Scientific Computing

Dr. Marco Caliari

a.a. 2012–2013

Chapter 1

Variational methods

1.1 Variational formulation of Poisson's problem

Let us consider the boundary value problem (Poisson's equation)

$$\begin{cases} -u''(x) = g(x), & x \in (0,1) \\ \quad u(0) = u(1) = 0 \end{cases} \tag{1.1}$$

where $g \in \mathcal{C}^0([0,1])$. We introduce the following space:

$$V = \{v\colon\ v \in \mathcal{C}^1([0,1]),\, v(0) = v(1) = 0\}$$

equipped with the scalar product

$$(u,v) = \int_0^1 u(x)v(x)\mathrm{d}x$$

Theorem 1 (Variational formulation)**.** *If $u(x)$ is the solution of* (1.1)*, then $u \in V$ and*

$$(u',v') = (g,v), \quad \forall v \in V \tag{1.2}$$

Proof. Let u be the solution of (1.1). Then, for any $v \in V$,

$$\int_0^1 -u''(x)v(x)\mathrm{d}x = \int_0^1 g(x)v(x)\mathrm{d}x = (g,v)$$

Integrating by parts,

$$\int_0^1 -u''(x)v(x)\mathrm{d}x = -u'(x)v(x)\Big|_0^1 + \int_0^1 u'(x)v'(x)\mathrm{d}x = (u',v')$$

since $v(0) = v(1) = 0$. □

Since problem (1.1) has a unique $u \in \mathcal{C}^2 \subset V$ solution, this is clearly also a solution for (1.2). Why do we introduce then the variational formulation? Let us consider instead the following problem

$$\begin{cases} -u''(x) = g_\varepsilon(x), & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{1.3}$$

where

$$g_\varepsilon(x) = \begin{cases} 0 & 0 \le x < \frac{1}{2} - \varepsilon \\ -\frac{1}{2\varepsilon} & \frac{1}{2} - \varepsilon \le x \le \frac{1}{2} + \varepsilon \\ 0 & \frac{1}{2} + \varepsilon < x \le 1 \end{cases}$$

Since g_ε is the load density, the total load is

$$\int_0^1 g_\varepsilon(x)\mathrm{d}x = -1$$

The "solution" of (1.3) is

$$u_\varepsilon(x) = \begin{cases} -\dfrac{1}{2}x & 0 \le x \le \dfrac{1}{2} - \varepsilon \\ \dfrac{1}{4\varepsilon}\left(x - \dfrac{1}{2}\right)^2 + \dfrac{\varepsilon - 1}{4} & \dfrac{1}{2} - \varepsilon \le x \le \dfrac{1}{2} + \varepsilon \\ -\dfrac{1}{2}(1-x) & \dfrac{1}{2} + \varepsilon \le x \le 1 \end{cases}$$

Since $u_\varepsilon''(1/2 \pm \varepsilon)$ does not exist, we cannot state that $-u_\varepsilon''(x) = g_\varepsilon(x)$, $x \in (0,1)$. But it is true that $u_\varepsilon \in V$ and

$$\begin{aligned} \int_0^1 u_\varepsilon'(x)v'(x)\mathrm{d}x &= \int_0^{\frac{1}{2}-\varepsilon} u_\varepsilon'(x)v'(x)\mathrm{d}x + \int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} u_\varepsilon'(x)v'(x)\mathrm{d}x + \int_{\frac{1}{2}+\varepsilon}^1 u_\varepsilon'(x)v'(x)\mathrm{d}x = \\ &= -\int_0^{\frac{1}{2}-\varepsilon} u_\varepsilon''(x)v(x)\mathrm{d}x - \int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} u_\varepsilon''(x)v(x)\mathrm{d}x - \int_{\frac{1}{2}+\varepsilon}^1 u_\varepsilon''(x)v(x)\mathrm{d}x = \\ &= -\int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} \frac{1}{2\varepsilon}v(x)\mathrm{d}x = \int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} g_\varepsilon(x)v(x)\mathrm{d}x = \int_0^1 g_\varepsilon(x)v(x)\mathrm{d}x \end{aligned}$$

that is $u_\varepsilon \in V$ is a solution of the variational formulation.

Coming back to problem (1.1), without any assumption on g, the classical $\mathcal{C}^2$ solution is called *strong solution* of (1.1), whereas the solution of (1.2) is the *weak solution* of (1.1). With the previous theorem and example, we showed that if the strong solution exists, it is also a weak solution. But the other way round is not true: the weak solution may exist, but not the strong one. Anyhow, if $u \in V$ is solution of (1.2) *and* $u \in \mathcal{C}^2([0,1])$ (notice that $\mathcal{C}^2([0,1]) \subset V$) and $g \in \mathcal{C}^0([0,1])$, then $0 = (u' - g, v) = (-u'' - g, v)$ for any $v \in V$. Since $u'' + g$ is continuous, we get $-u''(x) = g(x)$ for $0 < x < 1$. Hence, u is a strong solution, too.

The variational formulation (1.2) of (1.1) is in fact the more "physical": coming back to the problem of the beam, it allows to describe the case in which the load density $g(x)$ is not continuous. What it is necessary is just the existence of (g, v), $v \in V$. The weak solution, if it exists, is unique: in fact, if u_1 and u_2 are two solutions of (1.2), then

$$(u_1' - u_2', v') = 0, \quad \forall v \in V$$

and, in particular, for $v = u_1 - u_2$. Hence

$$\int_0^1 (u_1'(x) - u_2'(x))^2 \mathrm{d}x = 0$$

and $u_1'(x) - u_2'(x) = (u_1(x) - u_2(x))' = 0$. Hence $u_1 - u_2$ is constant and since $u_1(0) - u_2(0) = 0$, then $u_1(x) - u_2(x) = 0$.

In the general case, that is for more general problems than (1.1), the space V is not good for the variational formulation, essentially because it is not complete with respect to the above scalar product. With respect to problem (1.3), clearly the solution, for $\varepsilon \to 0$ converges to

$$u(x) = \begin{cases} -\dfrac{1}{2}x & 0 \le x \le \dfrac{1}{2} \\ -\dfrac{1}{2}(1-x) & \dfrac{1}{2} \le x \le 1 \end{cases}$$

which is not in V. We require that the integrals of $u'v'$ e gv exist. Therefore, we set

$$V = H_0^1(0,1) = \{v \in L^2(0,1),\ v' \in L^2(0,1),\ v(0) = v(1) = 0\}$$

equipped with the scalar product

$$\langle u, v \rangle_{H^1} = \int_0^1 u(x)v(x)\mathrm{d}x + \int_0^1 u'(x)v'(x)\mathrm{d}x$$

where the derivatives are distributional. Such a space is complete, and it is exactly the closure with respect to the scalar product of the space introduced above. Moreover it contains also not differentiable (in the classical sense) continuous functions, such as piecewise linear continuous functions. Now, if we assume $g \in L^2(0,1)$, than the variational formulation writes

$$\text{find } u \in H_0^1(0,1)\colon \int_0^1 u'(x)v'(x)\mathrm{d}x = \int_0^1 g(x)v(x)\mathrm{d}x, \quad \forall v \in H_0^1(0,1)$$

