MA2501 Fnite difference methods for two-point boundary value problems

Elena Celledoni

April 8, 2019

Contents

Chapter 1

Background material

1.1 Background on matrix theory

Let A be a $n \times n$-matrix with real (or complex) entries, we write $A \in \mathbf{R}^{n\times n}$ (or $A \in \mathbf{C}^{n\times n}$). We say that A *is diagonalizable* if it exist a matrix $X \in \mathbf{C}^{n\times n}$ such that

$$\Lambda = X^{-1}AX = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Every $\lambda_i \in \mathbf{C}$ is called *egenvalue* of A. The matrix X has n columns denoted by $X = [x_1, \ldots, x_n]$, and every $x_i \in \mathbf{C}^n$ is called *egenvector* (*associated to the egenvalue* λ_i). We write a diagonal matrix as Λ above, in the following form

$$\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

1.1.1 Jordan form

For any $A \in \mathbf{R}^{n\times n}$ (or $A \in \mathbf{C}^{n\times n}$) it exists a matrix $M \in \mathbf{C}^{n\times n}$ such that

$$M^{-1}AM = J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{bmatrix}, \qquad \text{(block-diagonal)}. \tag{1.1}$$

Here J_i is a $m_i \times m_i$-matrix, and $\sum_{i=1}^{k} m_i = n$. The *Jordan-blocks* J_i have the form

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}, \quad \text{if } m_i \geq 2$$

and $J_i = [\lambda_i]$ if $m_i = 1$. If all $m_i = 1$, then $k = n$ and the matrix is diagonalizable. If A has n distinct egenvalues, it is always diagonalizable. The converse is not true, that is a matrix can be diagonalizable even if it has multiple eigenvalues.

1.1.2 Symmetric matrices

When we talk about symmetric matrices, we mean normally *real* symmetric matrices. The *transpose* A^T of a $m \times n$-matrix A, is a $n \times m$-matrix with a_{ji} as the (ij)-element (a matrix whose columns are the rows of A). A $n \times n$ matrix is symmetric if $A^T = A$.

A symmetric $n \times n$ matrix has real eigenvalues $\lambda_1, \dots, \lambda_n$ and a set of real orthonormal eigenvectors $x_1, \dots, x_n$. Let $\langle \cdot, \cdot \rangle$ denote the standard inner-product on $\mathbf{C}^n$, then $\langle x_i, x_j \rangle = \delta_{ij}$ (Kronecker-delta).

A consequence of this is that the matrix of eigenvectors $X = [x_1, \dots, x_n]$ is real and orthogonal and its inverse is therefore the transpose

$$X^{-1} = X^T.$$

The diagonalization of A is given by

$$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n), \quad X = [x_1, \dots, x_n], \quad X^T X = I, \quad X^T A X = \Lambda \Leftrightarrow A = X \Lambda X^T$$

1.1.3 Positive definite matrices

If A is symmetric and $\langle x, Ax \rangle = x^T A x > 0$ for all $0 \neq x \in \mathbf{R}^n$ A is called *positive definite.*

A (symmetric) is positive semi-definite if $\langle x, Ax \rangle \geq 0$ for all $x \in \mathbf{R}^n$ and $\langle x, Ax \rangle = 0$ for at least a $x \neq 0$.

A positive definite $\Leftrightarrow$ A has only positive eigenvalues.

A positive semi-definite $\Leftrightarrow$ A has only non-negative eigenvalues, and at least a 0-eigenvalue.

1.1.4 Gershgorin's theorem

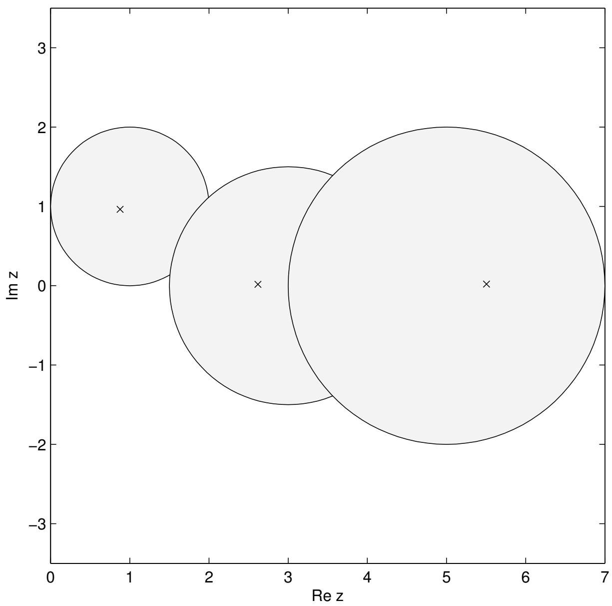
Gershgorin's theorem. Is given $A = (a_{ik}) \in \mathbf{C}^{n \times n}$. Define n disks S_j in the complex plane by

$$S_j = \left\{ z \in \mathbf{C} : |z - a_{jj}| \leq \sum_{k \neq j} |a_{jk}| \right\}.$$

The union $S = \bigcup_{j=1}^{n} S_j$ contains all the eigenvalues of A. For every eigenvalue λ of A there is a j such that $\lambda \in S_j$.

Example.

$$A = \begin{bmatrix} 1+i & 1 & 0 \\ 0.5 & 3 & 1 \\ 1 & 1 & 5 \end{bmatrix}.$$



,

□

Proof of Gershgorin's theorem: Let λ be a eigenvalue with associate eigenvector $x = [\xi_1, \ldots, \xi_n]^T \neq 0$. Choose ℓ among the indexes $1, \ldots, n$ such that $|\xi_\ell| \geq |\xi_k|,\ k = 1, \ldots, n$, and so $|\xi_\ell| > 0$. The equation $Ax = \lambda x$ has component ℓ:

$$\sum_{k=1}^{n} a_{\ell k} \xi_k = \lambda\, \xi_\ell \;\Rightarrow\; (\lambda - a_{\ell\ell})\xi_\ell = \sum_{k \neq \ell} a_{\ell k} \xi_k$$

Divide by $|\xi_\ell|$ on each side and take the absolute value

$$|\lambda - a_{\ell\ell}| \quad = \quad \left| \sum_{k \neq \ell} a_{\ell k} \frac{\xi_k}{\xi_\ell} \right| \quad \leq \quad \sum_{k \neq \ell} |a_{\ell k}| \frac{|\xi_k|}{|\xi_\ell|} \quad \leq \quad \sum_{k \neq \ell} |a_{\ell k}|$$

Then we get $\lambda \in S_\ell$.

Example. Diagonally dominant matrices with positive diagonal elements are positive definite. Why?

1.1.5 Vector and matrix norms

Consider a vector space X (real or complex). A norm $\| \cdot \| : X \to \mathbf{R}$ satisfies the following axioms

1. $\|x\| \geq 0$ for all x, $\|x\| = 0 \Leftrightarrow x = 0$.
2. $\|\alpha x\| = |\alpha| \|x\| \quad (\alpha \in \mathbf{R}\ (\mathbf{C}))$
3. $\|x + y\| \leq \|x\| + \|y\|$

Examples. $x = (\xi_k)$, $X = \mathbf{R}^n$.

$$\|x\|_1 = \sum_{k=1}^{n} |\xi_k|, \qquad \|x\|_2 = \left(\sum_{k=1}^{n} |\xi_k|^2 \right)^{1/2}, \qquad \|x\|_\infty = \max_{1 \leq k \leq n} |\xi_k|.$$

The matrix spaces $\mathbf{R}^{n\times n}$ and $\mathbf{C}^{n\times n}$ are also vector spaces over $\mathbf{R}$ ($\mathbf{C}$). We say that $\|\cdot\|$ is a matrix norm if for all $A, B \in \mathbf{R}^{n\times n}$ ($\mathbf{C}^{n\times n}$)

1. $\|A\| > 0$ for all A, $\|A\| = 0 \Leftrightarrow A = 0$,
2. $\|\alpha A\| = |\alpha|\|A\|, \quad (\alpha \in \mathbf{R}\ (\mathbf{C}))$,
3. $\|A + B\| \leq \|A\| + \|B\|$,
4. $\|A \cdot B\| \leq \|A\| \cdot \|B\|$.

