# Some notes on
# Advanced Numerical Analysis II (FEM)

Marco Caliari

a.y. 2016/17

These are the unofficial, not reliable, not complete and not correct notes for Advanced Numerical Analysis II (FEM). Use at your own risk.

# Contents

# Chapter 1

# Sobolev spaces

Some very nice examples are in [2].

## 1.1 $H^m(\Omega)$

### 1.1.1 $\Omega \subseteq \mathbb{R}$

$H^1(\Omega)$

We can define $H^1(\Omega)$ in the following way: it is the subspace of $L^2(\Omega)$ of functions $u$ for which there exists $g \in L^2(\Omega)$ such that

$$\int_\Omega u\varphi' = -\int_\Omega g\varphi \quad \forall \varphi \in C_c^\infty(\Omega)$$

We will denote $g$ by $u'$. This definition is equivalent to the definition with distributional derivatives.

**Theorem 1.** *If $u \in H^1(\Omega)$, there exists (unique) $\tilde{u} \in C(\bar{\Omega})$ such that*

$$u = \tilde{u} \textit{ almost everywhere}$$

*and*

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

We will call $\tilde{u}$ the continuous representative of the class of equivalence of $u$. We will often indicate it simply by $u$ when necessary. For instance, if we want to write $u(x_0)$, $x_0 \in \Omega$. The scalar product in $H^1(\Omega)$ is

$$(u, v) = \int_\Omega uv + \int_\Omega u'v'$$

with the inducted norm.

$H^m(\Omega)$

We can define $H^m(\Omega)$ in the following way: it is the subspace of $L^2(\Omega)$ of functions $u$ for which there exists $g1, g_2, \dots, g_m \in L^2(\Omega)$ such that

$$\int_\Omega u\varphi^{(j)} = (-1)^j \int_\Omega g_j \varphi \quad \forall \varphi \in C_c^\infty(\Omega)$$

We will denote $g_j$ by $u^{(j)}$ $(u', u'', \dots, u^{(m)})$.

$H_0^1(\Omega)$

$H_0^1(\Omega)$ is the closure of $C_c^1(\Omega)$ in $H^1(\Omega)$. If $\Omega = \mathbb{R}$, then $H_0^1(\mathbb{R}) = H^1(\mathbb{R})$. Since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, its closure is $H_0^1(\Omega)$ itself.

**Theorem 2.** *If $u \in H^1(\Omega)$, then $u \in H_0^1(\Omega)$ if and only if $u = 0$ on $\partial\Omega$.*

If $\Omega = (a, b)$, this is precisely a case in which the function $u$ such that $u(a) = u(b) = 0$ is the continuous representative (of the class of equivalence) of $u$. Another way to characterize $H_0^1(\Omega)$ is the following: $u \in H_0^1(\Omega)$ if and only if $\bar{u} \in H^1(\mathbb{R})$, where $\bar{u}(x) = u(x)$ if $x \in \Omega$ and $\bar{u}(x) = 0$ if $x \in \mathbb{R} \setminus \Omega$.

### 1.1.2 $\Omega \subseteq \mathbb{R}^n$, $n > 1$

$H^1(\Omega)$

The definition is analogous, we have to replace the derivative with all the partial derivatives. One main difference with the one-dimensional case is that there are functions, like the following

$$u(x, y) = \left(\log \frac{1}{\sqrt{x^2 + y^2}}\right)^k, \quad 0 < k < 1/2$$

that belongs to $H^1(\Omega)$, $\Omega = B(0, 1) \subset \mathbb{R}^2$, but it is noway possible to find a continuous representative $\tilde{u}$ for it. So, Theorem 1 does not hold.

### 1.1.3 $H_0^1(\Omega)$

$H_0^1(\Omega)$ is the closure of $C_c^1(\Omega)$ in $H^1(\Omega)$. If $\Omega = \mathbb{R}^n$, then $H_0^1(\mathbb{R}^n) = H^1(\mathbb{R}^n)$. Since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, its closure is $H_0^1(\Omega)$ itself.

Now, Theorem 2 cannot be stated in the same way: in general, there is no continuous representative which is zero at $\partial\Omega$. Still, it is correct to think to functions in $H_0^1(\Omega)$ as to the functions in $H^1(\Omega)$ which "are zero at $\partial\Omega$". Let us see in which sense.

**Theorem 3.** *If $\partial\Omega$ is sufficiently regular and $u \in H^1(\Omega) \cap C(\bar{\Omega})$, then $u \in H^1_0(\Omega)$ if and only if $u = 0$ on $\partial\Omega$.*

Moreover, it is possible to characterize $H^1_0(\Omega)$ as above: $u \in H^1_0(\Omega)$ if and only if $\bar{u} \in H^1(\mathbb{R}^n)$, where $\bar{u}(x) = u(x)$ if $x \in \Omega$ and $\bar{u}(x) = 0$ if $x \in \mathbb{R} \setminus \Omega$.

**Theorem 4.** *If $\Omega$ is a bounded open subset of $\mathbb{R}^n$ with $\partial\Omega$ sufficiently regular, then there exists a unique linear continuous operator $T\colon H^1(\Omega) \to L^2(\partial\Omega)$ such that*

$$\begin{aligned} &Tu = u|_{\partial\Omega} && \text{if } u \in H^1(\Omega) \cap C(\bar{\Omega}) \\ &\|Tu\|_{L^2(\partial\Omega)} \le C\|u\|_{H^1(\Omega)} \end{aligned}$$

*The operator $T$ is called* trace operator *and the function $Tu \in L^2(\partial\Omega)$ is called* trace *of $u$ on $\partial\Omega$.*

It is in fact not necessary to consider $\partial\Omega$: if $\Gamma \subset \mathbb{R}^{n-1}$ is sufficiently regular, it is possibile to define the trace operator

$$T\colon H^1(\Omega) \to L^2(\Gamma)$$

in the same way. The operator $T$ is not surjective on $L^2(\partial\Omega)$. The set of functions in $L^2(\partial\Omega)$ which are traces of functions in $H^1(\Omega)$ is a subspace of $L^2(\partial\Omega)$ denoted by $H^{1/2}(\partial\Omega)$. We have $H^1(\partial\Omega) \subseteq H^{1/2}(\partial\Omega) \subseteq H^0(\partial\Omega) = L^2(\partial\Omega)$. If $u$ is more regular, so is $u|_{\partial\Omega}$ in the sense that

$$T\colon H^k(\Omega) \to H^{k-1/2}(\partial\Omega) \subseteq H^{k-1}(\partial\Omega)$$

Given $u \in H^1(\Omega)$ and $u^{\mathrm{D}} \in H^{1/2}(\partial\Omega)$, if we require that "$u = u^{\mathrm{D}}$ on $\partial\Omega$" or "$u|_{\partial\Omega} = u^{\mathrm{D}}$", we really mean that $Tu = u^{\mathrm{D}}$ (almost everywhere). We notice that for $n > 1$, the functions in $H^{1/2}(\partial\Omega)$ may be discontinuous. In fact, if you take, for $n = 2$,

$$u(x, y) = \left(\ln \frac{1}{\sqrt{x^2 + y^2}}\right)^k, \quad 0 < k < \frac{1}{2}$$

then $u \in H^1(B(0, 1))$ and it is not conitnuous. Now you can consider as domain half of the disk $B(0, 1)$: then the trace of $u$ at the diameter is still not a continuous function. On the other hand, functions in $H^{1/2}(\partial\Omega)$ cannot have jumps.

Finally, we can define

$$H^1_0(\Omega) = \ker(T) = \{u \in H^1(\Omega)\colon Tu = 0\}$$

We can consider the following *line source* for $\Gamma_\ell$ a line in $\Omega \subset \mathbb{R}^2$ and $v \in H^1_0(\Omega)$

$$\ell(v) = \int_{\Gamma_\ell} Tv \mathrm{d}\gamma$$

Clearly, $\ell\colon H^1 \to \mathbb{R}$ and

$$|\ell(v)| \leq \int_{\Gamma_\ell} |Tv| \,\mathrm{d}\gamma \overset{\text{C.–S.}}{\leq} C \,\|Tv\|_{L^2(\Gamma_\ell)} \leq C \,\|v\|_{H^1(\Omega)}$$

therefore it is bounded.

## 1.2 Embeddings (see [1, p. 85])

Given an open bounded set $\Omega \subset \mathbb{R}^n$, we define $\mathcal{C}^m(\bar{\Omega})$ as the subset of $\mathcal{C}^m(\Omega)$ of the functions $u$ for which $D^\alpha u$ is bounded and uniformly continuous on $\Omega$ for $0 \leq |\alpha| \leq m$. It is a Banach space with the norm

$$\|u\|_{\mathcal{C}^m(\bar{\Omega})} = \max_{0\leq|\alpha|\leq m} \sup_{x\in\Omega}|D^\alpha u(x)|$$

If $\Omega \subset \mathbb{R}^n$ is "sufficiently regular" and $m > n/2$, then

$$H^{j+m}(\Omega) \to \mathcal{C}^j(\bar{\Omega}), \quad j = 0, 1, \dots$$

We mean that if $u \in H^{j+m}(\Omega)$ there exists $\tilde{u} \in \mathcal{C}^j(\bar{\Omega})$ such that $\tilde{u} = u$ in $H^{j+m}(\Omega)$ and

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

# Chapter 2

# Triangulations

## 2.1 Quasi-uniform onedimensional discretization

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

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

where $h_{\min} = \min_{1\leq i\leq n+1}\{x_i - x_{i-1}\}$ and $\varepsilon$ does not depend on $n$. For instance, the discretization given by $x_i = i/(n+1)$ is, of course, quasi-uniform. If we consider $x_i = (i/(n + 1))^2$, we have

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

and therefore it is not quasi-uniform. If we consider $x_i = 1/2 - \cos(i\pi/(n + 1))/2$ we have

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

from which

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

and therefore it is not quasi-uniform.

Finally, if we consider a discretization given by $n + 1$ intervals of length

$$h_0, h_0 r, \ldots, h_0 r^n$$

with $r > 1$, we find that

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

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

## 2.2 Two-dimensional triangulations

We have a set of sites $S = \{s_1, s_2, \ldots, s_m\}$ in $\mathbb{R}^2$, assuming that no four sites are co-circular.

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

A *graph* is a set of sites and arcs (called edges) connecting some of the sites.

A *planar graph* is a graph with no intersecting edges.

A *planar straight line graph* is a planar graph whose edges are segments.

A *connected graph* is a graph in which any pairs of sites is connected by a finite sequence of edges.

The *convex hull* of a set of sites is the smallest convex set containing all the points.

The bisectors between pairs of sites are straight lines that partition the plane into $m$ convex regions, one corresponding to each site $s_i$ (by induction). 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 a *Voronoi diagram* of $S$. The vertices and the edges of the convex regions are called *Voronoi vertices* and *Voronoi edges.*

**Proposition 1.** *The number of Voronoi vertices is $2(m-1)-h$ and the number of Voronoi edges is $3(m-1)-h$ where $h$ is the number of vertices on 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 conncected graph made of Voronoi vertices, Voronoi edges and the additional edges for each uonbounded 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 points. In fact, there are three edges departing from any point and in this way any edge is counted twice. Then, we use Euler's formula $V - E + F = 2$ getting $V = 2(F-2)$ and $E = 3(F-2)$. With respect to Figure 2.1 it is $V = 8$, $E = 12$ and $F = 6$. 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 2.** *The circle with center a Voronoi vertex and passing through the (three) sites that defines its center has no other sites in its interior.*

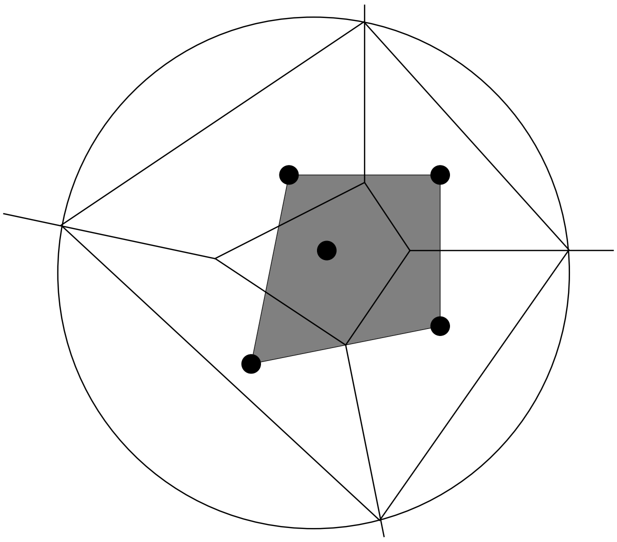


Figure 2.1: Voronoi's diagram.

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

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

A *triangulation* of sites is a set of straight line segments which intersect only at the sites and so that every region internal to the convex hull is a triangle.

**Proposition 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 triangles. 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.

## 2.3 Some other definitions

A family of triangulations $\mathcal{T}_h$ is said *regular* if there exists a constant $\delta > 0$, indipendent of $h$, such that

$$\frac{h_k}{\rho_k} \leq \delta, \quad \forall T_h^k \in \mathcal{T}_h$$

where $h_k$ is the diameter and $\rho_k$ the radius of the inscribed circle of the triangle $T_h^k$. Regularity excludes very deformed triangles, that is with small angles. In this sense, Delaunay triangulations are optimal.

### 2.3.1 Minimum angle of a triangle of a regular triangulation

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

$$A = (a+b+c)\rho = p\rho$$

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

$$\sin\frac{\alpha}{2} = \sqrt{\frac{1-\cos\alpha}{2}}, \quad \cos\frac{\alpha}{2} = \sqrt{\frac{1+\cos\alpha}{2}}, \quad \cos\alpha = \frac{b^2+c^2-a^2}{2bc}$$

we get

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

and therefore

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

Then

$$\frac{h}{(p-a)\tan\frac{\alpha}{2}} \leq \delta$$

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

$$\tan\frac{\alpha}{2} \geq \frac{2}{\delta}$$

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 insists with an angle (bounded as seen) on one of the three vertices. Therefore, the number of adjacent triangles is

$$n < \frac{5\pi}{2\arctan\frac{2}{\delta}} - 3$$

The number $5\pi$ is the sum of external angles of a triangle and $-3$ comes from the fact that the three triangle sharing an edge insist with two angles.

### 2.3.2 Constrained triangulations

A *constrained Delaunay triangulation* of a planar straight line graph is a triangulation in which each segment of the graph is present as a single edge in 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 also inserted 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.

## 2.4 Algorithms

There are several algorithms to make a Delaunay triangulation. It holds the folowing

**Proposition 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.*

See, for instance, the code [13].