1.1.1 Two-dimensional Poisson's problem

The problem, with homogeneous Dirichlet boundary conditions, is

$$\begin{cases} -\Delta u(x,y) = g(x,y), & (x,y) \in \Omega \\ u(x,y) = 0, & (x,y) \in \Gamma = \partial\Omega \end{cases}$$

We proceed as in the one-dimensional case, and using *Green's formula*, we end up with

$$-\int_\Omega \Delta u v \mathrm{d}\Omega = \int_\Omega \nabla u \cdot \nabla v \mathrm{d}\Omega - \int_\Gamma \frac{\partial u}{\partial n} v \mathrm{d}\Gamma = \int_\Omega g v \mathrm{d}\Omega$$

Since the test functions v are zero at the boundary, the integral on Γ disappears. The variational formulation then writes:

$$\text{find } u \in H_0^1(\Omega)\colon \int_\Omega \nabla u \cdot \nabla v \mathrm{d}\Omega = \int_\Omega g v \mathrm{d}\Omega, \quad \forall v \in H_0^1(\Omega)$$

1.2 Equivalence

We have seen that we further assumptions on u (essentially, $u \in \mathcal{C}^2$) the weak and the strong formulation are equivalent. Without that assumption, they still are equivalent in the sense of distributions. In fact, if we restrict the test functions to $\mathcal{D}(\Omega)$ (infinitely differentiable functions with compact support, $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$), we have

$$\int_\Omega \nabla u \cdot \nabla \varphi = \int_\Omega g \varphi \mathrm{d}\Omega, \quad \forall \varphi \in \mathcal{D}(\Omega)$$

Then we apply Green's formula

$$-\int_\Omega \Delta u \varphi \mathrm{d}\Omega + \int_\Gamma \frac{\partial u}{\partial n} \varphi \mathrm{d}\Gamma = \int_\Omega g \varphi \mathrm{d}\Omega, \quad \forall \varphi \in \mathcal{D}(\Omega)$$

where we mean

$$-\int_\Omega \Delta u \varphi = \langle -\Delta u, \varphi \rangle$$

$$\int_\Gamma \frac{\partial u}{\partial n} \varphi \mathrm{d}\Gamma = \langle \frac{\partial u}{\partial n}, \varphi \rangle$$

$$\int_\Omega g \varphi = \langle g, \varphi \rangle$$

Since $\varphi \in \mathcal{D}(\Omega)$, the boundary integral vanishes and

$$\langle -\Delta u - g, \varphi \rangle = 0$$

that is $-\Delta u - g$ is the null distribution in $\mathcal{D}'(\Omega)$.

In this context, taking the limit for $\varepsilon \to 0$ for problem (1.3), one gets the distribution

$$-\Delta u + \delta_{1/2}$$

where $\delta_{1/2}$ is Dirac's delta distribution. This is the model for a load concentrated in a single point on the beam.

1.3 Existence of a weak solution

Theorem 2 (Lax–Milgram's lemma plus corollary)**.** *Let V be a Hilbert space, $a(\cdot,\cdot)\colon V \times V \to \mathbb{R}$ a bilinear continuous and coercive form, $F(\cdot)\colon V \to \mathbb{R}$ a linear and continuous functional. Then, there exists a unique solution to the problem*

$$\textit{find } u \in V\colon a(u,v) = F(v), \quad \forall v \in V$$

Moreover

$$\|u\|_V \leq \frac{1}{\alpha}\|F\|_{V'}$$

where α is the coercitivity constant.

Proof. The classical Lax–Milgram theorem (Riesz's representation and Banach's close image) and (F is bounded)

$$\alpha\|u\|_V^2 \leq a(u,u) = F(u) \leq |F(u)| \leq \|F\|_{V'}\|u\|_V$$

□

As an exercise, we check that Poisson's problem falls in this case. We have $V = H_0^1(\Omega)$ with scalar product

$$\langle u,v\rangle_{H^1} = \int_\Omega uv\mathrm{d}\Omega + \int_\Omega \nabla u \cdot \nabla v\mathrm{d}\Omega$$

and $g \in L^2$

- $a(u,v) = \int_\Omega \nabla u \cdot \nabla v\mathrm{d}\Omega$ is bilinear (obvious), continuous

  $$|a(u,v)| \leq \|\nabla u\|_{L^2}\|\nabla v\|_{L^2} \leq \|\nabla u\|_{H^1}\|\nabla v\|_{H^1}$$

  (Cauchy–Schwarz's inequality), coercive

  $$\|u\|_{H^1}^2 = \|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2 \leq (C+1)\|\nabla u\|_{L^2}^2 \Rightarrow a(u,u) \geq \frac{1}{C+1}\|u\|_{H^1}^2$$

  (Poincaré's inequality)

- $F(v) = \int_\Omega gv\mathrm{d}\Omega$ is linear (obvious) and bounded

$$|F(v)| \leq \|g\|_{L^2}\|v\|_{L^2} \leq \|g\|_{L^2}\|v\|_{H^1} \Rightarrow \|F\| = \sup_{v\neq 0}\frac{|F(v)|}{\|v\|_{H^1}} \leq \|g\|_{L^2}$$

and hence continuous.

1.4 Variational approximation method

Let us take a finite dimensional subspace V_m of V. We then look for $\hat{u} \in V_m$ such that

$$a(\hat{u}, v) = F(v), \quad \forall v \in V_m \tag{1.4}$$

(Galerkin's method).

Theorem 3. *Problem* (1.4) *has a unique solution.*

Proof. It is still a consequence of Lax–Milgram theorem. For the case

$$a(u, v) = \int_0^1 u'(x)v'(x)\mathrm{d}x = (u', v') \quad \text{+homogeneous Dirichlet b.c.},$$

it is possible to directly prove the theorem and some more. Let $\{\phi_j\}_{j=1}^m$ be a basis of V_m. Then

$$\hat{u}(x) = \sum_{j=1}^m \hat{u}_j\phi_j(x)$$

and (1.4) rewrites, for $i = 1, 2, \ldots, m$,

$$\int_0^1 \hat{u}'(x)\phi_i'(x)\mathrm{d}x = \left(\left(\sum_{j=1}^m \hat{u}_j\phi_j\right)', \phi_i'\right) = \sum_{j=1}^m (\phi_j', \phi_i')\hat{u}_j = A\boldsymbol{u} = (g, \phi_i)$$

where $A = (a_{ij}) = (\phi_j', \phi_i')$ e $\boldsymbol{u} = [\hat{u}_1, \ldots, \hat{u}_m]^{\mathrm{T}}$. Let us compute $\boldsymbol{w}^{\mathrm{T}}A\boldsymbol{w}$ for $\boldsymbol{w} = [w_1, \ldots, w_m]^{\mathrm{T}}$. We have (since A is symmetric)

$$\boldsymbol{w}^{\mathrm{T}}A\boldsymbol{w} = \sum_{i=1}^m w_i\left(\sum_{j=1}^m (\phi_i', \phi_j')w_j\right)$$

and then, due to the per linearity of the scalar product

$$\boldsymbol{w}^{\mathrm{T}}A\boldsymbol{w} = \left(\sum_{i=1}^m w_i\phi_i'(x), \sum_{j=1}^m w_j\phi_j'(x)\right) = \int_0^1 \left(\sum_{j=1}^m w_j\phi_j'(x)\right)^2 \mathrm{d}x \geq 0$$

and the result is 0 only if $\sum w_j\phi_j(x)$ is constant (and, hence, null, because of the boundary conditions) Then, A is positive definite. □

