

1. **(Convergence of Linear FEM)** Consider the two point BVP that we studied in class, i.e.,

Given  $f \in C^0[0, 1]$ , find  $u \in C^2[0, 1]$  such

$$(D) : \begin{cases} -\frac{d^2u}{dx^2} = f; & 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases}$$

where  $f$  is chosen so that the exact solution is  $u(x) = \sin(\pi x)$ .

Use the linear FEM code that you have written with  $h = 2^{-j}$ ,  $j = 1, \dots, 10$  to find the rate of convergence of the following errors (use numerical integration):

- (a)  $\|u - u_k\|_{L^2(0,1)}$
- (b)  $\|u - u_k\|_E$
- (c)  $\max_i |u(x_i) - u_k(x_i)|$

where  $u_k \in V_k^1$  is the Galerkin solution with continuous piecewise linear finite elements and the  $x_i$  are the nodes of a partition of  $[0, 1]$ . For each case, plot the error as a function of the mesh step size  $h$  (loglog plot). Confirm the theoretical results discussed in class.

## Solution

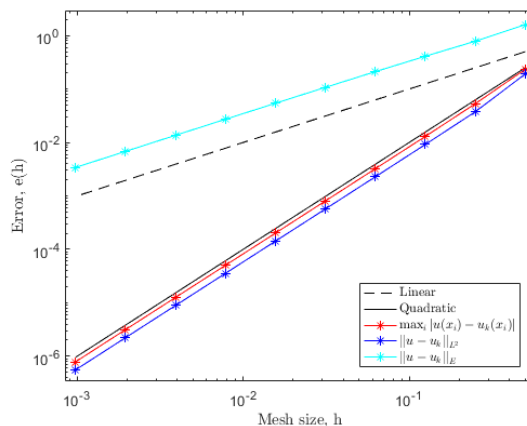


Figure 1: The FEM max and  $L_2$  error exhibits second-order behavior while the energy norm error is first-order.

The  $L_2$  and energy norms were computed using Simpson's 1/3 rule. Additionally, forward differences were used to calculate the derivative of  $u - u_k$

for the energy norm. The convergence plots of the max,  $L_2$ , and energy errors for ten different mesh sizes are plotted in Figure 1. This plot clearly shows that the  $L_2$  error is second-order while the energy error is instead first-order. This matches with the theoretical error estimates for linear finite elements.

## 2. (Relation between Finite Difference and Finite Element Methods

Consider problem 2 from your HW1 (discretization of 1D BVP) with (1) linear finite elements and (2) quadratic finite elements. In each case, assume  $f = 1$  and use a uniform (regular) mesh. On a regular mesh, the finite element approximation to the 1D BVP has the following feature: To each degree of freedom there corresponds a discrete equation. Do the following.

- For the case of linear finite elements, show that the  $i$ th row of the linear system  $AX = F$  is a centered finite difference discretization of the 1D BVP at the  $i$ th node of the mesh.
- For the case of quadratic finite elements show that the rows of the corresponding linear system can be sorted into two types of equations. Show that we obtain a discrete equation corresponding to an endpoint  $x_i$  of subinterval  $I_i$  in the form

$$\frac{1}{h^2} (14u_i(t) - 8(u_{i-1/2}(t) + u_{i+1/2}(t)) + u_{i+1}(t) + u_{i-1}(t)) = 1,$$

and another discrete equation corresponding to the midpoint  $x_{i+1/2}$  of the interval  $[x_i, x_{i+1}]$  given by

$$-\frac{4}{h^2} (u_i(t) - 2u_{i+1/2}(t) + u_{i+1}(t)) = 1,$$

where the  $u_i$ s are the degrees of freedom of the finite element solution.

## Solution

- The centered difference formula for

$$-\frac{d^2 u}{dx^2} = f$$

is given by

$$\frac{-u_{i-1}(t) + 2u_i(t) - u_{i+1}(t)}{h^2} = 1$$

The global stiffness matrix, load vector, and solution vector are given as follows for uniform mesh size,  $h$

$$A = \begin{bmatrix} \frac{2}{h} & \frac{-1}{h} & 0 & 0 & \dots & 0 & 0 \\ \frac{-1}{h} & \frac{2}{h} & \frac{-1}{h} & 0 & \dots & 0 & 0 \\ 0 & \frac{-1}{h} & \frac{2}{h} & \frac{-1}{h} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & \frac{-1}{h} & \frac{2}{h} \end{bmatrix}, \quad F = \begin{bmatrix} h \\ \dots \\ h \end{bmatrix}$$

$$X = \begin{bmatrix} u_0(t) \\ u_1(t) \\ \dots \\ u_i(t) \\ \dots \\ u_{k+1}(t) \end{bmatrix}$$

