TMA4212 Part 2: Introduction to finite element methods

Charles Curry

April 26, 2018

Contents

1 Introduction

In the first part of the course, we covered numerical solution of partial differential equations by finite difference methods. By introducing a grid of points and discretizing the equation, time-independent linear PDEs were reduced to linear systems

$$Au = F,$$

whilst time-dependent linear equations (evolution problems) gave rise to either ODEs (by semi-discretization) or linear evolution equations:

$$u' = Au + F \quad \text{or} \quad Au_{n+1} = B_n u_n + C_n$$

It is now time to introduce another important approach to the numerical solution of PDEs, namely the finite element methods. This technique is widely applied in industry, principally it because handles irregular domains with awkward boundaries much more comfortably than difference methods. Moreover, finite element methods are a particular type of Galerkin method, and thus provide an excellent introduction to a still broader vista of applied numerical analysis.

The verbal description of Galerkin methods is that the solution space is discretized, rather than the equation. Indeed, the idea is to specify a finite dimensional function space (such as piecewise linear functions, trigonometric polynomials up to a given degree, a finite set of orthogonal polynomials etc.), and look for solutions in this space. The key is that 'solution' in this sense is interpreted in some averaged or weak sense. This is made possible by the vast theory of weak formulations of PDEs. Constructing a weak form of an equation may be important in cases where the solution becomes insufficiently regular for the equation to make sense, as is seen in the development of a discontinuity (shock) in the following inviscid Burgers' equation.

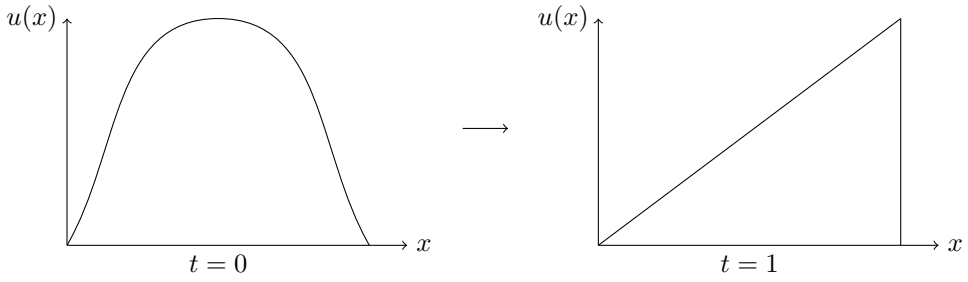

Figure 1: Shock development

This situation arises physically also, we need only think of sonic booms as aircraft break the sound barrier, or even the formation of traffic jams (recall that traffic flow can be modelled by PDEs). In this case it might be argued that the weak form is a better description of the physical system than the PDE which was derived first. On the other hand, in situations where there are no problems with

regularity of solutions (as often occurs with elliptic or parabolic problems), weak formulations may still have computational value. The Galerkin (and related) methods are among the most spectacular realizations of this phenomenon.

The heart of the material covered in this course lies in §2 and §5. In §2 we introduce the finite element method for second order ODE boundary value problems. This is extended to PDEs with two independent variables in §5, where the focus will be particularly on the Poisson equation (and related elliptic problems).

A difficulty that arises when first encountering finite element methods is that the underlying theory, grounded in functional analysis of PDEs, can be far more intimidating than the application and understanding of the methods themselves. An introduction to this side is found in §3-4 - we give in §3 the briefest treatment of the necessary results from functional analysis, together with the main theorem on existence and uniqueness of solutions to weak formulations of PDEs. This is followed in §4 by a demonstration of how stability and convergence is obtained for the finite element method in the ODE case, although the techniques and results of this section are much more widely applicable. These chapters should perhaps be read as appendices to §2 and §5, providing justification for the methods and a reference to some required concepts.

These notes conclude in §6 with a short discussion of the wider world of finite element and Galerkin methods. This is not curriculum material, but hopefully provides a bit of perspective, and perhaps an encouragement to learn more about the field in the TMA4220 finite element course!

2 Boundary value problems and the Galerkin method

We begin by consider the equation

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

At first we will assume homogeneous Dirichlet conditions $u(0) = u(1) = 0$, before examining other cases later in the same chapter. When applied to a linear PDE, the Galerkin method involves four stages:

1. *Setting up a weak (variational) form of the PDE*
2. *Choosing a finite dimensional function space where we will look for solutions, together with a basis.*
   For finite element methods, this involves specifying a discretization of the domain of the equation, which in the 1d (ODE) case is nothing other than a set of grid points in $[0, 1]$.
3. *'Assembling' a linear system*
   By expanding in terms of the basis, the (weak) PDE is reduced to a linear system $Au = b$ (or a linear ODE such as $M\dot{u}(t) = Au(t) + b$ in the case of evolution problems)
4. *Solution of the system*
   Whilst an important and nontrivial step in practice, the considerations are the same as those arising from finite difference methods so we will not comment further.

2.1 Setting up a variational form

Let v be an arbitrary element of a function space V (usually called a test function) - we will postpone discussion of the specific space V. We multiply (1) by v and integrate over the domain:

$$-\int_0^1 u''(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

Integration by parts results in the relation

$$\int_0^1 u'(x)v'(x)dx - \left[u'(x)v(x)\right]_0^1 = \int_0^1 f(x)v(x)dx \tag{2}$$

A characteristic of Galerkin methods is that the solutions u and test functions v are supposed to belong to the same function space V. In particular, we can assume that all elements of V satisfy the Dirichlet boundary conditions. In particular, $v(0) = v(1) = 0$, and the second term on the left hand side of the above equation vanishes. We are led to the following weak form of the PDE:

Find $u \in V$ such that for all $v \in V$,

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx \tag{3}$$

In this case the space V will be the Sobolev space $H_0^1([0,1])$, which consists of all functions on $[0,1]$ which satisfy the boundary conditions and are of sufficient regularity for all the terms in the above equation to make sense. This space is discussed in more detail in §3.1.

How closely are the equations (1) and (3) related? The above argument shows that any solution of (1) also solves (3). In fact, in this case it is possible to prove the converse, namely that any solution of (3) is sufficiently regular for the original ODE to make sense, and that the solution solves (1). However, as discussed in the introduction, we cannot expect such a result to hold for all PDEs, and indeed it may not always be relevant. Such questions are a major part of PDE theory and are treated at length in standard PDE texts, but will be avoided in this course.

2.2 Grids and basis functions

The fundamental idea of Galerkin methods is to restrict the functions u, v in the weak formulation (3) to lie in a finite dimensional function space $V_h \subset V$. If in addition we give a basis $\{\varphi_i\}$ for V_h, expanding u and v in terms of the basis will lead to a linear system of equations (see §2.3).

