Plan

On the next few pages you will find the syllabus for `MA3951` and `MA5352`. The listing of the syllabus includes a possible teaching schedule covering the first part of the material which is common to both modules to the end of 2004. Dr. Michael Warby (MKW) is taking the first part which is common to both `MA3951` and `MA5352` with Dr. Simon Shaw (SSS) taking the extra material which is in `MA5352` for the first part of the year and possibly into the spring as well. Simon Shaw will let you know the corresponding teaching schedule for his part for those taking `MA5352`. Dr. Francis Choi (CFC) will be taking the part common to both modules from January onwards.

For the first part of the module `MA3951` there will generally be 2 lectures per week and 1 hour which will be treated as a seminar or examples class. However, for the first 2 weeks all 3 hours will be used for lectures and then seminars will start from week 3 onwards. Students taking `MA5352` take all the hours of `MA3951` plus the additional hour taught by Simon Shaw. Francis Choi will let you know the arrangements for his part of the module when this starts.

The notes for the part taught by MKW will be handed out in the lectures and they will generally also made available from his 'home page area' on the www. This address together with other details are listed below.

Name: Dr. M. K. Warby

Room: `M301` (Mathematics building)

My number is `++44 1895 265604`
(The Brunel number is `++44 1895 274000`)

Mathematics department Fax: `++44 1895 203303`

Email: `Mike.Warby@brunel.ac.uk`

URL: `http://www.brunel.ac.uk/~icstmkw/home.html`

2004/2004

July

Su	Mo	Tu	We	Th	Fr	Sa
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

August

Su	Mo	Tu	We	Th	Fr	Sa
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

September

Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

October

Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

November

Su	Mo	Tu	We	Th	Fr	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

December

Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

January

Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

February

Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

March

Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

April

Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

May

Su	Mo	Tu	We	Th	Fr	Sa
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

June

Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

Material common to both MA3951 and MA5352.

27	Sep	--	17	Dec	Weeks 1--12	Lectures etc. by MKW
20	Dec	--	3	Jan	Weeks 13--14	*Christmas break*
4	Jan	--	24	Mar	Weeks 15--26	Lectures etc. by CFC
25	Mar	--	15	Apr	Week 27--29	*Easter break*
18	Apr	--	29	Apr	Weeks 30--31	Revision
3	May	--	20	May	Weeks 32--34	Exams

BRUNEL UNIVERSITY

DEPARTMENT OF MATHEMATICAL SCIENCES

SYLLABUS

Module Codes:	`MA3951/MA5352`
Module Title:	Numerical and Variational Methods for PDEs
Levels:	3 (`MA3951`), MSc (`MA5352`).
Lecturers:	Dr. M. K. Warby, Dr. S. Shaw(`MA5352` only) and Dr. F. Choi
Credit value:	20 (for `MA3951`), 30 (for `MA5352`).
Prerequisites:	Core level 1 calculus, algebra and numerical methods modules for `MA3951` students.
Formative feedback:	Participation in the seminars and questionnaires.
Method of assessment:	`MA3951:`Three hour unseen written examination worth 70% and 2 assignments worth in total 30%. `MA5352:` Three hour unseen written examination worth 70% and 3 assignments worth in total 30%.

Main aims of the module

(i) To reformulate two-point boundary value problems for ordinary differential equations (ODEs) and elliptic partial differential equations into weak form and to consider their subsequent solution by the Galerkin finite element method (FEM).

(ii) To introduce students to the classical calculus of variations.

(iii) To introduce finite difference methods for parabolic and hyperbolic partial differential equations and the method of characteristics for hyperbolic equations.

(iv) `MA5352` only:

To consider the theoretical background to the finite element method and its use and role in *a priori* and *a posteriori* error analysis.

Main topics of study

Preliminaries:

The spaces $C(\Omega), C^1(\Omega), C^2(\Omega), \cdots$ and an introduction to piecewise polynomial interpolation and approximation. A revision of differentiation and integration in $\mathbb{R}^2$ including the gradient vector, directional derivatives, double integrals, contour integrals, the divergence theorem and the transformation of double integrals.

One space dimension problems:

The reformulation of two point boundary value problems into a weak forms and as minimization problems for second and fourth order problems.
Essential and natural boundary conditions.
The calculus of variations and the Euler Lagrange equations for the stationary point of a functional.
The Ritz and Galerkin methods.
Weakening the continuity requirements and an introduction to Sobolev spaces $L_2(0,1) = H^0(0,1)$, $H^1(0,1)$ and $H^2(0,1)$. Conforming approximation spaces and the finite element spaces. Linear elements, quadratic elements and cubic Hermite elements. Element matrices, element vectors, assembly and dealing with boundary conditions. Properties of the element and global matrices.

An abstract overview of the finite element procedure:

Hilbert spaces, linear functionals, bilinear forms, the energy norm, Galerkin orthogonality and best approximation in the energy norm.

The finite element method in two space dimensions:

Poisson's equation with Dirichlet, Neumann and mixed boundary conditions.
The weak formulation of Poisson's equation with various boundary conditions and the case of natural boundary conditions.
The Sobolev spaces $H^0(\Omega)$, $H^1(\Omega)$ and $H^2(\Omega)$, the conforming condition and finite element spaces. This is to include isoparametric elements and in particular 3 noded linear triangles, 4 noded bilinear quadrilaterals, 6 noded quadratic triangles, 8 noded serendipity (quadrilateral) elements and the 9 noded quadratic quadrilateral. Transformation techniques and the computation of element matrices and element vectors. The assembly of element matrices and element vectors and the implementation of boundary conditions. Properties of the element and global matrices.
The residual in the case of piecewise linears and a discussion of *a-posteriori* error estimation and refinement.

Additional topics in `MA5352`

Further study of the weak formulation of the two point boundary value problem: $-(p(x)u'(x))' + q(x)u(x) = f(x)$ in $(0,1)$, subject to the Dirichlet boundary conditions: $u(0) = u(1) = 0$.
The Hilbert spaces $L_2(0,1)$ and $H^1(0,1)$. Linear and bilinear forms. The energy norm. Existence, uniqueness and stability by the Riesz Representation Theorem.
Finite element approximation. Interpolation-error estimates for piecewise polynomials. Galerkin scheme and Céa's lemma.
A-priori and $H^1(0,1)$ error estimates. Duality and *a priori* $L_2(0,1)$ error estimates. Green's function and *a-priori* maximum norm estimates.
A-posteriori energy, $H^1(0,1)$ and $L_2(0,1)$ error estimates. Adaptive error control.
Overview of the extra difficulties presented by two and three dimensional problems.

Finite difference methods for evolutionary problems

Parabolic equations. The simple diffusion equation in one space dimension; solution by the method of lines leading to sequential and parallel algorithms. Local truncation errors, consistency, A-stability, L-stability, convergence. Reaction-diffusion problems. Extension to two space variables and to nonlinear problems.
Hyperbolic equations. The method of characteristics for first-and second-order hyperbolic equations. Solution of hyperbolic equations (first-and second-order) by explicit and implicit finite-difference methods. Local truncation errors, consistency, stability and convergence.

Learning outcomes

On successful completion of the module, students will be able to do the following:

(i) reformulate ODEs and various elliptic PDEs with a variety of boundary conditions into weak form and to discretise the weak form using the Galerkin finite element method;

(ii) demonstrate knowledge of the basic properties of the different types of second order PDEs;

(iii) develop and analyse finite-difference methods for solving parabolic and hyperbolic PDEs and integrate along the characteristics of hyperbolic PDEs;

(iv) `MA5352` only:

apply techniques of modern error analysis.

Teaching/ learning methods/ strategies used to enable the achievement of learning outcomes

`MA3951`: 72 contact hours comprising lectures and example classes (seminars).
`MA5352`: 96 contact hours comprising lectures and example classes (seminars).

In lectures concepts are introduced, definitions are stated, results (theorems) and techniques are explained, often proved. All material is illustrated with practical examples. Example classes are contact hours where students can practice exercises, examples and ask questions in a supportive environment.

The University expects student learning hours for their studies to be a notional 35-hour week (over 31 weeks) including taught classes, private study, revision and assessment. In practice, students will spend time on each module that will vary from week to week, between modules and between students. In particular, attention should be paid to the assessment pattern (assignments, no final examination ...) and the delivery pattern (number of contact hours) that will have an impact on the organisation of the working weeks during the Academic Year.

`MA3951`: Average learning hours per week: 6 of which 3 hours are average student/tutor contact hours per week.

`MA5352`: Average learning hours per week: 9 of which 3 hours are average student/tutor contact hours per week.

Assessment methods which enable student to demonstrate the learning outcomes for the module

`MA3951`:

Continuous assessment consisting of 2 written assignments.	30%
A three hour unseen written examination.	70%

`MA5352`:

Continuous assessment consisting of 3 assignments.	30%
A three hour unseen written examination.	70%

Lecture notes

A paper version of notes, exercises and solutions to exercises are distributed during the lectures and seminars. These are also available on the www for the part of the module which is common to both `MA3951` and `MA5352` using the URL

`http://www.brunel.ac.uk/~icstmkw/ma3951/`

Reading list

Essential reading:

None. There is no core text for this module but several texts cover most parts of the module. Some of the texts are reasonably advanced. There are numerous texts concerned with finite element methods and in the library these may be in the mathematics section with codes starting with `QA` or in the engineering section with codes starting with `TA`.

Recommended reading for the finite element part

The following are a few of the texts that I consult regularly concerning the finite element method.

Axelsson, O. and Barker, V. A.:
Finite element solution of boundary value problems. SIAM (Society for Industrial and Applied Mathematics), 432 pages 2001.
ISBN `0 89871 499 0`, `QA 379.A9 2001`.
This book is a reproduction of the 1984 book of the same title published by Academic Press, Inc, Orlando, San Diego, San Fransisco.

Eriksson, K., Estep D., Hansbo, P. and Johnson, C:
Computational differential equations. Cambridge University Press, 538 pages, 1996.
ISBN `0 521 567 386`, `QA 371.5.D37C66`.

Johnson, C:
Numerical solution of partial differential equations by the finite element method. Cambridge University Press, 279 pages, 1987.
ISBN `0 521 347 580`, `TA 347.F5J64. 1987`

Reddy, J. N.:
An introduction to the finite element method. McGraw-Hill, 684 pages, 1993.
ISBN `0 07 051355 4`, `TA 347.F5R4`.

Recommended reading for the finite difference part

Iserles, A.:
A first course in the numerical analysis of differential equations. Cambridge University Press, 378 pages, 1996.
ISBN `0 521 55655`, `QA 371.I74`.
(This book also contains a chapter on the finite element method.)

Twizell, E. H.:
Computational methods for partial differential equations. Chichester: Ellis Horwood, 276 pages, 1984. `QA377.T84`.

Other texts relating to the finite element part

Numerous texts could be listed here and the following are just a few of them.

Braess, D.:
Finite elements, theory, fast solvers, and applications in solid mechanics. Cambridge University Press, Second edition, 352 pages, 2001.
ISBN `0521 01195 7`, `TA 347.F5B7313 2001`.

Brenner, S. C. and Scott, L. R.:
The mathematical theory of finite element methods. Second edition, Springer, 361 pages, 2002.
ISBN `0387 95451 1`, `QA 379.B74`,

Lewis, P.E. and Ward, J.P.:
The finite element method, principles and applications. Addison-Wesley Publishers Ltd, 421 pages, 1991.
ISBN `0-201-54415-6`, `QA 379.L49`.

Strang, G. and Fix, G.J.:
An analysis of the finite element method. Englewood Cliffs ; London: Prentice-Hall, 306 pages, 1973.
ISBN `0 961 40888X.01`, `TA 335.S77`.

Zienkiewicz, O.C. and Morgan, K.:
Finite elements and approximation. John Wiley and Sons, New York, 328 pages, 1983.
ISBN `0-471-89089-8`, `QA 297.5.Z53`.

Zienkiewicz, O.C. and Taylor, R.L.:
The finite element method, volume 1, the basis. Butterworth-Heinemann, fifth edition, 689 pages, 2000.
ISBN `0 7506 5049 4`, `TA 640.2.Z53 2000`.

Conference Proceedings relating to the finite element part

There are many conferences which involve finite elements and to advertise Brunel you should note that we run a conference here at Brunel about every three years called "The Mathematics of Finite Elements and Applications (MAFELAP)". The last such conference was the 11th and took place in June 2003. The previous 10 conferences have been here in 1972, 1975, 1978, 1981, 1984, 1987, 1990, 1993, 1996 and 1999. Most of the proceedings for these conferences (all edited by Professor Whiteman) are in the Brunel library under the number `QA 374.C65`.

MA3951 Finite element methods part, 2003–2004 Teaching Schedule

As there will be 3 lectures in weeks 1 and 2 the 20th lecture will be at the end of week 9. Please note this is just a guide.

Lecture number (approx)	Topics
1	Introduction to the entire module and an introduction to finite elements.
2,3	Preliminary material.
4–7	Reformulating ODEs into weak forms and as minimization problems.
8	Minimising functionals in general — the Euler equations.
9	The Galerkin and Ritz methods.
10	An introduction to Sobolev spaces.
11,12	Piecewise linears, quadratics and cubics. Element computations and assembly.
13,14	An abstract overview of the FEM.
15,16	Preliminaries for problems in $\mathbb{R}^2$ and reformulating Poisson's equation.
17–20	Piecewise linears on triangles. Bilinear quadrilateral elements.
21,22	6, 8 and 9 noded quadrilateral elements.
23–26	A discussion of additional topics and a revision of the module so far.

1 General comments about the module

This is the second time that the modules `MA3951` and `MA5352` have run in their current form and they relate approximately to the modules `MA3052A` and `MA3056S` for level 3 students and `MA5052A` and `MA5156S` for MSc students when the academic year was previously divided into an autumn and a spring semester. `MA3052A` and `MA5052A` were called "the numerical solution of partial differential equations (PDEs)" and `MA3056S` and `MA5156S` were respectively called "finite element methods" and "finite element methods and error analysis".

In `MA3951` and `MA5352` we consider both the finite element method and the finite difference method for the approximate solution of various ordinary differential equations (ODEs) and PDEs. In both cases the methods involve approximating a continuous problem by a problem involving only a finite number of parameters but the way that this is done is different in each case. In the finite difference method the idea is just to approximate partial derivatives by differences, e.g. if $u = u(x, t)$ then

$$\frac{\partial u}{\partial t}(x_i, t_j) \approx \frac{u(x_i, t_j + k) - u(x_i, t_j)}{k} \tag{1.1}$$

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_j) \approx \frac{u(x_i - h, t_j) - 2u(x_i, t_j) + u(x_i + h, t_j)}{h^2} \tag{1.2}$$

provided h and k are small. If this is done at every grid point in the domain under consideration then this generates a finite system of equations with the unknowns being the approximate solution at the grid points. The finite element method idea is to instead approximate the unknown solution u by a function

$$U = \sum_{1}^{n} c_i \phi_i \tag{1.3}$$

which is a finite linear combination of given basis functions $\phi_1, \cdots, \phi_n$ with the unknowns in this case being the coefficients $c_1, \cdots, c_n$. In the finite element method the functions $\phi_1, \cdots, \phi_n$ used are defined with respect to a mesh of the domain under consideration and the equations that determine $c_1, \cdots, c_n$ arise from approximately solving a weak problem associated with the original differential equation.

For many of the ODEs and PDEs encountered throughout this module both the finite difference method and the finite element method can be used although throughout usually only one of these methods will be used in any specific case. The choice as to which is described is influenced by factors such as which is "best" and if the best is very complicated then by which is "easiest" to describe. In the case of what are known as "steady-state" type equations the finite element method is widely regarded as the superior method to use and this is what is covered here. A standard example of this type is Poisson's equation

$$-\Delta u = \nabla^2 u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f, \quad (x, y)^T \in \Omega \subset \mathbb{R}^2, \quad u(x, y) = g \text{ on } \partial\Omega \tag{1.4}$$

where f and g are given and Ω is a given bounded domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. In the case of what are known as "evolutionary" equations the finite difference method is still quite widely used and it is simpler to describe and thus these are treated in this module by the finite difference method. Standard examples of these are the one space dimension heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \tag{1.5}$$

together with appropriate initial and boundary conditions, and the one space dimension wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \tag{1.6}$$

together with appropriate initial and boundary conditions. In (1.5) and (1.6) t denotes time in the physical situations that they describe which is where the term evolutionary comes from.

In the above the classification of the differential equations to be considered into steady-state and into evolutionary can be further classified as elliptic, parabolic and hyperbolic for respectively (1.4), (1.5) and (1.6). These equations are usually taken as the "simplest" examples of these types when the general second order quasi-linear equations of the form

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial xy} + c\frac{\partial^2 u}{\partial y^2} + e = 0 \tag{1.7}$$

are considered. In (1.7) the coefficients a, b, c and e may depend on x, y, u, $\partial u/\partial x$ and $\partial u/\partial y$ but not on any higher partial derivatives of u. In this general case the equation is classified as

(i) elliptic if $b^2 - 4ac < 0$,

(ii) parabolic if $b^2 - 4ac = 0$,

(iii) hyperbolic if $b^2 - 4ac > 0$.

The naming of this classification comes from the name of the conic sections that are described by

$$ax^2 + bxy + cy^2 + e = 0 \tag{1.8}$$

when a, b, c and e are constants. The reason for considering each case separately is that the solutions to (1.4), (1.5) and (1.6) have quite different features and this in turn influences how the numerical scheme is implemented. For example, as is shown later in the module, hyperbolic equations have "characteristic curves" which do not exist or are not a feature of the elliptic and the parabolic equations. This in turn leads to a method of characteristics which applies just to the hyperbolic case.

Thus to summarize, in this module the finite element method is described for the approximate solution of steady state elliptic ODEs and PDEs in the first part of the module and the finite difference method is described for the approximate solution of the parabolic and hyperbolic evolutionary type of PDEs.

2 Introduction to finite elements

The finite element method (FEM) is an established method for constructing an approximate solution to problems described by an ordinary differential equation (ODE), a system of ODEs, a partial differential equation (PDE) or a system of PDEs which arise in many areas of science and engineering. If Ω denotes the domain or region of the ODEs or PDEs involved (typically a region in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$) then the method involves, in brief, dividing the domain into non-overlapping subdomains $\Omega_1, \cdots, \Omega_{ne}$ and by constructing an approximate solution which is defined piece-by-piece. The division of Ω into the ne subdomains $\Omega_1, \cdots, \Omega_{ne}$ is known as the *mesh* with the component parts Ω_k known loosely as the *elements*. Reasons for the success of the method include the following.

(i) It is applicable to a wide range of problems.

(ii) It copes with complex geometries.

(iii) The elements can vary in size and thus the mesh can be adaptively refined until a desired accuracy in the approximation is achieved efficiently.

(iv) It copes with complex boundary conditions of physical interest.

It is because of such reasons that the method is considered superior to the simpler finite difference method when it can be applied and the success of the procedure is evident from the fact that it is the basis for several commercial finite element packages and computer aided design (CAD) systems which are widely used in industry. From these remarks and from a scan of the literature available on the FEM or on the use of the FEM it is evident that the FEM is a large topic and as a consequence only a fraction of the topic can be covered in this module. We outline next some of the things that will be done as part of `MA3951` and `MA5352`.

2.1 An overview of material in the finite element part of the module

The ODEs and PDEs that will be considered in this module will include the following two-point boundary value problem and elliptic PDE.

(i) Find $u : [0,1] \to \mathbb{R}$ satisfying

$$-(pu')' + qu = f, \quad 0 < x < 1, \quad u(0) = u(1) = 0, \tag{2.1}$$

where p, q and f are suitable given functions of x.

(ii) Find $u : [0,1] \to \mathbb{R}$ satisfying

$$u'''' = f, \quad 0 < x < 1, \quad u(0) = u'(0) = u(1) = u'(1) = 0, \tag{2.2}$$

for a given function f.

(iii) Given a connected open domain $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega$ find $u : \Omega \to \mathbb{R}$ satisfying

$$-\Delta u = \nabla^2 u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f, \quad (x,y)^T \in \Omega \subset \mathbb{R}^2, \quad u(x,y) = 0 \text{ on } \partial\Omega \tag{2.3}$$

for a given function f.

Further generalizations of these will be considered and in particular in the case of (iii) we will consider more complicated boundary conditions than the homogeneous Dirichlet conditions specified which use the normal derivative of u being specified on part of the boundary. Equation (2.3) is known as Poisson's equation.

A key first step to approximately solving these problems will be to first re-express them into what is known as a *weak form.* This will just involve multiplying the ODE or PDE by a suitable function v (known as a test function) and then integrating over the domain. In the case of (i) with a differentiable function v which satisfies $v(0) = v(1) = 0$ we get, by integration by parts,

$$\int_0^1 fv \, \mathrm{d}x = \int_0^1 (-(pu')' + qu)v \, \mathrm{d}x \tag{2.4}$$

$$= \int_0^1 (pu'v' + quv) \, \mathrm{d}x - [vpu']_0^1, \tag{2.5}$$

$$= \int_0^1 (pu'v' + quv) \, \mathrm{d}x, \quad \text{since } v(0) = v(1) = 0. \tag{2.6}$$

The weak problem then involves finding u satisfying the boundary conditions such that

$$\int_0^1 (pu'v' + quv) \, \mathrm{d}x = \int_0^1 fv \, \mathrm{d}x \quad \text{for all appropriate } v. \tag{2.7}$$

The key thing to note here is that in (i) we require u to have two derivatives (the term u'' which is part of $(pu')' = pu'' + p'u'$) whereas in (2.7) we only need u to have one derivative. In fact we only need u to have one derivative which can be integrated which is a lesser requirement. u' need not exist at every point but should exist sufficiently so that we can integrate it. This is where the term weak comes from. The weak formulation involves expressions with much weaker continuity requirements on the function u to be found.

The FEM is a method for approximately solving the weak problem and, under appropriate conditions, solving the weak problem is equivalent to solving the original ODE or PDE. If we define a bilinear form $a(.,.)$ and an inner product $(.,.)$ by

$$a(u,v) = \int_0^1 (pu'v' + quv) \, \mathrm{d}x \quad \text{and} \quad (f,v) = \int_0^1 fv \, \mathrm{d}x \tag{2.8}$$

then the weak problem involves finding u in an appropriate space of functions such that

$$a(u,v) = (f,v) \quad \text{for all appropriate } v. \tag{2.9}$$

If the appropriate space is labelled as V then the Galerkin method for obtaining an approximate solution involves constructing linearly independent functions $\phi_1, \cdots, \phi_n$ in V which gives a finite dimensional subspace $V_n := \mathrm{span}\{\phi_1, \cdots, \phi_n\} \subset V$ and with finding $U \in V_n$ such that

$$a(U,v) = (f,v) \quad \text{for all } v \in V_n. \tag{2.10}$$

The approximation U is hence of the form

$$U = \sum_1^n c_i \phi_i \tag{2.11}$$

and we can satisfy (2.10) by taking v to be each of $\phi_1, \cdots, \phi_n$. That is the n coefficients $c_1, \cdots, c_n$ satisfy the n equations

$$a\left(\sum_{i=1}^{n} c_i \phi_i, \phi_j\right) = (f, \phi_j), \quad j = 1, \cdots, n. \tag{2.12}$$

As $a(.,.)$ is a bilinear form (this will be considered properly later) these equations can be written as

$$\sum_{i=1}^{n} a(\phi_i, \phi_j) c_i = (f, \phi_j), \quad j = 1, \cdots, n. \tag{2.13}$$

The Galerkin finite element method, or simply the finite element method, is precisely this method with specific choices for the finite dimensional subspace V_n. The functions in V_n are piecewise polynomial with typically each function ϕ_i being zero over all but a few adjacent elements which for this one-dimensional problem will involve functions ϕ_i which are non-zero only on two adjacent subintervals, see Figure 2.1 for the case of piecewise linear functions which are also known as hat functions.

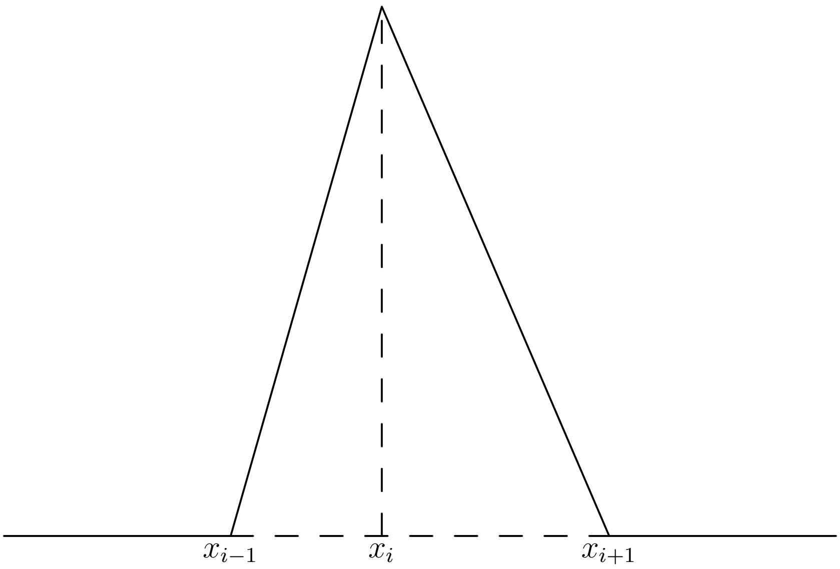


Figure 2.1: The graph of a hat function

For Poisson's equation in $\mathbb{R}^2$ with homogeneous Dirichlet boundary conditions given in (2.3) we can similarly obtain a weak form which turns out to be

$$a(u, v) = (f, v) \tag{2.14}$$

where now

$$a(u, v) = \iint_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right) \mathrm{d}x\mathrm{d}y \quad \text{and} \quad (f, v) = \iint_\Omega f v \, \mathrm{d}x\mathrm{d}y. \tag{2.15}$$

In the FEM applied to this problem the basis functions typically involve low order polynomials defined on triangular and/or quadrilateral elements Ω_k into which Ω is divided. The details concerning how this is done with the various element computations that are needed and how these are put together (the assembly procedure) is a reasonably large component of this module.

For the type of problems that are considered in this module the function u solving the weak problem will also minimise the corresponding quadratic functional

$$I(v) = \frac{1}{2}a(v,v) - (f,v) \tag{2.16}$$

over an appropriate space of functions, i.e. if V denotes this space then $u \in V$ is such that

$$I(u) \le I(v) \quad \text{for all } v \in V. \tag{2.17}$$

We obtain an approximate solution U by minimising this functional over a finite dimensional subspace V_n and this method of doing things is usually referred to as the Ritz method or the Rayleigh-Ritz method. If, as in the Galerkin method, $V_n = \text{span}\{\phi_1, \cdots, \phi_n\}$ then the Ritz method of doing things leads to exactly the same approximate solution U. Hence there is no advantage here in obtaining the same U from two different approaches. However, the equivalence of the weak formulation and the formulation in which we minimise a functional provides a firmer foundation for the method and it is used in error analysis which is concerned with how close U is to u for a given set of basis functions.

A further connection between the problems of solving a differential equation and of minimising a functional will be considered when we briefly consider what is known as "the calculus of variations" which takes the functional as its starting point. That is instead of obtaining the functional associated with the given ODE or PDE we consider the reverse problem of finding the ODE or PDE which characterises the function which minimises a given functional. This is how problems typically arises in physics and engineering when the functional involved represents some sort of energy. What we will obtain here is the Euler equations. With the quadratic functionals just considered the Euler equations are just the linear ODEs and PDEs that we started with but with more general functionals the Euler equations are non-linear ODEs and PDEs in general.

The planned route to covering the material in the first part of the module will involve first considering a few preliminary things which many of you may have done before. In particular, we consider spaces of continuous and continuously differentiable functions in which the exact and approximate solutions lie. We consider how well functions can be approximated by polynomials and we consider the divergence theorem from vector calculus which is needed to convert PDEs such as Poisson's equation into weak form. After the preliminary material we then consider many different aspects of the FEM for one dimensional problems so that full details can be covered without any complications such as difficult geometries. The integrations, the boundary conditions and the spaces are all simpler in one dimension. An abstract overview of the entire procedure will then be considered which is the basis of error estimation and this will be considered further in the `MA5352` lectures of Dr. S. Shaw. The remainder of the module, and probably the largest part of the module, will then concern the FEM for two dimensional problems such as Poisson's equation with various boundary conditions and various triangular and quadrilateral elements.

Problems in three space dimensions (i.e. in $\mathbb{R}^3$) and problems with systems of PDEs is beyond what will be covered in this module but it not such a big advance from what is covered here. It is systems of PDEs which describe how a structure deforms when loaded (the equilibrium equations of elasticity) where some of the early development of the FEM took place in the 1950s and 1960s with the corresponding development of computers. Some of the jargon of the subject reflects these roots with terms such as stiffness matrices and load vectors still being used in many different modelling situations including those which physically have nothing to do with elasticity.

3 Preliminaries

3.1 A discussion about meshing and continuity

In the introduction it was stated that the FEM involved dividing a domain Ω into non-overlapping subdomains $\Omega_1, \cdots, \Omega_{ne}$ and with constructing an approximate solution which is defined piece-by-piece. The details of this procedure will thus involve the form of the approximation on each piece and on how the approximation on each piece is related to the approximation on any neighbouring piece. This in turn will involve low degree polynomials and continuity. To consider these statements mathematically we consider first the domain Ω and we assume that it is a connected open bounded region which has a boundary which we denote by $\partial\Omega$ and let $\bar{\Omega} = \Omega \cup \partial\Omega$ denote the corresponding closed domain. For example, in one dimension a connected open bounded domain is simply an open interval such as $(0, 1)$, i.e.

$$\Omega = (0, 1), \quad \partial\Omega = \{0, 1\}, \text{ i.e. the boundary is the two edge points and } \bar{\Omega} = [0, 1]. \tag{3.1}$$

An example of a two dimensional domain is a $2:1$ rectangle

$$\Omega = \{\underline{x} : \ \underline{x} = (x, y)^T, \ 0 < x < 2, \ 0 < y < 1\}. \tag{3.2}$$

The closed domain is

$$\bar{\Omega} = \{\underline{x} : \ \underline{x} = (x, y)^T, \ 0 \le x \le 2, \ 0 \le y \le 1\}. \tag{3.3}$$

The boundary $\partial\Omega$ is the union of the 4 straight line segments Γ_1, Γ_2, Γ_3 and Γ_4, shown in Figure 3.1. With the convention that the subdomains Ω_k used in the division of Ω are open, the division of Ω into non-overlapping subdomains means, in set notation, that

$$\bar{\Omega} = \bigcup_{k=1}^{ne} \bar{\Omega}_k, \quad \Omega_i \cap \Omega_j = \text{empty set} \ , \quad \text{for } i \ne j. \tag{3.4}$$

The division of Ω into $\Omega_1, \cdots, \Omega_{ne}$ is known as the mesh with each Ω_k known as an element. (The term element usually refers to more than just the region of Ω_k as it usually also refers to the type of approximation that is used on the region. This will be considered later in the module.)

In one dimension a mesh of the domain $[0, 1]$ can be constructed easily. We simply take any collection of points x_i with

$$0 = x_0 < x_1 < \cdots < x_{ne} = 1, \quad \text{and} \quad \Omega_k = (x_{k-1}, x_k), \quad k = 1, \cdots, ne, \tag{3.5}$$

see Figure 3.2. In two dimensions see Figures 3.3 for possible meshes of the $2:1$ rectangle.

With a mesh of $\bar{\Omega}$ the finite element approximation of a problem involves a function, which we denote here by U, which is defined on $\bar{\Omega}$ and which is a low degree polynomial on each piece $\Omega_1, \cdots, \Omega_{ne}$. It is important to know how the polynomials on one piece relates to the polynomial on any neighbouring piece. For example, in one dimension the elements (x_{k-1}, x_k) and (x_k, x_{k+1}) both have $x = x_k$ as a boundary point. The global continuity of U defined on $\bar{\Omega}$ is concerned with whether or not

$$\lim_{\substack{x \to x_k \\ x \in (x_{k-1}, x_k)}} U(x) = \lim_{\substack{x \to x_k \\ x \in (x_k, x_{k+1})}} U(x) \tag{3.6}$$

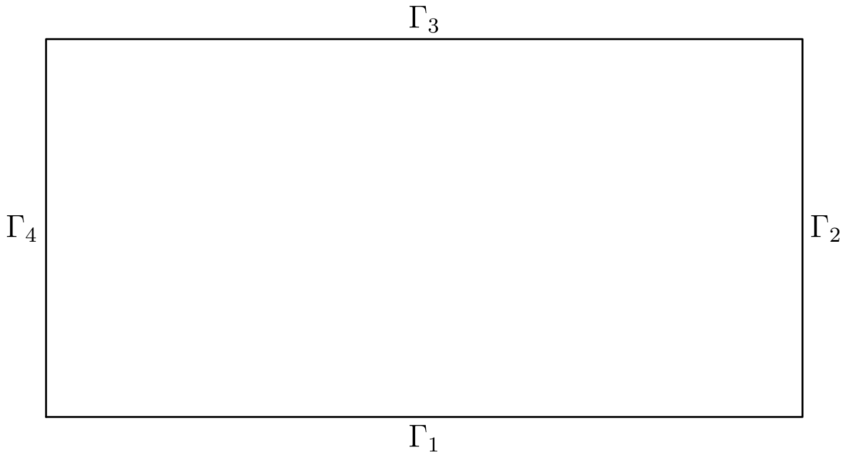


Figure 3.1: A 2 : 1 rectangle.

Figure 3.2: Meshes of $[0, 1]$.

is true and similarly for the first derivative, second derivative etc.. If (3.6) is true at every join point x_k then $U(x)$ is continuous on $\bar{\Omega}$. If

$$\lim_{\substack{x \to x_k \\ x \in (x_{k-1}, x_k)}} U'(x) = \lim_{\substack{x \to x_k \\ x \in (x_k, x_{k+1})}} U'(x) \tag{3.7}$$

is also true at every join point x_k then $U(x)$ is continuously differentiable on $\bar{\Omega}$. We will always seek approximations which are continuous. For most parts of the module (3.7) will not be true, i.e. U will not be continuously differentiable. The exception will be when we consider cubic Hermite functions in one dimension which give a continuously differentiable function but in general the second derivative will be discontinuous.

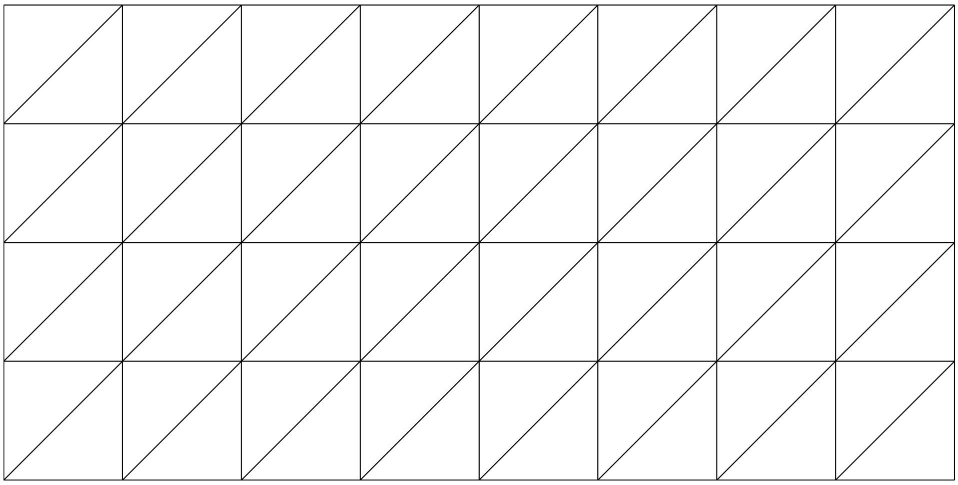
In two dimensions the common boundary of two adjacent triangles or two adjacent quadrilaterals is more than just one point and it is the edge of a triangle or quadrilateral. If $\underline{c} \in \mathbb{R}^2$ is a point on such a common edge of adjacent elements Ω_i and Ω_j then the continuity is again concerned with

$$\lim_{\substack{\underline{x} \to \underline{c} \\ \underline{x} \in \Omega_i}} U(\underline{x}) = \lim_{\substack{\underline{x} \to \underline{c} \\ \underline{x} \in \Omega_j}} U(\underline{x}). \tag{3.8}$$

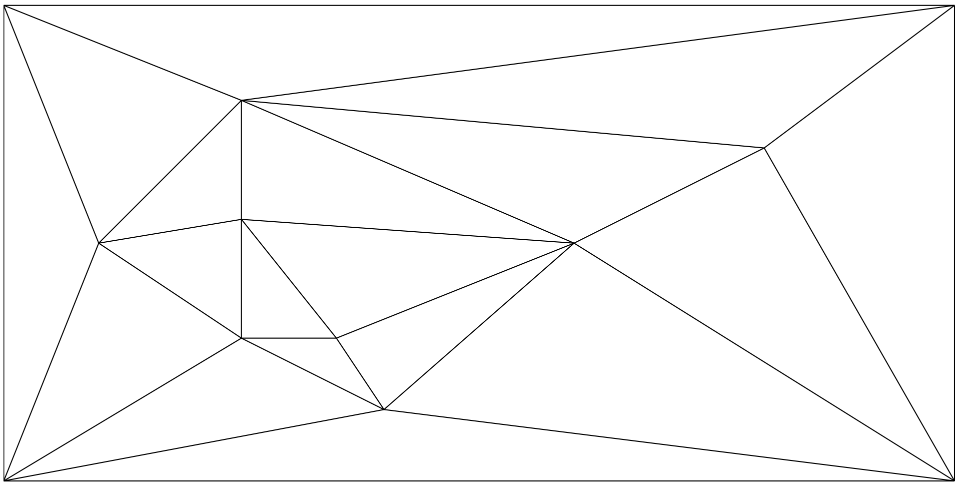
Here $\underline{x} \to \underline{c}$ means $|\underline{x} - \underline{c}| \to 0$ where $|.|$ here means the Euclidean distance. It is much harder in two dimensions to construct functions such that U is continuously differentiable across an edge as this requires that both $\partial U/\partial x$ and $\partial U/\partial y$ are continuous across the edge.

3.2 The spaces $C(0, 1)$, $C^1(0, 1)$, $C^2(0, 1)$ etc.

The discussion of the previous subsection was concerned with what is meant by continuity. When the domain of the function is $(0, 1)$ it is about functions $f : (0, 1) \to \mathbb{R}$ such that the



Uniform mesh



A non-uniform mesh

Figure 3.3: Triangular meshes of a 2 : 1 rectangle.

limit of $f(x)$ as $x \to c$ is the same as the function value $f(c)$. The collection of all functions $f : (0,1) \to \mathbb{R}$ that are continuous gives us a space of functions which we denote by $C(0,1)$. Technically the word *space* here means that if $f \in C(0,1)$ then the function $\alpha f \in C(0,1)$ for all $\alpha \in \mathbb{R}$ and if $f_1, f_2 \in C(0,1)$ then $f_1 + f_2 \in C(0,1)$. Functions which are continuous in the open interval $(0,1)$ can however be such that $|f(x)| \to \infty$ as $x \to 0$ or as $x \to 1$, e.g. $f(x) = 1/x$ or $f(x) = 1/(1-x)$. For functions which are defined at $x = 0$ and at $x = 1$ and are continuous at $x = 0$ and at $x = 1$ we use the notation $C[0,1]$. Functions which are in this space are bounded on $[0,1]$. To summarize the notation we have

$$
\begin{aligned}
C(0,1) &:= \{f : \quad f : (0,1) \to \mathbb{R} \text{ is continuous on } (0,1)\}, &(3.9)\\
C[0,1] &:= \{f : \quad f : [0,1] \to \mathbb{R} \text{ is continuous on } [0,1]\}. &(3.10)
\end{aligned}
$$

As we stated in the previous section, when the FEM is considered for approximately solving an ODE for an unknown function u we will always consider approximations which are in $C[0,1]$ and, in the simplest case, it will involve a piecewise linear approximation as illustrated in Figure 3.4. You should note that the piecewise linear function U shown in each figure is continuous but it has a discontinuous derivative at the points where two different linears join. The function U' is in fact a piecewise constant function. The discontinuities in U' do not however prevent us from being able to integrate such functions. The second derivative U'' cannot be defined in an ordinary sense and thus it does not make sense to substitute such a function U into a second order ODE to see if we can satisfy, or nearly satisfy, the ODE. However, as we will see this will not prevent us from constructing a piecewise linear function U approximating an unknown function u which satisfies an ODE involving second derivatives of the form

$$-u'' = f \quad \text{or more generally} \quad -(pu')' + qu = f \quad \text{for } x \in (0,1) \tag{3.11}$$

where p, q and f are given suitable smooth functions defined on $(0,1)$. As we will see in this module, a key feature in the method for approximately solving (3.11) will be in reformulating the problem in such a way that it makes sense to seek a piecewise polynomial function $U \in C[0,1]$ which approximates a smoother function u satisfying an ODE satisfying (3.11).

For functions which are smoother than those in $C(0,1)$ we use the notation

$$\begin{aligned} C^1(0,1) &:= \{f: \quad f, f' \in C(0,1)\}, & (3.12)\\ C^2(0,1) &:= \{f: \quad f, f', f'' \in C(0,1)\}, & (3.13)\\ C^1[0,1] &:= \{f: \quad f, f' \in C[0,1]\}, & (3.14)\\ C^2[0,1] &:= \{f: \quad f, f', f'' \in C[0,1]\}, & (3.15)\end{aligned}$$

that is for $k = 0, 1, 2, \cdots$ we have

$$\begin{aligned} C^k(0,1) &:= \{f: \quad f, f', \cdots, f^{(k)} \in C(0,1)\}, & (3.16)\\ C^k[0,1] &:= \{f: \quad f, f', \cdots, f^{(k)} \in C[0,1]\}. & (3.17)\end{aligned}$$

Similarly the notation $C^\infty(0,1)$ and $C^\infty[0,1]$ refers to functions which are infinitely differentiable on $(0,1)$ and $[0,1]$ respectively. These spaces are nested, i.e.

$$C^\infty(0,1) \subset \cdots \subset C^k(0,1) \subset \cdots \subset C^1(0,1) \subset C(0,1). \tag{3.18}$$

With this notation the problem

$$-(pu')' + qu = f \quad \text{for } x \in (0,1) \quad u(0) = u(1) = 0 \tag{3.19}$$

makes sense with $q, f \in C(0,1)$ and $p \in C^1(0,1)$ (we need to be able to differentiate p) giving, assuming a solution exists, a function $u \in C^2(0,1) \cap C[0,1]$. That is u is twice continuously differentiable in $(0,1)$ and continuous up to the edge in order to satisfy the boundary conditions $u(0) = u(1) = 0$. To repeat what has already been stated, the FEM, as usually implemented, leads to an approximation $U \in C[0,1]$.

In this module we will also briefly consider the fourth order problem of the form

$$\begin{aligned} &u'''' = f \quad \text{or more generally} \quad (pu'')'' = f \quad \text{in } x \in (0,1),\\ &u(0) = u'(0) = u(1) = u'(1) = 0 \end{aligned} \tag{3.20}$$

for suitably smooth functions p and f. As given the function $u \in C^4(0,1)$. We will show how such a problem can be reformulated in such a way that we can construct an approximation $U \in C^1[0,1]$. It is here that we use the cubic Hermite functions mentioned previously.

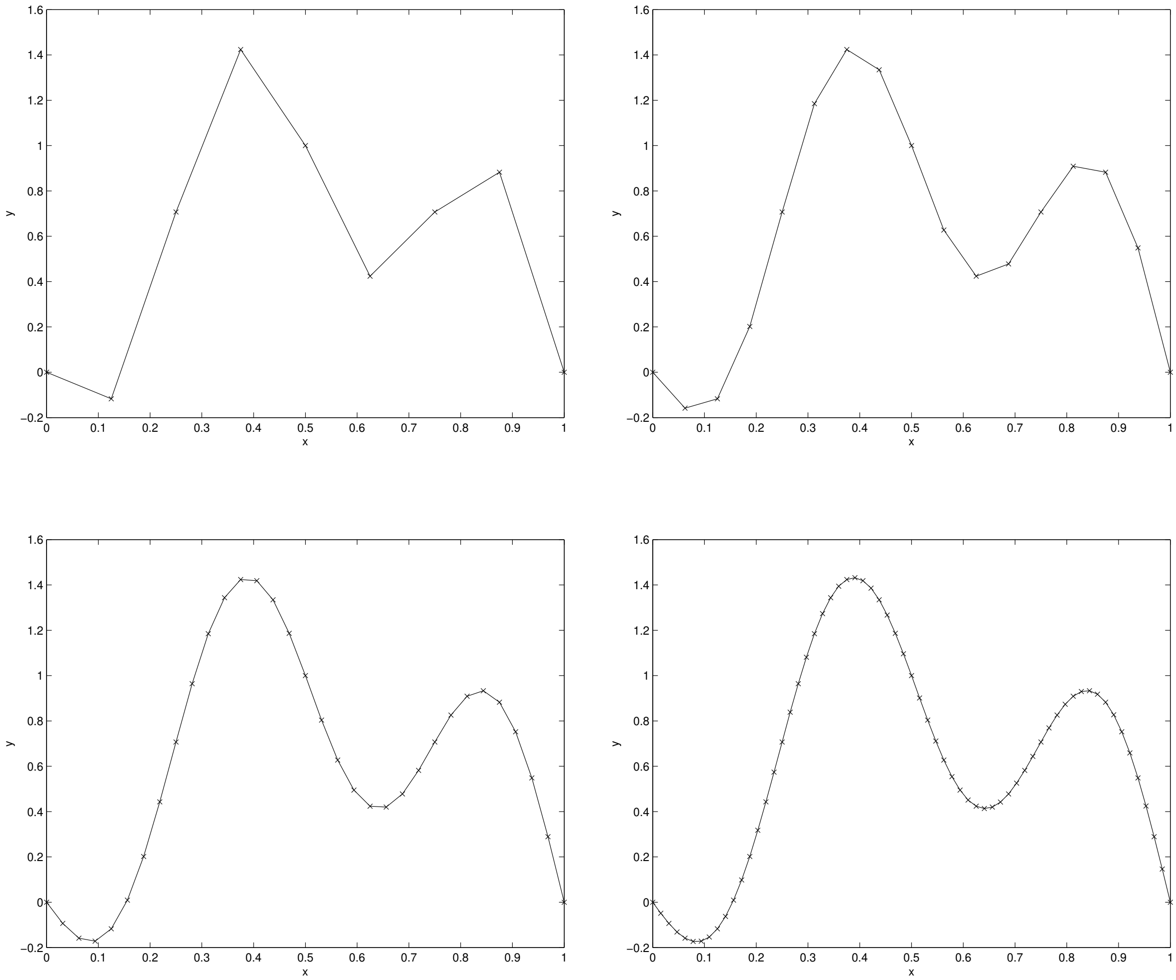


Figure 3.4: Graphs of piecewise linear functions. The graphs are in fact the piecewise linear interpolant of $y = \sin(\pi x) - \sin(4\pi x)/2$, for $0 \le x \le 1$, using respectively 9, 17, 33 and 65 equally spaced points.

3.3 The spaces $C(\Omega)$, $C^1(\Omega)$, $C^2(\Omega)$ etc. for $\Omega \subset \mathbb{R}^2$

When $\Omega \subset \mathbb{R}^2$ we similarly use the notation

$$
\begin{aligned}
C(\Omega) &:= \{u: \quad u \text{ is continuous on } \Omega\} && (3.21)\\
C(\bar{\Omega}) &:= \{u: \quad u \text{ is continuous on } \bar{\Omega}\} && (3.22)\\
C^1(\Omega) &:= \{u: \quad u,\ \partial u/\partial x \text{ and } \partial u/\partial y \text{ are continuous on } \Omega\} && (3.23)\\
C^1(\bar{\Omega}) &:= \{u: \quad u,\ \partial u/\partial x \text{ and } \partial u/\partial y \text{ are continuous on } \bar{\Omega}\} && (3.24)\\
C^2(\Omega) &:= \{u: \quad u,\ \partial u/\partial x,\ \partial u/\partial y,\ \partial^2 u/\partial x^2,\ \partial^2 u/\partial x\partial y, \text{ and } \partial^2 u/\partial y^2,\\
&\qquad\qquad \text{are continuous on } \Omega\} && (3.25)
\end{aligned}
$$

etc.

