Norwegian University of Science
and Technology
Department of Mathematical
Sciences

TMA4220 Numerical Solution of Partial Differential Equations Using Element Methods Fall 2023

Solutions to exercise set 2

1 **a)** We first write the bilinear form $a(u,v)$, which in our case is

$$a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x) \mathrm{d}x + c \int_\Omega u(x)v(x)\mathrm{d}x.$$

This is a coercive form, and in particular we have

$$a(u,u) \geq \min\{1,c\}\|u\|^2_{H^1(\Omega)} = \min\{1,c\}\left(|u|^2_{H^1(\Omega)} + \|u\|^2_{L^2(\Omega)}\right).$$

Applying Galerikin orthogonality we can conclude

$$\begin{aligned}\min\{1,c\}\|u-u_h\|^2_{H^1(\Omega)} &\leq a(u-u_h, u-u_h) = a(u-u_h, u-v_h)\\ &\leq \max\{1,c\}\|u-u_h\|_{H^1(\Omega)}\|u-v_h\|_{H^1(\Omega)}\end{aligned}$$

for a generic $v_h \in V_h$, and hence

$$\|u-u_h\|_{H^1(\Omega)} \leq \frac{\max\{1,c\}}{\min\{1,c\}}\|u-v_h\|_{H^1(\Omega)},$$

$$\|u-u_h\|_{H^1(\Omega)} \leq \frac{\max\{1,c\}}{\min\{1,c\}} \inf_{v_h \in V_h}\|u-v_h\|_{H^1(\Omega)}.$$

We remark that $u - v_h \in H^1_0(\Omega)$ and hence

$$\|u-v_h\|_{L^2(\Omega)} \leq C_\Omega \|\nabla u - \nabla v_h\|_{L^2(\Omega)},$$

which implies

$$\|u-u_h\|_{H^1(\Omega)} \leq \tilde{c}h^{s-1}\|u\|_{H^s(\Omega)}.$$

b) We now define the dual problem

$$\begin{cases} -\Delta\psi + c\psi = \varphi, & \Omega \\ \psi = 0, & \partial\Omega. \end{cases}$$

Let then $\psi_h \in V_h$ be a generic function, and consider

$$\begin{aligned}\int_\Omega (u(x)-u_h(x))\varphi(x)\mathrm{d}x &= \int_\Omega (u(x)-u_h(x))\left(-\Delta\psi(x) + c\psi(x)\right)\mathrm{d}x && (1)\\ &= \int_\Omega \nabla(u_h(x)-u(x))\cdot\nabla\psi(x)\mathrm{d}x && (2)\\ &\quad + c\int_\Omega (u_h(x)-u(x))\psi(x)\mathrm{d}x && (3)\\ &= a(u_h-u,\psi) = a(u_h-u,\psi-\psi_h) && (4)\\ &\leq \max\{1,c\}\|u_h-u\|_{H^1(\Omega)}\|\psi-\psi_h\|_{H^1(\Omega)} && (5)\\ &\leq \max\{1,c\}\tilde{c}h^{s-1}\|u\|_{H^s(\Omega)}\|\psi-\psi_h\|_{H^1(\Omega)} && (6)\end{aligned}$$

and by arbitrarity of ψ_h we have

$$\begin{aligned}\int_\Omega (u(x)-u_h(x))\varphi(x)\mathrm{d}x &\le \max\{1,c\}\tilde{c}h^{s-1}\|u\|_{H^s(\Omega)} \inf_{\psi_h\in V_h}\|\psi-\psi_h\|_{H^1(\Omega)},\\ &\le \max\{1,c\}\tilde{c}h^{s-1}\|u\|_{H^s(\Omega)}h\|\varphi\|_{L^2(\Omega)}\end{aligned}$$

where the last inequality comes from the elliptic Regularity assumption, and also the bound due to the regularity of the mesh and taking $s=2$. This allows to conclude the proof.

