MATH 536 - Numerical Solution of Partial Differential Equations

February 13, 2020

Contents

Chapter 1

Systems of Linear Equations

1.1 Direct Methods – Gaussian Elimination

1.1.1 Forward Elimination

(k-th step):

$$
\begin{aligned}
a_{ij}^{[k]} &= a_{ij}^{[k-1]} - a_{k-1,j}^{[k-1]}\ell_{i,k-1}^{[k-1]}, \qquad i = k,\ldots,n, \quad j = k-1,\ldots,n \\
\ell_{i,k-1}^{[k-1]} &= \frac{a_{i,k-1}^{[k-1]}}{a_{k-1,k-1}^{[k-1]}}, \qquad k = 2,\ldots,n \quad i = k,\ldots,n \\
b_{i}^{[k]} &= b_{i}^{[k-1]} - b_{k-1}^{[k-1]}\ell_{i,k-1}^{[k-1]}, \qquad i = k,\ldots,n
\end{aligned}
$$

In $n-1$ steps:

$$
\begin{bmatrix} a_{11}^{[1]} & a_{12}^{[1]} & \ldots & a_{1n}^{[1]} \\ 0 & a_{22}^{[2]} & \ldots & a_{2n}^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{nn}^{[n]} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{[1]} \\ b_2^{[2]} \\ \vdots \\ b_n^{[n]} \end{bmatrix}
$$

or

$$
Ux = b
$$

Operations Count

$$
\sum_{j=1}^{n-1} \left[\underbrace{(n-j)}_{\text{divisions}} + 2\underbrace{(n-j)\,(n-j+1)}_{\text{multiply and sum}} \right] = \frac{2n^3}{6} + \frac{n^2}{2} - \frac{7n}{6}
$$

Where we have used

$$
\sum_{j=1}^{m} j = \frac{m\,(m+1)}{2}, \quad \sum_{j=1}^{m} j^2 = \frac{m\,(m+1)\,(2m+1)}{6}
$$

Backward Substitution

$$
x_i = \frac{1}{a_{ii}^{[i]}} \left[b_i^{[i]} - \sum_{j=i+1}^{n} a_{ij}^{[i]} x_j \right]
$$

Operation Count

$$\sum_{i=1}^{n} \left[\underbrace{2\,(n-i)}_{\text{Multiplication and Subtraction}} + \underbrace{1}_{\text{division}} \right] = n^2$$

Matrix Representations

Define:

$$L_k = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ 0 & \dots & 1 & 0 \dots & 0 \\ 0 & \dots & -\ell_{k+1,k}^{[k]} & 1 \dots & 0 \\ \vdots & & \vdots & \ddots & \\ 0 & \dots & -\ell_{nk}^{[k]} & 0 \dots & 1 \end{bmatrix}$$

Note that:

$$L_k^{-1} = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ 0 & \dots & 1 & 0 \dots & 0 \\ 0 & \dots & \ell_{k+1,k}^{[k]} & 1 \dots & 0 \\ \vdots & & \vdots & \ddots & \\ 0 & \dots & \ell_{nk}^{[k]} & 0 \dots & 1 \end{bmatrix}$$

Lemma 1.1.1

$$A^{[2]} = L_1 A^{[1]}, \quad b^{[2]} = L_1 b^{[1]} \quad \left(A^{[1]} = A, b^{[1]} = b \right)$$

Therefore,

$$A^{[n]} = U = L_{n-1} L_{n-2} \cdots L_1 A$$

or

$$A = L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1} U = LU$$

Proposition 1.1.2

$$L = L_1^{-1} \cdots L_{n-1}^{-1}$$

is a lower triangular matrix with unit diagonal.

Definition A *Permutation matrix* is a matrix that has exactly one nonzero entry in each row and column, and this entry is equal to 1. The elementary permutation matrix (EPM) is defined as a matrix produced from the identity matrix by interchanging exactly one pair k, m of its columns. We denote it as $\underbrace{P_{km}}_{k<m} = P_k$.

For example:

$$P_{23} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Proposition 1.1.3 *If $A \in \mathbb{R}^{n \times n}$ and $P_{k,m} \in \mathbb{R}^{n \times n}$ is EPM, then $P_{k,m}A$ differs from A by an interchange of rows k and m. $AP_{k,m}$ differs from A by an interchange of columns k and m.*

1.2 Gaussian Elimination with Pivoting

$$A^{[n]} = U = L_{n-1}P_{n-1}\dots L_1P_1A$$

Where U is upper triangular, L_k is defined as before, and $P_k = P_{km}$ with $m \geq k$. Equivalently,

$$A = \underbrace{P_1L_1^{-1}P_2L_2^{-1}\cdots P_{n-1}L_{n-1}^{-1}}_{L^*}U$$

Lemma 1.2.1 *If $i \geq j > k$ and $P_j = P_{ji}$ then $P_jL_k^{-1}P_j$ is produced form L_k^{-1} by interchanging the j-th and i-th entry in the k-th column.*

Proof

$$\begin{bmatrix} \ddots & & & & & & \\ & 1 & & & & & \\ & \vdots & \ddots & & & & \\ & \ell_{jk}^{[k]} & & 1 & & & \\ & \vdots & & & \ddots & & \\ & \ell_{ik}^{[k]} & & & & 1 & \\ & \vdots & & & & & \ddots \end{bmatrix} \xrightarrow{P_jL_k^{-1}} \begin{bmatrix} \ddots & & & & & & \\ & 1 & & & & & \\ & \vdots & \ddots & & & & \\ & \ell_{ik}^{[k]} & & 0 & 1 & & \\ & \vdots & & & & \ddots & \\ & \ell_{jk}^{[k]} & & 1 & 0 & & \\ & \vdots & & & & & \ddots \end{bmatrix} \xrightarrow{P_jL_k^{-1}P_j} \begin{bmatrix} \ddots & & & & & & \\ & 1 & & & & & \\ & \vdots & \ddots & & & & \\ & \ell_{ik}^{[k]} & & 1 & & & \\ & \vdots & & & \ddots & & \\ & \ell_{jk}^{[k]} & & & & 1 & \\ & \vdots & & & & & \ddots \end{bmatrix}$$

Theorem 1.2.2 *Consider $P = P_{n-1}\dots P_1$. Then P is a permutation matrix and $PA = PL^*U = LU$, with L being a lower triangular matrix with unit diagonal.*

Proof

$$PL^* = PP_1L_1^{-1}P_2L_2^{-1}\cdots P_{n-1}L_{n-1}^{-1} =$$
$$P_{n-1}P_{n-2}\cdots \underbrace{P_1P_1}_{I} L_1^{-1}P_2L_2^{-1}\cdots P_{n-1}L_{n-1}^{-1} =$$
$$P_{n-1}P_{n-2}\cdots P_2L_1^{-1}P_2\underbrace{P_3\cdots P_{n-1}P_{n-1}\cdots P_3}_{I}L_2^{-1}\cdots P_{n-1}L_{n-1}^{-1} =$$
$$L_{1*}^{-1}P_{n-1}\cdots P_3L_2^{-1}P_3\cdots P_{n-1}L_{n-1}^{-1} = \cdots = L_{1*}^{-1}L_{2*}^{-1}\cdots L_{n-2*}^{-1}L_{n-1}^{-1} = L \quad \blacksquare$$

Definition If $A \in \mathbb{R}^{n\times n}$ and $1 \leq k \leq n$ then

$$\hat{A}_k = \begin{bmatrix} a_{11} & \dots & a_{1k} \\ a_{21} & \dots & a_{2k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \dots & a_{kk} \end{bmatrix}$$

is called the k^{th} principle sub-matrix of A

Theorem 1.2.3 *The pivot entries $a_{kk}^{[k]}$, $k = 1, 2, \dots, n-1$ are nonzero if and only if $\hat{A}_k$ are non-singular for $k = 1, \dots, n-1$ (note that $\hat{A}_n = A$ and we assume that A is nonsingular since otherwise the linear system is ill-posed).*

Proof (i) Suppose first that all $a_{kk}^{[k]} \neq 0, k = 1, \dots, n-1$. Then since $\hat{A}_k = \hat{L}_k\hat{U}_k$ it follows that $det(\hat{A}_k) = det(\hat{U}_k) = a_{11}^{[k]}\dots a_{kk}^{[k]} \neq 0$

(ii) Now suppose that $det(\hat{A}_k) \neq 0, k = 1, \dots, n-1$. Then, using a induction argument, we have:

(a) $a_{11}^{[1]} = \hat{A}_1$ and therefore $a_{11}^{[1]} \neq 0$.

(b) Assume that $a_{11}^{[1]}, \ldots, a_{kk}^{[k]} \neq 0$. If $\hat{A}_{k+1}^{[k+1]}$ is the $k+1$ principle sub-matrix of $A^{[k+1]}$ then it is easy to see that $\hat{A}_{k+1}^{[k+1]} = (\hat{L}_k)_{k+1} \ldots (\hat{L}_1)_{k+1} \hat{A}_{k+1}^{[1]}$ where we again assume that $(\hat{L}_m)_{k+1}, m = 1, \ldots, k$ are the $k+1$ principle sub matrices of L_m and $\hat{A}_{k+1}^{[1]}$ is the $k+1$ principle sub matrix of $A^{[1]} = A$. Therefore, $det(\hat{A}_{k+1}^{[k+1]}) = det(\hat{A}_{k+1}) \neq 0$. But $\hat{A}_{k+1}^{[k+1]}$ is upper triangular i.e.

$$\hat{A}_{k+1}^{[k+1]} = \begin{bmatrix} a_{11}^{[1]} & a_{12}^{[1]} & \ldots & a_{1,k+1}^{[1]} \\ 0 & a_{22}^{[2]} & \ldots & a_{2,k+1}^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{k+1,k+1}^{[k+1]} \end{bmatrix},$$

and then $det(\hat{A}_{k+1}^{[k+1]}) = a_{11}^{[1]} \ldots a_{k+1,k+1}^{[k+1]}$. This, together with the induction hypothesis yields that $a_{k+1,k+1}^{[k+1]} \neq 0$ which completes the proof.

Definition A matrix is *strictly diagonally dominant* if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

Corollary 1.2.4 *If a matrix is strictly diagonally dominant then no pivoting is necessary.*

Follows from theorem 1.2.3 and the following theorem due to Gershgorin:

Theorem 1.2.5 *The spectrum of $A \in \mathbb{C}^{n \times n}$, $S(A)$, is enclosed in the set:*

$$\left(\bigcup_{i=1}^{n} D_i \right),$$

where:

$$D_i = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}$$

Definition Matrix A is symmetric positive definite (spd) if

1. $A = A^T$
2. $v^T A v \geq 0$ for all $v \in \mathbb{R}^n$
3. $v^T A v = 0$ if and only if $v \equiv 0$

Corollary 1.2.6 *If a matrix is spd then no pivoting is necessary.*

Follows from the following theorem.

Theorem 1.2.7 *If $A \in \mathbb{R}^{n \times n}$ is symmetric the following statements are equivalent*

1. *A is positive definite,*
2. *$S(A)$ contains only positive real numbers, and*
3. *Every principle sub-matrix is positive definite.*

1.3 Special Cases

1.3.1 Cholesky Decomposition

Theorem 1.3.1 *If $A \in \mathbb{R}^{n\times n}$ is spd then there exists a lower triangular matrix C such that $CC^T = A$. Furthermore, the diagonal entries of C are positive.*

Proof Use induction on n:

$$n = 1 \quad : \quad C = [c_{11}] = \sqrt{a_{11}}$$

Assume the theorem holds for all matrices in $\mathbb{R}^{n\times n}$.

$$A = \begin{bmatrix} A_n & a \\ a^T & a_{n+1,n+1} \end{bmatrix}$$

where $a_{n+1,n+1} > 0$. By the induction hypothesis:

$$A_n = C_n C_n^T$$

Now try to find $[X]$ such that

$$CC^T = \begin{bmatrix} C_n & 0 \\ X^T & c \end{bmatrix} \begin{bmatrix} C_n^T & X \\ 0 & c \end{bmatrix} = \begin{bmatrix} A_n & a \\ a^T & a_{n+1,n+1} \end{bmatrix}$$

or such that

$$C_n X = a \quad \Rightarrow \quad X = C_n^{-1} a$$

and

$$X^T X + c^2 = a_{n+1,n+1} \quad \Rightarrow \quad c = \sqrt{a_{n+1,n_1} - X^T X}$$

But, is $c > 0$?

$$X^T X = \left[C_n^{-1} a\right]^T C_n^{-1} a = a^T \left[C_n^{-1}\right]^T C_n^{-1} a = a^T A_n^{-1} a$$

therefore

$$a_{n+1,n+1} - a^T A_n^{-1} a = a_{n+1,n+1} - X^T X$$

Then define $x := \left[\left[A_n^{-1} a\right]^T, -1\right]^T \neq 0$ so that

$$x^T A x = a_{n+1,n+1} - a^T A_n^{-1} a > 0 \quad \blacksquare$$

Cholesky Algorithm

1. $c_{11} \leftarrow \sqrt{a_{11}}$
2. For $i = 2, 3, \ldots, n$
3. $c_{i1} \leftarrow a_{i1}/c_{11}$
4. End i
5. For $j = 2, 3, \ldots, n-1$
6. $$c_{jj} \leftarrow \sqrt{a_{jj} - \sum_{k=1}^{j-1} c_{jk}^2}$$
7. For $i = j+1, \ldots, n$

8.
$$c_{ij} \leftarrow \frac{1}{c_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} c_{ik} c_{jk} \right)$$

9. End i

10. End j

11.
$$c_{nn} \leftarrow \sqrt{a_{nn} - \sum_{k=1}^{n-1} c_{nk}^2}$$

12. End

1.3.2 Thomas Algorithm

Definition A *band matrix* is of the form

$$\begin{bmatrix}
a_{11} & a_{1,2} & \dots & a_{1,l} & 0 & \dots & \dots & \dots & \dots & 0 \\
a_{21} & \ddots & \ddots & & \ddots & \ddots & & & & \vdots \\
\vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & & & \vdots \\
a_{k,1} & \dots & a_{k,k-1} & a_{kk} & a_{k,k+1} & \dots & a_{k,k+l-1} & 0 & \dots & 0 \\
0 & \ddots & & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & & \ddots & \ddots & \ddots & & \ddots & 0 \\
\vdots & & \ddots & \ddots & & \ddots & \ddots & \ddots & & a_{n-l+1,n} \\
\vdots & & & \ddots & \ddots & & \ddots & \ddots & \ddots & \vdots \\
\vdots & & & & \ddots & \ddots & & \ddots & \ddots & a_{n-1,n} \\
0 & \dots & \dots & \dots & \dots & 0 & a_{n,n-k+1} & \dots & a_{n,n-1} & a_{n,n}
\end{bmatrix}$$

A band matrix may be stored compactly as an $l + k - 1 \times n$ matrix of the form:

$$\begin{bmatrix}
\overbrace{0 \quad \dots \quad 0}^{k-1} & & & a_{11} & a_{12} & \dots & a_{1l} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
a_{k1} & \dots & a_{k,k-1} & a_{kk} & a_{k,k+1} & \dots & a_{k,k+l-1} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
a_{n,n-k+1} & \dots & a_{n,n-1} & a_{nn} & \underbrace{0 \quad \dots \quad 0}_{l-1} & &
\end{bmatrix}$$

Thomas Algorithm

For the linear system

$$\begin{bmatrix}
b_1 & c_1 & 0 & \dots & \dots & 0 \\
a_2 & b_2 & c_2 & 0 & \dots & 0 \\
0 & a_3 & b_3 & c_3 & \dots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & \ddots & \ddots & 0 \\
\vdots & & & \ddots & a_{n-1} & b_{n-1} & c_{n-1} \\
0 & \dots & \dots & 0 & a_n & b_n
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} =
\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix}$$

1.
$$b_j^{[f]} = b_j - a_j \frac{c_{j-1}^{[f]}}{b_{j-1}^{[f]}} \quad \text{and} \quad d_j^{[f]} = d_j - a_j \frac{d_{j-1}^{[f]}}{b_{j-1}^{[f]}}, \quad \text{for} \quad j = 2, \ldots, n$$

2.
$$x_{j-1} = d_{j-1}^{[f]} - b_j^{[f]} x_j, \quad \text{for} \quad j = n, \ldots, 2$$

1.4 Matrix Norms

Definition If $A \in \mathbb{R}^{n\times n}$ then the mapping $\mathbb{R}^{n\times n} \to \mathbb{R}$ is called a norm of $A := ||A||$ if and only if:

1. $||A|| \geq 0$ and $||A|| = 0$ if and only if $A = 0$,
2. $||\lambda A|| = |\lambda| \, ||A||$ for all $\lambda \in \mathbb{R}$,
3. $||A + B|| \leq ||A|| + ||B||$.

If in adddition, $||AB|| \leq ||A|| \, ||B||$ the norm is called multiplicative. In the sequel, we make use of multiplicative norms only, and therefore, the notion of a norm presumes a multiplicative norm.

Definition If $A \in \mathbb{R}^{n\times n}$, then for a *vector norm* $||\cdot|| \; : \; \mathbb{R}^n \to \mathbb{R}$,

$$||A|| := \sup_{||x||\neq 0} \frac{||Ax||}{||x||}$$

is called a *subordinate* matrix norm.

Theorem 1.4.1 *Let $A \in \mathbb{R}^{n\times n}$. Then,*

1.
$$||A||_\infty := \sup_{||x||_\infty \neq 0} \frac{||Ax||_\infty}{||x||_\infty} = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

2.
$$||A||_1 := \sup_{||x||_1 \neq 0} \frac{||Ax||_1}{||x||_1} = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

3.
$$||A||_2 := \sup_{||x||_2 \neq 0} \frac{||Ax||_2}{||x||_2} = \sqrt{\rho\left(A^T A\right)}$$

Proof

$$||Ax||_\infty = \max_{1\leq i\leq n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \left(\max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \right) \left(\max_{1\leq j\leq n} |x_j| \right) = M \, ||x||_\infty$$

therefore,

$$||A||_\infty \leq M$$

Let I be the index for which

$$\max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| = M$$

is achieved. Choose $x = \{x_j\}_{j=1}^n$

$$x_j := \begin{cases} a_{Ij} / |a_{Ij}| & : \quad a_{Ij} \neq 0 \\ 0 & : \quad a_{ij} = 0. \end{cases}$$

Then $||x||_\infty = 1$ and $||Ax||_\infty = M$. Therefore,

$$||A||_\infty = M \quad \blacksquare$$

1.5 Error Estimates

Consider two vectors x_1 and x_2 where

$$Ax_1 = b \quad \text{and} \quad Ax_2 = b + r$$

then

$$x_1 = A^{-1}b \quad \text{and} \quad x_2 = A^{-1}(b + r)$$

and it follows immediately that

$$||x_2 - x_1|| = \left|\left|A^{-1}r\right|\right|$$

Then,

$$\frac{||x_2 - x_1||}{||A||\,||x_1||} \leq \frac{||x_2 - x_1||}{||Ax_1||} = \frac{\left|\left|A^{-1}r\right|\right|}{||b||} \leq \frac{\left|\left|A^{-1}\right|\right|\,||r||}{||b||}$$

Therefore,

$$\frac{||x_2 - x_1||}{||x_1||} \leq \underbrace{||A||\,\left|\left|A^{-1}\right|\right|}_{c(A)} \frac{||r||}{||b||}$$

This provides means of estimating the error introduced in the solution of a linear system due to roundoff errors or uncertainty in the right-hand-side or the matrix of the system. The number $c(A)$ is called the condition number of A with respect to the norm $\|.\|$.

1.6 Iterative Method Preliminaries

Definition If $x_k \in \mathbb{R}^n$ then the sequence $\{x_k\}_{k=1}^{\infty}$ converges to $x \in \mathbb{R}^n$ in a norm $||\cdot||$ i.e.

$$\lim_{k\to\infty} x_k = x$$

if

$$\lim_{k\to\infty} ||x - x_k|| = 0.$$

Definition If $A_k \in \mathbb{R}^{n\times n}$ then the sequence $\{A_k\}_{k=1}^{\infty}$ converges to $A \in \mathbb{R}^{n\times n}$ in a norm $||\cdot||$ i.e.

$$\lim_{k\to\infty} A_k = A$$

if

$$\lim_{k\to\infty} ||A - A_k|| = 0.$$

Consider the matrix power series

$$\sum_{k=0}^{\infty} a_k A^k. \tag{1.1}$$

It is convergent if and only if

$$\lim_{K\to\infty} \sum_{k=0}^{K} a_k A^k = f(A),$$

where $f(A)$ is some matrix with a finite norm. In the following we will also need to consider the numerical power series

$$f(\lambda) \sim \sum_{k=0}^{\infty} a_k \lambda^k. \tag{1.2}$$

Theorem 1.6.1 *The matrix power series* (1.1) *is convergent if for all* $\lambda_i \in S(A)$ $|\lambda_i| < \rho$, *where* ρ *is the radius of convergence of* (1.2) *and* $S(A)$ *is the spectrum of* A. *It is divergent if* $|\lambda_i| > \rho$ *for some* i.

Proof From Jordan's theorem, there exists nonsingular C such that A can be transformed as follows:

$$B = C^{-1}AC, \;\; B = \operatorname{diag}(B_i), \;\; B_i = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots\dots & & 0 & \lambda_i \end{bmatrix}$$

where $B_i \in \mathbb{R}^{n_i \times n_i}$ and n_i is the multiplicity of the eigenvalue λ_i. Then, $\sum_{k=0}^{\infty} a_k A^k$ is convergent if and only if

$$\sum_{k=0}^{\infty} a_k B^k = \sum_{k=0}^{\infty} a_k C^{-1} A^k C = C^{-1} \left(\sum_{k=0}^{\infty} a_k A^k \right) C$$

is convergent as well. Note that $B^k = \operatorname{diag}\left(B_i^k\right)$ since B is a block-diagonal matrix. Now, consider the powers B_i^k of the i-th block of B. It can be shown by induction that:

$$B_i^2 = \begin{bmatrix} \lambda_i^2 & 2\lambda_i & 1 & \dots & 0 \\ 0 & \lambda_i^2 & 2\lambda_i & \dots & 0 \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ 0 & \dots\dots & & & \lambda_i^2 \end{bmatrix}, \dots, \quad B_i^k = \begin{bmatrix} \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \dots\dots & & \binom{k}{n_i-1}\lambda_i^{k-n_i+1} \\ 0 & \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \dots & \binom{k}{n_i-2}\lambda_i^{k-n_i+2} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ 0 & \dots\dots & & & \lambda_i^k \end{bmatrix}$$

so that the i-th block of the m-th partial sum of $\sum_{k=0}^{\infty} a_k B^k$ is

$$f_m(B_i) = \sum_{k=0}^{m} a_k B_i^k = \begin{bmatrix} f_m(\lambda_i) & \frac{1}{1!}f_m'(\lambda_i) & \dots & \frac{1}{(n_i-1)!}f_m^{(n_i-1)}(\lambda_i) \\ 0 & f_m(\lambda_i) & \dots & \frac{1}{(n_i-2)!}f_m^{(n_i-2)}(\lambda_i) \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ 0 & \dots & 0 & f_m(\lambda_i). \end{bmatrix}$$

Therefore, $f_m(B)$ is convergent if and only if

$$f_m(\lambda_i) = \sum_{k=0}^{m} a_k \lambda_i^k$$

(and therefore $f_m^{(l)}(\lambda_i)$) is convergent for $i = 1, \dots, n$ and $l = 1, \dots, n_i$. $f_m(\lambda_i)$ is convergent if $|\lambda_i|$ is smaller than ρ, the radius of convergence of (1.2). ∎

Corollary 1.6.2

$$f(A) = I + A + A^2 + \dots + A^m + \dots$$

is convergent if and only if $|\lambda_k| < 1$ *for all* $k = 1, \dots, n$ *and* $A \in \mathbb{R}^{n \times n}$.