The proof above can be done also in the case $a(\cdot,\cdot)$ symmetric. The coercivity of $a(\cdot,\cdot)$ gives the positive definiteness of the matrix $A = (a(\phi_j, \phi_i))$. Matrix A is called *(stiffness matrix)* and (g, ϕ_i) is the *(load vector)*.

Galerkin's method is *strongly consistent*, since

$$a(u, v) = F(v), \quad \forall v \in V_m \Rightarrow a(\hat{u} - u, v) = 0, \quad \forall v \in V_m$$

If $a(\cdot,\cdot)$ is symmetric, there is the following interpretation: $a(\cdot,\cdot)$ is a scalar

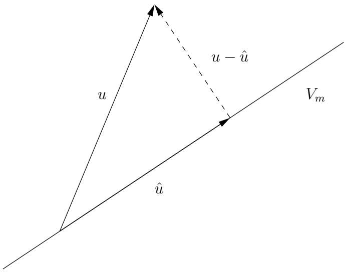


Figure 1.1: $\hat{u}$ as a projection onto V_m.

product in V and $\hat{u}$ is the orthogonal projection onto V_m of u. Then it is the best approximation in V_m of u. In fact, for $v \in V_m$

$$\begin{aligned} a(\hat{u} - u, \hat{u} - u) &= a(\hat{u} - u, \hat{u} - v + v - u) = a(\hat{u} - u, \hat{u} - v) + a(\hat{u} - u, v - u) = \\ &= a(\hat{u} - u, v - u) \end{aligned}$$

due to strong consistence. Using the coercivity and the continuity of $a(\cdot,\cdot)$

$$\alpha \|\hat{u} - u\|_V^2 \leq a(\hat{u} - u, \hat{u} - u) = |a(\hat{u} - u, v - u)| \leq M \|\hat{u} - u\|_V \|v - u\|_V$$

Hence

$$\|\hat{u} - u\|_V \leq \frac{M}{\alpha} \|v - u\|_V \Rightarrow \|\hat{u} - u\|_V \leq \inf_{v \in V_m} \frac{M}{\alpha} \|v - u\|_V$$

Now, we have to choose a subspace $V_m \subset V$ such that

$$\lim_{m \to \infty} \inf_{v \in V_m} \|v - u\|_V = 0$$

or, more in general,

$$\lim_{m \to \infty} \inf_{v \in V_m} \|v - w\|_V = 0, \quad \forall w \in V$$

In that case, it will be

$$\lim_{m \to \infty} \|\hat{u} - u\|_V = 0$$

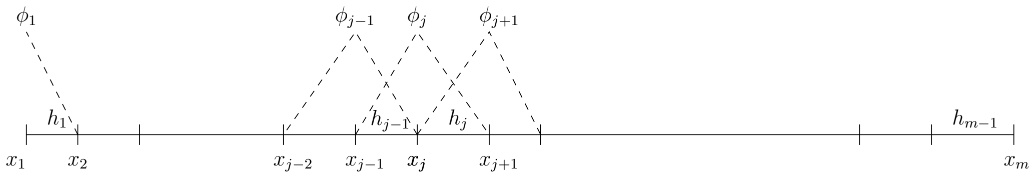


Figure 1.2: Hat functions

1.4.1 Finite Elements Method (FEM)

Let us introduce a discretization of the interval $[0, 1]$ with variable step, as in Figure 1.2. The space V_m is generated by the basis functions $\{\phi_j\}_{j=2}^{m-1}$, defined by

$$\phi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{h_{j-1}}, & x_{j-1} \le x \le x_j \\ \dfrac{x_{j+1} - x}{h_j}, & x_j \le x \le x_{j+1} \\ 0, & \text{elsewhere} \end{cases}$$

ans

$$\phi_j'(x) = \begin{cases} \dfrac{1}{h_{j-1}}, & x_{j-1} < x < x_j \\ -\dfrac{1}{h_j}, & x_j < x < x_{j+1} \\ 0, & \text{elsewhere} \end{cases}$$

However, in order to allow to deal with problems with different boundary conditions, we also consider

$$\phi_1(x) = \begin{cases} \dfrac{x_2 - x}{h_1}, & x_1 \le x \le x_2 \\ 0, & \text{elsewhere} \end{cases}$$

and

$$\phi_1'(x) = \begin{cases} -\dfrac{1}{h_1}, & x_1 < x < x_2 \\ 0, & \text{elsewhere} \end{cases}$$

ans

$$\phi_m(x) = \begin{cases} \dfrac{x - x_{m-1}}{h_{m-1}}, & x_{m-1} \le x \le x_m \\ 0, & \text{elsewhere} \end{cases}$$

and

$$\phi_m'(x) = \begin{cases} \dfrac{1}{h_{m-1}}, & x_{m-1} < x < x_m \\ 0, & \text{elsewhere} \end{cases}$$

Hence, in the approximation

$$\hat{u}(x) = \sum_{j=1}^{m} \hat{u}_j \phi_j(x)$$

the coefficients $\hat{u}_j$ are the values of $\hat{u}$ in the points x_j. Problem (1.4) rewrites

$$\int_0^1 \hat{u}'(x)\phi_i'(x)\mathrm{d}x = \sum_{j=1}^{m} \hat{u}_j \int_0^1 \phi_j'(x)\phi_i'(x)\mathrm{d}x = \sum_{j=1}^{m} \hat{u}_j \int_{x_i-h_{i-1}}^{x_i+h_i} \phi_j'(x)\phi_i'(x)\mathrm{d}x =$$

$$= \sum_{j=1}^{m} \hat{u}_j a_{ij} = \int_{x_i-h_{i-1}}^{x_i+h_i} g(x)\phi_i(x)\mathrm{d}x$$

In case of Neumann's conditions (for instance in $u'(0) = u_0'$), the weak formulation of the problem is

$$-\hat{u}'(x)\phi_i(x)\Big|_0^1 + \int_0^1 \hat{u}'(x)\phi_i'(x)\mathrm{d}x = \int_0^1 g(x)\phi_i(x)\mathrm{d}x, \quad 1 \le i \le m$$

For $i = 1$ we have

$$\hat{u}'(0) + \int_0^1 \hat{u}'(x)\phi_1'(x)\mathrm{d}x = \int_0^1 g(x)\phi_1(x)\mathrm{d}x$$

Hence, the first row of the linear system is

$$\int_0^1 \hat{u}'(x)\phi_1'(x)\mathrm{d}x = -u_0' + \int_0^1 g(x)\phi_1(x)\mathrm{d}x$$

Notice that the problem with two Neumann's conditions is not well-defined, since if $u(x)$ is a solution, then such is $u(x) + k$.