Remark. The last point requires that a matrix-matrix product is defined (this operation is not defined in a general vector space). In abstract terms the axioms 1–4 give an example of *Banach-algebra*.

Example. The Frobenius-norm of a matrix is defined as

$$\|A\|_F = \left(\sum_{j=1}^{n}\sum_{k=1}^{n}|a_{jk}|^2\right)^{1/2}.$$

1.1.6 Consistent and subordinate matrix norms

A given matrix norm is *consistent* with a given vector norm on $\mathbf{R}^n$ if

$$\|Ax\| \leq \|A\| \cdot \|x\| \quad \text{for all} \quad A \in \mathbf{R}^{n\times n},\ x \in \mathbf{R}^n.$$

A given matrix norm is *subordinate* to a given vector norm on $\mathbf{R}^n$ if

$$\|A\| = \max_{x\neq 0}\frac{\|Ax\|}{\|x\|}.$$

Examples. We give here as examples some of the most common subordinate matrix norms. We look for matrix norms subordinate to the three vector norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$.

1. Let $\|\cdot\|_1$ be the matrix norm subordinate to the vector norm $\|\cdot\|_1$. One can show that $A \in \mathbf{R}^{n\times n}$ ($\mathbf{C}^{n\times n}$) is

$$\|A\|_1 = \max_{1\leq k\leq n}\sum_{i=1}^{n}|a_{ik}|.$$

   In other words we can say that $\|A\|_1$ is the "maximal column-sum in A".

2. To find the matrix norm subordinate to the vector norm $\|\cdot\|_2$ we must define the *spectral radius* of a matrix $M \in \mathbf{R}^{n\times n}$ ($\mathbf{C}^{n\times n}$). If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of M, we denote the spectral radius of M by $\rho(M)$, and it is defined as

$$\rho(M) = \max_{1\leq k\leq n}|\lambda_k|. \tag{1.2}$$

   If we plot the eigenvalues of M in the complex plane, the spectral radius is the minimal radius of a circle centered in the origin and containing all eigenvalues of M.

   We define now the 2-norm of a matrix A as

$$\|A\|_2 = \sqrt{\rho(A^TA)}.$$

Note that $A^T A$ is positive (semi)definite, so all the eigenvalues are real and positive. Taking the square root of the biggest eigenvalue, we obtain $\|A\|_2$. Note also that the spectral radius of A can be very different from (the square root of) the spectral radius of $A^T A$. On the other hand if A is symmetric then $\|A\|_2 = \rho(A)$.

3. Let $\|\cdot\|_\infty$ be the matrix norm subordinate to the vector norm $\|\cdot\|_\infty$. We have

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{k=1}^{n} |a_{ik}|.$$

That is $\|A\|_\infty$ is the "maximal row-sum in A". Observe also that $\|A\|_1 = \|A^T\|_\infty$.

1.1.7 Matrix norms and spectral radius.

For any matrix norm $\|\cdot\|$ it is true that

$$\|A\| \ge \rho(A). \tag{1.3}$$

Proof: Let x be a eigenvector of A associated to an eigenvalue λ such that

$$Ax = \lambda x.$$

Let $y \in \mathbf{C}^n$ be arbitrary. Then we have

$$A(xy^T) = (Ax)y^T = \lambda(xy^T),$$

such that

$$\|A(xy^T)\| \le \|A\|\ \|xy^T\|.$$

As a consequence

$$|\lambda|\ \|xy^T\| = \|\lambda(xy^T)\| = \|A(xy^T)\| \le \|A\|\ \|xy^T\|.$$

Therefore $|\lambda| \le \|A\|$, and since this is true for every eigenvalue of A, it must be $\rho(A) \le \|A\|$.

Question to the reader: What is wrong with the following line of reasoning

$$|\lambda|\ \|x\| = \|\lambda\, x\| = \|A\, x\| \le \|A\|\ \|x\| \quad \text{etc.}$$

Convergent matrix. A matrix A is said to be *convergent* (to zero) if

$$A^k \to 0 \quad \text{when} \quad k \to \infty.$$

A sufficient criterion. If $\|A\| < 1$ for a particular matrix norm, A is convergent.

Proof.

$$\|A^k\| = \|A \cdot A^{k-1}\| \le \|A\| \cdot \|A^{k-1}\| \le \cdots \le \|A\|^k \ \to 0 \qquad \text{if } \|A\| < 1$$

Necessary and sufficient criterion. A is convergent if and only if the spectral radius $\rho(A)$, defined by (1.2), satisfies $\rho(A) < 1$.

Proof: We use Jordan form, and let $A = MJM^{-1}$ where $M \in \mathbf{C}^{n\times n}$ and J is like in (1.1). Then we have $A^2 = MJM^{-1}MJM^{-1} = MJ^2M^{-1}$, and by induction we get $A^k = MJ^kM^{-1}$. Now $A^k \to 0$ if and only if $J^k \to 0$. And $J^k \to 0$ if and only if every Jordan block $J_i^k \to 0$. Assume such a Jordan block has diagonal element λ and the $m_i \times m_i$-matrix F has its $(j, j+1)$ elements, for $j = 1, \ldots, m_i - 1$, equal to 1, and the other elements equal to zero. Then $J_i = \lambda I + F$ where I is the identity matrix. The matrix F is nihilpotent, i.e. $F^m = 0,\ m \geq n$. We assume that $k \geq n-1$ and compute

$$J_i^k = (\lambda I + F)^k = \sum_{m=0}^{k} \binom{k}{m} \lambda^{k-m} F^m = \sum_{m=0}^{n-1} \binom{k}{m} \lambda^{k-m} F^m = \sum_{m=0}^{n-1} \varphi_k^{(m)}(\lambda) F^m$$

where $\varphi_k(\lambda) = \lambda^k/k!$. When $k \to \infty$ then $\varphi_k^{(m)}(\lambda) \to 0$ for $0 \leq m \leq n-1$ if and only if $|\lambda| < 1$. This must be true for all Jordan blocks (i.e. eigenvalues of A) and this concludes the proof. □

1.2 Difference formulae

1.2.1 Taylor expansion

1 free variable. Let $u \in C^{n+1}(I)$ where $I \subset \mathbf{R}$ is a interval of the real line. This means that the $n+1$-th derivative of u exists and is continuous on the interval I. Then the following formula is valid.

