1 Introduction

1.1 An example

As a prototype PDE consider

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad \text{or} \quad u_t = u_{xx} \quad t > 0, 0 \leqslant x \leqslant 1. \tag{1.1a}$$

We solve this subject to the initial condition

$$u(x,0) = u^0(x) \qquad 0 \leqslant x \leqslant 1, \tag{1.1b}$$

and the Dirichlet boundary condition

$$u(0,t) = 1, \quad u(1,t) = 0, \quad t > 0, 0 \leqslant x \leqslant 1 \tag{1.1c}$$

where in general (1.1a–c) must be solved numerically.

We begin by choosing a time step k and a space step $h = 1/J$ (both "small") and then we make some approximations which we hope are sensible:

$$u_t(x,t) \approx \frac{u(x,t+k) - u(x,t)}{k} \quad \text{forward difference} \tag{1.2a}$$

$$u_x(x,t) \approx \delta u(x,t) := \frac{u(x+\frac{h}{2},t) - u(x-\frac{h}{2},t)}{h} \quad \text{central difference} \tag{1.2b}$$

$$u_t(x,t) \approx \frac{u(x,t) - u(x,t-k)}{k} \quad \text{backward difference} \tag{1.2c}$$

$$\tag{1.2d}$$

then using (1.2b) twice

$$\begin{aligned} u_{xx}(x,t) &\approx \frac{u_x(x+\frac{h}{2},t) - u(x-\frac{h}{2},t)}{h} \\ &\approx \frac{1}{h}\left[\frac{u(x+h,t) - u(x,t)}{h} - \frac{u(x,t) - u(x-h,t)}{h}\right] \\ &= \frac{1}{h^2}\left[u(x+h,t) - 2u(x,t) + u(x-h,t)\right] \end{aligned} \tag{1.3}$$

Substitute (1.2a) and (1.3) into (1.1a) and evaluate at $x = jh$ and $t = nk$ then defining $u_j^n = u(jh, nk)$ we have the approximate equation for u

$$\frac{u_j^{n+1} - u_j^n}{k} \approx \frac{1}{h^2}\left[u_{j+1}^n - 2u_j^n + u_{j-1}^n\right] \tag{1.4}$$

Now replacing the $\approx$ by $=$ yields a *Finite Difference Scheme* for (1.1a)

$$\frac{U_j^{n+1} - U_j^n}{k} = \frac{1}{h^2}\left[U_{j+1}^n - 2U_j^n + U_{j-1}^n\right] \tag{1.5a}$$

called the Forwards Euler method and we approximate the initial and boundary condition by

$$U_j^0 = u^0(jh) \quad U_0 = 1 \quad \text{and} \quad U_J = 0.$$

which we hope will approximate u well. Notice that we can rearrange (1.5a) into the form

$$U_j^{n+1} = \mu U_{j+1}^n + (1-2\mu)U_j^n + \mu U_{j-1}^n \quad \text{where } \mu = \frac{k}{h^2}. \tag{1.6}$$

If we set $\{\boldsymbol{U}^n\}_j = U_j^n \ j = 1, \cdots J-1$ then

$$\boldsymbol{U}^{n+1} = \begin{pmatrix} 1-2\mu & \mu & 0 & \cdots & 0 \\ \mu & 1-2\mu & \mu & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \mu & 1-2\mu & \mu \\ 0 & \cdots & 0 & \mu & 1-2\mu \end{pmatrix} \boldsymbol{U}^n + \begin{pmatrix} \mu \times 1 \\ 0 \\ \vdots \\ 0 \\ \mu \times 0 \end{pmatrix}.$$

Suppose we chose to use a *Backwards Euler* Difference method? Then (after setting "$n = n+1$") we would arrive at

$$\begin{pmatrix} 1+2\mu & -\mu & 0 & \cdots & 0 \\ -\mu & 1+2\mu & -\mu & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\mu & 1+2\mu & -\mu \\ 0 & \cdots & 0 & -\mu & 1+2\mu \end{pmatrix} \boldsymbol{U}^{n+1} = \boldsymbol{U}^n + \begin{pmatrix} \mu \times 1 \\ 0 \\ \vdots \\ 0 \\ \mu \times 0 \end{pmatrix}$$

The matrix is a strictly diagonally dominant[1] which we can solve using a numerical method — this is the case for all implicit methods.

This term, typical questions which we hope to answer are:

1. If $t = nk$, $x = jh$ are fixed as $k, h \to 0$ (obviously $n, j \to \infty$) does

   $$U_j^n \to u(x,t) \quad (\equiv u_j^n) \qquad \text{as } k, h \to 0$$

   This is called *Convergence* or rather more precisely *Uniform Convergence*.

2. Obviously as you decrease k, h the cost of computing to (x,t) increases. How does the error $u_j^n - U_j^n$ decrease as $k, h \to 0$? This is called *Accuracy*.

3. Are there better ways of solving (1.1a–c)? This is called *Efficiency*.

4. For k and h how does the long-term behaviour of U_j^n and $u(jh,t)$ compare as $n, t \to \infty$? This is called *Asymptotics*.

[1] Such matrices are invertible.

1.2 Truncation Error

The example in the previous section is typically of a more general case. Consider a PDE of the form

$$P(\frac{\partial}{\partial t}, \frac{\partial}{\partial x})u = f \tag{1.7}$$

where f is given and $P(\frac{\partial}{\partial t}, \frac{\partial}{\partial x})$ is a linear differential operator with respect to both x and t. Suppose this is approximated by the finite difference scheme

$$\underline{P}U_j^n = f_j^n \tag{1.8}$$

where $\underline{P}$ is a linear difference operator (i.e. $\underline{P}U_j^n$ is some linear combination of U_k^m, $m \geqslant 0$, $k \in \mathbb{Z}$.)

How do we estimate the error?

Definition 1.1
The Local truncation error (LTE) for (1.8) applied to (1.7) is defined to be

$$T_j^n = \underline{P}u_j^n - f_j^n \tag{1.9}$$

where $u_j^n = u(jh, nk)$ and u is the exact solution of (1.7).

T_j^n is the extent to which u fails to satisfy (1.8). Before we do an illustrative example we have a useful lemma.

Lemma 1.1
Assume u is analytic about (jh, nk) then

$$\frac{u_{j+1}^n - u_j^n}{h} = u_x + \frac{h}{2!}u_{xx} + \cdots \tag{1.10a}$$

$$\frac{u_j^{n+1} - u_j^n}{k} = u_t + \frac{k}{2!}u_{tt} + \cdots \tag{1.10b}$$

$$\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = u_{xx} + \frac{h^2}{12}u_{xxxx} + \cdots \tag{1.10c}$$

where the derivatives of u on the right-hand side are evaluated at (jh, nk).

Proof. We only prove (1.10c) as the others are a problem on the problem sheet. We could use Taylor's theorem and then we would only need a few derivative.

Since u is analytic

$$u_{j+1}^n = u((j+1)h, nk) = u + hu_x + \frac{h^2}{2!}u_{xx} + \frac{h^3}{3!}u_{xx} + \frac{h^4}{4!}u_{xx} + \cdots$$

$$u_{j-1}^n = u((j-1)h, nk) = u - hu_x + \frac{h^2}{2!}u_{xx} - \frac{h^3}{3!}u_{xx} + \frac{h^4}{4!}u_{xx} - \cdots$$

so adding the two equations and noting that $u_j^n = u(jh, nk)$ we find

$$u_{j+1}^n + u_{j-1}^n = 2u_j^n + h^2 u_{xx} + \frac{2h^4}{4!}u_{xx} + \cdots$$

and the result follows upon rearrangement

Example 1.1
In the previous section the PDE

$$P(\frac{\partial}{\partial t}, \frac{\partial}{\partial x})u = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 = f \tag{1.11}$$

is approximated by

$$\underline{P}U_j^n = \frac{U_j^{n+1} - U_j^n}{k} - \frac{1}{h^2}\left[U_{j+1}^n - 2U_j^n + U_{j-1}^n\right] = 0 = f_j^n$$

and

$$T_j^n = \frac{u_j^{n+1} - u_j^n}{k} - \frac{1}{h^2}\left[u_{j+1}^n - 2u_j^n + u_{j-1}^n\right] - 0$$

where $u_j^n = u(jh, nk)$ and u is the exact solution of (1.11). Equivalently

$$kT_j^n = u_j^{n+1} - (1 + \mu\delta^2)u_j^n,$$

which is the error incurred when one step of the Finite Difference Scheme is used to approximate u_j^{n+1} using the exact values for u_j^n at the n'th time level, where $\delta U_j^n := U_{j+1/2}^n - U_{j-1/2}^n$ so that $\delta^2 U_j^n = U_{j+1}^n - 2U_j^n + U_{j-1}^n$. We would like T_j^n to be small.

$$\begin{aligned} T_j^n &= \frac{u_j^{n+1} - u_j^n}{k} - \frac{1}{h^2}\left[u_{j+1}^n - 2u_j^n + u_{j-1}^n\right] \\ &= \left(u_t + \frac{k}{2!}u_{tt} + \cdots\right) - \left(u_{xx} + \frac{h^2}{12}u_{xxxx} + \cdots\right) \\ &= O(k) + O(h^2) \end{aligned}$$

where $O(k)$ is the order and it simply means that $O(k)/k$ remains bounded as $k \to 0$.

Definition 1.2
A finite difference scheme is consistent with the partial differential equation if

$$T_j^n \to 0 \quad \text{as } h, k \to 0$$

and is called convergent if

$$U_j^n \to u(x, t) \quad \text{as } h, k \to 0.$$

where $jh = x$ and $nk = t$ remain fixed.

Example (1.1) shows that the proposed finite difference scheme is consistent, but this does not always lead to convergence.

Example 1.2
Consider the first order equation

$$u_t + u_x = 0. \tag{1.12a}$$

In the region $t > 0$, $x \in \mathbb{R}$ (with no explicit boundary conditions), subject to the initial conditions

$$u(x,0) = u^0(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x \geqslant 0 \end{cases} \tag{1.12b}$$

note that the exact solution is

$$u(x,t) = u^0(x-t)$$

check it! Consider approximating (1.12a) by

$$\frac{U_j^{n+1} - U_j^n}{k} + \frac{U_{j+1}^n - U_j^n}{h} = 0.$$

Rewriting this leads to

$$U_j^{n+1} = (1+\lambda)U_j^n - \lambda U_{j+1}^n \quad \text{where } \lambda = \frac{k}{h} \tag{1.13a}$$

subject to the initial condition

$$U_j^0 = \begin{cases} 1 & \text{if } j < 0 \\ 0 & \text{if } j \geqslant 0 \end{cases}. \tag{1.13b}$$

We now check for consistency. Assuming that u is analytic from (1.10a) the local truncation error is

$$\begin{aligned} T_j^n &= (u_t + \frac{k}{2!}u_{tt} + \cdots) + (u_x + \frac{h}{2!}u_{xx} + \cdots) \\ &= O(k) + O(h) \to 0 \quad \text{as } k, h \to 0. \end{aligned}$$

That is the scheme is consistent. To see that (1.13a,b) does not converge to (1.12a,b) I will prove that $U_j^n = 0$ for all $j \geqslant 0$. Note that $U_j^0 = u^0(jh) = 0$ for all $j \geqslant 0$. We now use induction and assume the hypothesis. Let $j \geqslant 0$, then from (1.13a)

$$U_j^{n+1} = (1+\lambda)U_j^n - \lambda U_{j+1}^n = 0.$$

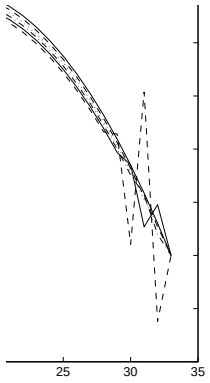
Now noting that $u(0,nk) = u^0(0-nk) = 1$ it follows that

$$0 = U_0^n \not\to 1 = u(0,nk) \quad \text{as } k \to 0,\ n \to \infty$$

where nk is fixed. So we have shown that we don't have convergence; note that u is not analytic. Look at how the characteristic moves.

Example 1.3
$u_t = u_{xx}$, $u(0,t) = u(1,t) = 0$, $u(x,0) = x(1-x)$. Using separation of variables you can show that

$$u(x,t) = \sum_{k \text{ odd}} \left(\frac{2}{k\pi}\right)^3 e^{-k^2\pi^2 t} \sin(k\pi x)$$



note the smoothing property and decay to zero. We solve $u_t = u_{xx}$ *using the forward Euler method in time and second order in space. We take* $k = C_\star h^2$ *and* $h = 1/32$ *where* $C_\star = 1.6, 0.4$*, notice that something goes wrong! Notice that if we solve the linear ODE,* $y_t = ay$*,* $y(0) = 1$ *using the forward Euler with* $k = T/n$ *then*

$$y^{n+1} = y^n + aky^n, \quad y^0 = 1 \Longrightarrow y^n = (1 + ak)^n$$

now

$$y^n = \left(1 + \frac{aT}{n}\right)^n \to e^{aT} = y(T) \text{ as } n \to \infty.$$

So there is something different between linear ODE's and linear PDE's.

2 Numerical Linear Algebra

Basics

- ⋆ *Matrix* A with (real or complex) elements a_{ij}: $A = (a_{ij})$.
- ⋆ The *Transpose* of the matrix A is denoted by A^T where $(A^T)_{ij} = a_{ji}$.
- ⋆ The *Hermitian conjugate* of the matrix A is denoted by A^H where $(A^H)_{ij} = \overline{a}_{ji}$.
- ⋆ The product of A and B which are $m \times n$ and $n \times p$ matrices respectively is

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

- ⋆ We denote the n element *column vector* $\boldsymbol{v}$; v_i are the elements of the column vector.
- ⋆ The *row vector* is denoted by $\boldsymbol{v}^T$.
- ⋆ The *inner product* for real/complex vectors is: $(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{n} \overline{u}_i v_i$.

- ⋆ A *Symmetric* matrix satisfies $A^T = A$.
- ⋆ A *Hermitian* matrix satisfies $A^H = A$.
- ⋆ A *Unitary* matrix satisfies $A^H A = I$.
- ⋆ An *Orthogonal* matrix satisfies $A^T A = I$.

Matrix Diagonalization

Two $n \times n$ matrices A and B are *similar* if there is an invertible matrix (i.e. non-singular) $n \times n$ matrix C such that

$$B = C^{-1}AC.$$

- ⋆ Similar matrices have the same eigenvalues.
- ⋆ A square matrix which is similar to a diagonal matrix is said to be *diagonizable*.
- ⋆ The eigenvalues of a diagonal matrix are its diagonal elements.
- ⋆ An $n \times n$ matrix is diagonizable if and only if it has n linearly independent eigenvectors.
- ⋆ A non-diagonizable matrix is said to be *defective* (i.e. doesn't have enough eigenvectors). Eigenvectors corresponding to different eigenvalues are linearly independent so a matrix can be defective if and only if it has at least one multiple eigenvalue.
- ⋆ An $n \times n$ real symmetric matrix has n real eigenvalues and n orthogonal eigenvectors. The same result is true of a Hermitian matrix.

THEOREM. 2.1 *(Schur decomposition) Let A be an $n \times n$ complex matrix with eigenvalues $\lambda_1 \to \lambda_n$. Then there is a unitary matrix, U, such that $U^H AU$ is upper triangular with diagonal elements $\lambda_1 \to \lambda_n$ in any order we choose.*

COROLLARY 2.2 *If A is a Hermitian matrix then A is diagonizable, has n real eigenvalues and n linearly independent eigenvectors.*

Finite Dimensional Matrix Norms

Let X be a vector space over a field $\mathbb{F}$ and $x \in X$. The *norm* of x is a non-negative number, $\|x\|$, with the properties

1. $\|x\| \geqslant 0 \ \forall \ x \in X$ and $\|x\| = 0$ iff $x = 0$.
2. $\|cx\| = |c|\|x\| \ \forall \ x \in X$ and $\forall \ c \in \mathbb{F}$.

3. $\|x+y\| \leqslant \|x\| + \|y\| \;\forall\; x, y \in X$.

EXAMPLES. Suppose $(X, \mathbb{F}) = (\mathbb{R}^n, \mathbb{R})$ or $(\mathbb{C}^n, \mathbb{C})$ with $\boldsymbol{x} = (x_1, \cdots, x_n)^T$, real or complex:

$$(a)\ \|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|, \qquad (b)\ \|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}, \qquad (c)\ \|\boldsymbol{x}\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|.$$

A *matrix norm* has the following properties:

1. $\|A\| \geqslant 0$ for all matrices A, and $\|A\| = 0$ if and only if $A = 0$.
2. $\|\alpha A\| = |\alpha| \|A\|$ for all matrices A and scalars α.
3. $\|A + B\| \leqslant \|A\| + \|B\|$ for all matrices A and B.
4. $\|AB\| \leqslant \|A\|.\|B\|$ for all matrices A and B (a desirable property).