The space V_m can be made of much more regular functions (such as polynomials of higher degree).

Let us see a general implementation strategy for FEM. Suppose we have l elements $\{\ell_j\}_{j=1}^{l}$ (in the one-dimensional case, the intervals) with the associate points. With respect to Figure 1.3, where $m = l + 1$, we have

$$\ell_{j,1} = j, \ \ell_{j,2} = j + 1, \quad 1 \le j \le l$$

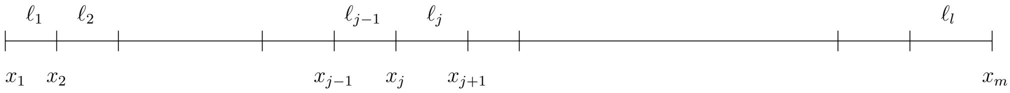


Figure 1.3: Points (bottom) and elements (top).

which means that points x_j and x_{j+1} are associate to element ℓ_j. The basis function which has value 1 on node $\ell_{j,k}$ and 0 on node $\ell_{j,3-k}$ has the form (on ℓ_j)

$$\begin{aligned}
\phi_{\ell_{j,1}} &= \frac{a_{\ell_{j,1}} + b_{\ell_{j,1}}x}{\Delta_j} = \begin{vmatrix} 1 & 1 \\ x & x_{\ell_{j,2}} \end{vmatrix} \Big/ \begin{vmatrix} 1 & 1 \\ x_{\ell_{j,1}} & x_{\ell_{j,2}} \end{vmatrix} = \frac{x_{\ell_{j,2}} - x}{x_{\ell_{j,2}} - x_{\ell_{j,1}}} = \frac{x_{\ell_{j,2}} - x}{h_j} \\
\phi_{\ell_{j,2}} &= \frac{a_{\ell_{j,2}} + b_{\ell_{j,2}}x}{\Delta_j} = \begin{vmatrix} 1 & 1 \\ x_{\ell_{j,1}} & x \end{vmatrix} \Big/ \begin{vmatrix} 1 & 1 \\ x_{\ell_{j,1}} & x_{\ell_{j,2}} \end{vmatrix} = \frac{-x_{\ell_{j,1}} + x}{x_{\ell_{j,2}} - x_{\ell_{j,1}}} = \frac{-x_{\ell_{j,1}} + x}{h_j}
\end{aligned}$$

and will contribute to the elements $a_{\ell_{j,k}\ell_{j,k}}$ and $a_{\ell_{j,k}\ell_{j,3-k}}$ (and its symmetric) of the stiffness matrix

$$\begin{aligned}
a_{\ell_{j,k}\ell_{j,k}} &= \int_0^1 \phi'_{\ell_{j,k}}(x)\phi'_{\ell_{j,k}}(x)\mathrm{d}x \\
a_{\ell_{j,k}\ell_{j,3-k}} &= \int_0^1 \phi'_{\ell_{j,k}}(x)\phi'_{\ell_{j,3-k}}(x)\mathrm{d}x
\end{aligned}$$

and to the element $g_{\ell_{j,k}}$ of the right hand side

$$g_{\ell_{j,k}} = \int_0^1 g(x)\phi_{\ell_{j,k}}(x)\mathrm{d}x$$

Hence

$$\begin{aligned}
a_{\ell_{j-1,2}\ell_{j-1,2}} &= a_{\ell_{j,1}\ell_{j,1}} = \int_{\ell_{j-1}} \phi'_{\ell_{j-1,2}}(x)\phi'_{\ell_{j-1,2}}(x)\mathrm{d}x + \int_{\ell_j} \phi'_{\ell_{j,1}}(x)\phi'_{\ell_{j,1}}(x)\mathrm{d}x = \\
&= \int_{\ell_{j-1}} \left(\frac{b_{\ell_{j-1,2}}}{\Delta_{j-1}}\right)^2 \mathrm{d}x + \int_{\ell_j} \left(-\frac{b_{\ell_{j,1}}}{\Delta_j}\right)^2 \mathrm{d}x = \\
&= \int_{\ell_{j-1}} \left(\frac{1}{h_{j-1}}\right)^2 \mathrm{d}x + \int_{\ell_j} \left(\frac{-1}{h_j}\right)^2 \mathrm{d}x = \frac{1}{h_{j-1}} + \frac{1}{h_j} = a_{jj} \\
a_{\ell_{j,2}\ell_{j,2}} &= a_{\ell_{j+1,1}\ell_{j+1,1}} = \int_{\ell_j} \phi'_{\ell_{j,2}}(x)\phi'_{\ell_{j,2}}(x)\mathrm{d}x + \int_{\ell_{j+1}} \phi'_{\ell_{j+1,1}}(x)\phi'_{\ell_{j+1,1}}(x)\mathrm{d}x = \\
&= \int_{\ell_j} \left(\frac{b_{\ell_{j,2}}}{\Delta_j}\right)^2 \mathrm{d}x + \int_{\ell_{j+1}} \left(-\frac{b_{\ell_{j+1,1}}}{\Delta_{j+1}}\right)^2 \mathrm{d}x = \\
&= \int_{\ell_j} \left(\frac{1}{h_j}\right)^2 \mathrm{d}x + \int_{\ell_{j+1}} \left(\frac{-1}{h_{j+1}}\right)^2 \mathrm{d}x = \frac{1}{h_j} + \frac{1}{h_{j+1}} = a_{j+1\,j+1} \\
a_{\ell_{j,1}\ell_{j,2}} &= a_{\ell_{j,2}\ell_{j,1}} = \int_{\ell_j} \phi'_{\ell_{j,1}}(x)\phi'_{\ell_{j,2}}(x)\mathrm{d}x = \int_{\ell_j} \frac{b_{\ell_{j,1}}}{\Delta_j}\frac{b_{\ell_{j,2}}}{\Delta_j}\mathrm{d}x = \int_{\ell_j} -\frac{1}{h_j}\frac{1}{h_j}\mathrm{d}x = \\
&= -\frac{1}{h_j} = a_{j\,j+1} = a_{j+1\,j} \\
g_{\ell_{j-1,2}} &= g_{\ell_{j,1}} = \int_{\ell_{j-1}} g(x)\phi_{\ell_{j-1,2}}(x)\mathrm{d}x + \int_{\ell_j} g(x)\phi_{\ell_{j,1}}(x)\mathrm{d}x = g_{\ell_{j-1}}\frac{h_{j-1}}{2} + g_{\ell_j}\frac{h_j}{2} \\
g_{\ell_{j,2}} &= g_{\ell_{j+1,1}} = \int_{\ell_j} g(x)\phi_{\ell_{j,2}}(x)\mathrm{d}x + \int_{\ell_{j+1}} g(x)\phi_{\ell_{j+1,1}}(x)\mathrm{d}x = g_{\ell_j}\frac{h_j}{2} + g_{\ell_{j+1}}\frac{h_{j+1}}{2}
\end{aligned}$$