# Chapter 3

# Strong, weak, distributional, $\delta$

## 3.1 Strong or distributional?

We may think to

$$-u_{xx} = f, \quad x \in \Omega = (0,1) \tag{3.1}$$

(which is usually called *strong formulation*) in a distributional sense. Let $v \in C_{\mathrm{c}}^{\infty}(\Omega)$ and $u \in H_0^1(\Omega)$. Then $L_u$ defined by

$$L_u(v) = \int_0^1 uv$$

is a distribution (it belongs to $(L^2(\Omega))'$, too). As usual, we identify $u$ and $L_u$. Since $u \in H_0^1(\Omega)$, it has a distributional derivative $u_x \in L^2(\Omega)$ such that

$$\langle D^1 L_u, v\rangle = -\langle L_u, v_x\rangle = -\int_0^1 uv_x = \int_0^1 u_x v$$

We consider now the distribution $L_{u_x}$ applied to $v_x$

$$\langle L_{u_x}, v_x\rangle = \int_0^1 u_x v_x = -\int_0^1 uv_{xx} = \langle -D^2 L_u, v\rangle$$

Moreover, if $f \in L^2(\Omega)$ we have

$$\int_0^1 fv = \langle L_f, v\rangle$$

Therefore, from the weak formulation (valid for $u \in H_0^1(\Omega)$ and $f \in L^2(\Omega)$) we have

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

of which (3.1) is just a short notation. This means also that, if $f \in L^2(\Omega)$, then $u \in H^2(\Omega)$. In this sense, we can also write

$$-u_{xx} = 4\delta(x - 1/2)$$

where we mean

$$\langle -D^2 L_u, v\rangle = \langle L_{\delta_{\frac{1}{2}}}, v\rangle = v\left(\tfrac{1}{2}\right)$$

## 3.2 FD for Poisson problem with $\delta$

In order to use Finite Differences with

$$-u_{xx} = 4\delta(x - 1/2), \quad x \in (0, 1)$$

we have to split the domain into its left and right parts. We get, easily,

$$\begin{aligned}
-u^{\mathrm{L}}_{xx} &= 0, & 0 < x < \frac{1}{2} \\
-u^{\mathrm{R}}_{xx} &= 0, & \frac{1}{2} < x < 1 \\
u^{\mathrm{L}}(0) &= u^{\mathrm{R}}(0) \\
u^{\mathrm{L}}\left(\tfrac{1}{2}\right) &= u^{\mathrm{R}}\left(\tfrac{1}{2}\right)
\end{aligned}$$

and then we take the weak form

$$\int_0^1 u_x v_x = 4v\left(\tfrac{1}{2}\right)$$

split the integral and integrate by parts

$$vu^{\mathrm{L}}_x\Big|_0^{\frac{1}{2}} - \int_0^{\frac{1}{2}} u^{\mathrm{L}}_{xx} v_x + vu^{\mathrm{R}}_x\Big|_{\frac{1}{2}}^1 - \int_{\frac{1}{2}}^1 u^{\mathrm{R}}_{xx} v_x = v\left(\tfrac{1}{2}\right) u^{\mathrm{L}}_x\left(\tfrac{1}{2}\right) - v\left(\tfrac{1}{2}\right) u^{\mathrm{R}}_x\left(\tfrac{1}{2}\right) = 4v\left(\tfrac{1}{2}\right)$$

from which

$$u^{\mathrm{L}}_x\left(\tfrac{1}{2}\right) - u^{\mathrm{R}}_x\left(\tfrac{1}{2}\right) = 4$$

```
m = 11; % odd number, at least 5
x = linspace(0,1,m)';
h = 1/(m-1);
A = toeplitz(sparse([1,1],[1,2],[2,-1]/h^2,1,m+1));
% the unknown is u = [uL;uR]
A(1,1:2) = [2/h^2,0]; % uL(0)=0
A((m+1)/2,:) = 0;
```

```
A((m+1)/2,(m+1)/2:(m+1)/2+1) = [-1,1]; % uL(1/2)=uR(1/2)
A((m+1)/2+1,:) = 0;
A((m+1)/2+1,(m+1)/2-2:(m+1)/2+3) = [1,-4,3,3,-4,1]/(2*h); % uL'(1/2)-uR'(1/2)=4
A(m+1,m:m+1) = [0,2/h^2]; % uR(1)=0
rhs = zeros(m+1,1);
rhs((m+1)/2+1) = 4;
u=A\rhs;
uL = u(1:(m+1)/2);
uR = u((m+1)/2+1:m+1);
plot(x(1:(m+1)/2),uL,'*',x((m+1)/2:m),uR,'o')
```

# Chapter 4

# Approximation theory

See [7].

## 4.1 One dimension

If $w \in H_0^1(0,1)$ and $w|_{T_h^k} \in \mathcal{C}^2(T_h^k)$ for each $T_h^k \in \mathcal{T}_h$, then by Rolle's theorem

$$\begin{aligned} \left|(w-\mathcal{I}_h w)'|_{T_h^k}(x)\right| = \left|\int_{z_k}^x (w-\mathcal{I}_h w)''|_{T_h^k}(t)\mathrm{d}t\right| = \left|\int_{z_k}^x w''|_{T_h^k}(t)\mathrm{d}t\right| \leq \\ \leq h_k \max_{x\in T_h^k}|w''(x)| \end{aligned} \tag{4.1}$$

By the way we have

$$\left|(w-\mathcal{I}_h w)'(x)|_{T_h^k}\right| \leq \int_{x_{k-1}}^{x_k} |w''(x)|\mathrm{d}x$$

which is true even if $w \in H^2(T_h^k)$, since $w'$ is *absolutely continuous.* Hence

$$|w-\mathcal{I}_h w|_{H^1}^2 = \sum_{k=1}^K \int_{T_h^k} |(w-\mathcal{I}_h w)'(x)|^2 \mathrm{d}x \leq \sum_{k=1}^K h_k \left(h_k \max_{x\in T_h^k}|w''(x)|\right)^2$$

If we take $h = \max_{1\leq k\leq K} h_k$, then

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

and hence

$$|w-\mathcal{I}_h w|_{H^1} \leq h \max_{1\leq k\leq K}\max_{x\in T_h^k}|w''(x)|$$

Then we can write

$$(w-\mathcal{I}_h w)|_{T_h^k}(x)=\int_{x_{k-1}}^{x}(w-\mathcal{I}_h w)'|_{T_h^k}(t)\mathrm{d}t \overset{(4.1)}{\leq} h_k\cdot h_k\max_{x\in T_h^k}|w''(x)|$$

and therefore

$$\|w-\mathcal{I}_h w\|_{L^2}^2=\sum_{k=1}^{K}\int_{T_h^k}|(w-\mathcal{I}_h w)(x)|^2\mathrm{d}x\leq\sum_{k=1}^{K}h_k\left(h_k^2\max_{x\in T_h^k}|w''(x)|\right)^2$$

and hence

$$\|w-\mathcal{I}_h w\|_{L^2}\leq h^2\max_{1\leq k\leq K}\max_{x\in T_h^k}|w''(x)|$$

Finally

$$\|w-\mathcal{I}_h w\|_{H^1(\Omega)}\leq h\sqrt{1+h^2}\max_{1\leq k\leq K}\max_{x\in T_h^k}|w''(x)|$$

Now we introduce the space

$$H^2(\Omega,\mathcal{T}_h)=\{w\in H_0^1(\Omega)\colon w|_{T_h^k}\in H^2(T_h^k)\ \forall k=1,\dots,K\}$$

with the *broken* seminorm and norm

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

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

We have

$$\begin{aligned}\left|(w-\mathcal{I}_h w)'|_{T_h^k}(x)\right|&\leq\int_{x_{k-1}}^{x_k}|w''(x)|\mathrm{d}x\overset{\text{C.–S.}}{\leq}\left(\int_{x_{k-1}}^{x_k}1^2\mathrm{d}x\right)^{1/2}\left(\int_{x_{k-1}}^{x_k}|w''|^2\mathrm{d}x\right)^{1/2}\leq\\&\leq h_k^{1/2}\left(\int_{x_{k-1}}^{x_k}|w''|^2\mathrm{d}x\right)^{1/2}\end{aligned}$$

and therefore

$$\begin{aligned}\int_{x_{k-1}}^{x_k}\left|(w-\mathcal{I}_h w)'(x)\right|^2\mathrm{d}x&\leq\int_{x_{k-1}}^{x_k}\left(h_k\int_{x_{k-1}}^{x_k}|w''(x)|^2\mathrm{d}x\right)\mathrm{d}x=\\&=h_k^2\int_{x_{k-1}}^{x_k}|w''(x)|^2\mathrm{d}x\end{aligned}\tag{4.2}$$

Hence

$$|w-\mathcal{I}_h w|_{H^1(\Omega)} \leq \sqrt{\sum_{k=1}^{K} h_k^2 |w|^2_{H^2(T_h^k)}} \leq h|w|_{H^2(\Omega,\mathcal{T}_h)} \leq h\|w\|_{H^2(\Omega,\mathcal{T}_h)}$$

Moreover

$$(w-\mathcal{I}_h w)|_{T_h^k}(x) = \int_{x_{k-1}}^{x} (w-\mathcal{I}_h w)'(t)\mathrm{d}t$$

and therefore

$$\left|(w-\mathcal{I}_h w)|_{T_h^k}(x)\right| \leq \int_{x_{k-1}}^{x_k} |(w-\mathcal{I}_h w)'(x)|\,\mathrm{d}x \overset{\text{C.–S.}}{\leq} h_k^{1/2}\left(\int_{x_{k-1}}^{x_k} |(w-\mathcal{I}_h w)'(x)|^2\mathrm{d}x\right)^{1/2} \overset{(4.2)}{\leq}$$

$$\leq h_k^{1/2}\left(h_k^2\int_{x_{k-1}}^{x_k} |w''(x)|^2\mathrm{d}x\right)^{1/2}$$

Hence

$$\int_{x_{k-1}}^{x_k} |w-\mathcal{I}_h w|^2\,\mathrm{d}x \leq h_k h_k^3 \int_{x_{k-1}}^{x_k} |w''(x)|^2\mathrm{d}x$$

and

$$\|w-\mathcal{I}_h w\|_{L^2} \leq \sqrt{\sum_{k=1}^{K} h_k^4 |w|^2_{H^2(T_h^k)}} \leq h^2|w|_{H^2(\Omega,\mathcal{T}_h)} \leq h^2\|w\|_{H^2(\Omega,\mathcal{T}_h)}$$

Finally,

$$\|w-\mathcal{I}_h w\|_{H^1(\Omega)} \leq \sqrt{\sum_{k=1}^{K} h_k^2(1+h_k^2) |w|^2_{H^2(T_h^k)}} \leq h\sqrt{1+h^2}|w|_{H^2(\Omega,\mathcal{T}_h)} \leq$$

$$\leq h\sqrt{1+h^2}\|w\|_{H^2(\Omega,\mathcal{T}_h)}$$

### 4.1.1 Example

If we consider

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

with

$$f_n = \begin{cases} 0 & x \leq \frac{1}{2} - \frac{1}{n} \\ 2n & \frac{1}{2} - \frac{1}{n} < x < \frac{1}{2} + \frac{1}{n} \\ 0 & x \geq \frac{1}{2} + \frac{1}{n} \end{cases}$$

then the corresponding solution $u_n \in \mathcal{C}^2(T_h^k)$ if the points $\frac{1}{2} \pm \frac{1}{n}$ are discretization points (even if $n \to \infty$), while $u_n \in H^2(\Omega)$ (and therefore $u_n \in H^2(\Omega, \mathcal{T}_h)$) for any set of discretization point, but, if $n \to \infty$, it is only in the space $H^2(\Omega, \mathcal{T}_h)$ if $\frac{1}{2}$ is a discretization point.

### 4.1.2 Nodal superconvergence

For the one-dimensional Poisson problem

$$a(u, v) = \int_0^1 u_x v_x = \ell(v)$$

we have $u_h = \mathcal{I}_h u$, where $u_h$ is the solution of

$$a(u_h, \varphi_i) = \ell(\varphi_i), \ \forall \varphi_i$$

In fact

$$\begin{aligned}
0 = \int_0^1 (u_h(x) - u(x))' \varphi_i'(x) \mathrm{d}x &= \sum_{k=1}^{K} \int_{T_h^k} (u_h(x) - u(x))' \varphi_i'(x) \mathrm{d}x = \\
= \sum_{k=1}^{K} (u_h(x) - u(x)) \varphi_i'(x) |_{x_{k-1}}^{x_k} &- \int_{T_h^k} (u_h(x) - u(x)) \varphi_i''(x) \mathrm{d}x = \\
&= \frac{(u_h(x_i) - u(x_i)) - (u_h(x_{i-1}) - u(x_{i-1}))}{h_i} + \\
&\quad - \frac{(u_h(x_{i+1}) - u(x_{i+1})) - (u_h(x_i) - u(x_i))}{h_{i+1}}
\end{aligned}$$

and therefore the vector $u_h(x_i) - u(x_i)$ satisfies the linear system

$$A_h \begin{bmatrix} u_h(x_1) - u(x_1) \\ \vdots \\ u_h(x_n) - u(x_n) \end{bmatrix} = 0$$

with $A_h$ the (SPD) stiffness matrix.

### 4.1.3 $H^1$ norm general bound

We have, for $w_h \in X_h$, and $a(\cdot, \cdot)$ coercive and continuous

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

Therefore

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

from which

$$\|u-u_h\|_{H^1} \leq \frac{\beta}{\alpha}\inf_{w_h\in X_h}\|u-w_h\|_{H^1}$$

If $a$ is symmetric, then we have

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

and thus

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

which is sharper than the previous since $\beta \geq \alpha$.

### 4.1.4 Output functionals

We have the following results:

- if $\ell^{\mathrm{O}} \in H^{-1}(\Omega)$ then $\left|\ell^{\mathrm{O}}(e)\right| \leq C\left\|e\right\|_{H^1(\Omega)} \leq Ch\left\|u\right\|_{H^2(\Omega,\mathcal{T}_h)}$
- if $\ell^{\mathrm{O}} \in L^2(\Omega)'$ then $\left|\ell^{\mathrm{O}}(e)\right| \leq C\left\|e\right\|_{L^2(\Omega)} \leq Ch^2\left\|u\right\|_{H^2(\Omega,\mathcal{T}_h)}$

They come from boundedness and common bounds on $e$. Moreover, with the help of the ajoint variables $\psi$ and $\phi_h$

- if $\ell^{\mathrm{O}} \in H^{-1}(\Omega)$ then $\left|\ell^{\mathrm{O}}(e)\right| \leq Ch^2\left\|u\right\|_{H^2(\Omega,\mathcal{T}_h)}\left\|\psi\right\|_{H^2(\Omega,\mathcal{T}_h)}$

that is, even in the more general case $\ell^{\mathrm{O}} \in H^{-1}(\Omega)$, $\ell^{\mathrm{O}}(e) \in \mathcal{O}(h^2)$.