Finite element methods are a specific type of Galerkin method where the function spaces V_0 are piecewise (or elementwise in higher dimensions) polynomials. Indeed, the domain Ω of the equation, here $[0,1]$ is partitioned into a set of elements K_i, here subintervals $[x_i, x_{i+1}]$, where $0 = x_0 < x_1 < \ldots < x_{M+1} = 1$ is a grid of points imposed on $[0,1]$ such as we have used in setting up finite difference methods. Having specified this set of elements, we let X_h^p be the space of functions that are polynomials of degree p when restricted to an element. Of special interest to us is the space X_h^1 of piecewise linear functions. There is a

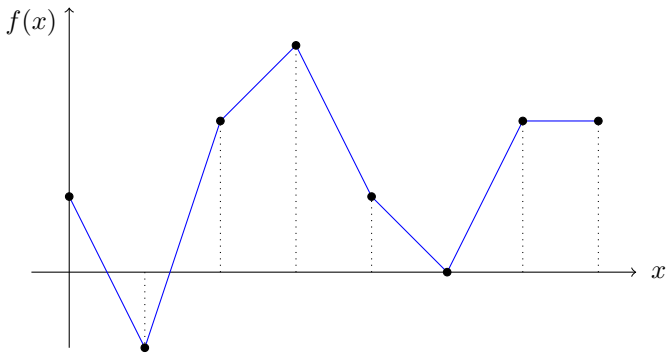

Figure 2: A representative element $f(x)$ of X_h^1

natural basis $\{\varphi_i\}_{i=0,\dots,M+1}$ for X_h^1, the nodal (or Lagrange) basis, where the basis functions $\varphi_i(x)$ are the unique elements of X_h^1 that evaluate to 1 at the point x_i, and 0 at all other grid points x_j. This basis is natural in the following

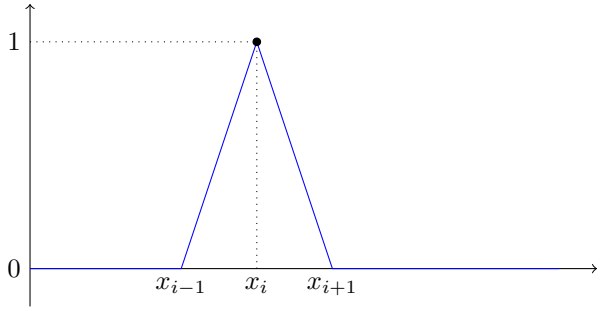

Figure 3: The basis function $\varphi_i(x)$

sense. As X_h^1 is a finite dimensional vector space, any element $v \in X_h^1$ can be expanded in terms of the basis, i.e. $v = \sum_{i=0}^{M+1} v_i \varphi_i(x)$, where v_i are uniquely determined coefficients. For this choice of basis, it follows that $v(x_i) = v_i$ for all $i = 0, \dots, M+1$, i.e. the coefficients in the expansion are simply the values of the function at each grid point (or node, hence nodal basis). This is convenient, as we will typically give the solution u in terms of its vector of components u_i, and hence the solution can be plotted easily.

The explicit formula for φ_i can be calculated:

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & x_{i-1} < x \leq x_i \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & x_i < x < x_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

For most computations involving the basis functions, we will restrict ourselves to a specified element $K_i = [x_i, x_{i+1}]$. In this case, we observe that all basis functions φ_j are identically zero on K_i except φ_i and φ_{i+1}:

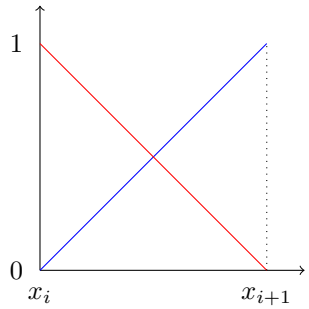

Figure 4: Basis functions φ_i, φ_{i+1} restricted to K_i

2.3 Assembly: setting up a linear system

We now return to the weak form (3), replacing the space V with V_h. As $u, v \in V_h$, they may be expanded in terms of the basis. Indeed, writing $u(x) = \sum_i u_i \varphi_i(x), v(x) = \sum_j v_j \varphi_j(x)$, the problem becomes:

Find a vector of coefficients u for which, for all vectors v

$$\sum_{i,j} u_i v_j \int_0^1 \varphi_i'(x)\varphi_j'(x)dx = v_j \int_0^1 f(x)\varphi_j(x)dx \tag{4}$$

If we define the 'stiffness matrix' A and 'load vector' F by

$$A_{ij} = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx, \quad F_j = \int_0^1 f(x)\varphi_j(x)dx, \tag{5}$$

equation (4) takes the form

$$v^T A u = v^T F.$$

As this must hold for all v, it is equivalent to the equation

$$Au = F.$$

In practice, a large part of the challenge of coding finite element solvers is the construction of the matrix A and vector F following (5). An important principle (especially for PDEs with spatial dimension 2 or greater) is to perform the construction *elementwise*, i.e. we compute

$$A_{ij} = \sum_k \int_{x_k}^{x_{k+1}} \varphi_i'(x)\varphi_j'(x)dx, \quad F_j = \sum_k \int_{x_k}^{x_{k+1}} f(x)\varphi_j(x)dx,$$

In view of the comments on restricting basis functions φ_j to a given element K_k (see figure 4), the only non-zero contributions to A_{ij} and F_j from K_k arise where $i, j = k$ or $k+1$. Let us begin with A, where from the explicit form of φ_i we find that (on K_k)

$$\varphi_k'(x) = -\frac{1}{x_{k+1} - x_k}, \quad \varphi_{k+1}'(x) = \frac{1}{x_{k+1} - x_k}$$

It follows that we can construct A by first initializing A to the matrix with all entries zero, and then adding the 2×2 matrix

$$\frac{1}{x_{k+1} - x_k} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

to the submatrix $[A_{ij}]_{i,j=k,k+1}$. For example, if the grid points are equidistant

so that $x_{k+1} - x_k = h$ for all k, the assembly process gives

$$A = \frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & 0 \\ 0 & -1 & 2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix} \tag{6}$$

This is not quite the correct matrix for the system, however, as we have neglected the boundary conditions. Indeed, the space X_h^1 is not a subspace of $H_0^1([0,1])$ as functions v in the latter space obey $v(0) = v(1) = 0$, which has not been imposed in the former space. Computationally, it is usually easier to assemble the matrix and then enforce the boundary conditions. Here, this is a simple matter, as any element $v \in X_h^1$ will satisfy the boundary conditions if and only if $v_0 = v_{M+1} = 0$. This constitutes a vector subspace of X_h^1 with the basis $\{\varphi_i\}_{i=1,\ldots,M}$. It follows that we obtain the correct matrix by removing the entries of A_{ij} where i, j are 0 or $M+1$, i.e. the first and last row and column, obtaining the matrix

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \tag{7}$$

The computation of F is more complicated, as the integrals involve not only the basis functions but also the forcing term $f(x)$. Indeed, the contribution from the element K_k will be the vector

$$\frac{1}{x_{k+1} - x_k} \begin{pmatrix} \int_{x_k}^{x_{k+1}} (x_{k+1} - x) f(x) dx \\ \int_{x_k}^{x_{k+1}} (x - x_k) f(x) dx, \end{pmatrix} \tag{8}$$

which should be added to the subvector $[F_j]_{j=k,k+1}$ in a similar manner as to the construction of A. Unless $f(x)$ takes a simple form, these integrals must be computed numerically, using a quadrature rule. Return to the case of a uniform grid, and suppose we evaluate these using the trapezoidal rule

$$\int_a^b g(x) dx \approx (b-a) \frac{g(a) + g(b)}{2}$$

The vector (8) becomes simply

$$\frac{1}{2} \begin{pmatrix} f(x_k) \\ f(x_{k+1}), \end{pmatrix}$$

hence before implementing boundary conditions we find

$$F = \left(\frac{1}{2}f(x_0), f(x_1), \ldots, f(x_M), \frac{1}{2}f(x_{M+1})\right)^T. \tag{9}$$

To implement the boundary conditions, we delete the entries corresponding to φ_0 and φ_{M+1}, giving

$$F = \big(f(x_1), \ldots, f(x_M)\big)^T$$

Thus solving (1) numerically by linear finite elements on a uniform grid with trapezoidal quadrature is equivalent (gives the same same linear system) to solving by three-point central difference approximation of the second derivative. It would however be more normal to use Gaussian quadrature on the integrals (8), in general

$$\int_a^b g(x)dx \approx \sum_i w_i g(x_i),$$

where w_i are the quadrature weights and x_i the quadrature nodes, particular values of which can be found in any textbook on numerical integration (or the wikipedia articles)

For time independent problems it is often the case that for regular geometries finite element methods can be shown to be equivalent to finite difference schemes. This is typically no longer the case for semidiscretization of evolution equations (see §6), but the main reason for employing finite element methods in practice is their ability to handle domains with irregular boundaries, as often arise in physical problems.