where we used the “barycentric” approximation

$$\int_{\ell_j} g(x)\phi_{\ell_{j,k}}(x)\mathrm{d}x \approx \frac{g(x_j)+g(x_{j+1})}{2}\int_{\ell_j} \phi_{\ell_{j,k}}(x)\mathrm{d}x = g_{\ell_j}\frac{h_j}{2} \tag{1.5}$$

This type of approximation is also used in the nonlinear case

$$\int_{\ell_j} g(\hat{u}(x))\phi_{\ell_{j,k}}(x)\mathrm{d}x \approx \frac{g(u_j)+g(u_{j+1})}{2}\frac{h_j}{2}$$

Hence, the *assembly* is done by

quadrature.m

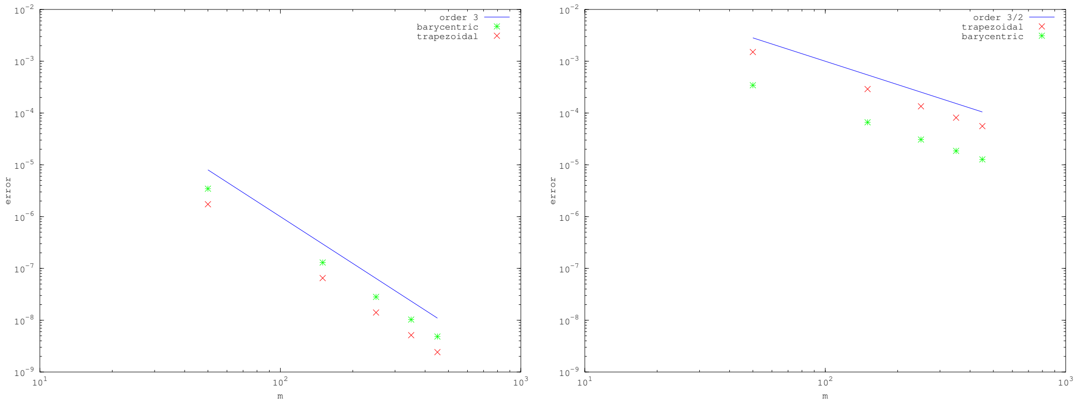


Figure 1.4: Maximum error over $i = 2, 3, \ldots, m-1$ between $\int_0^1 g(x)\phi_i(x)\mathrm{d}x$ and the trapezoidal and the barycentric formula, for $g(x) = |x - 1/2|^{5/2}$ (left) and $g(x) = |x - 1/2|^{1/2}$ (right).

```
• $a_{ij} = 0,\ 1 \leq i, j \leq m,\ g_i = 0,\ 1 \leq i \leq m$
• FOR $j = 1, \ldots, l$
      FOR $k = 1, 2$
          $a_{\ell_{j,k}\ell_{j,k}} = a_{\ell_{j,k}\ell_{j,k}} + \frac{1}{h_j},\ g_{\ell_{j,k}} = g_{\ell_{j,k}} + g_{\ell_j}\frac{h_j}{2}$
          FOR $i = k + 1, 2$
              $a_{\ell_{j,k}\ell_{j,i}} = a_{\ell_{j,k}\ell_{j,i}} - \frac{1}{h_j}$
              $a_{\ell_{j,i}\ell_{j,k}} = a_{\ell_{j,k}\ell_{j,i}}$
          END
      END
  END
```

The i-th row of the linear system turns out to be

$$\begin{bmatrix} 0 & \ldots & 0 & -\frac{1}{h_{i-1}} & \left(\frac{1}{h_{i-1}} + \frac{1}{h_i}\right) & -\frac{1}{h_i} & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \vdots \\ \hat{u}_{i-1} \\ \hat{u}_i \\ \hat{u}_{i+1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \frac{g_{\ell_{i-1}} h_{i-1} + g_{\ell_i} h_i}{2} \\ \vdots \end{bmatrix}$$

very similar to the discretization with finite differences with constant step size. Moreover, if we consider the trapezoidal rule with three points

$$\int_0^1 g(x)\phi_i(x)\mathrm{d}x = \int_{x_{i-1}}^{x_{i+1}} g(x)\phi_i(x)\mathrm{d}x \approx$$
$$\approx \frac{h_{i-1}}{2}g(x_i)\phi_i(x_i) + \frac{h_i}{2}g(x_i)\phi_i(x_i) = \frac{h_{i-1}+h_i}{2}g(x_i)$$

then the two formulations are equivalent. The stiffness matrix is symmetric, but imposing Dirichlet boundary conditions destroys its symmetry. An alternative (numerical) method (called *penalty* method) is to put a large number on the diagonal elements of the rows corresponding to Dirichlet nodes and modifying consequently the right hand side. For Poisson's problem, there is an interesting interpolation property:

Theorem 4. *If*

$$\hat{u}(x) = \sum_{j=1}^{m} \hat{u}_j\phi_j(x), \quad \boldsymbol{u} = [\hat{u}_1, \ldots, \hat{u}_m]^{\mathrm{T}}$$

is the weak solution of

$$\int_0^1 u'(x)v'(x)\mathrm{d}x = \int_0^1 g(x)v(x)\mathrm{d}x$$

where $\{\phi_j(x)\}_j$ are the hat functions, then

$$\hat{u}(x_i) = \hat{u}_i = u(x_i)$$

Proof. The i-th (inner) row of $A\boldsymbol{u}$ is

$$\sum_{j=1}^{m} \hat{u}_j \int_{x_{i-1}}^{x_{i+1}} \phi_j'(x)\phi_i'(x)\mathrm{d}x = -\frac{\hat{u}_{i-1}}{h_{i-1}} + \left(\frac{\hat{u}_i}{h_{i-1}} + \frac{\hat{u}_i}{h_i}\right) - \frac{\hat{u}_{i+1}}{h_i}$$

and it equals

$$\int_0^1 g(x)\phi_i(x)\mathrm{d}x$$

On the other hand,

$$\int_0^1 g(x)\phi_i(x)\mathrm{d}x = \int_0^1 u'(x)\phi_i'(x)\mathrm{d}x$$

and

$$\int_0^1 u'(x)\phi_i'(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} u'(x)\frac{1}{h_{i-1}}\mathrm{d}x - \int_{x_i}^{x_{i+1}} u'(x)\frac{1}{h_i}\mathrm{d}x =$$
$$= \frac{u(x_i)-u(x_{i-1})}{h_{i-1}} - \frac{u(x_{i+1})-u(x_i)}{h_i}$$

that is, $\boldsymbol{u}$ and $[u(x_1),\ldots,u(x_m)]^{\mathrm{T}}$ satisfy the same (positive definite) linear system. □