A matrix norm and vector norm are said to be *consistent* (compatible) if

$$\|A\boldsymbol{x}\| \leqslant \|A\|.\|\boldsymbol{x}\| \qquad \text{for every } n \times n \text{ matrix } A \text{ and } n\text{-vector } \boldsymbol{x}.$$

For each vector norm, an *induced matrix norm* (subordinate) is defined as

$$\|A\| = \sup_{\|\boldsymbol{x}\| \neq 0} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \sup_{\|\boldsymbol{x}\| \neq 0} \left\| \frac{A\boldsymbol{x}}{\|\boldsymbol{x}\|} \right\| = \max_{\|\boldsymbol{x}\| = 1} \|A\boldsymbol{x}\|.$$

A vector norm and its induced matrix norm are *always* consistent, see any NA book.

1. $\|A\|_\infty := \max_{\|\boldsymbol{x}\|_\infty = 1} \|A\boldsymbol{x}\|_\infty = \max_{i=1 \to n} \sum_{j=1}^{n} |a_{ij}|$.
2. $\|A\|_1 := \max_{\|\boldsymbol{x}\|_1 = 1} \|A\boldsymbol{x}\|_1 = \max_{j=1 \to n} \sum_{i=1}^{n} |a_{ij}|$.
3. $\|A\|_2 = \max_{\|\boldsymbol{x}\| = 1} \|A\boldsymbol{x}\|_2 = \max_{\|\boldsymbol{x}\| = 1} \left(\boldsymbol{x}^H A^H A \boldsymbol{x} \right)^{1/2} = \sqrt{\rho(A^H A)}$, where $\rho(B) = \max_i |\lambda_i(B)|$ is called the *spectral radius* of the matrix B.
4. The Frobenius norm $\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}$ is consistent with $\| \bullet \|_2$ but is not an induced norm.

PROOF. 1. Since $\|\boldsymbol{x}\|_\infty = \max_{i=1 \to n} |x_i| = 1$, then

$$\|A\boldsymbol{x}\|_\infty := \max_{i=1 \to n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \max_{i=1 \to n} \sum_{j=1}^{n} |a_{ij}|.\|\boldsymbol{x}\|_\infty = \max_{i=1 \to n} \sum_{j=1}^{n} |a_{ij}|.$$

The above inequality may be attained by a particular choice of $\boldsymbol{y}$. Let p be chosen so that $\sum_{j=1}^{n} |a_{pj}| = \max_{i=1\to n} \sum_{j=1}^{n} |a_{ij}|$. If A is real, then define $\boldsymbol{y}$ ($\|\boldsymbol{y}\|_\infty = 1$) by

$$y_j = \begin{cases} 1 & \text{if } a_{pj} \geqslant 0 \\ -1 & \text{if } a_{pj} < 0 \end{cases} \Longrightarrow (A\boldsymbol{y})_p = \sum_{j=1}^{n} a_{pj} y_j = \sum_{j=1}^{n} |a_{pj}| \leqslant \|A\|_\infty.$$

2. Since $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| = 1$, then

$$\|A\boldsymbol{x}\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|.|x_j| = \sum_{j=1}^{n} |x_j| \max_{j=1\to n} \sum_{i=1}^{n} |a_{ij}| = \max_{j=1\to n} \sum_{i=1}^{n} |a_{ij}|.$$

The above estimate is attainable. Let p satisfy $\sum_{i=1}^{n} |a_{ip}| = \max_{j=1\to n} \sum_{i=1}^{n} |a_{ij}|$, define

$$y_j = \begin{cases} 1 & \text{if } j = p \\ 0 & \text{if } j \neq p \end{cases} \Longrightarrow \|\boldsymbol{y}\|_1 = 1 \text{ and } \|A\boldsymbol{y}\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} y_j \right| = \sum_{i=1}^{n} |a_{ip}| \leqslant \|A\|_1.$$

3. $A^H A$ is Hermitian; it has n real non-negative eigenvalues $\lambda_1, \cdots, \lambda_n$ and n orthonormal eigenvectors $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n$. Every n-vector can be expressed as $\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{u}_i$. If $\|\boldsymbol{x}\|_2 = 1$ then $\sum_{i=1}^{n} |\alpha_i|^2 = 1$ and $A^H A\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \lambda_i \boldsymbol{u}_i$ so that

$$\boldsymbol{x}^H A^H A\boldsymbol{x} = \sum_{j=1}^{n} \overline{\alpha}_j \boldsymbol{u}_j^H \sum_{i=1}^{n} \alpha_i \lambda_i \boldsymbol{u}_i = \sum_{i=1}^{n} \lambda_i |\alpha_i|^2 \leqslant \max_{j=1\to n} \lambda_j \sum_{i=1}^{n} |\alpha_i|^2 = \max_{j=1\to n} \lambda_j.$$

Let $\lambda_p = \max_{j=1\to n} \lambda_j$ and $\boldsymbol{x} = \boldsymbol{u}_p$ then $\boldsymbol{x}^H A^H A\boldsymbol{x} = \boldsymbol{u}_p^H A^H A\boldsymbol{u}_p = \lambda_p = \rho(A^H A)$.

THEOREM. 2.3 **(a)** *For any consistent matrix norm* $\rho(A) \leqslant \|A\|$.

(b) *If* λ *is an eigenvalue of* A*, then* λ^m *is an eigenvalue for* A^m $(m = 0, 1, \ldots)$.

Convergence of sequences and series

Let A be a square matrix. The sequence $\{A^m\}$ *converges* iff $\lim_{m\to\infty} A^m = 0$.

THEOREM. 2.4 *(Convergence) Let $\|.\|$ be a consistent matrix norm. The sequence $\{A^m\}$ converges if and only if $\rho(A) < 1$.*

THEOREM. 2.5 *The series $I + A + A^2 + \cdots$ converges if and only if $\lim_{m\to\infty} A^m = 0$.*

COROLLARY 2.6 *Let $\|.\|$ be a consistent matrix norm. If $\|A\| < 1$ then $\{A^m\}$ converges and $I + A + A^2 + \cdots$ converges.*

3 Linear Equations

Indirect methods for solving $A\boldsymbol{x} = \boldsymbol{b}$

We consider iterative methods which, given $\boldsymbol{x}^{(0)}$, yield the sequence $\{\boldsymbol{x}^{(k)}\}$ generated by the linear one-point iteration scheme

$$\boldsymbol{x}^{(k+1)} = M\boldsymbol{x}^{(k)} + \boldsymbol{c} \qquad (k = 0, 1, \ldots).$$

An iteration is said to be *consistent* with the solution of $A\boldsymbol{x} = \boldsymbol{b}$ iff $\{\boldsymbol{x}^{(k)}\}$ converges to $\boldsymbol{x}$ and $\boldsymbol{x}$ is the stationary iterate.

If $a_{ii} \neq 0$ for $i = 1 \to n$ two consistent one-point methods are *Jacobi's iteration* and *Gauss-Seidel iteration* which are given by

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\right) \text{ and } x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right) \qquad i = 1 \to n$$

which can be written in terms of a matrix iteration:

$$M = -D^{-1}(L+U),\ \boldsymbol{c} = D^{-1}\boldsymbol{b} \text{ and } M = -(D+L)^{-1}U,\ \boldsymbol{c} = (D+L)^{-1}\boldsymbol{b}$$

where D, L and U are diagonal, lower triangular and upper triangular matrices associated with A.

EXAMPLE.

$$\begin{pmatrix} 10 & 1 & 1 \\ 1 & 10 & -1 \\ -1 & -1 & 10 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 10 \\ 8 \end{pmatrix} \text{ is solved by } \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Jacobi Iteration:
$$\begin{pmatrix} 10x_1^{(k+1)} \\ 10x_2^{(k+1)} \\ 10x_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} 12 \\ 10 \\ 8 \end{pmatrix} - \begin{pmatrix} x_2^{(k)} + x_3^{(k)} \\ x_1^{(k)} - x_3^{(k)} \\ -x_1^{(k)} - x_2^{(k)} \end{pmatrix}$$

$$\begin{array}{ccccccc} \boldsymbol{x}^{(0)} & \boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} & \boldsymbol{x}^{(3)} & \boldsymbol{x}^{(4)} & \boldsymbol{x}^{(5)} \\ \begin{pmatrix}0\\0\\0\end{pmatrix} & \begin{pmatrix}1.2\\1.0\\0.8\end{pmatrix} & \begin{pmatrix}1.02\\0.96\\1.02\end{pmatrix} & \begin{pmatrix}1.002\\1.000\\0.998\end{pmatrix} & \begin{pmatrix}1.0002\\0.9996\\1.0002\end{pmatrix} & \begin{pmatrix}1.00002\\1.00000\\0.99998\end{pmatrix} \end{array}$$

Gauss-Seidel Iteration: $$\begin{pmatrix}10x_1^{(k+1)}\\10x_2^{(k+1)}\\10x_3^{(k+1)}\end{pmatrix} = \begin{pmatrix}12\\10\\8\end{pmatrix} - \begin{pmatrix}x_2^{(k)} + x_3^{(k)}\\x_1^{(k+1)} - x_3^{(k)}\\-x_1^{(k+1)} - x_2^{(k+1)}\end{pmatrix}$$

$$\begin{array}{ccccc} \boldsymbol{x}^{(0)} & \boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} & \boldsymbol{x}^{(3)} & \boldsymbol{x}^{(4)} \\ \begin{pmatrix}0\\0\\0\end{pmatrix} & \begin{pmatrix}1.2\\0.88\\1.008\end{pmatrix} & \begin{pmatrix}1.0112\\0.99968\\1.00109\end{pmatrix} & \begin{pmatrix}0.99992\\1.00012\\1.000004\end{pmatrix} & \begin{pmatrix}0.99999\\1.000002\\0.999999\end{pmatrix} \end{array}.$$

So when does a consistent, one-point iterative converge, i.e. when does $\{\boldsymbol{x}^{(k)}\}$ converge to $\boldsymbol{x}$? Defining $\boldsymbol{e}^{(k)} = \boldsymbol{x} - \boldsymbol{x}^{(k)}$ it follows that

$$\boldsymbol{e}^{(k+1)} = M\boldsymbol{x} + \boldsymbol{c} - M\boldsymbol{x}^{(k)} - \boldsymbol{c} = M\boldsymbol{e}^{(k)} \Longrightarrow \boldsymbol{e}^{(k)} = M^k\boldsymbol{e}^{(0)}.$$

Hence picking *any* consistent matrix and vector norm, convergence follows iff $\lim_{k\to\infty} \boldsymbol{e}^{(k)} = \mathbf{0}$ iff $\rho(M) < 1$. A practical way to estimate the error:

$$\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \leqslant \frac{\|M\|^{k-1}}{1 - \|M\|}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|.$$

THEOREM. 3.1 *If A is a strictly diagonally dominant square matrix then both the Jacobi and Gauss-Seidel iterations converge.*

THEOREM. 3.2 *If A is a real, symmetric, positive definite matrix ($\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$), then the Gauss-Seidel iterates converge.*

Some other one-point iterative methods are the SOR method $\omega \in \mathbb{R}$

$$(I + \omega D^{-1}L)\boldsymbol{x}^{(k+1)} = \left((1-\omega)I - \omega D^{-1}U\right)\boldsymbol{x}^{(k)} + \omega D^{-1}\boldsymbol{b}$$

and the AOR $r, \omega \in \mathbb{R}$

$$(I + rD^{-1}L)\boldsymbol{x}^{(k+1)} = \left((1-\omega)I - (\omega - r)D^{-1}L - \omega D^{-1}U\right)\boldsymbol{x}^{(k)} + \omega D^{-1}\boldsymbol{b}.$$

THEOREM. 3.3 *A necessary condition for convergence of the SOR iteration is $0 < \omega < 2$.*

4 Convergence of Finite Difference Schemes

4.1 Basic theory

Recalling the notation from Section 1.2 we consider approximating

$$P(\frac{\partial}{\partial t}, \frac{\partial}{\partial x})u = f \tag{4.1a}$$

with initial conditions

$$u(x,0) = u^0(x), \quad x \in \mathbb{R}. \tag{4.1b}$$

By some finite difference scheme

$$\underline{P}U_j^n = f_j^n \tag{4.2a}$$

with initial conditions

$$U_j^0 = u^0(jh). \tag{4.2b}$$

If no boundary conditions are imposed then j is unbounded so $j \in \mathbb{Z}$. For each time level n, the solution values are a bi-infinite sequence $\left\{U_j^n | j \in \mathbb{Z}\right\}$. If boundary conditions are imposed then j runs through some finite set, say $j = 1, \cdots J-1$, and at time level n the solution values $\left\{U_j^n | j = 1, \cdots J-1\right\}$ is a vector in $\mathbb{R}^{J-1}$. Therefore at each time level the solution values lie in some vector space S. In which case it may be possible to reformulate the linear finite difference scheme into the computational form

$$\boldsymbol{U}^{n+1} = B\boldsymbol{U}^n + k\boldsymbol{f}^n. \tag{4.3}$$

In what follows $\|\cdot\|$ will be a norm associated with the vector space S.

Example 4.1
Suppose we take the example, from Section 1.1, (1.1a–c) with $v(t) \equiv w(t) \equiv 0$. A finite difference scheme for (1.1a) is (see (1.6))

$$U_j^{n+1} = \mu U_{j+1}^n + (1-2\mu)U_j^n + \mu U_{j-1}^n \quad \textit{where } j = 1, \cdots J-1 \quad \textit{and } \mu = \frac{k}{h^2} \tag{4.4a}$$

now (1.1c) leads to

$$U_0^n = 0 = U_J^n, \quad n \geqslant 1. \tag{4.4b}$$

Hence $\boldsymbol{U}^{n+1} = B\boldsymbol{U}^n$ where

$$B = \begin{pmatrix} 1-2\mu & \mu & 0 & \cdots & 0 \\ \mu & 1-2\mu & \mu & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \mu & 1-2\mu & \mu \\ 0 & \cdots & 0 & \mu & 1-2\mu \end{pmatrix}.$$

Now

$$\|B\boldsymbol{V}\|_\infty \leqslant [|\mu| + |1-2\mu| + |\mu|] \, \|\boldsymbol{V}\|_\infty \Longrightarrow \|B\|_\infty \leqslant \begin{cases} 1 & \mu \leqslant 1/2 \\ 4\mu - 1 & \mu \geqslant 1/2 \end{cases}$$

4.2 Stability Theory for PDE's

Definition 4.1
Consider solving (4.1a,b) by the method (4.2a,b) then:

1. *The Truncation error is*

$$T_j^n = \underline{P} u_j^n - f_j^n; \quad \textit{the vector } \boldsymbol{T}^n \in S.$$

2. *The method (4.2a,b) is consistent if*

$$\max_{m=0,\cdots n} \|\boldsymbol{T}^m\| \to 0 \quad \textit{as } k, h \to 0, \quad nk = T \textit{ fixed.}$$

3. *The method (4.2a,b) is convergent if*

$$\|\boldsymbol{U}^n - \boldsymbol{u}^n\| \to 0 \quad \textit{as } k, h \to 0, \quad nk = T \textit{ fixed.}$$

4. *The method (4.2a,b) is stable if*

   for all $T > 0$ there exists $C_T > 0$ such that for all initial conditions $\boldsymbol{u}^0$, $\boldsymbol{v}^0$ then the corresponding solutions at the n'th time level satisfy

$$\|\boldsymbol{U}^n - \boldsymbol{V}^n\| \leqslant C_T \|\boldsymbol{U}^0 - \boldsymbol{V}^0\| \quad \textit{when } nk \leqslant T.$$

We shall show that consistency and stability imply convergence; Lax equivalence theorem.