**Example**

Given the discretization $\{x_i\}_i$, is

$$\ell^{\mathrm{O}}(e) = \max_i |e(x_i)|$$

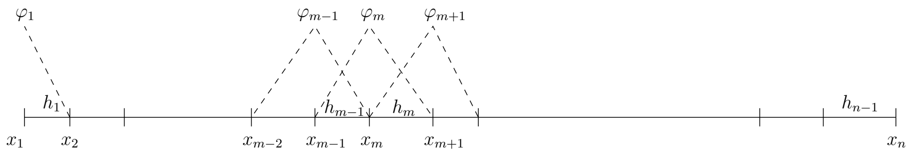
bounded in $H_0^1$? Yes,

$$\ell^{\mathrm{O}}(e) = e(x_{\bar{i}}) = \int_0^{x_{\bar{i}}} e'(x)\mathrm{d}x \overset{\text{C.–S.}}{\leq} \sqrt{x_{\bar{i}}}\left(\int_0^{x_{\bar{i}}}\left|e'(x)\right|^2\mathrm{d}x\right)^{1/2} \leq 1\cdot|e|_{H^1} \leq \|e\|_{H^1}$$

# Chapter 5

# Quadrature and assembly in one dimension

## 5.1 Quadrature formulas



Let us consider the problem to approximate

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

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

### 5.1.1 By interpolation (mass matrix)

If we consider to approximate

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

then

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

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

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

Therefore the global error is $\mathcal{O}(h^3)$.

### 5.1.2 By trapezoidal rule

We can approximate

$$\begin{aligned}\int_{x_{i-1}}^{x_{i+1}} f(x)\varphi_i(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} f(x)\varphi_i(x)\mathrm{d}x + \int_{x_i}^{x_{i+1}} f(x)\varphi_i(x)\mathrm{d}x \approx\\ \approx (0+f(x_i))\frac{h}{2} + (f(x_i)+0)\frac{h}{2}\end{aligned}$$

The quadrature error for the first intergral is

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

where $\xi_{i-1}$ is a point in $(x_{i-1}, x_i)$. For the second integral we have an error

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

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

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

Therefore a global error $\mathcal{O}(h^3)$ if $f \in \mathcal{C}^2$. The same consideration holds if we consider *quasi-uniform* meshes.

### Mass matrix by trapezoidal rule

Let us compute the mass matrix. We have

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

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

$$\begin{aligned}
M_{ij} &= \int_{x_{i-1}}^{x_{i+1}} \varphi_j(x)\varphi_i(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} \varphi_j(x)\varphi_i(x)\mathrm{d}x + \int_{x_i}^{x_{i+1}} \varphi_j(x)\varphi_i(x)\mathrm{d}x \approx \\
&\approx \varphi_j(x_i)\frac{h_{i-1}}{2} + \varphi_j(x_i)\frac{h_i}{2} = \delta_{ij}\frac{h_{i-1} + h_i}{2}
\end{aligned}$$

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

## 5.1.3 By barycentric formulas

Let us start with the midpoint rule to approximate

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

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

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

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

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

### 5.1.4 By 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 error for the approximation of

$$\int_{x_{i-1}}^{x_i} f(x)\varphi_i(x)\mathrm{d}x + \int_{x_i}^{x_{i+1}} f(x)\varphi_i(x)\mathrm{d}x$$

is $\mathcal{O}(h^5)$ if $f \in \mathcal{C}^4$. This formula is the qf2pE formula used by FreeFem++, while the default formula used by FreeFem++ is qf3pE, the Gauss–Legendre formula with nodes 0 and $\pm\sqrt{3/5}$ with associated weights 8/9 and 5/9.

## 5.2 Assembly

Let us see a general implementation strategy for FEM. Suppose we have $M$ elements $\{\ell_m\}_{m=1}^m$ (in the one-dimensional case, the intervals) with the associate points. With respect to Figure 5.1, where $n = M+1$, we have

$$\ell_{m,1} = m,\ \ell_{m,2} = m+1, \quad 1 \le m \le M$$

which means that points $x_m$ and $x_{m+1}$ are associate to element $\ell_m$. The two basis functions which have value 1 on node $\ell_{m,k}$ and 0 on node $\ell_{m,3-k}$, for

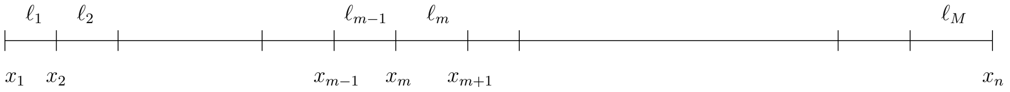


Figure 5.1: Points (bottom) and elements (top).

$k = 1, 2$, have the form (on $\ell_m$)

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

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

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

$$\left.\begin{aligned} a_{\ell_{m,k}\ell_{m,k}} &= \int_0^1 \varphi'_{\ell_{m,k}}(x)\varphi'_{\ell_{m,k}}(x)\mathrm{d}x \\ a_{\ell_{m,k}\ell_{m,3-k}} &= \int_0^1 \varphi'_{\ell_{m,k}}(x)\varphi'_{\ell_{m,3-k}}(x)\mathrm{d}x \end{aligned}\right\} \quad k = 1, 2$$

and to the element $\tilde{f}_{\ell_{m,k}}$ of the right hand side

$$\tilde{f}_{\ell_{m,k}} = \text{approximation of} \int_0^1 f(x)\varphi_{\ell_{m,k}}(x)\mathrm{d}x$$

in this way

$$a_{ij} = \sum_{\substack{\ell_{m,k}=i \\ \ell_{m,h}=j}} A_{\ell_{m,k}\ell_{m,h}}$$

$$\tilde{f}_i = \sum_{\ell_{m,k}=i} \tilde{F}_{\ell_{m,k}}$$

where

$$A_{\ell_{m,k}\ell_{m,h}} = \int_{\ell_m} \phi'_{\ell_{m,k}}(x)\phi'_{\ell_{m,h}}(x)\mathrm{d}x$$

$$\tilde{F}_{\ell_{m,k}} = \text{approximation of} \int_{\ell_m} f(x)\phi_{\ell_{m,k}}(x)\mathrm{d}x$$

Hence, for inner nodes,

$$\begin{aligned}
a_{\ell_{m,1}\ell_{m,1}} &= A_{\ell_{m-1,2}\ell_{m-1,2}} + A_{\ell_{m,1}\ell_{m.1}} = \\
&= \int_{\ell_{m-1}} \left(\frac{\beta_{m-1,2}}{\Delta_{m-1}}\right)^2 \mathrm{d}x + \int_{\ell_m} \left(\frac{\beta_{m,1}}{\Delta_m}\right)^2 \mathrm{d}x = \\
&= \int_{\ell_{m-1}} \left(\frac{1}{\Delta_{m-1}}\right)^2 \mathrm{d}x + \int_{\ell_m} \left(\frac{-1}{\Delta_m}\right)^2 \mathrm{d}x = \frac{1}{|\Delta_{m-1}|} + \frac{1}{|\Delta_m|} \\
a_{\ell_{m,1}\ell_{m,2}} &= a_{\ell_{m,2}\ell_{m,1}} = A_{\ell_{m,1}\ell_{m,2}} = \int_{\ell_m} \frac{\beta_{m,1}}{\Delta_m}\frac{\beta_{m,2}}{\Delta_m}\mathrm{d}x = \int_{\ell_m} -\frac{1}{\Delta_m}\frac{1}{\Delta_m}\mathrm{d}x = \\
&= -\frac{1}{|\Delta_m|} \\
\tilde{f}_{\ell_{m,1}} &= \tilde{F}_{\ell_{m-1,2}} + \tilde{F}_{\ell_{m,1}}
\end{aligned}$$

Hence, the *assembly* is done by

- $a_{ij} = 0,\ 1 \le i, j \le n,\ \tilde{f}_i = 0,\ 1 \le i \le n$
- FOR $m = 1, \ldots, M$

```
        FOR $k = 1, \ldots, 2$
            $a_{\ell_{m,k}\ell_{m,k}} = a_{\ell_{m,k}\ell_{m,k}} + A_{\ell_{m,k}\ell_{m,k}}$, $\tilde{f}_{\ell_{m,k}} = \tilde{f}_{\ell_{m,k}} + \tilde{F}_{\ell_{m,k}}$
            FOR $h = k+1, \ldots, 2$ ($h = 1, \ldots, 2,\ h \neq k$ non-symm. case)
                $a_{\ell_{m,k}\ell_{m,h}} = a_{\ell_{m,k}\ell_{m,h}} + A_{\ell_{m,k}\ell_{m,h}}$
                $a_{\ell_{m,h}\ell_{m,k}} = a_{\ell_{m,k}\ell_{m,h}}$ (only symm. case)
            END
        END
    END
```

### 5.2.1 Barycentric coordinates

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

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

The coordinates $\lambda_{\ell_{m,k}}(x)$ satisfy $\lambda_{\ell_{m,k}}(x_{\ell_{m,h}}) = \delta_{hk}$ and $\lambda_{\ell_{m,1}}(x)+\lambda_{\ell_{m,2}}(x) = 1$. Therefore, it is $\lambda_{\ell_{m,k}}(x) = \varphi_{\ell_{m,k}}(x)$.

## 5.3 Some examples

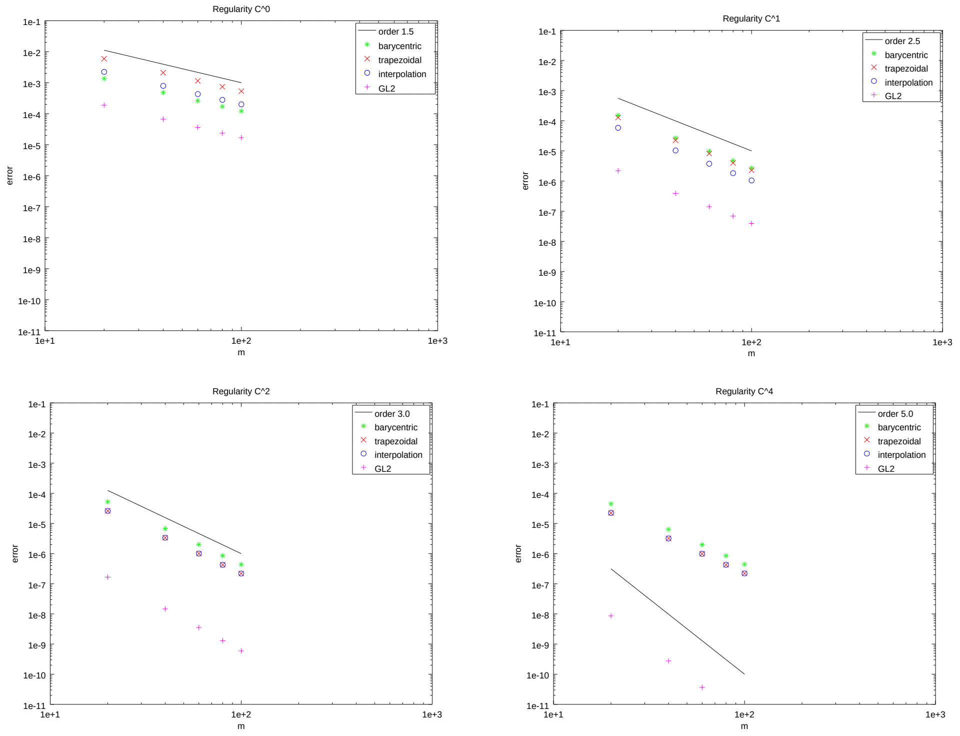


Figure 5.2: Maximum error over $i = 2, 3, \ldots, m-1$ between $\int_0^1 f(x)\varphi_i(x)\mathrm{d}x$ and the quadrature formulas, for $f(x) = |x - 1/2|^{\frac{1}{2}}$ and $f(x) = |x - 1/2|^{\frac{3}{2}}$ (top) and $f(x) = |x - 1/2|^{\frac{5}{2}}$ and $f(x) = |x - 1/2|^{\frac{9}{2}}$ (bottom).

We consider the family of functions $f(x) = |x - 1/2|^{p-1/2} \in \mathcal{C}^{p-1}(0, 1)$, $p = 1, 2, 3, 5$, and approximate $\int_0^1 f(x)\varphi_i(x)\mathrm{d}x$. The results are in Figure 5.2. We then consider the Poisson problem

$$\begin{cases} -u''(x) = f(x) & x \in (0, 1) \\ u(0) = u(1) = 0 \end{cases} \tag{5.1}$$

with $f(x) = |x - 1/2|^{p-1/2}$, $p = 1, 3$. The exact solution is

$$u(x) = -\frac{4}{(2p+3)(2p+1)}\left(\left|x - \frac{1}{2}\right|^{p+3/2} - \left(\frac{1}{2}\right)^{p+3/2}\right)$$

and it is $u \in H^{p+1}(0, 1)$ (and $u \in \mathcal{C}^{p+1}(0, 1)$, too). The results are in Figure 5.3. The right hand side was evaluated either exactly (high precision quadrature formula) or by the barycentric formulas.

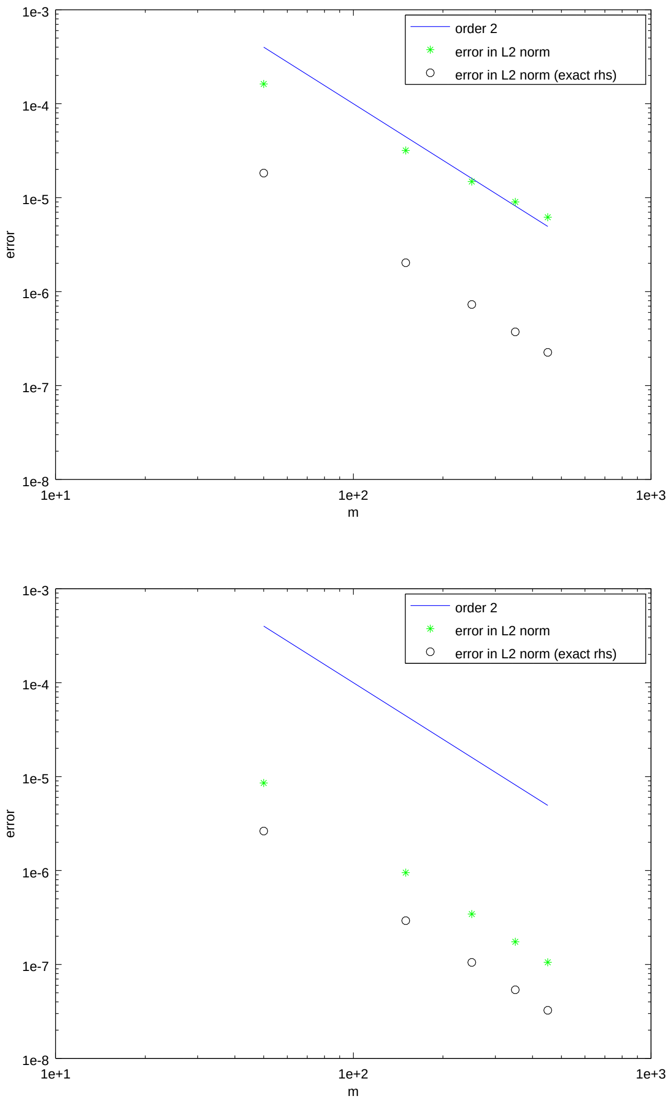


Figure 5.3: $L^2$ error for the Poisson problem 5.1, $p = 1$ (top) and $p = 3$ (bottom).