2.4 Neumann conditions

In contrast to finite difference methods, the implementation of Neumann conditions does not involve a substantial challenge. Let us consider equation (1) with the Neumann condition $u'(0) = a, u'(1) = b$. Multiplying by a test function v and integrating by parts gives the same equation

$$\int_0^1 u'(x)v'(x)dx - \big[u'(x)v(x)\big]_0^1 = \int_0^1 f(x)v(x)dx,$$

however we no longer enforce $v(0) = v(1) = 0$, indeed $v \in H^1([0,1])$ but not necessarily H_0^1. The second term on the left hand side no longer vanishes, but may be expressed in terms of the boundary conditions. The result is the following weak form: Find $u \in H^1([0,1])$ such that for all $v \in H^1([0,1])$,

$$\int_0^1 u'(x)v'(x)dx = bv(1) - av(0) + \int_0^1 f(x)v(x)dx \tag{10}$$

From here we proceed in the same manner as before - we first specify a finite dimensional subspace $V_h \subset V$ together with a basis $\{\varphi_i\}$, and arguing as per §2.3

reduces the problem to a linear system. Taking linear finite elements $V_h = X_h^1$ with the nodal (Lagrange) basis as before gives the equation

$$Au = (-a, 0, \ldots, b)^T + F,$$

where A and F are the same stiffness matrix and load vector (5), only this time we do not remove any rows or columns. For instance, taking a uniform grid results in A taking the form (6).

Note that if we evaluate the integrals in F by the trapezoidal rule as per (9), we obtain the same system as arises from the order 2 finite difference discretization that uses fictitious nodes x_{-1}, x_{M+2}, see case two from chapter 3 of the course notes on finite differences.

It is also possible to have mixed boundary conditions, e.g. $u'(0) = a, u(1) = 0$, in this case we multiply (1) by a function v and integrate by parts. The function space V comprises those elements of $H^1([0,1])$ for which $v(1) = 0$, and we proceed as before. The end result is a system

$$Au = (-a, 0, \ldots, 0)^T + F,$$

where we take the same A, F and remove the last row and column of A and final entry in F (all those corresponding to φ_{M+1}), but not the first (corresponding to φ_0).

2.5 Inhomogeneous Dirichlet conditions

Suppose we consider the problem (1) with the inhomogeneous Dirichlet conditions $u(0) = a, u(1) = b$. This requires a little care, because the space of functions $v \in H^1([0,1])$ which satisfy these conditions do not form a vector space - the sum of two such functions $v_1 + v_2$ will no longer satisfy the conditions. The usual way around this problem in the Galerkin framework is to relate the solution of the inhomogeneous and homogeneous problems. To illustrate the idea, suppose $\hat{u}$ solves (1) with homogeneous boundary conditions $\hat{u}(0) = \hat{u}(1) = 0$, and let

$$u(x) = \hat{u}(x) + R(x), \quad R(x) = a(1-x) + bx.$$

Then as $R''(x) = 0$, we see that u solves the same problem with inhomogeneous conditions $u(0) = a, u(1) = b$. We call the function R a *lifting* of the boundary data. Let us see how this idea translates to the weak formulation of the problem. Once again, we multiply (1) by a test function v, here taken from the space $H_0^1([0,1])$ - as we will be relating to the homogeneous problem we take test functions obeying $v(0) = v(1) = 0$, giving the equation

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx,$$

where here u is an element of $H^1([0,1])$ with $u(0) = a, u(1) = b$. Now suppose that $R_h(x)$ is an element of $H^1([0,1])$ satisfying the conditions $R_h(0) =$

$a, R_h(1) = b$, and define $\hat{u} := u - R_h$. It follows that $\hat{u} \in H_0^1([0,1])$, and

$$\int_0^1 \hat{u}'(x)v'(x)dx = \int_0^1 f(x)v(x)dx + \int_0^1 R_h'(x)v'(x)dx$$

where here $\hat{u}, v \in H_0^1([0,1])$. We can then solve this problem numerically in the familiar manner by introducing a finite dimensional subspace V_h of $H_0^1([0,1])$ together with a basis, and expanding in terms of the basis. The calculation of the right hand side is most convenient if $R_h(x) \in V_h$, and indeed in the case of linear finite elements $V_h = X_h^1$ with nodal basis $\{\varphi_i\}$ there is a natural choice:

$$R_h(x) = a\varphi_0(x) + b\varphi_{M+1}(x)$$

In this case the integral involving R_h' becomes

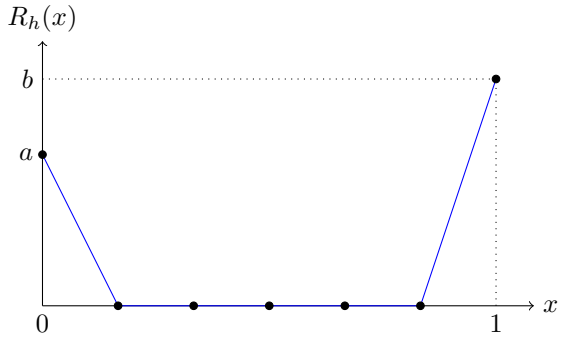

Figure 5: A canonical lifting R_h of the boundary conditions $u(0) = a$,$u(1) = b$

$$a\int_0^1 \varphi_0'(x)v'(x)dx + b\int_0^1 \varphi_{M+1}'(x)v'(x)dx,$$

which assumes a convenient form once we expand $v = \sum_j v_j\varphi_j(x)$:

$$\int_0^1 R_h'(x)v'(x)dx = A(a, 0, \ldots, 0, b)^T,$$

where A is the stiffness matrix from (5), *before boundary conditions are implemented*, e.g. in the case of a uniform grid it is (6) and not (7). The procedure is as follows: we set up the system

$$A\hat{u} = F + A(a, 0, \ldots, 0, b)^T,$$

computing the vector on the right hand side by matrix multiplication (it may be possible to code this more efficiently than a simple multiplication due to the large number of zeros in the vector to be multiplied), Only then do we implement the boundary conditions by removing the first row and column from A on the

left hand side and the first and last entries in the vector $F + A(a, 0, \ldots, 0, b)^T$. Finally, we take

$$u(x) = \hat{u}(x) + R_h(x),$$

which in vector terms consists of expanding the vector $\hat{u}$ from dimension M to $M + 2$ by including an additional a as the first entry, and b as the last entry.

2.6 Further examples

So far we have restricted ourselves to the simple equation $u''(x) = f(x)$, but similar reasoning is applicable to more general second order ODEs, for example the Sturm-Liouville type equation

$$-\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u = f(x), \quad a < x < b,$$

with appropriate boundary conditions (Neumann, Dirichlet, mixed etc). In this case, multiplying by a test function and integrating by parts results in the equation

$$\int_a^b p(x)u'(x)v'(x)dx + \int_a^b q(x)u(x)v(x)dx = -\left[p(x)u'(x)v(x)\right]_a^b + \int_a^b f(x)v(x)dx,$$

where the first term on the right hand side vanishes for the Dirichlet problem. Again, the problem of finding $u \in H^1((a, b))$ such that the above holds for all $v \in H^1((a, b))$ (or H_0^1 for the Dirichlet problem) is the appropriate weak form.