Corollary 1.6.3 *Since* $|\lambda_k| \leq ||A||$ *for any subordinate norm* $||\cdot||$ *then* $f(A)$ *is convergent if* $||A|| < 1$.

1.7 Iterative Methods

1.7.1 Matrix Splitting Methods

Consider

$$Ax = b$$

If $A = B + C$, so that $det(C) \neq 0$,then

$$Ax = b \quad \Leftrightarrow \quad x = x + C^{-1}(b - Ax)$$

and in this form we may solve the problem iteratively:

$$x^{k+1} = x^k + C^{-1}\left(b - Ax^k\right) \tag{1.3}$$

or equivalently:

$$Cx^{k+1} = b - Bx^k. \tag{1.4}$$

The last form clarifies why such methods are called matrix-splitting methods. Note, that C can be any non-singular matrix but not every choice is a good choice, of course. The essence of various iterative methods is in the choice of the matrix C that allows for the fast computation of a good approximation to the solution. If C is the identity matrix, the corresponding iteration is called simple (or Richardson) iteration. It is seldom a good choice. The form (1.3) suggests that the closer C is to A, the faster the convergence. On the other hand, the iteration requires the solution of a linear system with C and therefore, C should be such that this solution requires much less resources than the solution of the original system. These are very contradictory requirements and the choice of C depends on the properties of A and some other contraints like computer resources.

Theorem 1.7.1 *The iteration* (1.3) *is convergent if and only if* $\rho\left(I - C^{-1}A\right) < 1$.

Proof

$$\begin{aligned} x^{k+1} &= \underbrace{\left(I - C^{-1}A\right)}_{D} x^k + C^{-1}b = D^2x^{k-1} + (I + D)C^{-1}b = \ldots \\ &= D^{k+1}x^0 + \left(I + D + D^2 + \ldots + D^k\right)C^{-1}b \end{aligned}$$

$I + D + \ldots + D^k + \ldots$ is convergent iff $\rho\left(I - C^{-1}A\right) < 1$. Also, if it is convergent then

$$D^k \underset{k\to\infty}{\longrightarrow} 0 \quad \blacksquare$$

Definition $R(D) = -\log_{10}\rho(D)$ is called the *convergence rate* of the iteration.

Jacobi's Method

$$\begin{aligned} C &= \operatorname{diag}(A) \\ D = I - C^{-1}A &= \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ \vdots & & \ddots & & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \cdots & 0 \end{bmatrix} \end{aligned}$$

Theorem 1.7.2 *If A is strictly diagonally dominant then Jacobi is convergent.*

Proof

$$||D||_\infty = \max_i \sum_{j\neq i} \left|\frac{a_{ij}}{a_{ii}}\right| < 1.$$

Then for any eigenvalue λ_i of D we have:

$$|\lambda_i| \leq ||D||_\infty < 1 \quad \blacksquare.$$

Gauss-Seidel (GS) Iteration

$$C = L = \begin{bmatrix} a_{11} & & 0 \\ \vdots & \ddots & \\ a_{1n} & \dots & a_{nn} \end{bmatrix}$$
$$B = U = A - L$$
$$x^{k+1} = x^k + C^{-1}\left(b - Ax^k\right)$$

Theorem 1.7.3 *If A is strictly diagonally dominant the Gauss-Seidel iteration is convergent.*

Proof The formula for the i-th component of the $k+1$ iterate x^{k+1} is:

$$x_i^{k+1} = -\sum_{j<i} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j>i} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}}$$

The exact solution clearly is a fixed point of the iteration i.e.

$$x_i = -\sum_{j<i} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j>i} \frac{a_{ij}}{a_{ii}} x_j + \frac{b_i}{a_{ii}}$$

Therefore the error of the $k+1$ iteration, $\epsilon^{k+1} = x^{k+1} - x$, must satisfy:

$$\epsilon_i^{k+1} = -\sum_{j<i} \frac{a_{ij}}{a_{ii}} \epsilon_j^{k+1} - \sum_{j>i} \frac{a_{ij}}{a_{ii}} \epsilon_j^k$$

then

$$\left|\epsilon_i^{k+1}\right| \le \sum_{j<i} \left|\frac{a_{ij}}{a_{ii}}\right| \left|\epsilon_j^{k+1}\right| + \sum_{j>i} \left|\frac{a_{ij}}{a_{ii}}\right| \left|\epsilon_j^k\right| \tag{1.5}$$

Using induction on i, it is straightforward to show that:

$$\left|\epsilon_i^{k+1}\right| \le \left(\sum_{j\neq i} \left|\frac{a_{ij}}{a_{ii}}\right|\right) \left|\left|\epsilon^k\right|\right|_\infty < \left|\left|\epsilon^k\right|\right|_\infty .$$

Indeed, assuming that $\left|\epsilon_j^{k+1}\right| < \left|\left|\epsilon^k\right|\right|_\infty , \quad \forall j < i$, and substituting this into (1.5) yields that $\left|\epsilon_i^{k+1}\right| < \left|\left|\epsilon^k\right|\right|_\infty$.

Therefore,

$$\left|\left|\epsilon^{k+1}\right|\right|_\infty < \left|\left|\epsilon^k\right|\right|_\infty$$

so that the error must strictly decrease with each iteration. Then the error must go to zero because the exact solution is the unique fixed point of the iteration (why is it unique?). ∎

Successive Over-relaxation (SOR) Method

Suppose $\omega \in \mathbb{R}$ and consider the iteration:

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

Here $C = L + \frac{1}{\omega} D, D = \text{diag}\,(A)$, and being L the lower triangular part of A, not including the main diagonal. Then clearly $A = C + B$, where $B = (1 - \frac{1}{\omega})D + U$, U being the strictly upper triangular part of A (with zeros on the main diagonal). This iteration is similar to the GS iteration but has a free parameter that allows to better control the convergence rate.

Theorem 1.7.4 *(Ostrowski-Reich) If $A \in R^{n\times n}$ is spd then SOR is convergent iff $0 < \omega < 2$.*

1.7.2 Optimization-based Methods

From now on, until the end of this section we assume that the matrix A is s.p.d.

Theorem 1.7.5 *If $A \in \mathbb{R}^{n\times n}$ is spd then u is a solution to the linear system $Au = b$ if and only if u minimizes the function:*

$$F(v) = \frac{1}{2}v^T Av - b^T v$$

Proof Let u be the solution of the linear system i.e. $Au = b$. Then for any $v \in \mathbb{R}$ we have that:

$$\begin{aligned} F(v) - F(u) &= \frac{1}{2}v^T Av - b^T v - \frac{1}{2}u^T Au + b^T u \\ &= \frac{1}{2}v^T Av - u^T Av + \frac{1}{2}u^T Au = \frac{1}{2}(v-u)^T A(v-u) \geq 0 \end{aligned}$$

then

$$\frac{1}{2}(v-u)^T A(v-u) = 0 \quad \Leftrightarrow \quad v = u$$

so that $F(v) \geq F(u)$ and

$$F(v) = F(u) \quad \Leftrightarrow \quad v = u \quad \blacksquare$$

Definition If A is spd then for any two vectors $u, v \in \mathbb{R}^n$, $u^T Av$ defines the *energy inner product* induced by A denoted by:

$$\langle u, v\rangle_A = u^T Av.$$

Note that in the rest of the notes at some occasions $(u, v)_A$ is also used to denote the energy inner product.

Steepest Descent Method

Note that any iteration for finding a solution to a linear system $Au = b$ can be written in the form:

$$u^{k+1} = u^k + \alpha_k p_k$$

where $\alpha_k \in \mathbb{R}$ is called iteration step, $p_k \in \mathbb{R}^n$ is called the search direction. For example, in the case of matrix splitting methods the search direction is chosen to be $p_k = C^{-1}r_k$, with $r_k = b - Au^k$ being called the residual of the iterate u^k, and $\alpha_k = 1$. The step does not need to be constant and the gradient-type methods choose it at each iteration step so that this choice minimizes the function $F(v)$. The basic gradient-type iterative algorithm therefore can be written as:

1. Choose an initial search direction
$$p_0 = r_0 = b - Au_0$$
2. For $k = 1, 2, 3, \ldots$ do:
   (a) Find $\alpha_k \in \mathbb{R}$ such that
   $$u^{k+1} = u^k + \alpha_k p_k$$
   minimizes F over the line $u^k + \alpha p_k$
   (b) New iterate
   $$u^{k+1} = u^k + \alpha_k p_k$$
   (c) The new residual is
   $$r_{k+1} = b - Au^{k+1}$$
   (d) Find the new search direction p_{k+1}.
   (e) Let $k = k + 1$ and repeat item 2 until convergence.

To find α_k we minimize F along the search direction p_k:

$$\begin{aligned}\frac{d}{d\alpha_k}F\left(u^k+\alpha_k p_k\right) &= \nabla F\left(u^k+\alpha_k p_k\right)\cdot\frac{d}{d\alpha_k}\left(u^k+\alpha_k p_k\right)\\ &= \left[A\left(u^k+\alpha_k p_k\right)-b\right]^T p_k = \left[Au^k-b\right]^T p_k+\alpha_k p_k^T A p_k\\ &= -r_k^T p_k+\alpha_k p_k^T A p_k = 0\end{aligned}$$

so that

$$\boxed{\alpha_k = \frac{r_k^T p_k}{p_k^T A p_k}}$$

Note that for this choice of α_k we have:

$$\left[\underbrace{Au^{k+1}-b}_{-r_{k+1}}\right]^T p_k = \left[A\underbrace{\left(u^k+\alpha_k p_k\right)}_{u^{k+1}}-b\right]^T p_k = -r_k^T p_k+\alpha_k p_k^T A p_k = 0 \tag{1.6}$$

or (1.7)

$$-r_{k+1}^T p_k = \nabla F\left(u_{k+1}\right) p_k = 0 \tag{1.8}$$

In the steepest descent method we choose

$$p_k = r_k = b-Au^k.$$

Because this is the direction of the negative gradient of $F(v)$, this function is clearly non-increasing in this direction. Unfortunately, this choice of search direction is not optimal in some sense that will be elucidated in the next section.

Conjugate Gradient Method (CGM)

Definition u^k is optimal with respect to direction $p\neq 0$ iff $F\left(u^k\right)\leq F\left(u^k+\lambda p\right)$ for all $\lambda\in\mathbb{R}$.

Lemma 1.7.6 *u^k is optimal with respect to p iff*

$$p^T r_k = 0.$$

Proof u^k is optimal with respect to p iff $F\left(u^k+\lambda p\right)$ has a minimum at $\lambda=0$, i.e.

$$\left.\frac{\partial F}{\partial\lambda}\left(u^k+\lambda p\right)\right|_{\lambda=0} = p^T\underbrace{\left(Au^k-b\right)}_{-r_k}+\left.\lambda p^T A p\right|_{\lambda=0} = 0. \tag{1.9}$$

(Note that $\frac{\partial^2 F}{\partial\lambda^2} = p^T A p > 0$.)

(1.9) is obviously satisfied iff $p^T r_k = 0$.

We know that the choice of α_k guarantees that u_{k+1} is optimal with respect to p_k and this applies in particular to the steepest descent method for which $p_k = r_k$. However, u^{k+2} may not be optimal with respect to p_k i.e. the next iteration can undo some of the minimization work of the previous iteration. Indeed:

$$u^{k+2} = u^{k+1}+\alpha_{k+1}r_{k+1}$$

which gives

$$r_{k+2} = r_{k+1}-\alpha_{k+1}Ap_{k+1} \Rightarrow p_k^T r_{k+2} = \underbrace{p_k^T r_{k+1}}_{=0}-\alpha_{k+1}p_k^T Ap_{k+1} \quad\Rightarrow\quad p_k^T r_{k+2} = -\alpha_{k+1}p_k^T Ap_{k+1}.$$

For the steepest descent method $p_m = r_m$ and therefore, since $r_k^T r_{k+1} = 0$ (see (1.8)), $r_k^T A r_{k+1} \neq 0$, unless $A = cI$.

To find a different search direction p_{k+1} that maintains the optimality of u^{k+2} w.r.t. p_k, we need that:

$$0 = p_k^T r_{k+2} = \alpha_{k+1} p_k^T A p_{k+1} = 0.$$

For the conjugate gradient method we require that u^{k+2} is also optimal with respect to p_k i.e. we choose the search direction p_{k+1} from this condition.

Definition If $u, v \in \mathbb{R}^n$ are such that $\langle u, v \rangle_A = 0$ for some s.p.d. A then u, v are called *A-conjugate*.

This means that p_{k+1} should be *A-conjugate* to p_k. We search for p_{k+1} in the form $p_{k+1} = r_{k+1} + \beta_k p_k$, for some $\beta_k \in \mathbb{R}$, and then from the condition $p_k^T A p_{k+1} = 0$ we easily obtain that $\beta_k = -\dfrac{\langle r_{k+1}, p_k \rangle_A}{||p_k||_A^2}$

Algorithm (Basic CGM procedure) If $A \in \mathbb{R}^{n \times n}$ is spd and u^0 is an initial guess:

1. $r_0 \leftarrow b - Au^0$
2. $p_0 = r_0$
3. For $k = 1, 2, \ldots, m$:
4. $\alpha_{k-1} \leftarrow r_{k-1}^T p_{k-1} / \left|\left|p_{k-1}\right|\right|_A^2$
5. $u^k \leftarrow u^{k-1} + \alpha_{k-1} p_{k-1}$
6. $r_k \leftarrow r_{k-1} - \alpha_{k-1} A p_{k-1}$
7. $$p_k \leftarrow r_k - \frac{\langle r_k, p_{k-1} \rangle_A}{||p_{k-1}||_A^2} p_{k-1}$$
8. Next k
9. End

As it will be demonstrated below, if all arithmetic operations are exact, the algorithm computes the exact solution in a finite number of steps (finite termination property).

Later, we will show that the error decreases as:

$$\left|\left|u - u^k\right|\right|_A = \left|\left|\epsilon^k\right|\right|_A \leq 2 \left|\left|\epsilon_0\right|\right|_A \left[\frac{\sqrt{\mathrm{c}\,(A)} - 1}{\sqrt{\mathrm{c}\,(A)} + 1}\right]^k$$

If $\mathrm{c}\,(A) \gg 1$ then the convergence is slow. Therefore, the algorithm is often modified by formally multiplying the original system by a matrix that has a spectrum close to the spectrum of A and performing the CG method on the modified system which has the same solution as the original one. This process is called preconditioning.

Preconditioning

Instead of $Au = b$ we solve $\tilde{A}\tilde{u} = \tilde{b}$ where

$$\tilde{A} = \left(B^{-1}\right)^T A B^{-1}, \quad \tilde{u} = Bu, \quad \tilde{b} = \left(B^{-1}\right)^T b$$

Since the preconditioner $B^T B$ must be an approximation to the matrix A, a natural choice for B should be an approximation to the square root of A. Since A is s.p.d. its square root is also s.p.d. and therefore we assume that $B^T = B$. The preconditioner B^2 must satisfy somewhat contradicting requirements since on one hand it

should be a good approximation of A and on the other, it should be much easier to solve a linear system with a matrix B^2 than with A. The second requirement follows from the fact that, as it will become clear below, on each iteration of the CG method with preconditioning, we will need to solve one system with B^2.

For the system $Au = b$ we know that

$$u: \quad F(u) = \min_{v \in \mathbb{R}^n} F(v)$$

Now, if $\tilde{v} = Bv$ then

$$\begin{aligned} F(v) &= \frac{1}{2} v^T A v - b^T v = \frac{1}{2} \left[B^{-1}\tilde{v}\right]^T A \left[B^{-1}\tilde{v}\right] - b^T \left(B^{-1}\tilde{v}\right) \\ &= \frac{1}{2} \tilde{v}^T \underbrace{B^{-1} A B^{-1}}_{\tilde{A}} \tilde{v} - \underbrace{\left(B^{-1} b\right)^T}_{\tilde{b}^T} \tilde{v} = \tilde{F}(\tilde{v}) \end{aligned}$$

The preconditioned method should compute the new iterate from: $\tilde{u}^k = \tilde{u}^{k-1} + \tilde{\alpha}_{k-1}\tilde{p}_{k-1}$, with $\tilde{\alpha}_{k-1} = \frac{\tilde{r}_{k-1}^T \tilde{p}_{k-1}}{||\tilde{p}_{k-1}||_{\tilde{A}}}$. The search direction for the preconditioned method should be computed as $\tilde{p}_k = \tilde{\beta}_{k-1}\tilde{p}_{k-1} + \tilde{r}_k$ where the residual $\tilde{r}_k$ is given by $\tilde{r}_k = \tilde{r}_{k-1} - \tilde{\alpha}_{k-1}\tilde{A}\tilde{p}_{k-1}$, and $\tilde{\beta}_k = -\frac{\langle \tilde{r}_{k+1}, \tilde{p}_k \rangle_{\tilde{A}}}{||\tilde{p}_k||^2_{\tilde{A}}}$. These formulae are not convenient for practical computations since they require to somehow compute the inverse of the preconditioner, $(B^2)^{-1}$. Fortunately, it appears that the preconditioned method can be implemented almost exactly as the original CGM modifying only the computation of the search direction to:

$$p_k = \underbrace{\left(B^2\right)^{-1} r_k}_{z_k} - \frac{\left(\overbrace{\left(B^2\right)^{-1} r_k}^{z_k}\right)^T A p_{k-1}}{p_{k-1}^T A p_{k-1}} p_{k-1}, \tag{1.10}$$

and defining the initial iterate as $u^0 = B^{-1}\tilde{u}^0$ and the initial search direction as $p_0 = (B^2)^{-1} r_0$. With this modification, the residuals $r_k, \tilde{r}_k$, search directions $p_k, \tilde{p}_k$, and iterates $u^k, \tilde{u}^k$, of the modified CG algorithm and the preconditioned CG algorithm, verify the relations: $r_k = B\tilde{r}_k, p_k = B^{-1}\tilde{p}_k, u^k = B^{-1}\tilde{u}^k$. Indeed, using the assumption that $r_{k-1} = B\tilde{r}_{k-1}, p_{k-1} = B^{-1}\tilde{p}_{k-1}$, and $u^{k-1} = B^{-1}\tilde{u}^{k-1}$ (these assumptions are trivially verifiable at $k = 1$), we obtain that:

$$\tilde{\alpha}_{k-1} = \frac{\tilde{r}_{k-1}^T \tilde{p}_{k-1}}{||\tilde{p}_{k-1}||^2_{\tilde{A}}} = \frac{r_{k-1}^T B^{-1} B p_{k-1}}{\left(B p_{k-1}\right)^T B^{-1} A B^{-1} B p_{k-1}} = \frac{r_{k-1}^T p_{k-1}}{||p_{k-1}||^2_A} = \alpha_{k-1},$$

and subsequently that: $r_k = B\tilde{r}_k$ and $u^k = B^{-1}\tilde{u}^k$. For example,

$$\tilde{u}^k = \tilde{u}^{k-1} + \tilde{\alpha}_{k-1}\tilde{p}_{k-1} = B(u^{k-1} + \alpha_{k-1} p_{k-1}) = B u^k.$$

This immediately yields that $r_k = B\tilde{r}_k$. Then, if p_k is computed from (1.10) we obtain, multiplying it by B, that:

$$B p_k = B^{-1} r_k - \frac{r_k^T B^{-1} B^{-1} A B^{-1} B p_{k-1}}{p_{k-1}^T B B^{-1} A B^{-1} B p_{k-1}} B p_{k-1},$$

or, using the induction assumptions:

$$B p_k = \tilde{r}_k - \frac{\langle \tilde{r}_k, \tilde{p}_{k-1} \rangle_{\tilde{A}}}{||\tilde{p}_{k-1}||^2_{\tilde{A}}} \tilde{p}_{k-1} = \tilde{p}_k$$

Thus, using an induction argument we establish that $r_k = B\tilde{r}_k, p_k = B^{-1}\tilde{p}_k, u^k = B^{-1}\tilde{u}^k$, and therefore, $\tilde{\alpha}_k = \alpha_k$, for all positive integer k, Then, it is clear that the modified CG algorithm does not require the knowledge of $\tilde{A}$ to proceed. It needs only one extra step for the computation of the new search direction. Instead of using r_k for its computation, we need to use z_k that is a solution of the system $B^2 z_k = r_k$. The conjugate gradient method with preconditioning is given by:

Algorithm Preconditioned CGM:

1. $k \leftarrow 0$
2. $r_0 \leftarrow b - Au^0$
3. $p_0 = (B^2)^{-1} r_0$
4. For $k = 1, 2, \ldots, m$:
5. $\alpha_{k-1} \leftarrow r_{k-1}^T p_{k-1} / \left|\left|p_{k-1}\right|\right|_A^2$
6. $u^k \leftarrow u^{k-1} + \alpha_{k-1} p_{k-1}$
7. $r_k \leftarrow r_{k-1} - \alpha_{k-1} A p_{k-1}$
8. if $\left|\left|r_k\right|\right|_2 \geq \tau$ then
9. solve $B^2 z_k = r^k$
10. $$p_k \leftarrow z_k - \frac{\langle z_k, p_{k-1} \rangle_A}{\left|\left|p_{k-1}\right|\right|_A^2} p_{k-1}$$
11. $k = k + 1$
12. Go to (5)
13. End if
14. End

Analysis of the CGM

Lemma 1.7.7 *If $A \in \mathbb{R}^{n \times n}$ is spd, then for $m = 0, 1, 2, \ldots$ we have*

$$\operatorname{span}\{p_0, p_1, \ldots, p_m\} = \operatorname{span}\{r_0, r_1, \ldots, r_m\} = \operatorname{span}\{r_0, Ar_0, \ldots, A^m r_0\}$$

Proof

1. For $m = 0$, this is trivial.
2. Suppose that for $m = k$ we have
$$\operatorname{span}\{p_0, p_1, \ldots, p_k\} = \operatorname{span}\{r_0, r_1, \ldots, r_k\} = \operatorname{span}\left\{r_0, Ar_0, \ldots, A^k r_0\right\}.$$
3. To complete the proof we should show that the same is true for $m = k + 1$ i.e.
$$\operatorname{span}\{p_0, p_1, \ldots, p_{k+1}\} = \operatorname{span}\{r_0, r_1, \ldots, r_{k+1}\} = \operatorname{span}\left\{r_0, Ar_0, \ldots, A^{k+1} r_0\right\}.$$
We have shown that
$$\begin{aligned} r_{k+1} &= r_k - \alpha_k A p_k \\ p_k &\in \operatorname{span}\left\{r_0, Ar_0, \ldots, A^k r_0\right\} \\ Ap_k &\in \operatorname{span}\left\{Ar_0, A^2 r_0, \ldots, A^{k+1} r_0\right\} \end{aligned}$$
this implies
$$r_{k+1} \in \operatorname{span}\left\{r_0, Ar_0, \ldots, A^{k+1} r_0\right\}$$
so that
$$\operatorname{span}\{r_0, r_1, \ldots, r_{k+1}\} \subset \operatorname{span}\left\{r_0, Ar_0, \ldots, A^{k+1} r_0\right\} \tag{1.11}$$

Now,

$$\begin{aligned} A^k r_0 &\in \operatorname{span}\{p_0, p_1, \ldots, p_k\} \\ A^{k+1} r_0 &\in \operatorname{span}\{Ap_0, Ap_1, \ldots, Ap_k\} \\ r_{k+1} &= r_k - \alpha_k A p_k \end{aligned}$$

So that

$$A^{k+1} r_0 \in \operatorname{span}\left\{ \underbrace{Ap_0}_{(r_1-r_0)/\alpha_0}, \underbrace{Ap_1}_{(r_2-r_1)/\alpha_1}, \ldots, \underbrace{Ap_{k-1}}_{(r_k-r_{k-1})/\alpha_{k-1}}, r_{k+1} - r_k \right\} \subset \operatorname{span}\{r_0, r_1, \ldots r_{k+1}\}.$$

Therefore

$$A^{k+1} r_0 \in \operatorname{span}\{r_0, r_1, \ldots, r_{k+1}\}$$

and by induction

$$\operatorname{span}\left\{r_0, Ar_0, \ldots, A^{k+1} r_0\right\} \subset \operatorname{span}\{r_0, \ldots r_{k+1}\}.$$

The last inclusion together with (1.11) prove that:

$$\operatorname{span}\{r_0, Ar_0, \ldots, A^m r_0\} = \operatorname{span}\{r_0, \ldots r_m\}, \quad \forall m \geq 0.$$

$$\operatorname{span}\{p_0, p_1, \ldots, p_m\} = \operatorname{span}\{r_0, \ldots, r_m\}$$

is proven using

$$p_m = r_m + \beta_{m-1} p_{m-1}$$

and similar arguments as above. ∎

Definition $K_m := \operatorname{span}\{r_0, r_1, \ldots, r_{m-1}\}$ is the m-th *Krylov space* of A.