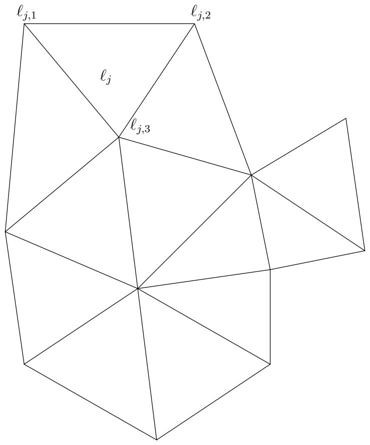


Figure 1.5: Two-dimensional mesh.

The construction in the two-dimensional case is not much different.

First of all, we consider the basis function $\phi_{\ell_{j,k}}$ which has value 1 on node $\ell_{j,k}$ and 0 on nodes $\ell_{j,h}$, $h\in\{1,2,3\}$, $h\neq k$ of the triangle ℓ_j. It has the form

$$\phi_{\ell_{j,k}}(x,y)=\frac{a_{\ell_{j,k}}+b_{\ell_{j,k}}x+c_{\ell_{j,k}}y}{2\Delta_j}=\begin{array}{c} k \\ \begin{vmatrix}1&1&1\\ x_{\ell_{j,1}}&x&x_{\ell_{j,3}}\\ y_{\ell_{j,1}}&y&y_{\ell_{j,3}}\end{vmatrix}\end{array}\Bigg/\begin{vmatrix}1&1&1\\ x_{\ell_{j,1}}&x_{\ell_{j,2}}&x_{\ell_{j,3}}\\ y_{\ell_{j,1}}&y_{\ell_{j,2}}&y_{\ell_{j,3}}\end{vmatrix}$$

where Δ_j is the area (with sign) of triangle ℓ_j. We need to compute

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

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

$$\int_{\ell_j}g(x,y)\phi_{\ell_{j,k}}(x,y)\mathrm{d}x\mathrm{d}y$$

for the right hand side. We have

$$\int_{\ell_j}\frac{\partial\phi_{\ell_{j,k}}(x,y)}{\partial x}\frac{\partial\phi_{\ell_{j,h}}(x,y)}{\partial x}\mathrm{d}x\mathrm{d}y=\int_{\ell_j}\frac{b_{\ell_{j,k}}}{2\Delta_j}\frac{b_{\ell_{j,h}}}{2\Delta_j}\mathrm{d}x\mathrm{d}y=\frac{b_{\ell_{j,k}}b_{\ell_{j,h}}}{4|\Delta_j|}$$

$$\int_{\ell_j}\frac{\partial\phi_{\ell_{j,k}}(x,y)}{\partial y}\frac{\partial\phi_{\ell_{j,h}}(x,y)}{\partial y}\mathrm{d}x\mathrm{d}y=\int_{\ell_j}\frac{c_{\ell_{j,k}}}{2\Delta_j}\frac{c_{\ell_{j,h}}}{2\Delta_j}\mathrm{d}x\mathrm{d}y=\frac{c_{\ell_{j,k}}c_{\ell_{j,h}}}{4|\Delta_j|}$$

and

$$\int_{\ell_j} g(x,y)\phi_{\ell_{j,k}}(x,y)\mathrm{d}x\mathrm{d}y \approx \frac{1}{3}\sum_{k=1}^{3} g(x_{\ell_{j,k}}, y_{\ell_{j,k}})\int_{\ell_j}\phi_{\ell_{j,k}}(x,y)\mathrm{d}x\mathrm{d}y = \\ = g_{\ell_j}\frac{|\Delta_j|}{3}$$

The algorithm for the assembly is

```
• $a_{ij} = 0$, $1 \leq i,j \leq m$, $g_i = 0$, $1 \leq i \leq m$

• FOR $j = 1,\ldots,l$

      FOR $k = 1,\ldots,3$
          $a_{\ell_{j,k}\ell_{j,k}} = a_{\ell_{j,k}\ell_{j,k}} + \frac{b_{\ell_{j,k}}b_{\ell_{j,k}}}{4|\Delta_j|} + \frac{c_{\ell_{j,k}}c_{\ell_{j,k}}}{4|\Delta_j|}$, $g_{\ell_{j,k}} = g_{\ell_{j,k}} + g_{\ell_j}\frac{|\Delta_j|}{3}$
          FOR $i = k+1,\ldots,3$
              $a_{\ell_{j,k}\ell_{j,i}} = a_{\ell_{j,k}\ell_{j,i}} + \frac{b_{\ell_{j,k}}b_{\ell_{j,i}}}{4|\Delta_j|} + \frac{c_{\ell_{j,k}}c_{\ell_{j,i}}}{4|\Delta_j|}$
              $a_{\ell_{j,i}\ell_{j,k}} = a_{\ell_{j,k}\ell_{j,i}}$
          END

      END

  END
```

Theorem 5. *Let be given the variational problem*

$$a(u,v) = F(v), \quad \forall v \in V$$

where $u\colon \Omega \to \mathbb{R}$, Ω polygonal, *is the solution and* $\hat{u}$ *its approximation by finite elements of degree* $r > 0$*. Then, under weak assumptions on the regularity of the mesh,*