A problem in $\mathbb{R}^2$ that we will consider is Poisson's equation

$$-\Delta u = \nabla^2 u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f, \quad (x, y)^T \in \Omega \subset \mathbb{R}^2, \quad u(x, y) = 0 \text{ on } \partial\Omega. \qquad (3.26)$$

With $f \in C(\Omega)$ this involves seeking a function $u \in C^2(\Omega) \cap C(\bar{\Omega})$. In Ω the function u is twice continuously differentiable in order to satisfy the PDE and it is continuous up to the boundary in order to take the specified boundary conditions. We will show how this problem can be reformulated in such a way that we can construct an approximation $U \in C(\bar{\Omega})$.

3.4 Lagrange polynomial interpolation in one dimension

How well the exact solution u can be approximated by a piecewise polynomial function U depends on how well functions can be approximated by polynomials. This can be answered by considering how closely a given function u can be interpolated by a polynomial. If $x_0 < x_1 < \cdots < x_{p-1} < x_p$ denote $p+1$ distinct points then an interpolating polynomial can be constructed after first constructing the Lagrange polynomials

$$l_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{p} \left(\frac{x - x_k}{x_i - x_k} \right) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_p)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_p)} \tag{3.27}$$

for $i = 0, 1, \cdots, p$. Each function $l_i(x)$ is a polynomial of degree p and the collection of these $p+1$ functions form a basis for polynomials of degree $\le p$ in the same way as $1, x, \cdots, x^p$ is a basis for polynomials of degree $\le p$. By construction these functions have the property

$$l_i(x_j) = \begin{cases} 1, & j = i, \\ 0, & j \neq i. \end{cases} \tag{3.28}$$

The interpolating polynomial u_I of degree $\le p$ which agrees with u at the $p+1$ points $x_0 < x_1 < \cdots < x_p$ is given by

$$u_I(x) = \sum_{i=0}^{p} u(x_i) l_i(x). \tag{3.29}$$

The error in this approximation is

$$e(x) = u(x) - u_I(x) \tag{3.30}$$

is hence 0 at the $p+1$ points $x_0, \cdots, x_p$ and we next show that we can express $e(x)$ in the form

$$e(x) = (\text{some function of } x) w(x), \quad \text{where } w(x) = (x-x_0)(x-x_1)\cdots(x-x_p). \tag{3.31}$$

This can be done as follows.

Let $\bar{x} \neq x_i$ for $i = 0, 1, \cdots, p$ so as to ensure that $w(\bar{x}) \neq 0$. Then we can define

$$g(x) = e(x) - e(\bar{x}) \frac{w(x)}{w(\bar{x})} \tag{3.32}$$

which by construction has a zero at $\bar{x}$ as well as at the $p+1$ points $x_0, \cdots, x_p$, i.e. it is zero at at least $p+2$ distinct points which we choose to label as $t_0 < t_1 < \cdots < t_{p+1}$. Now Rolle's theorem states that if f is differentiable in an interval (α, β) with $f(\alpha) = f(\beta)$ then $f'(\gamma) = 0$ at some point γ between α and β. Applying this theorem to g in each of the $p+1$ intervals (t_0, t_1), (t_1, t_2), $\cdots$, (t_p, t_{p+1}) leads us to conclude that g' has at least $p+1$ distinct zeros in $[t_0, t_{p+1}] = [x_0, x_p]$. Similar reasoning applied to g' leads us to conclude that g'' has at least p

distinct zeros, g''' has at least $p-1$ distinct zeros, etc. with in particular $g^{(p+1)}$ having at least one zero which we label as $\xi = \xi(\bar{x})$. Now the $(p+1)$th derivative of the polynomial $u_I(x)$, which is only of degree $\le p$, is 0 and hence $e^{(p+1)} = u^{(p+1)}$. Further

$$w(x) = (x - x_0)\cdots(x - x_p) = x^{p+1} + (\text{polynomial of degree} \le p) \tag{3.33}$$

and thus

$$w^{(p+1)} = (p+1)!. \tag{3.34}$$

Thus

$$0 = g^{(p+1)}(\xi) = u^{(p+1)}(\xi) - (p+1)!\frac{e(\bar{x})}{w(\bar{x})} \quad \text{giving} \quad e(\bar{x}) = \frac{u^{(p+1)}(\xi(\bar{x}))}{(p+1)!}w(\bar{x}). \tag{3.35}$$

Replacing $\bar{x}$ by x gives us our result

$$e(x) = \frac{u^{(p+1)}(\xi(x))}{(p+1)!}w(x). \tag{3.36}$$

In the case $p = 1$ corresponding to $u_I(x)$ being the linear interpolant we have

$$e(x) = u(x) - u_I(x) = \frac{u''(\xi)}{2}(x - x_0)(x - x_1) \tag{3.37}$$

where $\xi = \xi(x)$ is some point between x_0 and x_1. To bound this quantity let

$$M_2 = \max_{t\in[x_0,x_1]} |u''(t)| \quad \text{and let} \quad h = x_1 - x_0. \tag{3.38}$$

Now with $s = x - x_0$ and by noting that $x - x_1 = x - (x_0 + h) = s - h$ we have

$$(x - x_0)(x - x_1) = s(s - h) = s^2 - hs = (s - h/2)^2 - h^2/4 \ge -h^2/4. \tag{3.39}$$

On $[x_0, x_1]$ this lies between 0 and $-h^2/4$ and in magnitude $(x - x_0)(x - x_1)$ cannot exceed $h^2/4$ and thus an interpolation error bound for linear polynomials is

$$|u(x) - u_I(x)| \le \frac{M_2}{8}h^2. \tag{3.40}$$

In the case of polynomials of degree p and with

$$M_{p+1} = \max_{[x_0,x_p]} |u^{(p+1)}(t)| \tag{3.41}$$

we similarly obtain a result of the form

$$|u(x) - u_I(x)| \le (\text{const.})M_{p+1}h^{p+1}. \tag{3.42}$$

Also by modifying the proof that led to (3.36) it can be shown that the error in the derivatives satisfy a result of the form

$$|u^{(j)}(x) - u_I^{(j)}(x)| \le (\text{const.})M_{p+1}h^{p+1-j} \tag{3.43}$$

for $j = 0, 1, \cdots, p+1$. Thus when h is small increasing the degree of the polynomial will generally give a smaller error (provided u is sufficiently differentiable) and, as shown, each successive derivative gives us one power of h less accuracy.

Similar types of error bounds regarding the powers of h apply for the interpolation of functions in $\mathbb{R}^2$ but these are harder to obtain.

3.5 Differentiation and integration in $\mathbb{R}^2$

To solve a differential equation exactly, when this is possible, we usually have to integrate at some point and to reformulate our ODEs and PDEs in a way which requires weaker continuity requirements we also have to integrate. In order to keep this subsection reasonably short we just state the results concerning reformulating integrals in $\mathbb{R}^2$ (specifically the divergence theorem and changing variables in a double integral) as the details for these will be covered, when needed, later in the module.

For a function $f : \mathbb{R} \to \mathbb{R}$ of one space variable the derivative $f'(x)$ gives the slope of the curve of $y = f(x)$ at x and the integral of f from c to d ($c < d$), which we write as

$$\int_c^d f(x)\,\mathrm{d}x, \tag{3.44}$$

gives the area under the curve of f from c to d. In addition, one form of the fundamental theorem of calculus which links differentiation and integration is that

$$\int_c^d f'(x)\,\mathrm{d}x = [f(x)]_c^d = f(d) - f(c). \tag{3.45}$$

We now consider the corresponding situations when instead $f : \mathbb{R}^2 \to \mathbb{R}$, i.e. when f is a function of two space variables.

3.5.1 The gradient vector and directional derivatives

The first thing to note is that we now need partial derivatives instead of ordinary derivatives. We have

$$\frac{\partial f}{\partial x}(x,y) = \lim_{h\to 0}\frac{f(x+h,y) - f(x,y)}{h}, \tag{3.46}$$

$$\frac{\partial f}{\partial y}(x,y) = \lim_{h\to 0}\frac{f(x,y+h) - f(x,y)}{h}. \tag{3.47}$$

These partial derivatives give the slopes on the surface $z = f(x,y)$ in the x and y directions respectively. It is convenient to collect these two quantities together as a vector known as the gradient vector

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}. \tag{3.48}$$

We can similarly consider how f changes in any direction $\underline{n} = (n_1, n_2)^T$. If $\underline{n}$ is of unit length then this directional derivative is defined by

$$\frac{\partial f}{\partial n}(x,y) = \lim_{h\to 0}\frac{f(x+hn_1, y+hn_2) - f(x,y)}{h}. \tag{3.49}$$

Now

$$\frac{f(x+hn_1,y+hn_2) - f(x,y)}{h} = \frac{f(x+hn_1,y+hn_2) - f(x,y+hn_2)}{h} \tag{3.50}$$

$$+\frac{f(x,y+hn_2) - f(x,y)}{h}. \tag{3.51}$$

Now by the mean value theorem and the continuity of $\partial f/\partial x$ we have

$$\frac{f(x+hn_1,y+hn_2)-f(x,y+hn_2)}{h} = n_1\frac{\partial f}{\partial x}(x+\theta_1 h, y+hn_2) \to n_1\frac{\partial f}{\partial x}(x,y) \quad \text{as } h\to 0. \tag{3.52}$$

In the above line $0<\theta_1<1$. Similarly by the definition of the partial derivative

$$\frac{f(x,y+hn_2)-f(x,y)}{h} \to n_2\frac{\partial f}{\partial y}(x,y) \quad \text{as } h \to 0. \tag{3.53}$$

The result that we get is the chain rule of partial differentiation. In vector notation with $\underline{x}=(x,y)^T$ we have

$$\frac{\partial f}{\partial n}(\underline{x}) = n_1\frac{\partial f}{\partial x}(x,y)+n_2\frac{\partial f}{\partial y}(x,y) = \underline{n}\cdot\nabla f(\underline{x}) \tag{3.54}$$

where here $\cdot$ means the usual scalar or dot product of two vectors. We will also use the notation $\underline{n}^T\nabla u$, when convenient, to denote this operation.

3.5.2 Double integrals, i.e. area integrals

In the case of integration, the integral of f over a two dimensional region Ω, which we write as

$$\int_\Omega f(\underline{x})\,\mathrm{d}a \quad \text{or as} \quad \iint_\Omega f(\underline{x})\,\mathrm{d}x\mathrm{d}y, \tag{3.55}$$

gives the volume under the surface $z=f(\underline{x})=f(x,y)$ in the part $\underline{x}\in\Omega$. When the single integral notation is used the increment $\mathrm{d}a=\mathrm{d}x\mathrm{d}y$ indicates an area increment.

3.5.3 Contour integrals in $\mathbb{R}^2$

With functions $f(\underline{x})$ we can also consider contour integrals in the plane $\mathbb{R}^2$. If Γ is a curve in $\mathbb{R}^2$ which has a parametric representation

$$\Gamma = \{\underline{x}(t): \quad 0\le t\le L\} \tag{3.56}$$

then an incremental change dt in t gives a corresponding change $d\underline{x}=\underline{x}'(t)dt$ along the curve (i.e. tangential to the curve) which has length $ds=|\underline{x}'(t)dt|$. The (contour) integral of f along Γ is defined by

$$\int_\Gamma f(\underline{x})\,\mathrm{d}s = \int_0^L f(\underline{x}(t))|\underline{x}'(t)|\,\mathrm{d}t. \tag{3.57}$$

Note that the value of the integral along Γ here does not depends on the direction at which we go along Γ. This is in contrast to the integral

$$\int_L^0 g(x)\,\mathrm{d}x = -\int_0^L g(x)\,\mathrm{d}x \tag{3.58}$$

which has a slightly different meaning. In this latter case integrating g from L to 0 corresponds to integrating $-g$ from 0 to L. The direction aspect is taken care of within the expression in the integrand. This is the case in situations where we have contour integrals like (3.57) in the divergence theorem.

3.5.4 The divergence theorem

We now state the result concerning what happens when we integrate a partial derivative to get the two dimensional version of (3.45). This will be most conveniently expressed in what is known as the divergence theorem.

Let $\partial\Omega$ denote the boundary of $\Omega \subset \mathbb{R}^2$ taken in the anti-clockwise direction and let $\underline{n} = (n_1, n_2)^T$ denote the unit outward normal to $\partial\Omega$. It can be shown that

$$\iint_\Omega \frac{\partial f}{\partial x}(x,y)\, \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} f(\underline{x}) n_1 \,\mathrm{d}s, \tag{3.59}$$

$$\iint_\Omega \frac{\partial f}{\partial y}(x,y)\, \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} f(\underline{x}) n_2 \,\mathrm{d}s. \tag{3.60}$$

If $\underline{A} = (A_1, A_2)^T$ then these two relations imply that

$$\iint_\Omega \left(\frac{\partial A_1}{\partial x}(x,y) + \frac{\partial A_2}{\partial y}(x,y) \right) \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} A_1(\underline{x})n_1 + A_2(\underline{x})n_2 \,\mathrm{d}s \tag{3.61}$$

which in vector form is

$$\iint_\Omega \nabla \cdot \underline{A} \;\mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \underline{A} \cdot \underline{n} \,\mathrm{d}s \tag{3.62}$$

where

$$\nabla \cdot \underline{A} = \frac{\partial A_1}{\partial x} + \frac{\partial A_2}{\partial y} \quad \text{and} \quad \underline{A} \cdot \underline{n} = A_1 n_1 + A_2 n_2. \tag{3.63}$$

This results is known as the *divergence theorem* or as the *Gauss-divergence theorem*. The quantity $\nabla \cdot \underline{A}$ is known as the divergence of $\underline{A}$.

3.5.5 Some vector identities involving ∇

If you have previously studied courses on vector calculus then you will have met many identities involving ∇ in the context of the gradient, divergence and curl of various quantities. We only need one of the simpler of these relations here which is like the rule for differentiating a product in one dimension. Recall that this result is

$$(uv)' = uv' + u'v. \tag{3.64}$$

Now in two dimension with $v = v(x,y)$ being a scalar valued function and $\underline{A} = (A_1, A_2)^T$ being a vector valued function we have that $v\underline{A} = (vA_1, vA_2)^T$ and

$$\begin{aligned} \nabla \cdot (v\underline{A}) &= \frac{\partial (vA_1)}{\partial x} + \frac{\partial (vA_2)}{\partial y} && (3.65)\\ &= v\frac{\partial A_1}{\partial x} + \frac{\partial v}{\partial x}A_1 + v\frac{\partial A_2}{\partial y} + \frac{\partial v}{\partial y}A_2 && (3.66)\\ &= v\nabla \cdot \underline{A} + \nabla v \cdot \underline{A}. && (3.67) \end{aligned}$$

In the case that the vector $\underline{A}$ is the gradient of some function, i.e. $\underline{A} = \nabla u$ we have

$$\Delta u = \nabla^2 u = \nabla \cdot \nabla u = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{3.68}$$

which is know as the *Laplacian* of u. Thus using the previous identity and the divergence theorem we have

$$\nabla \cdot (v\nabla u) = v\Delta u + \nabla v \cdot \nabla u \tag{3.69}$$

and

$$\int_{\partial\Omega} v\nabla u \cdot \underline{n}\ \mathrm{d}s = \iint_{\Omega} \nabla \cdot (v\nabla u)\ \mathrm{d}x\mathrm{d}y = \iint_{\Omega} (v\Delta u + \nabla u \cdot \nabla v)\ \mathrm{d}x\mathrm{d}y. \tag{3.70}$$

Since

$$\nabla u \cdot \underline{n} = \frac{\partial u}{\partial n} \tag{3.71}$$

this can be written as

$$\iint_{\Omega} (v\Delta u + \nabla u \cdot \nabla v)\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} v\frac{\partial u}{\partial n}\ \mathrm{d}s. \tag{3.72}$$

If $v = 0$ on $\partial\Omega$ then the contour integral is 0 as the integrand is 0 and this reduces to

$$-\iint_{\Omega} v\Delta u\ \mathrm{d}x\mathrm{d}y = \iint_{\Omega} \nabla u \cdot \nabla v\ \mathrm{d}x\mathrm{d}y. \tag{3.73}$$

This is the key step in reformulating Poisson's equation

$$-\Delta u = f \tag{3.74}$$

as then we obtain

$$\iint_{\Omega} \nabla u \cdot \nabla v\ \mathrm{d}x\mathrm{d}y = \iint_{\Omega} fv\ \mathrm{d}x\mathrm{d}y \tag{3.75}$$

for all suitable functions v which vanish on the boundary $\partial\Omega$. This is the expression that we will need in the weak form of Poisson's equation.

3.5.6 Transforming a double integral

To complete this review of calculus in $\mathbb{R}^2$ we consider now how to transform an area integral which we will need to do later in the module when we map from a standard element to a general element. The corresponding situation in one dimension is when we make a substitution. For example, if $x : [0,1] \to \mathbb{R}$ is differentiable then

$$\int_{x(0)}^{x(1)} f(x)\,\mathrm{d}x = \int_0^1 f(x(t))x'(t)\,\mathrm{d}t. \tag{3.76}$$

The range of integration is $[x(0), x(1)]$ with respect to x and it is $[0,1]$ with respect to t with the adjustment in the integrand with the $x'(t)$ term. In $\mathbb{R}^2$ the corresponding situation involves a one-to-one and onto mapping $\underline{x} : T \to \tilde{T}$. The Jacobian matrix of the transformation is given by

$$\begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix} \tag{3.77}$$

with the determinant of the Jacobian being

$$J = \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s}. \tag{3.78}$$

The two-dimensional version of (3.76) is

$$\iint_{\tilde{T}} f(x,y)\,\mathrm{d}x\mathrm{d}y = \iint_{T} f(x(s,t),y(s,t))|J|\,\mathrm{d}s\mathrm{d}t. \tag{3.79}$$

Note that the determinant J defined here can be positive or negative and it is the absolute value $|J|$ which is needed in (3.79).

3.6 Summary

To summarize these preliminary results we have covered the following.

1. We have introduced the notation $C(\Omega)$, $C^1(\Omega)$, $C^2(\Omega)$ etc. for spaces of continuous functions.

2. We have discussed Lagrange polynomial interpolation in one dimension. A key result here is that if the interpolating points are contained in a region of size h then the error in the interpolating polynomial of degree p is $\mathcal{O}(h^{p+1})$ in magnitude. The error in the jth derivative is $\mathcal{O}(h^{p+1-j})$.

3. In $\mathbb{R}^2$ we introduced the gradient vector and the directional derivative

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}, \quad \frac{\partial f}{\partial n}(\underline{x}) = n_1 \frac{\partial f}{\partial x}(x,y) + n_2 \frac{\partial f}{\partial y}(x,y) = \underline{n} \cdot \nabla f(\underline{x}). \tag{3.80}$$

We also introduced the divergence operation and the divergence theorem

$$\nabla \cdot \underline{A} = \frac{\partial A_1}{\partial x} + \frac{\partial A_2}{\partial y} \quad \text{and} \quad \iint_{\Omega} \nabla \cdot \underline{A}\,\mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \underline{A} \cdot \underline{n}\,\mathrm{d}s \tag{3.81}$$

and we have noted the vector identity

$$\nabla \cdot (v\underline{A}) = v\nabla \cdot \underline{A} + \nabla v \cdot \underline{A}. \tag{3.82}$$

The Laplace operator in vector notation is

$$\Delta u = \nabla^2 u = \nabla \cdot \nabla u = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{3.83}$$

Integrals in $\mathbb{R}^2$ involving $\underline{x} : T \to \tilde{T}$ transform according to

$$\iint_{\tilde{T}} f(x,y)\,\mathrm{d}x\mathrm{d}y = \iint_{T} f(x(s,t),y(s,t))|J|\,\mathrm{d}s\mathrm{d}t. \tag{3.84}$$

Here J is the determinant of the Jacobian matrix

$$\begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix} \tag{3.85}$$

of the transformation $(s,t) \to (x,y)$.

4 Reformulating ODEs into a weaker form

4.1 Introduction

In this section we consider the second order ordinary differential equation (ODE)

$$-u'' = f \quad \text{and more generally} \quad -(pu')' + qu = f, \quad 0 < x < 1 \tag{4.1}$$

together with appropriate boundary conditions and we consider the fourth order ODE

$$u'''' = f \quad \text{and more generally} \quad (pu'')'' = f \tag{4.2}$$

together with appropriate boundary conditions. In both cases the right hand side function f and the other functions p and q are suitable functions of x by which we mean that they are sufficiently differentiable and bounded on $[0,1]$. Other specific conditions that we will need will be specified later. All the ODEs given are linear. That is in each case we can write the equation in the form

$$L(u) = f \quad \text{(which we will often abbreviate to } Lu = f) \tag{4.3}$$

where L is a linear differential operator with specifically

$$L(u) = \begin{cases} -u'' & \text{or } -(pu')' + qu \qquad \text{in (4.1)} \\ u'''' & \text{or } (pu'')'' \qquad\qquad\;\; \text{in (4.2)} \end{cases}. \tag{4.4}$$

In we let D_L denote the domain of the linear differential operator L then

$$D_L \subset C^2[0,1] \quad \text{for (4.1)} \quad \text{and} \qquad D_L \subset C^4[0,1] \quad \text{for (4.2)} \tag{4.5}$$

and the term linear means that

$$L(\alpha u) = \alpha L(u) \quad \text{for all } \alpha \in \mathbb{R} \text{ and for all } u \in D_L \tag{4.6}$$

and

$$L(u_1 + u_2) = L(u_1) + L(u_2) \quad \text{for all } u_1, u_2 \in D_L. \tag{4.7}$$

(For conveninece of the notation here we assume smoothness up to the edge points $x = 0$ and $x = 1$ although in practice we can usually replace a space such as $C^2[0,1]$ by $C[0,1] \cap C^2(0,1)$.) Despite the ODEs being linear it is generally not possible to express the solution in closed form in terms of standard functions and thus it is necessary to use a numerical method. The finite element method (FEM) is one of the methods that can be used here which involves solving a related weak or minimization problem and hence we first consider how to reformulate the ODE into these forms. As we will see, we get the weak problem by multiplying (4.3) by a suitable function v and by integrating over $(0,1)$, i.e. if u satisfies (4.3) then

$$\int_0^1 L(u)v \, \mathrm{d}x = \int_0^1 fv \, \mathrm{d}x. \tag{4.8}$$

We then use integration by parts to reexpress the left hand side into a form which involves lower derivatives on u than does Lu.

4.2 A remark about integration and continuity

Before we consider the weak forms for the problems (4.1) and (4.2) we consider what collection of functions v we need for (4.8) to hold to imply that $Lu = f$ for $0 < x < 1$. Now we have selected here a domain D_L in each case such that $r = Lu - f$ is continuous as a function on $(0,1)$ and thus the question concerns when does

$$\int_0^1 r(x)v(x)\,\mathrm{d}x = 0 \quad \text{for all appropriate } v \qquad \text{imply} \quad r(x) = 0, \quad 0 < x < 1? \tag{4.9}$$

Clearly, the larger the collection of functions v that we take the more likely this is to true and if the result holds for one set of functions then it holds for any larger set of functions containing this set. It is sufficient to consider any space of the form

$$V = \{v \in C^q[0,1] : \quad v(0) = v'(0) = \cdots = v^{(q)}(0) = v(1) = v'(1) = \cdots = v^{(q)}(1) = 0\} \tag{4.10}$$

for any $q \ge 0$. To show this we use a short proof by contradiction. Thus we start by assuming that $r(x) \ne 0$ at some point $x = c$. Since r is continuous at c it is non-zero in a neighbourhood of c. If we assume that $r(c) > 0$ then a consequence of the ϵ-δ definition of continuity is that there exists a $\delta > 0$ such that

$$r(x) > \frac{r(c)}{2} > 0 \quad \text{whenever} \quad |x - c| < \delta. \tag{4.11}$$

We can select $v \in V$ to be zero outside of this neighbourhood and positive inside it to get a contradiction, for example we can take

$$v(x) = \begin{cases} 0 & \text{if } |x-c| \ge \delta, \\ ((c+\delta) - x)^{q+1}(x - (c-\delta))^{q+1} & \text{if } |x-c| < \delta. \end{cases} \tag{4.12}$$

Note that by construction $v^{(k)}(x) = 0$ for $k = 0, 1, \cdots, q$ at the join points $x = c - \delta$ and $x = c + \delta$ to ensure that $v \in V$. With this choice of v we have

$$\int_0^1 r(x)v(x)\,\mathrm{d}x = \int_{c-\delta}^{c+\delta} r(x)v(x)\,\mathrm{d}x \ge \frac{r(c)}{2}\int_{c-\delta}^{c+\delta} v(x)\,\mathrm{d}x > 0. \tag{4.13}$$

This is our contradiction as we are contradicting the requirement that we should get 0 for all such v.

The previous proof shows that we can always take functions v which vanish at the edge points to establish that $Lu = f$ in $(0,1)$ and that we can restrict to functions v which have as many derivatives as we need. Obtaining a function u satisfying $Lu = f$ is however not the complete task as we also need the particular function which satisfies the boundary conditions and, as is shown below, this will lead us to also use functions v which are non-zero at the edge points in the case when we have natural boundary conditions.

4.3 Reformulating the problem $-u'' = f$ into weak form

To have a unique solution we need two linearly independent boundary conditions and typical ones are as follows.

(i) $u(0) = \alpha$, $u(1) = \beta$.

(ii) $u(0) = \alpha$, $u'(1) + \beta u(1) = g$.

In the first case u is known at both $x = 0$ and $x = 1$ (these are known as Dirichlet boundary conditions) whilst in the second case $u(1)$ is not known, we only have an equation connecting $u(1)$ and $u'(1)$. This is a mixed boundary condition. In fact in this case there needs to be some restriction on β to ensure that the problem has a unique solution and we will consider this later.

It is convenient in our description to have, when possible, homogeneous conditions and this can be achieved easily in both cases.

In the case (i) we can define

$$u_0 = \alpha + (\beta - \alpha)x \quad \text{so that} \quad u_0(0) = \alpha, \quad u_0(1) = \beta. \tag{4.14}$$

Then the function u that we seek can be written in the form

$$u = u_0 + w \quad \text{where } w \text{ satisfies} \quad w(0) = w(1) = 0 \tag{4.15}$$

and

$$Lu = f \quad \text{implies} \quad Lw = f_0 = f - Lu_0. \tag{4.16}$$

In fact with our choice of u_0 we have $Lu_0 = 0$ and $f_0 = f$, i.e. w satisfies the same ODE as u but with homogeneous boundary conditions.

In the case of (ii) we can similarly give a linear function u_0 which satisfies both boundary conditions. With

$$u_0(x) = \alpha + \gamma x, \quad g = u_0'(1) + \beta u_0(1) = \gamma + \beta(\alpha + \gamma) \quad \text{if} \quad \gamma = \frac{g - \alpha\beta}{1 + \beta} \tag{4.17}$$

provided $\beta \neq -1$. The function u that we seek is of the form

$$u = u_0 + w \quad \text{where} \quad w(0) = 0, \quad w'(1) + \beta w(1) = 0 \tag{4.18}$$

where w satisfies

$$Lw = f_0 = f - Lu_0 = f \quad (\text{as } Lu_0 = 0). \tag{4.19}$$

That is, as before, w satisfies the same ODE as u but with homogeneous boundary conditions.

From the previous constructions we thus do not lose much generality by supposing at the outset that the boundary conditions are of the form

(i) $u(0) = 0$, $u(1) = 0$.

(ii) $u(0) = 0$, $u'(1) + \beta u(1) = 0$.

For the case (i) we define the space

$$D_L = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\}. \tag{4.20}$$

The ODE hence involves finding $u \in D_L$ such that $Lu = f$. Following the previous discussion we can restrict $v \in D_L$. Then multiplying by v and integrating gives

$$\begin{aligned} 0 &= \int_0^1 (Lu - f)v \,\mathrm{d}x = \int_0^1 -u''v - fv \,\mathrm{d}x & (4.21)\\ &= \int_0^1 u'v' - fv \,\mathrm{d}x - [u'v]_0^1 & (4.22)\\ &= \int_0^1 u'v' - fv \,\mathrm{d}x & (4.23) \end{aligned}$$

because $v(0) = v(1) = 0$. A weaker problem is hence as follows:

$$\text{Find } u \in D_L \quad \text{such that} \quad \int_0^1 u'v' \, \mathrm{d}x = \int_0^1 fv \, \mathrm{d}x \quad \text{for all } v \in D_L. \tag{4.24}$$

(The term variational problem is also used. We will use the term weak problem when we consider problems such as this over larger spaces than D_L which only need derivatives of u and v to be integrable.)

For the case (ii) things are slightly more complicated due to the boundary condition $u'(1)+\beta u(1) = 0$ although, as we will show, this can actually be dealt with quite easily. For this problem we define the two nested spaces

$$D_L = \{v \in C^2[0,1] : \quad v(0) = 0,\ v'(1) + \beta v(1) = 0\} \tag{4.25}$$

and the larger space

$$\tilde{D}_L = \{v \in C^2[0,1] : \quad v(0) = 0\}. \tag{4.26}$$

The ODE involves finding $u \in D_L$ satisfying $Lu = f$. Proceeding as before we multiply the ODE by $v \in D_L$ and integrate by parts giving

$$\begin{aligned} 0 &= \int_0^1 (Lu - f)v \, \mathrm{d}x = \int_0^1 -u''v - fv \, \mathrm{d}x & (4.27)\\ &= \int_0^1 u'v' - fv \, \mathrm{d}x - [u'v]_0^1 & (4.28)\\ &= \int_0^1 u'v' - fv \, \mathrm{d}x - u'(1)v(1) & (4.29)\\ &= \int_0^1 u'v' - fv \, \mathrm{d}x + \beta u(1)v(1), \quad \text{using the boundary condition.} & (4.30) \end{aligned}$$

One version of the weaker problem is as follows:

$$\text{Find } u \in D_L \quad \text{such that} \quad \int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1) = \int_0^1 fv \, \mathrm{d}x \quad \text{for all } v \in D_L. \tag{4.31}$$

Notice that to get this expression we have used the boundary condition at $x = 1$ but also, as given, the boundary condition is part of the definition of the space D_L. If we consider a function $u \in D_L \subset \tilde{D}_L$ and any $v \in \tilde{D}_L$ then following the same procedure we similarly get

$$0 = \int_0^1 u'v' - fv \, \mathrm{d}x + \beta u(1)v(1). \tag{4.32}$$

Hence we can also consider the following weaker problem.

$$\text{Find } u \in \tilde{D}_L \quad \text{such that} \quad \int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1) = \int_0^1 fv \, \mathrm{d}x \quad \text{for all } v \in \tilde{D}_L. \tag{4.33}$$

The function that we wish to find satisfies this but as now as there is no condition on functions in $\tilde{D}_L$ at $x = 1$ we need to consider whether a solution of (4.33) must necessarily satisfy the boundary condition $u'(1) + \beta u(1) = 0$. The answer to this is yes which we can verify by reversing the procedure using integration by parts as follows. Since for $u, v \in \tilde{D}_L$

$$\int_0^1 u'v' \, \mathrm{d}x = [u'v]_0^1 - \int_0^1 u''v \, \mathrm{d}x = u'(1)v(1) - \int_0^1 u''v \, \mathrm{d}x \tag{4.34}$$

(4.33) gives

$$\int_0^1 (-u'' - f)v \, \mathrm{d}x + (u'(1) + \beta u(1))v(1) = 0. \tag{4.35}$$

This must be satisfied for all $v \in \tilde{D}_L$. Now we have already shown that by just considering all functions $v \in \tilde{D}_L$ which also satisfy $v(1) = 0$ is sufficient to ensure that

$$-u'' - f = 0, \quad 0 < x < 1. \tag{4.36}$$

(4.35) then reduces to

$$(u'(1) + \beta u(1))v(1) = 0 \tag{4.37}$$

and as this must hold for all $v(1) \neq 0$ we must have

$$u'(1) + \beta u(1) = 0. \tag{4.38}$$

To summarize, we have shown that solving the weaker problem (4.33) over the space $\tilde{D}_L$ which contains no boundary condition at $x = 1$ gives a solution u satisfying the boundary condition $u'(1) + \beta u(1) = 0$. This boundary condition is called a **natural boundary condition** of the weaker problem (4.33), i.e. it is a boundary condition that we do not need to impose on the function space but is instead the condition that the solution satisfies at that part of the boundary. For this problem the condition $v(0) = 0$ as part of the specification of $\tilde{D}_L$ at $x = 0$ is called an **essential boundary condition**.

As some final remarks, it is worth noting that the particular case $\beta = 0$ corresponds to $u'(1) = 0$ which is known as the **homogeneous Neumann boundary condition** and the weaker problem is as follows:

$$\text{Find } u \in \tilde{D}_L \quad \text{such that} \quad \int_0^1 u'v' \, \mathrm{d}x = \int_0^1 fv \, \mathrm{d}x \quad \text{for all } v \in \tilde{D}_L. \tag{4.39}$$

This looks the same as in (4.24) except that the spaces involved are different. In (4.24) the functions in D_L are 0 at $x = 1$ whereas here in $\tilde{D}_L$ there is no condition at $x = 1$.

As we will show later, the FEM can cope easily with natural boundary conditions as such conditions need not be imposed on the space of approximating functions used. The FEM generates an approximation which hence does not generally satisfy such boundary conditions but instead a sequence of converging approximations (which we get by refining the mesh used) more and more closely satisfy such conditions.

4.4 Some definitions

It is perhaps useful at this point to define a few terms which apply to the problem $-u'' = f$ just considered and which apply to most of the problems tackled in this module.

1. To repeat a definition already given, a differential operator L is said to be **linear** on a space V if

$$L(u_1 + u_2) = Lu_1 + Lu_2 \quad \text{and} \quad L(\alpha u) = \alpha Lu \quad \text{for all } u_1, u_2 \in V \text{ and } \alpha \in \mathbb{R}. \tag{4.40}$$

   This is the case for $Lu = -u''$ on the space $V = C^2[0, 1]$.

2. In the previous derivation we have repeatedly seen expressions of the form

$$\int_0^1 f(x)v(x)\ \mathrm{d}x. \tag{4.41}$$

This is the $L_2(0,1)$ **inner product** of the functions f and v. To represent this quantity we will use the usual inner product notation

$$(f,v) := \int_0^1 f(x)v(x)\ \mathrm{d}x. \tag{4.42}$$

Technically, for something to be an inner product the following properties need to hold. For all $u,v,w \in V$ and for all $\alpha \in \mathbb{R}$

$$(u,v) = (v,u), \tag{4.43}$$
$$(\alpha u,v) = \alpha(u,v), \tag{4.44}$$
$$(u+v,w) = (u,w)+(v,w), \tag{4.45}$$
$$(u,u) \geq 0 \quad \text{with } (u,u)=0 \text{ if and only if } u=0. \tag{4.46}$$

All of these hold in this case. The norm associated with this inner product is defined by

$$\|v\| := (v,v)^{1/2}. \tag{4.47}$$

A space in which we have an inner product is known as an **inner product space**.

3. A mapping $a(.,.) : V \times V \to \mathbb{R}$ is said to be **bilinear** if it is linear in each of its arguments, i.e.

$$\begin{array}{ll} a(\alpha u,v) = \alpha a(u,v), & a(u_1+u_2,v) = a(u_1,v)+a(u_2,v), \\ a(u,\alpha v) = \alpha a(u,v), & a(u,v_1+v_2) = a(u,v_1)+a(u,v_2). \end{array} \tag{4.48}$$

We have a **symmetric bilinear form** if we also have $a(u,v) = a(v,u)$ for all $u,v \in V$. All the weaker forms given in the previous sections involve symmetric bilinear forms. We have had

$$a(u,v) = \int_0^1 u'v'\ \mathrm{d}x \tag{4.49}$$

and we have had

$$a(u,v) = \int_0^1 u'v'\ \mathrm{d}x + \beta u(1)v(1). \tag{4.50}$$

4. A bilinear form $a(.,.)$ is said to be **positive definite** on a space V if

$$a(v,v) \geq 0 \quad \text{for all } v \in V \text{ with } a(v,v) = 0 \text{ if and only if } v = 0. \tag{4.51}$$

To check whether this condition holds for the bilinear forms given in (4.49) and (4.50) we need to consider the space V and the parameter β in the case of (4.50). In the first case we have

$$V = D_L = \{v \in C^2[0,1] : \quad v(0)=v(1)=0\}, \quad a(v,v) = \int_0^1 v'^2\ \mathrm{d}x \geq 0. \tag{4.52}$$

As v'^2 is continuous and $v'(x)^2 \ge 0$, $a(v,v) = 0$ implies that $v'(x) = 0$, $0 \le x \le 1$ which in turn implies that v is constant on $[0,1]$. But as $v(0) = v(1) = 0$ this constant is 0 and we have shown that $a(v,v) = 0$ implies that $v = 0$.

In the second problem we have

$$V = \tilde{D}_L = \{v \in C^2[0,1] : \quad v(0) = 0\}, \quad a(v,v) = \int_0^1 v'^2 \, \mathrm{d}x + \beta v(1)^2. \tag{4.53}$$

Clearly if $\beta \ge 0$ we have $a(v,v) \ge 0$ with $a(v,v) = 0$ only if $v = 0$ as before. In fact it is positive definite for $\beta > -1$. To show this you need to use the Cauchy Schwarz inequality which is a result which is frequently used in the MA5352 only part of the module. For $v \in V$ we have

$$\begin{aligned} v(1) &= \int_0^1 v'(x) \, \mathrm{d}x, && (4.54)\\ v(1)^2 &= \left(\int_0^1 v' \, \mathrm{d}x\right)^2 && (4.55)\\ &\le \left(\int_0^1 v'^2 \, \mathrm{d}x\right)\left(\int_0^1 \, \mathrm{d}x\right) = \int_0^1 v'^2 \, \mathrm{d}x. && (4.56) \end{aligned}$$

Thus when $-1 < \beta \le 0$ we have $\beta = -|\beta|$ and

$$a(v,v) = \int_0^1 v'^2 \, \mathrm{d}x - |\beta| v(1)^2 \ge (1 - |\beta|) \int_0^1 v'^2 \, \mathrm{d}x \ge 0. \tag{4.57}$$

As before $a(v,v) = 0$ only occurs if $v = 0$. The bilinear form is not positive definite for $\beta \le -1$ on the space S. When $\beta = -1$ we have $u'(1) - u(1) = 0$ and the function $u(x) = x$ satisfies both boundary conditions with $Lu = 0$. The homogeneous problem hence has a non-trivial solution which is why things break down at this point.

5. A differential operator L is said to be **symmetric** on the space D_L if

$$(Lu, v) = (u, Lv). \tag{4.58}$$

This was the case with all the problems so far considered. In the case $Lu = -u''$ we had

$$(Lu, v) = \int_0^1 u'v' \, \mathrm{d}x \quad \text{when } D_L = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\} \tag{4.59}$$

and we had

$$\begin{aligned} (Lu, v) &= \int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1) && (4.60)\\ \text{when } D_L &= \{v \in C^2[0,1] : \quad v(0) = 0,\ v'(1) + \beta v(1) = 0\}. && (4.61) \end{aligned}$$

When we have a symmetric L we get a symmetric bilinear form

$$a(u, v) = (Lu, v). \tag{4.62}$$

Note that in the mixed boundary condition case the operator L is not however symmetric on the larger space $\tilde{D}_L = \{v \in C^2[0,1] : \quad v(0) = 0\}$ that we use.

6. The differential operator L is said to be **positive definite** on the space D_L if

$$(Lv, v) \geq 0 \quad \text{for all } v \in D_L \text{ with } \quad (Lv, v) = 0 \quad \text{if and only if } v = 0. \tag{4.63}$$

To summarize all these terms, we are dealing with an ODE which involves a linear, symmetric and positive definite operator L and in the reformulation of the ODE

$$Lu = f \tag{4.64}$$

into weaker form we obtain

$$a(u, v) = (f, v) \tag{4.65}$$

where $a(.,.)$ is a symmetric and positive definite bilinear form.

4.5 The equivalent minimization problem

Up to this point we have just shown how to reformulate an ODE into a weaker form and we have in fact obtained the expression for the actual weak problem which results when we replace spaces such as $C^2[0,1]$ by Sobolev spaces which we discuss later in the module. We now show that the weak problem is itself equivalent to a corresponding minimization problem involving the quadratic functional

$$I(v) = \frac{1}{2}a(v, v) - (f, v). \tag{4.66}$$

That is, given any inner product space V and given any symmetric and positive definite bilinear form $a(.,.)$ the following two problems are equivalent.

- Weak problem: Find $u \in V$ such that $a(u, v) = (f, v)$ for all $v \in V$.
- Minimization problem: Find $u \in V$ such that $I(u) \leq I(v)$ for all $v \in V$.

To show the equivalence suppose first that $u \in V$ solves the weak problem. Then for any $w \in V$ we have

$$\begin{aligned}
I(u+w) &= \frac{1}{2}a(u+w, u+w) - (f, u+w), && (4.67)\\
&= \frac{1}{2}(a(u,u) + a(u,w) + a(w,u) + a(w,w)) - ((f,u) + (f,w)), && (4.68)\\
&= \frac{1}{2}(a(u,u) + a(w,w)) - (f,u), \quad \text{using } a(w,u) = a(u,w) = (f,w), && (4.69)\\
&= I(u) + \frac{1}{2}a(w,w). && (4.70)
\end{aligned}$$

Now as $a(.,.)$ is positive definite we get $a(w, w) > 0$ for all $w \neq 0$. This shows that $u \in V$ uniquely minimises the quadratic $I(v)$ over the space V.

Suppose instead that $u \in V$ minimises the quadratic over the space V, let $v \in V$ with $v \neq 0$ and consider

$$\begin{aligned}
g(t) = I(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - (f, u+tv) && (4.71)\\
&= \frac{1}{2}a(v,v)t^2 + (a(u,v) - (f,v))t + \left(\frac{1}{2}a(u,u) - (f,u)\right). && (4.72)
\end{aligned}$$

This is a quadratic in t and since u minimises the functional, this quadratic in t must have a minimum at $t = 0$, i.e. we must have $g'(0) = 0$. Hence

$$g'(0) = a(u, v) - (f, v) = 0. \tag{4.73}$$

As $v \in V$ is arbitrary it follows that

$$a(u, v) = (f, v) \quad \text{for all } v \in V. \tag{4.74}$$

We now list the spaces D_L, the bilinear forms $a(.,.)$ and the weaker and minimization problems considered so far.

1. ODE: Find $u \in D_L = \{v \in C^2[0, 1] : \quad v(0) = v(1) = 0\}$ such that $Lu = -u'' = f$.

   Bilinear form:
   $$a(u, v) = \int_0^1 u'v' \, \mathrm{d}x. \tag{4.75}$$

   Weaker form: Find $u \in D_L$ such that
   $$\int_0^1 u'v' \, \mathrm{d}x = \int_0^1 fv \, \mathrm{d}x \quad \text{for all } v \in D_L. \tag{4.76}$$

   Minimization problem: Find $u \in D_L$ such that $I(u) \le I(v)$ for all $v \in D_L$ where
   $$I(v) = \frac{1}{2} \int_0^1 v'^2 - 2fv \, \mathrm{d}x. \tag{4.77}$$

2. ODE: Find $u \in D_L = \{v \in C^2[0, 1] : \quad v(0) = 0, \; v'(1) + \beta v(1) = 0\}$ such that $Lu = -u'' = f$.

   Bilinear form:
   $$a(u, v) = \int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1). \tag{4.78}$$

   Weaker form (not using the natural boundary condition):
   Find $u \in D_L$ such that
   $$\int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1) = \int_0^1 fv \, \mathrm{d}x \quad \text{for all } v \in D_L. \tag{4.79}$$

   Weaker form (using the natural boundary condition):
   Find $u \in \tilde{D}_L = \{v \in C^2[0, 1] : \quad v(0) = 0\}$ such that
   $$\int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1) = \int_0^1 fv \, \mathrm{d}x \quad \text{for all } v \in \tilde{D}_L. \tag{4.80}$$

   Minimization problem (not using the natural boundary condition):
   Find $u \in D_L$ such that $I(u) \le I(v)$ for all $v \in D_L$ where
   $$I(v) = \frac{1}{2} \int_0^1 (v'^2 - 2fv) \, \mathrm{d}x + \beta v(1)^2. \tag{4.81}$$

Minimization problem (using the natural boundary condition):
Find $u \in \tilde{D}_L$ such that $I(u) \le I(v)$ for all $v \in \tilde{D}_L$ where

$$I(v) = \frac{1}{2}\int_0^1 (v'^2 - 2fv)\ \mathrm{d}x + \beta v(1)^2. \tag{4.82}$$

In this case with the mixed boundary condition at $x = 1$, the solution u minimises $I(v)$ over the space D_L and over the larger space $\tilde{D}_L$ which is another way to understand the term natural boundary condition.

4.6 Remarks about non-homogeneous boundary conditions

We started the section by showing that we could convert standard problems with non-homogeneous boundary conditions to equivalent problems with homogeneous boundary conditions. This was partly done for convenience so that we are searching for u in a space of functions rather than searching for u from a set of functions such as

$$S = \{v \in C^2[0,1] : \quad v(0) = \alpha,\ v(1) = \beta\} \tag{4.83}$$

in the case of the problem with Dirichlet boundary conditions. S here is a set if either α or β is non-zero because if $v_1, v_2 \in S$ then $v_1 + v_2 \notin S$. When $Lu = f$ we still however have that if $u \in S$ and $v \in D_L$, where

$$D_L = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\} \tag{4.84}$$

then

$$(Lu, v) = a(u, v) = \int_0^1 u'v'\ \mathrm{d}x. \tag{4.85}$$

Hence we still have a symmetric expression but S and D_L are not the same. Using the set S the weaker problem can be written in the form: Find $u \in S$ such that $a(u, v) = (f, v)$ for all $v \in D_L$.

For the problem with the mixed boundary condition $u'(1) + \beta u(1) = g$ we have

$$S = \{v \in C^2[0,1] : \quad v(0) = \alpha\} \tag{4.86}$$

and

$$\tilde{D}_L = \{v \in C^2[0,1] : \quad v(0) = 0\}. \tag{4.87}$$

With $u \in S$ and $v \in \tilde{D}_L$ the solution that we seek satisfies

$$(Lu, v) = \int_0^1 u'v'\ \mathrm{d}x - u'(1)v(1) = \int_0^1 u'v'\ \mathrm{d}x + (\beta u(1) - g)v(1). \tag{4.88}$$

If, as before, we define $a(.,.)$ by

$$a(u, v) = \int_0^1 u'v'\ \mathrm{d}x + \beta u(1)v(1) \tag{4.89}$$

then the weaker problem in terms of the set S involves finding $u \in S$ such that

$$a(u, v) = (f, v) + gv(1) \quad \text{for all } v \in D_L. \tag{4.90}$$

4.7 Reformulating a more general two-point BVP

If

$$Lu = -(pu')' + qu \tag{4.91}$$

then

$$\begin{aligned}(Lu, v) &= \int_0^1 (-(pu')'v + quv)\ \mathrm{d}x &(4.92)\\ &= \int_0^1 (pu'v' + quv)\ \mathrm{d}x - [pu'v]_0^1. &(4.93)\end{aligned}$$

This is symmetrical when $v(0) = v(1) = 0$ and thus we define the space D_L and the bilinear form by

$$D_L = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\}, \quad a(u, v) = \int_0^1 (pu'v' + quv)\ \mathrm{d}x. \tag{4.94}$$

Hence $Lu = f$ with boundary conditions $u(0) = u(1) = 0$ has the following weaker form.

$$\text{Find } u \in D_L \quad \text{such that} \quad a(u, v) = (f, v) \quad \text{for all } v \in D_L. \tag{4.95}$$

The bilinear form $a(u, v)$ here is symmetrical. If $p(x) \ge p_0 > 0$ on $[0, 1]$ and if $q(x) \ge 0$ then we can guarantee that it is positive definite as we have

$$a(v, v) \ge p_0 \int_0^1 v'^2\ \mathrm{d}x \ge 0 \tag{4.96}$$

and, if $a(v, v) = 0$ then, as in the previous example, $v = 0$. Hence corresponding to what was done for the problem $Lu = -u''$ we similarly have the following.