Introducing the subspace X_h^1 with its nodal basis then reduces the problem to the linear system

$$(A + M)u = F,$$

where

$$A_{ij} = \int_a^b p(x)\varphi_i'(x)\varphi_j'(x)dx, \quad M_{ij} = \int_a^b q(x)\varphi_i(x)\varphi_j(x)dx,$$

and the vector F is given by

$$F = \big(p(a)v(a)\sigma_a, 0, \ldots, 0, -p(b)v(b)\sigma_b\big) + \int_a^b \varphi(x)f(x)dx,$$

where $\varphi = (\varphi_0, \ldots, \varphi_{M+1})^T$, and the term in parentheses occurs only for the Neumann problem $u'(a) = \sigma_a, u'(b) = \sigma_b$. In the particular case $q(x) = 1$, the matrix M is called the *mass matrix*.

The assembly procedure is similar, except it is unlikely that the matrix A can be evaluated explicitly due to the presence of the weight $p(x)$. We perform the calculations elementwise, but in this case the integrals

$$\int_{x_i}^{x^{i+1}} p(x)\varphi_i'(x)\varphi_j'(x)$$

must be evaluated by quadrature; likewise the integrals comprising M.

Practical examples for both the 1D Poisson equation and the more general Sturm-Liouville problems are given in the codes published on the website.

3 Theory I: Variational problems

In this section, we present the keystone of the Galerkin method - the Lax-Milgram theorem concerning the existence and uniqueness of solutions to a wide variety of *variational problems* such as (3). The techniques used come from functional analysis, which takes us far afield from the course curriculum, so do not worry if you there are some claims you struggle to understand. The basic variational problem we consider takes the form

Definition 3.1. *Find $u \in V$ such that, for all $v \in V$,*

$$a(u, v) = F(v),$$

where a is a bilinear form (function $V \times V \to \mathbb{R}$), and F is a linear functional. We require that V be a so-called Hilbert space, the conditions for which do not concern us as they will be satisfied by all the function spaces we consider, but suffice to say that we need to specify an inner product and hence norm on V. For the important case of $H^1((0,1))$ we will give these, together with some inequalities that will be used in the sequel to justify the application of the Lax-Milgram theorem.

3.1 The space $H^1((0,1))$

Recall that we wished to derive weak forms such as

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \tag{11}$$

and pose them in a function space V for which all the terms above are defined. We begin by recalling that the function space $L^2((0,1))$ of square-integrable functions on $(0,1)$ is an inner product space with inner product $\langle f, g\rangle_{L^2}$ defined by

$$\langle f, g\rangle_{L^2} = \int_0^1 f(x)g(x)dx$$

If we define the associated norm $||f||_{L^2}^2 = \langle f, f\rangle_{L^2}$, the Cauchy-Schwarz inequality gives $\langle f, g\rangle_{L^2} \leq ||f||_{L^2}||g||_{L^2}$, hence

$$\int_0^1 f(x)g(x)dx \leq \left(\int_0^1 f(x)^2dx\right)^{\frac{1}{2}} \left(\int_0^1 g(x)^2dx\right)^{\frac{1}{2}}$$

In particular, if we assume that the source term $f \in L^2$, the integral on the right hand side of (11) exists if $v \in L^2$. By a similar argument, the integral on the left make sense provided $u', v' \in L^2$.

We should stop for a moment to reflect that elements of L^2 need not be continuous, and the values of discontinuous functions at any individual isolated point are irrelevant under the integral sign - indeed L^p spaces typically incorporate a quotient such that two functions which differ on a set of zero measure

only (such as an isolated point) are equivalent. This allows us to permit functions such as $f(x) = |x - \frac{1}{2}|$, which would normally be considered differentiable everywhere except at $x = \frac{1}{2}$. We say that such a function has a weak (or distributional) derivative, the formal definition of which is by integration by parts:

Definition 3.2 (Weak derivative)**.** *Let $u \in L^2((0,1))$. We say that $u'(x)$ is a weak derivative of u if, for all smooth functions v such that $v(0) = v(1) = 0$, we have*

$$\int_0^1 u'(x)v(x)dx = -\int_0^1 u(x)v'(x)dx$$

Note that if u is continuously differentiable then its weak derivative exists and coincides with the classical derivative. It can be checked that the function $f(x) = |x - \frac{1}{2}|$ has a weak derivative f' that evaluates to 1 if $x > \frac{1}{2}$ and -1 if $x < \frac{1}{2}$. On the other hand, we should not assume that (weak) differentiation can be performed by gluing together patches where the function is differentiable - observe that the weak derivative of the Heaviside step function is the Dirac delta function, which is not an element of L^2.

Theorem 3.3. *Let $H^1((0,1))$ be the space of functions $v \in L^2$ for which a weak derivative $v' \in L^2$ exists. Define the inner product by*

$$\langle f, g \rangle_{H^1} = \int_0^1 f(x)g(x)dx + \int_0^1 f'(x)g'(x)dx$$

Then $H^1((0,1))$ is a Hilbert space.

It is possible to define spaces H^k of k-times weakly differentiable functions with L^2 weak derivatives - the inner product includes all the additional integrals $\int_0^1 f^{(m)}g^{(m)}$ with $m \leq k$, and the result is still a Hilbert space. These spaces are called Sobolev spaces. For PDEs of order greater than second order (such as the biharmonic equation) this is necessary, however we will only require H^1 in these notes.

Most of the error estimates we obtain will be in the associated H^1 norm, defined by

$$||u||^2_{H^1} = \int_0^1 u(x)^2 dx + \int_0^1 u'(x)^2 dx \tag{12}$$

It will also be convenient to introduce the H^1 seminorm $|u|_{H^1}$, defined below: (the term seminorm is used as all the requirements for $|\cdot|_{H^1}$ to be a norm are satisfied except that $|u|_{H^1} = 0$ does not imply $u = 0$.)

$$|u|^2_{H^1} = \int_0^1 u'(x)^2 dx$$

The following result is convenient:

Lemma 3.4. *All functions in $H^1((0,1))$ are continuous.*

This is a particular case of the Sobolev embedding inequality, concerning the continuity of functions in H^k spaces and their derivatives - the general result is roughly that elements of H^k are m-times continuously differentiable (the case $m = 0$ giving continuity) provided $k > m + \frac{n}{2}$, where n is the dimension of the space. Note that in problems with 2 or 3 space dimensions, this no longer holds (although elements of H^2 are still continuous). This justifies the following:

Definition 3.5. *We define the space $H_0^1((0,1))$ to be the subspace of $H^1((0,1))$ comprising functions u for which $u(0) = u(1) = 0$*

If we could not guarantee continuity of u this definition would be meaningless in view of the comments on discontinuous functions in the paragraph immediately before the introduction of weak derivatives. There are ways round this (we will briefly discuss this in §5 when we come to the 2D Poisson equation), but it is better to avoid such complications where possible. The next result is critical to the application of the Lax-Milgram theorem and hence existence/uniqueness of solutions for all the examples we will consider:

Lemma 3.6 (Poincaré inequality)**.** *There exists a constant C such that, for all $v \in H_0^1((0,1))$, we have*

$$|v|_{H^1} \geq C||v||_{H^1}$$

3.2 The Lax-Milgram theorem

We are now in a position to state the promised result:

Theorem 3.7 (Lax-Milgram)**.** *Let V be a Hilbert space. Suppose that F is a continuous linear functional (it suffices that F be bounded), and that a is a continuous, coercive bilinear form, i.e. there exist M, α such that*

1. *(Continuous) $a(u,v) \leq M||u||_V||v||_V$ for all $u, v \in V$*
2. *(Coercive) $a(v,v) \geq \alpha||u||_V^2$ for all $v \in V$*

The the variational problem: find $u \in V$ such that, for all $v \in V$,

$$a(u,v) = F(v)$$

admits a unique solution.