## 5.4 Exercises

1. Consider the Poisson problem, written for semplicity in the strong form,

$$\begin{cases} \text{“}-\partial_{xx}u = -2\delta(x-1/2)\text{”} & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

(a) Find the analytical solution.

(b) What is the error in the energy norm ($H^1$ seminorm) proportional to when an approximate solution is computed on a uniform grid with an odd number of points? What is the error in the $L^2$ norm proportional to? Verify your answers by implementation.

(c) Do the same as above with an even number of points.

(d) Why in this case the classical error bound in the $H^2$ broken seminorm does not work?

# Chapter 6

# Quadrature and assembly in two dimensions

## 6.1 Quadrature formulas

### 6.1.1 By interpolation (mass matrix)

### 6.1.2 By trapezoidal rule

$$\int_{\ell_m} g(x,y)\mathrm{d}x\mathrm{d}y \approx |\Delta_m| \, \frac{g(x_{\ell_{j,1}}, y_{\ell_{m,1}}) + g(x_{\ell_{m,2}}, y_{\ell_{m,2}}) + g(x_{\ell_{m,3}}, y_{\ell_{m,3}})}{3}$$

**Mass matrix by trapezoidal rule**

Let us start computing

$$M_{ij} = \sum_{\substack{\ell_{m,k}=i \\ \ell_{m,h}=j}} \int_{\ell_m} \phi_{\ell_{m,h}}(x,y)\phi_{\ell_{m,k}}(x,y)\mathrm{d}x\mathrm{d}y = \begin{cases} \displaystyle\sum_{\ell_{m,k}=i} \frac{|\Delta_m|}{6}, & i = j \\ \displaystyle\sum_{\substack{\ell_{m,k}=i \\ \ell_{m,h}=j}} \frac{|\Delta_m|}{12}, & i \neq j \end{cases}$$

If we take the sum over $j$, we get

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

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

$$
\begin{aligned}
M_{ii} &\approx \sum_{\ell_{m,k}=i} \frac{|\Delta_m|}{3}\\
M_{ij} &\approx 0, \quad i \neq j
\end{aligned}
$$

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

### 6.1.3 By barycentric formulas

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

where

$$
\bar{f}_m = \frac{f(x_{\ell_{m,1}}, y_{\ell_{m,1}}) + f(x_{\ell_{m,2}}, y_{\ell_{m,2}}) + f(x_{\ell_{m,3}}, y_{\ell_{m,3}})}{3}
$$

### 6.1.4 By Gauss–Legendre quadrature

The first Gauss–Legendre quadrature formula is

$$
\int_{\ell_m} g(x,y)\mathrm{d}x\mathrm{d}y \approx |\Delta_m|\, g\left(\frac{x_{\ell_{m,1}} + x_{\ell_{m,2}} + x_{\ell_{m,3}}}{3}, \frac{y_{\ell_{m,1}} + y_{\ell_{m,2}} + y_{\ell_{m,3}}}{3}\right) = |\Delta_m|\, g(x_{\ell_m}, y_{\ell_m})
$$

which is exact for $g(x,y) \in \mathbb{P}_1$. In fact, due to the properties of the *centroid*,

$$
(\bar{x}_m, \bar{y}_m) = (x_{\ell_m}, y_{\ell_m}) = \left(\int_{\ell_m} x\mathrm{d}x\mathrm{d}y, \int_{\ell_m} y\mathrm{d}x\mathrm{d}y\right) / |\Delta_m|
$$

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

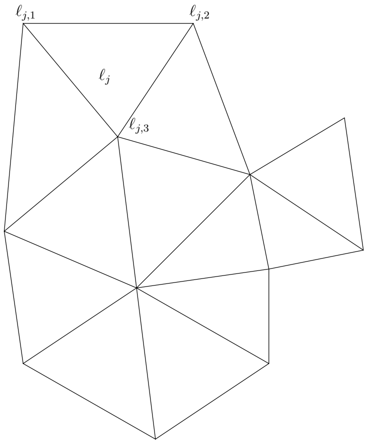
$$
g(x,y) = g(x_{\ell_m}, y_{\ell_m}) + \nabla g \cdot ((x,y) - (x_{\ell_m}, y_{\ell_m}))
$$

and therefore

$$
\begin{aligned}
\int_{\ell_m} g(x,y)\mathrm{d}x\mathrm{d}y &= \int_{\ell_m} g(x_{\ell_m}, y_{\ell_m})\mathrm{d}x\mathrm{d}y + \int_{\ell_m} \nabla g \cdot ((x,y) - (x_{\ell_m}, y_{\ell_m}))\,\mathrm{d}x\mathrm{d}y =\\
&= |\Delta_m|\, g(x_{\ell_m}, y_{\ell_m}) + \nabla g \cdot \int_{\ell_m} ((x,y) - (x_{\ell_m}, y_{\ell_m}))\,\mathrm{d}x\mathrm{d}y =\\
&= |\Delta_m|\, g(x_{\ell_m}, y_{\ell_m}) + \nabla g \cdot \left(|\Delta_m|\,(x_{\ell_m}, y_{\ell_m}) - \int_{\ell_m} (x_{\ell_m}, y_{\ell_m})\mathrm{d}x\mathrm{d}y\right) =\\
&= |\Delta_m|\, g(x_{\ell_m}, y_{\ell_m})
\end{aligned}
$$

There exist high order Gauss–Legendre quadrature formulas for triangles, involving three (qf2pT), seven (qf5pT, the default in FreeFem++) or more points.

## 6.2 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 function $\varphi_{\ell_{m,k}}$ which has value 1 on node $\ell_{m,k}$ and 0 on nodes $\ell_{m,h}$, $h \in \{1,2,3\}$, $h \neq k$ of the triangle $\ell_m$. It has the form

$$\phi_{\ell_{m,k}}(x,y)=\frac{\alpha_{m,k}+\beta_{m,k}x+\gamma_{m,k}y}{2\Delta_m}=\begin{array}{|ccc|} \multicolumn{1}{c}{} & \multicolumn{1}{c}{k} & \multicolumn{1}{c}{} \\ 1 & 1 & 1 \\ x_{\ell_{m,1}} & x & x_{\ell_{m,3}} \\ y_{\ell_{m,1}} & y & y_{\ell_{m,3}} \end{array} \Big/ \begin{vmatrix} 1 & 1 & 1 \\ x_{\ell_{m,1}} & x_{\ell_{m,2}} & x_{\ell_{m,3}} \\ y_{\ell_{m,1}} & y_{\ell_{m,2}} & y_{\ell_{m,3}} \end{vmatrix}$$

where $\Delta_m$ is the area (with sign) of triangle $\ell_m$. We need to compute

$$\int_{\ell_m}\left(\frac{\partial\phi_{\ell_{m,k}}(x,y)}{\partial x}\frac{\partial\phi_{\ell_{m,h}}(x,y)}{\partial x}+\frac{\partial\phi_{\ell_{m,k}}(x,y)}{\partial y}\frac{\partial\phi_{\ell_{m,h}}(x,y)}{\partial y}\right)\mathrm{d}x\mathrm{d}y,\quad h,k=1,2,3$$

for the stiffness matrix (and also derivatives with respect to $y$) and

$$\int_{\ell_m} f(x,y)\phi_{\ell_{m,k}}(x,y)\mathrm{d}x\mathrm{d}y$$

for the right hand side. We have

$$\int_{\ell_m}\frac{\partial\phi_{\ell_{m,k}}(x,y)}{\partial x}\frac{\partial\phi_{\ell_{m,h}}(x,y)}{\partial x}\mathrm{d}x\mathrm{d}y=\int_{\ell_m}\frac{\beta_{m,k}}{2\Delta_m}\frac{\beta_{m,h}}{2\Delta_m}\mathrm{d}x\mathrm{d}y=\frac{\beta_{m,k}\beta_{m,h}}{4\,|\Delta_m|}$$
$$\int_{\ell_m}\frac{\partial\phi_{\ell_{m,k}}(x,y)}{\partial y}\frac{\partial\phi_{\ell_{m,h}}(x,y)}{\partial y}\mathrm{d}x\mathrm{d}y=\int_{\ell_m}\frac{\gamma_{m,k}}{2\Delta_m}\frac{\gamma_{m,h}}{2\Delta_m}\mathrm{d}x\mathrm{d}y=\frac{\gamma_{m,k}\gamma_{m,h}}{4\,|\Delta_m|}$$

and their sum correspond to $A_{\ell_{m,k}\ell_{m,h}}$ and

$$\int_{\ell_m} f(x,y)\phi_{\ell_{m,k}}(x,y)\mathrm{d}x\mathrm{d}y\approx\tilde{F}_{\ell_{m,k}}$$

The algorithm for the assembly is

```
• $a_{ij}=0,\ 1\le i,j\le n,\ \tilde{f}_i=0,\ 1\le i\le n$
• FOR $m=1,\dots,M$
      FOR $k=1,\dots,3$
          $a_{\ell_{m,k}\ell_{m,k}}=a_{\ell_{m,k}\ell_{m,k}}+\frac{\beta_{m,k}\beta_{m,k}}{4|\Delta_m|}+\frac{\gamma_{m,k}\gamma_{m,k}}{4|\Delta_m|},\ \tilde{f}_{\ell_{m,k}}=\tilde{f}_{\ell_{m,k}}+\tilde{F}_{\ell_{m,k}}$
          FOR $h=k+1,\dots,3$ ($h=1,\dots,3,\ h\neq k$ non-symm. case)
              $a_{\ell_{m,k}\ell_{m,h}}=a_{\ell_{m,k}\ell_{m,h}}+\frac{\beta_{m,k}\beta_{m,h}}{4|\Delta_m|}+\frac{\gamma_{m,k}\gamma_{m,h}}{4|\Delta_m|}$
              $a_{\ell_{m,h}\ell_{m,k}}=a_{\ell_{m,k}\ell_{m,h}}$ (only symm. case)
          END
      END
  END
```

### 6.2.1 Barycentric coordinates

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

# Chapter 7

# Higher order basis functions

We consider, for semplicity, the homogeneous Dirichlet problem.

## 7.1 One-dimensional case

In the one dimensional case $\Omega$ is an open interval and $X = H_0^1(\Omega)$. We just consider the space $X_h^2 = \{v_h \in X \colon v_h|_{T_h} \in \mathbb{P}_2(T_h)\}$. A polynomial of degree two on a interval is defined by three points, usually the two extreme points and the middle point. Therefore, given an original set of nodes $\{y_j\}_{j=1}^m \subset \Omega$, we have to consider the new set of nodes $\{x_i\}_{i=1}^{2m-1} \subset \Omega$ given by

$$\begin{cases} x_i = y_{(i+1)/2}, & i \text{ odd} \\ x_i = \dfrac{y_{i/2} + y_{i/2+1}}{2}, & i \text{ even} \end{cases}$$

and the set of basis functions

$$\varphi_i(x) \in X_h^2,\ \varphi_i(x_j) = \delta_{ij}, \quad 1 \le i, j \le 2m-1$$

On the element $\ell_j$, with endpoints $\ell_{j,1}$ and $\ell_{j,3}$ and middle point $\ell_{j,2}$, the form of $\varphi_{\ell_{j,k}}$ is

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

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

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

Clearly now some of the basis function $\varphi_i$ shares its support with $\varphi_{i-2}, \varphi_{i-1}, \varphi_{i+1}, \varphi_{i+2}$ and therefore the stiffness matrix, for instance, is a pentadiagonal matrix.

### 7.1.1 Error estimates

The weak formulation is

$$\text{find } u \in H^1(\Omega) \text{ such that } a(u,v) = \ell(v),\ \forall v \in H^1(\Omega)$$

with $a$ bilinear, coercive, continuos and $\ell$ linear bounded. Therefore we assume that $u \in H^1(\Omega)$. Let us denote the generic triangle (edge) by $T_h^k$ and its length by $h_k$. The maximum length of the triangles is $h$.

#### $H^1$ norm, $X_h^r$ space

Let be $u_h \in X_h^r$. Then:

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

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

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

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

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

### $L^2$ norm, $X_h^r$ space

Let be $u_h \in X_h^r$. If from $\ell(v) = \ell_f(v) = \int_\Omega fv$ (therefore $f \in L^2(\Omega)$) it follows that $u \in H^2(\Omega)$ (it is called *elliptic regularity*, for instance, Poisson problem), then

- if $u \in H^{p+1}(\Omega, \mathcal{T}_h)$ and $s = \min\{p, r\}$
$$\|u_h - u\|_{L^2(\Omega)} \leq Ch^{s+1} |u|_{H^{s+1}(\Omega,\mathcal{T}_h)}$$
- if $u \in H^{p+1}(\Omega)$ and $s = \min\{p, r\}$
$$\|u_h - u\|_{L^2(\Omega)} \leq Ch^{s+1} |u|_{H^{s+1}(\Omega)}$$

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

## 7.2 Two-dimensional case

In the two-dimensional case $\Omega$ is a polygon and $X = H_0^1(\Omega)$. We just consider the space $X_h^2 = \{v_h \in X \cap \mathcal{C}^0(\bar{\Omega})\colon v_h|_{T_h} \in \mathbb{P}_2(T_h)\}$. 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 of the edges are taken. We introduce the *barycentric coordinates:* any point $x$ in a triangle $\ell_j$ with vertices $\{x_1, x_2, x_3\} \in \Omega$ can be written in a unique way as

$$x = \lambda_1(x)x_1 + \lambda_2(x)x_2 + \lambda_3(x)x_3, \quad \lambda_1(x) + \lambda_2(x) + \lambda_3(x) \equiv 1$$

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

Six distinct points in the plane and six correponding 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 a degree which takes the values $0, 0, 1$ in the three 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.** *Given three non-collinear points $x_1, x_2, x_3 \in \Omega$ and the corresponding middle points $x_{12}, x_{13}, x_{23}$, a polynomial $p(x)$ of total degree two is well defined by the values of $p(x)$ at the six points.*

*Proof.* It is enough to prove that if $p(x_1) = p(x_2) = p(x_3) = p(x_{12}) = p(x_{13}) = p(x_{23}) = 0$, than $p \equiv 0$. 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 0). 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 $x_{12}$, we have

$$0 = p(x_{12}) = \lambda_1(x_{12})\lambda_2(x_{12})\gamma = \frac{1}{2}\frac{1}{2}\gamma$$

and therefore $\gamma = 0$. □

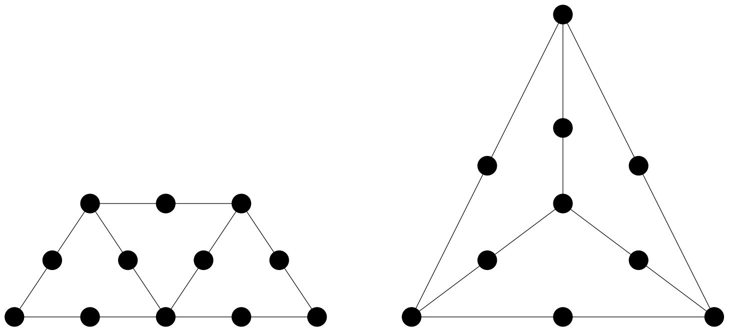


Figure 7.1: $m = 5$, $n = 3$ (right) and $m = 4$, $n = 3$ (left).