Remark

Notice that for stability $\boldsymbol{U}^n$ and $\boldsymbol{V}^n$ satisfy

$$\boldsymbol{U}^{n+1} = B\boldsymbol{U}^n + k\boldsymbol{f}^n \qquad \boldsymbol{V}^{n+1} = B\boldsymbol{V}^n + k\boldsymbol{f}^n \Longrightarrow \boldsymbol{U}^{n+1} - \boldsymbol{V}^{n+1} = B(\boldsymbol{U}^n - \boldsymbol{V}^n)$$

$$\Longrightarrow \boldsymbol{U}^n - \boldsymbol{V}^n = B^n(\boldsymbol{U}^0 - \boldsymbol{V}^0).$$

Thus, defining $\boldsymbol{W}^n = \boldsymbol{U}^n - \boldsymbol{V}^n$ where $\boldsymbol{W}^{n+1} = B\boldsymbol{W}^n$ (i.e. take $f \equiv 0$) the stability criterion boils down to

$$\|\boldsymbol{W}^n\| \leqslant C_T \|\boldsymbol{W}^0\| \qquad \text{or} \qquad \|B^n\| \leqslant C_T.$$

Lemma 4.1
Suppose $\boldsymbol{e}^n$, $\boldsymbol{t}^n \in S$ satisfy

$$\boldsymbol{e}^{n+1} = B\boldsymbol{e}^n + \boldsymbol{t}^n \quad n \geqslant 0 \tag{4.5}$$

where B is some linear transformation on S (i.e. a matrix), then

$$\boldsymbol{e}^n = B^n\boldsymbol{e}^0 + \sum_{m=0}^{n-1} B^{n-1-m}\boldsymbol{t}^m \qquad n \geqslant 1. \tag{4.6}$$

PROOF. Clearly (4.6) holds when $n = 1$. We now assume (4.6) holds for some $n \geqslant 1$ and prove the result by induction. Now

$$\begin{aligned} \boldsymbol{e}^{n+1} &= B\boldsymbol{e}^n + \boldsymbol{t}^n = B\left(B^n\boldsymbol{e}^0 + \sum_{m=0}^{n-1} B^{n-1-m}\boldsymbol{t}^m\right) + \boldsymbol{t}^n \\ &= B^{n+1}\boldsymbol{e}^0 + \sum_{m=0}^{n-1} B^{n-m}\boldsymbol{t}^m + B^{1+n-n}\boldsymbol{t}^n = B^{n+1}\boldsymbol{e}^0 + \sum_{m=0}^{n} B^{n-m}\boldsymbol{t}^m. \end{aligned}$$

Theorem 4.1
Assume that B in (4.3) is a linear transformation on S and that the method (4.2a,b) is stable. Then

$$\|\boldsymbol{U}^n - \boldsymbol{u}^n\| \leqslant TC_T \max_{m=0,\cdots n-1} \|\boldsymbol{T}^m\|$$

for all n such that $nk = T$.

Hence if the scheme is also consistent, then it converges and its rate of convergence is determined by how fast the quantity $\max_{m=0,\cdots n-1} \|\boldsymbol{T}^m\|$ approaches zero.

Proof. Recall (4.3)

$$\boldsymbol{U}^{n+1} = B\boldsymbol{U}^n + k\boldsymbol{f}^n. \tag{4.7}$$

was obtained from (4.2a)

$$\underline{P}U_j^n = f_j^n$$

by multiplication of k and rearrangement. Also note that by the definition of the truncation error

$$\underline{P}u_j^n = f_j^n + T_j^n$$

and it follows from a similar argument

$$\boldsymbol{u}^{n+1} = B\boldsymbol{u}^n + k\boldsymbol{f}^n + k\boldsymbol{T}^n. \tag{4.8}$$

Defining $\boldsymbol{e}^n = \boldsymbol{u}^n - \boldsymbol{U}^n$ then subtract (4.7) from (4.8) we obtain

$$\boldsymbol{e}^{n+1} = B\boldsymbol{e}^n + k\boldsymbol{T}^n \quad n \geqslant 0 \tag{4.9}$$

Then from Lemma 4.1

$$\boldsymbol{e}^n = B^n\boldsymbol{e}^0 + k\sum_{m=0}^{n-1} B^{n-1-m}\boldsymbol{T}^m.$$

However, noting that $\boldsymbol{e}^0 = 0$ by (4.2b), linearity of B, compatibility of the matrix and vector norms and the triangle inequality it follows by taking norms that

$$\begin{aligned} \|\boldsymbol{e}^n\| &= k\sum_{m=0}^{n-1} \|B^{n-1-m}\boldsymbol{T}^m\| \leqslant k\sum_{m=0}^{n-1} \|B^{n-1-m}\|\|\boldsymbol{T}^m\| && (4.10) \\ &\leqslant kn \max_{m=0,\cdots n-1} \|B^{n-1-m}\| \max_{m=0,\cdots n-1} \|\boldsymbol{T}^m\| && (4.11) \end{aligned}$$

Now noting the remark after Definition 4.1 yields the result.

Example 4.2
Returning to Example 1.1

$$U_j^{n+1} = (1+\mu\delta^2)U_j^n \quad U_j^0 = u^0(jk), \tag{4.12}$$

with no boundary conditions. There is a technique for assessing stability. Suppose the solution of (4.12) takes the form

$$U_j^n = g^n e^{ij\xi}, \quad j \in \mathbb{Z}, \quad n \geqslant 0 \tag{4.13}$$

for some $\xi \in [-\pi, \pi]^2$. *Substituting this into (4.12) yields*

$$g^{n+1}e^{ij\xi} = (1+\mu\delta^2)g^n e^{ij\xi} = g^n(1+\mu(e^{i\xi} - 2 + e^{-i\xi}))e^{ij\xi} = (1 - 4\mu\sin^2(\tfrac{\xi}{2}))g^n e^{ij\xi}.$$

Noting that $g \neq 0$ *we obtain that* $g = (1 - 4\mu\sin^2(\frac{\xi}{2}))$ *and* $U_j^{n+1} = gU_j^n$. *Turning to the question of stability, take* $\boldsymbol{V}^0 = \boldsymbol{0} = \boldsymbol{V}^n$, *then*

$$\|\boldsymbol{U}^n\| = |1 - 4\mu\sin^2(\tfrac{\xi}{2})|^n \|\boldsymbol{U}^0\|.$$

Hence the scheme is unstable if and only if

$$|1 - 4\mu\sin^2(\tfrac{\xi}{2})| > 1 \iff 1 - 4\mu\sin^2(\tfrac{\xi}{2}) > 1 \text{ or } -(1 - 4\mu\sin^2(\tfrac{\xi}{2})) > 1 \iff 2 < 4\mu\sin^2(\tfrac{\xi}{2})$$

holds for at least one $\xi \in [-\pi, \pi]$ *and so*

$$2 < 4\mu \iff \frac{1}{2} < \mu.$$

From linearity we have

$$\|\boldsymbol{U}^n - \boldsymbol{V}^n\| = |1 - 4\mu\sin^2(\tfrac{\xi}{2})|^n \|\boldsymbol{U}^0 - \boldsymbol{V}^0\|$$

and so for stability we require that for all $\xi \in [-\pi, \pi]$

$$\begin{aligned} &|1 - 4\mu\sin^2(\tfrac{\xi}{2})|^n \leqslant C_T \iff |1 - 4\mu\sin^2(\tfrac{\xi}{2})| \leqslant 1 \\ &\iff 1 - 4\mu\sin^2(\tfrac{\xi}{2}) \leqslant 1 \text{ and } 1 - 4\mu\sin^2(\tfrac{\xi}{2})) \geqslant -1 \iff 2 \geqslant 4\mu\sin^2(\tfrac{\xi}{2}) \end{aligned}$$

and so $\mu \leqslant \frac{1}{2}$.

[2]We assume $g \neq 0$ otherwise $U_j^n = 0$, or $U_j^0 = 0$, and this is a very special case!

4.3 Fourier Analysis

Let S be the space of bi-infinite sequences with norm

$$\|\cdot\|_2 = \left[\sum_{j\in\mathbb{Z}} h|V_j|^2\right]^{1/2}$$

which is bounded. We define the Fourier Transform of $\boldsymbol{V} \in S$ by

$$\widehat{\boldsymbol{V}}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{j\in\mathbb{Z}} hV_j \ \mathrm{e}^{-ijh\xi}, \quad \xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right].$$

Lemma 4.2
If $\boldsymbol{V} \in S$ then we have the "inversion formula"

$$V_k = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{ikh\xi} \widehat{\boldsymbol{V}}(\xi)\, d\xi$$

and "Parseval's identity"

$$\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |\widehat{\boldsymbol{V}}(\xi)|^2\, d\xi = \|\boldsymbol{V}\|_2^2$$

Proof. Assuming that we can interchange double sums, integrals and sums, etc.

$$\frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{ikh\xi} \widehat{\boldsymbol{V}}(\xi)\, \mathrm{d}\xi \ = \ \frac{1}{2\pi} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{ikh\xi} \sum_{j\in\mathbb{Z}} hV_j \ \mathrm{e}^{-ijh\xi}\, \mathrm{d}\xi = \frac{1}{2\pi} \sum_{j\in\mathbb{Z}} hV_j \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{i(k-j)h\xi}\, \mathrm{d}\xi.$$

If $k \neq j$ then

$$\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{i(k-j)h\xi}\, \mathrm{d}\xi \ = \ \frac{1}{i(k-j)h} \left[\ \mathrm{e}^{i(k-j)h\xi}\right]_{-\frac{\pi}{h}}^{\frac{\pi}{h}} = \frac{1}{i(k-j)h} \left[\ \mathrm{e}^{i(k-j)\pi} - \ \mathrm{e}^{-i(k-j)\pi}\right] = 0.$$

If $j = k$ then

$$\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{i(k-j)h\xi}\, \mathrm{d}\xi \ = \ 2\frac{\pi}{h}.$$

Hence

$$\frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{ikh\xi} \widehat{\boldsymbol{V}}(\xi)\, \mathrm{d}\xi \ = \ \frac{1}{2\pi} hV_k \times 2\frac{\pi}{h} = V_k.$$

Now to prove Parseval's equality

$$\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |\widehat{\boldsymbol{V}}(\xi)|^2\, \mathrm{d}\xi \ = \ \frac{1}{2\pi} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} h^2 V_j \overline{V_k} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \ \mathrm{e}^{-i(j-k)h\xi}\, \mathrm{d}\xi = \sum_{j\in\mathbb{Z}} h|V_j|^2\, \mathrm{d}\xi = \|\boldsymbol{V}\|_2^2$$

Theorem 4.2
Let $\underline{P}$ be a spatial linear finite difference operator.

1. *The difference scheme*
$$U_j^{n+1} = \underline{P}U_j^n \tag{4.14}$$
is satisfied by the trial solution
$$U_j^n = g^n \ \mathrm{e}^{ijh\xi}, \quad \xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right] \tag{4.15}$$
$g \neq 0$ if and only if $g = g(h\xi)$ satisfies
$$\widehat{\underline{P}\boldsymbol{V}}(\xi) = g\widehat{\boldsymbol{V}}(\xi), \quad \xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right] \tag{4.16}$$
for all $\boldsymbol{V} \in S$.

2. *If (4.15) satisfies (4.14) then any finite difference scheme of the form*
$$U_j^{n+1} = \underline{P}U_j^n + kf_j^n \tag{4.17}$$
is stable with respect to the $\|\cdot\|_2$ if and only if
$$|g| \leqslant 1 + Ck, \quad \xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right] \tag{4.18}$$
as $k \to 0$ for some fixed constant C.

Proof. Since $\underline{P}$ is a linear spatial finite difference operator,
$$\underline{P}V_j = \sum_{k=-p}^{q} a_k V_{j+k}$$
for some coefficients a_k. Noting that for k fixed
$$\widehat{\{V_{j+k}\}}(\xi) \ = \ \frac{1}{\sqrt{2\pi}} \sum_{j\in\mathbb{Z}} hV_{j+k} \ \mathrm{e}^{-ijh\xi} = \frac{1}{\sqrt{2\pi}} \sum_{j\in\mathbb{Z}} hV_j \ \mathrm{e}^{-i(j-k)h\xi} = \ \mathrm{e}^{ikh\xi}\widehat{\boldsymbol{V}}(\xi)$$
it follows since the Fourier transform and $\underline{P}$ are linear that
$$\widehat{\underline{P}\boldsymbol{V}}(\xi) = \left\{\widehat{\sum_{k=-p}^{q} a_k V_{j+k}}\right\}(\xi) = \sum_{k=-p}^{q} a_k \ \mathrm{e}^{ikh\xi}\widehat{\boldsymbol{V}}(\xi).$$
Now (4.15) satisfies (4.14), hence
$$\begin{aligned} g^{n+1} \ \mathrm{e}^{ijh\xi} \ &= \ \underline{P}U_j^n = \sum_{k=-p}^{q} a_k U_{j+k}^n = \sum_{k=-p}^{q} a_k g^n \ \mathrm{e}^{i(j+k)h\xi} \\ &= \ \left[\sum_{k=-p}^{q} a_k \ \mathrm{e}^{ikh\xi}\right] g^n \ \mathrm{e}^{ijh\xi} \Longleftrightarrow g = \left[\sum_{k=-p}^{q} a_k \ \mathrm{e}^{ikh\xi}\right]. \end{aligned}$$

For the second part, we only prove the result in one direction. Suppose $g \neq 0$ satisfies (4.18). Let $\boldsymbol{U}^n$ and $\boldsymbol{V}^n$ be two solutions coming from initial conditions $\boldsymbol{U}^0$ and $\boldsymbol{V}^0$ and set $\boldsymbol{W}^n = \boldsymbol{U}^n - \boldsymbol{V}^n$, then from linearity $W_j^{n+1} = \underline{P} W_j^n$. Hence applying the Fourier transform

$$\widehat{\boldsymbol{W}^{n+1}}(\xi) = g\widehat{\boldsymbol{W}^n}(\xi).$$

so by the previous Lemma,

$$\begin{aligned} \|\boldsymbol{W}^{n+1}\|_2^2 &= \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |\widehat{\boldsymbol{W}^{n+1}}(\xi)|^2 \,\mathrm{d}\xi = \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |g|^2 |\widehat{\boldsymbol{W}^n}(\xi)|^2 \,\mathrm{d}\xi \\ &\leqslant (1+Ck)^2 \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |\widehat{\boldsymbol{W}^n}(\xi)|^2 \,\mathrm{d}\xi = (1+Ck)^2 \|\boldsymbol{W}^n\|_2^2 \end{aligned}$$

hence taking square-roots and using induction

$$\|\boldsymbol{W}^n\|_2 \leqslant (1+Ck)^n \|\boldsymbol{W}^0\|_2 \leqslant \mathrm{e}^{Ckn} \|\boldsymbol{W}^0\|_2 \leqslant \mathrm{e}^{CT} \|\boldsymbol{W}^0\|_2$$

when $kn \leqslant T$ so the method is stable with respect to the $\|\cdot\|_2$ norm.

Example 4.3
In Example 2.2 we found

$$g = 1 - 4\mu \sin^2 \tfrac{\xi}{2}.$$

For μ fixed, g is independent of k and so stability is equivalent to

$$|g| \leqslant 1, \forall\ h \in [-\pi, \pi] \iff \mu \leqslant \frac{1}{2}.$$

Notice that this has the important implication that for stability $k \leqslant \frac{1}{2}h^2$. Also if the amplification factor, g, is independent of k then the previous theorem says that the scheme is stable if and only if $|g| \leqslant 1$. In Example 1.3, the oscillation were caused by instability.

Example 4.4
In Example 1.3 we showed that the finite difference scheme

$$U_j^{n+1} = (1+\lambda)U_j^n - \lambda U_{j+1}^n \quad \textit{where } \lambda = \frac{k}{h}$$

approximating $u_t + u_x = 0$ was consistent. Is it stable? Setting $U_j^n = g^n\ \mathrm{e}^{ij\xi}$ yields that

$$g = 1 + \lambda - \lambda\ \mathrm{e}^{i\xi} = 1 + \lambda - \lambda\cos\xi - i\lambda\sin\xi$$

and hence

$$\begin{aligned} |g|^2 &= (1+\lambda-\lambda\cos\xi)^2 + \lambda^2\sin^2\xi = (1+\lambda)^2 - 2\lambda(1+\lambda)\cos\xi + \lambda^2 \\ &= 1 + 2\lambda(1+\lambda)(1-\cos\xi) = 1 + 4\lambda(1+\lambda)\sin^2\tfrac{\xi}{2} > 1 \end{aligned}$$

and so for λ fixed the method is unstable for all $\lambda > 0$ which was expected due to the instability of the finite difference scheme.

Notice that the same methodology for proving stability can be applied to more general difference schemes.

In Section 1.1 we considered an explicit finite difference method to solve the heat equation. We now consider the *θ-method* which for $\theta > 0$ is implicit. Let $\theta \in [0,1]$ and consider the finite difference method

$$\frac{1}{k}(U_j^{n+1} - U_j^n) = \frac{1}{h^2}\left[\theta\delta^2 U_j^{n+1} + (1-\theta)\delta^2 U_j^n\right], \quad U_j^0 = u^0(jh). \tag{4.19}$$

With $\theta = 1, \frac{1}{2}, 0$ this is called the backwards/implicit Euler method, Crank-Nicholson method and forward/explicit Euler method.

In computational form one writes

$$(1 - \mu\theta\delta^2)U_j^{n+1} = (1 + \mu(1-\theta)\delta^2)U_j^n. \tag{4.20}$$

We shall see that when $\theta \geqslant \frac{1}{2}$ the method is stable.