2 **a)** We remark that a triangle is the convex hull of its vertices, i.e. the triangle T is defined as

$$T=\left\{p\in\mathbb{R}^2:\ p=\sum_{i=1}^3\xi_i p_i,\ 0\le\xi_i\le 1,\ \sum_{i=1}^3\xi_i=1\right\}=:\mathrm{conv}\{\mathrm{p}_1,\mathrm{p}_2,\mathrm{p}_3\}.$$

More precisely the interior $\mathring{T}$ is made by the points for which all the ξ_is are strictly positive. Otherwise if, let's say, $\xi_1=0$, then we have that p is in the convex hull of p_2 and p_3, i.e. on the segment connecting them which is an edge of T.

b) There are various ways to prove this result. Let us take $p\in\partial T$, and without loss of generality suppose p is on the edge connecting p_1 with p_2. Then it means we can express it as $\lambda p_1+(1-\lambda)p_2=p$ for some $\lambda\in[0,1]$. We can then conclude that ξ_3 has to be 0, and the same holds if we focus on the two other edges.

c) With the choice $\xi_3=1-\xi_1-\xi_2$ we can write

$$p=\xi_1p_1+\xi_2p_2+(1-\xi_1-\xi_2)p_3=p_3+\xi_1(p_1-p_3)+\xi_2(p_2-p_3)$$

which hence tells us that (ξ_1,ξ_2) are the coordinates of p in the reference frame $\{p_3,p_1-p_3,p_2-p_3\}$.

d) Let us start rewriting the generic form of a quadrature rule:

$$\int\int_{T_r} f(x,y)\mathrm{d}x\mathrm{d}y\approx\sum_{i=1}^p\omega_i f(x_i,y_i).$$

Define then the triangle $T=\mathrm{conv}\{\mathrm{p}_1,\mathrm{p}_2,\mathrm{p}_3\}$ made of points of the form $p=p_3+\xi_1(p_1-p_3)+\xi_2(p_2-p_3)=:\Phi(\xi_1,\xi_2)$. We notice that $p\in T$ if and only if

$$0\le\xi_1,\xi_2\le 1,\ 0\le 1-\xi_1-\xi_2\le 1.$$

This is equivalent to say that for points in T, the barycentric coordinates satisfy

$$0\le\xi_1\le 1,\ 0\le\xi_2\le 1-\xi_1.$$

We remark that

$$\{(\xi_1,\xi_2)\in\mathbb{R}^2:\ \xi_1\in[0,1],\ \xi_2\in[0,1-\xi_1]\}=T_r$$

is the reference triangle. Thus, the map Φ defines a map from T_r to T, and by computing its Jacobian we can apply the change of variables to compute the integral over T of a generic function f:

$$\int\int_T f(x,y)\mathrm{d}x\mathrm{d}y = \det(\mathrm{J}\Phi)\int\int_{\mathrm{T_r}} \mathrm{f}(\Phi(\xi_1,\xi_2))\mathrm{d}\xi_1\mathrm{d}\xi_2.$$

The Jacobian $J\Phi$ in our case is a the constant matrix

$$J\Phi = \begin{bmatrix} p_1 - p_3 & p_2 - p_3 \end{bmatrix}$$

and its determinant corresponds to the area of the parallelogram having $p_1 - p_3$ and $p_2 - p_3$ as edges. Equivalently, $\det(\mathrm{J}\Phi) = 2\mathrm{Area(T)}$. We conclude

$$\begin{aligned}\int\int_T f(x,y)\mathrm{d}x\mathrm{d}y &= 2\mathrm{Area(T)}\int\int_{\mathrm{T_r}} \mathrm{f}(\Phi(\xi_1,\xi_2))\mathrm{d}\xi_1\mathrm{d}\xi_2 \\ &\approx 2\mathrm{Area(T)}\sum_{\mathrm{i}=1}^{\mathrm{p}} \omega_\mathrm{i}\mathrm{f}(\Phi(\xi_{1\mathrm{i}},\xi_{2\mathrm{i}})),\end{aligned}$$

where $\{(\xi_{11},\xi_{12}),\ldots,(\xi_{1p},\xi_{2p})\}$ are the collocation nodes of the quadrature formula.