Thus, one row of this linear system may be written as

$$\begin{aligned} \frac{-u_{i-1}(t) + 2u_i(t) - u_{i+1}(t)}{h} &= h \\ \Rightarrow \frac{-u_{i-1}(t) + 2u_i(t) - u_{i+1}(t)}{h^2} &= 1 \end{aligned}$$

which is precisely the centered finite difference formula.

(b) I incorrectly stated the global stiffness matrix for quadratic finite elements in Homework 1. I restate it here along with the global load and solution vectors as

$$A = \begin{bmatrix} A_{22}^1 & A_{23}^1 & 0 & 0 & 0 & 0 & \dots \\ A_{32}^1 & A_{33}^1 + A_{11}^2 & A_{12}^2 & A_{13}^2 & 0 & 0 & \dots \\ 0 & A_{21}^2 & A_{22}^2 & A_{23}^2 & 0 & 0 & \dots \\ 0 & A_{31}^2 & A_{32}^2 & A_{33}^2 + A_{11}^3 & A_{12}^3 & A_{13}^3 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}, F = \begin{bmatrix} F_2^1 \\ F_3^1 + F_1^2 \\ F_2^2 \\ F_3^2 + F_1^3 \\ \dots \end{bmatrix}$$

$$X = \begin{bmatrix} u_0(t) \\ u_{\frac{1}{2}}(t) \\ u_1(t) \\ \dots \\ u_{i-\frac{1}{2}}(t) \\ u_i(t) \\ u_{i+\frac{1}{2}}(t) \\ \dots \\ u_{k+1}(t) \end{bmatrix}$$

where the local stiffness matrix and load vector are given as

$$A^l = \frac{1}{3h} \begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix}, F^l = \begin{bmatrix} \frac{h}{6} \\ \frac{2h}{3} \\ \frac{h}{6} \end{bmatrix}$$

Thus, two types of equations can be derived from the above system of equations, the first being

$$\begin{aligned}
A_{31}u_{i-1}(t) + A_{32}u_{i-\frac{1}{2}}(t) + (A_{33} + A_{11})u_i(t) + A_{12}u_{i+\frac{1}{2}}(t) + A_{13}u_{i+1}(t) &= F_3 + F_1 \\
\Rightarrow \frac{1}{3h} \left[ u_{i-1}(t) - 8u_{i-\frac{1}{2}}(t) + (7+7)u_i(t) - 8u_{i+\frac{1}{2}}(t) + u_{i+1}(t) \right] &= \frac{h}{3} \\
\Rightarrow \frac{1}{h^2} \left[ u_{i-1}(t) - 8u_{i-\frac{1}{2}}(t) + (7+7)u_i(t) - 8u_{i+\frac{1}{2}}(t) + u_{i+1}(t) \right] &= 1
\end{aligned}$$

the second being

$$\begin{aligned}
A_{21}u_i(t) + A_{22}u_{i+\frac{1}{2}} + A_{23}u_{i+1} &= F_2 \\
\Rightarrow \frac{1}{3h} \left[ -8u_i(t) + 16u_{i+\frac{1}{2}}(t) - 8u_{i+1}(t) \right] &= \frac{2h}{3} \\
\Rightarrow \frac{-4}{h} \left[ u_i(t) - 2u_{i+\frac{1}{2}}(t) + u_{i+1}(t) \right] &= 1
\end{aligned}$$

3. **(Static Condensation)** Consider problem 2 from your HW1 (discretization of 1D BVP with quadratic finite elements). Assume  $f = 1$  and consider a partition of  $[0, 1]$  into five elements ( $k = 4$  interior nodes).

- (a) Split the Galerkin system into two sets of equations

$$a(u_k, \phi_i) = \ell(\phi_i); 1 \leq i \leq 4 \quad (1)$$

$$a(u_k, \phi_{i-1/2}) = \ell(\phi_{i-1/2}); 1 \leq i \leq 5 \quad (2)$$

with  $u_k$ , the Galerkin solution, and where (1) are the equations corresponding to basis functions defined at interior nodes and (2) are the equations corresponding to basis functions defined at the midpoints of elements.