As before substitute $U_j^n = g^n \mathrm{e}^{ij\xi}$ in (4.20) then

$$(1 + 4\mu\theta\sin^2(\tfrac{\xi}{2}))g^{n+1} = (1 - 4\mu(1-\theta)\sin^2(\tfrac{\xi}{2}))g^n.$$

As we saw previously for stability we require that

$$\begin{aligned}
&-1 \leqslant \frac{1 - 4\mu(1-\theta)\sin^2(\frac{\xi}{2})}{1 + 4\mu\theta\sin^2(\frac{\xi}{2})} \leqslant 1 \\
&\iff -(1 + 4\mu\theta\sin^2(\tfrac{\xi}{2})) \leqslant 1 - 4\mu(1-\theta)\sin^2(\tfrac{\xi}{2}) \leqslant 1 + 4\mu\theta\sin^2(\tfrac{\xi}{2}) \\
&\iff 4\mu\theta\sin^2(\tfrac{\xi}{2})(1-2\theta) \leqslant 2 \text{ and } -4\mu\theta\sin^2(\tfrac{\xi}{2}) \leqslant 0.
\end{aligned}$$

Since μ is fixed the second inequality is always true. If $\theta \geqslant \frac{1}{2}$ then the first inequality is always true. If however $\theta < \frac{1}{2}$ then the method is stable if and only if $\mu \leqslant \frac{1}{2(1-2\theta)}$.

For the rest of this section we focus on the model hyperbolic problem $u_t + au_x = 0$.

We saw in Example 2.4 the problem described in Example 1.3 was unstable. To design a better scheme we note that on a $x-t$ grid information propagates along characteristics with positive gradient. This suggests the scheme

$$U_j^{n+1} = U_j^n - a\lambda(U_j^n - U_{j-1}^n). \tag{4.5}$$

If $a < 0$ we should use the scheme

$$U_j^{n+1} = U_j^n - a\lambda(U_{j+1}^n - U_j^n). \tag{4.6}$$

The Upwind scheme

For general a we can roll the previous two schemes together

$$U_j^{n+1} = U_j^n - \lambda\max\{a,0\}(U_j^n - U_{j-1}^n) - \lambda\min\{a,0\}(U_{j+1}^n - U_j^n). \tag{4.7}$$

This scheme is consistent with $T_j^n = O(k) + O(h)$. To check stability put $U_j^n = g^n \mathrm{e}^{ij\xi}$ then we find

$$|g|^2 = \begin{cases} 1 + 4a\lambda(1 + a\lambda)\sin^2\frac{\xi}{2} & \text{if } a < 0 \\ 1 - 4a\lambda(1 - a\lambda)\sin^2\frac{\xi}{2} & \text{if } a > 0 \end{cases}$$

and stability follows if $|a|\lambda \leqslant 1$. This is called the CFL condition (Courant, Friedrichs & Levy).

Dissipation

Preserving the quantitative feature of the solution to a PDE is desirable. If we suppose that the solution to $u_t + au_x = 0$ satisfies $u(x,t) \to 0$ as $|x| \to \infty$ for each fixed t then

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty} u^2\,\mathrm{d}x &= \lim_{L\to\infty}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_{-L}^{L} u^2(x,t)\,\mathrm{d}x = \lim_{L\to\infty}\int_{-L}^{L} u_t u\,\mathrm{d}x = -a\lim_{L\to\infty}\int_{-L}^{L} u_x u\,\mathrm{d}x \\ &= -a\frac{1}{2}\lim_{L\to\infty}\left[u^2(x,t)\right]_{x=-L}^{x=L} = 0 \Longrightarrow \int_{-\infty}^{\infty} u^2(x,t)\,\mathrm{d}x = \int_{-\infty}^{\infty}\left[u^0(x)\right]^2\,\mathrm{d}x \end{aligned}$$

we have we call this *non-dissipation*. We wish our numerical scheme to inherit this property. Recalling Theorem 2.2, if $|g| = 1$ then $\|\boldsymbol{U}^{n+1}\| = \|\boldsymbol{U}^n\|$ and the scheme is called non-dissipative. If $g < 1$ the scheme is called dissipative. Hence for the scheme just discussed we have non-dissipation if $|a|\lambda = 1$.

We would like $a = a(x)$ hence it is unlikely that non-dissipation can be achieved. Thus in general the upwind scheme may be considered to be dissipative (just choose λ small enough). Since the upwind scheme is dissipative and has truncation error $O(h) + O(k)$ we consider other schemes.

The Lax-Wendroff scheme

Assume a is constant and starting with the Taylor series

$$u(jh,(n+1)k) \approx u(jh,nk) + ku_t(jh,nk) + \frac{k^2}{2!}u_{tt}(jh,nk).$$

Now noting that

$$u_t(jh,nk) = -au_x(jh,nk) \approx -a\frac{u_{j+1}^n - u_{j-1}^n}{2h}$$

and

$$u_{tt}(jh,nk) = (-au_x(jh,nk))_t = a^2u_{xx}(jh,nk) \approx a^2\frac{\delta^2 u_j^n}{h^2} = a^2\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}.$$

our approximation becomes

$$U_j^{n+1} = U_j^n - \frac{a\lambda}{2}(U_{j+1}^n - U_{j-1}^n) + \frac{a^2\lambda^2}{2!}\delta^2 U_j^n.$$

It is easy to adapt this to variable a — see the problem sheet. We now consider the stability of this scheme. Assume that $U_j^n = g^n \mathrm{e}^{ij\xi}$ then

$$\begin{aligned} g &= 1 - \frac{a\lambda}{2}(\mathrm{e}^{i\xi} - \mathrm{e}^{-i\xi}) + \frac{a^2\lambda^2}{2!}(\mathrm{e}^{i\xi} - 2 + \mathrm{e}^{-i\xi}) \\ &= 1 - a\lambda i \sin\xi - 2a^2\lambda^2 \sin^2(\frac{\xi}{2}) \end{aligned}$$

Hence

$$\begin{aligned} |g|^2 &= (1 - 2a^2\lambda^2 \sin^2(\frac{\xi}{2}))^2 + a^2\lambda^2 \sin^2\xi \\ &= (1 - 2a^2\lambda^2 \sin^2(\frac{\xi}{2}))^2 + 4a^2\lambda^2 \sin^2\frac{\xi}{2}(1 - \sin^2\frac{\xi}{2}) \\ &= 1 - 4(1 - a^2\lambda^2)a^2\lambda^2 \sin^4(\frac{\xi}{2})) \end{aligned}$$

which is stable when $|a\lambda| \leqslant 1$. It has truncation error $O(k^2) + O(h^2)$, higher than for the upwind scheme. The dissipation in the Lax-Wendroff scheme is $4(1 - a^2\lambda^2)a^2\lambda^2 \sin^4(\frac{\xi}{2}))$ which is smaller than the dissipation for the upwind scheme $4a\lambda(1 - a\lambda)\sin^2\frac{\xi}{2}$.

Artificial Diffusion

$$\underbrace{U_j^{n+1} - U_j^n + \frac{a\lambda}{2}(U_{j+1}^n - U_{j-1}^n)}_{\text{A}} \underbrace{- \frac{a^2\lambda^2}{2}\delta^2 U_j^n}_{\text{B}} = 0. \tag{4.8}$$

A is unstable and B stabilizes the problem, c.f. Problem 3a.

Finite Volume

We first revisit some key points from vector calculus before introducing the Finite Element Method for two space dimensions.

Definition 4.2
If $V : \mathbb{R}^2 \to \mathbb{R}$ and $\boldsymbol{F} : \mathbb{R}^2 \to \mathbb{R}^2$ then

$$\nabla V = \left(\frac{\partial V}{\partial x_1}, \frac{\partial V}{\partial x_2}\right) \quad \textit{and} \quad \nabla \cdot \boldsymbol{F} = \left(\frac{\partial F_1}{\partial x_1}, \frac{\partial F_2}{\partial x_2}\right) \tag{4.9}$$

are called the gradient and divergence respectively.

Theorem 4.3 (Divergence)
If $\boldsymbol{F} : \Omega \to \mathbb{R}^2$ is a sufficiently differentiable vector valued function then

$$\int_\Omega (\nabla \cdot \boldsymbol{F})(\boldsymbol{x})\, dA = \int_\Gamma \boldsymbol{F}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) ds \tag{4.10}$$

where $\boldsymbol{n}$ is the unit normal pointing out of Ω.

A physical interpretation of the Divergence Theorem is if $\boldsymbol{F}$ is the velocity of a fluid then the left-hand integral is the amount of fluid lost from Ω, while the right-hand integral is the amount of fluid which has crossed Γ, i.e. this is a conservation law.

The Divergence Theorem in two-dimensions:

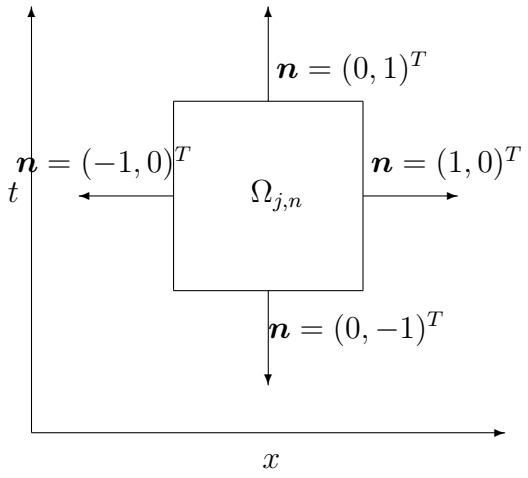
$$\iint_D (\nabla \cdot \boldsymbol{F})\,\mathrm{d}A = \int_{\partial D} \boldsymbol{F} \cdot \boldsymbol{n}\,\mathrm{d}s$$

is Green's theorem, so they say! Let the parameterisation be given by $\boldsymbol{r} = (x(s), y(s))^T$ where s is the arc-length ($|\boldsymbol{r}'| = 1$), then the unit tangent vector is (x', y') and so the unit outward pointing normal is $(y', -x')\mathrm{d}s = (\mathrm{d}y, -\mathrm{d}x)$ and hence from the Divergence theorem

$$\begin{aligned}\iint_D \left(\frac{\partial g}{\partial x} + \frac{\partial f}{\partial y}\right)\mathrm{d}\sigma &= \iint_D \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)\cdot(g, f)\mathrm{d}\sigma = \int_{\partial D}(g, f)\cdot(\mathrm{d}y, -\mathrm{d}x) \\ &= \int_{\partial D}(-f\,\mathrm{d}x + g\,\mathrm{d}y)\end{aligned}$$

where ∂D is the boundary of D described anti-clockwise, this is Green's theorem

We lay a grid on the solution region and then look at each grid cell and ask that the differential equation holds in some average sense over the cell. For example, let be a constant on an individual cell $\Omega_{j,n} = (x_j, x_{j+1}) \times (t_n, t_{n+1})$ then using the divergence theorem



$$\begin{aligned}0 &= \int_{\Omega_{j,n}} \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\mathrm{d}x\mathrm{d}t = \int_{\Omega_{j,n}} (\frac{\partial}{\partial x}, \frac{\partial}{\partial t}).(au, u)\mathrm{d}x\mathrm{d}t = \int_{\partial\Omega_{j,n}} (au, u).\boldsymbol{n}\mathrm{d}s \\ &= -a\int_{t_{n+1}}^{t_n} u(x_j, t)\,[-\mathrm{d}t] - \int_{x_j}^{x_{j+1}} u(x, t_n)\mathrm{d}x \\ &\quad +a\int_{t_n}^{t_{n+1}} u(x_{j+1}, t)\mathrm{d}t + \int_{x_{j+1}}^{x_j} u(x, t_{n+1})\,[-\mathrm{d}x]\,.\end{aligned}$$

We then approximate each of these integrals using the midpoint rule

$$\begin{aligned} 0 \;\approx\; & -\frac{ak}{2}\left[u(x_j,t_n)+u(x_j,t_{n+1})\right]-\frac{h}{2}\left[u(x_j,t_n)+u(x_{j+1},t_n)\right] \\ & +\frac{ak}{2}\left[u(x_{j+1},t_n)+u(x_{j+1},t_{n+1})\right]+\frac{h}{2}\left[u(x_j,t_{n+1})+u(x_{j+1},t_{n+1})\right] \end{aligned}$$

which setting $U_j^n = u(x_j,t_n)$, replacing the $\approx$ by $=$, multiplying by $2/h$ leads to the *cell-vertex* method

$$0 = -(a\lambda+1)U_j^n + (-1+a\lambda)U_{j+1}^n + (1-a\lambda)U_j^{n+1} + (1+a\lambda)U_{j+1}^{n+1}. \tag{4.11}$$

Leap-frog scheme

$$U_j^{n+1} = U_j^{n-1} - \lambda\left(a_{j+1}U_{j+1}^n - a_{j-1}U_{j-1}^n\right) \tag{4.12}$$

is consistent and is stable (when the CFL condition and a is constant).

Fourier Analysis gives rigorous stability criteria in the absence of boundary conditions, only in special cases will it be rigorous when boundary conditions are present. Even so, it is commonly used with any old boundary conditions. We have focused on stability in the $\|\cdot\|_2$ of course there are other norms on S however in general these don't have nice properties.

5 Boundary conditions

Consider the heat equation, the convection-diffusion equation or the reaction diffusion equation

$$u_t = u_{xx}, \quad u_t = u_{xx} + au_x, \quad u_t = u_{xx} + f(u) \tag{5.1}$$

where $f(u)$ is some function of u, i.e. e^u, $u^3 - u$, etc. Each of these equations is to be supplemented by an initial condition and a boundary condition.

Now we consider adding some non-zero boundary conditions and the practicalities.

Example 5.1
Consider replacing $u(0,t) = 1$ in Example **??** *by the zero Neumann boundary condition*

$$u_x(0,t) = 0, \quad t > 0. \tag{5.2}$$

*in (***??***). To handle this case we introduce the dummy variable, U_{-1}^n, at each time level. So that in the finite difference discretisation we replace $U_0^n = 1$ by*

$$\frac{U_1^n - U_{-1}^n}{2h} = 0 \Longrightarrow U_{-1} = U_1^n. \tag{5.3}$$

The matrix formulation is exactly the same except for the first line. Set $\{\boldsymbol{U}^n\}_j = U_j^n$ $j = 0, \cdots J-1$ *then*

$$\begin{pmatrix} 1+2\mu & -2\mu & 0 & \cdots & 0 \\ -\mu & 1+2\mu & -\mu & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\mu & 1+2\mu & -\mu \\ 0 & \cdots & 0 & -\mu & 1+2\mu \end{pmatrix} \boldsymbol{U}^{n+1} = \boldsymbol{U}^n + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \mu \times 0 \end{pmatrix}.$$

5.1 The maximum principle

The previous two examples are consistent and stable. A Fourier analysis ignores the boundary conditions. However, we have a direct proof of stability based on the *maximum principle*. We review the maximum principle for the previous example.

Let $C = \max\{0, \max_{x\in[0,1]} u^0(x)\}$. Multiply by $\max\{u(x,t) - C, 0\}$ then on noting that

$$\begin{aligned} \int_0^1 u_t \max\{u(x,t) - C, 0\}\,\mathrm{d}x &= \int_{x\in[0,1]:u(x,t)<C} (u-C)_t (u(x,t)-C)\,\mathrm{d}x \\ = \int_{x\in[0,1]:u(x,t)<C} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(u-C)^2\,\mathrm{d}x &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \left[\max\{u(x,t)-C,0\}\right]^2\,\mathrm{d}x \end{aligned}$$

Also

$$\begin{aligned} &\int_0^1 u_{xx} \max\{u(x,t) - C, 0\}\,\mathrm{d}x \\ &\quad = \left[u_x \max\{u(x,t)-C,0\}\right]_0^1 - \int_0^1 \left[\max\{u(x,t)-C,0\}\right]_x^2\,\mathrm{d}x \\ &\quad = -\int_0^1 \left[\max\{u(x,t)-C,0\}\right]_x^2\,\mathrm{d}x \end{aligned}$$

Hence,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \left[\max\{u(x,t)-C,0\}\right]^2\,\mathrm{d}x + \int_0^1 \left[\max\{u(x,t)-C,0\}\right]_x^2\,\mathrm{d}x = 0$$

so that integrating over $(0,t)$ and noting that $u(x,0) \leqslant C$ it follows that

$$\begin{aligned} 0 \leqslant &\frac{1}{2}\int_0^1 \left[\max\{u(x,t)-C,0\}\right]^2\,\mathrm{d}x \\ \leqslant\ &\frac{1}{2}\int_0^1 \left[\max\{u(x,t)-C,0\}\right]^2\,\mathrm{d}x - \frac{1}{2}\int_0^1 \left[\max\{u(x,0)-C,0\}\right]^2\,\mathrm{d}x \\ &+ \int_0^t \int_0^1 \left[\max\{u(x,s)-C,0\}\right]_x^2\,\mathrm{d}x\,\mathrm{d}s = 0 \end{aligned}$$

and so we conclude that

$$[\max\{u(x,t)-C,0\}]^2 \equiv 0 \Longleftrightarrow u(x,t)\leqslant C.$$

Similarly (multiplying by -1), it is possible to show that

$$-u(x,t)\leqslant \max\{0, \max_{x\in[0,1]} -u^0(x)\} \Longleftrightarrow u(x,t)\geqslant \min\{0, \min_{x\in[0,1]} u^0(x)\}$$

Hence, $u(x,t)$ is bounded above and below by the extremes attained by the initial data and boundary values.