- *if* $u \in H^{r+1}(\Omega)$

$$\|\hat{u} - u\|_{H^1} \leq \frac{M}{\alpha}Ch^r|u|_{H^{r+1}}$$

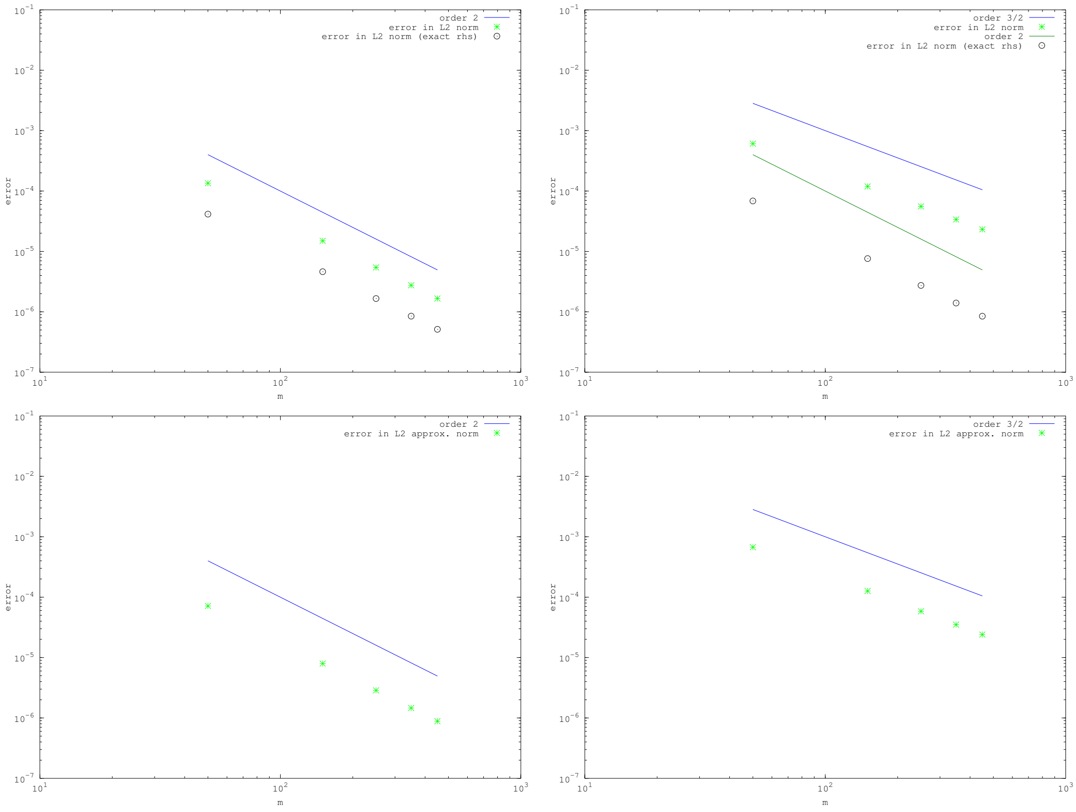
- *if* $u \in H^{p+1}(\Omega)$ *for some* $p > 0$

$$\|\hat{u} - u\|_{L^2} \leq Ch^{s+1}|u|_{H^{s+1}}, \quad s = \min\{r,p\}$$

where h *is the largest diameter of the elements and*

$$|u|^2_{H^{s+1}} = \sum_{|\alpha|=s+1}\int_\Omega |D^\alpha u|^2\mathrm{d}\Omega$$

Observe that for $r = p = 1$, we have the same order two of central finite difference. The requested regularity for u is much less (for finite differences $u \in \mathcal{C}^4$ is required), but the error estimate is in L^2 norm. Moreover, $F(v)$ (that is the terms $\int g\phi_{\ell_{j,k}}$) is assumed to be computed exactly or with sufficient accuracy. This may be not the case if g is not regular enough and the quadrature formula (1.5) is used (see Figure 1.6).



fem1d.m

fd1d.m

Figure 1.6: Error in the solution of (1.1) by FEM (top) and second order central FD (bottom), with $g(x)$ proportional to $|x - 1/2|^{5/2}$ (left) and to $|x - 1/2|^{1/2}$ (right), respectively.

Mesh generation

Let us consider the two-dimensional domain. Given a set of distinct points, there exists a "unique" *Delaunay triangulation*. It means that the disk circumscribed to each triangle does not properly contain any point. Among all the triangulations, it maximizes the minimum angle of the triangles (it is important for convergence). In fact, the weak assumption for the mesh is

$$\frac{h_{\ell_j}}{\rho_{\ell_j}} \leq \delta, \quad \forall \ell_j$$

where ρ_{ℓ_j} is the diameter of the disk inscribed into the triangle. In this case, the mesh is said *regular*. Moreover, the parameter δ enters in the constant C of the theorem above.

Linear systems

Since the arising matrices are sparse, usually Krylov methods (semi-iterative) are used for the linear systems, like the *conjugate gradient* method (for the symmetric positive definite case): given $x^{(1)}$, $r^{(1)} = b - Ax^{(1)}$, and $p^{(1)} = r^{(1)}$, the algorithm to compute $x^{(l+1)}$ is

$$\begin{aligned}
\alpha_l &= \frac{{r^{(l)}}^{\mathrm{T}} r^{(l)}}{{p^{(l)}}^{\mathrm{T}} A p^{(l)}} \\
x^{(l+1)} &= x^{(l)} + \alpha_l p^{(l)} \\
r^{(l+1)} &= r^{(l)} - \alpha_l A p^{(l)} \\
\beta_{l+1} &= \frac{{r^{(l+1)}}^{\mathrm{T}} r^{(l+1)}}{{r^{(l)}}^{\mathrm{T}} r^{(l)}} \\
p^{(l+1)} &= r^{(l+1)} + \beta_{l+1} p^{(l)}
\end{aligned} \tag{1.6}$$

The exit criterion is based on the norm of $r^{(l+1)}$. We notice that the algorithm does not require to know the matrix A, but just how to perform a matrix-vector product Av. Consider, for instance, the following diffusion-reaction problem

$$-\Delta u + u^2 \text{ on } \Omega$$

with appropriate boundary conditions. The Galerkin method form is

$$F_i(u) = \int_\Omega \nabla u \cdot \nabla \phi_i \mathrm{d}\Omega - \int_\Gamma \frac{\partial u}{\partial n} \phi_i \mathrm{d}\Gamma + \int_\Omega u^2 \phi_i \mathrm{d}\Omega = 0$$

We have to solve $F(u) = 0$ and Newton's method can be applied:

$$J_F(u^{(r)})\delta^{(r+1)} = -F(u^{(r)})$$

In order to solve this linear system with the conjugate gradient method it is enough to be able to compute the i-th component of $J_F(u^{(r)})v$, that is

$$\int_\Omega \nabla v \cdot \nabla \phi_i \mathrm{d}\Omega - \int_\Gamma \frac{\partial v}{\partial n} \phi_i \mathrm{d}\Gamma + \int_\Omega 2u^{(r)} v \phi_i \mathrm{d}\Omega$$