Taylor's formula with reminder. With $x \in I$, $x + h \in I$ is

$$u(x+h) = \sum_{m=0}^{n} \frac{h^m}{m!} u^{(m)}(x) + r_n$$

where

$$r_n = \frac{h^{n+1}}{(n+1)!} u^{(n+1)}(x + \theta h), \quad 0 < \theta < 1.$$

2 free variables. Assume now that $u \in C^{n+1}(\Omega)$ where $\Omega \subset \mathbf{R}^2$. It is convenient to use a operatornotation for the partial derivatives. We write $\mathbf{h} = [h, k]$, and let

$$\mathbf{h} \cdot \nabla := h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \quad \text{i.e.} \quad \mathbf{h} \cdot \nabla u = h\frac{\partial u}{\partial x} + k\frac{\partial u}{\partial y}$$

The operator produces the derivative of a function in the direction $\mathbf{h} = [h, k]$, and we find the *directional derivative.*

We can also define powers of the operator by for example

$$(\mathbf{h} \cdot \nabla)^2 = \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 = h^2\frac{\partial^2}{\partial x^2} + 2hk\frac{\partial^2}{\partial x \partial y} + k^2\frac{\partial^2}{\partial y^2}$$

The extension to the m-th power is obvious. then we can write

Taylor's formula with reminder for functions of two variables.

$$u(x+h, y+k) = \sum_{m=0}^{n} \frac{1}{m!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \right)^m u(x, y) + r_n \tag{1.4}$$

where

$$r_n = \frac{1}{(n+1)!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \right)^{n+1} u(x+\theta h, y+\theta k), \quad 0 < \theta < 1.$$

We have here assumed that the line segment between (x, y) and $(x + h, y + k)$ is included in Ω.

Derivation of the previous formula. We look at a function of one variable $\mu(t) = u(x+th, y+tk)$ for fixed x, y, h, k. Using Taylor's expansion with reminder for the case of one variable for $\mu(t)$ around $t = 0$, we obtain the two variables formula by setting $t = 1$.

1.2.2 Big $\mathcal{O}$-notation

Let ϕ be a function of h and p a positive integer. Then we have

$$\phi(h) = \mathcal{O}(h^p) \quad \text{when} \quad h \to 0$$

if there exist two constants $C, H > 0$ such that

$$|\phi(h)| \le C|h|^p \quad \text{when} \quad 0 < |h| < H.$$

If this holds, we say that $\phi(h)$ is of *order* p in the variable h.

The typical use of the big $\mathcal{O}$-notation is in connection with the local truncation error in numerical methods. For example in the Taylor expansion in one variable

$$|r_n| = \left| \frac{h^{n+1}}{(n+1)!} u^{(n+1)}(x+\theta h) \right| \le \frac{M}{(n+1)!} |h|^{n+1}, \quad M = \max_{y \in I} |u^{(n+1)}(y)|$$

where we know that the maximum exists if I is a closed, and bounded interval. So in this case we have $r_n = \mathcal{O}(h^{n+1})$.

Note that with the definition above for positive integers p we have

$$\phi(h) = \mathcal{O}(h^{p+1}) \quad \Rightarrow \quad \phi(h) = \mathcal{O}(h^p).$$

Therefore sometimes when we write $\phi(h) = \mathcal{O}(h^p)$ we mean that p is the biggest possible integer such that this is true. Often it is convenient to write $\mathcal{O}(h^p)$ in formulae with sums, like for example in the Taylor expansion of u above we replace r_n with $\mathcal{O}(h^{n+1})$ such that

$$u(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} u^{(m)}(x) + \mathcal{O}(h^{n+1}).$$

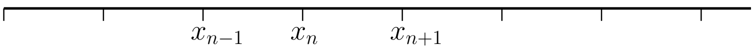
In general we have that $\phi(h) = \mathcal{O}(h^{p_\phi})$ and $\psi(h) = \mathcal{O}(h^{p_\psi})$, then

$$\psi(h) + \phi(h) = \mathcal{O}(h^q), \quad \text{where} \quad q = \min(p_\phi, p_\psi).$$

But sometimes one can get higher powers. An obvious example is when $\phi(h) = h^2$, $\psi(h) = h^3 - h^2$, each of them is $\mathcal{O}(h^2)$ but their sum is $\mathcal{O}(h^3)$. If you multiply a function $\phi(h)$ by a constant ($\neq 0$), the order does not change.

1.2.3 Difference approximations to the derivatives

We introduce a *grid* on $\mathbf{R}$ i.e. a monotone sequence of real numbers $\{x_n\}$ where $x_n \in \mathbf{R}$.



Assume that $u(x)$ is a given function, $u \in C^q(I)$, for a q which we will specify later. Let

$$u_n := u(x_n), \qquad u_n^{(m)} := u^{(m)}(x_n).$$

Assume the grid points x_n are equidistant, i.e. $x_{n+1} = x_n + h$ for all n, where $h \in \mathbf{R}$ is called *step-size.* We want to approximate $u_n^{(m)}$ with expressions of the type

$$\sum_{\ell=p}^{q} a_\ell u_{n+\ell}$$

$p \leq q$ are integers, and typically $p \leq 0$ and $q \geq 0$.

Truncation error. We define

$$\tau_n(h) = \sum_{\ell=p}^{q} a_\ell u_{n+\ell} - u_n^{(m)}.$$

The strategy is to choose p and q, and then compute the $q-p+1$ parameters $a_p, \ldots, a_q$ such that τ_n is "small".

By Taylor expansion we obtain

$$u_{n+\ell} = u(x_n + \ell h) = \sum_{k=0}^{\nu} \frac{(\ell h)^k}{k!} u_n^{(k)} + r_\nu,$$

where $r_\nu = \mathcal{O}(h^{\nu+1})$ and $\nu \geq m$, such that

$$\tau_n = \sum_{\ell=p}^{q} a_\ell \sum_{k=0}^{\nu} \frac{1}{k!} (\ell h)^k u_n^{(k)} - u_n^{(m)} + \mathcal{O}(h^{\nu+1}),$$

which can be rearranged in the form

$$\tau_n = \sum_{k=0}^{\nu} \frac{h^k}{k!} \left(\sum_{\ell=p}^{q} a_\ell \ell^k \right) u_n^{(k)} - u_n^{(m)} + \mathcal{O}(h^{\nu+1}).$$

We want that $\tau_n = \mathcal{O}(h^r)$ with r as big as possible. To approximate $u_n^{(m)}$ we need to impose conditions on p and q. Set $j := q - p$. In order to get consistent approximation formulae (i.e. such that $\tau_n(h) \to 0$ when $h \to 0$), we must require $j \geq m$. We choose then $a_p, \ldots, a_q$ such that

$$\frac{h^k}{k!} \sum_{\ell=p}^{q} \ell^k \, a_\ell = \begin{cases} 0 & 0 \leq k \leq m-1, \\ 1 & k = m, \\ 0 & m+1 \leq k \leq j. \end{cases} \tag{1.5}$$

Note that we have $q-p+1 = j+1$ free parameters $a_p, \ldots, a_q$ we can use, and the conditions in (1.5) must be satisfied for $0 \leq k \leq j$, this means a total of $j+1$ conditions. The system of equations has a unique solution for $h \neq 0$. If we choose a_ℓ from (1.5), and assume $\nu \geq j$, we obtain the following truncation error

$$\tau_n = \sum_{k=j+1}^{\nu} \frac{h^k}{k!} u_n^{(k)} \sum_{\ell=p}^{q} a_\ell \, \ell^k + \mathcal{O}(h^{\nu+1}).$$

This methods is called *the method of undetermined coefficients.*

Example. $m = 1$ (u'_n). Choose $p = -1,\ q = 1,\ j = 2$. We want to find a_{-1}, a_0, a_1. We write $j + 1 = 3$ equations i.e. $k = 0, 1, 2$ i (1.5).

$$\left.\begin{array}{lccccccc}
k=0 & a_{-1} & + & a_0 & + & a_1 & = & 0 \\
k=1 & -h\,a_{-1} & + & 0\cdot a_0 & + & h\,a_1 & = & 1 \\
k=2 & h^2\,a_{-1} & + & 0\cdot a_0 & + & h^2\,a_1 & = & 0
\end{array}\right] \Rightarrow \begin{array}{lcr}
a_{-1} & = & -\frac{1}{2h} \\
a_0 & = & 0 \\
a_1 & = & \frac{1}{2h}
\end{array}$$

Looking at the first terms in the local truncation error we obtain

$$\tau_n = \sum_{k=3} \frac{h^k}{k!} u_n^{(k)} \left(-\frac{1}{2h}(-1)^k + \frac{1}{2h}\, 1^k \right) = \sum_{s=1} \frac{u_n^{(2s+1)}}{(2s+1)!} h^{2s}.$$

In the last equality, we have used the fact that the terms with even k disappear, such that we can put $k = 2s + 1$ and let $s = 1, 2, \ldots$. We have omitted the upper limit value for the index s on purpose because the number of terms we include in the remainder depend on the circumstances. Since the first term in the expression for τ_n is of type h^2, we say that the formula is of *order* 2.