We illustrate the use of the above theorem by applying it to show existence and uniqueness to the homogeneous Dirichlet problem

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0, \quad f \in L^2((0,1))$$

cast in the weak form: find $u \in H_0^1((0,1))$ such that for all $v \in H_0^1((0,1))$, we have

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx.$$

Now as noted in §3.1, $H_0^1((0,1))$ is a Hilbert space, and hence by the linearity of the integral operator the above equation is of the variational form, where $a(u,v)$ is the left-hand side above and $F(v)$ is the right-hand side. To invoke the theorem, it suffices to show the continuity of a and F, and coercivity of a.

We begin with the continuity of F: it suffices that F be a bounded operator on H^1, i.e.

$$||F||_{(H^1)'} = \sup_{v \neq 0} \frac{|F(v)|}{||v||_{H^1}} < \infty$$

Now the Cauchy-Schwarz inequality in L^2 gives

$$F(v) = \langle f, v \rangle_{L^2} \leq ||f||_{L^2} ||v||_{L^2} \leq ||f||_{L^2} ||v||_{H^1},$$

i.e. $||F|| \leq ||f||_{L^2}$, hence we have the desired result.

Next, we show continuity of a: indeed

$$a(u,v) = \langle u', v' \rangle_{L^2} \leq ||u'||_{L^2} ||v'||_{L^2} = |u|_{H^1} |v|_{H^1} \leq ||u||_{H^1} ||v||_{H^1}$$

Finally, for the coercivity of a we use the Poincaré inequality:

$$a(v,v) = |v|_{H^1}^2 \geq \frac{1}{2}\big(C^2 ||v||_{L^2}^2 + |v|_{H^1}^2\big) \geq \alpha ||v||_{H^2},$$

where $\alpha = \frac{1}{2}\min(C^2, 1)$.

3.3 Rayleigh-Ritz method

We will briefly examine a numerical framework that is related to the Galerkin method, but less widely applicable. In practice these can lead to identical codes, but it is sometimes useful to be aware of this alternative. We illustrate the idea by recalling that if A is a symmetric, positive definite matrix, the solution of

$$Au = b$$

may be characterized as the unique minimizer of the function

$$\Phi(y) = \frac{1}{2} y^T A y - y^T b \tag{13}$$

Due to the coercivity requirement to guarantee well-posedness of the equation, positive definiteness is typically not such a strong imposition. On the other hand, asking for symmetry can be more problematic, although it is for instance satisfied for all the boundary value problems we considered from equation (1), see the form (6) for instance. We have the following:

Lemma 3.8. *Suppose a is a bilinear, symmetric from on V and F a linear form, for which the hypotheses of the Lax-Milgram theorem are satisfied. Then $u \in V$ solves*

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

If and only if it solves the minimization problem

$$u = \text{argmin}_{v \in V} \left(\frac{1}{2} a(v,v) - F(v) \right) \tag{14}$$

For instance, the weak form (3) can be cast as the minimization problem: find $u \in H^1((0,1))$ minimizing

$$\frac{1}{2}\int_0^1 u'(x)^2 dx - \int_0^1 u(x)^2 dx$$

The Rayleigh-Ritz method proceeds by finding the minimum of (14) in a finite dimensional subspace. If we introduce a basis for this subspace, it will lead to a quadratic minimization of the form (13), where A and b are the matrix and vector that would arise from the Galerkin method. Due to the equivalence of the minimization problem and linear system, the code used to implement the methods would be therefore be the same.

4 Theory II: Stability and convergence of Galerkin methods

4.1 Coercivity and stability

One benefit of the coercivity assumption in the Lax-Milgram theorem is that we obtain a general stability estimate immediately by functional analysis, which concerns both the exact and approximate solutions.

Lemma 4.1. *Let u solve a variational problem on V governed by the Lax-Milgram theorem. Then we have the bound*

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

This follows from the coercivity relation applied to the exact solution u:

$$||u||_V^2 \leq \frac{1}{\alpha}a(u,u) = \frac{1}{\alpha}F(u) \leq \frac{1}{\alpha}||u||_V||F||_{V'},$$

where the final inequality comes from the definition of the dual norm, see §3.2. This is a stability result in the sense that, for instance when applied to (3), it shows that the solutions of the Dirichlet problems

$$\begin{aligned} -u''(x) &= f(x), \qquad 0 < x < 1, \quad u(0) = u(1) = 0 \\ -v''(x) &= g(x), \qquad 0 < x < 1, \quad v(0) = v(1) = 0 \end{aligned}$$

are related by

$$||u - v||_V \leq \frac{1}{\alpha}||f - g||_{L^2}$$

Moreover, as the approximate solution is given by a variational problem in the subspace V_h to which the Lax-Milgram theorem still applies, the bound also concerns the approximate solutions, i.e. we would also obtain

$$||u_h - v_h||_V \leq \frac{1}{\alpha}||f - g||_{L^2}$$

4.2 Towards convergence: Galerkin orthogonality and Cea's lemma

The main ingredient in convergence proofs is Cea's lemma, which intuitively bounds the approximation error from the Galerkin method in terms of how close it is possible to approximate the solution.

Lemma 4.2. *(Galerkin orthogonality) Let u and u_h be the solutions of the infinite and finite dimensional variational problems respectively. Then*

$$a(u - u_h, v_h) = 0 \quad \forall v \in V_h$$

The proof is one line: we write

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = F(v_h) - F(v_h) = 0,$$

where as v_h is in both the spaces V and V_h we could use that u and u_h solve their variational problems to replace the terms in a with the corresponding terms in F.

Lemma 4.3. *(Cea's lemma) Let u and u_h be the solutions of the infinite and finite dimensional variational problems respectively, and suppose that the hypotheses of Lax-Milgram theorem are satisfied. Notably, we assume that a is continuous and coercive with constants M and α. Then*

$$||u - u_h||_V \leq \frac{M}{\alpha}||u - v_h||_V, \quad \forall v_h \in V_h$$

The proof begins by invoking coercivity:

$$||u - u_h||_V^2 \leq \frac{1}{\alpha}a(u - u_h, u - u_h)$$

The term on the right-hand side can be expanded,

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

where the second term on the right is zero due to Galerkin orthogonality (as we have $v_h - u_h \in V_h$). These can be combined to give

$$||u - u_h||_V^2 \leq \frac{1}{\alpha}a(u - u_h, u - v_h) \leq \frac{M}{\alpha}||u - u_h||_V||u - v_h||_V,$$

where we used continuity of a for the second inequality. The result follows after dividing both sides by $||u - u_h||_V$.

4.3 Interpolation estimates and convergence

We now specialise to variational problems where $V = H^1((0,1))$ and the approximating subspace $V_h = X_h^1$, the piecewise linear functions with respect to a given grid of maximum distance h between neighbouring points. In this case, the term v_h on the right-hand side of Cea's lemma can be taken to be the interpolating polynomial of u, and we can therefore use the following estimate on the accuracy of polynomial interpolation:

Lemma 4.4. *(Polynomial interpolation) Suppose $v \in H^{r+1}((0,1))$, and let v_h^r be its degree r polynomial interpolant on a given grid with maximum element size h. Then*

$$\begin{aligned} |v - v_h^r|_{H^1} &\leq C_{1,r} h^r |v|_{H^{r+1}} \\ ||v - v_h^r||_{L^2} &\leq C_{2,r} h^{r+1} |v|_{H^{r+1}} \end{aligned}$$

Combining the above estimates with Cea's lemma (recalling that $||u||_{H^1}^2 = ||u||_{L^2}^2 + |u|_{H^1}^2$) leads to:

Corollary 4.5. *Let u_h be the numerical approximation to the solution u of a variational problem on $H^1((0,1))$ obtained by the Galerkin method using the finite dimensional subspace X_h^1 with an associated grid on $[0,1]$ of maximum element size h. Assuming the hypotheses of the Lax-Milgram theorem, provided $u \in H^2((0,1))$, we have*

$$||u - u_h||_{H^1} \leq Ch|u|_{H^2},$$

i.e. first order convergence in H^1-norm.