1. For the problem with the Dirichlet boundary conditions $u(0) = u(1) = 0$ we have the following.

   ODE: Find $u \in D_L = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\}$ such that $-(pu')' + qu = f$.

   Weaker form: Find $u \in D_L$ such that

   $$\int_0^1 (pu'v' + quv)\ \mathrm{d}x = \int_0^1 fv\ \mathrm{d}x \quad \text{for all } v \in D_L. \tag{4.97}$$

   Minimization problem: Find $u \in D_L$ such that $I(u) \le I(v)$ for all $v \in D_L$ where

   $$I(v) = \frac{1}{2} \int_0^1 (pv'^2 + qv^2 - 2fv)\ \mathrm{d}x. \tag{4.98}$$

2. For the problem with a mixed boundary condition $u'(1) + \beta u(1) = 0$ we define

   $$\tilde{D}_L = \{v \in C^2[0,1] : \quad v(0) = 0\}. \tag{4.99}$$

   We have the following.

   ODE: Find $u \in \tilde{D}_L$ such that $-(pu')' + qu = f$ and $u'(1) + \beta u(1) = 0$.

Weaker form (using the natural boundary condition): Find $u \in \tilde{D}_L$ such that

$$\int_0^1 (pu'v' + quv)\ \mathrm{d}x + \beta p(1)u(1)v(1) = \int_0^1 fv\ \mathrm{d}x \quad \text{for all } v \in \tilde{D}_L. \tag{4.100}$$

Minimization problem (using the natural boundary condition): Find $u \in \tilde{D}_L$ such that $I(u) \le I(v)$ for all $v \in \tilde{D}_L$ where

$$I(v) = \frac{1}{2}\int_0^1 (pv'^2 + qv^2 - 2fv)\ \mathrm{d}x + \beta v(1)^2. \tag{4.101}$$

4.8 Reformulating a fourth order BVP

Consider now the fourth order boundary value problem (BVP)

$$Lu = u'''' = f, \quad 0 < x < 1, \quad u(0) = u'(0) = u(1) = u'(1) = 0. \tag{4.102}$$

Proceeding as before we define

$$D_L = \{v \in C^4[0,1] : \quad v(0) = v'(0) = v(1) = v'(1) = 0\}. \tag{4.103}$$

Then

$$\begin{aligned}
\int_0^1 (Lu)v\ \mathrm{d}x &= \int_0^1 u''''v\ \mathrm{d}x, && (4.104)\\
&= -\int_0^1 u'''v'\ \mathrm{d}x + [u'''v]_0^1 = -\int_0^1 u'''v'\ \mathrm{d}x, && (4.105)\\
&\qquad\qquad \text{because } v(0) = v(1) = 0, && (4.106)\\
&= \int_0^1 u''v''\ \mathrm{d}x + [u'''v']_0^1 = \int_0^1 u''v''\ \mathrm{d}x, && (4.107)\\
&\qquad\qquad \text{because } v'(0) = v'(1) = 0. && (4.108)
\end{aligned}$$

The differential operator L is symmetric on D_L and we define the symmetric bilinear form

$$a(u,v) = \int_0^1 u''v''\ \mathrm{d}x. \tag{4.109}$$

This is positive definite on D_L since

$$a(v,v) = \int_0^1 v''^2\ \mathrm{d}x \ge 0 \tag{4.110}$$

and if $a(v,v) = 0$ then the continuity of the integrand implies that $v'' = 0$ which in turn implies that v' is constant and v is linear. Since v must also satisfy the boundary conditions $v(0) = v(1) = 0$ we must have $v = 0$.

Thus the ODE, weaker problem and minimization problems are as follows.

ODE: Find $u \in D_L$ such that $u'''' = f$.
Weaker form: Find $u \in D_L$ such that

$$\int_0^1 u''v''\ \mathrm{d}x = \int_0^1 fv\ \mathrm{d}x \quad \text{for all } v \in D_L. \tag{4.111}$$

Minimisation problem: Find $u \in D_L$ such that $I(u) \le I(v)$ for all $v \in D_L$ where

$$I(v) = \frac{1}{2}\int_0^1 (v''^2 - 2fv)\ \mathrm{d}x. \tag{4.112}$$

Fourth order problems like this arise in the modelling of the deflection of beams. The equation that arises is

$$(EIu'')'' = f \tag{4.113}$$

where E is Young's modulus for the material of the beam, I is the moment of inertia of the cross-section of the beam and f is the force per unit length of the beam. E, I and f may all vary along a beam and thus may all be functions of x, $0 < x < 1$. The unknown $u = u(x)$ in this context is the deflection of the beam as a result of the loading given by $f = f(x)$ and the boundary conditions $u(0) = u'(0) = u(1) = u'(1) = 0$ correspond to a beam which is rigidly clamped at both ends. Following the derivation for $Lu = u''''$ the bilinear form here is

$$a(u, v) = \int_0^1 E(x)I(x)u''(x)v''(x)\ \mathrm{d}x. \tag{4.114}$$

4.9 Summary

1. In the creation of the weak form note that if $r = Lu - f \in C[0, 1]$ and

$$\int_0^1 r(x)v(x)\ \mathrm{d}x = 0 \quad \text{for all} \quad v \in C^q(0, 1), \quad v(0) = v(1) = 0 \tag{4.115}$$

   then $r(x) = 0$. The function v here is called a test function.

2. $$(u, v) = \int_0^1 u(x)v(x)\ \mathrm{d}x \tag{4.116}$$

   defines an inner product. We have given the definitions for a differential operator to be linear, symmetric and positive definite on an inner product space D_L.

3. $$a(u, v) = \int_0^1 u'v'\ \mathrm{d}x \tag{4.117}$$

   is an example of a bilinear form. We have given the definitions of symmetry and positive definite for $a(.,.)$.

4. Let

$$I(v) = \frac{1}{2}a(v, v) - (f, v). \tag{4.118}$$

$$a(u, v) = (f, v) \quad \forall v \in V \quad \text{implies} \quad I(u) \le I(v) \quad \forall v \in V. \tag{4.119}$$

$$I(u) \le I(v) \quad \forall v \in V \quad \text{implies} \quad a(u, v) = (f, v) \quad \forall v \in V. \tag{4.120}$$

   The weak problem and the minimization problem are equivalent.

5. ODE: $Lu = -u'' = f$, $0 < x < 1$, $u(0) = 0$, $u'(1) + \beta u(1) = 0$.

   Bilinear form: $a(u, v) = \int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1)$.

   Weaker form: Find $u \in D_L = \{v \in C^2[0, 1] : \quad v(0) = 0\}$ such that $a(u, v) = (f, v)$ for all $v \in D_L$.

   The Dirichlet condition $u(0) = 0$ is an essential boundary condition for the weaker form. The mixed boundary condition at $x = 1$ is a natural boundary condition for the weaker problem.

6. ODE: $Lu = -u'' = f$, $0 < x < 1$, $u(0) = \alpha$, $u'(1) + \beta u(1) = g$.

   Bilinear form: $a(u, v) = \int_0^1 u'v' \, \mathrm{d}x + \beta u(1)v(1)$.

   Weaker form: Find $u \in S = \{v \in C^2[0, 1] : \quad v(0) = \alpha\}$ such that $a(u, v) = (f, v) + gv(1)$ for all $v \in \{v \in C^2[0, 1] : \quad v(0) = 0\}$.

7. ODE: $Lu = u'''' = f$, $0 < x < 1$, $u(0) = u'(0) = u(1) = u'(1) = 0$.

   The weak form involves $\int_0^1 u''v'' \, \mathrm{d}x = \int_0^1 fv \, \mathrm{d}x$.

5 The Calculus of Variations

5.1 The Euler Lagrange equations — the simplest case

The previous sections have been concerned with reformulating a linear ODE $Lu = f$ into a weak form $a(u, v) = (f, v)$ or as a minimization problem involving a functional $I(v)$. We now briefly look at the reverse by starting with a functional $I(v)$ and consider determining the ODE that a function u which minimises the functional must satisfy. Specifically consider the following problem.

Let

$$S = \{v \in C^2[0,1] : \quad v(0) = \alpha,\ v(1) = \beta\} \quad \text{and} \quad D = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\} \tag{5.1}$$

and suppose that $I(v)$ is of the form

$$I(v) = \int_0^1 F(x, v(x), v'(x))\ \mathrm{d}x \tag{5.2}$$

for some function F. Now if $u \in S$ and $\eta \in D$, $\eta \neq 0$ then $u + t\eta \in S$ for all $t \in \mathbb{R}$ and we can consider the function

$$g(t) = I(u + t\eta) = \int_0^1 F(x, u + t\eta, u' + t\eta')\ \mathrm{d}x. \tag{5.3}$$

If we let

$$\tilde{u}(x,t) = u(x) + t\eta(x) \quad \text{then} \quad \tilde{u}'(x,t) = u'(x) + t\eta'(x) \tag{5.4}$$

where in the above $'$ means differentiation with respect to x with t fixed. With this notation

$$g(t) = I(\tilde{u}) = \int_0^1 F(x, \tilde{u}, \tilde{u}')\ \mathrm{d}x. \tag{5.5}$$

Differentiating and using the chain rule gives

$$g'(t) = \int_0^1 \frac{\partial}{\partial t} F(x, \tilde{u}, \tilde{u}')\ \mathrm{d}x \tag{5.6}$$

$$= \int_0^1 \left(\frac{\partial F}{\partial \tilde{u}} \frac{\partial \tilde{u}}{\partial t} + \frac{\partial F}{\partial \tilde{u}'} \frac{\partial \tilde{u}'}{\partial t} \right)\ \mathrm{d}x = \int_0^1 \left(\frac{\partial F}{\partial \tilde{u}} \eta + \frac{\partial F}{\partial \tilde{u}'} \eta' \right)\ \mathrm{d}x. \tag{5.7}$$

If we evaluate at $t = 0$ and use integration by parts on the second term in the integrand we obtain

$$g'(0) = \int_0^1 \left(\frac{\partial F}{\partial u} \eta + \frac{\partial F}{\partial u'} \eta' \right)\ \mathrm{d}x \tag{5.8}$$

$$= \int_0^1 \left(\frac{\partial F}{\partial u} - \frac{d}{dx} \left(\frac{\partial F}{\partial u'} \right) \right) \eta\ \mathrm{d}x \tag{5.9}$$

since $\eta(0) = \eta(1) = 0$. Now if the functional $I(.)$ has a minimum at $u \in S$ then $g(t)$ has a minimum at $t = 0$ and $g'(0) = 0$ for all choices of $\eta \in D$. Hence

$$\int_0^1 \left(\frac{\partial F}{\partial u} - \frac{d}{dx} \left(\frac{\partial F}{\partial u'} \right) \right) \eta\ \mathrm{d}x = 0 \quad \text{for all } \eta \in D \tag{5.10}$$

and, as we have already shown in the previous section, this implies that

$$\frac{\partial F}{\partial u} = \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) \tag{5.11}$$

which is known as the Euler Lagrange differential equation. (This equation actually describes any stationary point of the functional I.)

Examples of quadratic functionals

In all the cases that we have considered so far the function F is quadratic with the corresponding ODE being linear although in more general situations the ODE will be nonlinear. We have so far considered the following.

1. With

$$F(x,u,u') = \frac{1}{2}u'^2 - fu \tag{5.12}$$

we get

$$\frac{\partial F}{\partial u'} = u', \quad \frac{\partial F}{\partial u} = -f \quad \text{and the ODE is} \quad u'' = -f. \tag{5.13}$$

2. With

$$F(x,u,u') = \frac{1}{2}(pu'^2 + qu^2) - fu \tag{5.14}$$

we get

$$\frac{\partial F}{\partial u'} = pu', \quad \frac{\partial F}{\partial u} = qu - f \quad \text{and the ODE is} \quad (pu')' = qu - f. \tag{5.15}$$

5.2 The mixed boundary condition case in one space dimension

By modifying slightly the functional $I(v)$ considered above we can obtain the functional associated with the mixed boundary condition $u'(1) + \beta u(1) = g$. The functional we need to use is

$$I(v) = \int_0^1 F(x,v,v')\ \mathrm{d}x + \frac{1}{2}(\beta v(1)^2 - 2gv(1)). \tag{5.16}$$

Let

$$S = D = \{v \in C^2[0,1] : \quad v(0) = 0\}. \tag{5.17}$$

Following the previous derivation which led to the Euler Lagrange equations we have

$$I(u+t\eta) = \int_0^1 F(x, u+t\eta, u'+t\eta')\ \mathrm{d}x + \frac{1}{2}\left(\beta(u(1)+t\eta(1))^2 - 2g(u(1)+t\eta(1))\right). \tag{5.18}$$

$$\begin{aligned} \frac{d}{dt}I(u+t\eta)|_{t=0} &= \int_0^1 \left(\frac{\partial F}{\partial u}\eta + \frac{\partial F}{\partial u'}\eta'\right)\ \mathrm{d}x + (\beta u(1) - g)\eta(1), & (5.19)\\ &= \int_0^1 \left(\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right)\right)\eta\ \mathrm{d}x + \left(\frac{\partial F}{\partial u'} + \beta u(1) - g\right)\eta(1). & (5.20)\end{aligned}$$

For this to be zero for all $\eta \in D$ we must have

$$\frac{\partial F}{\partial u} = \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) \quad \text{and} \quad \frac{\partial F}{\partial u'}(1, u(1), u'(1)) + \beta u(1) = g. \tag{5.21}$$

The boundary condition at $x = 1$ is the natural boundary condition for this functional (5.16). That is the function u which minimises the functional over the space S satisfies the essential boundary condition $u(0) = 0$ (which is part of the specification of S) and the natural boundary condition given in (5.21).

5.3 The fourth order BVP in one space dimension

The functional associated with the fourth order BVPs are of the form

$$I(v) = \int_0^1 F(x, v, v', v'')\ \mathrm{d}x. \tag{5.22}$$

Let now

$$S = D = \{v \in C^4[0,1] : \quad v(0) = v'(0) = v(1) = v'(1) = 0\}. \tag{5.23}$$

Then with $u \in S$ and $\eta \in D$ we have

$$I(u + t\eta) = \int_0^1 F(x, u + t\eta, u' + t\eta', u'' + t\eta'')\ \mathrm{d}x \tag{5.24}$$

and

$$\frac{d}{dt} I(u + t\eta)|_{t=0} = \int_0^1 \left(\frac{\partial F}{\partial u}\eta + \frac{\partial F}{\partial u'}\eta' + \frac{\partial F}{\partial u''}\eta'' \right)\ \mathrm{d}x. \tag{5.25}$$

Integrating by parts once or twice as appropriate and making use of the boundary conditions satisfied by η gives

$$\frac{d}{dt} I(u + t\eta)|_{t=0} = \int_0^1 \left(\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial u''}\right) \right) \eta\ \mathrm{d}x. \tag{5.26}$$

For this to be zero for all $\eta \in D$ gives

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial u''}\right) = 0. \tag{5.27}$$

With the expression for F involving u'' terms and assuming that $\partial F/\partial u''$ still involves u'' terms the ODE that we obtain involves fourth order terms u''''.

In the example that we have considered we have

$$F(x, u, u', u'') = \frac{1}{2} EI u''^2 - fu \quad \text{giving} \quad (EIu'')'' = f. \tag{5.28}$$

You can extend this further to higher order problems. What we obtain is that if the functional has derivatives of order up to m then the corresponding Euler Lagrange differential equation has derivatives of order $2m$.

5.4 The Euler Lagrange equations in two space dimensions and for systems

The situation described in the previous sections involved functions of one variable and involved getting the ODE satisfied by the function which minimised a given functional. This generalises to situations in which we have functions of two or more variables and to situations in which we have functionals of more than one scalar valued function. In these cases the Euler Lagrange equations that we obtain are now partial differential equations or systems of partial differential equations. As much of this material goes slightly beyond the syllabus of `MA3951` and `MA5352` we mainly just state the results without too many details being given in most cases.

The basic two dimensional version of what was covered in the previous section involves functions $v = v(x, y)$ defined on $\Omega \subset \mathbb{R}^2$ and it involves functionals of the form

$$I(v) = \iint_\Omega F(x, y, v, v_x, v_y)\ \mathrm{d}x\mathrm{d}y, \quad v_x = \frac{\partial v}{\partial x}, \quad v_y = \frac{\partial v}{\partial y}. \tag{5.29}$$

If a function $u : \Omega \to \mathbb{R}$ minimises this over some set of functions S and $u + t\eta \in S$ for all

$$\eta \in D = \{v \in C^2(\bar{\Omega}) : \quad v = 0 \text{ on } \partial\Omega\} \tag{5.30}$$

then in a similar way to which the Euler Lagrange equations were derived in section 5.1 we have

$$\frac{d}{dt} I(u + t\eta)|_{t=0} = \iint_\Omega \left(\frac{\partial F}{\partial u}\eta + \frac{\partial F}{\partial u_x}\eta_x + \frac{\partial F}{\partial u_y}\eta_y \right)\ \mathrm{d}x\mathrm{d}y. \tag{5.31}$$

Observe that

$$\frac{\partial F}{\partial u_x}\eta_x + \frac{\partial F}{\partial u_y}\eta_y = \underline{A} \cdot \nabla\eta \quad \text{where} \quad \underline{A} = \left(\frac{\partial F}{\partial u_x}, \frac{\partial F}{\partial u_y} \right) \tag{5.32}$$

Now by using the vector identity

$$\nabla \cdot (\eta \underline{A}) = \eta \nabla \cdot \underline{A} + \underline{A} \cdot \nabla\eta \tag{5.33}$$

and the divergence theorem and that $\eta = 0$ on $\partial\Omega$ we obtain

$$\frac{d}{dt} I(u + t\eta)|_{t=0} = \iint_\Omega \left(\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) \right) \eta\ \mathrm{d}x\mathrm{d}y. \tag{5.34}$$

For this is to be equal to 0 for all $\eta \in D$ we get the Euler Lagrange equation which, in this case, is

$$\frac{\partial F}{\partial u} = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right). \tag{5.35}$$

This is a partial differential equation. We consider this later in the case

$$F = \frac{1}{2}(u_x^2 + u_y^2) - fu \tag{5.36}$$

which gives Poisson's equation

$$\Delta u = \nabla^2 u = u_{xx} + u_{yy} = -f. \tag{5.37}$$

In the case of a functional of two scalar functions of the form

$$I(u, v) = \iint_\Omega F(x, y, u, u_x, u_y, v, v_x, v_y)\ \mathrm{d}x\mathrm{d}y \tag{5.38}$$

we can similarly show that the functions which minimise this from an appropriate set of functions satisfy the system of PDEs

$$\frac{\partial F}{\partial u} = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right), \tag{5.39}$$

$$\frac{\partial F}{\partial v} = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial v_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial v_y}\right). \tag{5.40}$$

5.5 Some other problems from the calculus of variations

The following are outside of the syllabus of this module and are only mentioned to indicate some other areas in which problems from the calculus of variations arise.

5.5.1 Hamilton's principle in mechanics

This is concerned with the equations which determine the motion of a mechanical system. Suppose a system is described by N functions $q_1(t), \cdots, q_N(t)$ for $t_1 \le t \le t_2$ such that the kinetic energy T is given by

$$T = T(\dot{q}_1, \cdots, \dot{q}_N), \quad \text{where } \dot{q}_i = \frac{dq_i}{dt} \tag{5.41}$$

and the potential energy V is given by

$$V = V(q_1, \cdots, q_N). \tag{5.42}$$

The Lagrangian of the system is

$$L = T - V = L(q_1, \cdots, q_N, \dot{q}_1, \cdots, \dot{q}_N). \tag{5.43}$$

Hamilton's principle is that the actual motion of the system for time t in $t_1 \le t \le t_2$ is such as to render the integral

$$I = \int_{t_1}^{t_2} L \, \mathrm{d}t \tag{5.44}$$

an extremum. The Euler Lagrange equations in this case is the system of ODEs

$$\frac{\partial L}{\partial q_i} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right), \quad i = 1, \cdots, N. \tag{5.45}$$

These equations are known as the Lagrange equations of motion.

5.5.2 The Brachistochrone problem

One of the original problems from the calculus of variations is known as the Brachistochrone problem which was posed by J. Bernoulli in 1696. The problem concerns determining the shape of a curve joining a point $(0, h)$, $h > 0$ to a point $(b, 0)$, $b > 0$ such that a particle starting from rest at $(0, h)$ and moving only under the influence of gravity will reach $(b, 0)$ in the least time. To describe this suppose that the path can be described parametrically by

$$(x(t), y(t)), \quad 0 < t < T, \tag{5.46}$$

where t is time, $(x(0), y(0)) = (0, h)$, $(x(T), y(T)) = (b, 0)$ and the starting from rest condition is $(\dot{x}(0), \dot{y}(0)) = (0, 0)$. As the particle starts moving and increases its speed it gains kinetic energy and loses potential energy and under the assumption that the particle is moving along a frictionless path the conservation of energy leads to

$$\frac{1}{2}(\dot{x}^2 + \dot{y}^2) = g(h - y) \tag{5.47}$$

where g is the acceleration due to gravity. Now

$$\dot{y} = \frac{dy}{dt} = \frac{dy}{dx}\frac{dx}{dt} = y'\dot{x} \quad \text{giving} \quad \dot{x}^2 + \dot{y}^2 = \dot{x}^2(1 + y'^2). \tag{5.48}$$

Hence

$$\dot{x}^2(1+y'^2) = 2g(h-y) \quad \text{gives} \quad \mathrm{d}t = \left(\frac{(1+y'^2)}{2g(h-y)}\right)^{1/2} \mathrm{d}x. \tag{5.49}$$

With the curve described in the form $y = y(x)$, $0 \le x \le b$ we get an expression for the time T in the form

$$T = \int_0^T \mathrm{d}t = \int_0^b \left(\frac{(1+y'^2)}{2g(h-y)}\right)^{1/2} \mathrm{d}x. \tag{5.50}$$

This is of the form

$$T = \int_0^b F(y, y')\,\mathrm{d}x \quad \text{where} \quad F(y, y') = \left(\frac{(1+y'^2)}{2g(h-y)}\right)^{1/2}. \tag{5.51}$$

This is a functional defined for all sufficiently smooth functions for which $y(0) = h$ and $y(b) = 0$. As the function F does not depend explicitly on x we have

$$\frac{dF}{dx} = \frac{\partial F}{\partial y}y' + \frac{\partial F}{\partial y'}y'', \tag{5.52}$$

$$\frac{d}{dx}\left(F - y'\frac{\partial F}{\partial y'}\right) = \frac{\partial F}{\partial y}y' + \frac{\partial F}{\partial y'}y'' - y'\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) - \frac{\partial F}{\partial y'}y'' = 0 \tag{5.53}$$

where in the above we have used the product rule and that F satisfies the Euler Lagrange equations. Thus in these situations the Euler Lagrange equation reduces to

$$F - y'\frac{\partial F}{\partial y'} = \text{const.} \tag{5.54}$$

We omit the details here but it turns out that for $F(y, y')$ given in (5.51) these equations can be solved exactly (up to determining certain parameters) with the solution being a curve described parametrically in the form

$$x(\theta) = k\left(\theta - \frac{\sin(2\theta)}{2}\right), \tag{5.55}$$

$$y(\theta) = h - k\sin^2\theta = h - \frac{k}{2}(1-\cos 2\theta), \quad 0 \le \theta \le \theta_0 \tag{5.56}$$

where the constants k and θ_0 are such that

$$x(\theta_0) = b, \tag{5.57}$$

$$y(\theta_0) = 0. \tag{5.58}$$

For given values of $h > 0$ and $b > 0$ we have to solve these equations numerically.

5.5.3 Isoperimetric problems

Up to this point we have just considered minimising (or more generally finding the extremum) of a functional over an appropriate set of functions. A modification of this type of problem is that of finding a function u from an appropriate set of functions which minimising a functional subject to the constraint that another functional involving u has some specified value. That is, mathematically in the simplest case the problem is of minimising

$$I = \int_0^1 F(x, u, u')\,\mathrm{d}x \tag{5.59}$$

subject to the constraint that

$$\int_0^1 G(x,u,u')\ \mathrm{d}x = J \tag{5.60}$$

for some given number J. We can tackle this situation by the method of undetermined Lagrange multipliers by defining

$$\tilde{I} = \int_0^1 F(x,u,u')\ \mathrm{d}x - \lambda \left(\int_0^1 G(x,u,u')\ \mathrm{d}x - J \right). \tag{5.61}$$

If we minimise this with respect to the parameter λ and with respect to all functions from the appropriate set then

$$\frac{\partial \tilde{I}}{\partial \lambda} = 0, \quad \text{which gives} \quad \int_0^1 G(x,u,u')\ \mathrm{d}x = J, \tag{5.62}$$

and u satisfies the Euler Lagrange equations involving $\tilde{F} = F - \lambda G$ given by

$$\frac{\partial \tilde{F}}{\partial u} = \frac{d}{dx}\left(\frac{\partial \tilde{F}}{\partial u'} \right). \tag{5.63}$$

This type of problem where we minimise something with respect to something else remaining fixed is known as an *isoperimetric* problem. The name isoperimetric comes from one of the oldest problems of this type which we next outline.

The ancient Greeks posed the problem of determining the shape of a closed curve of a given fixed length L such that the enclosed area is as large as possible. The answer, as you are probably aware, is a circle. To describe the functionals involved which lead to this answer first let the closed curve be described parametrically by

$$(x(t), y(t)), \quad 0 \le t \le 1 \quad \text{with} \quad x(0) = x(1) \quad \text{and} \quad y(0) = y(1). \tag{5.64}$$

It can be shown that the area enclosed by the curve is given by

$$I = \int_0^1 \frac{1}{2}(x\dot{y} - y\dot{x})\,\mathrm{d}t, \quad \dot{x} = \frac{dx}{dt}, \quad \dot{y} = \frac{dy}{dt}. \tag{5.65}$$

(You get this by applying the divergence theorem to $\underline{A} = (x,y)/2$.) The length of the curve is given by

$$L = \int_0^1 (\dot{x}^2 + \dot{y}^2)^{1/2}\,\mathrm{d}t. \tag{5.66}$$

If $L > 0$ is fixed then the problem concerns determining suitable functions $x(t)$, $y(t)$ such that

$$\tilde{I} = \int_0^1 \left(\frac{1}{2}(x\dot{y} - y\dot{x}) - \lambda(\dot{x}^2 + \dot{y}^2)^{1/2} \right)\,\mathrm{d}t + \lambda L, \tag{5.67}$$

is minimised with respect to all such functions and the parameter λ. By setting up and solving the Euler Lagrange equations we get the familiar result that a curve which gives the minimum is a circle.

It is because the constraint in this problem involves a condition on the perimeter (i.e. on the boundary) of a region that we get the name isoperimetric problem.

5.6 Summary

1. The Euler Lagrange equation associated with the functional

$$I(u) = \int_0^1 F(x, u, u') \, \mathrm{d}x \tag{5.68}$$

is

$$\frac{\partial F}{\partial u} = \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right). \tag{5.69}$$

This is an ODE.

2. The Euler Lagrange equation associated with the functional

$$I(u) = \int_0^1 F(x, u, u', u'') \, \mathrm{d}x \tag{5.70}$$

is

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial u''}\right) = 0. \tag{5.71}$$

3. The Euler Lagrange equation associated with the functional

$$I(v) = \iint_\Omega F(x, y, v, v_x, v_y) \, \mathrm{d}x\mathrm{d}y. \tag{5.72}$$

is

$$\frac{\partial F}{\partial u} = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right). \tag{5.73}$$

This is a PDE.

4. The Euler Lagrange equations associated with the functional

$$I(u, v) = \iint_\Omega F(x, y, u, u_x, u_y, v, v_x, v_y) \, \mathrm{d}x\mathrm{d}y \tag{5.74}$$

are

$$\frac{\partial F}{\partial u} = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right), \tag{5.75}$$

$$\frac{\partial F}{\partial v} = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial v_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial v_y}\right). \tag{5.76}$$

5. Isoperimetric problems are concerned with minimising functionals over appropriate sets of functions subject to the constraint that another functional remains fixed. We can solve these using the method of undetermined Lagrange multipliers. In the simplest case this involves minimising expressions of the form

$$\tilde{I} = \int_0^1 F(x, u, u') - \lambda(G(x, u, u') \, \mathrm{d}x - J) \tag{5.77}$$

with respect to all functions u from some set and for all parameters λ.

6 The Galerkin and Ritz methods

6.1 The methods with a general set of basis functions

All the previous sections have been concerned with reformulating an ODE into a weak form or as an equivalent minimization problem. In abstract terms these weak forms are of the following form.

$$\text{Find } u \in V \quad \text{such that} \quad a(u,v) = (f,v) \quad \text{for all} \quad v \in V \tag{6.1}$$

where $a(.,.)$ is a symmetric and positive definite bilinear form and V is an appropriate space of functions. The practical task is then to attempt to approximately solve this weak problem which we do here by the Galerkin method.

The Galerkin method involves constructing linearly independent functions $\phi_1, \phi_2, \cdots$ in V from which we get an n-dimensional subspace

$$V_n := \operatorname{span}\{\phi_1, \cdots, \phi_n\} \subset V. \tag{6.2}$$

We then consider the approximate finite dimensional weak problem:

$$\text{Find } U_n \in V_n \quad \text{such that} \quad a(U_n, v) = (f,v) \quad \text{for all} \quad v \in V_n. \tag{6.3}$$

Now as the function $U_n \in V_n$ it is of the form

$$U_n = \sum_{j=1}^n c_j \phi_j. \tag{6.4}$$

To satisfy the condition "for all $v \in V_n$" it is necessary and sufficient to take v to be each of $\phi_1, \cdots, \phi_n$. The n unknown coefficients $c_1, \cdots, c_n$ hence satisfy the n equations

$$a(U_n, \phi_i) = (f, \phi_n) \quad \text{for } i = 1, 2, \cdots, n. \tag{6.5}$$

Hence

$$a(U_n, \phi_i) = a\left(\sum_j c_j\phi_j, \phi_i\right) = \sum_{j=1}^n a(\phi_j, \phi_i) c_j = (f, \phi_i), \quad i = 1, \cdots, n. \tag{6.6}$$

In matrix notation we have

$$K\underline{c} = \underline{b} \tag{6.7}$$

where

$$K = (k_{ij}), \quad k_{ij} = a(\phi_j, \phi_i), \quad \underline{b} = (b_i), \quad b_i = (f, \phi_i). \tag{6.8}$$

Note that since $a(.,.)$ is symmetric we have $k_{ij} = k_{ji} = a(\phi_i, \phi_j)$ and the matrix K is symmetric. Also note that if we consider a function

$$v = \sum_{j=1}^n c_j \phi_j \tag{6.9}$$

then the bilinear property of $a(.,.)$ gives

$$a(v,v) = a\left(\sum_1^n c_i\phi_i, \sum_1^n c_j\phi_j\right) = \sum_1^n \sum_1^n a(\phi_i, \phi_j) c_i c_j = \sum_1^n \sum_1^n k_{ij} c_i c_j. \tag{6.10}$$

This is a quadratic form and it can be written in matrix form as

$$a(v,v) = \underline{c}^T K \underline{c} \quad \text{where} \quad \underline{c} = (c_i). \tag{6.11}$$

As $a(.,.)$ is positive definite we have $a(v,v) > 0$ for all $v \neq 0$ and thus

$$a(v,v) = \underline{c}^T K \underline{c} > 0 \quad \text{for all} \quad \underline{c} \neq \underline{0}. \tag{6.12}$$

The matrix K is both symmetric and positive definite. To summarise, the Galerkin method solves a finite dimensional weak problem and gives an approximation

$$U_n = \sum_1^n c_j \phi_j \quad \text{where} \quad \underline{c} = (c_i) \quad \text{satisfies} \quad K\underline{c} = \underline{b}. \tag{6.13}$$

The Rayleigh-Ritz method (or just the Ritz method for short) is concerned with minimising the functional

$$I(v) = \frac{1}{2} a(v,v) - (f,v) \tag{6.14}$$

over the finite dimensional subspace $V_n \subset V$ and it gives the same approximate solution as the Galerkin method just described. This equivalence can be established as follows.

If $v \in V_n$ then it is of the form

$$v = \sum_{j=1}^n c_j \phi_j \quad \text{and} \quad a(v,v) = \underline{c}^T K \underline{c}. \tag{6.15}$$

Similarly

$$\begin{aligned} (f,v) &= \left(f, \sum_1^n c_j \phi_j \right) = \sum_1^n c_j (f,\phi_j) = \sum_1^n c_j b_j & (6.16) \\ &= (c_1, \cdots, c_n) \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \underline{c}^T \underline{b}. & (6.17) \end{aligned}$$

Thus

$$I(v) = \frac{1}{2} \underline{c}^T K \underline{c} - \underline{c}^T \underline{b}. \tag{6.18}$$

Minimising this quadratic is equivalent to solving a linear system. That is

$$\frac{\partial}{\partial c_i} I(v) = \frac{1}{2} \left(\underline{e}_i^T K \underline{c} + \underline{c}^T K \underline{e}_i \right) - \underline{e}_i^T \underline{b} \tag{6.19}$$

where $\underline{e}_i \in \mathbb{R}^n$ is the usual ith base vector, i.e. it is the ith column of the $n \times n$ identity matrix I. Since

$$\underline{e}_i^T K \underline{c} = \underline{c}^T K \underline{e}_i \tag{6.20}$$

we have

$$\frac{\partial}{\partial c_i} I(v) = \underline{e}_i^T (K\underline{c} - \underline{b}) = i\text{th component of } (K\underline{c} - \underline{b}). \tag{6.21}$$

Hence

$$\frac{\partial}{\partial c_i} I(v) = 0 \quad \text{for } i = 1, \cdots, n \text{ gives} \qquad K\underline{c} = \underline{b}. \tag{6.22}$$

6.2 Examples with globally defined basis functions

1. Consider the problem

$$-u'' + u = x, \quad u(0) = u(1) = 0. \tag{6.23}$$

We can solve this problem explicitly, the exact solution is

$$u(x) = x + A \sinh x \quad \text{with} \quad A = \frac{-1}{\sinh 1}. \tag{6.24}$$

To implement the Galerkin method to construct approximate solutions we need the corresponding weak problem which, in this case, involves

$$a(u,v) = (f,v), \quad a(u,v) = \int_0^1 u'v' + uv \,\mathrm{d}x, \quad (f,v) = \int_0^1 xv \,\mathrm{d}x \tag{6.25}$$

(i.e. $f(x) = x$) and the space

$$V = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\}. \tag{6.26}$$

We also need the functions $\phi_1, \phi_2, \cdots$. To use polynomials we need polynomials which have a factor $x(1-x)$ in order to satisfy the boundary conditions. The simplest approximation involves just taking $n = 1$ and $\phi_1(x) = x(1-x)$ giving

$$U_1 = c_1\phi_1 \quad \text{where } c_1 \text{ satisfies} \quad a(\phi_1, \phi_1)c_1 = (x, \phi_1). \tag{6.27}$$

We now need to evaluate the integrals involved.

$$\phi_1 = x(1-x) = x - x^2, \quad \phi_1' = 1 - 2x, \tag{6.28}$$

$$(x, \phi_1) = \int_0^1 x\phi_1 \,\mathrm{d}x = \int_0^1 x^2 - x^3 \,\mathrm{d}x = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \tag{6.29}$$

$$\begin{aligned}
a(\phi_1, \phi_1) &= \int_0^1 {\phi_1'}^2 + \phi_1^2 \,\mathrm{d}x = \int_0^1 (1-2x)^2 + x^2(1-2x+x^2) \,\mathrm{d}x && (6.30)\\
&= \int_0^1 (1 - 4x + 4x^2) + (x^2 - 2x^3 + x^4) \,\mathrm{d}x && (6.31)\\
&= \left(1 - \frac{4}{2} + \frac{4}{3}\right) + \left(\frac{1}{3} - \frac{2}{4} + \frac{1}{5}\right) && (6.32)\\
&= \frac{1}{3} + \frac{1}{30} = \frac{11}{30}. && (6.33)
\end{aligned}$$

Thus

$$c_1 = \frac{(x, \phi_1)}{a(\phi_1, \phi_1)} = \frac{5}{22} \quad \text{and} \quad U_1(x) = \frac{5}{22}x(1-x). \tag{6.34}$$

A plot of U_1 and the exact solution is shown in Figure 6.1.

By taking $\phi_k(x) = x^k(1-x)$, $k = 1, 2, \cdots$ we can also compute approximations U_2, U_3, etc.. The approximation U_2, obtained with a computer program, together with the exact solution u is shown in Figure 6.2. As is shown U_2 can hardly be distinguished from u in the plot.

2. Consider the problem

$$-u'' + u = 1, \quad u(0) = 0, \quad u'(1) = 0. \tag{6.35}$$

The exact solution in this case is given by

$$u(x) = 1 + A\sinh x + B\cosh x \quad \text{where now} \quad A = \frac{\sinh 1}{\cosh 1}, \quad B = -1. \tag{6.36}$$

The weak problem involves the same expression for $a(u, v)$ as in the first example but now $f(x) = 1$ and as we have the natural boundary condition at $x = 1$ we take

$$V = \{v \in C^2[0,1] : \quad v(0) = 0\}. \tag{6.37}$$

If we take $\phi_1(x) = x$, $\phi_2(x) = x^2$, $\cdots$ then these functions are in V. To compute the approximations note that $\phi_k' = kx^{k-1}$. Hence computing the terms involving ϕ_1 and ϕ_2 we have

$$a(\phi_1, \phi_1) = \int_0^1 {\phi_1'}^2 + \phi_1^2 \,\mathrm{d}x = \int_0^1 1 + x^2 \,\mathrm{d}x = \frac{4}{3}, \tag{6.38}$$

$$a(\phi_1, \phi_2) = \int_0^1 \phi_1'\phi_2' + \phi_1\phi_2 \,\mathrm{d}x = \int_0^1 2x + x^3 \,\mathrm{d}x = 1 + \frac{1}{4} = \frac{5}{4}, \tag{6.39}$$

$$a(\phi_2, \phi_2) = \int_0^1 {\phi_2'}^2 + \phi_2^2 \,\mathrm{d}x = \int_0^1 4x^2 + x^4 \,\mathrm{d}x = \frac{4}{3} + \frac{1}{5} = \frac{23}{15}, \tag{6.40}$$

$$(f, \phi_1) = \int_0^1 x \,\mathrm{d}x = \frac{1}{2}, \tag{6.41}$$

$$(f, \phi_2) = \int_0^1 x^2 \,\mathrm{d}x = \frac{1}{3}. \tag{6.42}$$

The approximation with $n = 1$ is $U_1 = c_1 x$ where

$$c_1 = \frac{(f, \phi_1)}{a(\phi_1, \phi_1)} = \frac{1/2}{4/3} = \frac{3}{8}. \tag{6.43}$$

A plot of U_1 and the exact solution is shown in Figure 6.3. The approximation with $n = 2$ is $U_2 = c_1 x + c_2 x^2$ where now

$$\begin{pmatrix} 4/3 & 5/4 \\ 5/4 & 23/15 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/3 \end{pmatrix}. \tag{6.44}$$

The solution to this 2×2 system is $c_1 = 0.72622$, $c_2 = -0.37464$ (to 5 significant figures). A plot of U_2 and the exact solution is shown in Figure 6.4.

Note that in this case $U_1'(1) = 3/8$ and $U_2'(1) = 0.72622 - 2(0.37464) = -0.02305$. Neither approximation satisfies the boundary condition of $u'(1) = 0$ at $x = 1$ but, as these values show and the plot shows the approximation U_2 is reasonably accurate.

3. In both the previous examples a low order polynomial gave an accurate approximate solution to problems in which we already knew the exact solution. Consider now the problem

$$-u'' + u = f = \exp(-10(x - 0.05)^2), \quad u(0) = 0, \quad u'(1) = 0. \tag{6.45}$$

This is more difficult and we no longer have an exact solution to assess how good is an approximation. With the functions $\phi_k(x) = x^k$, $k = 1, 2, \cdots$ we obtain, with a computer program, the approximations U_1, U_2, U_3 and U_4 shown in Figure 6.5. The integrals involved in computing (f, ϕ_k) have to be approximated with numerical quadrature which will be discussed briefly later.

4. Consider now the ODE

$$-u'' + u = \mathrm{e}^{3x}\sin(10x^2), \quad u(0) = 0, \quad u(1) = 0 \tag{6.46}$$

and suppose that we have the basis functions $\phi_k(x) = x^k(1-x)$, $k = 1, 2, \cdots$. Again this is an example which is not for hand calculations as we cannot, for example, do the integrals involving f exactly. Nevertheless we can generate U_1, U_2, $U_3, \cdots$ with a computer program and the plots of these are shown in Figure 6.6. The approximations are not converging so well here (none of the graphs shown are close to each other) and it is a case when it is better to use piecewise polynomials as in the FEM.

6.3 Summary

Let $V_n = \mathrm{span}\{\phi_1, \cdots, \phi_n\} \subset V$. The Galerkin method involves finding $U \in V_n$ such that $a(U, v) = (f, v)$ for all $v \in V_n$. We get

$$U = \sum_1^n c_i\phi_i, \quad K\underline{c} = \underline{b}, \quad K = (a(\phi_i, \phi_j)), \quad \underline{b} = ((f, \phi_i)). \tag{6.47}$$

K is symmetric and positive definite.

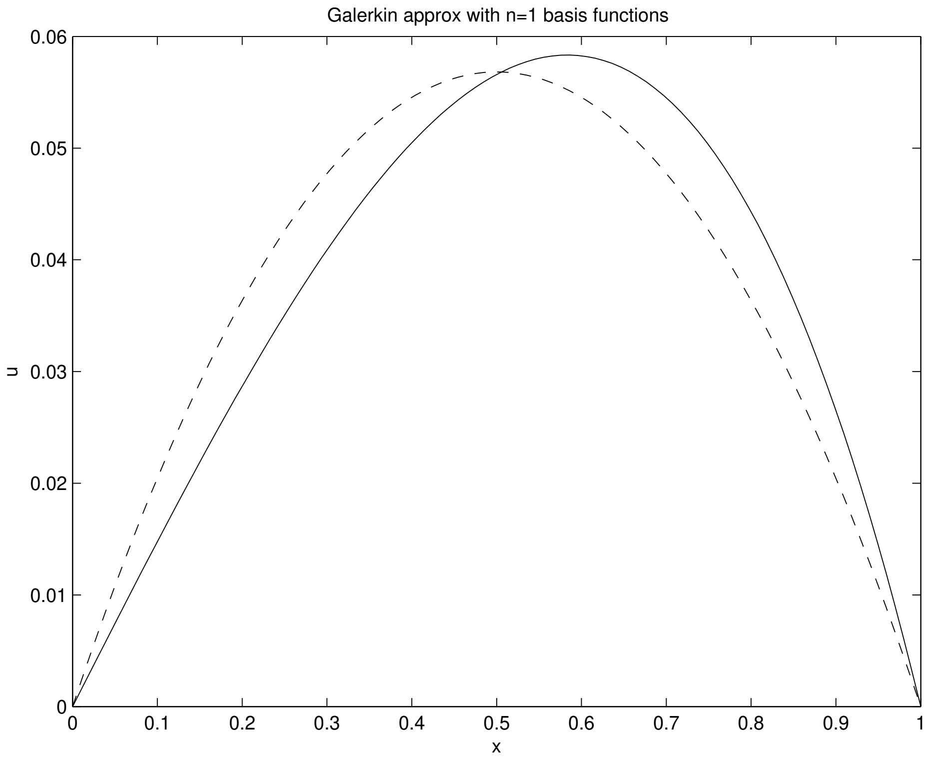


Figure 6.1: Galerkin approximation with $n = 1$ (dashed line) together with the exact solution (solid line).

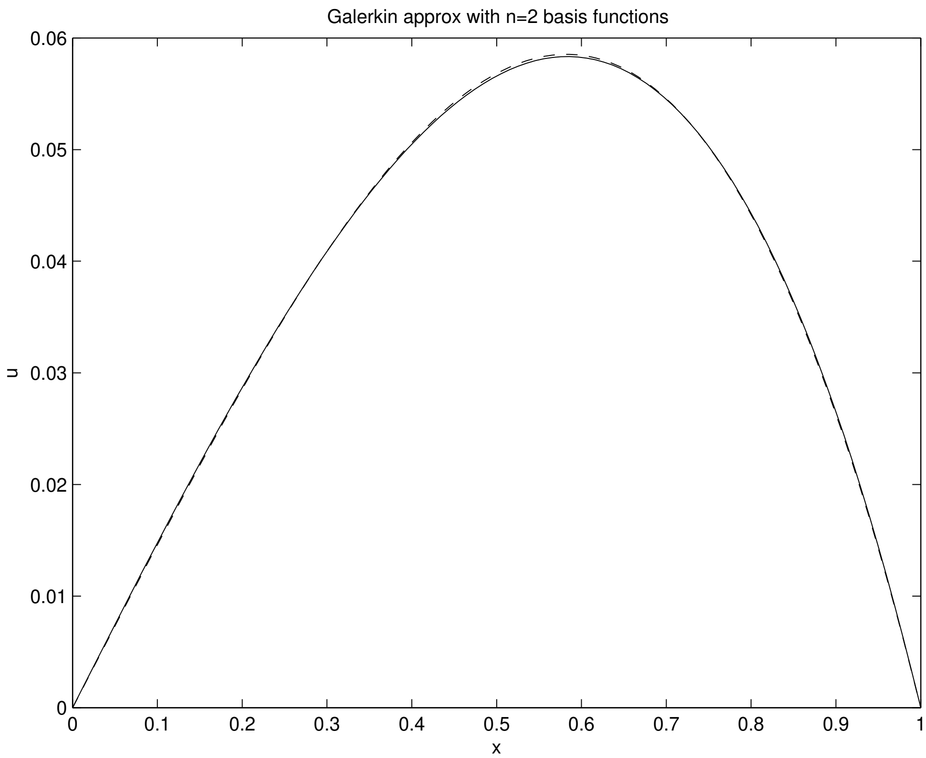


Figure 6.2: Galerkin approximation with $n = 2$ (dashed line) together with the exact solution (solid line).

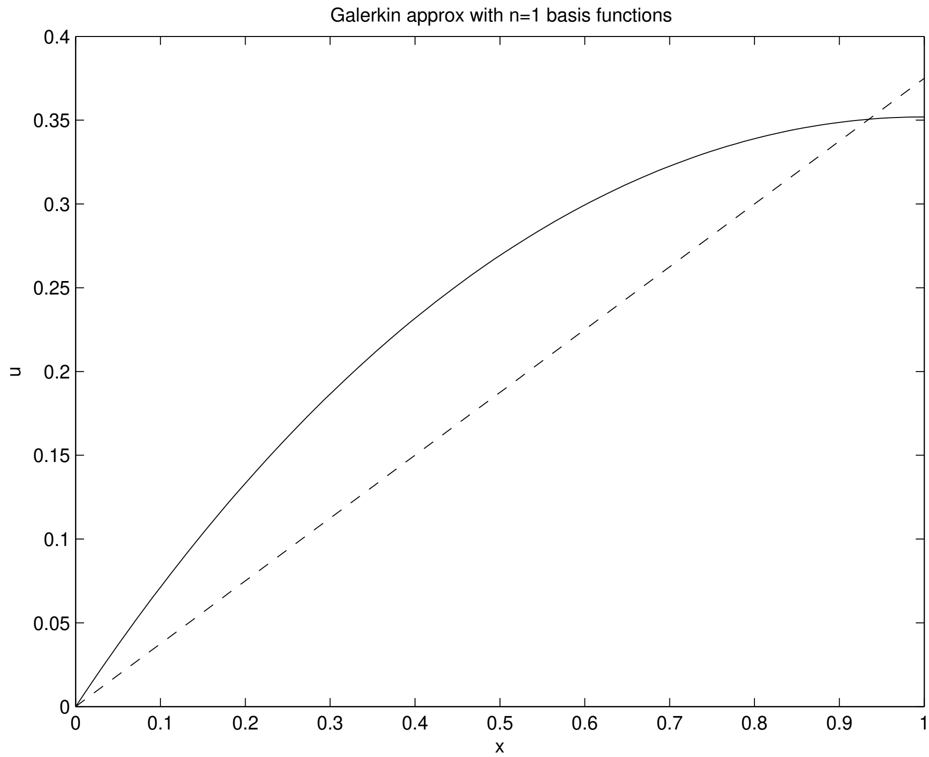


Figure 6.3: Galerkin approximation $U_1 = c_1 x$ (dashed line) together with the exact solution (solid line).