- (b) Write the system (1)-(2) in matrix/vector notation as

$$M_{11}\mathbf{u} + M_{12}\tilde{\mathbf{u}} = \mathbf{b}_1 \quad (3)$$

$$M_{21}\mathbf{u} + D_{22}\tilde{\mathbf{u}} = \mathbf{b}_2, \quad (4)$$

where  $\mathbf{u} = (u_1, u_2, u_3, u_4)^T$  is the vector of values of the Galerkin solution at the interior nodes, and  $\tilde{\mathbf{u}} = (u_{1/2}, u_{3/2}, u_{5/2}, u_{7/2}, u_{9/2})^T$  is the vector of values of the Galerkin solution at the midpoints. Show that the matrix  $M_{11} = (a(\phi_j, \phi_i))_{4 \times 4}$  is a tridiagonal matrix and the matrix  $D_{22} = (a(\phi_{j-1/2}, \phi_{i-1/2}))_{5 \times 5}$  is a diagonal matrix. Compute the matrix  $M_{12} = (a(\phi_j, \phi_i))_{4 \times 5}$  and show that  $M_{21} = M_{12}^T$ .

- (c) Show that we can eliminate the vector  $\tilde{\mathbf{u}}$  from the system (3)-(4) to obtain a system of the form

$$M\mathbf{u} = \mathbf{b}$$

with  $M$  a tridiagonal matrix. What is the relation of the matrix  $M$  and the vector  $\mathbf{b}$  in terms of the matrices and vectors in system (3)-(4)?

- (d) Solve the system  $M\mathbf{u} = \mathbf{b}$ . Also, solve the original Galerkin system and comment on the two solutions.
- (e) (optional) Show that the matrix  $M$  is positive definite.

## Solution

(a) For each  $i$  in  $a(u_k, \phi_i) = a(u_j \phi_j, \phi_i)$ , the corresponding equation is given by

$$a(u_j \phi_j, \phi_i) = \sum_{j=1}^4 u_j(\phi_j, \phi_i) + \sum_{j=1}^5 u_{j-\frac{1}{2}}(\phi_{j-\frac{1}{2}}, \phi_i) = l(\phi_i)$$

For each  $i$  in  $a(u_k, \phi_{i-\frac{1}{2}}) = a(u_j \phi_j, \phi_{i-\frac{1}{2}})$ , the corresponding equation is given by

$$a(u_j \phi_j, \phi_{i-\frac{1}{2}}) = \sum_{j=1}^4 u_j(\phi_j, \phi_{i-\frac{1}{2}}) + \sum_{j=1}^5 u_{j-\frac{1}{2}}(\phi_{j-\frac{1}{2}}, \phi_{i-\frac{1}{2}}) = l(\phi_{i-\frac{1}{2}})$$

(b) The matrices  $M_{11}$ ,  $M_{12}$ ,  $M_{21}$ , and  $D_{22}$  are found to be

$$M_{11} = \begin{bmatrix} (\phi_1, \phi_1) & (\phi_1, \phi_2) & 0 & 0 \\ (\phi_2, \phi_1) & (\phi_2, \phi_2) & (\phi_2, \phi_3) & 0 \\ 0 & (\phi_3, \phi_2) & (\phi_3, \phi_3) & (\phi_3, \phi_4) \\ 0 & 0 & (\phi_4, \phi_3) & (\phi_4, \phi_4) \end{bmatrix} = \frac{1}{3h} \begin{bmatrix} 14 & 1 & 0 & 0 \\ 1 & 14 & 1 & 0 \\ 0 & 1 & 14 & 1 \\ 0 & 0 & 1 & 14 \end{bmatrix}$$

$$M_{12} = \begin{bmatrix} (\phi_1, \phi_{\frac{1}{2}}) & (\phi_1, \phi_{\frac{3}{2}}) & 0 & 0 & 0 \\ 0 & (\phi_2, \phi_{\frac{3}{2}}) & (\phi_2, \phi_{\frac{5}{2}}) & 0 & 0 \\ 0 & 0 & (\phi_3, \phi_{\frac{5}{2}}) & (\phi_3, \phi_{\frac{7}{2}}) & 0 \\ 0 & 0 & 0 & (\phi_4, \phi_{\frac{7}{2}}) & (\phi_4, \phi_{\frac{9}{2}}) \end{bmatrix} = \frac{1}{3h} \begin{bmatrix} -8 & -8 & 0 & 0 & 0 \\ 0 & -8 & -8 & 0 & 0 \\ 0 & 0 & -8 & -8 & 0 \\ 0 & 0 & 0 & -8 & -8 \end{bmatrix}$$