Theorem 1.7.8 *If A is spd then:*

(A) $\langle p_k, p_m \rangle_A = 0$ *for* $m \neq k$ *and*

(B) $r_k^T r_m = 0$ *for* $m \neq k$.

Proof 1. For $k, m \leq 1$:

$$\begin{aligned} r_1^T r_0 &= 0 \\ \langle p_1, p_0 \rangle_A &= 0 \end{aligned}$$

2. Assume:
   (a) $\langle p_k, p_m \rangle_A = 0$ for $k \neq m$ and $k, m \leq l$
   (b) $r_k^T r_m = 0$ for $k \neq m$ and $k, m \leq l$
3. For $m < l$:

$$r_l^T p_m = r_l^T (c_0 r_0 + c_2 r_2 + \cdots + c_m r_m) = 0$$

from (b) and therefore

$$r_{l+1}^T p_m = (r_l - \alpha_l A p_l)^T p_m = 0.$$

For $m = l$ we have

$$r_{l+1}^T p_l = (r_l - \alpha_l A p_l)^T p_l = r_l^T p_l - \alpha_l \langle p_l, p_l \rangle_A = 0$$

by the definition of α_l. But $r_m \in \text{span}\{p_0, \ldots, p_m\}$ for $m = 1, \ldots, l$ and therefore

$$r_{l+1}^T r_m = 0, \quad m = 0, 1, \ldots, l,$$

which proves (B).

To prove (A), we use:

$$\begin{aligned} p_{l+1} &= r_{l+1} + \beta_l p_l \\ \langle p_{l+1}, p_m \rangle_A &= \langle r_{l+1}, p_m \rangle_A + \beta_l \langle p_l, p_m \rangle_A ; \langle p_l, p_m \rangle_A = 0, \quad m = 0, \ldots, l-1. \end{aligned}$$

But

$$Ap_m \in \text{span}\{r_0, r_1, \ldots, r_{m+1}\}$$

and from (B),

$$\langle r_{l+1}, p_m \rangle_A = 0 \quad \Rightarrow \quad \langle p_{l+1}, p_m \rangle_A = 0 \text{ for } m = 0, 1, \ldots, l-1.$$

By the choice of p_{l+1} in the CGM,

$$\langle p_{l+1}, p_l \rangle_A = 0$$

and so (A) holds.

From this theorem we can conclude that

$$\dim K_n \leq n.$$

Corollary 1.7.9 *If $A \in \mathbb{R}^{n \times n}$ is spd then for some $m \leq n$, $r_m = 0$.*

Theorem 1.7.10 *If A is spd then u^{k+1} minimizes the error $u - u^{k+1} = \epsilon_{k+1}$ over K_{k+1} in $||\cdot||_A$ i.e.*

$$||\epsilon_{k+1}||_A = \min_{v \in K_{k+1}} ||u - v||_A \quad (u \ : \ Au = b)$$

Proof Corollary 1.7.9 concluded that there exist some $m + 1 \leq n$ s.t. $r_{m+1} = 0$. Without loss of generality we can assume that $u^0 = 0$ since any change of the initial iterate can be interpreted as a change in the right hand side vector b: $A(u - u^0) = b - Au^0$, and therefore it does not affect the estimates that follow. Then the exact solution must be a linear combination of the search directions $p_0, \ldots, p_m$ i.e. $u = \sum_{i=0}^{m} \alpha_i p_i$. Consider some $v \in K_{k+1}, k < m + 1$. It must have the form $v = \sum_{i=0}^{k} \gamma_i p_i$. Then we have:

$$\begin{aligned} ||u - v||_A^2 = ||\sum_{i=0}^{k} (\alpha_i - \gamma_i) p_i + \sum_{i=k+1}^{m} \alpha_i p_i||_A^2 = \\ \sum_{i=0}^{k} |\alpha_i - \gamma_i|^2 ||p_i||_A^2 + \sum_{i=k+1}^{m} |\alpha_i|^2 ||p_i||_A^2, \end{aligned}$$

since $\langle p_i, p_j \rangle_A = 0$ if $i \neq j$. Therefore, $||u - v||_A$ has its minimum for $\alpha_i = \gamma_i, i = 1, \ldots, k$. But $u^{k+1} = \sum_{i=0}^{k} \alpha_i p_i$ which concludes the proof.

Definition Π_l denotes the set of all polynomials up to order l:

$$\Pi_l = \left\{ p^l \ : \ a_l x^l + \cdots + a_0 \right\}$$

Corollary 1.7.11 *If $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite with a spectrum $\{\lambda_i\}_{i=1}^n$ then:*

$$||\epsilon_{k+1}||_A \leq ||\epsilon_0||_A \max_{1 \leq j \leq n} |q(\lambda_j)|$$

for all polynomials $q \in \Pi_{k+1}$ such that $q(0) = 1$

Proof From theorem 1.7.10 we have that:

$$||\epsilon_{k+1}||_A \leq ||u-v||_A, \forall v \in K_{k+1}.$$

But

$$v \in K_{k+1} = \text{span}\left\{r_0, Ar_0, \ldots, A^k r_0\right\}$$

Without loss of generality, we can assume that $r_0 = b$ i.e. $u^0 = 0$ and then

$$v = \alpha_0 b + \alpha_1 Ab + \cdots + \alpha_k A^k b = p(A)\, b.$$

Therefore

$$||\epsilon_{k+1}||_A \leq ||u - p(A)\, b||_A, \forall p(A) \in \Pi_k$$

but $u^0 = 0$, $b = Au$, and $\epsilon_0 = u$ so that

$$\begin{aligned} ||\epsilon_{k+1}||_A &\leq \left|\left|\epsilon_0 - p(A)\, A\left(u - u^0\right)\right|\right|_A = ||\epsilon_0 (I - p(A)\, A)||_A \\ &\leq ||\epsilon_0|| \left|\left| \underbrace{I - p(A)\, A}_{\in \Pi_{k+1}} \right|\right|_A = ||\epsilon_0||\, ||q(A)||_A, \forall q(A) \in \Pi_{k+1} \text{ s.t. } q(0) = 1. \end{aligned}$$

Now we will prove that

$$||q(A)||_A = \max_j |q(\lambda_j)|.$$

Since A is symmetric, there exists an orthonormal set of eigenvectors $\{v_j\}_{j=1}^n$ of A i.e.

$$v_i^T v_j = \begin{cases} 0 & \text{if} \quad i \neq j \\ 1 & \text{if} \quad i = j \end{cases}$$

Take $x \in \mathbb{R}^n$ then

$$x = \sum_{i=1}^n c_i v_i$$

and therefore $||x||_A^2 = \sum_{i=1}^n c_i^2 \lambda_i$. Then we have

$$||q(A)||_A^2 = \sup_{||x||_A = 1} ||q(A)\, x||_A{}^2 = \sup_{\sum_{i=1}^n c_i^2 \lambda_i = 1} \left(q(A) \sum_{j=1}^n c_j v_j\right)^T A \left(q(A) \sum_{i=1}^n c_i v_i\right) =$$

$$\sup_{\sum_{i=1}^n c_i^2 \lambda_i = 1} \sum_{i=1}^n q^2(\lambda_i)\, c_i^2 \lambda_i \leq \max_{1 \leq i \leq n} q^2(\lambda_i).$$

If $\max_{1 \leq i \leq n} q^2(\lambda_i)$ is achieved for some index l then it is clear that the equality in the last relation would be achieved if we take $x = v_l / \sqrt{\lambda_l}$, and this concludes the proof of the corollary.

From corollary 1.7.11 it is clear that the sharpest error estimate would be obtained if we pick q to be such that $\max_{1 \leq j \leq n} |q(\lambda_j)|$ is minimal with respect to q. Let us denote the minimum and maximum eigenvalues of A by λ_{min} and λ_{max} correspondingly. Before we construct such a polynomial on the interval $[\lambda_{min}, \lambda_{max}]$ we first consider polynomials on $[-1, 1]$ whose maximum is minimal among all possible polynomials of the same degree i.e. polynomials with a minimal infinity norm. To remind you:

Definition The *infinity norm* of a polynomial on $[-1, 1]$ is defined as

$$||p||_{L^\infty[-1,1]} = \sup_{z \in [-1,1]} |p(z)|$$

If $p(0) = 1$ then $||p||_{L^\infty[-1,1]} \geq 1$ so the minimum for such polynomials would be 1.

The polynomials that have a unit infinity norm on $[-1, 1]$ are the *Chebyshev* polynomials. They are defined recursively as follows:

Definition

$$\begin{aligned} T_0(z) &= 1 \\ T_1(z) &= z \\ T_{n+1}(z) &= 2zT_n(z) - T_{n-1}(z) \end{aligned}$$

Alternatively, we may represent the Chebyschev polynomials as

$$T_k(z) = \frac{1}{2}\left[\left(z + \sqrt{z^2-1}\right)^k + \left(z - \sqrt{z^2-1}\right)^k\right],$$

or

$$T_k(z) = \cos(k \arccos(z)).$$

The polynomials with a minimum infinity norm on $[\lambda_{min}, \lambda_{max}]$, such that they equal 1 at 0, are given by the *Shifted Chebyshev* polynomials:

Definition

$$\tilde{T}_k(z) = \frac{T_k\left(1 - 2\frac{z-\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)}{T_k\left(1 + 2\frac{\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)}$$

Lemma 1.7.12

$$\left|\left|\tilde{T}_k(z)\right|\right|_{L^\infty[\lambda_{min},\lambda_{max}]} = \min_{p\in\Pi_k, p(0)=1} ||p(z)||_{L^\infty[\lambda_{min},\lambda_{max}]}.$$

Proof Since $T_k(z) = \cos(k \arccos(z))$ it is clear that T_k has $k+1$ extrema $T_k(z_i) = (-1)^i, i = 0, 1, \ldots, k$ in the nodes $z_i = \cos(i\pi/k)$ i.e it has $k+1$ alternating minima and maxima on $[-1, 1]$. $\tilde{T}_k$ is just a scaled and shifted version of T_k so it must have exactly the same number of alternating maxima and minima on $[\lambda_{min}, \lambda_{max}]$.

Now assume towards contradiction, that there is some $p_k(z)$ such that $p_k(0) = 1$ and $||p_k(z)||_{L^\infty[\lambda_{min},\lambda_{max}]} < \left|\left|\tilde{T}_k(z)\right|\right|_{L^\infty[\lambda_{min},\lambda_{max}]}$. Then, $p_k - \tilde{T}_k$ must also have at least the same number of alternating extrema as $\tilde{T}_k$ on $[\lambda_{min}, \lambda_{max}]$ since $\max|p_k| < \max|\tilde{T}_k|$ on this interval. This means that $p_k - \tilde{T}_k$ has at least k zeros on $[\lambda_{min}, \lambda_{max}]$. But $p_k(0) - \tilde{T}_k(0) = 0$ and $0 < \lambda_{min}$. So we reach the contradiction that the polynomial of k-th degree $p_k - \tilde{T}_k$ has at least $k+1$ zeros which concludes the proof.

From the corollary we have that

$$||\epsilon_k||_A \leq ||\epsilon_0||_A \max_{\lambda_i} |p(\lambda_i)|$$

for any $p(x) \in \Pi_k$ such that $p(0) = 1$. Therefore,

$$||\epsilon_k||_A \leq ||\epsilon_0||_A \sup_{\lambda_{\min}<z<\lambda_{\max}} |p(z)| = ||\epsilon_0||_A \|p\|_{L^\infty[\lambda_{min},\lambda_{max}]}.$$

Since p can be any polynomial in Π_k such that $p(0) = 1$, the sharpest estimate would be obtained for polynomials with a minimum infinity norm i.e. shifted Chebishev polynomials:

$$||\epsilon_k||_A \leq ||\epsilon_0||_A \max_{\lambda_{\min}<z<\lambda_{\max}} \left|\tilde{T}_k(z)\right|$$

Then,

$$\max_{\lambda_{\min}<z<\lambda_{\max}} \left|\tilde{T}_k(z)\right| = \max \left| \frac{T_k\left(1-2\frac{z-\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)}{T_k\left(1+2\frac{\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)} \right|$$
$$\leq \frac{1}{\left|T_k\left(1+2\frac{\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)\right|} \leq 2\left(\frac{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}\right)^k$$

where we have used

$$\left|T_k\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)\right| = \frac{1}{2}\left(\left(\frac{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}\right)^k + \left(\frac{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}\right)^k\right)$$
$$\geq \frac{1}{2}\left(\frac{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}\right)^k.$$

Then

$$||\epsilon_k||_A \leq 2\,||\epsilon_0||_A \left(\frac{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}\right)^k$$

If A is spd then

$$c_2(A) = ||A||_2 \left|\left|A^{-1}\right|\right|_2 = \frac{\lambda_{\max}}{\lambda_{\min}}$$

and

$$||\epsilon_k||_A \leq 2\,||\epsilon_0||_A \left(\frac{\sqrt{c_2(A)}-1}{\sqrt{c_2(A)}+1}\right)^k.$$

Let $p(\gamma)$ be the smallest integer $k>0$ such that $||\epsilon_k||_A \leq \gamma\,||\epsilon_0||_A$. Then, from the estimate above (which is sharp) it follows that

$$2\left(\frac{1-z}{1+z}\right)^k \leq \gamma, \forall z \in [0,1), \text{ where } z = 1/\sqrt{c_2(A)},$$

or

$$\frac{2}{\gamma} \leq \left(\frac{1+z}{1-z}\right)^k.$$

Note that $z=0$ corresponds to the limit $c_2(A)\to\infty$. Then we get

$$\ln\frac{2}{\gamma} \leq k\ln\frac{1+z}{1-z} = 2k(z+\frac{1}{3}z^3+\frac{1}{5}z^5+\dots), \quad z\in[0,1)$$

or

$$\frac{1}{2z}\ln\frac{2}{\gamma} \leq k(1+\frac{1}{3}z^2+\frac{1}{5}z^4+\dots) = k\frac{arctanh(z)}{z}.$$

But $arctanh(z)/z$ is a monotonically increasing function on $[0,1)$ and its minimum is 1. Therefore, we have;

$$\frac{1}{2}\sqrt{c_2(A)}\ln\frac{2}{\gamma} \leq k,$$

which means that $p(\gamma) \leq 1/2\sqrt{c_2(A)}\ln(2/\gamma)+1$, i.e. for all practical purposes, we need of order of $\sqrt{c_2(A)}$ iterations for satisfying the convergence condition $||\epsilon_k||_A \leq \gamma\,||\epsilon_0||_A$.

Chapter 2

Solutions to Partial Differential Equations

2.1 Classification of Partial Differential Equations

Suppose we have a pde of the form

$$\phi\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial x \partial y}, \frac{\partial^2 u}{\partial y^2}\right) = 0$$

then we classify the equation in terms of the second order derivatives, if they are not present we then classify in terms of the first order derivatives.

2.1.1 First Order linear PDEs

$$\alpha(t, x)\frac{\partial u}{\partial t} + \beta(t, x)\frac{\partial u}{\partial x} = \gamma(t, x)$$

Let us try to reduce it to an ODE over some path $(t(s), x(s))$ in the domain of the equation. We have that:

$$\frac{du}{ds} = \frac{\partial u}{\partial t}\frac{dt}{ds} + \frac{\partial u}{\partial x}\frac{dx}{ds}$$

If we can choose a path $\alpha = \frac{dt}{ds}$ and $\beta = \frac{\partial x}{\partial s}$ then

$$\frac{du}{ds} = \gamma$$

These paths are called *characteristics*.

2.1.2 Second Order PDE

$$\alpha\frac{\partial^2 u}{\partial x^2} + \beta\frac{\partial^2 u}{\partial x \partial y} + \gamma\frac{\partial^2 u}{\partial y^2} = \psi\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)$$

We use a similar idea as in the first order case but now we try to reduce the equation to a system of ODEs for the first derivatives of the solution i.e.

$$\frac{d}{ds}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial^2 u}{\partial x^2}\frac{dx}{ds} + \frac{\partial^2 u}{\partial x \partial y}\frac{dy}{ds}$$
$$\frac{d}{ds}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x \partial y}\frac{dx}{ds} + \frac{\partial^2 u}{\partial y^2}\frac{dy}{ds}$$

Then we may write this system as

$$\begin{bmatrix} \alpha & \beta & \gamma \\ \frac{dx}{ds} & \frac{dy}{ds} & 0 \\ 0 & \frac{dx}{ds} & \frac{dy}{ds} \end{bmatrix} \begin{bmatrix} \frac{\partial^2 u}{\partial x^2} \\ \frac{\partial^2 u}{\partial x \partial y} \\ \frac{\partial^2 u}{\partial y^2} \end{bmatrix} = \begin{bmatrix} \psi \\ \frac{d}{ds}\left(\frac{\partial u}{\partial x}\right) \\ \frac{d}{ds}\left(\frac{\partial u}{\partial y}\right) \end{bmatrix}$$

If we require, similarly to the first-order case, that the original equation is a linear combination of the two equations for the first partial derivative, i. e. this system is linearly dependent, then

$$\alpha\left(\frac{dy}{ds}\right)^2 - \beta\frac{dx}{ds}\frac{dy}{ds} + \gamma\left(\frac{dx}{ds}\right)^2 = 0 \quad \text{and} \quad \alpha\left(\frac{dy}{dx}\right)^2 - \beta\frac{dy}{dx} + \gamma = 0.$$

We classify the equations based on the discriminant

$$\begin{aligned} \beta^2 - 4\alpha\gamma > 0 \quad &\Rightarrow \quad \text{hyperbolic} \quad (2 \quad \text{real} \quad \text{characteristics}) \\ \beta^2 - 4\alpha\gamma = 0 \quad &\Rightarrow \quad \text{parabolic} \quad (1 \quad \text{real} \quad \text{characteristic}) \\ \beta^2 - 4\alpha\gamma < 0 \quad &\Rightarrow \quad \text{elliptic} \quad (0 \quad \text{real} \quad \text{characteristics}) \end{aligned}$$

2.2 Difference Operators

Suppose that we have a grid of nodes in an interval $[a, b]$, $\Delta = \{a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b\}$. The following mapping: $u_h : \Delta \to \mathbb{R}$ is called a grid function: $u_h = (u_{h,0}, \ldots, u_{h,N})^T$ with $u_{h,j} = u_h(x_j)$. A classical function $u : [a, b] \to \mathbb{R}$ gives rise to an associated grid function and, abusing notation somewhat, we denote this grid function by u and its value at a given node x_j by u_j. In addition we will make use of the following operators on such functions, that are used to approximate the corresponding derivatives:

$$\begin{aligned} \delta^+ u_j &= \frac{1}{h_{j+1}}\left(u_{j+1} - u_j\right), \qquad h_{j+1} = x_{j+1} - x_j \\ \delta^- u_j &= \frac{1}{h_j}\left(u_j - u_{j-1}\right) \\ \delta u_{j+\frac{1}{2}} &= \frac{1}{h_{j+1}}\left(u_{j+1} - u_j\right), \\ \delta^2 u_j &= \frac{1}{h_{j+\frac{1}{2}}}\left(\frac{u_{j+1} - u_j}{h_{j+1}} - \frac{u_j - u_{j-1}}{h_j}\right), \qquad h_{j+\frac{1}{2}} = \frac{1}{2}\left(h_j + h_{j+1}\right). \end{aligned}$$

We will also make use of the averaging operator:

$$u_{j+\frac{1}{2}} = \frac{1}{2}\left(u_j + u_{j+1}\right).$$

Assuming that $h_j = h, \forall j$, and enough regularity of the function u, and using Taylor expansions, we can derive the following estimates for the error of these approximations:

$$\begin{aligned} \delta^+ u_j &= \frac{1}{h}\left(u_{j+1} - u_j\right) = \frac{1}{h}\left(u_j + \frac{\partial u_j}{\partial x}h + \frac{\partial^2 u_j}{\partial x^2}\frac{h^2}{2} + \ldots - u_j\right) \\ &= \frac{\partial u_j}{\partial x} + h\frac{\partial^2 u_j}{\partial x^2} + \ldots = \frac{\partial u_j}{\partial x} + \mathcal{O}\left(h\right) \\ \delta^- u_j &= \frac{\partial u_j}{\partial x} + \mathcal{O}\left(h\right) \\ \delta^2 u_j &= \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = \frac{\partial^2 u_j}{\partial x^2} + \mathcal{O}\left(h^2\right) \\ u_{j+\frac{1}{2}} &= \frac{1}{2}\left(u_j + u_{j+1}\right) = u\left(x_{j+\frac{1}{2}}\right) + \mathcal{O}\left(h^2\right). \end{aligned}$$

Similar estimates can be derived for non-constant grid size h_j, however, in the rest of the notes we will consider only equidistant grids. The non-equidistant case requires more careful examination.

