0) = u_{0h}$ is an approximation pf u_0 in V_h finite dimensiona subspace of V. Such problem is called *semi-discretization* since the temporal variable has not yet been discretized.
To provide an algeraic interpretation of (8.2) we introduce a basis $\{\varphi_j\}$ for V_h and we observe that it suffices that (8.2) is verified for the basis functions in order to be satisfied by all the functions. Moreover, since for each $t > 0$ the solution to the Galerkin problem belongs to the subspace as well, we will have

$$u_h(\mathbf{x}, t) = \sum_{j=1}^{N_h} u_j(t)\varphi_j(\mathbf{x}),$$

where the coefficients $\{u_j(t)\}$ represent the unknowns of (8.2).

denoting by $\dot{u}_j(t)$ the time derivative of $u_j(t)$, (8.2) becomes

$$\int_\Omega \sum_{j=1}^{N_h} \dot{u}_j(t)\varphi_j\varphi_i \, d\Omega + a\left(\sum_{j=1}^{N_h} u_j(t)\varphi_j(t), \varphi_i\right) = \int_\Omega f(t)\varphi_i \, d\Omega, \qquad i = 1, 2, \ldots, N_h,$$

that is

$$\sum_{j=1}^{N_h} \dot{u}_j(t) \int_\Omega \varphi_j\varphi_i \, d\Omega + \sum_{j=1}^{N_h} u_j(t) a(\varphi_j, \varphi_i) = \int_\Omega f(t)\varphi_i \, d\Omega, \qquad i = 1, 2, \ldots, N_h,$$

i.e.

$$\sum_{j=1}^{N_h} \dot{u}_j(t) m_{i,j} + \sum_{j=1}^{N_h} u_j(t) a_{i,j} = f_i(t), \qquad i = 1, 2, \ldots, N_h.$$

If we define the vector of unknowns $u = [u_1(t), u_2(t), \ldots, u_{N_h}(t)]^T$, the mass matrix $M = [m_{i,j}]$, the mass matrix $A = [a_{i,j}]$ and the right hand side vector $f = [f_1(t), f_2(t), \ldots, f_{N_h}(t)]^T$, we have the following matrix form

$$M\dot{u}(t) + Au(t) = f(t).$$

For the solution of this ODE, we focus on the $\theta-$method. The latter discretizes the temporal derivative by a simple different quotient and replace the other terms with a linear combination of the values at time t^k and of the time t^{k+1}, depending on the real parameter $\theta \in [0, 1]$:

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

Let us see some particular cases

$\theta = 0$: forward Euler, order 1 with respect to Δt;

$\theta = 1$: backward Euler, order 1 with respect to Δt;

$\theta = 1/2$: Crank-Nicolson, order 2 with respect to Δt.

In the case $\theta = 0$, if we make M diagonal, we actually decouple the system. This operation is performed by the so-called lumping of the mass matrix. However, this scheme is not unconditionally stable and in the case where V_h is a subspace of finite elements we have the following stability condition: there exists $c > 0$ s.t. $\Delta t < ch^2$ for any $h > 0$. So Δt cannot be chosen irrespective of h.

In case $\theta > 0$, the system will have the form $Ku^{k+1} = g$, where g is the source term and $K = M/\Delta t + \theta A$. Such a matrix is invariant in time (the operator L, and therefore the matrix A, do not depend on time). If the mesh does not change, K can be factorized once for all.