e) We start by writing the basis functions for the triangle $T = \mathrm{conv}\{\mathrm{p_1},\mathrm{p_2},\mathrm{p_3}\}$. To do this, notice we need to invert a matrix since

$$\psi_i(x,y) = a_{ix}x + a_{iy}y + a_i = \left\langle \begin{bmatrix} a_{ix} \\ a_{iy} \\ a_i \end{bmatrix}, \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \right\rangle,\ i = 1,2,3$$

and we want to impose

$$\psi_i(p_j) = \begin{cases} 1,\ i = j \\ 0,\ i \neq j. \end{cases}$$

This can be re-expressed in the more convenient form

$$\begin{bmatrix} a_{1x} & a_{1y} & a_1 \\ a_{2x} & a_{2x} & a_2 \\ a_{3x} & a_{3x} & a_3 \end{bmatrix} \begin{bmatrix} p_1 & p_2 & p_3 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This means that

$$\begin{bmatrix} a_{1x} & a_{1y} & a_1 \\ a_{2x} & a_{2x} & a_2 \\ a_{3x} & a_{3x} & a_3 \end{bmatrix} = \begin{bmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ 1 & 1 & 1 \end{bmatrix}^{-1}.$$

One can express such inverse in various ways, and in numerical implementations often times these coefficients are found by solving linear systems for each choice of points p_1, p_2, p_3.

For our calculation we do not really need the explicit expression of $\psi_i(x,y)$, and this is one of the advantages of using local approaches in the finite element method, and building everything on the reference triangle T_r. Indeed $\psi_i(\Phi(\xi_1,\xi_2)) = \varphi_i(\xi_1,\xi_2)$, i.e. we can relate these basis functions to those on

the reference triangle T_r, of which we know the simple analytical expression. We thus can compute

$$\begin{aligned}
&\int\int_T \psi_1(x,y)\psi_2(x,y)\mathrm{d}x\mathrm{d}y = 2\mathrm{Area(T)}\int\int_{\mathrm{T_r}} \varphi_1(\xi_1,\xi_2)\varphi_2(\xi_1,\xi_2)\mathrm{d}\xi_1\mathrm{d}\xi_2 \\
&= \frac{\mathrm{Area(T)}}{3}\left(\varphi_1(1/2,0)\varphi_2(1/2,0) + \varphi_1(1/2,1/2)\varphi_2(1/2,1/2) + \varphi_1(0,1/2)\varphi_2(0,1/2)\right) \\
&= \frac{1}{4}\frac{\mathrm{Area(T)}}{3} = \frac{\mathrm{Area(T)}}{12} = \frac{1}{24}.
\end{aligned}$$

While the second integral writes

$$\begin{aligned}
&\int\int_T \psi_1(x,y)\sin\left(2\pi(x+y)\right)\mathrm{d}x\mathrm{d}y \\
&= 2\mathrm{Area(T)}\int\int_{\mathrm{T_r}} \varphi_1(\xi_1,\xi_2)\sin\left(2\pi(\mathrm{x}(\xi_1,\xi_2)+\mathrm{y}(\xi_1,\xi_2))\right)\mathrm{d}\xi_1\mathrm{d}\xi_2.
\end{aligned}$$

In the expression above, we have used the notation

$$\begin{bmatrix} x(\xi_1,\xi_2) \\ y(\xi_1,\xi_2)\end{bmatrix} = p_3 + \xi_1(p_1-p_3) + \xi_2(p_2-p_3) = \begin{bmatrix} p_3^x + \xi_1(p_1^x-p_3^x) + \xi_2(p_2^x-p_3^x) \\ p_3^y + \xi_1(p_1^y-p_3^y) + \xi_2(p_2^y-p_3^y)\end{bmatrix},$$

and in a more compact way one can get that given $\mathbf{1} = [1,1] \in \mathbb{R}^2$:

$$x(\xi_1,\xi_2) + y(\xi_1,\xi_2) = \mathbf{1}^T p_3 + \xi_1\mathbf{1}^T(p_1-p_3) + \xi_2\mathbf{1}^T(p_2-p_3).$$