Some more formulae. Other popular difference approximations are

$$\begin{array}{lccl}
m=1 & \dfrac{u_{n+1}-u_n}{h} & = & u'_n + \frac{1}{2!}h\,u''_n + \cdots \\[2ex]
m=1 & \dfrac{u_n-u_{n-1}}{h} & = & u'_n - \frac{1}{2!}h\,u''_n + \cdots \\[2ex]
m=2 & \dfrac{u_{n+1}-2u_n+u_{n-1}}{h^2} & = & u''_n + \frac{1}{12}h^2 u_n^{(4)} + \cdots
\end{array}$$

Which order have these formulae?

Exercises

1. Assume $m = 1$, $p = -2$, $q = 0$ use the method of undetermined coefficients to obtain an approximation of $u_n^{(1)}$ of the second order in h.
2. Assume $m = 1$, $p = -2$, $q = 1$ use the method of undetermined coefficients to obtain an approximation of $u_n^{(1)}$ of the third order in h.
3. Consider $m = 2$, $p = 0$, $q = 1$, so that $j = q - p = 1 < m$. Try to construct an approximation formula for $u_n^{(2)}$ using the method of undetermined coefficients. What happens?

1.2.4 Difference operators and other operators

Forward difference:	$\Delta u(x) = u(x+h) - u(x)$.
Backward difference:	$\nabla u(x) = u(x) - u(x-h)$.
Central difference:	$\delta u(x) = u(x+\frac{h}{2}) - u(x-\frac{h}{2})$.
Mean value:	$\mu u(x) = \frac{1}{2}\left(u(x+\frac{h}{2}) + u(x-\frac{h}{2})\right)$.
Shift:	$E u(x) = u(x+h)$.
Unity operator	$1\,u(x) = u(x)$.

Linearity. All the operators

$$\Delta,\ \nabla,\ \delta,\ \mu,\ E,\ 1,$$

are linear. This means that for $\alpha \in \mathbf{R}$, and with functions $u(x)$ and $v(x)$ we have

$$F(\alpha u(x) + v(x)) = \alpha F u(x) + F v(x),$$

where F can be any of the operators above. Let us verify this for $F = \Delta$.

$$\begin{aligned}\Delta(\alpha u(x) + v(x)) &= (\alpha u(x+h) + v(x+h)) - (\alpha u(x) + v(x)) \\ &= \alpha(u(x+h) - u(x)) + (v(x+h) - v(x)) = \alpha\Delta u(x) + \Delta v(x).\end{aligned}$$

Powers of the operators. Let F be one of the above defined operators. We can define powers of F as follows

$$F^0 = 1, \qquad F^k u(x) = F(F^{k-1}u(x)).$$

Example.

$$\begin{aligned}\delta u(x) &= u(x+\tfrac{h}{2}) - u(x-\tfrac{h}{2}), \\ \delta^2 u(x) &= \delta(\delta u(x)) = \delta u(x+\tfrac{h}{2}) - \delta u(x-\tfrac{h}{2}) = u(x+h) - u(x) - (u(x) - u(x-h)), \\ &= u(x+h) - 2u(x) + u(x-h).\end{aligned}$$

Another interesting example is the shift operator. We observe that $E^k u(x) = u(x+kh)$. In this case it is easy to extend the definition to include all possible real powers, simply by defining $E^s u(x) = u(x+sh)$ for all $s \in \mathbf{R}$. For example we have $E^{-1}u(x) = u(x-h)$ and this is the inverse of E since $Eu(x-h) = E^{-1}u(x+h) = u(x)$.

Relations between the difference operators.

$$\begin{aligned}\Delta u(x) &= u(x+h) - u(x) = Eu(x) - 1\,u(x) = (E-1)\,u(x), \\ \nabla u(x) &= u(x) - u(x-h) = 1\,u(x) - E^{-1}u(x) = (1 - E^{-1})\,u(x), \\ \delta u(x) &= u(x+\frac{h}{2}) - u(x-\frac{h}{2}) = (E^{1/2} - E^{-1/2})\,u(x), \\ \mu u(x) &= \frac{1}{2}\left(u(x+\frac{h}{2}) + u(x-\frac{h}{2})\right) = \frac{1}{2}(E^{1/2} + E^{-1/2})\,u(x).\end{aligned}$$

In a more compact notation we have

$$\begin{aligned}\Delta &= (E-1), \\ \nabla &= (1 - E^{-1}), \\ \delta &= (E^{1/2} - E^{-1/2}), \\ \mu &= \frac{1}{2}(E^{1/2} + E^{-1/2}).\end{aligned}$$

And now we have for example

$$\Delta^k = (E-1)^k = \sum_{\ell=0}^{k} \binom{k}{\ell} (-1)^{k-\ell} E^{\ell},$$

such that

$$\Delta^k u(x) = \sum_{\ell=0}^{k} \binom{k}{l} (-1)^{k-\ell} E^{\ell} u(x) = \sum_{\ell=0}^{k} \binom{k}{l} (-1)^{k-\ell} u(x+\ell h).$$

1.2.5 Differential operator.

Define

$$D = \frac{\mathrm{d}}{\mathrm{d}x} \quad \text{so that} \quad Du(x) = u'(x).$$

Let

$$D^m u(x) = u^{(m)}(x).$$

If $u(x)$ is analytic[1] in an interval containing $x, x+h$ we have

$$u(x+h) = \sum_{m=0}^{\infty} \frac{h^m}{m!} D^m u(x) = \left(\sum_{m=0}^{\infty} \frac{1}{m!} (hD)^m \right) u(x) = \mathrm{e}^{hD}\, u(x).$$

We think of this only as a *notation*. We have

$$Eu(x) = \mathrm{e}^{hD} u(x),$$

and then $E = \mathrm{e}^{hD}$.

Relation between D and the other operators.

$$\begin{aligned} \Delta &= E - 1 = \mathrm{e}^{hD} - 1 = \sum_{m=1}^{\infty} \frac{1}{m!} (hD)^m, \\ \Delta &= hD + \frac{1}{2!} (hD)^2 + \cdots \end{aligned}$$

We will see that under the extra assumption that u is analytic we can make manipulations with analytic functions in the way we are used to. The meaning is always that the final result is expanded with a Taylor expansion and is interpreted as a sum of powers of operators which are applied to a smooth function. The analyticity requirement can always be relaxed by considering a Taylor expansion with reminder, and requiring the function to be differentiable only a finite number of times.

We consider powers of Δ, and we obtain

$$\Delta^k = \left(\sum_{m=1}^{\infty} \frac{(hD)^m}{m!} \right)^k = h^k D^k + \frac{k}{2!} h^{k+1} D^{k+1} + \cdots$$

or

$$\Delta^k u(x) = h^k D^k u(x) + \frac{k}{2!} h^{k+1} D^{k+1} u(x) + \cdots$$

showing that Δ^k / h^k is a first order approximation (truncation error $\mathcal{O}(h)$) of the operator D^k.