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 triangularion). Therefore, the dimension of $X_h^2$ is $m+(m+n-1) = 2m+n-1$.

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

### 7.2.1 Bandwidth reduction

Even in the simplest case of piecewise linear basis function, an ordering of the nodes as in Figure 7.2 (left) would yield a sparsity pattern as in Figure 7.2 (right). The *degree* of a node is the number of 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 $Q$.

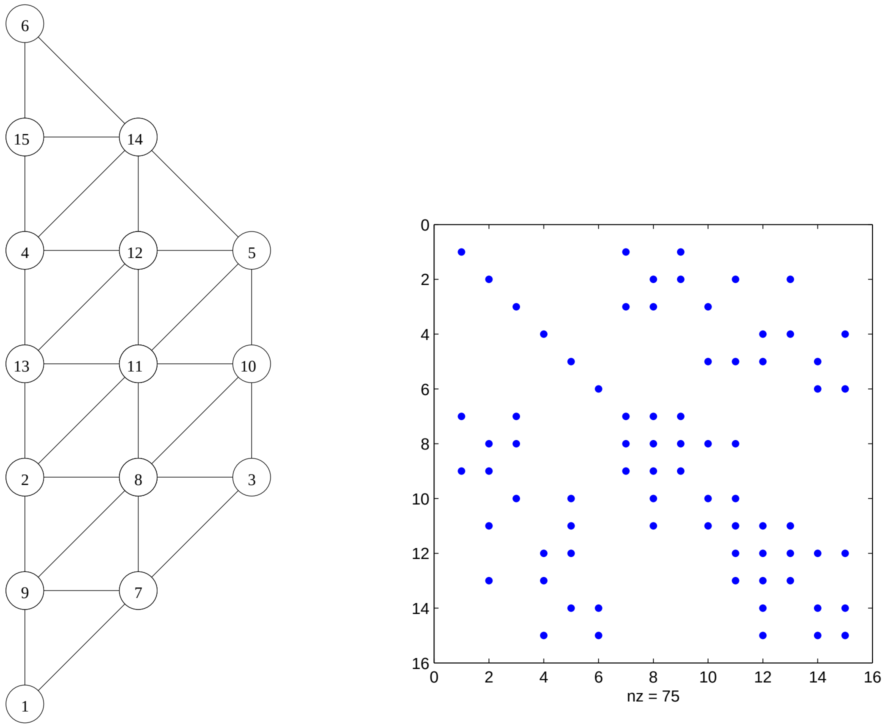


Figure 7.2: Unordered mesh and corresponding sparsity pattern.

- 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 in $Q$, in the increasing order of their degree,

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. The ordering produced by the reverse Cuthill–McKee algorithm with initial node 1 (a node with smallest degree) is shown in Figure 7.3.

### 7.2.2 Error estimates

The weak formulation is

$$\text{find } u \in H^1(\Omega) \text{ such that } a(u,v) = \ell(v),\ \forall v \in H^1(\Omega)$$

with $a$ bilinear, coercive, continuos and $\ell$ linear bounded. Therefore we assume that $u \in H^1(\Omega)$. Let us denote the generic triangle by $K$ and its

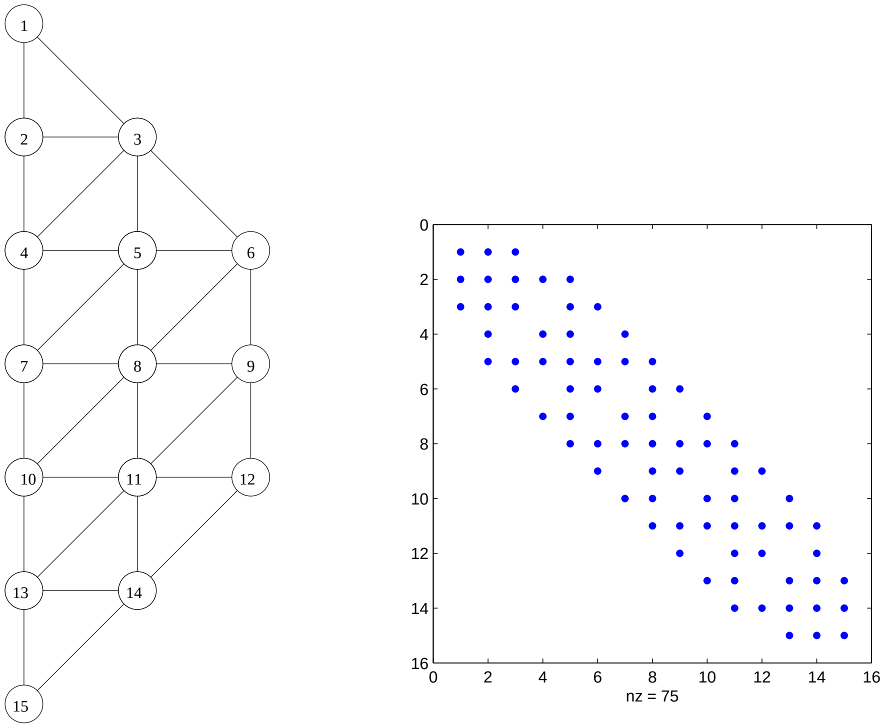


Figure 7.3: Reverse Cuthill–McKee ordered mesh and corresponding sparsity pattern.

diameter by $h_K$. The maximum diameter of the triangles is $h$.

### $H^1$ norm, $X_h^r$ space

Let be $\{\mathcal{T}_h\}_h$ a family of regular triangulations of $\Omega$, polygonal, convex and $u_h \in X_h^r$. Then Let be $u_h \in X_h^r$. Then:

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

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

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

$$\|u_h - u\|_{H^1(\Omega)} \leq C \sum_{T_h^k \in \mathcal{T}_h} \left( h_k^{2s} |u|^2_{H^{s+1}(T_h^k)} \right)^{1/2} \leq C h^s |u|_{H^{s+1}(\Omega)}$$

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

### $L^2$ norm, $X_h^r$ space

Let be $\{\mathcal{T}_h\}_h$ a family of regular triangulations of $\Omega$ polygonal, convex and $u_h \in X_h^r$. If from $\ell(v) = \ell_f(v) = \int_\Omega fv$ (therefore $f \in L^2(\Omega)$) and $\Omega$ convex it follows that $u \in H^2(\Omega)$ (it is called *elliptic regularity*, for instance, Poisson problem), then

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

$$\|u_h - u\|_{L^2(\Omega)} \leq Ch^{s+1} |u|_{H^{s+1}(\Omega,\mathcal{T}_h)}$$

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

$$\|u_h - u\|_{L^2(\Omega)} \leq Ch^{s+1} |u|_{H^{s+1}(\Omega)}$$

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

# Chapter 8

# Discontinuous Galerkin

## 8.1 mean and jump

Unfortunately, the notation for discontinuous Galerkin methods in [9] and [4] (based on [10]) is different. Let's try to understand. With reference to Figure 8.1 we have

$$[v] = v^+\mathbf{n}^+ + v^-\mathbf{n}^-, \quad \{\!\{\nabla u\}\!\} = \frac{(\nabla u)^+ + (\nabla u)^-}{2}$$

and these two terms are coupled with a minus sign in front, that is

$$\int_\Omega -\Delta u v = \int_\Omega \nabla u \cdot \nabla v - \sum_{e\in\mathcal{E}_h} \int_e [v] \cdot \{\!\{\nabla u\}\!\}$$

where $\mathcal{E}_h$ is the set of internal edges (with homogeneous Dirichlet b.c.).

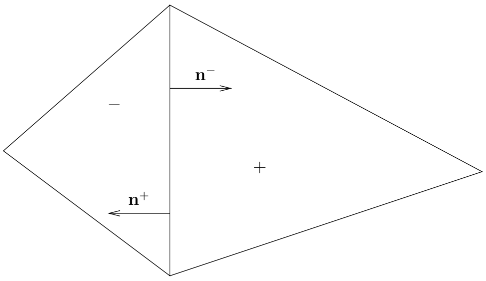


Figure 8.1: Adjacent triangles.

In FreeFem++, the average (called `mean`) is defined in the same way but the jump (called `jump`) is defined as *external* value minus *internal* value. On boundary edges, `mean` is simply the internal value and `jump` its opposite.

Each internal edge is counted twice in the command `intalledges`, each time with a different external normal, denoted by `N`. So, for the internal edge in the figure, if you call

```
intallegdes(Th)(mean(N.x*dx(u)+N.y*dy(u))*jump(v)/nTonEdge)
```

you get

$$\frac{1}{2}\left(\frac{\nabla u^{+}\cdot\mathbf{n}^{-}+\nabla u^{-}\cdot\mathbf{n}^{-}}{2}(v^{+}-v^{-})+\frac{\nabla u^{-}\cdot\mathbf{n}^{+}+\nabla u^{+}\cdot\mathbf{n}^{+}}{2}(v^{-}-v^{+})\right)= \\ =-\{\!\!\{\nabla u\}\!\!\}\cdot[v]$$

being `nTonEdge` equal to 2 on internal edges. For short, on internal edges, we can write

$$\frac{1}{2}\mathtt{mean}\left(\frac{\partial u}{\partial\mathtt{N}}\right)\mathtt{jump}(v)=-\{\!\!\{\nabla u\}\!\!\}\cdot[v]$$

In the same way, on internal edges,

$$\frac{1}{2}\mathtt{jump}(u)\mathtt{jump}(v)=[u]\cdot[v]$$

Now, on external edges, $\mathtt{mean}(u)$ is just internal value and $\mathtt{jump}(u)$ is the opposite of the interval value. The interior penalty method for Poisson's equation (with Dirichlet boundary conditions in Nitsche's way) writes

$$\int_{\Omega}\nabla u_{\delta}\cdot\nabla v_{\delta}+\sum_{e\in\mathcal{E}_h}\int_e\frac{1}{2}\mathtt{mean}\left(\frac{\partial u_{\delta}}{\partial\mathtt{N}}\right)\mathtt{jump}(v_{\delta})+\sum_{e\subseteq\partial\Omega}\int_e\mathtt{mean}\left(\frac{\partial u_{\delta}}{\partial\mathtt{N}}\right)\mathtt{jump}(v_{\delta})+ \\ +\tau\sum_{e\in\mathcal{E}_h}\int_e\frac{1}{2}\mathtt{mean}\left(\frac{\partial v_{\delta}}{\partial\mathtt{N}}\right)\mathtt{jump}(u_{\delta})+\tau\sum_{e\subseteq\partial\Omega}\int_e\mathtt{mean}\left(\frac{\partial v_{\delta}}{\partial\mathtt{N}}\right)\mathtt{jump}(u_{\delta})+ \\ \tau\sum_{e\subseteq\partial\Omega}\int_e g_{\delta}\frac{\partial v_{\delta}}{\partial\mathtt{N}}+\gamma\sum_{e\in\mathcal{E}_h}\frac{1}{|e|}\int_e\frac{1}{2}\mathtt{jump}(u_{\delta})\mathtt{jump}(v_{\delta})+ \\ +\gamma\sum_{e\subseteq\partial\Omega}\frac{1}{|e|}\int_e\mathtt{jump}(u_{\delta})\mathtt{jump}(v_{\delta})-\gamma\sum_{e\subseteq\partial\Omega}\frac{1}{|e|}\int_e g_{\delta}v_{\delta}=\int_{\Omega}fv_{\delta}$$

The integral over the boundary can be computed by `int1d(Th)` and the length of an edge is `lenEdge`.