Let us revisit Example 3.2. The j'th equation is

$$(1+2\mu)U_j^{n+1} = \begin{cases} U_0^n + 2\mu U_1^{n+1} & \text{if } j=0 \\ \mu U_{j-1}^{n+1} + U_j^n + \mu U_{j+1}^{n+1} & \text{if } j=1,\cdots J-1 \end{cases}.$$

Let $|U_j^{n+1}| = \|\boldsymbol{U}^{n+1}\|_\infty$ and for simplicity assume that $1\leqslant j\leqslant J-1$. Then since $1+2\mu>0$

$$\begin{aligned} (1+2\mu)\|\boldsymbol{U}^{n+1}\|_\infty &= |(1+2\mu)U_j^{n+1}| = |\mu U_{j-1}^{n+1} + U_j^n + \mu U_{j+1}^{n+1}| \\ &\leqslant \mu\|\boldsymbol{U}^{n+1}\|_\infty + \|\boldsymbol{U}^n\|_\infty + \mu\|\boldsymbol{U}^{n+1}\|_\infty \end{aligned}$$

or

$$\|\boldsymbol{U}^{n+1}\|_\infty \leqslant \|\boldsymbol{U}^n\|_\infty \leqslant \|\boldsymbol{U}^0\|_\infty$$

for all $n\geqslant 0$ and we have stability with respect to the $\|\cdot\|_\infty$ norm. Alternatively, taking $U_{j^\star}^{n+1} = \max_{j=0,\cdots J-1} U_j^{n+1}$ you can show that $\max_j U_j^{n+1} \leqslant \max_j U_j^n$ and similarly $\min_j U_j^{n+1} \geqslant \min_j U_j^n$. To show that the scheme is convergent, we also have to show it is consistent. The truncation error is

$$T_j^n = \frac{u_j^{n+1}-u_j^n}{k} - \begin{cases} (2u_1^{n+1} - 2u_0^{n+1})/h^2 & \text{if } j=0 \\ (u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1})/h^2 & \text{if } j=1,\cdots J-1 \end{cases}.$$

It is straightforward to show that (See the problem sheet)

$$T_j^n = O(k) + O(h^2) \quad j=1,\cdots J-1$$

We now turn to the case where $j=0$. Noting that $u_x(0,t)=0$

$$u_{xt}(0,kn) = \lim_{k'\to 0} \frac{u_x(0,kn+k') - u_x(0,kn)}{k'} = 0,$$

similarly for all higher order derivatives in t

$$u_{xxx} = u_{xt} = \frac{u_x(0,k(n+1)) - u_x(0,kn)}{k} + O(k) = O(k)$$

and

$$\begin{aligned} & u_1^{n+1} - u_0^{n+1} \\ &= hu_x + \frac{1}{2}(h^2u_{xx} + 2hku_{xt}) + \frac{1}{3}(h^3u_{xxx} + 3h^2ku_{xxt} + 3hk^2u_{xtt}) + O(h^4) + O(k^4) \\ &= \frac{1}{2}h^2u_{xx} + \frac{1}{3}(h^3u_{xxx} + 3h^2ku_{xxt}) + O(h^4) + O(k^4), \end{aligned}$$

so that

$$T_j^n = \frac{u_0^{n+1} - u_0^n}{k} - 2\frac{u_1^{n+1} - u_0^{n+1}}{h^2} = u_t + O(k) - u_{xx} + O(h^2) + O(k) = O(k) + O(h^2)$$

Suppose we want to solve $u_t + au_x = 0$ on $(0,1)$ where $a > 0$ is constant with $u(x,0) = u_0(x)$ and boundary condition

$$u(0,t) = v(t), \quad t > 0 \tag{5.4}$$

We will use a maximum principle argument to show that the upwind scheme is convergent in the presence of boundary conditions if the CFL condition ($a\lambda \leqslant 1$) holds.

The upwind scheme with appropriate boundary condition is for $j = 1, \cdots J$

$$U_j^{n+1} = U_j^n - a\lambda(U_j^n - U_{j-1}^n) \text{ with } U_j^0 = u^0(jh) \text{ and } U_0^n = v(nk).$$

Rewriting the difference scheme for $j = 1, \cdots, J$, taking the modulus and using the CFL condition $a\lambda \leqslant 1$

$$\begin{aligned} |U_j^{n+1}| &= |(1-a\lambda)U_j^n + a\lambda U_{j-1}^n| \leqslant (1-a\lambda)|U_j^n| + a\lambda|U_{j-1}^n| \\ &\leqslant (1-a\lambda)\max_{j=0,\cdots J}|U_j^n| + a\lambda \max_{j=0,\cdots J}|U_j^n| = \max_{j=0,\cdots J}|U_j^n|. \end{aligned}$$

Hence, taking the maximum over $j = 1, \cdots, J$ and noting that $|U_0^{n+1}| = |v((n+1)k)|$ we conclude that

$$\max_{j=0,\cdots J}|U_j^{n+1}| \leqslant \max\{\max_{j=0,\cdots J}|U_j^n|, |v((n+1)k)|\}$$

$$\Longrightarrow \max_{j=0,\cdots J}|U_j^n| \leqslant \max\{\max_{j=0,\cdots J}|U_j^0|, \max_{n=0,\cdots N}|v(nk)|\}$$

that is L^∞-stable and hence we conclude convergence.

6 Finite Element Methods

6.1 Introduction

So far we have only considered PDE's in one space dimension. Physical bodies are not one dimensional, and usually we want to solve one or two dimensional problems.

The two dimensional analogue of the Heat equation is

$$\frac{\partial u}{\partial t} - \nabla \cdot (c\nabla u) = f. \tag{6.1}$$

here, the conductivity $c = c(x)$ is taken to be a scalar function. The equation (6.1) may be solved for $u(\boldsymbol{x}, t)$ for $t > 0$ and $\boldsymbol{x} \in \Omega$ (Ω some bounded domain of $\mathbb{R}^2$ subject to initial

conditions $u(\boldsymbol{x}, 0) = u^0(\boldsymbol{x})$ and boundary conditions on $\Gamma = \partial\Omega$, the boundary of Ω). We will also consider the steady state of u, i.e. where $\frac{\partial u}{\partial t} = 0$, then

$$-\nabla \cdot (c\nabla u) = f \quad \text{on } \Omega. \tag{6.2}$$

Suppose we scale c so that $c \equiv 1$, then we get

$$-\Delta u \equiv -\nabla^2 u = -\nabla \cdot (\nabla u) = f \quad \text{on } \Omega \quad \text{or} \quad -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f. \tag{6.3}$$

This is called *Poisson's equations*. We accompany (6.2) by boundary conditions, for example

$$\text{Dirichlet} \qquad u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_D \tag{6.4}$$

$$\text{Neumann} \qquad \frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{x}) = \tilde{g}(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_N \tag{6.5}$$

where $\Gamma = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, g, and $\tilde{g}$ are given and $\frac{\partial u}{\partial \boldsymbol{n}}$ is the *normal derivative* of u on Γ, that is

$$\frac{\partial u}{\partial \boldsymbol{n}} = \nabla u(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) \tag{6.6}$$

where $\boldsymbol{n}(\boldsymbol{x})$ is the unit normal pointing outwards from Ω.

How do we approximate (6.2)?

If we start on a simple domain, i.e. $\Omega = [0,1] \times [0,1]$ with a uniform grid h in both directions, set $x_j = jh$, write $u_{j,k} = u(x_j, x_k)$ then we can approximate $-\nabla^2 u = f$ by a finite difference scheme

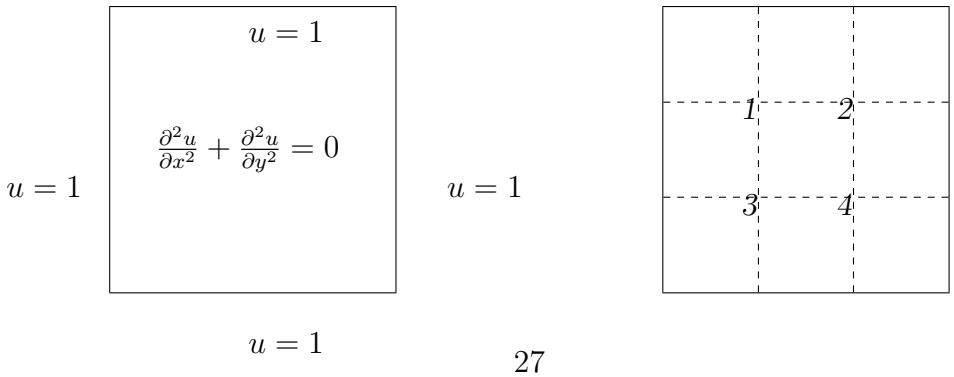
$$-\left\{\frac{U_{j+1,k} - 2U_{j,k} + U_{j-1,k}}{h^2}\right\} - \left\{\frac{U_{j,k+1} - 2U_{j,k} + U_{j,k-1}}{h^2}\right\} = f_{j,k}. \tag{6.7a}$$

If we have Dirichlet boundary condition (6.4), then

$$U_{j,k} = g_{j,k} \quad \text{when } (x_j, x_k) \in \Gamma \tag{6.7b}$$

and (6.7a) consists of a $(J-1)^2 \times (J-1)^2$ system of equations for the unknowns $U_{j,k}$.

Example 6.1
Set $h = 1/3$. At the point labelled 1



the following difference equation holds

$$\frac{u_2 - 2u_1 + 1}{(1/3)^2} + \frac{1 - 2u_1 + u_3}{(1/3)^2} = 0.$$

After some algebra the following equations hold:

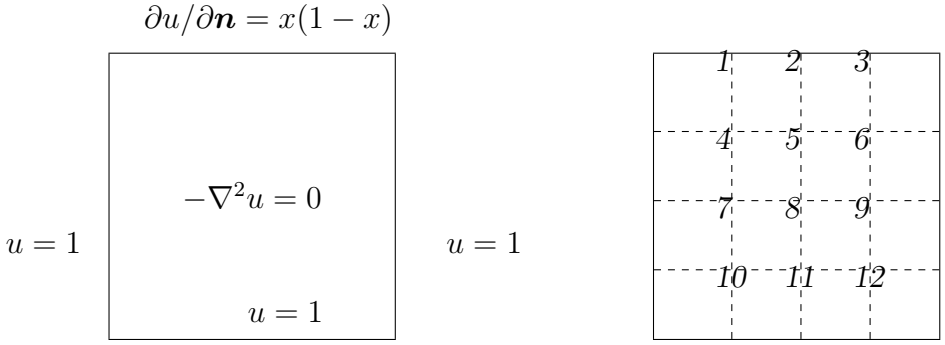
$$\begin{array}{rcl} u_2 - 4u_1 + u_3 + 2 &=& 0 \\ u_1 - 4u_2 + u_4 + 2 &=& 0 \\ u_1 - 4u_3 + u_4 + 2 &=& 0 \\ u_2 - 4u_4 + u_3 + 2 &=& 0 \end{array} \iff \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \end{pmatrix} \quad \textit{i.e. } A\boldsymbol{u} = \boldsymbol{2}$$

A is diagonally dominant and therefore invertible. The solution is given by

$$u_1 = u_2 = u_3 = u_4 = 1.$$

Using the method of separation of variables, it is easy to see that the exact solution is given by $u(x, y) = 1$.

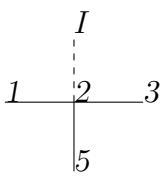
Example 6.2
Suppose we wish to approximate the problem



At internal points we use the five point difference operator, e.g. the point 8 gives us the difference equation:

$$\frac{u_9 - 2u_8 - u_7}{h^2} + \frac{u_5 - 2u_8 + u_{11}}{h^2} = 0.$$

To apply the boundary condition we introduce an imaginary point



then whatever the value of u at I is, we approximate it by the following equation:

$$\frac{1}{h^2}\left(u_I + u_1 + u_3 + u_5 - 4u_2\right) = 0$$

but also

$$x(1-x) = \frac{\partial u}{\partial \boldsymbol{n}} \approx \frac{u_I - u_5}{2h},$$

so taking $u_I = u_5$ yields

$$\frac{1}{h^2}\left(2u_5 + u_1 + u_3 + 2hx(1-x) - 4u_2\right) = 0.$$

We can repeat this for the points 1 and 3 to give us a matrix which is diagonally dominant and therefore invertible.

Notice that if we do not have a square, or some simple geometric shape for Ω we have to do considerable tinkering to adapt the finite difference method.

6.2 Piecewise Linear Finite Elements on non-uniform Intervals

Consider

$$-u'' = f, \quad \text{on } (0,1), \quad u'(0) = u(1) = 0.$$

Multiply this equation by some

$$v \in V = \left\{ v : \int_0^1 (v(x))^2 \,\mathrm{d}x + \int_0^1 (v'(x))^2 \,\mathrm{d}x \leqslant C, v(1) = 0 \right\}$$

then using integration by parts

$$\int_0^1 f \times v \,\mathrm{d}x = \int_0^1 [-u'']v \,\mathrm{d}x = [-u'v]_0^1 \,\mathrm{d}x + \int_0^1 u'v' \,\mathrm{d}x = \int_0^1 u'v' \,\mathrm{d}x.$$

This leads to the *weak formulation*: Find $u \in V$ such that

$$a(u,v) = \int_0^1 u'v' \,\mathrm{d}x = \int_0^1 fv \,\mathrm{d}x = L(v), \quad \forall v \in V.$$

Note that this has the important advantage that only first derivatives are required.

Divide $(0,1)$ into a *mesh* of J non-overlapping intervals, $\tau = (x_i, x_{i+1})$, and call this the *triangulation*, $\mathcal{T}^h$, of $(0,1)$. The vertices $x_0, \cdots x_J$ are called the *nodes*. For any interval $\tau \in \mathcal{T}^h$ define the local mesh parameter

$$h_\tau = \max\{|x-y| : x, y, \in \tau\} = x_{i+1} - x_i$$

and define $h = \max_\tau h_\tau$. h is a measure of the "fineness" of the triangulation. A function $v^h : (0,1) \to \mathbb{R}$ is called piecewise linear on $(0,1)$ if

$$v^h(x) = a + bx \quad \text{for } x \in \tau.$$

From this we define the space V^h

$$V^h = \left\{ v^h : [0,1] \to \mathbb{R},\ v^h \text{ is continuous on } (0,1), v^h \text{ is piecewise linear, } v^h(1) = 0 \right\}. \tag{6.8}$$

Since the hat functions

$$\phi_j(x) = \begin{cases} (x - x_{j-1})/(x_j - x_{j-1}) & \text{if } \in (x_{j-1}, x_j) \\ (x_{j+1} - x)/(x_{j+1} - x_j) & \text{if } \in (x_j, x_{j+1}) \in V^h \\ 0 & \text{otherwise} \end{cases}$$

form a basis for the piecewise linear function, where $\phi_j(x_i) = \delta_{ij}$, any function $v^h \in V^h$ may be written uniquely as

$$v^h(x) = \sum_{i=0}^{J-1} V_i \phi_i(x).$$

Note that $v^h(x_j) = V_j$. The *Piecewise Linear Finite Element Method* is: Find $u^h \in V^h$ such that

$$a(u^h, v^h) = L(v^h) \qquad \forall\ v^h \in V^h. \tag{6.9}$$

Noting that $u^h = \sum_{j=0}^{J-1} U_j \phi_j$ and taking $v^h = \phi_i\ i = 0, \cdots J-1$

$$\sum_{j=0}^{J-1} a(\phi_i, \phi_j) U_j = L(\phi_i), \quad i = 0, \cdots J-1$$

where we note the linearity of $a(\cdot, \cdot)$ in both variables. This is a $J \times J$ linear system

$$A\boldsymbol{U} = \boldsymbol{f} \tag{6.10}$$

where $A_{ij} = a(\phi_i, \phi_j)$ and $f_i = L(\phi_i)$.