Note that the Lax-Milgram theorem only guarantees that $u \in H^1((0,1))$, which is not sufficient for the above theorem (we need $u \in H^2$). In fact, the method will converge even in the absence of this assumption, but not necessarily with order 1.

Having shown that for the 1D Poisson equation the method essentially coincides with the central difference method on a finite grid, which is order 2, we may wonder how we lost an order of convergence. In fact, this is simply due to the choice of norm; it can be shown that in this case we obtain also order 2 convergence in the L^2 norm, but this is more difficult to prove, and requires use of some 'elliptic regularity' estimates on the exact solution u.

5 2D Poisson equation

We now move onto the 2D Poisson equation

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{15}$$

where $\Omega \subset \mathbb{R}^2$ is an open domain. The boundary conditions will be either Dirichlet (u given on the boundary $\partial\Omega$), Neumann ($\frac{\partial u}{\partial \mathbf{n}}$ given on $\partial\Omega$, where $\mathbf{n}$ is the outward normal to Ω), or some combination of the two.

The overall outline of the method is the same: we first write the problem in the familiar variational form: find $u \in V$ such that $a(u,v) = F(v)$ for all $v \in V$, and then reduce the problem to a linear system by introduce a finite dimensional subspace $V_h \subset V$ and expanding in a given basis $\{\varphi_i\}$. The biggest difference will be that suddenly in 2D a variety of geometries (and hence subspaces and bases) are possible. This will be dealt with in §5.2, but first we will derive the variational form.

5.1 The variational form

As before, we multiply (15) by a test function v and integrate:

$$-\int_\Omega v\Delta u \; d\Omega = \int_\Omega fv \; d\Omega$$

We now integrate by parts, which in the 2D case means using Green's identity (which is derived by using the Divergence theorem on the term $\nabla.(v\nabla u)$ together with the vector calculus identity $\nabla.(\varphi\mathbf{v}) = \varphi\nabla.\mathbf{v} + \mathbf{v}.\nabla\varphi$):

$$\int_\Omega v\Delta u \; d\Omega = -\int_\Omega \nabla u.\nabla v \; d\Omega + \int_{\partial\Omega} v\frac{\partial u}{\partial \mathbf{n}}$$

Inserting this into the original expression gives the relation,

$$\int_\Omega \nabla u.\nabla v \; d\Omega = \int_{\partial\Omega} v\frac{\partial u}{\partial \mathbf{n}} + \int_\Omega fv \; d\Omega \tag{16}$$

from which the variational forms for Neumann or homogeneous Dirichlet problems are readily derived. The functions u, v will be in the function space $H^1(\Omega)$, which is defined analogously to $H^1((0,1))$: it is the space of functions v for which v and its partial derivatives $\frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}$ are square-integrable (L^2). As before, the partial derivatives can be defined in the weak sense. The space $H^1(\Omega)$ is a Hilbert space when equipped with the inner product

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

We obtain a norm in the usual manner: $||u||_{H^1} = \sqrt{\langle u, u\rangle_{H^1}}$.

One aspect does not work out so nicely - as discussed earlier, Lemma 3.4 is no longer valid, i.e. functions $v \in H^1(\Omega)$ are not necessarily continuous. This causes problems when we attempt to introduce the space on functions $H_0^1(\Omega)$ which vanish on the boundary, as we would like to do when studying the Dirichlet problem. Suffice to say, provided the boundary $\partial\Omega$ is sufficiently regular it is possible to show the existence of a unique continuous linear operator

$$\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega)$$

which for continuous $v \in H^1(\Omega)$ coincides with evaluation of v on the boundary $\partial\Omega$. We can then say:

Definition 5.1. *Let $\Omega \subset \mathbb{R}^2$ be a bounded open domain of sufficiently regularity to permit the existence of the trace operator γ_0. Then the space $H_0^1(\Omega)$ is defined to be the subspace of functions $v \in H^1(\Omega)$ for which the trace $\gamma_0 v$ vanishes.*

At this stage we should not worry too much over the details of this construction. It transpires that the Poincaré inequality (Lemma 3.6) remains valid for this space, which permits proofs of coercivity and ultimately use of the Lax-Milgram theorem along similar lines as the 1D case. Combining the functional analytic aspects with the identiy (16) we arrive at the following:

Definition 5.2 (Weak form of homogeneous Dirichlet problem)**.** *Find* $u \in H_0^1(\Omega)$ *such that*

$$\int_\Omega \nabla u.\nabla v \; d\Omega = \int_\Omega fv \; d\Omega, \quad \forall v \in H_0^1(\Omega)$$

Definition 5.3 (Weak form of Neumann problem)**.** *Find* $u \in H^1(\Omega)$ *such that*

$$\int_\Omega \nabla u.\nabla v \; d\Omega = \int_{\partial\Omega} gv + \int_\Omega fv \; d\Omega, \quad \forall v \in H^1(\Omega),$$

where the boundary data is specified as $\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}) = g(\mathbf{x})$.

5.2 Triangulations and basis functions

In the 1D case, our function space X_h^1 consisted of functions which are linear polynomials when restricted to a given element K_i, where the elements $[x_i, x_{i+1}]$ partition the space $[0, 1]$. In 1D the decomposition into intervals is essentially the only reasonable way to cover the space in finite elements, but this is no longer the case for higher dimensions. We instead choose to work with triangulations, i.e. a splitting of the domain into non-overlapping triangles:

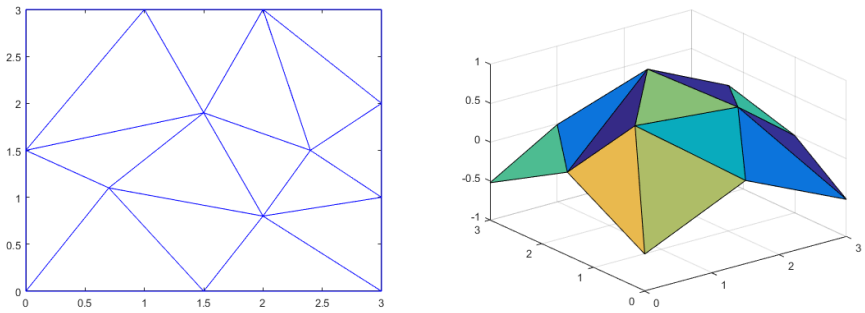

Figure 6: A triangulation of the square $[0, 3] \times [0, 3]$, and sample function in X_h^1

As shown above, it is possible to define the space X_h^1 in the same manner as before, requiring that any $f \in X_h^1$ be continuous and a linear polynomial when restricted to any triangle K_i, i.e.

$$f(x, y) = a_i + b_i x + c_i y, \quad x \in K_i$$

It follows that the value of f on K_i is determined uniquely by the values at three points (it has three *degrees of freedom*, as can be seen from the three constants a_i, b_i, c_i above), which can as before be taken to be the vertices. In view of the continuity requirement, any function $f \in X_h^1$ is uniquely determined by its values on the vertices of the triangulation of Ω, which motivates us to introduce once more the nodal basis $\{\varphi_j\}$. Let x_j be a vertex of the triangulation; the

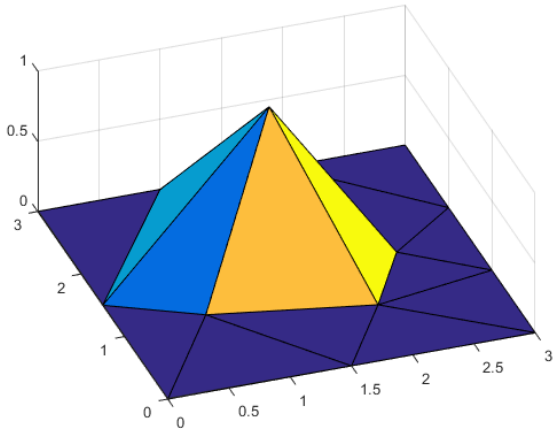

Figure 7: The nodal basis function φ_j

basis function φ_j is then defined as the unique element of X_h^1 that evaluates to 1 on x_j, and is zero on all other vertices:

Before beginning our analysis of the assembly procedure, it is important to introduce the canonical affine mapping of the *reference element* $\hat{K}$ with vertices at $\hat{\mathbf{x}}_1 = (1,0), \hat{\mathbf{x}}_2 = (0,1), \hat{\mathbf{x}}_3 = (0,0)$ to a given triangle K_i with arbitrary vertices $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. Let

$$\hat{\varphi}_1(\hat{\mathbf{x}}) = \hat{x}, \quad \hat{\varphi}_2(\hat{\mathbf{x}}) = \hat{y}, \quad \hat{\varphi}_3(\hat{\mathbf{x}}) = 1 - \hat{x} - \hat{y}, \tag{17}$$

these are readily seen to be the nodal basis functions on the reference element. The unique affine mapping of $\hat{K}$ to K_i is then given by

$$\mathbf{x}(\hat{\mathbf{x}}) = \mathbf{x}_1\hat{\varphi}_1(\hat{\mathbf{x}}) + \mathbf{x}_2\hat{\varphi}_2(\hat{\mathbf{x}}) + \mathbf{x}_3\hat{\varphi}_3(\hat{\mathbf{x}}) \tag{18}$$

It is often convenient to consider the values of the $\hat{\varphi}_i$ on $\hat{K}$ as a coordinate system, writing $(\lambda_1, \lambda_2, \lambda_3)$ for the point with $\hat{\varphi}_i = \lambda_i$. These are called *barycentric coordinates* due to the following geometric interpretation:

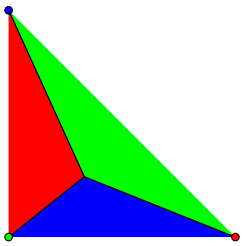

Figure 8: barycentric coordinates

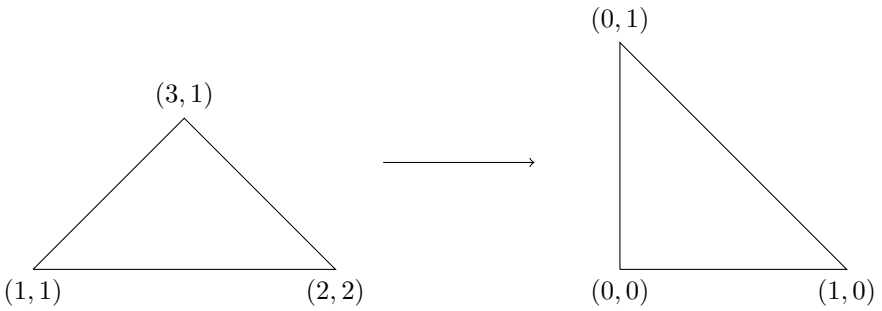

Figure 9: Mapping $K \to \hat{K}$ to the reference element

To find the barycentric coordinates of a point x, divide the triangle into three subtriangles with a common vertex at x. The barycentric coordinate $\lambda_i(x)$ associated to the vertex x_i is the ratio $\frac{A_i}{A}$, where A is the area of the large triangle (i.e., the element $\hat{K}$), and A_i is the area of the triangle of the same colour as the vertex in the diagram. This definition is not specific to the reference triangle $\hat{K}$, but can be given for an arbitrary triangle K_i, in this case we can map the barycentric coordinates λ to Cartesian coordinates following (18):

$$\mathbf{x}(\lambda) = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \lambda_3 \mathbf{x}_3 \tag{19}$$

This formula will be frequently used when computing quadratures, as quadrature formulae on triangles are typically given using nodes defined by barycentric coordinates.

5.3 Assembly

In this section we will show how to construct a linear system to find an approximate solution of the homogeneous Dirichlet problem (5.2). Assume that we have identified a triangulation $\mathcal{T}$ of our domain Ω, and now have the nodal basis $\{\varphi_j\}$ of X_h^1. We then solve (5.2) on the subspace X_h^1, i.e. find $u \in X_h^1$ such that

$$\int_\Omega \nabla u . \nabla v \, d\Omega = \int_\Omega f v \, d\Omega, \quad \forall v \in X_h^1$$

We have omitted to mention that we require u and v to be zero on the boundary $\partial\Omega$. In practice, we will proceed as before, constructing the linear system first without reference to this requiremenet and only later implementing these boundary conditions. Again, we expand u and v in terms of the basis; writing $u(x) = \sum_i u_i \varphi_i(x), v(x) = \sum_j v_j \varphi_j(x)$, the problem becomes:

Find a vector of coefficients u for which, for all vectors v

$$\sum_{i,j} u_i v_j \int_\Omega \nabla \varphi_i . \nabla \varphi_j \, d\Omega = v_j \int_\Omega f \varphi_j \, d\Omega$$

We therefore define the stiffness matrix A and load vector F by

$$A_{ij} = \int_\Omega \nabla\varphi_i.\nabla\varphi_j \, d\Omega, \quad F_j = \int_\Omega f\varphi_j \, d\Omega, \tag{20}$$

leading once again to the linear system

$$Au = F.$$

It is particularly important in the 2D (or higher dimensional) case to construct A and F elementwise, i.e. using the decomposition,

$$A_{ij} = \sum_k \int_{K_k} \nabla\varphi_i.\nabla\varphi_j \, d\Omega, \quad F_j = \sum_k \int_{K_k} f\varphi_j \, d\Omega,$$

where K_k are the elements (triangles) in the triangulation of Ω. In this case, there are three basis functions φ_j that are non-zero on each triangle K_k, namely those corresponding to nodes that are a vertex of K_k. In practice, this means that we will construct A and F by looping over the triangles, at each stage adding a 3×3 submatrix into A, and a 3 component subvector into F.

Lemma 5.4 (Stiffness submatrix)**.** *Let $K \subset \Omega$ be a triangle with vertices at $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, and let $\varphi_1, \varphi_2, \varphi_3$ be the corresponding nodal basis functions. Let J be the matrix*

$$J = [\mathbf{x}_1 - \mathbf{x}_3 | \mathbf{x}_2 - \mathbf{x}_3],$$

and let G be the solution of the linear system

$$GJ = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$$

Denoting the absolute value of the determinant of J by $|J|$, the 3×3 stiffness submatrix is given by

$$\int_{K_k} \nabla\varphi_i.\nabla\varphi_j \, d\Omega = \frac{GG^T|J|}{2!} \tag{21}$$

The proof of the above consists of transforming to the reference element and using the chain rule. Indeed, we begin by noting that

$$\nabla\varphi_i.\nabla\varphi_j = \sum_k \frac{\partial\varphi_i}{\partial x_k}\frac{\partial\varphi_j}{\partial x_k}$$

and make the substitution $x = x(\hat{x})$, $\varphi(x) = \hat{\varphi}(\hat{x})$ according to (18). The chain rule gives

$$\frac{\partial\varphi_i}{\partial x_k} = \sum_j \frac{\partial\hat{\varphi}_i}{\partial\hat{x}_j}\frac{\partial\hat{x}_j}{\partial x_k} = \left(\nabla_{\hat{\mathbf{x}}}\,\hat{\varphi}\right)J^{-1} = G,$$

where we have used

$$\nabla_{\hat{\mathbf{x}}}\,\hat{\varphi} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$$

It follows that $\nabla\varphi_i.\nabla\varphi_j = GG^T$ is constant on K_k. The result follows immediately by noting that the area of K_k is $\frac{|J|}{2!}$, which is of itself a standard formula for the area of a triangle, but can also be seen readily by noting that the area of the reference element is $\frac{1}{2!}$ and the Jacobian of the change of variable is $|J|$. The factorial in 2! is of course unnecessary, but has been left in to hint that the result generalizes straightforwardly to arbitrary dimension d, where the 2! is replaced by $d!$.