Note that for $s \in \mathbf{R}$ we have

$$E^s u(x) = u(x + sh) = \sum_{k=0}^{\infty} \frac{(sh)^k}{k!} D^k u(x) = \mathrm{e}^{shD} u(x)$$

which reflects known computational rules. For central differences we can therefore write

$$\delta = E^{1/2} - E^{-1/2} = \mathrm{e}^{\frac{1}{2}hD} - \mathrm{e}^{-\frac{1}{2}hD} = 2 \sinh \frac{hD}{2}.$$

[1]By analytic function on an interval we simply mean that its Taylor expansion converges in a neighborhood of any point of the interval.

We can also compute

$$\delta^k = \left(2\sinh\frac{hD}{2}\right)^k = \left(hD + \frac{2}{3!}\left(\frac{hD}{2}\right)^3 + \cdots\right)^k = (hD)^k + \frac{k}{24}(hD)^{k+2} + \cdots$$

that is

$$\delta^k u(x) = h^k D^k u(x) + \frac{k}{24} h^{k+2} D^{k+2} u(x) + \cdots$$

this shows that δ^k/h^k is a second order approximation of D^k.

In particular we find as before that

$$\delta^2 u(x) = u(x+h) - 2u(x) + u(x-h) = h^2 u''(x) + \frac{1}{12} h^4 u^{(4)}(x) + \cdots \tag{1.6}$$

It is tempting to manipulate further with analytic functions. We have seen that

$$\frac{\delta}{2} = \sinh\frac{hD}{2}.$$

We write therefore formally

$$D = \frac{2}{h}\sinh^{-1}\frac{\delta}{2}.$$

It is possible to expand $\sinh^{-1} z$ in a Taylor expansion

$$\sinh^{-1} z = z - \frac{1}{6}z^3 + \frac{3}{40}z^5 - \frac{5}{112}z^7 + \cdots$$

so $z = \delta/2$ and by multiplying by $2/h$ we obtain

$$D = \frac{1}{h}\left(\delta - \frac{1}{24}\delta^3 + \frac{3}{640}\delta^5 - \frac{5}{7168}\delta^7 + \cdots\right).$$

Since we know that $\delta^k = \mathcal{O}(h^k)$ we see that we can find approximations to the differential operator D of arbitrary high order by including enough terms in the expansion. The manipulation we have carried out is not rigorously justified here, but it turns out to be correct. For a detailed discussion on algebraic manipulations with differential operators, see the textbook by Arieh Iserles, *A first course in the numerical analysis of differential equations*, published by Cambridge University Press.

Chapter 2

Boundary value problems

2.1 A simple case example

We consider the boundary value problem

$$u_{xx} = f(x), \qquad 0 < x < 1, \quad u(0) = \alpha, \quad u(1) = \beta, \tag{2.1}$$

the exact solution can be obtained by integrating twice on both sides between 0 and 1, and then imposing the boundary conditions. We want to use this simple test problem to illustrate some of the basic features of finite difference discretization methods.

To obtain a finite difference discretization for this problem we consider the grid

$$x_m = m\,h,\ m = 0, \dots, M+1, \quad h = \frac{1}{M+1},$$

and the notation $u_m := u(x_m)$ such that $u_0 = \alpha$ and $u_{M+1} = \beta$. We denote with capital letters $U_m \approx u_m$ the numerical approximation to $u(x)$ at the grid point $x = x_m$.

By replacing derivatives with central difference approximations to the left hand side of (2.1) we obtain the so called discrete problem whose solution is the numerical approximation that we are seeking, this is

$$\frac{1}{h^2}\left(U_{m-1} - 2U_m + U_{m+1}\right) = f_m, \quad m = 1, \dots, M. \tag{2.2}$$

This is a linear system of equations

$$A_h \mathbf{U} = \mathbf{F}, \tag{2.3}$$

where A_h is a $M \times M$ matrix $\mathbf{U} \in \mathbf{R}^M$ and $\mathbf{F} \in \mathbf{R}^M$ and

$$A_h := \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_1 \\ \vdots \\ U_M \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} f_1 - \frac{\alpha}{h^2} \\ f_2 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

We know that $u''(x_m) = \frac{1}{h^2}\,\delta^2 u(x_m) + \mathcal{O}\left(h^2\right)$ and we want to deduce similar information about the error

$$e_m := U_m - u_m, \quad e_0 = 0,\ e_{M+1} = 0.$$

Let the error vector $\mathbf{e}_h$ be

$$\mathbf{e}_h := \mathbf{U} - \mathbf{u}, \qquad \mathbf{u} := \begin{bmatrix} u_1 \\ \vdots \\ u_M \end{bmatrix}.$$

We will in the sequel associate such vector to a piecewise constant function $e_h(x)$ defined on the interval $[0,1]$ as follows

$$e_h(x) = e_m, \quad x \in [x_m,\, x_{m+1}), \quad m = 1, \dots, M.$$

Because we are approximating a function, $u(x)$ solution of (2.1), it is appropriate to think of the numerical solution as a piecewise constant function approximating $u(x)$ and similarly for the error. We are therefore interested in measuring the norm of this piecewise constant error function rather than the norm of the corresponding error vector, however the two are closely related. In fact we can see that the following relationships hold:

- $\|e_h\|_\infty = \max_{0\le x\le 1} |e_h(x)| = \max_{1\le m\le M} |e_m| = \|\mathbf{e}_h\|_\infty$;
- $\|e_h\|_1 = \int_0^1 |e_h(x)|\, dx = \sum_{m=1}^M \int_{x_m}^{x_{m+1}} |e_h(x)|\, dx = h \sum_{m=1}^M |e_m| = h\|\mathbf{e}_h\|_1$;
- $\|e_h\|_2 = \left(\int_0^1 |e_h(x)|^2\, dx\right)^{\frac{1}{2}} = \left(h \sum_{m=1}^M |e_m|^2\right)^{\frac{1}{2}} = h^{\frac{1}{2}} \|\mathbf{e}_h\|_2$;

and we see that in these three popular cases the vector norm is related to the function norm of the corresponding piecewise constant function (with respect to the assumed grid) simply by a scaling factor. A similar result is true for the case of $\|\cdot\|_q$.

Truncation error. The truncation error is the vector that by definition has components

$$\tau_m := \frac{1}{h^2}\left(u_{m-1} - 2u_m + u_{m+1}\right) - f_m, \quad m = 1, \dots, M, \quad \tau_h := \begin{bmatrix} \tau_1 \\ \vdots \\ \tau_M \end{bmatrix}.$$

By using that $u''(x_m) = \frac{1}{h^2}\,\delta^2 u(x_m) - \frac{1}{12}h^2 u_m^{(4)} + \mathcal{O}\left(h^4\right)$ and $u''_m = f_m$, we obtain

$$\tau_m = u''_m + \frac{1}{12}h^2 u_m^{(4)} + \mathcal{O}\left(h^4\right) - f_m = \frac{1}{12}h^2 u_m^{(4)} + \mathcal{O}\left(h^4\right).$$

Equation for the error. The relationship between the error $\mathbf{e}_h$ and the truncation error is easily obtained: recall that by definition

$$\tau_h = A_h \mathbf{u} - \mathbf{F},$$

rearranging and subtracting this from $A_h \mathbf{U} = \mathbf{F}$ we obtain the important relation

$$A\mathbf{e}_h = -\tau_h \tag{2.4}$$

which can be also written componentwise as

$$\frac{1}{h^2}\left(e_{m-1} - 2e_m + e_{m+1}\right) = -\tau_m, \quad m = 1, \dots, M.$$

Definition. A method for the boundary value problem (2.1) is said to be *consistent* with the boundary value problem with respect to the norm $\|\cdot\|$ if and only if

$$\|\tau_h\| \to 0, \quad \text{when}\, h \to 0.$$

Consistence in the vector norm $\|\cdot\|$ implies that the corresponding piecewise constant function tends to zero as h tends to zero in the corresponding function norm, this is because of the relationship between vector and function norms as we have seen for $\|\cdot\|_1$, $\|\cdot\|_\infty$ and $\|\cdot\|_2$.