8.1 A priori estimates

In the weak problem, since the corresponding equations must hold for each $v \in V$, in particular hold for $v = u(t)$, yielding

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t)\, d\Omega + a(u(t), u(t)) = \int_\Omega f(t) u(t)\, d\Omega, \qquad \forall t > 0.$$

Concerning the individual terms,

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t)\, d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\left(\int_\Omega |u(t)|^2\, d\Omega\right) = \frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|^2_{L^2(\Omega)}.$$

If we assume $a(\cdot,\cdot)$ to be coercive,

$$a(u(t), u(t)) \geq \alpha \|u(t)\|_V^2.$$

Finally, by the Cauchy-Schwarz inequality,

$$\langle f(t), u(t)\rangle_{L^2(\Omega)} \leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)}.$$

Let us recall the Young's inequality

$$ab \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2, \qquad \forall a, b \in \mathbb{R}, \quad \varepsilon > 0.$$

Now, via Poincaré's and Young's inequality we get

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha\|\nabla u(t)\|^2_{L^2(\Omega)} &\leq \|f(t)\|_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)} \\ &\leq \frac{C_\Omega^2}{2\alpha}\|f(t)\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\nabla u(t)\|^2_{L^2(\Omega)}. \end{aligned} \tag{8.3}$$

Then, by integrating in time we obtain, for all $t > 0$, the following a priori energy estimate

$$\|u(t)\|^2_{L^2(\Omega)} + \alpha\int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)}\, ds \leq \|u_0\|^2_{L^2(\Omega)} + \frac{C_\Omega^2}{\alpha}\int_0^t \|f(s)\|^2_{L^2(\Omega)}\, ds.$$

Different kind of a priori estimates can be obtained as follows. Note that

$$\frac{1}{2}\frac{d}{dt}\|u\|^2_{L^2} = \|u\|_{L^2}\frac{d}{dt}\|u\|_{L^2},$$

then

$$\|u(t)\|_{L^2}\frac{d}{dt}\|u(t)\|_{L^2} + \frac{\alpha}{C_\Omega}\|u(t)\|_{L^2}\|\nabla u(t)\|_{L^2} \leq \|f(t)\|_{L^2}\|u(t)\|_{L^2}, \qquad t > 0.$$

If $\|u(t)\|_{L^2} \neq 0$, we divide by it and integrate in time to obtain

$$\|u(t)\|_{L^2} \leq \|u_0\|_{L^2} + \int_0^t \|f(s)\|_{L^2}\, ds, \qquad t > 0.$$

Let us now use (8.3) and integrate in time to yield

$$\begin{aligned} \|u(t)\|_{L^2} + 2\alpha\int_0^t \|\nabla u(s)\|^2_{L^2}\, ds &\leq \|u_0\|^2_{L^2} + 2\int_0^t \|f(s)\|_{L^2}\|u(s)\|_{L^2}\, ds \\ &\leq \|u_0\|_{L^2} + 2\int_0^t \|f(s)\|_{L^2}\left(\|u_0\|_{L^2} + \int_0^s \|f(\tau)\|_{L^2}\, d\tau\right) ds \\ &= \|u_0\|_{L^2} + 2\int_0^t \|f(s)\|_{L^2}\|u_0\|_{L^2} + 2\int_0^t \|f(s)\|_{L^2}\int_0^s \|f(\tau)\|_{L^2}\, d\tau\, ds \\ &= \left(\|u_0\|_{L^2} + \int_0^t \|f(s)\|_{L^2}\, ds\right)^2. \end{aligned}$$

The latter equality follows upon noticing that

$$\|f(s)\|_{L^2}\int_0^s \|f(\tau)\|_{L^2}\,d\tau = \frac{d}{ds}\left(\int_0^s \|f(\tau)\|_{L^2}\,d\tau\right)^2.$$

We therefore conclude with the additional a priori estimate

$$\sqrt{\|u(t)\|_{L^2}^2 + 2\alpha\int_0^t \|\nabla u(s)\|_{L^2}^2\,ds} \le \|u_0\|_{L^2} + \int_0^t \|f(s)\|_{L^2}\,ds, \qquad t>0.$$

We have the following a priori (stability) estimate for the solution to problem (8.2)

$$\|u_h(t)\|_{L^2}^2 + \alpha\int_0^t \|\nabla u_h(s)\|_{L^2}^2\,ds \le \|u_{0h}\|_{L^2}^2 + \frac{C_\Omega^2}{\alpha}\int_0^t \|f(s)\|_{L^2}^2\,ds, \qquad t>0.$$