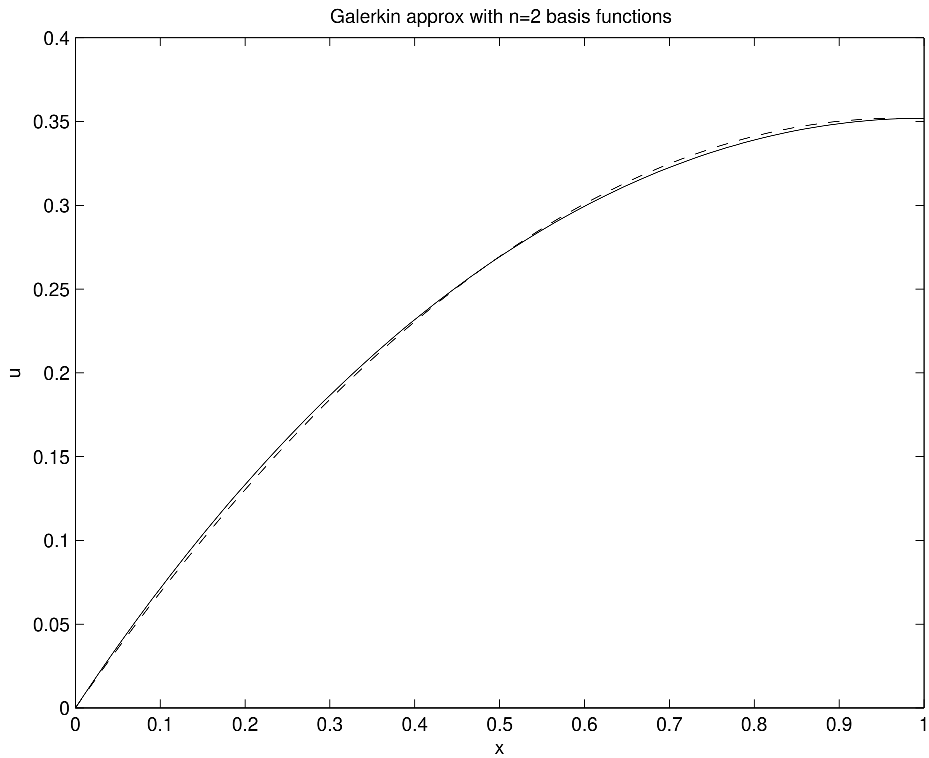


Figure 6.4: Galerkin approximation $U_2 = c_1 x + c_2 x^2$ (dashed line) together with the exact solution (solid line).

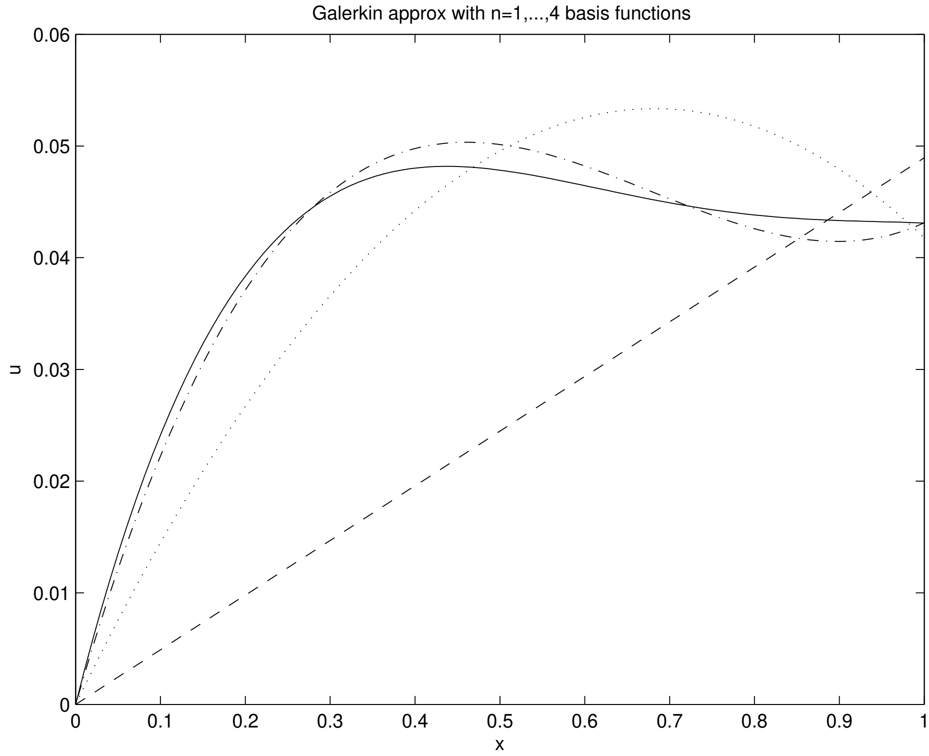


Figure 6.5: Galerkin approximations U_1, U_2, U_3 and U_4. Solid line for U_4.

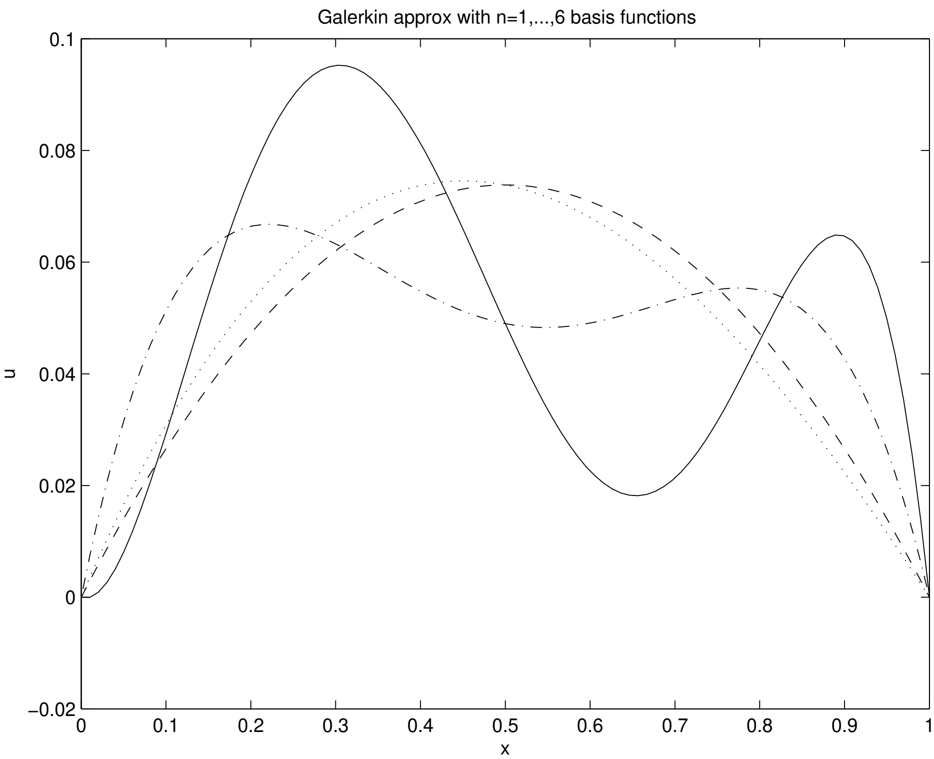


Figure 6.6: Galerkin approximations U_1, U_2, U_3 and U_4, `--` for U_1, dotted line for U_2, `-.` for U_3, solid line for U_4,

7 The weak form and the FEM in one space dimension

7.1 Weakening the continuity requirements: an introduction to Sobolev spaces

The finite element method which we start to consider in the next subsection involves using the Galerkin method with specific choices of the finite dimensional spaces V_n involving functions which are piecewise defined and for the second order ODEs that we consider these functions are not differentiable at all points in $[0,1]$. To attempt to justify the use of such approximating functions we consider again the weaker problems and minimisation problems derived earlier. For the ODE $Lu = -u'' = f$, $0 < x < 1$, $u(0) = u(1) = 0$, where $f \in C[0,1]$ we have $u \in V$ which is such that

$$a(u,v) = (f,v) \quad \text{for all} \quad v \in V \tag{7.1}$$

where

$$a(u,v) = \int_0^1 u'v' \,\mathrm{d}x \quad \text{and} \quad V = \{v \in C^2[0,1] : \quad v(0) = v(1) = 0\}. \tag{7.2}$$

The minimization problem involves

$$I(v) = \frac{1}{2}a(v,v) - (f,v) = \frac{1}{2}\int_0^1 v'^2 - 2fv \,\mathrm{d}x. \tag{7.3}$$

The ODE requires that u is twice continuously differentiable whereas the weaker and minimization problems both make sense provided we can integrate v' on $[0,1]$ and

$$\int_0^1 v'^2 \,\mathrm{d}x < \infty. \tag{7.4}$$

This is the case for piecewise defined functions that are used in the FEM and to attempt to justify this we argue as follows.

Suppose $0 = x_0 < x_1 < \cdots < x_{ne} = 1$ and consider a function v which is continuous on $[0,1]$ and which is differentiable in each interval (x_{i-1}, x_i), $i = 1, 2, \cdots, ne$ with in general discontinuities of v' at $x_1, x_2, \cdots, x_{ne-1}$. Despite the discontinuities in v' we can still integrate it. For example, given any smooth function ϕ satisfying $\phi(0) = \phi(1)$ we have

$$\int_0^1 v'\phi \,\mathrm{d}x = \sum_1^{ne} \int_{x_{i-1}}^{x_i} v'\phi \,\mathrm{dx} = \sum_1^{ne} \left([v\phi]_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} v\phi' \,\mathrm{dx} \right). \tag{7.5}$$

Now since v is continuous at each point x_i and since $\phi(x_0) = \phi(x_{ne}) = 0$ we have

$$\sum_1^{ne} [v\phi]_{x_{i-1}}^{x_i} = 0 \tag{7.6}$$

as the terms at x_0 and x_{ne} vanish and all the intermediate terms cancel. We thus get

$$\int_0^1 v'\phi \,\mathrm{d}x = -\sum_1^{ne} \int_{x_{i-1}}^{x_i} v\phi' \,\mathrm{dx} = -\int_0^1 v\phi' \,\mathrm{d}x. \tag{7.7}$$

That is for such functions v we still satisfy the integration by parts relation

$$\int_0^1 v'\phi \,\mathrm{d}x = -\int_0^1 v\phi' \,\mathrm{d}x \tag{7.8}$$

as do functions which are differentiable at all points in $[0, 1]$.

A function v for which there exists another function v' for which (7.8) holds for all smooth ϕ satisfying $\phi(0) = \phi(1) = 0$ is said to have a weak derivative. The weak derivative is the same as the classical derivative defined by

$$u'(x) = \lim_{h\to 0} \frac{u(x+h) - u(x)}{h} \tag{7.9}$$

whenever this exists but as the previous discussion has shown, the space of functions for which we have a weak derivative is a larger space as it also contains the piecewise defined functions that we use.

In the weaker problem and in the minimization problem we have

$$(f, v) = \int_0^1 fv \, \mathrm{d}x. \tag{7.10}$$

For this to make sense we need to be able to integrate f and v which leads us to define the following.

$$H^0(0,1) := L_2(0,1) := \{v : \quad v \text{ is integrable on } (0,1) \text{ and } \textstyle\int_0^1 v^2 \, \mathrm{d}x < \infty\} \tag{7.11}$$

$L_2(0, 1)$ is an inner product space with the inner product given by (7.10). We can similarly define

$$H^1(0,1) := \{v : \quad v \text{ and } v' \text{ are integrable on } (0,1) \text{ and } \textstyle\int_0^1 v^2 + v'^2 \, \mathrm{d}x < \infty\} \tag{7.12}$$

which is also an inner product space with the inner product defined by

$$(u, v)_1 := \int_0^1 uv + u'v' \, \mathrm{d}x. \tag{7.13}$$

There is a technical point here concerning what is meant by integrable which will not be considered in any detail here. The usual idea that the definite integral of a function gives the area under the curve of the function is sufficient for our needs. All the functions that concern us are in fact Riemann integrable. It turns out that to consider this rigorously we need to consider integrals in the sense of Lebesgue. When this is done the space $L_2(0, 1)$ is complete in the norm

$$\|v\|_{L_2(0,1)} := \left(\int_0^1 v^2 \, \mathrm{d}x\right)^{1/2} \tag{7.14}$$

and the space $H^1(0, 1)$ is complete in the norm

$$\|v\|_{H^1(0,1)} := \left(\int_0^1 v^2 + v'^2 \, \mathrm{d}x\right)^{1/2}. \tag{7.15}$$

The term complete means that if (v_n) is a Cauchy sequence with respect to the norm of the space then the sequence (v_n) converges to an entry of the space. A complete inner product space is known as a Hilbert space.

For the part of the module common to both MA3951 and MA5352 you do not need to be too concerned with all the jargon just introduced other than to note that the Hilbert space

$H^1(0,1)$ is the space of functions for which the derivative can be integrated with the integrals being finite. It is mathematically the largest space for which the expressions such as

$$a(u,v) = \int_0^1 u'v' \,\mathrm{d}x \quad \text{and} \quad a(u,v) = \int_0^1 pu'v' + quv \,\mathrm{d}x \tag{7.16}$$

make sense. It can be shown that functions in $H^1(0,1)$ must be continuous and as such we can consider pointwise values and define

$$H_0^1(0,1) = \{v \in H^1(0,1) : \quad v(0) = v(1) = 0\}. \tag{7.17}$$

With this space we can now consider the following weak problem.

$$\text{Find } u \in H_0^1(\Omega) \quad \text{such that} \quad a(u,v) = \int_0^1 fv \,\mathrm{d}x \quad \text{for all } v \in H_0^1(0,1). \tag{7.18}$$

This weak problem over $H_0^1(0,1)$ is equivalent to the following minimization problem.

$$\text{Find } u \in H_0^1(0,1) \quad \text{such that} \quad I(u) \le I(v) \quad \text{for all } v \in H_0^1(0,1) \tag{7.19}$$

where

$$I(v) = \frac{1}{2}a(v,v) - (f,v). \tag{7.20}$$

It is not difficult to show that these problems have unique solutions in $H_0^1(0,1)$ and this is done in the `MA5352` only part of the module.

The weak problem and the minimization problem on the space $H_0^1(0,1)$ are of course not completely equivalent to the original ODE involving $Lu = -u''$ or more generally $Lu = -(pu')' + qu$ as these only make sense if $u \in C^2[0,1]$ and thus it is relevant to wonder if by solving the weak or minimization problems we get a solution to the original ODE. The answer to this is yes when the ODE has a unique solution but there can also be situations in which the weak problem makes sense but the original differential equation has no meaning. To justify that the weak solution (that is the solution to the weak problem) is the classical solution, when this exists, note that if u solves the ODE and $v \in H_0^1(0,1)$ then this is sufficient for us to get the relation

$$(Lu,v) = a(u,v) \quad \text{and thus} \quad (Lu,v) = a(u,v) = (f,v). \tag{7.21}$$

Then consider

$$\begin{aligned} I(u+v) &= \frac{1}{2}(a(u,u) + 2a(u,v) + a(v,v)) - (f,u+v) && (7.22)\\ &= \left(\frac{1}{2}(a(u,u) - (f,u)\right) + \frac{1}{2}a(v,v) && (7.23)\\ &= I(u) + \frac{1}{2}a(v,v). && (7.24) \end{aligned}$$

In all cases considered $a(.,.)$ is positive definite and thus $a(v,v) > 0$ for all $v \neq 0$ and hence the smooth function u solving the ODE minimises the functional $I(v)$ over this larger function space. This is our justification for considering the larger space $H_0^1(0,1)$ for these one dimensional problems. With these larger spaces we give again the weak form problems associated with our ODEs.

Dirichlet boundary conditions at $x = 0$ and $x = 1$.

<u>ODE:</u> Find $u \in C^2[0,1]$ such that $-(pu')' + qu = f$ and $u(0) = u(1) = 0$.
<u>Weak form</u> Find $u \in H_0^1(0,1)$ such that

$$\int_0^1 (pu'v' + quv)\ \mathrm{d}x = \int_0^1 fv\ \mathrm{d}x \quad \text{for all } v \in H_0^1(0,1). \tag{7.25}$$

With a mixed boundary condition at $x = 1$

<u>ODE:</u> Find $u \in C^2[0,1]$ such that $-(pu')' + qu = f$ and $u(0) = 0$ and $u'(1) + \beta u(1) = 0$.
<u>Weak form</u> Find $u \in V = \{v \in H^1(0,1) : \quad v(0) = 0\}$ such that

$$\int_0^1 (pu'v' + quv)\ \mathrm{d}x + \beta u(1)v(1) = \int_0^1 fv\ \mathrm{d}x \quad \text{for all } v \in V. \tag{7.26}$$

As a remark here we cannot impose the condition $v'(1) + \beta v(1) = 0$ on the space of functions here as functions in $H^1(0,1)$ need not have an ordinary derivative at all points in $[0,1]$.

In the case of the fourth order ODE $Lu = (EIu'')'' = f$, $u(0) = u'(0) = u(1) = u'(1) = 0$ the weak form involves

$$a(u,v) = \int_0^1 EIu''v''\ \mathrm{d}x. \tag{7.27}$$

This makes sense for functions which have weak second derivatives and thus we analogously consider the space

$$H^2(0,1) := \{v : \quad v, v', v'' \in L_2(0,1)\}. \tag{7.28}$$

Functions which are in $H^2(0,1)$ can be shown to be continuously differentiable and hence we can define the space

$$H_0^2(0,1) := \{v \in H^2(0,1) : \quad v(0) = v'(0) = v(1) = v'(1) = 0\}. \tag{7.29}$$

Extending the intermediate weaker problems considered earlier to these larger spaces is crucial for the FEM as it enables us to construct piecewise defined functions which are continuous and are in $H^1(0,1)$ for the second order ODEs and which are continuously differentiable and are in $H^2(0,1)$ for the fourth order ODEs.

7.2 The finite element method: an introduction

The finite element method (FEM) (or the Galerkin FEM as it is sometimes called) involves approximately solving the weak problem as specified using the Hilbert space $H^1(0,1)$ for the second order ODEs and $H^2(0,1)$ for the fourth order ODEs. If V denotes the appropriate space then, as in the Galerkin method previously described, it approximates

$$a(u,v) = (f,v) \quad \text{for all } v \in V \tag{7.30}$$

by the problem

$$a(U,v) = (f,v) \quad \text{for all } v \in V_h \tag{7.31}$$

where V_h is a finite dimensional subspace of V. It is traditional to use the subscript h here as the subspace is usually constructed with respect to a mesh $0 = x_0 < x_1 < \cdots < x_{ne} = 1$ of the domain $[0,1]$ and $h_i = x_i - x_{i-1}$ is used as the notation for the length of a typical subinterval or element. Often we write $h = \max_i h_i$. A small value of h implies that the number of elements ne is large and correspondingly the dimension of V_h will be large. The functions in V_h are piecewise polynomials, i.e. if $v \in V_h$ then $v(x)$ is a polynomial in x in each interval (x_{i-1}, x_i) with the ne polynomials meeting appropriately to give a function in $C[0,1] \subset H^1(0,1)$ for the second order ODEs and giving a function in $C^1[0,1] \subset H^2(0,1)$ for the fourth order ODEs. To write these spaces mathematically the finite element spaces are of the form

$$V_h = \{v \in C[0,1] : \quad v \text{ is a polynomial of degree } p \text{ in } (x_{i-1}, x_i),\ i = 1, \cdots, ne\} \tag{7.32}$$

for the second order ODEs and

$$V_h = \{v \in C^1[0,1] : \quad v \text{ is a polynomial of degree } p \text{ in } (x_{i-1}, x_i),\ i = 1, \cdots, ne\} \tag{7.33}$$

for the fourth order ODEs. We restrict attention here to linears ($p = 1$) and quadratics ($p = 2$) for the case (7.32) and cubics ($p = 3$) for the case (7.33). It is not actually necessary to have the same degree polynomials in each piece although we will restrict to this case here.

7.3 Piecewise linear basis functions

A piecewise linear function is a function of the form shown in Figure 7.1. For a mesh $0 = x_0 < x_1 < \cdots < x_{ne} = 1$ we have $ne + 1$ mesh points $x_0, x_1, \cdots, x_{ne}$, known as **nodes**, and a piecewise linear function U based on these points is defined by its values $U(x_0), \cdots, U(x_{ne})$. We have

$$U(x) = \sum_{i=0}^{ne} U(x_i)\hat{\phi}_i(x) \tag{7.34}$$

where $\hat{\phi}_i(x)$, $i = 0, 1, 2, \cdots, ne$ are piecewise linear basis function satisfying

$$\hat{\phi}_i(x_j) = \begin{cases} 1, & \text{if } j = i, \\ 0, & \text{otherwise.} \end{cases} \tag{7.35}$$

See Figure 7.2 for a plot of such functions. The functions $\hat{\phi}_0$ and $\hat{\phi}_{ne}$ are non-zero over only one element with all other functions $\hat{\phi}_1, \cdots, \hat{\phi}_{ne}$ being non-zero over two adjacent elements. The function $\hat{\phi}_i$ is non-zero on (x_{i-1}, x_{i+1}) and we have

$$\hat{\phi}_i(x) = \begin{cases} (x - x_{i-1})/h_i, & \text{if } x_{i-1} \le x < x_i, \\ (x_{i+1} - x)/h_{i+1}, & \text{if } x_i \le x \le x_{i+1}. \end{cases} \tag{7.36}$$

where $h_i = x_i - x_{i-1}$ for $i = 1, 2, \cdots, ne$. We define

$$\bar{V}_h = \mathrm{span}\{\hat{\phi}_0, \cdots, \hat{\phi}_{ne}\} \tag{7.37}$$

as the space of all such piecewise linear functions based on these points and let

$$V_h = \mathrm{span}\{\hat{\phi}_1, \cdots, \hat{\phi}_{ne-1}\} = \{v \in \bar{V}_h : \quad v(0) = v(1) = 0\} \tag{7.38}$$

or

$$V_h = \mathrm{span}\{\hat{\phi}_1, \cdots, \hat{\phi}_{ne}\} = \{v \in \bar{V}_h : \quad v(0) = 0\} \tag{7.39}$$

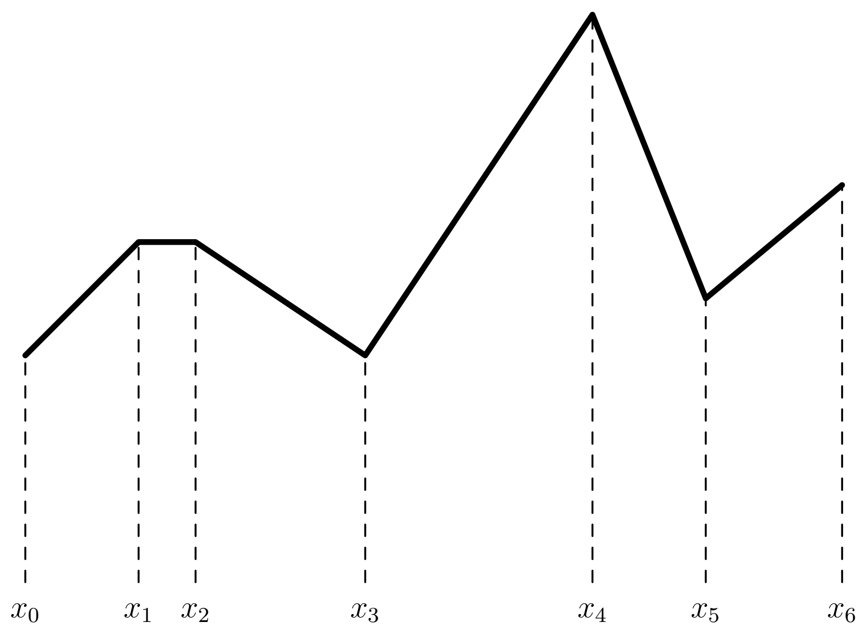


Figure 7.1: Plot of a piecewise linear function based on the points $x_0, x_1, \cdots, x_6$.

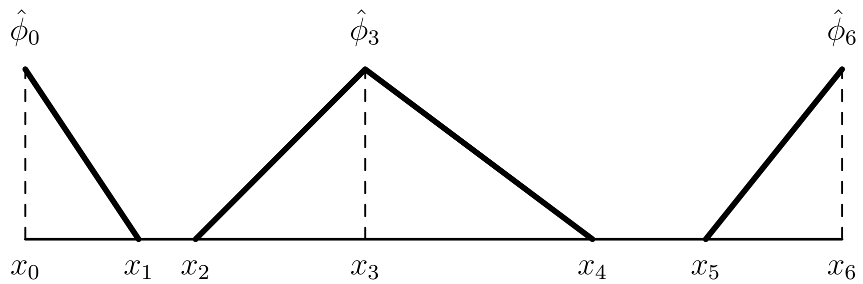


Figure 7.2: For $ne = 6$ a plot of $\hat{\phi}_0(x)$, $\hat{\phi}_3(x)$ and $\hat{\phi}_6(x)$.

depending on the problem being considered, i.e. depending on whether we have essential boundary conditions at both $x = 0$ and $x = 1$ (the first definition, i.e. (7.25)) or we have an essential boundary condition at $x = 0$ and a natural boundary condition at $x = 1$ (the second definition, i.e. (7.26)). Also let n denote the dimension of this space which is $n = ne - 1$ or $n = ne$ depending on which space is being used.

We now apply Galerkin's method as described in section 6 and start by noting some properties of the linear system of equations that are generated when these basis functions are used. As before, for any of the problems considered the approximate weak form involves the following

$$U = \sum_1^n c_j \hat{\phi}_j \quad \text{where} \quad a(U, \hat{\phi}_i) = (f, \hat{\phi}_i) \quad \text{for } i = 1, 2, \cdots, n \tag{7.40}$$

which gives the linear system

$$\sum_{j=1}^n a(\hat{\phi}_j, \hat{\phi}_i) c_j = (f, \hat{\phi}_i), \quad i = 1, 2, \cdots, n. \tag{7.41}$$

The coefficient $c_j = U(x_j)$. As before we can write the linear system in matrix-vector form as

$$K\underline{c} = \underline{b}, \quad K = (k_{ij}), \quad k_{ij} = a(\hat{\phi}_j, \hat{\phi}_i), \quad \underline{b} = (b_i), \quad b_i = (f, \hat{\phi}_i) \tag{7.42}$$

and, as before, the symmetry and positive definite properties of $a(.,.)$ gives us a symmetric and positive definite matrix K. There are more things that can be said about K by taking note of where the functions $\hat{\phi}_i$ are zero and non-zero. As $\hat{\phi}_i$ is only non-zero on (x_{i-1}, x_{i+1}) and $\hat{\phi}_j$ is only non-zero on (x_{j-1}, x_{j+1}) the term $a(\hat{\phi}_i, \hat{\phi}_j)$ is only non-zero when these two intervals intersect. Thus on row i of the matrix we only get 3 non-zero entries

$$k_{i,i-1} = a(\hat{\phi}_i, \hat{\phi}_{i-1}), \quad k_{i,i} = a(\hat{\phi}_i, \hat{\phi}_i), \quad k_{i,i+1} = a(\hat{\phi}_i, \hat{\phi}_{i+1}) \tag{7.43}$$

and the matrix K is thus tridiagonal. In full we have

$$K = \begin{pmatrix} k_{11} & k_{12} & \cdots & & & 0 \\ k_{21} & k_{22} & k_{23} & & & \vdots \\ \vdots & k_{32} & k_{33} & k_{34} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & k_{n-1,n} \\ 0 & \cdots & & & k_{n,n-1} & k_{nn} \end{pmatrix}. \tag{7.44}$$

To give more detail we consider the entries of K for specific forms of the bilinear expression $a(.,.)$ and we consider the organization of this in two different ways involving first a "direct approach" in which we consider each entry in the final system in turn and then we follow this by an "element-by-element" description which is usually done in practice to cope with any type of element. The element-by-element way of doing things involves ordering the calculations so that all the computations concerning an element (x_{i-1}, x_i) are done together. The different element contributions are then "assembled" to give the final system of equations. To keep things specific we consider the direct approach for the simplest case $Lu = -u''$ for which we have

$$a(u, v) = \int_0^1 u'v' \,\mathrm{d}x. \tag{7.45}$$

The direct approach for $Lu = -u'' = f$

The integrand involves the derivative of $\hat{\phi}_i$ which is the piecewise constant function given by

$$\hat{\phi}_i'(x) = \begin{cases} 1/h_i, & x_{i-1} < x < x_i \text{ where } h_i = x_i - x_{i-1}, \\ -1/h_{i+1}, & x_i < x < x_{i+1} \text{ where } h_{i+1} = x_{i+1} - x_i. \end{cases} \tag{7.46}$$

The row i entries of the matrix are thus

$$a(\hat{\phi}_i, \hat{\phi}_i) = \int_0^1 {\hat{\phi}_i'}^2 \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} {\hat{\phi}_i'}^2 \,\mathrm{dx} + \int_{x_i}^{x_{i+1}} {\hat{\phi}_i'}^2 \,\mathrm{dx} = \frac{1}{h_i} + \frac{1}{h_{i+1}}, \tag{7.47}$$

$$a(\hat{\phi}_i, \hat{\phi}_{i-1}) = \int_0^1 \hat{\phi}_i'\hat{\phi}_{i-1}' \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \hat{\phi}_i'\hat{\phi}_{i-1}' \,\mathrm{dx} = -\frac{1}{h_i} \tag{7.48}$$

$$a(\hat{\phi}_i, \hat{\phi}_{i+1}) = \int_0^1 \hat{\phi}_i'\hat{\phi}_{i+1}' \,\mathrm{d}x = \int_{x_i}^{x_{i+1}} \hat{\phi}_i'\hat{\phi}_{i+1}' \,\mathrm{dx} = -\frac{1}{h_{i+1}}. \tag{7.49}$$

The ith equation in (7.42) for the problem $-u'' = f$ is thus

$$-\frac{1}{h_i}U(x_{i-1}) + \left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right)U(x_i) - \frac{1}{h_{i+1}}U(x_{i+1}) = \int_{x_{i-1}}^{x_{i+1}} f(x)\hat{\phi}_i(x) \,\mathrm{dx}. \tag{7.50}$$

In the case of a uniform mesh we have $h_i = h$ for $i = 1, 2, \cdots, ne$ and this becomes, after dividing by h,

$$\frac{-U(x_{i-1}) + 2U(x_i) - U(x_{i+1})}{h^2} = \frac{1}{h}\int_{x_{i-1}}^{x_{i+1}} f(x)\hat{\phi}_i(x)\,\mathrm{dx}. \tag{7.51}$$

This is almost the same as what you get when you use the finite difference method to approximate $-u'' = f$ except that in the finite difference method with $U_i \approx u(x_i)$ we would usually take the scheme as

$$\frac{-U_{i-1} + 2U_i - U_{i+1}}{h^2} = f(x_i). \tag{7.52}$$

The difference is that the right hand side in (7.51) has an average of f in the vicinity (x_{i-1}, x_{i+1}) of x_i whereas the usual difference scheme just takes the function value at x_i.

The element-by-element approach for $Lu = -u'' = f$

An element-by-element way of doing things is to note that on each element (x_{i-1}, x_i) we only have two linear functions to consider given by

$$\tilde{\phi}_1(x) = \frac{x_i - x}{h_i}, \quad \tilde{\phi}_2(x) = \frac{x - x_{i-1}}{h_i} \quad \text{where} \quad h_i = x_i - x_{i-1}. \tag{7.53}$$

The element contributions just involve computing

$$a(\tilde{\phi}_k, \tilde{\phi}_l)|_i = \int_{x_{i-1}}^{x_i} \tilde{\phi}_k'\tilde{\phi}_l'\,\mathrm{dx}, \quad 1 \le k, l \le 2 \tag{7.54}$$

and

$$(f, \tilde{\phi}_l)|_i = \int_{x_{i-1}}^{x_i} f\tilde{\phi}_l\,\mathrm{dx}, \quad 1 \le l \le 2. \tag{7.55}$$

There are names for these quantities. The matrix

$$K_i = \begin{pmatrix} a(\tilde{\phi}_1, \tilde{\phi}_1)|_i & a(\tilde{\phi}_1, \tilde{\phi}_2)|_i \\ a(\tilde{\phi}_2, \tilde{\phi}_1)|_i & a(\tilde{\phi}_2, \tilde{\phi}_2)|_i \end{pmatrix} \tag{7.56}$$

is known as the **element matrix** and the vector

$$\underline{b}_i = \begin{pmatrix} (f, \tilde{\phi}_1)|_i \\ (f, \tilde{\phi}_2)|_i \end{pmatrix} \tag{7.57}$$

is known as the **element vector**. Since we have

$$\tilde{\phi}_1' = \frac{-1}{h_i} \quad \text{and} \quad \tilde{\phi}_2' = \frac{1}{h_i} \tag{7.58}$$

we get

$$\int_{x_{i-1}}^{x_i} {\tilde{\phi}_1'}^2\,\mathrm{dx} = \int_{x_{i-1}}^{x_i} {\tilde{\phi}_1'}^2\,\mathrm{dx} = (x_i - x_{i-1})\frac{1}{{h_i}^2} = \frac{1}{h_i}, \quad \text{and} \quad \int_{x_{i-1}}^{x_i} \tilde{\phi}_1'\tilde{\phi}_2'\,\mathrm{dx} = \frac{-1}{h_i}. \tag{7.59}$$

The element matrix is hence

$$K_i = \frac{1}{h_i}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{7.60}$$

This particular element matrix is known as the **element stiffness matrix** associated with the linear element.

The final matrix K of the linear system (7.42) can be obtained by appropriately combining all these element matrices K_i and likewise the vector $\underline{b}$ can be obtained by appropriately combining the 2×1 vectors $\underline{b}_i$. The "appropriately combining" operation is known as **assembly** and this depends on how the local functions $\tilde{\phi}_1$, $\tilde{\phi}_2$ are related to the globally defined functions $\hat{\phi}_i$, $i = 0, 1, \cdots, ne$. To get the precise form of K we also need to take account of the boundary conditions at the ends $x = 0$ and $x = 1$.

Now on (x_{i-1}, x_i), the two non-zero globally defined functions are $\hat{\phi}_{i-1}(x)$ and $\hat{\phi}_i(x)$ and the connection between these and $\tilde{\phi}_1(x)$ and $\tilde{\phi}_2(x)$ are as follows

$$\begin{aligned} \hat{\phi}_{i-1}(x) &= \tilde{\phi}_1(x), \quad x_{i-1} < x < x_i, && (7.61)\\ \hat{\phi}_i(x) &= \tilde{\phi}_2(x), \quad x_{i-1} < x < x_i. && (7.62) \end{aligned}$$

The 2 local basis functions $\tilde{\phi}_1$ and $\tilde{\phi}_2$ hence give the right hand part of $\hat{\phi}_{i-1}$ and the left hand part of $\hat{\phi}_i$ respectively. To illustrate the assembly in the case of $ne = 3$ elements we do the following. We form the matrix

$$\hat{K} = \begin{pmatrix} 1/h_1 & -1/h_1 & 0 & 0 \\ -1/h_1 & 1/h_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/h_2 & -1/h_2 & 0 \\ 0 & -1/h_2 & 1/h_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1/h_3 & -1/h_3 \\ 0 & 0 & -1/h_3 & 1/h_3 \end{pmatrix}. \quad (7.63)$$

That is we appropriately expand each of the element matrices K_1, K_2 and K_3 and then add the expanded matrices in the usual way. This gives a 4×4 matrix $\hat{K}$ corresponding to the 4 basis functions $\hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3$ associated with the 4 node points $0 = x_0 < x_1 < x_2 < x_3 = 1$. We can similarly construct a vector $\hat{\underline{b}}$ given by

$$\hat{\underline{b}} = \begin{pmatrix} b_{11} \\ b_{12} \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ b_{21} \\ b_{22} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ b_{31} \\ b_{32} \end{pmatrix}, \quad \text{here} \quad \underline{b}_1 = \begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix} \text{ etc..} \quad (7.64)$$

The matrix $\hat{K}$ and $\hat{\underline{b}}$ are not the matrix K and vector $\underline{b}$ of the final system (7.42) as they involve terms which are not part of (7.42). If (7.42) corresponds to the case when we have essential boundary conditions at $x = 0$ and $x = 1$ then $n = 2$ and if (7.42) corresponds to the case when we have essential boundary conditions at $x = 0$ and a mixed condition at $x = 1$ (i.e. the natural boundary condition case) then $n = 3$. In (7.63) and (7.64) we have combined everything irrespective of whether it is needed and specifically in the case of $\hat{K}$ we have actually computed the matrix

$$\begin{pmatrix} a(\hat{\phi}_0, \hat{\phi}_0) & a(\hat{\phi}_0, \hat{\phi}_1) & a(\hat{\phi}_0, \hat{\phi}_2) & a(\hat{\phi}_0, \hat{\phi}_3) \\ a(\hat{\phi}_1, \hat{\phi}_0) & a(\hat{\phi}_1, \hat{\phi}_1) & a(\hat{\phi}_1, \hat{\phi}_2) & a(\hat{\phi}_1, \hat{\phi}_3) \\ a(\hat{\phi}_2, \hat{\phi}_0) & a(\hat{\phi}_2, \hat{\phi}_1) & a(\hat{\phi}_2, \hat{\phi}_2) & a(\hat{\phi}_2, \hat{\phi}_3) \\ a(\hat{\phi}_3, \hat{\phi}_0) & a(\hat{\phi}_3, \hat{\phi}_1) & a(\hat{\phi}_3, \hat{\phi}_2) & a(\hat{\phi}_3, \hat{\phi}_3) \end{pmatrix}. \quad (7.65)$$

The matrix that we need is

$$\begin{pmatrix} a(\hat{\phi}_1, \hat{\phi}_1) & a(\hat{\phi}_1, \hat{\phi}_2) \\ a(\hat{\phi}_2, \hat{\phi}_1) & a(\hat{\phi}_2, \hat{\phi}_2) \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} a(\hat{\phi}_1, \hat{\phi}_1) & a(\hat{\phi}_1, \hat{\phi}_2) & a(\hat{\phi}_1, \hat{\phi}_3) \\ a(\hat{\phi}_2, \hat{\phi}_1) & a(\hat{\phi}_2, \hat{\phi}_2) & a(\hat{\phi}_2, \hat{\phi}_3) \\ a(\hat{\phi}_3, \hat{\phi}_1) & a(\hat{\phi}_3, \hat{\phi}_2) & a(\hat{\phi}_3, \hat{\phi}_3) \end{pmatrix} \quad (7.66)$$

which we obtain by removing the appropriate rows and columns of $\hat{K}$. Similarly the vector $\underline{\hat{b}}$ corresponds to

$$\begin{pmatrix} (f, \hat{\phi}_0) \\ (f, \hat{\phi}_1) \\ (f, \hat{\phi}_2) \\ (f, \hat{\phi}_3) \end{pmatrix} \tag{7.67}$$

and we similarly need to remove the appropriate rows to get

$$\begin{pmatrix} (f, \hat{\phi}_1) \\ (f, \hat{\phi}_2) \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} (f, \hat{\phi}_1) \\ (f, \hat{\phi}_2) \\ (f, \hat{\phi}_3) \end{pmatrix}. \tag{7.68}$$

Using a standard linear element

It is common in finite elements to express the local basis functions with respect to a standard element with every specific element being obtained by using a mapping from this standard element. For one dimensional problems there is not too much to choose between taking $[-1, 1]$ or $[0, 1]$ as the standard element. If we take $[0, 1]$ as the standard element then the 2 local basis functions are

$$\phi_1(s) = 1 - s \quad \text{and} \quad \phi_2(s) = s, \quad 0 < s < 1 \tag{7.69}$$

shown in Figure 7.3. These functions are known as the **linear shape functions**. We can map from this standard element $[0, 1]$ to a specific element $[x_{i-1}, x_i]$ by using

$$x(s) = x_{i-1} + (x_i - x_{i-1})s = x_{i-1} + h_i s. \tag{7.70}$$

With this set-up we have

$$\tilde{\phi}_1(x(s)) = \phi_1(s) = 1 - s \quad \text{and} \quad \tilde{\phi}_2(x(s)) = \phi_2(s) = s. \tag{7.71}$$

Now

$$\frac{d\tilde{\phi}_i}{dx} = \frac{ds}{dx}\frac{d\phi_i}{ds} = \frac{1}{dx/ds}\frac{d\phi_i}{ds} = \frac{1}{h_i}\phi_i'(s). \tag{7.72}$$

Thus doing things this way we have

$$a(\hat{\phi}_k, \hat{\phi}_l)|_i = \int_0^1 \frac{1}{h_i^2}\phi_k'\phi_l' h_i \, \mathrm{d}x = \frac{1}{h_i}\int_0^1 \phi_k'\phi_l' \, \mathrm{d}x \tag{7.73}$$

and

$$(f, \tilde{\phi}_l)|_i = \int_0^1 f(x(s))\phi_l(s) h_l \, \mathrm{d}x. \tag{7.74}$$

With this way of organising the computation of K_i and $\underline{b}_i$ the local basis functions are kept as simple as possible with the detail of the actual element being considered in the mapping (7.70).

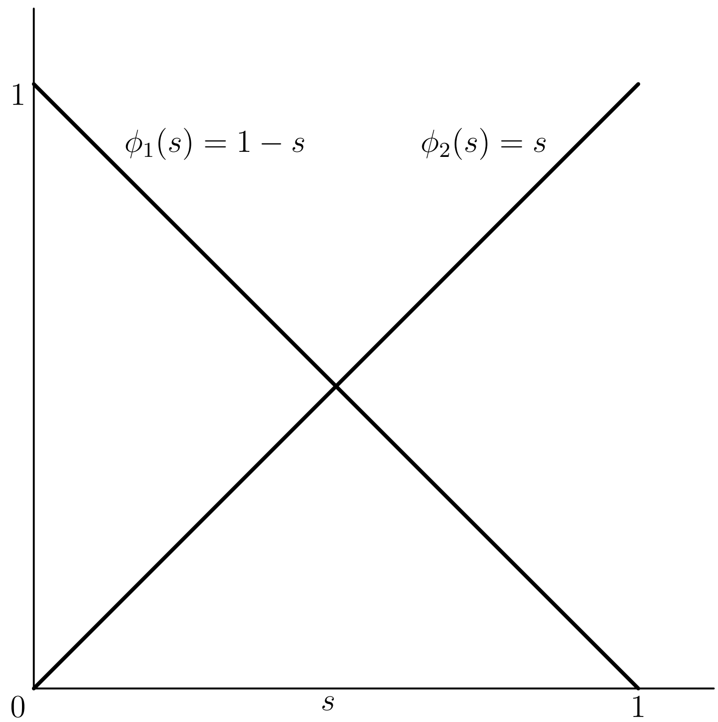


Figure 7.3: The linear shape functions on $[0, 1]$.

Element matrices for $Lu = -u'' + u = f$

For the problem $Lu = -u'' + u = f$, $u(0) = u(1) = 0$ the bilinear form is

$$a(u, v) = \int_0^1 u'v' + uv \, \mathrm{d}x. \tag{7.75}$$

The element matrix for an element $x_{i-1} < x < x_i$ now involves

$$a(\tilde{\phi}_k, \tilde{\phi}_l)|_i = \int_0^1 \left(\frac{1}{h_i^2} \phi_k' \phi_l' + \phi_k \phi_l \right) h_i \, \mathrm{d}x, \quad 1 \le k, l \le 2 \tag{7.76}$$

which naturally involves two parts. The first part involving $\phi_k'\phi_l'$ gives the 2×2 matrix

$$\frac{1}{h_i} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{7.77}$$

as before. This is the element stiffness matrix. The second part involving $\phi_k \phi_l$ gives the 2×2 matrix

$$h_i \int_0^1 \begin{pmatrix} (1-s)^2 & s(1-s) \\ s(1-s) & s^2 \end{pmatrix} \, \mathrm{d}x. \tag{7.78}$$

Now

$$\int_0^1 (1-s)^2 \, \mathrm{d}x = \int_0^1 s^2 \, \mathrm{d}x = \frac{1}{3}, \quad \int_0^1 s(1-s) \, \mathrm{d}x = \int_0^1 s - s^2 \, \mathrm{d}x = \frac{1}{2} = \frac{1}{3} = \frac{1}{6}. \tag{7.79}$$

This 2×2 matrix is hence

$$\frac{h_i}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{7.80}$$

This is known as the **element mass matrix**. The element matrix is the sum of these two matrices.

$$K_i = \frac{1}{h_i}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{h_i}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{7.81}$$

These matrices are then assembled, as before, to get the final matrix K of the linear system.

Example

Consider again the problem

$$-u'' + u = 1, \quad u(0) = 0, \quad u'(1) = 0 \tag{7.82}$$

discussed earlier and consider now a piecewise linear FEM approximation. To keep things manageable for hand calculation we use $ne = 1$ element and we use $ne = 2$ elements and we report the results obtained on a computer with $ne > 2$ elements. We also assume a uniform mesh with $h = 1/ne$.

1. $ne = 1$, $h = 1$.

   As the function $f = 1$ the element matrix (there is only 1 element matrix here) is

$$K_1 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{1}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \tag{7.83}$$

   and the element vector is

$$\underline{b}_1 = \int_0^1 \begin{pmatrix} 1-s \\ s \end{pmatrix}\,\mathrm{d}x = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{7.84}$$

   As $ne = 1$, $\hat{K} = K_1$ and we get the 1×1 matrix K that we need by applying the boundary condition at $x = 0$. By removing row 1 and column 1 we get

$$K = 1 + \frac{2}{6} = \frac{4}{3}, \quad \underline{b} = \frac{1}{2}. \tag{7.85}$$

   Hence

$$KU(x_1) = b_1 \quad \text{gives} \quad U(x_1) = 3/8. \tag{7.86}$$

   This is of course the same as what was done previously when we obtained $U(x) = 3x/8$.

2. $ne = 2$, $h = 1/2$.

   Both element matrices are the same and we have

$$K_1 = K_2 = 2\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{1}{12}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{7.87}$$

   Because f is constant the element vectors are the same. We have

$$\underline{b}_1 = \underline{b}_2 = \frac{1}{2}\int_0^1 \begin{pmatrix} 1-s \\ s \end{pmatrix}\,\mathrm{d}x = \frac{1}{4}\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{7.88}$$

Assembling the matrices gives

$$\begin{aligned}\hat{K} &= \begin{pmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 2 \end{pmatrix} + \frac{1}{12}\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \frac{1}{12}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 2 & -2 & 0 \\ -2 & 4 & -2 \\ 0 & -2 & 2 \end{pmatrix} + \frac{1}{12}\begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix}. \end{aligned} \tag{7.89}$$

Applying the boundary conditions gives

$$K = \begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix} + \frac{1}{12}\begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 13/3 & -23/12 \\ -23/12 & 13/6 \end{pmatrix}. \tag{7.90}$$

Similarly assembling the vector $\hat{\underline{b}}$ gives

$$\frac{1}{4}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}. \tag{7.91}$$

Applying the boundary conditions reduces this to

$$\underline{b} = \frac{1}{4}\begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{7.92}$$

The linear system is thus

$$\begin{pmatrix} 13/3 & -23/12 \\ -23/12 & 13/6 \end{pmatrix}\begin{pmatrix} U(x_1) \\ U(x_2) \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/4 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 52 & -23 \\ -23 & 26 \end{pmatrix}\begin{pmatrix} U(x_1) \\ U(x_2) \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \end{pmatrix}. \tag{7.93}$$

The numbers are messy now for hand calculation. We obtain

$$U(x_1) = 0.27339 \quad \text{and} \quad U(x_2) = 0.35723. \tag{7.94}$$

A plot of the piecewise linear function obtained with $ne = 2$ elements together with the exact solution is shown in Figure 7.4. The single quadratic does better here as a result of the fairly smooth form of the solution.

3. Results for $ne > 2$.

   Using a short computer program you can very quickly generate the approximations for $ne = 4, 8, 16, \cdots$ and the plot for $ne = 4$ is shown in Figure 7.5.