2.2.1 Poisson Equation

$$\begin{aligned} -\nabla^2 u(x,y) &= f(x,y), \quad (x,y) \in (0,1)\times(0,1) = \Omega \\ u(x,y) &= g(x,y), \quad (x,y) \in \partial\Omega \end{aligned}$$

Consider the grid:

$$\begin{aligned} \bar{\Omega}_h &= \{(x_j, y_l); x_j = jh,\ y_l = lh, \quad j,l \in 0,\ldots,N\} \\ \Omega_h &= \{(x_j, y_l); x_j = jh,\ y_l = lh, 1 \le j,l \le N-1\} \\ \partial\Omega_h &= \bar{\Omega}_h \setminus \Omega_h, \end{aligned}$$

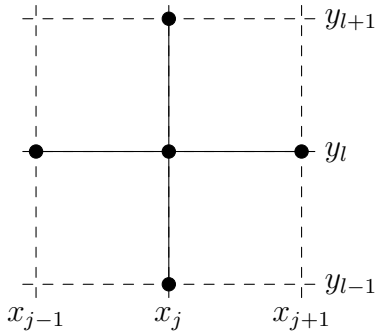
where $h = 1/N$. A second order scheme for an approximation to the solution u_h is given by:

$$\begin{aligned} -\left(\delta_x^2 u_{h,j,l} + \delta_y^2 u_{h,j,l}\right) &= f_{h,j,l}, \qquad (x_j, y_l) \in \Omega_h \\ u_{j,l} &= g_{h,j,l}, \qquad (x_j, y_l) \in \partial\Omega_h \end{aligned}$$

where $f_{h,j,l} = f_{j,l}, (x_j, y_l) \in \Omega_h$; $g_{h,j,l} = g_{j,l}, (x_j, y_l) \in \partial\Omega_h$ So that

$$-\left(u_{h,j-1,l} + u_{h,j,l-1} - 4u_{h,j,l} + u_{h,j,l+1} + u_{h,j+1,l}\right) = h^2 f_{h,j,l}$$

The left hand side is a discretization of $\nabla^2 u$ on the following stencil



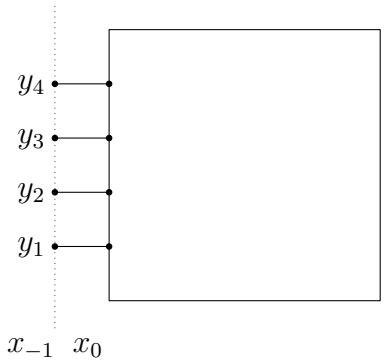
$$\begin{bmatrix} T & I & & & \\ I & T & I & & \\ & \ddots & & & \\ & & I & T & I \\ & & & I & T \end{bmatrix} \begin{bmatrix} u_{h,1} \\ u_{h,2} \\ \vdots \\ u_{h,N-2} \\ u_{h,N-1} \end{bmatrix} = -h^2 \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{N-2} \\ r_{N-1} \end{bmatrix}$$

where

$$T \equiv \begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & & & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix}, \quad u_{h,j} \equiv \begin{bmatrix} u_{h,j,1} \\ u_{h,j,2} \\ \vdots \\ u_{h,j,N-1} \end{bmatrix}, \quad f_j \equiv \begin{bmatrix} f_{j,1} \\ f_{j,2} \\ \vdots \\ f_{j,N-1} \end{bmatrix}, \quad r_j \equiv f_j + \frac{b_j}{h^2}$$

$$b_j \equiv \begin{bmatrix} g_{j,0} \\ 0 \\ \vdots \\ 0 \\ g_{j,N} \end{bmatrix} \text{ for } j = 2,\ldots,N-2, \quad b_1 \equiv \begin{bmatrix} g_{0,1} + g_{1,0} \\ g_{0,2} \\ \vdots \\ g_{0,N-2} \\ g_{0,N-1} + g_{1,N} \end{bmatrix}, \quad b_{N-1} \equiv \begin{bmatrix} g_{N,1} + g_{N-1,0} \\ g_{N,2} \\ \vdots \\ g_{N,N-2} \\ g_{N,N-1} + g_{N-1,N} \end{bmatrix}$$

2.2.2 Neumann Boundary Conditions



Suppose that we need to impose a Neumann condition

$$\left.\frac{\partial u}{\partial n}\right|_{ij} = g|_{ij}$$

on the left boundary of the domain sketched above. Then, we write the scheme for all nodes on this boundary and discretize the Neumann condition using a central difference scheme, introducing an additional layer of points (corresponding to level -1 in x direction)

$$\frac{u_{h,1,l} - u_{h,-1,l}}{2h} = g_{0,l}, \quad u_{h,-1,l} + u_{h,0,l-1} - 4u_{h,0,l} + u_{h,0,l+1} + u_{h,1,l} = -h^2 f_{0,l}.$$

In order to eliminate the additional layer of points, we combine these results to get

$$u_{h,0,l-1} - 4u_{h,0,l} + u_{h,0,l+1} + 2u_{h,1,l} = -h^2 f_{0,l} + 2hg_{0,l}$$

2.3 Consistency and Convergence

Consider the continuous problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega \end{aligned}$$

Now, consider the corresponding discrete problem

$$\begin{aligned} L_h u_h &= f_h \quad \text{in } \Omega_h \\ u_h &= g_h \quad \text{on } \partial\Omega_h. \end{aligned}$$

f_h, g_h are some sort of approximations of f, g and in these notes we assume that $f_h = f, g_h = g$ in the nodes of the discretization grid.

Definition For a given $\phi \in \mathcal{C}^\infty(\Omega)$ and $\boldsymbol{x}_h \in \Omega_h$

$$\tau_h(\boldsymbol{x}_h) \equiv (L - L_h)\,\phi(\boldsymbol{x}_h)$$

is called a *truncation* error of L_h. The *scheme* L_h is *consistent* with L if

$$\lim_{h \to 0} \tau_h(\boldsymbol{x}_h) = 0$$

for all $\boldsymbol{x}_h \in \Omega_h$ and $\phi \in \mathcal{C}^\infty(\Omega)$. If $\tau_h(\boldsymbol{x}_h) = \mathcal{O}(h^p)$, then L_h is *consistent to order p*.

Proposition 2.3.1

$$-\nabla_h^2 = -\delta_x^2 - \delta_y^2$$

is consistent to order 2 *with*

$$-\nabla^2 = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}$$

in Ω_h.

The consistency of a finite difference grid does not guarantee that the solution to the scheme is approximating well the exact solution of the continuous problem. In order to measure how close the solution to the continuous problem is to the solution of the discrete problem we introduce below the notion of convergence. Since the solution of the continuous problem is a function defined everywhere in Ω and the one of the discrete problem is known only in the set of nodes in $\bar{\Omega}_h$ and we need to compare them somehow, we can either extend the definition of u_h to the whole set of points inside $\bar{\Omega}$, or restrict somehow the function u to the nodes in $\bar{\Omega}_h$. The first approach requires the use of some interpolant of u_h, and then we can compare the two solutions in some continuous norm. In the second approach, we need to use some operator P from the space where u belongs, to the vector space of u_h. Assume that the grid $\bar{\Omega}_h$ contains the nodes $\boldsymbol{x}_k, k = 0, \ldots, K$. Then Pu is usually identified with the vector of values of u in all nodes $\boldsymbol{x}_k$ i.e. $Pu = (u(\boldsymbol{x}_0), \ldots, u(\boldsymbol{x}_K))^T$. To be more concise, we often denote $u(\boldsymbol{x}_k)$ by u_k and abusing notations somewhat, we often identify u with Pu if it appears under a discrete norm. Below we follow the second approach and in order to quantify the difference between u and u_h we introduce the notion of convergence as follows:

Definition u_h *converges* to u in a given norm $||.||$ if $\epsilon_h = Pu - u_h$ satisfies

$$\lim_{h \to 0} ||\epsilon_h|| = 0.$$

If $||\epsilon_h|| = \mathcal{O}(h^p)$, then p is the order of convergence.

In the section on finite difference methods we use exclusively the infinity (or maximum) norm: $||\epsilon_h|| = \max_j |\epsilon_{h,j}|$.

Note that consistency does not guarantee that the difference between the solutions of the continuous and discrete problems also tends to zero. It only guarantees that the action of the difference between the differential and discrete operators on a smooth enough function tends to zero.

Let us go back to the boundary value problem for the Poisson equation

$$\begin{aligned} -\nabla^2 u &= f \quad \text{in } \Omega \\ u &= g \quad \text{in } \partial\Omega \end{aligned} \tag{2.1}$$

with the corresponding discrete problem

$$\begin{aligned} -\nabla_h^2 u_h &= f_h \quad \text{in } \Omega_h \\ u_h &= g_h \quad \text{in } \partial\Omega_h. \end{aligned} \tag{2.2}$$

From the elliptic PDEs theory we know the following maximum principle: If $\nabla^2 u \geq 0$ in Ω then u achieves its maximum on $\partial\Omega$. Similarly the discrete Laplacian ∇_h^2 satisfies a Discrete Maximum Principle:

Theorem 2.3.2 (Discrete Maximum Principle) *If*

$$\nabla_h^2 v_{j,l} \geq 0$$

for all $(x_j, y_l) \in \Omega_h$, *then*

$$\max_{(x_j,y_l)\in\Omega_h} v_{j,l} \leq \max_{(x_j,y_l)\in\partial\Omega_h} v_{j,l}$$

Proof

$$\nabla_h^2 v_{j,l} = \frac{v_{j+1,l} + v_{j-1,l} + v_{j,l+1} + v_{j,l-1} - 4v_{j,l}}{h^2} \geq 0 \quad \Rightarrow \quad v_{j,l} \leq \frac{1}{4}\left(v_{j+1,l} + v_{j-1,l} + v_{j,l+1} + v_{j,l-1}\right) \quad \forall (x_j, y_l) \in \Omega_h$$

Suppose that the maximum of v is achieved in an internal point $(x_{\mathcal{J}}, y_{\mathcal{L}}) \in \Omega_h$, then:

$$v_{\mathcal{J},\mathcal{L}} \geq v_{\mathcal{J}-1,\mathcal{L}}, \quad v_{\mathcal{J},\mathcal{L}} \geq v_{\mathcal{J}+1,\mathcal{L}}, \quad v_{\mathcal{J},\mathcal{L}} \geq v_{\mathcal{J},\mathcal{L}-1}, \quad v_{\mathcal{J},\mathcal{L}} \geq v_{\mathcal{J},\mathcal{L}+1}$$

so that

$$4v_{\mathcal{J},\mathcal{L}} \geq v_{\mathcal{J}-1,\mathcal{L}} + v_{\mathcal{J}+1,\mathcal{L}} + v_{\mathcal{J},\mathcal{L}-1} + v_{\mathcal{J},\mathcal{L}+1} \quad \Rightarrow \quad 4v_{\mathcal{J},\mathcal{L}} = v_{\mathcal{J}-1,\mathcal{L}} + v_{\mathcal{J}+1,\mathcal{L}} + v_{\mathcal{J},\mathcal{L}-1} + v_{\mathcal{J},\mathcal{L}+1}$$

Therefore, v must attain its maximum in all neighbouring points too. Applying this argument repeatedly to the neighbours we eventually will reach a boundary point.

Corollary 2.3.3 *The following two results follow from the discrete maximum principle:*

1. $$\begin{aligned} \nabla_h^2 u_h &= 0 \quad \text{in } \Omega_h \\ u_h &= 0 \quad \text{on } \partial\Omega_h \end{aligned}$$

   has a unique solution $u_h = 0$ in $\Omega_h \cup \partial\Omega_h$.

2. *For given f_h and g_h,*

   $$\begin{aligned} \nabla_h^2 u_h &= f_h \quad \text{in } \Omega_h \\ u_h &= g_h \quad \text{on } \partial\Omega_h \end{aligned}$$

   has a unique solution.

Definition If

$$v: \ \Omega_h \cup \partial\Omega_h \to \mathbb{R}$$

then

$$\begin{aligned} ||v||_\Omega &\equiv \max_{(x_j, y_l) \in \Omega_h} |v_{j,l}| \\ ||v||_{\partial\Omega} &\equiv \max_{(x_j, y_l) \in \partial\Omega_h} |v_{j,l}| \end{aligned}$$

Lemma 2.3.4 *If $v_{j,l} = 0$ for all $(x_j, y_l) \in \partial\Omega_h$ then*

$$||v||_\Omega \leq M_\Delta \left|\left|\nabla_h^2 v\right|\right|_\Omega$$

for some $M_\Delta > 0$, that depends on Ω but not on h.

Proof Let

$$\left|\left|\nabla_h^2 v\right|\right|_\Omega = \nu \quad \Rightarrow \quad -\nu \leq \nabla_h^2 v \leq \nu,$$

and consider the function

$$w: \ w_{j,l} = \frac{1}{4}\left(x_j^2 + y_l^2\right) \geq 0.$$

Denote its norm on the boundary by M_Δ i.e.

$$||w||_{\partial\Omega} = M_\Delta > 0.$$

Note that $\nabla_h^2 w = 1$. Then

$$\left.\begin{array}{l} \nabla_h^2 (v + \nu w) \geq 0 \\ \nabla_h^2 (v - \nu w) \leq 0 \end{array}\right\} \text{ in } \Omega$$

From the maximum principle and since $v_h|_{\partial\Omega_h} = 0$ we have:

$$v_{j,l} \leq v_{j,l} + \nu w_{j,l} \leq \nu \left|\left|w_{j,l}\right|\right|_{\partial\Omega} = M_\Delta \left|\left|\nabla_h^2 v_{j,l}\right|\right|_\Omega$$
$$v_{j,l} \geq v_{j,l} - \nu w_{j,l} \geq -\nu \left|\left|w_{j,l}\right|\right|_{\partial\Omega} = -M_\Delta \left|\left|\nabla_h^2 v_{j,l}\right|\right|_\Omega$$

$\forall (x_j, y_l) \in \Omega_h$

Theorem 2.3.5 (error estimate) *Let u be the solution of the boundary value problem (2.1) and u_h is a solution to the discrete analog (2.2). Assuming that $u \in \mathcal{C}^4\left(\bar{\Omega}\right)$, then there exists a constant $k > 0$ such that:*

$$\left|\left|u - u_h\right|\right|_\Omega \leq kMh^2$$

where

$$M \equiv \max\left\{ \left|\left|\frac{\partial^4 u}{\partial x^4}\right|\right|_{L^\infty(\Omega)}, \left|\left|\frac{\partial^4 u}{\partial y^4}\right|\right|_{L^\infty(\Omega)} \right\}$$

Proof From the proposition:

$$\left(\nabla_h^2 - \nabla^2\right) u_{j,l} = \frac{h^2}{12}\left[\frac{\partial^4 u}{\partial x^4}(\xi_j, y_l) + \frac{\partial^4 u}{\partial y^4}(x_j, \eta_l)\right]$$

for some ξ_j, η_l s.t. $x_{j-1} \leq \xi_j \leq x_{j+1}, y_{l-1} \leq \eta_l \leq y_{l+1}$. Taking into account that u is the exact solution we have:

$$-\nabla_h^2 u_{j,l} = f_{j,l} - \frac{h^2}{12}\left[\frac{\partial^4 u}{\partial x^4}(\xi_j, y_l) + \frac{\partial^4 u}{\partial y^4}(x_j, \eta_l)\right]$$

Subtract $-\nabla_h^2 u_{h,j,l} = f_{h,j,l} = f_{j,l}$ and note that $u - u_h = 0$ on $\partial\Omega_h$. Then

$$\nabla_h^2 \left(u_{j,l} - u_{h,j,l}\right) = \frac{h^2}{12}\left[\frac{\partial^4 u}{\partial x^4}(\xi_j, y_l) + \frac{\partial^4 u}{\partial y^4}(x_j, \eta_l)\right]$$

From the Lemma we get

$$\left|\left|u - u_h\right|\right|_\Omega \leq M_\Delta \left|\left|\nabla_h^2 (u - u_h)\right|\right|_\Omega \leq \underbrace{\frac{2M_\Delta}{12}}_{=k} Mh^2 = kMh^2 \quad \blacksquare$$

2.4 Advection Equation

Consider the initial value problem

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad t > 0$$

with $v > 0$, $u(0, x) = f(x)$ for $-\infty < x < \infty$. The exact solution is given by

$$u(t, x) = f(x - vt)$$

Consider a grid $\Delta \equiv \Delta_t \times \Delta_x$:

$$\Delta_t \equiv \{nk \ : \ n = 0, 1, 2, \ldots\}$$
$$\Delta_x \equiv \{jh \ : \ j = 0, \pm 1, \pm 2, \ldots\}$$

One possible discretization is

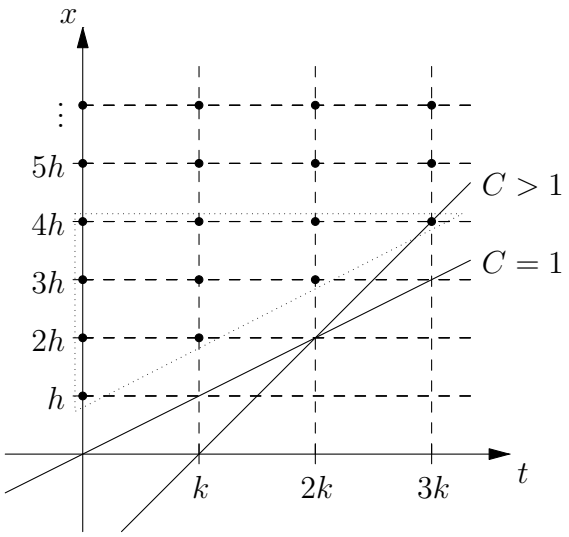
$$\left(\delta_t^+ + v\delta_x^-\right) u_{h,j}^n = \frac{u_{h,j}^n - u_{h,j}^{n-1}}{k} + v\frac{u_{h,j}^{n-1} - u_{h,j-1}^{n-1}}{h} = 0$$

or equivalently

$$u_{h,j}^n = u_{h,j}^{n-1} - C\left(u_{h,j}^{n-1} - u_{h,j-1}^{n-1}\right)$$

This is called the FB (Forwards Backwards) scheme where the Courant number is defined as

$$C \equiv \frac{vk}{h}$$



Assume for simplicity that $v = 1$. As suggested by the exact solution $u(t,x) = f(x-t)$, the solution at any point (t,x) depends only on the value of the initial condition at the point where the characteristic (having a slope of 1) crosses the vertical axes (see the figure above). The characteristic curve therefore is called a domain of dependance of the solution. On the other hand, the numerical solution at each point (t^n, x_j) depends only of the solution at (t^{n-1}, x_{j-1}) and (t^{n-1}, x_j). These values themselves depend on values at time t^{n-2} etc, so that the value at (t^n, x_j) depends on the values in a right angle triangle whose hypothenuse has a slope of C (see again the same figure), called numerical domain of dependance. Clearly, if the Courant number C is greater than 1, the numerical domain of dependance will not contain the point where the characteristic crosses the $x-$axes i.e. it will not account for the initial data that determines the exact solution at a given point. This should create troubles with the numerical solution and these troubles are quantified by the notion of stability. Informally speaking, the solution is stable if the numerical domain of dependance contains the exact domain of dependance. We will formally illustrate these concepts below.

Define S (shift operator) such that

$$Su_{h,j}^n = u_{h,j+1}^n$$

$$\begin{aligned} u_{h,j}^n &= \left(1 - C + CS^{-1}\right) u_{h,j}^{n-1} = \left(1 - C + CS^{-1}\right)^2 u_{h,j}^{n-2} = \cdots = \left(1 - C + CS^{-1}\right)^n u_{h,j}^0 = \left(1 - C + CS^{-1}\right)^n f_j \\ &= \sum_{m=0}^{n} \binom{n}{m} (1-C)^m \left(CS^{-1}\right)^{n-m} f_j = \sum_{m=0}^{n} \binom{n}{m} (1-C)^m C^{n-m} f_{j-n+m} \end{aligned}$$

Now, suppose that the initial condition is perturbed with an error ϵ_j i.e. instead of f_j the initial condition is given by$\hat{f}_j = f_j + \epsilon_j$. Then, the perturbed solution $\hat{u}_h$ is given by:

$$\hat{u}_{h,j} = \sum_{m=0}^{n} \binom{n}{m} (1-C)^m C^{n-m} \left(f_{j-n+m} + \epsilon_{j-n+m}\right)$$

and we have

$$\left|u_{h,j}^n - \hat{u}_{h,j}^n\right| = \left|\sum_{m=0}^{n} \binom{n}{m} (1-C)^m C^{n-m} \epsilon_{j-n+m}\right|$$
$$\leq \sum_{m=0}^{n} \binom{n}{m} |1-C|^m C^{n-m} |\epsilon_{j-n+m}| \leq ||\epsilon||_\infty \sum_{m=0}^{n} \binom{n}{m} |1-C|^m C^{n-m} = ||\epsilon||_\infty (|1-C|+C)^n$$

So that

$$\left|u_{h,j}^n - \hat{u}_{h,j}^n\right| \leq ||\epsilon||_\infty (|1-C|+C)^n$$

Then, if $C \leq 1$:

$$\left|u_{h,j}^n - \hat{u}_{h,j}^n\right| \leq ||\epsilon||_\infty$$

However, if $C > 1$:

$$\left|u_{h,j}^n - \hat{u}_{h,j}^n\right| \leq ||\epsilon||_\infty (|1-C|+C)^n = ||\epsilon||_\infty (2C-1)^n \overset{n\to\infty}{\to} \infty$$

Theorem 2.4.1 *If $C \leq 1$ and $u \in \mathcal{C}^2((0,t] \times \mathbb{R})$ then the FB scheme is convergent. That is*

$$E^n = ||u^n - u_h^n||_\infty \leq t^n \mathcal{O}(k+h)$$

where $t^n = kn$.

Proof

$$\delta_t^+ u(t^n, x_j) = \frac{\partial u}{\partial t}(t^n, x_j) + \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(\eta, x_j)$$
$$\delta_x^- u(t^n, x_j) = \frac{\partial u}{\partial x}(t^n, x_j) - \frac{h}{2}\frac{\partial^2 u}{\partial x^2}(t^n, \xi)$$
$$\delta_t^+ u + v\delta_x^- u = -\tau_{k,h}(t^n, x_j) = \mathcal{O}(k+h)$$
$$\delta_t^+ u_{h,j}^n + v\delta_x^- u_{h,j}^n = 0$$

and letting $\epsilon_h = u - u_h$ we get the following estimate:

$$\left|\epsilon_{h,j}^{n+1}\right| \leq \left|(1-C)\epsilon_{h,j}^n\right| + C\left|\epsilon_{h,j-1}^n\right| + k\left|\tau_{k,h}(t^n, x_j)\right|$$

Let

$$E^n = ||\epsilon_h^n||_\infty, \quad T^n = \max_j |\tau_{k,h}(t^n, x_j)|$$

Then from the last inequality we can conclude that:

$$E^{n+1} \leq (1-C)E^n + CE^n + k|\tau_{k,h}(t^n, x_j)| \leq E^n + kT^n \leq E^{n-1} + kT^n + kT^{n-1} \leq E^0 + k\sum_{m=1}^{n} T^m$$

If $E^0 = 0$ then

$$E^{n+1} \leq k\sum_{m=1}^{n} T^m \leq nk \max_{1\leq m\leq n} T^m = t^n \max_m T^m = t^n \mathcal{O}(k+h)$$

2.5 Von Neumann Stability Analysis

Definition Discrete L^2 norm:

$$||v||_2^2 \equiv h\sum_{j=-\infty}^{\infty} |v_j|^2$$

Definition A FD (Finite Difference) scheme for a time dependent PDE is *stable* if for any time $T > 0$, there exists $K > 0$ such that for any initial data u_h^0, the sequence $\{u_h^n\}$ satisfies:

$$||u_h^n||_2 \leq K \left|\left|u_h^0\right|\right|_2$$

for $0 \leq nk \leq T$. The constant K is independent of the time and space mesh size!