We now compute some useful values for our integration:

$$\begin{aligned}
x(1/2,1/2)+y(1/2,1/2) &= \mathbf{1}^T p_3 + \mathbf{1}^T(p_1-p_3)/2 + \mathbf{1}^T(p_2-p_3)/2 \\
&= 1 - 1/2 + 1/2 = 1 \\
x(1/2,0)+y(1/2,0) &= \mathbf{1}^T p_3 + \mathbf{1}^T(p_1-p_3)/2 = \\
&= 1 - 1/2 = 1/2 \\
x(0,1/2)+y(0,1/2) &= \mathbf{1}^T p_3 + \mathbf{1}^T(p_2-p_3)/2 = \\
&= 1 + 1/2 = 3/2 \\
\varphi_1(1/2,1/2) &= 0,\ \varphi_1(1/2,0) = 1/2,\ \varphi_1(0,1/2) = 1/2.
\end{aligned}$$

We conclude that the desired integral is approximated by 0 by the 3-point quadrature rule.

3 **a)** For the quartic Lagrange finite element we take a polynomial $p \in \mathcal{P}^4(T)$ and suppose $N_i(p) = 0$ for every $i = 1,\dots,15$. This means in particular that p is identically 0 on the three edges, and hence called L_1, L_2, L_3 the functions defining those segments, we can decompose $p(x)$ as $p(x) = q(x)L_1(x)L_2(x)L_3(x)$ for a polynomial $q \in \mathcal{P}^1(T)$. Let now L_4 be the function defining the line where two of the three internal points live. Since p has to vanish on those two points, it also follows $p(x) = cL_1(x)L_2(x)L_3(x)L_4(x)$ for $c \in \mathbb{R}$. This c has to be 0 since p vanishes also on the remaining point, which allows to conclude $p(x) \equiv 0$. In other words, we just proved that $\mathcal{N}_4$ determines $\mathcal{P}^4(T)$.

b) We now consider the Argyris finite element of degree 5. Consider a polynomial $p \in \mathcal{P}^5(T)$, and suppose $N_i(p) = 0$ for $i = 1, \dots, 21$. Let us focus on the edge connecting z_2 with z_3, i.e. the one determined by the function L_1 as $L_1(x) = 0$. Here, p is a polynomial of degree 5 in one variable and it vanishes at z_2 and z_3. Because of the additional degrees of freedom at z_2 and z_3, we also know the first and second derivatives of p on those vertices are zero. This means p has two triple roots at z_2 and z_3 Being p a polynomial of degree 5, this is possible only if $p|_{L_1} \equiv 0$. The same argument can be made for L_2 and L_3, which allows us to write $p(x) = L_1(x)L_2(x)L_3(x)q(x)$ where $q \in \mathcal{P}^2(T)$. We know that $\nabla p(z_1) = \nabla p(z_2) = \nabla p(z_3) = 0$ and hence all the directional derivatives are also zero on those vertices. Similarly, also the second directional derivatives are zero there. Let $\partial_{L_i} p$ be the directional derivative of p along the $i-$th edge of the triangle T. Call $\partial_{L_1} p(x) = \nabla p(x) \cdot (z_3 - z_2)$, i.e. the directional derivative of p along the edge defined by L_1, and do similarly for $\partial_{L_2} p$. This definition gives $\partial_{L_1} L_1 \equiv 0$. Furthermore, since $L_1(z_3) = L_2(z_3) = 0$, one gets

$$0 = \partial_{L_2}\partial_{L_1} p(z_3) = Q(z_3)L_3(z_3)\partial_{L_1} L_2 \partial_{L_2} L_1.$$

Since $L_3(z_3)\partial_{L_1} L_2 \partial_{L_2} L_1 \neq 0$, it follows $Q(z_3) = 0$. The same holds for $Q(z_1) = Q(z_2) = 0$, and one can get these results by similar arguments. To conclude, we call $\partial_{n_1} p(x) = \nabla p(x) \cdot n_1$ the normal derivative in the direction orthogonal to $L_1(x) = 0$, and do the same for the other 2 edges.