Notice that to implement (6.10) requires us to calculate the *stiffness matrix* for (6.9)

$$A_{ij} = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}x}\phi_i(x)\, \frac{\mathrm{d}}{\mathrm{d}x}\phi_j(x)\, \mathrm{d}x$$

and this will vanish unless the regions on which ϕ_i and ϕ_j are non-zero overlap (i.e. x_i and x_j share an interval). Since $(0,1)$ is the union of the intervals, τ, in $\mathcal{T}^h$, we can write

$$A_{ij} = \sum_{\tau \in \mathcal{T}^h} A_{ij}^\tau \quad \text{where } A_{ij}^\tau = \int_\tau \frac{\mathrm{d}}{\mathrm{d}x}\phi_i^\tau(x)\, \frac{\mathrm{d}}{\mathrm{d}x}\phi_j^\tau(x)\, \mathrm{d}x, \tag{6.11}$$

and ϕ_i^τ denotes the restriction of ϕ_i to τ. Let the nodes of the interval τ be labelled x_i and x_{i+1}, ordered from left to right, then observe that

$$\phi_i^\tau(x) = \frac{1}{|\tau|} \begin{vmatrix} 1 & x \\ 1 & x_{i+1} \end{vmatrix}$$

is linear and satisfies $\phi_i^\tau(x_j) = \delta_{ij}$. That is ϕ_i^τ is the restriction of the global basis function centred at x_i to τ. Also

$$\phi_{i+1}^\tau(x) = \frac{1}{|\tau|}\begin{vmatrix} 1 & x_i \\ 1 & x \end{vmatrix}.$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}x}\phi_i^\tau(x) = -\frac{1}{|\tau|} \text{ and } \frac{\mathrm{d}}{\mathrm{d}x}\phi_{i+1}^\tau(x) = \frac{1}{|\tau|}$$

and each of these is constant

$$A_{ij}^\tau = \int_\tau \frac{\mathrm{d}}{\mathrm{d}x}\phi_i^\tau(x)\,\frac{\mathrm{d}}{\mathrm{d}x}\phi_j^\tau(x)\,\mathrm{d}x$$

These are stored in the element matrix

$$\begin{pmatrix} A_{i,i}^\tau & A_{i,i+1}^\tau \\ A_{i+1,i}^\tau & A_{i,i+1}^\tau \end{pmatrix} = \begin{pmatrix} \frac{1}{|\tau|} & -\frac{1}{|\tau|} \\ -\frac{1}{|\tau|} & \frac{1}{|\tau|}. \end{pmatrix}$$

Note that all other elements of A^τ, the *element stiffness matrix*, are zero and A^τ is symmetric. Now we can assemble the matrix A where

$$A_{ij} = \sum_{\tau\in\mathcal{T}^h} A_{ij}^\tau$$

Example 6.3
If we have four intervals and we label the points sequentially $0, 1, 2, 3, 4$ *then*

$$A = \begin{pmatrix} \frac{1}{h_1} & -\frac{1}{h_1} & 0 & 0 \\ -\frac{1}{h_1} & \frac{1}{h_1}+\frac{1}{h_2} & -\frac{1}{h_2} & 0 \\ 0 & -\frac{1}{h_2} & \frac{1}{h_2}+\frac{1}{h_3} & -\frac{1}{h_3} \\ 0 & 0 & -\frac{1}{h_3} & \frac{1}{h_3}+\frac{1}{h_4} \end{pmatrix}$$

If we label the points sequentially $5, 1, 4, 2, 3$ *then*

$$A = \begin{pmatrix} \frac{1}{h_1}+\frac{1}{h_2} & 0 & -\frac{1}{h_2} & -\frac{1}{h_1} \\ 0 & \frac{1}{h_3}+\frac{1}{h_4} & -\frac{1}{h_3} & 0 \\ -\frac{1}{h_2} & -\frac{1}{h_3} & \frac{1}{h_2}+\frac{1}{h_3} & 0 \\ -\frac{1}{h_1} & 0 & 0 & \frac{1}{h_1} \end{pmatrix}$$

Reviewing how we calculate A_{ij}^τ, it will be advantageous to perform a change of variable in two dimensions, so lets do it in one dimension. Let us transform the interval $[x_i, x_{i+1}]$ on to the reference interval $\widehat{\tau} \equiv [0, 1]$. The appropriate transformation is

$$x = x_i(1-\xi) + x_{i+1}\xi \quad x \in [x_i, x_{i+1}],\ \xi \in [0, 1]$$

Then defining $\widehat{\phi}_i(\xi) = \phi_i^\tau(x)$, noting $\frac{\mathrm{d}}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}\xi}{\mathrm{d}x}$ and $\frac{\mathrm{d}\xi}{\mathrm{d}x} = \frac{1}{h_i}$ we find that when $\tau = [x_i, x_{i+1}]$

$$A_{ij}^\tau = \int_\tau \frac{\mathrm{d}}{\mathrm{d}x}\phi_i^\tau(x)\,\frac{\mathrm{d}}{\mathrm{d}x}\phi_j^\tau(x)\,\mathrm{d}x = \frac{1}{h_i}\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_i(\xi)\frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_j(x)\mathrm{d}\xi.$$

Notice that

$$\widehat{\phi}_i = 1-\xi,\ \widehat{\phi}_{i+1} = \xi \Longrightarrow \frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_i = -1,\ \frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_{i+1} = 1$$

then

$$\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_i(\xi)\frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_i(x)\mathrm{d}\xi = 1,\ \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_i(\xi)\frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_{1+1}(x)\mathrm{d}\xi = -1,$$
$$\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_{i+1}(\xi)\frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_i(x)\mathrm{d}\xi = -1,\ \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_{i+1}(\xi)\frac{\mathrm{d}}{\mathrm{d}\xi}\widehat{\phi}_{i+1}(x)\mathrm{d}\xi = 1,$$

and hence

$$\begin{pmatrix} A^\tau_{i,i} & A^\tau_{i,i+1} \\ A^\tau_{i+1,i} & A^\tau_{i,i+1} \end{pmatrix} = \frac{1}{h_i}\begin{pmatrix} 1 & -1 \\ -1 & 1. \end{pmatrix} = \begin{pmatrix} \frac{1}{|\tau|} & -\frac{1}{|\tau|} \\ -\frac{1}{|\tau|} & \frac{1}{|\tau|}. \end{pmatrix}$$

as before.

To compute $f_i = L(\phi_i) = \int_0^1 f(x)\phi_i\,\mathrm{d}x$ for any i we may wish to use numerical integration (for example the trapezium rule).

6.3 Piecewise Linear Finite Elements on Triangles

Suppose that $u, v : \mathbb{R}^2 \to \mathbb{R}$ and c are sufficiently differentiable functions then by setting $\boldsymbol{F} = vc\nabla u$ on noting that

$$\nabla\cdot(vc\nabla u) = \partial_i(vc\partial_i u) = c\partial_i v\partial_i u + v\partial_i(c\partial_i u) = c\nabla v\cdot\nabla u + v\nabla\cdot(c\nabla u)$$

we find from the Divergence Theorem

$$\iint_\Omega \left[c\nabla v\cdot\nabla u + v\nabla\cdot(c\nabla u)\right]\,\mathrm{d}A = \int_\Gamma vc\frac{\partial u}{\partial \boldsymbol{n}}\mathrm{d}s \tag{6.12}$$

which is just Green's Theorem and this is our starting point.

Example 6.4
Set $c \equiv 1$ in (6.2) with $\Omega = (0,1)\times(0,1)$, $g(x) \equiv 0$ on $\Gamma_D = \{0\}\times(0,1)\cup(0,1)\times\{0\}\cup\{1\}\times(0,1)$ in (6.4) and $\tilde{g}(x) \equiv x(1-x)$ on $\Gamma_N = (0,1)\times\{1\}$ in (6.5), i.e.

$$-\nabla^2 u = 2y \quad \textit{in } \Omega, \qquad u = 0 \quad \textit{on } \Gamma_D, \quad \frac{\partial u}{\partial \boldsymbol{n}} = x(1-x) \textit{on } \Gamma_N$$

(Exact solution: $u(x,y) = xy(1-x)$) then multiply by

$$v \in H^1_E(\Omega) = \left\{ v : \iint_\Omega (v(\boldsymbol{x}))^2\,dA + \iint_\Omega |\nabla v(\boldsymbol{x})|^2\,dA \leqslant C,\ v = 0 \textit{ on } \Gamma_D \right\},$$

integrate over Ω and use (6.12)

$$\iint_\Omega fv\,dA = -\iint_\Omega v\nabla\cdot(\nabla u)\,dA = -\int_\Gamma v\frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{x})\;ds + \iint_\Omega \nabla v\cdot\nabla u\,dA \tag{6.13}$$

then we obtain

$$a(u,v) = L(v) \quad \textit{where} \tag{6.14a}$$

$$a(u,v) = \iint_\Omega \nabla v\cdot\nabla u\,d\boldsymbol{x},\ \textit{and}\ L(v) = \iint_\Omega 2yv\,dA + \int_0^1 x(1-x)v(x,1)\,dx. \tag{6.14b}$$

This is called the weak formulation of the problem.

Note that (6.14a,b) has the important advantage over (6.13) that only first derivatives are required.

Here we restrict ourselves to the case when Ω is a polygon. Divide Ω into a *mesh* of non-overlapping triangles, τ, and call this the *triangulation*, $\mathcal{T}^h$, of Ω. The vertices are called the *nodes*. We assume that no node is an interior point of any edge. For any triangle $\tau\in\mathcal{T}^h$ define the local mesh parameter

$$h_\tau = \max\{|\boldsymbol{x}_1-\boldsymbol{x}_2| : \boldsymbol{x}_1,\boldsymbol{x}_2,\in\tau\} \quad \text{where } |\cdot| \text{ is the Euclidean length}$$

(h_τ is the longest side of τ) and define $h=\max_{\tau\in\mathcal{T}^h} h_\tau$. h is a measure of the "fineness" of the triangulation. A function $V:\Omega\to\mathbb{R}$ is called piecewise linear on $\mathcal{T}^h$ if

$$V(x,y) = a+bx+cy \quad \text{for } (x,y)\in\tau.$$

From this we define the space V^h

$$V^h = \left\{v^h:\Omega\to\mathbb{R},\ v^h \text{ is continuous on } \Omega, v^h \text{ is piecewise linear, } v^h=0 \text{ on } \Gamma\right\}. \tag{6.15}$$

Note that if you know $v^h\in V^h$ at the nodes on the mesh, then you know v^h everywhere on Ω.

Suppose we label the nodes of the triangulation $\boldsymbol{x}_1\to\boldsymbol{x}_J$. Then for any node $\boldsymbol{x}_i$ the *hat functions* $\phi_j\in V^h$ satisfy

$$\phi_j(\boldsymbol{x}_i) = \delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases}.$$

Note that $\text{span}\{\phi_i : i=1,\cdots J\}$ form a basis for V^h, hence any function $v^h\in V^h$ may be written uniquely as

$$v^h(x,y) = \sum_{i=1}^J V_i\phi_i(x,y) \text{ where } v(\boldsymbol{x}_j)=V_j.$$

Now we use V^h to approximate V in the weak form of our PDE.

Example 6.5
The Piecewise Linear Finite Element Method for Example 6.4 is: Find $u^h \in V^h$ such that

$$a(u^h, v^h) = L(v^h) \qquad \forall\ v^h \in V^h. \tag{6.16}$$

We proceed similarly to the previous subsection. Labelling all of the of the nodes of the mesh on interior and on Γ_N by $\boldsymbol{x}_1, \cdots \boldsymbol{x}_J$, setting $v^h = \phi_i$, noting $u^h = \sum_{j=1}^{J} U_j \phi_j$ and the symmetry and linearity of a we have (6.16) is equivalent to

$$\sum_{j=1}^{J} a(\phi_i, \phi_j) U_j = L(\phi_i), \quad i = 1, \cdots J.$$

This is a $J \times J$ linear system

$$A\boldsymbol{U} = \boldsymbol{f} \tag{6.17}$$

where $A_{ij} = a(\phi_i, \phi_j)$ and $f_i = L(\phi_i)$. Notice that to implement (6.17) requires us to calculate the stiffness matrix for (6.16)

$$A_{ij} = \iint_{\Omega} \nabla\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x})\, dA$$

and this will vanish unless the regions on which ϕ_i and ϕ_j are non-zero overlap (i.e. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ either share an edge or are coincident). Since Ω is the union of the triangles in $\mathcal{T}^h$, we can write

$$A_{ij} = \sum_{\tau \in \mathcal{T}^h} \iint_{\tau} \nabla\phi_i^\tau(\boldsymbol{x}) \cdot \nabla\phi_j^\tau(\boldsymbol{x})\, dA \tag{6.18}$$

where ϕ_i^τ denotes the restriction of ϕ_i to τ. If we define the element stiffness matrix for (6.16)

$$A_{ij}^\tau = \iint_{\tau} \nabla\phi_i^\tau(\boldsymbol{x}) \cdot \nabla\phi_j^\tau(\boldsymbol{x})\, dA$$

then this is zero except for a 3×3 sub-matrix corresponding to the nodes of τ. As we shall see, in practice a 3×3 version of A_{ij}^τ is assembled for each τ separately and then one assembles, A by

$$A_{ij} = \sum_{\tau \in \mathcal{T}^h} A_{ij}^\tau$$

Formulae for ϕ_i^τ and A_{ij}^τ

Let the nodes of the triangle τ be labelled $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$, ordered anti-clockwise, then observe that

$$\phi_i^\tau(x, y) = \frac{1}{2|\tau|} \begin{vmatrix} 1 & x & y \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}$$

is linear and satisfies $\phi_i^\tau(\boldsymbol{x}_j) = \delta_{ij}$. That is ϕ_i^τ is the restriction of the global basis function centred at $\boldsymbol{x}_i$ to τ. Since

$$\nabla\phi_i^\tau(\boldsymbol{x}) = \frac{1}{2|\tau|}\begin{pmatrix} -\begin{vmatrix} 1 & y_j \\ 1 & y_k \end{vmatrix} \\ \begin{vmatrix} 1 & x_j \\ 1 & x_k \end{vmatrix} \end{pmatrix} = \frac{1}{2|\tau|}\begin{pmatrix} y_j - y_k \\ -(x_j - x_k) \end{pmatrix} = \frac{1}{2|\tau|}\begin{pmatrix} d_{i2} \\ -d_{i1} \end{pmatrix}$$

and each of these are constant

$$A_{ij}^\tau = \iint_\tau \nabla\phi_i^\tau(\boldsymbol{x}) \cdot \nabla\phi_j^\tau(\boldsymbol{x})\,\mathrm{d}A = \frac{|\tau|}{4|\tau|^2}(d_{i1}d_{j1} + d_{i2}d_{j2})$$

Note that A^τ is symmetric. Also we could have made an affine transformation $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k$ maps to the triangle, $\widehat{\tau}$, with vertices, $(0,0)$, $(1,0)$, $(0,1)$

$$\boldsymbol{x} \mapsto \boldsymbol{\xi} = B(\boldsymbol{x} - \boldsymbol{x}_i)$$

where

$$B(\boldsymbol{x}_j - \boldsymbol{x}_i) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } B(\boldsymbol{x}_k - \boldsymbol{x}_i) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

or put alternatively

$$B\begin{pmatrix} \boldsymbol{x}_j - \boldsymbol{x}_i & \boldsymbol{x}_k - \boldsymbol{x}_i \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Notice that using elementary row operations

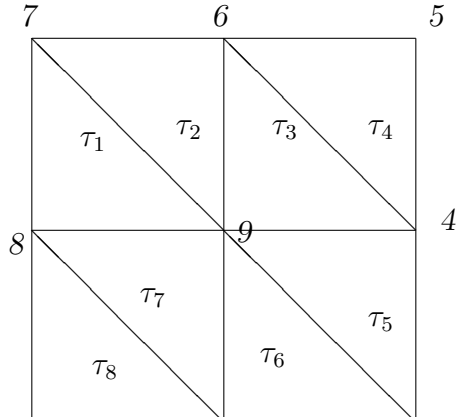
$$2|\tau| = \begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix} = \begin{vmatrix} \boldsymbol{x}_j - \boldsymbol{x}_i & \boldsymbol{x}_k - \boldsymbol{x}_i \end{vmatrix}$$

is non-zero and therefore B is uniquely defined, see the problem sheet. Then defining $\widehat{\phi}_i(\boldsymbol{\xi}) = \phi_i^\tau(\boldsymbol{x}) = \phi_i^\tau(B^{-1}\boldsymbol{\xi} + \boldsymbol{x}_i)$ we find that

$$A_{lm}^\tau = \iint_\tau \nabla\phi_l^\tau(\boldsymbol{x}) \cdot \nabla\phi_m^\tau(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \iint_{\widehat{\tau}} [B\nabla_{\boldsymbol{\xi}}\widehat{\phi}_l(\boldsymbol{\xi})] \cdot [B\nabla_{\boldsymbol{\xi}}\widehat{\phi}_m(x)]|B^{-1}|\mathrm{d}\boldsymbol{\xi}.$$

Note the Jacobian of the transformation is $|B^{-1}| = 2|\tau|$.