Consider the following grid on the real axes

$$\Delta_x = \{\ldots, -2h, -h, 0, h, 2h, \ldots\}.$$

Then the Fourier Transform of a given grid function v is defined as

$$(Fv)(\tau) \equiv \frac{h}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} e^{-ijh\tau} v_j$$

Inverse transform:

$$v_j = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{ijh\xi} (Fv)(\xi)\, d\xi$$

Theorem 2.5.1 (Parseval Identity)

$$||Fv||^2_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]} := \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} (Fv)^2\, d\xi = ||v||_2^2$$

Corollary 2.5.2 *A finite difference scheme is stable if and only if $\exists K > 0$, independent of k, h, s.t.:*

$$||Fu_h^n||_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]} \leq K \left|\left|Fu_h^0\right|\right|_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]}$$

Consider the forward backwards scheme. Since we have:

$$\begin{aligned}
u_{h,j}^n &= (1 - C)\, u_{h,j}^{n-1} + C u_{h,j-1}^{n-1} \\
u_{h,j}^n &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{ijh\xi} (Fu_h^n)(\xi)\, d\xi \\
u_{h,j-1}^{n-1} &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{i(j-1)h\xi} \left(Fu_h^{n-1}\right)(\xi)\, d\xi,
\end{aligned}$$

we obtain:

$$Fu_h^n = \underbrace{\left(1 - C + Ce^{-ih\xi}\right)}_{\text{Amplification Factor}} \left(Fu_h^{n-1}\right)(\xi)$$

$$A(h\xi) = 1 - C + Ce^{-ih\xi}$$

Then, since we have

$$||Fu_h^n||^2_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]} = \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} A(h\xi)^2 \left(Fu_h^{n-1}\right)^2 d\xi,$$

if $|A(h\xi)| \leq 1$, we obtain that:

$$||Fu_h^n||_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]} \leq \left|\left|Fu_h^{n-1}\right|\right|_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]}$$

and recursive application gives:

$$||Fu_h^n||_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]} \leq \left|\left|Fu_h^0\right|\right|_{L^2\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]}.$$

Since

$$A(\theta) = 1 - C + Ce^{-i\theta}$$

$$\begin{aligned}|A(\theta)|^2 &= (1 - C + C\cos\theta)^2 + (-C\sin\theta)^2 \\ &= 1 - 2C + C^2 + 2C\cos\theta - 2C^2\cos\theta + C^2\cos^2\theta + C^2\sin^2\theta \\ &= 1 - 2C(1 - \cos\theta) + 2C^2(1 - \cos\theta) = 1 + 2C(1 - \cos\theta)(C - 1) \\ &\leq 1 + 4C(C - 1)\end{aligned}$$

it follows that if $C > 1$ then $|A(\theta)| > 1$, and if $C \leq 1$ then $|A(\theta)| \leq 1$. Also,

$$\left|\left|1 - C + Ce^{-i\theta}\right|\right| \leq |1 - C| + C$$

If a numerical scheme requires a solution of a linear system with a non-diagonal matrix then we call it an *implicit* scheme, otherwise it is called an *explicit* scheme.

Consider the following "Backwards Backwards" scheme

$$\begin{aligned}\frac{u_{h,j}^n - u_{h,j}^{n-1}}{k} + v\frac{u_{h,j}^n - u_{h,j-1}^n}{h} &= 0 \\ (1 + C)u_{h,j}^n - Cu_{h,j-1}^n &= u_{h,j}^{n-1}\end{aligned}$$

This is an *implicit* scheme.

In order to study its stability (and the stability of any other FD scheme), it is enough, instead of considering the full Fourier transform of the solution, to consider the "action" of the scheme on a single mode of the form:

$$w^n e^{ijh\xi}$$

then plugging into the BB scheme:

$$\begin{aligned}w^n \underbrace{\left[1 + C\left(1 - e^{-ih\xi}\right)\right]}_{A^{-1}(h\xi)} &= w^{n-1} \\ w_n &= A(h\xi)\, w^{n-1} \\ A(\theta) &= \left[1 + C\left(1 - e^{-i\theta}\right)\right]^{-1}\end{aligned}$$

Then,

$$\left|1 + C\left(1 - e^{-ih\xi}\right)\right| \geq |1 + C| - \left|Ce^{-ih\xi}\right| = 1$$

so that

$$|A(\theta)| \leq 1 \quad \forall\theta$$

therefore the BB scheme is unconditionally stable.
We may outline a procedure for von Neumann stability analysis:

1. Consider a single mode $w^n e^{ijh\xi}$.
2. Derive conditions for A to satisfy $|A(h\xi)| \leq 1$

2.6 Sufficient Conditions for Convergence

More generally, if we approximate the solution of a well-posed initial-boundary or boundary value problem comprised by the equation:

$$Lu = f \text{ in } \Omega$$

and some initial and/or boundary conditions:

$$lu = g,$$

by a numerical scheme with a grid size h (this can be a vector of grid sizes in each direction of discretization):

$$\begin{aligned} L_h u_h &= f_h \text{ in } \Omega_h \\ l_h u_h &= g_h \end{aligned} \tag{2.3}$$

we need to generalize the notion of stability as follows. Assume that u_h, f_h, and g_h belong to three Banach spaces B_h^1, B_h^2, and B_h^3, with norms $\|.\|_{h,1}$, $\|.\|_{h,2}$, and $\|.\|_{h,3}$ correspondingly. Then we have the following generalized definition of stability:

Definition (Stable scheme) If the scheme (2.3) has a unique solution u_h, it is stable if $\exists C_1, C_2 > 0$, independent of h, s.t.:

$$\|u_h\|_{h,1} \leq C_1 \|f_h\|_{h,2} + C_2 \|g_h\|_{h,3}.$$

The essence of this definition is that it guarantees that a small perturbation in the data of the problem i.e. in f_h, g_h leads to a small perturbation in the solution u_h.

We can also generalize the notion of consistency as follows. Suppose that P^1 is an operator from a given Banach space B^1, containing the solution of the continuous problem u, into B_h^1. Then the discrete operators L_h and l_h are consistent with the continuous counterparts if:

$$\begin{aligned} L_h(P^1 u - u_h) &= \phi_h \text{ in } \Omega_h \\ l_h(P^1 u - u_h) &= \psi_h, \end{aligned} \tag{2.4}$$

and

$$\|\phi_h\|_{h,2} \underset{|h|\to 0}{\longrightarrow} 0, \quad \|\psi_h\|_{h,3} \underset{|h|\to 0}{\longrightarrow} 0.$$

Note that h can be a vector of grid sizes, if the problem involves more than one variable. If $\phi_h = O(|h|^m), \psi_h = O(|h|^k)$, then the scheme is consistent to order $min(k, m)$. Note that this definition is consistent with the previous definition that we used, provided that the the data of the discrete problem exactly matches the data of the continuous problem in the points of the grid.

The following theorem states that consistency and stability are sufficient for obtaining convergence.

Theorem 2.6.1 *Given a scheme that is consistent in a norm $\|.\|_{h,1}$, then it is convergent in this norm, if it is stable. If the scheme is consistent to order k then it is convergent to order k.*

Proof If the scheme is consistent we have that:

$$\begin{aligned} L_h(P^1 u - u_h) &= \phi_h \text{ in } \Omega_h \\ l_h(P^1 u - u_h) &= \psi_h. \end{aligned}$$

The stability guarantees that:

$$\|P^1 u - u_h\|_{h,1} \leq C_1 \|\phi_h\|_{h,2} + C_2 \|\psi_h\|_{h,3},$$

and this immediately yields convergence in the norm $\|.\|_{h,1}$.

It appears that the stability and consistency are also necessary conditions for convergence, as stated in the following theorem due to P. Lax. We will skip the proof in these notes.

Theorem 2.6.2 (Lax Theorem) *Given a consistent scheme for a well-posed initial boundary value problem, then stability is a necessary and sufficient condition for convergence.*

2.7 Parabolic PDEs – Heat Equation

$$\begin{cases} \frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}, & (t, x) \in (0, T] \times \mathbb{R} \\ u(0, x) = f(x), & x \in \mathbb{R} \end{cases}$$

Which has the solution

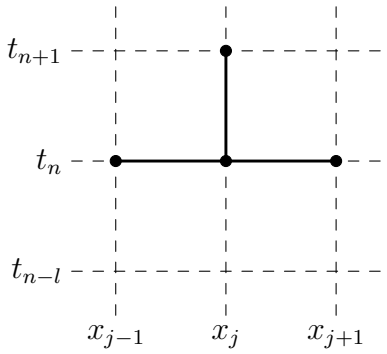
$$u(t, x) = \frac{1}{\sqrt{4\pi D t}} \int_{-\infty}^{\infty} f(\xi)\, e^{-\frac{(x-\xi)^2}{4Dt}}\, d\xi$$

2.7.1 FC Scheme

This is an explicit scheme.

$$\delta_t^+ u_{h,j}^n - D\delta_x^2 u_{h,j}^n = 0$$
$$u_{h,j}^{n+1} = (1 - 2\Gamma D)\, u_{h,j}^n + \Gamma D \left(u_{h,j+1}^n + u_{h,j-1}^n\right), \qquad \Gamma = \frac{k}{h^2}$$

where Γ is called a grid ratio.



The consistency error is found using

$$\partial_t u - \delta_t^+ u = \mathcal{O}(k)$$
$$\partial_{xx} u - \delta_x^2 u = \mathcal{O}\left(h^2\right)$$

so that

$$\tau_{k,h} = \mathcal{O}\left(k + h^2\right).$$

Stability is a central concern of parabolic and hyperbolic PDEs. Using the Neumann analysis idea we substitute the Fourier mode:

$$v_j^n = w^n e^{ij\theta}$$

into the scheme to get:

$$\frac{1}{k}\left(w^{n+1} - w^n\right) e^{ij\theta} = \frac{D}{h^2}\left(w^n e^{i\theta} - 2w^n + w^n e^{-i\theta}\right) e^{ij\theta}$$

This gives:

$$w^{n+1} = \underbrace{\left(1 - 4D\Gamma \sin^2\frac{\theta}{2}\right)}_{A(\theta)\le 1} w^n$$

So that

$$0 < D\Gamma \le \frac{1}{2} \quad \Leftrightarrow \quad |A(\theta)| \le 1 \quad \Rightarrow \quad \text{stability}$$

and we require:

$$\frac{Dk}{h^2} \le \frac{1}{2} \quad \Rightarrow \quad k \le \frac{1}{2}\frac{h^2}{D}$$

and we note that k and h are not of the same order. So, this is a conditionally stable scheme, but it is explicit and this is to be expected.

In this problem the domain Ω is the entire real axes and therefore it is convenient to use for measuring the error the following ∞-norm: $||v_h||_\Omega = \max_{-\infty<j<\infty} |v_{h,j}|$. Then the following theorem provides the convergence estimate for the FC scheme.

Theorem 2.7.1 *If $D\Gamma \leq \frac{1}{2}$, the exact solution u is sufficiently smooth, and if u_h satisfies exactly the initial condition then:*

$$||u - u_h||_\Omega = ||\epsilon_h||_\Omega = \mathcal{O}\left(k + h^2\right)$$

Proof From the scheme, if the exact solution u is sufficiently smooth*, we have:

$$u_j^{n+1} = D\Gamma u_{j-1}^n + (1 - 2D\Gamma)\, u_j^n + D\Gamma u_{j+1}^n + \mathcal{O}\left(k^2 + kh^2\right)$$

so that:

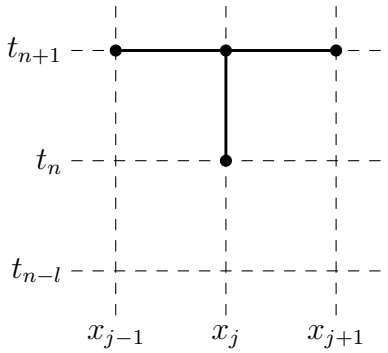
$$\epsilon_{h,j}^{n+1} = \underbrace{D\Gamma}_{\geq 0} \epsilon_{h,j-1}^n + \underbrace{(1 - 2D\Gamma)}_{\geq 0} \epsilon_{h,j}^n + \underbrace{D\Gamma}_{\geq 0} \epsilon_{h,j+1}^n + \mathcal{O}\left(k^2 + kh^2\right)$$

by the triangle inequality,

$$\left|\left|\epsilon_h^{n+1}\right|\right|_\Omega \leq ||\epsilon_h^n||_\Omega + \mathcal{O}\left(k^2 + kh^2\right) \leq \left|\left|\epsilon_h^0\right|\right|_\Omega + (n+1)\,\mathcal{O}\left(k^2 + kh^2\right) = \underbrace{(n+1)\,k}_{t^{n+1}}\,\mathcal{O}\left(k + h^2\right).$$

Note that the the right hand side of this estimate will blow up as $t \to \infty$ for fixed k, h.

2.7.2 BC scheme



$$\delta_t^- u_{h,j}^{n+1} = D\delta_x^2 u_{h,j}^{n+1}$$

This is a $\mathcal{O}\left(k + h^2\right)$ consistent scheme. It can be rewritten as:

$$-D\Gamma u_{h,j-1}^{n+1} + (1 + 2\Gamma D)\, u_{h,j}^{n+1} - D\Gamma u_{h,j+1}^{n+1} = u_{h,j}^n$$

So we must invert a tridiagonal matrix. Now we will study its stability. Substituting

$$v_j^n = w^n e^{ij\theta}$$

into the scheme we have:

$$w^{n+1} - w^n = D\Gamma\left(e^{i\theta} - 2 + e^{-i\theta}\right) w^{n+1}$$

So that:

$$w^{n+1} = \underbrace{\frac{1}{1 + 4D\Gamma \sin^2\frac{\theta}{2}}}_{A(\theta)\leq 1} w^n.$$

So, it is clear that the BC scheme is unconditionally stable.

*This term is used to require enough of the derivatives of the exact solution to be bounded, so that the coefficients in the function $\mathcal{O}\left(k^2 + kh^2\right)$ appearing in the consistency estimate below to be finite.

2.7.3 Crank-Nicolson Scheme

It is given by

$$\delta_t^- u_{h,j}^{n+1} = \frac{1}{2} D \left(\delta_x^2 u_{h,j}^{n+1} + \delta_x^2 u_{h,j}^n \right),$$

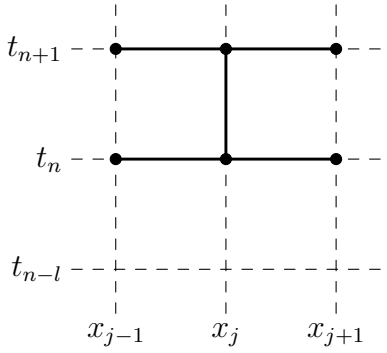
or

$$\frac{u_{h,j}^{n+1} - u_{h,j}^n}{k} = \frac{D}{2h^2} \left(u_{h,j-1}^{n+1} - 2u_{h,j}^{n+1} + u_{h,j+1}^{n+1} + u_{h,j-1}^n - 2u_{h,j}^n + u_{h,j+1}^n \right)$$

Consistency:

$$\begin{aligned} \frac{D}{2} \left(\delta_x^2 u_j^{n+1} + \delta_x^2 u_j^n \right) &= \frac{D}{2} \left(\left. \frac{\partial^2 u}{\partial x^2} \right|_j^{n+1} + \mathcal{O}\left(h^2\right) + \left. \frac{\partial^2 u}{\partial x^2} \right|_j^n + \mathcal{O}\left(h^2\right) \right) \\ &= D \left. \frac{\partial^2 u}{\partial x^2} \right|_j^{n+\frac{1}{2}} + \mathcal{O}\left(k^2 + h^2\right) \end{aligned}$$

$$\frac{u_j^{n+1} - u_j^n}{k} = \left. \frac{\partial u}{\partial t} \right|_j^{n+\frac{1}{2}} + \mathcal{O}\left(k^2\right)$$



Neumann stability analysis: First, rewrite the scheme in the following two-stage form

$$\frac{u_{h,j}^{n+\frac{1}{2}} - u_{h,j}^n}{\frac{k}{2}} = D \frac{u_{h,j-1}^n - 2u_{h,j}^n + u_{h,j+1}^n}{h^2} \rightarrow A_1(\theta) = 1 - 4D\frac{\Gamma}{2} \sin^2 \frac{\theta}{2}$$

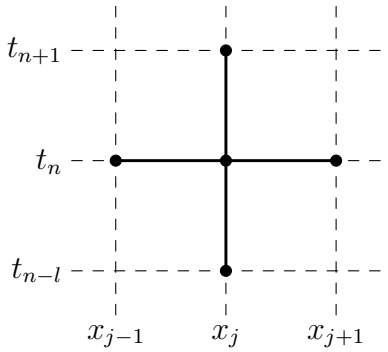
$$\frac{u_{h,j}^{n+1} - u_{h,j}^{n+\frac{1}{2}}}{\frac{k}{2}} = D \frac{u_{h,j-1}^{n+1} - 2u_{h,j}^{n+1} + u_{h,j+1}^{n+1}}{h^2} \rightarrow A_2(\theta) = \frac{1}{1 + 4D\frac{\Gamma}{2} \sin^2 \frac{\theta}{2}}.$$

Then,

$$w^{n+\frac{1}{2}} = A_1(\theta) w^n \quad w^{n+1} = A_2(\theta) w^{n+\frac{1}{2}} = \underbrace{A_1 A_2}_{A(\theta)} w^n$$

So that $A(\theta) \leq 1$, so the system is unconditionally stable, and $\mathcal{O}\left(h^2 + k^2\right)$.

2.7.4 Leapfrog Scheme



As we use central difference in both directions, the scheme is $\mathcal{O}\left(k^2+h^2\right)$. We may write the scheme as:

$$\frac{u_{h,j}^{n+1}-u_{h,j}^{n-1}}{2k}=D\frac{u_{h,j-1}^{n}-2u_{h,j}^{n}+u_{h,j+1}^{n}}{h^2}$$

Now we preform stability analysis: $v_j^n=w^n e^{ij\theta}$

$$w^{n+1}-w^{n-1}=2D\Gamma\left(e^{-i\theta}-2+e^{i\theta}\right)w^n$$

Then, assume that:

$$w^n=A\left(\theta\right)w^{n-1}\quad\text{ and }w^{n+1}=A\left(\theta\right)w^n$$

so that

$$\left(A^2-1\right)w^{n-1}=4D\Gamma\left(\cos\theta-1\right)Aw^{n-1}$$

this gives

$$A_{1,2}=-4D\Gamma\sin^2\frac{\theta}{2}\pm\sqrt{1+16D^2\Gamma^2\sin^4\frac{\theta}{2}}$$

clearly $|A_2|\geq 1$, and $|A_2|>1$ for some θ, so the scheme is unconditionally unstable.

2.7.5 DuFort-Frankel Scheme

$$\delta_t u_h^n=\frac{D}{h^2}\left[u_{h,j+1}^n-\left(u_{h,j}^{n+1}+u_{h,j}^{n-1}\right)+u_{h,j-1}^n\right]$$

Since

$$\frac{1}{2}\left(u_{h,j}^{n+1}+u_{h,j}^{n-1}\right)=u_{h,j}^n+\mathcal{O}\left(k^2\right)$$

this scheme can be considered as a stabilized version of the Leapfrog scheme. The truncation error is then:

$$\tau_{k,h}\left(t,x\right)=2\left(\frac{k}{h}\right)^2\frac{\partial^2 u}{\partial t^2}\left(t_n,x_j\right)+\frac{k^3}{3}\frac{\partial^2 u}{\partial t^2}\left(t_n,x_j\right)-\frac{h^2}{12}\frac{\partial^2 u}{\partial x^2}\left(t_n,x_j\right)+\mathcal{O}\left(k^3+h^3\right)$$

and the system is conditionally consistant if $\frac{k}{h}\to 0$. If k, h go to zero but $\frac{k}{h}\to c^2$, then DF is consistent with:

$$\frac{\partial u}{\partial t}+c^2\frac{\partial^2 u}{\partial t^2}=D\frac{\partial^2 u}{\partial x^2}.$$

The stability analysis reveals that this scheme is unconditionally stable.

2.8 Advection-Diffusion Equation

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = D\frac{\partial^2 u}{\partial x^2} \qquad (t,x) \in (0,T] \times (0,L)$$

and $v, D > 0$. Then define $\tau = \frac{vt}{L}$, $\xi = \frac{x}{L}$, and Peclet number $Pe = \frac{vL}{D}$. So the PDE becomes:

$$\frac{\partial u}{\partial \tau} + \frac{\partial u}{\partial \xi} = \frac{1}{Pe}\frac{\partial^2 u}{\partial \xi^2}, \quad (\tau,\xi) \in \left(0, \frac{vT}{L}\right] \times (0,1)$$

2.8.1 FC Scheme

$$\frac{u_{h,j}^{n+1} - u_{h,j}^n}{k} + \frac{u_{h,j+1}^n - u_{h,j-1}^n}{2h} - \frac{1}{Pe}\frac{u_{h,j-1}^n - 2u_{h,j}^n + u_{h,j+1}^n}{h^2} = 0$$

This scheme is consistant to $\mathcal{O}\left(k + h^2\right)$. Now preform stability analysis: $v_j^n = w^n e^{ij\theta}$

$$w^{n+1} - w^n + k\frac{e^{i\theta} - e^{-i\theta}}{2h}w^n - \frac{1}{Pe}k\frac{e^{-i\theta} - 2 + e^{i\theta}}{h^2} = 0$$

$$w^{n+1} = w^n \underbrace{\left(-Ci\sin\theta + \frac{2\Gamma}{Pe}\left(\cos\theta - 1\right) + 1\right)}_{A(\theta)}$$

So we may write,

$$\left|A\left(\theta\right)\right|^2 = \left(1 - 4\frac{\Gamma}{Pe}\sin^2\frac{\theta}{2}\right)^2 + C^2\sin^2\theta$$

If $\Gamma/Pe \leq 1/2$ the first term ≤ 1 and since $C^2 = k\Gamma$ then $C^2\sin^2\theta \leq Pek/2 = Mk, M > 0$, i.e.

$$\left|A\right|^2 \leq 1 + Mk$$

Theorem 2.8.1 *The FC scheme is stable if $\Gamma \leq Pe/2$ i.e. if $\left|A\left(\theta\right)\right|^2 \leq 1 + Mk$ for some $M > 0$.*

Proof

$$\left(Fu_h^n\right)^2(\xi) = A^2\left(h\xi\right)\left(Fu_n^{n-1}\right)^2(\xi) = A^{2n}\left(h\xi\right)\left(Fu_h^0\right)^2(\xi) \leq \left(1 + Mk\right)^n\left(Fu_h^0\right)^2(\xi)$$

Therefore,

$$||Fu_h^n||_2^2 \leq \left(1 + Mk\right)^n\left|\left|Fu_h^0\right|\right|^2 \leq \left(1 + Mk\right)^{\frac{nkM}{kM}}\left|\left|Fu_h^0\right|\right|_2^2 = \left(\left(1 + Mk\right)^{\frac{1}{kM}}\right)^{Mt^n}\left|\left|Fu_h^0\right|\right|_2^2 \leq e^{Mt^n}\left|\left|Fu_h^0\right|\right|_2^2, \quad \blacksquare$$

where $t^n = nk$.