Now we try to get in FreeFem++ the *upwind average*

$$\{\!\!\{\mathbf{b}u\}\!\!\}_{\mathbf{b}}=\begin{cases}\mathbf{b}u^{+} & \mathbf{b}\cdot\mathbf{n}^{+}>0\\ \mathbf{b}u^{-} & \mathbf{b}\cdot\mathbf{n}^{+}<0\\ \mathbf{b}\{u\} & \mathbf{b}\cdot\mathbf{n}^{-}=0\end{cases}$$

With reference to Figure 8.1, on internal edges it is

$$\{\!\{\mathbf{b}u\}\!\}_{\mathbf{b}} \cdot [v] = \frac{1}{2}\left(|\mathbf{b}\cdot\mathtt{N}|\,\mathtt{jump}(u)/2 - (\mathbf{b}\cdot\mathtt{N})\mathtt{mean}(u)\right)\mathtt{jump}(v)$$

On boundary edges, we have would like to have

$$\{\!\{\mathbf{b}u\}\!\}_{\mathbf{b}} = \begin{cases}\mathbf{b}u & \mathbf{b}\cdot\mathbf{n} > 0\\ 0 & \text{otherwise}\end{cases}$$

Since in this case, $\mathtt{jump}(u) = -u$, whereas $\mathtt{mean}(u) = u$, we should correct the previous formula with

$$\{\!\{\mathbf{b}u\}\!\}_{\mathbf{b}}\cdot[v] = \left(|\mathbf{b}\cdot\mathtt{N}|\,\mathtt{jump}(u)/2 - (\mathbf{b}\cdot\mathtt{N})\mathtt{mean}(u)/(3-\mathtt{nTonEdge})\right)\mathtt{jump}(v)/\mathtt{nTonEdge}$$

The DG formulation with upwind advection and Nitsche imposition of boundary conditions of the advection-diffusion problem

$$\begin{cases}\operatorname{div}(-\mu\nabla u + \boldsymbol{b}u) = f, & \Omega\\ u = g, & \Gamma = \Gamma^+\cup\Gamma^- = \partial\Omega\end{cases}$$

where $\Gamma^+ = \{e\in\partial\Omega\colon \mathbf{b}\cdot\mathbf{n}\geq 0\}$, is

$$\begin{aligned}\sum_{K\in\Omega}\int_K \mu\nabla u_\delta\cdot\nabla v_\delta - \sum_{e\in\mathcal{E}_h}\int_e [v_\delta]\cdot\{\!\{\mu\nabla u_\delta\}\!\} - \sum_{e\in\Gamma}\int_e \mu v_\delta\nabla u_\delta\cdot\mathbf{n}\\ -\tau\sum_{e\in\mathcal{E}_h}\int_e [u_\delta]\cdot\{\!\{\mu\nabla v_\delta\}\!\} - \tau\sum_{e\in\Gamma}\int_e \mu(u_\delta - g_\delta)\nabla v_\delta\cdot\mathbf{n}\\ +\sum_{e\in\mathcal{E}_h}\int_e\frac{\gamma}{|e|}[u_\delta]\cdot[v_\delta] + \sum_{e\in\Gamma}\int_e\frac{\gamma}{|e|}(u_\delta - g_\delta)v_\delta\\ -\sum_{K\in\Omega}\int_K u_\delta\mathbf{b}\cdot\nabla v_\delta + \sum_{e\in\mathcal{E}_h}\int_e\{\!\{\mathbf{b}u_\delta\}\!\}_{\mathbf{b}}\cdot[v_\delta] + \sum_{e\in\Gamma^+}\int_e u_\delta v_\delta\mathbf{b}\cdot\mathbf{n} + \sum_{e\in\Gamma^-}\int_e g_\delta v_\delta\mathbf{b}\cdot\mathbf{n}\\ -\sum_{K\in\Omega}\int f v_\delta = 0\end{aligned}$$

## 8.2 Basis functions

We first consider the space of discontinuous piecewise linear basis functions in one dimension. For the choice of basis functions, we have at least two possibilities. If we consider the intervals $\ell_m = [x_m, x_{m+1}]$ we can use $\phi_{\ell_{m,k}}$, $k = 1, 2$. Given a function $u(x)$, it can be approximated by

$$\hat{u}(x) = \sum_{m=1}^{M} \sum_{k=1}^{2} u_m^k \phi_{\ell_{m,k}}(x) \tag{8.1}$$

In order to recover the coefficients $u_m^k$, we need two conditions for each interval: for instance, we could prescribe interpolation at the two extreme points for each interval. Since in general the function to approximate is discontinuous, we can prescribe $u_m^1 = u(x_{\ell_{m,1}}^+)$ and $u_m^2 = u(x_{\ell_{m,2}}^-)$. But then

$$\hat{u}(x_2) = u_1^2 \phi_{\ell_{1,2}}(x_2) + u_2^1 \phi_{\ell_{2,1}}(x_2) = u_1^2 + u_2^1 = u(x_2^+) + u(x_2^-)$$

The problem is that at the common points between two adjacent intervals there are two basis functions which take value one. In one dimension, it would be possibile to remedy by restricting the basis functions $\phi_{\ell_{m,2}}$ to the interval $[x_{\ell_{m,1}}, x_{\ell_{m,2}})$, except the last $\phi_{\ell_{M,2}}$. But in a two-dimensional triangulation it is not possible to specifiy in an easy way which single basis function should take value one at a vertex. In this sense, representantion (8.1) should be understood as

$$\hat{u}(x)_{|\ell_m} = \sum_{k=1}^{2} u_m^k \phi_{\ell_{m,k}}(x)$$

In two dimensions, 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 three interpolation conditions at inner points close to the vertices. For instance, in one dimension,

$$y_{\ell_{m,k}} = \frac{x_{\ell_{m,1}} + x_{\ell_{m,2}}}{2} + 0.99 \cdot \left( x_{\ell_{m,k}} - \frac{x_{\ell_{m,1}} + x_{\ell_{m,2}}}{2} \right)$$

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. This approach is taken by [4]. 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 0 and 1, namely

$$L_0(x) = 1, \quad L_1(x) = \sqrt{3}(2x - 1)$$

They satisfy

$$\int_0^1 L_h(x) L_k(x) \mathrm{d}x = \delta_{hk}$$

Then, given the interval $I_i = [x_i, x_{i+1}]$, it is

$$\frac{1}{x_{i+1}-x_i}\int_{I_i} L_h\left(\frac{x-x_i}{x_{i+1}-x_i}\right) L_k\left(\frac{x-x_i}{x_{i+1}-x_i}\right)\mathrm{d}x = \delta_{hk}$$

We can therefore define

$$L_h^i(x)|_{I_j} = \delta_{ij}\frac{1}{\sqrt{x_{i+1}-x_i}}L_h\left(\frac{x-x_i}{x_{i+1}-x_i}\right)$$

and we have

$$\int_{I_l} L_h^i(x)L_k^j(x)\mathrm{d}x = \delta_{hk}\delta_{ij}\delta_{il}\delta_{jl}$$

The set $\{L_h^i(x)\}_{0\le h\le 1, 1\le i\le M}$ is the set of basis functions. Given $v(x) \in L^2([0,1])$, its approximation in the finite element space is

$$\hat{v}(x) = \sum_{m=1}^{M}\sum_{k=0}^{1}\hat{v}_m^k L_k^m(x)$$

with

$$\hat{v}_m^k = \int_{I_m} v(x)L_k^m(x)\mathrm{d}x$$

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 two dimensions, an appropriate orthonormal basis from $1, x, y$ can be obtained by the Gram–Schmidt procedure. The d.o.f. of a piecewise linear discontinuos elements is given by three times the number of triangles.

## 8.3 Computation of local error estimator

We consider the following local error estimator

$$\eta_K(u_h) = \sqrt{h_K^2\left\|f+\Delta u_h\right\|_{L^2(K)}^2 + \sum_{e\in K}|e|\left\|\left[\left[\frac{\partial u_h}{\partial \mathbf{n}}\right]\right]\right\|_{L^2(e)}^2}$$

used in [4] (similar to that used in [9]). The jump notation means $(\nabla u_h)^+ \cdot \mathbf{n}^+ + (\nabla u_h)^- \cdot \mathbf{n}^-$ on internal edges and its $L^2$ norm corresponds to the $L^2$ norm of $\texttt{jump}(\nabla u_h \cdot \texttt{N})$. In fact, with reference to Figure 8.1, if $K$ is the triangle denoted by $+$

$$\left|\left[\left[\frac{\partial u_h}{\partial \mathbf{n}}\right]\right]\right| = \left|\nabla u_h^+ \cdot \mathbf{n}^+ + \nabla u_h^- \cdot \mathbf{n}^-\right| = \left|(\nabla u_h^- - \nabla u_h^+)\cdot \mathbf{n}^+\right| = |\texttt{jump}(\nabla u_h \cdot \texttt{N})|$$

Each of the two terms under the square root can be computed at once using the basis functions of the space `P0` of piecewise contant functions on a triangulation, which are in fact $\chi_K$ functions.

$$h_K^2 \left\| f + \Delta u_h \right\|_{L^2(K)} = \int_\Omega h_K^2 \left| f + \Delta u_h \right|^2 \chi_K = \ell(\chi_K)$$

which is a linear functional on `P0` ($h_K$ can be recovered by `hTriangle`).

# Chapter 9

# Iterative methods for sparse linear systems

## 9.1 Direct methods can be unfeasible

Let us consider the Poisson equation on the square discretized by a regular triangulation with 40 discretization points on each edge. The stiffness matrix has size 1600 with a number of non-zero elements 7840. Its sparsity pattern is in Figure 9.1. If we compute a Cholesky factorization, the upper triangular factor $R$ has 68135 elements different from zero. So, we did not profit of the sparsity of the matrix.

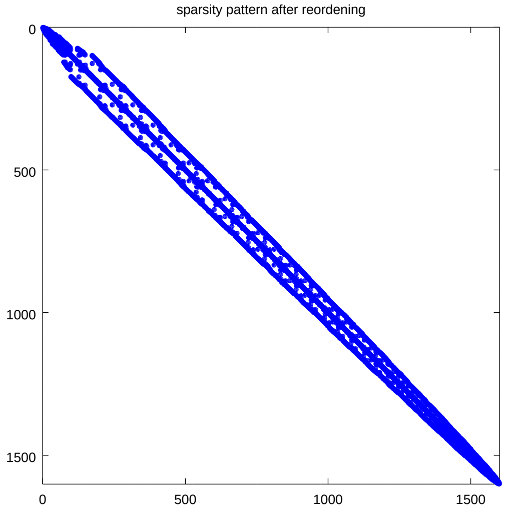


Figure 9.1: Sparsity pattern of the stiffness matrix.

## 9.2 Projection methods

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

$$Px \in M, \quad (x - Px, y)_H = 0 \quad \forall y \in L$$

If $L = M$, than $P$ 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, than 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 6.** *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*

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

*in and only if*

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

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

$$E(x) = (A(\bar{x} - x), \bar{x} - x)^{1/2} = (A(\bar{\delta} - \delta), \bar{\delta} - \delta)^{1/2}$$

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

$$\begin{aligned} E(\tilde{x}) = \min_{x \in x_0 + \mathcal{K}} E(x) &= \min_{\delta \in \mathcal{K}} (A(\bar{\delta} - \delta), \bar{\delta} - \delta)^{1/2} = \min_{\delta \in \mathcal{K}} (R(\bar{\delta} - \delta), R(\bar{\delta} - \delta))^{1/2} = \\ &= \min_{\delta \in \mathcal{K}} \left\|R(\bar{\delta} - \delta)\right\|_2 = \min_{\delta \in \mathcal{K}} \left\|R\bar{\delta} - R\delta\right\|_2 = \min_{w \in R\mathcal{K}} \left\|R\bar{\delta} - w\right\|_2 \end{aligned}$$

which 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 $w = R\delta$, $\delta \in \mathcal{K}$,

$$0 = (R\bar{\delta} - \tilde{w}, w) = (R(\bar{\delta} - \tilde{\delta}), w) = (A(\bar{\delta} - \tilde{\delta}), \delta) = (A(\bar{x} - \tilde{x}), \delta) = (b - A\tilde{x}, \delta).$$

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, $\tilde{\bar{\delta}}$ is the orthogonal projection of $\bar{\delta}$ onto $\mathcal{K}$ through the scalar product $(\cdot,\cdot)_A$. In fact,

$$(\bar{\delta}-\tilde{\bar{\delta}},\delta)_A = \delta^T A(\bar{x}-\tilde{x}) = (b-A\tilde{x},\delta) = 0 \quad \forall\delta\in\mathcal{K}$$

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

**Proposition 7.** *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*

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

*in and only if*

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

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

$$R(x) = \|b-Ax\|_2 = (b-Ax,b-Ax)^{1/2} = (A(\bar{x}-x),A(\bar{x}-x))^{1/2} = (A(\bar{\delta}-\delta),A(\bar{\delta}-\delta))^{1/2}$$

*Proof.* We have

$$\begin{aligned} R(\tilde{x}) &= \min_{x\in x_0+\mathcal{K}} R(x) = \min_{\delta\in\mathcal{K}}(A(\bar{\delta}-\delta),A(\bar{\delta}-\delta))^{1/2} = \\ &= \min_{\delta\in\mathcal{K}}\left\|A(\bar{\delta}-\delta)\right\|_2 = \min_{\delta\in\mathcal{K}}\left\|A\bar{\delta}-A\delta\right\|_2 = \min_{w\in\mathcal{L}}\left\|A\bar{\delta}-w\right\|_2 \end{aligned}$$

which is taken by $\tilde{w} = A\tilde{\bar{\delta}}$, where $\tilde{x} = x_0+\tilde{\bar{\delta}}$. 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 $w\in\mathcal{L}$,

$$0 = (A\bar{\delta}-\tilde{w},w) = (A(\bar{\delta}-\tilde{\bar{\delta}}),w) = (A(\bar{x}-\tilde{x}),w) = (b-A\tilde{x},w)$$

□

### 9.2.1 Conjugate Gradient (CG) method

See [12] for a “painless” introduction. Given a SPD matrix $A$ of dimension $n$, the idea is to solve

$$A\bar{x} = b$$

by minimizing the quadratic functional

$$J(x) = x^TAx - 2b^Tx$$

whose gradient is $\nabla J(x) = 2Ax - 2b = -2r(x)$. 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 $E(x) \geq 0$ and $E(\bar{x}) = 0$. 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

$$x_m = x_{m-1} + \alpha_{m-1} p_{m-1} \tag{9.1}$$

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

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

It is

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

and therefore the minimum of the parabola $E(x_{m-1} + \alpha p_{m-1})$ is taken at

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

**Proposition 8.** *If $\alpha_{m-1}$ is optimal, then*

$$r_m^T p_{m-1} = p_{m-1}^T r_m = 0 \tag{9.2}$$

*Proof.* First of all, we have

$$r_m = b - Ax_m = b - A(x_{m-1} + \alpha_{m-1} p_{m-1}) = r_{m-1} - \alpha_{m-1} A p_{m-1} \tag{9.3}$$

and then

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

□

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 and pointing inside.

Given $p_{m-1}$ and $\alpha_{m-1}$, 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

$$p_m = r_m + \beta_m p_{m-1} \tag{9.4}$$

with $\beta_m$ to find 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 $\alpha_m$

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

We observe that, using (9.2),

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

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 (9.2) 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 (9.4) of $p_{m-1}$ 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 (9.3) for $r_m$, if we multiply by $r_{m-1}^T$ we get

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

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, knowns as *Hestenes–Stiefel*

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

### Some properties of the CG method

It is possible to prove the following thorem

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

$$p_i^T r_{m-1} = 0 \qquad i < m-1 \tag{9.5}$$

$$p_i^T A p_{m-1} = 0 \qquad i < m-1 \tag{9.6}$$

$$r_i^T r_{m-1} = 0 \qquad i < m-1 \tag{9.7}$$

$$\text{span}\{r_0, r_1, \dots, r_{m-1}\} = \text{span}\{r_0, Ar_0, \dots, A^{m-1} r_0\} \tag{9.8}$$

$$\text{span}\{p_0, p_1, \dots, p_{m-1}\} = \text{span}\{r_0, Ar_0, \dots, A^{m-1} r_0\} \tag{9.9}$$

*Sketch of the proof.* First of all, we observe that if for a certain $i$ it is $r_i = 0$, then $x_i$ is the exact solution.