   None of the approximations shown in Figures 7.4 or 7.5 satisfies the natural boundary condition $u'(1) = 0$ but as the plots show they get closer and closer to satisfying this condition as the mesh is refined.

7.4 Quadratic elements

With quadratic elements the FEM approximation $U(x)$ is a quadratic polynomial on each of the ne elements $\Omega_1, \cdots, \Omega_{ne}$ and on each Ω_i we need 3 terms, e.g. $\alpha_0^{(i)} + \alpha_1^{(i)}x + \alpha_2^{(i)}x^2$, to

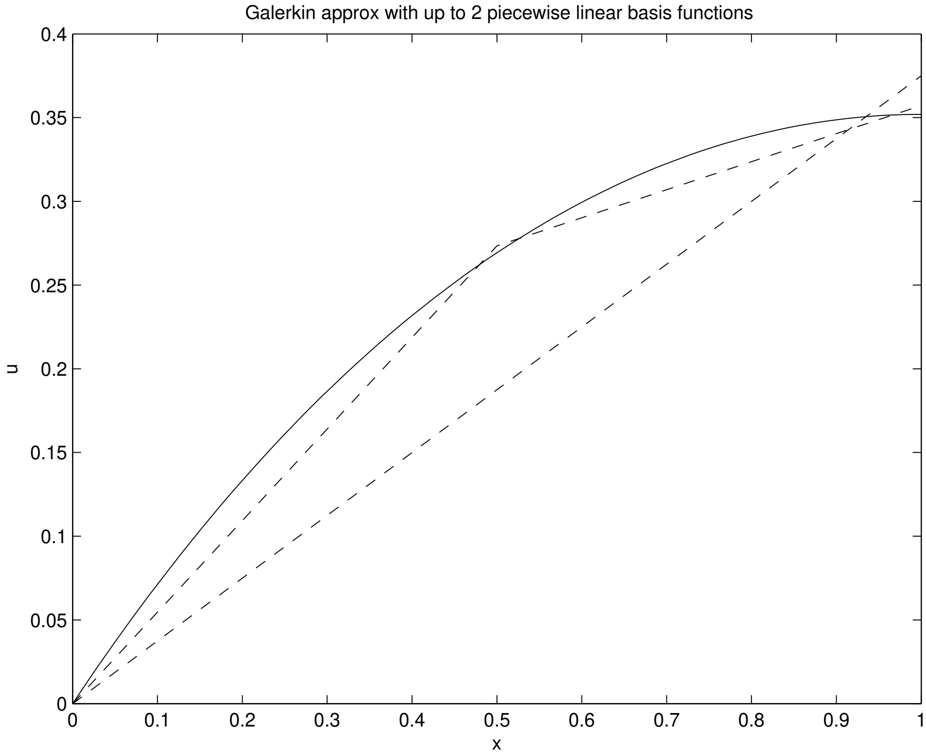


Figure 7.4: The finite element solution to $-u'' + u = 1$, $u(0) = 0$, $u'(1) = 0$ corresponding to $ne = 1$ and $ne = 2$. The exact solution is shown with a solid line, -- is used for the finite element solutions.

describe each of these quadratics. We can arrange for this by having 3 nodes associated with each element Ω_i. If we have in total $2ne + 1$ nodes

$$0 = x_0 < x_1 < x_2 < \cdots < x_{2ne} = 1 \tag{7.95}$$

then each element is

$$\Omega_i = (x_{2i-2}, x_{2i}). \tag{7.96}$$

We will assume here that the intermediate point x_{2i-1} is at the centre of the element, i.e.

$$x_{2i-1} = \frac{x_{2i-2} + x_{2i}}{2}. \tag{7.97}$$

The simplest way to describe the basis functions is to use a standard element which we choose to take here as $[-1, 1]$ with 0 as the mid-point. -1, 0 and 1 are the local coordinates of the nodes with respect to this standard element and the local basis functions are given by

$$\phi_1(s) = \frac{1}{2}s(s-1), \tag{7.98}$$

$$\phi_2(s) = 1 - s^2, \tag{7.99}$$

$$\phi_3(s) = \frac{1}{2}s(s+1). \tag{7.100}$$

These are the standard Lagrange basis functions taking the value 1 at one of the local nodes and 0 at all other nodes. Plots of these are shown in Figure 7.6.

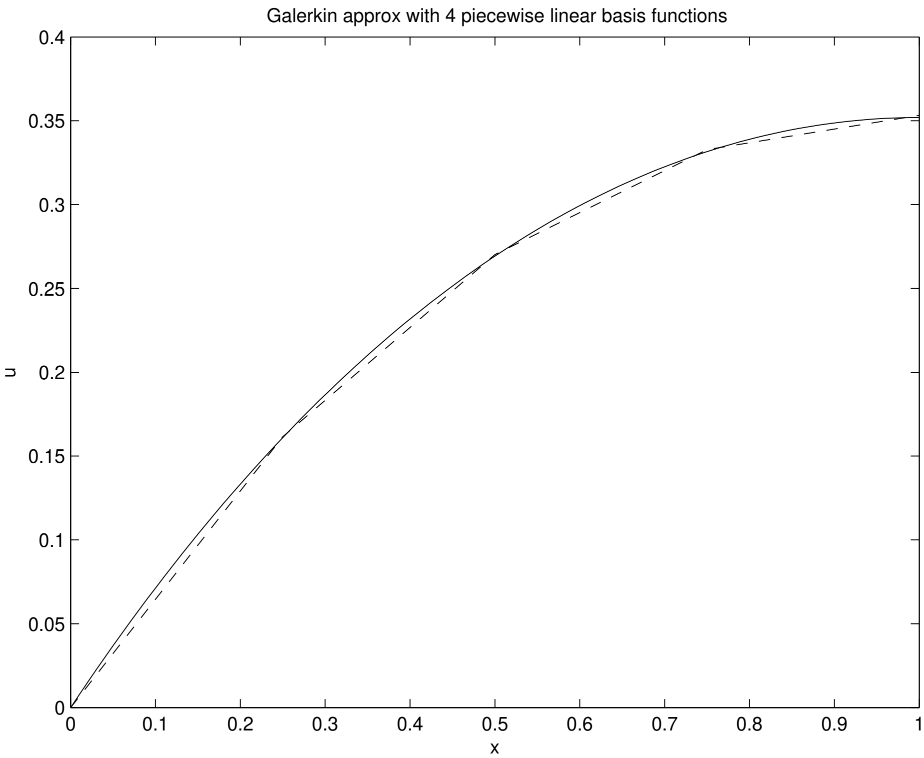


Figure 7.5: The finite element solution to $-u'' + u = 1$, $u(0) = 0$, $u'(1) = 0$ corresponding to $ne = 4$ elements. The exact solution is shown with a solid line, -- is used for the finite element solution.

The mapping between $[-1, 1]$ and $[x_{2i-1}, x_{2i}]$ can be written as

$$\begin{aligned} x(s) &= x_{2i-2}\phi_1(s) + x_{2i-1}\phi_2(s) + x_{2i}\phi_3(s) && (7.101)\\ &= x_{2i-2}\frac{s(s-1)}{2} + x_{2i-1}(1-s^2) + x_{2i}\frac{s(s+1)}{2} && (7.102)\\ &= x_{2i-1} + h_i s \quad \text{where} \quad h_i = \frac{x_{2i} - x_{2i-2}}{2}. && (7.103) \end{aligned}$$

The finite element function U can be similarly expressed in terms of the functions ϕ_1, ϕ_2, ϕ_3, we have

$$U(x(s)) = U(x_{2i-2})\phi_1(s) + U(x_{2i-1})\phi_2(s) + U(x_{2i})\phi_3(s). \qquad (7.104)$$

By comparing (7.101) and (7.104) we see that both the mapping between the standard element $[-1, 1]$ and the actual element $[x_{2i-2}, x_{2i}]$ and the form of the approximation on an element are both described in a similar way using the same standard basis functions ϕ_1, ϕ_2 and ϕ_3. All the constructions can be done using these functions ϕ_1, ϕ_2 and ϕ_3 and there is no need to actually explicitly form

$$\tilde{\phi}_j(x) = \phi_j(s(x)), \quad j = 1, 2, 3. \qquad (7.105)$$

Another point to note from (7.104) is that as $s \to 1$, $x(s) \to x_{2i}$ and the function value tends to $U(x_{2i})$. In the neighbouring element which has x_{2i} as an edge point the function similarly

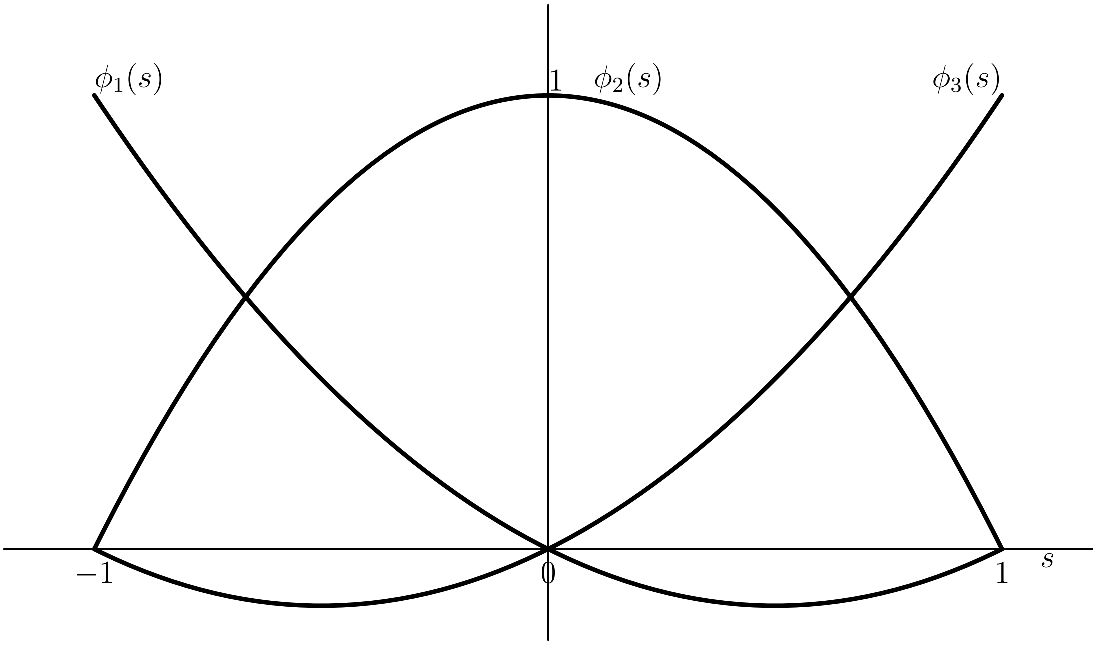


Figure 7.6: Plots of the quadratics ϕ_1, ϕ_2, ϕ_3.

tends to $U(x_{2i})$ as $x \to x_{2i}$ so ensuring that $U(x)$ is indeed continuous from an element to a neighbouring element.

We consider again the problem $Lu = -u'' = f$. The element stiffness matrix is now of size 3×3 with the k, l entry being

$$a(\tilde{\phi}_k, \tilde{\phi}_l)|_i = \int_{x_{2i-2}}^{x_{2i}} \tilde{\phi}_i' \tilde{\phi}_j' \,\mathrm{dx}. \tag{7.106}$$

From (7.105) the derivatives with respect to x can be expressed with respect to derivatives with respect to the local variable s, we have

$$\tilde{\phi}' = \frac{d\tilde{\phi}_j}{dx} = \frac{ds}{dx}\frac{d\phi_j}{ds} = \frac{1}{h_i}\phi_j'. \tag{7.107}$$

Hence we get

$$a(\tilde{\phi}_k, \tilde{\phi}_l)|_i = \frac{1}{h_i^2}\int_{-1}^{1} \phi_i'\phi_j' \frac{dx}{ds}\,\mathrm{ds} = \frac{1}{h_i}\int_{-1}^{1} \phi_i'\phi_j' \,\mathrm{ds}. \tag{7.108}$$

There is more work here to compute the element matrix by hand calculation than was the case with linear elements but it is just about manageable. To do this we need the derivatives of the functions and these are as follows.

$$\begin{array}{ll} \phi_1 = \frac{1}{2}s(s-1), & \phi_1' = s - \frac{1}{2}. \\ \phi_2 = 1 - s^2, & \phi_2' = -2s, \\ \phi_3 = \frac{1}{2}s(s+1), & \phi_3' = s + \frac{1}{2}. \end{array} \tag{7.109}$$

The integrands are thus

$$\begin{pmatrix} {\phi_1'}^2 & \phi_1'\phi_2' & \phi_1'\phi_3' \\ \phi_2'\phi_1' & {\phi_2'}^2 & \phi_2'\phi_3' \\ \phi_3'\phi_1' & \phi_3'\phi_2' & {\phi_3'}^2 \end{pmatrix} = \begin{pmatrix} (s-1/2)^2 & (-2s^2+s) & s^2 - 1/4 \\ (-2s^2+s) & 4s^2 & -2s^2 - s \\ s^2 - 1/4 & -2s^2 - s & (s+1/2)^2 \end{pmatrix}. \tag{7.110}$$

For the integrations we observe that

$$\int_{-1}^{1} \mathrm{ds} = 2, \quad \int_{-1}^{1} s\,\mathrm{ds} = 0, \quad \int_{-1}^{1} s^2\,\mathrm{ds} = \frac{2}{3}. \tag{7.111}$$

Hence for the diagonal entries

$$\int_{-1}^{1} (s-1/2)^2\,\mathrm{ds} = \int_{-1}^{1} (s+1/2)^2\,\mathrm{ds} = \frac{2}{3} + \frac{2}{4} = \frac{7}{6} \quad \text{and} \quad \int_{-1}^{1} 4s^2\,\mathrm{ds} = \frac{8}{3}. \tag{7.112}$$

For the off-diagonal entries, and note that we have symmetry,

$$\int_{-1}^{1} -2s^2 + s\,\mathrm{ds} = \int_{-1}^{1} -2s^2 - s\,\mathrm{ds} = -\frac{4}{3} \quad \text{and} \quad \int_{-1}^{1} s^2 - 1/4\,\mathrm{ds} = \frac{2}{3} - \frac{2}{4} = \frac{1}{6}. \tag{7.113}$$

Thus the element stiffness matrix is

$$\frac{1}{6h_i} \begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix}. \tag{7.114}$$

As before we need to assemble these 3×3 matrices to get a matrix corresponding to $\hat{K}$ and we then need to take account of the boundary conditions to get the matrix K as in the case of linear elements. As we now have $2ne + 1$ nodes $0 = x_0 < x_1 < \cdots < x_{2ne} = 1$ for the ne elements the matrix $\hat{K}$ is now of size $(2ne + 1) \times (2ne + 1)$ and as before it is symmetric and banded. If we let $\hat{\phi}_0, \hat{\phi}_1, \cdots, \hat{\phi}_{2ne}$ denote the "globally defined" basis functions satisfying the usual Lagrange condition

$$\hat{\phi}_k(x_j) = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise,} \end{cases} \tag{7.115}$$

a function such as $\hat{\phi}_{2i}$ is non-zero only on the two elements (x_{2i-2}, x_{2i}) and (x_{2i}, x_{2i+2}) and hence we only need to consider

$$a(\hat{\phi}_{2i}, \hat{\phi}_j), \quad j = 2i-2, 2i-1, 2i, 2i+1, 2i+2. \tag{7.116}$$

The matrix $\hat{K}$ and the final matrix K after the boundary conditions have been imposed can hence have up to 5 non-zero entries on each row and we get a matrix with a band width of 5. The greater approximating property of quadratics as opposed to linears is hence at the expense of getting a matrix with a larger bandwidth, the bandwidth was only 3 with linears.

7.5 Some remarks about element stiffness matrices

We have now constructed an element stiffness matrix for a linear element $[x_{i-1}, x_i]$ and for a quadratic element $[x_{2i-2}, x_{2i}]$ with x_{2i-1} at the mid-point. We have respectively

$$\frac{1}{h_i} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \frac{1}{6h_i} \begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix}. \tag{7.117}$$

Both matrices are symmetric because $a(.,.)$ is symmetric and in both cases the sum of the entries on each row is 0. This is always the case for an element stiffness matrix when we use basis functions of the Lagrange type. This is because if $\phi_1, \cdots, \phi_n$ are the locally defined basis functions then

$$\phi_1 + \cdots + \phi_n = 1 \tag{7.118}$$

and thus

$$\phi_1' + \cdots + \phi_n' = 0. \tag{7.119}$$

As the matrix of integrands for the element stiffness matrices involves

$$\begin{pmatrix} {\phi_1'}^2 & \phi_1'\phi_2' & \cdots & \phi_1'\phi_n' \\ \phi_2'\phi_1' & {\phi_2'}^2 & \cdots & \phi_2'\phi_n' \\ \vdots & \vdots & & \vdots \\ \phi_n'\phi_1' & \phi_n'\phi_2' & \cdots & {\phi_n'}^2 \end{pmatrix} \tag{7.120}$$

it thus follows that the sum of entries on a row is 0. Another way of stating this is that

$$K_i \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{7.121}$$

In particular this tells us that K_i is a singular matrix.

A similar remark applies to the global matrix $\hat{K}$ when it is assembled using only stiffness matrices. The final matrix K that we get after applying the boundary conditions is however non-singular.

7.6 Cubic elements for fourth order problems

For fourth order problems such as $Lu = u'''' = f$ we get a bilinear form

$$a(u, v) = \int_0^1 u''v'' \,\mathrm{d}x. \tag{7.122}$$

This makes sense for functions $u, v \in H^2(0,1)$ and for piecewise defined polynomials we need $u, v \in C^1[0,1]$. With nodal points $0 = x_0 < x_1 < \cdots < x_{ne} = 1$ and an interval (x_{i-1}, x_i) the easiest way to construct a function which is continuously differentiable across the element edges is to use cubic Hermite elements. In $x_{i-1} \le x \le x_i$ the approximation $U(x)$ is a cubic which can be expressed in the form

$$U(x) = U(x_{i-1})\tilde{\phi}_1(x) + U'(x_{i-1})\tilde{\phi}_2(x) + U(x_i)\tilde{\phi}_3(x) + U'(x_i)\tilde{\phi}_4(x) \tag{7.123}$$

where $\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3, \tilde{\phi}_4$ are cubics satisfying the conditions shown below.

$$\begin{array}{ccccc} & \tilde{\phi}_j(x_{i-1}) & \tilde{\phi}_j'(x_{i-1}) & \tilde{\phi}_j(x_i) & \tilde{\phi}_j'(x_i) \\ j=1 & 1 & 0 & 0 & 0 \\ j=2 & 0 & 1 & 0 & 0 \\ j=3 & 0 & 0 & 1 & 0 \\ j=4 & 0 & 0 & 0 & 1 \end{array} \tag{7.124}$$

To describe these functions we consider first a standard element $[-1, 1]$ and functions ϕ_1, ϕ_2, ϕ_3, ϕ_4 for this element.

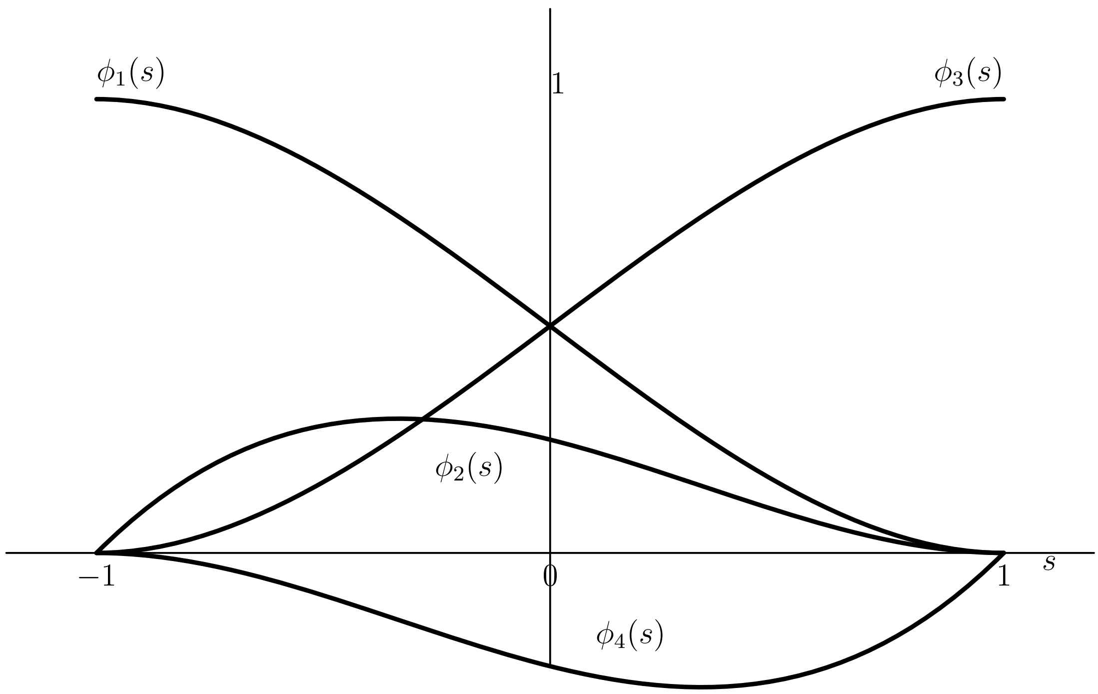


Figure 7.7: The 4 cubic basis functions on $[-1, 1]$.

First for the function ϕ_2 we satisfy the conditions $\phi_2(-1) = \phi_2(1) = \phi_2'(1) = 0$ if the function $\phi_2(s)$ has the factor $(s+1)(s-1)^2$. Now

$$\frac{d}{ds}(s+1)(s-1)^2|_{s=-1} = ((s+1)2(s-1) + (s-1)^2)|_{s=-1} = 4. \tag{7.125}$$

Hence

$$\phi_2(s) = \frac{1}{4}(s+1)(s-1)^2. \tag{7.126}$$

Similarly

$$\phi_4(s) = \frac{1}{4}(s+1)^2(s-1). \tag{7.127}$$

We now consider $\phi_1(s)$. To satisfy $\phi_1(1) = \phi_1'(1) = 0$ the function has the factor $(s-1)^2$ and hence it can be written in the form

$$\phi_1(s) = (s-1)^2(A(s+1) + B). \tag{7.128}$$

$1 = \phi_1(-1) = 4B$.

$$\phi_1'(s) = (s-1)^2 A + 2(s-1)(A(s+1) + B), \quad 0 = \phi_1'(-1) = 4(A - B). \tag{7.129}$$

This gives

$$\phi_1(s) = \frac{(s+2)(s-1)^2}{4} \quad \text{and similarly} \quad \phi_3(s) = \frac{(s+1)^2(2-s)}{4}. \tag{7.130}$$

Plots of $\phi_1, \phi_2, \phi_3, \phi_4$ are shown in Figure 7.7.

To map between $[-1, 1]$ and $[x_{i-1}, x_i]$ we have

$$x(s) = \frac{(x_{i-1} + x_i)}{2} + \frac{h_i}{2}s, \quad h_i = x_i - x_{i-1} \tag{7.131}$$

giving

$$\frac{dx}{ds} = \frac{h_i}{2}, \quad \frac{ds}{dx} = \frac{2}{h_i}. \tag{7.132}$$

The cubic functions $\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3, \tilde{\phi}_4$ are defined by the relations

$$\tilde{\phi}_1(x(s)) = \phi_1(s), \tag{7.133}$$

$$\tilde{\phi}_2(x(s)) = \frac{h_i}{2}\phi_2(s), \quad \text{so that } \frac{d\tilde{\phi}_2}{dx}|_{x=-1} = 1, \tag{7.134}$$

$$\tilde{\phi}_3(x(s)) = \phi_3(s), \tag{7.135}$$

$$\tilde{\phi}_4(x(s)) = \frac{h_i}{2}\phi_4(s), \quad \text{so that } \frac{d\tilde{\phi}_4}{dx}|_{x=1} = 1. \tag{7.136}$$

With 4 basis functions $\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3, \tilde{\phi}_4$ to consider on each element the element matrices are of size 4×4, the element vector is of size 4×1 and as a consequence the global matrices $\hat{K}$ and K have in general 7 non-zero entries on each row, i.e. $\hat{K}$ and K have a bandwidth of 7.

7.7 Some remarks about numerical integration

In many practical situations the integrals involved in the expressions for $a(\tilde{\phi}_k, \tilde{\phi}_l)|_i$ and $(f, \tilde{\phi}_l)|_i$ cannot be done exactly and must instead by approximated. This can be done using any of the standard techniques which produces a sufficiently accurate approximation. You can use a rule such as Simpson's rule if this is sufficiently accurate for the element being considered. The preferred technique is however to use Gauss-Legendre quadrature as this usually gives the best accuracy with the least function evaluations. With the use of standard elements and with $[-1, 1]$ as the standard interval the integrals to be approximated together with the approximation are of the form

$$\int_{-1}^{1} g(s)\,\mathrm{ds} \approx \sum_{i=1}^{nq} w_i g(s_i). \tag{7.137}$$

Here there are nq quadrature points $-1 \le s_1 < s_2 < \cdots < s_{nq} \le 1$ with $w_1, w_2, \cdots, w_{nq}$ denoting the corresponding weights. With Gauss-Legendre quadrature the points $s_1, s_2, \cdots, s_{nq}$ are the roots of the Legendre polynomial of degree nq and it can be shown that the approximation in (7.137) is exactly correct for all polynomials of degree less than or equal to $2nq - 1$. The weights and points for these methods are already known and computations such as these can be done easily on a computer. A popular rule to use corresponds to $nq = 2$. In this case the Legendre polynomial is a multiple of $3s^2 - 1$ and the points and weights are

$$s_1 = -\frac{1}{\sqrt{3}}, \quad s_2 = \frac{1}{\sqrt{3}}, \quad w_1 = w_2 = 1. \tag{7.138}$$

This will integrate cubics exactly. As a comparison, Simpson's rule will also integrate cubics exactly but this involves the 3 points -1, 0 and 1, i.e. it requires 3 function evaluations.

7.8 Summary

1. $$H^1(0,1) := \{v : \quad v \text{ and } v' \text{ are integrable on } (0,1) \text{ and } \int_0^1 v^2 + v'^2 \, dx < \infty\} \tag{7.139}$$

   is a an example of a Sobolev space, it is the largest function space for which the weak problem has meaning. $H^1(0,1)$ is also a Hilbert space.

2. The FEM is Galerkin's method using piecewise polynomial functions which are in $C[0,1]$ for the second order ODEs and which are in $C^1[0,1]$ for the fourth order ODEs. With these functions the matrix K is banded.

   With respect to the standard element $[0,1]$ the linear basis functions are

   $$\phi_1(s) = 1 - s, \quad \phi_2(s) = s. \tag{7.140}$$

   With respect to the standard element $[-1,1]$ using the 3 points -1, 0 and 1 the quadratic basis functions are

   $$\phi_1(s) = \frac{s(s-1)}{2}, \quad \phi_2(s) = 1 - s^2, \quad \phi_3(s) = \frac{s(s+1)}{2}. \tag{7.141}$$

3. If $x(s)$ is the mapping from the standard element to an actual element $[x_{i-1}, x_i]$ or $[x_{2i-2}, x_{2i}]$ as appropriate then

   $$\tilde{\phi}_i(x(s)) = \phi_i(s) \tag{7.142}$$

   relates the basis functions in terms of the local variable s to those in terms of the actual variable x.

4. The element matrix is the matrix $(a(\tilde{\phi}_k, \tilde{\phi}_l)|_i)$ involving all the local basis functions $\tilde{\phi}_l(x)$. It is of size 2×2 for linears, 3×3 for quadratics and 4×4 for cubics.

   The element stiffness matrix corresponds to $a(u,v) = \int_0^1 u'v' \, dx$. For linears we get

   $$\begin{pmatrix} a(\tilde{\phi}_1, \tilde{\phi}_1)|_i & a(\tilde{\phi}_1, \tilde{\phi}_2)|_i \\ a(\tilde{\phi}_2, \tilde{\phi}_1)|_i & a(\tilde{\phi}_2, \tilde{\phi}_2)|_i \end{pmatrix} = \int_{x_{i-1}}^{x_i} \begin{pmatrix} \tilde{\phi}_1'^2 & \tilde{\phi}_1'\tilde{\phi}_2' \\ \tilde{\phi}_2'\tilde{\phi}_1' & \tilde{\phi}_2'^2 \end{pmatrix} \mathrm{dx} = \frac{1}{h_i} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{7.143}$$

   The element mass matrix is

   $$\int_{x_{i-1}}^{x_i} \begin{pmatrix} \tilde{\phi}_1^2 & \tilde{\phi}_1\tilde{\phi}_2 \\ \tilde{\phi}_2\tilde{\phi}_1 & \tilde{\phi}_2^2 \end{pmatrix} \mathrm{dx} = \frac{h_i}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{7.144}$$

   The element vector is $((f, \tilde{\phi}_l)|_i)$ and it is of size 2×1 for linears, 3×1 for quadratics and 4×1 for cubics.

5. The process of combining the element matrices and the element vectors is known as assembly.

6. In practice the integrals are often approximated using Gauss Legendre quadrature.

8 An overview of the FEM and the convergence of the FEM

Up to this point in the module we have considered how to reformulate differential equations into weak form and we have described the Galerkin finite element method using piecewise polynomial functions. We now justify things a bit more by explaining why a sequence of approximations corresponding to successively refining the mesh used converges. This is quite straightforward to do in what is known as the energy norm but there is more to do to show convergence in other norms and to consider properties such as the existence, uniqueness and stability of the exact solution and of the finite element approximation.

For students taking `MA5352` some parts of these notes will overlap with what Dr. Simon Shaw does in his part of the module but the intention is not to repeat too much of what he covers in the lectures. Also, for everyone on the module (`MA3951` and `MA5352`) the intention is to to only cover the 'easier' parts in the lectures with the 'more difficult' parts included for completeness. Some of the 'harder' parts may however be covered in some way in the lectures of Simon Shaw.

To summarise the relevant terms and results which have already been introduced we have the following. We have a functions space V (typically a subspace of $H^1(0,1)$ or $H^2(0,1)$) of the weak form, we have a conforming approximation space $V_h \subset V$ and we have that the exact solution $u \in V$ (assuming that it exists) and the approximate solution $U_h \in V_h$ satisfy

$$\begin{aligned} a(u,v) &= F(v) \quad \text{for all } v \in V, && (8.1)\\ a(U_h,v) &= F(v) \quad \text{for all } v \in V_h. && (8.2) \end{aligned}$$

Here $a(.,.)$ is a symmetric and positive definite bilinear form and $F(.)$ is a linear functional. We have also shown that $u \in V$ solves the weak problem if and only if u minimises the quadratic functional

$$I(v) = \frac{1}{2}a(v,v) - F(v). \tag{8.3}$$

These terms are sufficient to start the analysis and to get a best approximation property in what is known as the energy norm but additional terms are needed later in this section in order to relate quantities to the 'usual' norm of the space V.

8.1 The energy norm and the Cauchy-Schwarz inequality

As $a(.,.)$ is a symmetric and positive definite bilinear form it defines an inner product on the space V known as the **energy inner product** and we can define a norm associated with this inner product by

$$\|v\|_E := a(v,v)^{1/2} \tag{8.4}$$

which is known as the **energy norm**. If $u, v \in V$ with $v \neq 0$ and $t \in \mathbb{R}$ then for $u - tv \in V$ we have

$$\begin{aligned} 0 \le \|u-tv\|_E^2 &= a(u-tv, u-tv) && (8.5)\\ &= a(v,v)t^2 - 2a(u,v)t + a(u,u) && (8.6)\\ &= a(v,v)\left(t - \frac{a(u,v)}{a(v,v)}\right)^2 + \left(a(u,u) - \frac{a(u,v)^2}{a(v,v)}\right). && (8.7) \end{aligned}$$

For this inequality to be true for all t, and in particular at the minimum, we must have

$$a(u,v)^2 \le a(u,u)a(v,v) = \|u\|_E^2 \|v\|_E^2 \tag{8.8}$$

which is the **Cauchy-Schwarz inequality**.

8.2 Galerkin orthogonality and the best approximation in the energy norm

From

$$\begin{aligned} a(u,v) &= F(v) \quad \text{for all } v \in V, & (8.9)\\ a(U_h,v) &= F(v) \quad \text{for all } v \in V_h \subset V & (8.10) \end{aligned}$$

and the bilinear property of $a(.,.)$ we get by subtracting one equation from the other that

$$a(u-U_h, v) = 0 \quad \text{for all } v \in V_h. \tag{8.11}$$

This is known as **Galerkin orthogonality**. The results says that the error $u - U_h$ in the Galerkin approximation is orthogonal to all functions in V_h in the energy inner product. Another way of saying this is that $U_h \in V_h$ is the **projection** of the exact solution u in the energy inner product.

Now as $U_h \in V_h$ it follows that for all $v \in V_h$ we have $U_h - v \in V_h$ is orthogonal to the error and thus

$$\begin{aligned} \|u-U_h\|_E^2 &= a(u-U_h, u-U_h) & (8.12)\\ &= a(u-U_h, u-U_h) + a(u-U_h, U_h - v) & (8.13)\\ &= a(u-U_h, u-v) & (8.14)\\ &\le \|u-U_h\|_E \|u-v\|_E & (8.15) \end{aligned}$$

where the last step is by the Cauchy Schwarz inequality. Thus

$$\|u-U_h\|_E \le \|u-v\|_E \quad \text{for all } v \in V_h. \tag{8.16}$$

This last result says that U_h is the best approximation to the exact solution u in the energy norm. This is a key step in proving convergence. It tells us that in particular that when V_h is a space of piecewise polynomials and $u_I \in V_h$ denotes the interpolant to u we have

$$\|u-U_h\|_E \le \|u-u_I\|_E \ . \tag{8.17}$$

Results concerning how well polynomials interpolate functions can then be used to show that $\|u-u_I\|_E \to 0$ as $h \to 0$ as the mesh is refined. The rate at which $\|u-u_I\|_E \to 0$ as $h \to 0$ depends on the continuity properties of u (i.e. how many derivatives are bounded) and it depends on the degree of the piecewise polynomials used. The details depend on the particular bilinear form being considered.

Examples

For the problem $-u'' = f$, $u(0) = u(1) = 0$ we have

$$a(u,v) = \int_0^1 u'v' \,\mathrm{d}x, \quad \|v\|_E^2 = \int_0^1 v'^2 \,\mathrm{d}x \quad \text{and} \quad \|u-u_I\|_E^2 = \sum_{i=1}^{ne} \int_{x_{i-1}}^{x_i} (u'-u_I')^2 \,\mathrm{d}x. \tag{8.18}$$

When V_h is the space of piecewise linears and $u \in C^2[0,1]$ we have the interpolation error representation

$$u(x) - u_I(x) = (x-x_{i-1})(x-x_i)\frac{u''(\xi(x))}{2} \quad \text{for some } \xi(x) \in (x_{i-1}, x_i). \tag{8.19}$$

From this it follows that $|u(x)-u_I(x)|$ is $\mathcal{O}(h_i^2)$ in magnitude, $h_i = x_i - x_{i-1}$, and $|u'(x)-u_I'(x)|$ is $\mathcal{O}(h_i)$ in magnitude. With $h = \max_i h_i$ this in turn implies that $\|u-u_I\|_E$ is $\mathcal{O}(h)$ in magnitude and we have convergence in the energy norm as $h \to 0$. Similarly if we use piecewise quadratics then the interpolation error in the function is $\mathcal{O}(h^3)$ in magnitude and the error in the derivative is $\mathcal{O}(h^2)$ in magnitude from which we get that $\|u-u_I\|_E$ is $\mathcal{O}(h^2)$ in magnitude and we have faster convergence in the energy norm. More generally when the piecewise polynomials are of degree $r \ge 1$ we have that $\|u-u_I\|_E$ is $\mathcal{O}(h^r)$ in magnitude.

Similar reasoning applies to the more general bilinear form

$$a(u,v) = \int_0^1 pu'v' + quv \,\mathrm{d}x. \tag{8.20}$$

In the case of the fourth order problem $u'''' = f$ the bilinear form is

$$a(u,v) = \int_0^1 u''v'' \,\mathrm{d}x \quad \text{with} \quad \|u\|_E = \int_0^1 v''^2 \,\mathrm{d}x. \tag{8.21}$$

With piecewise Hermite cubics $u(x) - u_I(x) = \mathcal{O}(h_i^4)$ on $x_{i-1} \le x \le x_i$ and as the energy norm here involves second derivatives we have $\|u - U_h\|_E = \mathcal{O}(h^2)$.

8.3 Further analysis

We now consider some additional properties of the bilinear form $a(.,.)$ and the linear functional $F(.)$ which we need to relate the energy norm to the 'usual' norm of V and which we need to discuss stability, existence and uniqueness of the exact solution u. We begin with the spaces used.

8.3.1 The spaces $L_2(0,1)$, $H^1(0,1)$ and $H^2(0,1)$

The spaces $L_2(0,1)$, $H^1(0,1)$ and $H^2(0,1)$ are all inner product spaces with inner products and norms as follows.

Space	Inner product	Norm
$L_2(0,1)$	$(u,v) = \int_0^1 uv \,\mathrm{d}x$	$\|v\| = \left(\int_0^1 v^2 \,\mathrm{d}x\right)^{1/2}$
$H^1(0,1)$	$(u,v) = \int_0^1 uv + u'v' \,\mathrm{d}x$	$\|v\| = \left(\int_0^1 v^2 + v'^2 \,\mathrm{d}x\right)^{1/2}$
$H^2(0,1)$	$(u,v) = \int_0^1 uv + u'v' + u''v'' \,\mathrm{d}x$	$\|v\| = \left(\int_0^1 v^2 + v'^2 + v''^2 \,\mathrm{d}x\right)^{1/2}$

As a technical point, the integrals here are in the sense of Lebesgue and it can be shown that all these spaces are **complete** in the given norm. This means that if we have a Cauchy sequence in any of these spaces then the sequence converges to an element of the space in the same way as a Cauchy sequence of real numbers is a convergent sequence of real numbers. If you have not met the term Cauchy sequence before then it refers to a sequence (v_n) in the space V with the property that for all $\epsilon > 0$ there exists a natural number $N = N(\epsilon)$ such that $\|v_n - v_m\| < \epsilon$ whenever $n, m \ge N$. That is the Cauchy sequence property is concerned with all the entries being eventually close to each other. The completeness property of the space is then concerned with the existence of an entry of V to which the sequence converges. That is the entries of the sequence are eventually close to each other as they are eventually close to the limit of the sequence which is an element of the space V.

As a final term, a complete inner product space is known as a **Hilbert space**. All the spaces that we are considering are hence Hilbert spaces.

8.3.2 Coercive and bounded/continuous properties

Throughout this module we have been considering bilinear forms $a(.,.)$ which are positive definite, i.e. $a(v, v) > 0$ for all $v \ne 0$. A stronger requirement than this is that the bilinear form is **coercive** in the space V which means that there exists a constant $\kappa_1 > 0$ such that

$$a(v, v) \ge \kappa_1 \|v\|^2 \quad \text{for all } v \in V. \tag{8.22}$$

This property is sometimes also called **elliptic** or V**-elliptic** where the V tag here is because the space being used has been labelled as V.

The bilinear form $a(.,.)$ is said to be **bounded**, or equivalently **continuous**, if there exists a constant $\kappa_2 > 0$ such that

$$|a(u, v)| \le \kappa_2 \|u\| \|v\| \quad \text{for all } u, v \in V. \tag{8.23}$$

The terms bounded and continuous are equivalent here because the bilinearity gives

$$|a(u_1, v) - a(u_2, v)| \le \kappa_2 \|u_1 - u_2\| \|v\| \tag{8.24}$$

and thus u_1 being close to u_2 implies that the bilinear form $a(u_1, v)$ is close to $a(u_2, v)$. With a bounded and coercive bilinear form we have

$$\kappa_1 \|v\|^2 \le \|v\|_E^2 = a(v, v) \le \kappa_2 \|v\|^2 \tag{8.25}$$

and hence

$$\sqrt{\kappa_1} \|v\| \le \|v\|_E \le \sqrt{\kappa_2} \|v\|^2 \quad \text{and} \quad \frac{1}{\sqrt{\kappa_2}} \|v\|_E \le \|v\| \le \frac{1}{\sqrt{\kappa_1}} \|v\|_E. \tag{8.26}$$

The energy norm and the norm of the space V are **equivalent**.

The linear functional $F(.)$ is similarly said to be bounded, or equivalently continuous, if there exists a constant $\kappa_3 > 0$ such that

$$|F(v)| \le \kappa_3 \|v\| \quad \text{for all } v \in V. \tag{8.27}$$

Example

For the problem $-u'' = f$, $u(0) = u(1) = 0$ we have

$$a(u,v) = \int_0^1 u'v' \, \mathrm{d}x, \quad F(v) = \int_0^1 fv \, \mathrm{d}x, \quad V = H_0^1(0,1) \subset H^1(0,1). \tag{8.28}$$

The norm of V and the energy norm are given by

$$\|v\|^2 = \int_0^1 v^2 + v'^2 \, \mathrm{d}x \quad \text{and} \quad \|v\|_E^2 = \int_0^1 v'^2 \, \mathrm{d}x. \tag{8.29}$$

We clearly have

$$\int_0^1 v^2 \, \mathrm{d}x \le \|v\|^2 \quad \text{and} \quad \int_0^1 v'^2 \, \mathrm{d}x \le \|v\|^2. \tag{8.30}$$

The bounded property
This is straightforward and just requires use of the Cauchy Schwarz inequality.

$$|F(v)|^2 = \left|\int_0^1 fv \, \mathrm{d}x\right|^2 \le \left(\int_0^1 f^2 \, \mathrm{d}x\right)\left(\int_0^1 v^2 \, \mathrm{d}x\right) \le \left(\int_0^1 f^2 \, \mathrm{d}x\right)\|v\|^2. \tag{8.31}$$

F is bounded provided

$$\kappa_3 = \left(\int_0^1 f^2 \, \mathrm{d}x\right)^{1/2} < \infty. \tag{8.32}$$

Similarly

$$|a(u,v)|^2 = \left|\int_0^1 u'v' \, \mathrm{d}x\right|^2 \le \int_0^1 u'^2 \, \mathrm{d}x \int_0^1 v'^2 \, \mathrm{d}x \le \|u\|^2\|v\|^2. \tag{8.33}$$

The bilinear form is bounded in V with the constant $\kappa_2 = 1$.

The coercive property(*more difficult*)
Showing that the coercive property holds is more difficult to do for most problems and it is even slightly tricky for this simplest of problems. To get a lower bound for $\|v\|_E^2$ which only involves v' in terms of $\|v\|^2$ which involves both v and v' requires that we make use of the boundary conditions $v(0) = v(1) = 0$ for functions in $v \in V = H_0^1(0,1)$. By just using $v(0) = 0$ and the Cauchy Schwarz inequality we can get a bound by using

$$\begin{aligned} v(x) &= \int_0^x v'(t) \, \mathrm{d}t && (8.34) \\ v(x)^2 &= \left(\int_0^x v'(t) \, \mathrm{d}t\right)^2 \le \int_0^x 1^2 \, \mathrm{d}t \int_0^x v'(t)^2 \, \mathrm{d}t && (8.35) \\ &\le x \int_0^1 v'(t)^2 \, \mathrm{d}x. && (8.36) \end{aligned}$$

Thus

$$\int_0^1 v(x)^2 \, \mathrm{d}x \le \frac{1}{2}\int_0^1 v'^2 \, \mathrm{d}x \quad \text{i.e.} \quad \int_0^1 v'^2 \, \mathrm{d}x \ge 2\int_0^1 v(x)^2 \, \mathrm{d}x. \tag{8.37}$$

This bound is not optimal. By using Fourier series (see a question on the exercise sheet) we actually have

$$\int_0^1 v'^2 \, \mathrm{d}x \ge \pi^2 \int_0^1 v^2 \, \mathrm{d}x. \tag{8.38}$$

To establish the coercive property the trick is to let $0 < \alpha < 1$ and to write

$$\int_0^1 v'^2 \, \mathrm{d}x = \alpha \int_0^1 v'^2 \, \mathrm{d}x + (1-\alpha) \int_0^1 v'^2 \, \mathrm{d}x \tag{8.39}$$

$$\geq \alpha\pi^2 \int_0^1 v^2 \, \mathrm{d}x + (1-\alpha) \int_0^1 v'^2 \, \mathrm{d}x, \tag{8.40}$$

i.e. we just bound below one of the terms in the above. To get the best result we then choose α so that the two terms 'balance' in the sense that $\alpha\pi^2 = 1-\alpha$. Thus

$$\alpha(1+\pi^2) = 1 \quad \text{and} \quad \alpha = \frac{1}{1+\pi^2}. \tag{8.41}$$

Thus we have for all $v \in V$ that

$$\|v\|_E^2 = \int_0^1 v'^2 \, \mathrm{d}x \geq \frac{\pi^2}{1+\pi^2} \left(\int_0^1 v^2 + v'^2 \, \mathrm{d}x \right) = \frac{\pi^2}{1+\pi^2} \|v\|^2. \tag{8.42}$$

Thus the coercive property is satisfied with $\kappa_1 = \pi^2/(1+\pi^2)$. This constant is optimal as the function $v(x) = \sin \pi x$ is in V and

$$\int_0^1 \sin^2 \pi x \, \mathrm{d}x = \int_0^1 \cos^2 \pi x \, \mathrm{d}x = \frac{1}{2}, \quad \|v\|_E^2 = \frac{\pi^2}{2} \quad \text{and} \quad \|v\|^2 = \frac{1+\pi^2}{2}. \tag{8.43}$$

8.3.3 Uniqueness of a solution in V

Assuming that we have a solution to the weak problem then we can quickly show that it is unique as follows. If u_1 and u_2 are in V and both solve the weak problem then

$$a(u_1, v) = a(u_2, v) = F(v) \quad \text{for all } v \in V \tag{8.44}$$

and thus

$$a(u_1 - u_2, v) = 0 \quad \text{for all } v \in V. \tag{8.45}$$

In particular this is true for $v = u_1 - u_2 \in V$ and using the coercive property we have

$$0 \leq \kappa_1 \|u_1 - u_2\|^2 \leq \|u_1 - u_2\|_E^2 = 0 \tag{8.46}$$

which shows that $u_1 = u_2$ as elements in V.

8.3.4 The stability of u and U_h

As the exact solution $u \in V$ satisfies

$$a(u, v) = F(v) \quad \text{for all } v \in V \tag{8.47}$$

the result is true if we take $v = u$. Hence using the coercive and bounded properties

$$\kappa_1 \|u\|^2 \leq \|u\|_E^2 = a(u, u) = F(u) \leq \kappa_3 \|u\| \tag{8.48}$$

so that

$$\|u\| \leq \frac{\kappa_3}{\kappa_1}. \tag{8.49}$$

This property that the solution u is bounded in the norm of V is known as a stability property.

We can similarly bound the approximation U_h in the norm of V. Recall that $U_h \in V_h$ satisfies

$$a(U_h, v) = F(v) \quad \text{for all } v \in V. \tag{8.50}$$

If we take $v = U_h \in V_h$ then we have

$$\kappa_1 \|U_h\|^2 \le \|U_h\|_E^2 = a(U_h, U_h) = F(U_h) \le \kappa_3 \|U_h\| \tag{8.51}$$

so that

$$\|U_h\| \le \frac{\kappa_3}{\kappa_1}. \tag{8.52}$$

8.3.5 The existence of the solution $u \in V$ (*more difficult*)

One way of showing that there actually exists a solution to the weak problem is to show that there exists $u \in V$ which minimises the functional

$$I(v) = \frac{1}{2} a(v, v) - F(v). \tag{8.53}$$

Now as $a(.,.)$ is coercive in V and $F(.)$ is bounded in V we can get an expression for a lower bound for $I(v)$ from

$$\begin{aligned} I(v) &\ge \text{(lower bound on } a(v,v)/2) - \text{(upper bound on } F(v)) &(8.54)\\ &= \frac{\kappa_1}{2}\|v\|^2 - \kappa_3\|v\|. &(8.55)\end{aligned}$$

Now the quadratic

$$\frac{\kappa_1}{2}t^2 - \kappa_3 t = \frac{\kappa_1}{2}\left(t^2 - 2\frac{\kappa_3}{\kappa_1}t\right) = \frac{\kappa_1}{2}\left(t - \frac{\kappa_3}{\kappa_1}\right)^2 - \frac{\kappa_3^2}{2\kappa_1} \ge -\frac{\kappa_3^2}{2\kappa_1}. \tag{8.56}$$

Hence for all function $v \in V$ the value of the functional $I(v)$ is at least as large as $-\kappa_3^2/(2\kappa_1)$. Another way of saying this is that the set of real numbers $\{I(v) : \ v \in V\}$ is bounded below and hence the set has a greatest lower bound or infimum β. Further, as a consequence of being an infimum there must be a decreasing sequence of numbers $(I(v_n))$ in the set with $I(v_n) \to \beta$ as $n \to \infty$. All that we need to do now is to show that the corresponding sequence of functions (v_n) converges to an element $u \in V$. The continuity properties of $a(.,.)$ and $F(.)$ then imply that $I(.)$ is continuous and $I(u) = \beta$ and u minimises the functional. We proceed as follows.