Theorem 2.8.2 *Assume:*

$$\frac{\Gamma}{Pe} \leq \frac{1}{2}$$

Then the solution of the FC scheme satisfies the maximum principle:

$$\max_j\left|u_{h,j}^{n+1}\right| \leq \max_j\left|u_{h,j}^n\right|$$

for all $n \geq 1$, if and only if $h \leq \frac{2}{Pe}$[†].

[†]Here we tacitly disregard the boundary conditions that are required since the equation involves a second derivative in space. It is quite clear, however, that the boundary conditions should be non-increasing functions of time. For example, if at the left edge of the domain we need to satisfy a Dirichlet condition that is an increasing function of time, and the approximation satisfies it exactly, then it cannot satisfy such a maximum principle in time.

Proof Let $h \leq \frac{2}{Pe}$, then:

$$\left|u_{h,j}^{n+1}\right| \leq \frac{\Gamma}{Pe}\left(1+\frac{1}{2}hPe\right)\left|u_{h,j-1}^{n}\right| + \left(1-2\Gamma Pe^{-1}\right)\left|u_{h,j}^{n}\right| + \frac{\Gamma}{Pe}\left(1-\frac{1}{2}hPe\right)\left|u_{h,j+1}^{n}\right| \leq \max_j \left|u_{h,j}^{n}\right|$$

because $h \leq \frac{2}{Pe}$, $\frac{\Gamma}{Pe} \leq \frac{1}{2}$.

Now, assume that

$$\max_j \left|u_{h,j}^{n+1}\right| \leq \max_j \left|u_{h,j}^{n}\right|.$$

Aiming towards a contradiction, assume that $h > \frac{2}{Pe}$. Then, choose the initial data as follows

$$u_{h,j}^{0} = \begin{cases} 1, & j = 0,1 \\ 0, & j > 1 \end{cases}$$

so that

$$u_{h,1}^{1} = \frac{\Gamma}{Pe}\left(1+\frac{1}{2}hPe\right) + 1 - 2\frac{\Gamma}{Pe} > \frac{\Gamma}{Pe}\left(1+\frac{1}{2}\frac{2}{Pe}Pe - 2\right) + 1 = 1$$

and this gives:

$$\max_j \left|u_{h,j}^{1}\right| > 1 = \max_j \left|u_{h,j}^{0}\right|$$

so, by contradiction, the theorem holds.

Since violation of the maximum principle leads to unphysical solution then we must choose $h \leq 2/Pe$ and so, if the Peclet number Pe is very large then we need to use very small h, of the order of Pe^{-1}. This corresponds to problems with boundary layers that need to be resolved by h. However, it is known from the properties of the corresponding boundary value problem that the thickness of such boundary layers is of the order of $Pe^{-1/2}$ i.e. it can be resolved by h being of order of $Pe^{-1/2}$. So, this scheme requires the use of a spatial step h much less than what is needed to resolve the actual solution. The next scheme cures this problem at the expense of loss of accuracy.

2.8.2 Upwinding Scheme (FB)

$$\frac{u_{h,j}^{n+1} - u_{h,j}^{n}}{k} + \frac{u_{h,j}^{n} - u_{h,j-1}^{n}}{h} = \frac{1}{Pe}\frac{u_{h,j-1}^{n} - 2u_{h,j}^{n} + u_{h,j+1}^{n}}{h^2}$$

Theorem 2.8.3 *Assume that $\frac{\Gamma}{Pe} \leq \frac{1}{2}$. Then the FB scheme satisfies a maximum principle provided that:*

$$2\frac{\Gamma}{Pe} + C \leq 1, \quad \left(C = \frac{k}{h}\right)$$

The proof is left as an exercise. This theorem implies that the maximum principle is satisfied if $k \leq (Peh^2)/(2+Peh)$. Note that in the limit $Pe \to \infty$, this restriction on k tends to h. On the other hand, the stability condition $k \leq Peh^2/2$ implies that if $Pe \to \infty$ but $Peh^2 \to const$ (i.e. $h = O(Pe^{-1/2})$), k is restricted by a constant. In conclusion, if the Peclet number Pe is very large, the scheme guarantees that the solution satisfies a maximum principle, similar to the exact solution, and is stable if $k \leq h$, but k can be of the order of h. And this is true even if $h = O(Pe^{-1/2})$ that is sufficient to resolve boundary layers.

The scheme has a consistency error

$$\tau_{k,h}(t,x) = \mathcal{O}(k+h)$$

as

$$\delta_t^+ u_j^n = \frac{\partial u}{\partial t}(t_n, x_j) + \mathcal{O}(k)$$

and

$$\delta_x^- u_j^n = \frac{\partial u}{\partial x}(t_n, x_j) - \frac{h}{2}\frac{\partial^2 u}{\partial x^2}(t_n, x_j) + \mathcal{O}\left(h^2\right)$$

and

$$\delta_x^2 u_j^n = \frac{\partial^2 u}{\partial x^2}(t_n, x_j) + \mathcal{O}\left(h^2\right).$$

Combining these results we have:

$$\delta_t^+ u_j^n + \delta_x^- u_j^n - \frac{1}{Pe}\delta_x^2 u_j^n = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} - \left(\frac{1}{Pe} + \frac{h}{2}\right)\frac{\partial^2 u}{\partial x^2} + \mathcal{O}\left(k + h^2\right)$$

i.e. up to terms that are $O(k + h^2)$, the scheme is consistent with the equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} - \left(\frac{1}{Pe} + \frac{h}{2}\right)\frac{\partial^2 u}{\partial x^2} = 0.$$

This means that the scheme adds an extra diffusion term that is of order of h. If the leading order term in the spatial error is proportional to a derivative of an even order, we call the scheme dissipative since such schemes usually dissipate sharp changes (large gradients) of the solution. If the leading order term of the error is proportional to a derivative of an odd order, the scheme is called dispersive. This is because odd derivatives in PDEs lead to the so called dispersion in the solution. In terms of a Fourier decomposition of the solution, this effect is manifested in the fact that due to the presence of odd derivatives different Fourier modes travel with a different speed.

Chapter 3

Introduction to Finite Elements

3.1 Weighted Residual Methods

3.1.1 Sobolev spaces

Sobolev space:

$$\mathbb{H}^m(\Omega) \equiv \left\{ v \,:\, \Omega \to \mathbb{R} \,:\, \int_\Omega \left(v^2 + \left(v^{(1)}\right)^2 + \ldots + \left(v^{(m)}\right)^2 \right) < \infty \right\}$$

then

$$\mathbb{H}^0(\Omega) = L^2(\Omega)$$

and $\mathbb{H}^m(\Omega)$ is a Hilbert space:

$$(u,v)_m = \int_\Omega \left(uv + u^{(1)}v^{(1)} + \ldots + u^{(m)}v^{(m)} \right) dx$$

and

$$||u||_m = \sqrt{\int_\Omega \left(u^2 + \ldots + \left(u^{(m)}\right)^2 \right) dx}$$

Proposition 3.1.1 *1.* $\mathbb{H}^{m+1}(\Omega) \subset \mathbb{H}^m(\Omega)$

2. *If* $v \in \mathbb{H}^m(\Omega)$ *then:*

$$||v||_m^2 = ||v||_0^2 + \left|\left|v^{(1)}\right|\right|_0^2 + \ldots + \left|\left|v^{(m)}\right|\right|_0^2$$

3. $||v||_{m+1} \geq ||v||_m$ *and* $||v||_{m+1} \geq \left|\left|v^{(1)}\right|\right|_m$.

4. *If* $v,\ w \in \mathbb{H}^m(\Omega)$ *then* $|(v,w)_m| \leq ||v||_m ||v||_m$, *this is the Cauchy-Schwartz inequality.*

5. *If* $v,\ w \in \mathbb{H}^m(\Omega)$ *then*

$$||v+w||_m \leq ||v||_m + ||w||_m$$

Proof We will prove the Cauchy-Schwartz inequality

$$|(u,v)_m| \leq ||u||_m ||v||_m$$

for $u,\ v \in \mathbb{H}^m(\Omega)$. For any $s \in \mathbb{R}$ we have

$$0 \leq (u - sv, u - sv)_m = ||u||_m^2 + s^2 ||v||_m^2 - 2s(u,v)_m .$$

Choose

$$s = \frac{(u,v)_m}{||v||_m^2}$$

so that:

$$0 \leq (u - sv, u - sv)_m = ||u||_m^2 + \frac{(u,v)_m^2}{||v||_m^2} - 2\frac{(u,v)_m^2}{||v||_m^2} = ||u||_m^2 - \frac{(u,v)_m^2}{||v||_m^2}$$

and, upon rearranging, we have:

$$(u,v)_m \leq ||u||_m \, ||v||_m$$

3.1.2 Weighted Residual Formulations

Consider

$$Lu = f, \quad \text{in } \Omega$$

with

$$u = 0, \quad \text{on } \partial\Omega$$

where

$$L = -\frac{d^2}{dx^2}$$

so that:

$$-\frac{d^2u}{dx^2} = f \quad \text{in } \Omega = (0,1)$$

with

$$u = 0, \quad \text{on } x = 0, 1$$

Define the "trial function" space, for example it can be chosen to be

$$U = \left\{ u \, : \, u \in \mathbb{H}^2(\Omega), \; u = 0 \text{ on } \partial\Omega \right\}$$

and we introduce the "weight functions" (also called test) space, for this choice of U it can be chosen to be

$$W = \left\{ w \, : \, w \in L^2(\Omega), \; w = 0 \text{ on } \partial\Omega \right\}$$

Now, find $u \in U$ such that

$$\int_\Omega \left(\frac{d^2u}{dx^2} + f \right) w \, d\Omega = 0, \quad f \in L^2(\Omega)$$

for all $w \in W$. This is one "weighted residual formulation" of the original problem. This particular formulation is also usually called a strong formulation. It still defines a continuous problem and in order to discretize it we need to discretize the corresponding functional spaces i.e. to define appropriate approximations for each function in them. We select a subset U_h of U:

$$U_h \subset U, \qquad U_h = \text{span}\left\{ \phi_0, \ldots, \phi_{n-1} \right\}$$

and W_h, a subset of W:

$$W_h \subset W, \qquad W_h = \text{span}\left\{ \psi_0, \ldots, \psi_{n-1} \right\}$$

and then, search for the discrete approximation u_h in the form:

$$u_h = \sum_{i=0}^{n-1} c_i \phi_i$$

so that:

$$\int_\Omega \left(\frac{d^2}{dx^2} \sum_{i=0}^{n-1} c_i \phi_i - f \right) \psi_j \, d\Omega = 0 \quad \Rightarrow \quad \sum_{i=0}^{n-1} c_i \int_\Omega \frac{d^2\phi_i}{dx^2} \psi_j = \int_\Omega f \psi_j \, d\Omega, \quad j = 0, 1, \ldots, n-1$$

This is a linear algebraic system $Lc = F$ where:

$$L_{ij} = \int_\Omega \frac{d^2\phi_i}{dx^2} \psi_j \, d\Omega, \qquad F_i = \int_\Omega f \psi_i \, d\Omega$$

3.1.3 Collocation Methods

Let

$$\psi_j(x) = \delta(x - x_j)$$

so that: $L_{ij} = \frac{d^2\phi_i}{dx^2}(x_j)$ and $F_j = f(x_j)$. Note that ψ_j are not in L^2 however the formulation still makes sense if $\phi_i \in C^0$ because the integrals are well defined. Next, we may approximate $\frac{d^2}{dx^2} \sim \delta_x^2$, and this gives us a finite difference method or we may select U_h as the space spanned by certain sets of orthogonal polynomials (Legendre, Chebyshev etc.), and this gives us spectral collocation.

3.2 Weak Methods

By far, the most popular weighted residual methods are based on the so called weak formulation of the classical problem. To obtain it we start from

$$\int_\Omega \left(\frac{d^2u}{dx^2} + f\right) w \, d\Omega = 0 \quad \forall\, w \in W$$

and choose the discrete test functions space to be

$$W = \left\{w \,:\, w \in \mathbb{H}_0^1(\Omega)\right\}.$$

Here we define

$$\mathbb{H}_0^1(\Omega) = \left\{u \in \mathbb{H}^1(\Omega),\ u = 0 \text{ on } \partial\Omega\right\}.$$

Then,

$$-\int_\Omega \frac{du}{dx}\frac{dw}{dx} + \underbrace{\int_{\partial\Omega} \frac{du}{dx} w \, ds}_{=0} = -\int_\Omega f w \, d\Omega$$

Then the problem is reformulated as: Find $u \in U$ such that:

$$\int_\Omega \frac{du}{dx}\frac{dw}{dx} = \int_\Omega f w \, d\Omega$$

A natural choice for U which makes all integrals well defined and incorporates the Dirichlet boundary conditions in the solution is $U = W$. Such formulation is called a *Galerkin formulation.* The space U can be discretized by means of piecewise polynomial functions, orthogonal polynomials, trigonometric polynomials, spline functions etc. These choices yield various weak methods. In the remainder of the notes we will focus on the piecewise polynomial approximations which yield the so called finite element methods.

Before we proceed with the discretization we shall prove some important results for the continuous Galerkin formulation: Find $u \in U = \mathbb{H}_0^1(\Omega)$ such that

$$\int_\Omega u'v' \, d\Omega = \int_\Omega f v \, d\Omega, \quad \forall\, v \in U.$$

Theorem 3.2.1 *If u is a solution of the classical formulation*

$$\begin{aligned} -u'' &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

then u is a solution of the Galerkin formulation. If u is a solution to the Galerkin Formulation then it is a solution to the classical formulation if $u \in \mathcal{C}^2([0,1])$ and $f \in \mathcal{C}^0([0,1])$.

Proof Suppose that u solves the differential formulation:

$$-u'' = f$$

then take $v \in U$ so that:

$$\begin{aligned} -\int_\Omega u'' v \, d\Omega &= \int_\Omega f v \, d\Omega \\ \int_\Omega u' v' \, d\Omega &= \int_\Omega f v \, d\Omega \\ (u', v')_0 &= (f, v)_0 \end{aligned}$$

Suppose that u solves the Galerkin formulation so that:

$$(u', v') = (f, v)$$

then

$$\int_\Omega (u'v' - fv) \, d\Omega = 0 \quad \Rightarrow \quad \int_\Omega (u'' + f) \, v \, d\Omega = 0$$

Assume towards contradiction that $u'' + f \neq 0$ in some point $x \in \Omega$. However, $u'' \in \mathcal{C}^0(\Omega)$ and $f \in \mathcal{C}^0(\Omega)$. Therefore,

$$u'' + f \in \mathcal{C}^0(\Omega) \Longrightarrow u'' + f \neq 0$$

in an entire open interval contained in Ω. Therefore, there exists $(x_0, x_1) \subset \Omega$ such that $u'' + f > 0$ for all $x \in (x_0, x_1)$ or $u'' + f < 0$ for all $x \in (x_0, x_1)$. We consider the first possibility and the second one can be considered in exactly the same way. Let us choose

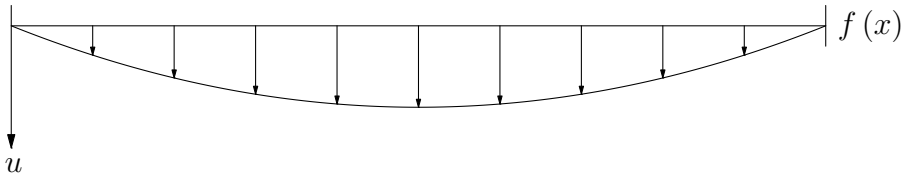
$$v = \begin{cases} -(x - x_0)(x - x_1) & \text{in } (x_0, x_1) \\ 0 & \text{otherwise.} \end{cases}$$

For this choice it is clear that $v \in \mathbb{H}_0^1(\Omega)$ and $v \geq 0$ in Ω. Now since $u'' + f > 0$ in (x_0, x_1) then

$$(u'' + f, v) > 0$$

but this contradicts the fact that u is a solution to $(u'' + f, v) = 0$ for all $v \in U$. Therefore, $u'' + f = 0$.

Example Consider the deformation of a rod from its equilibrium position under a given load f:



Then

$$E(u) = \frac{1}{2}(u', u')_0 - (f, u)_0$$

is the total energy of the system.

Therefore, E is called the energy functional of any system described by a second order elliptic differential equation.

Theorem 3.2.2 *$u \in \mathbb{H}_0^1(\Omega)$ is a solution to the Galerkin formulation if and only if u minimizes $E(v)$ over $\mathbb{H}_0^1(\Omega)$. That is, $E(u) \leq E(v)$ for all $v \in \mathbb{H}_0^1(\Omega)$.*

Proof 1. Suppose that u is a solution to the Galerkin formulation so that: $(u', v')_0 = (f, v)_0$ for all $v \in \mathbb{H}_0^1(\Omega)$ so that:

$$E(v) = E(u+w) = \frac{1}{2}\left((u+w)', (u+w)'\right)_0 - (f, u+w)_0 = \frac{1}{2}(u', u')_0 - (f, u)_0 + \frac{1}{2}(w', w')_0 - (f, w)_0 + (u', w')_0$$

Since

$$(u', w')_0 = (f, w)_0$$

we have

$$E(v) = E(u) + \frac{1}{2}||w'||_0^2 \geq E(u)$$

2. Suppose that $E(u) \leq E(v)$ for all $v \in \mathbb{H}_0^1(\Omega)$. Then if we take $s \in \mathbb{R}$, $E(u+sw)$ has a min at $s = 0$, and therefore:

$$\frac{d}{ds} E(u+sw)\Big|_{s=0} = 0$$

which we expand as:

$$\frac{d}{ds}\left(\frac{1}{2}\left((u+sw)', (u+sw)'\right)_0 - (f, u+sw)_0\right)\Big|_{s=0} = 0$$

$$\frac{d}{ds}\left(E(u) + s(u', w')_0 - s(f, w)_0 + \frac{1}{2}s^2(w', w')\right)\Big|_{s=0} = (u', w')_0 - (f, w)_0 + \underbrace{s(w', w')}_{=0}\Big|_{s=0} = 0$$

So that:

$$(u', w')_0 = (f, w)_0$$

for all $w \in \mathbb{H}_0^1(\Omega)$. Therefore, u solves the Galerkin Formulation.

Consider a Hilbert space V equipped with the product $(\cdot, \cdot)_V$ i..e. V is complete with respect to the norm $||\cdot||_V$ induced by this product. Now, consider the mapping $a: V \times V \to \mathbb{R}$ (further called a bilinear form) such that:

1. $a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$,
2. $a(w, \alpha u + \beta v) = \alpha a(w, u) + \beta a(w, v)$,
3. there exists $\beta > 0$ such that $|a(u, v)| \leq \beta ||u||_V ||v||_V$ (such bilinear form is called bounded w.r.t. $||.||_V$), and
4. there exists $\rho > 0$ such that
$$a(u, u) \geq \rho ||u||_V^2$$
(such bilinear form is called coercive w.r.t. $||.||_V$)

Now, suppose that $G(v)$ is a functional such that:

1. $G(\alpha u + \beta v) = \alpha G(u) + \beta G(v)$ (such a functional is called linear) and
2. there exists $\delta > 0$ such that $|G(u)| \leq \delta ||u||_V$ (such functional is called bounded).

The following theorem is one version of a very well known result from functional analysis.

Theorem 3.2.3 (Riesz Theorem) *If V is a Hilbert space with a given inner product $(.,.)_*$ and if $G(v)$ is a bounded linear functional on V (w.r.t. the norm induced by its inner product), then there is a unique element $\hat{u} \in V$ such that $(\hat{u}, v)_* = G(v)$ for all $v \in V$.*

This theorem almost directly yields the proof of a somewhat restricted version of the following fundamental result in PDE theory and their numerical analysis.

Lemma 3.2.4 (Lax-Milgram) *If $a(u,v)$ and $G(v)$ satisfy the six conditions stated above then there exists a unique solution $\hat{u} \in V$ for the following problem:* Find $u \in V$ s.t.

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

If $a(u,v) = a(v,u)$ (symmetric form) then $\hat{u}$ is the minimizer of

$$E(v) = \frac{1}{2}a(v,v) - G(v)$$

Proof We shall consider only the case where $a(u,v) = a(v,u)$.

- The proof of the first claim is a direct consequence of the Riesz theorem if we prove that $a(u,v)$ defines an inner product on V. We know that:
$$a(u,u) \geq \rho ||u||_V^2$$
which guarantees that
$$\sqrt{a(u,u)} = ||u||_a$$
is a norm on V. It is not difficult then to verify that $a(u,v)$ defines an inner product on V. Also, it is straightforward to show that the norm $||.||_a$ is equivalent to $||.||_V$ i.e.:
$$\sqrt{\rho}\,||u||_V \leq ||u||_a \leq \sqrt{\beta}\,||u||_V .$$
Therefore, since V is complete with respect to $||\cdot||_V$ then it is complete with respect to the norm $||.||_a$ i.e. V is a Hilbert space with respect to the product $a(.,.)$. As $G(v)$ is bounded with respect to $||\cdot||_V$, $G(v)$ is bounded with respect to $||\cdot||_a$. Since $a(.,.)$ is an inner product on V and $G(v)$ is bounded w.r.t. the norm induced by this inner product, the Riesz theorem implies the first claim of the lemma.
- Next we show that $\hat{u}$ minimizes $E(v) = \frac{1}{2}a(v,v) - G(v)$.
$$E(v) = E(\hat{u}+w) = E(\hat{u}) + 1/2\underbrace{a(w,w)}_{\geq 0} \geq E(\hat{u})$$

Example

$$\begin{cases} -\nabla^2 u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}.$$

Its Galerkin formulation reads: *Find $u \in \mathbb{H}_0^1(\Omega)$ s.t.:*

$$(\nabla u, \nabla v)_0 = (f,v)_0 \quad \forall v \in \mathbb{H}_0^1(\Omega). \tag{3.1}$$

Below we prove the so-called Poincaré inequality which will guarantee that $a(u,v) = (\nabla u, \nabla v)_0$ defines a bilinear form satisfying the conditions of the Lax-Milgram lemma.

Lemma 3.2.5 (Poincaré Inequality) *If Ω is a bounded domain and $v \in \mathbb{H}_0^1(\Omega)$ then $\exists C > 0$, depending only on the domain Ω, such that*

$$||v||_0^2 \leq C\,||\nabla v||_0^2 .$$

Proof Consider first $v \in \mathbb{C}_0^1(\Omega)$ i.e. v is a continuously differentiable function defined on Ω that vanishes on its boundary. We can always find $a \in \mathbb{R}$ large enough so that the cube $Q = \{x \in \mathbb{R}^n : |x_j| < a, 1 \leq j \leq n\}$ contains Ω. Integrating by parts in the x_1 direction and taking into account that the surface integral vanishes since $v = 0$ on $\partial\Omega$ we obtain:

$$\begin{aligned} ||v||_0^2 &= \int_\Omega v^2 dx = \int_\Omega 1 \cdot v^2 dx = -\int_\Omega x_1 \tfrac{\partial v^2}{\partial x_1} dx \\ &= -2\int_\Omega x_1 v \tfrac{\partial v}{\partial x_1} dx \leq 2a \int_\Omega |v| \left|\tfrac{\partial v}{\partial x_1}\right| dx. \end{aligned}$$

Using the Cauchy-Schwarz inequality for the L^2 inner product on Ω, we obtain:

$$||v||_0^2 \leq 2a \int\limits_{\Omega} |v| \left| \frac{\partial v}{\partial x_1} \right| dx \leq 2a \|v\|_0 \| \frac{\partial v}{\partial x_1} \|_0 \leq 2a \|v\|_0 \|\nabla v\|_0.$$

Dividing by $\|v\|_0$ gives the desired result with $C = (2a)^2$ and $v \in \mathbb{C}_0^1(\Omega)$. Due to some classical results in functional analysis (see for example Ern, A. and Guermond, J.-L., Theory and Practice of Finite Elements, Applied Mathematical Sciences, v. 159, Springer, 2004, p. 485), we can claim that for each $v \in \mathbb{H}_0^1(\Omega)$ we can choose a sequence of functions in $\{v_k\}_{k=1}^{\infty} \subset \mathbb{C}_0^1(\Omega)$ such that the sequence converges to v in the $\mathbb{H}^1\Omega)$ norm i.e. $\|v - v_k\|_0 \leq \|v - v_k\|_1 \to 0$ as $k \to \infty$ and similarly $\|\nabla v - \nabla v_k\|_0 \leq \|v - v_k\|_1 \to 0$. Using the triangular inequality in the form $\|u - w\|_0 \geq \|u\|_0 - \|w\|_0$ yields that $\|v_k\|_0 \to \|v\|_0$ and $\|\nabla v_k\|_0 \to \|\nabla v\|_0$. This allows us to take the limit in the Poincaré inequality for $v_k \in \mathbb{C}_0^1(\Omega)$ to derive the inequality for any function in $\mathbb{H}_0^1(\Omega)$.