Definition. A method for the boundary value problem (2.1) is said to be *convergent* in the (function) norm $\|\cdot\|$ if and only if

$$\|e_h\| \to 0, \quad \text{when}\, h \to 0.$$

Definition. *Stability.* Assume a difference method for the boundary value problem (2.1) is given by the discrete equation

$$A_h\mathbf{U} = \mathbf{F},$$

where h is the step-size of discretization. The method is *stable* in the norm $\|\cdot\|$ if there exist constants $C > 0$ and $H > 0$ such that

1. A_h^{-1} exists for all $h < H$;
2. $\|A_h^{-1}\| \leq C$ for all $h < H$.

The matrix norm in which we should prove stability is the one subordinate to the chosen function/vector norm in which we want to prove convergence.

Proposition. For the boundary value problem (2.1), stability and consistence with respect to the norm $\|\cdot\|$ imply convergence in the same norm.

Proof. We use (2.4) to obtain a bound for the norm of the error. Since we have stability A_h is invertible for all $h < H$ and therfore

$$\mathbf{e}_h = -A_h^{-1}\tau_h.$$

Then

$$\|e_h\|_q \leq h^{\frac{1}{q}}\|A_h^{-1}\|_q\|\tau_h\|_q = \|A_h^{-1}\|_q\|\tau_h\|_q \leq C\|\tau_h\|_q,$$

and we conclude that $\|e_h\|_q \to 0$ as $h \to 0$ because so does $\|\tau_h\|_q$.

For the case $\|\cdot\|_\infty$ the vector norm of $\mathbf{v} \in \mathbf{R}^M$ and the function norm of the corresponding piecewise constant function defined on the grid coincide, so to include this case we can conveniently adopt the notation $h^{\frac{1}{\infty}} := \lim_{q\to\infty} h^{\frac{1}{q}} = 1$.

2.1.1 2-norm stability for the case example

Observe that the eigenvalues of the matrix (2.3) are

$$\lambda_m = \frac{2}{h^2}\left(\cos(m\,\pi\,h) - 1\right), \quad m = 1, \ldots, M,$$

and the corresponding eigenvectors $\mathbf{v}^m$ have components

$$v_j^m = \sin(m\,\pi\,j\,h), \quad j = 1, \ldots, M.$$

Since by definition $\|A_h\|_2 = \sqrt{\rho(A_h^T A_h)}$ then because A_h is symmetric $\|A_h\|_2 = \rho(A_h)$. Denote with $\sigma(B)$ the spectrum of the $M \times M$ matrix B (the collection of all the eigenvalues of B), then $\|A_h\|_2 = \max_{\lambda\in\sigma(A_h)} |\lambda|$. Analogously

$$\|A_h^{-1}\|_2 = \rho(A_h^{-1}) = \max_{\lambda\in\sigma(A_h)} |\lambda^{-1}| = \frac{1}{\min_{\lambda\in\sigma(A_h)} |\lambda|}.$$

Using the series expansion $\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4!} + \mathcal{O}(x^6)$, with $x = m\,\pi\,h$ in the expression for the m-th eigenvalue we get

$$\lambda_m = \frac{2}{h^2}\left(-\frac{(m\,\pi\,h)^2}{2} + \frac{(m\,\pi\,h)^4}{4!} + \mathcal{O}(h^6)\right) = -m^2\pi^2 + \mathcal{O}(h^2),$$

and

$$\min_{\lambda\in\sigma(A_h)} |\lambda| = |\lambda_1| = -\pi^2 + \mathcal{O}(h^2),$$

such that

$$\|A_h^{-1}\|_2 = \frac{1}{|\lambda_1|} \to \frac{1}{\pi^2}, \quad \text{when } h \to 0,$$

and so there exist $C > 0$ and $H > 0$ such that $\|A_h^{-1}\|_2 = \frac{1}{|\lambda_1|} < C$ for all $h < H$, which proves stability of the proposed difference scheme in the 2-norm.

Exercise. Using the estimates for $\|A_h^{-1}\|_2$ and $\|\tau_h\|_2$ obtained in this section, prove that

$$\|e_h\| \leq \frac{1}{\pi^2}\,\frac{h^{2.5}}{12}\,\|f''\|_2 + \mathcal{O}(h^{3.5}),$$

assume f is twice differentiable.

2.1.2 Neumann boundary conditions

We consider now the boundary value problem

$$u_{xx} = f(x), \qquad 0 < x < 1, \quad u'(0) = \sigma, \quad u(1) = \beta, \tag{2.5}$$

and we propose three different discretizations of the left boundary condition which combined with the earlier consider discretization of the second derivative will lead to three different linear systems. In this case the matrices we obtain are no longer symmetric.

Case 1

We approximate $u'(0) = \sigma$ by

$$\frac{U_1 - U_0}{h} = \sigma,$$

and combine this discrete equation with (2.2), this leads to the $(M+1) \times (M+1)$ linear system

$$A_h \mathbf{U} = \mathbf{F},$$

where

$$A_h := \frac{1}{h^2} \begin{bmatrix} -h & h & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_0 \\ \vdots \\ U_M \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} \sigma \\ f_1 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

This leads to an overall method of order of consistency 1.

Case 2

To obtain order two we introduce a fictitious node external to the domain to the left of x_0 which we call x_{-1}. We approximate $u'(0) = \sigma$ by

$$\frac{U_1 - U_{-1}}{2h} = \sigma.$$

We use this equation together with (2.2) for $m = 0$, i.e.

$$\frac{1}{h^2}\left(U_{-1} - 2U_0 + U_1\right) = f_0,$$

and eliminate U_{-1}.

We are left with the equation

$$\frac{U_1 - U_0}{h} = \frac{f_0 h}{2} + \sigma$$

which we now combine with (2.2), this leads to a $(M+1) \times (M+1)$ linear system

$$A_h \mathbf{U} = \mathbf{F},$$

where A_h is as in the previous case and

$$\mathbf{F} := \begin{bmatrix} \frac{f_0 h}{2} + \sigma \\ f_1 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

Case 3

In this discretization we want to obtain order 2 without the use of fictitious nodes. We then consider the approximation of $u'(0) = \sigma$ by

$$-\frac{\frac{3}{2}U_0 - 2U_1 + \frac{1}{2}U_2}{h} = \sigma,$$

where the left hand side gives a discretization of order two of the first derivative of u at zero. Proceeding as in case 1, we obtain the $(M+1)\times(M+1)$ linear system

$$A_h\mathbf{U} = \mathbf{F},$$

where

$$A_h := \frac{1}{h^2}\begin{bmatrix} -\frac{3}{2}h & 2h & -\frac{h}{2} & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_M \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} \sigma \\ f_1 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

Exercise. Prove with the method of undetermined coefficients that

$$-\frac{\frac{3}{2}u(0) - 2u(h) + \frac{1}{2}u(2h)}{h} = u'(0) + \mathcal{O}(h^2).$$