The proof of all properties is by induction. The basic step of each statement is easy since $p_0 = r_0$. Then, it is important to assume all the statemets true for $m-1$ and prove them for $m$. □

**Definition 1.** *The space* $\mathcal{K}_m = \mathrm{span}\{r_0, Ar_0, \dots, A^{m-1}r_0\}$ *is called* Krylov *space.*

The set $\{r_0, r_1, \dots, r_{m-1}\}$ is an orthogonal basis for the Krylov space, which has therefore dimension $m$. It follows that the set $\{p_0, p_1, \dots, p_{m-1}\}$ is a set of linear 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 2.** *A set of vectors different from 0 and satisfying*

$$v_i^T A v_j = 0, \quad \text{for } i, j < m,\ i \neq j$$

*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 indipendently the following

**Proposition 9.** *A set of conjugate vectors 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

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

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

**Theorem 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 $p_i$, $i = 0, 1, \ldots, m-1$ are conjugate vectors

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

Now, we observe that

$$p_i^T r_0 = p_i^T (r_1 + \alpha_0 A p_0) = p_i^T r_1 = p_i^T (r_2 + \alpha_1 A p_1) = p_i^T r_2 = \ldots = p_i^T r_i$$

Therefore

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

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

This is a remarkable property: we started with looking for $u \in X$ such that

$$a(u, v) = \ell(v), \quad \forall v \in X$$

Then we selected a proper $X_h \subset X$ and discovered that $u_h$ staisfying

$$a(u_h, v) = \ell(v), \quad \forall v \in X_h$$

satisfyes

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

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

$$E(x) = (x - \bar{x})^T A (x - \bar{x}) = a(v_m - u_h, v_m - u_h) = |||v_m - u_h|||^2$$

The Conjugate Gradient method can find the infimum of $E(x)$ on the space $x_0 + \mathcal{K}_m$. Therefore, the solution $x_m$ of the CG method is the result of an orthogonal projection method onto $\mathcal{K}_m$ (see Proposition 6). This is clear also from the properties of the method, since

$$0 = r_m^T r_i = (b - A x_m, r_i), \quad 0 \leq i \leq m-1$$

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

**Proposition 10.** *The CG algorithm converges in $n$ iterations at maximum.*

*Proof.* The Krylov space $\mathcal{K}_m = \{p_0, p_1, \ldots, p_{m-1}\}$ has dimension $n$ at maximum. □

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

It is possible to prove the following convergence estimate

$$|||e(x_m)||| = \sqrt{E(x_m)} \leq 2\left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)^m |||e(x_0)|||$$

Here $\mathrm{cond}_2(A)$ is the condition number in the 2-norm, that is

$$\mathrm{cond}_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \sqrt{\rho(A^TA)} \cdot \sqrt{\rho(A^{-T}A^{-1})} = \frac{\lambda_{\max}}{\lambda_{\min}}$$

There exists a slightly better estimate

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

where $c = \frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}$ (see [12]).

### Computational costs

If we want to reduce the initial error $E_0$ by a quantity $\varepsilon$, we have to take

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

from which

$$m = \frac{\ln\frac{\varepsilon}{2}}{\ln\left(\frac{\sqrt{\mathrm{cond}_2(A)}-1}{\sqrt{\mathrm{cond}_2(A)}+1}\right)} = \frac{\ln\frac{\varepsilon}{2}}{\ln\left(1-\frac{2}{\sqrt{\mathrm{cond}_2(A)}+1}\right)} \approx \frac{\ln\frac{\varepsilon}{2}}{-\frac{2}{\sqrt{\mathrm{cond}_2(A)+1}}} \approx$$
$$\approx \frac{1}{2}\ln\frac{2}{\varepsilon}\sqrt{\mathrm{cond}_2(A)}$$

For a matrix with $\mathrm{cond}_2(A) \approx h^{-2}$ the number of expected iterations is therefore $\mathcal{O}(1/h)$. 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 $Av = A[v_1, v_2, \ldots, v_n]^T$ can be obtained by

$$\int_\Omega v_h'(x)\varphi_i'(x)\mathrm{d}x$$

## 9.3 Methods for nonsymmetric systems

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

### 9.3.1 Arnoldi's algorithm

It is possibile to factorize a matrix $A \in \mathbb{R}^{n\times n}$ into

$$AV_m = V_m H_m + w_m e_m^T \tag{9.10}$$

where the first column of $V_m$ is $v_1$ given, $V_m \in \mathbb{R}^{n\times m}$ such that $V_m^T V_m = I_m$ and $V_m^T w_m = 0$ and $H_m$ is superior Hessenberg (see [11, § 6.3]). The cost is $\mathcal{O}(m^2)$. From the relation

$$V_m^T A V_m = H_m$$

we get that if $A$ is symmetric, so is $H_m$ and since it is Hessenberg, it is in fact trigiadonal. Therefore, the Gram–Schmidt procedure is *short* and the cost is $\mathcal{O}(m)$.

### 9.3.2 Implicit restarted Arnoldi's algorithm

Let us analyze the method under the popular ARPACK [6] package for eigenvalue problems. It allows to compute "some" eigenvalues of large sparse matrices (such as the largest in magnitute, the smallest, ...). We start with an Arnoldi factorization

$$V_m^T A V_m = H_m$$

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

$$(v, Ax - \theta x) = 0, \quad \forall v \in \mathcal{K}$$

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

$$(V_m y, Ax - \theta x) = y^T V_m^T A V_m s - y^T V_m^T V_m s\theta = y^T (H_m s - \theta s) = 0$$

The couple $(\theta, x)$ is called Ritz pair and it is close to an eigenpair of $A$. In fact

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

where $\beta_m = \|w_m\|_2$. We can compute the eigenvalues of $H_m$, for instance by the QR method, and select an "unwanted" eigenvalue $\mu_m$ (which is an approximation of an eigenvalue $\mu$ of $A$). Then, we apply one iteration of shifted QR algorithm, that is

$$H_m - \mu_m I_m = Q_1 R_1, \quad H_m^+ = R_1 Q_1 + \mu_m I_m$$

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

$$AV_m Q_1 = V_m H_m Q_1 + w_m e_m^T Q_1 \tag{9.11}$$

With few manipulations

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

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, 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

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

Since $\mu_m$ is close to a $\lambda_{\bar{j}}$, $v_1^+$ lacks the component parallel to $x_{\bar{j}}$. Relation (9.11) can be rewritten as

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

ans 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 practice, 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 $Q = Q_1 Q_2 \dots Q_p$ and the first $k$ columns kept. This turns out to be an Arnoldi factorization. In fact

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

where now $V_m^+ = V_m Q$, 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 $Q_j$ are Hessenberg matrices, the last row of $Q$, that is $e_m^T Q$, has the first $k-1$ entries which are zero and then a value $\sigma$ (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

$$AV_k^+ = V_k^+ H_k^+ + w_k^+ e_k^T, \quad w_k^+ = (V_m^+ e_{k+1} h_{k+1,k} + w_m \sigma)$$

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

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

$$Ax = \lambda M x$$

If we are interested into eigenvalues around $\sigma$, first we notice that

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

from which

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

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

Suppose we want to compute in FreeFem++ the eigenvalues closest to $\sigma = 20$ of Poisson's problem

$$\int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega uv, \quad u \in H_0^1(\Omega)$$

We should set

```
real sigma = 20.0;
varf op(u,v) = int2d(Th)(dx(u)*dx(v)+dy(u)*dy(v)-sigma*u*v)
  +on(1,2,3,4,u=0);
varf b(u,v) = int2d(Th)(u*v);
matrix OP = op(Vh,Vh,solver=Crout,factorize=1);
matrix B = b(Vh,Vh);
int nev = 20;  // number of computed eigenvalues close to sigma
real[int] ev(nev); // to store nev eigenvalues
Vh[int] eV(nev);   // to store nev eigenvectors
int k = EigenValue(OP,B,sym=true,sigma=sigma,value=ev,vector=eV,tol=1e-10);
```

We have to pay attention that $OP$ is not positive definite, so a general factorization such as LU or Crout should be used.

### 9.3.3 Solution of overdetermined systems

Suppose we want to "solve" the linear system

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

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

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

Since $\nabla_y \left\|b - \bar{H}y\right\|_2^2 = -2\bar{H}^T b + 2\bar{H}^T\bar{H}y$, 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 usually *not* used in order to compute $y_m$. A second possibility is to compute the QR factorization of $\bar{H}$. If $\bar{H} = QR$, with $Q \in \mathbb{R}^{(m+1)\times(m+1)}$ orthogonal and $R \in \mathbb{R}^{(m+1)\times m}$ upper triangular (of rank $m$), then

$$\bar{H}^T Hy_m = \bar{H}^T b \Leftrightarrow R^T Q^T QRy_m = R^T Q^T b \Leftrightarrow R^T(Ry_m - Q^T b) = 0$$

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 (used by Matlab and GNU Octave) is to consider the SVD decomposition. We have

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

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

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

Therefore

$$\left\|b - \bar{H}y_m\right\|_2 = \left\|U^T(b - \bar{H}(VV^T)y_m)\right\|_2 = \left\|U^Tb - U^T\bar{H}V(V^Ty_m)\right\|_2 = \|f - Sz\|_2$$

where $z = V^Ty$ and $f = U^Tb$. Now, clearly

$$\arg\min \|f - Sz\|_2$$

has components $z_i = f_i/s_i$, $i = 1, 2, \dots, m$ and $y_m = Vz$.

## 9.4 Preconditioning

The idea is to change

$$A\bar{x} = b$$

into

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

in such a way that $P^{-1}A$ is better conditioned than $A$. The main problem for the CG algorithm is that even if $P$ is SPD, $P^{-1}A$ is not SPD. We can therefore factorize $P$ into $P = R^TR$ and consider the linear system

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

Now, $\tilde{A} = R^{-T}AR^{-1}$ is SPD and we can solve the system $\tilde{A}\bar{y} = \tilde{b}$, $\tilde{b} = R^{-T}b$, with the CG 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 CG 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 $m = 1, 2, \dots$ UNTIL $\|r_m\|_2 \leq \text{tol} \cdot \|b\|_2$

$$w_{m-1} = Ap_{m-1}$$

$$\alpha_{m-1} = \frac{z_{m-1}^T r_{m-1}}{p_{m-1}^T w_{m-1}}$$

```
        $x_m = x_{m-1} + \alpha_{m-1} p_{m-1}$
        $r_m = r_{m-1} - \alpha_{m-1} w_{m-1}$
        $P z_m = r_m$
        $\beta_m = \frac{z_m^T r_m}{z_{m-1}^T r_{m-1}}$
        $p_m = z_m + \beta_m p_{m-1}$
    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 $x_1 = A^{-1}b = \overline{x}$. 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 possibile to $A$, from the other it should be easy to "invert". The simplest choice is $P = \mathrm{diag}(A)$. It is called Jacobi preconditioner. This preconditioner is the default in FreeFem++, since quite effective due to the penalty method to impose Dirichlet boundary conditions (it is a sort of balancing of the rows of the matrix). If $P$ is not diagonal, usually it is factorized once and for all into $P = R^T R$, $R$ the triangular Cholesky factor, in such a way that $z_m$ can be recovered by two simple triangular linear systems. A possibile choice is the incomplete Cholesky factorization of $A$. That is, $P = \tilde{R}^T\tilde{R} \approx A$ where

$$\begin{cases} (A - \tilde{R}^T\tilde{R})_{ij} = 0 & \text{if } a_{ij} \neq 0 \\ \tilde{r}_{ij} = 0 & \text{if } a_{ij} = 0 \end{cases}$$

The preconditioned Conjugate Gradient 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 $P$).

### 9.4.1 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

$$\tilde{u}(x) = \sum_{j=1}^{n} \tilde{u}_j \psi_j(x), \quad n \leq m$$

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

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

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

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

We also have

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

and

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

Therefore

$$RQ \approx I_m, \quad QR \approx I_n$$

Thus, in order to solve the “difficult” problem

$$\bar{A}\bar{u} = \bar{b}$$

we may want to compute $\tilde{A}$ 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 $\bar{A}^{-1} \approx (R\tilde{A}Q)^{-1} \approx R\tilde{A}^{-1}Q$.

### 9.4.2 Algebraic preconditioners

# Chapter 10

# Optimization methods

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

## 10.1 Nonlinear conjugate gradient method

We can extend the Conjugate Gradient method for the minimization of $J^R(x)$, $x \in \mathbb{R}^{n\times 1}$ in the following way.