$$M_{21} = \begin{bmatrix} (\phi_{\frac{1}{2}}, \phi_1) & 0 & 0 & 0 \\ (\phi_{\frac{3}{2}}, \phi_1) & (\phi_{\frac{3}{2}}, \phi_2) & 0 & 0 \\ 0 & (\phi_{\frac{5}{2}}, \phi_2) & (\phi_{\frac{5}{2}}, \phi_3) & 0 \\ 0 & 0 & (\phi_{\frac{7}{2}}, \phi_3) & (\phi_{\frac{7}{2}}, \phi_4) \\ 0 & 0 & 0 & (\phi_{\frac{9}{2}}, \phi_4) \end{bmatrix} = \frac{1}{3h} \begin{bmatrix} -8 & 0 & 0 & 0 \\ -8 & -8 & 0 & 0 \\ 0 & -8 & -8 & 0 \\ 0 & 0 & -8 & -8 \\ 0 & 0 & 0 & -8 \end{bmatrix}$$

$$D_{22} = \begin{bmatrix} (\phi_{\frac{1}{2}}, \phi_{\frac{1}{2}}) & 0 & 0 & 0 & 0 \\ 0 & (\phi_{\frac{3}{2}}, \phi_{\frac{3}{2}}) & 0 & 0 & 0 \\ 0 & 0 & (\phi_{\frac{5}{2}}, \phi_{\frac{5}{2}}) & 0 & 0 \\ 0 & 0 & 0 & (\phi_{\frac{7}{2}}, \phi_{\frac{7}{2}}) & 0 \\ 0 & 0 & 0 & 0 & (\phi_{\frac{9}{2}}, \phi_{\frac{9}{2}}) \end{bmatrix} = \frac{1}{3h} \begin{bmatrix} 16 & 0 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 \\ 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

From this we can easily see that  $M_{11}$  is tridiagonal,  $M_{21} = M_{12}^T$ , and  $D_{22}$  is a diagonal matrix.

(c)  $\tilde{\mathbf{u}}$  may be eliminated from the system as follows

$$\begin{aligned} M_{21}\mathbf{u} + D_{22}\tilde{\mathbf{u}} &= \mathbf{b}_2 \\ \Rightarrow \tilde{\mathbf{u}} &= D_{22}^{-1}(\mathbf{b}_2 - M_{21}\mathbf{u}) \end{aligned}$$

which may be substituted into 3 to

$$\begin{aligned} M_{11}\mathbf{u} + M_{12}D_{22}^{-1}(\mathbf{b}_2 - M_{21}\mathbf{u}) &= \mathbf{b}_1 \\ \Rightarrow (M_{11} - M_{12}D_{22}^{-1}M_{21})\mathbf{u} &= \mathbf{b}_1 - M_{12}D_{22}^{-1}\mathbf{b}_2 \\ \Rightarrow M\mathbf{u} &= \mathbf{b} \end{aligned}$$

where  $M = M_{11} - M_{12}D_{22}^{-1}M_{21}$  and  $\mathbf{b} = \mathbf{b}_1 - M_{12}D_{22}^{-1}\mathbf{b}_2$ .

(d) Using the above results,  $M$  and  $\mathbf{b}$  may be computed to be

$$M = \frac{1}{h} \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

$$\mathbf{b} = \frac{5h}{6} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

(with  $h = 0.2$ ) which gives  $u = [0.0667 \quad 0.1000 \quad 0.1000 \quad 0.0667]^T$  when solved in MATLAB. Solving the original system gives

$$u_{full} = [0.0667 \quad 0.1000 \quad 0.1000 \quad 0.0667 \quad 0.0383 \quad 0.0833 \quad 0.1050 \quad 0.0833 \quad 0.0383]^T.$$

Thus, we can see that the same solution is found at the interior nodes of the quadratic finite elements. Additionally, the contributions of the basis functions defined at the midpoints of the elements are already incorporated into  $M$  and  $B$  when using static condensation to solve for the degrees of freedom at the interior nodes. Thus, this technique saves us from having to directly compute five extra degrees of freedom at the midpoints of each element. This technique will be valid so long as the midpoints basis functions are completely confined within one interval. It is this property that allowed for the formation of  $D_{22}$  which promoted the elimination of the vector  $\tilde{u}$  from the original system of equations.