Because of the linearity of $F(.)$ we have for any two functions $v_n, v_m \in V$

$$F(v_n) + F(v_m) - 2F((v_n + v_m)/2) = 0 \tag{8.57}$$

and

$$\begin{aligned} &I(v_n) + I(v_m) - 2I((v_n + v_m)/2) &(8.58)\\ &= \frac{1}{2}\left(a(v_n, v_n) + a(v_m, v_m) - \frac{1}{2}a(v_n + v_m, v_n + v_m)\right) &(8.59)\\ &= \frac{1}{4}\left(a(v_n, v_n) + a(v_m, v_m) - 2a(v_n, v_m)\right) &(8.60)\\ &= \frac{1}{4}a(v_n - v_m, v_n - v_m) = \frac{1}{4}\|v_n - v_m\|_E^2 &(8.61)\\ &\ge \frac{\kappa_1}{4}\|v_n - v_m\|^2. &(8.62)\end{aligned}$$

Thus

$$\|v_n - v_m\|^2 \le \frac{4}{\kappa_1}\left(I(v_n) + I(v_m) - 2I((v_n + v_m)/2)\right). \tag{8.63}$$

From the monotone convergence of $I(v_n) \to \beta$ we have for each $\epsilon > 0$ the existence of N such that $I(v_n) < \beta + \kappa_1\epsilon^2/8$ for all $n \ge N$. Now as $(v_n + v_m)/2 \in V$ we have $I((v_n + v_m)/2) \ge \beta$. It follows then that if $n, m \ge N$ we have

$$\|v_n - v_m\|^2 \le \frac{4}{\kappa_1}\left(\beta + \frac{\kappa_1\epsilon^2}{8} + \beta + \frac{\kappa_1\epsilon^2}{8} - 2\beta\right) = \epsilon^2. \tag{8.64}$$

This shows that the sequence (v_n) is a Cauchy sequence in V and because V is complete the sequence converges to an element $u \in V$ and the continuity of $I(.)$ implies that $I(u) = \beta$.

This result is known as the Lax-Milgram theorem. Note that we have only proved the existence of a solution u to the weak problem and the equivalent minimization problem from the function space V which in our applications corresponds to $H^1(0,1)$ for the second order problems, and corresponds to $H^2(0,1)$ for the fourth order problems. There is in general much more analysis to do for a general differential equation to show that $u \in V$ is also in function spaces of the type $C^m[0,1]$.

8.4 Summary

1. $a(u,v)$ is the energy inner product and the energy norm is $\|v\|_E = a(v,v)^{1/2}$.
2. The Cauchy Schwarz inequality is $|a(u,v)| \le \|u\|_E\|v\|_E$.
3. With $u \in V$ being the exact solution and $U_h \in V_h \subset V$ being the approximate solution the Galerkin orthogonality is $a(u - U_h, v) = 0$ for all $v \in V_h$.
4. $U_h \in V_h$ is the best approximation to $u \in V$ in the energy norm in the sense
$$\|u - U_h\|_E \le \|u - v\|_E \quad \text{for all } v \in V_h \quad . \tag{8.65}$$
5. For second order ODEs and piecewise linear approximations $\|u - U_h\|_E$ is $\mathcal{O}(h)$ in magnitude and hence $\|u - U_h\|_E \to 0$ as $h \to 0$.
6. Let $\|.\|$ be the norm of the space V. We have the following terms.

   Coercive: $\exists\kappa_1 > 0$ such that $a(v,v) \ge \kappa_1\|v\|^2$ for all $v \in V$.

   Bounded: $\exists\kappa_2 > 0$ such that $|a(u,v)| \le \kappa_2\|u\|\|v\|$ for all $u, v \in V$.

   Bounded: $\exists\kappa_3 > 0$ such that $|F(v)| \le \kappa_3\|v\|$ for all $v \in V$.
7. When $a(.,.)$ and $F(.)$ satisfies the properties in 6 there exists a unique solution $u \in V$ to the weak problem which is bounded in the norm of V. Similarly there exists a unique solution $U_h \in V_h$ to the approximate weak problem which is bounded in the norm of V.

9 Some mathematical preliminaries for problems with two space dimensions

The following duplicates a little of what was covered in the preliminary section at the start of the module and, for completeness, it extends what was covered earlier by including some details of the derivation of the divergence theorem and the result about transforming a double integral.

9.1 Differentiation and integration in $\mathbb{R}^2$ and the divergence theorem

To solve a differential equation exactly, when this is possible, we usually have to integrate at some point and to reformulate our ODEs and PDEs in a way which requires weaker continuity requirements we also have to integrate. We review now some of the definitions and results which will be needed.

For a function $f : \mathbb{R} \to \mathbb{R}$ of one space variable the derivative $f'(x)$ gives the slope of the curve of $y = f(x)$ at x and the integral of f from c to d ($c < d$), which we write as

$$\int_c^d f(x)\,\mathrm{d}x, \tag{9.1}$$

gives the area under the curve of f from c to d. In addition, one form of the fundamental theorem of calculus which links differentiation and integration is that

$$\int_c^d f'(x)\,\mathrm{d}x = [f(x)]_c^d = f(d) - f(c). \tag{9.2}$$

We now consider the corresponding situations when instead $f : \mathbb{R}^2 \to \mathbb{R}$, i.e. when f is a function of two space variables.

9.1.1 The gradient vector and directional derivatives

The first thing to note is that we now need partial derivatives instead of ordinary derivatives. We have

$$\frac{\partial f}{\partial x}(x,y) = \lim_{h\to 0} \frac{f(x+h,y) - f(x,y)}{h}, \tag{9.3}$$

$$\frac{\partial f}{\partial y}(x,y) = \lim_{h\to 0} \frac{f(x,y+h) - f(x,y)}{h}. \tag{9.4}$$

These partial derivatives give the slopes on the surface $z = f(x,y)$ in the x and y directions respectively. It is convenient to collect these two quantities together as a vector known as the gradient vector

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}. \tag{9.5}$$

We can similarly consider how f changes in any direction $\underline{n} = (n_1, n_2)^T$. If $\underline{n}$ is of unit length then this directional derivative is defined by

$$\frac{\partial f}{\partial n}(x,y) = \lim_{h\to 0} \frac{f(x+hn_1, y+hn_2) - f(x,y)}{h}. \tag{9.6}$$

Now

$$\frac{f(x+hn_1,y+hn_2)-f(x,y)}{h} = \frac{f(x+hn_1,y+hn_2)-f(x,y+hn_2)}{h} \tag{9.7}$$

$$+\frac{f(x,y+hn_2)-f(x,y)}{h}. \tag{9.8}$$

Now by the mean value theorem and the continuity of $\partial f/\partial x$ we have

$$\frac{f(x+hn_1,y+hn_2)-f(x,y+hn_2)}{h} = n_1\frac{\partial f}{\partial x}(x+\theta_1 h, y+hn_2) \to n_1\frac{\partial f}{\partial x}(x,y) \quad \text{as } h\to 0. \tag{9.9}$$

In the above line $0<\theta_1<1$. Similarly by the definition of the partial derivative

$$\frac{f(x,y+hn_2)-f(x,y)}{h} \to n_2\frac{\partial f}{\partial y}(x,y). \tag{9.10}$$

The result that we get is the chain rule of partial differentiation. In vector notation with $\underline{x}=(x,y)^T$ we have

$$\frac{\partial f}{\partial n}(\underline{x}) = n_1\frac{\partial f}{\partial x}(x,y)+n_2\frac{\partial f}{\partial y}(x,y) = \underline{n}\cdot\nabla f(\underline{x}) \tag{9.11}$$

where here $\cdot$ means the usual scalar or dot product of two vectors. We will also use the notation $\underline{n}^T\nabla u$, when convenient, to denote this operation.

9.1.2 Double integrals, i.e. area integrals

In the case of integration, the integral of f over a two dimensional region Ω, which we write as

$$\int_\Omega f(\underline{x})\ \mathrm{d}a \quad \text{or as} \quad \iint_\Omega f(\underline{x})\ \mathrm{d}x\mathrm{d}y, \tag{9.12}$$

gives the volume under the surface $z=f(\underline{x})=f(x,y)$ in the part $\underline{x}\in\Omega$. When the single integral notation is used the increment $\mathrm{d}a=\mathrm{d}x\mathrm{d}y$ indicates an area increment.

9.1.3 Contour integrals in $\mathbb{R}^2$

With functions $f(\underline{x})$ we can also consider contour integrals in the plane $\mathbb{R}^2$. If Γ is a curve in $\mathbb{R}^2$ which has a parametric representation

$$\Gamma = \{\underline{x}(t): \quad 0\le t\le L\} \tag{9.13}$$

then an incremental change dt in t gives a corresponding change $d\underline{x}=\underline{x}'(t)dt$ along the curve (i.e. tangential to the curve) which has length $ds=|\underline{x}'(t)dt|$. The (contour) integral of f along Γ is defined by

$$\int_\Gamma f(\underline{x})\,\mathrm{d}s = \int_0^L f(\underline{x}(t))|\underline{x}'(t)|\,\mathrm{d}t. \tag{9.14}$$

Note that the value of the integral along Γ here does not depends on the direction at which we go along Γ. This is in contrast to the integral

$$\int_L^0 g(x)\,\mathrm{d}x = -\int_0^L g(x)\,\mathrm{d}x \tag{9.15}$$

which has a slightly different meaning. In this latter case integrating g from L to 0 corresponds to integrating $-g$ from 0 to L. The direction aspect is taken care of within the expression in the integrand. This is the case in situations where we have contour integrals like (9.14) in the divergence theorem.

9.1.4 The divergence theorem

We now consider what happens when we integrate a partial derivative to get the two dimensional version of (9.2). This will be most conveniently expressed in what is known as the divergence theorem.

We consider the quantity

$$\iint_\Omega \frac{\partial f}{\partial x}\,\mathrm{d}x\mathrm{d}y. \tag{9.16}$$

We need to describe the region Ω in some way, and not to over complicate things, we will suppose that Ω is of the form shown in Figure 9.1. If we project Ω onto the $y-$axis then we have that each point $\underline{x} \in \Omega$ has a vertical component y in the range $\alpha < y < \beta$. For each $y \in (\alpha, \beta)$ the x value lies between $\gamma(y)$ and $\delta(y)$. Thus

$$\iint_\Omega \frac{\partial f}{\partial x}(x,y)\,\mathrm{d}x\mathrm{d}y = \int_{y=\alpha}^{\beta}\left(\int_{x=\gamma(y)}^{\delta(y)} \frac{\partial f}{\partial x}(x,y)\,\mathrm{d}x\right)\mathrm{d}y. \tag{9.17}$$

Now the fundamental theorem of calculus given in (3.45) applies to the inner integral since y is fixed giving

$$\iint_\Omega \frac{\partial f}{\partial x}(x,y)\,\mathrm{d}x\mathrm{d}y = \int_{y=\alpha}^{\beta}\left(f(\delta(y),y) - f(\gamma(y),y)\right)\mathrm{d}y. \tag{9.18}$$

Now with reference again to the Figure, $(\delta(y), y)$, $\alpha < y < \beta$ gives a parametric representation of the right hand part of the boundary $\partial\Omega$ transversed in the positive (i.e. anticlockwise) direction, which we label as Γ_1, and similarly $(\gamma(y), y)$, $\alpha < y < \beta$ gives a parametric representation of the left hand part of the boundary $\partial\Omega$ transversed in the opposite (i.e. clockwise) direction, which we label as Γ_2. The collection of all the points in Γ_1 and Γ_2 gives the complete boundary of $\partial\Omega$. To relate this to a contour integral note that on the right hand part Γ_1 of $\partial\Omega$ the vector $(\delta'(y), 1)^T$ is in the tangential direction to $\partial\Omega$ going in the anticlockwise direction giving us a unit tangent and a unit outward normal of respectively

$$\underline{t} = \frac{1}{(\delta'^2+1)^{1/2}}\begin{pmatrix}\delta'\\ 1\end{pmatrix} \quad \text{and} \quad \underline{n} = \begin{pmatrix}n_1\\ n_2\end{pmatrix} = \frac{1}{(\delta'^2+1)^{1/2}}\begin{pmatrix}1\\ -\delta'\end{pmatrix}. \tag{9.19}$$

Now to relate our integral involving $\mathrm{d}y$ to a contour integral as given in (9.14) which involves an arc length increment $\mathrm{d}s$ we have

$$\mathrm{d}s = (\delta'^2+1)^{1/2}\mathrm{d}y, \quad \text{so that} \quad \mathrm{d}y = \frac{\mathrm{d}s}{(\delta'^2+1)^{1/2}} = n_1\mathrm{d}s \tag{9.20}$$

and hence

$$\int_{y=\alpha}^{\beta} f(\delta(y),y)\,\mathrm{d}y = \int_{\Gamma_1} f(\underline{x})n_1\,\mathrm{d}s. \tag{9.21}$$

On the left part of the boundary the unit tangent corresponding to the anticlockwise direction and the unit outward normal are respectively

$$\underline{t} = \frac{-1}{(\gamma'^2+1)^{1/2}}\begin{pmatrix}\gamma'\\ 1\end{pmatrix} \quad \text{and} \quad \underline{n} = \begin{pmatrix}n_1\\ n_2\end{pmatrix} = \frac{-1}{(\gamma'^2+1)^{1/2}}\begin{pmatrix}1\\ -\gamma'\end{pmatrix}. \tag{9.22}$$

For the relation between the arc length increment $\mathrm{d}s$ and the increment $\mathrm{d}y$ we now have

$$\mathrm{d}s = (\gamma'^2+1)^{1/2}\mathrm{d}y \quad \text{so that} \quad \mathrm{d}y = \frac{\mathrm{d}s}{(\gamma'^2+1)^{1/2}} = -n_1\mathrm{d}s. \tag{9.23}$$

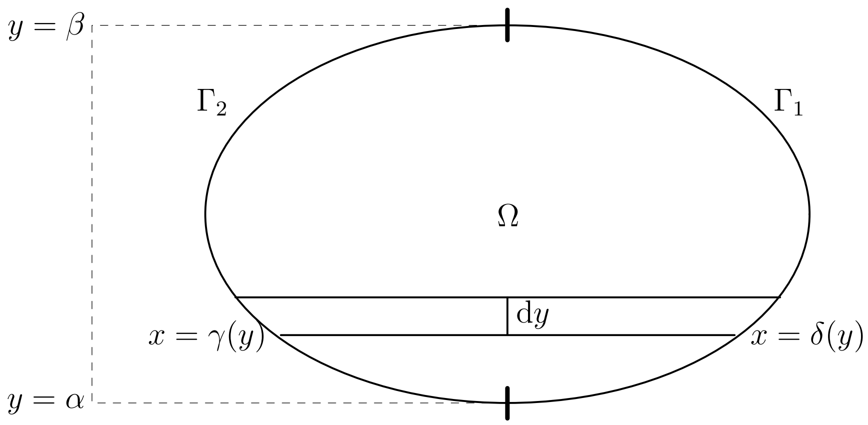


Figure 9.1: A convex simply connected domain Ω with boundary $\partial\Omega$ consisting of the right hand part Γ_1 and the left hand part Γ_2.

Hence we now get

$$-\int_{\alpha}^{\beta} f(\gamma(y), y)\,\mathrm{d}y = \int_{\Gamma_2} f(\underline{x}) n_1\,\mathrm{d}s. \tag{9.24}$$

Combining (9.18), (9.21) and (9.24) gives

$$\iint_{\Omega} \frac{\partial f}{\partial x}(x, y)\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} f(\underline{x}) n_1\,\mathrm{d}s. \tag{9.25}$$

A similar argument gives

$$\iint_{\Omega} \frac{\partial f}{\partial y}(x, y)\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} f(\underline{x}) n_2\,\mathrm{d}s. \tag{9.26}$$

If we consider a vector field $\underline{A} = (A_1, A_2)^T$ then the last two results imply that

$$\iint_{\Omega} \frac{\partial A_1}{\partial x}(x, y)\ \mathrm{d}x\mathrm{d}y \quad = \quad \int_{\partial\Omega} A_1(\underline{x}) n_1\,\mathrm{d}s, \tag{9.27}$$

$$\iint_{\Omega} \frac{\partial A_2}{\partial x}(x, y)\ \mathrm{d}x\mathrm{d}y \quad = \quad \int_{\partial\Omega} A_2(\underline{x}) n_2\,\mathrm{d}s. \tag{9.28}$$

As

$$\nabla \cdot \underline{A} = \frac{\partial A_1}{\partial x} + \frac{\partial A_2}{\partial y} \quad \text{and} \quad \underline{A} \cdot \underline{n} = A_1 n_1 + A_2 n_2 \tag{9.29}$$

(9.27) and (9.29) combine to give

$$\iint_{\Omega} \nabla \cdot \underline{A}\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \underline{A} \cdot \underline{n}\,\mathrm{d}s \tag{9.30}$$

which is known as the *divergence theorem* or as the *Gauss-divergence theorem*. The quantity $\nabla \cdot \underline{A}$ is known as the divergence of $\underline{A}$.

We have proved the divergence theorem with reference to a convex simply connected domain as shown in Figure 9.1. The result is however valid for a domain which is not convex, which has holes and with a boundary which is not smooth enough to have a normal derivative at every point as shown in Figure 9.2. A corresponding result also applies when $\Omega \subset \mathbb{R}^3$ which has a boundary $\partial\Omega$ which is a surface.

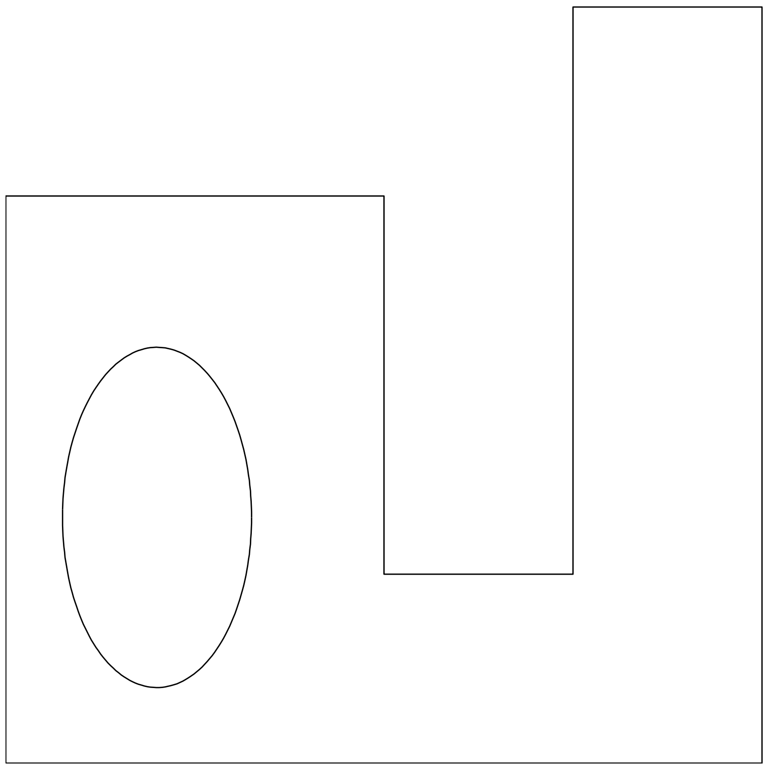


Figure 9.2: A doubly connected domain with a piecewise smooth boundary. The normal derivative does not exist at every point. The divergence theorem still holds with the boundary $\partial\Omega$ consisting of an outer boundary part and an inner boundary part.

9.1.5 Some vector identities involving ∇

If you have previously studied courses on vector calculus then you will have met many identities involving ∇ in the context of the gradient, divergence and curl of various quantities. We only need one of the simpler of these relations here which is like the rule for differentiating a product in one dimension. Recall that this result is

$$(uv)' = uv' + u'v. \tag{9.31}$$

Now in two dimension with $v = v(x, y)$ being a scalar valued function and $\underline{A} = (A_1, A_2)^T$ being a vector valued function we have that $v\underline{A} = (vA_1, vA_2)^T$ and

$$\begin{aligned} \nabla \cdot (v\underline{A}) &= \frac{\partial(vA_1)}{\partial x} + \frac{\partial(vA_2)}{\partial y} && (9.32)\\ &= v\frac{\partial A_1}{\partial x} + \frac{\partial v}{\partial x}A_1 + v\frac{\partial A_2}{\partial y} + \frac{\partial v}{\partial y}A_2 && (9.33)\\ &= v\nabla \cdot \underline{A} + \nabla v \cdot \underline{A}. && (9.34) \end{aligned}$$

In the case that the vector $\underline{A}$ is the gradient of some function, i.e. $\underline{A} = \nabla u$ we have

$$\Delta u = \nabla^2 u = \nabla \cdot \nabla u = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{9.35}$$

which is known as the *Laplacian* of u. Thus using the previous identity and the divergence theorem we have

$$\nabla \cdot (v\nabla u) = v\Delta u + \nabla v \cdot \nabla u \tag{9.36}$$

and

$$\int_{\partial\Omega} v\nabla u\cdot \underline{n}\ \mathrm{d}s = \iint_\Omega \nabla\cdot(v\nabla u)\ \mathrm{d}x\mathrm{d}y = \iint_\Omega (v\Delta u + \nabla u\cdot\nabla v)\ \mathrm{d}x\mathrm{d}y. \tag{9.37}$$

Since

$$\nabla u\cdot\underline{n} = \frac{\partial u}{\partial n} \tag{9.38}$$

this can be written as

$$\iint_\Omega (v\Delta u + \nabla u\cdot\nabla v)\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} v\frac{\partial u}{\partial n}\ \mathrm{d}s. \tag{9.39}$$

If $v=0$ on $\partial\Omega$ then the contour integral is 0 as the integrand is 0 and this reduces to

$$-\iint_\Omega v\Delta u\ \mathrm{d}x\mathrm{d}y = \iint_\Omega \nabla u\cdot\nabla v\ \mathrm{d}x\mathrm{d}y. \tag{9.40}$$

This is the key step in reformulating Poisson's equation

$$-\Delta u = f \tag{9.41}$$

as then we obtain

$$\iint_\Omega \nabla u\cdot\nabla v\ \mathrm{d}x\mathrm{d}y = \iint_\Omega fv\ \mathrm{d}x\mathrm{d}y \tag{9.42}$$

for all suitable functions v which vanish on the boundary $\partial\Omega$. This is the expression that we will need in the weak form of Poisson's equation.

9.1.6 Transforming a double integral

To complete this review of calculus in $\mathbb{R}^2$ we consider now how to transform an area integral which we will need to do later in the module when we map from a standard element to a general element. The corresponding situation in one dimension is when we make a substitution. For example, if $x:[0,1]\to\mathbb{R}$ is differentiable then

$$\int_{x(0)}^{x(1)} f(x)\,\mathrm{d}x = \int_0^1 f(x(t))x'(t)\,\mathrm{d}t. \tag{9.43}$$

The range of integration is $[x(0),x(1)]$ with respect to x and it is $[0,1]$ with respect to t with the adjustment in the integrand with the $x'(t)$ term. In $\mathbb{R}^2$ the corresponding situation involves a one-to-one and onto mapping $\underline{x}:T\to\tilde{T}$. We now consider the details concerning how we can change the region of integration from $\tilde{T}$ in the (x,y) plane to the region T in the (s,t) plane.

An incremental change ds in the s direction and an incremental change dt in the t direction in the (s,t) plane give incremental changes

$$\begin{aligned} \underline{c} &= \underline{x}(s+\mathrm{d}s,t)-\underline{x}(s,t) = \frac{\partial\underline{x}}{\partial s}\mathrm{d}s, & (9.44)\\ \underline{d} &= \underline{x}(s,t+\mathrm{d}t)-\underline{x}(s,t) = \frac{\partial\underline{x}}{\partial t}\mathrm{d}t, & (9.45)\end{aligned}$$

in the (x,y) plane as shown in Figure 9.3 with the rectangular shape in the (s,t) plane becoming a parallelogram in the (x,y) plane. To determine how the double integral changes we need to consider the area of this parallelogram. The area is given by the length of the vector $\underline{c}$ multiplied by the distance between the two sides parallel to this direction. The distance between the two parallel sides depends on the length of $\underline{d}$ and the angle θ between the two two vectors. To obtain an expression for this we can proceed as follows. If

$$\underline{c} = (c_1, c_2)^T \quad \text{and} \quad \underline{d} = (d_1, d_2)^T \tag{9.46}$$

then the area A is given by

$$A = |\underline{c}|\,|\underline{d}||\sin\theta| \quad \text{so that} \quad A^2 = |\underline{c}|\,|\underline{d}|\sin^2\theta = |\underline{c}|\,|\underline{d}|(1-\cos^2\theta). \tag{9.47}$$

Now as

$$\underline{c}\cdot\underline{d} = |\underline{c}|\,|\underline{d}|\cos\theta = c_1 d_1 + c_2 d_2 \tag{9.48}$$

we obtain

$$A^2 = (c_1^2 + c_2^2)(d_1^2 + d_2^2) - (c_1 d_1 + c_2 d_2)^2 \tag{9.49}$$

$$= (c_2^2 d_1^2 + c_1^2 d_2^2 - 2c_1 d_1 c_2 d_2) = (c_1 d_2 - c_2 d_1)^2. \tag{9.50}$$

Thus

$$A = |c_1 d_2 - c_2 d_1| \tag{9.51}$$

and

$$c_1 d_2 - c_2 d_1 = \det\begin{pmatrix} c_1 & d_1 \\ c_2 & d_2 \end{pmatrix} = \det\begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix} \mathrm{d}s\mathrm{d}t. \tag{9.52}$$

To summarize, for the transformation $(s,t) \to (x(s,t), y(s,t))$ the Jacobian matrix is given by

$$\begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix}. \tag{9.53}$$

The determinant of this matrix is

$$J = \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s} \tag{9.54}$$

and when transforming a double integral we have

$$\iint_{\tilde{T}} f(x,y)\,\mathrm{d}x\mathrm{d}y = \iint_T f(x(s,t), y(s,t))|J|\,\mathrm{d}s\mathrm{d}t. \tag{9.55}$$

Note that the determinant J defined here can be positive or negative and it is the absolute value $|J|$ which is needed in (9.55).

9.2 Summary

In $\mathbb{R}^2$ we introduced the gradient vector and the directional derivative

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}, \quad \frac{\partial f}{\partial n}(\underline{x}) = n_1 \frac{\partial f}{\partial x}(x,y) + n_2 \frac{\partial f}{\partial y}(x,y) = \underline{n}\cdot\nabla f(\underline{x}). \tag{9.56}$$

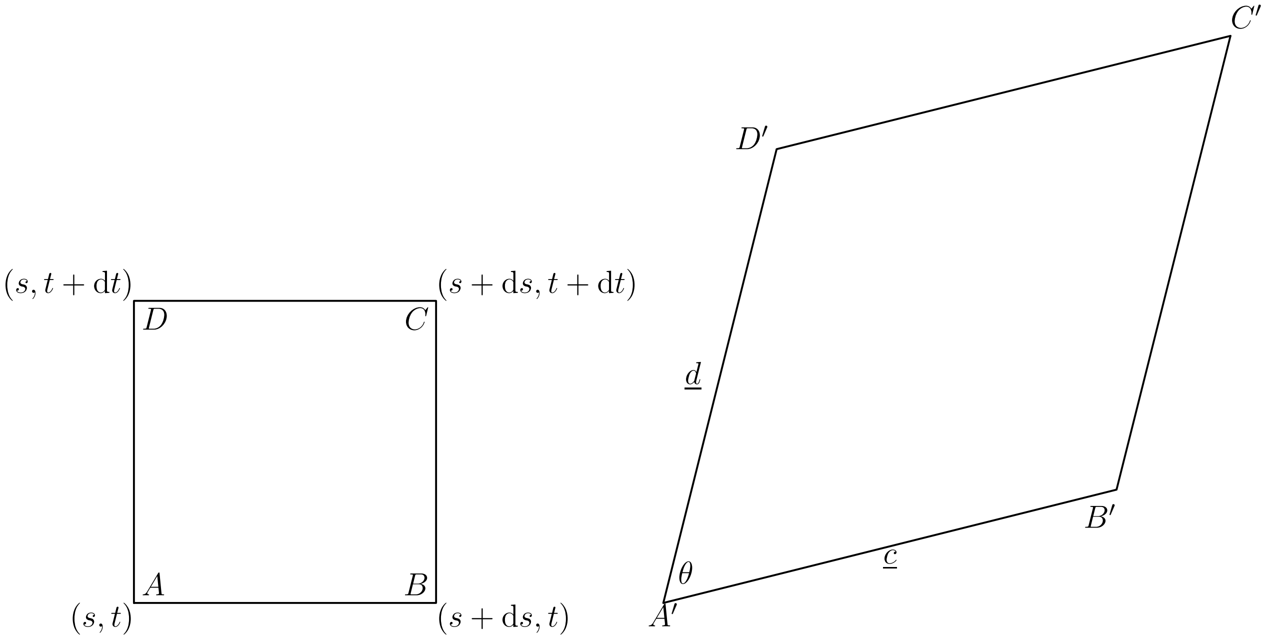


Figure 9.3: The rectangular area increment dsdt and it image in the x–y plane which is a parallelogram. The areas of the parallelogram is $|J|\mathrm{d}s\mathrm{d}t$ where J is the determinant of the Jacobian of the transformation $(s,t) \to (x(s,t), y(s,t))$.

We also introduced the divergence operation and the divergence theorem

$$\nabla \cdot \underline{A} = \frac{\partial A_1}{\partial x} + \frac{\partial A_2}{\partial y} \quad \text{and} \quad \iint_\Omega \nabla \cdot \underline{A}\, \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \underline{A} \cdot \underline{n}\, \mathrm{d}s \tag{9.57}$$

and we have noted the vector identity

$$\nabla \cdot (v\underline{A}) = v\nabla \cdot \underline{A} + \nabla v \cdot \underline{A}. \tag{9.58}$$

The Laplace operator in vector notation is

$$\Delta u = \nabla^2 u = \nabla \cdot \nabla u = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{9.59}$$

Integrals in $\mathbb{R}^2$ involving $\underline{x} : T \to \tilde{T}$ transform according to

$$\iint_{\tilde{T}} f(x,y)\, \mathrm{d}x\mathrm{d}y = \iint_T f(x(s,t), y(s,t))|J|\, \mathrm{d}s\mathrm{d}t. \tag{9.60}$$

Here J is the determinant of the Jacobian matrix

$$\begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix} \tag{9.61}$$

of the transformation $(s,t) \to (x,y)$.

10 Reformulating PDEs in $\mathbb{R}^2$

10.1 Introduction

In this section we consider a number of aspects relating to solving elliptic partial differential equations such as **Poisson's equation** which involves the equation

$$-\Delta u = -\nabla^2 u = -\nabla \cdot \nabla u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f \tag{10.1}$$

where f is some given function. If $\Omega \subset \mathbb{R}^2$ is the region in which this equation is to be satisfied and f is continuous on $\bar{\Omega}$ then a complete specification is of the following form.

$$\text{Find } u \in C^2(\bar{\Omega}) \quad \text{s.t.} \quad -\Delta u = f \quad \text{in } \Omega \quad \text{and } u \text{ satisfies a b.c. on } \partial\Omega. \tag{10.2}$$

As for the one-dimensional problems considered previously in the module we first reformulate the partial differential equation with various boundary conditions into a weaker form, we then weaken the continuity requirements by considering this over subspaces of the Sobolev space $H^1(\Omega)$ and we then consider Galerkin's method with various finite element spaces $V_h \subset H^1(\Omega)$ involving piecewise polynomial defined functions with respect to a triangular or quadrilateral meshes of Ω.

10.2 Reformulating PDEs such as Poisson's equation into a weaker form

In one space dimension the corresponding problem to Poisson's equation is $-u'' = f$ and the reformulation involves multiplying by a test function v and integrating over the domain. Integration by parts is then used to re-express the expression into a form which has less derivatives on u. Specifically we had

$$-\int_0^1 u''v \, \mathrm{d}x = -[u'v]_0^1 + \int_0^1 u'v' \, \mathrm{d}x \quad \text{resulting from } (u'v)' = u''v + u'v'. \tag{10.3}$$

In two space dimensions the gradient vector ∇u replaces the derivative u' and the $(u'v)'$ product rule relation is replaced by the vector identity

$$\nabla \cdot (v\nabla u) = v\Delta u + \nabla v \cdot \nabla u. \tag{10.4}$$

We can re-write this as

$$-v\Delta u = \nabla u \cdot \nabla v - \nabla \cdot (v\nabla u). \tag{10.5}$$

Integrating a divergence of a quantity over Ω gives the integral of this quantity around the boundary of Ω (the divergence theorem) and thus integrating (10.5) gives

$$-\iint_\Omega v\Delta u \, \mathrm{d}x\mathrm{d}y = \iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y - \int_{\partial\Omega} v\nabla u \cdot \underline{n} \, \mathrm{d}s \tag{10.6}$$

where $\underline{n}$ is the unit outward normal to $\partial\Omega$ and the contour integral is in the anticlockwise direction around $\partial\Omega$. The boundary term is usually re-written by noting that the directional derivative in the direction of $\underline{n}$ satisfies

$$\frac{\partial u}{\partial n} = \underline{n} \cdot \nabla u. \tag{10.7}$$

Thus

$$-\iint_{\Omega} v\Delta u \, \mathrm{d}x\mathrm{d}y = \iint_{\Omega} \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y - \int_{\partial\Omega} v\frac{\partial u}{\partial n} \, \mathrm{d}s. \tag{10.8}$$

To proceed further we need to consider the boundary conditions and we next consider a few standard cases.

10.2.1 Poisson's equation with homogeneous Dirichlet boundary conditions

Homogeneous Dirichlet boundary conditions means that $u = 0$ on $\partial\Omega$ and we have the problem

$$-\Delta u = f \quad \text{in } \partial\Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{10.9}$$

We hence seek $u \in V$ where

$$V = \{v \in C^2(\bar{\Omega}) : \quad v = 0 \text{ on } \partial\Omega\}. \tag{10.10}$$

If the test function v in (10.8) is restricted to V then we eliminate the contour integral term and we get the following weaker form.

$$\text{Find } u \in V \quad \text{such that} \quad \iint_{\Omega} \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y = \iint_{\Omega} fv \, \mathrm{d}x\mathrm{d}y \quad \text{for all } v \in V. \tag{10.11}$$

This is of the form $a(u, v) = (f, v)$ where now

$$a(u, v) = \iint_{\Omega} \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y \quad \text{and} \quad (f, v) = \iint_{\Omega} fv \, \mathrm{d}x\mathrm{d}y. \tag{10.12}$$

10.2.2 Poisson's equation with Neumann boundary conditions and Poisson's equation with a mixed boundary condition

A **Neumann boundary condition** is a condition of the form

$$\frac{\partial u}{\partial n} = g, \tag{10.13}$$

on part of the boundary, where g is some specified function. Poisson's equation often arises with a Dirichlet condition on part of the boundary and a Neumann condition, often a homogeneous Neumann condition, on the other part of the boundary. For example, if $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$ gives a partition of $\partial\Omega$ with $\partial\Omega_1$ being non-empty then a typical problem is

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \text{ on } \partial\Omega_1 \quad \text{and} \quad \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega_2. \tag{10.14}$$

If we consider again (10.8) then using the conditions on $\partial\Omega_2$ we get

$$-\iint_{\Omega} v\Delta u \, \mathrm{d}x\mathrm{d}y = \iint_{\Omega} \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y - \int_{\partial\Omega_1} v\frac{\partial u}{\partial n} \, \mathrm{d}s. \tag{10.15}$$

In this case we only need $v = 0$ on $\partial\Omega_1$ to eliminate the contour integral term and we seek $u \in V$ where now

$$V = \{v \in C^2(\bar{\Omega}) : \quad v = 0 \text{ on } \partial\Omega_1\}. \tag{10.16}$$

The weaker form is now as follows.

$$\text{Find } u \in V \quad \text{such that} \quad \iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y = \iint_\Omega fv \, \mathrm{d}x\mathrm{d}y \quad \text{for all } v \in V. \tag{10.17}$$

We have the same expression as before but V is now a larger space than before as there is no condition on the part $\partial\Omega_2$ of $\partial\Omega$. The homogeneous Neumann boundary condition on $\partial\Omega_2$ is the natural boundary condition of this problem.

To generalise the previous case we now consider a problem with a **mixed boundary condition** on $\partial\Omega_2$. We consider

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \text{ on } \partial\Omega_1 \quad \text{and} \quad \frac{\partial u}{\partial n} + \beta u = g \text{ on } \partial\Omega_2. \tag{10.18}$$

(We assume here that $\beta \ge 0$ which is sufficient to ensure that the resulting bilinear form is positive definite on the space under consideration.) Proceeding as before with again

$$V = \{v \in C^2(\bar{\Omega}) : \quad v = 0 \text{ on } \partial\Omega_1\} \tag{10.19}$$

we have that for the solution $u \in V$ (which thus satisfies the boundary conditions) and for any function $v \in V$

$$\begin{aligned} -\iint_\Omega v\Delta u \, \mathrm{d}x\mathrm{d}y &= \iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y - \int_{\partial\Omega_2} v\frac{\partial u}{\partial n} \, \mathrm{d}s \\ &= \iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y - \int_{\partial\Omega_2} v(g - \beta u) \, \mathrm{d}s. \end{aligned}$$

The weaker form is thus as follows. Find $u \in V$ such that

$$\iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} \beta uv \, \mathrm{d}s = \iint_\Omega fv \, \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} gv \, \mathrm{d}s \tag{10.20}$$

for all $v \in V$. This is of the form $a(u, v) = F(v)$ where now

$$a(u,v) = \iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} \beta uv \, \mathrm{d}s, \tag{10.21}$$

$$F(v) = \iint_\Omega fv \, \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} gv \, \mathrm{d}s. \tag{10.22}$$

The non-zero right hand side term g contributes to the linear functional $F(v)$ on right hand side of the weaker form.

10.2.3 The positive definite property of $a(.,.)$ for the Poisson's equation problem

In all the reformulations of Poisson's equation we get a symmetric bilinear form $a(u, v)$. We now show that the bilinear form is positive definite and to do this we observe that $a(v, v)$ is given by

$$a(v,v) = \begin{cases} \displaystyle\iint_\Omega (\nabla v)^2 \, \mathrm{d}x\mathrm{d}y, & \text{no mixed b.c. case,} \\ \displaystyle\iint_\Omega (\nabla v)^2 \, \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} \beta v^2 \, \mathrm{d}s, & \text{mixed b.c. case.} \end{cases} \tag{10.23}$$

In both cases we have integrands which are ≥ 0 and thus $a(v,v) \ge 0$ for all v. If $a(v,v) = 0$ then the continuity of the integrand when $V \subset C^2(\bar{\Omega})$ implies that $\nabla v = \underline{0}$ in Ω which in turn implies that v is a constant. As $\partial\Omega_1$ is non-empty and $v = 0$ on $\partial\Omega_1$ then this implies that $v = 0$ as required.

As a remark here it is necessary that there is a part of the boundary $\partial\Omega_1$ at which we have the condition $v = 0$ in order to have the positive definite property. If we do not have this then for the problem without the mixed condition we can can only conclude that v is a constant. The bilinear form $a(.,.)$ is only positive semi-definite. Related to this the problem

$$-\Delta u = f \quad \text{in } \Omega, \quad \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \tag{10.24}$$

does not have a unique solution. If u is a solution to (10.24) then $u + c$ is also a solution for all constants c. As a technical point here, when we weaken the continuity requirements and consider Sobolev spaces $H^1(\Omega)$ instead of the space $C^2(\bar{\Omega})$ we actually need the set $\partial\Omega_1$ to have positive measure which requires that it contains more than just one point.

10.2.4 Remarks about non-homogeneous Dirichlet boundary conditions

We have restricted attention to problems for which $u = 0$ on $\partial\Omega$ or on $\partial\Omega_1$. It is not difficult however to consider the non-homogeneous boundary conditions in practice. For example, consider the problem

$$-\Delta u = f \quad \text{in } \Omega, \quad u = g_1 \quad \text{on } \partial\Omega_1 \qquad \text{and} \quad \frac{\partial u}{\partial n} + \beta u = g_2 \quad \text{on } \partial\Omega_2. \tag{10.25}$$

Here $u = g_1$ on $\partial\Omega_1$ means that $u(x,y) = g_1(x,y)$ for all points $(x,y) \in \partial\Omega_1$. With $v \in V$ as before in the mixed boundary condition case we still have that u satisfies

$$a(u,v) = \iint_\Omega \nabla u \cdot \nabla v \;\mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} \beta uv \;\mathrm{d}s = \iint_\Omega fv \;\mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} gv \;\mathrm{d}s = F(v) \tag{10.26}$$

for all $v \in V$. If we have any function u_1 defined on $\bar{\Omega}$ such that $u_1|_{\partial\Omega_1} = g_1$ then the function u that we seek is of the form $u = u_1 + \tilde{u}$ where $\tilde{u} \in V$. Thus

$$a(u_1 + \tilde{u}, v) = F(v) \quad \text{giving} \quad a(\tilde{u}, v) = F(v) - a(u_1, v). \tag{10.27}$$

$\tilde{u}$ hence solves a problem with a homogeneous condition on $\partial\Omega_1$.

It is natural to ask here if is possible to explicitly find such a function u_1 defined on $\bar{\Omega}$ such that its restriction to $\partial\Omega_1$ is the given function g_1. The answer to this is no for the continuous problem as we cannot generally expect to get an explicit expression for u_1 unless $\partial\Omega_1$ and g_1 are fairly simple. This is however not a difficulty when we approximate u by U in a finite element space $\bar{V}_h$. We construct a function U_1 corresponding to u_1 which interpolates g_1 using the functions in $\bar{V}_h$ at the mesh points on $\partial\Omega_1$ and with $U = U_1 + \tilde{U}$ the unknown part $\tilde{U}$ of U satisfies

$$a(\tilde{U}, v) = F(v) - a(U_1, v) \tag{10.28}$$

for all $v \in \bar{V}_h$ which vanish on $\partial\Omega_1$. The span of all functions in $\bar{V}_h$ which vanish on $\partial\Omega_1$ is the space $V_h \subset V$.

10.2.5 Reformulating the biharmonic equation

To complete this section on reformulating partial differential equations into weaker forms we consider now the biharmonic equation $\Delta^2 u = f$.

In one dimension we reformulated $-u'' = f$ into a weak form by using integration by parts once and we reformulated $u'''' = f$ into a weak form by using integration by parts twice. Analogously we have just used the divergence theorem once to reformulate $-\Delta u = f$ into a weak form and we now show that doing this twice we can transform the biharmonic problem.

$$\Delta^2 u = \Delta(\Delta u) = f, \quad u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \tag{10.29}$$

into weak form. Let $w = \Delta u$ and note the vector identities

$$\begin{aligned} -v\Delta w &= \nabla v \cdot \nabla w - \nabla \cdot (v \nabla w), && (10.30)\\ -w\Delta v &= \nabla v \cdot \nabla w - \nabla \cdot (w \nabla v). && (10.31)\end{aligned}$$

Integrating over Ω for a function v satisfying $v = \partial v/\partial n = 0$ on $\partial\Omega$ we have

$$-\iint_\Omega v\Delta w \,\mathrm{d}x\mathrm{d}y = \iint_\Omega \nabla v \cdot \nabla w \,\mathrm{d}x\mathrm{d}y = -\iint_\Omega w\Delta v \,\mathrm{d}x\mathrm{d}y. \tag{10.32}$$

Thus by taking $w = \Delta u$ we have

$$\iint_\Omega v\Delta^2 u \,\mathrm{d}x\mathrm{d}y = \iint_\Omega \Delta u \Delta v \,\mathrm{d}x\mathrm{d}y. \tag{10.33}$$

With

$$V = \{v \in C^4(\bar{\Omega}) : \quad v = \frac{\partial v}{\partial n} = 0 \quad \text{on } \partial\Omega\} \tag{10.34}$$

the weaker form is as follows.

$$\text{Find } u \in V \quad \text{such that} \quad \iint_\Omega \Delta u \Delta v \,\mathrm{d}x\mathrm{d}y = \iint_\Omega f v \,\mathrm{d}x\mathrm{d}y \quad \text{for all } v \in V. \tag{10.35}$$

10.3 Weakening the continuity requirements — The Sobolev spaces in $\mathbb{R}^2$

The reformulation of Poisson's equation leads to a weaker form in which we only need to be able to integrate the first derivative over a two-dimensional region Ω. Similarly the reformulation of the biharmonic equation only needs that we are able to integrate the second partial derivatives over Ω. This leads us to consider the spaces

$$\begin{aligned} L_2(\Omega) &:= \{v : \quad \iint_\Omega v^2 \,\mathrm{d}x\mathrm{d}y < \infty\}, && (10.36)\\ H^1(\Omega) &:= \{v : \quad v, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y} \in L_2(\Omega)\}, && (10.37)\\ H^2(\Omega) &:= \{v : \quad v, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}, \frac{\partial^2 v}{\partial x^2}, \frac{\partial^2 v}{\partial xy}, \frac{\partial^2 v}{\partial y^2} \in L_2(\Omega)\}. && (10.38)\end{aligned}$$

$L_2(\Omega)$ has as its norm

$$\|v\| := \iint_\Omega v^2 \,\mathrm{d}x\mathrm{d}y < \infty. \tag{10.39}$$

$H^1(\Omega)$ has as its norm

$$\|v\| := \iint_\Omega v^2 + (\nabla v)^2 \, \mathrm{d}x\mathrm{d}y < \infty. \tag{10.40}$$

The norm for $H^2(\Omega)$ is similarly defined.

To make this rigorous here you need to say precisely what is meant by the derivatives existing sufficiently to be integrable over Ω which will lead us to consider weak derivatives in this context. We will not however do this here. For our needs it is sufficient to note that if the derivatives exist throughout Ω except on a finite line segments across which there is a finite jump discontinuity then the function is in the space. In particular, if $\bar{\Omega}$ is divided into pieces $\Omega_1, \cdots, \Omega_{ne}$ (which are open domains) with $v \in C(\bar{\Omega})$ being a polynomial in x and y on each Ω_i then ∇v is defined sufficiently for $v \in H^1(\Omega)$ with

$$\|v\|^2 = \sum_{i=1}^{ne} \iint_{\Omega_i} v^2 + (\nabla v)^2 \, \mathrm{d}x\mathrm{d}y < \infty. \tag{10.41}$$

For such functions the gradient vector ∇v has finite jump discontinuities across common edges $\bar{\Omega}_i \cap \bar{\Omega}_j$ between any two neighbouring domains Ω_i and Ω_j.

As a comment, the connection between the spaces of continuous functions and these Sobolev spaces is different in $\mathbb{R}^2$ than it was in $\mathbb{R}$. It was stated previously that $H^1(0,1) \subset C[0,1]$. However in $\mathbb{R}^2$ we have $H^1(\Omega) \not\subset C(\Omega)$ but it can be shown that $H^2(\Omega) \subset C(\Omega)$. That is $H^1(\Omega)$ contains functions which are not continuous. However, as in the one-dimensional case, piecewise polynomial functions which are in $C(\bar{\Omega})$ are also in $H^1(\Omega)$ which is what we need in the method.