```
• $x_0$ given, $d_0 = g_0 = -\nabla J^R(x_0)$
• FOR $m = 1, 2, \ldots$ UNTIL $\|d_{m-1}\|_2 \leq \mathrm{tol} \cdot \|d_0\|_2$
      $\alpha_{m-1} = \arg\min_{\alpha} J^R(x_{m-1} + \alpha d_{m-1})$
      $x_m = x_{m-1} + \alpha_{m-1} d_{m-1}$
      $g_m = -\nabla J^R(x_m)$
      $\beta_m = \frac{g_m^T \nabla J^R(x_m)}{g_{m-1}^T \nabla J^R(x_{m-1})}$
      $d_m = g_m + \beta_m d_{m-1}$
  END
```

It is in general not necessary to compute exactly $\alpha_{m-1}$. In this case we speak about *inexact linesearch*. It can be performed, for instance, by few steps of golden search of $g(\alpha) = J^R(x_{m-1} + \alpha d_{m-1})$. Or it is possible to look for the zero of $d_{m-1}^T \nabla J^R(x_{m-1} + \alpha d_{m-1})$. The choice of $\beta_m$ corresponds to Fletcher–Reeves. It is possible to use a preconditioner. In fact, suppose $-\nabla J^R(x_m)$ is of the form $b - A(x_m)$ for some $A\colon \mathbb{R}^n \to \mathbb{R}^n$ and $b \in \mathbb{R}^n$. It is then possible to use $\tilde{A}$ as preconditioner and $g_m$ is computed as

$$g_m = -\tilde{A}^{-1} \nabla J^R(x_m)$$

## 10.2 (Quasi)-Newton methods

It is possible to approximate $J^R$ (if regular enough and with an SPD Hessian) by a quadratic model

$$J^R(x) \approx J^R(x_0) + \nabla J^R(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T H_{J^R}(x_0)(x - x_0)$$

The minimum of the model is given by

$$x - x_0 = -H_{J^R}(x_0)^{-1} \nabla J^R(x_0)$$

and therefore we can define

$$x_1 = x_0 - H_{J^R}(x_0)^{-1} \nabla J^R(x_0)$$

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

$$\nabla J^R(x) = 0$$

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

## 10.3 An example: $p$-Laplacian problem

We are interested in the solution of

$$-\mathrm{div}(\left|\nabla u\right|^{p-2} \nabla u) = f$$

with $p > 2$ and homogeneous Dirichlet boundary conditions. We can compute

$$\begin{aligned}
J(u) &= \int_\Omega \frac{\left|\nabla u\right|^p}{p} - \int_\Omega fu \\
J'(u)v &= \int_\Omega \left|\nabla u\right|^{p-2} \nabla u \cdot \nabla v - \int_\Omega fv \\
J''(u)(w,v) &= \int_\Omega (p-2) \left|\nabla u\right|^{p-4} \nabla u \cdot \nabla w \nabla u \cdot \nabla v + \\
&+ \int_\Omega \left|\nabla u\right|^{p-2} \nabla w \cdot \nabla v
\end{aligned}$$

Given $u_h, v_h \in V_h$ and the basis functions $\{\varphi_i\}_{i=1}^m$, it is

$$J^R \colon \mathbb{R}^n \to \mathbb{R}, \quad J^R(\underline{u_h}) = \int_\Omega \frac{\left|\sum_i (u_h)_i \nabla \varphi_i\right|^p}{p} - f \sum_i (u_h)_i \varphi_i$$

$$\left(\nabla J^R(\underline{u_h})\right)_i = \int_\Omega \left|\nabla u_h\right|^{p-2} \nabla u_h \cdot \nabla \varphi_i - \int_\Omega f \varphi_i = J'(u_h)\varphi_i$$

and

$$\underline{v_h}^T \nabla J^R(\underline{u_h}) = \sum_i (v_h)_i \left( \int_\Omega |\nabla u_h|^{p-2} \nabla u_h \cdot \nabla \varphi_i - \int_\Omega f \varphi_i \right) =$$
$$= \int_\Omega |\nabla u_h|^{p-2} \nabla u_h \cdot \sum_i (v_h)_i \nabla \varphi_i - \int_\Omega f \sum_i (v_h)_i \varphi_i = J'(u_h) v_h$$

and therefore $J'(u_h)v_h$ is the scalar product in $\mathbb{R}^m$ of the vectors with components $J'(u_h)\varphi_i$ and $(v_h)_i$, respectively. Moreover, if $w_h \in V_h$, then

$$\left(H_{J^R}(\underline{u_h})\underline{w_h}\right)_i = J''(u_h)(w_h, \varphi_i) =$$
$$= \int_\Omega (p-2) |\nabla u_h|^{p-4} \nabla u_h \cdot \left( \sum_j (w_h)_j \nabla \varphi_j \right) \nabla u_h \cdot \nabla \varphi_i +$$
$$+ \int_\Omega |\nabla u_h|^{p-2} \left( \sum_j (w_h)_j \nabla \varphi_j \right) \cdot \nabla \varphi_i =$$
$$= \sum_j (w_h)_j \left( \int_\Omega (p-2) |\nabla u_h|^{p-4} \nabla u_h \cdot \nabla \varphi_j \nabla u_h \cdot \nabla \varphi_i + \right.$$
$$\left. + \int_\Omega |\nabla u_h|^{p-2} \nabla \varphi_j \cdot \nabla \varphi_i \right)$$

and therefore $J''(u_h)(w_h, \varphi_i)$ is the matrix-vector product between the (symmetric) matrix $H_{J^R}(\underline{u_h}) = J''(u_h)(\varphi_j, \varphi_i)$ and the vector with components $(w_h)_j$.

### 10.3.1 Minimization approach

The minimization approach is: find $u_h \in V_h$ such that

$$u_h = \arg \min_{v_h \in V_h} J(v_h)$$

In order to use, for instance, the nonlinear Conjugate Gradient method (see § 10.1) we just need $\nabla J^R(x_m) = J'((u_h)_m)\varphi_i$, where $x_m = (\underline{u_h})_m$ is the current approximation of $\underline{u_h}$. As a preconditioner it is possible to use, for a fixed $\overline{u_h}$, the matrix

$$\tilde{H}_{J^R}(\underline{\overline{u_h}})_{i,j} = \int_\Omega |\nabla \overline{u_h}|^{p-2} \nabla \varphi_j \cdot \nabla \varphi_i$$

which corresponds to a simplification of $H_{J^R}(\underline{\overline{u_h}})$.

### 10.3.2 Galerkin approach

The Galerkin approach is: find $u_h \in V_h$ such that

$$J'(u_h)v_h = 0 \quad \forall v_h \in V_h$$

We can consider the nonlinear function $F(u_h)$ with components

$$F_i(u_h) = J'(u_h)\varphi_i$$

In order to write Newton's method for it, we need its Jacobian applied to $\delta_h \in V_h$, whose $i$-th component is

$$\nabla_{u_h} F_i(u_h)\delta_h = J''(u_h)(\delta_h, \varphi_i)$$

and Newton's iteration writes

$$\begin{aligned}&\nabla_{u_h} F(u_h^r)\delta_h^r = -F(u_h^r)\\ &u_h^{r+1} = u_h^r + \delta_h^r\end{aligned}$$

# Chapter 11

# ADR equations

The Advection-Diffusion-Reaction equation (see [5]) is

$$\begin{cases} -\mathrm{div}(\mu\nabla u) + b\cdot\nabla u + \sigma u = f, & u \in \Omega \subset \mathbb{R}^n,\ n = 1, 2, 3 \\ \quad u = 0, & u \in \partial\Omega \end{cases}$$

## 11.1 Estimates

### 11.1.1 $\mathrm{div}(b), \sigma \in L^2(\Omega)$

In this case we have

$$\int_\Omega v b\cdot\nabla v \mathrm{d}\Omega + \int_\Omega \sigma v^2 \mathrm{d}\Omega = \int_\Omega v^2 \left(-\frac{1}{2}\mathrm{div}(b) + \sigma\right)\mathrm{d}\Omega$$

ans using Cauchy–Schwartz inequality

$$\int_\Omega \left| v^2 \left(-\frac{1}{2}\mathrm{div}(b) + \sigma\right)\right| \mathrm{d}\Omega \leq \left\| v^2 \right\|_{L^2(\Omega)} \left\| -\frac{1}{2}\mathrm{div}(b) + \sigma \right\|_{L^2(\Omega)}$$

Since $v \in H^1(\Omega)$, then $v \in L^4(\Omega)$ (see [9, § 2.5]). Moreover

$$\begin{aligned} \left|\int_\Omega \sigma u v \mathrm{d}\Omega\right| &\leq \|\sigma\|_{L^2(\Omega)} \|uv\|_{L^2(\Omega)} \leq \|\sigma\|_{L^2(\Omega)} \|u\|_{L^4(\Omega)} \|v\|_{L^4(\Omega)} \leq \\ &\leq C \|\sigma\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \end{aligned}$$

In fact $H^1(\Omega) \subset L^4(\Omega)$ with a continuous immersion.

### 11.1.2 $\mathrm{div}(b), \sigma \in L^\infty(\Omega)$

In this case we have

$$\int_\Omega vb \cdot \nabla v \mathrm{d}\Omega + \int_\Omega \sigma v^2 \mathrm{d}\Omega = \int_\Omega v^2 \left(-\frac{1}{2}\mathrm{div}(b) + \sigma\right) \mathrm{d}\Omega$$

ans using Hölder's inequality

$$\int_\Omega \left| v^2 \left(-\frac{1}{2}\mathrm{div}(b) + \sigma\right)\right| \mathrm{d}\Omega \leq \left\| v^2 \right\|_{L^1(\Omega)} \left\| -\frac{1}{2}\mathrm{div}(b) + \sigma \right\|_{L^\infty(\Omega)}$$

Therefore $v \in L^2(\Omega)$. Moreover

$$\begin{aligned}\left|\int_\Omega \sigma uv \mathrm{d}\Omega\right| &\leq \|\sigma\|_{L^\infty(\Omega)} \|uv\|_{L^1(\Omega)} \leq \|\sigma\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \\ &\leq \|\sigma\|_{L^\infty(\Omega)} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}\end{aligned}$$

## 11.2 One-dimensional AD problem

For the problem

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

the analytical solution is

$$u(x) = \begin{cases} x, & b = 0, \\ \dfrac{\exp\left(\frac{b}{\mu}x\right) - 1}{\exp\left(\frac{b}{\mu}\right) - 1}, & b \neq 0 \end{cases}$$

After the discretization with piecewise linear finite elements, for the difference equation we have

$$-(\mathrm{Pe} + 1) + 2\rho + (\mathrm{Pe} - 1)\rho^2 = 0$$

If $\mathrm{Pe} = 0$, $\rho_1 = \rho_2 = 1$, the general solution is

$$u_i = A_1 \rho_1^i + A_2 i \rho_2^i$$

from which, by imposing boundary conditions,

$$A_1 = 0, \quad A_2 = \frac{1}{M}$$

If Pe = 1, $\rho_1 = (\mathrm{Pe}+1)/2$. Therefore

$$u_i = A\rho^i$$

and by imposing boundary conditions,

$$A = 0$$

In the general case

$$\rho_1 = (1+\mathrm{Pe})/(1-\mathrm{Pe}), \quad \rho_2 = 1$$

We can try to find $\varepsilon$ such that $u(1-\varepsilon) \approx 0$. It is

$$f(\varepsilon) = u(1-\varepsilon) \approx f(0) + f'(0)\varepsilon = 1 + \frac{-\frac{b}{\mu}\exp\left(\frac{b}{\mu}\right)}{\exp\left(\frac{b}{\mu}\right)-1}\varepsilon = 0$$

from which

$$\varepsilon = \frac{\mu}{b}\frac{\exp\left(\frac{b}{\mu}\right)-1}{\exp\left(\frac{b}{\mu}\right)} = \mathcal{O}\left(\frac{\mu}{b}\right)$$

Therefore the boundary layer width is $\mathcal{O}\left(\frac{\mu}{b}\right)$.

### 11.2.1 Artificial diffusion

We want to find a function $\phi$ such as the new Péclet number is

$$\frac{\mathrm{Pe}}{1+\phi(\mathrm{Pe})}$$

We need

- $\phi(\mathrm{Pe}) \geq \mathrm{Pe} - 1$, but not too much
- $\phi(|b|\, h/2\mu) \in \mathcal{O}(h^2)$, for $h \to 0$, so that the new scheme is still second order in $h$

A possibile solution is

$$\phi(z) = z - 1 + e^{-z} = \frac{z^2}{2} + \mathcal{O}(z^3)$$

A better solution (Scharfetter–Gummel) is

$$\phi(z) = z - 1 + \frac{2z}{e^{2z}-1} = \frac{z^2}{3} + \mathcal{O}(z^4)$$

## 11.3 A time-dependent nonlinear ADR equation

We want to solve

$$u_t + \boldsymbol{b} \cdot \operatorname{div}(u) = \mu \Delta u + \rho u^2 (u-1), \quad \Omega$$

which can be rewritten as

$$u_t = Lu + g(u)$$

with the $\theta$-method with time step $k$. We want to make zero the function

$$F_i(u_h^{n+1}) = \int (u_h^{n+1} - u_h^n)\varphi_i - k\theta \left(a_L(u_h^{n+1}, \varphi_i) + g(u_h^{n+1})\varphi_i\right) - k(1-\theta)\left(a_L(u_h^n, \varphi_i) + g(u_h^n)\varphi_i\right)$$

We compute its Jacobian applied to $\delta_h$ (which is a bilinear form in $\delta_h$ and $\varphi_i$)

$$J_F(u_h^{n+1})\delta_h = \int \delta_h \varphi_i - k\theta \left(a_L(\delta_h, \varphi_i) + g'(u_h^{n+1})\delta_h \varphi_i\right)$$

Therefore, each Newton iteration is the solution of the week formulation

$$\begin{aligned} J_F(u_h^{n+1,r})\delta_h^r + F(u_h^{n+1,r}) &= 0 \\ u_h^{n+1,r+1} &= u_h^{n+1,r} + \delta_h^r \end{aligned}$$

with $u_h^{n+1,0} = u_h^n$. Such a week formulation requires boundary conditions. If $u_h^{n+1}|_{\partial\Omega} = u_h^n|_{\partial\Omega}$, then it is enough to set $\delta_h^r|_{\partial\Omega} = 0$. Otherwise, it is necessary first to set $u_h^{n+1,0}|_{\partial\Omega}$ to the proper boundary conditions.

If we consider the two-dimensional equation with $b_1 = b_2$, the exact solution is

$$u(t,x,y) = \frac{1}{1 + \exp(a(x+y-bt)+c)}$$

where $a = \sqrt{\rho/(4\mu)}$, $b = 2b_1 + \sqrt{\rho\mu}$ and $c = a(b-1)$ (see [5, § 10.6]). Of course, it requires time-dependent Dirichlet boundary conditions.

# Bibliography

[1] R. A. Adams and J. J. F Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, 2nd edition, 2003.

[2] M. Bramanti. *Introduzione alla formulazione debole dei problemi ai limiti per EDP per il Corso di Metodi Matematici per l'Ingegneria*. Politecnico di Milano, 2012. `www1.mate.polimi.it/~bramanti/corsi/pdf_metodi/sobolev2.pdf`.

[3] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer–Verlag, 2nd revised edition, 2008. `http://www.cs.uu.nl/geobook/interpolation.pdf`.

[4] F. Hecht. *Freefem++*. Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, 3rd edition, 2014.

[5] W. Hundsdorfer. Numerical Solution of Advection-Diffusion-Reaction Equations. Technical report, Thomas Stieltjes Institute, CWI, Amsterdam, 2000.

[6] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK User's Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. Software Environments Tools. SIAM, 1998.

[7] MIT OpenCourseWare. Finite element methods for elliptic problems, 2012. `http://www.ocw.nur.ac.rw/OcwWeb/Aeronautics-and-Astronautics/16-920JNumerical-Methods`

[8] A. Quarteroni. *Modellistica numerica per problemi differenziali*. Springer-Verlag, Italy, 5th edition, 2012.

[9] A. Quarteroni. *Numerical Models for Differential Problems*, volume 8 of *Modeling, Simulation and Applications*. Springer, second edition, 2014.

[10] B. Rivière, M. F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM J. Numer. Anal.*, 39(3):902–931, 2001.

[11] Y. Saad. *Iterative Methods for Sparse Linear systems.* Other Titles in Applied Mathematics. SIAM, 2nd edition, 2003.

[12] J. R. Shewchuck. An introduction to Cojugate Gradient method without the agonizing pain, 1994. http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.

[13] J. R. Shewchuk. Triangle: A two-dimensional quality mesh generator and delaunay triangulator, 2005. http://www.cs.cmu.edu/~quake/triangle.html.

[14] Voronoi diagram and Delaunay triangulation. http://asishm.myweb.cs.uwindsor.ca/cs557/F10/handouts/VDandDT.pdf.