(e) To quickly show that  $M$  is positive definite, the eigenvalues were calculated using MATLAB and found to be  $\lambda = [0.3820, 1.380, 2.6180, 3.6180]^T$ . Since all of the eigenvalues of this matrix are positive and real, the matrix  $M$  is indeed positive definite.



4. Write code for linear FEM using the reference element technique. Optionally, also implement quadratic finite elements using the reference element technique.

## Solution

The linear FEM with the reference element technique was coded. Setting  $f = \pi^2 \sin(\pi x)$  gave the following solutions for  $h = 0.5, 0.25, 0.125$ .

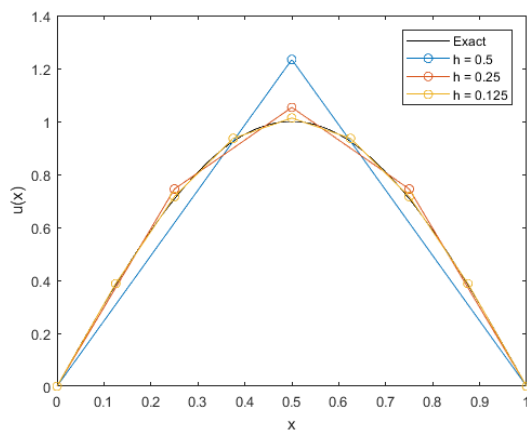


Figure 2: The linear reference element solution for  $\frac{d^2 u}{dx^2} = \pi^2 \sin(\pi x)$ ,  $u(0) = u(1) = 0$ .

This solution matched the solution given using the simple FEM code from Homework 1.

Finally, the quadratic FEM was implemented using the reference element technique. Figure 3 depicts the solution for  $N = 2, 4, 8$  elements. The convergence of the solution for  $\|u - u_k\|_{L^2(0,1)}$ ,  $\|u - u_k\|_E$ , and  $\max_i |u(x_i) - u_k(x_i)|$  norms was also plotted in Figure 3. From this, we see that the energy norm exhibits second-order convergence while the  $L_2$  and max norms exhibit fourth-order convergence.

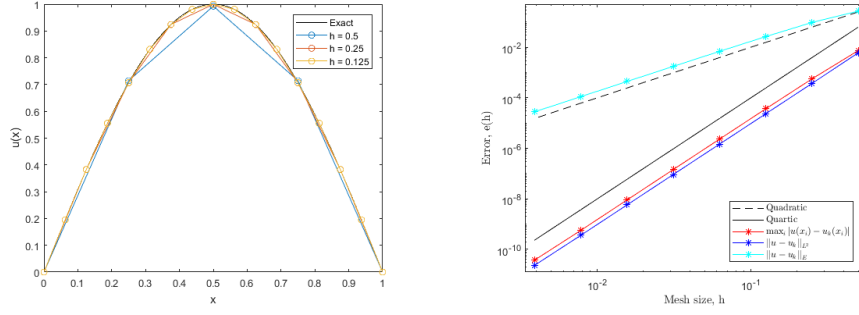


Figure 3: The quadratic reference element solution and convergence plots. The max and  $L_2$  error exhibits fourth-order convergence while the energy norm is second-order.

5. Consider the following finite element.

$$K = [-1, 1]^2, P_K = \text{span}\{1, x, y, xy\} \text{ (bilinear functions on } K),$$

$$\Sigma_K = \{l_i : P_K \rightarrow \mathbb{R}, i = 1, 2, 3, 4 \text{ associated to the four vertices of } K\}.$$

Do the following: (i) prove or disprove that the finite element is unisolvent. (ii) In the case that it is unisolvent, compute the basis of  $P_K$  consistent with the degrees of freedom. Draw plots of the basis functions. (iii) Construct the interpolation operator on  $C^0(K)$ . (iv) If we move the degrees of freedom to the midpoints of edges of  $K$ , is the FEM still unisolvent?