We compute

$$0 = (\partial_{n_1} p)(m_1) = Q(m_1)(\partial_{n_1} L_1)(m_1)L_2(m_1)L_3(m_1).$$

Since m_1 lives on the edge defined by $L_1(x) = 0$, we have $L_2(m_1)L_3(m_1) \neq 0$. Furthermore, $(\partial_{n_1} L_1)(m_1) \neq 0$ since this is the derivative normal to the subspace $L_1(x) = 0$. This, and the symmetric arguments, allow to get $Q(m_1) = Q(m_2) = Q(m_3) = 0$. We thus know Q vanishes at the 6 points of the quadratic Lagrangian triangular element, which means that $Q \equiv 0$ and hence $p \equiv 0$.

4 Let $e(x) = u(x) - u_h(x)$, $v \in H^1(\Omega)$ and $v_h \in V_h$.

a) The residual equation writes:

$$\begin{aligned} a(e, v) &= a(u, v) - a(u_h, v) = \ell(v) - a(u_h, v) \\ &= \int_\Omega f(x)v(x)\mathrm{d}x + \int_{\partial\Omega} g(x)v(x)\mathrm{d}s - \int_\Omega \nabla u_h(x) \cdot \nabla v(x)\mathrm{d}x - \int_\Omega u_h(x)v(x)\mathrm{d}x. \end{aligned}$$

b) By Galerkin orthogonality we simply have $a(e, v_h) = 0$.

c) We now expand $a(e,v)$ onto its contributions on the elements K_i:

$$\begin{aligned}
a(e,v) =& \sum_{i=1}^{N} \left(\int_{K_i} \left[f(x) - u_h(x) \right] v(x) \mathrm{d}x - \int_{K_i} \nabla u_h(x) \cdot \nabla v(x) \mathrm{d}x + \int_{\partial K_i \cap \partial \Omega} g(x) v(x) \mathrm{d}s \right) \\
=& \sum_{i=1}^{N} \left(\int_{K_i} \left[f(x) - u_h(x) \right] v(x) \mathrm{d}x + \int_{K_i} \Delta u_h(x) v(x) \mathrm{d}x \right. \\
& \left. - \int_{\partial K_i} \partial_{n_i} u_h(x) v(x) \mathrm{d}s + \int_{\partial K_i \cap \partial \Omega} g(x) v(x) \mathrm{d}s \right) \\
=& \sum_{i=1}^{N} \left(\int_{K_i} \left[f(x) - u_h(x) + \Delta u_h(x) \right] v(x) \mathrm{d}x + \right. \\
& \left. - \int_{\partial K_i \setminus \partial \Omega} \partial_{n_i} u_h(x) v(x) \mathrm{d}s + \int_{\partial K_i \cap \partial \Omega} \left[g(x) - \partial_{n_i} u_h(x) \right] v(x) \mathrm{d}s \right) \\
=& \sum_{i=1}^{N} \left(\int_{K_i} r(x) v(x) \mathrm{d}x + \int_{\partial K_i} R(x) v(x) \mathrm{d}s \right)
\end{aligned}$$

where

$$\begin{aligned}
r(x) &= f(x) - u_h(x) + \Delta u_h(x) \\
R(x) &= \begin{cases} -\partial_{n_i} u_h(x), & x \in \partial K_i \setminus \partial \Omega \\ g(x) - \partial_{n_i} u_h(x), & x \in \partial K_i \cap \partial \Omega. \end{cases}
\end{aligned}$$

d) We now use the interpolation operator $I_h : H^1(\Omega) \to V_h$ and by Galerkin orthogonality we get

$$\begin{aligned}
a(e,v) &= a(e,v) - a(e, I_{V_h} v) = a(e, v - I_{V_h} v) \\
&= \sum_{i=1}^{N} \left(\int_{K_i} r(x)(v - I_{V_h} v)(x) \mathrm{d}x + \int_{\partial K_i} R(x)(v - I_{V_h} v)(x) \mathrm{d}s \right).
\end{aligned}$$