Neumann boundary conditions on both sides

We consider now the boundary value problem

$$u_{xx} = f(x), \qquad 0 < x < 1, \quad u'(0) = \sigma_0, \quad u'(1) = \sigma_1, \tag{2.6}$$

and we propose a second order discretization of the boundary conditions using two fictitious nodes external to $[0,1]$, x_{-1} and x_{M+2}. The approximated boundary conditions become

$$\frac{U_1 - U_{-1}}{2h} = \sigma_0, \qquad \frac{U_{M+2} - U_M}{2h} = \sigma_1,$$

which we combine respectively with (2.2) for $m = 0$ and $m = M+1$ and obtain the linear system of equations $(M+2)\times(M+2)$ linear system

$$A_h\mathbf{U} = \mathbf{F},$$

where

$$A_h := \frac{1}{h^2}\begin{bmatrix} -h & h & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & 1 & -2 & 1 \\ & & & h & -h \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_0 \\ \vdots \\ U_{M+1} \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} \sigma_0 + \frac{h}{2}f_0 \\ f_1 \\ \vdots \\ f_M \\ f_{M+1}\frac{h}{2} - \sigma_1 \end{bmatrix}.$$

But now the matrix A_h is singular, in fact it is easy to verify that $\mathbf{v} := [1,\dots,1]^T \in \mathbf{R}^{M+2}$ is an eigenvector of A_h with eigenvalue 0. We want to determine the condition of existence of solutions for this linear system.

Before we do that we consider the solution of (2.6). Integrating on both sides between 0 and 1 we obtain

$$u'(1) - u'(0) = \int_0^1 f(x)\,dx,$$

that is

$$\sigma_1 - \sigma_0 = \int_0^1 f(x)\,dx, \tag{2.7}$$

this is a necessary condition for the existence of solutions of the boundary value problem (2.6).

We now turn to the analysis of the solutions of the discrete boundary value problem $A_h\mathbf{U} = \mathbf{F}$. Recall that

$$\text{Range}(A_h) := \text{span}\{A_h\mathbf{e}_1, \dots A_h\mathbf{e}_{M+2}\},$$

where we denote with $\mathbf{e}_j$ $j = 1, \dots, M+2$ the canonical basis of $\mathbf{R}^{M+2}$ and then $A_h\mathbf{e}_1, \dots A_h\mathbf{e}_{M+2}$ are the columns of A_h.

Recall that

$$\text{Ker}(A_h^T) = \{\mathbf{z} \in \mathbf{R}^{M+2} \,|\, A_h^T\mathbf{z} = 0\},$$

and $\text{Ker}(A_h^T)$ is a vector space.

Proposition. $A_h\mathbf{U} = \mathbf{F}$ admits a solution if and only if $\mathbf{F} \in \text{Range}(A_h)$

Proposition. $\mathbf{F} \in \text{Range}(A_h)$ if and only if $\mathbf{F} \perp \text{Ker}(A_h^T)$.

Proof The proof of these two propositions is by exercise.

Since $\mathbf{F} \perp \text{Ker}(A_h^T)$ if and only if

$$\forall \mathbf{z} \in \mathbf{R}^{M+2}, \text{ s.t. } A_h^T\mathbf{z} = 0, \ \mathbf{F}^T\mathbf{z} = 0,$$

we need to characterize $\mathbf{z} = [z_0, \dots, z_{M+1}]^T$ generic element of $\text{Ker}(A_h^T)$, this amounts at finding the solutions of the homogeneous linear system $A_h^T\mathbf{z} = 0$,

$$\frac{1}{h^2}\begin{bmatrix} -h & 1 & & & & \\ h & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & h \\ & & & & 1 & -h \end{bmatrix} \begin{bmatrix} z_0 \\ \vdots \\ \vdots \\ \vdots \\ z_{M+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

and we find that the solution is $\mathbf{z} = [z_0, hz_0, \dots, hz_0, z_0]^T \in \mathbf{R}^{M+2}$ for any value of $z_0 \in \mathbf{R}$. Then

$$\mathbf{F}^T\mathbf{z} = 0 \Leftrightarrow z_0(f_0\frac{h}{2} + \sigma_0) + hz_0\sum_{m=1}^{M} f_m + z_0(f_{M+1}\frac{h}{2} + \sigma_1) = 0,$$

for all z_0 and for $z_0 \neq 0$ we can divide out by z_0 to obtain the condition

$$\sigma_1 - \sigma_0 = f_0 \frac{h}{2} + h \sum_{m=1}^{M} f_m + f_{M+1} \frac{h}{2}.$$

We easily recognize that this condition is the discrete analogous of (2.7) and in fact the right hand side is the composite trapezoidal rule applied to the integral of $f(x)$ on the discretization grid.

2.2 Linear boundary value problems

Consider the general linear boundary value problem

$$a(x)u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad a \leq x \leq b, \quad u(a) = \alpha, \quad u(b) = \beta,$$

a discretization of this problem on the grid $x_m = a + m \cdot h$, $h = \frac{b-a}{M+1}$ is

$$a_m \frac{U_{m-1} - 2U_m + U_{m+1}}{h^2} + b_m \frac{U_{m+1} - U_{m-1}}{2h} + c_m U_m = f_m, \quad m = 1, \ldots, M$$

and $a_m := a(x_m)$, $b_m := b(x_m)$, $c_m := c(x_m)$ and $U_m \approx u_m := u(x_m)$. The corresponding linear system of equations has the form

$$\frac{1}{h^2} \begin{bmatrix} h^2 c_1 - 2a_1 & a_1 + h\frac{b_1}{2} & & & \\ a_2 - h\frac{b_2}{2} & h^2 c_2 - 2a_2 & a_2 + h\frac{b_2}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & & & a_{M-1} + h\frac{b_{M-1}}{2} \\ & & & a_m - h\frac{b_M}{2} & h^2 c_M - 2a_M \end{bmatrix} \begin{bmatrix} U_1 \\ \\ \vdots \\ \\ U_M \end{bmatrix} = \begin{bmatrix} f_1 - \alpha(\frac{a_1}{h^2} - \frac{b_1}{2h}) \\ f_2 \\ \vdots \\ f_{M-1} \\ f_M - \beta(\frac{a_M}{h^2} + \frac{b_M}{2h}) \end{bmatrix}$$

The linear system we obtain will in general be not symmetric.

2.2.1 A self-adjoint case

Suppose now we want to discretize

$$(\kappa(x)u'(x))' = f(x), \quad a \leq x \leq b, \quad u(a) = \alpha, \quad u(b) = \beta,$$

we could differentiate out the left hand side to obtain a problem of the class described above, i.e

$$\kappa(x)u''(x) + \kappa'(x)u'(x) = f(x)$$

this is however not the best strategy because, as we have seen, this leads in general to non symmetric linear systems. Consider now the differential operator $\mathcal{L}$ defined as follows

$$\mathcal{L}u(x) := (\kappa(x)u'(x))'$$

this operator is self-adjoint with respect to the $L_2([a,b])$ inner product,

$$\langle f, g \rangle := \int_a^b f(x)g(x)\,dx, \quad f, g \in L_2([a,b]).$$

This motivates the search for a symmetric discretization A_h of $\mathcal{L}$.