Instead of solving a linear system

$$Ax = b$$

one can try to find a matrix P (symmetric and positive definite) such that

$$P^{-1}Ax = P^{-1}b$$

is "easier" to solve (meaning less iterations). Notice that it is not possible to apply the conjugate gradient method to the problem above, since $P^{-1}A$ is not symmetric. Instead, one has to consider,

$$(R^{-\mathrm{T}}AR^{-1})y = R^{-\mathrm{T}}b, \quad y = Rx$$

where $P = R^{\mathrm{T}}R$. The application (and a rearrangement) of the conjugate gradient method writes

$$\begin{aligned}
&\alpha_l = \frac{{z^{(l)}}^{\mathrm{T}} r^{(l)}}{{p^{(l)}}^{\mathrm{T}} A p^{(l)}}\\
&x^{(l+1)} = x^{(l)} + \alpha_l p^{(l)}\\
&r^{(l+1)} = r^{(l)} - \alpha_l A p^{(l)}\\
&Pz^{(l+1)} = r^{(l+1)}\\
&\beta_{l+1} = \frac{{z^{(l+1)}}^{\mathrm{T}} r^{(l+1)}}{{z^{(l)}}^{\mathrm{T}} r^{(l)}}\\
&p^{(l+1)} = z^{(l+1)} + \beta_{l+1} p^{(l)}
\end{aligned} \tag{1.7}$$

where $Pz^{(1)} = r^{(1)}$. The following estimate for the error

$$\|e^{(l)}\|_A \leq \frac{2c^l}{1+c^{2l}} \|e^{(1)}\|_A$$

holds, where

$$c = \frac{\sqrt{\mathrm{cond}_2(P^{-1}A)} - 1}{\sqrt{\mathrm{cond}_2(P^{-1}A)} + 1}, \quad \|v\|_A = v^{\mathrm{T}}Av$$

It is clear that the more c is small (that is, the more $\mathrm{cond}_2(P^{-1}A)$ is close to 1), the better is. Notice that there is an addition linear system $Pz^{(l+1)} = r^{(l+1)}$ to solve which is in general easy, since $P = R^{\mathrm{T}}R$ and R is triangular. Also in this case, it is not strictly necessary to know the matrix P, but it is enough to know the result of P^{-1} applied to a vector. The choice minimizing the number of iterations is, of course, if $P = A$, but P has also to be "easy to invert". In practice, because of the bad scaling of the matrix due to the penalty method, it is always necessary to use a preconditioned. A simple choice is $P = \mathrm{diag}(A)$ and a more effective choice is the *incomplete* Cholesky factorization with no fill-in of A. That is, $P = \tilde{R}^{\mathrm{T}}\tilde{R} \approx A$ where

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

The incomplete factorization is chosen because for a sparse matrix A, its complete factors are generally not sparse. We report here the algorithm for the general incomplete LU factorization with no fill-in:

```
function [L,U] = luinckij(A)
% Incomplete LU with no fill-in, kij variant, no pivoting
m = length(A);
for k = 1:m-1
  for i = k+1:m
    if (A(i,k) ~= 0)
      A(i,k) = A(i,k) / A(k,k);
      for j = k+1:m
        if (A(i,j) ~= 0)
          A(i,j) = A(i,j) - A(i,k) * A(k,j);
        end
      end
    end
  end
end
U = triu(A);
L = tril(A,-1) + speye(m);
```

If we remove the two `if` clauses, it is a complete LU factorization. Usually, for a row-contiguous data structure (such as C), the *ikj*-variant is used and for column-contiguous data structure (such as Fortran, Matlab, Octave) the *jki*-variant is used, which minimizes memory accesses. The incomplete Cholesky factorization can be obtained from the incomplete LU by scaling the rows of U by the square root of the diagonal

```
R = diag(sqrt(diag(U))) \ U
```

In general, the incomplete factorization with no fill-in is not used: instead, a certain fill-level is allowed, in order to have $\tilde{L}\tilde{U}$ closer to A. From these two examples, it appears clear that the more the *bandwidth* of A is small, the better is. The bandwidth depends only on the *topology* of the mesh.

Since usually the mesh points are given and the triangulation is unique, the bandwidth essentially depends only on the numbering of the points. There are algorithms, such as the *symmetric reverse Cuthill–McKee* one, which heuristically minimize the bandwidth of a matrix.

For non-symmetric problems, the conjugate gradient method has to be extended to the *bi-conjugate gradient method.* Another possible choice is GMRES. As preconditioner, one can consider the *incomplete LU factorization.*

1.4.2 Evolution problems and mass matrix

For the heat equation

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x,y) = \Delta u(t,x,y) & (x,y) \in \Omega \\ u(0,x,y) = u_0(x,y) & (x,y) \in \Omega \\ \textit{+boundary conditions} \end{cases}$$

the procedure is almost the same, considering the approximation

$$u(t,x,y) \approx \sum_{j=1}^{m} u_j(t)\phi_j(x,y)$$

The weak formulation leads to the computation of

$$\int_{\ell_j} \phi_{\ell_{j,k}}(x,y)\phi_{\ell_{j,h}}(x,y)\mathrm{d}x\mathrm{d}y = \begin{cases} \frac{|\Delta_j|}{6} & \text{if } k = h \\ \frac{|\Delta_j|}{12} & \text{if } k \neq h \end{cases}$$

We notice that in the one-dimensional case it is

$$\int_{\ell_j} \phi_{\ell_{j,k}}(x)\phi_{\ell_{j,h}}(x)\mathrm{d}x = \begin{cases} \frac{h_j}{3} & \text{if } k = h \\ \frac{h_j}{6} & \text{if } k \neq h \end{cases}$$

The FEM approximation is then

$$Pu'(t) = -Au(t)$$

where P is the *mass matrix* (symmetric and positive definite). At this point a method for systems of ODEs can be applied, such as the θ-method

$$\begin{aligned} Pu^{n+1} - Pu^n &= -\Delta t(1-\theta)Au^n - \Delta t\theta Au^{n+1} \\ (P + \Delta t\theta A)u^{n+1} &= (P - \Delta t(1-\theta))u^n \end{aligned}$$

Due to the presence of the mass matrix, even explicit methods (such as Euler's) requires the solution of linear systems. This is not a real drawback, since usually for the *stiff* heat equations explicit methods are not a choice.

About exponential methods, they require to compute P^{-1} as well

$$u^{n+1} = \exp(\Delta t P^{-1} A) u^n$$

A solution to avoid the possible expensive computation of $P^{-1}A$ is to replace P with the *mass-lumped* diagonal matrix P_L which contains in the diagonal element of each row the sum of all the elements of the row. Its usage does not compromise the order or the accuracy of the method, since it is the result of

$$\int_{\ell_j} \phi_{\ell_{j,k}}(x,y)\phi_{\ell_{j,h}}(x,y)\mathrm{d}x\mathrm{d}y \approx \begin{cases} \frac{|\Delta_j|}{3} & \text{if } k = h \\ 0 & \text{if } k \neq h \end{cases}$$

whenever the trapezoidal formula is used to approximate the integrals. The trapezoidal formula does not introduce an error greater than the error already done by using linear basis functions.

Bibliography

[1] Y. Saad, Iterative Methods for Sparse Linear Systems, SIAM, 2003, `www-users.cs.umn.edu/~saad/IterMethBook_2ndEd.pdf`

[2] O. Pironneau, F. Hecht, A. Le Hyaric, J. Morice, *FreeFem++*, `http://www.freefem.org`.

[3] A. Quarteroni, Modellistica Numerica per Problemi Differenziali, Springer, 2006.

[4] MIT, Lecture Notes on *Numerical Methods for Partial Differential Equations*, MITOPENCOURSEWARE, 2003.