Using this inequality we easily prove the following proposition.

Proposition 3.2.6 *$a(u,v) = (\nabla u, \nabla v)_0$ is a bilinear, bounded, and coercive form in $H_0^1(\Omega)$; $G(v) = (f,v)_0$ is bounded in $H_0^1(\Omega)$ if $||f||_0 < \infty$.*

Proof

$$|a(u,v)| = \left| \int\limits_{\Omega} \nabla u \nabla v dx \right| \leq ||\nabla u||_0 \, ||\nabla v||_0 \leq ||u||_1 \, ||v||_1$$

as $||u||_1 = ||\nabla u||_0 + ||u||_0 \geq ||\nabla u||_0$. Then,

$$a(u,u) = (\nabla u, \nabla u) = ||\nabla u||_0^2 = \frac{1}{2}\left(||\nabla u||_0^2 + ||\nabla u||_0^2 \right) \geq \frac{1}{2}\left(||u||_0^2 + ||\nabla u||_0^2 \right) = \frac{1}{2} ||u||_1^2$$

The last proposition guarantees that the bilinear form $a(u,v) = \int\limits_{\Omega} \nabla u \nabla v dx$ and the functional $G(v) = (f,v)_0$ satisfy the conditions of the Lax-Milgram lemma if $||f||_0 < \infty$. This automatically guarantees that the corresponding weak formulation (3.1) has a unique solution in $H_0^1(\Omega)$.

3.3 Finite Element Method (FEM)

Let us define a grid:

$$\Delta = \{x_0, \ldots, x_M\}$$

and the following linear space of continuous piecewise linear functions on Δ:

$$M_0^1(\Delta) = \text{span}\{l_0, l_1, \ldots, l_M\} \in \mathbb{H}^1(\Omega)$$

where

$$l_i = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}}, & x_{i-1} \leq x \leq x_i \\ \frac{x - x_{i+1}}{x_i - x_{i+1}}, & x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise.} \end{cases}.$$

Define the discrete space:

$$V_h = \left\{ v \in M_0^1(\Delta), \; v(0) = v(1) = 0 \right\} \subset \mathbb{H}_0^1(\Omega)$$

which is equivalent to

$$V_h = \text{span}\{l_1, \ldots, l_{M-1}\}$$

Let us now consider the following discrete problem: *Find $u_h \in V_h$ such that*:

$$(u_h', v_h')_0 = (f, v_h)_0$$

for all $v_h \in V_h$. It is clear that:

$$u_h = \sum_{j=1}^{M-1} u_j l_j$$

and

$$(u_h', l_i')_0 = (f, l_i)_0, \quad 1 \le i \le M-1$$

so that:

$$\left(\sum_{j=1}^{M-1} u_j l_j', l_i' \right)_0 = (f, l_i)_0 \quad \Rightarrow \quad \sum_{j=1}^{M-1} u_j \left(l_j', l_i' \right)_0 = (f, l_i)_0, \quad 1 \le i \le M-1.$$

The last set of equations clearly constitute a linear algebraic system:

$$Au_h = F$$

where $A_{ij} = \left(l_i', l_j' \right)_0, F_i = (f, l_i)_0$, and $u_{h,i} = u_i$.

Now, as

$$0 \le (v_h', v_h')_0 = \left(\sum_{i=1}^{M-1} v_i l_i', \sum_{j=1}^{M-1} v_j l_j' \right)_0 = \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} v_i \left(l_i', l_j' \right)_0 v_j = v^T A v$$

the Lax-Milgram lemma guarantees that this system has a unique solution if $f(x) \in L^2(\Omega)$.

3.4 Gaussian Quadrature

We may approximate the integral of function $\phi(x)$ over the range $x \in (-1, 1)$ using Gaussian integration:

$$\int_{-1}^{1} \phi(x)\, dx \approx \sum_{i=1}^{n} A_i^{(n)} \phi\left(x_i^{(n)} \right)$$

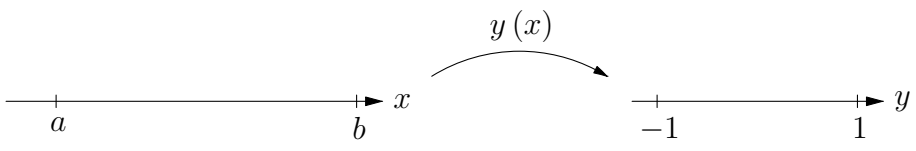
where $A_i^{(n)}$ are the weights and $x_i^{(n)}$ are the nodes of the quadrature. n is a positive integer that controls the accuracy. The weights and nodes are determined from the condition that the quadrature is exact for all polynomials of the highest possible degree. We have $2n$ undetermined coefficients (n weights and n nodes) and so we can make the quadrature exact for polynomials of $2n-1$ degree (that have $2n$ coefficients).

The following table contains the weights and Gauss points for the first fourGauss quadratures:

n	$A_i^{(n)}$	$x_i^{(n)}$	$2n-1$
1	2	0	1
2	1, 1	$-\frac{1}{\sqrt{3}}$, $\frac{1}{\sqrt{3}}$	3
3	$\frac{5}{9}$, $\frac{8}{9}$, $\frac{5}{9}$	$-\sqrt{\frac{3}{5}}$, 0, $\sqrt{\frac{3}{5}}$	5
4	0.347854, 0.652145, 0.347854, 0.652145	−0.861136, −0.339981, 0.339981, 0.861136	7

.

Remark We may approximate the integral over arbitrary bounds a and b using Gaussian Quadrature by applying the substitution:



$$y(x) = \frac{2}{b-a}x + \frac{a+b}{a-b}$$

So that $y(a) = -1$, $y(b) = 1$, and $dx = \frac{b-a}{2}\,dy$. Therefore:

$$\int_a^b \phi(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} \phi\left(\frac{(b-a)\,y + a + b}{2}\right)\,dy$$

3.5 Error Estimates

Theorem 3.5.1 *If u_h is the finite element approximation to the exact solution u of the boundary value problem*

$$-u'' = f, \qquad u(0) = u(1) = 0$$

then there exists $k > 0$ such that:

$$||u - u_h||_1 \leq k\,||u - v_h||_1$$

for all $v_h \in V_h$.

Proof u is the exact solution, so that

$$a(u, w_h) = (f, w_h)$$

for all $w_h \in V_h$, and

$$a(u_h, w_h) = (f, w_h)$$

So that

$$a\left(\underbrace{u - u_h}_{\epsilon_h}, w_h\right) = 0 \quad \Rightarrow \quad a(\epsilon_h, w_h) = 0$$

Then,

$$\frac{1}{2}\,||\epsilon_h||_1^2 \leq a(\epsilon_h, \epsilon_h) + a(\epsilon_h, w_h) \leq a(\epsilon_h, \epsilon_h + w_h) = a(\epsilon_h, u - u_h + w_h) = a(\epsilon_h, u - v_h) \leq ||\epsilon_h||_1\,||u - v_h||_1$$

So that

$$||\epsilon_h||_1 \leq 2\,||u - v_h||_1$$

Corollary 3.5.2 *If u and u_h are as in the theorem, then we have*

$$||u - u_h||_1 \leq ch$$

where $c > 0$ is independent of h.

Proof For the time being we will assume that $u \in \mathcal{C}^2(\Omega)$ i.e. u is a solution to the classical problem for $f \in C^0(\Omega)$. Let $\hat{u}$ be the piecewise linear interpolant of u in V_h i.e. $\hat{u} = \sum_{i=1}^{M-1} u_i l_i$, where u_i are the values of u in the nodes of the grid x_i. Using Taylor expansion it is possible to prove the following estimate for the interpolation error:

$$||u' - \hat{u}'||_{L^\infty} \leq ||u''||_{L^\infty}\,h$$

Recall that:

$$||u - u_h||_1^2 \leq 4\,||u - v||_1^2$$

for all $v \in V_h$ and as $\hat{u} \in V_h$ we have:

$$||u - u_h||_1^2 \leq 4\,||u - \hat{u}||_1^2 = 4\left(||u - \hat{u}||_0^2 + ||u' - \hat{u}'||_0^2\right) \leq 8\,||u' - \hat{u}'||_0^2$$

where, in the last step, we applied the Poincaré inequality. Continuing, we have:

$$||u-u_h||_1^2 \leq 8\int_0^1 (u'-\hat{u}')^2\,dx \leq 8\,||u'-\hat{u}'||_{L^\infty}^2 \leq 8\,||u''||_{L^\infty}^2\,h^2$$

where we have used the interpolation estimate $||u'-\hat{u}'||_{L^\infty}^2 \leq ||u''||_{L^\infty}^2\,h^2$. This yields

$$||u-u_h||_1 \leq \sqrt{8}\,||u''||_{L^\infty}\,h = \mathcal{O}(h).$$

Applying Poincaré's inequality, we also have that:

$$||u-u_h||_1^2 = ||u-u_h||_0^2 + ||u'-u_h'||_0^2 \geq 2\,||u-u_h||_0^2$$

so that we obtain the estimate in the L^2 norm:

$$||u-u_h||_0 \leq ch\,||u''||_{L^\infty}.$$

This is a suboptimal estimate because from interpolation theory we know that u can be approximated with a piecewise linear function with a second order accuracy in the L^2 norm.

3.6 Optimal Error Estimates

Lemma 3.6.1 *If $\hat{u}$ is the finite element interpolant of $u \in \mathbb{H}^2(\Omega)$, i.e. $\hat{u} = \sum_{j=1}^{M-1} u(x_j) l_j$, then:*

1. $||u-\hat{u}||_0 \leq \left(\frac{h}{\pi}\right)^2 ||u''||_0$
2. $||u'-\hat{u}'||_0 \leq \frac{h}{\pi}\,||u''||_0$

Proof The domain is subdivided into elements at points:

$$x_0 = 0,\ x_1,\ x_2,\ \ldots x_{i-1},\ x_i,\ \ldots x_{M-1},\ x_M = 1$$

Consider the difference

$$u-\hat{u} \quad \text{in } [0,h] = [x_0, x_1]$$

and define $\eta(x) = u-\hat{u} \in \mathbb{H}^1([0,h])$. Clearly,

$$\eta(0) = \eta(h) = 0.$$

Since $\eta \in C^0[0,h]$ it can be expanded in a uniformly convergent Fourier sine series:

$$\eta(x) = \sum_{n=1}^{\infty} \eta_n \sin\left(\frac{n\pi x}{h}\right)$$

Then, using the Parseval identity,

$$\int_0^h \eta^2\,dx = ||\eta||_{L^2[0,h]}^2 = \frac{h}{2}\sum_{n=1}^{\infty}\eta_n^2.$$

Differentiating term wise we get:

$$\eta'(x) = \sum_{n=1}^{\infty} \eta_n \frac{n\pi}{h}\cos\left(\frac{n\pi x}{h}\right) \quad \Rightarrow \quad ||\eta'||_{L^2[0,h]}^2 = \frac{h}{2}\sum_{n=1}^{\infty}\left(\eta_n\frac{n\pi}{h}\right)^2$$

and differentiating again we have:

$$\eta''(x) = -\sum_{n=1}^{\infty} \eta_n \frac{n^2\pi^2}{h^2}\sin\left(\frac{n\pi x}{h}\right) \quad \Rightarrow \quad ||\eta''||_{L^2[0,h]}^2 = \frac{h}{2}\sum_{n=1}^{\infty}\eta_n^2\left(\frac{n\pi}{h}\right)^4$$

So that

$$||\eta'||^2_{L^2[0,h]} = \frac{h}{2}\sum_{n=1}^{\infty}\eta_n^2\left(\frac{n\pi}{h}\right)^2\left(\frac{n\pi}{h}\right)^2\frac{h^2}{n^2\pi^2} \leq \frac{h^2}{\pi^2}\frac{h}{2}\sum_{n=1}^{\infty}\eta_n^2\left(\frac{n\pi}{h}\right)^4 = \frac{h^2}{\pi^2}||\eta''||^2_{L^2[0,h]} = \frac{h^2}{\pi^2}||u''||^2_{L^2[0,h]}$$

and

$$||\eta||^2_{L^2[0,h]} = \frac{h}{2}\sum_{n=1}^{\infty}\eta_n^2 \leq \frac{h^2}{\pi^2}\frac{h}{2}\sum_{n=1}^{\infty}\eta_n^2\left(\frac{n\pi}{h}\right)^2 = \frac{h^2}{\pi^2}||\eta'||^2_{L^2[0,h]} \leq \left(\frac{h}{\pi}\right)^4||u''||^2_{L^2[0,h]}$$

Applying the same for any subsequent subinterval and summing the resulting inequalities we complete the proof.

Corollary 3.6.2 *1.* $||u-\hat{u}||_1 \leq \frac{h\sqrt{2}}{\pi}||u''||_0$

2. $||u-u_h||_1 \leq \frac{\sqrt{8}h}{\pi}||u''||_0$

Proof 1.

$$||u-\hat{u}||_1^2 = ||u-\hat{u}||_0^2 + ||u'-\hat{u}'||_0^2 \leq 2||u'-\hat{u}'||_0^2 \leq 2\left(\frac{h}{\pi}\right)^2||u''||_0^2$$

2. for all $v_h \in V_h$ in the finite element space we have:

$$||u-u_h||_1 \leq 2||u-v_h||_1$$

Take $v_h = \hat{u}$, then:

$$||u-u_h||_1 \leq 2||u-\hat{u}||_1 \leq \frac{h\sqrt{8}}{\pi}||u''||_0$$

So that

$$||u-u_h||_1 \leq Ch||u''||_0$$

where $C > 0$ is independent of h.

Theorem 3.6.3 (L^2 lifting theorem) *If $u_h \in V_h$ is the Galerkin finite element approximation of:*

$$u \; : \; \begin{cases} -u'' = f(x), & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

then there exists $\Gamma > 0$ such that

$$||u-u_h||_0 \leq \Gamma h^2||u''||_0$$

Proof Define:

$$\epsilon_h = u - u_h$$

then,

$$(u', v_h')_0 = (f, v_h)_0$$

for all $v_h \in V_h$ or,

$$a(u, v_h) = (u', v_h')_0$$

so that

$$\left.\begin{array}{l} a(u,v_h) = (f,v_h)_0 \\ a(u_h,v_h) = (f,v_h)_0 \end{array}\right\} \quad a(\epsilon_h, v_h) = 0 \tag{3.2}$$

Now, consider the auxiliary problem

$$\begin{cases} -\phi'' = \epsilon_h, & 0 < x < 1 \\ \phi(0) = \phi(1) = 0 \end{cases}$$

Then,

$$||\epsilon_h||_0^2 = (\epsilon_h, \epsilon_h)_0 = -(\epsilon_h, \phi'')_0 = (\epsilon_h', \phi')_0 = a(\epsilon_h, \phi)$$

Now, taking $v_h = \hat{\phi}$ in (3.2), we get:

$$a\left(\epsilon_h, \hat{\phi}\right) = 0,$$

and subtracting it from the above equation we obtain:

$$\begin{aligned} ||\epsilon_h||_0^2 = a(\epsilon_h, \phi) = a(\epsilon_h, \phi) - a\left(\epsilon_h, \hat{\phi}\right) = a\left(\epsilon_h, \phi - \hat{\phi}\right) = \left(\epsilon_h', \phi' - \hat{\phi}'\right)_0 = \int_0^1 \epsilon'\left(\phi' - \hat{\phi}'\right) dx \\ \leq ||\epsilon_h'||_0 \left|\left|\phi' - \hat{\phi}'\right|\right|_0 \leq ||\epsilon_h'||_0 \left|\left|\phi - \hat{\phi}\right|\right|_1 \leq ||\epsilon_h'||_0 \frac{\sqrt{2}h}{\pi} ||\phi''||_0 = ||\epsilon_h'||_0 \frac{h\sqrt{2}}{\pi} ||\epsilon_h||_0 \\ \leq ||\epsilon_h||_1 \frac{h\sqrt{2}}{\pi} ||\epsilon_h||_0 \leq \Gamma h^2 ||u''||_0 ||\epsilon_h||_0 \end{aligned}$$

So that

$$||\epsilon_h||_0 \leq \Gamma h^2 ||u''||_0$$

for some $\Gamma > 0$, independent of h.

3.6.1 Other Boundary Conditions

Let us consider an elliptic problem with more general boundary conditions:

$$\begin{cases} -u''(x) = f, & 0 < x < 1 \\ u(0) = \beta_1, & u'(1) = \beta_2. \end{cases}$$

Since we need to satisfy non-homogeneous boundary conditions of Dirichlet and Neumann type, this time we discretize the solution with the expansion:

$$u_h = \beta_1 l_0(x) + \sum_{j=1}^{M} u_j l_j(x) \tag{3.3}$$

so that

$$u_h(0) = \beta_1 l_0(0) = \beta_1.$$

Now, consider the original equation:

$$-u'' = f$$

multiply it by $v \in V = \mathbb{H}^1(\Omega)$, and integrate over Ω to obtain:

$$(-u'', v)_0 = (f, v)_0 \quad \Rightarrow \quad (u', v')_0 + u'(0) v(0) - u'(1) v(1) = (f, v)_0.$$

If we approximate u with (3.3) and take the test functions to be $v_h \in V_h = \text{span}\{l_1, \ldots, l_M\}$, we have:

$$\left(u_h', l_j'\right)_0 - u_h'(1) l_j(1) = (f, l_j)_0, \quad j = 1, \ldots, M,$$

or

$$\left(\beta_1 l_0' + \sum_{n=1}^{M} u_n l_n', l_j'\right)_0 - \beta_2 l_j(1) = (f, l_j)_0.$$

This yields:

$$\left(\beta_1 l_0' + \sum_{n=1}^{M} u_n l_n', l_1'\right)_0 = (f, l_1)_0,$$

$$\left(\sum_{n=1}^{M} u_n l_n', l_j'\right)_0 = (f, l_j)$$

for $j = 2, \ldots, M-1$ and finally,

$$\left(\sum_{n=1}^{M} u_n l_n', l_M'\right)_0 - \beta_2 = (f, l_M)_0 .$$

3.7 Transient Problems

Consider the Initial Boundary Value Problem given by:

$$\begin{cases} \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left[a(x)\frac{\partial u}{\partial x}\right] = f(x) & \text{for} \quad (t,x) \in (0,T] \times (0,1) \\ u(t,0) = u(t,1) = 0 & \text{for} \quad t > 0 \\ u(0,x) = g(x) & \text{for} \quad x \in (0,1) \end{cases}$$

For $a(x) \in \mathcal{C}^1(\bar{\Omega})$ such that $0 < \alpha \leq a(x) \leq A < \infty$ and $||f(x)||_0 \leq L < \infty$.

The Galerkin formulation of the problem is given by: Find $u \in \mathbb{H}_0^1(\Omega)$ such that

$$\left(\frac{\partial u}{\partial t}, v\right)_0 + b(u,v) = (f(x), v)_0$$

for all $v \in \mathbb{H}_0^1(\Omega)$, where $b(u,v) = \left(a(x)\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\right)_0$. Note that $b(u,v)$ is a coercive and bounded bilinear form on $H_0^1(\Omega)$ (why?). In order to discretize it we define

$$V_h = \text{span}\{l_1, l_2, \ldots l_{M-1}\}$$

and search for $u_h(t,x) = \sum_{j=1}^{M-1} u_j(t) l_j(x) \in V_h$ such that:

$$\left(\frac{\partial u_h}{\partial t}, v_h\right)_0 + b(u_h, v_h) = (f, v_h)_0$$

for all $v_h \in V_h$. Then, taking into account that l_j form a basis of V_h, it is sufficient to take $v_h = l_j, j = 1, \ldots, M-1$ and we obtain the linear system

$$\frac{\partial}{\partial t}(\mathbf{M}\mathbf{u}_h) + \mathbf{S}\mathbf{u}_h = \mathbf{F}$$

where

$$A_{ij} = \underbrace{\int_\Omega (l_i, l_j)\, dx}_{\text{mass matrix } \mathbf{M}} + \underbrace{\int_\Omega a(x) \frac{\partial l_i}{\partial x}\frac{\partial l_j}{\partial x}\, dx}_{\text{stiffness } \mathbf{S}}$$

for each $i,\ j = 1, \ldots, M-1$, and

$$F_j = \int_\Omega f l_j\, dx, j = 1, \ldots, M-1; \quad \mathbf{u}_h = (u_1(t), \ldots, u_{M-1}(t))^T.$$

Then, as $\mathbf{M}$ is time-independent,

$$\mathbf{M}\frac{\partial}{\partial t}(\mathbf{u}_h) + \mathbf{S}\mathbf{u}_h = \mathbf{F}$$

We may discretize this ODE system using for example a backward difference scheme (a good choice for such problems, as we know from the previous section) to obtain the final linear system that yields the solution at time level n:

$$\mathbf{M}\frac{\mathbf{u}_h^n - \mathbf{u}_h^{n-1}}{k} + \mathbf{S}\mathbf{u}_h^n = \mathbf{F}^n.$$

Define

$$\mathbf{A} = \left(\frac{\mathbf{M}}{k} + \mathbf{S} \right),$$

where :

$$M_{ij} = \int_{\Omega} l_i l_j$$

and

$$S_{ij} = \int_{\Omega} a\left(x\right) \frac{\partial l_i}{\partial x} \frac{\partial l_j}{\partial x}\, dx \geq \alpha \int_{\Omega} \frac{\partial l_i}{\partial x} \frac{\partial l_j}{\partial x}\, dx.$$

M, A are spd (why?). If u is the exact solution, then

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(a\left(x\right) \frac{\partial u}{\partial x} \right) = f$$

Now we multiply by $v_h \in V_h$ and integrate:

$$\left(\frac{\partial u}{\partial t}, v_h \right)_0 + b\left(u, v_h\right) = \left(f, v_h\right)_0$$

Adding and subtracting $\left(\frac{u^n - u^{n-1}}{k}, v_h \right)_0$, we get:

$$\left(\frac{u^n - u^{n-1}}{k}, v_h \right)_0 + b\left(u^n, v_h\right)_0 = \left(f^n, v_h\right)_0 - \left(\frac{\partial u}{\partial t}^n, v_h \right)_0 + \left(\frac{u^n - u^{n-1}}{k}, v_h \right)_0$$

which we may rewrite as:

$$\left(\frac{u^n - u^{n-1}}{k}, v_h \right)_0 + b\left(u^n, v_h\right) = \left(f^n, v_h\right)_0 + \left(\tau_k, v_h\right)_0,$$

where $\tau_k = -\frac{\partial \mathbf{u}}{\partial t}^n + \frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{k}$ is the truncation error of the backward difference approximation of the first derivative in time.