Example 6.6
We discretize $\Omega = (0,1)^2$ as follows



with $h = \frac{1}{2}$ *and* $|B^{-1}| = 2|\tau| = h^2$. *Notice that on the reference triangle*

$$\widehat{\phi}_1 = 1 - \xi_1 - \xi_2,\ \widehat{\phi}_2 = \xi_1, \widehat{\phi}_3 = \xi_2$$

$$\Longrightarrow \nabla_{\boldsymbol{\xi}}\widehat{\phi}_1 = \begin{pmatrix}-1\\-1\end{pmatrix},\ \nabla_{\boldsymbol{\xi}}\widehat{\phi}_2 = \begin{pmatrix}1\\0\end{pmatrix},\ \nabla_{\boldsymbol{\xi}}\widehat{\phi}_3 = \begin{pmatrix}0\\1\end{pmatrix}.$$

Noting that

$$B = \begin{pmatrix}\frac{1}{h} & 0\\ 0 & \frac{1}{h}\end{pmatrix} \text{ or } B = \begin{pmatrix}-\frac{1}{h} & 0\\ 0 & -\frac{1}{h}\end{pmatrix}$$

it follows that

$$A_{ij}^{\tau} = \iint\limits_{\widehat{\tau}} [B\nabla_{\boldsymbol{\xi}}\widehat{\phi}_i(\boldsymbol{\xi})] \cdot [B\nabla_{\boldsymbol{\xi}}\widehat{\phi}_j(x)]|B^{-1}|d\boldsymbol{\xi}. = \iint\limits_{\widehat{\tau}} \nabla_{\boldsymbol{\xi}}\widehat{\phi}_i(\boldsymbol{\xi}) \cdot \nabla_{\boldsymbol{\xi}}\widehat{\phi}_j(x)d\boldsymbol{\xi}$$

and

$$\iint\limits_{\widehat{\tau}} \nabla_{\boldsymbol{\xi}}\widehat{\phi}_1(\xi) \cdot \nabla_{\boldsymbol{\xi}}\widehat{\phi}_1(x)d\xi = \int_0^1 d\xi_2 \int_0^{\xi_2} 2\, d\xi_1 = 1$$

$$\iint\limits_{\widehat{\tau}} \nabla_{\boldsymbol{\xi}}\widehat{\phi}_1(\xi) \cdot \nabla_{\boldsymbol{\xi}}\widehat{\phi}_2(x)d\xi = \int_0^1 d\xi_2 \int_0^{\xi_2} -1\, d\xi_1 = -\frac{1}{2}$$

$$\iint\limits_{\widehat{\tau}} \nabla_{\boldsymbol{\xi}}\widehat{\phi}_1(\xi) \cdot \nabla_{\boldsymbol{\xi}}\widehat{\phi}_3(x)d\xi = \int_0^1 d\xi_2 \int_0^{\xi_2} -1\, d\xi_1 = -\frac{1}{2}$$

$$\iint\limits_{\widehat{\tau}} \nabla_{\boldsymbol{\xi}}\widehat{\phi}_2(\xi) \cdot \nabla_{\boldsymbol{\xi}}\widehat{\phi}_2(x)d\xi = \iint\limits_{\widehat{\tau}} \nabla_{\boldsymbol{\xi}}\widehat{\phi}_3(\xi) \cdot \nabla_{\boldsymbol{\xi}}\widehat{\phi}_3(x)d\xi = \int_0^1 d\xi_2 \int_0^{\xi_2} d\xi_1 = \frac{1}{2}$$

$$\iint\limits_{\widehat{\tau}} \nabla_{\boldsymbol{\xi}}\widehat{\phi}_2(\xi) \cdot \nabla_{\boldsymbol{\xi}}\widehat{\phi}_3(x)d\boldsymbol{\xi} = 0$$

Hence, the 3×3 *non-zero matrix entries in*

$$A_{ij}^{\tau} = \begin{pmatrix}1 & -\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2} & 0\\ -\frac{1}{2} & 0 & \frac{1}{2}\end{pmatrix}$$

as before.

On τ_1 *the points in anti-clockwise order starting from the right-angled corner are 8,9,7 then*

$$A = \begin{pmatrix}1+\frac{1}{2}+\frac{1}{2} & -\frac{1}{2}-\frac{1}{2}\\ -\frac{1}{2}-\frac{1}{2} & \frac{1}{2}+\frac{1}{2}+1+\frac{1}{2}+\frac{1}{2}+1\end{pmatrix}.$$

The complete list of entries for $(9, 9)$ *is*

$$A_{2,2}^{(1)} + A_{3,3}^{(2)} + A_{1,1}^{(3)} + A_{2,2}^{(5)} + A_{3,3}^{(6)} + A_{1,1}^{(7)}.$$

We note that the integration rule

$$\iint\limits_{\widehat{\tau}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx \frac{1}{6}\left[f(\frac{1}{2},0)+f(0,\frac{1}{2})+f(\frac{1}{2},\frac{1}{2})\right]$$

is exact for quadratics.

Note that

$$f_6 \;=\; L(\phi_6) = \iint\limits_{\Omega} 2y\phi_6(x,y)\,\mathrm{d}x\mathrm{d}y + \int_0^1 x(1-x)\phi_6(x,1)\,dx = \frac{5}{24}+\frac{5}{48} = \frac{5}{16}$$

and

$$f_9 = L(\phi_9) = \iint\limits_{\Omega} 2y\phi_6(x,y)\,\mathrm{d}x\mathrm{d}y = \frac{3}{8}$$

Hence the system of equations to solve is

$$\begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}\begin{pmatrix} U_6 \\ U_9 \end{pmatrix} = \begin{pmatrix} \frac{25}{96} \\ \frac{3}{8} \end{pmatrix} \Longrightarrow \begin{pmatrix} U_6 \\ U_9 \end{pmatrix} = \frac{1}{112}\begin{pmatrix} 26 \\ 17 \end{pmatrix} \approx \begin{pmatrix} 0.232 \\ 0.152 \end{pmatrix}$$

to 3 d.p. which compares with the exact answer of 0.25 *and* 0.125.

With $h = 1/3$ *and labelling* $(2/3,(4-i)/3)$ P_i $(i = 1 \to 3)$ *and* $(1/3,(i-3)/3)$ P_i $(i = 4 \to 6)$ *we find*

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & -\frac{1}{2} \\ -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & -1 & 4 & -1 & 0 & 0 \\ 0 & 0 & -1 & 4 & -1 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 \\ -\frac{1}{2} & 0 & 0 & 0 & -1 & 2 \end{pmatrix}.$$

Now

$$f_i = \iint\limits_{\Omega} 2y\phi_i(x,y)\,\mathrm{d}A \quad i = 2 \to 5, \qquad f_1 = \iint\limits_{\Omega} 2y\phi_1(x,y)\,\mathrm{d}A + \int_0^1 x(1-x)\phi_1(x,1)\,\mathrm{d}x$$

$$\textit{and} \quad f_6 = \iint\limits_{\Omega} 2y\phi_6(x,y)\,\mathrm{d}A + \int_0^1 x(1-x)\phi_6(x,1)\,\mathrm{d}x$$

So

$$f_1 = \frac{8}{81}+\frac{13}{324}+\frac{1}{36} = \frac{1}{6} = f_6, \qquad f_2 = f_5 = \frac{4}{27}, \quad f_3 = f_4 = \frac{2}{27}$$

which yields

$$\boldsymbol{U} = (0.205761, 0.141975, 0.072016, 0.072016, 0.141975, 0.205761)^T$$

the values of the exact solution at these points are:

$$\boldsymbol{u} = (0.222222, 0.148148, 0.074074, 0.074074, 0.148148, 0.222222)^T.$$

Variations on a theme

1. We could seek continuous piecewise quadratic polynomials

$$a + bx + cy + dxy + ex^2 + fy^2$$

as the polynomial approximation of the triangle τ.

2. Instead of using triangles, we could use quadrilateral elements to approximate the domain Ω and modify piecewise linear functions by bi-linear functions of the form:

$$a + bx + cy + dxy.$$

One could also change the polynomials from being piecewise bi-linear to piecewise bi-quadratic.

6.4 Elementary Error Analysis for Finite Element Methods

Theorem 6.1
Let a be an inner-product on V and $L : V \to \mathbb{R}$ be a linear operator. Let V^h be a finite dimensional subspace of V. Suppose u solves the problem:
Find $u \in V$ such that

$$a(u, v) = L(v) \quad \forall\ v \in V \tag{6.19}$$

then the problem:
Find $u^h \in V^h$ such that

$$a(u^h, v^h) = L(v^h) \quad \forall\ v^h \in V^h \tag{6.20}$$

has a unique solution, and

$$\|u - u^h\|_a = \min_{v^h \in V^h} \|u - v^h\|_a \tag{6.21}$$

where $\|\cdot\|_a^2 = a(\cdot, \cdot)$.

Note that u^h is the best approximation to u in V^h with respect to the norm $\|\cdot\|_a$.

Proof. Let $\phi_1, \cdots, \phi_J$ be a basis for V^h, so $u^h = \sum_{j=1}^J U_j \phi_j$ then noting linear property of the inner-product

$$b_i = L(\phi_i) = a(u^h, \phi_i) = a(\sum_{j=1}^J U_j \phi_j, \phi_i) = \sum_{j=1}^J a(\phi_i, \phi_j) U_j = \{A\boldsymbol{U}\}_i \tag{6.22}$$

where $A_{ij} = a(\phi_j, \phi_i)$ and $\{\boldsymbol{U}\}_j = U_j$.

Since $A_{ij} = a(\phi_j, \phi_i) = a(\phi_i, \phi_j) = A_{ji}$, then A is symmetric. Now, noting the definition of the inner-product a

$$\boldsymbol{U}^T A \boldsymbol{U} = a(u^h, u^h) \geqslant 0$$

and is equal to zero only when $u^h \equiv 0 \Longleftrightarrow \boldsymbol{U} = 0$ implies that A is positive definite. Hence (6.22) has a unique solution.

We now turn our attention to proving (6.21). On noting that $a(u-u^h, v^h) = a(u, v^h) - a(u^h, v^h) = L(v^h) - L(v^h) = 0$ and the Cauchy-Schwarz inequality

$$\begin{aligned} \|u-u^h\|_a^2 &= a(u-u^h, u-u^h) = a(u-u^h, u) - a(u-u^h, u^h) \\ &= a(u-u^h, u) - a(u-u^h, v^h) = a(u-u^h, u-v^h) \\ &= \|u-u^h\|_a \|u-v^h\|_a \end{aligned}$$

the result follows on dividing by $\|u-u^h\|_a$.

Example 6.7
We now apply this abstract theory to

$$-\nabla^2 u = 2y \quad \text{on } \Omega \text{ which is polygonal} \tag{6.23a}$$

$$u = 0 \quad \text{on } \Gamma_D, \tag{6.23b}$$

$$\tfrac{\partial u}{\partial \boldsymbol{n}} = x(1-x) \quad \text{on } \Gamma_N, \tag{6.23c}$$

The weak from is

$$a(u,v) = \iint_\Omega \nabla u \cdot \nabla v \, dA = L(v). \tag{6.24}$$

Clearly $a(\cdot,\cdot)$ is symmetric and linear. We need to check the final property of an inner-product.

$$a(u,u) = \iint_\Omega |\nabla u|^2 \, dA \geqslant 0. \tag{6.25}$$

Intuitively $a(u,u) = 0$ must mean that $|\nabla u| = 0$, i.e. u is a constant. However, noting the boundary conditions $u = 0$ on Γ_D this must mean that the constant is zero, i.e. $u = 0$. Using this definition then $a(\cdot,\cdot)$ and $L(\cdot)$ satisfy the assumptions of Theorem 6.1 with $V \equiv H_E^1(\Omega)$. Moreover with V^h taken to be piecewise linears which vanish on Γ_D then $V^h \subset H_E^1(\Omega)$. Hence the solution u^h satisfies

$$\iint_\Omega |\nabla(u-u^h)|^2 \, dA \leqslant \iint_\Omega |\nabla(u-v^h)|^2 \, dA \quad \forall \; v^h \in V^h. \tag{6.26}$$

One function which we hope will be close to u^h is $\pi^h u \in V^h$ where

$$\pi^h u(\boldsymbol{x}_i) = u(\boldsymbol{x}_i) \quad \text{where } \boldsymbol{x}_i \in \tau \text{ are nodes } \forall \; \tau \in \mathcal{T}^h.$$

Definition 6.1
Let $v : \Omega \to \mathbb{R}$, then we define the norms

$$\|v\|_\infty = \max_{\boldsymbol{x}\in\Omega} |v(\boldsymbol{x})|, \quad \|v\|_{\infty,\tau} = \max_{\boldsymbol{x}\in\tau} |v(\boldsymbol{x})|. \tag{6.27}$$

Lemma 6.1
Assume that $\max_{j=1,2} \|\frac{\partial v}{\partial x_j}\|_\infty \leqslant C$ *then*

$$\|v - \pi^h v\|_\infty \leqslant Ch.$$

Furthermore, if $\max_{i,j=1,2} \|\frac{\partial^2 v}{\partial x_i \partial x_j}\|_\infty \leqslant C$ *and* $|\tau| \geqslant Ch_\tau^2$ *then*

$$\|\nabla v - \nabla \pi^h v\|_\infty \leqslant Ch$$

PROOF. Let us restrict ourselves to the triangle, τ, and label the nodes in an anti-clockwise direction $1, 2, 3$, then

$$\pi^h v(\boldsymbol{x}) = \sum_{i=1}^{3} v(\boldsymbol{x}_i)\phi_i(\boldsymbol{x}) \quad \text{and} \quad v(\boldsymbol{x}) = \sum_{i=1}^{3} v(\boldsymbol{x})\phi_i(\boldsymbol{x}).$$

Hence

$$\|v - \pi^h v\|_{\infty,\tau} \leqslant \sum_{i=1}^{3} \|v(\boldsymbol{x}) - v(\boldsymbol{x}_i)\|_{\infty,\tau}.$$

Now using Taylor's theorem

$$v(\boldsymbol{x}) = v(\boldsymbol{x}_i) + (\boldsymbol{x} - \boldsymbol{x}_i)^T \nabla v(\boldsymbol{\xi}_i)$$

hence

$$\|v(\boldsymbol{x}) - v(\boldsymbol{x}_i)\|_{\infty,\tau} = \|(\boldsymbol{x} - \boldsymbol{x}_i)^T \nabla v(\boldsymbol{\xi}_i)\|_{\infty,\tau} \leqslant 2h_\tau \max_{j=1,2} \|\frac{\partial v}{\partial x_j}\|_{0,\infty,\tau}$$

this is true over each triangle and the result follows. Now define

$$v^{\tau,i}(\boldsymbol{x}) := v(\boldsymbol{x}_i) + (\boldsymbol{x} - \boldsymbol{x}_i)^T \nabla v(\boldsymbol{x}_i)$$

which satisfies Taylor's theorem

$$v(\boldsymbol{x}) - v^{\tau,i}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}_i)^T \begin{pmatrix} \frac{\partial^2 v}{\partial x^2} & \frac{\partial^2 v}{\partial x \partial y} \\ \frac{\partial^2 v}{\partial x \partial y} & \frac{\partial^2 v}{\partial y^2} \end{pmatrix} (\boldsymbol{\zeta}_i)(\boldsymbol{x} - \boldsymbol{x}_i)^T$$

and hence assuming that $\max_{i,j=1,2} \|\frac{\partial^2 v}{\partial x_i \partial x_j}\|_{0,\infty} \leqslant C$ then

$$\|v(\boldsymbol{x}) - v^{\tau,i}(\boldsymbol{x})\|_{\infty,\tau} \leqslant Ch_\tau^2.$$

Now

$$\nabla v(\boldsymbol{x}) - \nabla v^{\tau,i}(\boldsymbol{x}) = \nabla v(\boldsymbol{x}) - \nabla v(\boldsymbol{x}_i)$$

hence when $\max_{i,j=1,2} \|\frac{\partial^2 v}{\partial x_i \partial x_j}\|_{0,\infty} \leqslant C$ we again have by Taylor's theorem

$$\|\nabla v(\boldsymbol{x}) - \nabla v^{\tau,i}(\boldsymbol{x})\|_{\infty,\tau} \leqslant Ch_\tau.$$

Recalling that

$$\phi_1(x,y) = \frac{1}{2|\tau|}\begin{vmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}$$

Then

$$\nabla\phi_1(x,y) = \frac{1}{2|\tau|}\begin{pmatrix} y_2 - y_3 \\ x_3 - x_2 \end{pmatrix}$$

and similarly

$$\nabla\phi_2(x,y) = \frac{1}{2|\tau|}\begin{pmatrix} y_3 - y_1 \\ x_1 - x_3 \end{pmatrix}, \quad \nabla\phi_3(x,y) = \frac{1}{2|\tau|}\begin{pmatrix} y_1 - y_2 \\ x_2 - x_1 \end{pmatrix}.$$

Hence

$$\|\nabla\pi^h v\|_{\infty,\tau} \leqslant C\frac{h_\tau}{|\tau|}\|v\|_{\infty,\tau}.$$

Hence, noting that $\pi^h v^{\tau,i} = v^{\tau,i}$

$$\begin{aligned} \|\nabla v - \nabla\pi^h v\|_{\infty,\tau} &\leqslant \|\nabla v - \nabla v^{\tau,i}\|_{\infty,\tau} + \|\nabla\pi^h(v - v^{\tau,i})\|_{\infty,\tau} \\ &\leqslant Ch_\tau + \frac{h_\tau}{|\tau|}\|v - v^{\tau,i}\|_{\infty,\tau} \leqslant Ch_\tau + \frac{h_\tau^3}{|\tau|} \leqslant Ch_\tau. \end{aligned}$$

Theorem 6.2
For Example 6.6 suppose that $u, \frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i \partial x_j} \in C(\Omega)$, *then*

$$\left[\iint_\Omega |\nabla(u - u^h)|^2 \, dA\right]^{1/2} \leqslant Ch. \tag{6.28}$$

Proof. Noting that the conditions of the previous Lemma hold

$$\begin{aligned} \iint_\Omega |\nabla(u - \pi^h u)|^2 \, \mathrm{d}A &\leqslant \sum_{\tau\in\mathcal{T}^h} \iint_\tau |\nabla(u - \pi^h u)|^2 \, \mathrm{d}A \\ &\leqslant C\sum_{\tau\in\mathcal{T}^h} |\tau|\|\nabla(u - \pi^h u)\|^2_{\infty,\tau} \leqslant \sum_{\tau\in\mathcal{T}^h} |\tau|h^2 \leqslant C|\Omega|h^2. \end{aligned}$$

7 Algebraic eigenvalue problems

One way of calculating the eigenvalues of A is to solve the equation $\det(A - \lambda I)$, this may be a lengthy process. We already know some simple bound for the eigenvalues $\rho(A) = \max_i |\lambda_i| \leqslant \|A\|$ for any consistent matrix norm.