10.4 Summary

To transform Poisson's equation into weak form we use the vector identity

$$\nabla \cdot (v\nabla u) = v\Delta u + \nabla u \cdot \nabla v \tag{10.42}$$

and the divergence theorem

$$\int_{\partial\Omega} v\frac{\partial u}{\partial n} \, \mathrm{d}s = \iint_\Omega \nabla \cdot (v\nabla u) \, \mathrm{d}x\mathrm{d}y. \tag{10.43}$$

The weak form for Poisson's equation can be considered over the Sobolev space

$$H^1(\Omega) := \{v : \quad v, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y} \in L_2(\Omega)\}. \tag{10.44}$$

Piecewise polynomial functions in $C(\bar{\Omega})$ have finite jump discontinuities in the gradient vector ∇v are in $H^1(\Omega)$. Using these Sobolev's spaces the following weak forms are obtained.

1. **Homogeneous Dirichlet boundary conditions**

   PDE

   $$-\Delta u = f \quad \text{in } \partial\Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{10.45}$$

   Weak form

   $$\text{Find } u \in V \quad \text{such that} \quad \iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y = \iint_\Omega fv \, \mathrm{d}x\mathrm{d}y \quad \text{for all } v \in V \tag{10.46}$$

   where

   $$V = \{v \in H^1(\Omega) : \ v = 0 \text{ on } \partial\Omega\}. \tag{10.47}$$

2. A **Neumann boundary condition** case.

   PDE

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \text{ on } \partial\Omega_1 \quad \text{and} \quad \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega_2. \qquad (10.48)$$

   Weak form

$$\text{Find } u \in V \quad \text{such that} \quad \iint_\Omega \nabla u \cdot \nabla v \; \mathrm{d}x\mathrm{d}y = \iint_\Omega f v \; \mathrm{d}x\mathrm{d}y \quad \text{for all } v \in V \qquad (10.49)$$

   where

$$V = \{v \in H^1(\Omega) : \; v = 0 \text{ on } \partial\Omega_1\}. \qquad (10.50)$$

   The boundary condition on $\partial\Omega_2$ is a natural boundary condition.

3. A **mixed boundary condition** on $\partial\Omega_2$.

   PDE

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{on } \partial\Omega_1 \quad \text{and} \quad \frac{\partial u}{\partial n} + \beta u = g \text{ on } \partial\Omega_2. \qquad (10.51)$$

   Weak form Find $u \in V$ such that

$$\iint_\Omega \nabla u \cdot \nabla v \; \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} \beta u v \; \mathrm{d}s = \iint_\Omega f v \; \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega_2} g v \; \mathrm{d}s \qquad (10.52)$$

   for all $v \in V$ where

$$V = \{v \in H^1(\Omega) : \; v = 0 \text{ on } \partial\Omega_1\}. \qquad (10.53)$$

   The boundary condition on $\partial\Omega_2$ is a natural boundary condition for this weak form.

4. The **biharmonic equation**.

   PDE

$$\Delta^2 u = \Delta(\Delta u) = f, \quad u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \qquad (10.54)$$

   Weak form

$$\text{Find } u \in V \quad \text{such that} \quad \iint_\Omega \Delta u \Delta v \; \mathrm{d}x\mathrm{d}y = \iint_\Omega f v \; \mathrm{d}x\mathrm{d}y \quad \text{for all } v \in V \qquad (10.55)$$

   where

$$V = \{v \in H^2(\Omega) : \; v = \frac{\partial v}{\partial n} = 0 \text{ on } \partial\Omega\}. \qquad (10.56)$$

11 The FEM in $\mathbb{R}^2$ — the piecewise polynomial basis functions

11.1 Remarks about meshing

The finite element spaces V_h that can be used are defined with respect to a mesh of the domain Ω of the problem and before the details of the functions used are described we consider first what is needed to describe a mesh of a two dimensional domain Ω.

Given any domain Ω with a boundary $\partial\Omega$ which is polygonal (i.e. it involves only straight line segments) it is possible to subdivide Ω exactly into non-overlapping triangles $\Omega_1, \cdots, \Omega_{ne}$ so that

$$\bar{\Omega} = \bigcup_1^{ne} \bar{\Omega}_i, \quad \Omega_i \cap \Omega_j = \text{empty set}, \; i \neq j, \tag{11.1}$$

see Figures 11.1 for examples. If $\partial\Omega$ has curved sides then we can approximate it by a polygon and then mesh the polygon. In fact with a boundary $\partial\Omega$ which involves curved sides we can approximate it more accurately using curved elements which we can obtain using any of the quadratic elements. This will be mentioned when the 6-noded triangle and the 8 and 9 noded quadrilaterals are described.

The mesh generation capabilities of a finite element package is usually a significant component of these packages and it is something which is quite difficult to do in the case of 3 dimensional problems. We will not however consider any of the techniques which are used to generate a mesh for general shaped domains and we will restrict attention here to just making comments about what the mesh generator creates. In the case of a mesh of triangles it generates **nodes** $\hat{\underline{x}}_1, \cdots, \hat{\underline{x}}_M$ and a matrix of size $ne \times 3$ with row r containing 3 numbers i_{r_1}, i_{r_2} and i_{r_3} which gives the 3 nodes associated with the rth element. That is $\hat{\underline{x}}_{i_{r_1}}$, $\hat{\underline{x}}_{i_{r_2}}$ and $\hat{\underline{x}}_{i_{r_3}}$ are the 3 vertices describing the rth triangle. The matrix of size $ne \times 3$ gives what is known as the **connectivity** information. In the case of a mesh of 4-noded quadrilateral elements we similarly get a list of nodes together with a connectivity matrix which instead has 4 columns as there are 4 nodes associated with each quadrilateral. Other elements such as the 6-noded triangle, 8-noded quadrilateral and 9-noded quadrilateral give connectivity matrices with respectively 6, 8 and 9 columns.

In Figure 11.2 we show an example of a uniform mesh of a rectangle which can generated easily with the node numbers and the element numbers displayed. What this shows is that, for example, element number 19 has nodes 12, 18 and 13. It also shows that node 13 is a node of elements 11, 12, 14, 19, 21 and 22. When the computations are done element-by-element we get contributions to the matrix entries relating to node 13 from each of these elements.

11.2 The finite element basis functions and the finite element equations

With piecewise linears defined on a mesh of triangles the piecewise linear function $U(\underline{x})$ is completely defined by the values $U(\hat{\underline{x}}_i)$, $i = 1, \cdots, M$ at the nodes. We can associate with each node $\hat{\underline{x}}_i$ a piecewise linear function $\hat{\phi}_i(\underline{x})$ which satisfies the Lagrange interpolation condition

$$\hat{\phi}_i(\hat{\underline{x}}_j) = \begin{cases} 1, & j = i, \\ 0, & \text{if } j \neq i. \end{cases} \tag{11.2}$$

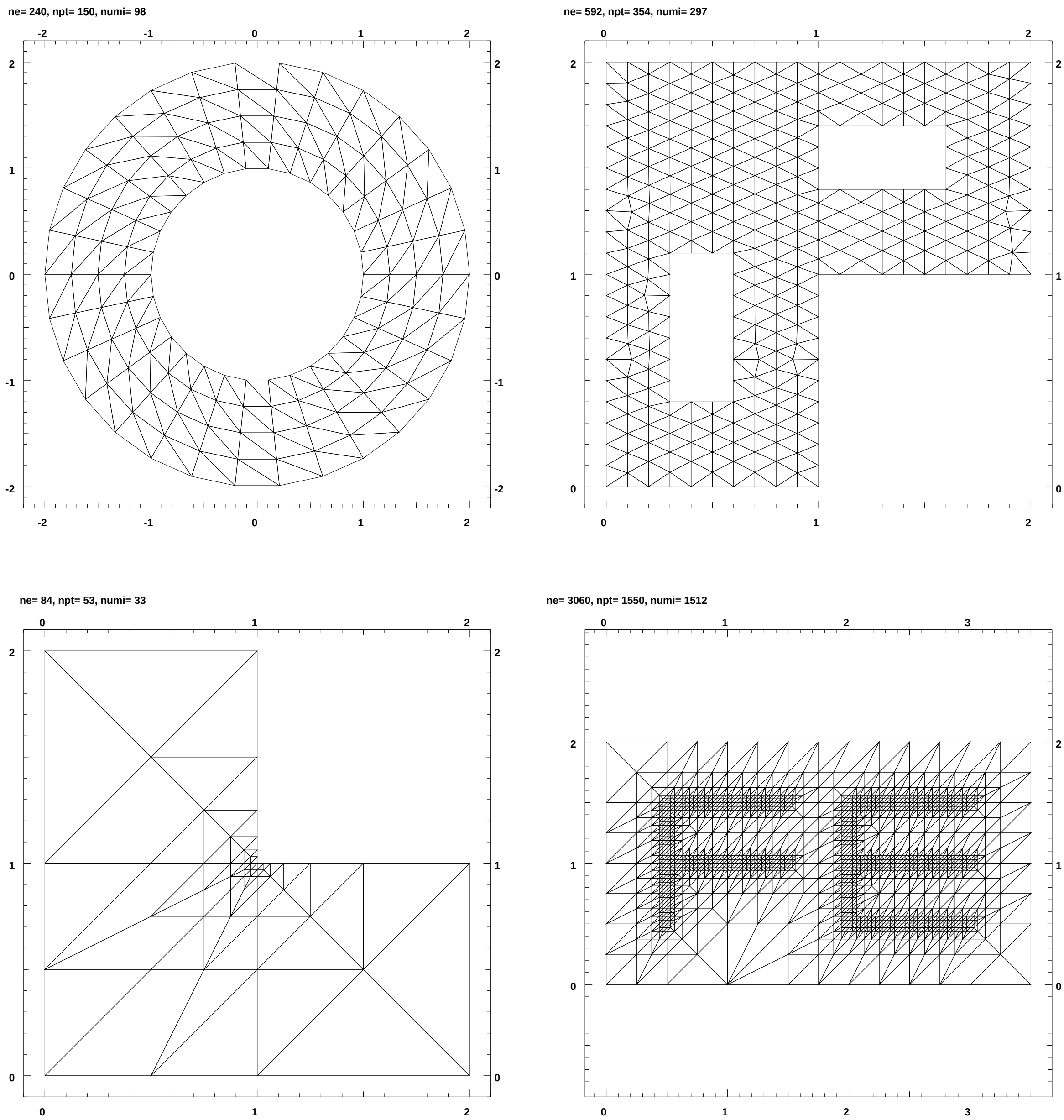


Figure 11.1: Examples of triangular meshes of polygonal domains.

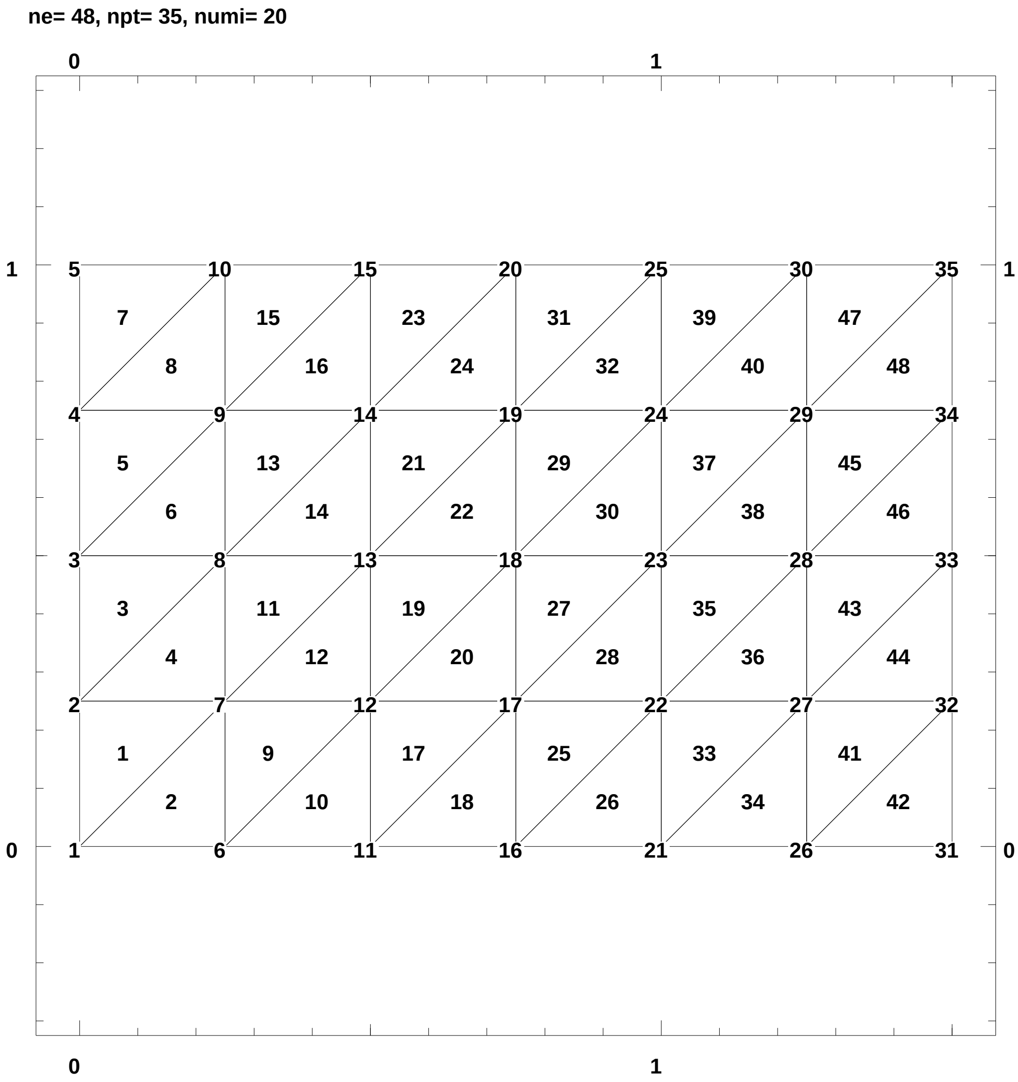


Figure 11.2: A mesh with the nodes and elements numbered.

A surface plot of such a function is shown in Figure 11.3. With these functions we can define

$$\bar{V}_h = \text{span}\{\hat{\phi}_1, \cdots, \hat{\phi}_M\}. \tag{11.3}$$

This set-up of one basis function associated with each node in the mesh will also be the case for all the elements considered in these notes and it was also the case for all the one-dimensional elements described earlier except for the case of the cubic Hermite elements. We have one basis function associated with each node when we have one unknown parameter at each node. To state this another way, the approximating function $U(\underline{x})$ is defined once all the nodal values are known.

For a given weak problem under consideration we need the subspace $V_h \subset \bar{V}_h$ involving functions which vanish on $\partial\Omega_1$ (which may be the entire boundary $\partial\Omega$) on which there is an essential boundary condition. If for the convenience of this description the nodes at which $U(\underline{x})$ is unknown are numbered so that $\hat{\underline{x}}_1, \cdots, \hat{\underline{x}}_{M_1}$ (these are the points not on $\partial\Omega_1$) and

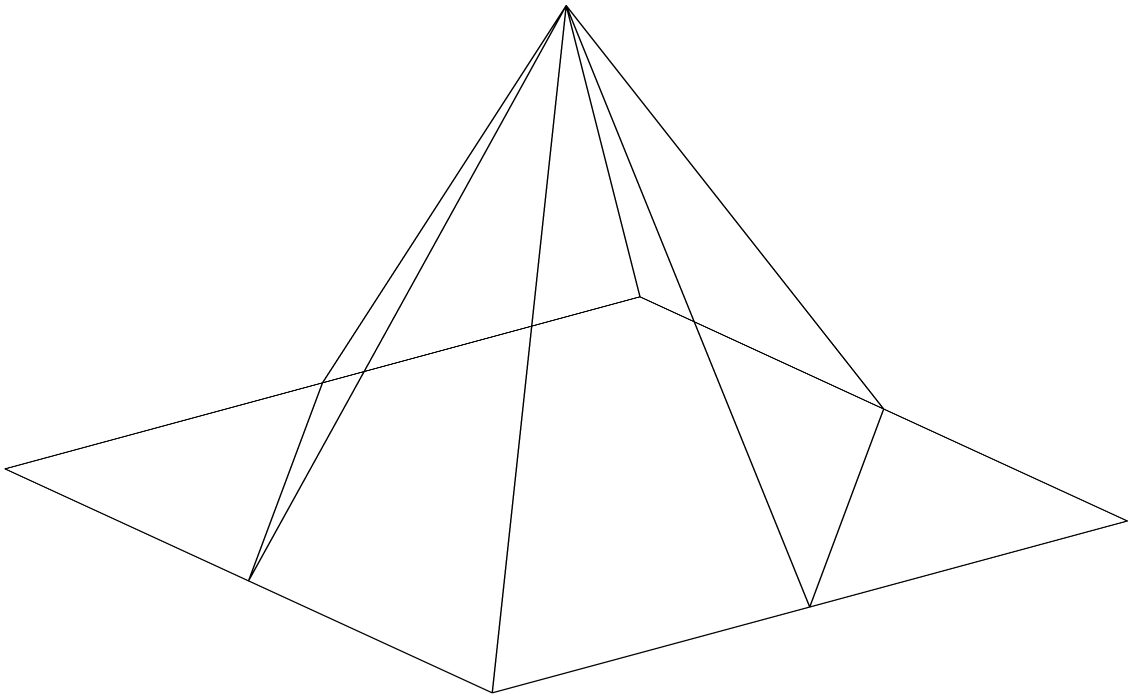


Figure 11.3: Surface plot of a typical function $\hat{\phi}_i$.

$\underline{\hat{x}}_{M_1+1}, \cdots, \underline{\hat{x}}_M$ are the points of $\partial\Omega_1$ then

$$V_h = \text{span}\{\hat{\phi}_1, \cdots, \hat{\phi}_{M_1}\}. \tag{11.4}$$

Assuming that we have a homogeneous Dirichlet condition on $\partial\Omega_1$ then the Galerkin method approximation is of the form

$$U(x) = \sum_{i=1}^{M_1} c_j \hat{\phi}_j(x) \tag{11.5}$$

where the Lagrange interpolation condition gives $c_j = U(\underline{\hat{x}}_j)$ and U satisfies

$$a(U, \hat{\phi}_i) = (f, \hat{\phi}_i), \quad i = 1, \cdots, M_1 \tag{11.6}$$

which in matrix-vector form gives

$$K\underline{c} = \underline{b}, \quad \underline{c} = (c_i), \quad K = (a(\hat{\phi}_i, \hat{\phi}_j)) \quad \text{and} \quad \underline{b} = ((f, \hat{\phi}_i)). \tag{11.7}$$

An extra computation that arises with the two dimensional problems as compared with one-dimensional problems is that to compute $a(\hat{\phi}_i, \hat{\phi}_j)$ and $(f, \hat{\phi}_i)$ generally involves several elements instead of just 1 or 2 neighbouring elements as in the one-dimensional case. The computations are always organised in an element-by-element way. For the Poisson problem which gives

$$a(u, v) = \iint_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y \quad \text{and} \quad (f, v) = \iint_\Omega fv \, \mathrm{d}x\mathrm{d}y \tag{11.8}$$

we define

$$a(u, v)_r = \iint_{\Omega_r} \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y \quad \text{and} \quad (f, v)_r = \iint_{\Omega_r} fv \, \mathrm{d}x\mathrm{d}y \tag{11.9}$$

so that

$$a(u,v) = \sum_{r=1}^{ne} a(u,v)_r \quad \text{and} \quad (f,v)_r = \sum_{r=1}^{ne} (f,v)_r. \tag{11.10}$$

The element-by-element way of organising things is to do every calculation which involves Ω_r and to put these together according to (11.10). The calculations to consider on element Ω_r involve all the basis functions which are non-zero on Ω_r and to describe these basis functions it is advantageous to refer everything to a standard element T. With a one-to-one and onto mapping $\underline{x} : T \to \Omega_r$ we associate with each point $(s,t) \in T$ a unique point $\underline{x}(s,t) \in \Omega_r$ and for each function $\phi_i = \phi_i(s,t)$ defined on T we have a related function $\tilde{\phi}_i(\underline{x})$ defined on Ω_r such that

$$\tilde{\phi}_i(\underline{x}(s,t)) = \phi_i(s,t). \tag{11.11}$$

We now consider the details of this for a few of the common elements that are used in practice.

11.3 Linear triangles

Let Ω_r denote a triangle with nodes $\underline{x}_1 = \underline{\hat{x}}_{i_{r_1}}$, $\underline{x}_2 = \underline{\hat{x}}_{i_{r_2}}$ and $\underline{x}_3 = \underline{\hat{x}}_{i_{r_3}}$ and let T denote the right angled triangle shown in Figure 11.4 which has vertices $(0,0)$, $(1,0)$ and $(0,1)$. T is a standard triangle. Linear basis functions satisfying the Lagrange interpolation condition defined on this standard triangle are given by

$$\phi_1(s,t) = 1 - s - t, \quad \phi_2(s,t) = s, \quad \phi_3(s,t) = t. \tag{11.12}$$

Using these we can describe both the mapping $\underline{x} : T \to \Omega_r$ and the approximating function U. An element with these properties is known as an **isoparametric element**. All the elements considered in this module are of this type. To describe the mapping and the approximation let $U_1 = U(\underline{x}_1)$, $U_2 = U(\underline{x}_2)$ and $U_3 = U(\underline{x}_3)$. We have

$$\begin{aligned} \underline{x}(s,t) &= \underline{x}_1\phi_1(s,t) + \underline{x}_2\phi_2(s,t) + \underline{x}_3\phi_3(s,t) = \underline{x}_1 + (\underline{x}_2 - \underline{x}_1)s + (\underline{x}_3 - \underline{x}_1)t, && (11.13)\\ U(\underline{x}(s,t)) &= U_1\phi_1(s,t) + U_2\phi_2(s,t) + U_3\phi_3(s,t) = U_1 + (U_2 - U_1)s + (U_3 - U_1)t. && (11.14) \end{aligned}$$

To verify that the mapping $\underline{x} = \underline{x}(s,t)$ does have the required mapping properties we can immediately observe that $\underline{s}_i \to \underline{x}_i$, $i = 1,2,3$ (i.e. the corners map to the corners), the sides of T map to the sides of Ω_r and the interior of ∂T maps to the interior of $\partial\Omega_r$. A mapping of the form (11.13) is known as an **affine transformation**, it is a combination of a linear transformation followed by a translation.

With everything referred to the standard triangle the functions are kept as simple as possible in terms of s and t. We can define the corresponding functions of x and y by using the mapping and we can relate these to the globally defined functions $\hat{\phi}_{i_{r_1}}$, $\hat{\phi}_{i_{r_2}}$ and $\hat{\phi}_{i_{r_3}}$ by the relations

$$\hat{\phi}_{i_{r_j}}(\underline{x}(s,t)) = \tilde{\phi}_j(\underline{x}(s,t)) = \phi_j(s,t), \quad j = 1,2,3. \tag{11.15}$$

That is $\tilde{\phi}_j$ is defined on Ω_r, ϕ_j is defined on T and (11.15) gives the value of $\hat{\phi}_{i_{r_j}}(\underline{x})$ at a point $\underline{x} \in \Omega_r$. The function $\hat{\phi}_{i_{r_j}}$ is non-zero on every triangle which has $\underline{\hat{x}}_{i_{r_j}}$ as one of its vertices.

The element matrix, which is the element stiffness matrix in this case, and the element vector are

$$K_r = \begin{pmatrix} a(\tilde{\phi}_1,\tilde{\phi}_1)_r & a(\tilde{\phi}_1,\tilde{\phi}_2)_r & a(\tilde{\phi}_1,\tilde{\phi}_3)_r \\ a(\tilde{\phi}_2,\tilde{\phi}_1)_r & a(\tilde{\phi}_2,\tilde{\phi}_2)_r & a(\tilde{\phi}_2,\tilde{\phi}_3)_r \\ a(\tilde{\phi}_3,\tilde{\phi}_1)_r & a(\tilde{\phi}_3,\tilde{\phi}_2)_r & a(\tilde{\phi}_3,\tilde{\phi}_3)_r \end{pmatrix} \quad \text{and} \quad \underline{b}_r = \begin{pmatrix} (f,\tilde{\phi}_1)_r \\ (f,\tilde{\phi}_2)_r \\ (f,\tilde{\phi}_3)_r \end{pmatrix}. \tag{11.16}$$

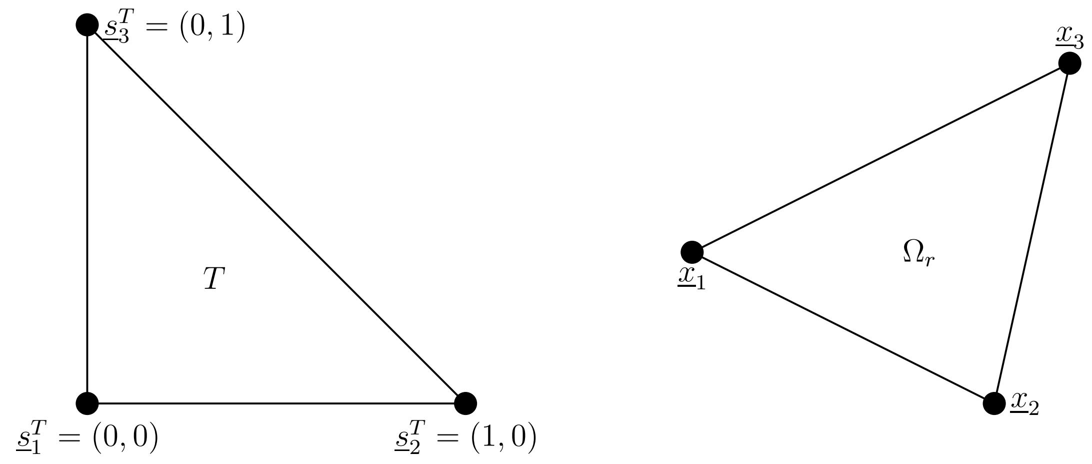


Figure 11.4: The standard triangle T and an actual triangle Ω_r.

We can determine K_r explicitly in this case as follows. The matrix entries are

$$a(\tilde{\phi}_i, \tilde{\phi}_j)_r = \iint_{\Omega_r} \nabla\tilde{\phi}_i \cdot \nabla\tilde{\phi}_j \,\mathrm{d}x\mathrm{d}y = \iint_{\Omega_r} \left(\frac{\partial\tilde{\phi}_i}{\partial x}\frac{\partial\tilde{\phi}_j}{\partial x} + \frac{\partial\tilde{\phi}_i}{\partial y}\frac{\partial\tilde{\phi}_j}{\partial y} \right) \mathrm{d}x\mathrm{d}y, \quad 1 \le i,j \le 3. \tag{11.17}$$

Now as each $\tilde{\phi}_j$ is linear in x and y the gradient $\nabla\tilde{\phi}_i$ is a constant vector and the integrand here is constant. Thus

$$a(\tilde{\phi}_i, \tilde{\phi}_j)_r = (\text{area of } \Omega_r) \left(\frac{\partial\tilde{\phi}_i}{\partial x}\frac{\partial\tilde{\phi}_j}{\partial x} + \frac{\partial\tilde{\phi}_i}{\partial y}\frac{\partial\tilde{\phi}_j}{\partial y} \right). \tag{11.18}$$

Now if we let

$$B = \begin{pmatrix} \frac{\partial\tilde{\phi}_1}{\partial x} & \frac{\partial\tilde{\phi}_2}{\partial x} & \frac{\partial\tilde{\phi}_3}{\partial x} \\ \frac{\partial\tilde{\phi}_1}{\partial y} & \frac{\partial\tilde{\phi}_2}{\partial y} & \frac{\partial\tilde{\phi}_3}{\partial y} \end{pmatrix} \in \mathbb{R}^{2,3} \tag{11.19}$$

then B^TB is a 3×3 matrix with

$$\begin{aligned} (B^TB)_{ij} &= (\text{row } i \text{ of } B^T)(\text{column } j \text{ of } B) & (11.20) \\ &= \text{product of columns } i \text{ and } j \text{ of } B = \nabla\tilde{\phi}_i \cdot \nabla\tilde{\phi}_j. & (11.21) \end{aligned}$$

Hence we can describe the matrix K_r as

$$K_r = (\text{area of } \Omega_r) B^T B. \tag{11.22}$$

We now relate things to the coordinates $\underline{x}_1 = (x_1, y_1)^T$, $\underline{x}_2 = (x_2, y_2)^T$ and $\underline{x}_3 = (x_3, y_3)^T$ of the vertices of the triangle. From the transformation $\underline{x} = \underline{x}(s,t)$ we can get the area of Ω_r and we can relate the derivatives with respect to x and y to those in terms of s and t. The mapping is

$$\underline{x}(s,t) = \underline{x}_1 + (\underline{x}_2 - \underline{x}_1)s + (\underline{x}_3 - \underline{x}_1)t = \underline{x}_1 + \bar{\underline{x}}_2 s + \bar{\underline{x}}_3 t = \underline{x}_1 + \tilde{J}\begin{pmatrix} s \\ t \end{pmatrix} \tag{11.23}$$

where $\bar{\underline{x}}_2 = \underline{x}_2 - \underline{x}_1$ and $\bar{\underline{x}}_3 = \underline{x}_3 - \underline{x}_1$ and where

$$\tilde{J} = \begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix} = \begin{pmatrix} \bar{x}_2 & \bar{x}_3 \\ \bar{y}_2 & \bar{y}_3 \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \tag{11.24}$$

is the constant Jacobian matrix. With $J = \det(\tilde{J})$ the magnitude $|J|$ of J gives the ratio of an area increment in Ω_r to the corresponding increment in T and since J is constant on T and the area of T is $1/2$ it follows that

$$\text{the area of } \Omega_r = \frac{|J|}{2}. \tag{11.25}$$

For the relation between the derivatives we have for any function $\tilde{v}(x, y) = v(s, t)$ the chain rule

$$\frac{\partial v}{\partial s} = \frac{\partial x}{\partial s}\frac{\partial \tilde{v}}{\partial x} + \frac{\partial y}{\partial s}\frac{\partial \tilde{v}}{\partial y} \tag{11.26}$$

$$\frac{\partial v}{\partial t} = \frac{\partial x}{\partial t}\frac{\partial \tilde{v}}{\partial x} + \frac{\partial y}{\partial t}\frac{\partial \tilde{v}}{\partial y} \tag{11.27}$$

and we can write this in matrix-vector form as

$$\begin{pmatrix} \frac{\partial v}{\partial s} \\ \frac{\partial v}{\partial t} \end{pmatrix} = \begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial y}{\partial s} \\ \frac{\partial x}{\partial t} & \frac{\partial y}{\partial t} \end{pmatrix} \begin{pmatrix} \frac{\partial \tilde{v}}{\partial x} \\ \frac{\partial \tilde{v}}{\partial y} \end{pmatrix} = \tilde{J}^T \begin{pmatrix} \frac{\partial \tilde{v}}{\partial x} \\ \frac{\partial \tilde{v}}{\partial y} \end{pmatrix}. \tag{11.28}$$

Thus

$$\begin{pmatrix} \frac{\partial \tilde{v}}{\partial x} \\ \frac{\partial \tilde{v}}{\partial y} \end{pmatrix} = \tilde{J}^{-T} \begin{pmatrix} \frac{\partial v}{\partial s} \\ \frac{\partial v}{\partial t} \end{pmatrix} \tag{11.29}$$

where $\tilde{J}^{-T}$ means the transpose of the inverse $\tilde{J}^{-1}$. Taking $\tilde{v}$ in turn as $\tilde{\phi}_1$, $\tilde{\phi}_2$ and $\tilde{\phi}_3$ we have

$$B = \begin{pmatrix} \frac{\partial \tilde{\phi}_1}{\partial x} & \frac{\partial \tilde{\phi}_2}{\partial x} & \frac{\partial \tilde{\phi}_3}{\partial x} \\ \frac{\partial \tilde{\phi}_1}{\partial y} & \frac{\partial \tilde{\phi}_2}{\partial y} & \frac{\partial \tilde{\phi}_3}{\partial y} \end{pmatrix} = \tilde{J}^{-T} \begin{pmatrix} \frac{\partial \phi_1}{\partial s} & \frac{\partial \phi_2}{\partial s} & \frac{\partial \phi_3}{\partial s} \\ \frac{\partial \phi_1}{\partial t} & \frac{\partial \phi_2}{\partial t} & \frac{\partial \phi_3}{\partial t} \end{pmatrix} = \tilde{J}^{-T} \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}. \tag{11.30}$$

For the matrix K_r we have

$$K_r = \frac{|J|}{2} B^T B = \frac{|J|}{2} \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tilde{J}^{-1} \tilde{J}^{-T} \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}. \tag{11.31}$$

Now

$$\tilde{J} = \begin{pmatrix} \bar{x}_2 & \bar{x}_3 \\ \bar{y}_2 & \bar{y}_3 \end{pmatrix} \quad \text{gives} \quad \tilde{J}^{-1} = \frac{1}{J} \begin{pmatrix} \bar{y}_3 & -\bar{x}_3 \\ -\bar{y}_2 & \bar{x}_2 \end{pmatrix} \quad \text{and} \quad J = \bar{x}_2 \bar{y}_3 - \bar{x}_3 \bar{y}_2 \tag{11.32}$$

and we can also write this relation for K_r as

$$K_r = \frac{1}{2|J|} A^T A \quad \text{where} \quad A^T = \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_3 & -\bar{x}_3 \\ -\bar{y}_2 & \bar{x}_2 \end{pmatrix}. \tag{11.33}$$

Given the coordinates (x_1, y_1), (x_2, y_2) and (x_3, y_3) of the 3 vertices of Ω_r this gives a formula for computing K_r.

Before we consider the matrix K_r for two specific triangles we first make some observations about what exactly the matrix K_r depends on. From the expressions for K_r we see that it only depends on the differences $\bar{\underline{x}}_2 = \underline{x}_2 - \underline{x}_1$ and $\bar{\underline{x}}_3 = \underline{x}_3 - \underline{x}_1$ and hence if we translate the coordinates then the matrix does not change. Also if we do a uniform magnification by the factor $\alpha > 0$ then the Jacobian matrix $\tilde{J}_\alpha$ of the new mapping is $\tilde{J}_\alpha = \alpha\tilde{J}$ with $J_\alpha = \det(\tilde{J}_\alpha) = \alpha^2 \det(\tilde{J})$ and

$$|J_\alpha|\tilde{J}_\alpha^{-1}\tilde{J}_\alpha^{-T} = |J|\tilde{J}^{-1}\tilde{J}^{-T}. \tag{11.34}$$

This shows that the element stiffness matrix is the same for all uniform scalings. Also note that if the triangle is subjected to any orthogonal transformation (a rotation or a reflection) then this corresponds to an affine transformation with a Jacobian matrix of the form

$$\tilde{J}_Q = Q\tilde{J} \tag{11.35}$$

where Q is a 2×2 orthogonal matrix, i.e. $Q^TQ = QQ^T = I$, $Q^{-1} = Q^T$. Now $Q^TQ = I$ implies that $(\det(Q))^2 = 1$ so that $|J_Q| = |J|$. Also

$$\tilde{J}_Q^{-1} = \tilde{J}^{-1}Q^{-1} = \tilde{J}^{-1}Q^T \quad \text{so that} \quad \tilde{J}_Q^{-1}\tilde{J}_Q^{-T} = \tilde{J}^{-1}Q^TQ\tilde{J}^{-T} = \tilde{J}^{-1}\tilde{J}^{-T}. \tag{11.36}$$

Thus K_r does not change if we we rotate or reflect the triangle. The invariance of K_r to translation, scalings, rotations and reflections means that K_r only depends on the angles of the triangle.

We next consider two specific triangles.

11.3.1 K_r when $\Omega_r = T$, the standard triangle

When Ω_r coincides with the standard triangle with $\underline{x}_1^T = (0,0)$, $\underline{x}_2^T = (1,0)$ and $\underline{x}_3^T = (0,1)$ the mapping is the identity mapping, i.e. $\tilde{J} = I$ and $J = \det(\tilde{J}) = 1$ and we have

$$K_r = \frac{1}{2}\begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}. \tag{11.37}$$

11.3.2 K_r for an equilateral triangle

We now consider the equilateral triangle shown in Figure 11.5 with vertices at $\underline{x}_1 = (0,0)^T$, $\underline{x}_2 = (1,0)^T$ and $\underline{x}_3 = (1/2, \sqrt{3}/2)^T$. The area of the triangle is $\sqrt{3}/4$, and $|J| = \sqrt{3}/2$. Here

$$\begin{pmatrix} \bar{y}_3 & -\bar{x}_3 \\ -\bar{y}_2 & \bar{x}_2 \end{pmatrix} = \begin{pmatrix} y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix} = \begin{pmatrix} \sqrt{3}/2 & -1/2 \\ 0 & 1 \end{pmatrix} \tag{11.38}$$

giving

$$A^T = \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \sqrt{3}/2 & -1/2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -\sqrt{3}/2 & -1/2 \\ \sqrt{3}/2 & -1/2 \\ 0 & 1 \end{pmatrix} \tag{11.39}$$

and

$$K_r = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix} = \frac{1}{2\sqrt{3}}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}. \tag{11.40}$$

We get the same matrix here for all orders of the points $\underline{x}_1$, $\underline{x}_2$ and $\underline{x}_3$.

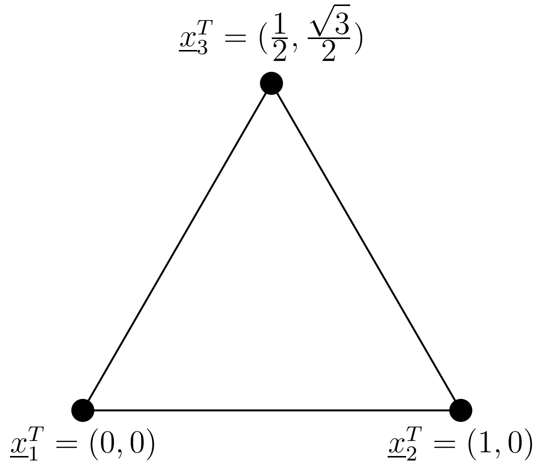


Figure 11.5: An equilateral triangle.

11.4 The element vector and quadrature

The 3×1 vector which involves integrals of the right hand side function f over the element Ω_r cannot usually be done exactly unless f is fairly simple and it is instead approximated. To perform the computations we first map the region of integration Ω_r to the standard triangle T and then use a quadrature rule. We have

$$\begin{aligned}
\underline{b}_r &= \iint_T f(\underline{x}(s,t)) \begin{pmatrix} \phi_1(s,t) \\ \phi_2(s,t) \\ \phi_3(s,t) \end{pmatrix} |J| \,\mathrm{d}s\mathrm{d}t && (11.41)\\
&= |J| \int_{t=0}^{t=1} \int_{s=0}^{s=1-t} f(\underline{x}(s,t)) \begin{pmatrix} 1-s-t \\ s \\ t \end{pmatrix} \mathrm{d}s\mathrm{d}t && (11.42)\\
&\approx |J| \sum_{k=1}^{nq} w_k f(\underline{x}(s_k,t_k)) \begin{pmatrix} 1-s_k-t_k \\ s_k \\ t_k \end{pmatrix} && (11.43)
\end{aligned}$$

where $(s_k, t_k) \in T$, $k = 1, \cdots, nq$ are the quadrature points and $w_1, \cdots w_{nq}$ are the weights. Possible schemes that can we used are given in Table 11.1. The centroid $(1/3, 1/3)$ of the standard triangle T maps to the centroid of the actual triangle and the simplest scheme corresponding to $nq = 1$ just involves evaluating f at this point. To check the claim that it does integrate exactly 1, s and t observe that

$$\begin{aligned}
\int_0^1 \int_0^{1-t} \mathrm{d}s\mathrm{d}t &= 1/2, && (11.44)\\
\int_0^1 \int_0^{1-t} s\,\mathrm{d}s\mathrm{d}t &= \int_0^1 \frac{1}{2}(1-t)^2\,\mathrm{d}t = \left[\frac{(t-1)^3}{6}\right]_0^1 = \frac{1}{6}, && (11.45)\\
\int_0^1 \int_0^{1-t} t\,\mathrm{d}s\mathrm{d}t &= \int_0^1 t(1-t)\,\mathrm{d}t = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}. && (11.46)
\end{aligned}$$

In the first case we get the area of T. In the case of integrating s and t the value of these at the centroid is $1/3$ and $(1/2)(1/3) = 1/6$.

Table 11.1: Quadrature points and weights for integrating over the standard triangle

nq	Points	Weights	Degree of polynomial in s and t which is integrated exactly by the formula.
1	$(1/3, 1/3)$	$1/2$	1 (i.e. 1, s and t)
3	$(1/2, 0)$, $(1/2, 1/2)$, $(0, 1/2)$.	$1/6$, $1/6$, $1/6$.	2 (i.e. 1, s, t, s^2, st and t^2)

11.5 Assembling the element contributions and the global matrix and the global vector

In a finite element calculation using linear triangles for Poisson's problem with a mesh of ne triangles and M nodes we get element matrices K_r and element vectors $\underline{b}_r$ for $r = 1, \cdots, ne$ and these have to be appropriately combined to create the $M \times M$ global matrix $\hat{K}$ and the $M \times 1$ global vector $\hat{\underline{b}}$ given by

$$\hat{K} = (a(\hat{\phi}_i, \hat{\phi}_j)) \quad \text{and} \quad \hat{\underline{b}} = ((f, \hat{\phi}_i)). \tag{11.47}$$

This operation is known as **assembly** and this is perhaps most easily described in the form of a computer algorithm. What is done is we start with $\hat{K}$=zero matrix and $\hat{\underline{b}} = \underline{0}$ and then proceed as follows.

$$
\begin{aligned}
&\text{for } r = 1, \cdots, ne \qquad (\textit{element loop})\\
&\qquad \hat{\underline{b}}_{i_{r_k}} \leftarrow \hat{\underline{b}}_{i_{r_k}} + (\underline{b}_r)_k, \quad 1 \le k \le 3.\\
&\qquad \hat{K}_{i_{r_k}, i_{r_l}} \leftarrow \hat{K}_{i_{r_k}, i_{r_l}} + (K_r)_{kl}, \quad 1 \le k, l \le 3.
\end{aligned}
$$

The rth element has nodes i_{r_1}, i_{r_2} and i_{r_3} with the locally numbered functions ϕ_1, ϕ_2 and ϕ_3 corresponding to part of the specification of the globally defined functions $\hat{\phi}_{i_{r_1}}$, $\hat{\phi}_{i_{r_2}}$ and $\hat{\phi}_{i_{r_3}}$. Adding the contributions involves adding the part resulting from the integrals over Ω_r and once everything has been done the final values correspond to the integrals over the entire domain Ω.

The matrix $\hat{K}$ that is created is sparse but the non-zero entries are not tightly banded about the diagonal as in the case in one space dimension with the precise structure depending on how the nodes have been numbered. The entries on row i are all connected with the function $\hat{\phi}_i$ and this is non-zero only on the elements which have i as one of the nodes. It is the collection of all the nodes on all of these elements which gives the possible non-zero entries on row i of the matrix. For the uniform mesh shown in Figure 11.2 this involves, for example, node 13 being connected to nodes 7, 8, 12, 13, 14, 18 and 19 and thus row 13 has up to 7 non-zero entries. This is fairly typical with triangular meshes for this type of problem.

11.6 Dealing with homogeneous boundary conditions

With a mesh with M nodes and with $\hat{\phi}_i(\underline{x})$, $i = 1, \cdots, M$ denoting the piecewise linear basis functions all piecewise linear functions defined with respect to this mesh of Ω are in the space

$$\hat{V}_h = \text{span}\{\hat{\phi}_1, \cdots, \hat{\phi}_M\}. \tag{11.48}$$

For a problem with the boundary condition $u = 0$ on $\partial\Omega_1$ and with $\mathcal{N}$ denoting the set of nodes on $\partial\Omega_1$ and with $\mathcal{N}' = \{1, \cdots, M\}\backslash\mathcal{N}$ denoting the other nodes then the finite element space to use is

$$V_h = \text{span}\{\hat{\phi}_i : \quad i \in \mathcal{N}'\}. \tag{11.49}$$

The finite element approximation that we seek is of the form

$$U(\underline{x}) = \sum_1^M U(\hat{\underline{x}}_i)\hat{\phi}_i(\underline{x}) = \sum_{i\in\mathcal{N}'} U(\hat{\underline{x}}_i)\hat{\phi}_i(\underline{x}) \in V_h \tag{11.50}$$

and this satisfies

$$a(U, \hat{\phi}_i) = (f, \hat{\phi}_i), \quad i \in \mathcal{N}'. \tag{11.51}$$

If there are M_1 nodes in $\mathcal{N}'$ then this gives us a system of M_1 equations in the M_1 unknown parameters of $U(\underline{x})$. If this system is written as $K\underline{c} = \underline{b}$ then we can obtain K and $\underline{b}$ from $\hat{K}$ and $\hat{\underline{b}}$ by removing the appropriate rows and columns in the case of K and just the rows in the case of $\underline{b}$ corresponding to entries in the set $\mathcal{N}$. For example, if the nodes in $\mathcal{N}'$ are the first M_1 nodes then

$$K = (a(\hat{\phi}_i, \hat{\phi}_j)), \quad \underline{b} = ((f, \hat{\phi}_i)), \quad 1 \le i, j \le M_1. \tag{11.52}$$

11.7 Example

Consider the square $\bar{\Omega} = [0, 1] \times [0, 1]$ which is divided into the 4 right angled isosceles triangles shown in Figure 11.6 which involves one interior node at $(1/2, 1/2)$ and 4 boundary nodes. The connectivity data for this mesh can be taken as follows

elem numbers	node numbers		
1	3	2	1
2	3	1	4
3	3	5	2
4	3	4	5

For each triangle the first local node corresponds to node 3 where we have the right angle and as each triangle can be obtained by rotating one of the other triangles the element stiffness matrix in each case is

$$K_r = \frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad r = 1, 2, 3, 4. \tag{11.53}$$

We need to assemble these to create the 4×4 matrix $\hat{K}$. $\hat{K}_{33}$ has contributions from all 4 elements and we get $\hat{K}_{33}=4$. For the other diagonal entries we have contributions from just two elements and we get

$$(\hat{K})_{11}=(\hat{K})_{22}=(\hat{K})_{44}=(\hat{K})_{55}=1. \tag{11.54}$$

There is no contribution to the $(\hat{K})_{21}$ entry from element 1 and hence $(\hat{K})_{21}=0$. For the $(\hat{K})_{31}$ entry we get contributions of $-1/2$ from element 1 and element 2 giving $(\hat{K})_{31}=-1$. All other possibilities correspond to one of these cases and we get the 5×5 matrix

$$\hat{K}=\begin{pmatrix} 1 & 0 & -1 & 0 & 0\\ 0 & 1 & -1 & 0 & 0\\ -1 & -1 & 4 & -1 & -1\\ 0 & 0 & -1 & 1 & 0\\ 0 & 0 & -1 & 0 & 1 \end{pmatrix}. \tag{11.55}$$

We now consider the approximate solution of $-\Delta u = f = 1$ on Ω in the case of two different boundary conditions using the mesh of these 4 triangles. In both cases we need the element vectors. As the right hand side function $f=1$, each triangle has area of $1/4$ and the value of each basis function is $1/3$ at the centroid we have

$$\underline{b}_r=\frac{1}{4}\begin{pmatrix}\frac{1}{3}\\ \frac{1}{3}\\ \frac{1}{3}\end{pmatrix}=\frac{1}{12}\begin{pmatrix}1\\1\\1\end{pmatrix}, \quad r=1,2,3,4. \tag{11.56}$$

Assembling these gives the global vector

$$\underline{\hat{b}}=\frac{1}{12}\begin{pmatrix}2\\2\\4\\2\\2\end{pmatrix}=\frac{1}{6}\begin{pmatrix}1\\1\\2\\1\\1\end{pmatrix}. \tag{11.57}$$

1. If the problem is $-\Delta u=1$ in Ω with the boundary condition $u=0$ on $\partial\Omega$ then with this coarse mesh U is only unknown at node number 3, i.e. at $(1/2,1/2)$. The solution is of the form
$$U(\underline{x})=U_3\hat{\phi}_3(\underline{x}), \quad \text{where } U_3=U(\underline{\hat{x}}_3). \tag{11.58}$$
We hence have the single equation
$$4U_3=(\underline{\hat{b}})_3=\frac{4}{12} \quad \text{giving } U_3=\frac{1}{12}. \tag{11.59}$$

2. Suppose now that the problem is $-\Delta u=1$ in Ω with the boundary conditions
$$u(x,0) = 0, \quad 0\le x\le 1, \tag{11.60}$$
$$u(0,y) = 0, \quad 0\le y\le 1, \tag{11.61}$$
$$\frac{\partial u}{\partial n} = 0, \quad \text{on the remainder of the boundary } \partial\Omega. \tag{11.62}$$
U is now unknown at nodes 3 and 5, the approximation is of the form
$$U(\underline{x})=U_3\hat{\phi}_3(\underline{x})+U_5\hat{\phi}_5(\underline{x}) \tag{11.63}$$

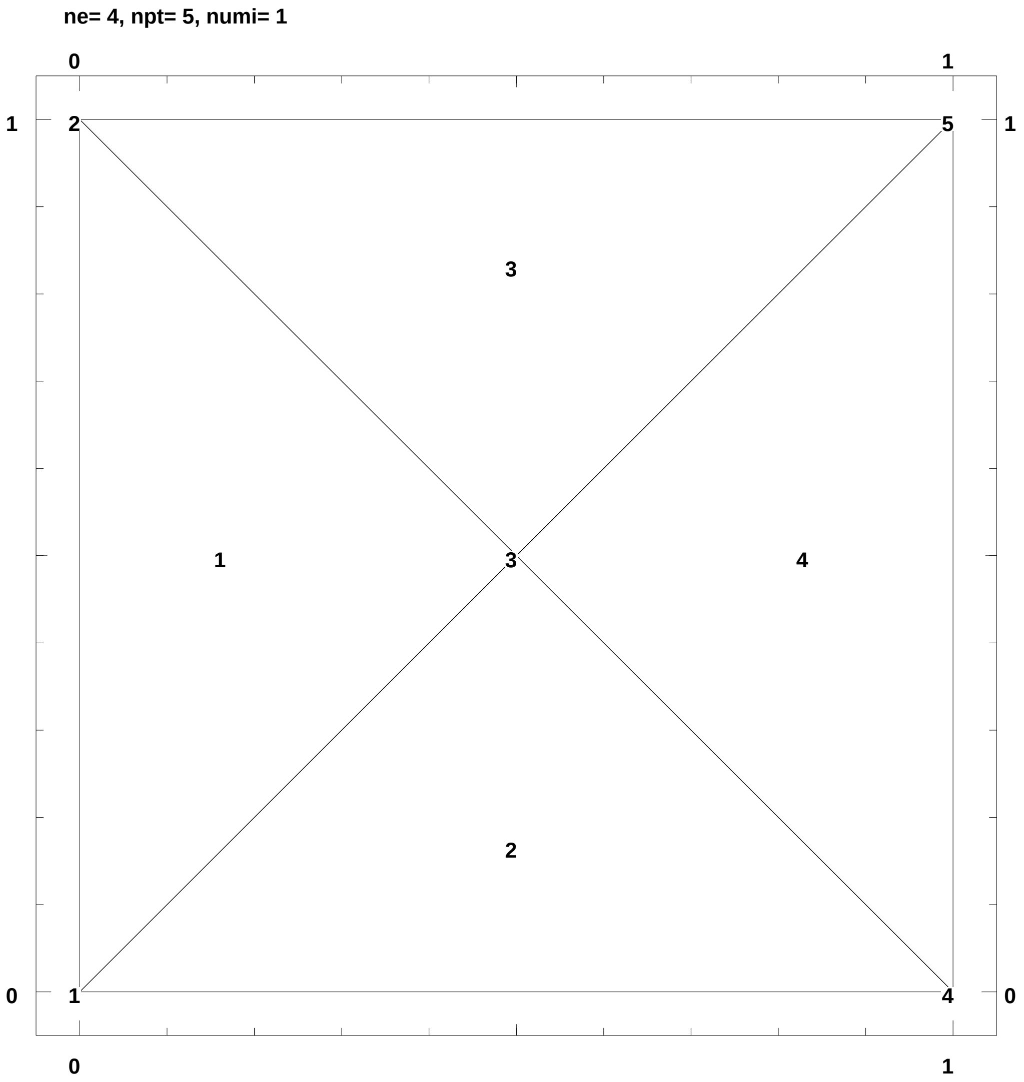


Figure 11.6: A mesh of $[0,1]\times[0,1]$ involving 4 triangles.

and we get a 2×2 system

$$\begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} U_3 \\ U_5 \end{pmatrix} = \begin{pmatrix} (\hat{\underline{b}})_3 \\ (\hat{\underline{b}})_5 \end{pmatrix} = \frac{1}{12}\begin{pmatrix} 4 \\ 2 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{11.64}$$

The solution is given by

$$\begin{pmatrix} U_3 \\ U_5 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}\frac{1}{6}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \frac{1}{18}\begin{pmatrix} 3 \\ 6 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{11.65}$$

With this coarse mesh the approximate solution at $(1/2,1/2)$ is $1/6$ and the approximate solution at $(1,1)$ is $1/3$.