To achieve this we consider the mid-points of the grid $x_{m-\frac{1}{2}} = x_m - \frac{h}{2}$ and $x_{m+\frac{1}{2}} = x_m + \frac{h}{2}$, we then consider the following approximations

$$\kappa_{m-\frac{1}{2}} u'_{m-\frac{1}{2}} = \kappa_{m-\frac{1}{2}} \frac{u_m - u_{m-1}}{h} + \mathcal{O}(h^2)$$

and

$$\kappa_{m+\frac{1}{2}} u'_{m+\frac{1}{2}} = \kappa_{m+\frac{1}{2}} \frac{u_{m+1} - u_m}{h} + \mathcal{O}(h^2).$$

We use these two to produce

$$(\kappa\, u')'\big|_{x_m} = \frac{1}{h^2}\left[\kappa_{m-\frac{1}{2}} u_{m-1} - (\kappa_{m-\frac{1}{2}} + \kappa_{m+\frac{1}{2}}) u_m + \kappa_{m+\frac{1}{2}} u_{m+1}\right] + \mathcal{O}(h^2).$$

The methods is of second order and the linear system we have to solve is

$$\frac{1}{h^2}\begin{bmatrix} -(\kappa_{\frac{1}{2}} + \kappa_{\frac{3}{2}}) & \kappa_{\frac{3}{2}} & & & \\ \kappa_{\frac{3}{2}} & -(\kappa_{\frac{3}{2}} + \kappa_{\frac{5}{2}}) & \kappa_{\frac{5}{2}} & & \\ & \ddots & \ddots & \ddots & \\ & & & & \kappa_{M-\frac{1}{2}} \\ & & & \kappa_{M-\frac{1}{2}} & -(\kappa_{M-\frac{1}{2}} + \kappa_{M+\frac{1}{2}}) \end{bmatrix} \begin{bmatrix} U_1 \\ \vdots \\ \vdots \\ U_M \end{bmatrix} = \begin{bmatrix} f_1 - \alpha\frac{\kappa_{\frac{1}{2}}}{h^2} \\ f_2 \\ \vdots \\ f_{M-1} \\ f_M - \beta\frac{\kappa_{M+\frac{1}{2}}}{h^2} \end{bmatrix}.$$

2.3 A nonlinear example

We consider the pendulum equations

$$\theta''(t) + \sin(\theta(t)) = 0, \quad 0 < t < T, \quad \theta(0) = \alpha,\ \theta(T) = \beta, \tag{2.8}$$

with numerical discretization on the grid $t_m = m\,h$, $m = 0, \dots, M+1$ and $h = \frac{T}{M+1}$ leading to the discretized equations

$$\frac{1}{h^2}(\Theta_{m-1} - 2\Theta_m + \Theta_{m+1}) + \sin(\Theta_m) = 0, \quad m = 1, 2, \dots, M.$$

The system of equations is now nonlinear and we denote it in a vector form as

$$G_h(\Theta) = 0, \quad \Theta = \begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_M \end{bmatrix},$$

where $G_h(\Theta) = A_h\Theta + \sin(\Theta)$, $\sin(\Theta) := [\sin(\Theta_1), \dots, \sin(\Theta_M)]^T$ and A_h is the same as in (2.3).

The numerical solution of this system must be addressed iteratively, for example with a Newton method:

$$\Theta^{[k+1]} = \Theta^{[k]} - J_h(\Theta^{[k]})^{-1} G_h(\Theta^{[k]}),$$

where $J_h(\Theta)$ is the Jacobian of $G_h(\Theta)$, and it is easily seen that $J_h(\Theta)$ is a tridiagonal symmetric matrix such that

$$(J_h)_{i,j}(\Theta) = \begin{cases} \frac{1}{h^2} & \text{for} \quad j = i-1 \text{ or } j = i+1, \\ \frac{-2}{h^2} + \cos(\Theta_i) & \text{for} \quad j = i, \\ 0 & \text{otherwise,} \end{cases}$$

so

$$J_h(\Theta) = A_h + C(\Theta), \qquad C(\Theta) := \mathrm{diag}(\cos(\Theta_1), \ldots, \cos(\Theta_M)).$$

The truncation error is

$$\tau_h := G_h(\overrightarrow{\theta}), \qquad \overrightarrow{\theta} := \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}, \quad \theta_j := \theta(t_j).$$

It is easily shown by Taylor expansion that the components of τ_h satisfy

$$\tau_j = \frac{1}{12} h^2 \theta^{(4)}(t_j) + \mathcal{O}(h^4), \quad j = 1, \ldots, M,$$

this ensures that the method is consistent of order 2. As usual we want to use the connection between error $E_h := \Theta - \overrightarrow{\theta}$ and truncation error τ_h in order to prove convergence. In general this connection is a bit less manageable in the case of nonlinear problems, but in this particular example it is not too complicated.

We combine the discrete equation $G_h(\Theta) = 0$ and the equation for the truncation error $G_h(\overrightarrow{\theta}) = \tau_h$ to obtain

$$G_h(\Theta) - G_h(\overrightarrow{\theta}) = -\tau_h. \tag{2.9}$$

From (2.9), using $G_h(\Theta) = A_h\Theta + \sin(\Theta)$ we get

$$A_h E_h + \sin(\Theta) - \sin(\overrightarrow{\theta}) = -\tau_h,$$

and by Taylor expansion

$$\sin(\Theta) = \sin(\overrightarrow{\theta}) + C(\hat{\theta})E_h,$$

where $C(\hat{\theta})$ is the diagonal matrix earlier defined, and the components $\hat{\theta}_i$ of the vector $\hat{\theta}$ belong to the open intervals $(\Theta_i, \theta(t_i))$.

Due to the stability already proven for the linear case, A_h is invertible for all $h < H$ and so

$$E_h + A_h^{-1} C(\hat{\theta}) E_h = -A_h^{-1}\tau_h.$$

We here use the same notation for the vectors E_h and τ_h and the corresponding piecewise constant functions defined on the discretization grid, the norms we consider in the sequel are function norms. Assuming we operate in the 2-norm, we get

$$\|E_h\|_2 \leq \|A_h^{-1}\|_2 \left[\|C(\hat{\theta})E_h\|_2 + \|\tau_h\|_2\right] \leq \|A_h^{-1}\|_2 \left[\max_{1\leq m\leq M} |\cos(\hat{\theta}_m)| \, \|E_h\|_2 + \|\tau_h\|_2\right]$$

so

$$(1 - \|A_h^{-1}\|_2) \, \|E_h\|_2 \leq \|A_h^{-1}\|_2 \|\tau_h\|_2.$$

We know from earlier analysis that $\|A_h^{-1}\|_2 = \frac{1}{|\lambda_1|}$ where λ_1 is the eigenvalue of A_h with minimum absolute value. We also obtained the estimate $|\lambda_1| = \pi^2 + \mathcal{O}(h^2)$ and so we can also deduce that $\|A_h^{-1}\|_2 = \frac{1}{\pi^2} + \mathcal{O}(h^2)$ and for h small enough $\|A_h^{-1}\|_2 < 1$ and we get

$$\|E_h\|_2 \leq \frac{\|A_h^{-1}\|_2}{1 - \|A_h^{-1}\|_2} \|\tau_h\|_2.$$

Using again the estimate for $\|A_h^{-1}\|_2$ we can obtain

$$\frac{\|A_h^{-1}\|_2}{1 - \|A_h^{-1}\|_2} = \frac{1}{\pi^2 - 1} + \mathcal{O}(h^2).$$

$$\|E_h\|_2 \leq \frac{1}{\pi^2 - 1}\|\tau_h\|_2 + \mathcal{O}(h^2)$$

and recalling that $\|\tau_h\|_2 = \frac{1}{12}h^2\|\theta^{(4)}\|_2 + \mathcal{O}(h^4)$ we get finally

$$\|E_h\|_2 \leq \frac{1}{\pi^2 - 1}\left(\frac{1}{12}h^2\|\theta^{(4)}\|_2\right) + \mathcal{O}(h^4)$$

which guarantees convergence when h goes to zero, as it is easy to see that $\theta^{(4)}$ is bounded by differentiating the equation twice.