At the global level, the vertices of the triangle K_k will not generally be numbered x_1, x_2, x_3, but rather $x_{k_1}, x_{k_2}, x_{k_3}$. In this case, the submatrix (21) should accordingly be added into A_{ij} such that $i, j = k_1, k_2, k_3$, reflecting the identifiers of the basis functions φ associated to the nodes x_{k_i}. The implementation of this is discussed further in the following section §5.4.

It remains to discuss the computation of the load vector,

$$F_j = \int_{K_k} f\varphi_j \, d\Omega,$$

In practice this is accomplished by quadrature, using formulae of the form

$$F_j \approx \frac{|J|}{2!} \sum w_i \lambda_j^i f(\mathbf{x}(\lambda^i)), \quad j = 1, 2, 3 \tag{22}$$

where w_i are a set of given quadrature weights, and $\lambda^i = (\lambda_1^i, \lambda_2^i, \lambda_3^i)$ are the quadrature nodes given in barycentric coordinates. The computation of $\mathbf{x}(\lambda^i)$ is according to (19). Tables of quadrature rules for triangles can be found online or in standard texts on numerical analysis, the simplest sensible rule consists of evaluation at a single node, the barycentre $\lambda = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, with associated weight $w = 1$.

As before, it remains to tackle the boundary conditions. At present, a number of nodes x_i will lie on the boundary, and hence we wish to enforce $u_i = 0$. In the 1D case, there were only two such points numbered 0 and $M + 1$, such that it was a simple matter to simply remove the first and last rows and columns of the system. In this case, the boundary points x_i will be scattered throughout the ordering $x_1, \ldots, x_M$, so resizing the matrix like this is complicated and computational inefficient. Once these points x_i are identified, it is usually better to set $F_i = 0$, and the ith row of A to be the ith row of the identity matrix of the same size, i.e. $A_{ii} = 1$, $A_{ij} = 0$ if $j \neq i$, which achieves the same goal. Another possibility requiring even less computational effort is to set $F_i = 0$ and A_{ii} to be a very large number, which will achieve approximately the same result. The identification of boundary is more complicated - in fact it is usually best to label the boundary points whilst constructing the triangulation, and pass this to the program as part of the triangulation data.

5.4 Coding practicalities

The most straightforward way of handling the specification of a triangulation is to collect the coordinates of the vertices $x_1, \ldots, x_M$ in a $M \times 2$ matrix

$$X = [x_1| \cdots |x_M]^T,$$

together with an $N \times 3$ *connectivity* matrix T (N being the total number of elements in the triangulation), for which the kth row gives the identifying number of the three vertices making up the triangle K_k. For example, for the following simple triangulation we have

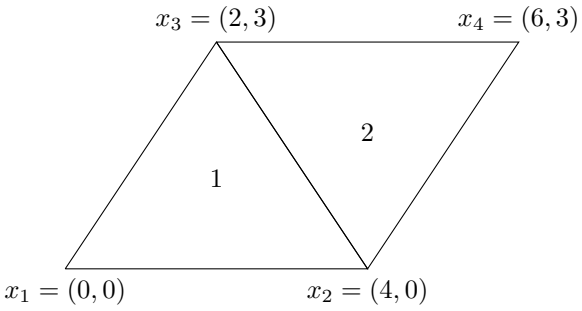

Figure 10: A simple triangulation

$$X = \begin{pmatrix} 0 & 0 \\ 4 & 0 \\ 2 & 3 \\ 6 & 3 \end{pmatrix}, \qquad T = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{pmatrix}.$$

The assembly procedure begins by initializing an $M \times M$ matrix A and an M-dimensional vector F. We then loop through the N elements. At each stage we extract the three vertices $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, which are used to compute J and hence the 3×3 matrix (21) and the 3 component vector (22).

At this stage, we then look up the row $t = T[k, :]$, which identifies the vertices of the triangle. This is then used to know where to insert these submatrices. Indeed, the procedure is

```
A[t,t] = A[t,t] + A_new;
F[t] = F[t] + F_new;
```

The sample codes available on the website give examples of the practical implementation of this procedure.

5.5 Further comments

It is possible to prove the first order convergence in H^1-norm of the finite element method using X_h^1 by means of Cea's lemma. The additional ingredients are the

bounds for the error of polynomial interpolation in 2D, which will not be given here. In this sense, h will be the maximum diameter of the elements of the triangulation; the nature of the interpolation estimates require an additional assumption that the triangulations parametrized by h form a sufficiently nice family, but this is not so restrictive in practice.

More general equations than the Poisson equation are of course possible, for instance it can be checked that discretization of the equation

$$-\Delta u(\mathbf{x}) + \sigma u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega$$

leads to linear systems of the form

$$A + \sigma M = F,$$

where A and F are as before, and M is the mass matrix

$$M_{ij} = \int_\Omega \varphi_i \varphi_j \, d\Omega.$$

Generalizing further, we can consider equations

$$-\nabla.\big(p(\mathbf{x})\nabla u(\mathbf{x})\big) + \sigma(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x}),$$

where as per the ODE case, the mass and stiffness matrices become weighted by the coefficient:

$$A_{ij} = \int_\Omega p\nabla\varphi_i.\nabla\varphi_j \, d\Omega, \quad M_{ij} = \int_\Omega \sigma\varphi_i\varphi_j \, d\Omega.$$

So far all of the problems have led to symmetric variational forms, this will no longer hold if we consider the stationary diffusion-transport-reaction problem

$$-\nabla.(p\nabla u) + \mathbf{b}.\nabla u + \sigma u = f,$$

as the middle term (the transport or convection term) will lead to a term

$$\int_\Omega \mathbf{b}\varphi_j.\nabla\varphi_i \, d\Omega.$$

6 Outlook

We have focused on applications of the finite element method to time independent equations, but the techniques also allow for treatment of evolution equations. Consider for instance the heat equation

$$u_t - \Delta u = f, \quad (x,t) \in \Omega \times [0,T],$$

multiplying by a test function and integrating leads to variational problems of the form

$$\int_\Omega v\frac{\partial u}{\partial t} \, d\Omega + a\big(u(t), v\big) = \int_\Omega v f(t) \, d\Omega,$$

where a is the bilinear form coming from the variational form of the Poisson equation. The Galerkin method can be applied, leading to an ODE of the form

$$M\dot{u}(t) + Au(t) = F(t).$$

This is a semidiscretization of the heat equation. Other strategies for evolution problems exist, for example for hyperbolic equations it is possible to derive a finite element Lax-Wendroff method by first Taylor expanding in time, and then discretizing in space by finite elements; this is an example of a *Taylor-Galerkin* method.

We conclude by mentioning that it is possible to obtain schemes of higher order by taking spaces X_h^p of pth order elementwise polynomials as an approximating space. Alternatively using spaces of orthogonal polynomials in the Galerkin scheme leads to *spectral methods*. Sometimes solution spaces must be constructed with special regard to the equations themselves, as can occur in problems in electromagnetism and fluid mechanics, for instance. Moreover it may be helpful to consider forms of *generalized Galerkin method*, where for instance the test function and solution do not belong to the same space, or a regularizing term is added to the discrete equation. Other important topics include grid generation and grid adaptivity, and of course how to program solvers as efficiently as possible.

For those wishing to take the next steps in finite element methods, the course TMA4220 is recommended!