THEOREM. 7.1 *(Gerschgorin theorems) 1. The eigenvalues of A lie in the union of the following disks in the complex plane:*

$$|z - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}| \qquad i = 1 \to n,\ z \in \mathbb{C}.$$

2. If m of the disks form a connected region, being isolated from all the other disks, then precisely m eigenvalues lie in this region.

EXAMPLES. Similarity transformation can help dilate the disks to obtain better bounds on the eigenvalues.

THEOREM. 7.2 *(Rayleigh Quotient) Let A be a symmetric real matrix, then all of the eigenvalues of A satisfy:*

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} \leqslant \lambda_i \leqslant \max_{\boldsymbol{x} \in \mathbb{R}^n} \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

The Power Method

The *power method* is a technique designed to estimate the dominant eigenvalue and corresponding eigenvector of a real $n \times n$ matrix A.

Let A be diagonisable, hence the n eigenvectors of A, $\{\boldsymbol{u}_i\}$, are linearly independent. Order the corresponding eigenvalues $\{\lambda_i\}_{i=1}^n$ and assume that

$$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n| \geqslant 0.$$

Since λ_1 is distinct, it is real and $\boldsymbol{u}_1$ can be chosen to be real. Further $\boldsymbol{u}_1^T \boldsymbol{u}_k = 0$ for $k = 2, \cdots, n$.

THEOREM. 7.3 *(Power Method) Let $\boldsymbol{v}$ be any real vector such that $\boldsymbol{v}^T \boldsymbol{u}_1 \neq 0 \neq \boldsymbol{u}_1^T \boldsymbol{x}^{(0)}$.*

$$\textit{Define } \boldsymbol{y}^{(k+1)} = A\boldsymbol{x}^{(k)}, \quad S_k = \frac{\boldsymbol{v}^T \boldsymbol{y}^{(k+1)}}{\boldsymbol{v}^T \boldsymbol{x}^{(k)}} \quad \textit{and} \quad \boldsymbol{x}^{(k+1)} = \frac{\boldsymbol{y}^{(k+1)}}{\|\boldsymbol{y}^{(k+1)}\|} \qquad \textit{for } k = 0, 1, \ldots$$

Then $\lim_{k \to \infty} S_k = \lambda_1$ and either $\lim_{k \to \infty} \boldsymbol{x}^{(k)} = \pm \boldsymbol{u}_1$ if $\lambda_1 > 0$ or $\lim_{k \to \infty} (-1)^k \boldsymbol{x}^{(k)} = \pm \boldsymbol{u}_1$ if $\lambda_1 < 0$.

PROOF. Any vector $\boldsymbol{x}^{(0)}$ can be expressed as $\boldsymbol{x}^{(0)} = \sum_{k=1}^{n} a_k \boldsymbol{u}_k$ where $a_1 = \boldsymbol{u}_1^T \boldsymbol{x}^{(0)} \neq 0$. We are going to prove by induction that for $k \geqslant 1$

$$\boldsymbol{x}^{(k)} = \frac{a_1\lambda_1^k \boldsymbol{u}_1 + \lambda_2^k a_2 \boldsymbol{u}_2 + \cdots + \lambda_n^k a_n \boldsymbol{u}_n}{\mu_k} \text{ where } \mu_k := \left\| a_1\lambda_1^k \boldsymbol{u}_1 + \lambda_2^k a_2 \boldsymbol{u}_2 + \cdots + \lambda_n^k a_n \boldsymbol{u}_n \right\|.$$

Notice that $\boldsymbol{y}^{(1)} = a_1\lambda_1^1 \boldsymbol{u}_1 + \lambda_2^1 a_2 \boldsymbol{u}_2 + \cdots + \lambda_n^1 a_n \boldsymbol{u}_n$ and as $\boldsymbol{x}^{(1)} = \boldsymbol{y}^{(1)}/\|\boldsymbol{y}^{(1)}\|$, the hypothesis is true for $k = 1$. Assume the hypothesis to be true for k

$$\begin{aligned} \boldsymbol{y}^{(k+1)} &= A\boldsymbol{x}^{(k)} = \frac{a_1\lambda_1^{k+1} \boldsymbol{u}_1 + \lambda_2^{k+1} a_2 \boldsymbol{u}_2 + \cdots + \lambda_n^{k+1} a_n \boldsymbol{u}_n}{\mu_k}, \\ \boldsymbol{x}^{(k+1)} &= \frac{\boldsymbol{y}^{(k+1)}}{\|\boldsymbol{y}^{(k+1)}\|} = \frac{(a_1\lambda_1^{k+1} \boldsymbol{u}_1 + \cdots + \lambda_n^{k+1} a_n \boldsymbol{u}_n) \times \mu_k}{\mu_k \times \left\| a_1\lambda_1^{k+1} \boldsymbol{u}_1 + \cdots + \lambda_n^{k+1} a_n \boldsymbol{u}_n \right\|}, \end{aligned}$$

therefore we have proven the induction hypothesis. Now for $k \geqslant 1$

$$\begin{aligned} S_k &:= \frac{\boldsymbol{v}^T \boldsymbol{y}^{(k+1)}}{\boldsymbol{v}^T \boldsymbol{x}^{(k)}} = \frac{\boldsymbol{v}^T A \boldsymbol{x}^{(k)}}{\boldsymbol{v}^T \boldsymbol{x}^{(k)}} = \frac{\boldsymbol{v}^T (a_1\lambda_1^{k+1} \boldsymbol{u}_1 + \cdots + \lambda_n^{k+1} a_n \boldsymbol{u}_n)\mu_k}{\mu_k \boldsymbol{v}^T (a_1\lambda_1^k \boldsymbol{u}_1 + \cdots + \lambda_n^k a_n \boldsymbol{u}_n)} \\ &= \frac{a_1\lambda_1 \boldsymbol{v}^T \boldsymbol{u}_1 + \cdots + \lambda_n \left(\frac{\lambda_n}{\lambda_1}\right)^k a_n \boldsymbol{v}^T \boldsymbol{u}_n}{a_1 \boldsymbol{v}^T \boldsymbol{u}_1 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^k a_n \boldsymbol{v}^T \boldsymbol{u}_n} \to \frac{a_1\lambda_1 \boldsymbol{v}^T \boldsymbol{u}_1}{a_1 \boldsymbol{v}^T \boldsymbol{u}_1} = \lambda_1 \square \end{aligned}$$

What if one of the assumptions is not valid?

- ⋆ If $\boldsymbol{u}_1^T \boldsymbol{x}^{(0)} = 0$, $|\lambda_2| > |\lambda_i|$ for $i \neq 1, 2$, $\boldsymbol{u}_2^T \boldsymbol{x}^{(0)} \neq 0$ and our arithmetic is exact then the sequence $\{\boldsymbol{x}^{(k)}\}$ will converge to $\pm \boldsymbol{u}_2$. However, in real computations, rounding errors are likely to introduce a small multiple of $\boldsymbol{u}_1$, so we will eventually get convergence to the dominant eigenvalue and eigenvector.
- ⋆ If $\boldsymbol{v}^T \boldsymbol{u}_1 = 0$, $|\lambda_2| > |\lambda_i|$ and $\boldsymbol{v}^T \boldsymbol{u}_2 \neq 0 \neq \boldsymbol{u}_2^T \boldsymbol{x}^{(0)}$, we find that $S_k \to \lambda_2$.
- ⋆ If $|\lambda_2| = |\overline{\lambda}_1| > |\lambda_3|$ then defining $\boldsymbol{z}^{(k)} = A^k \boldsymbol{z}^{(0)}$, we find

  $$\boldsymbol{z}^{(k)} = a_1\lambda_1^k \boldsymbol{u}_1 + a_2\overline{\lambda}_1^k \boldsymbol{u}_2 + \sum_{j=3}^{n} a_j\lambda_j^k \boldsymbol{u}_j,$$

  neglecting $|\lambda_j|^k$ in comparison with $|\lambda_1|^k$, for $j > 2$, we have $\boldsymbol{z}^{(k)} \approx a_1\lambda_1^k \boldsymbol{u}_1 + a_2\overline{\lambda}_1^k \boldsymbol{u}_2$ for k sufficiently large. If λ_1, $\overline{\lambda}_1$ are the roots of the equation $\lambda^2 + b\lambda + c$, and $\boldsymbol{z}^{(k)} = a_1\lambda_1^k \boldsymbol{u}_1 + a_2\overline{\lambda}_1^k \boldsymbol{u}_2$ then

  $$\boldsymbol{z}^{(k+2)} + b\boldsymbol{z}^{(k+1)} + c\boldsymbol{z}^{(k)} = a_1\lambda_1^k(\lambda_1^2 + b\lambda_1 + c)\boldsymbol{u}_1 + a_2\overline{\lambda}_1^k(\overline{\lambda}_1^2 + b\overline{\lambda}_1 + c)\boldsymbol{u}_2 = 0.$$

  From this set of linear equations, b and c may be found and hence λ_1 estimated.
- ⋆ If λ_1 is a multiple eigenvalue (of multiplicity r) with r linearly independent eigenvectors, the power method may be modified.

A reference is Burden & Faires p. 455

EXAMPLE.

$$\text{Let } A = \begin{pmatrix} -3 & 1 & 0 \\ 1 & -3 & -3 \\ 0 & -3 & 4 \end{pmatrix}, \quad \boldsymbol{v} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \boldsymbol{x}^{(0)} \qquad \text{where to 5 s.f.} \quad \begin{aligned} \lambda_1 &= 5.1248, \\ \lambda_2 &= -4.6477, \\ \lambda_3 &= -2.4771. \end{aligned}$$

Since A is symmetric, the three eigenvalues are real and the eigenvectors are orthogonal. As $\boldsymbol{x}^{(0)}$ and $\boldsymbol{v}$ are clearly *not* eigenvectors then $\boldsymbol{v}^T\boldsymbol{u}_1 \neq 0 \neq \boldsymbol{u}_1^T\boldsymbol{x}^{(0)}$.

$$\boldsymbol{y}^{(k)} = \left\{ \begin{pmatrix} 0 \\ -3 \\ 4 \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \right\}$$

estimates for $\lambda_1 = 4, 6.25, 4.36, 5.90, 4.53, \cdots$.

Some ways of speeding up the convergence are

- ⋆ Using Aitken's Δ^2-method we can accelerate the convergence.
- ⋆ *Origin Shift*: Let A have eigenvalues $\lambda_1 \to \lambda_n$ then the eigenvalues of $B := A - bI$ ($b \in \mathbb{R}$) are $\mu_i = \lambda_i - b$ ($i = 1 \to n$). Suppose for simplicity, $n = 2$ and $\lambda_1 > \lambda_2$. Then the dominant eigenvalue of B is either μ_1 or μ_2 depending on the choice of b. If the dominant eigenvalue is μ_1 then the convergence of the power method depends on $|\mu_2/\mu_1| = |\lambda_2 - b|/|\lambda_1 - b|$, choose b so that this quantity is as small as possible.

$$\text{Let } B = A + 4I = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & -3 \\ 0 & -3 & 8 \end{pmatrix} \quad \boldsymbol{v} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \boldsymbol{x}^{(0)}.$$

$$\boldsymbol{y}^{(k)} \left\{ \begin{pmatrix} 0 \\ -3 \\ 8 \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \right\}$$

estimates for $\mu_1 = 8, 9.13, 9.11, 9.12, 9.12, 9.12$. Convergence is much better than before.

This method can be used to shift the other eigenvalues and make them largest in magnitude.

Sturm sequence property

Let A be a real symmetric tridiagonal $n \times n$ matrix. Defining

$$f_0(\lambda) = 1, \quad f_k(\lambda) = \det \begin{pmatrix} \alpha_1 - \lambda & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 - \lambda & \beta_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \beta_{k-2} & \alpha_{k-1} - \lambda & \beta_{k-1} \\ 0 & \cdots & 0 & \beta_{k-1} & \alpha_k - \lambda \end{pmatrix}$$

we get the recursion relation

$$f_k(\lambda) = (\alpha_k - \lambda) f_{k-1}(\lambda) - \beta_{k-1}^2 f_{k-2}(\lambda) \qquad k = 2 \to n.$$

Assuming that $\beta_k \neq 0$ $(k = 1 \to n-1)$, then the number of roots of f_n which are greater than a is given by the number of agreements in sign of consectutive members of the sequence $\{f_k(a)\}_{k=0}^n$; if $f_k(a) = 0$, let its sign be opposite to that of $f_{k-1}(a)$.

Jacobi's method

Given a real symmetric matrix A, *Jacobi's method* uses similarity transformations to eliminate a_{pq} where $p \neq q$. Let U be the orthogonal matrix defined by

$$u_{ij} = \begin{cases} \delta_{ij} & \text{if } i \neq j \text{ and } \{i,j\} \neq \{p,q\} \\ \cos\theta & \text{if } i = j = p \text{ or } i = j = q \\ \sin\theta & \text{if } i = p \text{ and } j = q \\ -\sin\theta & \text{if } i = q \text{ and } j = p \end{cases}$$

Noting that $U^T A U$ has the following form

$$A^{(1)} = U^T A U = \left(\begin{array}{ccc} & | & \\ \text{---} & | & \text{---} \\ & | & \\ \text{---} & | & \text{---} \\ & | & \end{array} \right)$$

where the lines represent elements in the p'th or q'th row/column which have been transformed (all other elements of A remain unchanged). The elements where the lines cross are given by

$$a_{pp}^{(1)} = a_p p \cos^2\theta - 2a_{pq}\cos\theta\sin\theta + a_{qq}\sin^2\theta, \;\; a_{qq}^{(1)} = a_p p \cos^2\theta + 2a_{pq}\cos\theta\sin\theta + a_{qq}\sin^2\theta$$

$$\text{and } a_{pq}^{(1)} = (a_{pp} - a_{qq})\cos\theta\sin\theta + a_{pq}(\cos^2\theta - \sin^2\theta).$$

A calculation shows that θ satisfying

$$\cot(2\theta) = \frac{a_{qq} - a_{pp}}{2a_{pq}} \Longrightarrow a_{pq} = 0. \tag{7.29}$$

Jacobi's method is decribed iteratively as follows:

1. Set $A^{(0)} = A$ and $k = 1$.
2. At the k'th step, select $a_{pq}^{(k-1)}$ to be the largest in absolute value off-diagonal element of $A^{(k-1)}$.
3. Set $U^{(k)}$ to be the orthogonal matrix described above where θ satisfies (7.29).
4. Set $A^{(k)} = (U^{(k)})^T A^{(k-1)} U^{(k)}$.
5. Let $k = k + 1$. Either `STOP` if the sum of the squares off-diagonal elements is small or `GOTO` 2.

THEOREM. 7.4 *Let A be a real, symmetric $n \times n$ matrix. The sequence $A^{(k)}$ described above converges to a diagonal matrix D.*