Applying Cauchy-Schwartz one gets

$$a(e,v) \leq \sum_{i=1}^{N} \left(\|r\|_{L^2(K_i)} \|v - I_{V_h} v\|_{L^2(K_i)} + \|R\|_{L^2(\partial K_i)} \|v - I_{V_h} v\|_{L^2(\partial K_i)} \right).$$

We now call $\tilde{K}_i$ the patch of K_i, and hence we can use the estimates

$$\begin{aligned}
\|v - I_{V_h} v\|_{L^2(K_i)} &\leq C h_i \|v\|_{H^1(\tilde{K}_i)} \\
\|v - I_{V_h} v\|_{L^2(\partial K_i)} &\leq C h_i^{1/2} \|v\|_{H^1(\tilde{K}_i)},
\end{aligned}$$

which allows us to say

$$\begin{aligned}
a(e,v) &\leq C \sum_{i=1}^{N} \left(\|r\|_{L^2(K_i)} h_i \|v\|_{H^1(\tilde{K}_i)} + \|R\|_{L^2(\partial K_i)} h_i^{1/2} \|v\|_{H^1(\tilde{K}_i)} \right) \\
&= C u^T v
\end{aligned}$$

where $u, v \in \mathbb{R}^N$ are vectors with with entries

$$u_i = \|r\|_{L^2(K_i)} h_i + \|R\|_{L^2(\partial K_i)} h_i^{1/2}, \; v_i = \|v\|_{H^1(\tilde{K}_i)}.$$

By Cauchy-Schwartz we hence have

$$\begin{aligned} a(e,v) &\leq C\|u\|_2\|v\|_2 \\ &\leq \tilde{C}\left(\sum_{i=1}^{N}\|r\|^2_{L^2(K_i)}h_i^2 + \|R\|^2_{L^2(\partial K_i)}h_i\right)^{1/2}\left(\sum_{i=1}^{N}\|v\|^2_{H^1(\tilde{K}_i)}\right)^{1/2}. \end{aligned}$$

For the first term we used the relation

$$(a+b)^2 \leq 2(a^2+b^2),$$

and hence $\tilde{C} = \sqrt{2}C$. By mesh regularity, the number of elements in a patch can be bounded by an n independent on N. Furthermore, saying that in a patch there are at most n elements, implies that every element appears in at most n patches. This allows us to obtain

$$a(e,v) \leq \sqrt{n}\tilde{C}\|v\|_{H^1(\Omega)}\left\{\sum_{i=1}^{N}\left(\|r\|^2_{L^2(K_i)}h_i^2 + h_i\|R\|^2_{L^2(\partial K_i)}\right)\right\}^{1/2},$$

since $\|\cdot\|_{H^1(K_i)}$ appears at most n times in the sum. Setting $v = e$, by coercivity we get

$$\|e\|^2_{H^1(\Omega)} \leq \sqrt{n}\tilde{C}\sum_{i=1}^{N}\left(\|r\|^2_{L^2(K_i)}h_i^2 + h_i\|R\|^2_{L^2(\partial K_i)}\right).$$

e) The local estimator η_{K_i} in our case writes

$$\eta^2_{K_i} := \|r\|^2_{L^2(K_i)}h_i^2 + h_i\|R\|^2_{L^2(\partial K_i)}.$$

f) We define

$$\eta = \left(\sum_{i=1}^{N}\eta^2_{K_i}\right)^{1/2}.$$

This estimator is sharp and reliable if there exist constants $C_1, C_2 > 0$ such that

$$C_1\|e\|_{H^1(\Omega)} \leq \eta \leq C_2\|e\|_{H^1(\Omega)}. \tag{7}$$

This is important because it guarantees that the η term tends to 0 at the same rate as the true error. To verify (7) holds, we can design a problem with known analytical solution u, and proceed by refining the mesh, i.e. distributing properly the errors through the elements, and check if the rate at which $\|u - u_h\|_{H^1(\Omega)}$ goes to 0 is the same of the one at which η goes to 0.