### Solution

(i) We begin by defining a general bilinear basis function

$$\psi_i(x, y) = a_i + b_i x + c_i y + d_i xy; \quad x, y \in [-1, 1]$$

From this, we define the degrees of freedom at the vertices of the square as

$$l_1(\psi_1) = \psi_1(-1, -1) = a_1 - b_1 - c_1 + d_1$$

$$l_2(\psi_2) = \psi_2(-1, 1) = a_2 - b_2 + c_2 - d_2$$

$$l_3(\psi_3) = \psi_3(1, -1) = a_3 + b_3 - c_3 - d_3$$

$$l_4(\psi_4) = \psi_4(1, 1) = a_4 + b_4 + c_4 + d_4$$

this may be rewritten in matrix form as

$$A = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

which has  $\det(A) = -16$ .

(ii) Since the determinant is nonzero, this finite element is unisolvent. From here, using the Kronecker property, the finite element can be rewritten into system form as

$$\begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Which when solved, gives the following expressions for the four bilinear basis functions

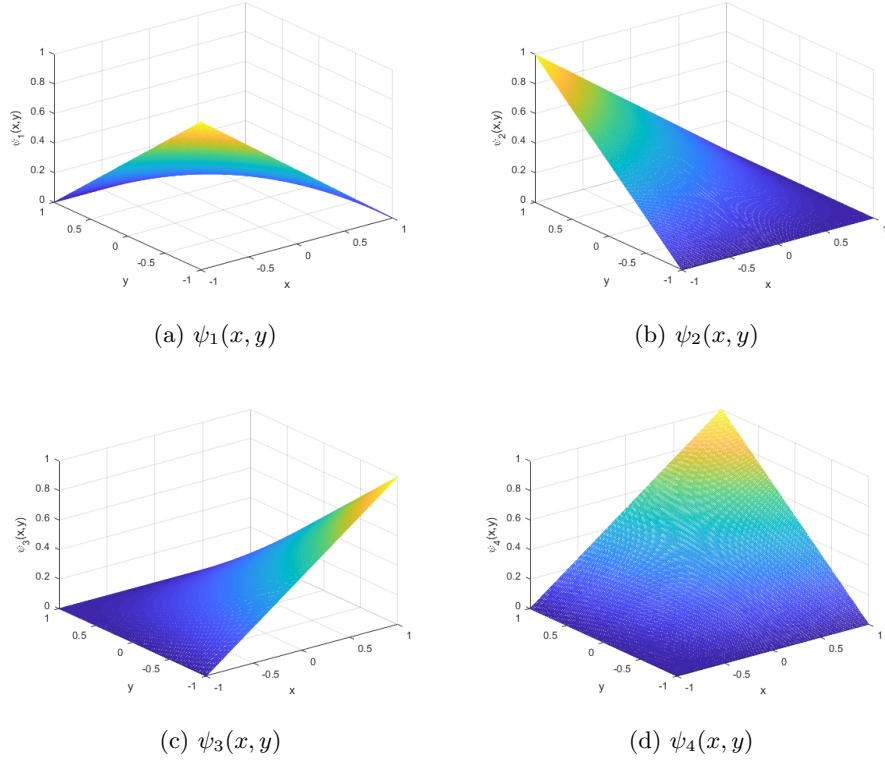


Figure 4: The four bilinear basis functions defined on the vertices of a square.

$$\psi_1(x, y) = \frac{1}{4}(1 - x - y + xy)$$

$$\psi_2(x, y) = \frac{1}{4}(1 - x + y - xy)$$

$$\psi_3(x, y) = \frac{1}{4}(1 + x - y - xy)$$

$$\psi_4(x, y) = \frac{1}{4}(1 + x + y + xy)$$

which are plotted above in Figure 4.

(iii) The interpolation operator on  $C^0(K)$  may be written as

$$\begin{aligned} \pi_K(v)(x, y) = & \frac{1}{4}[v(-1, 1)(1 - x - y + xy) \\ & + v(-1, -1)(1 - x + y - xy) \\ & + v(1, -1)(1 + x - y - xy) \\ & + v(1, 1)(1 + x + y + xy)] \end{aligned}$$

(iv) Moving the degrees of freedom to the midpoints of the edges of  $K$  gives the following expressions for the degrees of freedom

$$\begin{aligned}l_1(\psi_1) &= \psi_1(-1, 0) = a_1 - b_1 \\l_2(\psi_2) &= \psi_2(1, 0) = a_2 + b_2 \\l_3(\psi_3) &= \psi_3(0, -1) = a_3 - c_3 \\l_4(\psi_4) &= \psi_4(0, 1) = a_4 + c_4\end{aligned}$$

which may be rewritten into matrix form as

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

which has a determinant equal to zero. Thus, this finite element is not unisolvent.