On the other hand the finite element approximation to the solution satisfies:

$$\left(\frac{u_h^n - u_h^{n-1}}{k}, v_h \right)_0 + b\left(u_h^n, v_h\right) = \left(f^n, v_h\right)_0.$$

Letting

$$u^n - u_h^n = \epsilon_h^n$$

and subtracting the two preceding equations equations, we get:

$$\left(\frac{\epsilon_h^n - \epsilon_h^{n-1}}{k}, v_h \right) + b\left(\epsilon_h^n, v_h\right) = \left(\tau_k, v_h\right)_0$$

Now, define w_h such that

$$b\left(w_h, v_h\right) = b\left(u, v_h\right) \tag{3.4}$$

for all $v_h \in V_h$. w_h is called the *elliptic projection of u onto V_h*. Then we split the error ϵ_h^n into $\epsilon_h^n = \eta^n + \xi^n$ where $\eta^n = u^n - w_h^n$ is the error of the elliptic projection of u. Therefore, as we have already established for the solution of (3.4)

$$||u^n - w_h^n||_0 \leq \Gamma h^2 \left|\left| \frac{\partial^2 u}{\partial x^2} \right|\right|_0$$

But then, as $\frac{\partial w_h}{\partial t}$ is the elliptic projection of $\frac{\partial u}{\partial t}$ onto V_h,

$$\left|\left|\left(\frac{\partial u}{\partial t}\right)^n - \left(\frac{\partial w_h}{\partial t}\right)^n\right|\right|_0 \leq \Gamma h^2 \left|\left|\frac{\partial^3 u}{\partial t \partial^2 x}\right|\right|_0$$

Now,

$$\left(\xi^n - \xi^{n-1}, v_h\right)_0 + kb\left(\xi^n, v_h\right) = k\left(\tau_k^n, v_h\right)_0 - \left(\eta^n - \eta^{n-1}, v_h\right)_0 - kb\left(\eta^n, v_h\right)$$

But, as $\xi^n \in V_h$,

$$\left(\xi^n - \xi^{n-1}, \xi^n\right)_0 + kb\left(\xi^n, \xi^n\right) = k\left(\tau_k^n, \xi^n\right)_0 - \left(\eta^n - \eta^{n-1}, \xi^n\right)_0 - kb\left(\eta^n, \xi^n\right)$$

and, since since $b\left(\xi^n, \xi^n\right) \geq 0$ and $b\left(\eta^n, \xi^n\right) = 0$, this leads to the inequality:

$$\left(\xi^n - \xi^{n-1}, \xi^n\right)_0 \leq k\left(\tau_k^n, \xi^n\right)_0 - \left(\eta^n - \eta^{n-1}, \xi^n\right)_0$$

which we may rewrite as:

$$||\xi^n||_0^2 \leq \left(\xi^{n-1}, \xi^n\right)_0 + k\left(\tau_k^n, \xi^n\right)_0 - \left(\eta^n - \eta^{n-1}, \xi^n\right)_0. \tag{3.5}$$

Lemma 3.7.1

$$\left(\xi^{n-1}, \xi^n\right)_0 \leq \frac{1}{2}\left|\left|\xi^{n-1}\right|\right|_0^2 + \frac{1}{2}||\xi^n||_0^2$$

Proof

$$0 \leq \left|\left|\xi^n - \xi^{n-1}\right|\right|_0^2 = ||\xi^n||_0^2 - 2\left(\xi^n, \xi^{n-1}\right)_0 + \left|\left|\xi^{n-1}\right|\right|_0^2$$

Upon rearranging, this yields the desired result.

Theorem 3.7.2 (Young Inequality)

$$(u, v)_0 \leq \alpha^2 ||u||_0^2 + \frac{1}{4\alpha^2}||v||_0^2$$

Lemma 3.7.3

$$k\left(\tau_k^n, \xi^n\right) \leq \frac{k^2}{2}\int_{(n-1)k}^{nk}\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_0^2 dt + \frac{k}{2}||\xi^n||_0^2$$

Proof

$$k\left(\tau_k^n, \xi^n\right) \leq \frac{k}{2}||\tau_k^n||_0^2 + \frac{k}{2}||\xi^n||_0^2.$$

Now using the integral form of the truncation error: $\tau_k^n = \frac{\partial u^n}{\partial t} - \frac{u^n - u^{n-1}}{k} = \frac{1}{k}\int_{(n-1)k}^{nk}\left[t - (n-1)k\right]\frac{\partial^2 u}{\partial t^2}\,dt$, and the Cauchy-Schwartz inequality for the integral in time, we obtain:

$$\begin{aligned}
||\tau_k^n||_0^2 &= \left|\left|\frac{\partial u^n}{\partial t} - \frac{u^n - u^{n-1}}{k}\right|\right|_0^2 = \left|\left|\frac{1}{k}\int_{(n-1)k}^{nk}\left[t - (n-1)k\right]\frac{\partial^2 u}{\partial t^2}\,dt\right|\right|_0^2 \\
&\leq \left|\left|\frac{1}{k}\sqrt{\int_{(n-1)k}^{nk}\left[t - (n-1)k\right]^2 dt}\sqrt{\int_{(n-1)k}^{nk}\left(\frac{\partial^2 u}{\partial t^2}\right)^2 dt}\right|\right|_0^2 \\
&\leq \left|\left|\frac{\sqrt{k^3}}{k}\sqrt{\int_{(n-1)k}^{nk}\left(\frac{\partial^2 u}{\partial t^2}\right)^2 dt}\right|\right|_0^2 = k\int_\Omega\int_{(n-1)k}^{nk}\left(\frac{\partial^2 u}{\partial t^2}\right)^2 dt\,dx \\
&= k\int_{(n-1)k}^{nk}\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_0^2 dt
\end{aligned}$$

So then,

$$k\left(\tau_k^k,\xi^n\right)_0 \le \frac{k}{2}\left|\left|\tau_k^n\right|\right|_0^2 + \frac{k}{2}\left|\left|\xi^n\right|\right|_0^2 \le \frac{k^2}{2}\int_{(n-1)k}^{nk}\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|^2 dt + \frac{k}{2}\left|\left|\xi^n\right|\right|_0^2$$

Lemma 3.7.4

$$\left|\left(\eta^n-\eta^{n-1},\xi^n\right)_0\right| \le \frac{\Gamma^2 h^4}{2}\int_{(n-1)k}^{nk}\left|\left|\frac{\partial^3 u}{\partial t\partial x^2}\right|\right|_0^2 dt + \frac{k}{2}\left|\left|\xi^n\right|\right|_0^2$$

Proof Using the Cauchy-Schwartz inequality for the integral in time and subsequently, the Young's inequality we obtain

$$\begin{aligned}\left|\left(\eta^n-\eta^{n-1},\xi^n\right)_0\right| &= \left|\left(\int_{(n-1)k}^{nk}\frac{\partial\eta}{\partial t}\,dt,\xi^n\right)_0\right| \le \left|\left(\sqrt{\int_{(n-1)k}^{nk}\left(\frac{\partial\eta}{\partial t}\right)^2 dt},\sqrt{k}\xi^n\right)_0\right| \\ &\le \frac{1}{2}\int_\Omega\int_{(n-1)k}^{nk}\left(\frac{\partial\eta}{\partial t}\right)^2 dt + \frac{k}{2}\left|\left|\xi^n\right|\right|_0^2 = \frac{1}{2}\int_{(n-1)k}^{nk}\int_\Omega\left(\frac{\partial\eta}{\partial t}\right)^2 dt + \frac{k}{2}\left|\left|\xi^n\right|\right|_0^2 \\ &= \frac{1}{2}\int_{(n-1)k}^{nk}\left|\left|\frac{\partial\eta}{\partial t}\right|\right|_0^2 dt + \frac{k}{2}\left|\left|\xi^n\right|\right|_0^2\end{aligned}$$

Then, applying the result

$$\left|\left|\frac{\partial w_h}{\partial t}-\frac{\partial u}{\partial t}\right|\right|_0^2 \le \Gamma^2 h^4\left|\left|\frac{\partial^2}{\partial x^2}\left(\frac{\partial u}{\partial t}\right)\right|\right|_0^2 = \Gamma^2 h^4\left|\left|\frac{\partial^3 u}{\partial x^2\partial t}\right|\right|_0^2,$$

we obtain the final claim of the lemma.

Lemma 3.7.5 *Let $k \le \frac{1}{4}$, then:*

$$\left|\left|\xi^n\right|\right|_0^2 \le 2\left|\left|\xi^0\right|\right|_0^2 + 2k^2\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_{L^2(0,T,L^2(\Omega))}^2 + 2h^4\Gamma^2\left|\left|\frac{\partial^3 u}{\partial x^2\partial t}\right|\right|_{L^2(0,T,L^2(\Omega))}^2 + 4k\sum_{m=0}^{n-1}\left|\left|\xi^m\right|\right|_0^2$$

Where we define

$$\left|\left|f\left(t,x\right)\right|\right|_{L^2(0,T,L^2(\Omega))} = \int_0^T\int_\Omega f^2, \qquad f\in L^2\left((0,T)\times\Omega\right).$$

Proof Subbing the results of the last three lemmas into (3.5) we easily obtain:

$$\frac{1}{2}\left|\left|\xi^m\right|\right|_0^2 \le \frac{1}{2}\left|\left|\xi^{m-1}\right|\right|_0^2 + \frac{k^2}{2}\int_{(m-1)k}^{mk}\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_0^2 dt + \frac{\Gamma^2 h^4}{2}\int_{(m-1)k}^{mk}\left|\left|\frac{\partial^3 u}{\partial x^2\partial t}\right|\right|_0^2 dt + k\left|\left|\xi^m\right|\right|_0^2,$$

or:

$$\left|\left|\xi^m\right|\right|_0^2 \le \left|\left|\xi^{m-1}\right|\right|_0^2 + k^2\int_{(m-1)k}^{mk}\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_0^2 dt + \Gamma^2 h^4\int_{(m-1)k}^{mk}\left|\left|\frac{\partial^3 u}{\partial x^2\partial t}\right|\right|_0^2 dt + 2k\left|\left|\xi^m\right|\right|_0^2.$$

Now we sum for $m=1,\ldots,n$ to get:

$$\begin{aligned}\left|\left|\xi^n\right|\right|_0^2 &\le \left|\left|\xi^0\right|\right|_0^2 + k^2\int_0^{nk}\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_0^2 dt + \Gamma^2 h^4\int_0^{nk}\left|\left|\frac{\partial^3 u}{\partial t\partial x^2}\right|\right|_0^2 dt + 2k\sum_{m=1}^{n-1}\left|\left|\xi^m\right|\right|_0^2 + 2k\left|\left|\xi^n\right|\right|_0^2 \\ (1-2k)\left|\left|\xi^n\right|\right|_0^2 &\le \left|\left|\xi^0\right|\right|_0^2 + k^2\int_0^{nk}\left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_0^2 + \Gamma^2 h^4\int_0^{nk}\left|\left|\frac{\partial^3 u}{\partial x^2\partial t}\right|\right|_0^2 dt + 2k\sum_{m=1}^{n-1}\left|\left|\xi^m\right|\right|_0^2.\end{aligned}$$

Since $k \leq 1/4$ then $1 - 2k \geq 1/2$ we obtain:

$$||\xi^n||_0^2 \leq 2\left|\left|\xi^0\right|\right|_0^2 + 2k^2 \int_0^{nk} \left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_0^2 + 2\Gamma^2 h^4 \int_0^{nk} \left|\left|\frac{\partial^3 u}{\partial x^2 \partial t}\right|\right|_0^2 dt + 4k \sum_{m=0}^{n-1} ||\xi^m||_0^2 .$$

and this competes the proof.

Lemma 3.7.6 (Gronwall Lemma) *If the sequence* $\{\xi^n\}_{n=1}^N$ *satisfies*

$$|\xi^n| \leq \nu + \mu \sum_{m=0}^{n-1} |\xi^m| , \qquad \nu, \mu \geq 0$$

then,

$$|\xi^n| \leq e^{\mu N}\left(\nu + \mu\left|\xi^0\right|\right), \quad \forall n \in \{1, \ldots, N\}$$

Proof Consider the sequence z_n defined by

$$z^n = \nu + \mu \sum_{m=0}^{n} |\xi^m| \quad \Rightarrow \quad z^n - z^{n-1} = \mu\,|\xi^n|$$

so that,

$$\begin{aligned} z^n &\leq (1+\mu)\, z^{n-1} \\ z^{n-1} &\leq (1+\mu)\, z^{n-2} \leq \ldots \leq (1+\mu)^{\frac{(n-1)\mu}{\mu}} z^0 \leq e^{\mu(n-1)} z^0 \leq e^{\mu N} z^0 \end{aligned}$$

because:

$$(1+\mu)^{\frac{1}{\mu}} \leq e.$$

Then,

$$|\xi^n| \leq z^{n-1} \leq e^{\mu N} z^0 = e^{\mu N}\left(\nu + \mu\left|\xi^0\right|\right)$$

for $n = 1, \ldots, N$.

Now let $\mu = 4k$ and

$$\nu = 2\left|\left|\xi^0\right|\right|_0^2 + 2k^2 \left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_{L^2(0,T;L^2(\Omega))}^2 + 2\Gamma^2 h^4 \left|\left|\frac{\partial^3 u}{\partial t \partial x^2}\right|\right|_{L^2(0,T;L^2(\Omega))}^2 .$$

Denoting the final time instant by $T = nk$ and using the results of lemma 3.7.5 we find that

$$||\xi^n||_0^2 \leq e^{4T}\left(3\left|\left|\xi^0\right|\right|_0^2 + 2k^2 \left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_{L^2(0,T;L^2(\Omega))}^2 + 2h^4\Gamma^2 \left|\left|\frac{\partial^3 u}{\partial t \partial x^2}\right|\right|_{L^2(0,T;L^2(\Omega))}^2\right)$$

Let's choose $u_h^0 = w_h^0$ i..e.

$$b\left(u_h^0, v_h\right) = b\left(g, v_h\right)$$

for all $v_h \in V_h$. Then $\xi^0 = 0$. Therefore,

$$||\xi^n||_0^2 \leq e^{4T}\left(2k^2 \left|\left|\frac{\partial^2 u}{\partial t^2}\right|\right|_{L^2(0,T;L^2(\Omega))}^2 + 2\Gamma^2 h^4 \left|\left|\frac{\partial^3 u}{\partial t \partial x^2}\right|\right|_{L^2(0,T;L^2(\Omega))}^2\right) = \mathcal{O}\left(k^2 + h^4\right)$$

Theorem 3.7.7 *Consider the scheme:*

$$\mathbf{M}\frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{k} + \mathbf{A}\mathbf{u}^n = \mathbf{f}$$

If $k < \frac{1}{4}$ *and* $u_h^0 = w_h^0$*, then*

$$||\epsilon_h^n||_0 \leq \mathcal{O}\left(k + h^2\right)$$

Proof

$$||\xi^n||_0^2 \leq a^2k^2 + b^2h^4 \leq \left(ak + bh^2\right)^2$$

where we define:

$$a^2 = 2e^{4T} \left|\left| \frac{\partial^2 u}{\partial t^2} \right|\right|^2_{L^2(0,T;L^2(\Omega))}$$

$$b^2 = 2e^{4T}\Gamma^2 \left|\left| \frac{\partial^3 u}{\partial t \partial x^2} \right|\right|^2_{L^2(0,T;L^2(\Omega))}$$

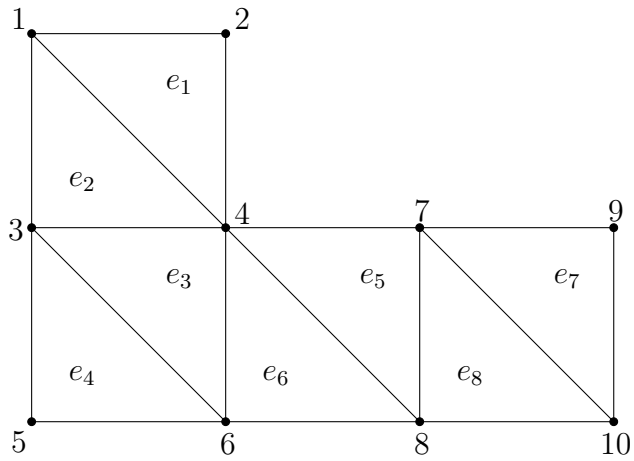
Then,

$$||\epsilon_h^n||_0 \leq ||\eta^n||_0 + ||\xi^n||_0 \leq \sqrt{2}e^{2T} \left|\left| \frac{\partial^2 u}{\partial t^2} \right|\right|_{L^2\left(0,T;L^2(\Omega)\right)} k + h^2\Gamma \left[\max_{0\leq n\leq N} \left|\left| \frac{\partial^2 u^n}{\partial x^2} \right|\right|_0 + \sqrt{2}e^{2T} \left|\left| \frac{\partial^3 u}{\partial t \partial x^2} \right|\right|_{L^2\left(0,T;L^2(\Omega)\right)} \right]$$

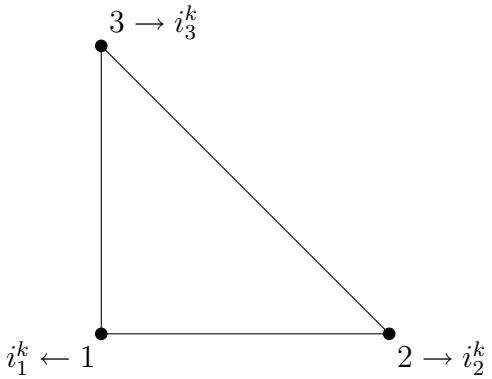
3.8 Finite Elements in Two Dimensions

Consider the model problem:

$$\begin{cases} -\left(pu_x\right)_x - \left(pu_y\right)_y + qu = g & \text{in } \Omega \\ u = \gamma & \text{on } \partial\Omega \end{cases}$$



This is the global numbering, the local numbering (for $i = 1, 2, 3$) is:



where k is the element number and i_j^k is the global number of the j-th local point in the k-th element.

3.8.1 Finite Element Basis Functions

With each global node i we associate one basis funcion $\phi_i(x, y)$. If N_j is the j-th point in the grid, we require:

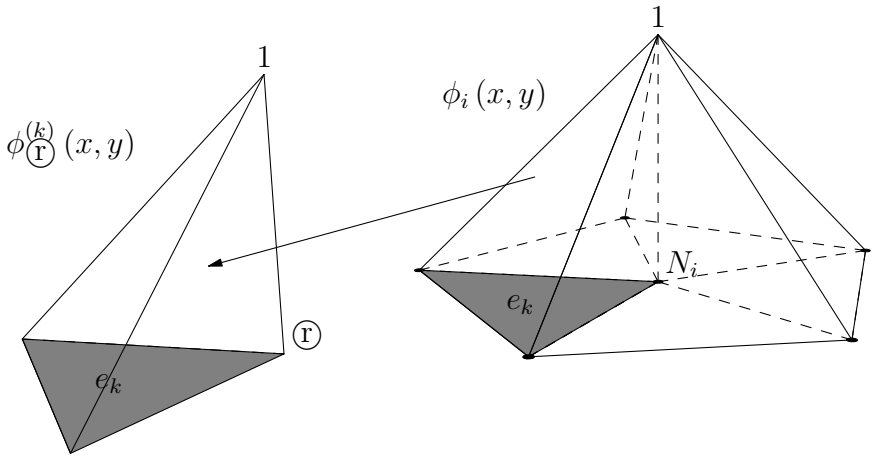
$$\phi_i(N_j) = \delta_{ij}$$

where $\phi_j|_{e_k} \in P^1$, the space of all first degree polynomials. Then,

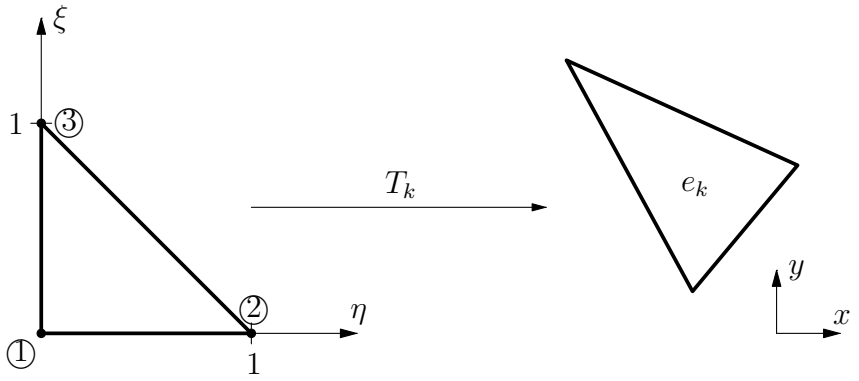
$$V_h = \text{span}\{\phi_1, \ldots, \phi_M\}$$

where M is the total number of points in the grid.

$$u_h(x, y) = \sum_{i=1}^{M} u_i \phi_i(x, y)$$



where $i^k_r = N_i$ and r can be 1, 2, or 3 (local numbers).



Now, we define:

$$T_k \equiv \begin{cases} x = x(\xi, \eta) = a_1\xi + a_2\eta + a_3 \\ y = y(\xi, \eta) = b_1\xi + b_2\eta + b_3 \end{cases}$$

Then,

$$\phi^k_{①}(\xi, \eta) = 1 - \xi - \eta, \quad \phi^k_{②}(\xi, \eta) = \xi, \quad \phi^k_{③}(\xi, \eta) = \eta$$

So that:

$$\int_\Omega \phi_i(x, y)\, \phi_j(x, y)\, d\Omega = \sum_{k=1}^{K_e} \int_{e_k} \phi_i(\xi, \eta)\, \phi_j(\xi, \eta)\, de_k$$

Where K_e is the total number of finite elements, $de_k = |\mathcal{J}_{T_k}|\, d\xi d\eta$, and $\mathcal{J}_{T_k}$ is the Jacobian of T_k.

3.8.2 Discrete Galerkin Formulation

Find $u_h \in V_h$ such that:

$$\int_\Omega p u_{h,x} v_{h,x} + \int_\Omega p u_{h,y} v_{h,y} + \int_\Omega q u_h v_h = \int_\Omega g v_h$$

and, for simplicity, we choose $\gamma \equiv 0$, $p =$ constant and $q =$ constant. Then, substitutiting

$$u_h = \sum_{i=1}^{M} u_i \phi_i$$

and $v_h = \phi_j$ for $j = 1, \ldots, M$ yields the linear system:

$$(p\mathbf{S} + q\mathbf{M})\,\mathbf{u}_h = \mathbf{G}$$

where:

$$S_{ij} = p \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, d\Omega \quad \text{(the stiffness matrix)}$$
$$M_{ij} = q \int_\Omega \phi_i \phi_j \, d\Omega \quad \text{(the mass matrix)}$$