11.8 Evaluating the approximation $U(\underline{x})$ and the gradient $\nabla U(\underline{x})$

Once we have solved the linear system $K\underline{c} = \underline{b}$ we get, using $\underline{c}$ and the boundary conditions, the values $U(\hat{\underline{x}}_i)$, $i = 1, \cdots, M$, i.e. we have the value of the approximation at all the mesh points. To evaluate $U(\underline{x})$ at any other point we need to first determine which element the point $\underline{x}$ lies. If we find that $\underline{x} \in \Omega_r$ and Ω_r has nodes $\underline{x}_1 = \hat{\underline{x}}_{i_{r_1}}$, $\underline{x}_2 = \hat{\underline{x}}_{i_{r_2}}$ and $\underline{x}_3 = \hat{\underline{x}}_{i_{r_3}}$ and $U_i = U(\underline{x}_i)$, $i = 1, 2, 3$ then

$$\begin{aligned} \underline{x} &= \underline{x}_1\phi_1(s,t) + \underline{x}_2\phi_2(s,t) + \underline{x}_3\phi_3(s,t) && (11.66)\\ &= \underline{x}_1 + (\underline{x}_2 - \underline{x}_1)s + (\underline{x}_3 - \underline{x}_1)t = \underline{x}_1 + \tilde{J}\begin{pmatrix} s \\ t \end{pmatrix}, && (11.67)\\ U(\underline{x}(s,t)) &= U_1\phi_1(s,t) + U_2\phi_2(s,t) + U_3\phi_3(s,t) && (11.68)\\ &= U_1 + (U_2 - U_1)s + (U_3 - U_1)t. && (11.69) \end{aligned}$$

That is given $\underline{x}$ we determine s and t from

$$\begin{pmatrix} s \\ t \end{pmatrix} = \tilde{J}^{-1}(\underline{x} - \underline{x}_1) \tag{11.70}$$

and then we obtain the value of $U(\underline{x})$ from (11.69).

Observe that if $t = 0$ then $U(\underline{x}(s,0)) = U_1 + (U_2 - U_1)s$, if $s = 0$ then $U(\underline{x}(0,t)) = U_1 + (U_3 - U_1)t$, and if $t = 1 - s$ then $U(\underline{x}(s, 1-s)) = U_3 + (U_2 - U_3)s$. On all the edges of the element the approximation varies linearly and it depends only on the nodal values of U at the nodes on the edge. It is because of this that U is continuous across an element edge and we have $U \in C(\bar{\Omega})$.

As a consequence of (11.69) and (11.70) $U(\underline{x})$ is linear in x and y on the triangle and this implies that the gradient vector ∇U is constant on each triangle Ω_r. Thus $\nabla U(\underline{x})$ is a piecewise constant vector valued function on Ω. In general ∇U has a discontinuity across an element edge which is why in general $U \notin C^1(\Omega)$. To determine the constant vector ∇U on the rth element we need to proceed in a similar way to that used to determine the element matrix K_r. We have

$$\begin{pmatrix} \dfrac{\partial U}{\partial s} \\ \dfrac{\partial U}{\partial t} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial x}{\partial s} & \dfrac{\partial y}{\partial s} \\ \dfrac{\partial x}{\partial t} & \dfrac{\partial y}{\partial t} \end{pmatrix} \begin{pmatrix} \dfrac{\partial U}{\partial x} \\ \dfrac{\partial U}{\partial y} \end{pmatrix} = \tilde{J}^T \begin{pmatrix} \dfrac{\partial U}{\partial x} \\ \dfrac{\partial U}{\partial y} \end{pmatrix} \tag{11.71}$$

so that

$$\nabla U = \begin{pmatrix} \dfrac{\partial U}{\partial x} \\ \dfrac{\partial U}{\partial y} \end{pmatrix} = \tilde{J}^{-T} \begin{pmatrix} \dfrac{\partial U}{\partial s} \\ \dfrac{\partial U}{\partial t} \end{pmatrix} = \tilde{J}^{-T} \begin{pmatrix} U_2 - U_1 \\ U_3 - U_1 \end{pmatrix} \tag{11.72}$$

where as before

$$\tilde{J} = \begin{pmatrix} \dfrac{\partial x}{\partial s} & \dfrac{\partial x}{\partial t} \\ \dfrac{\partial y}{\partial s} & \dfrac{\partial y}{\partial t} \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}. \tag{11.73}$$

This tells us how to calculate ∇U and it shows us explicitly that ∇U depends on all the nodal values U_1, U_2 and U_3. We show next what this involves for each of two adjacent right angles triangles shown in Figure 11.7. The derivative of U in the direction tangential to the element edge is the same for both triangles but the other component of ∇U involves the value of U at the other node of the triangle in each case and in general $U_3 - U_1 \neq U_1 - U_3'$.

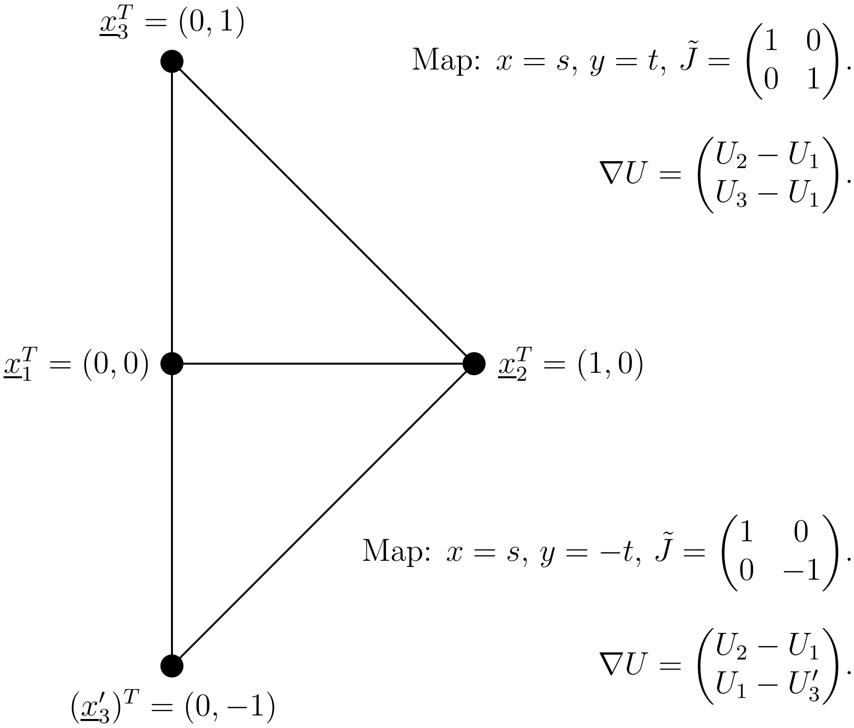


Figure 11.7: The gradient vector ∇U on two adjacent triangles.

11.9 4-noded bilinear elements

Instead of partitioning a domain Ω into triangles we may instead partition it into quadrilateral elements and use an approximating function which involves bilinear functions. As before we refer to a standard element which we choose here to take as $-1 \le s, t \le 1$. The standard element and a typical actual element Ω_r are shown in Figure 11.8.

The bilinear functions involve combinations of 1, s, t and st and with quadrilateral elements the Lagrange basis functions can be obtained by simply combining the linear functions $(1-s)/2$ and $(1+s)/2$ in the s direction with $(1-t)/2$ and $(1+t)/2$ in the t direction. We have

$$\phi_1(s,t) = \frac{1}{4}(1-s)(1-t), \tag{11.74}$$

$$\phi_2(s,t) = \frac{1}{4}(1+s)(1-t), \tag{11.75}$$

$$\phi_3(s,t) = \frac{1}{4}(1+s)(1+t), \tag{11.76}$$

$$\phi_4(s,t) = \frac{1}{4}(1-s)(1+t). \tag{11.77}$$

In matrix form we can also write this as

$$\begin{pmatrix} \phi_4 & \phi_3 \\ \phi_1 & \phi_2 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 1+t \\ 1-t \end{pmatrix} \begin{pmatrix} 1-s & 1+s \end{pmatrix}. \tag{11.78}$$

With $U_i = U(\underline{x}_i)$, $i = 1, 2, 3, 4$ the isoparametric property of the mapping and the form of U

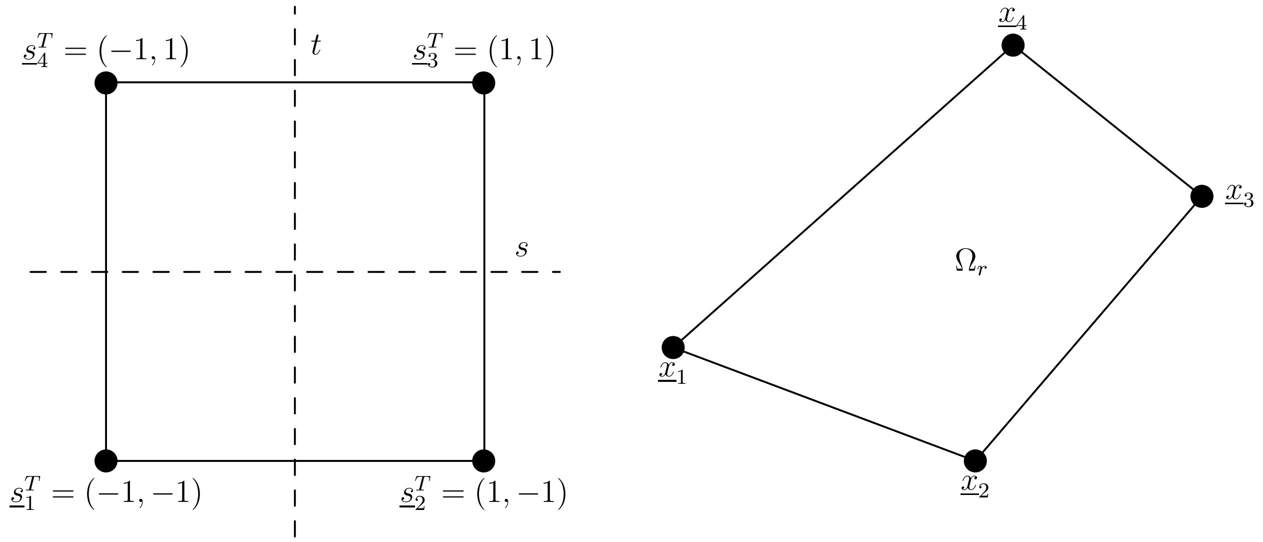


Figure 11.8: The standard bilinear element $-1 \le s, t \le 1$ and an actual element.

is that

$$\underline{x}(s,t) = \sum_1^4 \underline{x}_i \phi_i(s,t), \tag{11.79}$$

$$U(\underline{x}(s,t)) = \sum_1^4 U_i \phi_i(s,t). \tag{11.80}$$

On an element edge such as $t = -1$ we have $\phi_3(s,-1) = \phi_4(s,-1) = 0$ and both the map and the approximation are linear in s and we have

$$\underline{x}(s,-1) = \underline{x}_1 \left(\frac{1-s}{2}\right) + \underline{x}_2 \left(\frac{1+s}{2}\right) \tag{11.81}$$

$$U(\underline{x}(s,-1)) = U_1 \left(\frac{1-s}{2}\right) + U_2 \left(\frac{1+s}{2}\right). \tag{11.82}$$

On an element edge we hence have the same situation as with linear triangles and it is because of this that we can have a mesh with a 4-noded bilinear element next to a 3-noded linear triangle.

As a comment, the approximation $U(\underline{x}(s,t))$ is bilinear in the local variables s and t but it might not be polynomial in x and y. From (11.81), this is the case if the coefficient of st is non-zero in the expression for $\underline{x}(s,t)$ then we no longer have an affine mapping and the Jacobian matrix is no longer constant for $-1 \le s, t \le 1$. For the mapping to be an affine mapping we need

$$\underline{x}_1 - \underline{x}_2 + \underline{x}_3 - \underline{x}_4 = \underline{0} \tag{11.83}$$

which is the case when the actual quadrilateral is a rectangle or more generally a parallelogram.

We now consider what is involved to compute the element stiffness matrix. With the 4 basis functions the element stiffness matrix for Poisson's equation is of size 4×4 and it can be determined by the same procedure as used in the case of linear triangles. With $\tilde{\phi}_1$, $\tilde{\phi}_2$, $\tilde{\phi}_3$

and $\tilde\phi_4$ denoting the basis functions in terms of x and y satisfying

$$\tilde\phi_i(\underline{x}(s,t)) = \phi_i(s,t), \quad i = 1,2,3,4, \tag{11.84}$$

we have

$$\begin{pmatrix} \dfrac{\partial\phi_1}{\partial s} & \dfrac{\partial\phi_2}{\partial s} & \dfrac{\partial\phi_3}{\partial s} & \dfrac{\partial\phi_4}{\partial s} \\ \dfrac{\partial\phi_1}{\partial t} & \dfrac{\partial\phi_2}{\partial t} & \dfrac{\partial\phi_3}{\partial t} & \dfrac{\partial\phi_4}{\partial t} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial x}{\partial s} & \dfrac{\partial y}{\partial s} \\ \dfrac{\partial x}{\partial t} & \dfrac{\partial y}{\partial t} \end{pmatrix} \begin{pmatrix} \dfrac{\partial\tilde\phi_1}{\partial x} & \dfrac{\partial\tilde\phi_2}{\partial x} & \dfrac{\partial\tilde\phi_3}{\partial x} & \dfrac{\partial\tilde\phi_4}{\partial x} \\ \dfrac{\partial\tilde\phi_1}{\partial y} & \dfrac{\partial\tilde\phi_2}{\partial y} & \dfrac{\partial\tilde\phi_3}{\partial y} & \dfrac{\partial\tilde\phi_4}{\partial y} \end{pmatrix} \tag{11.85}$$

$$= \tilde J^T B \tag{11.86}$$

where B is the 2×4 matrix of all the partial derivatives of $\tilde\phi_1, \tilde\phi_2, \tilde\phi_3, \tilde\phi_4$. Hence

$$B = \tilde J^{-T} \begin{pmatrix} \dfrac{\partial\phi_1}{\partial s} & \dfrac{\partial\phi_2}{\partial s} & \dfrac{\partial\phi_3}{\partial s} & \dfrac{\partial\phi_4}{\partial s} \\ \dfrac{\partial\phi_1}{\partial t} & \dfrac{\partial\phi_2}{\partial t} & \dfrac{\partial\phi_3}{\partial t} & \dfrac{\partial\phi_4}{\partial t} \end{pmatrix} \tag{11.87}$$

$$= \frac{1}{4}\tilde J^{-T} \begin{pmatrix} t-1 & 1-t & t+1 & -(t+1) \\ s-1 & -(s+1) & s+1 & 1-s \end{pmatrix}. \tag{11.88}$$

The 4×4 element matrix is

$$K_r = (a(\tilde\phi_i, \tilde\phi_j)_r) = \iint_{\Omega_r} B^T B \,\mathrm{d}x\mathrm{d}y = \int_{-1}^1\int_{-1}^1 B^T B|J| \,\mathrm{d}s\mathrm{d}t. \tag{11.89}$$

If the Jacobian matrix $\tilde J$ is not a constant matrix on $-1 \le s, t \le 1$ then the components of B are rational functions of s and t and the integrals are at best very awkward to do exactly. Instead we approximate the matrix using a quadrature rule of the form

$$K_r \approx \sum_{i=1}^{nq} w_i (B^T B|J|)(s_i, t_i) \tag{11.90}$$

where (s_i, t_i), $i = 1, \cdots, nq$ are the quadrature points in $[-1,1]\times[-1,1]$ and w_i, $i = 1, \cdots, nq$ are the associated weights. We similarly approximate the element vector $\underline{b}_r$ using the quadrature

$$\underline{b}_r = \iint_{\Omega_r} f(\underline{x}) \begin{pmatrix} \tilde\phi_1(\underline{x}) \\ \tilde\phi_2(\underline{x}) \\ \tilde\phi_3(\underline{x}) \\ \tilde\phi_4(\underline{x}) \end{pmatrix} \mathrm{d}x\mathrm{d}y = \int_{-1}^1\int_{-1}^1 f(\underline{x}(s,t)) \begin{pmatrix} \phi_1(s,t) \\ \phi_2(s,t) \\ \phi_3(s,t) \\ \phi_4(s,t) \end{pmatrix} |J(s,t)| \,\mathrm{d}s\mathrm{d}t \tag{11.91}$$

$$\approx \sum_{i=1}^{nq} w_i f(\underline{x}(s_i,t_i)) \begin{pmatrix} \phi_1(s_i,t_i) \\ \phi_2(s_i,t_i) \\ \phi_3(s_i,t_i) \\ \phi_4(s_i,t_i) \end{pmatrix} |J(s_i,t_i)|. \tag{11.92}$$

A quadrature rule which is adequate here involves generalising the one-dimensional 2-point Gauss Legendre rule

$$\int_{-1}^1 g(s)\,\mathrm{d}s \approx g(-\frac{1}{\sqrt3}) + g(\frac{1}{\sqrt3}) \tag{11.93}$$

to the 4-point rule

$$\int_{-1}^{1}\int_{-1}^{1} g(s,t)\,\mathrm{d}s \approx g(-\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}}) + g(\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}}) + g(\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}}) + g(-\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}}). \qquad (11.94)$$

We can compute the element matrix K_r exactly when the geometry of the element Ω_r is such that (11.83) holds, i.e. if Ω_r is a parallelogram. To keep things as simple as possible for hand calculation we restrict to the case that Ω_r is a square and as the matrix that we get is invariant to uniform scalings and translations we may as well assume that Ω_r is $[-1,1]\times[-1,1]$ so that the mapping $(s,t)\to(x,y)$ is the identity mapping giving $\tilde{J}=I$. The B matrix is

$$B = \frac{1}{4}\begin{pmatrix} t-1 & 1-t & t+1 & -(t+1) \\ s-1 & -(s+1) & s+1 & 1-s \end{pmatrix}. \qquad (11.95)$$

Now B^TB is the 4×4 matrix involving integrations of the product of columns on B with other columns of B and this involves integrals of the following type.

$$\int_{-1}^{1}\int_{-1}^{1}(t-1)^2\,\mathrm{d}s\mathrm{d}t = \int_{-1}^{1}\mathrm{d}s\int_{-1}^{1}(t-1)^2\,\mathrm{d}t = 2\left[\frac{(t-1)^3}{3}\right]_{-1}^{1} = 2\left(\frac{8}{3}\right) = \frac{16}{3}. \qquad (11.96)$$

Similarly for all the other integrals of this type, i.e.

$$\int_{-1}^{1}\int_{-1}^{1}(t+1)^2\,\mathrm{d}s\mathrm{d}t = \int_{-1}^{1}\int_{-1}^{1}(s-1)^2\,\mathrm{d}s\mathrm{d}t = \int_{-1}^{1}\int_{-1}^{1}(s+1)^2\,\mathrm{d}s\mathrm{d}t = \frac{16}{3}. \qquad (11.97)$$

Hence

$$(K_r)_{ii} = \int_{-1}^{1}\int_{-1}^{1}(B^TB)_{ii}\,\mathrm{d}s\mathrm{d}t = \frac{1}{16}\left(\frac{16}{3}+\frac{16}{3}\right) = \frac{2}{3} \quad \text{for } i=1,2,3,4. \qquad (11.98)$$

As part of the calculation of the $2,1$ entry we need to consider an integral

$$\int_{-1}^{1}\int_{-1}^{1}(1-s^2)\,\mathrm{d}s\mathrm{d}t = 4 - \int_{-1}^{1}\mathrm{d}t\int_{-1}^{1}s^2\,\mathrm{d}s = 4 - 2\left(\frac{2}{3}\right) = \frac{8}{3}. \qquad (11.99)$$

This gives

$$(K_r)_{21} = \frac{1}{16}\left(-\frac{16}{3}+\frac{8}{3}\right) = -\frac{1}{6}. \qquad (11.100)$$

Similarly $(K_r)_{41} = (K_r)_{43} = -1/6$. This is sufficient to determine all of the element stiffness matrix K_r as the matrix is a symmetric matrix and the entries on each row must add to 0. We have

$$K_r = \frac{1}{6}\begin{pmatrix} 4 & -1 & -2 & -1 \\ -1 & 4 & -1 & -2 \\ -2 & -1 & 4 & -1 \\ -1 & -2 & -1 & 4 \end{pmatrix}. \qquad (11.101)$$

11.10 Higher order elements — some comments

In one dimension we discussed linear elements and then later discussed quadratic elements and cubic Hermite elements. The linear and the quadratic elements give an approximating function in $C[0,1]$ and give an approximating function which is completely described by the values $U(x_i)$ at the nodal points x_i. To achieve the approximating function in $C^1[0,1]$ in the cubic Hermite case requires that we have two parameters $U(x_i)$ and $U'(x_i)$ at each nodal point and involves 4 cubic basis functions on each element. Elements in two space dimensions which give a $C^1(\bar{\Omega})$ function are known and could be included in this module but they are a bit cumbersome. With quadrilateral shaped elements this is possible with an element which has 16 parameters involving U, $\partial U/\partial x$, $\partial U/\partial y$ and $\partial^2 U/\partial x \partial y$ at each of the 4 corners and there are triangular elements which give a $C^1(\bar{\Omega})$ function which involve even more parameters than this. We will not consider any of these any further in this module and we will restrict attention instead to popular higher order $C(\bar{\Omega})$ elements on triangles and on quadrilaterals.

Before the 6 noded quadratic triangle, the 9 noded quadrilateral and the 8 noded quadrilateral are described it is worth making some comments indicating what we require from an element and where the elements described here fit into the hierarchy of all possible elements. For the requirements we need an element which gives a function which is in $C(\bar{\Omega})$ and we need an element which has the capability of generating an approximation which is close to the exact solution. This last aspect is concerned with the element containing all the terms of as high as possible complete degree of polynomial in s and t. The terms that this involves for complete polynomials of degree 0, 1, 2 etc. is shown below in the form of a triangle.

$$
\begin{array}{lccccccccccc}
\text{For degree 0:} & & & & & & 1 & & & & & \\
\text{For degree 1:} & & & & & s & & t & & & & \\
\text{For degree 2:} & & & & s^2 & & st & & t^2 & & & \\
\text{For degree 3:} & & & s^3 & & s^2t & & st^2 & & t^3 & & \\
\text{For degree 4:} & & s^4 & & s^3t & & s^2t^2 & & st^3 & & t^4 & \\
\text{For degree 5:} & s^5 & & s^4t & & s^3t^2 & & s^2t^3 & & st^4 & & t^5
\end{array}
$$

For a complete polynomial of degree 1 we need 3 terms $(1, s, t)$ which is what we get with the linear triangle. For a complete polynomial of degree 2 we need 6 terms which is what we get with the 6 noded triangle. The 8 noded and the 9 noded quadrilateral elements also contain all the complete polynomial of degree 2 terms plus a few more times. As we will see the 9-noded element contains all the terms $s^i t^j$, $0 \le i, j \le 2$ whereas the 8 noded element contains all the terms for a complete polynomial of degree 2 with in addition $s^2 t$ and st^2. The 8 noded element is slightly more efficient than the 9 noded element in that it involves the least number of terms for the quadratic shaped element which contains all the terms of a complete polynomial of degree ≤ 2 as well as giving a $C(\bar{\Omega})$ function.

11.11 The 6-noded triangle

The standard 6 noded quadratic triangle is shown in Figure 11.9 together with two possible actual triangles.

Corresponding to how the linear and bilinear elements were described we now have 6 Lagrange basis functions and if $U_i = U(\underline{x}_i)$ then the form of the mapping from T to Ω_r and the form of the approximation are given by

$$\underline{x}(s,t) = \sum_{i=1}^{6} \underline{x}_i \phi_i(s,t), \tag{11.102}$$

$$U(\underline{x}(s,t)) = \sum_{i=1}^{6} U_i \phi_i(s,t). \tag{11.103}$$

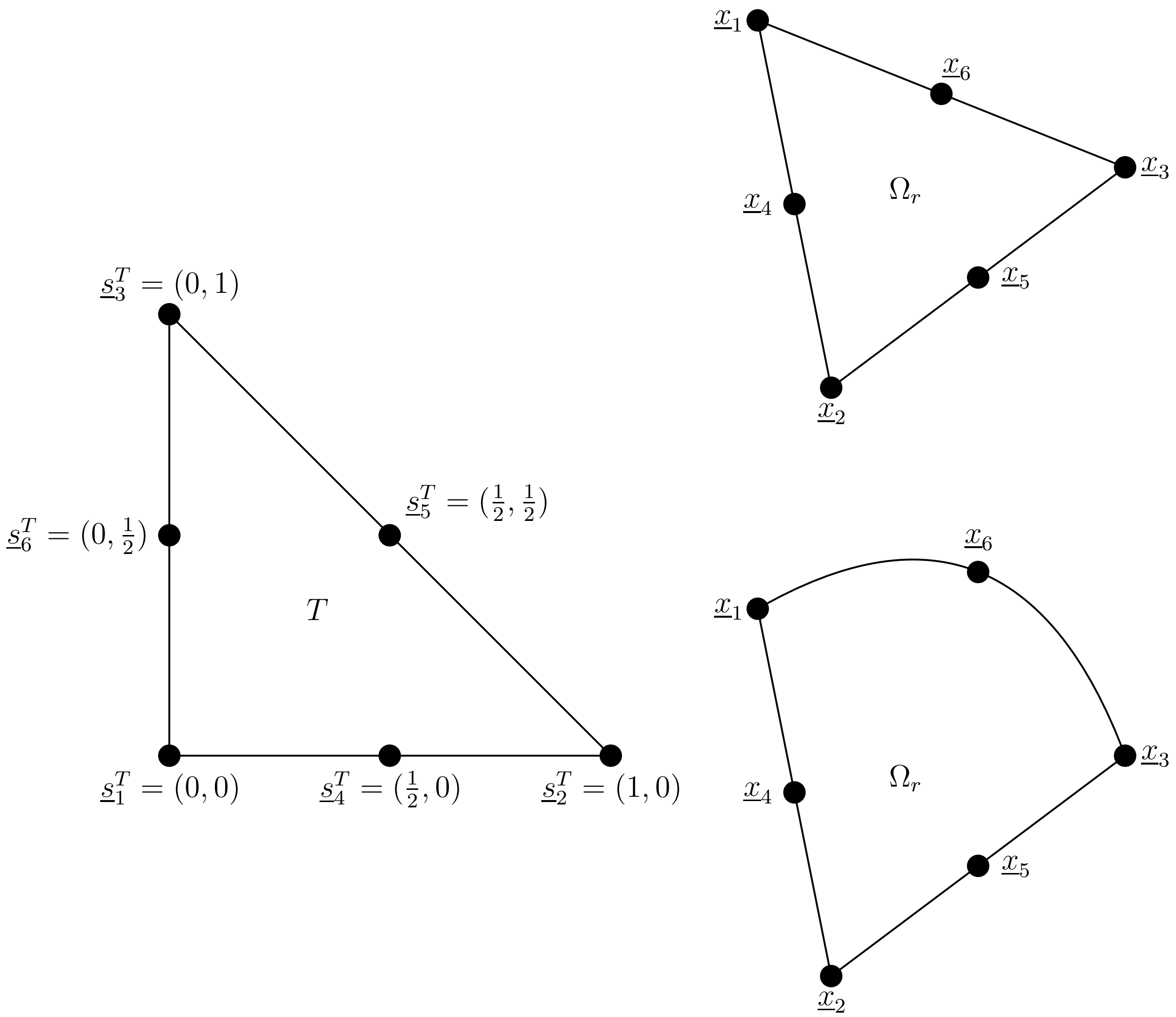


Figure 11.9: The standard 6 noded triangle and two possible actual triangles.

Depending on the positions of $\underline{x}_1, \cdots, \underline{x}_6$ the mapping can involve curve segments in the x, y plane. This is the case for the curved side $\underline{x}_3, \underline{x}_6, \underline{x}_1$ shown in one of the actual elements shown in Figure 11.9. The curved side shown in Figure 11.9 has a parametric description

$$\underline{x}(0, 1-t) = \underline{x}_3\phi_3(0, 1-t) + \underline{x}_6\phi_6(0, 1-t) + \underline{x}_1\phi_1(0, 1-t), \quad 0 < t < 1. \tag{11.104}$$

If the boundary $\partial\Omega$ of Ω is curved then we can exploit this feature of the 6 noded triangle to approximate such a curve reasonably accurately by a relatively small number of such elements instead of using a much larger number of smaller 3-noded linear triangles to obtain the same accuracy.

In the case that the mid-side points $\underline{x}_4$, $\underline{x}_5$ and $\underline{x}_6$ are positioned exactly at the mid-points of each of the sides then the mapping reduces to an affine transformation as in the case of linear triangles. This is what is usually done when a mesh of 6 noded triangles is usually generated for all the triangles in the interior of the region Ω. When $\underline{x}(s,t)$ has no quadratic term the Jacobian matrix $\tilde{J}$ is constant and the approximation U is quadratic in x and y.

As we have just stated, if the mid-points are located exactly in the middle of the sides then the mapping $\underline{x}(s,t)$ has a constant Jacobian matrix $\tilde{J}$ which will be non-singular provided

$J = \det(\tilde{J}) \neq 0$. When this is the case we are guaranteed that the mapping $\underline{x}(s,t)$ is one-to-one and onto on T and it is hence invertible on T. However, when the mid-side nodes are moved away from this middle position the matrix $\tilde{J}$ is no longer constant on T and it may be necessary to check that the mapping is still invertible throughout T which may be awkward to do in practice. This would have to be done on a case by case basis.

We now come to the Lagrange basis functions which are relatively straightforward to derive. What we need are functions which are in span$\{1, s, t, s^2, st, t^2\}$, which are zero at 5 of the points and take the value 1 at the other point. We can arrange for the function to be zero at 5 different points by taking a function which is zero on a line passing through 3 of the points and which is also zero on another line passing through the other 2 points. This determines the function up to a scaling factor which we determine from the condition $\phi_i(\underline{s}_i) = 1$. In the case of ϕ_1 this leads to

$$\phi_1(s,t) = (1-s-t)(1-2s-2t). \tag{11.105}$$

Similar reasoning applies to $\phi_2, \cdots, \phi_6$ and the complete set of basis functions is

$$\begin{aligned}
\phi_1(s,t) &= (1-s-t)(1-2s-2t). && (11.106)\\
\phi_2(s,t) &= s(2s-1), && (11.107)\\
\phi_3(s,t) &= t(2t-1), && (11.108)\\
\phi_4(s,t) &= 4s(1-s-t), && (11.109)\\
\phi_5(s,t) &= 4st, && (11.110)\\
\phi_6(s,t) &= 4t(1-s-t). && (11.111)
\end{aligned}$$

For Poisson's equation the element matrices with this 6-noded triangle are 6×6 and the element vector is 6×1. These are approximated using numerical quadrature.

11.12 The 9-noded quadrilateral

On the standard element $[-1,1] \times [-1,1]$ this involves the 9 points shown in Figure 11.10. As in the case of the 4-noded bilinear quadrilateral the basis functions are each products of a one-dimensional basis function in s with a one-dimensional basis function in t. For the bilinears we had products of linears whereas now we have products of quadratics. The functions $\phi_1, \cdots, \phi_9$ can be compactly written in the matrix form

$$\begin{pmatrix} \phi_4 & \phi_7 & \phi_3 \\ \phi_8 & \phi_9 & \phi_6 \\ \phi_1 & \phi_5 & \phi_2 \end{pmatrix} = \begin{pmatrix} \dfrac{t(t+1)}{2} \\ 1-t^2 \\ \dfrac{t(1-t)}{2} \end{pmatrix} \begin{pmatrix} \dfrac{s(s+1)}{2} & 1-s^2 & \dfrac{s(1-s)}{2} \end{pmatrix}. \tag{11.112}$$

With these basis functions the mapping and the approximation are given by

$$\begin{aligned}
\underline{x}(s,t) &= \sum_{i=1}^{9} \underline{x}_i \phi_i(s,t), && (11.113)\\
U(\underline{x}(s,t)) &= \sum_{i=1}^{9} U_i \phi_i(s,t) && (11.114)
\end{aligned}$$

where $U_i = U(\underline{x}_i)$. As with the 6-noded triangles the shape of the actual element is the image of this standard element under the mapping (11.113) and we only get straight edges if the mid-side points are exactly in the middle positions. Also, on an edge the mapping and the approximation are both quadratics in a single variable as is the case with the 6-noded triangle and as such we can have 6-noded triangles next to 9-noded quadrilaterals in a mesh of Ω.

11.13 The 8-noded serendipity quadrilateral

On the standard element $[-1,1]\times[-1,1]$ this involves the 8 nodes $\underline{s}_1,\cdots,\underline{s}_8$ shown in Figure 11.11 which are all on the edge of the element. To derive the 8 basis functions which must only contain terms in 1, s, t, s^2, st, t^2, s^2t and st^2 but not s^2t^2 is slightly harder than for the 9-noded element just described. To ensure that we have no term in s^2t^2 we consider functions which are the product of 3 linear factors and the 3 lines that this corresponds to contains the 7 points at which a particular basis function needs to take the value 0. For ϕ_1 we take the 3 lines as $s=1$, $t=1$ and $t=-s-1$. For ϕ_5 the lines are $s=\pm1$ and $t=1$. By considering all the cases we obtain the following functions.

$$\phi_1(s,t) = \frac{1}{4}(1-s)(1-t)(-1-s-t), \tag{11.115}$$
$$\phi_2(s,t) = \frac{1}{4}(1+s)(1-t)(-1+s-t), \tag{11.116}$$
$$\phi_3(s,t) = \frac{1}{4}(1+s)(1+t)(-1+s+t), \tag{11.117}$$
$$\phi_4(s,t) = \frac{1}{4}(1-s)(1+t)(-1-s+t), \tag{11.118}$$
$$\phi_5(s,t) = \frac{1}{2}(1-s^2)(1-t), \tag{11.119}$$
$$\phi_6(s,t) = \frac{1}{2}(1-t^2)(1+s), \tag{11.120}$$
$$\phi_7(s,t) = \frac{1}{2}(1-s^2)(1+t), \tag{11.121}$$
$$\phi_8(s,t) = \frac{1}{2}(1-t^2)(1-s). \tag{11.122}$$

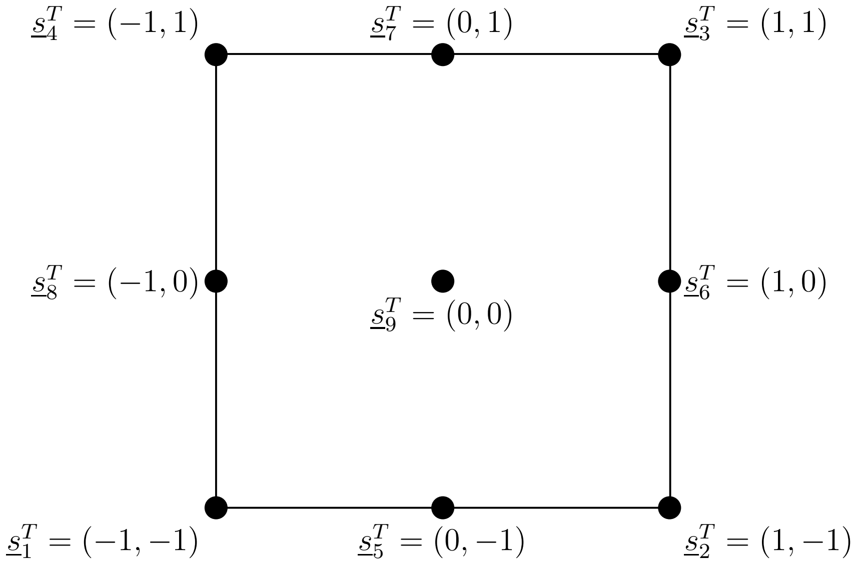


Figure 11.10: The standard 9-noded quadrilateral.

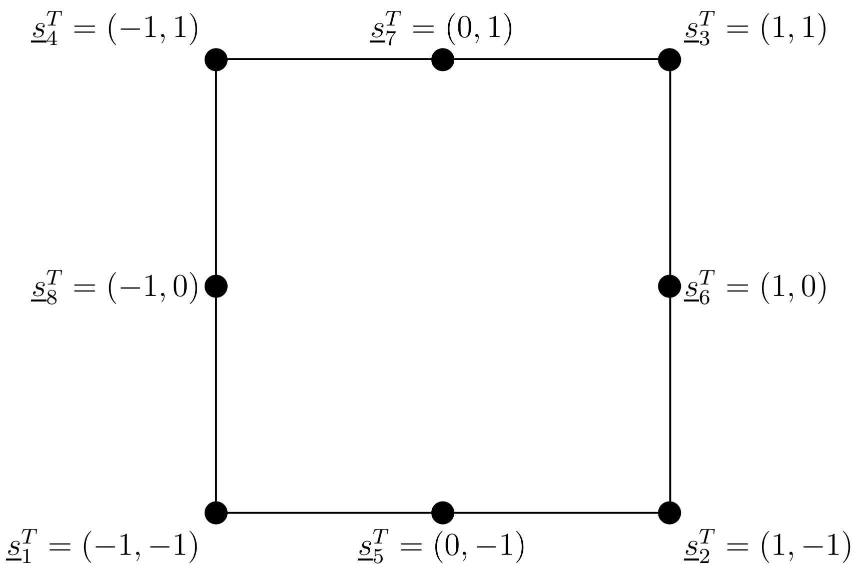


Figure 11.11: The standard 8-noded quadrilateral.

11.14 Summary

1. A mesh is a partition of the domain which mathematically means that

$$\bar{\Omega} = \bigcup_{i=1}^{ne} \bar{\Omega}_i, \quad \Omega_i \cap \Omega_j = \text{empty set for } i \neq j. \tag{11.123}$$

   This involves a collection of nodal points $\underline{\hat{x}}_1, \cdots, \underline{\hat{x}}_M$ and a connectivity matrix which tells us which nodes are on which element.

2. If $V_h = \text{span}\{\hat{\phi}_1, \cdots, \hat{\phi}_{M_1}\} \subset V$ is the finite element space then the Galerkin finite element equations are

$$K\underline{c} = \underline{b}, \quad K = (a(\hat{\phi}_i, \hat{\phi}_j)), \quad \underline{b} = ((f, \hat{\phi}_i)). \tag{11.124}$$

3. With

$$a(u, v)_r = \iint_{\Omega_r} \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y, \quad (f, v)_r = \iint_{\Omega_r} fv \, \mathrm{d}x\mathrm{d}y \tag{11.125}$$

   the rth element matrix K_r and the rth element vector $\underline{b}_r$ are

$$K_r = (a(\tilde{\phi}_i, \tilde{\phi}_j)_r) \quad \text{and} \quad \underline{b}_r = ((f, \tilde{\phi}_i)_r) \tag{11.126}$$

   where $\tilde{\phi}_1, \cdots, \tilde{\phi}_m$ are the Lagrange basis functions on Ω_r.

4. With isoparametric elements we have a standard element T, a mapping $\underline{x} : T \to \Omega_r$ and and approximating function $U(\underline{x})$ of the form

$$\underline{x}(s, t) = \sum_{i=1}^{m} \underline{x}_i \phi(s, t), \tag{11.127}$$

$$U(\underline{x}(s, t)) = \sum_{i=1}^{m} U_i \phi(s, t). \tag{11.128}$$

   For linears $m = 3$, for bilinears $m = 4$, for the 6-noded triangle $m = 6$, for the 8-noded quadrilateral $m = 8$ and for the 9-noded quadrilateral $m = 9$. In the case $m = 3$ the mapping is an affine transformation and

$$\underline{x}(s, t) = \underline{x}_1 + \tilde{J} \begin{pmatrix} s \\ t \end{pmatrix} \quad \text{where } \tilde{J}\text{=Jacobian matrix.} \tag{11.129}$$

5. For the linear triangle $\phi_1 = 1 - s - t$, $\phi_2 = s$, $\phi_3 = t$ and

$$K_r = \frac{1}{2}|J|B^T B, \quad B = \begin{pmatrix} \dfrac{\partial\tilde{\phi}_1}{\partial x} & \dfrac{\partial\tilde{\phi}_2}{\partial x} & \dfrac{\partial\tilde{\phi}_3}{\partial x} \\ \dfrac{\partial\tilde{\phi}_1}{\partial y} & \dfrac{\partial\tilde{\phi}_2}{\partial y} & \dfrac{\partial\tilde{\phi}_3}{\partial y} \end{pmatrix} = \tilde{J}^{-T} \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \tag{11.130}$$

   where here $J = \det(\tilde{J})$. K_r is invariant to translation, uniform scalings, rotations and reflections.

6. Most integrals have to be approximated. The simplest approximation for triangles is the one point rule

$$\iint_{\Omega_r} g(x,y)\,\mathrm{d}x\mathrm{d}y \approx (\text{area of } \Omega_r) g(x_c, y_c), \quad \text{where } (x_c, y_c) = \text{centroid of } \Omega_r. \tag{11.131}$$

7. The assembly operation corresponds to performing the calculation

$$a(\hat\phi_i, \hat\phi_j) = \sum_{r=1}^{ne} a(\hat\phi_i, \hat\phi_j)_r, \quad (f, \hat\phi_i) = \sum_{r=1}^{ne} (f, \hat\phi_i)_r. \tag{11.132}$$

8. K and $\underline{b}$ in the system $K\underline{c} = \underline{b}$ are obtained by extracting from $\hat K = (a(\hat\phi_i, \hat\phi_j))$ and $\hat{\underline{b}} = ((f, \hat\phi_i))$, $1 \le i, j \le M$ all the indices i at which U is unknown at $\hat{\underline{x}}_i$.

9. On $[-1,1] \times [-1,1]$ the 4 bilinear functions are

$$\begin{pmatrix} \phi_4 & \phi_3 \\ \phi_1 & \phi_2 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 1+t \\ 1-t \end{pmatrix} \begin{pmatrix} 1-s & 1+s \end{pmatrix}. \tag{11.133}$$

10. On the standard triangle for the 6-noded triangular element the basis functions are

$$\phi_1(s,t) = (1-s-t)(1-2s-2t), \tag{11.134}$$
$$\phi_2(s,t) = s(2s-1), \tag{11.135}$$
$$\phi_3(s,t) = t(2t-1), \tag{11.136}$$
$$\phi_4(s,t) = 4s(1-s-t), \tag{11.137}$$
$$\phi_5(s,t) = 4st, \tag{11.138}$$
$$\phi_6(s,t) = 4t(1-s-t). \tag{11.139}$$

11. On $[-1,1] \times [-1,1]$ the 9 quadratic basis functions for the 9-noded element are

$$\begin{pmatrix} \phi_4 & \phi_7 & \phi_3 \\ \phi_8 & \phi_9 & \phi_6 \\ \phi_1 & \phi_5 & \phi_2 \end{pmatrix} = \begin{pmatrix} \frac{t(t+1)}{2} \\ 1-t^2 \\ \frac{t(t-1)}{2} \end{pmatrix} \begin{pmatrix} \frac{s(s+1)}{2} & 1-s^2 & \frac{s(s-1)}{2} \end{pmatrix}. \tag{11.140}$$

12. On $[-1,1] \times [-1,1]$ the 8-noded element has all of its nodes on the edge of the element. The 8 quadratic basis functions are appropriate linear combinations of 1, s, t, s^2, st, t^2, s^2t and st^2 and can each be written as a product of 3 linear factors.